# Expanding our Participatory Democracy Toolkit using Algorithms, Social Choice, and Social Science

**Bailey Flanigan**

CMU-CS-24-130

May 2024

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

**Thesis Committee:**
Ariel Procaccia (Harvard), Chair
Nihar Shah
Anupam Gupta (New York University)
Nika Haghtalab (UC Berkeley)
Ashish Goel (Stanford)

*Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.*



This thesis was supported by an NSF GRFP Fellowship, a Siebel Scholarship, and a Fannie and John Hertz Foundation Fellowship.

**Keywords:** Social Choice, Algorithms, Voting, Deliberative Democracy, Sortition

*For the storytellers: those who generously share their experiences in order to educate, inspire, and advocate for the interests of others.*

# Abstract

In most of the world's democracies, policy decisions are primarily made by elected political officials. However, under mounting dissatisfaction with representative government due to issues ranging from social inequality to public distrust, a new proposal is taking off: to augment representative democracy with mechanisms by which the public can *directly participate in policymaking*.

The guiding application of this thesis will be one particular model of participation, *deliberative minipublics* (DMs), though we will argue that our contributions may apply to many models of direct participation. In a DM, a panel of citizens is selected by lottery from the population; then, this panel convenes around a particular policy issue to study background information, deliberate amongst themselves, and then weigh in on the issue. DMs have been gaining momentum over the past decade, and they are now being used at national and supranational levels, and are even being integrated into representative governments.

Motivated by this application domain, we make the following main contributions: In **Part I**, we design algorithms for performing the random selection of DM participants, a process known as *sortition*. Our sortition algorithms permit users to make optimal trade-offs between descriptive representation and other desirable properties conferred by randomness, and we characterize these tradeoffs using game theory, optimization, and empirics. In **Part II**, we use a novel social choice theory framework to investigate a notion of representation that departs from descriptive representation in a key way: it accounts for the political reality that people may be affected to *widely varying degrees* by any given policy decision. In **Part III**, we study an important hypothesized impact of deliberation: increasing the extent to which participants consider how *others in their society* may be affected by different policies. In Part **IV**, we highlight how the enclosed research illustrates new ways to combine tools from political science and computer science.

# Acknowledgments

The person who I foremost want to thank is Ariel Procaccia, who has been truly been the best advisor and collaborator I could have asked for. During PhD application season, working with Ariel was my ultimate goal. When I got to CMU, I found that the reasons for my excitement held true: Ariel is indeed a master of asking the right question, defining the right model, and writing papers that cut to the heart of the issue. What I have appreciated by even more over the past five years, though, are our discussions: we have debated research directions, motivating arguments, political scientific theories, potential faculty jobs, paper titles, and even books. I feel so lucky to have spent five years being challenged in my thinking by someone so astute; to debate and usually lose, but always be welcome to come back and try again (usually after writing 40 pages of Overleaf notes). Thank you Ariel for always believing in me, pushing me to sharpen my perspective, and making everyone else look bad by responding to my emails in an average of 3 minutes.

This thesis has also been made possible in large part by the other incredible teachers and mentors I have had along the way. Foremost among them is Paul Gölz, who during the first few years of my PhD spent countless hours after each research meeting, patiently answering all of my questions about what on earth we were talking about. I am continuously grateful for your willingness to let me pick your brain, and you remain one of the technical experts I trust the most. In a similar vein, I want to thank Anupam Gupta as another key teacher in my path, having taught me much of the algorithmic thinking that I still often use today. I want to thank Mor Harchol-Balter for being a wonderful friend, role model, and unconditional supporter for me throughout the many ups and downs of my PhD; Zico Kolter, for taking on CS-JEDI and always being on my team; Vasilis Gkatzelis, for introducing me to and supporting my initial interest in theoretical computer science; Michael Neblo and Kevin Esterling, for supporting my work in political science from the beginning; Steven Wright, for believing in my potential long before I did (and always being up for an evening of pep talks and way too many appetizers); John Puccinelli, for having endless time to support me through my undergraduate degree; and Matt Bowman, for being one of my most enduring teaching role models. Finally, I want to thank my committee members, Anupam, Nika, Ariel, Nihar, and Ashish, for the multitude of ways in which they have each supported me throughout my PhD.

I also want to acknowledge the wonderful community by which I was surrounded throughout my PhD. Chief among them are my extremely talented collaborators, who not only made this research possible, but also made it incredibly fun. I look forward to many more years of enjoying the friendships we forged through fixing last-minute bugs in proofs, pulling together last-minute EC submissions, and going for beers to celebrate wins. I am also forever grateful to the Hertz Foundation for introducing me to Alex Atanasov, Dolev Bluvstein, Robbee Kosak, Philip Welkhoff, Ben Eysenbach, and Katherine Van Kirk, who I could always count during graduate school for equal parts support and fun. In the past year on the job market, I have been so grateful to the friends from our community who have shown up for me, helped prepared me, and commiserated with me, including Lily Xu, Kira Goldner, Jess Sorrell, Kate Donahue, Serina Chang, Serena Wang, Sean Sinclair, Alex Jacquillat, Jessie Finocchiaro, Elisabeth Paulson, Thodoris Lykouris, Ellen Vitercik, Victoria Dean, and many others. Finally and most importantly, I have to thank my core CMU PhD squad, Sara McAllister, Catalina Vajiac, and Ananya

Joshi: you have been my closest friends, strongest supporters, and partners in crime throughout my time at CMU.

Outside of the academic community, I am so fortunate to be supported by an endlessly supportive network of lifelong friends. In the past years, I have been drawn home to the midwest by the promise of morning coffee walks with Susan Miller, and I always look forward an evening of guacamole and beers with our family friends Lisa, Rodney, Elizabeth, and Kurt. Other key people in my support system have been my DC contingent of Jacob, Polly, Aaron, and Cristina; my Philadelphia contingent of Maryann, Matt, Zach, and Freddi; and my lifelong best friend, Rose Brown. Finally, to Grace, Anne, Donita, and Charlie: thank you for being my second family :).

Perhaps the most important thank you is to my family—my parents, Rob and Mona, and my brother Will. Thank you for supporting my loves of learning, logic puzzles, and books long before it was clear that they would turn into anything useful. You have always been there to support my education, listen to me complain, celebrate the wins, and give me great advice. I am so grateful to be able to count my family among my best friends!

Finally, I want to thank my partner, Chara, who also belongs among my collaborators, job market supporters, and best friends. In addition to supporting me through the ups and downs of the work in this thesis, thank you for being my biggest fan, best buddy, greek food chef, running partner, travel companion, and so much more. Every day with you and Terra is an adventure;
είμαι πολύ ενθουσιασμένη που ζω τη ζωή μαζί σας!

# Contents

# Listing of figures

# List of Tables

## 0.1 Introduction

Democracy as a form of governance—despite its long-debated and evolving imperfections—remains prized for its foundational principle: *citizens should have the right to participate actively in governing their society.* In democratic countries around the world, this philosophy tends to be implemented via *representative* government, in which citizens elect representatives to make decisions on their behalf and, in theory, in their interests. In recent decades, many have begun sounding the alarm about a "crisis in democracy", spurred by mounting evidence that many feel decreasingly represented by, trusting in, and able to influence their governments [74, 168, 217, 282, 283]. Others might argue that the more enduring crisis is that, over the past centuries, many of the world's democratic powers have consistently under-served the interests of entire subsets of their constituents, including indigenous populations, racial and ethnic minorities, people with disabilities, and the unhoused (e.g., [20, 171, 202, 228, 236, 262]). Either perspective leads to the same basic conclusion: there is a need to create more effective and inclusive access to governing power.

In this thesis, we focus on a proposed solution that is now gaining widespread momentum: augmenting representative democracy with processes that permit *direct citizen participation* in policymaking.[1] Our aim will be to contribute tools that support the principled design, implementation, and proliferation of these processes.

To start, in Section 0.1.1 we will design from the ground up a *hypothetical* process for facilitating direct participation in policymaking. Our goal here is not to build a perfect process, or even survey all possible design choices. Rather, we aim to illustrate some of the major challenges associated with involving citizens directly in policymaking, and to motivate the tools and ideas we will study in response.

### 0.1.1 Facilitating Direct Participation, Hypothetically

We start from what might seem to many like the biggest hurdle: *Do everyday people have the expertise to make high-quality policy recommendations?* If we imagine an average member of the public being asked to recommend a policy under everyday conditions, a reasonable answer might be *No*: there is significant data supporting that people are susceptible to misinformation and propaganda [271], polarized [265], and politically disengaged [73, 108].

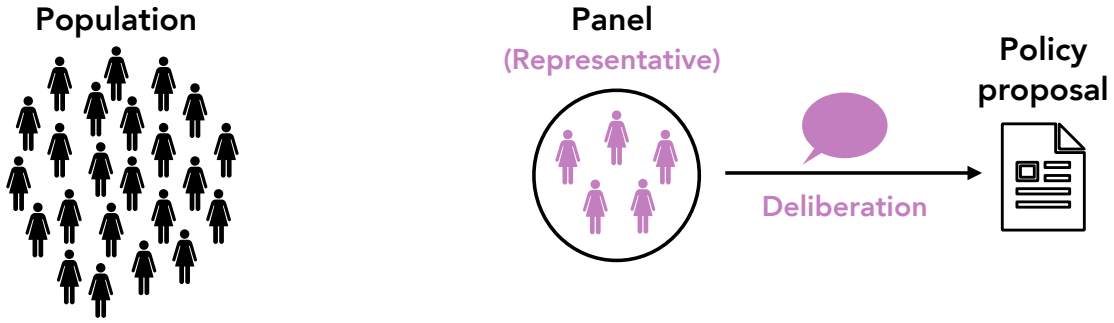
Fortunately, we are not bound to everyday conditions. This the proposal of *democratic deliberation*: to have people reason about politics through a structured discussion that is grounded in evidence and reasoning.[2] A growing community experts have high hopes for deliberation, seeing it as a way to facilitate high-quality political reasoning among everyday people even under conditions of polarization, misinformation, and political distrust [3, 100, 123, 222]. We add this tool to our hypothetical process so that prior to weighing in on what should be done, the citizens weighing in will engage in informed deliberation.

[1] We will use "citizen" to refer to *any* constituent of a democracy, implying nothing about legal citizenship status.

[2] Underlying this simple definition, there is significant research dedicated to the precise definition of *deliberation* in a democratic context (e.g., [200]).


For all its potential benefits, democratic deliberation has the downside that it is time- and resource-intensive: it often takes place over several days, in-person, and participants are necessarily compensated and reimbursed. As a result, not everyone in the population can participate. In practice, the solution is usually to choose a smaller group of citizens, which we will call a *Panel*, that ultimately participates in the discussion.

As soon as we need to select a smaller panel from within the population, we encounter a second major challenge: *whose should be given a seat at the table?* This challenge engages the concept of *representation*, on which there is a rich scholarship investigating who can, and who should, represent who? (e.g., [199, 229]). In many real-world decision-making contexts, it is popular to aim for *descriptive* representation of the population — that is, proportional representation of population subgroups — at least with respect to a predefined set of identities. We will adopt this goal for now, though we will revisit it later.




Finally, regardless of the notion of representation we want to ensure, we must decide: *how should we select our representative panel*? Here, we take inspiration from a centuries-old example of direct citizen participation: in ancient Athenian democracy, political representatives were selected directly from the population by lottery—a concept known as *sortition* [272]. Randomly selection may seem unprincipled compared to, e.g., choosing citizens who are especially qualified by some criteria. However, it is precisely this *absence of reasons* for which many advocate sortition, arguing that a uniform lottery gives everyone a fair chance to participate; mitigates perverse and filtration mechanisms produced by elections; and as an added bonus, produces descriptive representation [71].

Unfortunately, adding sortition to our process will be a bit less straightforward than running

a uniform lottery: when participation is voluntary (as it typically is in models of direct democratic participation), those who are willing to partake tend to be highly skewed demographically and ideologically compared to the underlying population. Under this circumstance, known as *selection bias*, a uniform lottery will faithfully replicate this skew, failing to ensure the descriptive representation we desire. We therefore must implement a random selection procedure that strikes a desirable tradeoff between the representation we want, and the randomness that originally conferred sortition's defining benefits. We punt, for now, on how to do this, as it will be a central focus of this thesis.

Figure 1 depicts our resulting hypothetical process for facilitating direct citizen participation in governance.

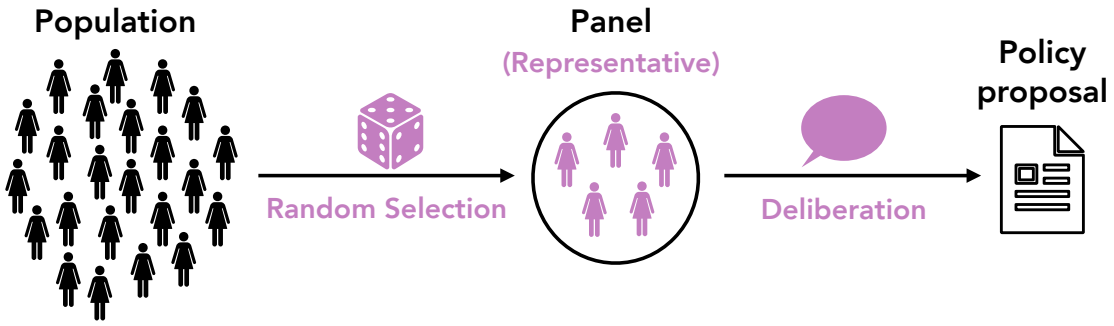



Figure 1: Our hypothetical process for facilitating direct citizen participation in governance.

### 0.1.2 From a Hypothetical Process to Real Participation Models

Although this process is hypothetical, in developing it we had to posit solutions for challenges that are somewhat fundamental to the task of facilitating direct participation in policymaking: to ensure participants can make well-reasoned decisions, processes often need to be somewhat intensive. When processes are intensive and the underlying population is large, it is often necessary to choose only a subset of the public to participate. When we must choose a subset of the public, we must make judgements about whose representation should be ensured, and how to choose participants in a way that best serves the goals of direct participation.

Given the fundamental nature of these challenges, it is perhaps unsurprising that the process in Figure 1 has many hallmarks in common with real participatory models being adopted around the world:

– In *participatory budgeting*, a subset of citizens are appointed as *budget delegates*, and they convene to review evidence and collectively decide on how to divide a public budget over candidate public-interest projects. Although budget delegates are not typically selected randomly, many participatory budgeting resources reference the importance of representation among participants, especially from communities that are marginalized in standard politics [223, 244, 255]. In many cases, delegates engage in deliberation or other similar modes of learning about and

collaboratively weighing project alternatives [89, 192]. One guide even proposes *deliberative participatory budgeting*, suggesting the use of sortition to select a representative deliberative body [36]. Since 2013, participatory budgeting has been used to allocate billions of dollars (or other units of currency) [16]. In 2019 alone, there were upwards of 11,000 participatory budgeting events worldwide spanning 71 countries [89].

– In *deliberative town halls*, citizens have moderated many-on-one discussions with their elected officials [211]. Deliberative town halls are built around goal of facilitating deliberative interaction between constituents and officials [212]. While they are sometimes open to all members of the community, in high-stakes applications with large populations, participants have been randomly selected with measures to ensure descriptive representation [178]. Deliberative town halls, often run through OSU's *Connecting to Congress* initiative [13], were recently used as part of Chile's national effort to amend their constitution [8, 9], and this participation model is now being scaled up through the online platform *Prytaneum* [15].

– In *independent redistricting commissions*, a group of voters is convened to draw the boundaries of voting districts in a way that is hopefully more impartial than those produced through partisan gerrymandering. It is often paramount that the members of these commissions are representative on dimensions like political leaning; in some cases, the selection of participants involves randomizing; and many times, the process of collaboratively drawing new maps can involve substantial discussion [66, 245]. Independent redistricting commissions have been increasing in uptake over the past few years, having been used recently to draw congressional districts in Michigan, New York, Virginia, and Colorado [83].

– Finally, the participation model that perhaps most closely resembles Figure 1 is the *deliberative minipublic* (DM). A DM proceeds much like our hypothetical model: a representative sample of the public is chosen by lottery to serve on a panel. Then, this panel convenes around a policy issue for several days, learning from experts, deliberating, and then finally weighing in on what should be done. DMs are actually an entire category of democratic paradigms, encompassing *citizens' assemblies*, *citizens' panels*, *citizens' juries*, *deliberative polls*, and more.[1] DMs constitute one of the most rapidly-growing models of citizen participation globally, with hundreds having been run around the world in just the past decade [4, 225]. DMs have been used at the national level in many countries including Mongolia [1], South Korea [7], Ireland [169], France [2, 65], and Germany [5]; they have even been used at the supranational scale (e.g., in the COP 26 Global Climate Assembly [11]). In the past few years, several instances have begun charting a path toward formal integration of deliberative minipublics into representative government. Citizens' assemblies are being integrated as permanent arms of governing bodies in major regions and cities, including Ostbelgian [215], Madrid [29], and Brussels [6], and there is now a law in Mongolia requiring deliberative polls before making certain kinds of constitutional amendments [1].

[1]The main distinction deliberative polls and many other DMs is how opinions are elicited post-deliberation: deliberative polls end with an anonymous poll of participants [125], while many other DMs end with participants collaboratively forming a policy proposal.

## 0.2 Thesis Overview

In this thesis, we will consider each of the three key components depicted in Figure 1: in Part I, we study procedures for randomly sampling a representative panel when there is selection bias; in Part II, we consider the implications of descriptive representation and explore an alternative; and in Part III, we examine potential impacts of deliberation on participants' political reasoning. Our work on these tools will be guided primarily by the application of deliberative minipublics, and sometimes citizens' assemblies in particular. However, given these tools' relevance to even just the participation models that already exist—and the fact that new ones are emerging all the time [12, 14, 91, 289]—we hope the ideas in this thesis can be applied to many other participatory processes as the landscape of citizen participation evolves.

Across Parts I - III, this thesis contains eight completed original papers [32, 43, 128, 130, 131, 134, 135], each in their own chapter.[1] Each of these Parts is laid out as follows:

- A **background chapter** outlines the problems we will solve, their context and motivation, modeling notions that apply across chapters, and an overview of the chapters themselves.
- Several **research chapters** each enclose the body of a single original paper.
- An **ongoing and future work chapter** discusses limitations of the existing work and some of the follow-up questions needed to address them.

Part IV is the Discussion, and Part V contains the appendices of all enclosed papers.

## 0.3 A Final Comment on Motivation

The primary motivation for the enclosed research, as discussed above, is advancing and supporting deliberative minipublics and other emerging models of direct citizen engagement. However, it is important to acknowledge that the political potential of these processes is yet unclear: multiple of these paradigms have so far shown mixed impacts [167, 220, 244, 267], and in many cases there remains the question of how to balance giving adequate authority to citizens' input while also maintaining robustness against bad actors [268, 284].

I tend to interpret these concerns as an indication that direct participation is a necessarily complex solution to a complex problem, and while it may have great potential to work, a lot of research is required to get there. The uncertainty of the present, however, requires us to address the question: *what does this research contribute, in the event that direct citizen participation does not ultimately find a path to integration into representative government?*

[1] It omits four additional published papers [88, 132, 133, 279], two of which relate to topics in this thesis but were deemed insufficiently relevant to democratic participation to be included. The first of these studies voting axiom satisfaction under a smoothed model of preferences, delineating classes of axioms and voting rules by whether semi-random noise is enough to escape axiomatic impossibilities [132]. The second covers a student-designed discussion-based course on diversity and inclusion in computer science [133], the discussion portion of which was influenced by—and has influenced—my thinking around democratic deliberation.

Fortunately, the ultimate implementation of these democratic innovations is not required for their study to be worthwhile. As we will illustrate throughout this thesis, the process of trying to understand these innovations offers new angles from which to study concepts that are fundamental to how citizens engage with democracy in general. Moreover, direct participation models offer a *uniquely* good settings in which to study these concepts: first, they tend to bring people together in a location for long periods, allowing more in-depth inquiry of people's political knowledge, opinions, and patterns of political reasoning. Second, these processes are not yet entrenched in institutions, and the resulting fluidity of their design makes them fertile ground for experimentation. Several ideas in this thesis will make use of – or propose new ways for others to make use of – these features for future research.

# Part I

# *Sortition*: Representation by Random Representatives

# 1 Background

### 1.0.1 Why Sortition (In Theory)?

There is a large body of political science scholarship arguing for the use of sortition over other methods for selecting political representatives. Here, we overview three of the arguments considered must centrally in this literature. We then distill these arguments into four technical ideals, numbered **(i)-(iv)**, which we will pursue algorithmically throughout Part I. The arguments we discuss here were originally laid out in 1989 by Fredrik Engelsted [112] and have since been expanded upon by several scholars [98, 124, 259**?** ]. Importantly, this body of literature conceives of sortition as a *uniform lottery* over the population, so for the purpose of interpreting the following arguments, we will adopt this conception for now.

The first argument in favor of sortition, articulated here by Carson and Martin, is that "those chosen [by sortition] are far more likely to be a typical cross section of the population, with the same sort of distribution according to sex, age, ethnicity, income, occupation, and so forth" [71]. Here, Carson and Martin refer to the fact that a uniform lottery will in expectation (and *ex post*, with high probability) choose a panel that is proportionally representative of all population cross-sections. Fishkin makes a similar argument, specifically with regards to how sortition can support the legitimacy of deliberative democracy: "We can only know [what the people would think] if we start the deliberations with a good microcosm, as representative as possible in both demographics and attitudes." [124] We distill these points into the following ideal:

**(i)** ***Descriptive Representation:*** *The panel should be (at least nearly) proportionally representative of all population cross-sections.*

A second popular argument for sortition has to do with its equal treatment of potential partici-

pants. Carson and Martin talk about the importance of *equality of opportunity*: "[Sortition] gives everyone an equal chance of being chosen, whereas in elections, factors such as funding, appearance, speaking ability, threats, and promises play a big role" [**?** ]. Peter Stone offers a *allocative justice* as a normative argument in favor of this type of equality: "Allocative justice...is to treat public office as a type of good to which citizens might have various claims...Random selection is...appropriate...when all citizens have equal claims to that office" [259]. We summarize these these arguments as the following ideal:

**(ii)** ***Fairness:*** *All people should have an equal opportunity to participate.*

A third benefit of sortition is its ability to protected against subversion. As put by Oliver Dowlen, "[Sortition's] primary political potential is its ability to protect the public process of selection from subversion by those who might...use it for their own private or partisan ends" [98]. To distill technically well-defined ideals from this argument, we first ask: *who* might want to subvert the sortition process to their own ends? Peter Stone identifies two such parties: the organizers who select the panel, and the potential panel participants themselves [259].

According to Stone, sortition avoids potential subversion by the former group because "If [the agent who must select officials] selects randomly, then she must act on the basis of no reasons, and therefore cannot be influenced by corrupting or dominating interests even if she would like to be" [259]. However, we note a caveat to this argument: it relies on the fact that the public *can confirm* the selection was random. Otherwise, what is to stop the organizers from hand-picking a panel behind the scenes, and then *claiming* the selection was random? This can be avoided if there is *transparency:*

**(iii)** ***Transparency:*** *Observers of the panel selection process should be able to confirm their probability of selection using only simple intuition about probability.*

Fortunately, transparency is not hard to achieve when participants are selected by uniform lottery: one can just run a public lottery using simple physical randomness (e.g., drawing balls from bins).

The second entity who may want to subvert the sortition process are the *participants themselves*: they might engage dishonestly in the selection process in order to stack the panel with people supporting their interests. For reasons that will soon become clear, we will be concerned with one particular method of dishonesty: misrepresenting one's attributes during the selection process (and/or convincing others to do so). Of course, because a uniform lottery treats everyone equally and independently, it ensures that this issue avoided. To capture this potential issue, we define the ideal of *manipulation-robustness*:

**(iv)** ***Manipulation-Robustness:*** Potential participants should not be able to affect their own or others' probabilities of selection by misrepresenting their identities in the selection process.

### 1.0.2 Sortition in Practice

Today, one of the most prominent (and growing!) use cases of sortition is to choose participants of deliberative minipublics. The practical reality of this use case, however, differs in a crucial way from the ideal. In idealized sortition, it is assumed that any member of the population selected by uniform lottery will participate. In contrast, organizers of modern deliberative minipublics *cannot* compel participants to participate, and must rely on people to opt in. The typical rate of opting in among the general population is around 2-5% [128], and as we will illustrate with data shortly, those who agree to participate are usually highly demographically skewed. A simple uniform lottery would replicate this skew, producing a panel that is far from descriptively representative. Seeing this as an important issue for public and normative legitimacy, practitioners—at least of citizens' assemblies, which will be the primary application of Part I—perform random selection via the following two-stage process (depicted in Figure 1.1).

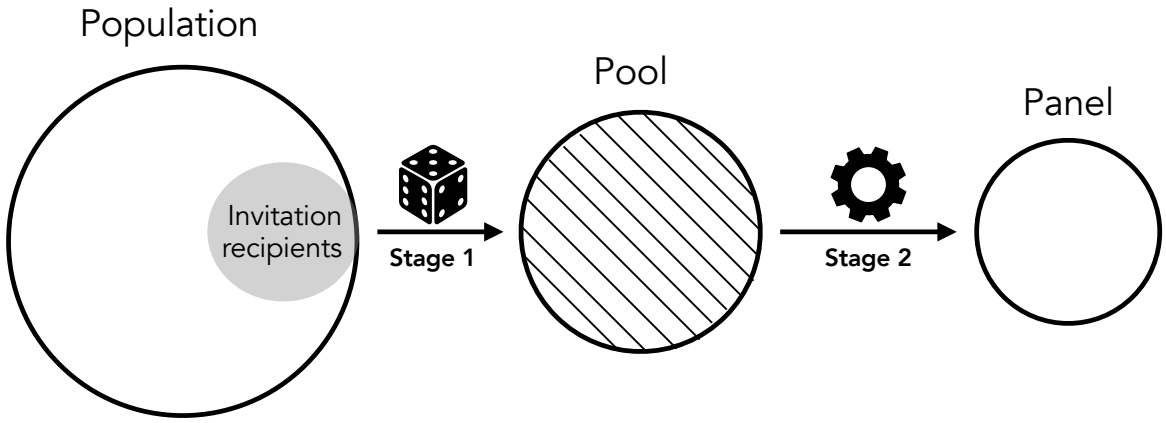



Figure 1.1: The two-stage panel selection process commonly used to select citizens' assemblies in practice. The dashed lines through the pool represent the fact that, the while the panel is designed to resemble the population, the pool may be very skewed demographically and ideologically due to selection bias.

Stage 1 (Uniform Lottery Invitations). Practitioners invite participants uniformly randomly from the underlying population (usually via either letters or phone calls). Those who receive an invitation and respond affirmatively form the *Pool* of volunteers. Upon volunteering (i.e., joining the pool), all pool members must fill out a survey about their demographic and ideological attributes, which will be used in Stage 2 to ensure that the panel is representative.

Stage 2 (Panel Selection). The *Panel* is selected from within the pool. In practice, this panel must satisfy two main requirements deterministically:

- ***Panel size.*** The panel must contain exactly $k$ members, where $k$ is chosen by the panel organizers. This requirement arises from budgetary constraints, as practitioners must cover some per-participant cost.
- ***Representative quotas.*** The panel must satisfy upper and lower *quotas* on a set of pre-

defined attributes, again chosen by practitioners. More precisely, these quotas are structured in the following way:

Let $F$ be a set of attribute categories, which we will call *features*; for example, in a UK-wide assembly on climate change in 2020 (our running example for the remainder of this section), $F$ included the features *education level, gender, age, climate concern level, race/ethnicity, geography 1,* and *geography 2* [232]. For each feature $f \in F$, practitioners define a set of mutually exclusive and exhaustive values $V_f$, which we call *feature-values*. For example, in the UK climate assembly, $V_{\textit{climate concern level}}$ contained the values *Not concerned, not very concerned, fairly concerned, very concerned*.

Upper and lower quotas are imposed at the feature-value level, denoted respectively as $\ell_{f,v}, u_{f,v}$ for feature-value $f, v$. Take the example of the feature-value *gender, female*: lower and upper quotas of $\ell_{gender,female} = 8$ and $u_{gender,female} = 12$ would mean that the panel would be required to contain between 8 and 12 women. Typically, upper and lower quotas are imposed on all feature-values in $\bigcup_{f \in F} V_f$, and enforce that each feature-value-defined group receives a number of panel seats near-proportional to their share of the population. For example, if women comprise 49% of the population, then quotas on a panel of size $k = 100$ might require between $\ell_{f,v} = 48$ and $u_{f,v} = 50$ women.[1]

Let $N$ denote the pool, let $f(i) \in V_f$ denote a pool member $i$'s value for feature $f$, and let $FV := \bigcup_{f \in F} V_f$ denote the set of feature-values on which quotas are imposed. $\mathbb{I}(\cdot)$ will be the indicator function. Then, an *instance* of the panel problem is defined by four quantities, $N, (\ell_{f,v} | f, v \in FV), (u_{f,v} | f, v \in FV), k$. In any given instance, the set of *valid panels* is

$$\left\{ K : K \subseteq N \ \wedge \ |K| = k \ \wedge \ \sum_{i \in K} \mathbb{I}\left(f(i) = v\right) \in [\ell_{f,v}, u_{f,v}] \text{ for all } f, v \in FV \right\}.$$

An instance of the panel selection problem is solved by a *selection algorithm*, which is any procedure (mapping) that intakes an instance of the panel selection task and outputs a valid panel, provided at least one valid panel exists.

### 1.1 Our Task: Sortition *Subject to Quotas*, *Under Selection Bias*

Notably absent from this discussion so far has been any discussion of *randomness*, the hallmark of sortition. As the previous section suggests, in our version of sortition, we must randomize *within the quotas*. This restriction already precludes one aspect of idealized sortition: we can no longer *independently* sample the panel members. The question is then whether we can still retain the

[1] Occasionally, practitioners impose quotas on combinations of attributes as well, though of course the extent to which this is possible is limited by the combinatorial explosion of attribute combinations relative to the small panel. To implement quotas on arbitrary combinations of attributes within this model, one would just make each combination they care about a feature-value. For example, if you wanted to enforce representation on all intersections of age and height, you would define the feature $f = \textit{age} \times \textit{height}$ with values $V_{\textit{age} \times \textit{height}} = \{$ young & short, young & tall, old & short, old & tall $\}$.

other, perhaps more defining property of idealized sortition: giving people *equal probability of selection*. Unfortunately, in the practical case, the answer is *No*. This is because there is selection bias in who opts into the pool from the population, meaning that the pool is far from representative of the population on the dimensions on which we impose quotas. To illustrate this point, Figure 1.2 shows the compositions of the *Population, Pool* and *Panel* in the UK Climate Assembly for two features: $f$ = *education level* whose values we re $V_{\textit{education level}} = \{0/1, 2/3, 4+\}$ where higher levels correspond to more education, and $f$ = *climate concern level*, whose values were $V_{\textit{climate concern level}} = \{\textit{Not concerned, not very concerned, fairly concerned, very concerned}\}$. In the figure below, we group the two lowest levels of climate concern.

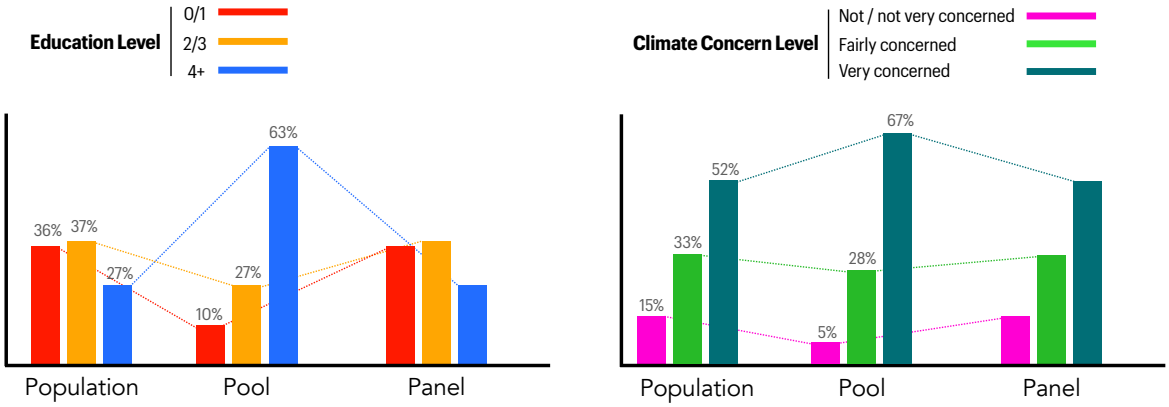



Figure 1.2: Selection bias in the UK Climate Assembly across values of two features, *education level* and *climate concern level*. Percentages are omitted above *Panel* bars because by design, they are essentially the same as those for the *Population*.

To see the selection bias in Figure 1.2, we must compare the composition of the population versus the pool. In making this comparison, note that because invitations to participate were sent out to members of the population uniformly at random, the pool is effectively a uniform random sample of *people who would opt if invited*. Examining the features of *education* and *climate concern level* as in Figure 1.2, we see that less educated groups and those who are less concerned about climate change are dramatically underrepresented among those who opted into the pool.

To understand the implications of this selection bias for the panel selection problem, we now compare the pool versus the panel. Comparing pool members with education level 0/1 versus those with level 4+, notice that there are *way more panel seats per person* reserved for the former group: there are about 1/12 as many pool members in the former group, and they are entitled to about 1.5x as many panel seats. They key consequence is that *as a result, those with education level 0/1 must be selected for the panel with higher probability, on average, than those with education level 4+*. This example illustrates a very general impossibility that grounds our subsequent work:

**Key Impossibility:** When we must satisfy representative quotas under the condition of selection bias, we cannot select all pool members with equal probability.

Things now seem a bit bleak: in practical sortition, we cannot give people equal probabilities of

selection, and thus cannot achieve the defining property of a uniform lottery. Our last hope is that despite these impossibilities, we can still achieve ideals (i)-(iv) — i.e., the reasons for using sortition at all — *to at least a reasonable degree*. This will be the goal of Part I, whose chapters we now overview. For most of this part, we will consider the first ideal, *(i) Descriptive Representation*, to be automatically satisfied by the requirement of quotas. We examine the extent to which this an oversimplification in our ongoing and future work (Chapter 7).

**Remark 1.1.1 (The Case of Mandatory Participation).** One might argue that these algorithms—designed around the challenge of selection bias—would become obsolete if participation in direct democratic processes became required. The assumption implicit in this argument is that requiring participation would eliminate selection bias and restore the viability of a uniform lottery; we now examine this assumption via the case study of U.S. jury selection, in which participation is legally required upon being summoned—*except if the person meets one of the many criteria for excused absence*, commonly including being above a certain age, being a student, or having a dependent child (precise regulations vary by jurisdiction, e.g., [138, 235]). These exemptions are necessary to avoid placing undue burdens on citizens, and similar exemptions would undoubtedly be needed if mandating participation in deliberative minipublics were mandated. Unsurprisingly, these exemptions—along other issues with people simply failing to appear—result in documented issues with certain populations being underrepresented among those who report for jury duty [270], posing a concern for the equity of our justice system's verdicts [26]. Based on this case study, it seems unlikely that a participation mandate—or any other intervention—could completely eliminate selection bias in deliberative minipublics, and moreover, failing to account for what bias remains would likely lead to systematic exclusion of certain groups. Our algorithms can ensure representation in the presence of selection bias, while retaining key properties of uniform lotteries to the greatest extent possible. Of course, as we will discuss in this thesis, efforts to decrease selection bias are extremely important; our algorithms align with this goal, improving in performance as selection bias decreases in severity.

## 1.2 Overview of Chapters

We begin by posing the question of whether our Key Impossibility above is really so fundamentally problematic. That impossibility says that we cannot give all *pool members* an equal chance of selection; however, it seems like to imitate idealized sortition, we should care about giving all *population members* an equal chance of participating, rather than all pool members. We now consider the implications of this distinction for the most directly-related ideal, *(ii) Fairness*. We make the distinction explicit by delineating two possible interpretations of fairness:

*Fairness of **outcome**.* All population members should have the same probability of *participating*.

*Fairness of **opportunity**.* All population members should have the same probability of *receiving the opportunity to participate*.

To unpack these notions, we first define the central object of Part I: a pool member's *selection probability* is their probability of being chosen for the panel. These two notions of fairness, then,

differ fundamentally in what they mean for the ideal selection probabilities. The first notion amounts to giving all *population members* an equal probability of ending up on the panel, which would be achieved by choosing each pool member with a probability *inversely proportional* to their chance of opting into the pool. The second notion amounts giving equal selection probability to all pool members, because the first stage is a uniform lottery, which gives everyone the same probability of being invited; then, to maintain equal *opportunity* to participate, we need to give all those who opted in an equal chance of being chosen for the panel in the second stage.

For reasons that will become apparent throughout Part I, we will focus primarily on the second notion of *Fairness*. However, the first notion *a priori* perhaps seems more theoretically natural, so in Chapter 2, we explore what it would take to design a selection algorithm that achieves it.

> **Chapter 2: A Selection Algorithm for Explicitly Reversing Self-Selection Bias**
>
> Based on *Neutralizing Self-Selection Bias in Sampling for Sortition* [128].
>
> In this chapter, we design a selection algorithm that achieves the first notion of *(ii) Fairness* via the intuition above: if each population member $i$ opts into the pool (conditional on being invited) with probability $q_i$, we want to choose them with probability proportional to $1/q_i$ in the second stage. This will give all population members an equal probability of ending up on the panel end-to-end, regardless of their opt-in probability.
>
> The key technical challenges here are (1) *knowing* individuals' unobservable opt-in probabilities $q_i$, so that our algorithm can set probabilities correctly; (2) determining conditions under which we can set all pool members' selection probabilities proportionally $1/q_i$ such that they remain in $[0, 1]$; and finally, (3) designing a procedure for turning these probabilities into a final panel that a. preserves these probabilities and b. is deterministically guaranteed to satisfy descriptively representative quotas.
>
> Our approach to challenge (1) begins with the observation that, by comparing the composition of the *pool* to that of the *population*, one can make inferences about which types of people tend to participate. For example, if the population is 50% women but the pool is only 20% women, you might infer that being a woman decreases one's chance of participating. We formalize this intuition by using these data to fit a model predicting the $q_i$'s via maximum likelihood estimation. Addressing challenge (2) requires an assumption that no participation probability is too low, and our guarantees depend on the extent to which this assumption holds. Finally, we address challenge (3) by designing a dependent rounding procedure based on a celebrated discrepancy theorem by Beck and Fiala [40]. This dependent rounding procedure preserves the $1/q_i$-proportional selection probabilities, while also guaranteeing that the final panel does not deviate from perfect proportional representation by more than $\pm|F|$.

While the selection algorithm presented in Chapter 2 is theoretically appealing, we argue that a main takeaway of this chapter is that *trying to explicitly reverse self-selection bias in practice is fraught with risks*. First, the levels of *Descriptive Representation* and *Fairness* achieved by this method hinge on one's ability to accurately estimate pool members' individual opt-in probabilities. In practice, this estimation must be done by comparing the pool versus the population

composition, which makes estimations fundamentally susceptible to errors for several reasons: (1) the population data serves as the "control group", but it truly contains many people who might have opted in if invited, so we have no data on exclusively people who *decline to participate*; (2) the available population data may be limited, especially at the level of *combinations* of attributes;[1] and (3) practitioners typically know very few feature-values about pool members, so predictions would likely be based on only *some* of the features that are crucial to the decision to opt-in.

Even if we could perfectly estimate pool members' opt-in probabilities, our rounding-based approach does not give strong enough guarantees on ex-post representation: in contrast to the custom quotas practitioners prefer to use, this rounding method may relax representation by $|F|$, which in practice typically ranges from 4-8; for groups that are small (which is not uncommon in real instances), this relaxation can dramatically weaken or even eliminate any guarantee of their inclusion. While it might be possible to drop this bound to $\sqrt{|F|}$ based on a conjecture by Beck and Fiala [40], it seems unlikely that it could be improved further due to a lower bound previously proven by Olson and Spencer [221]. In principle, one could altogether abandon the approach of first determining selection probabilities and then rounding them, though there is not a clear alternative approach that would reverse the selection bias while meaningfully circumventing this issue.

For the remainder of Part I, we will for now abandon the goal of *Fairness of* ***outcome***, and instead pursue *Fairness of* ***opportunity***, which we will henceforth refer to as *Fairness*. Our goal will now be to design selection algorithms that achieve the ideals outlined above—*(ii) Fairness*, *(iii) Transparency*, and *(iv) Manipulation Robustness*,—to the greatest degree possible subject to custom practitioner-defined quotas (standing in for *(i) Descriptive Representation*).

Our approach begins from the observation that these ideals were originally conferred by giving people *equal* selection probabilities. Given that this is impossible by our Key Impossibility, we pursue the next best goal: to make pool members' selection probabilities *as equal as possible*, subject to the quotas. Solving this technical task is the purpose of Chapter 3, whose primary contribution is an algorithmic framework that will serve as the basis of all later chapters.

> **Chapter 3. An Algorithmic Framework for Maximal Equality**.
>
> Based on *Fair Algorithms for Selecting Citizens' Assemblies* [130].
>
> Before designing any algorithms of its own, this paper's first contribution was to evaluate the selection algorithms that were being used in practice at the time. These pre-existing algorithms were greedy heuristics whose main goal was to find *any* valid panel, randomizing in ad-hoc fashion where possible along the way (this was a reasonable first goal, given that finding any single valid panel is NP-hard [130]). In our empirical evaluation, we found that a popular such algorithm—whose probabilistic properties had never been characterized—was prone to giving a large portion of the pool near-zero chance of selection, posing a significant

[1] For example, the European Social Survey data (the public population-level data corresponding to the UK climate assembly) is missing 335 of the 762 unique feature-value combinations that appear in the pool.

issue from the perspective of *Fairness*.

In response, our goal was to design a selection algorithm that made pool members selection probabilities *maximally equal*, subject to practitioner-defined quotas. A natural question, of course, is how one should measure "maximally equal". The algorithms we present in this paper will permit the use of *any convex function* that intakes a vector of selection probabilities and outputs a real-numbered measurement of their level of equality—a class which encompasses well-known notions like *Maximin* (no one receives too little selection probability), *Nash Welfare* (the geometric product of selection probabilities), or the *Gini Coefficient* (a popular measure of inequality).

The algorithmic framework we present, at a high level, first computes a distribution over valid panels, which we call a *panel distribution*, and then samples the final panel from that distribution. Note that by taking this approach, we ensure that the resulting panel is valid. Then, since any panel distribution implies selection probabilities[a], our task boils down to finding an *optimal* panel distribution—i.e., one that makes pool members' selection probabilities maximally equal.

The major technical challenge in computing an optimal panel distribution is that *a priori*, any optimal distribution might need to place selection probability on all valid panels, of which in practice there are astronomically many. Fortunately, we show that due to our equality objective being a function of just the selection probabilities (a very low-dimensional object relative to the space of all valid panels), by Caratheodory's theorem there must exist an optimal solution over only very few panels. Unfortunately, this theorem does not tell us how to *find* such a small-support distribution.

Our main contribution is an algorithmic framework for finding an optimal panel distribution over a practicable (but not theoretically bounded, in the worst case) number of panels. To understand this algorithm, it is useful to envision the primal program: we have one variable per every possible valid panel, corresponding to the probability we will place on that panel. Our goal is to find values of these (astronomically many) variables that optimize our convex objective function, which captures the equality of the selection probabilities implied by our panel distribution. The framework solves this massive program by solving its dual, using column generation to iteratively add panels to the support (corresponding to adding constraints to the dual) until a stopping condition is reached. As we prove, this stopping condition is sufficient for the existence of a KKT-condition-satisfying solution to the full dual program, corresponding to the current randomization being optimal among all randomizations over *all valid panels*.

[a] Given a panel distribution, the selection probability of any pool member is simply the probability of drawing a panel containing them.

With this algorithmic framework in hand, the question is then, *How can we best use it to serve our ideals of Fairness, Manipulation-Robustness, and Transparency?* This is the main question that will occupy us for the rest of Part I.

In the paper from Chapter 3, we instantiated our framework with an equality objective chosen to promote ***Fairness***. To translate this ideal into a mathematical objective, we began from the allocative justice perspective [259] that the chance to hold public office is a good to which people are *entitled* to their fair share. Accordingly, we defined "maximal fairness" according to the objective *Maximin*, which when optimized maximizes the *minimum* selection probability received by any pool member, corresponding to ensuring that no one receives too much less than their share. Our implementation (of a slight refinement of this objective called *Leximin*) is publicly available on Panelot.org and on the widely-used selection tool of the *Sortition Foundation*, a major organizer of citizens' assemblies [163]. Since 2020, this algorithm has been used to select high-profile assemblies, including the Global Climate Assembly in 2020, Michigan's statewide assembly on COVID-19 in 2021, Scotland's national climate assembly in 2022, Germany's National Assembly on Nutrition in 2023, and Ostbelgien's permanent assembly in 2023.

An additional contribution of Chapter 3 is a proposal for an algorithmic add-on targeting the goal of ***Transparency***. In its standard implementation, our algorithmic framework from Chapter 3 is not very transparent; it computes a complicated distribution over a few thousand panels, and then samples this distribution. In principle, one could publish this distribution, and then sample it by publicly running code that generates a random number. Needless to say (but we will say it anyway), this would be virtually impossible for the average person to understand. To remediate this, we proposed the following approach:

1. *Round* the optimal panel distribution produced by our algorithmic framework so that all probabilities in the panel distribution are multiples of 1/1000 (or some other integer denominator of the user's choice), while ensuring it remains a valid distribution.
2. *Number* these probability blocs from 000...999. Note that each bloc corresponds to a panel (with multiple blocs potentially corresponding to the same panel). Now, one has a list of 1000 panels (with duplicates).
3. *Uniformly sample* this list of panels to choose the final panel. By construction, this corresponds to sampling the rounded distribution from step 1.

This method is transparent in an important sense: if the list of panels and their (anonymized) members can be made public—which in practice, it has been [130]—this uniform lottery over panels allows the public to observe their own and other pool members' selection probabilities by the same simple reasoning required to understand that by buying more lottery tickets, you increase your chance of winning the lottery.[1]

Although Chapter 3 *proposed* this method, it left out an important open question: *Does the rounding of the optimal panel distribution significantly compromise its optimality?* In Chapter 4, we prove that it does not.

[1] If someone sees that they are on 20 out of 1000 panels, they immediately see that their chance of selection is 20/1000 = 2%.

**Chapter 4. *Transparency*.**

Based on *Fair Sortition Made Transparent* [131].

In this paper, we propose several algorithms for rounding the panel distribution output by the algorithm in Chapter 3. For each rounding algorithm, we upper-bound the maximum extent that it can change *any individual selection probability* in order to quantize any panel distribution. Though we extend these bounds to bound the optimality loss for only two equality objectives—*Maximin* and *Nash Welfare*—our bounds on changes to individual selection probabilities are general enough to bound optimality loss for most reasonable equality objectives. Finally, we empirically evaluate these rounding algorithms in real citizens' assembly datasets. This analysis identifies a simple and fast rounding procedure that almost exactly retains the optimality of the original panel distribution across instances. We conclude that *Transparency* comes at essentially no cost to maximal equality in practice, and at a practically bounded cost in theory.

Since its proposal, this rounding method has been used in conjunction with our algorithmic framework to select multiple citizens' assemblies, including aforementioned assemblies in Michigan and Germany.

We now turn our attention to the ideal of ***Manipulation Robustness***. In an unfortunate turn of events, our study of this ideal will reveal some bad news about the equality objectives we have studied so far.

**Chapter 5. *Manipulation Robustness***

Based on *Manipulation-Robust Citizens' Assembly Selection* [135].

In Chapter 3, translating the conceptual ideal of *Fairness* to a mathematical equality objective to plug into our algorithmic framework was relatively straightforward. However, the analogous transformation for *Manipulation Robustness* is less straightforward: to understand what equality objective minimizes incentives for pool members to misreport their features, we first need to define a game theoretic model. Instead of defining just one such model, we define three, each corresponding to different potential motive for misreporting one's features: to increase one's own selection probability, decrease someone else's, or to *steal seats*—that is, you might impersonate another group so that if you are selected, you will have taken a panel seat reserved for that group.

Our first finding is quite troubling: we show that *Maximin* and *Nash Welfare*, the two objectives we've studied for their prioritization of *Fairness*, are *arbitrarily manipulable*—that is, they permit a pool member, by misreporting their features, to gain selection probability 1. This is a worst-case result, but we show that this finding also holds in real datasets, even for very rudimentary manipulations requiring no knowledge of the algorithm.

The most striking aspect of this impossibility is that it persists *even as the pool grows arbitrarily large relative to the panel.*[a] To see why this is surprising, let $n$ be the pool size; as $n$ grows relative to $k$, the *average* selection probability $k/n$ should go down. Then, it seems that there

is less probability available per person, so shouldn't *everyone's* selection probability decrease, both pre- and post-manipulation, thereby decreasing the amount of probability that can be gained by manipulating? The key to understanding this impossibility is the intuition people can misreport combinations of features that *do not exist in the pool*, and which can make them "unicorns" to the objectives *Maximin* and *Nash Welfare*—both which almost exclusively care about making sure the lowest probability is not too low.[b] Here, someone is a "unicorn" when giving them more selection probability makes it feasible to raise the lowest selection probabilities. When someone misreports a combination of features that makes them a unicorn, both *Maximin* and *Nash Welfare* may pile probability onto them to the greatest extent possible, bringing their selection probability up to 1.

Based on the intuition that *high* selection probabilities are a problem for manipulation robustness, it should not come as a surprise that the equality objective *Minimax*, which minimizes the maximum selection probability, provably minimizes manipulation incentives.[c] We show that the manipulation incentives induced by *Minimax* decline at a rate of $O(k/n)$ as $n$ grows relative to $k$, which is the optimal possible rate for any selection algorithm.

[a] "Growing the pool" just means sending out more letters in the first stage, so the composition of the pool (and thus the level of selection bias) remains relatively constant.

[b] Maximin does this by definition. Nash Welfare, by being the *product* of selection probabilities, is relatively unaffected by probabilities near 1, but is extremely affected by even a single probability near 0.

[c] In the paper, we do not strictly study the objective *Minimax*, but rather the $\ell_p$ norms of selection probabilities, which effectively converge to *Minimax* as $p \to \infty$ (a regime we characterize). In subsequent work, we will consider *Minimax* in place of the $\ell_\infty$ norm.

**Remark 1.2.1** (Revisiting Our Approach From Chapter 2)**.** With Chapters 4 and 5 under our belt, we can now identify another reason why the approach taken in Chapter 2—to make pool members' probabilities inversely proportional to their chance of opting in—is practically dicey. The key reason is that this method will make pool members' probabilities widely different—far more different, potentially, than algorithms maximizing their equality. This is a problem for *Transparency*: if we make selection probabilities visible to the public and they are extremely disparate, this may create the sentiment that the process is very unfair (and the estimates upon which these probabilities are based are hard to soundly and transparently justify). Second, these disparate problems are an even *bigger* problem for *Manipulation Robustness*: the higher the probability someone can receive based on their features, the stronger the incentives for manipulation.

We have now studied *Fairness*, *Transparency*, and *Manipulation Robustness*; in our final research chapter in Part I, we will study the extent to which we can achieve these ideals *simultaneously*. If we can achieve this, we will complete our original goal: to design a sortition algorithm that achieves ideals (i)-(iv) to the greatest extent possible given the non-ideal conditions of real-world sortition.

**Chapter 6. *Fairness, Manipulation-Robustness,* and *Transparency***
Based on *Fair, Manipulation-Robust, and Transparent Sortition* [31].

Setting aside *Transparency* for a moment, the previous chapters reveal a potential tradeoff

between *Fairness* and *Manipulation Robustness*: low probabilities are a problem for the former (essentially by definition), and high probabilities are a problem for the latter. No equality objective we have studied achieves anywhere close to both ideals: *Maximin/Leximin* and *Nash Welfare* control only low probabilities, making them very fair but arbitrarily manipulable; by controlling only high probabilities, *Minimax* is optimally manipulation-robust but arbitrarily unfair, giving many pool members zero chance of selection.

In this chapter, we propose a new equality objective, called *Goldilocks*, that aims to achieve these ideals simultaneously by controlling both high and low selection probabilities, and which can be optimized via the framework in Chapter 3. The fundamental challenge in controlling high and low probabilities simultaneously is that manipulating coalitions, by misreporting, can affect the *quality of available lotteries* by reporting features between which there must be fundamental gaps between the maximum and minimum probability. Thus, in order to analyze our algorithm, we must first characterize the extent to manipulation can damage the space of feasible solutions, and *then* we can analyze the ability of our algorithm to recover good solutions despite this.

After circumventing these challenges, we give theoretical bounds (many of them tight) on the extent to which *Goldilocks* achieves *Fairness* and *Manipulation Robustness*, finding that in a very important sense, *Goldilocks* recovers among the best available solutions in a given instance. We then extend these theoretical bounds to the case where the output of *Goldilocks* is transformed to achieve a third goal, *Transparency*. Our empirical analysis of *Goldilocks* in real data is even more promising: we find that this objective achieves nearly instance-optimal minimum and maximum selection probabilities *simultaneously* in most real instances — an outcome not even guaranteed to be possible for any algorithm.

Although there is always room for future work here, in many respects, *Goldilocks* closes the question of whether we can simultaneously achieve three key ideals of sortition — *Fairness*, *Manipulation Robustness*, and *Transparency* — and contributes a practicable algorithm for doing so. Now that we better understand what is possible in lottery design regarding ideals *(ii)-(iv)*, in ongoing and future work, we can circle back to the very first ideal—*(i) Descriptive Representation*—and consider the implications of how quotas are used to enforce it.

**Chapter 7: Ongoing and Future Work.** In our first ongoing project, we are trying to help practitioners set more principled quotas. Currently, practitioners hand-design quotas without insight into how small changes in the quotas may change the lottery — changes that are difficult to predict, given the combinatorial relationship between the quotas and the space of possible lotteries. To close this gap, we are designing a deployable tool for fine-tuning quotas that are optimized to permit better lotteries. Empirically, our preliminary results show that small changes in quotas can permit significantly more uniform lotteries.

Sometimes our quota-tuning method relaxes a quota, e.g., reserving 22 instead of 24 seats for women, in favor of a far more uniform lottery. Often, we think of loosening quotas as *definitively harming* representation; however, expanding our conception of representation beyond

just the few features protected by quotas, it is not actually clear: a more uniform lottery can *also* support representation by ensuring that no group *unprotected* by quotas is systematically given low selection probability, plus it can decrease incentives for representation-corrupting manipulation. We are investigating this ambiguity by studying how different trade-offs between quota tightness and lottery uniformity affects *representation of unprotected groups* and the *diversity* of the resulting panel.

Additional ongoing and future work aims to alleviate other bottlenecks in the selection process, including the need to select alternate panel members to handle dropout.

# 2

# A Selection Algorithm for Explicitly Reversing Selection Bias

*Neutralizing Self-Selection Bias in Sampling for Sortition* [128].
Bailey Flanigan, Paul Gölz, Anupam Gupta, and Ariel D. Procaccia.
*NeurIPS* 2020.

## 2.1 Introduction

What if political decisions were made not by elected politicians but by a randomly selected panel of citizens? This is the core idea behind *sortition*, a political system originating in the Athenian democracy of the 5th century BC [272]. A *sortition panel* is a randomly selected set of individuals who are appointed to make a decision on behalf of population from which they were drawn. Ideally, sortition panels are selected via uniform sampling without replacement — that is, if a panel of size $k$ is selected from a population of size $n$, then each member of the population has a $k/n$ probability of being selected. This system offers appealing fairness properties for both individuals and subgroups of the population: First, each individual knows that she has the same probability of being selected as anyone else, which assures her an equal say in decision making. The resulting panel is also, in expectation, *proportionally representative* to all groups in the population: if a group comprises $x\%$ of the population, they will in expectation comprise $x\%$ of the panel as well. In fact, if $k$ is large enough, concentration of measure makes it likely that even a group's *ex post* share of the panel will be close to $x\%$. Both properties stand in contrast to the status quo of electoral democracy, in which the equal influence of individuals and the fair participation of minority groups are often questioned.

Due to the evident fairness properties of selecting decision makers randomly, sortition has seen

a recent surge in popularity around the world. Over the past year, we have spoken with several nonprofit organizations whose role it is to sample and facilitate sortition panels [75]. One of these nonprofits, the *Sortition Foundation*, has organized more than 20 panels in about the past year.[1] Recent high-profile examples of sortition include the Irish Citizens' Assembly,[2] which led to Ireland's legalization of abortion in 2018, and the founding of the first permanent sortition chamber of government,[3] which occurred in a regional parliament in the German-speaking community of Belgium in 2019.

The fairness properties of sortition are often presented as we have described them — in the setting where panels are selected *from the whole population* via uniform sampling without replacement. As we have learned from practitioners, however, this sampling approach is not applicable in practice due to limited participation: typically, only between 2 and 5% of citizens are willing to participate in the panel when contacted. Moreover, those who do participate exhibit self-selection bias, i.e., they are not representative of the population, but rather skew toward certain groups with certain features.

To address these issues, sortition practitioners introduce additional steps into the sampling process. Initially, they send a large number of invitation letters to a random subset of the population. If the recipients are willing to participate in a panel, they can opt into a *pool* of volunteers. Ultimately, the panel of size $k$ is sampled from the pool. Naturally, the pool is unlikely to be representative of the population, which means that uniformly sampling from the pool would yield panels whose demographic composition is unrepresentative of that of the population. To prevent grossly unrepresentative panels, many practitioners impose quotas on groups based on orthogonal demographic features such as gender, age, or residence inside the country. These quotas ensure that the ex-post number of panel members belonging to such a group lies within a narrow interval around the proportional share. Since it is hard to construct panels satisfying a set of quotas, practitioners typically sample using greedy heuristics. While these heuristics tend to be successful at finding valid panels, the probability with which an individual is selected is not controlled in a principled way.

Since individual selection probabilities are not deliberately chosen, the current panel selection procedure gives up most of the fairness guarantees associated with sortition via sampling from the whole population. Where uniform sampling selects each person with equal probability $k/n$, currently-used greedy algorithms do not even guarantee a minimum selection probability for members of the *pool*, let alone fair "end-to-end" probabilities with which members of the population will end up on the panel. As a further downside, the greedy algorithms we have seen being applied may need many attempts to produce a valid panel and might take exponential time to produce a valid panel even if one exists.

[1] https://www.youtube.com/watch?v=hz2d_8eBEKg at 8:53.

[2] https://2016-2018.citizensassembly.ie/en/

[3] https://www.politico.eu/article/belgium-democratic-experiment-citizens-assembly/

### 2.1.1 Our Techniques and Results

The main contribution of this paper is a more principled sampling algorithm that, even in the setting of limited participation, retains the individual fairness of sampling without replacement while allowing the deterministic satisfaction of quotas. In particular, our algorithm satisfies the following desiderata:

- *End-to-End Fairness:* The algorithm selects the panel via a process such that all members of the population appear on the panel with probability asymptotically close to $k/n$. This also implies that all groups in the population have near-proportional expected representation.
- *Deterministic Quota Satisfaction:* The selected panel satisfies certain upper and lower quotas enforcing approximate representation for a set of specified features.
- *Computational Efficiency:* The algorithm returns a valid panel (or fails) in polynomial time.

Deterministic quota satisfaction is a guarantee of group fairness, while end-to-end fairness, which recovers most of the ex ante guarantees of sampling without replacement, can be seen primarily as a guarantee of individual fairness. The phrase *end-to-end* refers to the fact that we are fair to individuals with respect to their probabilities of going from *population* to *panel*, across the intermediate steps of being invited, opting into the pool, and being selected for the panel.

The key challenge in satisfying these desiderata is self-selection bias, which can result in the pool being totally unrepresentative of the population. In the worst case, the pool can be so skewed that it contains no representative panel — in fact, the pool might not even contain $k$ members. As a result, no algorithm can produce a valid panel from every possible pool. However, we are able to give an algorithm that succeeds with high probability, under weak assumptions mainly relating the number of invitation letters sent out to $k$ and the minimum participation probability over all agents.

Crucially, any sampling algorithm that gives (near-)equal selection probability to all members of the population must reverse the self-selection bias occurring in the formation of the pool. We formalize this self-selection bias by assuming that each agent $i$ in the population agrees to join the pool with some positive participation probability $q_i$ when invited. If these $q_i$ values are known for all members of the pool, our sampling algorithm can use them to neutralize self-selection bias. To do so, our algorithm selects agent $i$ for the panel with a probability (close to) proportional to $1/q_i$, conditioned on $i$ being in the pool. This compensates for agents' differing likelihoods of entering the pool, thereby giving all agents an equal end-to-end probability. On a given pool, the algorithm assigns marginal selection probabilities to every agent in the pool. Then, to find a distribution over valid panels that implements these marginals, the algorithm randomly rounds a linear program using techniques based on discrepancy theory. Since our approach aims for a fair *distribution* of valid panels rather than just a single panel, we can give probabilistic fairness guarantees.

As we mentioned, our theoretical and algorithmic results take the probabilities $q_i$ of all pool members $i$ as given in the input. While these values are not observed in practice, we then show that

they can be estimated from available data. We cannot directly train a classifier predicting participation, however, because practitioners collect data only on those who *do* join the pool, yielding only positively labeled data. In place of a negatively labeled control group, we use publicly available survey data, which is unlabeled (i.e., includes no information on whether its members would have joined the pool). To learn in this more challenging setting, we use techniques from *contaminated controls*, which combine the pool data with the unlabeled sample of the population to learn a predictive model for agents' participation probabilities. Finally, we use data from a real-world sortition panel to show that plausible participation probabilities can be learned and that the algorithm produces panels that are close to proportional across features. For a synthetic population produced by extrapolating the real data, we show that our algorithm obtains fair end-to-end probabilities.

### 2.1.2 Related Work

Our work is broadly related to existing literature on fairness in the areas of *machine learning*, *statistics*, and *social choice*. Through the lens of fair machine learning, our quotas can be seen as enforcing approximate statistical fairness for protected groups, and our near-equal selection probability as a guarantee on individual fairness. Achieving simultaneous group- and individual-level fairness is a commonly discussed goal in fair machine learning [49, 151, 166], but one that has proven somewhat elusive. To satisfy fairness constraints on orthogonal protected groups, we draw upon techniques from discrepancy theory [34, 40], which we hope to be more widely applicable in this area.

Our paper addresses self-selection bias, which is routinely faced in statistics and usually addressed by sample reweighting. Indeed, our sampling algorithm can be seen as a way of reweighting the pool members under the constraint that weights must correspond to the marginal probabilities of a random distribution. While reweighting is typically done by the simpler methods of post-stratification, calibration [165], and sometimes regression [233], we use the more powerful tool of learning with contaminated controls [185, 277] to determine weights on a more fine-grained level.

Our paper can also be seen as a part of a broader movement towards statistical approaches in social choice [195, 197, 252]. The problem of selecting a representative sortition panel can be seen as a fair division problem, in which $k$ indivisible copies of a scarce resource must be randomly allocated such that an approximate version of the proportionality axiom is imposed. Our group fairness guarantees closely resemble the goal of apportionment, in which seats on a legislature are allocated to districts or parties such that each district is proportionally represented within upper and lower quotas [33, 58, 157].

So far, only few papers in computer science and statistics directly address sortition [44, 246, 274]. Only one of them [44] considers, like us, how to sample a representative sortition panel. Unfortunately, their stratified sampling algorithm assumes that all agents are willing to participate, which, as we address in this paper, does not hold in practice.

## 2.2 Model

**Agents.** Let $N$ be a set of $n$ agents, constituting the underlying population. Let $F$ be a set of *features*, where feature $f \in F$ is a function $f : N \to V_f$, mapping the agents to a set $V_f$ of possible values of feature $f$. For example, for the feature *gender*, we could have $V_{gender} = \{male, female, non\text{-}binary\}$. Let the *feature-value pairs* be $\bigcup_{f \in F} \{(f, v) \mid v \in V_f\}$. In our example, the feature-value pairs are (*gender*, *male*), (*gender*, *female*), and (*gender*, *non-binary*). Denote the number of agents with a particular feature-value pair $(f, v)$ by $n_{f,v}$.

Each agent $i \in N$ is described by her *feature vector* $F(i) := \{(f, f(i)) \mid f \in F\}$, the set of all feature-value pairs pertaining to this agent. Building on the example instance, suppose we add the feature *education-level*, so $F = \{gender, education\ level\}$. If *education level* can take on the values *college* and *no college*, a college-educated woman would have the feature-vector $\{(gender, female), (education\ level, college)\}$.

**Panel Selection Process.** Before starting the selection process, organizers of a sortition panel must commit to the panel's parameters. First, they must choose the number of *recipients* $r$ who will be invited to potentially join the panel, and the required *panel size* $k$. Moreover, they must choose a set of features $F$ and values $\{V_f\}_{f \in F}$ over which quotas will be imposed. Finally, for all feature-value pairs $(f, v)$, they must choose a *lower quota* $\ell_{f,v}$ and an *upper quota* $u_{f,v}$, implying that the eventual panel of $k$ agents must contain *at least* $\ell_{f,v}$ and *at most* $u_{f,v}$ agents with value $v$ for feature $f$. Once these parameters are fixed, the panel selection process proceeds in three steps:

$$population \xrightarrow{\textbf{STEP 1}} recipients \xrightarrow{\textbf{STEP 2}} pool \xrightarrow{\textbf{STEP 3}} panel$$

In **STEP 1**, the organizer of the panel sends out $r$ letters, inviting a subset of the population — sampled with equal probability and without replacement — to volunteer for serving on the panel. We refer to the random set of agents who receive these letters as $Recipients$. Only the agents in $Recipients$ will have the opportunity to advance in the process toward being on the panel.

In **STEP 2**, each letter recipient may respond affirmatively to the invitation, thereby opting into the pool of agents from which the panel will be chosen. These agents form the random set $Pool$, defined as the set of agents who received a letter and agreed to serve on the panel if ultimately chosen. We assume that each agent $i$ joins the pool with some *participation probability* $q_i > 0$. Let $q^*$ be the lowest value of $q_i$ across all agents $i \in N$. A key parameter of an instance is $\alpha := q^* r/k$, which measures how large the number of recipients is relative to the other parameters. Larger values of $\alpha$ will allow us the flexibility to satisfy stricter quotas.

In **STEP 3**, the panel organizer runs a *sampling algorithm*, which selects the panel from the pool. This panel, denoted as the set $Panel$, must be of size $k$ and satisfy the predetermined quotas for all feature-value pairs. The sampling algorithm may also fail without producing a panel.

We consider the first two steps of the process to be fully prescribed. The focus of this paper is to develop a sampling algorithm for the third step that satisfies the three desiderata listed in the

introduction: end-to-end fairness, deterministic quota satisfaction, and computational efficiency.

## 2.3 Sampling Algorithm

In this section, we give an algorithm which ensures, under natural assumptions, that every agent ends up on the panel with probability at least $\big(1 - o(1)\big)\, k/n$ as $n$ goes to infinity.[1] Furthermore, the panels produced by this algorithm satisfy non-trivial quotas, which ensure that the ex-post representation of each feature-value pair cannot be too far from being proportional.

Our algorithm proceeds in two phases: *I. assignment of marginals*, during which the algorithm assigns a marginal selection probability to every agent in the pool, and *II. rounding of marginals*, in which the marginals are dependently rounded to 0/1 values, the agents' indicators of being chosen for the panel. As we discussed previously, our algorithm succeeds only with high probability, rather than deterministically; it may fail in phase I if the desired marginals do not satisfy certain conditions. We refer to pools on which our algorithm succeeds as *good pools*. A good pool, to be defined precisely later, is one that is highly representative of the population — that is, its size and the prevalence of all feature values within it are close to their respective expected values. We leave the behavior of our algorithm on bad pools unspecified: while the algorithm may try its utmost on these pools, we give no guarantees in these cases, so the probability of representation guaranteed to each agent must come only from good pools and valid panels. Fortunately, under reasonable conditions, we show that the pool will be good with high probability. When the pool is good, our algorithm always succeeds, meaning that our algorithm is successful overall with high probability.

Our algorithm satisfies the following theorem, guaranteeing close-to-equal end-to-end selection probabilities for all members of the population as well as the satisfaction of quotas.

**Theorem 2.3.1.** *Suppose that $\alpha \to \infty$ and $n_{f,v} \geq n/k$ for all feature-value pairs $f, v$. Consider a sampling algorithm that, on a good pool, selects a random panel, $Panel$, via the randomized version of lemma 2.3.3, and else does not return a panel. This process satisfies, for all $i$ in the population, that*

$$\mathbb{P}[i \in Panel] \geq (1 - o(1))\, k/n.$$

*All panels produced by this process satisfy the quotas $\ell_{f,v} := (1 - \alpha^{-.49})\, k\, n_{f,v}/n - |F|$ and $u_{f,v} := (1 + \alpha^{-.49})\, k\, n_{f,v}/n + |F|$ for all feature-value pairs $f, v$.*

The guarantees of the theorem grow stronger as the parameter $\alpha = q^* r/k$ tends toward infinity, i.e., as the number $r$ of invitations grows. Note that, since $r \leq n$, this assumption requires that $q^* \gg k/n$. We defer all proofs to appendix A.2 and discuss the preconditions in appendix A.2.1.

### 2.3.1 Algorithm Part I: Assignment of Marginals

To afford equal probability of panel membership to each agent $i$, we would like to select agent $i$ with probability inversely proportional to her probability $q_i$ of being in the pool. For ease of

[1] We allow $k \geq 1$ and $r \geq 1$ to vary arbitrarily in $n$ and assume that the feature-value pairs are fixed.

notation, let $a_i := 1/q_i$ for all $i$. Specifically, for agent $i$, we want $\mathbb{P}[i \in Panel \mid i \in Pool]$ to be proportional to $a_i$. Achieving this exactly is tricky, however, because each agent's *selection probability* from pool $P$, call it $\pi_{i,P}$, must depend on those of all other agents in the pool, since their marginals must add to the panel size $k$. Thus, instead of reasoning about an agent's probability across all possible pools at once, we take the simpler route of setting agents' selection probabilities for each pool separately, guaranteeing that $\mathbb{P}[i \in Panel \mid i \in P]$ is proportional to $a_i$ across all members $i$ of a good pool $P$. For any good pool $P$, we select each agent $i \in P$ for the panel with probability

$$\pi_{i,P} := k\, a_i / \textstyle\sum_{j \in P} a_j.$$

Note that this choice ensures that the marginals always sum up to $k$.

**Definition of Good Pools.** For this choice of marginals to be reasonable and useful for giving end-to-end guarantees, the pool $P$ must satisfy three conditions, whose satisfaction defines a *good pool* $P$. First, the marginals do not make much sense unless all $\pi_{i,P}$ lie in $[0, 1]$:

$$0 \leq \pi_{i,P} \leq 1 \quad \forall i \in P. \tag{2.1}$$

Second, the marginals summed up over all pool members of a feature-value pair $f, v$ should not deviate too far from the proportional share of the pair:

$$(1 - \alpha^{-.49})\, k\, n_{f,v}/n \leq \textstyle\sum_{i \in P : f(i) = v} \pi_{i,P} \leq (1 + \alpha^{-.49})\, k\, n_{f,v}/n \quad \forall f, v. \tag{2.2}$$

Third, we also require that the term $\sum_{i \in P} a_i$ is not much larger than $\mathbb{E}[\sum_{i \in Pool} a_i] = r$, which ensures that the $\pi_{i,P}$ do not become to small:

$$\textstyle\sum_{i \in P} a_i \leq r/(1 - \alpha^{-.49}). \tag{2.3}$$

Under the assumptions of our theorem, pools are good with high probability, even if we condition on any agent $i$ being in the pool:

**Lemma 2.3.2.** *Suppose that $\alpha \to \infty$ and $n_{f,v} \geq n/k$ for all $f, v$. Then, for all agents $i \in Population$, $\mathbb{P}[Pool$ is good $\mid i \in Pool] \to 1$.*

Note that only constraint (2.1) prevents Phase II of the algorithm from running; the other two constraints just make the resulting distribution less useful for our proofs. In practice, if it is possible to rescale the $\pi_{i,P}$ and cap them at 1 such that their sum is $k$, running phase II on these marginals seems reasonable.

### 2.3.2 Algorithm Part II: Rounding of Marginals

The proof of Theorem 2.3.1 now hinges on our ability to implement the chosen $\pi_{i,P}$ for a good pool $P$ as marginals of a distribution over panels. This phase can be expressed in the language of randomized dependent rounding: we need to define random variables $X_i = \mathbb{1}\{i \in Panel\}$ for each $i \in Pool$ such that $\mathbb{E}[X_i] = \pi_{i,P}$. This difficulty of this task stems from the ex-post requirements on the pool, which require that $\sum_i X_i = k$ and that $\sum_{i : f(i) = v} X_i$ is close to $k\, n_{f,v}/n$ for all feature-value

pairs $f, v$. While off-the-shelf dependent rounding [78] can guarantee the marginals and the sum-to-$k$ constraint, it cannot simultaneously ensure small deviations in terms of the representation of all $f, v$.

Our algorithm uses an iterative rounding procedure based on a celebrated theorem by Beck and Fiala [40]. We sketch here how to obtain a deterministic rounding satisfying the ex-post constraints; the argument can be randomized using results by Bansal [34] or via column generation (Appendix A.2.4).[1] The iterated rounding procedure manages a variable $x_i \in [0, 1]$ for each $i \in Pool$, which is initialized as $\pi_{i,P}$. As the $x_i$ are repeatedly updated, more of them are fixed as either 0 or 1 until the $x_i$ ultimately correspond to indicator variables of a panel. Throughout the rounding procedure, it is preserved that $\sum_i x_i = \sum_i \pi_{i,P} = k$, and the equalities $\sum_{i:f(i)=v} x_i = \sum_{i:f(i)=v} \pi_{i,P}$ are preserved until at most $|F|$ variables $x_i$ in the sum are yet to be fixed. As a result, the final panel has exactly $k$ members, and the number of members from a feature-value pair $f, v$ is at least $\sum_{i:f(i)=v} \pi_{i,P} - |F| \geq (1 - \alpha^{-.49})\, k\, n_{f,v}/n - |F|$ (symmetrically for the upper bound).[2] As we show in appendix A.2.4,

**Lemma 2.3.3.** *There is a polynomial-time sampling algorithm that, given a good pool $P$, produces a random panel $Panel$ such that (1) $\mathbb{P}[i \in Panel] = \pi_{i,P}$ for all $i \in P$, (2) $|Panel| = k$, and (3) $\sum_{i:f(i)=v} \pi_{i,P} - |F| \leq |\{i \in Panel \mid f(i) = v\}| \leq \sum_{i:f(i)=v} \pi_{i,P} + |F|$.*

Our main theorem follows from a simple argument combining Lemmas 2.3.2 and 2.3.3 (Appendix A.2.5).

While the statement of theorem 2.3.1 is asymptotic in the growth of $\alpha$, the same proof gives bounds on the end-to-end probabilities for finite values of $\alpha$. If one wants bounds for a specific instance, however, bounds uniquely in terms of $\alpha$ tend to be loose, and one might want to relax Condition (2.2) of a good pool in exchange for more equal end-to-end probabilities. In this case, plugging the specific values of $n, r, k, q^*, n_{f,v}$ into the proof allows to make better trade-offs and to extract sharper bounds.

## 2.4 Learning Participation Probabilities

The algorithm presented in the previous section relies on knowing $q_i$ for all agents $i$ in the pool. While these $q_i$ are not directly observed, we can estimate them from data available to practitioners.

First, we assume that an agent $i$'s participation probability $q_i$ is a function of her feature vector $F(i)$. Furthermore, we assume that $i$ makes her decision to participate through a specific generative model known as *simple independent action* [119, as cited in [280]]. First, she flips a coin with

[1] Bansal [34] gives a black-box polynomial-time method for randomizing our rounding procedure. We found column-generation-based algorithms to be faster in practice, with guarantees that are at least as tight.

[2] Observe that our Beck-Fiala-based rounding procedure only increases the looseness of the quotas by a constant additive term beyond the losses to concentration. The concentration properties of standard dependent randomized rounding do not guarantee such a small gap with high probability. Moreover, our bound does not directly depend on the number of quotas (i.e., twice the number of feature-value pairs) but only depends on the number of features, which are often much fewer.

probability $\beta_0$ of landing on heads. Then, she flips a coin for each feature $f \in F$, where her coin pertaining to $f$ lands on heads with probability $\beta_{f,f(i)}$. She participates in the pool if and only if all coins she flips land on heads, leading to the following functional dependency:

$$q_i = \beta_0 \prod_{f \in F} \beta_{f,f(i)}.$$

We think of $1 - \beta_{f,v}$ as the probability that a reason specific to the feature-value pair $f, v$ prevents the agent from participating, and of $1 - \beta_0$ as the baseline probability of her not participating for reasons independent of her features. The simple independent action model assumes that these reasons occur independently between features, and that the agent participates iff none of the reasons occur.

If we had a representative sample of agents — say, the recipients of the invitation letters — labeled according to whether they decided participate ("positive") or not ("negative"), learning the parameters $\beta$ would be straightforward. However, sortition practitioners only have access to the features of those who enter the pool, and not of those who never respond. Without a control group, it is impossible to distinguish a feature that is prevalent in the population and associated with low participation rate from a rare feature associated with a high participation rate. Thankfully, we can use additional information: in place of a negatively-labeled control group, we use a *background sample* — a dataset containing the features for a uniform sample of agents, but without labels indicating whether they would participate. Since this control group contains both positives and negatives, this setting is known as *contaminated controls*. A final piece of information we use for learning is the fraction $\overline{q} := |Pool|/r$, which estimates the mean participation probability across the population. In other applications with contaminated controls, including $\overline{q}$ in the estimation increased model identifiability [277].

To learn our model, we apply methods for maximum likelihood estimation (MLE) with contaminated controls introduced by Lancaster and Imbens [185]. By reformulating the simple independent action model in terms of the logarithms of the $\beta$ parameters, their estimation (with a fixed value of $\overline{q}$) reduces to maximizing a concave function.

**Theorem 2.4.1.** *The log-likelihood function for the simple independent action model under contaminated controls is concave in the model parameters.*

By this theorem, proven in Appendix A.3, we can directly and efficiently estimate $\beta$. Logistic models, by contrast, require more involved techniques for efficient estimation [277].

## 2.5 Experiments

**Data.** We validate our $q_i$ estimation and sampling algorithm on pool data from *Climate Assembly UK*,[1] a national-level sortition panel organized by the Sortition Foundation in 2020. The panel consisted of $k = 110$ many UK residents aged 16 and above. The Sortition Foundation invited all members of 30 000 randomly selected households, which reached an estimated $r = 60\,000$ eligible

[1] https://www.climateassembly.uk/

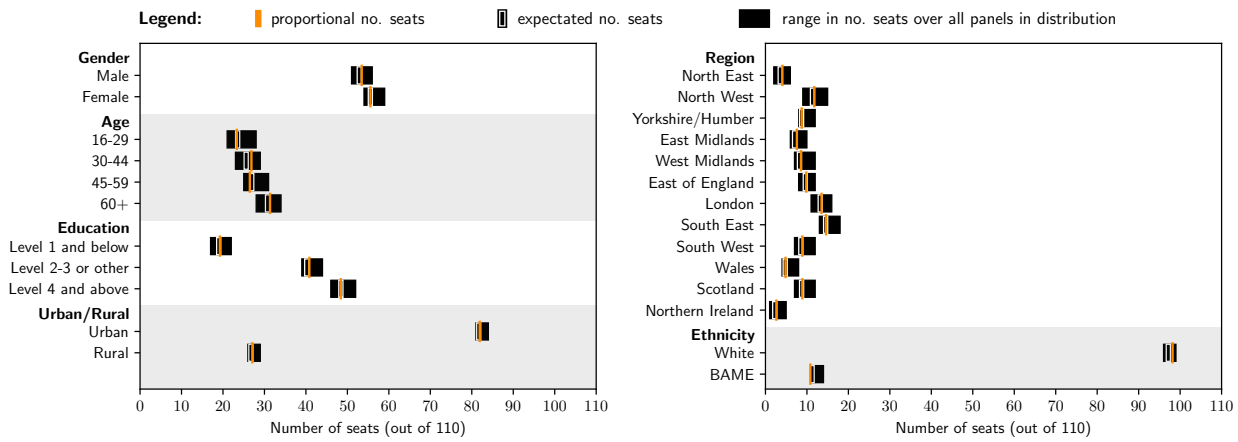



Figure 2.1: Expected and realized numbers of panel seats our algorithm gives each feature-value pair in the Climate Assembly pool.

participants.[1] Of these letter recipients, 1 715 participated in the pool,[2] corresponding to a mean participation probability of $\overline{q} \approx 2.9\%$. The feature-value pairs used for this panel can be read off the axis of fig. 2.1. We omit an additional feature *climate concern level* in our main analysis because only 4 members of the pool have the value *not at all concerned*, whereas this feature-value pair's proportional number of panel seats is 6.5. To allow for proportional representation of groups with such low participation rates, $r$ should have been chosen to be much larger. We believe that the merits of our algorithm can be better observed in parameter ranges in which proportionality can be achieved. For the background sample, we used the 2016 European Social Survey [216], which contains 1,915 eligible individuals, all with features and values matching those from the panel. Our implementation is based on PyTorch and Gurobi, runs on consumer hardware, and its code is available on github. Appendix A.4 contains details on Climate Assembly UK, data processing, the implementation, and further experiments (including the climate concern feature).

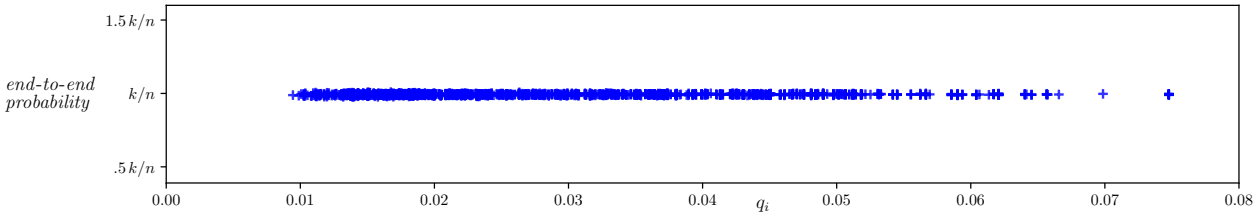
**Estimation of $m\beta$ Parameters.** We find that the baseline probability of participation is $\beta_0 = 8.8\%$. Our $\beta_{f,v}$ estimates suggest that (from strongest to weakest effect) highly educated, older, urban, male, and non-white agents participate at higher rates. These trends reflect these groups' respective levels of representation in the pool compared to the underlying population, suggesting that our estimated $\beta$ values fit our data well. Different values of the remaining feature, region of residence, seem to have heterogeneous effects on participation, where being a resident of the South West gives substantially increased likelihood of participation compared to other areas. The lowest participation probability of any agent in the pool, according to these estimates, is $q^* = 0.78\%$, implying that $\alpha \approx 4.25$. See Appendix A.4.4 for detailed estimation results and validation.

[1] Note that every person in the population has equal probability (30 000/#households) of being invited. We ignore correlations between members of the same household.

[2] Excluding 12 participants with gender "other" as no equivalent value is present in the background data.

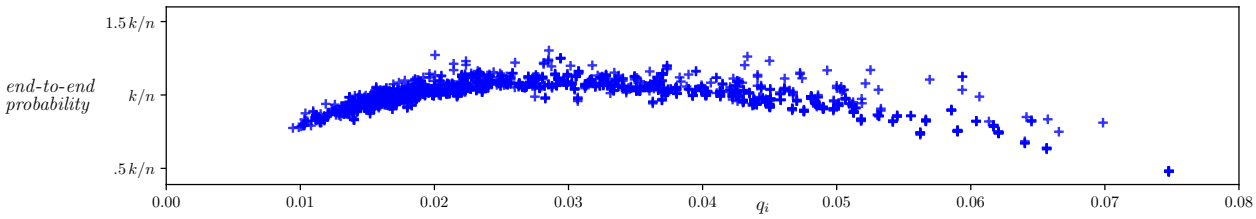
**Running the Sampling Algorithm on the Pool.** The estimated $q_i$ allow us to run our algorithm on the Climate Assembly pool and thereby study its fairness properties for non-asymptotic input sizes. We find that the Climate Assembly pool is good relative to our $q_i$ estimates, i.e., that it satisfies eqs. (2.1) to (2.3). As displayed in fig. 2.1, the marginals produced by Phase I of our algorithm give each feature-value pair $f, v$ an expected number of seats, $\sum_{i \in P, f(i)=v} \pi_{i,P}$, within *one seat* of its proportional share of the panel, $k\, n_{f,v}/n$. By lemma 2.3.3, Phase II of our algorithm then may produce panels from these marginals in which $f, v$ receives up to $|F| = 6$ fewer or more seats than its expected number. However, as the black bars in fig. 2.1 show, the actual number of seats received by any $f, v$ across *any panel* produced by our algorithm on this input never deviates from its expectation by more than 4 seats. As a result, while theorem 2.3.1 only implies lower quotas of $.51\, k\, n_{f,v}/n - |F|$ and upper quotas of $1.49\, k\, n_{f,v}/n + |F|$ for this instance, the shares of seats our algorithm produces lie in the much narrower range $k\, n_{f,v}/n \pm 5$ (and even $k\, n_{f,v}/n \pm 3$ for 18 out of 25 feature-value pairs). This suggests that, while the quotas guaranteed by our theoretical results are looser than the quotas typically set by practitioners, our algorithm will often produce substantially better ex-post representation than required by the quotas.

**End-to-End Probabilities.** In the previous experiments, we were only able to argue about the algorithm's behavior on a single pool. To validate our guarantees on individual end-to-end probabilities, we construct a synthetic population of size 60 million by duplicating the ESS participants, assuming our estimated $q_i$ as their true participation probabilities. Then, for various values of $r$, we sample a large number of pools. By computing $\pi_{i,P}$ values for all agents $i$ in each pool, we can estimate each agent's end-to-end probability of ending up on the panel. Crucially, we assume that our algorithm does not produce any panel for bad pools, analogously to theorem 2.3.1. As shown in the following graph, for $r = 60\,000$ (as was used in Climate Assembly UK), all agents in our synthetic population, across the full range of $q_i$, receive probability within $.1\, k/n$ of $k/n$ (averaged over 100 000 random pools):




That these end-to-end probabilities are so close to $k/n$ also implies that bad pools are exceedingly rare for this value of $r$. As we show in appendix A.4.6, we see essentially the same behavior for values of $r$ down to roughly 15 000, when $\alpha \approx 1$. For even lower $r$, most pools are bad, so end-to-end probabilities are close to zero under our premise that no panels are produced from bad pools.

To demonstrate that our algorithm's theoretical guarantees lead to realized improvements in individual fairness over the state-of-the-art, we re-run the experiment above, this time using the Sortition Foundation's greedy algorithm to select a panel from each generated pool. Since their algorithm requires explicit quotas as input, we set the lower and upper quotas for each feature-value group to be the floor and ceiling of that group's proportional share of seats. This is a popular way of setting quotas in current practice.




The results of this experiment show that the individual end-to-end probabilities generated by the currently-used greedy algorithm range from below $0.5\,k/n$ up to $1.3\,k/n$. In comparison to the end-to-end probabilities generated by our algorithm, those generated by the greedy algorithm are substantially skewed, and tend to disadvantage individuals with either low or high participation probabilities. One might argue that the comparison between our algorithm and the greedy is not quite fair, since the greedy algorithm is required to satisfy stronger quotas. However, looser quotas do not improve the behavior of the greedy algorithm; they simply make it behave more similarly to uniform sampling from the pool, which further disadvantages agents with low participation probability (for details, see appendix A.4.5).

Taken together, these results illustrate that, although greedy algorithms like the one we examined achieve proportional representation of a few pre-specified groups via quotas, they do not achieve fairness to individuals or to groups unprotected by quotas. Compared to the naive solution of uniform sampling from the pool, greedily striving for quota satisfaction does lead to more equal end-to-end probabilities, as pool members with underrepresented features are more likely to be selected for the panel than pool members with overrepresented features. However, this effect does not neutralize self-selection bias when there are multiple features, even when selection bias acts through the independent-action model as in our simulated population. Indeed, in this experiment, the greedy algorithm insufficiently boosts the probabilities of agents in the intersection of multiple low-participation groups (the agents with lowest $q_i$), while also too heavily dampening the selection probability of those in the intersection of multiple high-participation groups (with highest $q_i$). These observations illustrate the need for panel selection algorithms that explicitly control individual probabilities.

## 2.6 Discussion

In a model in which agents *stochastically* decide whether to participate, our algorithm guarantees similar end-to-end probabilities to all members of the population. Arguably, an agent's decision to participate when invited might not be random, but rather *deterministically* predetermined.

From the point of view of such an agent $i$, does our algorithm, based on a model that doesn't accurately describe her (and her peers') behavior, still grant her individual fairness? If $i$ deterministically *participates*, the answer is yes (if not, of course she cannot be guaranteed anything). To see why, first observe that, insofar as it concerns $i$'s chance of ending up on the panel, all other agents might as well participate randomly.[1] Indeed, from agent $i$'s perspective, the process looks

[1]Fix a group of agents who, assuming the stochastic model, will participate if invited with probability $q$. Then, sampling letter recipients from this set of agents in the stochastic model is practically equivalent to sampling recip-

like the stochastic process where every other agent $j$ participates with probability $q_j$, where $i$ herself always participates, and where the algorithm erroneously assumes that $i$ joins only with some probability $q_i$. Therefore, the pool is still good with high probability conditioned on $i$ being in it, as argued in lemma 2.3.2. Even if the algorithm knew that $q_i = 1$, $i$'s end-to-end probability would be at least $\big(1 - o(1)\big)\, k/n$, and the fact that the algorithm underestimates her $q_i$ only increases her probability of being selected from the pool. It follows that $i$'s end-to-end probability in this setting still must be at least around $k/n$.

Thus, in a deterministic model of participation, our individual guarantees are reminiscent of the axiom of population monotonicity in fair division: *If the whole population always participated when invited, every agent would reach the panel with probability $k/n$. The fact that some agents do not participate cannot (up to lower-order terms) decrease the selection probabilities for those who do.*

ients from this group in the deterministic model, if a $q$ fraction of the group deterministically participate.

# 3

# A Framework of Sortition Algorithms

*Fair Algorithms for Selecting Citizens Assemblies* [130].
Bailey Flanigan, Paul Gölz, Anupam Gupta, Brett Hennig, & Ariel D. Procaccia.
*Nature,* 2021.

*This exposition is adapted from the slightly modified version of this paper enclosed in [152].*

## 3.1 Introduction

In representative democracies, political representatives are usually selected by election. However, over the last 35 years, an alternative selection method has been gaining traction among political scientists [71, 87, 90] and practitioners [126, 203, 222, 225]: *sortition*, the random selection of representatives from the population. The chosen representatives form a panel, commonly called a *citizens' assembly*, which convenes to deliberate on a policy question. Citizens' assemblies are now being administered by around 50 organizations in over 25 countries[92], and just one of these organizations, the Sortition Foundation in the UK, recruited 29 panels in 2020. While many citizens' assemblies are initiated by civil-society organizations, [71, 87, 90, 92, 126, 203, 222, 225] they are also increasingly being commissioned by public authorities on municipal, regional, national, and supranational levels [222]. In fact, since 2019, multiple regional parliaments in Belgium and the Council of Paris have internally established permanent sortition bodies [137, 215]. Citizens' assemblies' growing utilization by governments is giving their decisions a more direct path to policy impact. For example, two recent citizens' assemblies commissioned by Ireland's national legislature led to the legalization of same-sex marriage and abortion [169].

Ideally, a citizens' assembly selected via sortition acts as a microcosm of society: its participants are representative of the population, and thus its deliberation simulates the entire population

convening "under conditions where it can really consider competing arguments and get its questions answered from different points of view" [124]. Whether this goal is realized in practice, however, depends on exactly how assembly members are chosen.

Panel selection is generally done in three stages: first, thousands of randomly chosen constituents are invited to participate. Second, a subset of the invited constituents opt into a *pool* of volunteers. Third, a panel of pre-specified size is randomly chosen from the pool via some fixed procedure, which we call a *selection algorithm* [82, 170, 211, 218]. As the final and most complex component of the selection process, the selection algorithm has great power in deciding who will be chosen to represent the population. In this chapter, we introduce selection algorithms that preserve the key desirable property of existing algorithms, while also more fairly distributing the sought-after opportunity [82, 170, 211, 218] of being a representative.

To our knowledge, all of the selection algorithms used in practice aim to satisfy one particular property, known as *descriptive representation*, the idea that the panel should reflect the composition of the population [124]. Unfortunately, the pool from which the panel is chosen tends to be far from representative. Specifically, it tends to overrepresent groups whose members are more likely to accept an invitation to participate, such as high educational attainment. To ensure descriptive representation despite the biases of the pool, selection algorithms require that the panels they output satisfy upper and lower *quotas* on a set of specified features, which are roughly proportional to each feature's population rate (e.g. quotas might require that a 40-person panel contain between 20 and 21 women). These quotas are generally imposed on feature categories delineated by gender, age, education level, and other attributes relevant to the policy issue at hand. We note that quota constraints of this form are more general than those achievable via *stratified sampling*, a common technique for drawing representative samples.

Selection algorithms that pre-date this work focused solely on satisfying quotas, leaving unaddressed a second property that is also central to sortition: that all individuals should have an *equal chance* of being chosen for the panel. Several political theorists present equality of selection probabilities as a central advantage of sortition, stressing its role in promoting the ideals such as *equality of opportunity* [71, 224], *democratic equality* [123, 124, 224, 258], and *allocative justice* [257, 258]. In fact, Engelstad, who introduced an influential model of sortition's benefits, argues that this form of equality constitutes "The strongest normative argument in favor of sortition" [112]. (See Appendix B.4 for more details on sortition desiderata from political theory.) In addition to political theorists, major practitioner groups have also advocated for equal selection probabilities [21, 201]. However, they face the fundamental hurdle that, in practice, the quotas almost always necessitate selecting people with somewhat unequal probabilities, as individuals from groups that are underrepresented in the pool must be chosen with disproportionately high probabilities to satisfy the quotas.

Though it is generally impossible to achieve *perfectly* equal probabilities, the reasons to strive for equality also motivate a more gradual version of this goal: making probabilities as equal as possible, subject to the quotas. We refer to this goal as *maximal fairness*. We find that our benchmark, a selection algorithm representing the previous state of the art, falls far short of this

goal, giving volunteers drastically unequal probabilities across several real-world instances. This algorithm even consistently selects certain types of volunteers with *near-zero* probability, thereby excluding them in practice from the chance to serve. We further show that, in these instances, it is possible to give all volunteers probability well above zero while still satisfying the quotas, demonstrating that the level of inequality produced by the benchmark is avoidable.

In this chapter, we close the gaps we have identified, both in theory and in practice. We first introduce not just one selection algorithm that achieves maximal fairness, but a more general (I) algorithmic framework for producing such algorithms. Motivated by the multitude of possible ways to quantify the fairness of an allocation of selection probabilities, our framework gives a maximally fair selection algorithm for any measure of fairness with a certain functional form. Notably, such measures include the most prominent from the literature on *fair division* [55, 206], and we show that these well-established metrics can be applied to our setting by casting the problem of assigning selection probabilities as one of fair resource allocation. Then, to bring this innovation into practice, we implement a (II) deployable selection algorithm, which is maximally fair according to one specific measure of fairness. We evaluate this algorithm and find that it is substantially fairer than the benchmark on several real-world datasets and by multiple fairness measures. Our algorithm is now in use by a growing number of sortition organizations around the world, making it one of only a few [62, 114, 150, 261] deployed applications of fair division.

## 3.2 Contribution I: Algorithmic Framework

### 3.2.1 Definitions

We begin by introducing necessary terminology. We refer to the input to a selection algorithm — a pool of size $n$, a set of quotas, and the desired panel size $k$ — as an *instance* of the panel selection problem. Given an instance, a selection algorithm randomly selects a *panel*, which is a quota-compliant set of $k$ pool members. We define the algorithm's *output distribution* on an instance as the distribution specifying the probabilities with which the algorithm outputs each possible panel. Then, a pool member's *selection probability* is the probability that they are on a panel randomly drawn from the output distribution. We refer to the mapping from pool members to their selection probabilities as the *probability allocation*, which we aim to make as fair as possible. Finally, a *fairness measure* is a function that maps a probability allocation to a fairness "score" (e.g. the geometric mean of probabilities, where higher is fairer). An algorithm is called *optimal* with respect to a fairness measure if, on any instance, the fairness of the algorithm's probability allocation is at least as high as that of any other algorithm.

### 3.2.2 Formulating the Optimization Task

To inform our approach, we first analyze the algorithms pre-dating ours. Those we have seen in use all have the same high-level structure: they select individuals for the panel one-by-one, in each step randomly choosing whom to add next from among those who, according to a myopic heuristic, seem unlikely to produce a quota violation later. Since finding a quota-compliant panel

is an algorithmically hard problem,[1] it is already an achievement that such simple algorithms find *any* panel in most practical instances. Due to their focus on finding any panel at all, however, these algorithms do not tightly control *which* panel they output, or more precisely, their output distribution (the probabilities with which they output different panels). Since an algorithm's output distribution directly determines its probability allocation, existing algorithms' probability allocations are also uncontrolled, leaving room for them to be highly unfair. In contrast to these

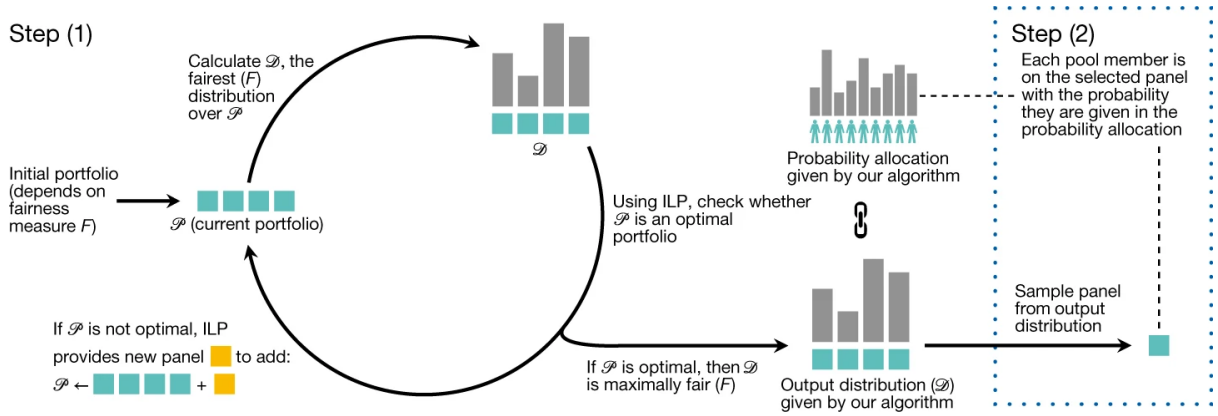



Figure 3.1: The steps of the algorithm optimizing the fairness measure $F$. The left-hand panel shows the implementation of step (1): constructing a maximally fair output distribution over panels (denoted by white boxes), which is done by iteratively building an optimal portfolio of panels and computing the fairest distribution over that portfolio. The right-hand panel shows step (2): sampling the distribution to select a final panel.

existing algorithms, which have output distributions that arise implicitly from a sequence of myopic steps, the algorithms in our framework (1) *explicitly* compute their own output distribution, and then (2) sample from that distribution to select the final panel (fig. 3.1). Crucially, the maximal fairness of the output distribution found in the first step makes our algorithms optimal. To see why, note that the behavior of *any* selection algorithm on a given instance is described by some output distribution; thus, since our algorithm finds the *fairest possible* output distribution, it is always at least as fair as any other algorithm.

Since step (2) of our selection algorithm is simply a random draw, we have reduced the problem of finding an optimal selection algorithm to the optimization problem in step (1) — finding a maximally fair distribution over panels. Now, to fully specify our algorithm, it remains only to solve this optimization problem.

[1] See supplementary information 6.

### 3.2.3 Solving the Optimization Task

A priori, it would seem that computing a maximally fair distribution might require constructing *all possible* panels, since achieving optimal fairness might necessitate assigning non-zero probability to all of them. Such an approach would be impracticable, however, as the number of panels in most instances is intractably large. Fortunately, since we measure fairness according to only individual selection probabilities, there must exist an *optimal portfolio* — a set of panels over which there exists a maximally fair distribution — containing few panels by Carathéodory's theorem:

**Proposition 3.2.1.** *Fix an arbitrary instance and a fairness measure $F$ for this instance. If there exists any maximally fair distribution over panels for $F$, there exists a maximally fair output distribution whose support includes at most $n + 1$ panels.*

*Proof.* Consider the hypercube $[0, 1]^n$, and associate each dimension with one pool member. A panel $P$ can be embedded into this space by its characteristic vector $\vec{v}_P \in \{0, 1\}^n$, whose $i$th component is one exactly if pool member $i$ is contained in $P$.

Fix a maximally fair output distribution, let $\mathcal{P}$ denote its support, and let $\{\lambda_P\}_{P\in\mathcal{P}}$ denote its probability mass function. Note that

$$\vec{p} \coloneqq \sum_{P\in\mathcal{P}} \lambda_P \vec{v}_P$$

is a probability allocation maximizing $F$, and that it is a convex combination of the $\{\vec{v}_P\}_{P\in\mathcal{P}}$. By Carathéodory's theorem, there is a subset $\mathcal{P}' \subseteq \mathcal{P}$ of size at most $n+1$ such that $\vec{p}$ still lies in the convex hull of this smaller set. Thus, there are nonnegative real numbers $\{\lambda'_P\}_{P\in\mathcal{P}'}$ adding up to one such that

$$\vec{p} = \sum_{P\in\mathcal{P}'} \lambda'_P \vec{v}_P.$$

These $\lambda'_P$ form the probability mass function of a distribution over at most $n + 1$ panels, which has the same probability allocation $\vec{p}$ as the original maximally fair distribution, which implies that the new distribution is also maximally fair for $F$. □

This result brings a practical algorithm within reach, and shapes the goal of our algorithm: to find an optimal portfolio while constructing as few panels as possible.

We accomplish this goal using an algorithmic technique called *column generation*, where, in our case, the "columns" being generated correspond to panels. A more in-depth discussion and formal description of this algorithm, as well as proofs of correctness, can be found in supplementary information 8. As shown in fig. 3.1, our algorithms find an optimal portfolio by iteratively adding panels to a portfolio $\mathcal{P}$, in each iteration alternating between two subtasks: (i) finding the optimal distribution $\mathcal{D}$ over only the panels currently in $\mathcal{P}$ and (ii) adding a panel to $\mathcal{P}$ that, based on the gradient of the fairness measure, will move the portfolio furthest towards optimality. This second subtask makes use of *integer linear programming*, which we use to generate quota-compliant panels despite the theoretical hardness of the problem. Eventually, the panel with the most promising

gradient will already be in $\mathcal{P}$, in which case $\mathcal{P}$ is provably optimal and $\mathcal{D}$ must be a maximally fair distribution. In practice, we observe that this procedure terminates after few iterations.

Our techniques extend column generation methods that are typically applied to linear programs, allowing them to be used to solve a large set of convex programs.This extension allows our framework to be used with a wide range of fairness measures — essentially any for which the fairest distribution over a portfolio can be found via convex programming. Supported measures include those most prominent in the fair division literature: egalitarian welfare [111], Nash welfare [206], Gini inequality [110, 187], and the Atkinson indices [110, 247][247]. Our algorithmic approach also has the benefit of easily extending to organization-specific constraints beyond quotas; for example, practitioners can prevent multiple members of the same household from appearing on the same panel. Due to its generality, our framework even applies to domains outside of sortition, including the allocation of classrooms to charter schools [182] and kidney exchange [243].

### 3.3 Contribution II: Deployable Selection Algorithm

To bring fair panel selection into practice, we develop an efficient implementation of one specific selection algorithm, which we call LexiMin (formally defined in supplementary information 10). LexiMin optimizes the well-established fairness measure *leximin* [51, 182, 206], a fairness measure that is sensitive to the very lowest selection probabilities. In particular, leximin is optimized by maximizing the lowest selection probability, then breaking ties between solutions in favor of probability allocations with highest second-lowest probability, and so on. This choice of fairness measure is motivated by the fact that, as we show in this section and in supplementary information 13, Legacy gives some pool members a near-zero probability when much more equal probabilities are possible. This type of unfairness is especially pressing because, if it consistently impacted pool members with certain combinations of features, these individuals and their distinct perspectives would be "systematically excluded from participation" [250], which runs counter to a key promise of random selection.

To increase the accessibility of LexiMin, we made its implementation available through an existing open-source panel selection tool [164] and on Panelot [153], a website where anyone can run the algorithm without installation. LexiMin has since been deployed by several organizations, including *Cascadia* (US), the *Danish Board of Technology* (Denmark), *Nexus* (Germany), *of by for ** (US), *Particitiz* (Belgium), and the *Sortition Foundation* (UK). As of July 2021, the Sortition Foundation alone had already used LexiMin to select more than 40 panels.

We measure the impact of adopting LexiMin over pre-existing algorithms by comparing its fairness to that of a benchmark, Legacy (supplementary information 11), the algorithm used by the Sortition Foundation prior to their adoption of LexiMin. We choose Legacy as a benchmark because it was widely used prior to this work, it is similar to several other selection algorithms used in practice (see supplementary information 13) and it is the only existing algorithm we found that was fully specified by an official implementation. We compare the LexiMin and Legacy on ten datasets from real-world panels, with respect to several fairness measures including the minimum probability (table 3.1), the Gini coefficient, and the geometric mean. In this analysis, we find that

| instance | $n$ | $k$ | # of features | $k/n$ | Legacy min. probability (sampled)[1] | LexiMin min. probability (exact) | LexiMin running time |
|---|---|---|---|---|---|---|---|
| sf(a) | 312 | 35 | 6 | 11.2% | $\leq$ 0.32% | 6.7% | 20 sec |
| sf(b) | 250 | 20 | 6 | 8.0% | $\leq$ 0.17% | 4.0% | 9 sec |
| sf(c) | 161 | 44 | 7 | 27.3% | $\leq$ 0.15% | 8.6% | 6 sec |
| sf(d) | 404 | 40 | 6 | 9.9% | $\leq$ 0.11% | 4.7% | 46 sec |
| sf(e) | 1727 | 110 | 7 | 6.4% | $\leq$ 0.03% | 2.6% | 67 min |
| cca | 825 | 75 | 4 | 9.1% | $\leq$ 0.03% | 2.4% | 7 min |
| hd | 239 | 30 | 7 | 12.6% | $\leq$ 0.09% | 5.1% | 37 sec |
| mass | 70 | 24 | 5 | 34.3% | $\leq$ 14.9% | 20.0% | 1 sec |
| nexus | 342 | 170 | 5 | 49.7% | $\leq$ 2.24% | 32.5% | 1 min |
| obf | 321 | 30 | 8 | 9.3% | $\leq$ 0.03% | 4.7% | 3 min |

Table 3.1: List of instances used in our experiments. For the *instances* we study, panels were recruited by the following organisations. sf(a-e): Sortition Foundation; cca: Center for Climate Assemblies; hd: Healthy Democracy; mass: MASS LBP; nexus: Nexus; obf: of by for * (At the request of practitioners, topics, dates, and locations of the panels are not identified.) $n$ is the pool size, $k$ is the panel size, and consequently, $k/n$ is the mean selection probability. The *# of features* is $|F|$, where each $f \in F$ has between 2 and 49 possible values (with the typical range being 2-5).

LexiMin is fairer on all instances we examine, and substantially so in nine out of ten.

### 3.4 Effect of Adopting LexiMin over Legacy

We study datasets from ten sortition panels, organized by six different sortition organizations in Europe and North America. As Table 3.1 shows, our instances are diverse in panel size (range: 20–170, median: 37.5) and number of quota categories (range: 4–8). On consumer hardware, the run-time of our algorithm is well within the time available in practice.

Out of concern about low selection probabilities, we first compare the minimum selection probabilities given by Legacy and LexiMin, summarized in the second and third columns from the right in Table 3.1. Strikingly, in all instances except mass (an outlier in that its quotas only mildly restrict the fraction of panels that are feasible), Legacy chooses some pool members with probability close to zero. In fact, we can identify combinations of features that lead to low selection probabilities *across all instances*,[2] raising the concern that Legacy may in fact systematically exclude some groups from participation. By contrast, LexiMin selects no one nearly so infrequently, with minimum selection probabilities ranging from 26% to 65% (median: 49%) of $k/n$, the "ideal" probability individuals would receive in the absence of quotas. One might wonder whether this increased minimum probability achieved by LexiMin affects only a few pool members most disadvantaged by Legacy. This is not the case: As shown in Figure 3.2 by the shaded

[2]See methods section "Individuals rarely selected by Legacy" of the full version.

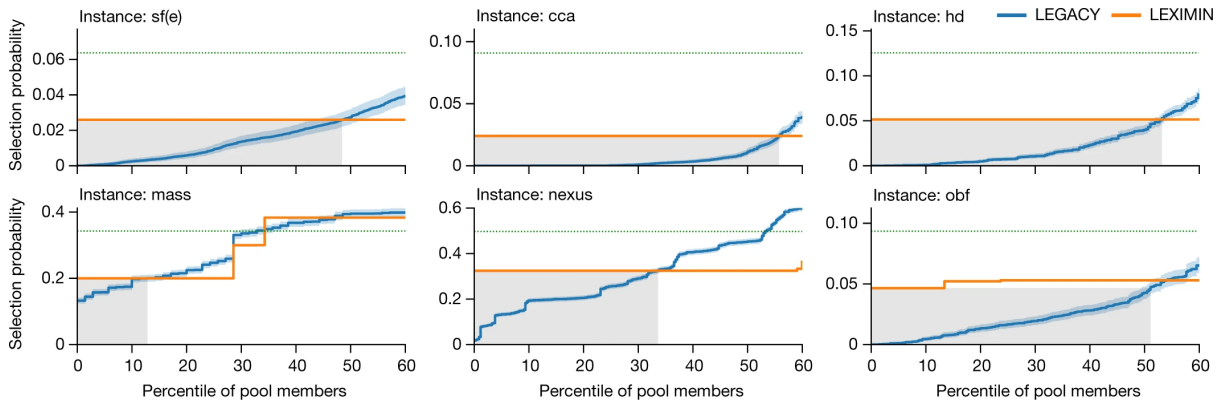



Figure 3.2: Selection probabilities given by Legacy and LexiMin to the bottom 60% of pool members on six representative instances, where pool members are ordered in order of increasing probability given by the respective algorithms. Shaded boxes denote the range of pool members whose selection probability given by Legacy is lower than the minimum probability given by LexiMin. Legacy probabilities are estimated over 10,000 random panels and are indicated with 99% confidence intervals (see methods section "Statistics" of the full version). For corresponding graphs for all other instances and up to the 100th percentile, see Figures 3.4 and 3.5 respectively in Section 3.6.

boxes, between 13% and 56% of pool members (median 46%) across instances receive probability from Legacy lower than the *minimum given to anyone* by LexiMin (Table 3.7). Thus, even just the first stage of LexiMin, i.e., maximizing the *minimum* probability, provides a sizable section of the pool with more equitable access to the panel.

We have so far compared Legacy and LexiMin over only the lower end of selection probabilities, as this is the range in which LexiMin prioritizes being fair. However, even considering the *entire* range of selection probabilities, we find that LexiMin is quantifiably fairer than Legacy on all instances by two established metrics of fairness, namely the Gini Coefficient and the geometric mean (Table 3.6). For example, across instances excluding `mass`, LexiMin decreases the Gini coefficient, a standard measure of inequality, by between 5 and 16 percentage points (median: 12; negligible improvement on `mass`). Strikingly, the 16-point improvement in the Gini coefficient achieved by LexiMin on the instance `obf` (from 59% to 43%) approximately reflects the gap between relative income inequality in Namibia (59% in 2015) and the United States (42% in 2019) [285].

## 3.5 Discussion

As the recommendations made by citizens' assemblies increasingly impact public decision-making, so grows the urgency that selection algorithms distribute this power fairly across constituents.

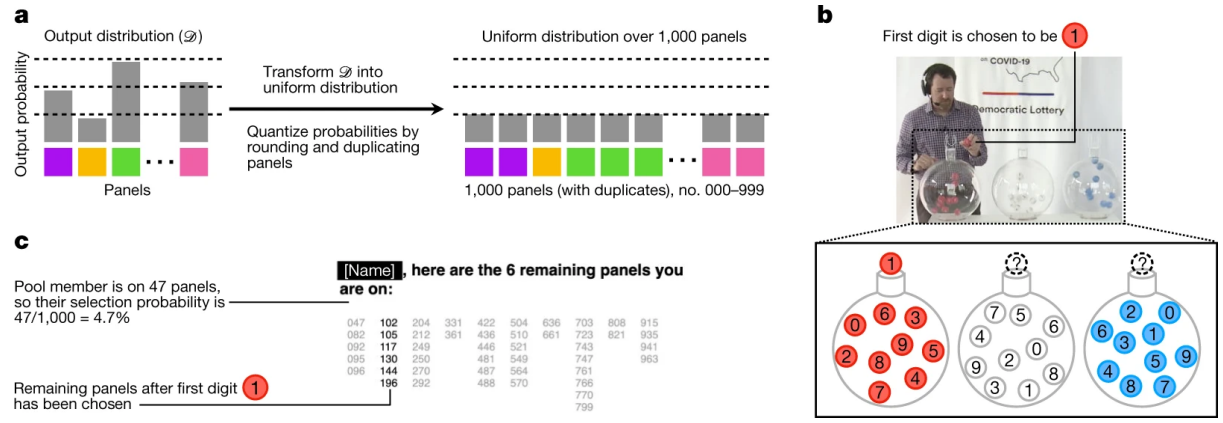



Figure 3.3: How LexiMin's output was used to select a panel via a live uniform lottery. (a) First, the output distribution was transformed into a uniform distribution over 1,000 panels, numbered 000–999. (b) The three digits determining the final panel were drawn from lottery machines, making each panel observably selected with equal probability. (c) The personalized interface (screen-captured with (b)) shows each pool member the number of panels out of 1,000 they are on, allowing them to verify their own and others' selection probabilities. Screenshots credit: *of by for **.

We have made substantial progress on this front: the optimality of our algorithmic framework conclusively resolves the search for fair algorithms for a broad class of fairness measures, and the deployment of LexiMin puts an end to some pool members being virtually never selected in practice.

Beyond these immediate benefits to fairness, the exchange of ideas we have initiated between practitioners and theorists presents continuing opportunities to improve panel selection in areas such as transparency. For example, for an assembly in Michigan, we assisted *of by for ** in selecting their panel via a live lottery in which participants could easily observe the probabilities with which each pool member was selected. This is an advance over the transparency possible with previous selection algorithms. We found that, in this instance, the output distribution of LexiMin could be transformed into a simple lottery without meaningful loss of fairness (fig. 3.3). Subsequent work by Flanigan et al. [131] developed general procedures and bounds for this transformation.

The *Organisation for Economic Co-operation and Development (OECD)* describes citizens' assemblies as part of a broader democratic movement to "give citizens a more direct role in [...] shaping the public decisions that affect them" [222]. By bringing mathematical structure, increased fairness, and greater transparency to the practice of sortition, research in this area promises to put practical sortition on firmer foundations, and to promote citizens' assemblies' mission to give everyday people a greater voice.

## 3.6 Additional Methods and Empirical Analysis

### 3.6.1 Extended results for Figure 3.2,Table 3.1

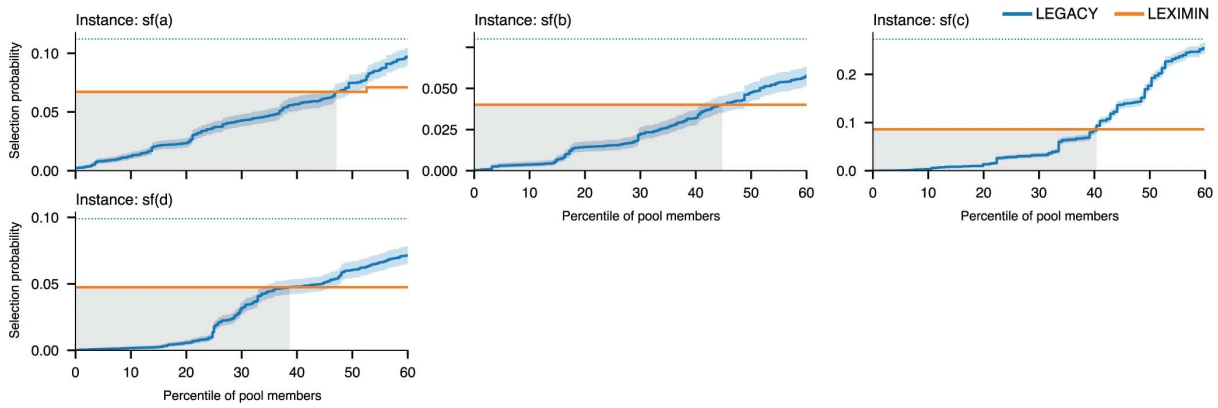



Figure 3.4: Selection probabilities given by Legacy and LexiMin to the bottom 60% of pool members on the 4 instances that are not shown in Figure 3.2. Pool members are ordered across the x axis in order of increasing probability given by the respective algorithms. Shaded boxes denote the range of pool members with a selection probability given by Legacy that is lower than the minimum probability given by LexiMin. Legacy probabilities are estimated over 10,000 random panels and are indicated with 99% confidence intervals (as described in Statistics in the Methods). Green dotted lines show the equalized probability $(k/n)$.

### 3.6.2 Individuals rarely selected by Legacy

The empirical results in Table 3.1 demonstrate that, in most instances, Legacy selects some pool members with very low probability. However, in any given citizens assembly, this does not automatically imply that these individuals had low probability of serving on the panel. Indeed, if such an individual would have been selected by Legacy with higher probability in most other pools that could have formed (as a result of other sets of agents being randomly invited alongside this individual), then the individual might still have had a substantial overall probability of serving on the citizens assembly.

In this section, we show how our data suggest that this is not the case, and that some people do in fact seem to have very low likelihood overall of ending up on the panel when Legacy is used. We make this case by demonstrating two separate points. First, we show that, across instances, Legacy tends to give very low selection probabilities to agents who have many features that are overrepresented in the observed pool relative to the quotas. Second, we discuss why it is likely that, across possible pools for the same citizens assembly, it is usually the same agents who have many overrepresented features. These two points, taken together, suggest that agents who have many overrepresented features in the pools we observe are rarely selected by Legacy overall.

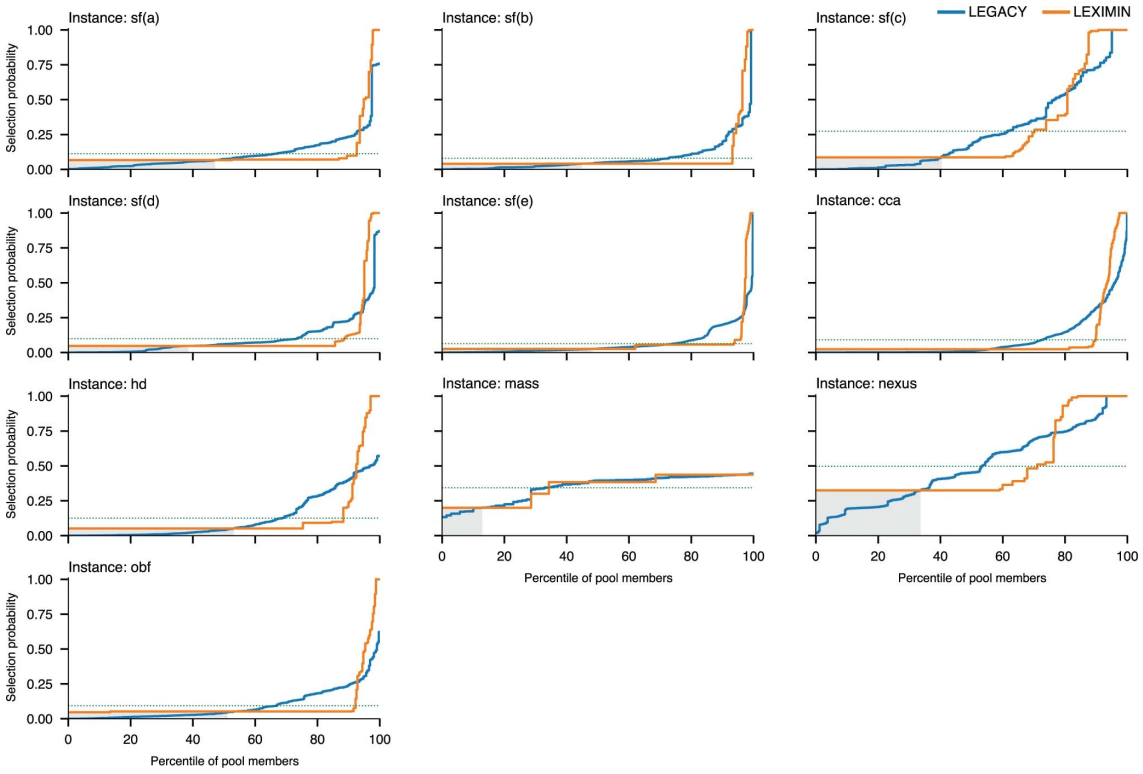



Figure 3.5: Selection probabilities given by Legacy and LexiMin on all ten instances. Pool members are ordered across the x axis in order of increasing probability given by the respective algorithms. In contrast to Figure 3.2 and Figure 3.4, this graph shows the full range of selection probabilities (up to the 100th percentile). Shaded boxes denote the range of pool members with a selection probability given by Legacy that is lower than the minimum probability given by LexiMin. Legacy probabilities are estimated over 10,000 random panels and are indicated with 99% confidence intervals (as described in Statistics in the Methods). Green dotted lines show the equalized probability $(k/n)$.

**Relationship between overrepresentation of features and selection probability.** To measure the relationship between the level of overrepresentation of an agents features and that agents selection probability by Legacy, we first construct a simple indicator called the ratio product, which measures the level of overrepresentation of a given agents set of features in the pool. The ratio product is composed of, for each of the features of an agent, the ratio between the fraction of this feature in the pool and the fraction of the quotas of the feature (specifically, the mean of lower and upper quota) in the panel. That is, if we denote the set of pool members with a feature $f$ by $N_f$ and if we denote the lower and upper quotas of the feature by $\ell_f$ and $u_f$, respectively,

| Instance | Gini coefficient of LEGACY (lower is fairer) | Gini coefficient of LEXIMIN (lower is fairer) | Geometric mean of LEGACY (higher is fairer) | Geometric mean of LEXIMIN (higher is fairer) |
|---|---|---|---|---|
| *sf(a)* | 51.2% | 37.3% | 6.5% | 8.1% |
| *sf(b)* | 59.6% | 47.4% | 3.5% | 4.8% |
| *sf(c)* | 57.0% | 52.5% | 8.3% | 16.3% |
| *sf(d)* | 59.3% | 48.7% | 3.5% | 6.0% |
| *sf(e)* | 64.4% | 51.2% | 2.2% | 3.9% |
| *cca* | 75.3% | 67.8% | 0.7% | 3.5% |
| *hd* | 64.5% | 52.9% | 3.1% | 7.3% |
| *mass* | 14.9% | 14.8% | 32.6% | 32.7% |
| *nexus* | 30.8% | 25.4% | 40.9% | 44.2% |
| *obf* | 58.9% | 42.7% | 3.7% | 6.2% |

Figure 3.6: Gini coefficient and geometric mean of probability allocations of both algorithms, for each instance. On every instance, LEGACY has a lower Gini coefficient and a larger geometric mean. For computing the geometric mean, we slightly correct upward empirical selection probabilities of LEGACY that are close to zero (as described in Statistics in the Methods).

| Instance | Share selected by LEGACY with probability below LEXIMIN minimum selection probability |
|---|---|
| *sf(a)* | 47.1% |
| *sf(b)* | 44.8% |
| *sf(c)* | 40.4% |
| *sf(d)* | 38.6% |
| *sf(e)* | 48.4% |
| *cca* | 55.8% |
| *hd* | 53.1% |
| *mass* | 12.9% |
| *nexus* | 33.6% |
| *obf* | 51.1% |

Figure 3.7: For each instance, the share of pool members selected with lower probability by LEGACY than the minimum selection probability of LEXIMIN is shown. This corresponds to the width of the shaded boxes in Figures 3.2, 3.4 and 3.5.

then the ratio product of an agent $i$ is defined as:

$$\prod_{\text{features } f \text{ of } i} \frac{|N_f| \,/\, n}{(\ell_f + u_f) \,/\, 2k}.$$

Given that the quotas are typically set in proportion to the share of the feature in the population, we say that agents with a high ratio product have many overrepresented features. Using this indicator, we find that there is a clear negative relationship in all instances between the ratio product of an individual and their selection probability by LEGACY (Figure 3.8). Most importantly, as this trend would suggest, we find that the pool members with the largest ratio products consistently have some of the lowest selection probabilities.

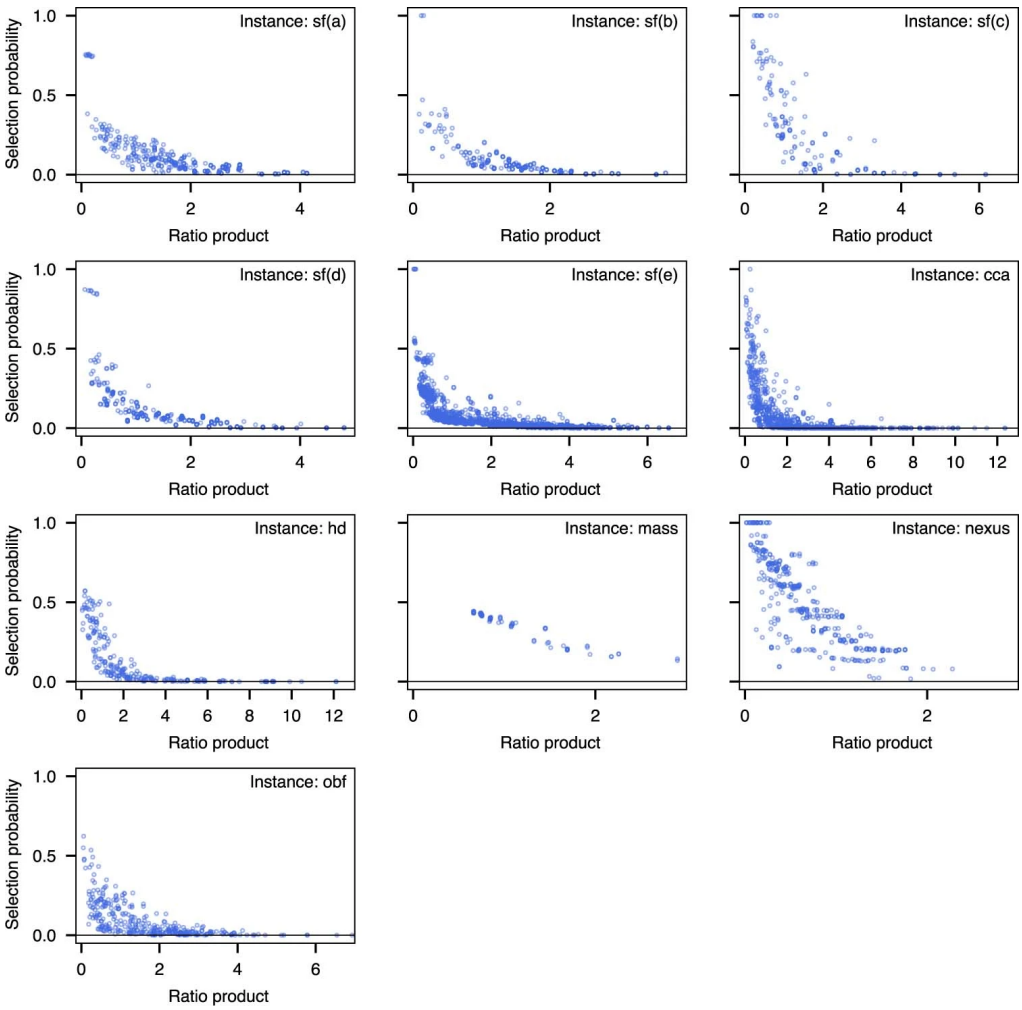



Figure 3.8: Relationship between how overrepresented the features of an agent are and how likely they are to be chosen by the Legacy algorithm. The level of overrepresentation is quantified as the ratio product (as described in Individuals rarely selected by Legacy in the Methods); agents further to the right are more overrepresented. Across instances, pool members with high ratio product are consistently selected with very low probabilities.

**The same agents probably have many overrepresented features across most possible pools.** Recall that we define an instance with respect to a single pool. However, this observed pool is only one among several hypothetical pools that could have resulted from the random process of sending out invitation letters. We define the ratio product of an agent with respect to a single instance and, therefore, a single observed pool. Then, if a different hypothetical pool (including that agent) had instead been drawn during the invitation process, the ratio product of the same agent with respect to that pool would probably be different, depending on which constituents were invited to join the pool alongside them. As the quotas and the target panel size k would be the same for all these hypothetical instances, the differences in ratio product

would be due to different values of $|N_f|$, for all features f of the agent. Here, $|N_f|$ — a random variable, the value of which is determined during the random invitation process — essentially follows a hypergeometric distribution, because it is simply the number of invitations sent to constituents who both have feature f and are willing to participate. Consequentially, all $|N_f|$ are well-concentrated, from which it follows that the ratio product of an individual should not vary much across all hypothetical pools containing them. The ratio product should be especially concentrated when all of an individuals features tend to be overrepresented, and thus all factors of the ratio product are large.

**Interpretation of results.** The analysis so far suggests that Legacy selects individuals with many overrepresented features with low probability. Even so, one might consider the possibility that these individuals are more likely to join the pool if invited (given that they are overrepresented in the pool), and that, therefore, their lower selection probability by Legacy in the panel-selection stage is outweighed by their higher probability of entering the pool in the pool-formation stage. This raises the question of whether the low selection probabilities given to these individuals by Legacy are necessarily inconsistent with a scenario in which the probabilities of people going from population to panel (their end-to-end probabilities [128]) are actually equal.

A back-of-the-envelope calculation suggests that this is not the casethat, in fact, the end-to-end probabilities are probably far from equal when using Legacy. Across instances, the median ratio between the average selection probability k/n and (the upper confidence bound on) the minimum selection probability given by Legacy is larger than 100. If the selection probability of an individual conditioned on appearing in some pool is indeed 100 times lower than that of an average citizen, the individual would have to enter the pool 100 times more frequently than this average citizen to serve on the panel with equal end-to-end probability. Given that average response rates are typically between 2 and 5%, someone opting into the pool 100 times more frequently than an average citizen is simply not possible.

Although we have demonstrated that Legacy underrepresents a specific group (agents with many overrepresented features), we do not have reason to believe that Legacy would exclude groups defined by intersections of few features (for example, young women or conservatives with a university degree are the intersection of two features). In Supplementary Information section 14, we investigate the representation of such groups for one instance, sf(e). There, we find that Legacy and LexiMin represent intersectional groups to similar degrees of accuracy (Figure 3.9), explore factors determining the representation of an intersectional group, and describe how the accuracy of intersectional representation could be improved using our algorithmic framework.

### 3.6.3 Instance-data preprocessing

At the request of practitioners, we pseudonymize the features of each dataset. This does not affect the analysis, as both Legacy and LexiMin are agnostic to this information.

For data from Healthy Democracy (instance hd), of by for* (instance obf) and MASS LBP (instance mass), and for the instance sf(e) from the Sortition Foundation, respondent data and quotas were taken without modification. For privacy reasons, pool members with non-binary gender in the in-

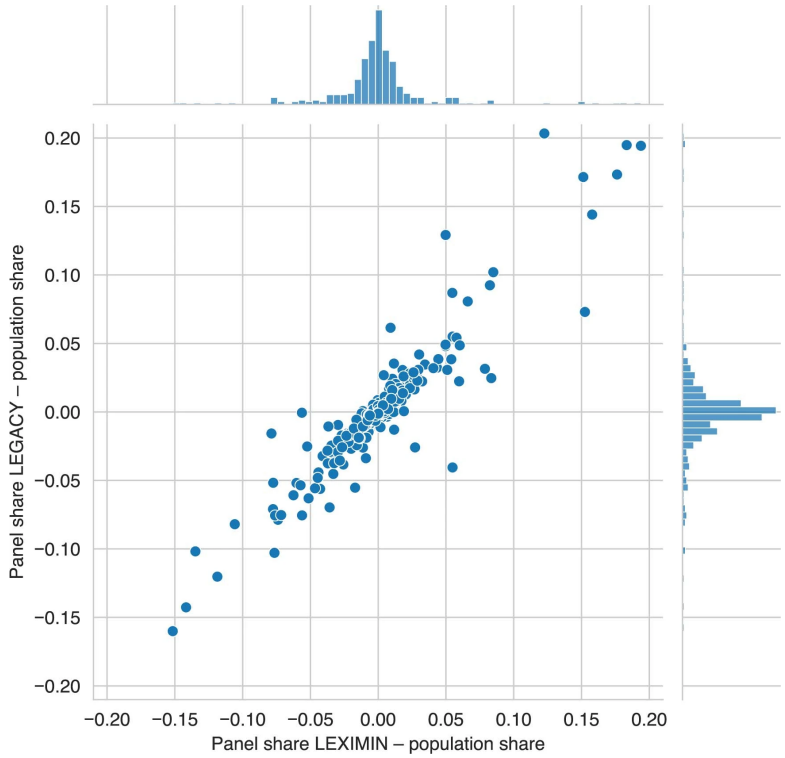



Figure 3.9: For all intersections of two features on the instance sf(e), how far the expected number of group members selected by Legacy or LexiMin differs from the proportional share in the population is shown. Although many intersectional groups are represented close to accurately, some groups are over- and underrepresented by more than 15 percentage points by either algorithm. Which groups get over- and underrepresented is highly correlated between both algorithms. Panel shares are computed for a pool of size 1,727, and population shares are based on a survey with 1,915 respondents after cleaning.

stances sf(a) to sf(d) were randomly assigned female or male gender with equal probability. In two of these instances (sf(a) and sf(d)), the originally used quotas were not recorded in the data, but we reconstructed them according to the procedures of the Sortition Foundation for constructing quotas from the population fractions. The panel from the Center for Climate Assemblies (instance cca) did not formally use upper and lower quotas; instead, exact target values for each feature were given (which could not simultaneously be satisfied) as well as a priority order over which targets were more important than others. We set quotas by identifying the minimal relaxation to the lowest-priority target that could be satisfied. For the Nexus instance (instance nexus), the region of one pool member was missing and inferred from their city of residence. Because Nexus only used lower quotas, the upper quotas of each feature were set to the difference between k and the sum of lower quotas of all other features of the same category. Such a change does not influence the output distribution of either Legacy or LexiMin but makes the ratio product defined in Individuals rarely selected by Legacy above more meaningful. Because Nexus permitted k to range between 170 and 175, we chose 170 to make their lower quotas as tight as possible.

### 3.6.4 Statistics

The selection probabilities of LexiMin are not empirical estimates, but rather exact numbers generated by the algorithm, computed from its output distribution.

By contrast, the selection probabilities given to each agent by Legacy (as used in the numbers in the text and tables) refer to the fraction of 10,000 sampled panels in which the agent appears (in which each sample is from a single run of Legacy on the same instance).

In Figures 3.2, 3.4 and 3.5, when plotting the line representing Legacy, agents are sorted along the $x$ axis in order of this empirical estimate of their selection probability by Legacy, and this is the selection probability given on the $y$ axis. As, for each agent, the number of panels on which they appear across runs of Legacy is distributed as a binomial variable with 10,000 trials and unknown success probability, we indicate Jeffreys intervals for each of these success probabilities (that is, selection probabilities) with 99% confidence. These are confidence intervals on the selection probability of a specific agent, not on the selection probability of a specific percentile of the agents.

In addition to reporting two-sided 99% confidence intervals on each agents selection probability by Legacy, in Table 3.1, we report a 99% confidence upper bound on the minimum selection given to any agent by Legacy per instance. We cannot simply set this upper bound equal to the smallest upper end of the two-sided confidence interval of any agent as computed above because out of these many confidence intervals, some are likely to lie entirely below the true selection probability of the respective agent. Instead, we compute the upper bound on the minimum probability using the confidence interval for a single agent, by running two independent sets of 10,000 samples: In the first set of samples (the one discussed two paragraphs prior), we identify a single agent who was least frequently chosen to the panel in this set; then, we count how often this specific agent is selected across the second set of samples and calculate an upper bound based on a one-sided Jeffreys interval as follows: if the specific agent was selected in s out of the 10,000 panels, the confidence bound is the 99th percentile of the distribution $\beta(1/2+s, 1/2+10,000-s)$. (The bound would be 1 if $s = 10,000$, but this does not happen in any of the instances.) With 99% confidence, this is an upper bound on the selection probability of the specific agent, and thus also an upper bound with 99% confidence on the minimum selection probability.

As the magnitudes of the two-sided confidence intervals in Figures 3.2, 3.4 and 3.5 show, the empirical estimates we get of the selection probabilities of agents by Legacy are likely to be close to their true values. Moreover, two of the three statistics we report are not very sensitive to sampling errors: For Gini inequality, additive errors in the estimate of selection probabilities translate into additive errors in the Gini coefficient; and, when we report the number of agents whose selection probability by Legacy lies under the minimum selection probability of LexiMin, Figures 3.2, 3.4 and 3.5 show that the confidence intervals of most agents lie either below or above this threshold. Therefore, our analysis of Legacy selection probabilities should not be substantially affected by the fact that we can only use empirical estimates of selection probabilities rather than the ground-truth selection probabilities themselves. The one exception is the geometric mean, for which the error in estimating small selection probabilities can severely affect the measure. In particular, in

all instances in which one individual appeared in 0 out of 10,000 sampled panels, the geometric mean of empirical selection probabilities would be 0. Thus, when computing the geometric mean for Legacy in Table 3.6 and in the body, we erred on the side of being generous to Legacy by setting the selection probabilities of these individuals to 1/10,000 instead of 0.

The running times of LexiMin were measured on a 2017 Macbook Pro with a 3.1-GHz dual-core Intel i5 processor. Although the running time should not depend on random decisions in the algorithm, the running time of calls to the optimization library Gurobi depends on how the operating system schedules different threads. Reported times are medians of three runs, and are rounded to the nearest second if below 60 s, or to the nearest minute otherwise.

# 4

# Fairness & Transparency

*Fair Sortition Made Transparent* [131].
Bailey Flanigan, Gregory Kehne, & Ariel D. Procaccia.
*NeurIPS* 2021.

## 4.1 Introduction

In a *citizens' assembly*, a panel of randomly chosen citizens is convened to deliberate and ultimately make recommendations on a policy issue. The defining aspect of citizens' assemblies is the randomness of the process, *sortition*, by which participants are chosen. In practice, the sortition process works as follows: first, volunteers are solicited via thousands of letters or phone calls, which target individuals chosen uniformly at random. Those who respond affirmatively form the *pool* of volunteers, from which a final panel will be chosen. Finally, a *selection algorithm* is used to randomly select some pre-specified number $k$ of pool members for the panel. To ensure adequate representation of demographic groups, the chosen panel is often constrained to satisfy some upper and lower quotas on feature categories such as age, gender, and ethnicity. We call a quota-satisfying panel of size $k$ a *feasible panel*. As this process illustrates, citizens' assemblies offer a way to involve the public in informed decision-making. This potential for civic participation has recently spurred a global resurgence in the popularity of citizens assemblies; they have been commissioned by governments and led to policy changes at the national level [130, 169, 222].

Prompted by the growing impact of citizens' assemblies, there has been a recent flurry of computer scientific research on sortition, and in particular, on the fairness of the procedure by which participants are chosen [46, 128, 130]. The most practicable result to date is a family of selection algorithms proposed by Flanigan et al. [130], which are distinguished from their predecessors by their use of randomness toward the goal of fairness: while previously-used algorithms selected

pool members in a random but ad-hoc fashion, these new algorithms are *maximally fair*, ensuring that pool members have as equal probability as possible of being chosen for the panel, subject to the quotas.[1] To encompass the many interpretations of "as equal as possible," these algorithms permit the optimization of any fairness objective with certain convexity properties. There is now a publicly available implementation of the techniques of Flanigan et al. [130], called *Panelot*, which optimizes the egalitarian notion that no pool member has too little selection probability via the *Leximin* objective from fair division [129, 206]. This algorithm has already been deployed by several groups of panel organizers, and has been used to select dozens of panels worldwide.

Fairness gains in the panel selection process can lend legitimacy to citizens' assemblies and potentially increase their adoption, but only insofar as the public trusts that these gains are truly realized. Currently, the potential for public trust in the panel selection process is limited by multiple factors. First, the latest panel selection algorithms select the final panel via behind-the-scenes computation. When panels are selected in this manner, observers cannot even verify that any given pool member has *any* chance of being chosen for the panel. A second and more fundamental hurdle is that randomness and probability, which are central to the sortition process, have been shown in many contexts to be difficult for people to understand and reason about [196, 240, 281]. Aiming to address these shortcomings, we propose and pursue the following notion of transparency in panel selection:

> *Transparency:* Observers should be able to, without reasoning in-depth about probability, (1) understand the probabilities with which each individual will be chosen for the panel *in theory*, and (2) verify that individuals are actually selected with these probabilities *in practice*.

In this paper, we aim to achieve transparency and fairness simultaneously: this means advancing the defined goal of transparency, while preserving the fairness gains obtained by maximally fair selection algorithms. Although this task is reminiscent of existing AI research on trade-offs between fairness or transparency with other desirable objectives [47, 48, 120, 269], to our knowledge, this is the first investigation of the trade-off between fairness and transparency.

Setting aside for a moment the goal of fairness, we consider a method of random decision-making that is already common in the public sphere: the uniform lottery. To satisfy quotas, a uniform lottery for sortition must randomize not over individuals, but over entire feasible panels. In fact, this approach has been suggested by practitioners, and was even used in 2020 to select a citizens' assembly in Michigan. The following example, which closely mirrors that real-world pilot,[2] illustrates that panel selection via uniform lottery is naturally consistent with the transparency notion we pursue.

Suppose we construct 1000 feasible panels from a pool (possibly with duplicates), numbered 000-999, and publish an (anonymized) list of which pool members are on each panel. We then inform

[1] Quotas can preclude giving individuals exactly equal probabilities: if the panel must be 1/2 men, 1/2 women but the pool is split 3/4 men, 1/4 women, then some women must be chosen more often than some men.

[2] Of By For's pilot of live panel selection via lottery can be viewed at https://vimeo.com/458304880#t=17m59s from 17:59 to 21:23. For a more detailed description, see Figure 3 and surrounding text in [130].

spectators that we will choose each panel with equal probability. This satisfies criterion (1): spectators can easily understand that all panels will be chosen with the same probability of 1/1000, and can easily determine each individual's selection probability by counting the number of panels containing the individual. To satisfy criterion (2), we enact the lottery by drawing each of the three digits of the final panel number individually from lottery machines. Lottery spectators can confirm that each ball is drawn with equal probability; this provides confirmation that panels are indeed being chosen with uniform probabilities, thus confirming the enactment of the proposed individual selection probabilities. In addition to its conventionality as a source of randomness, decision-making via drawing lottery balls invites an exciting spectacle, which can promote engagement with citizens' assemblies.

This simple method neatly satisfies our transparency criteria, but it has one obvious downside: a uniform lottery over an arbitrary set of feasible panels does not guarantee any measure of equal probabilities to individuals. In fact, it is not even clear that the *fairest possible* uniform lottery over $m$ panels, where $m$ is a number conducive to selection by physical lottery (e.g. $m$ =1000), would not be significantly less fair than maximally fair algorithms, which sample the fairest possible unconstrained distribution over panels. For example, if $m$ is too small, there may be *no* uniform lottery which gives all individuals non-zero selection probability, even if each individual appears on some feasible panel (and so can attain a non-zero selection probability under an unconstrained distribution).

Fortunately, empirical evidence suggests that there is hope: in the 2020 pilot mentioned above, a uniform lottery over $m$ =1000 panels was found that nearly matched the fairness of the maximally fair distribution generated by Panelot. Motivated by this anecdotal evidence, we aim to understand whether such a fair uniform lottery is guaranteed to exist in general, and if it does, how to find it. We summarize this goal in the following research questions:

> *Does there exist a uniform lottery over m panels that nearly preserves the fairness of the maximally fair unconstrained distribution over panels?* And,
> *Algorithmically, how do we compute such a uniform lottery?*

**Results and Contributions.** After describing the model in Section 4.2, in Section 4.3 we prove that it is possible to round an (essentially) arbitrary distribution over panels to a uniform lottery while preserving *all* individuals' selection probabilities up to only a small bounded deviation. These results use tools from discrepancy theory and randomized rounding. Intuitively, this bounded change in selection probabilities implies bounded losses in fairness; we formalize this intuition in Section 4.4, showing that there exists in general a uniform lottery that is nearly maximally fair, with respect to multiple choices of fairness objective. Although we would ideally like to give such bounds for the *Leximin* fairness objective, due to its use practice, we cannot succinctly represent bounds for this objective because it is not scalar valued. We therefore give bounds for *Maximin*, a closely related egalitarian objective which only considers the minimum selection probability given to any pool member [81]. We discuss in Section 4.4 why bounds on loss in Maximin fairness are, in the most meaningful sense, also bounds on loss in Leximin fairness. We additionally give upper bounds on the loss in *Nash Welfare* [206], a similarly well-established

fairness objective that has also been implemented in panel selection tools [163].

Finally, in Section 4.5, we consider the algorithmic question in practice: given a maximally fair distribution over panels, can we actually *find* nearly maximally fair uniform lotteries that match our theoretical guarantees? To answer this question, we implement two standard rounding algorithms, along with near-optimal (but more computationally intensive) integer programming methods, for finding uniform lotteries. We then evaluate the performance of these algorithms in 11 real-world panel selection instances. We find that in all instances, we can compute uniform lotteries that nearly exactly preserve not only fairness with respect to both objectives, but *entire sets* of Leximin-optimal marginals, meaning that from the perspective of individuals, there is essentially no difference between using a uniform lottery versus the optimal unconstrained distribution sampled by the latest algorithms. We discuss these results, their implications, and how they can be deployed directly into the existing panel selection pipeline in Section 4.6.

## 4.2 Model

Panel Selection Problem. First, we formally define the task of panel selection for citizens' assemblies. Let $N = [n]$ be the *pool* of volunteers for the panel—individuals from the population who have indicated their willingness to participate in response to an invitation. Let $F = \{f_t\}_t$ denote a fixed set of *features* of interest. Each feature $f_t : N \to \Omega_t$ maps each pool member to their value of that feature, where $\Omega_t$ is the set of $f_t$'s possible values. For example, for feature $f_t$ = "gender", we might have $\Omega_t$ = {"male","female", "non-binary"}. We define individual $i$'s *feature vector* $F(i) = (f_t(i))_t \in \prod_t \Omega_t$ to be the vector encoding their values for all features in $F$.

As is done in practice and in previous research [128, 130], we impose that the chosen panel $P$ must be a subset of the pool of size $k$, and must be representative of the broader population with respect to the features in $F$. This representativeness is imposed via *quotas:* for each feature $f$ and corresponding value $v \in \Omega$, we may have lower and upper quotas $l_{f,v}$ and $u_{f,v}$. These quotas require that the panel contain between $l_{f,v}$ and $u_{f,v}$ individuals $i$ such that $f(i) = v$.

In terms of these parameters, we define an instance of the panel selection problem as: given $(N, k, F, l, u)$—a pool, panel size, set of features, and sets of lower and upper quotas—randomly select a *feasible panel*, where a feasible panel is any set of individuals $P$ from the collection $\mathcal{K}$:

$$\mathsf{K} := \left\{P \in \binom{N}{k} : \; l_{f,v} \leq |\{i \in P : f(i) = v\}| \leq u_{f,v} \text{ for all } f, v\right\}.$$

Maximally Fair Selection Algorithms. A *selection algorithm* is a procedure that solves instances of the panel selection problem. A selection algorithm's level of fairness on a given instance is determined by its *panel distribution* $p$, the (possibly implicit) distribution over $\mathcal{K}$ from which it draws the final panel. Because we care about fairness to individual pool members, we evaluate the fairness of $p$ in terms of the fairness of selection probabilities, or *marginals*, that $p$ implies for all pool members.[1] We denote the vector of marginals implied by $p$ as $\pi$, and we

[1] A panel distribution $p$ implies a unique vector of marginals $\pi$ as follows: fixing $p$, $\pi$, a pool member $i$'s marginal selection probability $\pi_i$ is equal to the probability of drawing a panel from $p$ containing that pool member. For a

will sometimes specify a panel distribution as $p$, $\pi$ to explicitly denote this pair. We say that $\pi$ is *realizable* if it is implied by some distribution $p$ over the feasible panels $\mathsf{K}$.

*Maximally fair* selection algorithms are those which solve the panel selection problem by sampling a specifically chosen $p$: one which implies marginals $\pi$ that allocate probability as fairly as possible across pool members. The fairness of $p, \pi$ is measured by a *fairness objective* $\mathcal{F}$, which maps an allocation—in this case, of selection probability to pool members—to a real number measuring the allocation's fairness. Fixing an instance, a fairness objective $\mathcal{F}$, and a panel distribution $p$, we express the fairness of $p$ as $\mathcal{F}(p)$. Existing maximally fair selection algorithms can maximize a wide range of fairness objectives, including those considered in this paper.

Leximin, Maximin, and Nash Welfare. Of the three fairness objectives we consider in this paper, Maximin and Nash Welfare (NW) have succinct formulae. For $p, \pi$ they are defined as follows, where $\pi_i$ is the marginal of individual $i$:

$$\text{Maximin}(p) := \min_{i \in N} \pi_i, \qquad \text{NW}(p) := \left( \prod_i \pi_i \right)^{1/n}.$$

Intuitively, NW maximizes the geometric mean, prioritizing the marginal $\pi_i$ of each individual $i$ in proportion to $\pi_i^{-1}$. Maximin maximizes the marginal probability of the individual least likely to be selected. Finally, Leximin is a refinement of Maximin, and is defined by the following algorithm: first, optimize Maximin; then, fixing the minimum marginal as a lower bound on any marginal, maximize the second-lowest marginal; and so on.

Our task: quantize a maximally fair panel distribution with minimal fairness loss. We define a $1/m$-*quantized* panel distribution as a distribution over all feasible panels $\mathcal{K}$ in which all probabilities are integer multiples of $1/m$. We use $\bar{p}$ to denote a panel distribution with this property. Formally, while an (unconstrained) panel distribution $p$ lies in $\mathcal{D} := \{p \in \mathsf{R}_+^{|\mathsf{K}|} : \|p\|_1 = 1\}$, a $1/m$-quantized panel distribution in $\bar{p}$ lies in $\overline{\mathcal{D}} := \{\bar{p} \in (\mathbb{Z}_+/m)^{|\mathsf{K}|} : \|\bar{p}\|_1 = 1\}$. Note that a $1/m$-quantized distribution $\bar{p}$ immediately translates to a physical uniform lottery of over $m$ panels (with duplicates): if $\bar{p}$ assigns probability $\ell/m$ to panel $P$, then the corresponding physical uniform lottery would contain $\ell$ duplicates of $P$. Thus, if we can compute a $1/m$-quantized panel distribution $\bar{p}$ with fairness $\mathcal{F}(\bar{p})$, then we have designed a physical uniform lottery over $m$ panels with that same level of fairness.

Our goal follows directly from this observation: we want to show that given an instance and desired lottery size $m$, we can compute a $1/m$-quantized distribution $\bar{p}$ that is nearly as fair, with respect to a fairness notion $\mathcal{F}$, as the maximally fair panel distribution in this instance $p^* \in \arg\max_{p \in \mathcal{D}} \mathcal{F}(p)$. We define the *fairness loss* in this quantization process to be the difference $\mathcal{F}(p^*) - \mathcal{F}(\bar{p})$. We are aided in this task by the existence of practical algorithms for computing $p^*$ [130], which allows us to use $p^*$ as an input to the quantization procedure we hope to design. For intuition, we illustrate this quantization task in Figure 4.1, where $\pi^*$, $\bar{\pi}$ are the marginals implied

more detailed introduction to the connection between panel distributions and marginals, we refer readers to [130].

by $p^*, \bar{p}$, respectively. Since the fairness of $p^*, \bar{p}$ are computed in terms of $\pi^*$, $\bar{\pi}$, it is intuitive that a quantization process that results in small *marginal discrepancy*, defined as the maximum change in any marginal $\|\pi - \bar{\pi}\|_\infty$, should also have small fairness loss. This idea motivates the upcoming section, in which we give quantization procedures with provably bounded marginal discrepancy, forming the foundation for our later bounds on fairness loss.

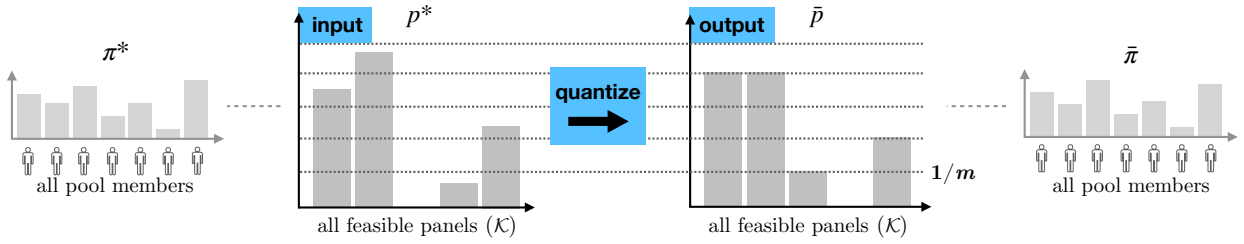



Figure 4.1: The quantization task takes as input a maximally fair panel distribution $p^*$ (implying marginals $\pi^*$), and outputs a $1/m$-quantized panel distribution $\bar{p}$ (implying marginals $\bar{\pi}$).

## 4.3 Theoretical Bounds on Marginal Discrepancy

Here we prove that for a fixed panel distribution $p, \pi$, there exists a uniform lottery $\bar{p}, \bar{\pi}$ such that $\|\pi - \bar{\pi}\|_\infty$ is bounded. Preliminarily, we note that it is intuitive that bounds on this discrepancy should approach 0 as $m$ becomes large with respect to $n$ and $k$. To see why, begin by fixing some distribution $p, \pi$ over panels: as $m$ becomes large, we approach the scenario in which a uniform lottery $\bar{p}$ can assign panels arbitrary probabilities, providing increasingly close approximations to $p$. Since the marginals $\pi_i$ are continuous with respect to $p$, as $\bar{p} \to p$ we have that $\bar{\pi}_i \to \pi_i$ for all $i$.

While this argument demonstrates convergence, it provides neither efficient algorithms nor tight bounds on the rate of convergence. In this section, our task is therefore to bound the rate of this convergence as a function of $m$ and the other parameters of the instance. All omitted proofs of results from this section are included in appendix C.2.

### 4.3.1 Worst-Case Upper Bounds

Our first set of upper bounds result from rounding Standard LP, the LP that most directly arises from our problem. This LP is defined in terms of a panel distribution $p, \pi$, and $M$, an $n \times |\mathcal{K}|$ matrix describing which individuals are on which feasible panels: $M_{i,P} = 1$ if $i \in P$ and $M_{i,P} = 0$ otherwise.

Standard LP

$$Mp = \pi \tag{4.1}$$

$$\|p\|_1 = 1 \tag{4.2}$$

$$p \geq 0.$$

Here, (4.1) specifies $n$ total constraints. Our goal is to round $p$ to a uniform lottery $\bar{p}$ over $m$ panels (so the entries $\bar{p}$ are multiples of $1/m$) such that (4.2) is maintained exactly, and no constraint in (4.1) is relaxed by too much, i.e., $\|Mp - M\bar{p}\|_\infty = \|\pi - \bar{\pi}\|_\infty$ remains small.

Randomized rounding is a natural first approach. Any randomized rounding scheme satisfying negative association (which includes several that respect (4.2)) yields the following bound: For any realizable $\pi$, we may efficiently randomly generate $\bar{p}$ such that its marginals $\bar{\pi}$ satisfy

$$\|\pi - \bar{\pi}\|_\infty = O\left(\frac{\sqrt{n \log n}}{m}\right).$$

Fortunately, there is potential for improvement: randomized rounding does not make full use of the fact that $M$ is $k$-column sparse, due to each panel in $\mathsf{K}$ containing exactly $k$ individuals. We use this sparsity to get a stronger bound when $n \gg k^2$, which is a practically significant parameter regime. The proof applies a dependent rounding algorithm based on a theorem of Beck and Fiala [40], to which a modification ensures the exact satisfaction of constraint (4.2). For any realizable $\pi$, we may efficiently construct $\bar{p}$ such that its marginals $\bar{\pi}$ satisfy

$$\|\pi - \bar{\pi}\|_\infty \leq k/m.$$

This bound is already meaningful in practice, where $k \ll m$ is insured by the fact that $m$ is pre-chosen along with $k$ prior to panel selection. Note also that $k$ is typically on the order of 100 (Appendix D.4.1), whereas a uniform lottery can in practice be easily made orders of magnitude larger, as each additional factor of 10 in the size of the uniform lottery requires drawing only one more ball (and there is no fairness cost to drawing a larger lottery, since increasing $m$ allows for uniform lotteries which better approximate the unconstrained optimal distribution).

### 4.3.2 Beyond-Worst-Case Upper Bounds

As we will demonstrate in section 4.3.3, we cannot hope for a better worst-case upper bound than $\text{poly}(k)/m$. We thus shift our consideration to instances which are "simple" in their feature structure, having a small number of features (theorem C.2.7), a limited number of unique feature vectors in the pool (section 4.3.2), or multiple individuals that share each feature vector present (theorem C.2.8). The beyond-worst-case bounds given by section 4.3.2 and theorem C.2.8 asymptotically dominate our worst-case bounds in section 4.3.1 and section 4.3.1, respectively. Moreover, section 4.3.2 dominates all other upper bounds in 10 of the 11 practical instances studied in section 4.5.

We note that while our worst-case upper bounds implied the near-preservation of *any* realizable set of marginals $\pi$, some of our beyond-worst-case results apply to only realizable $\pi$ which are *anonymous*, meaning that $\pi_i$ are equal for all $i$ with equal feature vectors. We contend that any reasonable set of marginals should have this property,[1] and furthermore that the "anonymization"

[1]The class of all anonymous marginals $\pi$ includes the maximizers $\pi^*$ of all reasonable fairness objectives, and second, this condition is satisfied by all existing selection algorithms used in practice, to our knowledge.

of any realizable $\pi$ is also realizable (claim C.2.6); hence this restriction is insignificant. Our beyond-worst-case bounds also differ from our worst-case bounds in that they depart from the paradigm of rounding $p$, instead randomizing over panels that may fall outside the support of $p$.

The main beyond-worst-case bound we give, stated below, is parameterized by $|C|$, where $C$ is the set of unique feature vectors that appear in the pool. All omitted proofs and other beyond worst-case results are stated and proven in appendix C.2.

If $\pi$ is anonymous and realizable, then we may efficiently construct $\bar{p}$ such that its marginals $\bar{\pi}$ satisfy

$$\|\pi - \bar{\pi}\|_\infty = O\left(\frac{\sqrt{|C| \log |C|}}{m}\right).$$

$|C|$ is at most $n$, so this bound dominates section 4.3.1. In 10 of the 11 real-world instances we study, $|C|$ is also smaller than $k^2$ (appendix C.1), in which case this bound also dominates section 4.3.1.

At a high level, our beyond-worst-case upper bounds are obtained not by directly rounding $p$, but instead using the structure of the sortition instance to abstract the problem into one about "types." For this bound we then solve an LP in terms of "types," round that LP, and then reconstruct a rounded panel distribution $\bar{p}, \bar{\pi}$ from the "type" solution. In particular, the *types* of individuals are the feature vectors which appear in the pool, and *types* of panels are the multisets of $k$ feature vectors that satisfy the instance quotas. Fixing an instance, we project some $p$ into type space by viewing it as a distribution $\mathfrak{p}$ over types of panels $\mathfrak{K}$, inducing marginals $\tau_c$ for each type individuals $c \in C$.

To begin, we define the Type LP, which is analogous to eq. (4.1). We let $Q$ be the type analog of $M$, so that entry $Q_{cj}$ is the number of individuals $i$ with $F(i) = c$ contained in panels of type $j \in \mathfrak{K}$.[1] Then,

Type LP

$$Q\,\mathfrak{p} = \tau \tag{4.3}$$

$$\|\mathfrak{p}\|_1 = 1 \tag{4.4}$$

$$\mathfrak{p} \geq 0.$$

We round $\mathfrak{p}$ in this LP to a panel type distribution $\bar{\mathfrak{p}}$ while preserving (4.4). All that remains, then, is to construct some $\bar{p}, \bar{\pi}$ such that $p$ is consistent with $\bar{\mathfrak{p}}$ and $\|\pi - \bar{\pi}\|_\infty$ is small. This $\bar{p}$ is in general supported by panels outside of $supp(p)$, unlike the $\bar{p}$ obtained by section 4.3.1. It is the anonymity of $\pi$ which allows us to construct these new panels and prove that they are feasible for the instance.

### 4.3.3 Lower Bounds

This method of using bounded discrepancy to derive nearly fairness-optimal uniform lotteries has its limits, since there are even sparse $M$ and fractional $x$ for which no integer $\bar{x}$ yields nearby

[1] Completing the analogy, $C, \mathfrak{K}, Q, \mathfrak{p}, \bar{\mathfrak{p}}, \tau$ are the "type" versions of $N, \mathcal{K}, M, p, \bar{p}, \pi$ from the original LP.

$M\bar{x}$. In the worst case, we establish lower bounds by modifying those of Beck and Fiala [253]: There exist $p, \pi$ for which for all uniform lotteries $\bar{p}, \bar{\pi}$,

$$\min_{\bar{p} \in \overline{\mathcal{D}}} \|\pi - \bar{\pi}\|_\infty = \Omega\left(\frac{\sqrt{k}}{m}\right).$$

Our $k$-dependent upper and lower bounds are separated by a factor of $\sqrt{k}$, matching the current upper and lower bounds of the Beck-Fiala conjecture as applied to linear discrepancy (also known as the lattice approximation problem [254]). The respective gaps are incomparable, however, since for a given $x \in [0,1]^n$, the former problem aims to minimize $\|M(x - \bar{x})\|_\infty$ over $\bar{x} \in \{0,1\}^n$, while we aim to do the same over a subset of the $\bar{x} \in \mathbb{Z}^n$ for which $\sum_j x_j = \sum_j \bar{x}_j$ (see lemma C.2.4).

## 4.4 Theoretical Bounds on Fairness Loss

Since the fairness of a distribution $p$ is determined by its marginals $\pi$, it is intuitive that if uniform lotteries incur only small marginal discrepancy (per section 4.3), then they should also incur only small fairness losses. This should hold for any fairness notion that is sufficiently "smooth" (i.e., doesn't change too quickly with changing marginals) in the vicinity of $p, \pi$.

Although our bounds from section 4.3 apply to any reasonable initial distribution $p$, we are particularly concerned with bounding fairness loss from *maximally fair* initial distributions $p^*$. Here, we specifically consider such $p^*$ that are optimal with respect to Maximin and NW. We note that, since there exist anonymous $p^*, \pi^*$ that maximize these objectives, we can apply any upper bound from section 4.3 to upper bound $\|\pi^* - \bar{\pi}\|_\infty$. We defer omitted proofs to appendix C.3.

### 4.4.1 Maximin

Since Leximin is the fairness objective optimized by the maximally fair algorithm used in practice, it would be most natural to start with a $p^*$ that is Leximin-optimal and bound fairness loss with respect to this objective. However, the fact that Leximin fairness cannot be represented by a single scalar value prevents us from formulating such an approximation guarantee. Instead, we first pursue bounds on the closely-related objective, Maximin. We argue that in the most meaningful sense, a worst-case Maximin guarantee *is* a Leximin guarantee: such a bound would show limited loss in the minimum marginal, and it is Leximin's *lexicographically first priority* to maximize the minimum marginal.

First, we show there exists some $\bar{p}, \bar{\pi}$ that gives bounded Maximin loss from $p^*, \pi^*$, the Maximin-optimal unconstrained distribution. This bound follows from Theorems 4.3.2 and C.2.8, using the simple observation that $\bar{p}$ can decrease the lowest marginal given by $p^*$ by no more than $\|\pi^* - \bar{\pi}\|_\infty$. Here $n_{min} := \min_c n_c$ denotes the smallest number of individuals which share any feature vector $c \in \mathsf{C}$.

By section 4.3.2 and C.2.8, for Maximin-optimal $p^*$, there exists a uniform lottery $\bar{p}$ that satisfies

$$\text{Maximin}(p^*) - \text{Maximin}(\bar{p}) = \frac{1}{m} \cdot O\left(\min\left\{\sqrt{|\mathsf{C}| \log |\mathsf{C}|},\ \frac{k}{n_{min}} + 1\right\}\right).$$

Section 4.3.3 demonstrates that we cannot get an upper bound on Maxmin loss stronger than $O(\sqrt{k}/m)$ using a uniform bound on changes in all $\pi_i$. However, since Maximin is concerned only with the smallest $\pi_i$, it seems plausible that better upper bounds on Maximin loss could result from rounding $\pi$ while tightly controlling only losses in the smallest $\pi_i$'s, while giving freer reign to larger marginals. We show that this is not the case by further modifying the instances from section 4.3.3 to obtain the following lower bound on the Maximin loss: There exists a Maximin-optimal $p^*$ such that, for all uniform lotteries $\bar{p}$,

$$\text{Maximin}(p^*) - \text{Maximin}(\bar{p}) = \Omega\left(\frac{\sqrt{k}}{m}\right).$$

### 4.4.2 Nash Welfare

As NW has also garnered interest by practitioners and is applicable in practice [163], we upper-bound the NW fairness loss. Unlike Maximin loss, an upper bound on NW loss does not immediately follow from one on $\|\pi - \bar{\pi}\|_\infty$, because decreases in smaller marginals have larger negative impact on the NW. As a result, the upper bound on NW resulting from section 4.3 is slightly weaker than that on Maximin:

For NW-optimal $p^*$, there exists a uniform lottery $\bar{p}$ that satisfies

$$\text{NW}(p^*) - \text{NW}(\bar{p}) = \frac{k}{m} \cdot O\left(\min\left\{\sqrt{|\mathsf{C}| \log |\mathsf{C}|},\ \frac{k}{n_{min}} + 1\right\}\right).$$

We give an overview of the proof of section 4.4.2. To begin, fix a NW-optimizing panel distribution $p^*, \pi^*$. Before applying our upper bounds on marginal discrepancy from section 4.3, we must contend with the fact that if this bounded loss is suffered by already-tiny marginals, the NW may decrease substantially or even go to 0. Thus, we first prove Lemmas 4.4.2 and 4.4.2, which together imply that no marginal in $\pi^*$ is smaller than $1/n$.

For NW-optimal $p^*$ over a support of panels $supp(p^*)$, there exists a constant $\lambda \in \mathsf{R}^+$ such that, for all $P \in supp(p^*)$, $\sum_{i \in P} 1/\pi_i^* = \lambda$.

For NW-optimal $p^*, \pi^*$, we have that $\pi_i^* \geq 1/n$ for all $i \in N$.

Section 4.4.2 follows from the fact that the partial derivative of NW with respect to the probability it assigns a given panel must be the same as that with respect to any other panel at $p^*$ (otherwise,

mass in the distribution could be shifted to increase the NW). Section 4.4.2 then follows by the additional observation that $\mathbb{E}_{P\sim p^*}\left[\sum_{i\in P} 1/\pi_i^*\right] = n$.

Finally section 4.4.2 follows from the fact that Section 4.4.2 limits the potential multiplicative, and therefore additive, impact on the NW of decreasing any marginal by $\|\pi - \bar{\pi}\|_\infty$:

For NW-optimal $p^*, \pi^*$, there exists a uniform lottery $\bar{p}, \bar{\pi}$ that satisfies $\mathrm{NW}(p^*) - \mathrm{NW}(\bar{p}) \leq k\,\|\pi^* - \bar{\pi}\|_\infty$. As the NW-optimal marginals $\pi^*$ are anonymous, we can apply the upper bounds given by section 4.3.2 and theorem C.2.8 to show the existence of a $\bar{p}, \bar{\pi}$ satisfying the claim of the theorem.

## 4.5 Practical Algorithms for Computing Fair Uniform Lotteries

**Algorithms.** First, we implement versions of two existing rounding algorithms, which are implicit in our worst-case upper bounds.[1] The first is Pipage rounding [141], or Pipage, a randomized rounding scheme satisfying negative association [102]. The second is Beck-Fiala, the dependent rounding scheme used in the proof of section 4.3.1. To benchmark these algorithms against the highest level of fairness they could possibly achieve, we use integer programming (IP) to compute the fairest possible uniform lotteries over $supp(p^*)$, the panels over which $p^*$ randomizes.[2] We define IP-Maximin and IP-NW to find uniform lotteries over $supp(p^*)$ maximizing Maximin and NW, respectively. We remark that the performance of these IPs is still subject to our theoretical upper and lower bounds. We provide implementation details in appendix C.4.1.

One question is whether we should prefer the IPs or the rounding algorithms for real-world applications. Although IP-Maximin appears to find good solutions at practicable speeds, IP-NW converges to optimality prohibitively slowly in some instances (see appendix C.4.2 for runtimes). At the same time, we find that our simpler rounding algorithms give near-optimal uniform lotteries with respect to both fairness objectives. Also in favor of simpler rounding algorithms, many randomized rounding procedures (including Pipage rounding) have the advantage that they exactly preserve marginals over the combined steps of randomly rounding to a uniform lottery and then randomly sampling it—a guarantee that is much more challenging to achieve with IPs.

**Uniform lotteries nearly exactly preserve Maximin, Nash Welfare fairness.** We first measure the fairness of uniform lotteries produced by these algorithms in 11 real-world panel selection instances from 7 different organizations worldwide (instance details in appendix C.1). In all experiments, we generate a lottery of size $m = 1000$. This is fairly small; it requires drawing only 3 balls from lottery machines, and in one instance we have that $m < n$. We nevertheless see excellent performance of all algorithms, and note that this performance will only improve with larger $m$.

[1] We do not implement the algorithm implicit in section 4.3.2 because our results already present sufficient alternatives for finding excellent uniform lotteries in practice.

[2] Note that these lotteries are not necessarily universally optimal, as they can randomize over only $supp(p^*)$; conceivably, one could find a fairer uniform lottery by also randomizing over panels not in $supp(p^*)$. However, Pipage and Beck-Fiala are also restricted in this way, and thus must be weakly dominated by the IP.

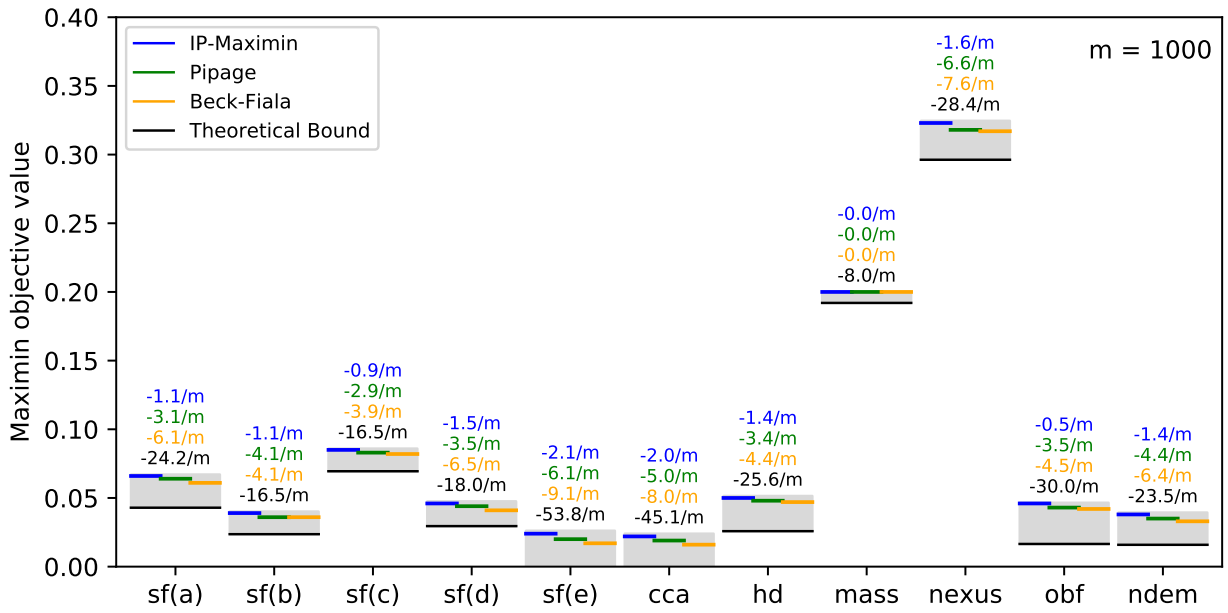



Figure 4.2: $m = 1000$. Shaded regions extend from Maximin($p^*$), the fairness of the optimal unconstrained distribution, down to the minimum fairness implied by the tightest theoretical upper bound in that instance (in all instances but "obf" section 4.3.2 is tightest). Each algorithm or bound's loss relative to Maximin($p^*$) is written above in the corresponding color. We show a representative run of Pipage, a randomized algorithm.

Figure 4.2 shows the Maximin fairness of the uniform lottery computed by Pipage, Beck-Fiala, and IP-Maximin for each instance. For intuition, recall that the level of Maximin fairness given by any lottery is exactly the minimum marginal assigned to any individual by that lottery. The upper edges of the gray boxes in fig. 4.2 correspond to the optimal fairness attained by an unconstrained distribution $p^*$. These experiments reveal that the cost of transparency to Maximin-fairness is practically non-existent: across instances, the quantized distributions computed by IP-Maximin decrease the minimum marginal by at most $2.1/m$, amounting to a loss of no more than 0.0021 in the minimum marginal probability in any instance. Visually, we can see that this loss is negligible relative to the original magnitude of even the smallest marginals given by $p^*$. Surprisingly, though Pipage and Beck-Fiala do not aim to optimize any fairness objective, they achieve only slightly larger losses in Maximin fairness, with Pipage outperforming Beck-Fiala. Finally, the heights of the gray boxes indicate that our theoretical bounds are often meaningful in practice, giving lower bounds on Maximin fairness well above zero in nine out of eleven instances. We note these bounds only tighten with larger $m$. We present similarly encouraging results on NW loss in appendix C.4.3.

**Uniform lotteries nearly preserve all Leximin marginals.** We still remain one step away from practice: our examination of Maximin does not address whether uniform lotteries can attain the finer-tuned fairness properties of the Leximin-optimal distributions currently used in prac-

tice. Fortunately, our results from section 4.3 imply the existence of a quantized $\bar{p}$ that closely approximates *all marginals* given by the Leximin-optimal distribution $p^*, \pi^*$. We evaluate the extent to which Pipage and Beck-Fiala preserve these marginals in fig. 4.3. They are benchmarked against a new IP, IP-Marginals, which computes the uniform lottery over $supp(p^*)$ minimizing $\|\pi^* - \bar{\pi}\|_\infty$.

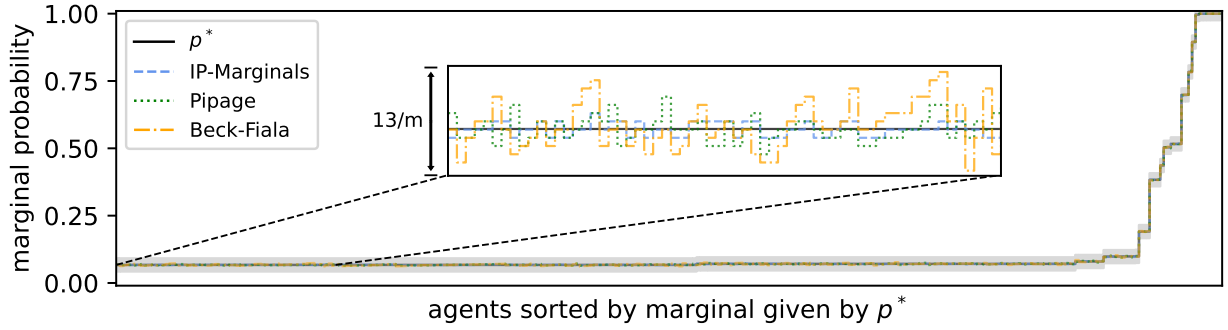



Figure 4.3: Instance = sf(a), $m = 1000$. Line plot shows the Leximin-optimal marginals $\pi^*$ (implied by panel distribution $p^*$), along with marginals given by all algorithms sorted according to $\pi^*$. Note that each $x$ coordinate then corresponds to an individual. The zoomed box shows the magnitude of marginal discrepancy around $\pi^*$. The surrounding shaded region shows the tightest theoretical bound on the marginal discrepancy, in this case from section 4.3.2, around the optimal marginals. We show a representative run of Pipage, a randomized algorithm.

Figure 4.3 demonstrates that in the instance "sf(a)", all algorithms produce marginals that deviate negligibly from those given by $\pi^*$. Analogous results on remaining instances appear in appendix C.4.4 and show similar results. As was the case for Maximin, we see that our theoretical bounds are meaningful, but that we can consistently outperform them in real-world instances.

## 4.6 Discussion

Our aim was to show that uniform lotteries can preserve fairness, and our results ultimately suggest this, along with something stronger: that in practical instances, uniform lotteries can reliably almost exactly replicate *the entire set of marginals* given by the optimal unconstrained panel distribution. Our rounding algorithms can thus be plugged directly into the existing panel selection pipeline with essentially no impact on individuals' selection probabilities, thus enabling translation of the output of Panelot (and other maximally fair algorithms) to a nearly maximally fair *and* transparent panel selection procedure. We note that our methods are not just compatible with ball-drawing lotteries, but any form of uniform physical randomness (e.g. dice, wheel-spinning, etc.).

Although we achieve our stated notion of transparency, a limitation of this notion is that it focuses on the final stage of the panel selection process. A more holistic notion of transparency might require that onlookers can verify that the panel is not being intentionally stacked with certain

individuals. This work does not fully enable such verification: although onlookers can now observe individuals' marginals, they still cannot verify that these marginals are *actually maximally fair* without verifying the underlying optimization algorithms. In particular, in the common case where quotas require even maximally fair panel distributions to select certain individuals with probability near one, onlookers cannot distinguish those from unfair distributions engineered such that one or more pool members are chosen with probability near one.

In research on economics, fair division, and other areas of AI, randomness is often proposed as a tool to make real-world systems fairer [63, 139, 157]. Nonetheless, in practice, these systems (with a few exceptions, such as school choice [213]) remain stubbornly deterministic. Among the hurdles to bringing the theoretical benefits of randomness into practice is that allocation mechanisms fare best when they can be readily understood, and that randomness can be perceived as undesirable or suspect. Sortition is a rather unique paradigm at the heart of this tension: it relies centrally on randomness, while in the public sphere it is attaining increasing political influence. It is therefore a uniquely high-impact domain in which to study how to combine the benefits of randomness, such as fairness, with transparency. We hope that this work and its potential for impact will inspire the investigation of fairness-transparency tradeoffs in other AI applications.

**Acknowledgements.** We would foremost like to thank Paul Gölz for helpful technical conversations and insights on the practical motivations for this research. We also thank Anupam Gupta for helpful technical conversations. Finally, several organizations for supplying real-world citizens' assembly data, including the Sortition Foundation, the Center for Climate Assemblies, Healthy Democracy, MASS LBP, Nexus Institute, Of by For, and New Democracy.

# 5

# Manipulation-Robustness

*Manipulation-Robust Citizens' Assembly Selection* [128].
Bailey Flanigan, Jennifer Liang, Ariel D. Procaccia, & Sven Wang.
*AAAI* 2024.

## 5.1 Introduction

In a *citizens' assembly*, a panel of randomly-chosen constituents convenes to make a policy recommendation on a political issue. Although citizens' assembly participants are not career politicians, their recommendations are informed by an extensive process of learning from experts and deliberating with one another. As such, citizens' assemblies are appealing because they combine the goals of engaging everyday citizens in democratic decision-making, while also facilitating informed decisions. Citizens' Assemblies are now being used to make increasingly high-profile decisions around the world [225]; for example, France recently ran a national-level assembly on the topic of assisted dying, and its outcome is slated to affect policy on palliative care [65].

Because the participants of a citizens' assembly represent their entire underlying constituency, the process by which they are selected is crucial to whether the policy recommendation they produce is perceived as trustworthy. The importance of this selection process has motivated a growing body of research on *selection algorithms* [104, 106, 128, 130, 131], which solve the following task: from among a *pool* of volunteers, randomly sample a *panel* that is (at least approximately) *descriptively representative* of the underlying population. This means that if the population is 48% women, the panel should be approximately 48% women. Because exact representation of all identities cannot be achieved with a finite-size panel, practitioners' main goal is to achieve representation with respect to a handful of key features, such as gender, age, geographic location, education level, and opinion on the issue at hand.

The main algorithmic challenge in selecting descriptively representative participants is *self-selection bias*: different demographic groups agree to participate at vastly different rates, so the pool of volunteers from which the panel is sampled is demographically skewed compared to the underlying population. Consequently, simple sampling techniques do not produce the desired descriptive representation.

Existing work has circumvented the challenge of achieving representation to a large degree. The first selection algorithms, developed by practitioners, were heuristics that searched for representative panels, injecting randomness wherever possible. More recent work has contributed algorithms that not only find representative panels, but do so in a way that achieves other desiderata simultaneously. For example, Flanigan et al. [130] presents a framework of algorithms that are *maximally fair* to individual pool members: that is, they make pool members' probabilities of being selected as equal as possible, subject to representation constraints. One algorithm within this framework, called *Leximin* [130], is now widely used in practice.

Beyond the desiderata of representation and maximal fairness, follow-up work has contributed methods for additionally achieving *transparency* [131]. However, at the current frontier of research on selection algorithms, a key desideratum remains yet untouched: their *manipulability*.

In this paper, we initiate the study of selection algorithms' vulnerability to perhaps the most salient type of potential manipulation: *volunteers misreporting their features*. With Example 5.1.1, we now illustrate in detail why the selection process, as it commonly works in practice, can permit — and strongly incentivize — such manipulation.

**Example 5.1.1.** *We want to select a panel of 10 people to convene on climate policy. We care about descriptive representation of one feature only: people's level of concern about climate change. This feature has two possible values: those who are less concerned (20% of the population) and more concerned (80% of the population). Thus, we will reserve 2 and 8 panel seats for these respective groups.*

*Stage 1: Recruiting the pool of volunteers. We send out invitations to 1000 uniformly sampled households in our constituency. In response, 100 people volunteer to participate, but they are strongly self-selected: only 4 are truly less concerned, and 96 of them are truly more concerned.*[1] *In preparation for selection, we ask all 100 volunteers to report which group they belong to. Among these volunteers, suppose there is one strategic agent $i$ who is truly more concerned, but is willing to misreport their group membership if it increases their chance of being on the panel.*

*Stage 2: Panel selection. Given this pool of volunteers and their self-reported group memberships, a selection algorithm is then used to choose a panel. We assume nothing about this algorithm except that it treats people in the same group uniformly, and it produces a panel with 2 seats for less concerned people and 8 seats for more concerned people.*

*It is not hard to see that, in this example, $i$ benefits significantly from misreporting their group membership. If $i$ truthfully reports they are more concerned, they will join a group of 96 people for*

[1] These numbers are based on a real-world panel selection task (instance *sf-e* in our empirical analysis).

*whom the panel has 8 seats, and thus will be chosen with probability* $8/96 \approx 8\%$*. If* $i$ *reports that they are less concerned, they will join a group of 5 people for whom the panel has 2 seats, and will be chosen with probability* $2/5 = 40\%$*. By misreporting that they are less concerned,* $i$ *can increase their selection probability by almost* $32\%$*. Moreover, with probability 40%,* $i$ *will be given a panel seat reserved for less concerned people, thereby giving the group of more concerned people an extra panel seat.*

Example 5.1.1 illustrates why such manipulation is of practical concern: the nature of self-selection bias in this example would be fairly easy for constituents to anticipate — surely, people who care less about climate change will be less likely to volunteer — making the optimal manipulation public knowledge.[1] Moreover, we cannot always prevent manipulation through verification; here, people's opinions would be impossible to check. As citizens' assemblies are used for increasingly higher-profile decisions, the political power associated with participating — and thus the incentive to manipulate — will only increase. Example 5.1.1 also shows a fundamental impossibility: when there is self-selection bias, achieving descriptive representation *necessitates* giving different probabilities to different groups, thereby permitting manipulability. In other words, *no* selection algorithm can achieve representation while eliminating manipulation incentives. This motivates our research question:

**Research question:** What aspects of the selection process can we adjust in practice to *limit* agents' incentives to misreport their features?

**Approach.** We focus on two main aspects of the selection process that can be changed in practice: *the size of the pool of volunteers* $n$, and *the choice of selection algorithm*. The intuition for why increasing $n$ could help is simple: as the pool grows, there are more volunteers per available panel seat. For the correct choice of selection algorithm, this could permit the decrease of *all* volunteers' selection probabilities, thereby diluting the potential gains of manipulation.

Among selection algorithms, we consider only algorithms that achieve maximal fairness, because per Example 5.1.1, manipulation incentives arise from *inequality* in selection probabilities (thus, the goal of equalizing selection probabilities is aligned with limiting manipulation). Specifically, we introduce and study *rounding-based* selection algorithms — a class of maximally fair algorithms that generalizes an algorithm of Flanigan et al. [128]. As discussed in Section 5.2, rounding-based algorithms closely reflect those used in practice, but enforce a slightly relaxed notion of representation.

Each rounding-based algorithm optimizes a different *fairness objective*: a function measuring *how fairly* the chance to participate is spread over volunteers. We study several such functions: *Leximin,* the objective most commonly used in real-world panel selection [130]; *Nash Welfare*, which has known fairness and transparency properties and is available online for practical use [131]; and all $\ell_p$ norms, which we newly introduce to the citizens' assembly setting.

**Results and Contributions. (1) Manipulation model.** Our first contribution is to formally model three realistic manipulation incentives in the assembly selection context: increasing one's

[1] More generally, there are clear patterns across real-world instances of which groups tend to be most underrepresented among volunteers (e.g., those with less education).

own probability of selection, changing someone else's, and — as we saw in Example 5.1.1 — misappropriating seats from other groups. **(2) Impossibilities for existing algorithms.** We then show that, somewhat alarmingly, the state-of-the-art objectives *Leximin* and *Nash Welfare* are *arbitrarily manipulable* on multiple of these counts. Even as $n$ grows large, they permit agents to gain *probability 1* by misreporting, and they allow coalitions to misappropriate a constant fraction of the panel seats. These lower bounds give a key insight: fairness objectives are manipulable when they permit some agents to receive very high selection probabilities. **(3) An optimal selection algorithm.** Motivated by this finding, we study $\ell_p$ norms, which heavily penalize high probabilities due to their strong convexity. We show that even when agents can costlessly misreport any vector of features, the manipulability of the $\ell_p$-norm declines in $n$ at a rate $n^{-(1-1/p)}$, a rate which holds for all three notions of manipulability. We further show that *any selection algorithm* must suffer manipulability at least $\Omega(1/n)$; as $p \to \infty$, our upper bound approaches this lower bound, implying that the $\ell_\infty$ norm — the objective that minimizes the maximum selection probability — achieves optimal convergence. As a bonus, our analysis handles coalitions of size up to $\Theta(n)$. **(4) Empirical results.** We complement these theoretical results with experiments in eight real-world panel selection datasets. Our empirical results closely track our theory, showing that *Leximin* and *Nash Welfare* suffer high manipulability even as $n$ grows, while the manipulability of the $\ell_2$ and $\ell_\infty$ norms declines quickly.

## 5.2 Model

### 5.2.1 Foundations of Selection Algorithms

At a high level, a *selection algorithm* must select a panel of $k$ agents from the pool of $n$ agents. This panel must be representative of the population with respect to a predefined set of *features* $F$, where each $f \in F$ has a predefined set of possible *values* $V_f$. For example, the feature $f = age$ might have possible values $V_{age} = \{18\text{-}40, 41\text{-}60, 61+\}$. We assume that for each feature $f$, its possible values $V_f$ are exhaustive and mutually exclusive. We define $FV := \bigcup_{f \in F} V_f$ to contain all *feature-value pairs*, $(f, v)$ for all $f \in F, v \in V_f$. For all $(f, v)$, $p_{(f,v)}$ is the fraction of the underlying population with value $v$ for feature $f$. Then, a *representative* panel contains $p_{(f,v)} \cdot k$ agents with value $v$ for feature $f$, for all $(f, v) \in FV$. Let $p := (p_{(f,v)} | f \in F, v \in V_f)$.

An *instance* of the panel selection task is then composed of population rates $p$; a desired panel size $k$; and the *pool* $N$, which is defined by all $n$ agents' *true* values of each feature. To define these values, we let $f(i)$ denote $i$'s value for $f$, thereby implicitly treating each feature as a function $f : [n] \to V_f$. $i$'s values across features are summarized in their *feature vector* $w(i) := (f(i) | f \in F)$. The *pool* of volunteers $N := (w(i) | i \in [n])$ is then an $n$-tuple containing all agents' feature vectors. We let $\mathcal{W} := \prod_{f \in F} V_f$ be the collection of all possible feature vectors (i.e., all possible *intersections* of feature-value pairs). A generic feature vector is $w \in \mathcal{W}$. We will often reason only about *fractional composition* of a pool $N$, called $\boldsymbol{\nu}(N)$. This vector is indexed by feature-vector, with $w$-th entry $\nu_w(N) := |\{i \in [n] : w(i) = w\}|/|N|$ representing the fraction of the pool with vector $w$.

In practice, organizers must rely on agents to *report* their feature vectors. Agent $i$'s *reported*

feature vector is denoted $\tilde{w}(i) \in \mathcal{W}$; in general, we will use tilde $\tilde{\cdot}$ throughout the paper to distinguish reported values from true values. The *reported* pool is then denoted as $\tilde{N} = (\tilde{w}(i) | i \in [n])$. In an instance $p, k, N$, a *selection algorithm* $\mathcal{A}$ actually receives as input $p, k, \tilde{N}$, and must map it to a panel $K \subseteq \tilde{N}$.

In the next subsection, we will formally define three motives with which an agent might misreport their feature vector. All these motives revolve around controlling a particular resource: *selection probability*. Agent $i$'s selection probability is $\mathbb{P}[i \in K]$, the probability $i$ is chosen for the panel. We define $\pi_i^{\mathcal{A}}(p, k, \tilde{N})$ to be the selection probability given to agent $i$ by algorithm $\mathcal{A}$ on input $p, k, \tilde{N}$. Accordingly, the vector of agents' selection probabilities is $\pi^{\mathcal{A}}(p, k, \tilde{N})$. Since $p$ and $k$'s true values are known to the algorithm, we simply write $\pi^{\mathcal{A}}(\tilde{N})$. A generic vector of selection probabilities is $\pi$. Note that there are $k$ available seats for $n$ people, so the average selection probability over agents must be $k/n$.

### 5.2.2 Manipulation of Selection Algorithms

In the game we study, we permit all agents to costlessly misreport any feature vector in $\mathcal{W}$. We assume that agents report their feature vector $\tilde{w}(i)$ with knowledge of the entire instance $p, k, N$, plus full access to the selection algorithm.[1] While the assumption that agents exactly know the true pool $N$ is slightly adversarial, our study of simple manipulation heuristics in Section 5.5 will shed light the potential for manipulation using less detailed information about the pool.

We do not commit to a specific utility function for agents, because they might manipulate with a variety of different goals. Instead, we define the three measures of manipulability below, each corresponding to a different motive: the *internal* manipulability $\textsc{manip}_{\text{int}}$ captures how much a coalition can increase the selection probability of its members; the *external* manipulability $\textsc{manip}_{\text{ext}}$ captures how much a coalition can harm a non-member; and the *composition* manipulability $\textsc{manip}_{\text{comp}}$ captures how many seats (in expectation) a coalition can misappropriate from any feature-value group. We denote a coalition as $C$, and we let $N_{-C}$ denote the pool with the feature vectors of $i \in C$ removed. In instance $p, k, N$, the manipulability of $\mathcal{A}$ by any coalition of size $c$ is defined, per notion, as follows, where $\divideontimes := \max_{C \subseteq [n], |C|=c} \max_{\tilde{\boldsymbol{w}} \in \mathcal{W}^{|C|}}$ is shorthand for taking the worst possible coalition of size $c$ and worst possible strategic reports of its members.

$$\textsc{manip}_{\text{int}}(N, \mathcal{A}, c) := \divideontimes \max_{i \in C} \ \pi_i^{\mathcal{A}}(N_{-C} \cup \tilde{\boldsymbol{w}}) - \pi_i^{\mathcal{A}}(N),$$

$$\textsc{manip}_{\text{ext}}(N, \mathcal{A}, c) := \divideontimes \max_{i \notin C} \ \pi_i^{\mathcal{A}}(N) - \pi_i^{\mathcal{A}}(N_{-C} \cup \tilde{\boldsymbol{w}}),$$

$$\textsc{manip}_{\text{comp}}(N, \mathcal{A}, c) :=$$

$$\divideontimes \max_{(f,v) \in FV} \sum_{i: f(i)=v} \pi_i^{\mathcal{A}}(N_{-C} \cup \tilde{\boldsymbol{w}}) - \sum_{i: f(i)=v} \pi_i^{\mathcal{A}}(N).$$

[1] It is realistic to assume agents know $p$ and $k$, and can access the selection algorithm: $p$ is found in census data, and for transparency, $k$ might be public and the selection algorithm would be open-sourced. Assuming agents know $N$ is somewhat adversarial, because in practice, the agents report their features simultaneously; however, this assumption reflects the concern that, by comparing census data and the compositions of past pools, agents could infer who tends to participate, and thus the likely composition of $N$.

### 5.2.3 Rounding-Based Selection Algorithms

We study the manipulability of a class of selection algorithms which we call *rounding-based* selection algorithms. Each rounding-based algorithm is specified by a convex function $g : [0,1]^n \to \mathbb{R}$; we will refer to the algorithm defined by function $g$ simply as $g$. Algorithm $g$ proceeds in two steps: Step 1 computes selection probabilities that minimize $g$, subject to some constraints; then, Step 2 dependently rounds these probabilities to produce a final panel. Since selection probability is the resource sought by manipulating agents — and the selection probabilities are fully determined in Step 1 — only the Step 1 will be of interest in this paper.

*Step 1. Find g-optimal selection probabilities.* Given instance $p, k, N$, in this step the algorithm optimizes $g$ over the polytope $\mathcal{R}(N)$, defined such that $\pi \in \mathcal{R}(N) \iff \pi$ satisfies the following constraints:

$$\sum_{i \in N : f(i) = v} \pi_i = k p_{(f,v)} \quad \text{for all } (f,v) \in FV \tag{C1}$$

$$\sum_{i \in N} \pi_i = k \tag{C2}$$

$$\pi \in [0,1]^n \tag{C3}$$

(C1) requires *ex-ante* representation for all feature-value pairs; (C2) requires that the panel is the correct size in expectation (required for Step 2), and (C3) requires $\pi$ to contain valid probabilities. Formally, in step 1 the algorithm $g$ solves the following convex program:

$$\min_{\pi} \; g(\pi) \quad \text{s.t. } \; \pi \in \mathcal{R}(N) \tag{OPT-PROB}$$

Note that without loss of generality, we can assume that the solution of this convex program assigns the same probability to all agents with the same feature vector, since as any feasible solution can be transformed into such a solution, per the definition of $\mathcal{R}(N)$. We will consider only such solutions throughout the paper.

*Step 2: Randomized-rounding.* This step intakes the selection probabilities found in the previous step, called $\pi^g$, and samples a panel $K$ of size $k$ using the discrepancy-based rounding procedure of Flanigan et al. [128]. For our purposes, the key property of this rounding procedure is that it preserves the selection probabilities $\pi^g$; we defer the details of this procedure to Appendix D.1.1.

**Specific choices of $g$.** We will instantiate the rounding-based algorithms above with various convex functions $g$ — all which, when minimized, tend to make selection probabilities more equal. We analyze two choices of $g$ that serve as benchmarks: *Nash Welfare*, and *Leximin*. Nash Welfare is the geometric mean of selection probabilities:

$$\textsc{Nash}(\pi) := -\prod_{i \in [n]} \pi_i.$$

Leximin is not itself strictly a function, but a refinement of the objective *Maximin*, which maximizes the minimum selection probability given to any agent:

$$\mathsf{maximin}(\pi) := -\min_{i \in [n]} \pi_i.$$

The *Leximin*-optimal solution is computed iteratively: optimize maximin, fix the minimum entry of that solution as a lower bound on any entry of $\pi$, then maximize the second-lowest entry; repeat until all entries are fixed.

Finally, we study all $\ell_p$ norms for $p > 1$, which measure the distance between $\pi$ and the vector of exactly equal selection probabilities $(k/n, k/n, \ldots, k/n)$:

$$\ell_p(\pi) := \|\pi - (k/n, \ldots, k/n)\|_p^p.$$

**Connections to existing algorithms.** With rounding-based algorithms defined, we can now compare them to existing selection algorithms. The most closely-related algorithm is that of Flanigan et al. [128]. Their algorithm computes selection probabilities within $\mathcal{R}$ as in our in Step 1, and then rounds them via the same procedure as in our Step 2. The main difference is that their algorithm manually sets selection probabilities to specific values in Step 1 in a way that ends up satisfying the constraints, while algorithm $g$ within our class sets them by optimizing the function $g$.

Slightly further afield are the most widely-implemented maximally fair algorithms, as introduced by Flanigan et al. [130]. These algorithms differ from ours only in that they enforce representation slightly differently: instead of *ex ante* representation, they require the satisfaction of hard upper and lower demographic quotas *ex post* (e.g., quotas might require that a panel of 10 people contains between 4 and 6 women). As we show in Proposition D.1.2, our algorithms are formally equivalent to a continuous relaxation of these quota-based algorithms where agents are *divisible*. Moreover, our rounding-based algorithms do, in fact, achieve a relaxed version of these ex-post quotas: they are guaranteed to produce a panel containing within $\pm|F|$ of $kp_{(f,v)}$ agents with each value $v$ of each feature $f$ (Lemma 9, Flanigan et al. [128]). This panel is found via a rounding scheme based on a discrepancy theorem due to Beck and Fiala [41].

### 5.3 Leximin and Nash are Highly Manipulable

We begin by analyzing the two objectives most closely tied to practice. Strikingly, Theorem 5.3.1 shows that both Leximin and Nash are extremely manipulable: using either algorithm, an individual agent can gain selection probability 1 by misreporting, and a coalition can *deterministically* misappropriate (approaching) *half* of all panel seats for their own group. The proof of this theorem is found in Appendix D.2.1; we give a proof sketch below.

**Theorem 5.3.1.** *For an arbitrarily large $n$ and for all $c \in [1, k/2)$, there exists an instance $p, k, N$, $|N| = n$ such that*

$$\begin{aligned} \mathit{manip}_{\mathit{int}}(N, \textsc{Leximin}, 1) &= 1 \text{ and} \\ \mathit{manip}_{\mathit{int}}(N, \textsc{Nash}, 1) &= 1; \text{ moreover,} \\ \mathit{manip}_{\mathit{comp}}(N, \textsc{Leximin}, c) &= c \text{ and} \\ \mathit{manip}_{\mathit{comp}}(N, \textsc{Nash}, c) &= c. \end{aligned}$$

*Proof sketch.* Fix a $c \in [1, k/2)$. All claims are proven by a single instance $p, k, N$ with features $f_1$, $f_2$ that take on binary values $\{0, 1\}$ (so the possible feature vectors are $00, 01, 10, 11$). In this instance, we let the population rates of all feature-values be balanced: $p_{f_1,0} = p_{f_1,1} = p_{f_2,0} = p_{f_2,1} = 1/2$. We construct $N$ with the following fractional composition, where $v^*$ should be thought of as a quantity shrinking in $c$: $\nu_{00}(N) = \nu_{11}(N) = v^*$, $\nu_{10}(N) = 1 - 2v^*$, and $\nu_{01}(N) = 0$. We let this pool have some size $|N| = n \geq k^2$, such that its fractional composition can be realized.

First, observe that in this instance, all agents with vector 10 must receive zero selection probability *due to the constraints*: giving them any probability would induce a constraint-violating imbalance in the probability given to agents with $f_1 = 0$ versus $f_2 = 0$, which cannot be re-balanced because the complementary vector 01 does not exist in $N$. This suggests a manipulation strategy: an agent with 10 could misreport 01, thereby permitting greater fairness by allowing agents with 10 to receive some probability.

Let $i$ with $w(i) = 10$, and define $\tilde{N} := N_{-i} \cup \{01\}$ as the pool resulting from $i$ using the proposed strategy. In instance $p, k, \tilde{N}$, agents with 10 can receive probability; the catch is that, for every unit of probability given to such an agent, a unit must also be given to $i$, meaning that $i$ must receive $|N|\nu_{10}$ times the probability of any agent with 10. The key observation is that both Leximin and Nash prioritize ensuring the *minimum* probability is not too small, with little consideration for what happens to the highest probability. For this reason, both algorithms give $i$ selection probability 1 in the instance $p, k, \tilde{N}$. $i$ has gained probability 1 by misreporting, implying the bounds on $\mathsf{manip}_{int}(N, \textsc{Leximin}, 1)$ and $\mathsf{manip}_{int}(N, \textsc{Nash}, 1)$. This argument extends to an entire coalition of $c < k/2$ such agents, implying the bounds on $\mathsf{manip}_{comp}(N, \textsc{Leximin}, c)$ and $\mathsf{manip}_{comp}(N, \textsc{Nash}, c)$. □

**Takeaway: strongly convex objectives.** The key takeaway from this proof is that objectives that do not penalize high selection probabilities can be highly manipulable. A natural class of objectives that *do* penalize high probabilities are *strongly convex* objectives — we formalize this intuition in Proposition D.2.1. This insight suggests that in future study of selection algorithms, it may be desirable to focus on such objectives. This finding also motivates our focus on $\ell_p$ norms — a natural class of strongly-convex objectives.

### 5.4 $\ell_p$-Norms Approach Optimal Manipulability as $p \to \infty$

We now present upper-bounds on all three measures of manipulability for all rounding-based algorithms $\ell_p$ with $p > 1$. These upper bounds will hold for any instance whose pool satisfies Assumption 5.4.1, which conceptually requires that the pool has a minimal level of feature vector richness.

**Assumption 5.4.1** (Pool richness)**.** *$N$ contains some set of feature-vectors $\mathcal{W}^* \subseteq \mathcal{W}$ such that*

1. *there is a constant $\kappa^* > 0$ such that $\nu_w(N) \geq \kappa^* + k/n$ for all $w \in \mathcal{W}^*$, and*
2. *$\mathcal{R}(N)$ contains a solution $\pi^*$ such that $\pi_i = 0$ for all $i : w(i) \notin \mathcal{W}^*$.*

This assumption is likely to hold in practice; in fact, due to how the pool is sampled, *every* feature-vector group's presence in the pool should grow approximately linearly in $n$. We expand on this in Appendix D.3.1. Also, note that the pool used to prove Theorem 5.3.1 satisfies Assumption 5.4.1 (Proposition D.3.1), thus demonstrating a genuine gap between the manipulability of all $\ell_p$ norms and LEXIMIN, NASH.

**Theorem 5.4.2.** *Let $p > 1$, and let $N$ be any pool of size $n$ satisfying Assumption 5.4.1 with $\mathcal{W}^*, \kappa^*, \pi^*$. Let $\kappa \in (0, \kappa^*)$; then, for any coalition size $c \leq \kappa n$, we have that*

$$\begin{aligned} \textsc{manip}_{int}(N, \ell_p, c) &\in O\left(k/n^{1-1/p}\right), \\ \textsc{manip}_{ext}(N, \ell_p, c) &\in O\left(k/n^{1-1/p}\right), \text{ and} \\ \textsc{manip}_{comp}(N, \ell_p, c) &\in O\left(ck/n^{1-1/p}\right). \end{aligned}$$

*Proof.* Fix a pool $N$ with $\mathcal{W}^*, \kappa^*, \pi^*$, as in the theorem statement. Fix any coalition $C \subseteq N$ of size $c \leq \kappa n$. Let $\tilde{N} := N_{-C} \cup \{\tilde{w}(i) | i \in C\}$ be the manipulated pool. For convenience, we will again work with feature-vector-indexed objects. We will again use $\nu_w(N)$ as the frequency of $w$ in $N$. We also define $t_w(\pi) : \sum_{i:w(i)=w} \pi_i$ as the total probability $\pi$ gives to agents with vector $w$. Let the vector of these totals be $t(\pi) = (t_w(\pi) | w \in \mathcal{W})$. We can now reformulate the constraints defining $\mathcal{R}(N)$ in terms of the variable $t$: let $\mathcal{T}(N) \subseteq \mathbb{R}^{|\mathcal{W}|}$ such that $t(\pi) \in \mathcal{T}(N)$ iff

$$\sum_{w:w_f=v} t_w(\pi) = k p_{(f,v)} \text{ for all } (f, v) \in FV \tag{C1'}$$

$$\sum_{w} t_w(\pi) = k \tag{C2'}$$

$$\frac{t_w(\pi)}{n\nu_w(N)} \in [0, 1] \text{ for all } w \in \mathcal{W} \tag{C3'}$$

Let $\pi^* \in \mathcal{R}(N)$ be the feasible solution assumed to exist by Assumption 5.4.1. Then, construct the vector $\tilde{\pi}$ as follows:

$$\tilde{\pi}_i = t_{w(i)}(\pi^*) \,/\, n\nu_{w(i)}(\tilde{N}) \quad \text{for all } i \in N.$$

In effect, the *total* probability assigned to each vector group from $\pi^*$ to $\tilde{\pi}$ is maintained, despite the potentially changing number of agents in that group from $N$ to $\tilde{N}$. Formally:

*Claim 1:* For all $w \in \mathcal{W}$, $t_w(\pi^*) = t_w(\tilde{\pi})$. *Proof:*

$$t_w(\tilde{\pi}) = \sum_{i:w(i)=w} \tilde{\pi}_i = \sum_{i:w(i)=w} \frac{t_w(\pi^*)}{n\nu_w(\tilde{N})} = t_w(\pi^*).$$

*Claim 2:* $\tilde{\pi} \in \mathcal{R}(N)$. *Proof:* We prove this by equivalently showing that $t(\tilde{\pi}) \in \mathcal{T}(\tilde{N})$. The satisfaction of constraints C1' and C2' follow from Claim 1. Moreover, by definition $\frac{t_w(\tilde{\pi})}{n\nu_w(\tilde{N})} \geq 0$

for all $w$. Then, it just remains to show C3':

$$\frac{t_w(\tilde{\pi})}{n\nu_w(\tilde{N})} = \frac{t_w(\pi^*)}{n\nu_w(\tilde{N})} \leq \frac{t_w(\pi^*)}{n(\nu_w(N) - \kappa)}$$
$$\leq \frac{t_w(\pi^*)}{n(\kappa^* + k/n - \kappa)} \leq \frac{k}{k + n(\kappa^* - \kappa)} \leq 1.$$

Now, we will show that the vectors of probabilities $\pi^*$, $\tilde{\pi}$ have maximum entry on the order $1/n$:

*Claim 3:* $\|\pi^*\|_\infty \leq k/\kappa^* n$ and $\|\tilde{\pi}\|_\infty \leq k/(\kappa^* - \kappa)n$. *Proof:* For all $i$ with $w(i) \notin \mathcal{W}^*$, $\pi_i^* = \tilde{\pi}_i = 0$ by definition. For $i$ with $w(i) \in \mathcal{W}^*$, we have that

$$\pi_i^* = \frac{t_w(\pi^*)}{n\nu_w(N)} \leq \frac{k}{n\kappa^*} \text{ and } \tilde{\pi}_i = \frac{t_w(\pi^*)}{n\nu_w(\tilde{N})} \leq \frac{k}{n(\kappa^* - \kappa)}.$$

Now, we relate the infinity-norms of any feasible solution and the $\ell_p$-optimal solution of OPT-PROB:

*Claim 4:* For all $\pi \in \mathcal{R}(N)$, $\|\pi^{\ell_p}(N)\|_\infty \leq n^{1/p}\|\pi\|_\infty + 2kn^{-\frac{p-1}{p}}$. *Proof:* By the optimality of $\pi^{\ell_p}(N)$, we have that $\ell_p(\pi^{\ell_p}(N))^{1/p} \leq \ell_p(\pi(N))^{1/p}$. Then, using properties of norms, and the triangle inequality (twice), we obtain that

$$\begin{aligned}\|\pi^{\ell_p}(N)\|_\infty &\leq \ell_p(\pi^{\ell_p}(N))^{1/p} + \|k/n1\|_p \\ &\leq \ell_p(\pi)^{1/p} + \|k/n1\|_p \\ &\leq \|\pi\|_p + 2\|k/n1\|_p \leq n^{1/p}\|\pi\|_\infty + 2kn^{\frac{1-p}{p}}.\end{aligned}$$

Using that $\pi^* \in \mathcal{R}(N)$, $\tilde{\pi} \in \mathcal{R}(\tilde{N})$, Claims 3 and 4 together imply that $\|\pi^{\ell_p}(N)\|_\infty \leq k/(\kappa^* n^{1-1/p}) + 2k/n^{1-1/p}$ and likewise, $\|\pi^{\ell_p}(\tilde{N})\|_\infty \leq k/((\kappa^* - \kappa)n^{1-1/p}) + 2k/n^{1-1/p}$. Using that the entries of all $\pi$ are nonnegative, it follows that

$$\|\pi^{\ell_p}(\tilde{N}) - \pi^{\ell_p}(N)\|_\infty \leq \left(\frac{1}{\kappa^* - \kappa} + 2\right)\frac{k}{n^{1-1/p}}. \tag{5.1}$$

We've now shown an upper bound on how many any $i$'s probability changes between pool $N$ and pool $\tilde{N}$. This immediately implies the upper bounds on $\text{MANIP}_{\text{int}}(N, \ell_p, c)$ and $\text{MANIP}_{\text{ext}}(N, \ell_p, c)$. Our upper bound on $\|\pi^{\ell_p}(\tilde{N})\|_\infty$ further implies that post-defection, the members of the coalition can have at most $O(ck/n^{1-1/p})$ total selection probability, giving our upper bound on $\text{MANIP}_{\text{comp}}(N, \ell_p, c)$. □

We now show a lower bound that applies to *any* rounding-based algorithm. It shows that up to constants, the manipulability of $\ell_\infty$ decreases at the *optimal* rate in $n$.

**Theorem 5.4.3.** *There is some $\eta > 0$ such that there exist pools $N$ of arbitrarily large size $n$ which, for any coalition size $c \leq 5n/64$ and all objectives $g$, satisfy*

$$manip_{int}(N, g, c) \geq \eta\, k/n, \quad manip_{ext}(N, g, c) \geq \eta\, k/n,$$
$$manip_{comp}(N, g, c) \geq \eta\, ck/n.$$

*The same pools also satisfy Assumption 5.4.1.*

The proof is in Appendix D.3.3 and relies on an example exactly like Example 5.1.1: there is one binary feature, where $v_1$ is severely underrepresented in the pool. The bounds arise from agents with $v_0$ misreporting $v_1$.

## 5.5 Manipulability of Real-World Instances

Now we compare the manipulability of Leximin, Nash, $\ell_2$ and $\ell_\infty$ in eight real-world panel selection instances. Instance details are provided in Appendix D.4.1. We present here two representative instances, called *sf(a)* and *hd*, and defer the rest to Appendix D.4. The datasets were obtained from groups of assembly organizers based in the UK and US, respectively. Each real-world instance consists of $p, k, N$. To study how manipulability changes as we increase the pool size, we simply copy the pool, leaving $p$ and $k$ fixed. In each instance, we copy the pool until $n \geq 100k$, as practitioners often specify their target pool size in multiples of $k$.

We will test our selection algorithms against an *individual* manipulator — that is, we measure how much selection probability any agent can gain by misreporting their feature vector. The most powerful individual manipulator could gain $\mathsf{manip}_{int}(N, \mathcal{A}, 1)$ probability against $\mathcal{A}$ — the quantity to which our theoretical bounds apply. Given the computational difficulty of calculating the optimal manipulation (each agent has $|\mathcal{W}| \in \Omega(2^{|F|})$ possible strategies), we test our algorithms against three practically-motivated heuristic strategies: *OPT-1, MU,* and *HP*, defined below. The results are summarized in Figure 5.1.

**OPT-1:** ***Optimal misreport of one feature.*** An agent playing strategy *OPT-1* reports the feature vector that benefits them most, *subject to misreporting their value for at most one feature.* This strategy, in practice, might correspond to a practical setting in which only a few features cannot be validated. When comparing across algorithms, we think of OPT-1 as a proxy for the optimal individual manipulation. As column 1 of Figure 5.1 shows, the manipulability of $\ell_2$ and $\ell_\infty$ against OPT-1 declines quickly in $n$, while Leximin and Nash remain arbitrarily susceptible to manipulation. The fact that Leximin and Nash are so manipulable *even when agents are willing to misreport only one feature* was not implied by our lower bounds, and shows the findings in our theoretical lower bounds are of practical relevance.

**MU:** ***Most underrepresented.*** Let $\eta_{(f,v)}(N) := |\{i|f(i) = v\}|/|N|$ be the fraction of agents with value $v$ for feature $f$. An agent playing strategy *MU* reports the vector containing the most underrepresented value of each feature $f$ — that is, $\tilde{w}_f := \arg\max_{v \in V_f} p_{(f,v)}/\eta_{(f,v)}(N)$. Again, Leximin and Nash are arbitrarily manipulable against *MU*, even for large $n$. The vulnerability of Leximin and Nash here is of especially high practical concern, because the *MU* manipulation strategy is perhaps the most likely to be used in practice by less sophisticated manipulators: it is intuitive and requires only ordinal information about (the only $O(|F|)$ many) feature-value frequencies

and no access to the algorithm (in contrast, *OPT-1* and *HP* require algorithm access *and* information about the pool's vector-level composition).

**HP: *Highest-Probability.*** Another reasonable heuristic a manipulator $i$ might use would be to report the vector $\tilde{w}$ that receives the highest selection probability in the true pool; we call this heuristic *HP*. That this strategy's efficacy declines in $n$ intuitively makes sense: misreporting a vector that is already in the pool means joining a vector group whose size is growing linearly in $n$ (at least in these experiments, where we are duplicating $N$). This intuition alludes to the insight that the most problematic misreports for suboptimal algorithms are those of vectors that do not already exist in the pool — an intuition supported by both the proof of our lower bound in Theorem 5.3.1, and the fact that the most underrepresented vector (targeted by the much more effective strategy *MU*) is not in the original pool of any instance we study.

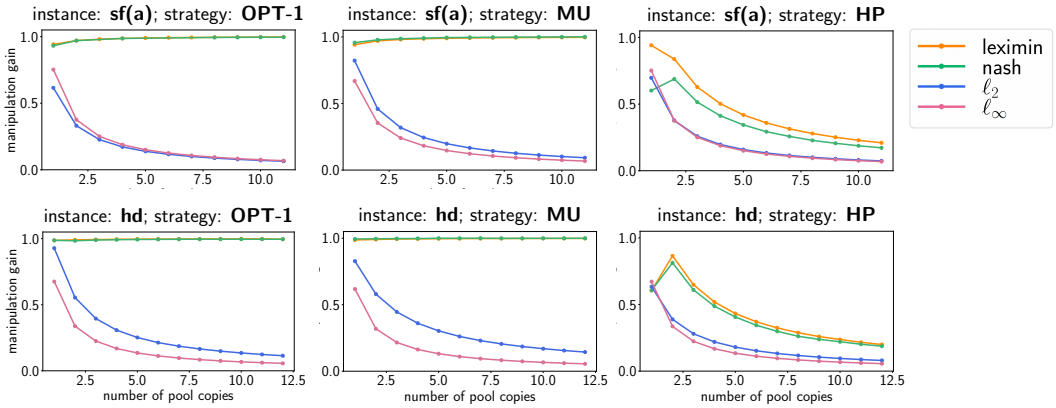



Figure 5.1: Rounding-based algorithms Leximin , Nash , $\ell_2$, and $\ell_\infty$ versus each manipulation strategy in instances *sf(a)* and *hd*.

### 5.5.1 Extension: Manipulability and *Selection Bias*

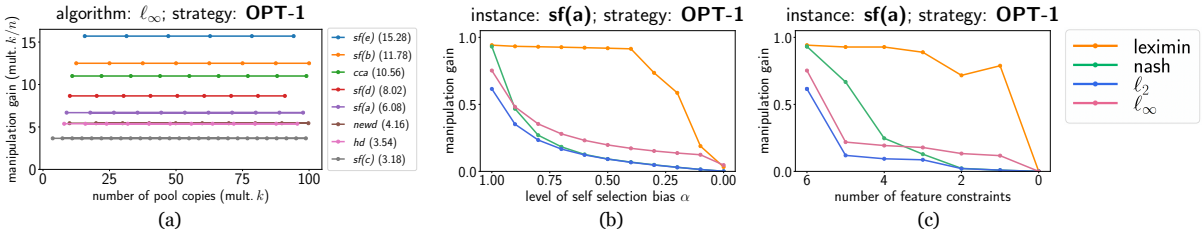



Figure 5.2: The impact of self-selection bias on the manipulability of Leximin , Nash , $\ell_2$ and $\ell_\infty$ by an agent playing *OPT-1* strategy.

While $n$ is much easier to change in practice than the level of self-selection bias (SSB), the SSB could be decreased by a more targeted recruitment process, motivating our study of this would impact the manipulability. We introduce a measure of SSB in an instance, which roughly captures

how severely the algorithm must skew selection probabilities to satisfy the constraints:

$$\Delta_{p,k,N} := \max_{(f,v)\in FV} \frac{p_{(f,v)}}{\eta_{(f,v)}(N)} - \min_{(f,v)\in FV} \frac{p_{(f,v)}}{\eta_{(f,v)}(N)}$$

Figure 5.2(a) shows that this measure of SSB is highly predictive of manipulability: across instances, the manipulation gain of OPT-1 (scaled by $k/n$, for standardization) against $\ell_\infty$ corresponds closely with instances' $\Delta_{p,k,N}$ values, as listed in the figure legend. Proceeding with this measure, we evaluate the impact of decreasing it in two ways. First, in Figure 5.2(b), we decrease the SSB smoothly by *interpolating* between the original pool $N$ and the "nearest" (by Euclidean distance) pool $N'$ with $\Delta_{p,k,N'} = 0$. Second, in Figure 5.2(c), we decrease the SSB by successively dropping features from the instance in decreasing order of their *feature-level* SSB, defined as $\Delta_{p,k,N}$ restricted to the values of a given feature. Using either approach, in *sf(a)*, the manipulability of all algorithms except Leximin against *OPT-1* drops quickly, while Leximin remains manipulable until extremely low levels of SSB are reached. We defer the details of these methods, plus results for the remaining instances, to Appendix D.4.3.

## 5.6 Discussion

Our work illuminates a tradeoff between two goals: ensuring that no one gets too *little* selection probability (as pursued in the related work [130]), and ensuring that no one gets too *much* probability (which we show is important for limiting manipulation incentives). Leximin and Nash prioritize the first goal but, as we show, perform poorly on the second. In contrast, we show that $\ell_p$ norms can be optimal in regards to the second goal, but they perform poorly on the first: we find that both $\ell_2$ and $\ell_\infty$ give at least one agent zero probability in all eight instances we study (see Appendix D.4.5). This begs the question: *is there an objective that both prevents high probabilities (thereby limiting manipulability) as well as low probabilities?* An objective with *optimal* dependency on $n$ for both desiderata at once would give all agents $\Theta(1/n)$ probability.[1]

Another first-order technical extension of this work would be to repeat this analysis within *quota-based* algorithms, as they implement the notion of representation most commonly used [130]. Because the separation between Leximin , Nash versus $\ell_p$ norms is due to fundamental properties of these objectives, we expect them to exhibit roughly similar behavior in quota-based algorithms. However, the combinatorial structure of quotas may make quota-based algorithms much *more* manipulable in the worst case.

Even without this extension to quota-based algorithms, our work raises some practical insights. First, it suggests that in general, algorithms permitting high selection probabilities come with risks of manipulability — a property that can be tested in any selection algorithm, maximally fair or not. If one *does* maximize a carefully chosen fairness objective, our work reveals practicable strategies for limiting manipulation incentives: decreasing the SSB (even simply by dropping features that one expects to be highly self-selected), or recruiting a larger pool. Based on our

[1] $\Theta(1/n)$ is the optimal rate at which manipulability can decline (Theorem 5.4.3); because any algorithm must divide $k$ probability over $n$ people, the minimum probability can be at most $\Theta(1/n)$.

empirical results, even doubling the pool sizes currently used in practice would substantially decrease manipulability.

Beyond the application of assembly selection, our problem is conceptually reminiscent of *strategic classification*, in which agents may misreport their features to increase their probability of receiving a desirable prediction from a machine-learned classifier [19, 80, 95, 161]. Within the strategic classification framework, we can view a selection algorithm as a *constrained* classifier: one which classifies agents as either on or off the panel with some probability based on their features, while satisfying demographic representation constraints on who receives a positive classification. While some existing work is tangentially related [190], to our knowledge this precise problem has not been studied in the strategic classification literature. Our notions of manipulability, and our technical results on the stability of our convex program, may be of interest for this domain.


## Acknowledgements

We thank Thibaut Horel and Paul Gölz for helpful technical discussions; the reviewers for their excellent feedback; and with the several organizations that provided real-world citizens assembly data including the Sortition Foundation, the Center for Blue Democracy, Healthy Democracy, and New Democracy.

# 6

# Fairness, Manipulation-Robustness, & Transparency

*Fair, Manipulation-Robust, and Transparent Sortition* [31].
Carmel Baharav & Bailey Flanigan.
*submitted 2024.*

## 6.1 Introduction

In a *citizens' assembly*, a panel of *randomly-selected* everyday people is convened to discuss and collectively weigh in on a policy issue. Each year, more and more cities, regions, countries, and even supranational bodies are turning to citizens' assemblies[1] to involve the public in policymaking; prominent recent examples include multiple national citizens' assemblies in France [65, 146], Scotland's national climate assembly [249], and a permanent assembly instated in the Ostebelgian government [219].

The subject of this paper is the process used to randomly select the panel members, called *sortition*. Broadly defined, sortition just means "random selection", and it is often thought of as a simple uniform lottery over the population. In practice, however, the task of sortition is more complicated: practitioners require the panel to satisfy custom *quotas*, which enforce near-proportional representation of key population sub-groups. These groups are usually defined by individual features (e.g., *women* or *right-leaning voters*), but can be defined by intersections of features as well. While representation of groups would in theory be achieved by a uniform lottery, there is *selec-*

[1] *Citizens' assemblies* belong to a broader category of closely-related methods called *deliberative minipublics*, which consist also of citizens' juries, citizens' panels, deliberative polls, and other processes of similar form. We will discuss citizens' assemblies as the primary application domain of this paper.

*tion bias*: different subgroups tend to agree to participate at very different rates, meaning that a simple lottery would produce a panel that is far from representative. To ensure representation despite selection bias, in practice panels are selected via the following two-stage process:

(1) First, a uniform sample of the population is invited to participate. Those who respond affirmatively form the *pool of volunteers*. Due to selection bias, this pool is typically very skewed compared to the population.

(2) All pool members are asked to report their values of the features on which quotas will be imposed. Then, a *selection algorithm* is used to find a panel within the pool, which must satisfy the practitioner-defined quotas and be of predetermined size $k$.

Our focus is the design of the selection algorithm in Stage (2), whose task it is to sample a representative panel from a skewed pool. The skew of the pool relative to the panel prevents any selection algorithm from randomizing over the pool members perfectly uniformly, as would a simple lottery.[1] However, recent work by Flanigan et al. [130] has made it possible to randomize *as equally as possible* over pool members: they introduce an algorithmic framework that can make volunteers' probabilities *maximally equal* subject to the quotas, as measured by any convex *equality objective* $\mathcal{E}$ (i.e., any mapping from a vector of pool members' selection probabilities to a real number measuring how equal they are). The ability to make selection probabilities maximally equal is desirable because it offers hope of retaining — at least to a maximum degree possible — the normative ideals granted by a simple lottery, such as ***Fairness, Manipulation Robustness,*** and ***Transparency*** (to be defined shortly). The question is then: *what equality objective $\mathcal{E}$ should we optimize, in order to maximally achieve these ideals*? Subsequent work, which we overview now, has revealed how the choice of $\mathcal{E}$ can have important consequences for these ideals.

The originally proposed equality objective was *Leximin* (a refinement of *Maximin*), which measures equality according to the *minimum* selection probability, thereby ensuring that no one gets *too little* selection probability. This choice of objective was motivated by the ideal of ***Fairness***: *that every willing participant is entitled to their fair share of the chance to participate.*

*Leximin* made very real fairness gains over the existing state-of-the-art algorithms, but subsequent work identified a major weakness of this objective: both in theory and in practice, it allows pool members to ensure they are deterministically selected for the panel by misreporting their features at the beginning of Stage (2). Flanigan et al. [135] named this behavior *manipulation*; while they show that some incentives for manipulation are unavoidable, they say a selection algorithm exhibits ***Manipulation Robustness*** if it *minimizes agents' incentives to manipulate*. Flanigan *et al* diagnose the reason *Leximin* is so vulnerable to manipulation: it raises low probabilities without regard for high probabilities, so if the manipulator can guess which identities are essential to raising low probabilities, the algorithm may push their selection probability all the way up to 1. Further, Flanigan *et al* show that the popular equality objective *Nash Welfare* — the geometric mean of selection probabilities — suffers the same problem for the same reason. Motivated by these negative results, they propose a new equality objective, *Minimax*, which minimizes

[1] If a group is disproportionately overrepresented in the pool compared to their quota-allotted fraction of the panel, satisfying the quotas requires giving at least one member of this group below-average chance of selection.

the maximum selection probability. They show that *Minimax*, by controlling high probabilities, minimizes agents' incentives for manipulating, thereby achieving optimal *Manipulation Robustness*. Unfortunately, they find that *Minimax* has essentially the opposite problem as *Leximin*: because *Minimax* does not control *low* probabilities, it often gives many people zero selection probability, thereby performing unacceptably poorly on the ideal of *Fairness*.

From the related work, we distill three observations: low probabilities are a problem for *Fairness*; high probabilities are a problem for *Manipulation Robustness*; and no known objective controls both simultaneously. These observations motivate the first two questions we will tackle in this paper:

***Question 1:*** Can we design an equality objective $\mathcal{E}$ that ensures optimal simultaneous lower and upper bounds on selection probabilities? and consequently,

***Question 2:*** Do these bounds on $\mathcal{E}$ permit simultaneous guarantees on $\mathcal{E}$'s *Fairness* and *Manipulation Robustness*?

Finally, we investigate a third ideal, *Transparency*—colloquially, the idea that public should be able to confirm that the selection process is actually *random*, and the organizers are not just stacking the panel behind the scenes. Flanigan, Kehne, and Procaccia [131] proposed a more precise definition of the ideal of ***Transparency:*** that *without reasoning in-depth about probability, the public should be able to observe all volunteers' chances of selection.* Flanigan et al. [130] also proposed a method that targets this ideal: a rounding algorithm that, with only slight modification to the optimal probabilities, permits panel selection to be done via a *live uniform lottery* over (potentially duplicated) panels [130]. Then, given a public (anonymized) list of panels each pool member is on, people can tabulate pool members' selection probabilities by simply counting the number of panels they are on and dividing by the total number of panels (usually 1000 in practice). Flanigan et al. [131] proved bounds on how much optimality can be lost in the rounding required to produce the uniform lottery, with respect to the objectives *Leximin* and *Nash Welfare*. Given that this uniform lottery approach is used in practice [130], for a new equality objective to be viable, we must ensure that it can be rounded to a uniform lottery without too much loss in *Fairness* or *Manipulation Robustness*. This motivates our third question:

***Question 3:*** If we achieve *Transparency* by rounding the output of our $\mathcal{E}$-optimal algorithm to a uniform lottery, to what extent does $\mathcal{E}$ still achieve *Fairness* and *Manipulation Robustness*?

### 6.1.1 Approach and Contributions

**Unification of existing models and new equality objectives (Section 6.2).** Before addressing these questions, there is considerable work to do in unifying existing models of fairness, manipulation robustness, and transparency. This requires new algorithm performance metrics, such as manip-fairness, which captures the worst-case fairness of an algorithm *in the presence of manipulating coalitions*. Next, given the insufficiency of known equality objectives, we propose two new ones that explicitly aim to control high and low selection probabilities simultaneously.

To design these objectives, we draw from multi-objective optimization: we combine our two goals into one objective, with a scalar $\gamma$ that determines the extent to which prioritize them relative to each other. The objectives are as follows, where (slightly informally for now) $\boldsymbol{\pi}$ is an assignment of selection probabilities to pool members, and $\max(\boldsymbol{\pi}), \min(\boldsymbol{\pi})$ are variables describing the maximum and minimum selection probability, respectively:

$$\textit{Linear}_\gamma(\boldsymbol{\pi}) : \max(\boldsymbol{\pi}) - \gamma \cdot \min(\boldsymbol{\pi}) \qquad \text{and} \qquad \textit{Goldilocks}_\gamma(\boldsymbol{\pi}) : \max(\boldsymbol{\pi}) + \gamma \cdot {}^{1}/_{\min(\boldsymbol{\pi})}.$$

**Impossibilities for existing equality objectives, plus *Linear*$_\gamma$ (Section 6.3).** First, we show that all objectives studied in past work — *Maximin*, *Leximin*, *Nash Welfare*, and *Minimax*[1] — are either arbitrarily manipulable (unnecessarily giving manipulators probability 1) or unfair (unnecessarily giving some agents probability 0). Surprisingly, we show that despite its explicit prioritization of both high and low probabilities, *Linear*$_\gamma$ also suffers these problems: for high $\gamma$, *Linear*$_\gamma$ is also arbitrarily manipulable; for lower $\gamma$, it is unnecessarily unfair. Conceptually, this is because *Linear*$_\gamma$ does not penalize low probabilities relative to high ones steeply enough. This finding motivates our study of *Goldilocks*$_\gamma$, whose gradient in the minimum probability is steeper.

**Main Results: Bounds on the *Fairness*, *Manipulation Robustness* and *Transparency* of *Goldilocks*$_1$ (Sections 6.4 and 6.5).** As in previous work [135], we study three manipulation incentives: increasing one's own selection probability, decreasing someone else's, or misappropriating panel seats from other groups. As in past work, we permit manipulating coalitions of up to linear size (in the pool size $n$), and we permit agents to misreport any features costlessly with full knowledge of the selection algorithm and the pool's composition.

*Key technical challenge.* *Goldilocks*$_1$'s ability to guarantee *Fairness* and *Manipulation Robustness* simultaneously depends on its ability to ensure that no selection probability is too low (to ensure fairness) nor too high (to ensure manipulation robustness). However, the extent to which *any* algorithm can do this depends on the quality of solutions available in the instance, *which we observe can be affected by manipulation.* Concretely, we show that a coalition of agents can misreport their features in a way that eliminates *all feasible solutions* in which agents' probabilities are close together. In previous work on the manipulation robustness of *Minimax* — an objective which cares only about *high* selection probabilities — this was not a problem, because *Minimax* could simply give groups inducing such probability gaps zero selection probability. In other words, manipulating coalitions did not affect the set of potentially *optimal* solutions. In contrast, *Goldilocks* (or any objective controlling both high and low probabilities) *must respond* to such fundamental gaps in selection probabilities, so manipulating coalitions can affect the set of potentially optimal solutions.

*Approach.* Our approach consists of two steps: first, we show that *Goldilocks*$_1$ guarantees lower and upper bounds on selection probabilities that scale naturally — in the many relevant cases, tightly — with the quality of available feasible solutions, which can depend on the existence of

[1]For simplicity, we study *Minimax* in place of the $\ell_p$-norm equality objectives studied in [135]. As we will see, the behavior of these classes of objectives is essentially the same, with both strongly penalizing high probabilities.

coalitions. Then, we bound the extent to which a manipulating coalition can affect the set of potentially *Goldilocks*-optimal solutions.

*Results.* In regards to manipulation robustness, we find that no manipulating coalition of size $c$ can increase a single member's selection probability by more than order $\sqrt{c}/n$ — a quantity that diminishes quickly in $n$, even if $c$ grows linearly with $n$. We show that this bound is tight. We give similar bounds for the other two manipulation incentives, and discuss their tightness. Regarding *Fairness*, we find that, even when a coalition of size $c$ manipulates, *Goldilocks*$_1$ guarantees *Maximin* fairness (i.e., minimum probability) of at least order $1/(\sqrt{c}n)$. We show that this bound is tight. To enable these results to be (approximately) achieved alongside *Transparency*, in Section 6.5 we extend our bounds to hold for output of *Goldilocks*$_1$ after it is rounded to a uniform lottery.

**Empirical study of *Goldilocks*$_1$ (Section 6.6).** Finally, we analyze *Goldilocks*$_1$ in real citizens' assembly datasets, and we find that it performs even better than our bounds guarantee. Our first key finding is that *Goldilocks*$_1$ achieves near *Leximin*-optimal minimum probabilities *and* *Minimax*-optimal maximum probabilities — an outcome whose possibility by *any* algorithm was not guaranteed. On our ideals, we compare *Goldilocks*$_1$'s performance to the other equality notions previously analyzed — *Leximin*, *Nash Welfare*, and *Minimax* — as well as *Legacy*, a heuristic standing in for the wide variety of heuristics still used in practice today. We find that *Goldilocks* performs nearly as well as *Leximin* on fairness and *Minimax* on manipulability, and far outperforms all other algorithms in its ability to achieve both these goals at once. Finally, we find that *Goldilocks*$_1$ can be made transparent with little-to-no cost to the maximum and minimum selection probabilities.

### 6.1.2 Related Work

In addition to the existing work on fairness [130], manipulation robustness [135], and transparency [131] on which we directly build, there is a growing body of work pursuing selection algorithms achieving similar ideals. There is especially a wealth of literature considering the interplay of two ideals: *fairness* (as we define it), and proportional representation of the underlying population, which we enforce with quotas. However, much of this work is done in the distinct model of sortition where it is possible to sample the population directly, and all chosen will participate (i.e., there is no selection bias). For example, Ebadian and Micha [104] study how to achieve exact fairness and deterministic proportional representation simultaneously; closely related is work by Benadè et al. [46], which focuses on uniform-like stratified sampling while preserving subgroup-level representation. Ebadian et al. [106] ask richer questions about the nature of representation that that can be achieved when individual people can serve as representatives for others to varying extents. Outside the uniform selection model, Gąsiorowska [143] does a qualitative survey across many selection process case studies, evaluating them on the basis of *randomness* (closely related to our ideal of *fairness*) as well as representation. Beyond related work on sortition, the existing theoretical results we build on in this paper use tools from across several fields, including randomized rounding [141], discrepancy theory [41], and optimization

of large linear programs [54].

### 6.2 Model

We use $\Delta(S)$ to represent the set of all distributions over the elements of set $S$. Let $N = [n]$ be the *pool*, where $i \in N$ is an individual agent. $N$ is formed by inviting a uniform sample of the population to participate; the agents in $N$ are those who responded affirmatively to this invitation.

**Features, feature-values, and feature-vectors.** Let $F$ be a predefined set of *features*, where each feature $f \in F$ can take on some predefined set of values $V_f$. For example, $F$ could be {*age*, *gender*}, and $V_{age}$ might be {18 - 40, 41 - 60, 61 +}. We call each $v \in V_f$ a *feature-value* and each $f, v$ a *feature-value pair*. $FV := \{(f, v) | f \in F, v \in V_f\}$ is the set of all feature-value pairs.

We assume that for each feature $f$, its possible values $V_f$ are exhaustive and mutually exclusive, so every agent has exactly one value $v \in V_f$ for every feature $f$. We denote $i$'s value for feature $f$ as $f(i)$, thereby using each $f$ as a function $f : [n] \to V_f$. We let $i$'s *feature vector* $w(i) := (f(i) | f \in F)$ summarize their feature-values, and let $\mathcal{W} := \prod_{f \in F} V_f$ be the set of all possible feature vectors. We will often reason about a pool according to the number of agents it contains with each feature vector; for all $w \in \mathcal{W}$, let $N_w$ denote the number of agents in $N$ with vector $w$. Because the pool will generally not contain all possible feature vectors, we will abuse notation slightly and let $\mathcal{W}_N \subseteq \mathcal{W}$ denote the set of unique feature vectors present in $N$.

**The panel selection task.** Our task is to choose a *panel* $K \subset N$ of some pre-chosen size $k \in \mathbb{N}$. The main constraint on $K$ is that it must satisfy *upper and lower quotas* on all feature-values. Formally, for each $f, v \in FV$, we define lower and upper quotas $\ell_{f,v} \in \mathbb{N}^+$ and $u_{f,v} \in \mathbb{N}^+$ We summarize these quotas in $\boldsymbol{\ell} = \{\ell_{f,v} | f, v \in FV\}$ and $\boldsymbol{u} = \{u_{f,v} | f, v \in FV\}$. The set of all *valid panels* — i.e., those satisfying all requirements — is then

$$\mathcal{K} := \left\{K : K \subseteq N \quad \wedge \quad |K| = k \quad \wedge \quad \ell_{f,v} \leq |\{i \in K : f(i) = v\}| \leq u_{f,v} \; \forall f, v \in FV\right\}.$$

An *instance* of the panel selection task is defined as $\mathcal{I} := (N, k, \boldsymbol{\ell}, \boldsymbol{u})$. Given an instance, the panel selection task is to output a valid panel $K \in \mathcal{K}$.

**Panel distributions and selection probabilities.** In instance $\mathcal{I}$ with valid panels $\mathcal{K}$, $\Delta(\mathcal{K})$ is the set of all possible randomizations over valid panels. We call each $\mathbf{d} \in \Delta(\mathcal{K})$ a *panel distribution*, where $d_K$ then denotes the probability of drawing $K$ from $\mathbf{d}$. Any given $\mathbf{d}$ must imply some *selection probability* for each agent $i \in N$, defined as

$$\pi_i(\mathbf{d}) := \textstyle\sum_{K \in \mathcal{K} : i \in K} d_K \qquad \text{for all } i \in [n].$$

In words, $\pi_i(\mathbf{d})$ is the probability that $i$ is included on the panel when the panel is drawn from $\mathbf{d}$. We refer to $\boldsymbol{\pi}(\mathbf{d}) := (\pi_i(\mathbf{d}) | i \in [n])$ as an *assignment* of selection probabilities to all agents in the pool. A generic selection probability assignment will be $\boldsymbol{\pi}$. We use the shorthand $\max(\boldsymbol{\pi}) := \max_{i \in [n]} \pi_i$ and $\min(\boldsymbol{\pi}) := \min_{i \in [n]} \pi_i$ to respectively represent the maximum and minimum selection probability assigned by $\boldsymbol{\pi}$ to any agent.

In any instance $\mathcal{I}$, the space of all *realizable* selection probability assignments is $\Pi(\mathcal{I}) := \{\boldsymbol{\pi}(\mathbf{d}) : \mathbf{d} \in \Delta(\mathcal{K})\}$. In words, $\Pi(\mathcal{I})$ is the set of all selection probability assignments that are implied by some randomization over exclusively valid panels. Observe that for any $\boldsymbol{\pi} \in \Pi(\mathcal{I})$, $\sum_{i\in[n]} \pi_i = k$. Therefore, in any given instance, the selection probability assignment that gives all agents equal selection probability must be $\boldsymbol{\pi} = k/n\mathbf{1}^n$, the $n$-length vector in which every entry is $k/n$ (note that in most instances, this selection probability assignment will not be in $\Pi(\mathcal{I})$).

Finally, we say that $\boldsymbol{\pi}$ is *anonymous* iff it gives all agents with the same feature vector the same selection probability — that is, for all $w \in \mathcal{W}$, there exists a constant $z_w$ such that $\pi_i = z_w$ for all $i : w(i) = w$. As we will typically work with anonymous selection probability assignments, we define *vector-indexed selection probabilities* $\mathrm{p}_w(\boldsymbol{\pi}) = z_w$. Let $\mathrm{p}(\boldsymbol{\pi}) = (\mathrm{p}_w(\boldsymbol{\pi})|w \in \mathcal{W})$. When $\boldsymbol{\pi}$ is clear from context or when we work with arbitrary vector-indexed probabilities, we simply write p.

**Equality objectives.** Let an *equality objective* $\mathcal{E} : [0,1]^n \to \mathbb{R}$ be a function that intakes a selection probability assignment and outputs a scalar measure of how *equal* the selection probabilities within it are. All equality objectives we will consider will be convex, and will have the property that $\boldsymbol{\pi}$ is "more equal" than $\boldsymbol{\pi}'$ according to $\mathcal{E}$ if $\mathcal{E}(\boldsymbol{\pi}) \leq \mathcal{E}(\boldsymbol{\pi}')$. Then, a selection probability assignment $\boldsymbol{\pi}$ is *maximally equal* in $\mathcal{I}$ iff $\boldsymbol{\pi} \in \arg\inf_{\boldsymbol{\pi}\in\Pi(\mathcal{I})} \mathcal{E}(\boldsymbol{\pi})$. The set of all maximally equal selection probability assignments in $\mathcal{I}$, as measured by $\mathcal{E}$, is

$$\Pi^{\mathcal{E}}(\mathcal{I}) := \arg\inf_{\boldsymbol{\pi}\in\Pi(\mathcal{I})} \mathcal{E}(\boldsymbol{\pi}) \subseteq \Pi(\mathcal{I}).$$

The equality objectives we study are defined below. Of these objectives, we introduce $Linear_\gamma$ and $Goldilocks_\gamma$; all others have been studied in past work on sortition [130, 131, 135]. Here, *Nash* is the *Nash Welfare*. By convention, we define all objectives to be minimized.

$$Maximin(\boldsymbol{\pi}) := -\min(\boldsymbol{\pi}), \quad Minimax(\boldsymbol{\pi}) := \max(\boldsymbol{\pi}), \quad Nash(\boldsymbol{\pi}) := -\left(\textstyle\prod_{i\in[n]} \pi_i\right)^{1/n}.$$

We also study *Leximin*, which is not strictly an equality objective, but is a refinement of *Maximin*. *Leximin* first maximizes the minimum selection probability (i.e., finds the *Maximin*-optimal solution), then maximizes the *second-lowest* selection probability, then the third-lowest, and so on. The two new equality objectives we introduce, $Linear_\gamma$ and $Goldilocks_\gamma$, are both designed to simultaneously ensure that no one gets too *little* or too *much* selection probability. In either objective, $\gamma \in \mathbb{R}_{\geq 0}$ controls the relative priority placed on each goal.

$$Linear_\gamma(\boldsymbol{\pi}) := \max(\boldsymbol{\pi}) - \gamma \min(\boldsymbol{\pi}), \qquad Goldilocks_\gamma(\boldsymbol{\pi}) := n/k \ \max(\boldsymbol{\pi}) + \gamma \cdot \frac{1}{n/k \ \min(\boldsymbol{\pi})}.$$

In addition to being convex (proposition E.1.1), all objectives we consider in this paper[1] satisfy two other natural axioms — *conditional equitability* (Proposition E.1.2) and *anonymity* (Proposition E.1.4), both weak requirements reflecting that these objectives truly measure the level of *equality* of selection probabilities. In words, *conditional equitability (CE)* requires $\mathcal{E}$ to consider

[1]Because *Leximin* is not an equality objective, it cannot formally satisfy these properties. However, as will be clear throughout the paper, *Leximin effectively* satisfies these properties to the extent we need it to.

$\boldsymbol{\pi} = k/n\mathbf{1}^n$ the most equal possible probability assignment, and *anonymity* requires that $\mathcal{E}$ does not penalize giving identical agents identical selection probabilities.

**Axiom 6.2.1** (CE). *$\mathcal{E}$ is conditionally equitable iff for all $\mathcal{I}$, $k/n\mathbf{1}^n \in \Pi(\mathcal{I}) \implies k/n\mathbf{1}^n \in \Pi^{\mathcal{E}}(\mathcal{I})$.*

**Axiom 6.2.2** (Anonymity). *$\mathcal{E}$ is anonymous iff for all $\mathcal{I}$, there exists an anonymous $\boldsymbol{\pi} \in \Pi^{\mathcal{E}}(\mathcal{I})$.*

Because all objectives $\mathcal{E}$ we consider satisfy anonymity, we will without loss of generality redefine $\Pi(\mathcal{I})$ and $\Pi^{\mathcal{E}}(\mathcal{I})$ to contain *only anonymous selection probability assignments.*

### 6.2.1 Selection algorithms

A *selection algorithm* $\mathrm{A} : \mathcal{I} \to \mathcal{K}$ is any (potentially randomized) mapping from an instance to a valid panel $K \in \mathcal{K}$. Note that in a given instance, any selection algorithm must induce a panel distribution; we denote the panel distribution implied by A in $\mathcal{I}$ as $\mathbf{d}^{\mathrm{A}}(\mathcal{I}) \in \Delta(\mathcal{K})$. Its implied selection probability assignment is then $\boldsymbol{\pi}(\mathbf{d}^{\mathrm{A}}(\mathcal{I}))$; for simplicity of notation, when the panel distribution is not directly relevant, we will shorten this to $\boldsymbol{\pi}^{\mathrm{A}}(\mathcal{I})$.

A selection algorithm A is *maximally equal* with respect to $\mathcal{E}$ iff $\boldsymbol{\pi}^{\mathrm{A}}(\mathcal{I}) \in \Pi^{\mathcal{E}}(\mathcal{I})$ for all $\mathcal{I}$. Fortunately, the optimization framework proposed by Flanigan et al. [130] gives an algorithmic implementation for any maximally equal selection algorithm whose corresponding equality objectives $\mathcal{E}$ is convex, which we will use to optimize the equality objectives defined above. At a high level, their algorithmic approach works in two steps: first, it explicitly computes a panel distribution implying *maximally equal* selection probabilities per $\mathcal{E}$; then, it draws the final panel from this panel distribution, thereby realizing those maximally equal selection probabilities. As shorthand, we will refer to the algorithm from this framework optimizing $\mathcal{E}$ as E (e.g., the algorithm optimizing *Maximin* is called Maximin).

To simplify our exposition, we make two weak assumptions about the instances and the maximally equal selection algorithms we study.

**Assumption 6.2.3.** *(1)* ***Feasibility:*** *All $\mathcal{I}$ have non-empty corresponding $\mathcal{K}$. (2)* ***Unincludable Agents:*** *If in $\mathcal{I}$, there exists any $i \in N$ for which $\{K : K \in \mathcal{K} \wedge i \in K\} = \emptyset$, then the algorithm will first identify these agents, remove them from the instance, and act on the resulting instance in which all agents exist on a valid panel.*

### 6.2.2 Ideals: *Manipulation Robustness*, *Fairness*, and *Transparency*

*Manipulation Robustness.* To capture the fact that agents may misreport their feature-values to the algorithm, we denote $i$'s *reported* feature vector $\tilde{w}(i) \in \mathcal{W}$, which may differ from $w(i)$. When $i$ misreports their feature vector, this changes the composition of the pool given to the selection algorithm; we denote the *reported* pool as $N_{-i} \cup \tilde{w}(i)$. Abusing notation slightly, if an entire *coalition* of agents $C \subset N$ misreports their feature vectors as $\tilde{\boldsymbol{w}} \in \mathcal{W}^{|C|}$, we denote the new pool as $N_{-C} \cup \tilde{\boldsymbol{w}}$. We will denote a manipulated pool as $\tilde{N}$ and the resulting manipulated instance as $\tilde{\mathcal{I}} := (\tilde{N}, k, \boldsymbol{\ell}, \boldsymbol{u})$.

As in past work [135], we assume that agents or coalitions can costlessly misreport any feature vector in $\mathcal{W}$, and they do so with full information about the selection algorithm and pool $N$. We consider three incentives for doing so: $\text{manip}_{\text{int}}$ captures how much a coalition can *increase the selection probability of someone internal to the coalition*; $\text{manip}_{\text{ext}}$ measures how much a coalition can *decrease the selection probability of someone external to the coalition*; and $\text{manip}_{\text{comp}}$ measures how many seats a coalition can, in expectation, misappropriate from another group. In the formal definitions of these measures, $\divideontimes := \max_{C\subseteq[n],|C|=c} \max_{\tilde{\boldsymbol{w}}\in\mathcal{W}^{|C|}}$ is shorthand for taking the worst possible coalition of size $c$ and worst possible strategic misreports of its members, and $\tilde{\mathcal{I}} := (N_{-C} \cup \tilde{\boldsymbol{w}}, k, \boldsymbol{\ell}, \boldsymbol{u})$ is the instance that results from a coalition $C$ misreporting as $\tilde{\boldsymbol{w}}$.

$$\begin{aligned}
\text{manip}_{\text{int}}(\mathcal{I}, \text{A}, c) &:= \divideontimes \max_{i\in C} \ \pi_i^{\text{A}}(\tilde{\mathcal{I}}) - \pi_i^{\text{A}}(\mathcal{I}),\\
\text{manip}_{\text{ext}}(\mathcal{I}, \text{A}, c) &:= \divideontimes \max_{i\notin C} \ \pi_i^{\text{A}}(\mathcal{I}) - \pi_i^{\text{A}}(\tilde{\mathcal{I}}),\\
\text{manip}_{\text{comp}}(\mathcal{I}, \text{A}, c) &:= \divideontimes \max_{(f,v)\in FV} \sum_{i:f(i)=v} \left(\pi_i^{\text{A}}(\tilde{\mathcal{I}}) - \pi_i^{\text{A}}(\mathcal{I})\right).
\end{aligned}$$

These measures can be interpreted as Nash equilibrium-style measures, capturing how much a coalition can gain if everyone else is truthful. From a formal game theoretic perspective, the argument of each maximum above can be thought of as a utility function, which an agent or coalition may aim to maximize.

*Fairness.* In accordance with past work on fairness in sortition [130, 131], we evaluate the fairness of a selection algorithm A in instance $\mathcal{I}$ by the *Maximin* fairness objective — that is, a *fairer* algorithm makes the minimum selection probability higher. Formally, A's *fairness* in $\mathcal{I}$ is

$$\text{fairness}(\mathcal{I}, \text{A}) := \mathit{Maximin}(\boldsymbol{\pi}^{\text{A}}(\mathcal{I})).$$

Because we want to guarantee fairness and manipulation robustness *simultaneously*, we will also want to measure fairness *when our pool may be corrupted by a manipulating coalition.* Defining $\divideontimes$ and $\tilde{\mathcal{I}}$ as before, the fairness of A in $\mathcal{I}$ *permitting a manipulating coalition of up to size c* is

$$\text{manip-fairness}(\mathcal{I}, \text{A}, c) := \divideontimes \mathit{Maximin}\left(\pi_i^{\text{A}}(\tilde{\mathcal{I}})\right).$$

*Transparency.* We now formally define the components of Flanigan et al. [131]'s algorithmic approach to transparency. For a given set of valid panels $\mathcal{K}$, let $m \in \mathbb{Z}^+$ and define the set of all *m-uniform lotteries* $\overline{\Delta}_m(\mathcal{K}) := (\mathbb{Z}^+/m)^{|\mathcal{K}|} \cap \Delta(\mathcal{K})$ as the set of all panel distributions in which all probabilities are multiples of $1/m$. $\overline{\mathbf{d}} \in \overline{\Delta}_m$ is called an $m$-uniform lottery due to the following key observation: $\overline{\mathbf{d}}$ contains exactly $m$ discrete blocs of $1/m$ probability mass, so we can sample a panel from $\overline{\mathbf{d}}$ via a *uniform lottery over m panels (with duplicates)* by numbering these probability blocs $1 \ldots m$, and then uniformly drawing a number from $[m]$. For example, if $m = 1000$, we can execute this uniform lottery physically, by drawing balls from bins corresponding to drawing 3 digits between 0 and 9, as in Figure 3 of Flanigan et al. [130]. Finally, we define $\overline{\Pi}_m(\mathcal{I}) :=$

$\{\boldsymbol{\pi}(\bar{\mathbf{d}})|\bar{\mathbf{d}} \in \overline{\Delta}_m(\mathcal{K})\}$ as the set of all selection probability assignments realizable by $m$-uniform lotteries in $\mathcal{I}$.

An $m$-uniform lottery is created by a *rounding algorithm* $\mathcal{R}_m : \Delta(\mathcal{K}) \to \overline{\Delta}_m(\mathcal{K})$, which is any (possibly randomized) mapping from a panel distribution into an $m$-uniform lottery. We apply a rounding algorithm $\mathcal{R}_m$ *in conjunction* with maximally fair algorithm E as follows: first, run E to compute panel distribution $\mathbf{d}^{\mathrm{E}}(\mathcal{I})$; then, use $\mathcal{R}_m$ to round $\mathbf{d}^{\mathrm{E}}(\mathcal{I})$ to an $m$-uniform lottery $\mathcal{R}_m(\mathbf{d}^{\mathrm{E}}(\mathcal{I}))$. Note that $\mathcal{R}_m \circ \mathrm{E}$ is itself a selection algorithm, mapping $\mathcal{I}$ to an $m$-uniform lottery $\mathbf{d}^{\mathcal{R}_m \circ \mathrm{E}}(\mathcal{I})$. We will define specific rounding algorithms as needed.

### 6.2.3 Key Assumption: Pool is Growing Linearly

Our positive results will be proven under the following assumption on the *true* pool $N$. For comparability, the truthful instances in our lower bounds will also satisfy this assumption.

**Assumption 6.2.4.** *In all instances $\mathcal{I} = (N, k, \boldsymbol{\ell}, \boldsymbol{u})$ where $N$ is the true pool,*

1. *There exists some constant $\kappa^* > 0$ such that $N_w \geq n\kappa^* + k$ for all $w \in \mathcal{W}_N$.*
2. *For all $i \in N$, there exists some $K \in \mathcal{K}$ such that $i \in K$.*

*(2)* is very weak, and holds in all real-world datasets we study. Conceptually, *(1)* requires two things: (i) that every vector group present in $N$ grows linearly as $N$ grows, and (ii) that $N$ is large enough so that there are at least $k$ people of each vector type. (i) is true in practice in expectation,[1] and should be true reliably as soon as $n$ is sufficiently large for variance effects to diminish. Given that $k$ is fixed relative to $n$, (ii) will hold with high probability for large enough $n$; however, it does not hold at the $n$ values in the instances we study.

## 6.3 Impossibilities for Maximin / Leximin, Nash, Minimax, and Linear$_\gamma$

To motivate our ultimate study of Goldilocks$_\gamma$, we now show that all previously-studied objectives — Maximin/Leximin, Nash and Minimax — all perform poorly on the dimension of either *Fairness* or *Manipulation Robustness*. Moreover, we surprisingly find the same issue with perhaps the most natural candidate objective for our purposes, Linear$_\gamma$, which combines our two goals linearly. Although similar impossibilities have already been proven for the previously-studied objectives, they were proven in the *relaxation* of the panel selection task studied by [135], so these negative results do not immediately carry over to our setting.

First, Theorem 6.3.1 shows that Maximin/Leximin and Nash are highly manipulable. While we defer the arithmetic aspects of the proof to Appendix E.2.1, we present and analyze the constructions of the pre and post-manipulation instances here. We will name them $\mathcal{I}^=$ and $\mathcal{I}_c^*$ respectively,

[1] The pool is recruited by inviting uniformly-sampled members of the population. If a vector group $w$'s average rate of entering the pool conditioned on being invited is some constant $r_w > 0$, then over the randomness of receiving and accepting the invitation, group $w$ will compose in expectation a $r_w$ fraction of the pool. If $r_w = 0$, then they will never enter the pool, and our growth assumption does not apply to them.

as we will use them again to prove lower bounds later in the paper. These instances are closely related to those used to prove lower bounds in [135].

**Theorem 6.3.1 (Lower Bound).** *For all $E \in \{\textsc{Maximin}, \textsc{Leximin}, \textsc{Nash}\}$, there exists $\mathcal{I}$ satisfying Assumption 6.2.4 such that*

$$\mathsf{manip}_{int}(\mathcal{I}, E, 2n/k + 1) = 1 - k/n.$$

*Proof.* Let there be two features, $f_1$ and $f_2$, and suppose that they are binary, so $V_{f_1} = \{0, 1\}$ and $V_{f_2} = \{0, 1\}$. Then, there are four unique possible feature vectors: $\mathcal{W} = \{00, 01, 10, 11\}$. Set the quotas $\boldsymbol{\ell}, \boldsymbol{u}$ so that $\ell_{f_1,0} = \ell_{f_2,0} = u_{f_1,0} = u_{f_2,0} = k/2$, meaning that any valid panel will be perfectly split between values on both features. Fix $k$ (we will set it carefully in the proof).

**Instance $\mathcal{I}^=$.** Define instance $\mathcal{I}^= := (N, k, \boldsymbol{\ell}, \boldsymbol{u})$ be our truthful instance, which inherits the features, quotas, and $k$ defined above. We define the pool $N$ such that $N_{00} = N_{11} = n/2$ and $N_{01} = N_{10} = 0$.

**Observation 6.3.2.** We make two key deductions about $\mathcal{I}^=$.

6.3.2.1 $\mathcal{I}^=$ satisfies Assumption 6.2.4 with associated $\kappa^*$ so long as $n \geq 2k/(1 - \kappa^*)$; we can set $n$ and $k$ such that this is the case.

6.3.2.2 $k/n\mathbf{1} \in \Pi(\mathcal{I}^=)$. By observation, we can give all $i \in N$ equal selection probability by randomizing uniformly over all panels containing $k/2$ agents with vector 00 and $k/2$ agents with vector 11. It follows that for any conditionally equitable algorithm E, $\boldsymbol{\pi}^E(\mathcal{I}^=) = k/n\mathbf{1}$.

**A manipulating coalition.** Now, let $C \subseteq N$ be a coalition of size $c$, composed of $c/2$ agents with $w(i) = 00$ and $c/2$ agents with $w(i) = 11$. Let some $i^* \in C$ misreport the vector $\tilde{w}(i^*) = 01$; for all other agents $i \in C \setminus \{i^*\}$, let $\tilde{w}(i) = 10$. We name the resulting instance $\mathcal{I}_c^*$, as we will use this construction later on.

**Instance $\mathcal{I}_c^*$.** The resulting instance is $\mathcal{I}_c^* := (\tilde{N}, k, \boldsymbol{\ell}, \boldsymbol{u})$, where $\tilde{N}_{00} = \tilde{N}_{11} = \frac{n-c}{2}$, $\tilde{N}_{10} = c - 1$, and $\tilde{N}_{01} = 1$ and $k, \boldsymbol{\ell}, \boldsymbol{u}$ are the same as in $\mathcal{I}^=$. Now, we make several useful observations about $\mathcal{I}_c^*$, which we will reference each time we analyze it:

**Observation 6.3.3.** We deduce the following about $\mathcal{I}_c^*$, whose associated valid panels we call $\mathcal{K}_c^*$:

6.3.3.1 $\mathcal{K}_c^*$ contains two types of valid panels:

- **Type 1:** Panels containing $k/2$ agents with vector 00 and $k/2$ agents with vector 11
- **Type 2:** Panels containing $k/2 - 1$ agents with vector 00, $k/2 - 1$ agents with vector 11, $i^*$, and 1 agent with vector 10.

6.3.3.2 Fix any $\mathbf{d} \in \Delta(\mathcal{K}_c^*)$, and let $d_1, d_2$ represent the total probability $\mathbf{d}$ places on panels of Types 1 and 2, respectively. Then, by simply dividing the expected panel seats given to

agents with each vector $w$ divided by the total number of pool members with vector $w$, the resulting selection probabilities (assumed to be anonymous) in terms of $d_1, d_2$ are:

$$\mathsf{p}_{00} = \mathsf{p}_{11} = d_1 \frac{k/2}{(n-c)/2} + d_2 \frac{k/2-1}{(n-c)/2}, \qquad \mathsf{p}_{10} = d_2 \frac{1}{c-1}, \qquad \mathsf{p}_{01} = d_2. \tag{6.1}$$

Equation (6.1) reveals that in any realizable selection probability allocation over panel types 1 and 2, $\mathsf{p}_{01}$ *must be* $c-1$ *times as large as* $\mathsf{p}_{10}$. Because any selection algorithm must suffer this gap, then the question remaining is: how do Maximin, Leximin, and Nash prioritize the maximum and minimum selection probabilities when positioning this gap? As we show in the full proof in Appendix E.2.1, Maximin, Leximin, and Nash, all being highly sensitive to *low* probabilities, will position this gap as high as possible to mitigate low probabilities, thereby driving $i^*$'s probability all the way to roughly $ck/n$. □

Moving onto Minimax, we expect this algorithm to perform poorly with respect to fairness *Fairness*, becauseit considers only the highest probabilities and can thus unnecessarily give some agents selection probability 0. Observing that Minimax is the special case of $Linear_\gamma = \max(\boldsymbol{\pi}) - \gamma \cdot \min(\boldsymbol{\pi})$ where $\gamma = 0$, we will prove a negative result on Minimax in the course of proving a negative result for the objective $Linear_\gamma$. To show that $Linear_\gamma$ does not adequately control high and low selection probabilities, we will show that no matter how we set $\gamma$, we can find an instance where this objective is either arbitrarily unfair (low $\gamma$), essentially arbitrarily manipulable (high $\gamma$), or unfair to a degree that we will later show is sub-optimal (intermediate $\gamma$).

**Theorem 6.3.4 (Lower Bound).** *For all $\gamma \in [0, 1)$, there exists $\mathcal{I}$ satisfying Assumption 6.2.4 such that*

$$\mathsf{fairness}(\mathcal{I}, \textsc{Linear}_\gamma) = 0.$$

*For all $\gamma \in [1, n/3-1)$, there exists an instance $\mathcal{I}'$ satisfying Assumption 6.2.4 in which*

$$\mathsf{manip\text{-}fairness}(\mathcal{I}', \textsc{Linear}_\gamma, n/6) = \tfrac{9k}{2(n^2-9)} \in O(k/n^2).$$

*For all $\gamma \in [n/3-1, \infty)$, there exists an instance $\mathcal{I}'$ satisfying Assumption 6.2.4 such that*

$$\mathsf{manip}(\mathcal{I}', \textsc{Linear}_\gamma, n/6) = 1 - k/n.$$

*Proof sketch.* We defer the full proof to Appendix E.3.1, but the main ideas will give useful intuition. The truthful pool is the same across all three cases of the proof: $N_{00} = N_{11} = n/3$ and $N_{01} = N_{10} = n/6$. The quotas we set, however, differ across cases; we describe each one separately.

*Large/Intermediate $\gamma$.* In this case, our quotas are set as in $\mathcal{I}_c^*$, so $\ell_{f_1,0} = \ell_{f_2,0} = u_{f_1,0} = u_{f_2,0} = k/2$. For both large and intermediate $\gamma$, we pursue bounds assuming manipulation by a coalition of size $n/6$. The coalition $C$ is the same in both cases: it consists of all agents with vector 01. Of these agents, one agent $i^*$ reports $\tilde{w}(i^*) = w(i^*) = 01$, and the rest $i \in C \setminus \{i^*\}$ misreport $\tilde{w}(i) = 10$. The resulting instance is exactly $\mathcal{I}^*_{n/6}$, as defined in the proof of Theorem 6.3.1. Per Observation 6.3.3, there is a fundamental gap in selection probabilities of $n/6-1$ between agents with vector 10 and

01. Intuitively, to control maximum and minimum probabilities to the greatest extent possible, we would like to place this multiplicative gap squarely over $k/n$, so probabilities are bounded in $[k/n\sqrt{n}, k\sqrt{n}/n]$. We conclude these two cases by showing that neither intermediate nor large $\gamma$ places the gap as we want. When $\gamma \geq n/3 - 1$, we find that $\text{LINEAR}_\gamma$ acts like MAXIMIN, prioritizing low probabilities too much, to the point that $i^*$ receives probability nearly 1 — a problem for *Manipulation Robustness*. In contrast, when $\gamma < n/3 - 1$, we find that $\gamma$ does not place *enough* weight on low probabilities, and thus places this probability gap too low, guaranteeing minimum probabilities of at best order $k/n^2$ — a problem for *Fairness*.

*Small $\gamma$.* Finally, for small $\gamma \in [0, 1]$ we consider the fairness of $\text{LINEAR}_\gamma$ in the instance with the pool $N$ above, but now with *skewed* quotas: we let $\ell_{f_1,0} = 2k/3$ and $\ell_{f_2,0} = k/3$. Then, there are more panel seats available for agents with value 0 for the first feature and value 1 for the second feature. This skew means that we must give agents with vector 01 higher selection probability than those with 10, where this fundamental gap is additive and of order $k/n$. As $\gamma$ gets large, $\text{LINEAR}_\gamma$ starts behaving like MINIMAX (indeed, $\lim_{\gamma \to 0} Linear_\gamma = Minimax$) — prioritizing ensuring that the maximum selection probability $\text{LINEAR}_\gamma$ places this gap very low, giving some agents 0 probability. □

Although our lower bound for intermediate $\gamma$ is already adequate to show a separation between $\text{LINEAR}_\gamma$ and our proposed algorithm $\text{GOLDILOCKS}_\gamma$, we suspect that a more elaborate construction can permit an even more extreme lower bound for intermediate $\gamma$ (to see why, note the discontinuity in our lower bounds at $\gamma = 1$; we expect a truly worst-case construction to allow continuous bounds across this juncture).

### 6.4 Analysis of Goldilocks

Finding all previous equality objectives insufficient to control high and low selection probabilities to the degree we want, we now study $Goldilocks_\gamma(\boldsymbol{\pi}) = n/k \max(\boldsymbol{\pi}) + \gamma \frac{1}{n/k \min(\boldsymbol{\pi})}$. At the cost of being nonlinear, this objective addresses this issue with $Linear_\gamma$ by *more steeply* penalizing decreases in the minimum probability relative to the maximum. We focus on GOLDILOCKS with $\gamma = 1$, as this will be sufficient both in theory and in practice. We now prove our main result:

**Theorem 6.4.1 (Upper Bound).** *For any instance $\mathcal{I}$ in which $N$ satisfies Assumption 6.2.4 with $\kappa^*$, then for any constant $\kappa \in (0, \kappa^*)$ and all $c \leq \kappa n/\sqrt{k}$,*

$$\begin{aligned}
\text{manip}_{int}(\mathcal{I}, \text{GOLDILOCKS}_1, c) &\in O\left(k^2\sqrt{c}/n\right) \\
\text{manip}_{ext}(\mathcal{I}, \text{GOLDILOCKS}_1, c) &\in O\left(k/n \cdot \left(1 - 1/(k\sqrt{c})\right)\right) \\
\text{manip}_{comp}(\mathcal{I}, \text{GOLDILOCKS}_1, c) &\in \max_{(f,v) \in FV} (u_{f,v} - \ell_{f,v}) + O\left(k^2 c\sqrt{c}/n\right) \\
\text{manip-fairness}(\mathcal{I}, \text{GOLDILOCKS}_1, c) &\in \Omega\left(1/n \cdot 1/\sqrt{c}\right).
\end{aligned}$$

To get some intuition for the meaning of this result before proving it, recall the problem we identified with $Linear_\gamma$ for intermediate $\gamma$, where a coalition of size $n/6$ induced an order $n$ gap in agents' selection probabilities. We wanted $Linear_\gamma$ to place this gap directly over $k/n$, guaranteeing probabilities in order $[1/n\sqrt{n}, \sqrt{n}/n]$ (treating $k$ as constant relative to $n$). While $Linear_\gamma$ failed to do this, Theorem 6.4.1 alludes to the fact that GOLDILOCKS$_1$ will succeed: in that instance, with coalition of size $c$ of order $n$, GOLDILOCKS$_1$ will achieve manip-fairness of order $1/n\sqrt{n}$ and manip$_{int}$ (a proxy for the maximum probability) of order $\sqrt{n}/n$, as desired.

### 6.4.1 Proof of Theorem 6.4.1

The proof of Theorem 6.4.1 will rely on three technical lemmas. These lemmas center around the instance-wise parameter $\delta(\mathcal{I})$, which conceptually measures the quality of feasible solutions that exist in instance $\mathcal{I}$. Formally, we measure the quality of a given $\boldsymbol{\pi}$ by two values,

$$\delta_{below}(\boldsymbol{\pi}) := \frac{k/n}{\min(\boldsymbol{\pi})} \qquad \text{and} \qquad \delta_{above}(\boldsymbol{\pi}) := \frac{\max(\boldsymbol{\pi})}{k/n},$$

which respectively capture how much any selection probability in $\boldsymbol{\pi}$ deviates below and above $k/n$. Then, we define $\delta(\mathcal{I})$ to capture the quality of the "best" feasible solution available in $\mathcal{I}$:

$$\delta(\mathcal{I}) := \min_{\boldsymbol{\pi} \in \Pi(\mathcal{I})} \max\{\delta_{below}(\boldsymbol{\pi}), \delta_{above}(\boldsymbol{\pi})\}.$$

Now, we state and prove our key lemmas. First, Lemma 6.4.2 gives instance-dependent bounds on the maximum and minimum selection probability given by GOLDILOCKS. The instance-dependence of these bounds is reflected in their dependence on $\delta(\mathcal{I})$. This is really our key lemma; it shows that GOLDILOCKS$_1$ will recover among the best solutions available in any given instance.

**Lemma 6.4.2.** *For all instances $\mathcal{I}$, $\boldsymbol{\pi}^{\text{GOLDILOCKS}_1}(\mathcal{I}) \in \left[\frac{k/n}{2\,\delta(\mathcal{I})},\ k/n \cdot 2\delta(\mathcal{I})\right]^n$.*

*Proof.* Fix instance $\mathcal{I}$ and the selection probability assignment $\boldsymbol{\pi}' \in \Pi(\mathcal{I})$ such that $\max\{\delta_{below}(\boldsymbol{\pi}'), \delta_{above}(\boldsymbol{\pi}')\} = \delta(\mathcal{I})$. For this proof, we will use the shorthand $\delta_{below} = \delta_{below}(\boldsymbol{\pi}')$, $\delta_{above} = \delta_{above}(\boldsymbol{\pi}')$ and $\boldsymbol{\pi}^* = \boldsymbol{\pi}^{\text{GOLDILOCKS}_1}(\mathcal{I})$.

First, we upper bound the optimal objective value using our feasible solution $\pi$:

$$\begin{aligned} \text{GOLDILOCKS}_1(\boldsymbol{\pi}^*) \le n/k\ \max(\boldsymbol{\pi}') + \frac{1}{n/k\ \min(\boldsymbol{\pi}')} &= \delta_{above} + \delta_{below} \\ &\le 2\max\{\delta_{below}, \delta_{above}\} = \delta(\mathcal{I}). \end{aligned} \tag{6.2}$$

Now, suppose there exists $i \in [n]$ such that $\boldsymbol{\pi}^*_i > k/n \cdot 2\delta(\mathcal{I})$. Then,

$$\text{GOLDILOCKS}_1(\boldsymbol{\pi}^*) > n/k \cdot k/n \cdot 2\delta(\mathcal{I}) + 0 = 2\delta(\mathcal{I}),$$

which is a contradiction to (6.2). We conclude that $\pi^*_i \le k/n \cdot 2\delta(\mathcal{I})$ for all $i \in [n]$. Likewise, suppose that there exists $i \in [n]$ such that $\boldsymbol{\pi}^*_i < k/n \cdot 1/2\delta(\mathcal{I})$. Then,

$$\text{GOLDILOCKS}_1(\boldsymbol{\pi}^*) > 0 + \frac{1}{n/k \cdot k/n \cdot 1/2\delta(\mathcal{I})} = 2\delta(\mathcal{I}).$$

This is again a contradiction to (6.2), and we conclude that $\pi_i^* \geq k/n \cdot 1/2\delta(\mathcal{I})$ for all $i \in [n]$. □

Now, we want to use Lemma 6.4.2 to draw conclusions about the *absolute* maximum and minimum probabilities guaranteed by GOLDILOCKS. The functional form of Lemma 6.4.2 dictates that doing so requires characterizing what $\delta(\mathcal{I})$ can be. First, in Lemma 6.4.3, we answer this question for *truthful* instances. We show that given Assumption 6.2.4 on the true pool, if agents are truthful, then $\delta(\mathcal{I}) \in O(1)$. This is the best case scenario, and GOLDILOCKS$_1$ will make use of it: by Lemma 6.4.2, in this case GOLDILOCKS$_1$ will give all agents probabilities in $\Theta(k/n)$.

**Lemma 6.4.3.** *In instance $\mathcal{I}$, if $N$ satisfies Assumption 6.2.4, then $\delta(\mathcal{I}) \in O(1)$.*

We defer the proof of this claim to Appendix E.4.3, because it uses notation and machinery that will not be useful throughout the rest of the paper, and we will drag constants through the argument. We sketch the proof here: First, showing that we can give all agents $O(k/n)$ probabilities is simple: all vector groups present in the pool are of size order $n$ by Assumption 6.2.4, and the total probability given to all agents in any given vector group cannot exceed $k$. Spreading at most $k$ probability over $n$ members of a vector group ensures that no agent gets more than $O(k/n)$ selection probability. Proving that probabilities are lower-bounded as $\Omega(k/n)$ is less obvious, and has to do with the limited number of possible feature vectors in the pool — a consequence of Assumption 6.2.4.

Although excellent solutions are possible — and recovered by GOLDILOCKS$_1$ — when agents are truthful, such solutions are no longer necessarily available when agents or coalitions can misreport their vectors. For example, take the example $\mathcal{I}_c^*$, which we know can arise from a manipulating coalition of size $c$; in that instance, one agent must receive $(c-1)$ times as much probability as another agent (Equation (6.1)), meaning that $\delta(\mathcal{I}_c^*)$ must be at least $\sqrt{c-1}$. We formalize this lower bound in Appendix E.4.4. Fortunately, we now show that although coalitions can drive up $\delta$, this simple lower bound is actually tight (up to a factor of $k$):

**Lemma 6.4.4.** *If in instance $\mathcal{I} = (N, k, \boldsymbol{\ell}, \boldsymbol{u})$, $N$ satisfies Assumption 6.2.4, then for any constant $\kappa \in (0, \kappa^*)$ and coalition size $c \leq n\kappa/\sqrt{k}$,*

$$\delta(N_{-C} \cup \tilde{\boldsymbol{w}}, k, \boldsymbol{\ell}, \boldsymbol{u}) \in O(k\sqrt{c}) \quad \text{for all } C \subseteq [n], |C| = c \text{ and } \tilde{w} \in \mathcal{W}^c.$$

This is our most technically demanding result, and we defer the formal proof to Appendix E.4.5, because it again requires significant notation and machinery. The key idea is that we can modify the solution in the truthful instance, redistributing some probability from truthful agents to define a panel distribution in the post-manipulation instance. More concretely, the steps of the proof are as follows: first, we show that all *types* of panels (defined by the fraction of seats taken up by each feature vector) that exist in the truthful instance still exist in the post-manipulation instance. This is the critical consequence of Assumption 6.2.4: that although a coalition can shift the space of panel types, *it can only expand it.* We then modify our solution in the truthful instance, which we know gives all agents $\Theta(k/n)$ selection probability by Lemma 6.4.3. We modify this solution by shifting a small amount of probability away from the panel types available in

the original solution, and onto panels containing the coalition members. We precisely tune the amount of probability we redistribute in order to ensure that, although there may be requisite gaps in coalition members' probabilities, we place these gaps to be symmetric around $k/n$. We then argue that the panel distribution we construct gives all agents $i \in [n]$ selection probabilities within $\pi_i \in \left[\Omega\left(1/\sqrt{c}n\right), O\left(k+k\sqrt{c}/n\right)\right]$. This gives us the desired bound on $\delta(\mathcal{I})$.

**Finally, we apply Lemmas 6.4.2, 6.4.3, and 6.4.4 to complete the proof.** Fix a truthful instance $\mathcal{I}$ with pool $N$; let it satisfy Assumption 6.2.4 with constant $\kappa^*$. Let $\boldsymbol{\pi}^* = \boldsymbol{\pi}^{\text{Goldilocks}}(\mathcal{I})$ denote the optimal probabilities given by $\text{Goldilocks}_1$ on this instance. By Lemma 6.4.3, we know that $\delta(\mathcal{I}) \in O(1)$, and hence by Lemma 6.4.2, we have that $\pi_i^* \in \Theta(k/n)$ for all $i \in [n]$. Now, fix an arbitrary coalition $C \subset N$ of size $c \leq \kappa n/\sqrt{k}$ for some constant $\kappa \in (0, \kappa^*)$, and suppose they misreport vectors $\tilde{w}(i) | i \in C$, creating a new instance $\tilde{\mathcal{I}}$.

Let $\tilde{\pi}^* = \pi^{\text{Goldilocks}}(\tilde{\mathcal{I}})$ be shorthand for the optimal probabilities given by Goldilocks on this post-manipulation instance. By Lemma 6.4.4, we know that no matter what the members of this coalition misreport, $\delta(\tilde{\mathcal{I}}) \in O(k\sqrt{c})$. By Lemma 6.4.2, the probabilities in $\tilde{\boldsymbol{\pi}}^*$ are bounded as

$$\tilde{\pi}_i^* \in \left[\Omega(1/n\sqrt{c}), O(k^2\sqrt{c}/n)\right] \quad \text{for all } i \in [n].$$

From this we can draw conclusions about the extent to which $\text{Goldilocks}_1$ satisfies *Manipulation Robustness* and *Fairness* in this instance. The maximum gain in probability for any individual agent can be upper bounded by the maximum marginal in $\tilde{\pi}^*$. It follows that

$$\text{manip}_{int}(\mathcal{I}, \text{Goldilocks}_1, c) \in O\left(k^2/n \cdot \sqrt{c}\right).$$

The largest probability decline for any agent, regardless of whether they are in $C$, is bounded in terms of the minimum guaranteed probability as $O(k/n - 1/n\sqrt{c}) = O(k/n \cdot (1 - 1/k\sqrt{c}))$. Hence,

$$\text{manip}_{ext}(\mathcal{I}, \text{Goldilocks}_1, c) \in O(k/n \cdot (1 - 1/k\sqrt{c})).$$

Analyzing $\text{manip}_{comp}$ requires a little more care. Let $\tilde{f} : N \to V_f$ map each agent $i$ to their *reported* feature. Fix an $f, v$; we will divide the quantity we want to bound into two quantities, where the first represents the probability garnered among people who actually report value $v$ for feature $f$ in the post-manipulation pool, and the second represents probability garnered among people who do *not* report value $v$ for feature $f$ in the post-manipulation pool, but truly possess that feature.

$$\begin{aligned}
\sum_{i:f(i)=v} \left(\pi_i^{\text{A}}(\tilde{\mathcal{I}}) - \pi_i^{\text{A}}(\mathcal{I})\right) &= \sum_{i:f(i)=v \wedge \tilde{f}(i)=v} \left(\pi_i^{\text{A}}(\tilde{\mathcal{I}}) - \pi_i^{\text{A}}(\mathcal{I})\right) + \sum_{i:f(i)=v \wedge \tilde{f}(i)\neq v} \left(\pi_i^{\text{A}}(\tilde{\mathcal{I}}) - \pi_i^{\text{A}}(\mathcal{I})\right) \\
&\leq u_{f,v} - \ell_{f,v} + \sum_{i:f(i)=v \wedge \tilde{f}(i)\neq v} \left(\pi_i^{\text{A}}(\tilde{\mathcal{I}}) - \pi_i^{\text{A}}(\mathcal{I})\right) \\
&\leq u_{f,v} - \ell_{f,v} + O(k^2 c\sqrt{c}/n),
\end{aligned}$$

Here, the final step holds because every $i$ contributing to the second sum must be in the manipulating coalition; there can be at most $c$ of these $i$'s, and each of them can have at most $O(k^2\sqrt{c}/n)$

total probability in $\tilde{\mathcal{I}}$ (Lemmas 6.4.2 and 6.4.4). Therefore, we have that

$$\mathsf{manip}_{comp} \in \max_{(f,v)\in FV} u_{f,v} - \ell_{f,v} + O\left(k^2 c\sqrt{c}/n\right).$$

Finally, we get manip-fairness directly from the lower bound on marginals in $\tilde{\pi}^*$ given by Lemmas 6.4.2 and 6.4.4 together, and we conclude the proof:

$$\mathsf{manip\text{-}fairness} \in \Omega(1/n\sqrt{c}).$$

### 6.4.2 On the tightness and implications of Theorem 6.4.1

First, we give tight lower bounds on $\mathsf{manip}_{int}$ and manip-fairness (up to $k$), representing the ideals *manipulation robustness* and *fairness*. These lower bounds are proven with truthful instance $\mathcal{I}^{=}$ and manipulated instance $\mathcal{I}_c^*$; the full proof is in Appendix E.4.6.

**Theorem 6.4.5 (Lower Bound).** $\mathsf{manip}_{int}(\mathcal{I}^{=}, \textsc{Goldilocks}_1, c) = k/n \cdot (\sqrt{c-1} - 1)$ *and* $\mathsf{manip\text{-}fairness}(\mathcal{I}^{=}, \textsc{Goldilocks}_1, c) = k/n \cdot 1/\sqrt{c-1}$.

Now, we discuss our bounds in Theorem 6.4.1 item-wise. First, our upper bound on $\mathbf{manip}_{int}$ is encouraging: even up to linear-size coalitions, this upper bound declines at a rate of $O(1/\sqrt{n})$, meaning we can improve $\textsc{Goldilocks}_1$'s manipulation robustness by recruiting more pool members. This upper bound also shows a clear separation between $\textsc{Goldilocks}_1$ and Maximin, Leximin, Nash and $\textsc{Linear}_\gamma$ with large $\gamma$; for all those objectives, linear-size coalitions could drive the selection probability of a member of a manipulating coalition up to (essentially) 1 (Theorem 6.3.1, Theorem 6.3.4).

Our positive result for **manip-fairness** is also encouraging, showing that even linear-size coalitions can drive the minimum probability only as low as order $\frac{k}{n\sqrt{n}}$. This bound shows a clear separation between $\textsc{Goldilocks}_1$ versus Minimax and $\textsc{Linear}_\gamma$ with $\gamma < n/3 - 1$, where linear-size manipulating coalitions can drive the fairness down to $O(k/n^2)$ or even 0 (Theorem 6.3.4). We remark that while we proved in Theorem 6.4.5 that our positive result on manip-fairness is tight, in that lower-bound instance, it is *only members of the coalition* who receive probability $\frac{k/n}{\sqrt{c-1}}$—all truthful agents actually receive probability in $\Theta(k/n)$ in the post-manipulation instance. With this in mind, it might be tempting to say that we want to be fair *only to honest agents*, in which case an even stronger positive result on manip-fairness might be possible. However, we caution that while such a bound could hold in theory under Assumption 6.2.4, trying to detect and protect only truthful agents would be risky in practical instances where Assumption 6.2.4 is unlikely to exactly hold.

Finally, despite the foregoing good news, one may notice something concerning about our upper bound on $\mathbf{manip}_{comp}$: if $c \in \Omega(n^{2/3})$, this bound is *constant* and could be as large as $k$. Conceivably, then, a coalition could misappropriate the entire panel. However, it is actually not at all clear that this upper bound is tight, at least based on the lower bound constructions we have used so far. In all of our lower bounds, manipulating coalitions drive gaps in selection probabilities by dividing into two subgroups who report *complementary feature vectors* 10 and 01. While this

can drive up the probabilities of agents in one such group at the expense of the other, the *total probability* garnered by each feature-value group must remain roughly commensurate due to the quotas. Because these feature vectors are complements, the total probability misappropriated from any feature-value by such strategies is actually roughly 0. Thus, to either tighten our upper bound on $\mathsf{manip}_{comp}$ or prove a useful lower bound, new ideas are required, and we leave this for future work.

### 6.5 Analysis of *Transparent* Goldilocks

We now seek a rounding algorithm $\mathcal{R}_m$ that allows us to extend our guarantees on Goldilocks$_1$ to $\mathcal{R}_m \circ$ Goldilocks$_1$. For this, we directly apply the rounding algorithms studied in Flanigan et al. [131], which are guaranteed to round any panel distribution to an $m$-uniform lottery while changing no agent's selection probability by more than a bounded amount:

**Theorem 6.5.1** (Thms 3.2 and 3.3, [131])**.** *For all $\mathcal{I}$ and $m \in \mathbb{Z}^+$, there exists an $\mathcal{R}_m$ such that for all $\mathbf{d}$ with corresponding $\boldsymbol{\pi}$, it holds for $\mathcal{R}_m(\mathbf{d}) = \overline{\mathbf{d}}$ with corresponding $\overline{\boldsymbol{\pi}}$ that*

$$\|\boldsymbol{\pi} - \overline{\boldsymbol{\pi}}\|_\infty \leq O\left(\min\left\{k, \sqrt{|\mathcal{W}_N| \log(|\mathcal{W}_N|)}\right\}/m\right).$$

However, one may notice a problem with directly applying this bound to our setting: by Theorem 6.4.5 that the minimum probability may be dropping at a rate of $1/n\sqrt{c}$ in $n$, which can be as low as order $1/n\sqrt{n}$. In contrast, the above bound does not shrink in $n$, meaning that as $n$ grows (which is beneficial for manipulation robustness), the minimum probability given to any agent will eventually become so small that this upper bound will be larger than the minimum probability in the pre-rounded instance, resulting in a fairness guarantee of 0. We solidify this concern by proving a lower bound showing that this dependency on $m$ is indeed unavoidable: Proposition 6.5.2 shows that in the worst case, rounding the Goldilocks-optimal solution (using that Goldilocks is conditionally equitable) may result in decreasing the minimum probability by up to $\sqrt{k}/m$:

**Proposition 6.5.2.** *There exists an instance $\mathcal{I}$ such that for all conditionally equitable algorithms $E$ and for all $\bar{\boldsymbol{\pi}} \in \overline{\Pi}_m(\mathcal{I})$, $\min\left(\boldsymbol{\pi}^E(\mathcal{I})\right) - \min\left(\overline{\boldsymbol{\pi}}\right) \geq \sqrt{k}/m$.*

*Proof.* We defer the construction of the instance to Flanigan et al. [131], about which they show that for all $\bar{\boldsymbol{\pi}} \in \overline{\Pi}_m(\mathcal{I})$, $Maximin(\boldsymbol{\pi}) - Maximin(\overline{\boldsymbol{\pi}}) \geq \sqrt{k}/m$. We simply generalize their result to all conditionally equitable objectives with the following simple observation. In their construction, the original instance $\mathcal{I}$ is such that $k/n\mathbf{1}^n \in \Pi(\mathcal{I})$, and in the original panel distribution they consider in fact implies $\boldsymbol{\pi} = k/n\mathbf{1}^n$. By the definition of conditional equitability, $\boldsymbol{\pi}$ must be maximally equal with respect to any conditionally equitable objective $\mathcal{E}$. □

In order to avoid this issue, we need $m$ to grow at least at a rate of $\Omega(n\sqrt{n})$. The good news is that in practice, it is much lower cost to scale up $m$ than to scale up $n$. For example, scaling up $n$ by a factor of 10 requires sending out 10 times as many letters; multiplying $m$ by 10 just requires

adding another lottery bin, thereby permitting the panels to be numbered 0000 - 9999 instead of 000 - 999. Thus, we assume that $m \geq n\sqrt{n}$. Finally, we are permitting manipulation here, which means that the number of unique vectors in the pool (a parameter of the bound we will apply) may be larger than $|\mathcal{W}_N|$. We thus use the observation that a manipulating coalition of size $c$ can change the number of unique feature vectors by at most $c$, so $\mathcal{W}_{\tilde{N}} \leq \mathcal{W}_N + c$. By combining this observation, Theorem 6.5.1, and Theorem 6.4.1, we conclude the following bounds on the simultaneous manipulation robustness and fairness of $\mathcal{R}_m \circ \textsc{Goldilocks}_1$ for any $m \geq n\sqrt{n}$:

**Theorem 6.5.3 (Upper Bound).** *There exists an $\mathcal{R}_m$, $m \geq n\sqrt{n}$ such that for all $\mathcal{I}$,*

$$\begin{aligned}
\mathsf{manip}_{int}(\mathcal{I}, \mathcal{R}_m \circ \textsc{Goldilocks}_1, c) &\in O\left(k\sqrt{c}/n + \min\left\{k, \sqrt{(|\mathcal{W}_N|+c)\log(|\mathcal{W}_N|+c)}\right\}/n\sqrt{n}\right), \\
\mathsf{manip}_{ext}(\mathcal{I}, \mathcal{R}_m \circ \textsc{Goldilocks}_1, c) &\in O\left(k(1-1/\sqrt{c})/n + \min\left\{k, \sqrt{(|\mathcal{W}_N|+c)\log(|\mathcal{W}_N|+c)}\right\}/n\sqrt{n}\right), \\
\mathsf{manip}_{comp}(\mathcal{I}, \mathcal{R}_m \circ \textsc{Goldilocks}_1, c) &\in \max_{(f,v)\in FV}(u_{f,v} - \ell_{f,v}) \\
&\quad + O\left(kc\sqrt{c}/n + \min\left\{k, \sqrt{(|\mathcal{W}_N|+c)\log(|\mathcal{W}_N|+c)}\right\}/n\sqrt{n}\right), \\
\mathsf{manip\text{-}fairness}(\mathcal{I}, \mathcal{R}_m \circ \textsc{Goldilocks}_1, c) &\in \Omega\left(k/n\sqrt{c} - \min\left\{k, \sqrt{(|\mathcal{W}_N|+c)\log(|\mathcal{W}_N|+c|)}\right\}/n\sqrt{n}\right).
\end{aligned}$$

## 6.6 Empirical Evaluation

**Instances.** We analyze 9 instances of real-world panel selection data identified only by number (their sources are anonymized). The relevant properties of these instances are in Appendix E.5.2.

**Algorithms.** We evaluate four maximally fair selection algorithms Leximin, Minimax, Goldilocks$_1$[1], Nash. For our most computationally-intensive experiments, we replace Leximin with Maximin, as it runs much faster on large instances but behaves similarly with respect to the properties we aim to test. We also analyze the selection algorithm Legacy, which is a greedy heuristic that was used widely in practice, and serves here as a benchmark representing greedy algorithms that remain in use (see Appendix E.5.3 for details). In some analyses, we consider only a key subset of these algorithms: Minimax, Leximin, and Goldilocks$_1$. When optimizing *Goldilocks* via Flanigan et al. [130]'s algorithmic framework, we run into the issue that it is not differentiable, as is needed to apply the framework. We thus instead implement the following differentiable version of *Goldilocks*$_1$:

$$n/k \cdot \left(\sum_{i\in[n]} \pi_i^p\right)^{1/p} + 1/n/k \cdot \left(\sum_{i\in[n]} 1/\pi_i^p\right)^{1/p}.$$

In Appendix E.5.4, we show that this objective converges quickly to Goldilocks$_1$ as $p \to \infty$ (Proposition E.5.1); we characterize this convergence rate precisely in Lemma E.5.2. We set $p =$

[1]It seems plausible that an *instance-specfic* setting of $\gamma$ might perform better than a fixed $\gamma$ across instances. We additionally define and evaluate two natural instance-specific definitions of $\gamma$, but find that they make very little difference to the performance of Goldilocks. These $\gamma$ values are defined in Appendix E.5.1, and the corresponding results are in Table E.1.

100 in our analysis. Appendix E.5.4 also describes our implementation of Flanigan et al. [130]'s framework.

### 6.6.1 Maxes and mins

We first compare algorithms in their ability to control the maximum and minimum probability simultaneously. In Table 6.1, for each algorithm A we report entries of the form $\left(\frac{\min(\pi^{\text{A}})}{\min(\pi^{\text{Maximin}})}, \frac{\max(\pi^{\text{A}})}{\max(\pi^{\text{Minimax}})}\right)$. In words, we are reporting the multiplicative approximations achieved by A to the optimal minimum probability (given by Maximin) and optimal maximum probability (given by Minimax).

| | **Equality Notions** | | | | | |
|---|---|---|---|---|---|---|
| **Instances** | Legacy | Minimax | Maximin | Leximin | Nash | Goldilocks$_1$ |
| 1 | (0.0, 1.14) | (0.0, 1.0) | (1.0, 2.0) | (1.0, 2.0) | (0.62, 2.0) | (0.73, 1.14) |
| 2 | (0.03, 1.01) | (0.0, 1.0) | (1.0, 1.33) | (1.0, 1.33) | (0.67, 1.33) | (0.9, 1.0) |
| 3 | (0.0, 1.0) | (0.0, 1.0) | (1.0, 1.0) | (1.0, 1.0) | (0.61, 1.0) | (0.99, 1.0) |
| 4 | (0.01, 1.0) | (0.0, 1.0) | (1.0, 1.0) | (1.0, 1.0) | (0.61, 1.0) | (0.97, 1.0) |
| 5 | (0.0, 1.02) | (0.0, 1.0) | (1.0, 1.17) | (1.0, 1.17) | (0.57, 1.17) | (0.92, 1.0) |
| 6 | (0.66, 1.11) | (0.25, 1.0) | (1.0, 1.5) | (1.0, 1.11) | (0.9, 1.08) | (1.0, 1.09) |
| 7 | (0.0, 2.18) | (0.0, 1.0) | (1.0, 3.5) | (1.0, 3.5) | (0.46, 3.5) | (0.7, 1.45) |
| 8 | (0.0, 1.0) | (0.0, 1.0) | (1.0, 1.0) | (0.98, 1.0) | (0.78, 1.0) | (0.96, 1.0) |
| 9 | (0.0, 1.0) | (0.0, 1.0) | (1.0, 1.0) | (1.0, 1.0) | (0.45, 1.0) | (0.94, 1.0) |

Table 6.1: Approximations to the optimal minimum, maximum probabilities across algorithms and instances.

What we see is already encouraging: in 7 out of 9 instances, Goldilocks$_1$ simultaneously achieves within 10% of optimal maximum and minimum probabilities. This is striking, because it was not even clear a priori that this would be possible for *any* algorithm. In contrast, we see that Legacy and Minimax perform poorly on low probabilities, and Maximin/Leximin perform poorly on high probabilities, and Nash performs poorly on both.

However, these results do not paint a complete picture: in several instances (3, 4, 8, 9), *all* algorithms achieve the optimal maximum probability *simply because the quotas require an agent to receive probability 1.* Thus, to fully compare the performance of these algorithms' on controlling high probabilities, we must examine their performance on less constrained — but still realistic — instances. To do so, we study these algorithms' maximum and minimum probabilities as we successively drop features from each instance in decreasing order of their selection bias, as in [135], Figure 1c. We discuss the formal feature dropping procedure and provide results for omitted instances in Appendix E.5.5.

What we see in Figure 6.1 is striking: Goldilocks$_1$ hugs the gray region almost perfectly above and below, thus maintaining near optimality as features are dropped. This is in contrast to Maximin and Minimax, which continue to perform poorly on high and low probabilities respectively, even as the instance is loosened and better probabilities are possible. Together, these results show

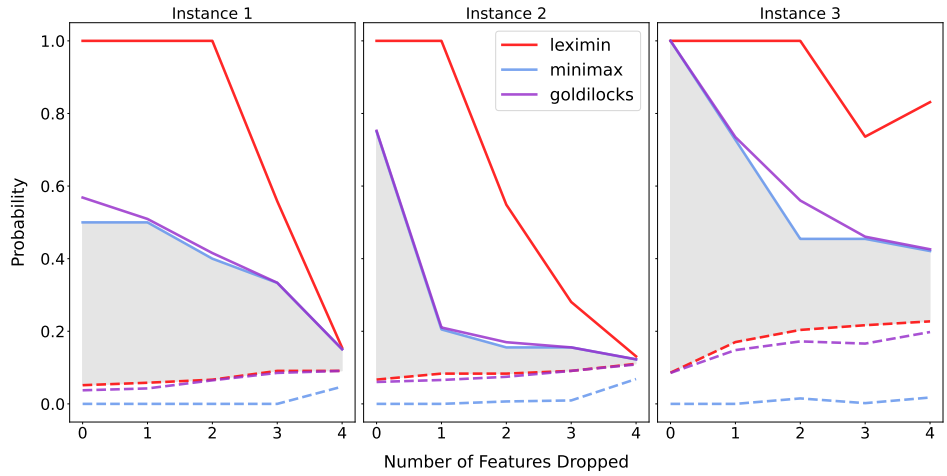



Figure 6.1: The solid, dashed lines represent maximum, minimum probabilities per algorithm, respectively. The shaded region lies between the optimal maximum probability and optimal minimum probability, establishing the region where no algorithm's extremal probabilities can exist.

that controlling high and low probabilities simultaneously is generally possible to a great extent, and all previously explored algorithms were leaving a lot on the table with respect to this goal.

### 6.6.2 Fairness, Manipulation Robustness, and Transparency

**Fairness.** While the above results already show the performance of all algorithms on *Maximin* fairness, there are other normatively justified notions of fairness, such as the *Gini Coefficient*, defined as $Gini(\boldsymbol{\pi}) := \frac{\sum_{i,j\in[n]} |\pi_i - \pi_j|}{2\sum_{i,j\in[n]} \pi_i \pi_j}$. Figure 6.2 shows the Gini Coefficient achieved across algorithms and instances. Note that a smaller Gini Coefficient reflects greater fairness, as *Gini* measures *inequality*.

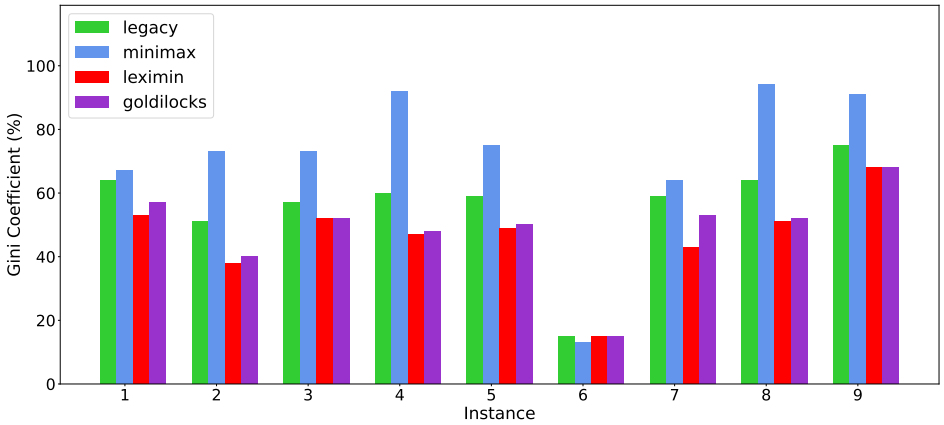



Figure 6.2: Gini coefficient across algorithms and instances. Lower Gini Coefficient means greater fairness.

Figure 6.2 shows similar algorithmic behavior across instances: Legacy and Minimax — which

we expect to be very unfair — tend to have high inequality per *Gini*. In contrast, GOLDILOCKS$_1$ and LEXIMIN perform far better (unsurprisingly, LEXIMIN is slightly better, as it prioritizes fairness alone).

**Manipulation Robustness.** In accordance with previous work, we evaluate manipulation robustness by measuring a weakened version manip$_{int}$ with a single manipulator. In particular, we measure the maximum probability gainable by any *Most Underrepresented (MU)* manipulator, reports the value of each feature that is most disproportionately underrepresented in the pool (as studied in Flanigan et al. [135]). We evaluate how this probability changes as $n$ grows, which we simulate by simply duplicating the pool. Details on these experiments are found in Appendix E.5.6.

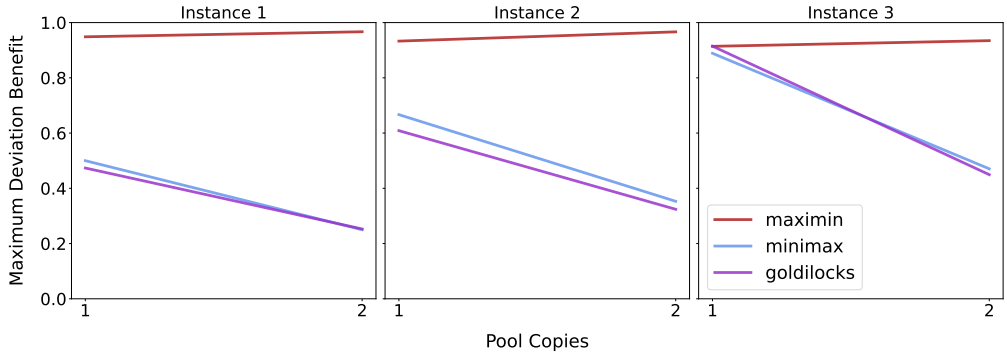



Figure 6.3: The maximum amount of probability any single *MU* manipulator can gain, for 1 and 2 pool copies.

In Figure 6.3, we see that in instances 1 and 2, GOLDILOCKS$_1$ is far less manipulable than MAXIMIN; in instance 3, we know the quotas require some agents to receive probability 1. However, we see that as the pool is duplicated, GOLDILOCKS$_1$ makes use of this and the manipulation drops; in contrast, across instances, MAXIMIN remains just as manipulable. From Figures 6.3 and 6.2, we conclude that GOLDILOCKS$_1$ achieves meaningful gains in Manipulation Robustness over LEXIMIN — the practical state-of-the-art — without any meaningful cost to fairness, as desired.

**Transparency.** Finally, we evaluate the extent to which we can round GOLDILOCKS$_1$-optimal panel distributions to $m$-uniform lotteries without losing too much on high or low probabilities. In this analysis, we use $m = 1000$. Although this is lower than $n\sqrt{n}$ as our theory dictates, we will find that this practicable number of panels is sufficient for good performance.

We consider two rounding algorithms. First, *ILP* is the integer program which finds the $\mathcal{E}$-optimal $m$-uniform lottery. Second *Pipage* rounding [141] is a simple randomized dependent rounding procedure. Although this algorithm does not come with any formal guarantees on how much its rounded distribution will change agents' selection probabilities, *Pipage* is fast; already implemented in practice for the purposes of transparent sortition [163]; and has the added advantage that, over the randomness of the rounding *and* sampling, it *perfectly* preserves the selection probabilities, thereby exactly maintaining our guarantees in Theorem 6.4.1 end-to-end. Details on our

rounding algorithms and experimental methods are in Appendix E.5.7

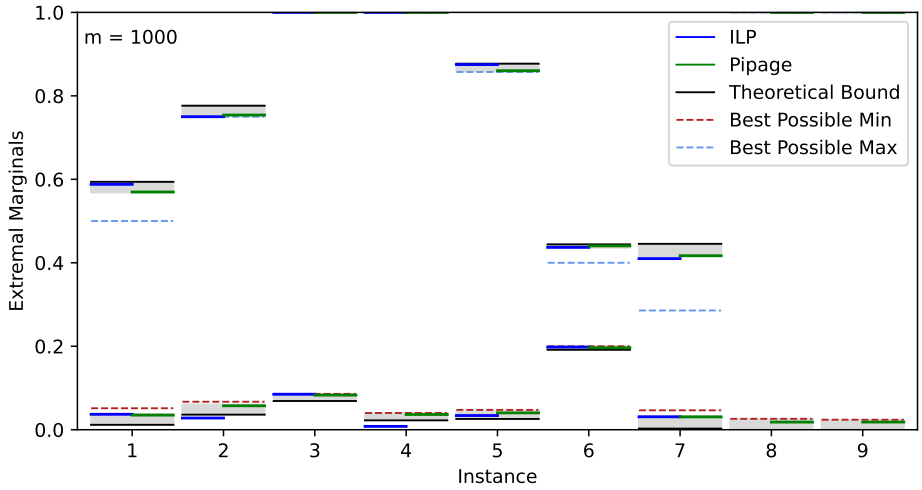



Figure 6.4: Deviations from GOLDILOCKS$_1$-optimal selection probability assignments by *Pipage* and *ILP*. The values for *Pipage* correspond to averages of minimum, maximum probability per run over 1000 runs. Error bars are plotted to indicate standard deviation, but they are so small that they are not visible. Gray boxes extend vertically from the minimum (resp. maximum) probability given by GOLDILOCKS$_1$ to the "theoretical bound", as given by Theorem 6.5.1. Optimal minimum, maximum probabilities per instance are shown for reference.

Figure 6.4 shows good news: while *ILP* tends to lose a lot on either the maximum or minimum probability in many instances, *Pipage* is reliably leaving GOLDILOCKS$_1$'s optimal selection probabilities essentially unchanged. This is great news, because it means that we can have the best of both words: we can achieve a high-quality uniform lottery *while* preserving our fairness and manipulation robustness guarantees from Theorem 6.4.1 exactly, end-to-end.

## 6.7 Discussion

From our empirical analysis, we see that the good behavior of GOLDILOCKS$_1$ suggested by our main results (Theorem 6.4.1) is borne out in real data: GOLDILOCKS$_1$ indeed performs well on *fairness*, *manipulation robustness*, and *transparency* simultaneously. However, this was not actually guaranteed; while apparently reflective of reality, our bounds in Theorem 6.4.1 do not directly apply to most real-world instances because real pools are too small to satisfy Assumption 6.2.4. To prove bounds that truly give guarantees in practice, one would need to prove results like in Theorem 6.4.1 without Assumption 6.2.4. The key technical challenge here is that manipulating coalitions can not only *expand* the set of feasible panel types; they can also diminish it. This may enable qualitatively different kinds of strategies, which may reveal new potential vulnerabilities of selection algorithms — and algorithmic solutions that can resolve them.

# 7

# Ongoing & Future Work

## 7.1 A Second Look At Representation/Randomness Trade-offs

The prevailing narrative in applied sortition is that *tighter quotas* equals *better descriptive representation*. Under this interpretation, descriptive representation must fundamentally trade off with the *equality* of the lottery, due to the simple mathematical fact that tighter quotas more strongly constrain the randomness available.

However, we now call this narrative into question with the following simple observation: *there are many groups whose representation might matter that cannot be protected with quotas.* Such groups may exist for various reasons: organizers may not *know* these groups are important to the conversation *a priori*; it may be legally or publicly controversial to impose quotas on a certain groups; one may not want to impose quotas on groups whose defining feature(s) cannot be easily confirmed, raising risks of manipulation; or, it may simply be that the number of groups we care about are too numerous, and imposing quotas on all of them would cause infeasibility or too severely limit the randomness.

Clearly, tightening the quotas will improve the representation of *groups protected by quotas*. However, assuming there are some subgroups that may matter but who cannot be protected by quotas, it becomes far less clear that tightening the quotas will definitely improve representation overall. We illustrate how tightening the quotas can harm representation (and diversity) via the following concrete example.

### 7.1.1 Motivating Example

Let $k = 8$. Let $n = 64$. Suppose we have three binary features, each which can take on values 0 or 1. Suppose on every feature, 3/4 of the population has value 0 and 1/4 has value 1. Suppose the pool contains the following unique vectors: 000, 100, 010, 111. The numbers of each type don't matter much for most of the example; for now, let's just assume that there are at least $k = 8$ people in each group so that our counterexamples won't be due to simply running out of people.

*Perfectly tight quotas.* Suppose first that we set exactly tight quotas, so for each feature, we need exactly 6 people with value 0 and 2 people with value 1. Then, the only feasible panel is

$$\text{Panel Type 1: } 6 \times (000), \quad 2 \times (111).$$

To see this, first observe that *any valid panel must contain two people with vector 111* in order to get two people with 1 values for the third feature; then the rest of the panel must be 6 people with vector 000. Thus, people with vectors 100 and 010 cannot be included on any feasible panel, and must receive 0 probability.

*Relaxation of 1.* Now, suppose we permit a tolerance of 1 on all quotas: we need 5-7 people with value 0 and 1-3 people with value 1 for each feature. Now, the set of valid panels includes the following three panels:

- Panel Type 2: $5 \times (000), \quad 1 \times (111), \quad 1 \times (100), \quad 1 \times (010)$
- Panel Type 3: $5 \times (000), \quad 2 \times (111), \quad 1 \times (100)$
- Panel Type 4: $5 \times (000), \quad 2 \times (111), \quad 1 \times (010)$

**Observation 7.1.1.** *Tighter quotas can harm **intersectional representation**.*

*A strong case can be made that Panel 2 is more representative at the intersectional level than panels 1, 3, or 4. This panel is only available to us if we loosen the quotas.*

**Observation 7.1.2.** *Tighter quotas can harm **hidden feature representation**.*

*Suppose that there is another hidden feature, which we cannot even observe in the selection process. Suppose that that value 0 for this feature perfectly correlates with vectors 100 and 010. Then, tightening quotas excludes those with value 0 for the hidden feature, thereby also excluding this group.*

**Observation 7.1.3.** *Tighter quotas can harm representation on groups **protected by quotas**.*

*Suppose the pool contains only 2 people with vector 111. Then, both of them will have to be chosen with probability 1 in the original example, and a manipulator from group 000 can misreport vector 111 and be selected with probability 2/3. Then, in expectation, people with 0-values for any of the three features will receive (6+2/3)/8 = 83% of the seats instead of their allotted 75%.*

This example illustrates how tightening the quotas can actually *harm* representation of **all three kinds of groups**: those protected by features, those defined by intersections of protected features, and those defined by features we cannot observe.

Once the quotas are loosened, this problem with representation being fixed is not guaranteed: Panel Type 1 is still valid, and in principle, a selection algorithm could randomize over only panels of that type, causing exactly the same problems as above. However, we can ensure some measure of representation by carefully designing the lottery. To see this, suppose our lottery uniformly randomizes over Type 2 panels. This lottery is like a (possibly suboptimal) version of the *Goldilocks* objective from Chapter 6, in that it ensures no one gets too little or too much selection probability. Here is how our three kinds of groups now fare with regards to representation:

- ***Intersectional groups:*** Vectors 100 and 010 are guaranteed 1 panel seat each, whereas before they were receiving 0.
- ***Hidden groups*:** People with value 0 for our hidden feature are now guaranteed two panel seats, whereas before they were receiving 0.
- ***Groups protected by quotas:*** There is now 1 seat reserved for people of vector 111, so misreporting vector 111 as a 000 person will mean you are selected with probability 1/3. Now, 0-value groups for the first two features receive in expectation (5+1/3)/8 = 67% of their due 75% of seats (partly due to the relaxed quotas), and those with 0 for the third feature receives 79% of panel seats (which is still over their allotted 75%, but less so).

### 7.1.2 Research Questions

Based on the example above, we define *rich representation* as proportional representation of *all population subgroups*, including those defined by combinations of observed features and those defined by features we cannot observe. The above example illustrates why sacrificing the equality of the lottery in favor of tighter quotas will not necessarily improve representation when its conception goes beyond the narrow version defined by quotas. On the contrary, a more equal lottery can provide guarantees on representation of unprotected groups, which can be *harmed* by tightening the quotas.

In this project, we are examining—both in theory and in practice—how both the uniformity of the lottery *and* the quotas can together help support rich representation—and which balance of lottery uniformity and tightness of quotas (a true mathematical trade-off) to achieve the best possible rich representation. The relationships between these quantities must fundamentally depend on how features correlate with one another in the population versus the pool. Our goal is to understand (1) what representation guarantees are possible regardless of these correlations, which often cannot be observed, and (2) whether the quotas currently being used in practice are tight enough to be significantly *harming* rich representation.

### 7.1.3 Practical Impact: A Tool for Quota-Tuning.

We present a tool to help practitioners to fine-tune their quotas in a way that accounts for the equality of the lottery, as measured by any equality objective $\mathcal{E}$ that can be used in the framework discussed in Chapter 3. This tool is motivated by the research above, which illustrates why it can

be advantageous for representation to slightly *loosen* the quotas in favor of a far more uniform lottery.

**Inputs.** When practitioners use this tool, they must supply the *equality objective* $\mathcal{E}$ they want the lottery to optimize, the pool, and their desired panel size $k$, as usual. Then, instead of exact quotas, they give the following inputs for every feature-value $f, v$ on which they want to ultimately want quotas to be imposed:

- $p_{f,v}$, the *ideal* number of panel seats given to people with value $v$ for feature $f$.
- $L_{f,v}$ and $U_{f,v}$, integer-valued *hard limits* on the number of seats given to group $f, v$. Note that these hard limits are distinct from quotas imposed at the panel selection stage: they are considerably looser, as they are just the non-negotiable bounds.
- $\lambda_{f,v}^{\text{upper}}$ and $\lambda_{f,v}^{\text{lower}}$ (optional): describes the *priority* on the upper (resp. lower) quota on $f, v$. This dictates how important it is for that quota to be very close to the ideal proportion $p_{f,v}$, relative to other quotas. If not set, these values default to 1.

**A Human-In-The-Loop Algorithm.**

Step 1: Generate candidate quota settings. First, three possible quota settings are determined by optimizing the following program. We initially run it for three different $Q$ values (e.g., $Q \in \{1, 10, 20\}$), where $Q$ controls the extent to which tight quotas are prioritized over the quality of the lottery. We are experimenting with $z \in \{1, 2\}$.

$$\min_{\pi,\, \ell_{f,v},\, u_{f,v}} \quad \mathcal{E}(\pi) \; + \frac{Q}{|FV|} \sum_{f,v} \left( \lambda_{f,v}^{\text{upper}} \left( \frac{u_{f,v} - p_{f,v}}{p_{f,v}} \right)^z + \lambda_{f,v}^{\text{lower}} \left( \frac{p_{f,v} - \ell_{f,v}}{p_{f,v}} \right)^z \right)$$

$$\begin{aligned} \text{Subject to:} \quad & \ell_{f,v} \in [L_{f,v}, p_{f,v}] \\ & u_{f,v} \in [p_{f,v}, U_{f,v}] \\ & \sum_{i:\, f(i)=v} \pi_i \in [\ell_{f,v}, u_{f,v}] \qquad \forall f, v \\ & \sum_{i \in [n]} \pi_i = k \end{aligned}$$

For any given $Q$, $z$, the quotas proposed by this program are the values of variables $\ell_{f,v}, u_{f,v}$. The first two constraints require that these proposed quotas fall within the hard limits set by practitioners. The last two constraints are those from the *continuous relaxation* of the quotas and panel size restrictions, in which people are treated as divisible (as studied in Chapter 5). These constraints require that the lottery, specified by the variable $\pi$, satisfies the quotas in expectation. Subject to these restrictions, the objective function balances two goals according to $Q$: optimizing the equality of the lottery, and tightening the quotas.

Step 2: Generate lotteries based on candidate quota settings. From running the above program for $Q \in \{1, 5, 20\}$, we have three proposed ways to set the quotas. Because these quotas may be fractional, we round them *outwards* to the nearest integer; that is, $\ell_{f,v}$ values get rounded

down, $u_{f,v}$ values get rounded up.[1] Then, for each $Q \in \{1, 10, 20\}$ we run the algorithmic framework from Chapter 3 with the corresponding rounded quotas and the chosen equality objective $\mathcal{E}$. This produces lotteries corresponding to each of these quota settings, which we display on the following interface. Here, quota settings and corresponding $\mathcal{E}$-optimal lotteries are ordered on a line in increasing order of $Q$:

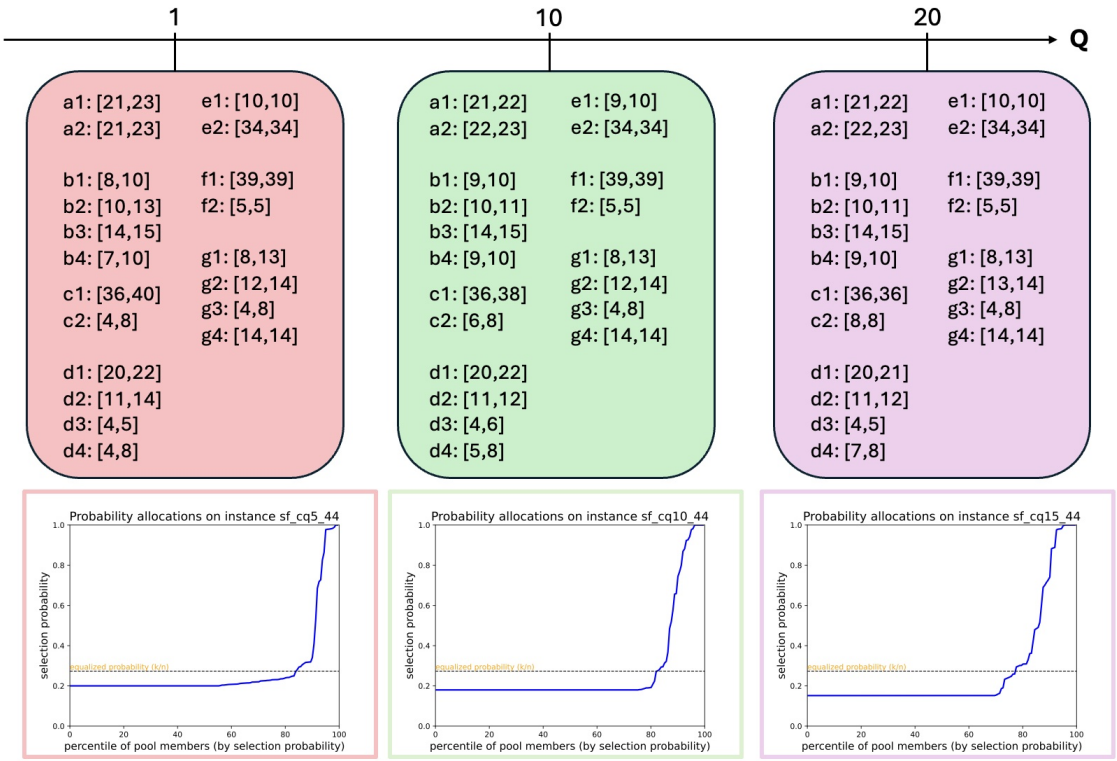



Figure 7.1: Caption

Step 3: Human Selection of $Q$. Practitioners see the interface above, and they can review the proposed quota settings and the corresponding lotteries, each which represents a point near[2] the Pareto frontier between quota tightness and lottery equality. They then have 3 options: (a) If they are happy with one of the proposed options, they can directly select it. (b) If they are *mostly* happy with one of the proposed options but want to tweak it, they can modify a quota setting, and then immediately generate the associated lottery to make sure they're ok with it. (c) If they want to explore a $Q$ value that isn't shown, they can click on the $Q$ line to select a new value of $Q$

[1]We round outwards because these quotas were optimized in the *continuous relaxation* of the panel selection problem, whereas they will be used the more constrained integer version. Thus, for there to be hope of retaining the good properties of the lottery determined in Step 1, we need to round outward to loosen the problem. Note that this rounding will produce quotas that still fall within $[L_{f,v}, U_{f,v}]$ because these bounds are integers.

[2]The trade-off is not optimal, because the quotas are optimized in the continuous relaxation of the problem. Part of this project will be investigating, both in theory and in practice, how much the lottery can change from the continuous to integer version. Empirically, it looks to change very little.

and the corresponding lottery will be generated and placed on the $Q$ line. They iterate through this process until they find a quota setting they are happy with.

**Preliminary Results.** Here, we show two different instances where we compare the original quotas hand-tuned practitioners to quotas proposed by our method for $q \in \{1, 10, 20\}$, where $\mathcal{E}$ is GOLDILOCKS. By inspection, the quotas chosen by our algorithm are by and large qualitatively very similar to the original quotas, but we can permit significantly more uniform lotteries.

| original | q=1.0 | q=10.0 | q=20.0 |
|---|---|---|---|
| (2, 4) | (2, 5) | (2, 5) | (2, 4) |
| (2, 4) | (2, 4) | (2, 4) | (2, 5) |
| (2, 4) | (2, 4) | (2, 4) | (2, 4) |
| (2, 4) | (2, 4) | (2, 4) | (2, 4) |
| (2, 4) | (2, 5) | (2, 5) | (2, 5) |
| (2, 4) | (2, 4) | (2, 4) | (2, 4) |
| (2, 4) | (2, 4) | (2, 4) | (2, 4) |
| (2, 4) | (1, 4) | (2, 4) | (2, 4) |
| (2, 4) | (2, 4) | (2, 4) | (2, 4) |
| (13, 15) | (13, 15) | (13, 15) | (13, 15) |
| (15, 17) | (15, 17) | (15, 17) | (15, 17) |
| (8, 10) | (5, 10) | (5, 10) | (7, 10) |
| (4, 6) | (3, 6) | (4, 6) | (4, 6) |
| (4, 6) | (4, 6) | (4, 6) | (4, 6) |
| (2, 4) | (2, 5) | (2, 5) | (2, 4) |
| (2, 3) | (2, 4) | (2, 3) | (2, 3) |
| (3, 5) | (3, 5) | (3, 5) | (3, 5) |
| (2, 4) | (2, 4) | (2, 4) | (2, 4) |
| (13, 15) | (13, 15) | (13, 15) | (13, 15) |
| (13, 15) | (13, 15) | (13, 15) | (13, 15) |
| (2, 2) | (1, 3) | (1, 3) | (1, 3) |
| (2, 2) | (1, 3) | (1, 3) | (1, 3) |
| (2, 2) | (1, 3) | (1, 3) | (1, 3) |
| (4, 6) | (2, 6) | (2, 6) | (4, 6) |
| (2, 4) | (1, 4) | (2, 4) | (2, 4) |
| (2, 2) | (1, 3) | (1, 3) | (1, 3) |
| (16, 18) | (16, 22) | (16, 20) | (16, 18) |
| (4, 6) | (4, 6) | (4, 6) | (4, 6) |
| (24, 26) | (24, 26) | (24, 26) | (24, 26) |
| (2, 3) | (1, 3) | (2, 3) | (2, 3) |
| (4, 6) | (3, 6) | (4, 6) | (4, 6) |
| (11, 13) | (10, 13) | (11, 13) | (11, 13) |
| (10, 12) | (10, 15) | (10, 12) | (10, 12) |

| original | q=1.0 | q=10.0 | q=20.0 |
|---|---|---|---|
| (19, 20) | (19, 20) | (19, 20) | (19, 20) |
| (20, 21) | (20, 21) | (20, 21) | (20, 21) |
| (7, 7) | (6, 8) | (6, 8) | (6, 7) |
| (8, 9) | (8, 9) | (8, 9) | (8, 9) |
| (13, 14) | (13, 14) | (13, 14) | (13, 14) |
| (11, 12) | (11, 12) | (11, 12) | (11, 12) |
| (11, 12) | (11, 12) | (11, 12) | (11, 12) |
| (7, 8) | (7, 8) | (7, 8) | (7, 8) |
| (3, 4) | (2, 4) | (2, 4) | (2, 4) |
| (4, 4) | (2, 5) | (2, 5) | (2, 5) |
| (4, 4) | (3, 5) | (3, 5) | (3, 5) |
| (11, 12) | (11, 12) | (11, 12) | (11, 12) |
| (20, 20) | (19, 21) | (19, 21) | (19, 20) |
| (20, 20) | (19, 21) | (19, 21) | (19, 20) |
| (20, 21) | (20, 23) | (20, 23) | (20, 21) |
| (13, 14) | (13, 14) | (13, 14) | (13, 14) |
| (6, 6) | (4, 7) | (4, 7) | (5, 7) |
| (37, 38) | (37, 38) | (37, 38) | (37, 38) |
| (3, 3) | (2, 4) | (2, 4) | (2, 4) |

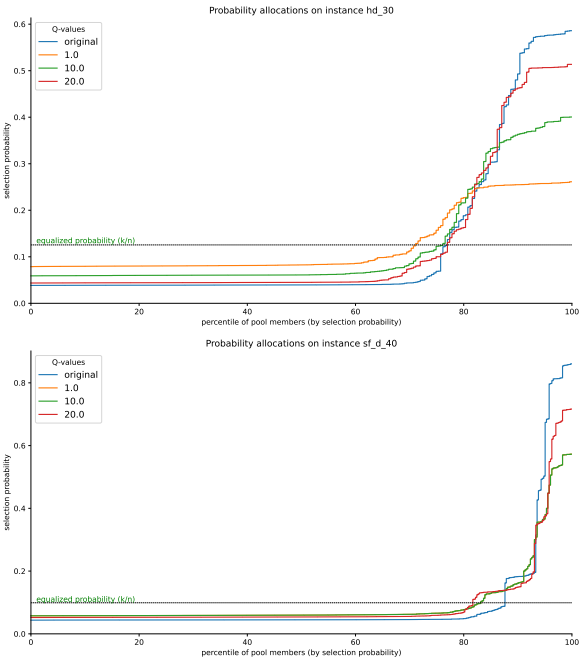



Figure 7.2: Caption

### 7.1.4 EXTENSION TO DIVERSITY

Another goal one might have for the panel, beyond rich representation, is *diversity*—roughly, how many *different kinds of people* are on the panel. In fact, the intuition from our example in SECTION also applies to this goal:

**Observation 7.1.4.** *Tighter quotas can harm **diversity**.*

*By any reasonable metric, Panel 2 is more diverse than Panels 1, 3, or 4, whether or not we consider our hidden feature. This panel is only available to us if we loosen the quotas.*

We plan to repeat our analyses proposed above to study the impact of quotas and the equality of the lottery on *diversity*, in addition to rich representation.

## 7.2 Holistically Designing the Participant Recruitment Process

So far, the research discussed only addresses Stage 2 of the two-stage participant recruitment process, as defined in Chapter 1. Now, we zoom out to consider the entire recruitment process end-to-end, which involves the two stages defined before, plus a third stage we omitted above for simplicity. The unifying aim of the open directions in this section is to design these three stages *together*, so that each stage works in concert with the other stages. When possible, we propose new ways to utilize data from past selection processes to make predictions, which is available in practice but is currently unused.

*Stage (1): Pool recruitment.* People from the population are recruited into a group that has said they will participate if chosen. In reality, while these people have a far higher probability of participating if chosen than the general population, they are *not guaranteed* to ultimately say yes. We can assume that, unlike people in the population, we know the relevant features of the people in the pool, which they are asked to report.

*Stage (2): Panel selection.* The panel is typically selected from the pool by lottery, and must satisfy demographic quotas.

*Stage (3): Panel re-selection post-dropout.* After the panel is selected, some subset of people from the panel will drop out. They may drop out as early as the moment they are notified of their selection, or as late as the first day of deliberation (in which case they just do not show up to the process). Their seats need to be filled with people who are as close as possible to their feature-values, while also trying to preserve properties of the lottery.

We present open research directions roughly in the order we think they should be completed: working backwards from stage (3) back to stage (1), so that our approaches in earlier stages can account for the needs of later stages. The first direction, in Section 7.2.1, is already ongoing.

### 7.2.1 Replacing panelists who drop out in stage (3)

**Remark:** Many questions posed in this section can be studied with or without access to predictions about who will drop out, which can be made based on existing data from past panels. Using access to predictions would prompt questions beyond those listed below regarding ensuring robustness to bad predictions, as is the common goal in the algorithms with predictions literature (e.g., [193]).

In practice, replacement panelists can be sourced in one of two ways. Sometimes, practitioners sample them directly from the remaining members of the pool. When this approach is taken, practitioners typically run the algorithmic framework from Chapter 3. This approach poses several questions, including: *Are there ways to design the selected panel to make it more likely that we will be able to retain representation after dropout?* and *If we run a lottery, lose some people to*

*dropout, and then run it again, how closely do the selection probabilities approximate those that would have been achieved in a one-shot lottery where the dropouts had not signed up in the first place?*

In higher-stakes settings, it is especially important to the integrity of the deliberative process for all panelists to attend the first day of the event, when crucial on-boarding information is conveyed. To prepare for the inevitable scenario that some panelists do not show up for the first day, sometimes practitioners pre-select a set of *alternate* panelists and pay them to attend the first day. The constraint, then, is that selecting more alternates is more costly. The task of selecting these panelists raises several additional questions, the most obvious being *how should these alternates be selected, potentially based on predictions*, *how many alternates do we need to select to achieve guarantees on the availability of replacements*, and *should these alternates be selected in conjunction with the panel itself, to maximize the chances of suitable replacements being available?*

Finally, when analyzing any algorithm(s) for handling dropout, we need to consider the potential for a new kind of manipulation: *opting in with the intent of dropping out.* This type of manipulation is possible, because all else held constant, it is possible for someone *change the lottery* by, instead of just declining to participate, opting in and then dropping out. Our task is to characterize the extent to which such manipulations can be successful, and under what conditions.

### 7.2.2 Characterizing what pools are "good" for Stages (2) and (3).

Chapters 3 to 6 gave us several deployable algorithmic solutions for stage (2). The ongoing work described in Section 7.2.1, once completed, will also provide algorithmic solutions for stage (3). However these algorithms, no matter how sophisticated, are fundamentally limited by the quality of the pool from which they can choose participants. When we talk about impossibilities, the theme that arises again and again is *selection bias*—how the pool is imbalanced compared to the population, and resultingly, compared to the quotas we ultimately want to satisfy.

The algorithms designed in the work so far are built to take the pool as given, assuming no control over its composition. Ultimately, we want to take a more active role in improving Stage (1), pool recruitment. Before we do this, though, we must understand: *what kind of pool are we trying to create?* In other words, *what kinds of selection bias are the biggest problem in practice, for stages (2) and (3)?* This is difficult to answer, because there is no clear metric of selection bias: one can measure it at the *feature vector* level, or the *quota level*, or in any other number of ways; in any case, it is hard to reduce such a complex and combinatorial property into a single-dimensional measurement capturing the extent the skew of the pool will prevent us from achieving our original ideals throughout Stages (2) and (3).

Some work in this thesis has examined one approach in a limited capacity: in Chapter 5 Figure 2(a), our empirical analysis revealed a very simple measure of selection bias was highly predictive of (asymptotically) optimal manipulability across datasets :

$$\max_{f,v \in FV} \frac{p_{f,v}}{\eta_{f,v}} - \min_{f,v \in FV} \frac{p_{f,v}}{\eta_{f,v}},$$

where $p_{f,v}$ and $\eta_{f,v}$ are the fraction of the population and pool, respectively, with value $v$ for feature $f$. Then, the first term is a proxy for the maximum selection probability due to selection bias (proportional to the seats per person for the most underrepresented feature-value group), and the second is a proxy for the minimum. As such, this difference intuitively *roughly* captures the gap in selection probabilities, though not precisely. While this is a natural metric, its usefulness has so far been established only in a very limited capacity, and it is just one of many reasonable measurements.

In this project, we would consider many such metrics, studying—both theoretically and empirically—which are most robustly likely to compromise our ideals. This line of questioning would also involve interrogating the stability of the relationship between our metric and our sortition ideals to small random perturbations in the quotas or pool composition, to understand how robustly it is possible to meaningfully measure the selection bias.

#### 7.2.3 Improving recruitment in Stage (1)

With the understanding of which metrics best capture the problematic types of selection bias, we can then turn to the task of understanding how to better-design Stage (1). Our goals in doing so should really be two-fold: first, we want to decrease selection bias along our metrics, which will measure the representativeness of the pool only based on the attributes protected by quotas. While this is important for the properties of the lottery, from a normative perspective we also want to try to reach hard-to-reach groups—even those defined by attributes that will not be protected in the lottery—because they may have important perspectives to add *even by virtue* of being hard to reach (e.g., because they might be distrustful of government). We propose a range of ideas here, which achieve these two goals to varying degrees.

**Door-knocking plus two-stage selection.** Some groups (e.g., [10]) advocate and deploy panel selection via *door-knocking*, which departs significantly from the two-stage process we have studied. In the door-knocking approach, a random lottery is done, and then for every person selected, the organizers go to their door and engage in a conversation with them about participating. Although this process is significantly can be resource-intensive—especially at scale—it can reach hard-to-reach groups who would not typically participate, resulting in richer diversity.

One challenge with door-knocking alone is that, even though it can significantly bring up participation rates, they can remain far from 100% and still result in a panel that is far from proportionally representative on salient demographic and ideological dimensions. For contexts where this is a concern, one could consider doing a two-wave process to try to get the best of both worlds: do door-knocking first at a small scale, from which you would get a group that is not proportionally representative, but would include populations who would otherwise not opt in. Then, run the two-stage process above to select the remainder of the panel, using the quotas to correct deviations from proportional representation.

**Directed sampling with imperfect data.** Suppose for a moment that you had perfect demographic data about every person in the underlying population. Then, it would be possible to simply do adaptive sampling of the population until a representative pool (at least on the ob-

served dimensions) was achieved. Even doing this is not perfectly trivial: without knowing a priori who will accept, one has to be thoughtful about the order in which people are invited to avoid risking having to recruit an enormous number of people before the pool is well-balanced. How to do this while retaining good probabilistic properties and a reasonably-sized pool (and number of phone calls) is already an interesting question, with or without predictions of who will accept. Regardless of how it's implemented, however, any such sampling process that produces a balanced pool must be effectively shifting the unequal probability of selection from the second stage to the first.

The more interesting question is how to do this adaptive sampling when data about the population is *limited*. For example, practitioners may have aggregate level data about the residents of neighborhoods or zip codes, perhaps about a limited set of features (e.g., race, income). Based on this data, plus an understanding of how these known features *correlate with unknown features* based on census data, one could try to design a probabilistic process for adaptively sampling. The questions are: *can we build a process that is flexible enough to use diverse kinds of data?* and *how closely can such a process—depending on the data quality—approximate the optimal sampling procedure in the case where perfect population data is available?*

Furthermore, being able to make more intentional sampling decisions based on the limited population data available in practice opens up the door for other kinds of more targeted approachs. One example that might be fruitful would be to extend the hybrid door-knocking approach above to more continuously interpolate between the opting-in process and the door-knocking process. For example, suppose you could run the two-stage process we study, but based on predictions (using limited data) you could decide to more intensively recruit certain invitees – the ones you believe for whom it will make a difference – by door-knocking.

**Fine-tuning the invitation format.** Across organizations and time, the invitations sent out in Stage (1) have varied in format and medium. Using these invitations — and data on the pools that were recruited with them — one could ask: *What formats of invitations lead to more balanced pools?* This question would require feature extraction, possibly using LLMs, of the text of invitations. It is reminiscent of work done in marketing on how to format solicitations of donations, and tools from that regime might be useful; the qualitative difference here is that our goal is not maximizing the number of the responses, but the *diversity*, which might lead to qualitatively different prediction methods and conclusions.

# Part II

# Beyond Descriptive Representation

# 8

# Background

In part I, we discussed how one of the key arguments of sortition, in its "ideal" form is a uniform lottery, is that it yields proportional descriptive representation. Underlying the commitment to this ideal is a philosophical judgement that is very popular in computational social choice: *that regardless of the decision at hand, everyone should be entitled to the same influence.* This principle founds not just sortition, but most popular decision methods in computational social choice, in which all voters are given a single vote to cast (or in the case of liquid democracy [159], delegate). In this section, we question this philosophical principle, and explore an intuitive alternative notion of representation that accounts for the extent to which people are affected by the decision at hand.

Consider the following scenario. Suppose we have a city that is making decisions about major changes to the public transit system. 90% of people in the city can afford their own private transit, and rarely or never use public transit. In contrast, the remaining 10% cannot afford private transit and depend on public transit for all travel. Now, suppose we are choosing a panel of $k = 50$ people to deliberate on some decision about their local transit system. Let us be in the "ideal" case where we can do a uniform lottery and all chosen will partake. Then, the resulting panel will contain roughly 5 people who use public transit regularly, and 45 people who essentially never use it. If we imagine this deliberation, it seems somewhat strange: the large majority of participants will not be impacted at all by changes to the public transit system. Even worse, it could be the case that those dependent on public transit tend to have a fundamentally different opinion about what should be done than those who do not; with such a small fraction of seats on the panel, they may be unsuccessful in advocating for their own interests, despite being the ones who will have to live with the ultimate decision that is made on a day-to-day basis.

At face value, this is an example of a classical phenomenon, the *tyranny of the majority* (e.g., see [57]). However, we pose that this example actually illustrates a more insidious *sub-case* of this general phenomenon. In one version of the tyranny of the majority, the majority group *and* the minority group can both benefit significantly, but from opposing policy options. In this case, someone has to lose—this is just a fundamental impossibility of politics, which often requires us to choose one policy that will apply to all. However, the example above presents a much more problematic version: the minority can gain significantly by getting their way, *while they majority cannot.* This is even worse than the previous case: now, when the majority gets their way, they barely stand to benefit, when a different outcome could have substantially helped the population much more overall.

We will refer to this phenomenon here as the *tyranny of the* ***less affected*** *majority.* It is not hard to think about salient examples of issues in which this problem could occur: Consider masking requirements and immunocompromised populations, or accessibility features (e.g., wheelchair ramps) and those who rely on them. This idea can also apply to issues where "minority" status is not statistical, but rather political: consider the global discussions around climate change, in which the global south has been underrepresented despite projections that they will be disproportionately affected [260].

The tyranny of the less affected majority scenario —and why it can lead to democratic outcomes with suboptimal social benefit — can be understood without committing to any formal model or democratic mechanism. It does, however, connect to an important folklore impossibility in the computational social choice literature on *distortion* (Theorem 8.0.1). It is therefore from within this model that we will begin our study of this problem.

**The distortion model.** The distortion literature traditionally studies a model in which voters' preferences are defined by latent *utilities*: that is, each voter $i \in [n]$ has a nonnegative utility $u_i(a)$ for each alternative (policy or candidate) $a \in [m]$, which we will think of as the extent to which $i$ benefits if $a$ is implemented. Voters "vote" by providing a complete ranking in order of their utilities: that is, $i$ prefers $a$ to $b$ if and only if $u_i(a) > u_i(b)$. A *voting rule* $f$ is any (possibly randomized) mapping from a set of $n$ such rankings to a single winning alternative, typically denoted in this thesis as $a'$. We evaluate the "societal benefit" of any given alternative via its *utilitarian social welfare*, $\mathrm{sw}(a) := \sum_{i\in[n]} u_i(a)$, the sum of voters' utilities for it. In a given election, let $a^* := \arg\max_{a\in[m]} \mathrm{sw}(a)$ be the highest-welfare alternative (i.e., the most socially beneficial outcome available). The *distortion* is the competitive ratio between $\mathrm{sw}(a^*)$ and $\mathrm{sw}(a')$, quantifying the suboptimality of the winner relative to the best possible outcome. Note that because it is defined as $\mathrm{sw}(a^*)/\mathrm{sw}(a')$, larger distortion reflects a more sub-optimal outcome.

**Theorem 8.0.1** (folklore)**.** *The distortion of any deterministic voting rule is unbounded.*

*Proof.* This can be proven via our transit decision example above, once embedded into the distortion model above. Suppose there are two possible transit system policies, $a$ and $b$. The 10% who depend on public transit are significantly benefited by $a$ but not at all by $b$; let their vector of utilities for $a$ and $b$ be $(10, 0)$. Let the remaining 90% benefit slightly from $b$ and not at all from $a$; accordingly, their utilities are $(0, \epsilon)$. The former 10% of voters will vote for $a$, all the rest will

vote for $b$, making $b$ the winner, with welfare proportional to $0.9 \cdot \epsilon < \epsilon$. In contrast, $a$ is the highest-welfare alternative, with welfare proportional to $0.1 \cdot 10 = 1$. The distortion is therefore at least $1/\epsilon$. □

This resounding impossibility has spawned substantial work trying to get around it. Usually this literature does so via one of two kinds of assumptions, both which will connect to our work in Part II. First, there is the *unit-sum utilities* assumption [231], where each voters' utilities are assumed to add to 1—that is, for all $i \in [n]$, $\sum_{a \in [m]} u_i(a) = 1$. Another assumes the ability to query utilities. We will discuss these assumptions and other related assumptions in more detail in Chapter 9, where we present a different, intuitive solution: suppose we could represent people in proportion to how often they use public transit. This solution — while very intuitive — pursues a kind of representation that is *fundamentally different* from descriptive representation; we call it *stakes-based representation*. This representation concept connects to multiple political scientific and philosophical theories; for example, Harry Brighouse and Marc Fleurbaey define the principle of democratic *proportionality* as the idea that those who are more affected by the decision are entitled to greater representation [57]. A more detailed discussion of these connections is provided in Chapter 9.

## 8.1 Overview of Chapters

**Chapter 9: A Voting Framework for *Stakes-Based* Representation [132].** In this chapter, we investigate whether stakes-based representation can address the impossibility of unbounded distortion, and in the process circumvent the tyranny of the less affected majority. We study this question in perhaps the most canonical method of democratic decision-making: ranking based voting. In this setting, we design a general social choice framework that captures the intuition of accounting for stakes through the notion of *stakes-based representation*, whereby voters are reweighted according to their stakes. In this framework, voters' stakes are measured via *stakes functions*—any function mapping each voter's vector of utilities to a real number measuring the extent to which they stand to lose or gain from the decision at hand.

Within this framework, we then derive tight bounds on achievable distortion under stakes-based representation when perfect information about stakes is available. When stakes can only be estimated, we show that our distortion bounds hold approximately, implying that even a coarse estimate of stakes can bring about a large reduction in distortion. Finally, for the setting where stakes are completely unknown, we develop a proof-of-concept mechanism that ties together multiple elections so that the behavior of voters reveals their stakes, leading to bounded distortion in each election.

Our exploration of future work embarks on the significant task of exploring what it would take to bring this theory of stakes into the practice of democracy. When we think about bringing stakes-based representation in practice, we must let go of the goal of precise stakes-proportional representation, as it is likely impossible to assign any single "correct" number to someone's util-

ities or stakes. We are therefore just aiming to account for stakes within orders of magnitude, a goal toward which our results already go a long way, especially for issues where voters' stakes are extremely disparate. In Chapter 10, we discuss two distinct extensions of Chapter 9, both which pertain to the goal of bringing the theory of stakes into practice.

**Chapter 10: Ongoing and Future Work.** The first approach we consider endeavors to analyze *Quadratic Voting* (QV), an increasingly deployed voting mechanism, within our stakes framework from Chapter 9. Although the proposed work in this part is theoretical, it takes a pragmatic perspective, considering whether a core QV assumption holds in practice; whether an existing practical fix has the desired effect; and how to apply multi-issue mechanisms from Chapter 9 with QV ballots.

Our second course of ongoing/future study is less economic and more political. In the context of deliberative town halls, we aim to (a) devise a publicly acceptable approach to identifying key high-stakes groups on the issue, and then (b) designing the town hall to give these groups "substantively sufficient" influence in the democratic process—even if it means deviating from descriptive representation. We hypothesize there is hope for achieving goal (a), which comes from the (so far anecdotally-supported) idea that when asked who should be in the room when deciding a given issue, people seem to gravitate toward listing high-stakes groups (For example, in response to the question *"whose perspectives absolutely need to be considered in the conversation about COVID-19 masking requirements?"*, anecdotally people commonly identify frontline workers, medical staff, and those who are immunocompromised). Goal (b) is much trickier to achieve, because it relies on understanding what constitutes "substantively sufficient" influence—and measuring it. Conceptually, we use "substantive influence" to mean that a group's interests are accurately accounted for in the final decision. Whether this is achieved is a function of who is in the room *and* what happens in the room. The latter will be considered in Part III.

# 9

# A Voting Framework for Stakes-Based Representation

*Voters with Stakes can Ward Off Bad Outcomes* [32]
Bailey Flanigan, Ariel D. Procaccia, and Sven Wang
*submitted* 2024 (and *FORC* 2023, non-archival)

## 9.1 Introduction

In the standard model of voting, voters express their *ordinal* preferences by ranking a set of alternatives. It is reasonable to assume, however, that there *exist* cardinal utilities that voters associate with alternatives, and a voter's ranking is consistent with those latent utilities. From this viewpoint, a natural goal for a voting rule — which aggregates the given rankings — would be to select a good alternative in terms of utilitarian social welfare (the sum of utilities), despite having access only to ordinal information. The notion of *distortion* [231] measures how far a given voting rule is from achieving this goal. It is the worst-case ratio between the social welfare of the optimal alternative and the social welfare of the alternative selected by the rule, where the worst case is taken over utilities.

Without any additional assumptions, any deterministic voting rule must have unbounded distortion. Intuitively, even in an election with two voters, one preferring $a$ to $b$ and the other preferring $b$ to $a$, it could be the case that one voter has arbitrarily high utility for their preferred alternative whereas the other has low utility for both alternatives, but it is impossible to differentiate between $a$ and $b$ based on the ordinal information alone. To circumvent this obstacle, the rich literature on distortion has explored several approaches, one of which is to assume access to limited cardinal information (see Section 9.1.2 for details). Our work is inspired by this approach;

we are interested in a specific and very natural type of cardinal information: *stakes*.

To motivate the idea of stakes, let us consider a concrete example. Cambridge, Massachusetts, like many US cities, seeks public input on the questions surrounding affordable housing. In past years, the city has proposed to permit building taller buildings — a move that would increase affordable housing, but would change the skyline of Cambridge and potentially cast shadows over existing homes. City Councilor Burhan Azeem, who proposed the amendments, said "I understand that tall buildings are something that people are sensitive to, but this comes down to which should we care more about. How tall a building is? Or the people who dont have stable housing?" [60]. Councilor Azeem is pointing out a contrast of *stakes*: while both homeowners and those who don't have stable housing are affected by the decision of whether to permit taller buildings, the latter group is *far more affected*.

Now, suppose the decision is put to public referendum; placing this example in the utility model above (using numbers to illustrate the intuition), suppose that 10% of residents are renters who are at risk of losing their home due to rising rental prices, and 90% are homeowners who are worried that affordable housing would lead to a decrease in the value of their property. Let the former group have utility 100 for accepting the proposal and utility 0 for rejecting it; let the latter group have utility 1 for rejecting the proposal and utility 0 for accepting it. Based on their utilities, 90% of residents will vote to reject the proposal, and any majority-consistent voting rule must confirm their choice. This is a severely suboptimal outcome, as the social welfare of accepting the proposal is more than 10 times that of rejecting it.

Notice that, in the foregoing example, the chosen numbers do not matter much: the key problem is that a minority of residents have disproportionately high stakes, but they lack the voting power to sway the election. By contrast, the majority stands to gain little, so when they get their way, little value is generated for the population. It is not hard to think of salient examples of real political decisions with this property: consider masking requirements and immunocompromised populations, accessibility features (e.g., wheelchair ramps) and those who rely on them, or the design of public transit and those who cannot afford private transportation. Although the impossibility that all deterministic rules have unbounded distortion is often thought of as a theoretical one, these examples make the issue seem practically pressing: given that people *are* likely to have disparate stakes in real issues — and sometimes minority groups have much higher stakes than the majority — these examples suggest that such welfare loss can occur in *real elections*.

Motivated by this problem, we pursue bounded distortion by assuming that voters' stakes are known, at least approximately, to the voting rule. Unlike a significant branch of the distortion literature that assumes voters' utilities are normalized, we will allow voters' utilities to be arbitrary and unknown to ensure that our model captures the problematic examples above. One may wonder whether it is plausible to know stakes information but not precise utilities. To see why knowing stakes is far easier, consider the affordable housing example: conceptually, stakes captured *how affected each voter is relative to other voters*. Stakes information, then, is just a single number measuring the extent to which a voter can gain or lose depending on the decision outcome, *relative to other voters in the election*. As the number of alternatives in an election grows,

the gap in difficulty widens between the doing a coarse-grained assessment of how relatively affected a voter is (which in the examples above is not hard to do) versus understanding individuals' fine-grained preferences over all alternatives. As we discuss below, our results permit having just approximate stakes information; going further, we then initiate a study of how voters' stakes can be revealed in their behavior, potentially permitting our positive results to hold even in cases where legitimate estimates of voters' stakes are unavailable.

### 9.1.1 Approach and contributions

**How should we measure stakes? (Section 9.2).** First, we must embed a model of stakes in the standard model of voting with latent utilities. The affordable housing example illustrates that, intuitively, a voter's stakes are captured in their *utility vector* — that is, their utilities across alternatives. In that example, it seems natural to measure voters' stakes as the difference between their utilities (so the respective stakes of renters and homeowners would be 100 and 1). However, how to measure stakes becomes less obvious when $m > 2$. (Consider the utility vectors $(1, 1, 0)$ and $(1, 0, 0)$. Which reflects higher stakes?) We thus define and study general *stakes functions* $s$: any mapping from a voter's utility vector to a scalar measure of their stakes in the election.

**What can we do with *perfect* stakes information? (Section 9.3).** In the affordable housing example, an intuitive idea for addressing high distortion would have been to re-weight votes according to voters' stakes. Taking this approach, we characterize the distortion possible, across stakes functions $s$, by any deterministic or randomized rule when votes are re-weighted in a stakes-proportional way.

*Deterministic rules.* Here we find that knowing stakes information — even according to a surprisingly simple stakes function — can drop the distortion from $\infty$ to the number of alternatives $m$. We first prove a lower bound showing that no deterministic voting rule, when reweighted by *any stakes function* $s$ according to *any reweighting scheme*, can achieve distortion lower than $m$ (Theorem 9.3.1). We then prove a general upper bound on the distortion of *any* deterministic voting rule, when votes are reweighted proportionally via *any* stakes function $s$ (Theorem 9.3.4). We use this bound to identify a stakes function, voting rule pair that matches our lower bound: we show that the rule Plurality achieves optimal distortion $m$ when reweighted according to the stakes as measured by *maximum utility* or the *difference between maximum and minimum utilities* (Proposition 9.3.7). We henceforth refer to these stakes functions as $s = \max$ and $s = \text{range}$, respectively. This bound further implies proportional re-weighting is sufficient to achieve optimality. Beyond Plurality, we surprisingly find that among deterministic voting rules, most are not helped by knowledge of stakes — a result which has both positive and negative interpretations.

*Randomized rules (plus, an independently interesting lemma).* We repeat this analysis for randomized rules, showing that if a stakes-reweighted voting rule is permitted to be randomized, the distortion can be as low as $O(\sqrt{m})$. We show that the rule Stable Lottery [105] achieves $O(\sqrt{m})$ distortion when given stakes measured by $s = \max$, $s = \text{range}$, or $s = \text{sum}$ (the *sum* of a voter's utilities) (Theorem 9.3.12). The lemma used to prove this upper bound may be of independent interest, as it shows a surprising and technically useful connection between our

setting and the popular setting of distortion assuming *unit-normalized utilities* (i.e., where voters' utilities are assumed to sum to 1). We observe, first, that assuming voters' utilities sum to 1 is akin to assuming they have *identical stakes*, as measured by the stakes function $s = \text{sum}$. Then, we show that this assumption is *equivalent*, from a distortion perspective, to permitting arbitrary utilities but reweighting votes by stakes measured by $s = \text{sum}$. In fact, we show this equivalence holds for *any 1-homogeneous stakes function* (Lemma 9.3.13). This result establishes a formal link — and permits the transfer of bounds — between two key assumption classes in the distortion literature: unit-normalized utilities and queries of cardinal information (our setting). Finally, we conclude our analysis with a lower bound proving that Stable Lottery's distortion is within a log $m$ factor of optimal across randomized rules and stakes functions (Theorem 9.3.11).

**What if we just have stakes information in *orders of magnitude*? (Section 9.4).** In the motivating examples above, we can identify one or more population groups who are substantially disproportionately affected by the decision. However, assigning any single number to the *extent* to which their stakes are higher is difficult; in reality, people's stakes may be observable (or even fundamentally measurable) only at the resolution of orders of magnitude. Fortunately, in Section 9.4 we find that even very approximate stakes information can help: we give an instance-wise upper bound showing that for any 1-homogeneous stakes function $s$ (a natural class encompassing all $s$ discussed so far), if our stakes information is $\delta$-approximately correct (and adversarially-designed otherwise), the distortion of any rule $f$ scales simply by $\delta$ (Theorem 9.4.1). Given that the distortion of deterministic voting is otherwise unbounded, this result means that even just coarsely accounting for voters' differing stakes can offer substantial improvements in distortion.

**What if we don't have *any stakes information at all?* (Section 9.5).** In some cases, we may not even be able to guess voters' stakes at the resolution of orders of magnitude — or, even more likely, one may encounter cases where it is not possible to make a publicly acceptable case that a certain group should receive disproportionate representation in a given democratic decision. However, in many such cases, it is still desirable to account for stakes; this is an extremely sticky problem, which merits more exploration than we can do in a single paper. We present an initial study of this problem in Section 9.5, where we propose and explore a class of mechanisms that aims to *reveal voters' information in their behavior* and, in the process, accounts for stakes automatically. This class of mechanisms, called *multi-issue mechanisms*, is based on a simple idea: it requires voters to decide how to allocate a total allotment of voting power across elections. In forcing voters to make such trade-offs, this class of mechanisms exploits a special property of stakes information: unlike other kinds of cardinal utility information, stakes-information is *action-relevant*, describing, roughly, how much a voter may care about the outcome of a given election.

After defining this class of mechanisms, we perform a proof-of-concept analysis of a mechanism in this class. This example illustrates mathematically how voters' stakes can be revealed in their behavior, and how our results from the previous sections can be applied to analyze such a mechanism. Our distortion analysis of this mechanism shows that although any election run individually within the mechanism could have unbounded distortion, the distortion of each elec-

tion in our mechanism is at most $\delta m^2$, where $\delta$ corresponds to the extent to which voters have the same *total stakes* across the issues we place on the multi-issue ballot (Theorem 9.5.4).

### 9.1.2 Related work

The literature on distortion is quite rich; for an overview, we refer the reader to the survey by Anshelevich et al. [25]. At a high level, there are at least three avenues to achieving meaningful bounds on distortion. The first (and by far the most common) is to restrict the utilities, e.g., by assuming that voters have the same sum of utilities [53, 231], or by assuming that the utilities are induced by an underlying metric space [24]. The second is to consider *public spirit*, in the sense that voters seek to optimize social welfare in addition to their own utilities [134]. And the third, which is most relevant to our work, is to assume the availability of limited cardinal information.

In the context of the last approach, the work of Amanatidis et al. [22] is most closely related to ours. They study deterministic voting rules with access to one of two kinds of queries: *value* queries, where the voting mechanism can directly ask agents about any one of their utilities; and *comparison* queries, where the voting mechanism can ask agents: "for alternatives $a$ and $b$, is your utility for $a$ at least $\tau$ times your utility for $b$?" Among other results, they prove that constant distortion is achievable using $O(\log^2 m)$ queries per voter, where $m$ is the number of alternatives. To achieve their upper bounds, they construct an approximate utility profile via the queries and maximize social welfare with respect to these estimated utilities. This is conceptually very different from our approach of employing common voting rules and accounting for stakes through stakes-proportionality. Nevertheless, there are a few technical connections between the work of Amanatidis et al. [22] and ours, which comment on below.

Another (more distantly) related paper in the same vein is that of Abramowitz et al. [17]. Their main results pertain to a setting where answers to the above comparisons queries are given for every pair of alternatives, either with respect to a single fixed threshold $\tau$ or multiple fixed thresholds; their distortion bounds are parameterized by these thresholds. They additionally assume an underlying metric space and therefore their results are technically incomparable to ours.

Although standard social choice mechanisms do not account for stakes,[1] the concept has been conceived of in multiple disciplines. From the social sciences, there is the philosophical notion of *proportionality* — the idea that "power should be distributed in proportion to peoples stakes in the decision under consideration" [57]. In the context of binary decisions, **?** ] explores the welfare cost of treating agents symmetrically when they have different stakes, and the results of Fleurbaey [136] suggest that accounting for stakes can help: they show that in elections over two alternatives, reweighting each voter's vote by the difference of their utilities — a measurement of their "stakes" — increases the welfare of majority voting. Our work can be seen as generalizing the latter analysis substantially, permitting $m$ alternatives, any stakes function $s$, any stakes-reweighting scheme, and any voting rule.

The idea that those with higher stakes in a decision should have greater political influence arises

[1] We know of two social choice papers that use the term "stakes" [23, 172]; both use the term differently and explore unrelated questions.

in many other theories in political science as well, such as the *principle of affected interests* [140]; the concept of *empowered inclusion* [38, 278]; and the concept of *precarity* [210], which describes the idea that socially or economically vulnerable populations may be more affected by political issues due to their inability to adjust to decisions that are sub-optimal for them. Our work can be seen as a technical companion to this literature in three ways: it (1) offers a formal framework for modeling stakes, (2) explores the impact of accounting for stakes as proposed by these theories, and (3) identifies a class of mechanisms which can be explored further as a method for accounting for stakes.

Our mechanism design approach requires assumptions about how voters will engage with the mechanism, which can be informed by existing social science research on how voter behavior depends on preference intensity [99, 115]. Moreover, our proposed class of mechanisms relates to (but is not encompassed by) two existing voting mechanisms, *quadratic voting* [230] and *storable votes* [72]. We defer a detailed discussion of these connections to Sections 9.5 and 9.6.1, where we can establish them more concretely in comparison to the mechanisms we study.

## 9.2 Model

We introduce the model in two parts. Section 9.2.1 establishes the standard voting model; Section 9.2.2 embeds our model of voters' stakes within it. Throughout the paper, we use the shorthand $\mathbf{1}_\ell\mathbf{0}_{\ell'}$ to mean a vector containing $\ell$ ones followed by a string of $\ell'$ zeros. We let $\mathbb{I}(\cdot)$ be the indicator function.

### 9.2.1 The voting model

In an election, there are $n$ voters and $m$ *alternatives*. We let voters $i \in [n]$ and alternatives $a \in [m]$ have some fixed numbering. Voters' underlying preferences over $[m]$ are modeled with *utilities*: each voter $i$ has a utility $u_i(a) \in \mathbb{R}_{\geq 0}$ for each alternative $a \in [m]$. Let $\mathbf{u}_i = (u_i(a) | a \in [m])$ be $i$'s *utility vector*, and let $\mathbf{u}$ be a generic utility vector. We summarize all voters' utilities in a *utility matrix* $U \in \mathbb{R}_{\geq 0}^{n \times m}$.

**The voting process.** Each voter $i$ expresses their preferences via a complete *ranking* over (i.e., permutation of) $[m]$. Letting $S_m$ be the set of all permutations of $[m]$, a generic ranking is $\pi \in S_m$. We use $a \succ_\pi a'$ to denote that $a$ precedes $a'$ in $\pi$, reflecting that $a$ is preferred over $a'$. Abusing notation slightly, we let $\pi(j)$ denote the alternative ranked in the $j$-th position in ranking $\pi$. When voter $i$ "votes", they submit a ranking $\pi_i$. This ranking is determined by $\mathbf{u}_i$: $i$ ranks alternatives in decreasing order of their utilities, so that $u_i(a) > u_i(a') \implies a \succ_{\pi_i} a'$ for all $a, a' \in [m]$.[1]

A collection of $n$ voters' rankings is called a preference *profile* $\boldsymbol{\pi}$. As in prior work such as that of Xia [287], instead of working with profiles $\boldsymbol{\pi}$, we will work with *histograms*, which summarize collections of rankings by their frequencies. A generic preference *histogram* is a vector indexed by

[1] For simplicity of our lower bounds, we will assume worst-case rankings when $u_i(a) = u_i(a')$; one could instead tie-break explicitly by perturbing the utilities by arbitrarily small amounts.

rankings, $\mathbf{h} = (h_\pi | \pi \in S_m)$, where $h_\pi \in [0,1]$ is the fraction of rankings in a given collection equal to $\pi$. As such, $\|\mathbf{h}\|_1 = 1$. The space of all possible preference histograms is thus the simplex of all valid distributions over $S_m$, which we call $\Delta(S_m) := \left\{\mathbf{h} \in [0,1]^{S_m} : \sum_{\pi \in S_m} h_\pi = 1\right\}$. Connecting profiles and histograms, we say $\boldsymbol{\pi}$ is *consistent* with $\mathbf{h}$ if each $\pi \in S_m$ appears in $\boldsymbol{\pi}$ exactly $n \cdot h_\pi$ times. Let $\Pi^{\mathbf{h}}$ be the set of all profiles consistent with a histogram $\mathbf{h}$. Note that $\Pi^{\mathbf{h}}$ is non-empty iff all entries of $\mathbf{h}$ are rational.

Since voters' rankings are fully implied by $U$, we let $U$ constitute an instance. We denote the histogram implied by $U$ as $\mathsf{hist}(U)$, whose $\pi$-th entry is given by

$$\mathsf{hist}_\pi(U) := 1/n \sum_{i \in [n]} \mathbb{I}\{\pi_i = \pi\}, \quad \text{for all } \pi \in S_m.$$

**Voting rules.** Let $\Delta([m])$ denote the set of all probability distributions over the alternatives $[m]$. Then, a *voting rule* is a function $f : \Delta(S_m) \to \Delta([m])$ that maps a preference histogram to a distribution over winning alternatives.[1] We refer to this class of functions as *randomized* rules to distinguish them from their sub-class, *deterministic* voting rules, which map a histogram to a distribution with singleton support. Among deterministic rules, the only specific rule we study is Plurality, whose winner is the alternative that is ranked first by the most voters. Among randomized rules, we consider the Stable Lottery rule [105], which draws a winner either at random or from a *stable lottery* — a distribution over a subset of $[m]$ that is preferred by voters to other such subsets. We will not apply this rule's precise definition, so we defer it to Appendix F.2.9.

**Distortion.** Let an alternative $a$'s utilitarian *social welfare* be $\mathsf{sw}(a, U) := \sum_{i \in [n]} u_i(a)$. We benchmark the social welfare of the winner against that of $a^* := \arg\max_{a \in [m]} \mathsf{sw}(a, U)$, the highest-welfare alternative. For any rule $f$, the value of this approximation ratio in instance $U$ is called the *instance-specific distortion*, defined as

$$\mathsf{dist}_U(f) := \frac{\mathsf{sw}(a^*, U)}{\mathbb{E}[\mathsf{sw}(f(\mathsf{hist}(U)), U)]},$$

where the expectation is over the draw of the winner from the distribution $f(\mathsf{hist}(U))$. As is standard, we evaluate $f$ via its overall *distortion*, $\mathsf{dist}(f)$, which is the worst-case approximation ratio over all possible instances $U$:

$$\mathsf{dist}(f) := \sup_{n \geq 1} \sup_U \ \mathsf{dist}_U(f).$$

The supremum over $n$ is just to more conveniently deal with the fact that in worst-case instances, $n$ must be large enough relative to $m$ to realize utility matrices with $m$-dependent fractional compositions. We consider the distortion to be a function of $m$, as is standard in the literature.

### 9.2.2 A stakes framework within the voting model

**Measuring stakes via *stakes functions*.** A *stakes function* is any function $s : \mathsf{R}^m_{\geq 0} \to \mathsf{R}$ that maps a utility vector to a scalar measure of the stake it reflects. Intuitively, a voter's stake should

[1] Histograms are inherently anonymized, so we study only anonymous voting rules (encompassing all common voting rules).

depend on the relative magnitudes of their utilities, but not *which* alternatives they prefer; we thus restrict to functions $s$ which are *permutation invariant*. For example, utility vectors $(0, 1)$ and $(1, 0)$ reflect the same stake. We often apply this invariance to evaluate voters' stakes on a *sorted* version of their utility vector.

In some results, we restrict our consideration to stakes functions that are 1-*homogeneous*, i.e., for all scalars $\alpha$, $s(\alpha \mathbf{u}) = \alpha s(\mathbf{u})$. This applies to $\alpha = 0$, implying that for all 1-homogeneous $s$, $s(\mathbf{0}) = 0$. This restriction on $s$ is natural in that it makes our notion of *accounting for stakes*, formalized below, invariant to rescaling $U$. Although many of our results apply for generic stakes functions, three in particular will come up frequently, so we define shorthand for them:

$$\mathsf{range}(\mathbf{u}) := \max_a u(a) - \min_a u(a), \qquad \mathsf{max}(\mathbf{u}) := \max_a u(a), \qquad \mathsf{sum}(\mathbf{u}) := \textstyle\sum_a u(a)$$

**Stakes-reweighting of votes.** Colloquially, we say that a voting process "accounts for stakes" if it grants voters representation to an extent that depends on their relative stakes. We can think of this as a form of stakes-dependent reweighting: instead of voter $i$'s ranking contributing to the $\pi_i$-th entry of the histogram with weight $1/n$, its contribution is additionally weighted by some function of $s(\mathbf{u}_i)$. We can also think of this as *recomposing* the electorate, by effectively duplicating voters in proportion to some function of $s(\mathbf{u}_i)$. For convenience of intuition and notation, we adopt the second interpretation. Formally, let $r : \mathsf{R}_{\geq 0} \to \mathsf{R}_{\geq 0}$ be a generic *recomposition function*. Then, the $(r, s)$-recomposed histogram arising from $U$ has $\pi$-th entry

$$\mathsf{hist}_\pi^{r \circ s}(U) = \frac{\sum_{i \in [n]} r(s(\mathbf{u}_i)) \cdot \mathbb{I}(\pi_i = \pi)}{\sum_{i \in [n]} r(s(\mathbf{u}_i))} \quad \forall \pi \in S_m,$$

representing the fraction of voters in the $(r, s)$-recomposed electorate with ranking $\pi$. In this recomposed histogram, each voter $i$'s ranking is represented with weight $r(s(\mathbf{u}_i))/\sum_{i \in [n]} r(s(\mathbf{u}_i))$.

***Proportional* stakes-reweighting of votes.** In this paper, we will focus on perhaps the simplest reweighting scheme: *stakes-proportionality*, where $r$ is the identity function $I$. In the *$s$-proportional* histogram $\mathsf{hist}^{I \circ s}(U)$, voters' votes are reweighted *in proportion* to their stakes, i.e., by $s(\mathbf{u}_i)/\sum_{i \in [n]} s(\mathbf{u}_i)$. For notational simplicity, we will shorten the name of $\mathsf{hist}^{I \circ s}$ to $\mathsf{hist}^s$ henceforth. We will use *stakes proportionality* (or $s$-proportionality) to refer to the condition under which votes are reweighted in this way.

**Distortion under stakes-proportionality.** In a given instance, reweighting an electorate to be stakes-proportionate will potentially change the distortion. We define the *$s$-distortion* of $f$ as its distortion in $\mathsf{hist}^s(U)$, the $s$-proportional electorate arising from $U$:

$$\mathsf{dist}_U^s(f) := \frac{\max_{a \in [m]} \mathsf{sw}(a, U)}{\mathbb{E}[\mathsf{sw}(f(\mathsf{hist}^s(U)), U)]}, \text{ and } \mathsf{dist}^s(f) := \sup_{n \geq 1} \sup_U \mathsf{dist}_U^s(f).$$

## 9.3 What if we have *perfect* stakes information?

### 9.3.1 Deterministic voting rules

We begin by analyzing deterministic voting rules, which without stakes information have infinite distortion (the proof is essentially the housing example in Section 9.1; for a formal proof, see Appendix F.2.1). Motivated by this impossibility, we will study what $s$-distortion is possible across deterministic rules and stakes functions.

#### Lower bound for all $f, s$.

Our first result is a lower bound that shows that the best possible $s$-distortion achievable by any deterministic rule $f$, given stakes information according to any stakes function $s$, is at least $m-1$.

**Theorem 9.3.1 (lower bound).** *For all $s$ and deterministic $f$,*

$$\mathrm{dist}^{\mathbf{s}}(f) \geq m-1.$$

*Proof sketch.* Our approach is to define two instances, $U$ and $U'$, and show that all deterministic rules $f$ must have at least $m-1$ distortion in one of these two instances. To construct $U, U'$, first set aside one alternative $a'$; let the remaining alternatives be $a_1 \dots a_{m-1}$. For any $\ell \in [m-1]$, define $A_\ell = \{a_j | j \in [m] \setminus \{\ell\}\}$. When we write $A_\ell$ in a ranking, it represents a ranking over all the alternatives within it *in increasing order of index.* Now, we define voters' rankings $\boldsymbol{\pi}$ and two possible underlying utilities $U$ and $U'$. Divide voters in into $m-1$ groups, and consider a voter $i$ in group $\ell$. Let them rank alternatives as $\pi_i = a_\ell \succ a' \succ A_\ell$. Their underlying utility vectors as given by $U$ and $U'$, called $\mathbf{u}_i$ and $\mathbf{u}'_i$, are defined below:

| alternative: | $a_\ell$ | $\succ$ | $a'$ | $\succ$ | $A_\ell$ |
|---|---|---|---|---|---|
| $\mathbf{u}_i$ for $i \in$ group $\ell$: | 1 | | 1 | | $0 \dots 0$ |
| $\mathbf{u}'_i$ for $i \in$ group $\ell$: | 1 | | 0 | | $0 \dots 0$ |

With this construction, the proof follows from three observations. *Observation 1:* First, because both $U$ and $U'$ result in the same ranking for each voter $i$, $\mathsf{hist}(U) \equiv \mathsf{hist}(U')$, and these two underlying utility matrices are indistinguishable to any voting rule. *Observation 2:* Within each utility matrix, all voters have the same ordered utility vector, and thus have the same stakes; formally, $\mathsf{hist}^s(U) \equiv \mathsf{hist}(U)$ and $\mathsf{hist}^s(U') \equiv \mathsf{hist}(U')$. *Observation 3:* In both $U$ and $U'$, the social welfare of any $a_\ell$ is equal to $n/(m-1)$; however, in $U$, $a'$ has high social welfare ($\mathsf{sw}(a', U) = n$) while in $U'$, $a'$ has low social welfare ($\mathsf{sw}(a', U') = 0$).

Now, when a voting rule $f$ receives the profile $\boldsymbol{\pi}$, either $f(\boldsymbol{\pi}) = a'$ or $f(\boldsymbol{\pi}) = a_\ell$ for some $\ell \in [m-1]$. If the former is true, *Observations 1-3* imply that for any $s$, $\mathrm{dist}^s_{U'}(f) = \frac{n/(m-1)}{0} = \infty$. If the latter is true, *Observations 1-3* imply that for any $s$, $\mathrm{dist}^s_{U'}(f) = \frac{n}{n/(m-1)} = m-1$. We spell out this argument in full formality in Appendix F.2.2. ◻

**Remark 9.3.2** (Extension to arbitrary recomposition functions)**.** *In the instance giving Theorem 9.3.1, either utility matrix gave all voters identical stakes. If $s(\mathbf{u}_i)$ is equal across voters, then for any recomposition function $r$, $r(s(\mathbf{u}_i))$ will be equal across voters. Observation 2 still holds, and the lower bound applies. Given that we will find this lower bound to be tight, it implies that in the worst case, applying a recomposition function other than the identity function will not improve the $s$-distortion of any deterministic voting rule, for any stakes function $s$.*

**Remark 9.3.3** (Connection to existing results)**.** *Theorem 7 of Amanatidis et al. [22] shows that any single value query of utilities (where the voting mechanism can directly ask agents about any one of their utilities) can enable at best $\Omega(m)$ distortion. Our lower bound in Theorem 9.3.1 generalizes this lower bound, showing that any system of queries yielding the value of a scalar-valued stakes function, when paired with a deterministic voting rule, can achieve at best $\Omega(m)$ distortion.*[1]

Upper bound for all $f, s$.

Now, we will prove upper bound on the $s$-distortion of *any* deterministic voting rule $f$, for *any* stakes function $s$. To reason about all voting rules $f$ and stakes functions $s$ at once, we must determine: *Given any pair of $s, f$, what properties of $s$ and $f$ will lead to low distortion?* We now introduce two such properties. First, $\beta_f$ is the minimum fraction of voters that must rank the winner by $f$ in first position:

$$\beta_f := \min_{\mathbf{h}\in\Delta(S_m)} \sum_{\pi\in S_m} h_\pi \cdot \ \mathbb{I}\{\pi(1) = f(\mathbf{h})\}.$$

Second, $\kappa$-upper$(s)$ and $\kappa$-lower$(s)$ measure the extent to which $s$ can over- or under-estimate $\max(\mathbf{u})$, respectively:

$$\kappa\text{-upper} := \sup_{\mathbf{u}} \tfrac{s(\mathbf{u})}{\max(\mathbf{u})}, \quad \kappa\text{-lower}(s) := \inf_{\mathbf{u}} \tfrac{s(\mathbf{u})}{\max(\mathbf{u})}.$$

While bounds in terms of other properties of $s, f$ are conceivable, these quantities will permit optimal upper bounds.

In terms of these quantities, Theorem 9.3.4 gives an upper bound on the $s$-distortion for any $s$ and any deterministic $f$. The proof relies on the insight that $\beta_f$ and the $\kappa$ values are linked: $\beta_f$ lower bounds how often the winner is ranked first, while the $\kappa$'s links the stakes and maximum utility of any voter who ranks the winner first. This connection implies a lower-bound on the social welfare of the winner.

**Theorem 9.3.4 (upper bound).** *For all $s$ and deterministic $f$,*

$$\text{dist}^{\mathbf{s}}(f) \le \beta_f^{-1} \cdot \kappa\text{-upper}(s) \,/\, \kappa\text{-lower}(s).$$

*Proof.* Fix an instance $U$, a stakes function $s$, and a deterministic rule $f$. Let $a' = f(\text{hist}^s(U))$ be the winner of the $s$-proportional election. First, we have that the social welfare of any alternative

[1] Note: a necessary step in showing that our lower bound subsumes theirs is arguing that our lower bound actually applies to *any stakes-dependent electoral recomposition*, not just proportional recomposition, which we do.

$a$ is upper-bounded:

$$\mathrm{sw}(a, U) \leq \sum_{i \in [n]} \max(\mathbf{u}_i) \leq \sum_{i \in [n]} s(\mathbf{u}_i) \,/\, \kappa\text{-lower}(s). \tag{9.1}$$

Now, let $N_{a'}$ be the set of voters who rank $a'$ first. All $i \in N_{a'}$ must have at least some utility for $a'$:

$$u_i(a') = \max_a u_i(a) \geq s(\mathbf{u}_i)/\kappa\text{-upper}(s). \tag{9.2}$$

Also, since $a'$ is the winner, $N_{a'}$ composes at least a $\beta_f$ fraction of the stakes-proportional electorate:

$$\textstyle\sum_{i \in N_{a'}} s(\mathbf{u}_i) \,/\, \sum_{i \in [n]} s(\mathbf{u}_i) \geq \beta_f.$$

This fact, combined with Equation (9.2), gives that

$$\mathrm{sw}(a', U) \geq \sum_{i \in N_{a'}} u_i(a') \geq \sum_{i \in N_{a'}} s(\mathbf{u}_i) \,/\, \kappa\text{-upper}(s) \geq \beta_f \sum_{i \in [n]} s(\mathbf{u}_i) \,/\, \kappa\text{-upper}(s).$$

Combining this with Equation (9.1) and denoting the maximum welfare alternative by $a^*$, we obtain that

$$\mathrm{dist}^s_U(f) = \frac{\mathrm{sw}(a^*, U)}{\mathrm{sw}(a', U)} \leq \beta_f^{-1} \cdot \frac{\kappa\text{-upper}(s)}{\kappa\text{-lower}(s)}. \qquad \square$$

**Remark 9.3.5.** *Theorem 9.3.4 also holds if in the definitions of $\kappa$-upper$(s)$ and $\kappa$-lower$(s)$, max is replaced with range. This is because the worst-case distortion can always be realized by instances where every voter has minimum utility 0, in which case max = range. We prove this in Appendix F.2.3.*

### Optimality of Plurality and $s$ =max

In our first extension of these results, we identify a voting rule-stakes function pair that attains the best possible $s$-distortion $m$, matching our lower bound in Theorem 9.3.1. We do this by minimizing the upper bound in Theorem 9.3.4, which amounts to choosing $s$ and $f$ to minimize both $\kappa$-upper$(s)/\kappa$-lower$(s)$ and $\beta_f^{-1}$.

It is easy to see that $s = \max$ achieves the minimal value $\kappa$-upper$(s)/\kappa$-lower$(s) = 1$. For $\beta_f$, the answer is more subtle; in Appendix F.2.4, we prove the following lemma, which shows that the maximal attainable value is $\beta_f = 1/m$, and that this maximum is achieved by Plurality.

**Lemma 9.3.6.** *For any deterministic voting rule $f$, $\beta_f \leq 1/m$, and $\beta_{\textsc{Plurality}} = 1/m$.*

With this lemma in hand, we now apply Theorem 9.3.4 to conclude the following upper bound, which matches the lower bound in Theorem 9.3.1.

**Proposition 9.3.7.** *$\mathrm{dist}^{max}(\textsc{Plurality}) \leq m$.*

The above result suggests using Plurality with $s = \max$ as a promising choice. However, in some motivating contexts — e.g., where stakes-proportionality arises from voters' behavior — we may not be able to control which stakes function is used. To characterize the $s$-distortion of Plurality, we show that it is essentially tight with respect to Theorem 9.3.4 for all $s$, except that $\kappa$-lower is replaced with $\tilde{\kappa}$-lower, in which the supremum is taken over only utility vectors $\mathbf{u}$ with the same first and second entry. See Appendix G.1.10 for the formal definition of $\tilde{\kappa}$-lower and subsequent proof.

**Proposition 9.3.8.** *For all $s$, $\mathrm{dist}^{\mathbf{s}}(\textsc{Plurality}) \geq (m-1) \cdot \frac{\kappa\text{-upper}(s)}{\tilde{\kappa}\text{-lower}(s)}$.*

**Remark 9.3.9** (Connection to existing results)**.** *Theorem 1 of Amanatidis et al. [22] shows that their voting mechanism 1-PRV — which is equivalent to Plurality under stakes-proportionality with respect to max — gives distortion $O(m)$. This result corresponds to our upper bound on the max-distortion of Plurality, proven via Proposition 9.3.7.*

### Beyond Plurality

To understand the $s$-distortion of other deterministic voting rules $f$, we first observe that when $\beta_f = 0$, there is an unbounded gap between the lower bound in Theorem 9.3.1 and the upper bound in Theorem 9.3.4. Unfortunately, the $s$-distortion for such $f$ is indeed unbounded — a finding that is practically significant because, as we prove in Appendix F.2.6, most popular voting rules have $\beta_f = 0$.

**Proposition 9.3.10.** *For all stakes functions $s$ and all deterministic rules $f$ with $\beta_f = 0$, $\mathit{dist}^s(f) = \infty$.*

*Proof.* Let $f$ satisfy $\beta_f = 0$, and fix a histogram $\mathbf{h}$ in which the winner $f(\mathbf{h})$ is never ranked first. Then, set the underlying $U$ to realize this histogram while setting each voter's ordered utility vector to $\mathbf{1}_1\mathbf{0}_{m-1}$. Since the winner is never ranked first, it must get 0 average utility. Since each voter gives their respective first-ranked alternative utility 1, at least one alternative must have at least $1/m$ average utility; thus, $\mathrm{dist}_U(f) = \infty$ is unbounded. Because all voters have identical utility vectors, for all $s$, $\mathsf{hist}(U) = \mathsf{hist}^s(U) \implies f(\mathsf{hist}(U)) = f(\mathsf{hist}^s(U)) \implies \mathrm{dist}_U(f) = \mathrm{dist}^s_U(f) = \infty$. □

Proposition 9.3.7 and Proposition 9.3.10 together point to Plurality-like rules as uniquely promising when stakes are accounted for. This may seem strange, as Plurality is often considered a "bad rule" due to its lack of expressiveness. One positive interpretation of this finding is that Plurality, when stakes are accounted for, actually accounts for the most critical information; this is good news, as Plurality-like voting methods are widely used. Another possibility is that the ranking-based ballot format is insufficiently expressive; as we discuss in Section 9.6, our model and approach extend easily to richer ballot formats.

It is also important to acknowledge that although $\beta_f = 0$ for many non-Plurality voting rules, this is a *worst case* result. It could very well be that in typical instances, the winners chosen by

other voting rules are often ranked first by many voters. In this case, our impossibility in Proposition 9.3.10 would be quite pessimistic. This possibility motivates beyond-worst-case analysis, and/or simple sufficient conditions under which a broader range of voting rules can make use of stakes information. We leave these directions to future work.

### 9.3.2 Randomized voting rules

Next, we give an analogous but brief analysis of randomized rules. First, we lower-bound the $s$-distortion across all $f$ and all 1-homogeneous $s$, characterizing what $s$-distortion stakes-proportionality will permit. For comparison, randomized rules have at best $\Omega(m)$ distortion without accounting for stakes (see Appendix F.2.10 for details).

**Theorem 9.3.11 (lower bound).** *For all 1-homogeneous $s$, randomized $f$,* $\text{dist}^s(f) \geq \frac{\sqrt{m}}{10+3\log m}$.

*Proof sketch.* The construction of this lower bound is rather intricate, and its full proof is deferred to Appendix F.2.7. The main idea is to identify indices of the utility vector over which large gaps in utilities have the smallest effects on the stakes. More formally, we try to identify a small $z \in [m]$ such that the following quantity is upper-bounded:

$$\frac{s(\mathbf{1}_{z+1}\mathbf{0}_{m-(z+1)})}{s(\mathbf{1}_z\mathbf{0}_{m-z})}.$$

Then, we can exploit this lack of sensitivity of the stakes function to drive large gaps in alternatives' utilities over this index. If such a small $z$ doesn't exist, we know that placing utility gaps over all early indices of a utility vector will create a lower-bounded gap in stakes, which we can then exploit via a different construction. As in previous lower bounds, in the instance we design, all voters have identical stakes. □

Our next result proves that a known randomized rule, Stable Lottery, paired with one of a few stakes functions $s$, achieves $O(\sqrt{m})$ $s$-distortion, matching our lower bound up to a log factor. This rule was originally introduced with the goal of achieving low distortion in the distinct setting assuming normalized utilities; there, Stable Lottery was shown to achieve distortion $O(\sqrt{m})$ when utilities were restricted to have have unit sum, i.e., $\sum_{a\in[m]} u_i(a) = 1$ for all $i$ [105]. The theorem is proven via Lemma 9.3.13, which shows a connection between our model and the popular normalized utilities model that may be of independent interest.

**Theorem 9.3.12 (upper bound).** *For $s \in \{sum, max, range\}$,* $\text{dist}^s(\text{Stable Lottery}) \in O(\sqrt{m})$.

*Proof.* We prove this upper bound by connecting stakes information to the popular normalized utility model, in which it is assumed that voters' utilities are normalized to sum to 1 [69, 70, 231] (or, in recent work, have *maximum* utility 1 [105]). Placed within our model, observe that such assumptions amount to assuming that voters have *identical stakes* as measured by sum (resp. max).[1]

[1]This interpretation is already informative, as it raises substantive questions about whether these widely-used restrictions on the utilities are likely to hold in practice, where people may have dramatically differing stakes.

Of course, one could assume this normalization with respect to *any* stakes function $s$; we call this general class of assumptions *$s$-unit-stakes* assumptions.

We now show a surprising equivalence: Assuming sum-unit stakes is *equivalent*, from a distortion perspective, to the sum-distortion (i.e., the distortion achievable under $s$-proportionality when $s = \text{sum}$). In fact, we show this correspondence to hold for *any 1-homogeneous stakes function*, as stated informally in Lemma 9.3.13.

**Lemma 9.3.13** (informal)**.** *For all 1-homogeneous $s$, the distortion of $f$ under the $s$-unit-stakes assumption is equivalent to its $s$-distortion.*

The bidirectional reduction that proves this lemma is pictured in Figure 9.1. We prove it formally over two theorems in Appendix F.2.8: Theorem F.2.5 handles rational-valued histograms, and Theorem F.2.7 extends the claim to real-valued histograms (at the cost of a mild technical condition on the voting rule). With this reduction in hand, we can now transfer lower and upper

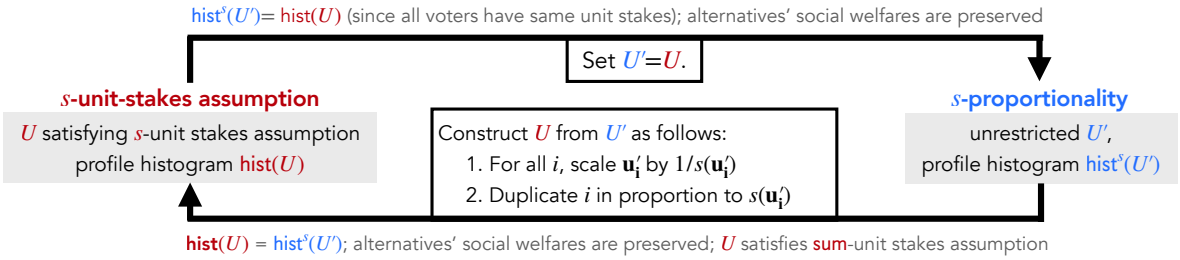



Figure 9.1: Constructions for reducing between the $s$-unit stakes assumption (existing model) and $s$-proportionality (our model).

bounds between the $s$-unit stakes setting and ours. We conclude the proof of Theorem 9.3.12 by transferring bounds on Stable Lottery proven under the sum-, max-unit stakes assumptions [105, Theorem 3.4]. The case of range requires a minor technical extension — see Appendix F.2.9 for details. □

**Implications for the $s$-unit stakes literature.** This reduction proves a formal connection between two distinct branches of the literature: distortion under normalized utilities, and distortion with auxiliary utility information. Using the fact that this reduction permits transferring bounds across these settings, in Appendix F.1.1 we illustrate how our results recover — and even sometimes strengthen — existing results, often via simpler arguments.

Beyond existing results, we can also use our reduction to immediately derive new results for the $s$-unit stakes assumption setting. Although there is a vast space of possible *$s$-unit-stakes* assumptions (one per possible stakes function $s$), the literature has explored few; to our knowledge, only max and sum. This inspires the open question: *would assuming unit stakes with respect to a different stakes function permit better distortion bounds?* Our results immediately close this question for deterministic rules:

**Corollary 9.3.14** (of Lemma 9.3.13)**.** *For all 1-homogeneous stakes functions $s$ and all deterministic voting rules $f$, the best achievable distortion of any $f$ under any $s$-unit stakes assumption is $m-1$ (Theorem 9.3.1), and this is achieved by assuming* max-unit stakes or range-unit stakes and using the rule Plurality (Proposition 9.3.7, Remark 9.3.5).

Our results also close this question within a factor of $\log(m)$ for randomized rules:

**Corollary 9.3.15** (of Lemma 9.3.13)**.** *For all 1-homogeneous stakes functions $s$ and all randomized voting rules $f$, the best achievable distortion of any $f$ under any $s$-unit stakes assumption is at least $\Omega(\sqrt{m}/\log(m))$ (Theorem 9.3.11), and this distortion is achieved within a $\log(m)$ factor by assuming* max-unit stakes, range-unit stakes, or sum-unit stakes and using the rule Stable Lottery (Theorem 9.3.12).

## 9.4 What if we have *approximate* stakes information?

In many cases, we may know voters' stakes only coarsely, at the level of orders of magnitude. To understand what is possible in this case, we now extend our results from previous sections to the case where our estimates of voters' stakes are incorrect.

Let us formally define these errors. Suppose we achieve $s$-proportionality according to an incorrect estimate of each voter $i$'s stakes $\hat{s}(\mathbf{u}_i) := \delta_i\, s(\mathbf{u}_i)$, where $\delta_i \geq 1$ is the factor by which we overestimate $i$'s stakes.[1] Let $\boldsymbol{\delta} := (\delta_i | i \in [n])$ be the vector of all such errors. Given $U$ and $\boldsymbol{\delta}$, we denote the $\boldsymbol{\delta}$-approximately stakes-proportional histogram as $\mathsf{hist}_\pi^{\boldsymbol{\delta},s}(U)$, with $\pi$-th entry

$$\mathsf{hist}_\pi^{\boldsymbol{\delta},s}(U) := \frac{\sum_{i\in[n]} \hat{s}(\mathbf{u}_i)\cdot \mathbb{I}(\pi_i = \pi)}{\sum_{i\in[n]} \hat{s}(\mathbf{u}_i)}.$$

For $\delta := \max_i \delta_i$, the $\delta, s$-distortion of $f$ is then given as

$$\mathsf{dist}^{\delta,s}(f) = \sup_{n\geq 1,\ U,\ \boldsymbol{\delta}\in[1,\delta]^n} \frac{\max_a \mathsf{sw}(a,U)}{\mathsf{E}[\mathsf{sw}(f(\mathsf{hist}^{\boldsymbol{\delta},s}(U)),U)]}.$$

Note that for fixed $\delta$, by this definition, $\boldsymbol{\delta} \in [1,\delta]^n$ is chosen adversarially. We now prove strong, instance-wise robustness to such errors.

**Theorem 9.4.1.** *For all $f$, 1-homogeneous $s$, $U$, and $\delta \geq 1$,* $\mathsf{dist}_U^{\delta,s}(f) \leq \delta\, \mathsf{dist}_U^s(f)$.

The intuition is simple: since $s$ is 1-homogeneous, mis-estimating $i$'s stakes by up to $\delta$ is the same as overestimating voters' *utilities* by up to $\delta$. Such overestimates can change the distortion by at most a $\delta$ factor. We now formalize this intuition.

*Proof of Theorem 9.4.1.* Fix a utility matrix $U$, a 1-homogeneous stakes function $s$, and an error vector $\boldsymbol{\delta} \in [1,\delta]^n$. Let $\tilde{U}$ be the utility matrix where voter $i$'s utility vector is scaled by a factor of $\delta_i$, i.e., $\tilde{\mathbf{u}}_i = \delta_i \mathbf{u}_i$. Then, since $s$ is 1-homogeneous, we have that $s(\tilde{\mathbf{u}}_i) = \delta_i s(\mathbf{u}_i)$, and therefore

[1] We can realize any type of errors with $\delta \geq 1$, because the weighting of the resulting electorate is relative.

$\mathsf{hist}^s(\tilde{U}) \equiv \mathsf{hist}^{\boldsymbol{\delta},s}(U)$, directly implying that $f(\mathsf{hist}^s(\tilde{U})) \equiv f(\mathsf{hist}^{\boldsymbol{\delta},s}(U))$. Moreover, for every alternative $a$, it holds that $\mathsf{sw}(a,\tilde{U}) \in [\mathsf{sw}(a,U), \delta\mathsf{sw}(a,U)]$. It follows that

$$\mathsf{E}[\mathsf{sw}(f(\mathsf{hist}^{\boldsymbol{s}}(\tilde{U})),\tilde{U})] \leq \delta \cdot \mathsf{E}[\mathsf{sw}(f(\mathsf{hist}^{\boldsymbol{\delta},s}(U)),U)],$$

from which we deduce that

$$\begin{aligned}\frac{\max_a \mathsf{sw}(a,U)}{\mathsf{E}[\mathsf{sw}(f(\mathsf{hist}^{\boldsymbol{\delta},s}(U)),U)]} \leq \delta\frac{\max_a \mathsf{sw}(a,U)}{\mathsf{E}[\mathsf{sw}(f(\mathsf{hist}^{s}(\tilde{U})),\tilde{U})]} &= \frac{\max_a \mathsf{sw}(a,U)}{\max_a \mathsf{sw}(a,\tilde{U})} \cdot \delta \cdot \frac{\max_a \mathsf{sw}(a,\tilde{U})}{\mathsf{E}[\mathsf{sw}(f(\mathsf{hist}^{s}(\tilde{U}),\tilde{U})]} \\ &\leq \delta \cdot \mathsf{dist}^s(f).\end{aligned}$$

Taking suprema on the left hand side then completes the proof. □

### 9.5 What if we have *no* stakes information?

In this section, we are interesed in the case where voters may have extremely different stakes (in which case our results dictate that accounting for them is important), but either (1) we cannot even coarsely guess them, or (2) we cannot explicitly re-weight votes without public backlash. In this section, we propose a mechanism design concept for *revealing voters' stakes through their behavior*. We propose a class of mechanisms, called *multi-issue mechanisms*, based on a simple idea: if voters must decide how to allocate voting power over *multiple entire elections*, they will exert influence in the decisions that most affect them, thereby revealing their stakes. We define this class of mechanisms here, where *(MD)* indicates a decision by the mechanism designer:

**Multi-issue mechanisms:**

*Setup.* Each voter $i \in [n]$ is presented with a slate of $k$ entire elections *(MD)*, where each election $\ell$ is over its own set of alternatives $A^\ell$. Voter $i$ has utility vector $\mathbf{u}_i^\ell$ in election $\ell$.

*Voting.* To vote, each voter $i$ submits to each election two things: a ballot of some format *(MD)*, and a scalar *weight* $w_i^\ell \in [0,1]$ describing the weight $i$ wishes to place on election $\ell$. These weights are restricted such that $\sum_\ell c(w_i^\ell) \leq 1$ for all $i$, where $c : \mathbb{R} \to \mathbb{R}^+$ is a cost function *(MD)* describing how much a voter is charged per unit of weight placed on an election.

*Aggregation.* Each voter $i$'s ballot in each election $\ell$ is weighted by $w_i^\ell$, and a voting rule *(MD)* is then used to aggregate these weighted ballots into a winner.

We next perform a proof-of-concept analysis of a multi-issue mechanism. We emphasize that this analysis is meant to build intuition rather than to be the final word on the design of multi-issue mechanisms. In particular, the analysis will illustrate (1) how voters' individual incentives can reveal their stakes according to a natural stakes function, (2) what assumptions are required for analyzing such a mechanism, and (3) how our results from Section 9.3 and Section 9.4 can be applied to bound the distortion of such a mechanism. Then, with the intuition from this analysis in hand, in Section 9.6.1, we will discuss generalizations of this mechanism and connections to

the existing mechanisms *storable votes* [72] and *quadratic voting* [230], which use similar types of trade-offs to elicit richer utility information.

To fully define our multi-issue mechanism, we must instantiate the four aspects left up to the mechanism designer (those flagged with *(MD)*): the elections to be included, the ballot format, the cost function, and and the voting rule.

*Elections to be included.* We allow the inclusion of an arbitrary number of elections $k \geq 2$. For ease of exposition, we assume each election $\ell$ contains the same number of alternatives $m$ (this can easily be achieved by simply adding dummy alternatives).

*Ballot format.* Voters "vote" by submitting complete rankings as was assumed throughout the paper, in order to apply our results to analyze the mechanism.

*Cost function.* We let $c : \mathbb{R}^k \mapsto \mathbb{R}$ be quadratic,[1] so that given $\mathbf{x} := (x_i | i \in [k]), c(\mathbf{x}) = \sum_{i \in [k]} x_i^2$.

*Voting rule.* We will aggregate votes via Plurality, guided by our results showing its optimality among deterministic rules.

For the purposes of analyzing our mechanism, we must also define with what motives — and based on what information — voters translate their utilities into actions (votes). Our assumptions prioritize simplicity.

**Assumption 9.5.1.** *We assume that voters satisfy the following properties.*

- *Honest:* voters will submit rankings that are true to their underlying utilities.
- *Oblivious*: voters have uniform priors over all sets of $n-1$ other rankings in every election, i.e., they make the *impartial culture assumption* [145].
- *Not mathematicians:* voters believe their probability of pivotality in election $\ell$ increases linearly in $w_i^\ell$, the weight they place on their vote in that election.[2]
- *Utility-maximizing:* voters submit weights across elections to maximize their total expected individual utility (where the expectation is over the randomness of others' votes).

Finally, we need to introduce formal notation that will allow us to analyze multi-issue mechanisms. In each election $\ell \in [k]$, the underlying $n \times m$ utility matrix is denoted $U^\ell$; correspondingly, voters' utility vectors are $\mathbf{u}_i^\ell$ and individual utilities are $u_i^\ell(a)$. For a given stakes function $s$, we use shorthand $\boldsymbol{s}_i := s(\mathbf{u}_i^\ell)_{\ell \in [k]}$ to summarize $i$'s stakes across all $k$ elections.

[1] One may wonder if there is a correspondence between this mechanism and Quadratic Voting (QV). Technically, the cost quadraticity serves the same purpose, but our mechanism is over *multiple elections*, while QV occurs within a single election.

[2] While this is technically inconsistent with their uniform priors over others' votes, we see it as at least as reasonable as assuming voters compute their precise probabilities of pivotality: even under uniform priors, computing the exact marginal increase in probability of pivotality per unit of weight is a combinatorial calculation that we cannot expect voters to do. In light of this, linearity is a natural assumption.

Because we use Plurality, we are concerned only with each voter's first-ranked alternative, so we will consider histograms that are indexed by $a \in m$ rather than $\pi \in S_m$. Let $b_i^\ell \in [m]$ be $i$'s first-ranked alternative in election $\ell$; abusing notation, we will represent $i$'s vote in this election using $\boldsymbol{e}_{b_i^\ell}$, the $m$-length alternative-indexed basis vector with a 1 at the $b_i^\ell$-th index. We can then represent $i$'s weighted vote in election $\ell$ as the weighted basis vector $w_i^\ell \boldsymbol{e}_{b_i^\ell}$. Let $\boldsymbol{w}_i := (w_i^\ell | \ell \in [k])$ be the weights $i$ submits across elections. We let the resulting histogram in election $\ell$ be $\mathbf{h}^\ell := \sum_{i\in[n]} w_i^\ell \boldsymbol{e}_{b_i^\ell} / \sum_{i\in[n]} w_i^\ell$. Let $\mathbf{h}^\ell_{-i}$ be the histogram not including $i$'s vote, i.e., $\mathbf{h}^\ell_{-i} := \sum_{i\in[n]\setminus\{i\}} w_i^\ell \boldsymbol{e}_{b_i^\ell} / \sum_{i\in[n]} w_i^\ell$. We write $\mathbf{h} \sim \mathcal{I}$ to denote a histogram that reflects a profile drawn from the Impartial Culture model.

Now, we apply results from Sections 9.3 and 9.4 to show that although any election encompassed by the mechanism, if run in isolation, could have had unbounded distortion, within this mechanism all elections are guaranteed to have distortion at most order $m^2$ (Theorem 9.5.4). This analysis proceeds in 3 steps. In **Step 1**, we show how voters' stakes are revealed in how they allocate weights across elections, and we characterize the particular stakes function that arises from their behavior. Then, in **Step 2**, we show that this stakes function is quite natural, and per our previous results, behaves similarly enough to range to permit near optimal distortion when paired with Plurality. Finally, in **Step 3**, we characterize the distortion of each election in the mechanism, which reveals an important feature of our choice of the $k$ elections around which we design the mechanism.

**Step 1: A natural stakes function $s^*$ arises from voter behavior.** First, we show that the stakes function arising from voters' utility-maximizing behavior is simple and intuitive:

$$s^*(\mathbf{u}) := \max(\mathbf{u}) - \frac{\mathsf{sum}(\mathbf{u}) - \max(\mathbf{u})}{m-1},$$

the gap between their maximum utility and their average utility for the other alternatives.

**Lemma 9.5.2.** *In each election $\ell$, each voter $i$ weights their vote by $\hat{w}_i^\ell = s^*(\mathbf{u}_i^\ell) \,/\, \|\mathbf{s}_i^*\|_2$.*

*Proof.* Fix an $i$, whose behavior in the mechanism we will analyze. Let the alternatives in each election $\ell$ be $A^\ell$, and for all $\ell \in [k]$, let $b_i^\ell$ be $i$'s favorite alternative in the $\ell$-th election. Now, define the following events, some which depend on $\boldsymbol{w}_i$:

- $X^\ell$: $b_i^\ell$ wins among only the votes in $\mathbf{h}^\ell_{-i}$ (and thus also wins with $i$'s ranking weighted by $w_i^\ell$ added to the profile).
- $Z^\ell(\boldsymbol{w}_i)$: Some $a \neq b_i^\ell$ wins among only the votes in $\mathbf{h}^\ell_{-i}$ by a margin small enough that $i$'s vote for $b_i^\ell$ is *pivotal*.
- $\neg X^\ell \wedge \neg Z^\ell(\boldsymbol{w}_i)$: Some alternative $a \neq b^\ell$ wins among the votes in $\mathbf{h}^\ell_{-i} + w_i^\ell \boldsymbol{e}_{b_i^\ell}$ (i.e., regardless of whether $i$ votes for $b_i^\ell$ with weight $w_i^\ell$).

Note that these events are mutually exclusive. Now, we compute voter $i$'s expected reward in terms of these quantities:

$$\mathsf{E}_{\mathbf{h}_{-i}\sim\mathcal{I}}\left[\sum_{\ell\in[k]} u_i(\textsc{Plurality}(\mathbf{h}_{-i}+w_i^\ell \boldsymbol{e}_{b_i^\ell}))\right]$$
$$=\sum_{\ell\in[k]}\left(\left(\mathbf{Pr}[X^\ell]+\mathbf{Pr}\left[Z^\ell(\boldsymbol{w}_i)\right]\right)\cdot u_i(b_i^\ell)+Pr\left[\neg X^\ell\wedge\neg Z^\ell(\boldsymbol{w}_i)\right]\cdot\sum_{a\in A^\ell\setminus\{b_i^\ell\}}\frac{u_i(a)}{m-1}\right)$$

Unpacking this expression, if either event $X^\ell$ or $Z^\ell$ occurs, $i$ gets the utility associated with $b_i^\ell$. If neither occurs, $i$ expects some other alternative to win, and has uniform priors over other voters' rankings; thus their expected utility in this event is the average of all alternatives in $A^\ell\setminus\{b_i^\ell\}$. Again using $i$'s uniform priors over other voters' behavior, $Pr[X^\ell]=1/m$. Simplifying,

$$=\sum_{\ell\in[k]}\left(\left(1/m+\mathbf{Pr}[Z^\ell(\boldsymbol{w}_i)]\right)u_i(b_i^\ell)+\left(1-1/m-Pr[Z^\ell(\boldsymbol{w}_i)]\right)\cdot\sum_{a\in A^\ell\setminus\{b_i^\ell\}}\frac{u_i(a)}{m-1}\right)$$
$$=\sum_{\ell\in[k]}\left(\sum_{a\in A^\ell}\frac{u_i(a)}{m}+Pr[Z^\ell(\boldsymbol{w}_i)]\left(u_i(b_i^\ell)-\sum_{a\in A^\ell\setminus\{b_i^\ell\}}\frac{u_i(a)}{m-1}\right)\right).$$

The first term of the summand is fixed in the instance, so it is not decision relevant. Then, we have deduced that

$$\max_{\boldsymbol{w}_i}\mathsf{E}_{\mathbf{h}_{-i}\sim\mathcal{I}}\left[\sum_{\ell\in[k]} u_i(\textsc{Plurality}(\mathbf{h}_{-i}+w_i^\ell \boldsymbol{e}_{b_i^\ell}))\right]$$
$$=\max_{\boldsymbol{w}_i}\sum_{\ell\in[k]}Pr[Z^\ell(\boldsymbol{w}_i)]\left(u_i(b_i^\ell)-\sum_{a\in A^\ell\setminus\{b_i^\ell\}}\frac{u_i(a)}{m-1}\right)$$
$$=\max_{\boldsymbol{w}_i}\sum_{\ell\in[k]}Pr[Z^\ell(\boldsymbol{w}_i)]\left(\max(\mathbf{u}_i^\ell)-\frac{\mathsf{sum}(\mathbf{u}_i^\ell)-\max(\mathbf{u}_i^\ell)}{m-1}\right)$$
$$=\max_{\boldsymbol{w}_i}\sum_{\ell\in[k]}Pr[Z^\ell(\boldsymbol{w}_i)]\cdot s^*(\mathbf{u}_i^\ell)$$

Finally, by Assumption 9.5.1, voters act as though their probability of pivotality increases linearly with each additional unit of weight placed behind their vote. Applying this assumption,

$$=\max_{\boldsymbol{w}_i}\sum_{\ell\in[k]}w_i^\ell\cdot s^*(\mathbf{u}_i^\ell).$$

Then, subject to to the constraint that each voter has total weight 1, the final problem voters are solving is as follows, where $\hat{\boldsymbol{w}}_i$ describes their optimal weighting across elections:

$$\hat{\boldsymbol{w}}_i := \arg\max_{\boldsymbol{w}_i} \sum_{\ell \in [k]} w_i^\ell \cdot s^*(\mathbf{u}_i^\ell) \quad \text{s.t.} \quad \sum_{\ell \in [k]} (w_i^\ell)^2 \le 1. \tag{9.3}$$

We can compute the optimizer of this program by simply projecting the vector $\boldsymbol{s}_i^*$ onto the unit sphere, concluding that

$$\hat{\boldsymbol{w}}_i = \frac{\boldsymbol{s}_i^*}{\|\boldsymbol{s}_i^*\|_2}. \qquad \square$$

**Step 2: This stakes function $s^*$ is nearly optimal.** Next, using our previous distortion bounds, we find that $s^*$ is *nearly optimal* across all stakes functions when paired with Plurality:

**Lemma 9.5.3.** $\mathsf{dist}^{s^*}(\textsc{Plurality}) \le m^2$.

*Proof.* We will use the alternative definitions of $\kappa$-upper and $\kappa$-lower from Remark 9.3.5, where max is replaced with range. Characterizing these $\kappa$ values, $\kappa\text{-upper}(s^*) = 1$, realized by the vector $\mathbf{u} = \mathbf{1}_1\mathbf{0}_{m-1}$, and $\kappa\text{-lower}(s^*) = 1/(m-1)$, realized by the vector $\mathbf{u} = \mathbf{1}_{m-1}\mathbf{0}_1$. Using that $\beta_{\textsc{Plurality}} = 1/m$ (Lemma 9.3.6), Theorem 9.3.4 gives us that $\mathsf{dist}^{s^*}(\textsc{Plurality}) \le m \cdot \frac{1}{1/(m-1)} \le m^2$. $\square$

**Step 3: Bounding the distortion of the mechanism.** Finally, we bound the distortion in each election within the mechanism. The remaining issue to deal with is that although each *individual voter* will spread their votes across elections in proportion to their stakes (Lemma 9.5.2), if some voters have substantially higher *total stakes across the k elections*, the uniform budgets across voters will under-count these voters' stakes. It may seem that we are back where we started — where uniform voting power fundamentally cannot account for stakes. However, unlike before, we have another lever at our disposal: *the design of our slate of k elections*. Specifically, we can choose this slate of elections such that all voters are affected to a relatively high degree by at least one election. While we cannot hope for this approach to perfectly equalize voters' total stakes, it may bring them *closer*, e.g., within a factor of $\delta$. Then, per Theorem 9.4.1, all elections in our mechanism will have distortion at most $\delta m^2$.

**Theorem 9.5.4.** *Fix a $\delta \ge 1$ such that for all pairs of voters $i, i' \in [n]$, $\|\boldsymbol{s}_i^*\|_2 \,/\, \|\boldsymbol{s}_{i'}^*\|_2 \le \delta$. Then, for all $\ell \in [k]$,*

$$\mathsf{dist}^{s^*}_{U^\ell}(\textsc{Plurality}) \le \delta m^2.$$

*Proof.* Let $\alpha := \max_{i \in [n]} \|\boldsymbol{s}_i^*\|_2$ be the maximum total stakes of any voter. Fix an $\ell \in [k]$. Then, per Lemma 9.5.2, for all $i$, and for some $\delta_i \in [1, \delta]$, we have that $w_i^{*\ell} = s^*(\mathbf{u}_i^\ell) \,/\, \|\boldsymbol{s}_i^*\|_2 = \delta_i \cdot s^*(\mathbf{u}_i^\ell)/\alpha$. Since $\alpha$ is constant across voters, we get $s^*$-proportionality in election $\ell$ with respect to the mis-estimate of $i$'s stakes $\tilde{s}^*(\mathbf{u}^\ell) = \delta_i \cdot s^*(\mathbf{u}^\ell)$. This is the precondition of Theorem 9.4.1; using this and Lemma 9.5.3,

$$\mathsf{dist}^{\delta, s^*}_{U^\ell}(\textsc{Plurality}) \le \delta\, \mathsf{dist}^{s^*}_{U^\ell}(\textsc{Plurality}) \le \delta m^2. \ \square$$

## 9.6 Discussion

We conclude with an in-depth discussion of questions raised by multi-issue mechanisms, followed by opportunities for accounting for stakes in emerging democratic paradigms.

### 9.6.1 The multi-issue mechanism design space & connections to other mechanisms

While we focused on ranking-based voting, the same mechanistic approach we introduced for multi-issue mechanisms could be used with essentially any election format: simply place multiple elections on the same ballot, across which voters must trade off a total allotment of voting power. In fact, there is a known mechanism that roughly does this: *storable votes* [72], which allows voters to save up votes over a sequence of elections and spend them on the later elections they care about most. Like the mechanism above, storable votes is just one within a massive design space, whose many levers we explore below. As we go, we point out a wealth of open questions.

**Slate of issues.** As captured by $\delta$ in our analysis above, the performance of a multi-issue mechanism relies on choosing a slate of issues that *roughly balance* voters' total stakes. In some cases this will not be hard: consider a case where we want to decide a design aspect of our public transit systems, and want to ensure we sufficiently account for the interests of those who cannot afford cars. We might choose our second issue to be where to fix potholes — a decision that will *primarily* affect those who drive. In this case, voters who do/do not have cars will have high stakes in opposite elections, getting closer to balanced total stakes than in either election alone.

However, this gets complicated quickly when we want to consider more than two groups of stakeholders. For example, suppose there is a third group in our example above — people who drive, but also live by candidate transit stops. These people may have high stakes in *both decisions*, due to their potential to avoid pothole damage to their cars *and* their potential to be affected by loud public transportation outside their house. To balance stakes across these voters too, we need a third issue, which may in turn imbalance the stakes of the first two groups.

To state this question of issue design formally: Suppose there is a universe of potential elections $L$ and a set of voters $[n]$, where for each election $\ell \in L$, each voter $i$ has stakes $s(\mathbf{u}_i^\ell)$. For any given $K \subseteq L$, let $\boldsymbol{s}_i^K := (s(\mathbf{u}_i^\ell) | \ell \in K)$. Then, for any given $\delta \geq 1$, we want to know: *what is the smallest slate of elections $K$ that ensures that for all $i, i' \in [n]$, $\|\boldsymbol{s}_i^*\|_2 \,/\, \|\boldsymbol{s}_{i'}^*\|_2 \leq \delta$?* As illustrated by our analysis above, the precise stakes function and method of totaling stakes for which we want to achieve this bound is mechanism-dependent.

**Ballot format.** In our paper, we assumed voters submit their preferences as complete rankings. However, there are many other ballot formats that could reveal richer information about voters' utilities. One promising candidate is the ballots used in Quadratic Voting (QV) [230], where a voter has a budget of total voting power to allocate *over the alternatives in a single election*, and they are charged quadratically per unit of weight they place on a given alternative. For example, in an election over alternatives $a, b, c$ where each voter has 10 votes total, they could place 3 on alternative $a$, 1 on $b$, and 0 on $c$ to spend their total budget of $3^2 + 1^2 + 0^2 = 10$.

Notice that this is different from how we use quadraticity in our mechanism: we charge voters quadratically for power allocated toward a given *election*, while QV considers a single election and charges voters for power allocated toward a single *alternative*. However, the quadraticity works for the same reason, encouraging voters to spread their power over elections or alternatives *in proportion* to their stakes or relative utilities, respectively. In summary, QV ballots — at least under the (strong) assumptions made by Posner and Weyl [230] — allow the recovery of information about voters' *relative utilities* (i.e., $i$ likes $a$ 3 times as much as $b$), while rankings lose this information, thereby losing $m$ distortion even when stakes are known. To design a multi-issue mechansim using QV ballots, then, one must answer: *how can we set up voting budgets to recover stakes-proportionality when voters are charged quadratically per unit of votes for a single alternative* and *weight toward a single election?*

Designing new ballots will open up the possible space of voting rules — an enticing prospect, given that with ranking-based ballots, our results show that the space of rules where stakes information is helpful is very limited.

**Voter model.** Ultimately, the success of a multi-issue mechanism depends on how voters will *behave* within it. This depends on several things, including (1) voters' beliefs about their chances of pivotality across elections — whether they believe some elections are close whereas they have no chance of swaying others, (2) how voters may strategize in misreporting their preferences, and most importantly, (3) how voters' preference intensities relate to both their *ability* and their *desire* to vote in a given election. A more holistic mechanism design approach here would build on the wealth of empirical evidence on these topics [99, 115].

#### Other kinds of trade-offs.

In a multi-issue mechanism, voters' stakes are revealed in how they make trade-offs between voting power in election $\ell$ with voting power in a different election. However, one could also imagine designing a mechanism in which voters must trade off voting power in election $\ell$ with *some other resource*. One natural option, from a mechanism design perspective, would be currency (or something with similar external economic value). This is precisely the proposal of quadratic voting, which in its theoretical conception requires voters to purchase votes with *money*, thereby revealing voters' stakes in how many votes they buy [230]. However, we emphasize that such monetary mechanisms may not recover stakes at all — to do so, they require voters convert their value for voting power to their value for money *at the same rate*, when in reality, less wealthy voters may have higher marginal value for money, in which case they will purchase fewer votes and their stakes will be underestimated.

### 9.6.2 Stakes in liquid and deliberative democracy

Our model of *stakes functions* applies in any voting model with latent utilities, and while we only formally define the reweighting of *ranking-based* ballots, our model can easily be extended to study the impact of stakes-dependent reweighting under other ballot formats or decision mechanisms. One extension where studying stakes may especially be of interest is *liquid democracy*

[154], where each voter receives one unit of voting power but may delegate it to others. Conceivably, voters might naturally delegate their votes to others in ways that at least partially depend on stakes; in the extreme, one could imagine a pro-social society in which voters delegate their votes to those with the highest stakes, out of a belief that they deserve the most influence. In this case, we might not need to reweight votes explicitly, motivating an interesting question: *if voters delegate their votes in a way that accounts for others' stakes, does the outcome have better distortion?*

One can also conceive of accounting for stakes in ways that go beyond asking who receives the most *voting* power. For example, there is a growing body of work in computer science studying *sortition*, the random process of selecting a "representative" body of citizens to deliberate and make a collective policy recommendation on a given issue [130]. Although in practice "representation" is typically designed to be proportional to population composition, one could instead implement progressive representation, where representation targets for population groups are at least in part determined by their stakes. In fact, *this is already being tried*: in a deliberative poll in Australia on how to facilitate reconciliation between Indigenous and non-Indigenous groups, Indigenous people — a very small fraction of the overall population, but affected by this decision to an outsize degree — were intentionally over-represented in some deliberation groups [176]. More generally, on the topic of representation in the deliberative context, the idea of accounting for the interests of highly affected groups in how representation is decided is already being discussed and advocated [176].

# 10

# Ongoing and Future Work

## 10.1 A Stakes-Based Analysis of Quadratic Voting Ballots

Quadratic voting (QV) [107] has been increasing in popularity over the past five years, being used in polling and decision-making around the world.[1] Its popularity is due at least in part to its apparent ability to circumvent impossibilities faced by most other popular voting mechanisms: QV is marketed as an incentive-compatible voting system which achieves optimal utilitarian social welfare with high probability as the number of voters becomes large.

At a high level, QV as proposed works as follows: each voter can purchase unlimited votes using some currency. For reasons we will discuss, the key feature of this currency is that it has real economic value outside the election. Each vote purchased can be *for* or *against* any alternative. The core idea (and namesake) of QV is that a voter pays *quadratically* in the number of votes they purchase for or against a given alternative; that is, to cast 3 votes for alternative $a$ and 3 votes against $b$, a voter pays $3^2 + 3^2 = 18$ units of currency; to cast 6 votes for $a$ alone, a voter pays $6^2 = 36$ units. After votes are cast, the total votes (positive and negative) for each alternative are summed to yield a vote total, at which point a voting rule is applied to choose a winner. Two different voting rules have been studied in the QV literature: earlier work, which permitted elections over only two alternatives, used deterministic majority voting to aggregate votes [184]. In more recent work, Eguia *et al* present a QV procedure that applies to $m$ alternatives, which uses a new, randomized aggregation mechanism in which an alternative wins with probability exponential in the vote total [107]. We will focus here on the latter QV mechanism.

[1] For an overview of instances in which QV has been applied, see the the Quadratic Voting Wikipedia page: https://en.wikipedia.org/wiki/Quadratic_voting.

The feature of QV essential to its optimal social welfare and incentive compatibility is its quadratic cost function. Conceptually, this quadraticity means that the *derivative* of the overall utility function voters are optimizing — describing their marginal benefit from purchasing an additional vote for a given alternative — is changing linearly in the cost of that vote. Thus, a voter's utility-maximizing strategy is to purchase a number of votes for each alternative that is approximately "proportional", in a loose sense, to their utility. This amounts to the mechanism getting almost complete information about voters' utilities, as revealed through the number of votes they purchase for each alternative, thereby permitting the mechanism to achieve optimal social welfare.[1]

### 10.1.1 *Stakes* Information and *Relative Utilities* Information

First, we observe that for any stakes function $s$, a voter's utility information $\mathbf{u}$ can be decomposed into two types of information: their *stakes* $s(\mathbf{u})$—how much they care about the outcome overall—and their *relative utilities* $\mathbf{u}/s(\mathbf{u})$ — i.e., how strongly they prefer each alternative *relative to each other alternative*. If $s$ is 1-homogeneous, we can perfectly recover $i$'s utility vector from these two types of information.

We now conjecture connections between these quantities to the behavior of QV. For the sake of intuition, throughout this section we will use phrase "proportional to" loosely. Technically, it reflects a linear transformation of the utilities by an invertible matrix, rather than just multiplication by a scalar as "proportional to" would typically suggest.

**Conjectures:**

- **Stakes.** When voters vote via QV ballots, stakes information is encoded in the total number of votes each voter purchases. That is, voter $i$ purchases a total number of votes proportional to their stakes.
- **Relative utilities.** When voters vote via QV ballots, relative utility information is encoded in how a voter spreads these total votes over alternatives. That is, the number of votes $i$ casts for $a$ is proportional to their utility for $a$.

### 10.1.2 What information do we lose when QV assumptions do not hold?

We now study: which types of information are *lost* — and how is the social welfare consequently impacted — under two practically-motivated ways in which QV may deviate from the assumptions made in Eguia *et al*'s theoretical model?

***Assumption Deviation 1:*** *What if voters don't the have the same marginal utility per unit of currency?*

Intuitively, the cost of voting causes people's stakes to be revealed through their behavior: because voting is costly, people who care less about the decision will purchase fewer votes, thereby

[1] Suppose, in contrast, the costs were linear in the number of votes: then the derivative of the objective with respect to each alternative would be constant, and the voter would allocate all their votes to their favorite alternative, giving the mechanism much less rich utility information.

leading to something that looks like stakes-based representation among the votes cast. However, that QV elicits voters' stakes in this way relies on the assumption that all voters have the *same marginal value* for a unit of the currency. Because any externally-valuable currency is essentially money,[1] this assumption seems unlikely to hold in practice: people with greater wealth will have lower marginal value for currency, will purchase more votes per additional unit of stakes, and the result will no longer be stakes-proportional.

We therefore instead suppose that voter $i$ has idiosyncratic marginal utility per unit of currency $c_i$, and propose the following conjectures:

**Conjectures:**

- **Stakes information is lost.** Voter $i$ will purchase a total number of votes (approximately) proportional to their stakes *divided by* $c_i$, where $c_i$ can be arbitrary and is unknown to the mechanism.
- **Relative utility information is retained.** Voters will still spread the total votes they purchase over the alternatives proportionally to their relative utilities.
- **Distortion is unbounded.** Without stakes information, the distortion can be unbounded due to negative examples similar to those in Chapter 9.

***Assumption Deviation 2:*** *Uniform vote budgets.*

Perhaps because of the practical issue discussed in Assumption Deviation 1, the implementation of QV in practice deviates from the use of externally valuable currency, as studied in theory. Instead of paying for potentially unlimited votes with externally valuable currency, in practice voters may be given the same fixed budget of votes. While this a priori appears to solve the unequal marginal values problem, it is still unclear whether this mechanism elicits stakes — after all, it stands to reason that all voters, regardless of how much they care, will cast all their votes.

Supposing that all voters receive a divisible vote budget of $B$, we propose the following conjectures:

**Conjectures:**

- **Stakes information is lost.** Voter $i$ will purchase a total of $B$ votes, regardless of their stakes.
- **Relative utility information is retained.** Voters will still spread the total votes they purchase over the alternatives proportionally to their relative utilities.
- **Distortion is unbounded.** Without stakes information, the distortion can be unbounded due to negative examples similar to those in Chapter 9.

[1] If a currency that is not strictly monetary has societal value, it can (and will likely) be monetized in a market.

### 10.1.3 Proposed Solution: QV within A Multi-Issue Mechanism

Among the vast space of possible multi-issue mechanisms, one instantiation that seems especially promising for getting constant distortion is one which uses QV ballots to elicit preferences within each election. We hypothesize that the following intuition would hold: due to the multi-issue nature of the mechanism, voters will spread their votes across elections according to some reasonable stakes function, as was the case in Chapter 9. Thus, at the stage where voters decide how many votes to cast in each election, **stakes information is recovered**—at least to the degree that our slate of issues balances voters' total stakes. Then, due to the quadraticity of the ballots, voters will spread the votes they have allocated to any given election proportionally to their relative utilities over alternatives in that election. Thus, at the stage where voters cast their votes in each election, **relative utilities information is recovered.**

## 10.2 Implementing Stakes-Based Representation in *Deliberative Town Halls*

For the duration of this section, for consistency with the literature we will often use the term "affected interests" interchangeably with *stakes*, referencing Archon Fung's Theory of Affected Interests [140].

There are two main reasons why deliberative town halls (DTHs) are an excellent participatory model in which to experiment with bringing stakes-based representation into practice.

- *In the status quo, representativeness of any kind is not always ensured.* Unlike citizens' assemblies where it is standard to enforce descriptive representation, DTHs are often open to anyone in the community.[1] This has the benefit of broadening participation, but a major downside is that community members with more spare time and resources tend to be over-represented with respect to attendance and speaking time. This status quo offers a clear opportunity to improve representation by moving toward representation of people based on the unique ways in which they are *impacted* by the issue at hand.
- *DTHs are currently scaling up in an AI-enabled way.* Researchers at UC Riverside, led by Kevin Esterling, are in the process of building and testing *Prytaneum*, an online platform that streamlines the process of running DTHs [15]. The existence of an online platform naturally provides infrastructure for deploying surveys, advertising to potential participants, balancing and guiding the conversation, and tracking which affected interests are being discussed.

The task of implementing stakes-based representation in DTHs presents many experimental design choices and research questions. For example, "representation" can mean multiple things, including who is actually in the room *and* who is given speaking time; we have the potential to influence both. If we are to *recruit participants* based on affected interests, the question is how to recruit participants in order to reach hard-to-reach groups who may be very affected, but less predisposed to participating. If we are to moderate *speaking time* based on affected interests, we

[1] Especially in larger use-cases, DTH participants are sometimes recruited to be approximately descriptively representative [178].

may need to structure the conversation more intentionally around how participants are affected—a change which may itself have positive impacts on outcomes. On this note, another lever we can pull is what *impacts to measure*: does improving stakes-based representation materially change... the nature of the discussion? ...participants' or elected officials' perceptions of the usefulness or legitimacy of the event? ...what participants' or elected officials learn about others in their community?

In order to pursue any of these lines of research, we first need a clear and practically-applicable definition of *who exactly qualifies as having an affected interest.* In this future work section, we approach this task by decomposing "affected interests" into *types* of affected interests that may exist for any given political issue. We propose them only as a starting point, and do not claim they are exhaustive. To ensure our definitions are concrete, we will apply them to a running example as we go. This example — a decision about how to allocate funding to support mental health services in k-12 schools — is based on an actual DTH that will be run in June 2024 using *Prytaneum.*

**Defining types of affected interests in practical cases**

We begin from the premise that the minimum requirement for an individual to have an "affected interest" in a policy decision is that it is possible to articulate what that interest or stake *is*, i.e., how the individual is affected. In other words, one should be able to provide a *reason* for why others should see the individual as having an interest that is affected. Here are some possible reasons someone could be affected by a policy decision.

People with **primary affected interests** in a decision are those whose basic needs being met (healthcare, shelter, food, physical and mental safety...) are impacted by that decision.

> *Example*: K-12 students would be an example of a population that has primary affected interests, as they are the people being served by mental health services.
>
> Note that the *degree* of primary affected interests across K-12 students can vary significantly depending on things like precarity  does the student have access to mental health services outside of school, or do they rely on school-provided mental health services?

People with **secondary affected interests** in a decision are those who are impacted because they care about the basic needs of someone who has primary affected interests.

> *Example:* Parents of K-12 students would be an example of people with secondary affected interests, because they care deeply about the well-being of their children. They are affected because they meaningfully suffer when they see their children suffer.
>
> Note that parents could additionally have primary affected interests, because without access to mental health services at their childrens school, they may have to sacrifice means of meeting basic needs to get their children mental health care elsewhere.

People with **professional affected interests** in a decision are those whose professional responsibility it is to attend to the basic needs of people with primary affected interests.

> *Example 1:* Pediatricians are professionally responsible for ensuring that their patients are having their basic needs met.
>
> *Example 2:* A school mental health provider will be professionally responsible for enacting the policy to the best of their ability.
>
> Note that pediatricians or school mental health providers may also have secondary affected interests, as they may care deeply about the well-being of their patients.

People with **group-based affected interests** in a decision are those who have a sense of *linked fate* with a group within whom some members may be primarily affected.

> *Example 1:* Someone of a certain demographic group may care about the impact of the decision on other members of their own demographic group.
>
> *Example 2:* A teacher might identify as a union member, and the decision might have an impact on the community of unionized teachers.

People with **abstract, ideological, or conceptual affected interests** in a decision are those who have a categorical or principle-based concern with the decision, even if it does not impact them personally.

> *Example 1:* Someone who does not (and will never) have kids in k-12 school might object on principle to the idea of increasing in-school services to support students in the queer community.

# Part III

# Deliberation, *Public Spirit*, and the Quality of Democratic Outcomes

# 11

# Background

In an important sense, deliberative minipublics are built around *deliberation*: a discussion between constituents, usually informed by expert-derived information, about a political decision at hand. In a deliberative minipublic, deliberation is theorized to serve a key purpose: to help participants learn about and carefully weigh competing, evidence-based arguments *before* expressing their opinions about what should be done.

Now, we interrogate a key anecdotal claim underlying the study and practice of deliberative democracy: *that deliberation leads to better democratic outcomes.* We posit and investigate one particular hypothesis as to why: that deliberation improves democratic outcomes *by shifting participants' opinions toward outcomes that more strongly prioritize the good of their society, even if it means deprioritizing their own interests.* In fact, many practitioners and political scientists had already identified this "social consciousness" as a key condition cultivated by deliberation; we call it *public spirit* after [276], but similar concepts go by many names (e.g., sociotropism, collectivism).

In this line of work, we consider two main questions about public spirit: *I. Does deliberation truly cultivate public spirit?* and *II. If people are more public-spirited, are democratic outcomes guaranteed to be "better"?* These questions target implications I and II in the diagram below, which, if demonstrated, would together support the overall claim that deliberation, by cultivating public spirit, improves democratic outcomes.

$$\text{Deliberation} \xrightarrow{\textit{Implication I}} \text{Public Spirit} \xrightarrow{\textit{Implication II}} \text{Better Democratic Outcomes}$$

The completed research in Part III, covered in Chapters 12 and 13, addresses Implication II. This implication is most naturally studied in a theoretical model, where we can assume a practically-

unobservable ground truth about outcomes' relative social benefits, and then test whether greater public spirit reliably leads to more socially-beneficial outcomes. In Chapter 12, we capture public-spirited voting behavior via a natural generalization of the standard voting model with latent utilities. Recall from Part II that in the standard version of this model, voters rank alternatives according to their utilities. In our generalization, we permit voters to weigh each alternative's *utilitarian social welfare* to an idiosyncratic degree in addition to their own utility when they rank alternatives. We call the relative weight a voter places on the utilitarian social welfare their *level of public spirit.*

In this model, in Chapter 12 we show that in most senses, increasing voters' public spirit levels *does* dramatically improves the welfare of voting outcomes. Notably, we find that public spirit permits constant welfare loss in voting *without restricting (or assuming the election designer has knowledge of) voters' latent utilities*—a first, to our knowledge.

> **Chapter 12: *Public Spirit:* Voting Beyond Self Interest** [134] In this chapter, we show that the issue of unbounded distortion, as discussed in Part II, is mitigated by voters being *public-spirited*: that is, when deciding how to rank alternatives, voters weigh the common good in addition to their own interests. We first generalize the standard voting model to capture this public-spirited voting behavior. In this model, we show that public-spirited voting can substantially — and in some senses, monotonically — reduce the distortion of several voting rules. Notably, these results include the finding that if voters are *at all* public-spirited, some voting rules have *constant distortion* in the number of alternatives — a highly sought-after result in the social choice literature. Further, we demonstrate that these benefits are robust to adversarial conditions likely to exist in practice. Taken together, our results suggest an implementable approach to improving the welfare outcomes of elections: *democratic deliberation*, an already-mainstream practice that is believed to increase voters' public spirit.

In Chapter 13, we then extend our model of public-spirited voting to the more general setting of participatory budgeting (PB). We show that again, public spirit permits fundamental improvements in welfare loss without structural restrictions on or knowledge of voters' utilities — and reveals a new voting rule that may make better use of public spirit in the PB setting.

> **Chapter 13: Extensions to Participatory Budgeting [43].** The paper begins by closing a question about standard voting left open in Chapter 12: what is the best possible distortion achievable by *any* ranking-based deterministic voting rule when voters are public-spirited? We answer this question, giving a lower bound that matches the upper bound from Chapter 12 on the voting rule Copeland when $m$ is large, and that for the voting rule Plurality when $m$ is small. For the the first time, we then extend these results to study what is possible with *randomized* voting rules. We characterize the optimal distortion, giving a lower bound and providing an optimal randomized rule to match.
>
> Moving onto the strictly more general setting of PB, we study the public-spirited distortion of various common PB ballot formats such as rankings by value, rankings by value for money,

$k$-approvals, knapsack votes, and threshold approval votes. We prove that multiple of these ballot formats achieve distortion linear in $m$, but unfortunately, none of these ballot formats can break that linear barrier. We then design a novel and practical PB ballot format which, we prove, achieves sublinear distortion in $m$ (and even logarithmic, if voting over two rounds is possible).

**Chapter 14: Ongoing and Future Work.**

The theory above suggests that public spirit can lead to far better democratic outcomes. Then, the question is: *does* deliberation cultivate public spirit, and what does it look like? This question targets *Implication I*, which we examine in our first stream of ongoing work. This implication is best studied experimentally, as we are trying to understand what public spirit looks like in real deliberative contexts. In designing these experiments, our theoretical public spirit model is valuable, because it decomposes "public spirited" behavior into three outcome-relevant components that can be measured experimentally: the extent to which deliberants prioritize societal benefit over their own; with what information deliberants *evaluate* alternatives' "social benefit"; and what notion of "social benefit" deliberants actually care about (e.g., some may be *utilitarian*, caring about total prosperity, while others may be *egalitarian*, caring about minimizing inequality).

In this chapter, we also push the theoretical frontier of this line of work, discussing multiple ways to go beyond one very strong assumption made by our public-spirited model: that at the time of deliberation, there is *one ground-truth policy* that is best for society.

# 12

# *Public Spirit:* Voting Beyond Self Interest

*Distortion Under Public-Spirited Voting* [134]
Bailey Flanigan, Ariel D. Procaccia, & Sven Wang
*EC* 2023

## 12.1 Introduction

Consider an election with two alternatives, $a$ and $b$; of the 100 voters, 50 prefer $a$ to $b$ and 50 prefer $b$ to $a$. Since the preference profile is symmetric, let us assume that $a$ is elected. Although their rankings are symmetric, voters may have highly asymmetric underlying *intensities* of preferences, perhaps capturing that they are affected to differing degrees by the outcome of the election. We capture these preference intensities with *utilities*, which can be interpreted as measuring the value a voter gains from a given alternative. In this case, suppose the supporters of $a$ are affected similarly by the alternatives, having utility 1 for $a$ and 0 for $b$, whereas the admirers of $b$ are, by comparison, affected much more disparately by the alternatives, having utilities 0 for $a$ and 100 for $b$.

From a societal benefit standpoint, $b$ would have been the better choice, as it would yield substantially more utility to voters overall. This intuition is captured by the *utilitarian social welfare*, defined as the sum of voters' utilities for a given alternative: $a$ (the winner) is severely suboptimal in terms of this measure, its social welfare being 100 times lower than that of $b$ (the alternative with optimal social welfare). This ratio can be made arbitrarily large by, say, making the supporters of $a$ arbitrarily unaffected by the decision.

The simple example above implies an alarming conclusion: that *any* deterministic rankings-based voting procedure will, in some instances, choose an alternative that yields arbitrarily suboptimal value for the population. Moreover, while this is just a theoretical example, what makes it patho-

logical — that people can be affected to differing degrees by a given decision — is almost surely a property of real elections, suggesting that such welfare loss *could* occur in practice. From a technical standpoint, this welfare loss arises due to information lost between cardinal utilities and ordinal preferences; this was first observed by [231], who quantified this loss with the notion of *distortion*. Assuming voters report rankings that are consistent with their underlying utilities, the distortion of a voting rule is the worst-case (over latent utilities) ratio between the utilitarian social welfare of the highest-welfare alternative and that of the elected alternative. By the example above, then, *all deterministic voting rules* must have unbounded distortion.

A natural question, then, is: under what assumptions *is* the distortion bounded? The rich literature on distortion — overviewed in an excellent recent survey by [25] — has largely taken one of two approaches to achieving bounded distortion. One line of research, originating from the work of [231], assumes that each voter's utilities sum to 1, thereby eliminating the possibility of voters being affected by widely differing degrees by the decision. Another line of research, originating from the work of [24], assumes that voters' preferences are induced by distances in an underlying metric space.

Both of these lines of work rely on assumptions that restrict voters' possible latent utilities (or analogously in some models, costs). However, it is not clear whether we can rely on such assumptions to hold in practice. This is perhaps most directly illustrated by the fact that the core problem in our example above cannot be ruled out as a potential feature of real-world elections: the utilities are such that there is a minority group that is much more affected by the issue than a majority group with decisive voting power. Moreover, it seems unlikely that we can *promote* such conditions on the utilities, because voters' utilities — how much they fundamentally gain from a given election outcome — would likely arise from features that are difficult to change with simple interventions.

In this paper, we take a different approach to attaining bounded distortion. This approach begins from the realization that while underlying utilities like those in our example might unfortunately be realistic, the *behavioral model by which voters translate utilities into rankings* might be too pessimistic. The standard behavioral assumption made in the literature is that voters rank alternatives according to only the order of their own utilities. However, as many social scientists have observed, this model is unrealistic in a way that can potentially help us: voters can be *public-spirited* — that is, when they vote, they weigh not only how they themselves are impacted by each alternative, but also each alternative impacts their society as a whole.[1] This behavior of balancing self and societal interest can be captured in a natural generalization of the standard behavioral model of voters: instead of ranking alternatives according to only their own utilities, a $\gamma$-*public spirited* voter ranks alternatives according to values that place weight $1 - \gamma$ on their own utilities, and weight $\gamma$ on each alternative's utilitiarian social welfare. It is then intuitive why public spirited voting could help decrease the distortion: it will cause voters to more highly rank higher-welfare alternatives, potentially increasing the social welfare of the election winner.

[1] Public-spirited behavior among voters has been demonstrated empirically [177, 290] and has long featured in economic theories of how people make decisions [42, 173].

While existing work suggests that voters are willing and able to be public-spirited, we need not assume that these conditions are satisfied by default; instead, we can intentionally cultivate them within the democratic process. One promising innovation on this front that is currently gaining momentum globally[1] is *democratic deliberation*, summarized by [204] as dialogues in which "people rely on reasons that speak to the needs or principles of everyone affected by the matter at hand." This description of deliberation already alludes to some of its key potential benefits, which roughly correspond to promoting our conditions. For instance, deliberation is theorized to lead to "citizens [being] more enlightened about their own and others' needs and experiences" [204] — akin to promoting more accurate estimates of alternatives' welfares, and to "an increased willingness to recognize community values and to compromise in the interest of the common good" [175] — akin to promoting voters' levels of public spirit. These theorized benefits are supported by empirical evidence showing, for example, that deliberation can increase public-spiritedness [276], lead to more egalitarian values [144], and increase empathy for members of social outgroups [158].

This evidence suggests that public-spirited voting behavior can be cultivated (or may already exist) among voters. This motivates our research question, which, if answered affirmatively, would lead to an actionable approach to decreasing deterministic voting rules' otherwise unbounded distortion:

> *To what extent is public-spirited voting guaranteed to decrease the distortion, and for which voting rules?*[2]

We aim to formally answer this question with the tools of social choice theory, as outlined in the results and contributions below. In our analysis, we focus on deterministic voting rules, owing to the several political hurdles to implementing randomized rules. We leave the study of randomized rules in our model to future work.

### 12.1.1 Results and Contributions

Throughout the rest of the paper, we will often use *PS* to refer to the concept of public spirit.

**Section 12.2: A model of public-spirited voters.** A precursor to answering the question above is formally modeling public-spirited voting behavior. Our model is a simple generalization of the standard model: voter $i$ has public spirit level $\gamma_i \in [0, 1]$, where higher $\gamma_i$ corresponds to more public spirit. Then, voter $i$ ranks alternatives in order of their *PS-value* for each alternative $a$,

[1] Democratic deliberation is commonly implemented through *deliberative polls* or *citizens' assemblies*, of which hundreds have been run in the past few years [225]. Such processes have played a key role in major political decisions: for example, citizens' assemblies commissioned by Ireland's national legislature recently led to amending the Irish constitution on the issues of same-sex marriage and abortion [169].

[2] A natural question here is, if *constituents* can learn alternatives' social welfares via, e.g., deliberation, why can't the *election designer* learn these values and directly select the highest-welfare alternative? One reason is that the election designer imposing such "complete" public spirit could be perceived as undemocratic and illegitimate. Underlying this point is the premise that in a democracy, it is voters' prerogative to decide *how strongly* to account for the social good, an interpretation which views deliberation as a process of *clarifying for voters* how much public spirit their values dictate they should have.

called $v_i(a, \boldsymbol{\gamma}, U)$. This value is a convex combination of their utility $u_i(a)$ and $a$'s social welfare $\text{sw}(a, U)$ — the sum of all voters' utilities for $a$, summarized in the utility matrix $U$:

$$v_i(a, \boldsymbol{\gamma}, U) = (1 - \gamma_i)u_i(a) + \gamma_i \cdot \text{sw}(a, U)/n.$$

The standard behavioral model is then the special case of our model where $\gamma_i = 0$ for all $i$.

**Section 12.3: Distortion bounds for voting rules.** We begin by proving our key lemma, which upper bounds the extent to which the social welfare of an alternative $a$ can exceed that of another alternative $b$ — a bound which is decreasing in the fraction of voters who rank $b$ ahead of $a$ in the election, along with the minimum level of public spirit among voters, $\gamma_{min} := \min_i \gamma_i$. We then use this result, plus other techniques, to give tight bounds on the distortion of several popular voting rules. For consistency with the distortion literature, we consider these bounds asymptotic in $m$, the number of alternatives in the election. The main takeaway from these bounds is that when voters have *any* public spirit (i.e., if $\gamma_{min} > 0$), several voting rules' distortion drops from unbounded to linear (for the rules Borda, Plurality, Maximin) or even *constant* (for the rules Copeland and Slater). We emphasize that our bounds asymptotically — and for some settings of $\gamma_{min}$, non-asymptotically — either match or beat those possible in both aforementioned models, and moreover do so without any assumptions on voters' underlying utilities.

**Section 12.4: PS-Monotonicity.** The upper and lower bounds we give in Section 12.3 are decreasing in $\gamma_{min}$, hinting at a weak form of *PS-monotonicity* — i.e., that the distortion decreases as voters' public spirit increases. Although it seems intuitive that this property should hold, we show that, while some notions of PS-monotonicity are guaranteed, other natural notions do not hold. Working from weaker to stronger notions, we show first that if public spirit increases *uniformly* among voters, then the worst-case distortion of *all voting rules* decreases monotonically. Given that in reality voters' $\gamma_i$ levels are unlikely to be uniform, we then show that for Copeland and Plurality, the worst-case distortion decreases even if voters' public spirit is increased heterogeneously. This implies that cultivating greater public spirit among any voters to any extent is guaranteed to decrease the worst-case distortion over possible utility profiles — already a useful guarantee, since we cannot observe voters' initial levels of public spirit. Given that utilities are also unobservable, one might hope that PS-monotonicity holds for all fixed utility matrices *and* initial levels of public spirit. We soundly resolve this question by showing this is too much to hope for: applying classic axiomatic impossibilities by Muller and Satterthwaite, we prove that *no weakly unanimous, non-dictatorial voting rule* exhibits PS-monotonicity on an instance-by-instance basis.

**Section 12.5: Robustness of distortion bounds.** There are two key weaknesses, from a practical perspective, of our upper bounds in the Section 12.3. First, they are vacuous if $\gamma_{min} = 0$, and second, they *a priori* rely on voters using accurate and internally-consistent inputs to our model of PS-values, $\gamma_i$, $u_i(a)$, and $\text{sw}(a, U)$. We provide robustness results that address both of these gaps. First, we show that our upper bounds degrade by only a constant factor if up to some fraction of voters has $\gamma_i = 0$; for Copeland this fraction is quite large — up to 1/2 of voters. Second, we generalize our model to allow voters to deviate arbitrarily from correct and/or internally-consistent

values of *any model input* $\gamma_i$, $u_i(a)$, and $\text{sw}(a, U)$. We then extend our distortion upper bounds to this generalized model, showing that our original bounds are robust to *all such deviations*: that is, our upper bounds degrade smoothly, by constant factors, in the magnitude of these deviations.

### 12.1.2 Related Work

**Distortion under existing models.** As discussed in the introduction, the main body of work achieving bounded distortion does so by assuming regularity conditions on voters' utilities. Under the assumption that voters' preferences can be embedded in a metric space, the well-known rule Copeland has distortion of 5 and there are deterministic voting rules that achieve the best possible distortion of 3 [147, 180]. Under the assumption that each voter's utilities sum to 1, all deterministic rules have distortion at least $\Omega(m^2)$, where $m$ is the number of alternatives; the popular rule Plurality achieves a matching upper bound [67]. More distantly, there is some work that achieves bounded distortion by assuming additional access to some cardinal information about voters' utilities (see Sec. 5 of [25] for an overview). In contrast to these lines of existing work, our distortion bounds require neither regularity conditions on voters' utilities, nor any information from voters beyond their rankings. Nonetheless, we can match or improve upon the metric model's upper bound of 5 on Copeland's distortion when $\gamma_{min} \geq (\sqrt{5} - 1)/2 \approx 0.61$, and we can show that Plurality, along with several other deterministic rules, have linear or sublinear distortion, improving upon the distortion achievable in the unit sum model by at least a factor of $m$ (Table 13.2).

**Related behavioral models.** Our model of public-spirited voting is a direct analog of a model used in the study of congestion games by [79], who in turn attribute the idea to Ledyard [186, p. 154]. Additionally, similar ideas appear in literature exploring altruistic behavior by agents in decision-making systems: for instance, [189] model agents as giving some linear weight $\alpha > 0$ to the interests of another entity as a form of altruism. We remark, however, that *altruism* in this work is distinct from public spirit, because it may involve accounting for only the interests of population subgroups or specific agents for strategic reasons, rather than arising from the motive of benefiting society at large. Other related models include Fehr and Schmidt's model of how economic agents incorporate inequality into their utilities [116], Austen-Smith and Feddersen's model of how voters may be inequality-averse [28], and political economy models of *sociotropic* voters, who weigh the economic interests of their country over their own [179]. [39] even aim to estimate from data *how* sociotropic voters are, corresponding to estimating $\gamma_i$ parameters in our model.

## 12.2 Model

### 12.2.1 Public-Spirited Voting Behavior

There are $n$ voters and $m$ alternatives. We refer to the set of voters as $[n]$ and alternatives as $[m]$. By default, individual voters and alternatives are denoted $i \in [n]$ and individual alternatives are denoted $a \in [m]$.

**Public spirit (PS)** We represent voters' levels of public spirit with the *PS-vector* $\boldsymbol{\gamma} \in [0,1]^n$, whose $i$-th entry $\gamma_i$ is voter $i$'s level of public spirit (higher $\gamma_i$ means more public spirit). Our upper bounds will be in terms of the minimum level of public spirit possessed by any voter, $\gamma_{min} := \min_{i\in[n]} \gamma_i$. We will also sometimes restrict our consideration to *uniform* PS-vectors $\boldsymbol{\gamma} = \gamma\mathbf{1}$, in which all voters have the same public spirit level $\gamma \in [0,1]$.

**Utilities.** We define a *utility matrix* $U \in [0,1]^{n\times m}$ such that its $(i,a)$-th entry is $i$'s utility for $a$, called $u_i(a)$. Let $\mathrm{sw}(a,U)$ denote the utilitarian *social welfare* of $a$ based on $U$, i.e.,

$$\mathrm{sw}(a,U) := \sum_{i\in[n]} u_i(a).$$

When $U$ is clear, we may denote the highest-welfare alternative in $U$ as $a^* := \arg\max_{a\in[m]} \mathrm{sw}(a,U)$.

**PS-values.** Together, a pair $\boldsymbol{\gamma}$, $U$ imply a *PS-values matrix* $V(\boldsymbol{\gamma},U)$, containing the values for alternatives by which voters decide how to vote. A voter $i$'s *PS-value* for $a$ weighs their own utility $u_i(a)$ to a $(1-\gamma_i)$ extent, and $a$'s social welfare $\mathrm{sw}(a,U)$ to a $\gamma_i$ extent:

$$v_i(a,\boldsymbol{\gamma},U) = (1-\gamma_i)u_i(a) + \gamma_i \mathrm{sw}(a,U)/n. \tag{12.1}$$

Note that $\mathrm{sw}(a,U)/n$ is interpreted as voters' *average utility* for $a$. Per this equation, the mathematical interpretation of a voter's public spirit level is the weight they place on the average utility versus their own in this convex combination.

**Rankings.** A *ranking* $\pi$ is a permutation of $[m]$. Voter $i$ expresses their preferences over alternatives as a strict, complete ranking $\pi_i$. We denote that $i$ ranks $a$ ahead of $b$ by $a \succ_{\pi_i} b$. We say that $\pi_i(j)$ is the alternative that voter $i$ ranks in the $j$-th position.

**Preference profiles.** A *preference profile* $\boldsymbol{\pi}$ is the $n$-tuple of all $n$ voters' rankings: $\boldsymbol{\pi} := (\pi_i : i \in [n])$. We let $\Pi$ be the set of all preference profiles. To compare how two alternatives' relative positions compare within a profile $\boldsymbol{\pi}$, we denote the number of voters in $\boldsymbol{\pi}$ who prefer $a$ to $b$ as $|\{i : a \succ_{\pi_i} b\}|$. A *pairwise election* between $a$ and $b$ in $\boldsymbol{\pi}$ compares $|\{i : a \succ_{\pi_i} b\}|$ and $|\{i : b \succ_{\pi_i} a\}|$; we say that *$a$ pairwise-dominates $b$* if $|\{i : a \succ_{\pi_i} b\}| > n/2$, and we add *weakly* if the inequality is weak. We say that $a$ is a *Condorcet winner* in $\boldsymbol{\pi}$ if $a$ pairwise-dominates all $b \neq a$ (note: not all profiles have a Condorcet winner).

**Translating instances to preference profiles.** In any instance $(\boldsymbol{\gamma},U)$, its associated PS-values matrix $V(\boldsymbol{\gamma},U)$ naturally implies a preference profile in which alternatives are ordered in decreasing order of PS-value; formally, for any voter $i$,

$$v_i(a,\boldsymbol{\gamma},U) > v_i(b,\boldsymbol{\gamma},U) \implies a \succ_{\pi_i} b. \tag{12.2}$$

We do not specify the ranking implied when $v_i(a,\boldsymbol{\gamma},U) = v_i(b,\boldsymbol{\gamma},U)$; rather, we allow there to be *multiple* profiles consistent with the same $V(\boldsymbol{\gamma},U)$. We let $\Pi_{V(\boldsymbol{\gamma},U)}$ be the set of all profiles consistent with $V(\boldsymbol{\gamma},U)$.

### 12.2.2 Voting Rules

A preference profile maps to a winning alternative via a (resolute) *voting rule* $f : \Pi \to [m]$. Then, $f(\boldsymbol{\pi}) = a$ means that on profile $\boldsymbol{\pi}$, rule $f$ chooses $a$ as the winner. We study two main classes of voting rules, *uncovered set rules* and *positional scoring rules*, defined below.[1] All of our examples will be strict, so we need not specify tie-breaking methods.

**Uncovered Set Rules.** The *uncovered set* of a given profile $\boldsymbol{\pi}$ is the set of all alternatives $a$ such that there is no $b$ that pairwise-dominates both $a$ and all alternatives pairwise-dominated by $a$. *Uncovered set rules* are all voting rules whose winner lies in the uncovered set, for all profiles. From this class, we primarily study the well-known rule Copeland, where the score of an alternative is the number of alternatives it pairwise-dominates, and an alternative with maximum score is the Copeland winner. We also study Slater, which selects the ranking that is inconsistent with the outcomes of as few pairwise elections as possible.

**Positional Scoring Rules.** Positional scoring rules are defined by a score vector $\mathbf{s}$ of weakly decreasing scores $s_1 \geq \cdots \geq s_m$, where (without loss of generality) $s_1 = 1$ and $s_m = 0$. The winner by positional scoring rule $f_{\mathbf{s}}$ is the alternative that receives the most points, where $a$ receives $s_j$ points for every voter that ranks it $j$th. We will study three standard positional scoring rules, Plurality with score vector $\mathbf{s} = (1, 0, \ldots, 0)$, Borda with score vector $\mathbf{s} = (1, 1 - 1/m-1, 1 - 2/m-1, \ldots, 1/m-1, 0)$, and Veto with score vector $\mathbf{s} = (1, \ldots, 1, 0)$. We will also define a new positional scoring rule Piecewise in Section 12.3, which will achieve better distortion than any of the previous three.

**Other rules and axioms.** We characterize one additional rule, Maximin, which chooses the alternative with the lowest minimax score, defined for $a$ as the magnitude of $a$'s most severe pairwise domination, i.e., $\max_{\tilde{a} \neq a} |\{i : \tilde{a} \succ_{\pi_i} a\}|$. We also sometimes discuss the axiom *Condorcet consistency*, where $f$ is Condorcet consistent if it selects the Condorcet winner in all profiles in which one exists. Of the rules we study, Copeland, Slater, and Maximin are Condorcet consistent.

### 12.2.3 Distortion of Voting Rules

The distortion of a voting rule $f$ in an instance $(\boldsymbol{\gamma}, U)$, called $\text{dist}(f, \boldsymbol{\gamma}, U)$, is the ratio between the respective welfares of the highest-welfare alternative $a^*$ and the winner $f(\boldsymbol{\pi})$. As is standard, we use *distortion*, called $\text{dist}(f, \boldsymbol{\gamma})$, to mean the *worst-case* such ratio over all $U$ (here, for a fixed $\boldsymbol{\gamma}$).

$$\text{dist}(f, \boldsymbol{\gamma}, U) := \sup_{\boldsymbol{\pi} \in \Pi_{V(\boldsymbol{\gamma}, U)}} \frac{\text{sw}(a^*, U)}{\text{sw}(f(\boldsymbol{\pi}), U)}, \quad \text{and} \quad \text{dist}(f, \boldsymbol{\gamma}) := \sup_{U \in \mathbb{R}_{\geq 0}^{n \times m}} \text{dist}(f, \boldsymbol{\gamma}, U).$$

[1] The rules we study are standard, defined in, e.g., [84] (Slater) and [286] (all others).

## 12.3 Distortion Bounds for Voting Rules

We now analyze the distortion of several voting rules under the condition that $\gamma_{min}$, the minimum level of public spirit among voters, is positive. First, in Section 12.3.1, we prove our key lemma, which founds our analysis of specific voting rules and gives intuition for why public spirit *should* limit the distortion. In Section 12.3.2, we will apply this lemma in various forms to upper bound the distortion of several standard voting rules. Section 12.3.3 contains our lower bounds for these rules, which match in almost all cases. We summarize these bounds in Table 13.2. Most include exact constants; the few asymptotic results we give are asymptotic in $m$, as is standard in the distortion literature.

| **Rule** | **Upper bounds** | | **Lower bounds** | |
|---|---|---|---|---|
| Uncovered set rules | $(2z_{\gamma_{min}}+1)^2$ | (Thm. 12.3.3) | | |
| Copeland | $(2z_{\gamma_{min}}+1)^2$ | | $(2z_\gamma+1)^2$ | (Prop. 12.3.9) |
| Slater | $(2z_{\gamma_{min}}+1)^2$ | | $(2z_\gamma+1)^2$ | (Prop. 12.3.10) |
| Positional scoring rules | | | $\Omega(\sqrt{m})$ | (Thm. 12.3.11) |
| Plurality | $mz_{\gamma_{min}}+1$ | (Prop. 12.3.5) | $mz_\gamma+1$ | (Prop. 12.3.15) |
| Borda | $mz_{\gamma_{min}}+1$ | (Prop. 12.3.6) | $(m-1)z_\gamma+1$ | (Prop. 12.3.13) |
| Veto | | | infinite | (Prop. 12.3.14) |
| Piecewise | $O(m^{2/3})$ | (Prop. 12.3.7) | $\Omega(\sqrt{m})$ | |
| Maximin | $mz_{\gamma_{min}}+1$ | (Prop. 12.3.8) | $(m-1)z_\gamma+1$ | (Prop. 12.3.16) |

Table 12.1: Bounds on the distortion of voting rules. Upper bounds hold for all $\boldsymbol{\gamma}$; lower bounds hold for all uniform $\boldsymbol{\gamma} = \gamma\mathbf{1}$. As shorthand, we let $z_\gamma = {}^{(1-\gamma)}/_{\gamma}$. Gray-text results are inherited from more general results.

### 12.3.1 Key Lemma

**Lemma 12.3.1.** *For all $U$, all $a, b \in [m]$ with $sw(a,U) > 0$, all $\boldsymbol{\gamma}$ with $\gamma_{min} > 0$, and all $\boldsymbol{\pi} \in \Pi_{V(\boldsymbol{\gamma},U)}$,*

$$\frac{sw(b,U)}{sw(a,U)} \leq \frac{1-\gamma_{min}}{\gamma_{min}} \cdot \frac{n}{|\{i : a \succ_{\pi_i} b\}|} + 1.$$

Conceptually, Lemma 12.3.1 states that for arbitrary alternatives $a, b$, the more voters who rank $a$ ahead of $b$, the less the welfare of $b$ can exceed that of $a$ (assuming $\gamma_{min} > 0$). The intuition for the proof, below, is that any voter $i$ who ranks $a \succ_{\pi_i} b$ must have utility for $a$ that exceeds $b$ sufficiently to close the countervailing gap $\mathrm{sw}(b,U) - \mathrm{sw}(a,U)$, which is weighted by $\gamma_i$ in $i$'s PS-value. This fact implies a lower bound on $i$'s utility for $a$, which grows in $\gamma_i$; summing over all voters $i$, we get a lower bound on $\mathrm{sw}(a,U)$ relative to $\mathrm{sw}(b,U)$, which grows stronger in $\gamma_{min}$.

*Proof.* Fix a $U$, $\boldsymbol{\gamma}$, and let $\boldsymbol{\pi} \in \Pi_{V(\boldsymbol{\gamma},U)}$. Let $N_{a \succ b}$ be the set of voters in $\boldsymbol{\pi}$ who rank $a$ ahead of $b$, and let $i \in N_{a \succ b}$. The fact that $a \succ_{\pi_i} b$ means that $v_i(a, \boldsymbol{\gamma}, U) \geq v_i(b, \boldsymbol{\gamma}, U)$, implying that

$$(1-\gamma_i)u_i(a) + \gamma_i \frac{\mathrm{sw}(a,U)}{n} = v_i(a,\boldsymbol{\gamma},U) \geq v_i(b,\boldsymbol{\gamma},U) = (1-\gamma_i)u_i(b) + \gamma_i \frac{\mathrm{sw}(b,U)}{n} \geq \gamma_i \frac{\mathrm{sw}(b,U)}{n}.$$

Now, dividing both sides by $\gamma_i$ and then adding up both sides over all $i \in N_{a \succ b}$:

$$\sum_{i \in N_{a \succ b}} \left( \frac{1-\gamma_i}{\gamma_i} u_i(a) + \frac{\mathrm{sw}(a,U)}{n} \right) \geq \sum_{i \in N_{a \succ b}} \frac{\mathrm{sw}(b,U)}{n}.$$

Using that $\frac{1-\gamma_i}{\gamma_i}$ is decreasing in $\gamma_i$ and making simplifications,

$$\implies |N_{a \succ b}|/n \cdot \mathrm{sw}(a,U) + \frac{1-\gamma_{min}}{\gamma_{min}} \sum_{i \in N_{a \succ b}} u_i(a) \geq |N_{a \succ b}|/n \cdot \mathrm{sw}(b,U).$$

Finally, we use that $\sum_{i \in N_{a \succ b}} u_i(a) \leq \mathrm{sw}(a,U)$ to conclude the claim. □

In the next sections, we will apply this lemma to upper bound the distortion of various voting rules. Although we apply it in different ways across voting rules, the key idea is always the same: as long as enough voters rank the election winner $a'$ ahead of the highest-welfare alternative $a^*$ (or an alternative with social welfare comparable to $a^*$), then $\mathrm{sw}(a^*, U)$ cannot exceed $\mathrm{sw}(a', U)$ by more than a bounded amount, bounding the distortion. Intuitively, for "reasonable" voting rules, the number of voters who prefer $a'$ to some such alternative *should* be lower-bounded — otherwise, $a'$ would not be the winner. We will formalize this intuition as we prove our upper bounds.

### 12.3.2 Upper bounds

#### Uncovered Set Rules

We will now show that, when $\gamma_{min} > 0$, all uncovered set rules — most notably including Copeland and Slater — have *constant* distortion. To prove this, we apply Lemma 12.3.1 in two different ways: in the first case, we use it to directly compare $a'$, the winner, and $a^*$. In the second and more interesting case, we apply the lemma twice, first to compare $a'$ with some intermediate alternative $a$, and then to compare $a$ with $a^*$. The choice of this intermediate alternative $a$ arises from a known[1] property of the uncovered set:

**Lemma 12.3.2** ([205]). *If $a'$ is in the uncovered set then for all $a \neq a'$, $a'$ either weakly pairwise-dominates $a$, or there exists some $a''$ such that $a'$ weakly pairwise-dominates $a''$ and $a''$ weakly pairwise-dominates $a$.*

[1] Our framing slightly adapts the classic result [205] to permit pairwise ties. We remark that this result was also used to prove the constant distortion of uncovered set rules under metric preferences [24, Thm. 5].

**Theorem 12.3.3.** *For all uncovered set rules $f$ and all $\boldsymbol{\gamma}$ with $\gamma_{min} > 0$,*

$$dist(f, \boldsymbol{\gamma}) \leq \left(\frac{2(1-\gamma_{min})}{\gamma_{min}} + 1\right)^2.$$

*Proof.* Let $f$ be an uncovered set rule, and fix arbitrary $U$, $\boldsymbol{\gamma}$ and $\boldsymbol{\pi} \in \Pi_{V(\boldsymbol{\gamma},U)}$. Let $a^*$ be the highest-welfare alternative in $U$, and let $a'$ be the winner by $f$, i.e., $a' = f(\boldsymbol{\pi})$. Then, we know $a'$ is in the uncovered set. If $a'$ weakly pairwise-dominates $a^*$, then $|\{i : a' \succ_{\pi_i} a^*\}|/n \geq 1/2$ and by applying Lemma 12.3.1 with $a' = a, a^* = b$, we immediately obtain an upper bound stronger than the claim. Else, by Lemma 12.3.2, there exists some $a$ such that $a'$ weakly pairwise-dominates $a$, and $a$ weakly pairwise-dominates $a^*$. Fix this $a$. Then, by Lemma 12.3.1, both $\mathrm{sw}(a^*)/\mathrm{sw}(a)$ and $\mathrm{sw}(a)/\mathrm{sw}(a')$ are at most $2(1-\gamma_{min})/\gamma_{min} + 1$. Multiplying these inequalities implies the claim. □

### Positional scoring rules

In giving upper bounds on the distortion of positional scoring rules, we will establish an upper bound on the distortion of *all* voting rules — one which will turn out to be tight for not only key positional scoring rules, but also some Condorcet consistent rules (e.g., Maximin, as analyzed in Section 12.3.2). This upper bound will be a corollary of Lemma 12.3.1, derived by using the lemma to compare the social welfares of $a'$ directly with $a^*$.

Formally, we deduce this corollary by plugging in $a = f(\boldsymbol{\pi})$ (for any $\boldsymbol{\pi} \in \Pi_{V(\boldsymbol{\gamma},U)}$) and $b = a^*$. Then, for a given $f$, we need only to bound the quantity $|\{i : f(\boldsymbol{\pi}) \succ_{\pi_i} a^*\}|/n$. We thus define the parameter $\kappa_f(m)$, the minimum fraction of voters who must rank the winner $f(\boldsymbol{\pi})$ ahead of *any* other given alternative, in *any* profile $\boldsymbol{\pi}$.

$$\kappa_f(m) := \min_{\boldsymbol{\pi}} \min_{a \neq f(\boldsymbol{\pi})} |\{i : f(\boldsymbol{\pi}) \succ_{\pi_i} a\}|/n. \tag{12.3}$$

Although this quantity is often function of $m$, for brevity we will write it as $\kappa_f$. For a fixed $f$, we then have by definition that $|\{i : f(\boldsymbol{\pi}) \succ_{\pi_i} a^*\}|/n \geq \kappa_f$ for all instances $(\boldsymbol{\gamma}, U)$ and corresponding $\boldsymbol{\pi} \in \Pi_{V(\boldsymbol{\gamma},U)}$, as needed. From this we conclude the following corollary of Lemma 12.3.1, which we emphasize is an upper bound on the distortion of *any* voting rule $f$:

**Corollary 12.3.4** (Universal Upper Bound)**.** *For all rules $f$ and all $\boldsymbol{\gamma}$ with $\gamma_{min} > 0$,*

$$dist(f, \boldsymbol{\gamma}) \leq \frac{1-\gamma_{min}}{\gamma_{min} \cdot \kappa_f} + 1.$$

To apply this corollary to upper bound the distortion of a specific $f$, we must simply lower bound $\kappa_f$. One useful observation, before doing so, is that for *all* $f$, $\kappa_f \leq 1/m$; thus, Corollary 12.3.4 can be used to prove linear distortion at best.[1]

[1]To see why $\kappa_f \leq 1/m$ for all $f$, divide $[n]$ into $m$ equal-sized groups $G_1, ..., G_m$. Then, for each group $G_k$, suppose the voters have rankings $k \succ k+1 \succ ... \succ m \succ 1 \succ ...k-1$. In this case, every alternative's worst pairwise defeat is to be ranked behind another alternative by an $(m-1)/m$ voters. Hence, $\kappa_f \leq 1/m$.

Now, we prove upper bounds on the standard positional scoring rules Borda and Plurality by characterizing their respective $\kappa_f$ values and applying Corollary 12.3.4:

**Proposition 12.3.5.** *$\kappa_{\text{Plurality}} = 1/m$, so for all $\boldsymbol{\gamma}$ with $\gamma_{min} > 0$, dist(Plurality, $\boldsymbol{\gamma}$) $\leq m\frac{1-\gamma_{min}}{\gamma_{min}} + 1$.*

**Proposition 12.3.6.** *$\kappa_{\text{Borda}} = 1/m$, so for all $\boldsymbol{\gamma}$ with $\gamma_{min} > 0$, dist(Borda, $\boldsymbol{\gamma}$) $\leq m\frac{1-\gamma_{min}}{\gamma_{min}} + 1$.*

For Veto, we cannot apply the same approach, because $\kappa_{\text{Veto}}$ is $1/n$ — i.e., there exists an instance in which just one voter must rank the winner ahead of any other alternative — and thus the upper bound given by Corollary 12.3.4 is unbounded in $n$. It will turn out that, as Corollary 12.3.4 would suggest, the distortion of Veto is truly unbounded, shown via an instance in which the Veto-winner is almost never ranked ahead of the highest-welfare alternative.

So far, we have not found a positional scoring rule that has sub-linear distortion, prompting the question: does one exist? We answer this question in the affirmative with Piecewise, a voting rule we newly define. It can be seen as a hybrid of Plurality and Borda, defined by a score vector with $m^{2/3}$ non-zero entries: $\mathbf{s} = \left(1, 1 - 1/m^{2/3}, 1 - 2/m^{2/3}, \ldots, 1/m^{2/3}, 0, \ldots, 0\right)$. We now show that, when $\gamma_{min}$ is any nonzero constant, Piecewise suffers at most $O(m^{2/3})$ distortion. Here, we depart from the approach of directly applying Corollary 12.3.4 (as we must in order to obtain a sub-linear bound).

**Proposition 12.3.7.** *For all $\boldsymbol{\gamma}$ with (fixed) $\gamma_{min} > 0$, dist(Piecewise, $\boldsymbol{\gamma}$) $\in O(m^{2/3})$.*

The proof of this proposition, found in Appendix G.1.4, again applies our key lemma, but in a more intricate fashion than in the preceding bounds. Similarly to the proof of Theorem 12.3.3, the argument considers one case comparing the Piecewise winner $a'$ directly to $a^*$, and another comparing $a'$ to some intermediate alternative(s) other than $a^*$. The first case is invoked in profiles where at least half of voters rank $a^*$ in the first $m^{2/3}$ positions; then, normalizing $a^*$'s social welfare to be constant, $a'$ must have social welfare $\Omega(m^{-2/3})$ in order to win the election. In the second case, over half the voters must rank $a^*$ in the *last* $m - m^{2/3}$ positions, implying that each of these voters must rank at least $m^{2/3}$ many alternatives ahead of $a^*$. In order for $a'$ to win the election over these other alternatives, $a'$ must again have social welfare $\Omega(m^{-2/3})$.

### Maximin

Given that $\kappa_f$ is not meaningfully lower-bounded for Copeland and Slater (indeed, per the instance giving Proposition 12.3.9, it can be arbitrarily small), one might think that this is the case for all Condorcet consistent rules. On the contrary, here we show that $\kappa_{\text{Maximin}} = 1/m$, and thus Corollary 12.3.4 gives a useful distortion upper bound for Maximin — in fact, it will turn out that this upper bound is tight. The proof of this proposition is found in Appendix G.1.5.

**Proposition 12.3.8.** *$\kappa_{\text{Maximin}} = 1/m$, so for all $\boldsymbol{\gamma}$ with $\gamma_{min} > 0$, dist(Maximin, $\boldsymbol{\gamma}$) $\leq m\frac{1-\gamma_{min}}{\gamma_{min}} + 1$.*

### 12.3.3 Lower bounds

We give matching lower bounds for all voting rules analyzed in Section 12.3.2 except Piecewise. The lower bound we give for Piecewise is $\Omega(\sqrt{m})$ (thus leaving an asymptotic gap of $m^{1/6}$) is implied by Theorem 12.3.11, which shows that even when voters are public-spirited, *all* positional scoring rules must suffer at least $\Omega(\sqrt{m})$ distortion. The proofs of all our lower bounds proceed by fixing an arbitrary uniform PS-vector $\boldsymbol{\gamma} = \gamma\mathbf{1}$, and then constructing a utility matrix $U$ whose entries depend on $\gamma$, in which the election winner $a'$ has far lower social welfare than $a^*$.

#### Uncovered set rules

**Proposition 12.3.9.** *For all uniform $\boldsymbol{\gamma} = \gamma\mathbf{1}$, $\gamma \in [0, 1]$, $dist(\textsc{Copeland}, \boldsymbol{\gamma}) \geq \left(\frac{2(1-\gamma)}{\gamma} + 1\right)^2$.*

**Proposition 12.3.10.** *For all uniform $\boldsymbol{\gamma} = \gamma\mathbf{1}$, $\gamma \in [0, 1]$, $dist(\textsc{Slater}, \boldsymbol{\gamma}) \geq \left(\frac{2(1-\gamma)}{\gamma} + 1\right)^2$.*

The proofs of these propositions are found in Appendices G.1.6 and G.1.7, respectively. Both use the same instance, constructed so that $a'$ pairwise-dominates every alternative except $a^*$, and $a^*$ pairwise-dominates all but two alternatives, $a_1,\ a_2 \neq a'$. (We use two such alternatives here only to ensure that $a^*$ is *not* contained in the uncovered set, and thus the winner $a'$ is unique. Proving the bound requires reasoning about $a_1$ *or* $a_2$; here, we explain the bound via $a_1$.) Normalizing the average utility of $a^*$ to be 1, observe that because at least half of voters rank $a_1$ ahead of $a^*$, $a_1$ must have average utility at least $2(1-\gamma)/\gamma$. In turn, because at least half of voters rank $a'$ ahead of $a_1$, $a'$ must have average utility of at least $(2(1-\gamma)/\gamma)^2$. Then, the $U$ that minimizes $a'$'s social welfare relative to $a^*$ while also realizing the above profile makes all these inequalities tight, giving the lower bound.

#### Positional scoring rules

First, in Theorem 12.3.11, we show that whenever $\gamma_{min} < 1$, *all* positional scoring rules must have distortion at least $\Omega(\sqrt{m})$. Note that this result implies a fundamental separation between positional scoring rules and uncovered set rules, which per Theorem 12.3.3 have at most constant distortion for fixed values of $\gamma_{min} > 0$.

**Theorem 12.3.11.** *For all positional scoring rules $f$ and uniform $\boldsymbol{\gamma} = \gamma\mathbf{1}$ with (fixed) $\gamma \in [0, 1)$,*

$$dist(f, \boldsymbol{\gamma}) \in \Omega(\sqrt{m}).$$

The key observation underlying this lower bound, proven formally in Appendix G.1.8, is that in any positional scoring rule's score vector, there exists some position $t$ amongst the first $\sqrt{m}$ entries in the score vector — that is, $t \in \{1, \ldots, \sqrt{m}\}$ — such that the gap $s_t - s_{t+1}$ between the scores for positions $t$ and $t+1$ is at most $1/\sqrt{m}$ (this is simply by averaging). Then, for fixed $\gamma$ and corresponding PS-vector $\boldsymbol{\gamma} = \gamma\mathbf{1}$, one can use this fact to construct an instance $(\boldsymbol{\gamma}, U)$ which realizes order-$\sqrt{m}$ distortion. The construction works as follows: Divide voters into two groups, a small group of size $O(1/\sqrt{m})$, and the remainder of the electorate. Let all voters in the larger

group rank $a^*$ in the $t$-th position and the winner $a'$ in the $(t+1)$-st position. In the small group, $a'$ is ranked first and $a^*$ is ranked last, thereby compensating for $a'$'s 'scoring' deficit in the larger group and allowing it to win the election. Because $a'$ is so rarely ranked ahead of $a^*$ in this profile, it can be realized by a utility matrix in which $a^*$ has constant average utility, while all voters have utility $O(1/\sqrt{m})$ for the winner $a'$, resulting in a distortion of order $O(\sqrt{m})$.

It turns out that many positional scoring rules have distortion far exceeding $\Omega(\sqrt{m})$ distortion; this is true, for instance, for all voting rules with a small value of $\Delta_f := s_1 - s_2$, the gap in scores of the first two ranking positions:

**Lemma 12.3.12.** *For all positional scoring rules $f$ and uniform $\boldsymbol{\gamma} = \gamma\mathbf{1}$, $\gamma \in [0,1]$, $dist(f, \boldsymbol{\gamma}) \geq \frac{1-\gamma}{\gamma\Delta_f} + 1$.*

In the proof of this proposition, found in Appendix G.1.9, we again construct an instance in which as few voters as possible rank the winner $a'$ ahead of $a^*$. To illustrate why smaller $\Delta_f$ permits fewer voters to rank $a'$ ahead of $a^*$, we will describe this construction. Divide voters into two groups: voters in the first group rank $a'$ first and $a^*$ last, and voters in the second group rank $a^*$ and $a'$ adjacently over the first two positions. In order for $a'$ to win this election, the first group must contain at least $\Delta_f$ voters; moreover, only these voters must have non-negligible utility for $a'$. Note that the use of the gap over the *first two positions* is essential: if we placed $a^* \succ a'$ over a smaller adjacent gap elsewhere, $a'$ would be ranked below several other alternatives by many voters, and we could no longer guarantee that it wins the election.

We can now directly apply Lemma 12.3.12 to lower bound the distortion of Borda and Veto, using that $\Delta_{\text{Borda}} = 1/(m-1)$ and $\Delta_{\text{Veto}} = 0$.

**Proposition 12.3.13.** *For all uniform $\boldsymbol{\gamma} = \gamma\mathbf{1}$, $\gamma \in [0,1]$, $dist(\textsc{Borda}, \boldsymbol{\gamma}) \geq (m-1) \cdot \frac{1-\gamma}{\gamma} + 1$.*

**Proposition 12.3.14.** *For all uniform $\boldsymbol{\gamma} = \gamma\mathbf{1}$, $\gamma \in [0,1]$, $dist(\textsc{Veto}, \boldsymbol{\gamma}) = \infty$.*

For Plurality, $\Delta_{\text{Plurality}} = 1$, so Lemma 12.3.12 does not give a useful lower bound. However, we can get a tight lower bound using a similar construction: We let a $1/m + \epsilon$ fraction of voters rank $a'$ first, and all other voters rank $a'$ last. Only the former group of voters must have non-negligible utility for $a'$, while all other alternatives can receive non-negligible utility from the much larger second group of voters, yielding linear distortion. The full proof is found in Appendix G.1.10.

**Proposition 12.3.15.** *For all uniform $\boldsymbol{\gamma} = \gamma\mathbf{1}$, $\gamma \in [0,1]$, $dist(\textsc{Plurality}, \boldsymbol{\gamma}) \geq m \cdot \frac{1-\gamma}{\gamma} + 1$.*

### Maximin

Finally, we show that our upper bound on Maximin's distortion was, indeed, tight.

**Proposition 12.3.16.** *For all uniform $\boldsymbol{\gamma} = \gamma\mathbf{1}$, $dist(\textsc{Maximin}, \boldsymbol{\gamma}) \geq (m-1) \cdot \frac{1-\gamma}{\gamma} + 1$.*

The formal proof of this proposition is found in Appendix G.1.11. The construction is somewhat involved, but it intuitively works as follows: voters are divided into two groups. In Group 1, containing a $1/(m-1)$ fraction of voters, the election winner $a'$ is ranked first; in Group 2,

composed of the remaining voters, $a'$ is ranked last. The relative ranking of alternatives other than $a'$ is 'cyclical' — that is, all voters order them identically, up to a shift. There are $m-1$ possible such shifts, and each shifted ranking occupies a $1/(m-1)$ fraction of the voters. In this profile, $a'$'s greatest pairwise defeat is by $(m-2)/(m-1)$ fraction of voters, and the cyclical treatment of all other alternatives ensures that each suffers a pairwise defeat at least as severe as $a'$, making $a'$ the winner. This profile can be realized with a utility matrix in which all alternatives besides $a'$ get utility 1 from all voters in Group 2, while $a'$ only gets utility from Group 1.

## 12.4 PS-monotonicity

Given that increasing voters' public spirit can only promote higher-welfare alternatives in their rankings, it seems natural that distortion should decrease as voters' public spirit increases. We refer to this general property of voting rules — decreasing distortion with increasing public spirit — as *public-spirit monotonicity* (for short, *PS-monotonicity*). Our upper (and matching lower) bounds from Section 12.3 already hint at a weak form of PS-monotonicity, as they are decreasing in $\gamma_{min}$.

In this section, we pursue stronger forms of PS-monotonicity, which ask for monotonicity not just in $\gamma_{min}$, but in voters' individual levels of public spirit. To this end, we define and analyze three notions of PS-monotonicity, from weakest to strongest. We first study *uniform PS-monotonicity*, which requires that distortion decreases as public spirit increases uniformly across voters. We find that this property holds for *all voting rules* — i.e., it is a fundamental property of the model. We next study a much stronger notion, *nonuniform PS-monotonicity*, which requires that the distortion decreases as voters' public spirit increases heterogeneously. We show that this notion holds for all voting rules when $m \leq 3$, and it holds for arbitrary $m$ for Copeland and Plurality.

These first two notions examine monotonicity in the *worst-case* distortion. Even more optimistically, one might hope that public spirit would decrease the distortion on an *instance-wise* basis: i.e. in a fixed instance, if all voters' public spirit levels weakly increase, the welfare of the chosen outcome should only increase. We refer to this property as *instance-wise PS-monotonicity*. Unfortunately, we prove via classical voting axioms that no reasonable voting rule satisfies this notion: specifically, any weakly unanimous voting rule that satisfies instance-wise PS-monotonicity must be a dictatorship.

### 12.4.1 Uniform PS-monotonicity

**Definition 12.4.1** (Uniform PS-monotonicity)**.** *A voting rule $f$ exhibits uniform PS-monotonicity if, for all $\gamma' \geq \gamma$ and associated uniform $\boldsymbol{\gamma} = \gamma\mathbf{1}, \boldsymbol{\gamma}' = \gamma'\mathbf{1}$, $dist(f, \boldsymbol{\gamma}') \leq dist(f, \boldsymbol{\gamma})$.*

**Theorem 12.4.2.** *All voting rules are uniform PS-monotonic.*

*Proof.* We will prove this theorem by showing that, given arbitrary $U$ and $\gamma_{\text{big}} \geq \gamma_{\text{small}}$, we can find $\tilde{U}$ such that $\text{dist}(f, \gamma_{\text{big}}, U) = \text{dist}(f, \gamma_{\text{small}}, \tilde{U})$: roughly, under a lower level of public spirit, there exists a utility matrix with distortion at least as high. In fact, this distortion-preserving $\tilde{U}$

will simply be $U$ with some carefully-chosen amount of public spirit applied:

$$\tilde{U} := V(\gamma^*, U), \quad \text{where} \quad \gamma^* := \frac{\gamma_{\text{big}} - \gamma_{\text{small}}}{1 - \gamma_{\text{small}}}. \tag{12.4}$$

We begin by considering two $n \times m$ matrices: $U$ (which we can interpret as a matrix) and $W_U$, whose columns contain the column sums of $U$:

$$W_U = \begin{bmatrix} \text{sw}(a_1, U)/n & \dots & \text{sw}(a_m U)/n \\ \vdots & & \vdots \\ \text{sw}(a_1, U)/n & \dots & \text{sw}(a_m, U)/n \end{bmatrix}.$$

We think of applying an arbitrary $\gamma$ to $U$ as a linear transformation on $U$, where varying $\gamma$ from 0 to 1 interpolates between the matrices $U$ and $W_U$: applying $\gamma = 0$ returns $U$, applying $\gamma = 1$ returns $W_U$, and there is an infinite sequence of matrices in between ranging over $\gamma \in [0, 1]$, where the $\gamma$-th matrix is equal to a convex combination of $U$ and $W_U$ — that is, $V(\gamma, U) = (1 - \gamma)U + \gamma W_U$.

A key property of this transformation is that it is *column-sum-preserving*, so all matrices in this sequence have the same column sums; that is, for all $\gamma \in [0, 1]$, $W_{V(\gamma,U)} = W_U$. We use this fact to make the general observation that applying public spirit $\gamma_1$ and then $\gamma_2$ in succession is the same as applying $\gamma_1 + \gamma_2 - \gamma_1\gamma_2$ public spirit all at once:

**Lemma 12.4.3.** *For arbitrary $U$ and arbitrary $\gamma_1, \gamma_2 \in [0, 1]$, $V(\gamma_2, V(\gamma_1, U)) = V(\gamma_1 + \gamma_2 - \gamma_1\gamma_2, U)$.*

*Proof of Lemma 12.4.3:*

$$\begin{aligned} V\left(\gamma_2, V(\gamma_1, U)\right) &= \gamma_2 W_{V(\gamma_1,U)} + (1 - \gamma_2)V(\gamma_1, U) \\ &= \gamma_2 W_{V(\gamma_1,U)} + (1 - \gamma_2)\big((1 - \gamma_1)U + \gamma_1 W_U\big) \\ &= \gamma_2 W_U \; + \; (1 - \gamma_2)((1 - \gamma_1)U + \gamma_1 W_U) \\ &= (1 - \gamma_1)(1 - \gamma_2)U + \big(\gamma_1 + \gamma_2 - \gamma_1\gamma_2\big)W_U \\ &= V\left(\gamma_1 + \gamma_2 - \gamma_1\gamma_2, U\right) \end{aligned}$$ □

Because applying public spirit is column-sum-preserving, we can set $\tilde{U}$ to *any* matrix $V(\gamma, U)$, $\gamma \in [0, 1]$ and be certain that $\tilde{U}$ will give the same welfares to all alternatives as $U$. We will carefully choose this $\gamma = \gamma^*$ according to Lemma 12.4.3: $\gamma^* = \gamma_1$, $\gamma_{\text{small}} = \gamma_2$, and $\gamma_{\text{big}} = \gamma_1 + \gamma_2 - \gamma_1\gamma_2$, which means setting $\gamma^*$ as in Equation (12.4). This setting of $\gamma^*$ then ensures that the rankings are preserved:

$$V(\gamma_{\text{small}}, V(\gamma^*, U)) = V(\gamma_{\text{big}}, U) \implies \Pi_{V(\gamma_{\text{small}}, \tilde{U})} = \Pi_{V(\gamma_{\text{big}}, U)}.$$

Thus, across $(U, \gamma_{\text{big}})$ and $(\tilde{U}, \gamma_{\text{small}})$, the rankings (and therefore the winner) and social welfares are identical. The distortion must then be the same across the instances, proving the claim. □

### 12.4.2 Nonuniform PS-monotonicity

Here, we define the ordering of vectors in the standard way: $\boldsymbol{\gamma}' \geq \boldsymbol{\gamma}$ iff $\gamma_i' \geq \gamma_i$ for all $i \in [n]$.

**Definition 12.4.4** (Nonuniform PS-monotonicity)**.** *A voting rule $f$ exhibits nonuniform PS-monotonicity if for all $\boldsymbol{\gamma}, \boldsymbol{\gamma}'$ where $\boldsymbol{\gamma}' \geq \boldsymbol{\gamma}$, $dist(f, \boldsymbol{\gamma}') \leq dist(f, \boldsymbol{\gamma})$.*

First, we show that nonuniform PS-monotonicity holds for *all* voting rules when $m \leq 3$.

**Proposition 12.4.5.** *If $m \leq 3$, then all voting rules exhibit nonuniform monotonicity.*

We defer the proof of this proposition to Appendix G.2.1, as it is fairly involved. The main intuition behind the proof is as follows: If $U$ is a utility matrix, $\boldsymbol{\gamma}$ is some PS-vector and $\tilde{\boldsymbol{\gamma}}$ arises from *lowering* an entry in $\boldsymbol{\gamma}$, then we can explicitly construct another utility matrix $\tilde{U}$ such that the profile(s) implied by $(\boldsymbol{\gamma}, U)$ and $(\tilde{\boldsymbol{\gamma}}, \tilde{U})$ are identical (i.e., $\Pi_{V(\boldsymbol{\gamma},U)} = \Pi_{V(\tilde{\boldsymbol{\gamma}},\tilde{U})}$), and the social welfares of all alternatives are preserved (i.e., $\mathrm{sw}(a, U) = \mathrm{sw}(a, \tilde{U})$ for all $a$). Across these instances, the election winner, and thus the distortion, must be the same.

The construction used to show Proposition 12.4.5 is already considerably complicated when $m = 3$; proving the claim for all (or a broad class of) voting rules when $m \geq 4$ remains an interesting open problem. However, we do affirmatively resolve this question for two specific voting rules, showing that Copeland and Plurality both satisfy nonuniform PS-monotonicity for arbitrary $m$.

**Proposition 12.4.6.** *Copeland is nonuniform PS-monotonic.*

**Proposition 12.4.7.** *Plurality is nonuniform PS-monotonic.*

These propositions are proven in Appendices G.2.2 and G.2.3, respectively. Although the constructions used to analyze Copeland and Plurality are different, both reflect the argument from Proposition 12.4.5: given $U, \boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}$ where $\tilde{\boldsymbol{\gamma}} \leq \boldsymbol{\gamma}$, we construct a $\tilde{U}$ such that the election winner and welfares are preserved instances. Note that these arguments can be simpler than the proof of Proposition 12.4.5 because, given that we are not reasoning about *all* voting rules, preserving these features across instances does not necessitate preserving the full preference profile. As such, in the analysis of Copeland, $\tilde{U}$ just preserves the relevant aspects of the uncovered set; in the analysis of Plurality, $\tilde{U}$ just preserves the first-ranked alternatives.

### 12.4.3 Instance-wise PS-Monotonicity

**Definition 12.4.8** (Instance-wise PS-monotonicity)**.** *A voting rule $f$ is instance-wise PS-monotonic iff, for all $U$ and all $\boldsymbol{\gamma}, \boldsymbol{\gamma}'$ where $\boldsymbol{\gamma}' \geq \boldsymbol{\gamma}$, $dist(f, \boldsymbol{\gamma}', U) \leq dist(f, \boldsymbol{\gamma}, U)$.*

Unfortunately, Theorem 12.4.11 shows that no reasonable — i.e., *weakly unanimous* (Definition 12.4.9) and *non-dictatorial* (Definition 12.4.10) — voting rule satisfies this property. Although the proof is involved, the intuition is simple: consider three alternatives in order of decreasing welfare, $a, b, c$. Suppose $a$ wins initially, but after increasing voters' public spirit, all voters promote $b$ over $c$ but no other relative rankings change. For any monotonic and otherwise reasonable voting rule, $b$ — whose welfare is lower than $a$'s — must in some cases be able to become the winner.

**Definition 12.4.9** (weakly unanimous)**.** *A voting rule $f$ is weakly unanimous iff for every profile $\boldsymbol{\pi}$, if there is a pair of alternatives $a, b$ such that $a \succ_{\pi_i} b$ for all voters $i$, then $f(\boldsymbol{\pi}) \neq a$.*

**Definition 12.4.10** (dictatorship)**.** *Voter $i$ is a dictator with respect to $f$ if $f$ always selects is top choice: for every profile $\boldsymbol{\pi}$, $f(\boldsymbol{\pi}) = a$ iff for all $a' \neq a$, $a \succ_{\pi_i} a'$. $f$ is a dictatorship if it has a dictator.*

**Theorem 12.4.11.** *If $m \geq 3$ and $f$ is weakly unanimous and instance-wise monotonic, $f$ is a dictatorship.*

We prove Theorem 12.4.11 at the end of this subsection by showing that instance-specific PS-montonicity implies an increasingly strong series of voting axioms. We build up this system of axiomatic implications until they meet the preconditions of a known result by [208] implying that $f$ is a dictatorship. Below, we step through each of these axiomatic implications, defining the relevant axioms as we go.

First, Lemmas 12.4.13 and 12.4.15 (proven in Appendix G.2.4 and Appendix G.2.5) show that for all weakly unanimous $f$, instance-wise PS-monotonicity implies *monotonicity* (Definition 12.4.12), the standard voting axiom, and *swap invariance* (Definition 12.4.14), which we newly define.

**Definition 12.4.12** (monotonic)**.** *A voting rule rule $f$ is monotonic iff, for every profile $\boldsymbol{\pi}$ such that $f(\boldsymbol{\pi}) = a$, and for every $i \in [n]$, if $\boldsymbol{\pi}'$ is identical to $\boldsymbol{\pi}$ except that in ranking $\pi_i'$, $a$ is promoted (with one adjacent swap) compared in $\pi_i$, then $f(\boldsymbol{\pi}') = a$.*

**Lemma 12.4.13.** *If $f$ is weakly unanimous and instance-wise PS-monotonic, then it is monotonic.*

**Definition 12.4.14** (swap invariant)**.** *A voting rule $f$ satisfies swap invariance iff, for every profile $\boldsymbol{\pi}$ such that $f(\boldsymbol{\pi}) = a$, every $i \in [n]$, and every pair of alternatives $b, c \in [m]$ where $b, c \neq a$, if $\boldsymbol{\pi}'$ is identical to $\boldsymbol{\pi}$ except $b$ and $c$ are adjacently swapped in $\pi_i'$, then $f(\boldsymbol{\pi}') = a$.*

**Lemma 12.4.15.** *If $f$ weakly unanimous and monotonic, then if $f$ is instance-wise PS-monotonic, it must also be swap-invariant.*

Next, Lemma 12.4.17 (proven in Appendix G.2.6) shows that together, monotonicity and swap invariance imply a stronger notion of monotonicity known as *Maskin monotonicity* (Definition 12.4.16).

**Definition 12.4.16** (Maskin-monotonic)**.** *A voting rule $f$ is Maskin-monotonic iff, for every preference profile $\boldsymbol{\pi}$ such that $f(\boldsymbol{\pi}) = a$, if $\boldsymbol{\pi}'$ is another profile such that $a \succ_{\pi_i'} b$ whenever $a \succ_{\pi_i} b$ for every voter $i$ and every alternative $b$, then $f(\boldsymbol{\pi}') = a$.*

**Lemma 12.4.17.** *If $f$ is monotonic and swap-invariant, then it is Maskin-monotonic.*

Finally, we apply Theorem 12.4.18, a known result by Muller and Satterthwaite, which shows that any voting rules that is weakly unanimous and Maskin-monotonic must also be a dictatorship.

**Theorem 12.4.18** ([208])**.** *When $m \geq 3$, if $f$ is weakly unanimous and Maskin-monotonic, it is also dictatorial.*

We prove Theorem 12.4.11 by applying these lemmas in sequence.

*Proof of Theorem 12.4.11.*

$f$ is *weakly unanimous* and *instance-wise PS-monotonic* $\implies$ $f$ is *monotonic* (Lemma 12.4.13)

$f$ is *weakly unanimous*, *monotonic*, and *instance-wise PS-monotonic*
$\implies$ $f$ is *swap-invariant* (Lemma 12.4.15)

$f$ is *monotonic* and *swap-invariant* $\implies$ $f$ is *Maskin-monotonic* (Lemma 12.4.17)

$f$ is *weakly unanimous* and *instance-wise PS-monotonic*
$\implies$ $f$ is *weakly unanimous* and *Maskin-monotonic*
$\implies$ $f$ is a dictatorship. (Theorem 12.4.18)

□

## 12.5 Robustness of Distortion Bounds

So far, we have considered the distortion of voting rules under two ideal conditions, which we will now relax: (a) $\gamma_{min}$, the minimum public spirit level, is bounded away from zero, and (b) voters act according to precise and internally-consistent values of the model inputs $u_i(a)$, $\gamma_i$, and $\mathrm{sw}(a, U)$. We will show that the distortion is asymptotically maintained — and degrades smoothly by constant factors — as we relax these conditions (up to an extent, for (a)).

When proving robustness to violation of (a), we essentially work within our model; to study deviations from (b), we meaningfully generalize our model to encompass a variety of errors. Our arguments for both types of robustness follow the same structure, paralleling the main upper bound results from Sections 12.3.1 and 12.3.2 in the robust setting. In particular, for both (a) and (b), we first prove a "robust" version of Lemma 12.3.1, and then deduce corresponding "robust" distortion upper bounds via the same arguments used to deduce our original upper bounds from Lemma 12.3.1.

### 12.5.1 Robustness to a Non-Public-Spirited Contingent

Here, we show that our upper bounds from Section 12.3 continue to hold up to constants as long as the number of non-public-spirited voters $i$, i.e. with $\gamma_i = 0$, is not too large. We begin by proving a "robust" version of Lemma 12.3.1, with respect to this form of robustness:

**Lemma 12.5.1.** *Let $U$ be any utility matrix, and let $\boldsymbol{\gamma}$ be such that $\gamma_{min} > 0$. Then, for any $c < 1$, any alternatives $b$, $a$ with $sw(a, U) > 0$ and any $\tilde{\boldsymbol{\gamma}}$ which arises from setting the public spirit of at most any $c \cdot |\{a \succ_{\pi_i} b\}|$ voters in $\boldsymbol{\gamma}$ to zero, it holds that*

$$\frac{sw(b, U)}{sw(a, U)} \leq \frac{1 - \gamma_{min}}{\gamma_{min}} \cdot \frac{n}{|\{i : a \succ_{\pi_i} b\}|(1 - c)} + 1.$$

*Proof.* Let us denote the set of voters who both have at least $\gamma_{min}$ public spirit and rank $a$ ahead of $b$ by $\tilde{N}_{a \succ b} := |\{i : a \succ_{\pi_i} b \text{ and } \gamma_i \geq \gamma_{min}\}|$. Then, we can follow the same arguments as in the

proof of Lemma 12.3.1 with $\tilde{N}_{a>b}$ in place of $N_{a>b}$ to obtain the inequality

$$\frac{|\tilde{N}_{a>b}|}{n}\mathrm{sw}(a,U) + \frac{1-\gamma_{min}}{\gamma_{min}}\mathrm{sw}(a,U) \geq \frac{|\tilde{N}_{a>b}|}{n}\mathrm{sw}(b,U).$$

Dividing both sides by $\mathrm{sw}(a,U) \cdot |\tilde{N}_{a>b}|/n$ yields

$$\frac{\mathrm{sw}(b,U)}{\mathrm{sw}(a,U)} \leq \frac{1-\gamma_{min}}{\gamma_{min}}\frac{n}{|\tilde{N}_{a>b}|} + 1.$$

By assumption, $|\tilde{N}_{a>b}| \geq (1-c)|\{i : a \succ_{\pi_i} b\}|$, and the claim follows. □

Then, since $\kappa_f$ lower bounds the fraction of agents who must rank ahead the winner (which we think of as $a$ in the lemma above) ahead of the maximum welfare alternative (which we think of as $b$), Lemma 12.5.1 immediately implies the following corollary, as Lemma 12.3.1 implied Corollary 12.3.4.

**Corollary 12.5.2.** *Let $f$ be any voting rule, and let $\boldsymbol{\gamma}$ with $\gamma_{min} > 0$. Then, for any $c < 1$ and any $\tilde{\boldsymbol{\gamma}}$ created by setting the public spirit of at most $c\kappa_f \cdot n$ many voters in $\boldsymbol{\gamma}$ to zero,*

$$dist(f, \tilde{\boldsymbol{\gamma}}) \leq \frac{1-\gamma_{min}}{\gamma_{min}(1-c)\kappa_f} + 1.$$

Similarly, for Uncovered Set Rules, Lemma 12.5.1 implies the following corollary (analogously to Lemma 12.3.1 implying Theorem 12.3.3).

**Corollary 12.5.3.** *Let $f$ be an uncovered set rule, and let $\boldsymbol{\gamma}$ with $\gamma_{min} > 0$. Then, for any $c < 1/2$ and for any $\tilde{\boldsymbol{\gamma}}$ created by setting the public spirit of a $c$-fraction of voters in $\boldsymbol{\gamma}$ to zero,*

$$dist(f, \tilde{\boldsymbol{\gamma}}) \leq \left(\frac{1-\gamma_{min}}{\gamma_{min}(1/2-c)} + 1\right)^2.$$

### 12.5.2 Robustness to inaccurate or internally-inconsistent voter behavior

Our model assumes that voters know (or can come to know) their own utilities and the respective welfares of all alternatives, to which they then uniformly apply some level of public spirit. However, voters almost certainly do not maintain precise internal values of $\gamma_i$ and $u_i(a)$, $\mathrm{sw}(a,U)$ for all $a$, and then vote by tabulating their PS-values. In fact, it is dubious whether a voter, if asked, could even assign useful numeric values to these quantities. As a result, the best we can probably hope for in practice is that voters have some internal *sense* of these quantities, which may be subject to errors, biases, and internal inconsistencies.

This motivates our extension of our upper bounds to the case where voters may deviate from our model with respect to *any input* to Equation (12.1). First, we allow voters to misestimate their utilities, and likewise the social welfares, by some bounded multiplicative error. Since we can always

rescale utilities by a multiplicative factor without changing the voting outcome or the distortion, we can without loss of generality only consider the case where voters *overestimate* these quantities. Second, voters may apply different levels of public spirit to different alternatives. These are not "errors", per se, because voters' levels of public spirit do not factor into the utilitiarian social welfare (our benchmark) and thus do not necessarily have a ground-truth value. These deviations can rather be seen as internal inconsistencies — or even natural behaviors — where voters are partial to the nature of certain alternatives' social benefit over that of others.

To formalize these errors, we assume voter $i$ applies multiplicative errors $\delta_i(a) \geq 1$ to their utility for $a$ and $\eta_i(a) \geq 1$ to the social welfare of $a$. We define $\delta^* = \max_{i\in[n],a\in[m]} \delta_i(a)$ and $\eta^* = \max_{i\in[n],a\in[m]} \eta_i(a)$ as the maximum such errors across all voters and alternatives, and we let $\boldsymbol{\delta} \in [1,\delta^*]^{n\times m}$, $\boldsymbol{\eta} \in [1,\eta^*]^{n\times m}$ be the matrices of these errors across voters and alternatives. We also assume $i$ applies public spirit level $\gamma_i(a)$ to each alternative $a$, and we let the *PS-matrix* $\Gamma \in [0,1]^{n\times m}$ be the matrix of these $\gamma$ over all voters and alternatives. We now let $\gamma_{min} = \min_{i\in[n],a\in[m]} \gamma_i(a)$ be the minimum level of public spirit in $\Gamma$.

Incorporating these deviations, voter $i$'s *effective* PS-value is then

$$\tilde{v}_i(a,\Gamma,U,\boldsymbol{\delta},\boldsymbol{\eta}) := (1-\gamma_i(a))\cdot\delta_i(a)u_i(a) + \gamma_i(a)\cdot\eta_i(a)\mathrm{sw}(a,U)/n.$$

Correspondingly, we let $\tilde{V}(\Gamma,U,\boldsymbol{\delta},\boldsymbol{\eta})$ be the matrix of all voters' effective PS-values. Finally, we define distortion under such errors, bounded above by $\delta^*,\eta^*$ respectively, as

$$\mathrm{dist}^{\delta^*,\eta^*}(f,\Gamma) := \sup_{U\in\mathbb{R}^{n\times m}_{\geq 0},\ \boldsymbol{\delta}\in[1,\delta^*]^{n\times m},\ \boldsymbol{\eta}\in[1,\eta^*]^{n\times m}} \sup_{\boldsymbol{\pi}\in\Pi_{\tilde{V}(\Gamma,U,\boldsymbol{\delta},\boldsymbol{\eta})}} \frac{\mathrm{sw}(a^*,U)}{\mathrm{sw}(f(\boldsymbol{\pi}),U)}$$

*A priori*, it seems that the distortion of a voting rule might not be at all robust to such errors, because even a minimal deviation could cause a pivotal switch in two alternatives, changing the winner and causing a jump in distortion. Surprisingly, however, we find that we can give distortion upper bounds on any voting rule that increase smoothly in $\delta^*$ and incur merely an additive term of $\eta^*/\gamma_{min}$. At a high level, this holds because Lemma 12.3.1 must still upper-bound the ratio of the *estimated* social welfares of the winner and $a^*$, which in turn bounds the ratio of the *true* welfares, given that the estimates are not too far off. We formalize this intuition below in a generalized "robust" version of Lemma 12.3.1 that incorporates these errors.

**Lemma 12.5.4.** *Fix utility matrix $U$, $\delta^*,\eta^*$, errors $\boldsymbol{\delta} \in [1,\delta^*]^{n\times m}$ and $\boldsymbol{\eta} \in [1,\eta^*]^{n\times m}$, and a PS-matrix $\Gamma$ with $\gamma_{min} > 0$. Then, for any alternatives $a$, $b$ with $sw(a,U) > 0$ and any $\boldsymbol{\pi} \in \Pi_{\tilde{V}(\Gamma,U,\boldsymbol{\delta},\boldsymbol{\eta})}$,*

$$\frac{sw(b,U)}{sw(a,U)} \leq \frac{\delta^*\cdot(1-\gamma_{min})}{\gamma_{min}} \cdot \frac{n}{|\{i : a \succ_{\pi_i} b\}|} + \frac{\eta^*}{\gamma_{min}}$$

*Proof.* We take the same approach as in the proof of Lemma 12.3.1, this time accounting for all

deviations. For any voter $i$ ranking $a \succ b$, and thus having $\tilde{v}_i(a, \Gamma, U, \boldsymbol{\delta}, \boldsymbol{\eta}) \geq \tilde{v}_i(b, \Gamma, U, \boldsymbol{\delta}, \boldsymbol{\eta})$,

$$\begin{aligned}(1-\gamma_i(a)) \cdot \delta^* u_i(a) + \gamma_i(a)\frac{\eta^* \mathrm{sw}(a,U)}{n} &\geq (1-\gamma_i(a)) \cdot \delta_i(a) u_i(a) + \gamma_i(a) \cdot \eta_i(a)\frac{\mathrm{sw}(a,U)}{n} \\ &\geq (1-\gamma_i(b)) \cdot \delta_i(b) u_i(b) + \gamma_i(b) \cdot \eta_i(b)\frac{\mathrm{sw}(b,U)}{n} \\ &\geq \gamma_i(b)\frac{\mathrm{sw}(b,U)}{n}.\end{aligned}$$

Then, following through the same rearrangements as in the proof Lemma 12.3.1 and summing over $N_{a \succ b}$ (shorthand for $\{i : a \succ_{\pi_i} b\}$), we conclude the proof:

$$\begin{aligned}&\eta^* \cdot \mathrm{sw}(a,U)\frac{|N_{a\succ b}|}{n}\frac{\gamma_i(a)}{\gamma_i(b)} + \delta^* \cdot \frac{1-\gamma_i(a)}{\gamma_i(b)}\mathrm{sw}(a,U) \geq \frac{|N_{a\succ b}|}{n}\mathrm{sw}(b,U) \\ \implies &\frac{\mathrm{sw}(b,U)}{\mathrm{sw}(a,U)} \leq \delta^* \cdot \frac{1-\gamma_i(a)}{\gamma_i(b)}\frac{n}{|N_{a\succ b}|} + \eta^* \cdot \frac{\gamma_i(a)}{\gamma_i(b)} \leq \frac{\delta^*(1-\gamma_{min})}{\gamma_{min}}\frac{n}{|N_{a\succ b}|} + \frac{\eta^*}{\gamma_{min}}.\end{aligned}$$

□

Now, we conclude the robust versions of our original distortion upper bounds. Lemma 12.5.4 implies Corollary 12.5.5 just as Lemma 12.3.1 implied Corollary 12.3.4. Similarly, Lemma 12.5.4 implies Corollary 12.5.6 just as Lemma 12.3.1 implied Theorem 12.3.3.

**Corollary 12.5.5.** *For all voting rules $f$, all $\delta^*, \eta^* \geq 1$ and PS-matrices $\Gamma \in [0,1]^{n\times m}$ with $\gamma_{min} > 0$,*

$$dist^{\delta^*,\eta^*}(f,\Gamma) \leq \frac{\delta^*(1-\gamma_{min})}{\gamma_{min}\,\kappa_f} + \frac{\eta^*}{\gamma_{min}}.$$

**Corollary 12.5.6.** *For all uncovered set rules $f$, all $\delta^*, \eta^* \geq 1$ and PS-matrices $\Gamma \in [0,1]^{n\times m}$ with $\gamma_{min} > 0$,*

$$dist^{\delta^*,\eta^*}(f,\Gamma) \leq \left(\frac{2\delta^*(1-\gamma_{min})}{\gamma_{min}} + \frac{\eta^*}{\gamma_{min}}\right)^2.$$

**A remark about tightness.** Most of the upper bounds derived from Lemma 12.3.1 were tight for constant PS-vectors $\boldsymbol{\gamma} = \gamma\mathbf{1}$ (Section 12.3.3). Thus, one may wonder whether the upper bounds in this subsection are likewise tight for constant PS-matrices. This question merits formal theoretical treatment, because one must construct a separate lower bound for each voting rule, as in Section 12.3.3. However it does seem that tightness should hold via the following simple construction: let $a'$ be the election winner. Then, construct a profile in which, for all voters $i$, $\delta_i(a') = \delta^*$ and $\eta_i(a') = \eta^*$ and $\delta_i(a) = \eta_i(a) = 1$ for all other $a \neq a'$. Intuitively, this construction allows $a'$ to win the election with the smallest true utility possible, and should yield lower bounds corresponding to those in Section 12.3.3 as follows: if a lower bound on the standard model is, for some functions $h, g$, of the form $h(g(m)\cdot(1-\gamma)/\gamma+1)$, then it should be $h(g(m)\cdot\delta^*(1-\gamma)/\gamma+\eta^*)$ in the generalized model.

## 12.6 Discussion

A key contribution of our work is to establish *cultivating voters' public spirit* as a new approach to increasing the welfare of democratic decision-making — an approach which can be operationalized via publicly-palatable interventions like deliberation. In the introduction, we discussed why increasing the welfare of voting outcomes is a pressing goal; however, regardless of how pressing one believes this goal to be, our results suggest that in many senses, interventions that promote public spirited voting can only help.

Of course, these results arise from a theoretical model, so their practical implications depend on how our model may capture — or fail to capture — reality. On this note, our robustness results in Section 12.5.2 cover a wide range of plausible behavioral deviations: they allow voters to, e.g., assess their utilities on different scales, overestimate their own utilities compared to others', underestimate the interests of certain groups due to biases, apply different levels of public spirit to different alternatives, or even more coarsely, just maintain a ranking over alternatives rather than any sense of these quantities (this corresponds to arbitrary errors in utilities and social welfares). Our results in Section 12.5.1 also allow for participants who exhibit *no* public spirit; however, a key issue we sidestep is the case where some participants not only lack public spirit, but are actually adversarial to the process. We address this in part 2 of the future work below.

### 12.6.1 Future work

In addition to the theoretical directions identified below, we remark that our work motivates further experiments studying how voters' public spirit changes over the course of deliberation. In turn, with a more detailed understanding of the structure of voters' deviations from our model, one can get more fine-grained robustness bounds than we achieve in Section 12.5.2.

**1 Identifying the *optimal* deterministic voting rule.** Although we exactly characterize the distortion of several popular deterministic voting rules, this work leaves open: 'what is the *welfare-optimal* deterministic voting rule when voters are public-spirited?' More precisely, observe that our lower bound on Plurality's distortion (Proposition 12.3.15) can be directly extended to prove that *any* deterministic voting rule must suffer distortion at least $2(1-\gamma)/\gamma + 1$ when $\boldsymbol{\gamma} = \gamma\mathbf{1}$.[1] It is not clear whether this lower bound is tight, however, because no voting rule we study has distortion matching this bound. A natural extension of this work, then, would be to prove a tight lower bound on the distortion all deterministic voting rules under public-spirited voting, and find a voting rule whose distortion matches this bound.

**2 Strategic voters among public spirited voters.** As always, in our setting there is the potential for manipulation — perhaps more so here because some voters are prioritizing the collective rather than acting in rational self-interest. The possibility of some voters being strategic opens

[1] Let $f$ be any deterministic voting rule, and consider the instance used to prove Proposition 12.3.15 with $\epsilon = 0$ and $m = 2$. Let the two alternatives be $a, b$, corresponding to the utilities of groups (A) and (B) in the instance). Then, exactly half of agents rank $a \succ b$ and half rank $b \succ a$. Since $a$ and $b$ are symmetric from the perspective of $f$, wlog let $f$ choose $a$. This gives distortion exactly $2(1-\gamma)/\gamma + 1$.

several questions, such as: 'Does public spirit among most voters make the voting process more or less robust to a few manipulators?' and 'Given that the presence of strategic voters might pose a risk to others, how might voters who would otherwise intend to be public-spirited respond?'

**3 Sufficient conditions for (approximate) instance-wise monotonicity.** While in many respects, our results suggest that increased public spirit is beneficial, Theorem 12.4.11 shows an extremely fundamental impossibility: that in general, public spirit may not help on an instance-by-instance basis. This begs the question: can we establish sufficient conditions on instances — ideally which are roughly detectable in practice — under which we *can* be certain that increasing the public spirit will improve outcomes? Moreover, even if we cannot hope for *exact* monotonicity, can we show approximate notions, e.g., in which the social welfare increases up to bounded fluctuations?

**4 Extensions to other notions of social welfare.** In this paper, we assume that public-spirited voters determine how positively an alternative impacts society according to its *utilitarian* social welfare. However, voters might just as easily take an *egalitarian* perspective, thus quantifying an alternative's social welfare by how it affects the person it benefits the *least*. Even further, there is no guarantee that public spirited voters apply the *same* priorities when assessing the social welfare. These points open questions such as, if voters are public-spirited *but quantify the social good via different objectives*, does public spirit still increase the welfare of the outcome?

**5 Other collective decision mechanisms.** Our results identify public-spirited voting behavior as a powerful, practically-motivated beyond-worst-case assumption. We have demonstrated this specifically for deterministic voting mechanisms where voters express preferences as complete rankings. However, there are many other well-studied collective decision mechanisms — e.g., randomized voting rules, approval voting, multi-winner elections, liquid democracy, participatory budgeting — that could potentially benefit from public spirit, too. To initiate the study of public spirit in other mechanisms, we remark that all the aforementioned mechanisms can be analyzed in the same utilitarian social welfare framework: one needs only to specify a model of how voters translate their underlying utilities into ballot responses — analogous to our Equations (12.1) and (12.2) — that allows voters to weigh their own interests against the common good.

# 13

# Extensions to Participatory Budgeting

*The Distortion of Public-Spirited Participatory Budgeting* [43]
Mark Bedaywi, Bailey Flanigan, Mohamad Latifian, & Nisarg Shah
*submitted* 2024

## 13.1 Introduction

Governments at all scales regularly face the question: *With a limited budget, which public-good projects — e.g., building bike paths or installing streetlamps — should they fund?* To make such decisions democratically, governments are increasingly using *participatory budgeting* (PB), in which constituents vote on which projects they would like to see funded. In PB, the government supplies a budget $B$ and a list of $m$ potential projects $a \in \{1, \dots, m\}$ with corresponding costs $c_1, \dots, c_m$. Voters submit their preferences via *ballots*, and then these ballots are aggregated via an *aggregation rule* to select a set of projects to be funded, whose total cost must be at most $B$. PB is now used all over the world to decide allocations of public funds[1] [89, 225, 275].

When designing the PB process described above, one goal that many consider important is ensuring that the ultimate allocation of funds has high societal benefit. As have many others (e.g., Benadè et al. [45]), we formalize the "societal benefit" of an allocation by its *utilitarian social welfare*: the total utility it gives to all voters combined. In using this measurement, we adopt the standard model of latent additive utilities: each voter $i$ has *utility* $u_i(a) \in \mathbb{R}_{\geq 0}$ for each project $a$, and their total utility for a set of projects $S$ being funded is $u_i(S) = \sum_{a \in S} u_i(a)$. Then, the *social welfare* of $S$ is equal to $\mathrm{sw}(S) = \sum_{i \in N} u_i(S)$.

If voters' utilities were observable, choosing the maximum-welfare allocation would amount to

[1] See https://en.wikipedia.org/wiki/List_of_participatory_budgeting_votes for a list of use cases.

solving the knapsack problem. However, in practice voters' preferences can only be elicited more coarsely through *ballots*. For example, popular ballot formats in PB include *rankings by value*, where voters are asked to rank the individual projects, or *$k$-approval votes*, where voters are asked to approve their favorite $k$ alternatives. It is not hard to see that such ballot formats lose far too much information about voters' utilities to allow deterministic selection of a high-welfare solution: suppose there are two projects, $a$ and $b$, both costing $B$ so we must simply choose one or the other to fund. If half the population has utilities $1, 0$ for $a, b$ and the other half has utilities $0, 1000$ for $a, b$ (so the welfare of $b$ is 1000 times that of $a$). Although $b$ has far higher social welfare, any ordinal ballot format where voters only compare sets of alternatives will produce symmetric ballots, leading to any deterministic aggregation rule—i.e., any deterministic mapping from $n$ ballots to an allocation funds—to choose (without loss of generality) $a$; the best thing we can do here is to randomize uniformly over the two options.

This example illustrates a prohibitive impossibility: in the worst case, *any* deterministic aggregation rule over any ordinal PB ballot format will select an outcome with arbitrarily sub-optimal social welfare, simply because these PB ballot formats do not contain enough information about voters' cardinal preferences. Formally, this sub-optimality is captured with the *distortion*: the worst-case (over possible latent utilities) ratio of the best possible social welfare that of the outcome. Existing work sidesteps this impossibility by assuming that each voter's utilities are restricted to add up to 1 [45]. Although this permits bounded distortion in theory, it remains unclear whether these bounds apply in practice: For example, this assumption may not hold in the likely case that the public goods will more greatly impact lower-income constituents.

Fortunately, recent work by Flanigan et al. [134] offers a source of hope: under unrestricted utilities, they achieve low distortion in single-winner elections by leveraging the idea that voters may be *public-spirited*: when casting their ballots, voters consider others' interests in addition to their own. While it is not clear that such behavior would be reliably present in the wild, as Flanigan *et al* point out, research suggests that public spirit can be cultivated via democratic deliberation — *a practice that is already commonplace in PB elections* [89, 225]. The possibility of cultivating public spirit among PB participants motivates our main research question:

> **Question:** *If voters are public-spirited, do there exist ballot formats and associated aggregation rules that achieve small distortion, without any restrictions on voters' utilities?*

An affirmative answer to this question would suggest a practicable approach — democratic deliberation — to achieving higher-welfare outcomes in PB elections. In the process of pursuing this question, we close an open question for the single-winner voting setting left open by Flanigan et al. [134], and introduce a new ballot format which makes better use of voters' public spirit to break an important distortion barrier in the PB context. We overview these contributions below.

### 13.1.1 Results and contributions.

We study the distortion of PB with public-spirited participants by adopting Flanigan et al. [134]'s model of public-spirited voting, extending it as needed to new ballot formats. In this model, each

voter $i$ evaluates each alternative $a$ not just according to her her own utility $u_i(a)$, but by her *public-spirited (PS) value*: the convex combination of her utility for $a$ and its social welfare. This convex combination is weighted by her *public spirit level* $\gamma_i \in [0, 1]$, where higher $\gamma_i$ means she more strongly weighs the social welfare. As in Flanigan et al. [134], our distortion bounds, summarized in Tables 13.1 and 13.2, are parameterized by $\gamma_{\min} = \min_i \gamma_i$, the minimum public spirit level of any voter.

**Contribution 1: Tight bounds for single-winner voting with *ranking by value* ballots.** Building directly from Flanigan et al. [134], we begin by studying *ranking-by-value* ballots, where voters rank the alternatives in $[m]$ in decreasing order of their public-spirited values. Before analyzing the performance of this ballot format in the PB context, we first study it in the single-winner context — a significant strict restriction of the PB setting where all projects cost $B$. We begin with the single-winner setting because, although this is precisely the setting studied by Flanigan *et al*, there remain two important open questions, which we close in order to build upon their answers later.

*1.1 What is the best distortion achievable by any **deterministic** voting rule over ranking-by-value ballots?* The lowest-distortion voting rule identified by Flanigan *et al* is Copeland, achieving constant (in $m$) distortion of exactly $(1 + {}^{2(1-\gamma_{\min})}/_{\gamma_{\min}})^2$; in contrast, their results lead to a lower bound on any deterministic rule of at most $1+{}^{2(1-\gamma_{\min})}/_{\gamma_{\min}}$. To close this gap, we design a nontrivial construction to prove a stronger lower bound. This lower bound is tight to known upper bounds in its dependency on both $m$ and $\gamma_{\min}$, thereby closing the question of what level of distortion is possible in single-winner public-spirited deterministic voting. This analysis reveals that in fact, the rule Copeland is optimal (except when $m$ is small relative to $1/\gamma_{\min}$, in which case Plurality is optimal).

*1.2 What is the best distortion achievable by any **randomized** voting rule over ranking-by-value ballots?* Flanigan et al. [134] did not study randomized voting rules at all. Thus, here we must prove lower and upper bounds anew. Our lower bound arises from the same construction as described above. We identify a novel optimal voting rule for this case, whose distortion matches our lower bound in both $m$ and $\gamma_{\min}$. Its distortion is $\Theta(\min\{m, 1/\gamma_{\min}\})$, the best distortion possible in single-winner public-spirited randomized voting.

**Contribution 2: Distortion bounds for PB with *ranking-by-value* ballots.** Next, we generalize our results from the single-winner to the PB setting, again with the goal of identifying optimal aggregation rules and proving matching lower bounds.

*2.1 Lower bounds.* First, we extend our lower bounds from the single-winner case to prove that in PB, the distortion of any deterministic rule must be in $\Omega(m/\gamma_{\min})$, and that of any randomized rule must be in $\Omega(\log m)$.

*2.2 Upper bounds via reductions from single-winner voting to PB.* For both deterministic and randomized rules, we prove our upper bounds via direct reductions relating any voting rule's dis-

tortion in the single-winner setting to its performance in the PB setting. Such a reduction was previously known for deterministic rules, incurring a factor of at most $m$ in the distortion from single-winner to PB. Via this method, we find that COPELAND is again optimal as before, with distortion matching our lower bound in both its dependency on $m$ and $\gamma_{\min}$.

For randomized rules, no such reduction existed, so we extend the previous reduction to the randomized case. Via this reduction, we incur a factor of order at most $\log m$ in the distortion from single-winner to PB. Then, we apply this reduction to give an upper bound on our single-winner randomized rule above. In the PB setting, this voting rule achieves distortion with optimal dependency on $m$, and within a factor of at most $\gamma_{\min}$ of optimal dependency on $\gamma_{\min}$.

**Contribution 3: Approval-style ballot formats.** A practically important type of ballot format in the PB context are $k$-approval ballots. In our model, this means voters submit the set of $k$ alternatives for which they have the highest public-spirited values. Due to their practical importance, we now repeat our analysis for this entirely new ballot format. Our first key finding is that if $k$ is larger than one or more maximal budget-feasible sets of projects, the distortion can be unbounded because voters' approval sets can be budget-infeasible, thus giving us no information about their preferences over budget-feasible sets. This is clearly avoided when $k = 1$; accordingly, we give matching lower and upper bounds on the distortion of 1-approval ballots of $\Theta(m^2/\gamma_{\min})$

The issue of $k$-approval ballots permitting budget-infeasible approval sets motivates another ballot format often considered in the PB literature — *knapsack* ballots. Knapsack ballots again allow each voter to approve a set of items, but *only if that set is budget-feasible.* Perhaps the most striking finding in our analysis of knapsack ballots is that while they have at best exponential distortion $\Omega(2^m/\sqrt{m})$ under the unit sum utilities assumption, we show via a novel approach of comparing entire subsets of alternatives that under public-spirited voting, these ballots have polynomial distortion of at most order $O\left(m^3\right)$.

**Contribution 4: Ballot formats that breaks the $m$ distortion barrier.** In the previous sections, our lower bounds show us that across the ballot formats we study — plus two others whose analysis we relegate to the appendix — *no ballot format can achieve distortion with sublinear dependency on m with deterministic aggregation rules (which is the practical case of interest).* This barrier also exists under the unit-sum utilities assumption [45]. Motivated by this, we ask: *is public spirit powerful enough to permit a **any** practical ballot format to break this barrier?*

We find that in fact, the answer is yes. We define a new, simple ballot format, which pre-partitions the alternatives into (at most $m$) feasible sets of alternatives, and requires voters to rank them rather than the individual alternatives. We show that by carefully bundling the alternatives in the ballot, we can get $O\left(\sqrt{m}/\gamma_{\min}^2\right)$ distortion. If a second stage of elicitation is allowed, we show that the distortion can be further reduced to $O\left(\log m/\gamma_{\min}^4\right)$ using this ballot format. These results show that our new ballot format is significantly more efficient, while also being thrifty and practical. These results also point to the exciting open question of whether any ordinal ballot that only asks voters to compare polynomially many sets of alternatives can reduce the distortion all the way down to a constant.

| | | Public-Spirit | Unit-Sum |
|---|---|---|---|
| SW | Deterministic | $\Theta\left(1/\gamma_{\min} \cdot \min\{m, 1/\gamma_{\min}\}\right)$ | $\Theta(m^2)$ |
| | Randomized | $\Theta\left(\min\{m, 1/\gamma_{\min}\}\right)$ | $\Theta(\sqrt{m})$ |
| PB | Deterministic | $\Omega\left(m/\gamma_{\min}\right)$, $O\left(m/\gamma_{\min} \cdot \min\{m, 1/\gamma_{\min}\}\right)$ | $\Theta(m^2)$ |
| | Randomized | $\Omega\left(\log m\right)$, $O\left(\min\{m, (\log m)/\gamma_{\min}\}\right)$ | $\Omega(\sqrt{m})$, $O\left(\sqrt{m}\log m\right)$ |

Table 13.1: Asymptotic (in $m, \gamma_{\min}$) distortion bounds for rankings-by-value, comparing results for Single-winner (SW) and Participatory Budgeting (PB) ballots. The unit-sum results are derived in Benadè et al. [45] and are included for comparison.

| | Public-Spirit | Unit-Sum |
|---|---|---|
| *k-approvals* ($k > 1$) | $\infty$ | $\infty$ |
| 1-*approval* | $\Theta\left(m^2/\gamma_{\min}\right)$ | $\Theta(m^2)$ |
| *Knapsack* | $\Omega\left(m/\gamma_{\min}\right)$, $O\left(m^3/\gamma_{\min}^2\right)$ | $\Omega(2^m/\sqrt{m})$, $O\left(m2^m\right)$ |
| *Single Round rbp* | $O\left(\sqrt{m}/\gamma_{\min}^2\right)$ | $\Omega(m^2)$ |
| *Two Round rbp* | $O\left((\log m)/\gamma_{\min}^4\right)$ | $\Omega(m^2)$ |

Table 13.2: Asymptotic (in $m, \gamma_{\min}$) deterministic distortion bounds across ballot formats other than ranking-by-value. The colored rows indicate new ballots introduced in this paper. The unit-sum results are derived in Benadè et al. [45] and are included for comparison.

### 13.1.2 Related work

Our work directly builds on the works of Benadè et al. [45], who analyzed distortion in PB, and Flanigan et al. [134], who introduced the public-spirit model. Our results eliminate the unit-sum assumption made in the former work, and generalize the latter work from single-winner elections (selecting a single alternative) to the more general problem of PB, where multiple alternatives are selected subject to a budget constraint and there are multiple reasonable ballot formats to consider.

Procaccia and Rosenschein [231] introduce the distortion framework in single-winner elections under the unit-sum assumption. We now know that the best distortions achievable by deterministic and randomized rules for this special case are $\Theta(m^2)$[67, 68] and $\Theta(\sqrt{m})$ [53, 103], respectively. Optimal distortion bounds have also been identified for $k$-committee selection [52, 68], which still remains a special case of PB. As an alternative to the unit-sum assumption, unit-range utilities or metric costs have been studied [24, 117], but all of these place some restriction on voter preferences. For further details, we suggest the survey of Anshelevich et al. [25].

Multiple approaches other than distortion have been studied for PB. The axiomatic approach has been used to identify aggregation rules satisfying desirable axioms such as various monotonicity

properties Baumeister et al. [37], Rey et al. [239], Talmon and Faliszewski [263]. Another important consideration in PB is whether the allocation of funds is fair with respect to (groups of) voters [59, 113, 227]. For further details, we suggest the survey of Rey and Maly [238] and the book chapter of Aziz and Shah [30].

## 13.2 Model

We introduce the most general framework of participatory budgeting (PB) first, and later introduce single-winner and multiwinner voting as its special cases.

There is a set $N$ of $n$ *voters* and a set $A$ of $m$ *alternatives* (projects). We denote voters by $i, j$ and alternatives by $a, b$. There is a total budget of $B$, which is normalized to 1 without loss of generality, and a cost function $c : A \to [0, 1]$, where $c(a)$ is the *cost* of $a$. Slightly abusing notation, we use $c(S) = \sum_{a \in S} c_a$ as the total cost of alternatives in $S$. Let $\mathcal{F} = \{S \subseteq A : c(S) \leq B\}$ be the set of *budget-feasible* subsets of alternatives. The goal is to select such a budget-feasible subset by eliciting and aggregating voter preferences.

Special cases. We note that *$k$-committee selection* is a special case of PB, where the cost of each alternative is $1/k$, so $\mathcal{F}$ consists of all subsets of alternatives of size $k$. We use "$k$-committee rule" to refer to a rule for this special case. Further, *single-winner selection* is a special case of $k$-committee selection where $k = 1$; we use "single-winner rule" to refer to a rule for this special case.

Utilities. Each voter $i \in N$ has a *utility* for each alternative $a \in A$ denoted by $u_i(a) \in \mathsf{R}_{\geq 0}$. Together, these utilities form a utility matrix $U \in \mathsf{R}_{\geq 0}^{n \times m}$. Define the *social welfare* of an alternative $a \in A$ w.r.t. utility matrix $U$ as $\mathsf{sw}(a, U) = \sum_{i \in N} u_i(a)$; for a subset of alternatives $S \subseteq A$, define $\mathsf{sw}(S, U) = \sum_{a \in S} \mathsf{sw}(a, U)$. We use $\mathsf{sw}(a)$ or $\mathsf{sw}(S)$ when $U$ is clear from context.

PS-values. Following the model introduced by Flanigan et al. [134], we assume that each voter $i \in N$ has a *public spirit (PS) level* $\gamma_i \in [0, 1]$ and together these PS-levels form the PS-vector $\vec{\gamma} \in [0, 1]^n$. Our results depend on the minimum public spirit level of the voters $\gamma_{\min} \triangleq \min_{i \in N} \gamma_i$.

Each voter submits her preferences according to not her personal utilities, but her *PS-values*, which she computes by taking a $\gamma_i$-weighted convex combination of her personal utilities and the average utility of all voters. Formally, the *PS-value* of voter $i$ for alternative $a$ is

$$v_i(a) = (1 - \gamma_i) \cdot u_i(a) + \gamma_i \cdot \mathsf{sw}(a)/n.$$

Together, these PS-values form the *PS-value matrix* $V_{\vec{\gamma},U} \in \mathsf{R}_{\geq 0}^{n \times m}$. PS-values are additive across alternatives, so that for each $S \subseteq A$, $v_i(S) = \sum_{a \in S} v_i(a)$.

Note that PS-values have the same scale as utilities because $\mathsf{sw}(a) = \sum_{i \in N} u_i(a) = \sum_{i \in N} v_i(a)$ for each $a \in A$. We show that this transformation allows us to get rid of the unit-sum assumption ($\sum_{i \in N} u_i(a) = 1, \forall a \in A$) required by much of the prior work [45].

Elicitation. Since it is cognitively burdensome for voters to report numeric PS-values, it is common to elicit their preferences using discrete ballots. Following the model of Benadè et al. [45], a *ballot format* $\mathsf{X} : \mathsf{R}^m_{\geq 0} \times [0,1]^m \to \mathcal{L}_\mathsf{X}$ turns every PS-value function into a "vote", which takes values from a (usually finite) set $\mathcal{L}_\mathsf{X}$, sometimes using the cost function over the alternatives. Under this ballot format, each voter $i$ submits the vote $\rho_i = \mathsf{X}(v_i)$; together, these votes form the *input profile* $\vec{\rho} = \{\rho_1, \ldots, \rho_n\}$. We use $V_{\vec{\gamma},U} \rhd_\mathsf{X} \vec{\rho}$ to indicate that PS-value matrix $V_{\vec{\gamma},U}$ induces input profile $\vec{\rho}$ under ballot format $\mathsf{X}$. Alternatively, we say that $\vec{\rho}$ is consistent with $V_{\vec{\gamma},U}$. We omit $\mathsf{X}$ when it is clear from the context.

We study four ballot formats also studied by Benadè et al. [45], namely rankings by value, rankings by value for money, knapsack votes, and threshold approval votes, as well as a new ballot format we introduce, namely ranking of predefined bundles; we define them in their respective sections.

Aggregation Rules. Let $\Delta(\mathcal{F})$ be the set of all distributions over $\mathcal{F}$. A (randomized) *aggregation rule* $f : \mathcal{L}^n_\mathsf{X} \times [0,1]^m \to \Delta(\mathcal{F})$ for ballot format $\mathsf{X}$ takes an input profile $\vec{\rho} \in \mathcal{L}^n_\mathsf{X}$ and a cost function over alternatives $c \in [0,1]^m$ as input, and outputs a distribution over feasible sets of alternatives in $\mathcal{F}$. We say that $f$ is deterministic if its output always has singleton support.

Distortion. The *distortion* measures the efficiency of a voting system, composed of a ballot format and an aggregation rule for that ballot format. For a ballot format $\mathsf{X}$ and minimum public spirit level $\gamma_{\min} \in [0,1]$, the distortion of an aggregation rule $f$ on input profile $\vec{\rho}$ in format $\mathsf{X}$ and cost function $c$ is the following worst-case ratio:

$$\mathrm{dist}_\mathsf{X}(f, \vec{\rho}, c) = \sup_{\substack{U, \vec{\gamma}: \\ \min_{i \in N} \gamma_i = \gamma_{\min}, \\ V_{\vec{\gamma},U} \rhd \vec{\rho}}} \frac{\max_{S \in \mathcal{F}} \mathrm{sw}(S, U)}{\mathsf{E}_{S' \sim f(\vec{\rho})} \mathrm{sw}(S', U)}.$$

The (overall) distortion of $f$ is obtained by taking the worst case over all instances $(\vec{\rho}, c)$ and all $n$:

$$\mathrm{dist}_\mathsf{X}(f) = \sup_{n \geq 1} \sup_{\vec{\rho} \in \mathcal{L}^n_\mathsf{X},\ c \in [0,1]^m} \mathrm{dist}_\mathsf{X}(f, \vec{\rho}, c).$$

The resulting distortion is a function of $m$ and $\gamma_{\min}$; we fix arbitrary $m \geq 2$ and $\gamma_{\min} \in (0,1]$ throughout the paper. We are interested in the lowest distortion enabled by each ballot format, across all aggregation rules for that ballot format. This is a measure of the usefulness of the information contained in the ballot format for social welfare maximization.

Supporting results. Let us state a lemma that we use throughout the paper. This is a simple generalization of Lemma 3.1 of Flanigan et al. [134]; the proof is in Appendix H.3.1.

**Lemma 13.2.1.** *Let $A_1, A_2 \subseteq A$ be two arbitrary subsets of alternatives. Fix any $\alpha \geq 0$ and define $N_{A_1 \succ A_2} = \{i \in N : \alpha \cdot v_i(A_1) \geq v_i(A_2)\}$. Then:*

$$\frac{sw(A_2)}{sw(A_1)} \leq \alpha \cdot \left( \frac{1 - \gamma_{\min}}{\gamma_{\min}} \frac{n}{|N_{A_1 \succ A_2}|} + 1 \right).$$

Finally, for comparison, we remark that for all ballot formats we consider, when there is no public spirit and the utilities are unrestricted, all deterministic voting rules have unbounded distortion and the randomized rules have at best $m$ distortion (Appendix H.3.2).

## 13.3 Single-Winner Voting

As mentioned before, single-winner voting can be seen as a special case of participatory budgeting problem in which all the alternatives have a cost equal to the budget, so only a single alternative can be selected. Flanigan et al. [134] analyze the distortion of various deterministic voting rules for this single-winner case under public-spirited voting. In this section we give lower bounds on the distortion of any deterministic and randomized voting rule in this setting, and also design rules that match the lower bound. For the results in this section, we consider, as do Flanigan et al. [134], the prominent ballot format of *rankings by value* (rbv). In this ballot format, each voter ranks the alternatives in a non-increasing order of her values for them. Formally, $\mathcal{L}_{\mathsf{rbv}}$ is the set of all rankings of the alternatives, and each voter $i$ submits a ranking $\rho_i \in \mathcal{L}_{\mathsf{rbv}}$ such that for every $a, b \in A$ with $v_i(a) > v_i(b)$, we have $a \succ_{\rho_i} b$ (i.e., $a$ appears above $b$ in the ranking $\rho_i$); the voter can break ties among equal-PS-valued alternatives arbitrarily.

### 13.3.1 Lower bounds

We start by proving the lower bound for the *deterministic* rules.

**Theorem 13.3.1** (Lower Bound - Deterministic)**.** *Any deterministic single-winner voting rules $f$ with ranked preferences has distortion*

$$dist_{rbv}(f) \geq 1 + 2\frac{1-\gamma_{\min}}{\gamma_{\min}} \cdot \frac{m^2}{2\gamma_{\min} + \gamma_{\min} m^2 + (2-3\gamma_{\min})m} \in \Omega\left(\frac{1}{\gamma_{\min}} \cdot \min\left\{ m, \frac{1}{\gamma_{\min}} \right\}\right).$$

*Proof Sketch.* Our construction consists of $m$ types of voters, equally distributed with $n/m$ voters of each type. Let $N_k$ be the set of voters of type $k$. Suppose each voter type votes as follows,

$$\begin{array}{cccccccccc}
N_1 & : & a_1 & \succ & a_2 & \succ & \dots & \succ & a_{m-1} & \succ & a_m \\
N_2 & : & a_2 & \succ & a_3 & \succ & \dots & \succ & a_m & \succ & a_1 \\
\vdots & & & & & & & & & & \\
N_{m-1} & : & a_{m-1} & \succ & a_m & \succ & \dots & \succ & a_{m-3} & \succ & a_{m-2} \\
N_m & : & a_m & \succ & a_1 & \succ & \dots & \succ & a_{m-2} & \succ & a_{m-1}
\end{array}$$

so that $N_i$ prefers alternative $a_i$ most, and cycles through the rest. We use this instance to prove the lower bound. □

The full proof can be found in Appendix H.4.1. We include the instance that gives this lower bound here, because versions of it will be used to prove lower bounds throughout the paper. Using a similar instance, we can prove a lower bound on the distortion of any *randomized* voting rule. The full proof of this theorem is in Appendix H.4.2.

**Theorem 13.3.2** (Lower Bound - Randomized)**.** *Any randomized single-winner voting rules $f$ with ranked preferences has distortion*

$$dist_{rbv}(f) \in \Omega\left(\min\left\{m, \frac{1}{\gamma_{\min}}\right\}\right).$$

### 13.3.2 Upper Bounds

In this section we focus on designing voting rules with distortion matching the lower bounds. First, in the deterministic case, we give a deterministic voting rule that directly combines upper bounds from Flanigan et al. [134].

**Corollary 13.3.3** (Upper Bound - Deterministic)**.** *The deterministic single-winner rule $f_{PC}$ that runs Plurality if $m \leq 1/\gamma_{\min}$ and Copeland otherwise, has distortion at most*

$$dist_{\mathsf{rbv}}(f_{PC}) \leq \min\left\{\frac{m}{\gamma_{\min}} - m, \left(\frac{2}{\gamma_{\min}} - 1\right)^2\right\} \in O\left(\frac{1}{\gamma_{\min}} \cdot \min\left\{m, \frac{1}{\gamma_{\min}}\right\}\right).$$

*Proof.* Per Proposition 3.5 and Theorem 3.3 of Flanigan et al. [134] respectively, $\mathsf{dist}_{\mathsf{rbv}}(f_{\mathsf{Plurality}}) \leq m/\gamma_{\min} - m$ and that of $\mathsf{dist}_{\mathsf{rbv}}(f_{\mathsf{Copeland}}) \leq (2/\gamma_{\min} - 1)^2$. Thus, by defining the rule that chooses the Plurality winner when $m \leq 1/\gamma_{\min}$ and the Copeland winner otherwise, we can guarantee achievement of the desired distortion. □

Now, we endeavor to find an optimal randomized voting rule. Since Flanigan et al. [134] does not study randomized rules, we cannot apply their bounds. Here, we turn to *maximal lottery*, a randomized voting rule that was originally proposed by Kreweras [181] and rediscovered numerous times in the social choice literature [121, 122, 183, 241]. Curiously, Charikar et al. [76] recently use this rule to derive a breakthrough result in the related setting of metric distortion. There are various alternative formulations of this rule, but the one most useful to us is the following.

**Definition 13.3.4** (Maximal Lottery)**.** *Define the domination graph to be a directed graph $G$ with alternatives in $A$ as the vertices and an edge between every pair of vertices, oriented so that if $a$ beats $b$ in a pairwise election, then the edge goes from $a$ to $b$. In the case of ties, we may pick orientation arbitrarily. The maximal lottery rule returns a distribution $p$ over the vertices such that for any vertex $v \in A$, the probability of picking $v$ or a vertex adjacent to $v$ is at least $1/2$. The existence of such a distribution can be inferred from, e.g., Farkas' lemma (see Theorem 2.4 of Harutyunyan et al. [162]).*

**Theorem 13.3.5** (Upper Bound - Randomized)**.** *There exists a randomized single-winner voting rule $f$ with distortion at most*

$$dist_{\mathsf{rbv}}(f) \leq \min\{m, 2(2/\gamma_{\min} - 1)\} \in O\left(\min\left\{m, \frac{1}{\gamma_{\min}}\right\}\right).$$

*Proof.* To match our piecewise lower bound, we must again decide between two voting rules: the voting rule which chooses an alternative uniformly at random (thereby achieving $m$ distortion) and the maximal lottery rule, which we prove has distortion at most $2/\gamma_{\min} - 1$.

Indeed, let $a^*$ be the optimal alternative. If we pick $a^*$ or an alternative $b$ that beats $a^*$ in a pairwise election, by Lemma 13.2.1 we get distortion:

$$\frac{\mathrm{sw}(a^*)}{\mathrm{sw}(b)} \leq 2\frac{1-\gamma_{\min}}{\gamma_{\min}} + 1.$$

Let the set of such alternatives be $A' = \{\, b \in A : |\, \{\, i \in N : b \succ_i a^* \,\}\, | \geq n/2 \,\}$. Then, the distortion of our rule is:

$$\begin{aligned}\frac{\mathrm{sw}(a^*)}{\sum_{a\in A} p(a)\mathrm{sw}(a)} &\leq \frac{\mathrm{sw}(a^*)}{\sum_{a\in A'} p(a)\mathrm{sw}(a)} \leq \frac{\mathrm{sw}(a^*)}{(\min_{a\in A'} \mathrm{sw}(a)) \sum_{a\in A'} p(a)} \\ &\leq 2\frac{\mathrm{sw}(a^*)}{\min_{a\in A'} \mathrm{sw}(a)} \leq 4\frac{1-\gamma_{\min}}{\gamma_{\min}} + 2 = \frac{4}{\gamma_{\min}} - 2.\end{aligned}$$

□

Importantly, because $\gamma_{\min}$ is unobservable to the voting rule, *implementing* these piecewise voting rules (for both the randomized and deterministic cases) is not quite practicable, ut the intuition — that for small $m$, Plurality is desirable, and for large $m$, Copeland is better — is.

## 13.4 Rankings by Value

We now move on to the more general setting of participatory budgeting (PB). To begin with, we examine how powerful the same rankings by value ballot format is for PB. Note that while voters still rank individual alternatives by value, the fact that a (feasible) set of alternatives can be funded can significantly affect the power of this ballot format.

### 13.4.1 Deterministic Rules

First, we show that for rbv ballots, deterministic rules must incur a distortion at least $(m-1)\gamma_{\min}^{-1}$. The intuition for this bound is as follows: PB is easy when cheap alternatives are always ranked higher than costly ones, there is never any reason to pick the costly alternatives. So, to construct hard instances, have voters rank costly alternatives highly.

**Theorem 13.4.1** (lower bound)**.** *For rankings by value, every deterministic rule $f$ has distortion*

$$dist_{\mathsf{rbv}}(f) \geq \frac{m-1}{\gamma_{\min}} \in \Omega\left(\frac{m}{\gamma_{\min}}\right).$$

Now, we show how to build directly on results from the single-winner case to give optimal rules for the much more general setting of PB. Specifically, to prove upper bounds, in both the deterministic case and the randomized case, we show how to construct a PB rule from any deterministic single-winner rule while losing an only a factor of $m$ on the distortion.

**Lemma 13.4.2** (Single-Winner → PB - Deterministic)**.** *For any $d \geq 1$, any deterministic rule $f$ with distortion $d$ in the single-winner case has distortion $dist_{\mathsf{rbv}}(f) \leq m \cdot d$ in participatory budgeting.*

*Proof.* Fix any instance and let $f$ return the singleton set $\{a\}$. Let $A^*$ be an optimal budget-feasible set. Then,

$$\frac{\mathrm{sw}(A^*)}{\mathrm{sw}(a)} = \sum_{a^* \in A^*} \frac{\mathrm{sw}(a^*)}{\mathrm{sw}(a)} \leq m \cdot \max_{a^* \in A^*} \frac{\mathrm{sw}(a^*)}{\mathrm{sw}(a)} \leq m \cdot d. \qquad \square$$

We now use this lemma to translate known results from the single-winner setting to PB. In single winner elections, Flanigan et al. [134] show that Plurality has distortion at most $m(\gamma_{\min}^{-1}-1)+1$ and Copeland's rule has distortion at most $\left(2\gamma_{\min}^{-1}-1\right)^2$. Plugging these bounds into Lemma 13.4.2, we conclude upper bounds for the PB setting:

**Theorem 13.4.3** (upper bound)**.** *For rankings by value,*

$$dist_{\mathsf{rbv}}(f_{\mathsf{Plurality}}) \leq m^2(\gamma_{\min}^{-1}-1)+m, \text{ and}$$
$$dist_{\mathsf{rbv}}(f_{\mathsf{Copeland}}) \leq m\left(2\gamma_{\min}^{-1}-1\right)^2.$$

*Hence, there exists a deterministic rule $f$ with distortion*

$$dist_{\mathsf{rbv}}(f) \in O\left(\frac{m}{\gamma_{\min}} \cdot \min\left\{m, \frac{1}{\gamma_{\min}}\right\}\right).$$

**Remark 13.4.4.** *Note that there remains a gap between our upper and lower bounds (in Theorem 13.4.3 and Theorem 13.4.1, respectively): Plurality achieves the optimal dependence on $\gamma_{\min}$, Copeland achieves the optimal dependence on $m$, but neither achieves both. Also, the "best" rule in Theorem 13.4.3 is again a piecewise rule that depends on $\gamma_{\min}$ to decide which of plurality and Copeland to execute. However, it is unclear if a $\gamma_{\min}$-agnostic rule can achieve the same (or even a better) distortion bound.*

### 13.4.2 Randomized Rules

**Theorem 13.4.5** (upper bound)**.** *For rankings by value, there exists a randomized rule $f$ with distortion*

$$dist_{\mathsf{rbv}}(f) \leq 4\left(\frac{2}{\gamma_{\min}}-1\right) \cdot \left(\lceil \log_2(m) \rceil + 1\right) \in O\left(\frac{\log(m)}{\gamma_{\min}}\right).$$

To prove this bound, we will derive another general-purpose reduction — this time for randomized rules — from PB to $k$-committee selection (Lemma 13.4.6), and then from $k$-committee selection to single-winner selection (Lemma 13.4.7). The first will suffers $O(\log m)$ overhead; the latter suffers none (asymptotically). To apply this reduction, we want to plug in bounds on randomized single-winner rules; unfortunately, no such results exist in the public spirit model.

In response, we give in Theorem 13.3.5 a novel randomized single-winner rule with asymptotically optimal (in both $m$ and $\gamma_{\min}$) distortion of at most $4\gamma_{\min}^{-1} - 2$. We now state and prove these results in succession, before applying them to prove Theorem 13.4.5.

**Lemma 13.4.6** (Committee $\to$ PB - Randomized)**.** *Fix any $d \geq 1$. If there exists a randomized $k$-committee selection rule $f_{m',k}$ with distortion at most $d$ for each $m' \leq m$ and $k \in [m']$, then there exists a randomized participatory budgeting rule $f$ for rankings by value with distortion at most $2d \cdot (\lceil \log_2(m) \rceil + 1)$.*

*Proof.* Fix any PB instance. Split the alternatives into buckets $A_0, A_1, \ldots, A_{\lceil \log_2(m) \rceil}$, where $A_0 = \{a \in A : c_a \leq 1/m\}$ and for $i \neq 0$, $A_i = \left\{a \in A : 2^{i-1}/m < c_a \leq 2^i/m\right\}$.

The randomized PB rule $f$ is as follows:

1. Sample $j \in \{0, 1, \ldots, \lceil \log_2(m) \rceil\}$ uniformly.
2. Consider the restricted instance with only the alternatives in $A_j$. That is, with $m' = |A_j|$ and $k = \min(m', \lfloor \frac{m}{2^j} \rfloor)$, use the $k$-committee selection rule $f_{m',k}$ to pick a set of $k$ alternatives and return it.

Let $A^*$ be the optimal budget-feasible subset of the alternatives, $L_j^*$ be the optimal $\lfloor \frac{m}{2^j} \rfloor$-committee of $A_j$, and $L_j$ be the one selected by the $k$-committee rule. For $j \neq 0$, $A^* \cap A_j$ is of size at most $\frac{m}{2^{j-1}}$. That means $\mathrm{sw}(A^* \cap A_j) \leq 2\mathrm{sw}(L_j^*)$ for any $j \neq 0$.

In addition, for $j = 0$, $L_0^* = A_0$ which implies $\mathrm{sw}(A^* \cap A_j) \leq \mathrm{sw}(L_j^*)$. Since the $k$-committee selection rule has distortion of $d$ for any $j$, we have $\mathrm{sw}(L_j^*) \leq d\mathrm{sw}(L_j)$, implying that $\mathrm{sw}(A^* \cap A_j) \leq 2d\mathrm{sw}(L_j)$. Letting $\delta$ be the distribution of the mechanism output, we deduce the desired bound:

$$\begin{aligned}
\mathsf{E}_{L\sim\delta}[\mathrm{sw}(L)] &= \frac{1}{\lceil \log_2(m) \rceil + 1} \sum_{j=0}^{\lceil \log_2(m) \rceil} \mathrm{sw}(L_j) \\
&\geq \frac{1}{\lceil \log_2(m) \rceil + 1} \sum_{j=0}^{\lceil \log_2(m) \rceil} \frac{\mathrm{sw}(A^* \cap A_j)}{2d} \geq \frac{\mathrm{sw}(A^*)}{2d(\lceil \log_2(m) \rceil + 1)}.
\end{aligned}$$

□

Next, we reduce $k$-committee selection to single-winner selection without any asymptotic overhead. The idea is to simply add an alternative to the committee using the single-winner randomized rule, then remove the selected alternative, and repeat the procedure $k$ times.

**Lemma 13.4.7** (Single-Winner $\to$ Committee)**.** *Fix any $k \in [m]$ and $d \geq 1$. If there exists a single-winner rule with distortion at most $d$ for each $m' \leq m$, then there exists a $k$-committee selection rule with distortion at most $d$. The committee selection rule is deterministic if the underlying rule is deterministic, and it is randomized if the underlying rule is randomized.*

The deterministic case is proved in Theorem 8 of Goel et al. [149]. Their key idea is to repeatedly pick alternatives using the single winner rule $k$ times. We extend their result to the randomized case using the same argument. We include the proof in Appendix H.5.2.

Having reduced the PB problem to that of single-winner selection, we now use the novel randomized single-winner rule presented in Theorem 13.3.5 to prove the desired bound.

*Proof of Theorem 13.4.5.* Finally, we apply Lemmas 13.4.6 and 13.4.7 and theorem 13.3.5 to prove Theorem 13.4.5. By Lemma 13.3.5, there exists a randomized single-winner rule (for any $m$) that achieves distortion at most $4\gamma_{\min}^{-1} - 2$. Thus, by Lemma 13.4.7, we get a randomized $k$-committee selection rule (for any $m$ and $k \in [m]$) that achieves distortion at most $4\gamma_{\min}^{-1} - 2$. Finally, by Lemma 13.4.6, we get a randomized PB rule with the desired distortion. □

We prove that this is asymptotically optimal as a function of $m$ in Theorem 13.4.8, thereby proving that our reduction is, in a sense, tight. Deriving the optimal dependence on $\gamma_{\min}$ is left as an open question.

**Theorem 13.4.8** (Lower Bound)**.** *For rankings by value, every randomized rule $f$ has distortion*

$$dist_{\mathsf{rbv}}(f) \geq \ln(m)/2 \in \Omega(\log(m)).$$

*Proof.* Define $k = \lceil \sqrt{m} \rceil - 1$ and partition the alternatives into $k+1$ buckets $A_1, \ldots, A_k, B$ such that for $\ell \in [k]$, $A_\ell$ consists of $\ell$ alternatives with cost $1/\ell$ each, and $B$ includes the rest of the alternatives with cost 1 each. Note that each $A_\ell$ is a feasible subset.

Suppose that all the voters have the same ranking where they rank every alternative in $A_\ell$ higher than every alternative in $A_{\ell'}$ for all $\ell < \ell'$ (and breaks ties within each $A_\ell$ arbitrarily), and rank members of $B$ at the end of their ranking.

Consider any aggregation rule. For each $a \in A$, let $p_a$ denote the marginal probability of alternative $a$ being included in the distribution returned by the rule on this profile. For each $\ell \in [k]$, define $\bar{p}_\ell = \frac{1}{\ell} \sum_{a \in A_\ell} p_a$ as the average of the marginal probabilities of alternatives in $A_\ell$ being chosen. Since the rule returns a distribution over budget-feasible subsets of alternatives (with total cost at most 1), the expected cost under this distribution is also at most 1. Due to additivity of cost and linearity of expectation, the expected cost can be written as

$$\sum_{a \in A} p_a \cdot c_a \geq \sum_{\ell \in [k]} \left( \frac{1}{\ell} \sum_{a \in A_\ell} p_a \right) = \sum_{\ell \in [k]} \bar{p}_\ell \leq 1. \tag{13.1}$$

Next, fix an arbitrary $t \in [k]$. Consider the following consistent utility function of the agent (which, in this case, is also her PS-value function): $v(a) = u(a) = 1$ if $a \in \cup_{\ell \in [t]} A_\ell$ and $v(a) = u(a) = 0$ otherwise. It is evident that the budget-feasible subset with the highest social welfare (i.e., one which contains the highest number of alternatives of value 1 to the agent) is $A_t$, and $\mathrm{sw}(A_t) = t$. In contrast, using the additivity of the utility function over the alternatives and linearity of expectation, we can write the expected social welfare under the rule as $\sum_{a \in \cup_{\ell \in [t]} A_\ell} p_a \cdot 1 = \sum_{\ell \in [t]} \ell \cdot \bar{p}_\ell$, which means the distortion is at least

$$D_t = \frac{t}{\sum_{\ell \in [t]} \ell \cdot \bar{p}_\ell}.$$

Because $t \in [k]$ was fixed arbitrarily, we get that the distortion is at least $D = \max_{t \in [k]} D_t$. Our goal is to show that $D = \Omega(\log m)$.

Note that for each $t \in [k]$, we have

$$\frac{t}{\sum_{\ell \in [t]} \ell \cdot \bar{p}_\ell} \leq D \Rightarrow \sum_{\ell \in [t]} \ell \cdot \bar{p}_\ell \geq \frac{t}{D}.$$

Dividing both sides by $t(t+1)$, we have that

$$\sum_{\ell \in [t]} \frac{\ell}{t(t+1)} \cdot \bar{p}_\ell \geq \frac{1}{D \cdot (t+1)}, \forall t \in [k].$$

Taking the sum over $t \in [k]$, the right hand side sums to $(H_{k+1} - 1)/D$. In the left hand side, the coefficient of each $\bar{p}_\ell$ is

$$\ell \cdot \sum_{t=\ell}^{k} \frac{1}{t(t+1)} = \ell \cdot \left( \sum_{t=\ell}^{k} \frac{1}{t} - \frac{1}{t+1} \right) = \ell \cdot \left( \frac{1}{\ell} - \frac{1}{k+1} \right) \leq 1.$$

Hence, the left hand side sums to at most $\sum_{\ell \in [k]} \bar{p}_\ell \leq 1$. Since the left hand side is at least the right hand side, we have that

$$1 \geq \frac{H_{k+1} - 1}{D} \Rightarrow D \geq H_{k+1} - 1 = H_{\lceil \sqrt{m} \rceil} - 1,$$

which completes the proof after observing that $H_{\lceil \sqrt{m} \rceil} \geq \ln(\lceil \sqrt{m} \rceil) \geq \ln(\sqrt{m}) = \frac{1}{2} \ln(m)$. □

**Remark 13.4.9** (Rankings by value-for-money)**.** *Another ranking-based ballot format considered in the PB literature is rankings by value-for-money, which force voters to consider the cost-benefit analysis of different alternatives, rather than just the benefits. In Appendix H.1, we give analogous upper and lower bounds for this ballot format, showing unbounded deterministic distortion in Theorem H.1.1, and randomized distortion analogous to ranking by value $O\left({}^{(\log m)}/_{\gamma_{\min}}\right)$ in Theorem H.1.2. We demote this ballot format to the appendix because it can be difficult for voters to compute, and in the deterministic case it is bad; in the randomized case, it behaves similarly to pure rankings-by-value.*

## 13.5 Approval-Based Ballots

Another popular type of ballot — especially in participatory budgeting — is to ask voters to simply *approve* their favorite items, rather than rank items relative to one another. The most common type of approval-based ballots in practice is the *k-approval ballot*, in which voters "vote" by identifying their $k$ favorite alternatives. However, this ballot format has an important limitation in the PB context: as we show, it allows voters to approve items or sets of items that are *not budget-feasible*. In the worst case, this can leave the voting rule with little or no information about which

*budget-feasible* allocations are desirable, in which case it can do nothing better than making an arbitrary choice.

A natural potential fix for this is allowing voters to approve *only sets of items that are budget-feasible*. This is can be achieved by either restricting our use to 1-approval ballots (and removing all items which individually exceed the budget), or using *Knapsack ballots*, an approval-based ballot format in which voters can approve any set of projects whose total cost does not exceed the budget. We explore both these directions.

### 13.5.1 $k$-approval ballots

For the ballot format $k$-approval (k-app), the set of possible ballots $\mathcal{L}_{\text{k-app}}$ is the set of all subsets of size $k$ of $A$. That means each voter submits the set of her top $k$ alternatives (breaking the ties arbitrarily). We start by showing that asking voters to approve more than one alternative leads to an unbounded distortion.

**Theorem 13.5.1** (LB - Deterministic)**.** *For $k$-approval ballot format with $k \geq 2$, any deterministic PB rule has unbounded distortion.*

*Proof.* Suppose we are using $k$-approval ballots. Let $A$ be the alternatives, and suppose that each $a \in A$ has cost $\frac{1}{k-1}$. Suppose all agents have the same utilities, where $\epsilon > 0$ is arbitrarily small, giving 1 utility to $a_1$, $\epsilon$ utility for all of $a_2 \dots a_k$, and 0 for all $A \setminus \{a_1, \dots, a_k\}$. Then, everyone's public-spirited values are identical to their utilities. All agents approve $a_1, \dots, a_k$, and the deterministic rule must pick $k-1$ of these arbitrarily. Let the deterministic rule pick $a_2 \dots a_k$. The best possible welfare is $n$, achieved by any $k-1$-subset including $a_1$; the winner has welfare $\epsilon n$, making the distortion $\frac{1}{\epsilon}$ (unbounded). □

These lower bounds were for $k \geq 2$; one can also realize the same bounds with $k = 1$, where all voters approve items whose costs exceed 1, giving the voting rule no information about which budget-feasible set to choose. However, an obvious fix for this is to remove all items ahead of time that exceed the budget. If we assume every *individual* item has cost at most 1, then 1-approval ballots ensure that voters can only approve budget-feasible sets, escaping the problem described above. Then, 1-approval-based ballots are akin to plurality voting, and they permit the following positive result:

**Proposition 13.5.2** (UB, 1-app, Deterministic)**.** *If all alternatives have cost at most* 1*, then for* 1*-approval ballot format, there exists a deterministic voting rule $f$ with distortion*

$$dist_{1\text{-}app}(f) \in O\left(\frac{m^2}{\gamma_{\min}}\right).$$

*Proof.* Pick the most approved alternative $a$. This is in fact the plurality winner and by Theorem 13.4.3, the plurality rule achieves the claimed distortion. □

The following proposition shows that this is the best we can hope for. The full proof of Proposition 13.5.3 is available in Appendix H.6.1.

**Proposition 13.5.3** (LB, 1-app, Deterministic)**.** *For* 1*-approval ballot format, every deterministic rule $f$ has distortion*

$$dist_{1\text{-}app}(f) \in \Omega\left(\frac{m^2}{\gamma_{\min}}\right).$$

*Proof Sketch.* Consider an instance with $\frac{m}{2}$ alternatives of cost 1 where each of them are approved by $\frac{2}{m}$ voters. In addition the remaining $\frac{m}{2}$ alternatives have cost $\frac{m}{2}$, and are never approved by any voter, .

Any PB rule must pick one of the approved alternative, since otherwise we can take the underlying utility profile that gives the unapproved alternatives utility zero. In this case, we can make unapproved alternatives to appear in the second to the $m/2+1$-th position of every voter which gives us the claimed bound. □

**Remark 13.5.4.** *While not explicitly studied in Benadè et al. [45], a deterministic distortion of $\Theta(m^2)$ in the 1-approval ballot format follows from their analysis of the ranking by value ballot format immediately, as it simply uses a plurality rule to aggregate voter preferences.*

While 1-approval ballot sounds practical, it does not yield a good distortion since the basic potential of PB (which is selecting multiple alternatives if the budget allows) is not used. However, this is really the best we can hope for with $k$-approval ballots. This motivates the consideration of knapsack ballots, which elicits the top budget-feasible subset from each voter's perspective.

### 13.5.2 Knapsack ballots

For the ballot format *knapsack* (knap), the set of possible ballots $\mathcal{L}_{\text{knap}} = \mathcal{F}$ is the set of all budget-feasible subsets of $A$. Each voter $i$ submits the subset she values most: $\rho_i \in \arg\max_{S\in\mathcal{F}} v_i(S)$. This amounts to asking each voter to solve her own personal knapsack problem.

Unfortunately, similar to what happens with 1-app ballots, an instance similar to the one in Proposition 13.5.3 also applies to knapsack ballots, since voters are only permitted to approve budget-feasible allocations, which all consist of one single item.

**Corollary 13.5.5** (LB, knap, Deterministic)**.** *For knapsack ballot format, every deterministic rule $f$ has distortion*

$$dist_{knap}(f) \geq m\gamma_{\min}^{-1} - m + 1 \in \Omega\left(\frac{m}{\gamma_{\min}}\right).$$

For randomized rules, we prove a slightly weaker lower bound that is $\gamma_{\min}$ times our lower bound for deterministic rules. As $\gamma_{\min}$ goes from 0 to 1, the lower bound for deterministic rules goes from unbounded to 1 while that for randomized rules goes from $m$ to 1. It is easy to observe that both lower bounds are tight at both extremes, but there may be room for improvement for intermediate values of $\gamma_{\min}$. The proof is in Appendix H.7.1.

**Theorem 13.5.6** (LB, knap, Randomized)**.** *For knapsack ballot format, every randomized rules $f$ has distortion*

$$dist_{knap}(f) \geq m(1 - \gamma_{\min}) + \gamma_{\min}.$$

This lower bound is trivially tight in $m$. We show this by having $m$ alternatives of cost 1 each, and $\frac{n}{m}$ voters approving each one.

**Remark 13.5.7** (UB, knap, Randomized)**.** *The voting rule $f$ which ignores all the ballots and simply picks a single alternative uniformly at random trivially yields an upper bound of $dist_{knap}(f) \leq m$.*

Finally, we present upper bounds for knapsack due to its importance in the literature. In the unit-sum model, Benadè et al. [45] give exponential lower bounds for the knapsack ballot format. We are able to prove that in the public-spirit model, it is possible to break this exponential barrier, showing that the worst-case instances for knapsack in the unit-sum model rely on potentially infeasible voter preferences. In doing so, we rely on new techniques for aggregating knapsack votes. This illustrates how public spirit can be much more powerful than that pervasive assumption (which is hard to justify) in mitigating distortion, especially when the number of alternatives is at all large.

**Theorem 13.5.8** (UB, knap, Deterministic)**.** *For knapsack votes, there exists a deterministic rule $f$ with distortion*

$$dist_{knap}(f) \leq 4m^3(\gamma_{\min}^{-2} - \gamma_{\min}^{-1}) + 3m \in O\left(\frac{m^3}{\gamma_{\min}^2}\right).$$

*Proof.* For any subset of alternatives $S \subseteq A$, let $n_S := \sum_{i \in N} \mathbb{I}(S \subseteq \rho_i)$ be the number of voters whose knapsack set contains $S$. We use shorthand $n_a := n_{\{a\}}$ and $n_{a,b} := n_{\{a,b\}}$ for all $a, b \in A$. Then, informally, $n_{a,b}$ is the number of voters who vote for both $a$ and $b$.

For an arbitrary input, define $A_0 := \{a \in A : n_a \geq \frac{n}{2m}\}$ and initialize $A^- = A_0$ and $A^+ = \emptyset$. We will return $A^+$ after running the following until $A^-$ is empty:

1. Remove the alternative $b$ with the highest cost in $A^-$ and add it to $A^+$.
2. Remove from $A^-$ all alternatives $a$ such that

$$\frac{n_{a,b}}{n_b} \leq \frac{m-1}{m}.$$

First, we will prove that this algorithm always returns a budget-feasible subset. Suppose for the sake of contradiction that at some point, the max-cost item in $A^-$, call it $a^{\mathsf{m}}$, is no longer within budget: i.e., $c_{a^{\mathsf{m}}} + \sum_{b \in A^+} c_b > 1$. We will show that there exists some $b \in A^+$ such that $\frac{n_{b,a^{\mathsf{m}}}}{n_b} \leq \frac{m-1}{m}$.

Let $b^{\mathsf{m}} \in A^+$ be the first alternative added to $A^+$, so that it has maximum cost. Then, for all $b \in A^+ \setminus \{b^{\mathsf{m}}\}$, because $b$ wasn't pruned in step 2 directly after adding $b^{\mathsf{m}}$, it must be that $\frac{n_{b,b^{\mathsf{m}}}}{n_{b^{\mathsf{m}}}} > \frac{m-1}{m}$. By the same reasoning, the same must be true for $a^{\mathsf{m}}$ —that is, $\frac{n_{a^{\mathsf{m}},b^{\mathsf{m}}}}{n_{b^{\mathsf{m}}}} > \frac{m-1}{m}$. Summing over

these inequalities, we get that:

$$n_{a^m,b^m} + \sum_{b \in A^+ \setminus \{b^m\}} n_{b^m,b} > n_{b^m} \left[ \frac{m-1}{m} + \frac{m-1}{m} \left( |A^+| - 1 \right) \right] = n_{b^m} \frac{m-1}{m} |A^+|.$$

Notice that the left hand side is at most the number of voters who voted for $b^m$, multiplied by the number of other alternatives in $\{a^m\} \cup |A^+|$ they could have voted for. Since $\{a^m\} \cup A^+$ is an infeasible set, no voter could have voted for all of them. Thus, each voter can only vote for $|A^+|$ alternatives in $\{a^m\} \cup |A^+|$, and so only $|A^+| - 1$ alternatives other than $b^m$. The left hand side is then at most $(|A^+| - 1) n_{b^m}$, and therefore

$$(|A^+| - 1) n_{b^m} > n_{b^m} \frac{m-1}{m} |A^+|.$$

Simplifying, we can see that this is impossible, as this is equivalent to the inequality:

$$|A^+| - 1 > |A^+| - |A^+|/m.$$

We have encountered a contradiction, so our premise — that we added an $a$ to $A^+$ that exceeded the budget — must have been false.

Now, we will show that if an $a \in A^-$ is pruned in Step 2, then $\frac{\mathrm{sw}(a)}{\mathrm{sw}(A^+)} \leq 2m^2 \frac{1-\gamma_{\min}}{\gamma_{\min}} + 1$. Indeed, because we prune it, there exists some $b \in A^+$ such that:

$$\frac{n_{a,b}}{n_b} \leq \frac{m-1}{m}.$$

Since $b \in A_0$, we have $n_b \geq n/2m$ and so $n_b - n_{a,b}$, the number of voters that vote for $b$ but not $a$, is at least $n/(2m^2)$:

$$n_b - n_{a,b} \geq n_b - \frac{m-1}{m} n_b \geq \frac{n}{2m^2}.$$

Notice that because we pick the highest cost alternative $b$ in each iteration, any alternative pruned later by the algorithm must have a cost lower than $c_b$. Therefore, any time a voter votes for $b$ but not $a$, they could have replaced $b$ with $a$ and have gotten another feasible set. The fact that they did not means that they prefer $b$ to $a$. We have at least $n/(2m^2)$ of such voters (that prefer $b$ to $a$), by Lemma 13.2.1 we can conclude that $\frac{\mathrm{sw}(a)}{\mathrm{sw}(A^+)} \leq 2m^2 \frac{1-\gamma_{\min}}{\gamma_{\min}} + 1$, as needed.

Extending this result, define $m_0 := |A_0|$, we get that

$$\frac{\mathrm{sw}(A_0)}{\mathrm{sw}(A^+)} \leq m_0 \left( 2m^2 \frac{1-\gamma_{\min}}{\gamma_{\min}} + 1 \right).$$

On the other hand, for alternatives outside of $A_0$, the distortion must be small. Let $A^*$ be the optimal budget-feasible set of alternatives. Then:

$$\frac{\mathrm{sw}(A^* \setminus A_0)}{\mathrm{sw}(A^+)} = \frac{\mathrm{sw}(A^* \setminus A_0)}{\mathrm{sw}(A_0)} \cdot \frac{\mathrm{sw}(A_0)}{\mathrm{sw}(A^+)}.$$

It remains to bound $\frac{\mathrm{sw}(A^* \setminus A_0)}{\mathrm{sw}(A_0)}$. Because at most $n/(2m)$ voters include each alternative in $A \setminus A_0$ in their knapsack set, and there are at most $m - m_0$ such alternatives, we know that at most $n(m-m_0)/2m$ voters vote for alternatives in $A \setminus A_0$, that is at least $n(m+m_0)/2m$ voters only vote for alternatives in $A_0$. Observing that $A^* \setminus A_0 \in \mathcal{F}$ (since $A^* \in \mathcal{F}$), it must be that for all $n(m+m_0)/2m$ voters $i$ who vote for only alternatives in $A_0$, $v_i(A_0) \geq v_i(\rho_i) \geq v_i(A^* \setminus A_0)$ for each $a \in A \setminus A_0$. Therefore, by Lemma 13.2.1,

$$\frac{\mathrm{sw}(A^* \setminus A_0)}{\mathrm{sw}(A_0)} \leq \frac{2m}{m+m_0} \cdot \frac{1-\gamma_{\min}}{\gamma_{\min}} + 1.$$

Thus,

$$\begin{aligned}
\frac{\mathrm{sw}(A^*)}{\mathrm{sw}(A^+)} &\leq \frac{\mathrm{sw}(A_0)}{\mathrm{sw}(A^+)} + \frac{\mathrm{sw}(A^* \setminus A_0)}{\mathrm{sw}(A^+)} = \frac{\mathrm{sw}(A_0)}{\mathrm{sw}(A^+)} + \frac{\mathrm{sw}(A^* \setminus A_0)}{\mathrm{sw}(A_0)} \cdot \frac{\mathrm{sw}(A_0)}{\mathrm{sw}(A^+)} \\
&\leq \frac{\mathrm{sw}(A_0)}{\mathrm{sw}(A^+)} \left(1 + \frac{m}{m_0} \cdot \frac{1-\gamma_{\min}}{\gamma_{\min}} + 1\right) \\
&\leq m_0 \left(2m^2 \frac{1-\gamma_{\min}}{\gamma_{\min}} + 1\right) \left(\frac{m}{m_0} \cdot \frac{1-\gamma_{\min}}{\gamma_{\min}} + 2\right) \\
&\leq 2m^3 \left(\frac{1-\gamma_{\min}}{\gamma_{\min}}\right)^2 + 4m^3 \frac{1-\gamma_{\min}}{\gamma_{\min}} + m \frac{1-\gamma_{\min}}{\gamma_{\min}} + 2m \\
&\leq 4m^3 \left(\gamma_{\min}^{-2} - \gamma_{\min}^{-1}\right) + 3m. \qquad \square
\end{aligned}$$

It's possible that for general Knapsack voting, this cannot be improved to match the lower bound that is achieved in the case that reduces to plurality voting. This is because in the general case where people can approve more than 1 alternative, although we have *budget-feasible information*, we don't know what people's *favorite* element is in their approval set if it is greater than size 1.

**Remark 13.5.9.** *For the special case of committee selection, we show in Appendix H.7.2 that this bound can be improved to* $m^2(\gamma_{\min}^{-1} - 1) + m \in O\left(m^2/\gamma_{\min}\right)$.

**Remark 13.5.10** (Threshold approvals)**.** *Another approval-based ballot format considered in the literature is threshold approvals, which are categorically different than knapsack and k-approvals: instead of approving a limited set of alternatives, voters approve any alternative for which their utility exceeds a certain threshold. In Appendix H.2, we give analogous upper and lower bounds for this ballot format. For deterministic rules, we show unbounded deterministic distortion for a fixed choice of threshold in Proposition H.2.1 and* $\Omega(m)$ *and* $O\left(m^2/\gamma_{\min}\right)$ *distortion when the threshold is variable in Theorems H.2.3 and H.2.2. For randomized rules, we show* $\Omega(\sqrt{m})$ *with fixed thresholds and* $\Omega(\log m)$ *with variable thresholds in Theorems H.2.4 and H.2.5 using the ideas in Benadè et al. [45]. We demote this ballot format to the appendix due to its limited practicability: even if people can assign internally-consistent numeric values to their utilities, they may not consider their utilities on the same scale, making it hard for people to reliably approve alternatives according a given threshold.*

## 13.6 A Thrifty Ordinal Ballot Gets Sublinear Distortion

Let us revisit the story so far for deterministic aggregation rules, which is the more practical case. Rankings by value allowed us to achieve $O\left(m/\gamma_{\min}^2\right)$ distortion, and approval-based ballots, which could outperform rankings by value in the unit-sum model [45], fail to do so in the public spirit model, leaving our quest of achieving distortion sublinear in $m$ (via a practical ballot format) unfulfilled.

In this section, we introduce a new (family of) ballot format(s), *ranking of predefined bundles (rpb)*, which meets both these desiderata. Not only does it allow achieving sublinear distortion via a deterministic aggregation rule, it is also extremely practical in participatory budgeting due to four reasons:

- *Explainable:* It simply asks voters to rank bundles of projects by value instead of individual projects.
- *Ordinal*: It asks voters to only ordinally compare bundles of projects.
- *Thrifty*: The number of bundles that voters rank is at most $m$, making the number of bits of information elicited from each voter polynomial in $m$.
- *Reduction to single-winner voting:* The bundles we create below are budget-feasible (so voters can realistically imagine them being implemented) and pairwise disjoint (so voters can easily compare them). Further, the subset of projects funded in the end is precisely one of the bundles on the ballot. This creates a reduction to single-winner voting, where voters understand that they are effectively expressing preferences over possible final outcomes. This also opens up the possibility of using well-known aggregation rules from single-winner voting (such as our use of Copeland's rule below), which voters may already be familiar with.

Specifically, an rpb ballot is characterized by a set $\mathcal{P} = \{P_1, \ldots, P_\ell\}$ of $\ell$ feasible subsets of $A$. We suggest that $\ell$ should be at most polynomial in $m$. Thus, $\mathcal{L}_{\mathsf{rpb}(\mathcal{P})}$ is the set of all rankings over $\mathcal{P}$. Each voter $i$ submits a ranking $\rho_i \in \mathcal{L}_{\mathsf{rpb}(\mathcal{P})}$ such that for all bundles $P, P' \in \mathcal{P}$ with $v_i(P) > v_i(P')$, we have $P \succ_{\rho_i} P'$. An aggregation rule $f$ for this format gets $\vec{\rho} \in \mathcal{L}^n_{\mathsf{rpb}(\mathcal{P})}$ as input.

We show how to use the rpb ballot to achieve $O\left(\sqrt{m}/\gamma_{\min}^2\right)$ distortion in a one-round voting system, and an even better $O\left((\log m)/\gamma_{\min}^4\right)$ distortion in a two-round voting system.

### 13.6.1 Sublinear Distortion in One Round

Let us describe our proposed voting system, which comprises of an rpb ballot we term *high-low bundling* (HLB) along with a deterministic aggregation rule (Copeland's rule).

Ballot: rpb with high-low bundling (HLB). We initialize an rpb ballot with the set $\mathcal{P}^{\text{HLB}}$ constructed as follows. Let $L = \{a \in A : c(a) \leq 1/\sqrt{m}\}$ be the set of *low-cost* alternatives, and $H = \{a \in A : c(a) > 1/\sqrt{m}\}$ be the set of *high-cost* alternatives. $\mathcal{P}^{\text{HLB}}$ consists of an arbitrary

partition of $L$ into at most $\sqrt{m}$ feasible bundles[1] and an arbitrary partition of $H$ into feasible bundles.[2] Note that $|\mathcal{P}| \leq |H| + |L| = m$.[3] The voters are asked to rank the bundles in $\mathcal{P}^{\text{HLB}}$, which generates an input profile $\vec{\rho}$.

Aggregation rule. We simply run Copeland's rule on $\vec{\rho}$, treating each bundle as an alternative in single-winner voting, to select one of the feasible bundles as the final output.

**Theorem 13.6.1** (Upper Bound)**.** *The distortion of (deterministic) Copeland's aggregation rule $f_{Copeland}$ applied to the HLB ballot is*

$$dist_{\text{rpb}(\mathcal{P}^{HLB})}(f_{Copeland}) \leq \frac{2\sqrt{m}}{\gamma_{\min}^2} \in O\left(\frac{\sqrt{m}}{\gamma_{\min}^2}\right).$$

*Proof.* Let $A^*$ be an optimal budget-feasible subset of alternatives. The elements of $A^*$ are distributed among $L$ and $H$, so $\text{sw}(L \cap A^*) + \text{sw}(H \cap A^*) = \text{sw}(A^*)$, implying that either $\text{sw}(L \cap A^*) \geq \frac{1}{2}\text{sw}(A^*)$ or $\text{sw}(H \cap A^*) \geq \frac{1}{2}\text{sw}(A^*)$. We claim that there exists a bundle $P^* \in \mathcal{P}^{\text{HLB}}$ for which $\text{sw}(P^*) \geq \frac{\text{sw}(A^*)}{2\sqrt{m}}$.

Suppose $\text{sw}(L) \geq \text{sw}(L \cap A^*) \geq \frac{1}{2}\text{sw}(A^*)$. Since $L$ is partitioned into at most $\sqrt{m}$ bundles in $\mathcal{P}^{\text{HLB}}$, there exists $P^* \in \mathcal{P}^{\text{HLB}}$ such that $\text{sw}(P^*) \geq \frac{\text{sw}(L)}{\sqrt{m}} \geq \frac{\text{sw}(A^*)}{2\sqrt{m}}$.

Next, suppose $\text{sw}(H \cap A^*) \geq \frac{1}{2}\text{sw}(A^*)$. Since each alternative in $H \cap A^*$ has cost more than $\frac{1}{\sqrt{m}}$ and lies in the budget-feasible set $A^*$, we have that $|H \cap A^*| \leq \sqrt{m}$. Thus, there exists an alternative $a^* \in H \cap A^*$ with $\text{sw}(a^*) \geq \frac{\text{sw}(H \cap A^*)}{\sqrt{m}} \geq \frac{\text{sw}(A^*)}{2\sqrt{m}}$. Hence, for the bundle $P^* \in \mathcal{P}^{\text{HLB}}$ containing $a^*$, we have $\text{sw}(P^*) \geq \frac{\text{sw}(A^*)}{2\sqrt{m}}$.

Note that Copeland's rule receives rankings over bundles in $\mathcal{P}^{\text{HLB}}$ as input to pick a bundle $P$. Using its distortion bound (from single-winner voting), we know that

$$\text{sw}(P) \geq \gamma_{\min}^2 \cdot \text{sw}(P^*) \geq \gamma_{\min}^2 \cdot \frac{\text{sw}(A^*)}{2\sqrt{m}},$$

yielding distortion at most $\frac{2\sqrt{m}}{\gamma_{\min}^2} \in O\left(\frac{\sqrt{m}}{\gamma_{\min}^2}\right)$. □

**Remark 13.6.2.** *In the unit-sum model of Benadè et al. [45] (without public spirit), the distortion of any deterministic aggregation rule on any rpb ballot remains $\Omega(m^2)$ due to single-winner instances (as a special case of PB). When each bundle is budget-feasible, this creates precisely a single-winner*

[1] This is possible because $|L| \leq m$ and any subset of $\sqrt{m}$ alternatives from $L$ is feasible.

[2] One can use this flexibility of partitioning $L$ and $H$ arbitrarily to make the resulting bundles meet practical desiderata, e.g., including a diverse set of projects. Alternatively, one can also create partitions of $L$ and $H$ into the fewest feasible bundles to reduce the size of the ballot.

[3] In practice, with many low-cost projects, we expect $|\mathcal{P}|$ to be much smaller.

*instance. And it is easy to see that grouping any two alternatives together can lead to infinite distortion if the voters unanimously find that bundle the most preferable but we may pick the bad alternative in that bundle which the voters have zero value for.*

### 13.6.2 Logarithmic Distortion in Two Rounds

Next, we describe a two-round voting system, which beats even the sublinear distortion achieved above and yields a logarithmic distortion.

First ballot: rankings by value. Simply use the rankings by value ballot, where voters are asked to rank the alternatives in $A$.

Second ballot: rpb with tiered-cost bundling (TCB). For $r \in \{0, 1, \ldots, \lceil \log_2 m \rceil\}$, define tiers of costs as

$$T_r = \begin{cases} \{ a \in A : c(a) \leq 1/m \} & \text{if } r = 0, \\ \{ a \in A : 2^{r-1}/m < c(a) \leq 2^r/m \} & \text{if } r > 0. \end{cases}$$

For each $r \in \{0, 1, \ldots, \lceil \log_2 m \rceil\}$, use the committee selection rule from Lemma 13.4.7 to pick $P_r \subseteq T_r$ of size $t_r = \lfloor \min(|T_r|, \max(1, m/2^r)) \rfloor$. Note that each $P_r$ is budget-feasible. Our rpb ballot in the second stage is now defined by $\mathcal{P}^{\text{TCB}} = (P_0, \ldots, P_{\lceil \log_2 m \rceil})$. Each voter submits a ranking $\rho_i$ over $\mathcal{P}^{\text{TCB}}$.

Aggregation rule. Run (deterministic) Copeland's rule on the input $\vec{\rho}$ and return the bundle $P \in \mathcal{P}^{\text{TCB}}$ that it picks.

**Theorem 13.6.3.** *The distortion of the two-round voting system that uses rankings by value, then the rpb ballot with tiered-cost bundling, and then Copeland's rule is at most* $2(\lceil \log_2 m \rceil + 1) \cdot \left(2\gamma_{\min}^{-1} - 1\right)^4$.

*Proof.* Let $A^*$ be an optimal budget-feasible subset of the alternatives. Fix any $r \in \{0, 1, \ldots, \lceil \log_2 m \rceil\}$. Let $P_r^*$ be the optimal $t_r$-sized subset of $T_r$ (note that this is feasible by the definition of $t_r$). Using the distortion bound of the committee selection rule from Lemma 13.4.7, we have $\text{sw}(P_r^*) \leq \left(2\gamma_{\min}^{-1} - 1\right)^2 \cdot \text{sw}(P_r)$. Since $A^*$ is feasible, $|A^* \cap T_r| \leq 2t_r$, so $A^* \cap T_r$ can be partitioned into two feasible subsets of $T_r$ of size at most $t_r$ each, yielding $\text{sw}(A^* \cap T_r) \leq 2 \cdot \text{sw}(P_r^*) \leq 2\left(2\gamma_{\min}^{-1} - 1\right)^2 \cdot \text{sw}(P_r)$.

Since $T_0, \ldots, T_{\lceil \log_2 m \rceil}$ partitions the set of alternatives $A$, we have

$$\text{sw}(A^*) = \textstyle\sum_{r \in \{0,1,\ldots,\lceil \log_2 m \rceil\}} \text{sw}(A^* \cap T_r) \leq 2(\lceil \log_2 m \rceil + 1)\left(2\gamma_{\min}^{-1} - 1\right)^2 \cdot \max_{r \in \{0,1,\ldots,\lceil \log_2 m \rceil\}} \text{sw}(P_r).$$

Using the distortion bound of Copeland's rule, we have that for the bundle $P$ picked by the rule,

$$\text{sw}(P) \geq \frac{\max_{r \in \{0,1,\ldots,\lceil \log_2 m \rceil\}} \text{sw}(P_r)}{\left(2\gamma_{\min}^{-1} - 1\right)^2} \geq \frac{\text{sw}(A^*)}{2(\lceil \log_2 m \rceil + 1) \cdot \left(2\gamma_{\min}^{-1} - 1\right)^4}. \qquad \square$$

We remark that there are no known lower bounds that prohibit one from achieving even constant distortion using a one-round voting system that uses an rpb (or some other fully ordinal) ballot format with only polynomially many comparisons. We leave this as a major open question that can have implications for PB ballot design in practice.

## 13.7 Discussion

Our work lays out several interesting open questions as in some cases, our upper and lower bounds do not asymptotically match (see Tables 13.1 and 13.2) in either $m$, $\gamma_{\min}$ or both.

Our work posits, based on prior research, that democratic deliberation in real-world PB may cause voters to be public-spirited. However, modeling the exact level of public spirit achieved and using this to in turn optimize the design of the deliberation process itself would be an important direction for future research. More broadly, distortion has been studied in models beyond voting, such as matching [118] and fair division [160], to which the public-spirit model can also be applied. Finally, under the public-spirit model, participants take the utilitarian welfare into account when submitting their preferences, which works well since the goal is to optimize the utilitarian welfare as well. But the idea of distortion has been extended to other objectives such as the Nash welfare or proportional fairness [103], which raises the question: what form of public-spirit can be helpful in optimizing such objectives and how can it be cultivated?

# 14

# Ongoing and Future Work

## 14.1 Public Spirit in the Wild

Chapters 12 and 13 asked the question, *does public spirited voting behavior improve the social good of democratic outcomes?* We find that in both standard voting and the more difficult setting of participatory budgeting, the answer is resoundingly yes. However, these findings are only as practically relevant as the model in which they were proven. Thus, we now turn to the question: *what does public spirit actually look like in practice?* Though in principle public spirit could be measured in any political context, we focus specifically on public spirit as cultivated by *democratic deliberation*. We make this choice based on the existing evidence that deliberation cultivates public spirit (see Chapter 12), and the fact that deliberation creates a controlled environment whose main features can be repeated.

In guiding our measurement of public spirit in practice, our theoretical model serves an important purpose: it decomposes "public spirited" behavior into three outcome-relevant components that can be measured experimentally. These components are (a) the extent to which deliberants prioritize societal benefit over their own; (b) with what information deliberants *evaluate* alternatives' "social benefit"; and (c) what notion of "social benefit" deliberants actually care about. Regarding (c), our theoretical model assumed that people are *utilitarian*, by encoding the sociotropic component of their preferences as the utilitarian social welfare. In reality, people may have multiple diverse notions of social good: some people may care about minimizing inequality, or minimizing harm to the worst-off people; others may be more *deontological*, caring about upholding principles regardless of the outcome in any specific policy context.

In my ongoing work, I am running survey-based studies to measure changes in (a), (b), and (c)

throughout deliberation. To deploy these surveys in real deliberative events, I am collaborating with political scientists and groups of practitioners who are running deliberative town halls and citizens' assemblies. Contexts studied so far include a national-level deliberative town hall on the reform of the Chilean constitution, a deliberation-based political science course at the University of Houston, and a citizens' assembly in Australia on renewable energy pricing.

### 14.2 Beyond a Deterministic Highest-Welfare Alternative

The question studied in Chapters 12 and 13 can be phrased as follows: *when voters are public spirited, how closely does the welfare of the outcome they collectively choose approximate that of the highest-welfare alternative?* In other words, we are assuming that there is a ground-truth best outcome, and we are asking how well voters can recover it.

In the future work proposed here, we now consider: *what if there is no single ground-truth best outcome?* There are many reasons that it is hard to imagine there being a truly "best" policy in a real political decision. We unpack two such reasons here, showing how to extend the public-spirited model to capture these added complexities. In these generalized models, we then propose to study the following hypothesis: *that when there are multiple alternatives that are defensibly the best, deliberation will at least uncover one of them, eliminating the "clearly bad" alternatives.*

**Competing notions of "social good".** One key reason that there might be multiple defensibly best outcomes is *philosophical*, boiling down to the question of whether there is one "right" way to make difficult trade-offs. If all outcomes help some people and harm others, how can we say which one is best for society overall? Making these kinds of trade-offs either implicitly or explicitly requires committing to a notion of social good, of which there are many that are arguably justified. For example, we can imagine versions of our model where instead of the utilitarian social welfare, voters are concerned about inequality ($v_{ineq}$), those who are the worst off ($v_{maximin}$), or even the Nash Welfare ($v_{NW}$):[1]

$$
\begin{aligned}
v_{ineq}(a) &:= (1-\gamma_i)u_i(a) - \gamma_i \left(\max_{j\in[n]} u_j(a) - \min_{j'\in[n]} u_{j'}(a)\right) \\
v_{maximin}(a) &:= (1-\gamma_i)u_i(a) + \gamma_i \min_{j'\in[n]} u_{j'}(a) \\
v_{NW}(a) &:= (1-\gamma_i)u_i(a) + \gamma_i \left(\prod_{j\in[n]} u_j(a)\right)^{1/n}
\end{aligned}
$$

While it may seem strange that people would be computing something as nonlinear as the Nash product, there is some limited evidence that the Nash product is actually more aligned with people's intuitive conception of social good than the utilitarian social welfare [288].

Any of these models—or a convex combination of them—could be a more accurate representation of a given voter's considerations, and which model is truest likely depends on the voter. The

[1] An alternative definition of $v_{NW}(a)$ might be $u_i(a)^{1/\gamma_i} \cdot \prod_{j\in[n]} u_j(a)$, in which case $\gamma_i \in [0, \infty)$.

ongoing experiments discussed in Section 14.1 are designed to shed light on people's conceptions of social good in practice.

We can then address our motivating question in the following way: using formalisms like those above, we can allow voters to form their public-spirited preferences based on diverse conceptions of social good. Likewise, we can redefine the notion of distortion to quantify welfare via different notions. Then, we can ask: *if voters have different conceptions of social good, can we guarantee a good approximation to at least one of them?*

**The unknown future.** Suppose we set aside these considerations, for now maintaining the utilitarian social welfare as our notion of social good. Even with this assumption in place, there is another reason a single highest-welfare alternative is unrealistic: at the time of the decision, policy impacts are *not deterministic*. Rather, how much a given person will harmed or benefited by a given policy can depend on how the future unfolds, which may be random and unpredictable at the time of the policy decision.

We can formalize this intuition by generalizing the latent utilities model: let utility matrix $U$ be a random variable, drawn from a distribution over *entire* $n \times m$ matrices of utilities. We draw entire matrices instead of individual utilities independently because whatever future shocks occur will apply to *everyone simultaneously*, meaning that people's utilities should be correlated. In this generalized model, there is no longer a single outcome that is decisively the most socially beneficial. Rather, each alternative has some associated *distribution* over possible levels of societal benefit. Then, even if an alternative has high *expected* social benefit, it may also come with significant *risk*, having some chance of being extremely harmful.

Now that there is no longer any single "best alternative" with respect to social good, we can again consider our question: does deliberation at least eliminate the "obviously bad" alternatives? One precise interpretation of this question is, suppose we design a measure of *social benefit* (e.g., expected social welfare), and another of *risk*. Then, there is a Pareto frontier, and some alternatives are Pareto-dominated by others. We can define an "obviously bad" alternative as one that is Pareto-dominated by another alternative (i.e., it is both higher-risk and lower benefit).

# Part IV

# Discussion

Although this is officially a computer science thesis, the enclosed work has drawn heavily also from theoretical and empirical political science. In this discussion, we delve deeper into the ways the work in this thesis connects political science and theoretical computer science, hopefully distilling some ideas that can help others do the same.

## 14.3 Translating normative ideals into computational tools

One avenue for combining political science and theoretical computer science is to **implement normative ideals from political theory in the computational tools we build**, or put another way, use computational tools *in service* of achieving political scientific ideals. An example of this approach is our work in Part I, where we designed an algorithmic framework, plus optimization objectives and algorithmic augmentations, that provably serve mathematical formalizations of key sortition ideals put forth by political theorists.

The way this approach can contribute both academically and practically is illustrated by this line of work. First, by formalizing these conceptual ideals mathematically, we proved that under selection bias – a condition that we argue is essentially inevitable in practice (Remark 1.1.1) – there are provable trade-offs between ideals whose *simultaneous* satisfaction has often been used to argue for sortition in the past. By illuminating, characterizing, and algorithmically optimizing these trade-offs, our work opens new questions about the political implications of striking these tradeoffs in different ways, and offers computational tools for empirically studying them. In turn, the political science research on sortition helps clarify a broader and clearer picture of what goals our algorithms do and do not serve, allowing users to make more informed choices.

There is already some work applying this approach to unifying political science and computer science (e.g., see [50]). However, there are many more candidate problem settings that could benefit from it: algorithms are increasingly being proposed to assist with decisions impacting our political systems (e.g., [27, 35, 191, 194, 207, 226, 264]), and formally specifying these algorithms' goals — or deciding whether to deploy these algorithms altogether — often requires making decisions that encode moral trade-offs. When making such decisions, there is work to do in understanding the relevant political scientific frameworks and then carefully considering, from a technical standpoint, how algorithms may serve or impact these goals. Even in cases when these political scientific frameworks do not reveal a clear answer, they can clarify key trade-offs, risks of unintended consequences, benchmarks for evaluation, and questions that are essential to answer before deployment. In turn, designing tools whose exposition explicitly engages with political science frameworks can spur richer multidisciplinary engagement with how to improve these algorithms further.

## 14.4 Part II: Grounding computational social choice models in political scientific research

The fields of theoretical and computational social choice regularly rely on models of electoral systems. These models make assumptions about how voters interact with different ballots, how

they might strategize and with what information, how their preferences may be formed (e.g., worst-case, with randomness, etc), or even how their cardinal preferences (captured as utilities) might be structured. Models are necessary, even if they are always wrong in some way; however, our work in Part II illustrates how it can be informative to investigate how to **relate — and better ground — computational social choice models in frameworks and evidence from political science.**

An example of this approach is our work in Part II, where we propose to capture voters' differing *stakes* — an idea introduced in multiple political scientific theories — within a computational social choice model. In doing so, we prove an equivalence between *stakes* and the widely-used computational social choice assumption that each voter's utility sums to 1. By connecting these literatures, our work suggests an interpretation of the unit-sum utilities assumption: that all voters have the same stakes in the decision. Although this interpretation calls into question whether the unit-sum utilities assumption holds in practice, its equivalence with stakes gives bounds under that assumption added purpose: if we can mechanistically incentivize voters to engage with democratic systems in ways that respond to their stakes, we can (at least approximately, realistically) achieve a condition equivalent to the unit-sum utilities assumption.

Our conclusions in this paper allude to many further opportunities to apply this approach. First, understanding how real voters might engage with our mechanism prompts deeper study of several fundamental assumptions in voting theory: how voters decide in which elections to vote; how they assess the degree to which they are impacted by a decision, and their own probability of pivotality; and how we should practically interpret the utilities used in computational social choice models. Grounding our social choice models in these aspects would serve not only the future work proposed in this thesis, but also open up new avenues for re-examining the many prohibitive impossibilities arising in worst-case social choice models.

One particular modeling assumption that is worth digging into further is the preference model, encapsulating models of voters' ballot submissions and, sometimes, the underlying utilities that shape them. Assuming voters' utilities preferences are adversarial potentially eliminates consideration of an entire spectrum of less pessimistic, still realistic preference structures. While beyond-worst-case models of voter preferences do exist, many assume *independence* between voters (e.g., see this expected distortion model [155], or the growing line of work on smoothed analysis of voting [132, 287]). If we believe that Part II's motivating scenario—the *tyranny of the less affected majority*—could be realistic, then we have reason to be suspicious of this independence assumption: if voters' preferences are independent, it becomes very difficult to recover this kind of problematic election scenario, because doing so relies precisely on the correlations between beliefs that can emerge in practice due to homophily, segregation, and filter bubbles. A fundamental problem that remains open, then, is building theoretical social choice models that more accurately replicate the problems — and the opportunities for positive results — that arise in real elections.

## 14.5 Part III: Distilling where theoretical models have a comparative advantage — and where they should give way to support empirical research and measurement

Our work in Part III can be seen as an example of the approach discussed in the previous section: we propose a model of voter behavior that reflects — if coarsely, so far — evidence of voters going beyond individual rationality in favor of sociotropism, especially under deliberative conditions. However, Part III also exemplifies yet another way that theoretical computer science can contribute to political science research, and vice versa: by **helping support empirical research and measurement**. Doing so, however, requires acute attention to where theoretical models have a comparative advantage.

Suppose for a moment that you, as a computer science researcher, had the same intuition that inspired our work in Part III: that democratic deliberation improved democratic outcomes. In other words, you want to study the following causal relationship:

$$\text{Deliberation} \implies \text{Better Democratic Outcomes}$$

Maybe you decide to approach this question in theory because you identify the same key challenge we did: in any practical case study, it is hard to argue that any given outcome is absolutely better than another (at least in any way that avoids subjective judgement). Theory is a natural approach, because it permits making claims about the quality of the outcome (in our case, its utilitarian social welfare) that hold *irrespective* of underlying model primitives that cannot be observed in practice (in our case, utilities).

Now that we have determined why theory may have a comparative advantage in studying this question, we might proceed via the most direct theoretical approach: to study the entire implication, end-to-end, in theory. This means building a theoretical model of how people interact during deliberation, and then characterizing what kinds of conditions lead to what kinds of outcomes. While this approach is not in any way wrong, it is not clear that studying the *entire implication*—from deliberation to outcome quality—maximizes theory's comparative advantage. After all, there are many aspects of deliberation that are so complicated that they are even difficult to measure, let alone capture precisely in a theoretical model. As a result, any conclusions we reach may be based on a model that departs from reality in ways we cannot even measure, and it will be hard to act on them with certainty.

In Part III, we took a different approach, which we now aim to generalize here. We first expanded our model of this causal relationship into a more general conjectured causal graph, identifying mediators based on existing theories and empirical evidence. In Part III, our conjectured causal graph was quite simple:

$$\text{Deliberation} \implies \textit{Public Spirit} \implies \text{Better Democratic Outcomes}$$

After expanding our causal relationship into a causal graph, we then revisited our question: which of these implications require theory, and which are best suited for study in practice? Our response, as discussed in Part III, was that while the second implication is better-studied in theory due to the aforementioned challenge of measuring absolute outcome quality, *the first implication*

*is more easily measured in practice.* Conveniently, the first implication collapses immeasurably complicated deliberation dynamics into a much simpler condition, meaning that our theoretical model could be considerably simpler than one capturing the entirety of deliberative dynamics. This may in general be an advantage of the approach of identifying causal intermediates: it can simplify the requisite theoretical models, because we can study the impacts of intermediates in isolation — or at least be transparent about why we cannot.

Zooming out, this decomposition of into multiple logical links can be valuable in facilitating richer interplay between empirical study and theoretical study. As was the case in Part III, our theory informed our experiments by identifying outcome-relevant quantities to measure, and our experiments will in turn help inform more accurate theoretical models. This can be true in applications of this method that lead to more general causal graphs as well. In practice, a research approach using this decomposition method can also lead to new *outcome-based evaluation techniques* — even in settings like deliberation, where measuring the quality of outcomes directly is challenging. As an example, consider our work in Part III: based on our approach, one could evaluate a deliberative process on the basis of whether it cultivated public spirit — a condition which we know leads to higher-welfare democratic outcomes, subject to whatever caveats our theoretical study of its impact requires.

# References


[1] Deliberative polling for constitutional change in mongolia: An unprecedented experiment, September 2017. URL https://constitutionnet.org/news/deliberative-polling-constitutional-change-mongolia-unprecedented-experiment. Accessed: 2024-04-19.

[2] Citizens assembly on climate, July 2019. URL https://missionspubliques.org/pf/citizens-assembly-on-climate/?lang=en. Accessed: 2024-04-19.

[3] Some assembly required. *The Economist*, 436(9212):57–58, Sep 19 2020. URL https://www.proquest.com/magazines/some-assembly-required/docview/2444107633/se-2.

[4] Mini publics, 2021. URL https://sfb1265.github.io/mini-publics/. Accessed: 2024-04-19.

[5] Citizens' assembly on nutrition decided, 2023. URL https://www.buergerrat.de/en/news/citizens-assembly-on-nutrition-decided/. Accessed: 2024-04-19.

[6] Permanent climate assembly in brussels, 2023. URL https://www.buergerrat.de/en/news/permanent-climate-assembly-in-brussels/. Accessed: 2024-04-19.

[7] National deliberative poll® in south korea shows wide support for electoral reform, June 2023. URL https://cddrl.fsi.stanford.edu/news/national-deliberative-pollr-south-korea-shows-wide-support-electoral-reform. Accessed: 2024-04-19.

[8] Chile dialoga (dth): Evento 1, 2024. URL https://fundacionpiensa.cl/estudio/chile-dialoga-dth/. Accessed: 2024-04-19.

[9] Chile dialoga (dth): Evento 2, 2024. URL https://fundacionpiensa.cl/estudio/chile-dialoga-dth-evento-2/. Accessed: 2024-04-19.

[10] Das aufsuchende losverfahren - es geht los. https://esgehtlos.org/geloste-burger-rate/aufsuchendes-losverfahren, 2024. Accessed: 2024-04-23.

[11] Global assembly, 2024. URL https://globalassembly.org/. Accessed: 2024-04-19.

[12] Projects - mit center for constructive communication, 2024. URL https://www.ccc.mit.edu/projects/. Accessed: 2024-04-19.

[13] Connecting to congress, 2024. URL https://democracyinstitute.osu.edu/projects/connecting-congress. Accessed: 2024-04-19.

[14] Polis, 2024. URL https://pol.is/home. Accessed: 2024-04-19.

[15] About us - prytaneum, 2024. URL https://prytaneum.io/aboutus. Accessed: 2024-04-19.

[16] List of participatory budgeting votes, 2024. URL https://en.wikipedia.org/wiki/List_of_participatory_budgeting_votes. Accessed: 2024-04-19.

[17] Ben Abramowitz, Elliot Anshelevich, and Wennan Zhu. Awareness of voter passion greatly improves the distortion of metric social choice. In *Web and Internet Economics: 15th International Conference, WINE 2019, New York, NY, USA, December 10–12, 2019, Proceedings 15*, pages 3–16. Springer, 2019.

[18] L. Adleman. Two theorems on random polynomial time. In *Proceedings of the Symposium on Foundations of Computer Science (FOCS)*, pages 75–83. doi: 10/b28vdn.

[19] Saba Ahmadi, Hedyeh Beyhaghi, Avrim Blum, and Keziah Naggita. The strategic perceptron. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 6–25, 2021.

[20] Michelle Alexander. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. The New Press, New York, 2010.

[21] Allianz Vielfältige Demokratie and Bertelsmann Foundation. Citizens' participation using sortition, 2018. URL http://aei.pitt.edu/102678/1/181102_Citizens__Participation_Using_Sortition_mb_-_Copy.pdf.

[22] Georgios Amanatidis, Georgios Birmpas, Aris Filos-Ratsikas, and Alexandros A Voudouris. Peeking behind the ordinal curtain: Improving distortion via cardinal queries. *Artificial Intelligence*, 296:103488, 2021.

[23] Jørgen Juel Andersen, Jon H Fiva, and Gisle James Natvik. Voting when the stakes are high. *Journal of Public Economics*, 110:157–166, 2014.

[24] E. Anshelevich, O. Bhardwaj, E. Elkind, J. Postl, and P. Skowron. Approximating optimal social choice under metric preferences. *Artificial Intelligence*, 264:27–51, 2018.

[25] E. Anshelevich, A. Filos-Ratsikas, N. Shah, and A. A. Voudouris. Distortion in social choice problems: The first 15 years and beyond. arXiv:2103.00911, 2021.

[26] Shamena Anwar, Patrick Bayer, and Randi Hjalmarsson. The impact of jury race in criminal trials. *The Quarterly Journal of Economics*, 127(2):1017–1055, 2012.

[27] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

[28] David Austen-Smith and Timothy Feddersen. Deliberation and voting rules. In *Social choice and strategic decisions: Essays in honor of Jeffrey S. Banks*, pages 269–316. Springer, 2005.

[29] Ayuntamiento de Madrid. El pleno aprueba el reglamento del observatorio de la ciudad. 2019. URL https://diario.madrid.es/blog/notas-de-prensa/el-pleno-aprueba-el-reglamento-del-observatorio-de-la-ciudad/. Accessed: 2024-05-23.

[30] Haris Aziz and Nisarg Shah. Participatory budgeting: Models and approaches. In Tamás Rudas and Gábor Péli, editors, *Pathways Between Social Science and Computational Social Science: Theories, Methods, and Interpretations*, pages 215–236. Springer, 2021.

[31] Carmel Baharav and Bailey Flanigan. Fair, manipulation-robust, and transparent sortition. 2024.

[32] Sven Wang Bailey Flanigan, Ariel D. Procaccia. Accounting for stakes in democratic decisions. 2023.

[33] M. L. Balinski and H. P. Young. *Fair Representation: Meeting the Ideal of One Man, One Vote*. Brookings Institution Press, 2010.

[34] Nikhil Bansal. On a generalization of iterated and randomized rounding. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1125–1135, 2019.

[35] Jake Barrett, Kobi Gal, Paul Gölz, Rose M Hong, and Ariel D Procaccia. Now were talking: Better deliberation groups through submodular optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5490–5498, 2023.

[36] Hanne Bastiaensen, Annie Cook, Kelly McBride, and Daniela Amann. Guide to deliberation: Participatory budgeting, 2021. URL https://www.demsoc.org/uploads/store/mediaupload/560/file/Guide%20to%20Deliberation-%20Participatory%20Budgeting.pdf. Illustrations by Proudfoot, Jenny.

[37] Dorothea Baumeister, Linus Boes, and Tessa Seeger. Irresolute approval-based budgeting. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1774–1776, 2020.

[38] Edana Beauvais and Mark E Warren. What can deliberative mini-publics contribute to democratic systems? *European Journal of Political Research*, 58(3):893–914, 2019.

[39] M. M. Bechtel and R. Liesch. Reforms and redistribution: Disentangling the egoistic and sociotropic origins of voter preferences. *Public Opinion Quarterly*, 84(1):1–23, 2020.

[40] József Beck and Tibor Fiala. "Integer-making" theorems. *Discrete Applied Mathematics*, 3 (1):1–8, 1981.

[41] József Beck and Tibor Fiala. integer-making theorems. *Discrete Applied Mathematics*, 3(1): 1–8, 1981.

[42] G. S. Becker. Altruism, egoism, and genetic fitness: Economics and sociobiology. *Journal of economic Literature*, 14(3):817–826, 1976.

[43] Mark Bedaywi, Bailey Flanigan, Mohamad Latifian, and Nisarg Shah. The distortion of public-spirited participatory budgeting.

[44] G. Benadè, P. Gölz, and A. D. Procaccia. No Stratification without Representation. In *Proceedings of the 20th ACM Conference on Economics and Computation*, pages 281–314, 2019.

[45] G. Benadè, S. Nath, A. D. Procaccia, and N. Shah. Preference elicitation for participatory budgeting. *Management Science*, 65(5):2813–2827, 2021.

[46] Gerdus Benadè, Paul Gölz, and Ariel D. Procaccia. No stratification without representation. In *2019*, pages 281–314, 2019.

[47] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50: 3–44, 2018.

[48] Dimitris Bertsimas, Vivek F. Farias, and Nikolaos Trichakis. On the efficiency-fairness trade-off. *Management Science*, 58(12):2234–2250, 2012.

[49] R. Binns. On the Apparent Conflict between Individual and Group Fairness. In *Proceedings of the 3rd Annual ACM Conference on Fairness, Accountability, and Transparency*, pages 514–524, 2020.

[50] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency*, pages 149–159. PMLR, 2018.

[51] A. Bogomolnaia and H. Moulin. Random matching under dichotomous preferences. *Econometrica*, 72:257–279, 2004.

[52] Allan Borodin, Daniel Halpern, Mohamad Latifian, and Nisarg Shah. Distortion in voting with top-t preferences. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, pages 116–122, 2022.

[53] C. Boutilier, I. Caragiannis, S. Haber, T. Lu, A. D. Procaccia, and O. Sheffet. Optimal social choice functions: A utilitarian view. *Artificial Intelligence*, 227:190–213, 2015.

[54] Stephen P. Bradley, Arnoldo C. Hax, and Thomas L. Magnanti. *Applied Mathematical Programming*. Addison-Wesley Pub. Co.

[55] S. J. Brams and A. D. Taylor. *Fair Division: From Cake-Cutting to Dispute Resolution*. Cambridge University Press, 1996.

[56] Petter Brändén and Johan Jonasson. Negative dependence in sampling. *Scandinavian Journal of Statistics*, 39(4):830–838, 2012.

[57] H. Brighouse and M. Fleurbaey. Democracy and proportionality. *Journal of Political Philosophy*, 18(2):137–155, 2010.

[58] M. Brill, P. Gölz, D. Peters, U. Schmidt-Kraepelin, and K. Wilker. Approval-Based Apportionment. In *Proceedings of the 34th Annual AAAI Conference on Artificial Intelligence*, 2020.

[59] Markus Brill, Stefan Forster, Martin Lackner, Jan Maly, and Jannik Peters. Proportionality in approval-based participatory budgeting. In *Thirty-Seventh AAAI Conference on Artificial Intelligence*, pages 5524–5531, 2023.

[60] Andrew Brinker. In cambridge, a battle over affordable housing revives long-standing political tensions. https://www.bostonglobe.com/2023/10/15/business/affordable-housing-cambridge/, 2023. Accessed: 2024-02-12.

[61] Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. 16(2):101–117. doi: 10/btx9nv.

[62] Eric Budish, Gérard P. Cachon, Judd B. Kessler, and Abraham Othman. Course match: A large-scale implementation of approximate competitive equilibrium from equal incomes for combinatorial allocation. 65(2):314–336. doi: 10/f982dz.

[63] Eric Budish, Yeon-Koo Che, Fuhito Kojima, and Paul Milgrom. Designing random allocation mechanisms: theory and applications. *American Economic Review*, 103(2):585–623, 2013.

[64] B. Bukh. An Improvement of the Beck–Fiala Theorem. *Combinatorics, Probability and Computing*, 25(3):380–398, 2016. ISSN 0963-5483, 1469-2163.

[65] Bürgerrat. French citizens' assembly supports assisted dying. Available at https://www.buergerrat.de/en/news/french-citizens-assembly-supports-assisted-dying/ (2024/02/11), 2023.

[66] Campaign Legal Center. Independent Redistricting Commissions. https://campaignlegal.org/democracyu/accountability/independent-redistricting-commissions. Accessed: April 27, 2024.

[67] I. Caragiannis and A. D. Procaccia. Voting almost maximizes social welfare despite limited communication. *Artificial Intelligence*, 175(9–10):1655–1671, 2011.

[68] I. Caragiannis, S. Nath, A. D. Procaccia, and N. Shah. Subset selection via implicit utilitarian voting. *Journal of Artificial Intelligence Research*, 58:123–152, 2017.

[69] Ioannis Caragiannis and Ariel D Procaccia. Voting almost maximizes social welfare despite limited communication. *Artificial Intelligence*, 175(9-10):1655–1671, 2011.

[70] Ioannis Caragiannis, Swaprava Nath, Ariel D Procaccia, and Nisarg Shah. Subset selection via implicit utilitarian voting. *Journal of Artificial Intelligence Research*, 58:123–152, 2017.

[71] Lyn Carson and Brian Martin. *Random Selection in Politics*. Praeger, 1999.

[72] A. Casella. Storable votes. *Games and Economic Behavior*, 51(2):391–419, 2005.

[73] Pew Research Center. Many around the world are disengaged from politics, 2018. URL https://www.pewresearch.org/global/wp-content/uploads/sites/2/2018/10/Pew-Research-Center_-International-Political-Engagement-Report_2018-10-17.pdf.

[74] Pew Research Center. Americans dismal views of the nations politics. Technical report, Pew Research Center, September 2023. URL https://www.pewresearch.org/wp-content/uploads/sites/20/2023/09/PP_2023.09.19_views-of-politics_REPORT.pdf.

[75] Center For Climate Assemblies, Healthy Democracy, Of By For *, and Sortition Foundation. Personal Communication, 2019–2020.

[76] Moses Charikar, Prasanna Ramakrishnan, Kangning Wang, and Hongxun Wu. Breaking the metric voting distortion barrier. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1621–1640, 2024.

[77] B. Chazelle. *The Discrepancy Method: Randomness and Complexity*. Cambridge University Press, 2001.

[78] C. Chekuri, J. Vondrak, and R. Zenklusen. Dependent Randomized Rounding via Exchange Properties of Combinatorial Structures. In *Proceedings of the 51st IEEE Annual Symposium on Foundations of Computer Science*, pages 575–584, 2010.

[79] P.-A. Chen, B. de Keijzer, D. Kempe, and G. Schäfer. Altruism and its impact on the price of anarchy. *ACM Transactions on Economics and Computation*, 2(4): article 17, 2014.

[80] Yiling Chen, Yang Liu, and Chara Podimata. Learning strategy-aware linear classifiers. *Advances in Neural Information Processing Systems*, 33:15265–15276, 2020.

[81] Yann Chevaleyre, Ulle Endriss, Jérôme Lang, and Nicolas Maudet. A short introduction to computational social choice. In *Proceedings of the 33rd Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM)*, pages 51–69, 2007.

[82] Citizens' Panel on COVID-19. Final report of the citizens' panel on covid-19, 2020. URL https://joinofbyfor.org/wp-content/uploads/2020/11/Final-Report-of-the-Citizens-Panel-on-COVID-19.pdf.

[83] Congressional Research Service. Redistricting commissions for congressional districts. CRS Report IN11053, Congressional Research Service, 2023. URL https://crsreports.congress.gov/product/pdf/IN/IN11053.

[84] V. Conitzer. Computing Slater rankings using similarities among candidates. In *21st*, pages 613–619, 2006.

[85] Vincent Conitzer, Rupert Freeman, and Nisarg Shah. Fair public decision making. In *Proceedings of the ACM Conference on Economics and Computation (EC)*, pages 629–646. doi: 10/gqcq68.

[86] Dimitri Courant. Sortition and Democratic Principles: A Comparative Analysis. In *Legislature by Lot: Transformative Designs for Deliberative Governance*, The Real Utopias Project. Verso.

[87] R. A. Dahl. *After the Revolution?: Authority in a Good Society*. Yale University Press, 1990.

[88] Jessica Dai, Bailey Flanigan, Meena Jagadeesan, Nika Haghtalab, and Chara Podimata. Can Probabilistic Feedback Drive User Impacts in Online Platforms? In *International Conference on Artificial Intelligence and Statistics*, pages 2512–2520. PMLR, 2024.

[89] Michiel S De Vries, Juraj Nemec, and David Špaček. International trends in participatory budgeting. *Cham: Palgrave Macmillan*, 2022.

[90] G. Delannoi and O. Dowlen, editors. *Sortition: Theory and Practice*. Imprint Academic, 2010.

[91] Deliberations.US. Deliberations. https://www.deliberations.us/, 2024. Accessed: 2024-04-22.

[92] Democracy R&D. About the network, 2019. URL https://democracyrd.org/about/.

[93] John P. Dickerson, David F. Manlove, Benjamin Plaut, Tuomas Sandholm, and James Trimble. Position-indexed formulations for kidney exchange. In *Proceedings of the ACM Conference on Economics and Computation (EC)*, pages 25–42. doi: 10/gqcq69.

[94] Benjamin Doerr. Roundings respecting hard constraints. *Theory of Computing Systems*, 40 (4):467–483, 2007.

[95] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

[96] Oliver Dowlen. *The Political Potential of Sortition: A Study of the Random Selection of Citizens for Public Office*. Imprint Academic, .

[97] Oliver Dowlen. Sorting Out Sortition: A Perspective on the Random Selection of Political Officers. 57(2):298–315, . ISSN 0032-3217, 1467-9248. doi: 10/dk7xk7.

[98] Oliver Dowlen. Sorting out sortition: A perspective on the random selection of political officers. *Political Studies*, 57(2):298–315, 2009.

[99] Anthony Downs. An economic theory of democracy. *Harper and Row*, 28, 1957.

[100] John S Dryzek, André Bächtiger, Simone Chambers, Joshua Cohen, James N Druckman, Andrea Felicetti, James S Fishkin, David M Farrell, Archon Fung, Amy Gutmann, et al. The crisis of democracy and the science of deliberation. *Science*, 363(6432):1144–1146, 2019.

[101] Devdatt Dubhashi and Desh Ranjan. Balls and bins: A study in negative dependence. *BRICS Report Series*, 3(25):30, 1996.

[102] Devdatt Dubhashi, Johan Jonasson, and Desh Ranjan. Positive influence and negative dependence. *Combinatorics, Probability & Computing*, 16(1):29, 2007.

[103] S. Ebadian, A. Kahng, D. Peters, and N. Shah. Optimized distortion and proportional fairness in voting. In *23rd*, 2022. Forthcoming.

[104] Soroush Ebadian and Evi Micha. Boosting sortition via proportional representation. Manuscript, 2023.

[105] Soroush Ebadian, Anson Kahng, Dominik Peters, and Nisarg Shah. Optimized distortion and proportional fairness in voting. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 563–600, 2022.

[106] Soroush Ebadian, Gregory Kehne, Evi Micha, Ariel D Procaccia, and Nisarg Shah. Is sortition both representative and fair? *Advances in Neural Information Processing Systems*, 35, 2022.

[107] Jon X Eguia, Nicole Immorlica, Steven P Lalley, Katrina Ligett, Glen Weyl, and Dimitrios Xefteris. Efficiency in collective decision-making via quadratic transfers. *arXiv preprint arXiv:2301.06206*, 2023.

[108] Elise Uberoi, Neil Johnston. Business statistics and policy analysis, 2022. URL https://researchbriefings.files.parliament.uk/documents/CBP-7501/CBP-7501.pdf.

[109] Edith Elkind, Piotr Faliszewski, Piotr Skowron, and Arkadii Slinko. Properties of multiwinner voting rules. 48(3):599–632. doi: 10/f92cx3.

[110] Ulle Endriss. Reduction of economic inequality in combinatorial domains. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. URL http://dl.acm.org/citation.cfm?id=2484951.

[111] Ulle Endriss. Lecture notes on fair division. 2010. URL https://arxiv.org/abs/1806.04234.

[112] Fredrik Engelstad. The assignment of political office by lot. 28(1):23–50. ISSN 0539-0184, 1461-7412. doi: 10/cggrj2.

[113] Brandon Fain, Kamesh Munagala, and Nisarg Shah. Fair allocation of indivisible public goods. In *Proceedings of the ACM Conference on Economics and Computation (EC)*, pages 575–592. doi: 10/gqcq6w.

[114] Fair Outcomes. Fair proposals, n.d. URL https://web.archive.org/web/20220331040954/https://www.fairproposals.com/.

[115] Timothy J Feddersen and Wolfgang Pesendorfer. The swing voter's curse. *The American economic review*, pages 408–424, 1996.

[116] Ernst Fehr and Klaus M Schmidt. A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868, 1999.

[117] Aris Filos-Ratsikas and Peter Bro Miltersen. Truthful approximations to range voting. In *Proceedings of the 10th Conference on Web and Internet Economics (WINE)*, pages 175–188, 2014.

[118] Aris Filos-Ratsikas, Søren Kristoffer Stiil Frederiksen, and Jie Zhang. Social welfare in one-sided matchings: Random priority and beyond. In *International Symposium on Algorithmic Game Theory*, pages 1–12, 2014.

[119] D. J. Finney. *Probit Analysis*. Cambridge University Press, 3rd edition edition.

[120] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *2016*, pages 144–152, 2016.

[121] Peter C Fishburn. Probabilistic social choice based on simple voting comparisons. *The Review of Economic Studies*, 51(4):683–692, 1984.

[122] David C Fisher and Jennifer Ryan. Tournament games and positive tournaments. *Journal of Graph Theory*, 19(2):217–236, 1995.

[123] J. S. Fishkin. *When the People Speak: Deliberative Democracy and Public Consultation*. Oxford University Press, 2009.

[124] James S. Fishkin. *Democracy When the People Are Thinking: Revitalizing Our Politics Through Public Deliberation*. Oxford University Press. doi: 10.1093/oso/9780198820291.001.0001.

[125] James S Fishkin. *Democracy and deliberation: New directions for democratic reform*. Yale University Press, 1991.

[126] James S. Fishkin and Robert C. Luskin. Experimenting with a democratic ideal: Deliberative polling and public opinion. 40(3):284–298. doi: 10/dmkk5q.

[127] B. Flanigan, P. Gölz, A. Gupta, and A. D. Procaccia. Neutralizing self-selection bias in sampling for sortition. In *34th*, 2020.

[128] Bailey Flanigan, Paul Gölz, Anupam Gupta, and Ariel D Procaccia. Neutralizing self-selection bias in sampling for sortition. In *34th*, pages 6528–6539, 2020.

[129] Bailey Flanigan, Paul Gölz, Anupam Gupta, Ariel D. Procaccia, and Gili Rusak. Panelot, 2020. http://www.panelot.org/.

[130] Bailey Flanigan, Paul Gölz, Anupam Gupta, Brett Hennig, and Ariel D Procaccia. Fair algorithms for selecting citizens assemblies. *Nature*, 596(7873):548–552, 2021.

[131] Bailey Flanigan, Gregory Kehne, and Ariel D Procaccia. Fair sortition made transparent. *Advances in Neural Information Processing Systems*, 34:25720–25731, 2021.

[132] Bailey Flanigan, Daniel Halpern, and Alexandros Psomas. Smoothed analysis of social choice revisited. In *International Conference on Web and Internet Economics*, pages 290–309. Springer, 2023.

[133] Bailey Flanigan, Ananya A Joshi, Sara McAllister, and Catalina Vajiac. CS-JEDI: Required DEI Education, by CS PhD Students, for CS PhD Students. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, pages 87–93, 2023.

[134] Bailey Flanigan, Ariel D Procaccia, and Sven Wang. Distortion under public-spirited voting. In *Proceedings of the 24th ACM Conference on Economics and Computation*, page 700, 2023.

[135] Bailey Flanigan, Jennifer Liang, Ariel D Procaccia, and Sven Wang. Manipulation-robust selection of citizens assemblies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[136] Marc Fleurbaey. Weighted majority and democratic theory. Manuscript, 2008.

[137] Organisation for Economic Co-operation and Development. Eight ways to institutionalise deliberative democracy.

[138] U.S. District Court for the District of Massachusetts. Jury duty excuses, 2023. URL https://www.mad.uscourts.gov/jurors/jury-duty-excused.htm. Accessed 2024-05-23.

[139] Rupert Freeman, Nisarg Shah, and Rohit Vaish. Best of both worlds: ex-ante and ex-post fairness in resource allocation. In *21st*, pages 21–22, 2020.

[140] Archon Fung. The principle of affected interests: An interpretation and defense. *Representation: Elections and beyond*, page 236, 2013.

[141] Rajiv Gandhi, Samir Khuller, Srinivasan Parthasarathy, and Aravind Srinivasan. Dependent rounding and its applications to approximation algorithms. *Journal of the ACM (JACM)*, 53(3):324–360, 2006.

[142] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. A Series of Books in the Mathematical Sciences. Freeman, 1979.

[143] Adela Gąsiorowska. Sortition and its principles: Evaluation of the selection processes of citizens assemblies. *Journal of Deliberative Democracy*, 19(1), 2023.

[144] J. Gastil, C. Bacci, and M. Dollinger. Is deliberation neutral? patterns of attitude change during the deliberative polls. *Journal of public deliberation*, 6(2), 2010.

[145] William V Gehrlein. Condorcet's paradox and the likelihood of its occurrence: different perspectives on balanced preferences. *Theory and decision*, 52:171–199, 2002.

[146] Louis-Gaëtan Giraudet, Bénédicte Apouey, Hazem Arab, Simon Baeckelandt, Philippe Begout, Nicolas Berghmans, Nathalie Blanc, Jean-Yves Boulin, Eric Buge, Dimitri Courant, et al. co-construction in deliberative democracy: lessons from the french citizens convention for climate. *Humanities and Social Sciences Communications*, 9(1):1–16, 2022.

[147] V. Gkatzelis, D. Halpern, and N. Shah. Resolving the optimal metric distortion conjecture. In *61st*, pages 1427–1438, 2020.

[148] Ambros Gleixner, Gregor Hendel, Gerald Gamrath, Tobias Achterberg, Michael Bastubbe, Timo Berthold, Philipp M. Christophel, Kati Jarck, Thorsten Koch, Jeff Linderoth, Marco Lübbecke, Hans D. Mittelmann, Derya Ozyurt, Ted K. Ralphs, Domenico Salvagnin, and Yuji Shinano. MIPLIB 2017: Data-Driven Compilation of the 6th Mixed-Integer Programming Library. doi: 10/gmb3xb.

[149] Ashish Goel, Reyna Hulett, and Anilesh K. Krishnaswamy. Relating metric distortion and fairness of social choice rules. In *Proceedings of the 13th Workshop on Economics of Networks, Systems and Computation*, NetEcon '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359160. doi: 10.1145/3230654.3230658. URL https://doi.org/10.1145/3230654.3230658.

[150] Jonathan Goldman and Ariel D. Procaccia. Spliddit: Unleashing fair division algorithms. 13(2):41–46. doi: 10/gn8t3j.

[151] P. Gölz, A. Kahng, and A. D. Procaccia. Paradoxes in Fair Machine Learning. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, pages 8340–8350, 2019.

[152] Paul Gölz. *Social Choice for Social Good: Proposals for Democratic Innovation from Computer Science*. PhD thesis, Carnegie Mellon University, 2022.

[153] Paul Gölz and Gili Rusak. Panelot, 2020. URL http://www.panelot.org/.

[154] Paul Gölz, Anson Kahng, Simon Mackenzie, and Ariel D Procaccia. The fluid mechanics of liquid democracy. *ACM Transactions on Economics and Computation*, 9(4):1–39, 2021.

[155] Yannai A Gonczarowski, Gregory Kehne, Ariel D Procaccia, Ben Schiffer, and Shirley Zhang. The distortion of binomial voting defies expectation. *Advances in Neural Information Processing Systems*, 36, 2024.

[156] Jacek Gondzio, Pablo González-Brevis, and Pedro Munari. Large-scale optimization with the primal-dual column generation method. 8(1):47–82. ISSN 1867-2949, 1867-2957. doi: 10/gqcq9r.

[157] G. Grimmett. Stochastic Apportionment. *The American Mathematical Monthly*, 111(4): 299–307, 2004.

[158] K. Grönlund, K. Herne, and M. Setälä. Empathy in a citizen deliberation experiment. *Scandinavian Political Studies*, 40(4):457–480, 2017.

[159] Paul Gölz, Anson Kahng, Simon Mackenzie, and Ariel D. Procaccia. The Fluid Mechanics of Liquid Democracy. In *Proceedings of the Conference on Web and Internet Economics (WINE)*.

[160] Daniel Halpern and Nisarg Shah. Fair and efficient resource allocation with partial information. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 224–230, 2021.

[161] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 111–122, 2016.

[162] Ararat Harutyunyan, Tien-Nam Le, Alantha Newman, and Stéphan Thomassé. Domination and fractional domination in digraphs. *arXiv preprint arXiv:1708.00423*, 2017.

[163] Brett Hennig and Paul Gölz. Sortition foundation stratification application, 2020. https://github.com/sortitionfoundation/stratification-app.

[164] Brett Hennig and Paul Gölz. StratifySelect, 2021. URL https://github.com/sortitionfoundation/stratification-app.

[165] D. Holt and D. Elliot. Methods of Weighting for Unit Non-Response. *The Statistician*, 40 (3):333, 1991. ISSN 00390526.

[166] L. Hu and Y. Chen. Fair Classification and Social Welfare. In *Proceedings of the 3rd Annual ACM Conference on Fairness, Accountability, and Transparency*, pages 535–545, 2020.

[167] David Imamura. The rise and fall of redistricting commissions, 2022. URL https://www.americanbar.org/groups/crsj/publications/human_rights_magazine_home/economics-of-voting/the-rise-and-fall-of-redistricting-commissions/.

[168] Ipsos. The state of democracy. Technical report, 2023. URL https://www.ipsos.com/sites/default/files/ct/news/documents/2023-12/Ipsos-KnowledgePanel-TheStateOfDemocracy.pdf.

[169] The Irish Citizens' Assembly Project, 2019. http://www.citizenassembly.ie/work/.

[170] Vincent Jacquet. The role and the future of deliberative mini-publics: A citizen perspective. 67(3):639–657. doi: 10/gqcq9c.

[171] Samara Jones and Guillem Fernàndez. Mean streets: A report on the criminalisation of homelessness in europe. *European Journal of Homelessness _ Volume*, 8 (2), 2014. URL https://www.housingrightswatch.org/sites/default/files/Mean%20Streets%20-%20Full.pdf.

[172] Anson Kahng and Gregory Kehne. Worst-case voting when the stakes are high. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5100–5107, 2022.

[173] O. E. Kangas. Self-interest and the common good: The impact of norms, selfishness and context in social policy opinions. *The Journal of Socio-Economics*, 26(5):475–494, 1997.

[174] Richard M. Karp and Richard J. Lipton. Some connections between nonuniform and uniform complexity classes. In *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*, pages 302–309. doi: 10/djmpjb.

[175] C. F. Karpowitz and T. Mendelberg. An experimental approach to citizen deliberation. *Cambridge Handbook of Experimental Political Science*, pages 258–272, 2011.

[176] Christopher F Karpowitz and Chad Raphael. Ideals of inclusion in deliberation. *Journal of Deliberative Democracy*, 12(2), 2016.

[177] C. Kendall and J. Matsusaka. The common good and voter polarization. Technical report, Mimeo, University of Southern California, 2021.

[178] Ryan Kennedy, William Minozzi, Michael Neblo, and Bailey Flanigan. Chile deliberative town halls - government perceptions. https://osf.io/bejm9, 2024. Accessed: 2024-04-20.

[179] D. R. Kinder and D. R. Kiewiet. Sociotropic politics: the american case. *British Journal of Political Science*, 11(2):129–161, 1981.

[180] F. E. Kizilkaya and D. Kempe. Plurality veto: A simple voting rule achieving optimal metric distortion. In *31st*, pages 349–355, 2022.

[181] Germain Kreweras. Aggregation of preference orderings. In *Mathematics and Social Sciences I: Proceedings of the seminars of Menthon-Saint-Bernard*, pages 73–79, 1965.

[182] D. Kurokawa, A. D. Procaccia, and N. Shah. Leximin allocations in the real world. *ACM Transactions on Economics and Computation (TEAC)*, 6(3-4):1–24, 2018.

[183] Gilbert Laffond, Jean-Francois Laslier, and Michel Le Breton. The bipartisan set of a tournament game. *Games and Economic Behavior*, 5(1):182–201, 1993.

[184] S. Lalley and E. G. Weyl. Quadratic Voting: How Mechanism Design Can Radicalize Democracy. SSRN Scholarly Paper ID 2003531, Social Science Research Network, Rochester, NY, December 2017.

[185] T. Lancaster and G. Imbens. Case-Control Studies with Contaminated Controls. *Journal of Econometrics*, 71(1-2):145–160, 1996.

[186] J. Ledyard. Public goods: A survey of experimental research. In J. Kagel and A. Roth, editors, *Handbook of Experimental Economics*. Princeton University Press, 1997.

[187] Julien Lesca and Patrice Perny. LP Solvable Models for Multiagent Fair Allocation Problems. In *ECAI*, volume 2010, pages 393–398.

[188] Dominique Leydet. Which conception of political equality do deliberative mini-publics promote? 18(3):349–370. ISSN 1474-8851, 1741-2730. doi: 10/gf4m6v.

[189] A. Lindbeck and J. W. Weibull. Altruism and time consistency: the economics of fait accompli. *Journal of Political Economy*, 96(6):1165–1182, 1988.

[190] Lydia T Liu, Nikhil Garg, and Christian Borgs. Strategic ranking. In *International Conference on Artificial Intelligence and Statistics*, pages 2489–2518, 2022.

[191] Ping Liu, Karthik Shivaram, Aron Culotta, Matthew A Shapiro, and Mustafa Bilgic. The interaction between political typology and filter bubbles in news recommendation algorithms. In *Proceedings of the Web Conference 2021*, pages 3791–3801, 2021.

[192] Virpi Lund and Soile Juujärvi. Deliberation in workshops in the participatory budgeting process. 2022.

[193] Thodoris Lykouris and Sergei Vassilvitskii. Competitive caching with machine learned advice. *Journal of the ACM (JACM)*, 68(4):1–25, 2021.

[194] Thodoris Lykouris and Wentao Weng. Learning to defer in content moderation: The human-ai interplay. *arXiv preprint arXiv:2402.12237*, 2024.

[195] M. Magdon-Ismail and L. Xia. A Mathematical Model for Optimal Decisions in a Representative Democracy. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pages 4702–4711, 2018.

[196] Sam J. Maglio and Evan Polman. Revising probability estimates: Why increasing likelihood means increasing impact. *Journal of Personality and Social Psychology*, 111(2):141, 2016.

[197] D. Mandal, A. D. Procaccia, N. Shah, and D. Woodruff. Efficient and Thrifty Voting by Any Means Necessary. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, pages 7178–7189, 2019.

[198] Jane Mansbridge. Should blacks represent blacks and women represent women? A contingent yes. 61(3):628–657. doi: 10/fmcb36.

[199] Jane Mansbridge. Should blacks represent blacks and women represent women? a contingent" yes". *The Journal of politics*, 61(3):628–657, 1999.

[200] Jane Mansbridge. A minimalist definition of deliberation. *Deliberation and development: Rethinking the role of voice and collective action in unequal societies*, pages 27–50, 2015.

[201] MASS LBP. How to run a Civic Lottery: Designing fair selection mechanisms for deliberative public processes.

[202] Kimberly Matheson, Ann Seymour, Jyllenna Landry, Katelyn Ventura, Emily Arsenault, and Hymie Anisman. Canadas colonial genocide of indigenous peoples: A review of the psychosocial and neurobiological processes linking trauma and intergenerational outcomes. *International journal of environmental research and public health*, 19(11):6455, 2022.

[203] Spencer McKay and Peter MacLeod. MASS LBP and Long-Form Deliberation in Canada. 5 (2):108–113. ISSN 2332-8894, 2332-8908. doi: 10/gqcq9p.

[204] T. Mendelberg. The deliberative citizen: Theory and evidence. *Political Decision Making, Deliberation and Participation*, 6(1):151–193, 2002.

[205] H. Moulin. Choosing from a tournament. *Social Choice and Welfare*, 3(4):271–291, 1986.

[206] Hervé Moulin. *Fair division and collective welfare*. MIT press, 2003.

[207] Ishmael Mugari and Emeka E Obioha. Predictive policing and crime control in the united states of america and europe: trends in a decade of research and the future of predictive policing. *Social sciences*, 10(6):234, 2021.

[208] E. Muller and M. A. Satterthwaite. The equivalence of strong positive association and strategy-proofness. *Journal of Economic Theory*, 14(2):412–418, 1977.

[209] Dritan Nace and James B. Orlin. Lexicographically Minimum and Maximum Load Linear Programming Problems. 55(1):182–187. ISSN 0030-364X, 1526-5463. doi: 10/cw6gfh.

[210] Sofia Näsström and Sara Kalm. A democratic critique of precarity. *Global Discourse*, 5(4): 556–573, 2015.

[211] Michael A Neblo, Kevin M Esterling, Ryan P Kennedy, David MJ Lazer, and Anand E Sokhey. Who wants to deliberate – and why? *American Political Science Review*, 104 (3):566–583, 2010.

[212] Michael A Neblo, Kevin M Esterling, and David MJ Lazer. *Politics with the people: Building a directly representative democracy*, volume 555. Cambridge University Press, 2018.

[213] New York City Department of Education. Random selection in admissions, 2021. https://www.schools.nyc.gov/enrollment/enroll-grade-by-grade/how-students-get-offers-to-doe-public-schools/random-numbers-in-admissions.

[214] newDemocracy Foundation and United Nations Democracy Fund. Enabling National Initiatives to Take Democracy Beyond Elections.

[215] Christoph Niessen and Min Reuchamps. Designing a permanent deliberative citizens' assembly: The ostbelgien modell in belgium. 2019.

[216] NSD - Norwegian Centre for Research Data. European Social Survey Round 8 Data, 2016. Data file edition 2.1.

[217] OECD. Government at a glance 2019. Technical report, Organisation for Economic Co-operation and Development (OECD), Paris, 2019. URL ,https://doi.org/10.1787/8ccf5c38-en.

[218] Of By For *. Democratic lottery — the citizens' panel on covid-19, 2020. URL https://joinofbyfor.org/panel/.

[219] OIPD. The ostbelgien model: a long-term citizens' council combined with short-term citizens' assemblies. Available at https://oidp.net/en/practice.php?id=1237 (2024/02/11), 2024.

[220] Naomi O'Leary. The myth of the citizen's assembly, 2019. URL https://www.politico.eu/article/the-myth-of-the-citizens-assembly-democracy/.

[221] John E Olson and Joel H Spencer. Balancing families of sets. *Journal of Combinatorial Theory, Series A*, 25(1):29–37, 1978.

[222] Organisation for Economic Co-operation and Development. *Innovative Citizen Participation and New Democratic Institutions: Catching the Deliberative Wave*. OECD. doi: 10.1787/339306da-en.

[223] Organizing Engagement. Participatory budgeting, 2023. URL https://organizingengagement.org/models/participatory-budgeting/.

[224] Joel Matthew Parker. Randomness and legitimacy in selecting democratic representatives.

[225] Participedia. Participedia. https://participedia.net/search?selectedCategory=case&query=deliberation, 2023.

[226] Dominik Peters and Piotr Skowron. Equal shares, 2023. URL https://equalshares.net/. Accessed on 2024-05-23.

[227] Dominik Peters, Grzegorz Pierczyński, and Piotr Skowron. Proportional participatory budgeting with additive utilities. *Advances in Neural Information Processing Systems*, 34: 12726–12737, 2021.

[228] David Pettinicchio. Why disabled americans remain second-class citizens. *The Washington Post*, July 2019. URL https://www.washingtonpost.com/outlook/2019/07/23/why-disabled-americans-remain-second-class-citizens/.

[229] Hanna F Pitkin. *The concept of representation*. Univ of California Press, 2023.

[230] Eric A Posner and E Glen Weyl. Voting squared: Quadratic voting in democratic politics. *Vand. L. Rev.*, 68:441, 2015.

[231] A. D. Procaccia and J. S. Rosenschein. The distortion of cardinal preferences in voting. In *10th*, pages 317–331, 2006.

[232] Ariel Procaccia. Citizens' assemblies are upgrading democracyfair algorithms are part of the program. *Scientific American*, 2022. URL https://www.scientificamerican.com/article/citizens-assemblies-are-upgrading-democracy-fair-algorithms-are-part-of-the-program/. Accessed: 2024-04-22.

[233] G. Raab, K. Buckner, S. Purdon, and I. Waterston. Adjusting for Non-Response by Weighting. In *Practical Exemplars for Survey Analysis*. 2009.

[234] Douglas W. Rae. *Equalities*. Harvard University Press.

[235] Randall County. Frequently asked questions. https://www.randallcounty.gov/Faq.aspx?QID=102, 2024. Accessed: 2024-04-22.

[236] Sara K Rankin. Punishing homelessness. *New Criminal Law Review*, 22(1):99–135, 2019.

[237] Min Reuchamps. Belgium's experiment in permanent forms of deliberative democracy, 2020. URL http://constitutionnet.org/news/belgiums-experiment-permanent-forms-deliberative-democracy.

[238] Simon Rey and Jan Maly. The (computational) social choice take on indivisible participatory budgeting. arXiv:2303.00621, 2023.

[239] Simon Rey, Ulle Endriss, and Ronald de Haan. Designing participatory budgeting mechanisms grounded in judgment aggregation. In *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning*, pages 692–702, 2020.

[240] Valerie F Reyna and Charles J Brainerd. Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, 18(1):89–107, 2008.

[241] Ronald L Rivest and Emily Shen. An optimal single-winner preferential voting system based on game theory. In *Proc. of 3rd International Workshop on Computational Social Choice*, pages 399–410, 2010.

[242] A. E. Roth, T. Sönmez, and M. U. Ünver. Pairwise kidney exchange. *Journal of Economic Theory*, 125:151–188, 2005.

[243] Alvin E. Roth, Tayfun Sönmez, and M. U. Ünver. Pairwise kidney exchange. 125(2):151–188. ISSN 00220531. doi: 10/b4vd2n.

[244] Zachary Roth. Making participatory budgeting work: Experiences from the front lines, 2022. URL https://www.brennancenter.org/our-work/analysis-opinion/making-participatory-budgeting-work-experiences-front-lines.

[245] Sara Sadhwani. Independent redistricting: An insider's view. *The Forum*, 2023. URL https://www.degruyter.com/document/doi/10.1515/for-2022-2063/html.

[246] R. Saran and N. Tumennasan. Whose Opinion Counts? Implementation by Sortition. *Games and Economic Behavior*, 78:72–84, 2013.

[247] Sebastian Schneckenburger, Britta Dorn, and Ulle Endriss. The Atkinson Inequality Index in Multiagent Resource Allocation. In *AAMAS*, pages 272–280. URL https://www.ifaamas.org/Proceedings/aamas2017/pdfs/p272.pdf.

[248] A. Schrijver. *Theory of Linear and Integer Programming*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley, 1986.

[249] Scotland's Climate Assembly. Recommendations for action, 2021. URL https://webarchive.nrscotland.gov.uk/web/20220321142709/https://www.climateassembly.scot/sites/default/files/2021-09/620640_SCT0521502140-001_Scotland%E2%80%99s%20Climate%20Assembly_Final%20Report%20Goals_WEB%20ONLY%20VERSION.pdf.

[250] Graham Smith. *Democratic Innovations: Designing Institutions for Citizen Participation*. Theories of Institutional Design. Cambridge University Press.

[251] Graham Smith and Maija Setälä. *Mini-Publics and Deliberative Democracy*, pages 299–314. Oxford University Press. doi: 10.1093/oxfordhb/9780198747369.013.27.

[252] H. A. Soufiani, D. C. Parkes, and L. Xia. A Statistical Decision-Theoretic Framework for Social Choice. In *Proceedings of the 27th Conference on Neural Information Processing Systems*, pages 3185–3193, 2014.

[253] J. Spencer. Six Standard Deviations Suffice. *Transactions of the American Mathematical Society*, 289(2):679–706, 1985.

[254] Aravind Srinivasan. Improving the discrepancy bound for sparse matrices: Better approximations for sparse lattice approximation problems. In *8th*, pages 692–701, 1997.

[255] Christina Stacy, Martha Fedorowicz, and Rebecca Dedert. Best practices for inclusive participatory budgeting, August 2022. URL https://www.urban.org/sites/default/files/2022-09/Best%20Practices%20for%20Inclusive%20Participatory%20Budgeting.pdf.

[256] Daniel Steel, Naseeb Bolduc, Kristina Jenei, and Michael Burgess. Rethinking Representation and Diversity in Deliberative Minipublics. 16(1):46–57. ISSN 2634-0488. doi: 10/gm7wnc.

[257] P. Stone. *The Luck of the Draw: The Role of Lotteries in Decision Making*. Oxford University Press, 2011.

[258] Peter Stone. Sortition, voting, and democratic equality. 19(3):339–356. ISSN 1369-8230, 1743-8772. doi: 10/gddffq.

[259] Peter Stone. *The luck of the draw: The role of lotteries in decision making*. Oxford University Press, 2011.

[260] Erika Strazzante, Stéphanie Rycken, and Vanessa Winkler. Global north and global south: How climate change uncovers global inequalities. *Generation Climate Europe*, 2022.

[261] Albert Sun. Divide your rent fairly. URL https://www.nytimes.com/interactive/2014/science/rent-division-calculator.html.

[262] Peter Sutton. The politics of suffering: Indigenous policy in australia since the 1970s. In *Anthropological Forum*, volume 11, pages 125–173. Taylor & Francis, 2001.

[263] Nimrod Talmon and Piotr Faliszewski. A framework for approval-based budgeting methods. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*, pages 2181–2188, 2019.

[264] Zhanna Terechshenko, Fridolin Linder, Vishakh Padmakumar, Michael Liu, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. A comparison of methods in political science text classification: Transfer learning language models for politics. *Available at SSRN 3724644*, 2020.

[265] Andrew O'Donohue Thomas Carothers. How to understand the global spread of political polarization, 2019. URL https://carnegieendowment.org/2019/10/01/how-to-understand-global-spread-of-political-polarization-pub-79893.

[266] William Thomson. Introduction to the theory of fair allocation., 2016.

[267] Tierney Sneed. Redistricting commission takeaways: Successes and failures, 2022. URL https://www.cnn.com/2022/06/18/politics/redistricting-commission-takeaways-success/index.html.

[268] Jason Torchinsky and Dennis W Polio. How independent is too independent?: Redistricting commissions and the growth of the unaccountable administrative state. *Geo. JL & Pub. Pol'y*, 20:533, 2022.

[269] Andrew Trask, Emma Bluemke, Ben Garfinkel, Claudia Ghezzou Cuervas-Mons, and Allan Dafoe. Beyond privacy trade-offs with structured transparency. arXiv 2012.08347, 2020.

[270] Seattle University. Washington state jury summons demographic study interim report. Technical report, Seattle University, College of Arts and Sciences, Department of Criminal Justice, Crime and Justice Research Center, 2022. URL https://www.seattleu.edu/media/college-of-arts-and-sciences/departments/criminaljustice/crimeandjusticeresearchcenter/documents/Washington-State-Jury-Summons-Demographic-Study-Interim-Report-2022.pdf.

[271] Sander Van Der Linden. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature medicine*, 28(3):460–467, 2022.

[272] D. Van Reybrouck. *Against Elections: The Case for Democracy*. Random House, 2016.

[273] D. Wajc. Negative Association – Definition, Properties, and Applications. 2017.

[274] T. Walsh and L. Xia. Lot-Based Voting Rules. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, pages 603–610, 2012.

[275] Brian Wampler, Stephanie McNulty, and Michael Touchton. *Participatory budgeting in global perspective*. Oxford University Press, 2021.

[276] R. Wang, J. S. Fishkin, and R. C. Luskin. Does deliberation increase public-spiritedness? *Social Science Quarterly*, 101(6):2163–2182, 2020.

[277] G. Ward, T. Hastie, S. Barry, J. Elith, and J. R. Leathwick. Presence-Only Data and the EM Algorithm. *Biometrics*, 65(2):554–563, 2009.

[278] Mark E Warren. A problem-based approach to democratic theory. *American Political Science Review*, 111(1):39–53, 2017.

[279] Janith Weerasinghe, Bailey Flanigan, Aviel Stein, Damon McCoy, and Rachel Greenstadt. The pod people: Understanding manipulation of social media popularity via reciprocity abuse. In *Proceedings of The Web Conference 2020*, pages 1874–1884, 2020.

[280] C. R. Weinberg. Applicability of the Simple Independent Action Model to Epidemiologic Studies Involving Two Factors and a Dichotomous Outcome. *American Journal of Epidemiology*, 123(1):162–173, 1986. ISSN 1476-6256, 0002-9262.

[281] Sean Jeremy Westwood, Solomon Messing, and Yphtach Lelkes. Projecting confidence: How the probabilistic horse race confuses and demobilizes the public. *The Journal of Politics*, 82(4):1530–1544, 2020.

[282] Richard Wike and Janell Fetterolf. Global public opinion in an era of democratic anxiety. Technical report, Pew Research Center, December 2021. URL https://www.pewresearch.org/global/2021/12/07/global-public-opinion-in-an-era-of-democratic-anxiety/.

[283] Richard Wike, Janell Fetterolf, Maria Smerkovich, Sarah Austin, Sneha Gubbala, and Jordan Lippert. Representative democracy remains a popular ideal, but people around the world are critical of how its working. Technical report, Pew Research Center, February 2024. URL https://www.pewresearch.org/global/wp-content/uploads/sites/2/2024/02/gap_2024.02.28_democracy-closed-end_report.pdf.

[284] Rich Wilson and Claire Mellier. Getting real about citizens' assemblies: A new theory of change for citizens' assemblies, 2023. URL https://europeandemocracyhub.epd.eu/getting-real-about-citizens-assemblies-a-new-theory-of-change-for-citizens-assemblie

[285] World Bank Group. Gini index (world bank estimate), 2022. URL https://data.worldbank.org/indicator/SI.POV.GINI.

[286] L Xia and V Conitzer. Determining possible and necessary winners under common voting rules given partial orders. a longer unpublished version of [38], 2010.

[287] Lirong Xia. The smoothed possibility of social choice. *Advances in Neural Information Processing Systems*, 33:11044–11055, 2020.

[288] Mingzhu Yao and Donggen Wang. Modeling household relocation choice: An egalitarian bargaining approach and a comparative study. *Journal of Transport and Land Use*, 14(1): 625–645, 2021.

[289] Moira L Zellner. Participatory modeling for collaborative landscape and environmental planning: From potential to realization. *Landscape and Urban Planning*, 247:105063, 2024.

[290] I. Zettler, B. E. Hilbig, and J. Haubrich. Altruism at the ballots: Predicting political attitudes and behavior. *Journal of Research in Personality*, 45(1):130–133, 2011.

# Part V

# Appendices

# A

# Chapter 2 Appendix

## A.1 Notation Glossary

| | |
|---|---|
| **Sets of Agents** | |
| $N$ | Set of agents in the population |
| $Recipients$ | Set of agents who receive invitation letters (random variable) |
| $Pool$ | Set of agents in the pool (random variable) |
| $Panel$ | Set of agents on the panel (random variable) |
| **Sortition Panel Parameters** | |
| $n$ | Size of the population |
| $r$ | Number of invitation letters sent out |
| $k$ | Size of the panel |
| $F$ | Set of all features |
| $V_f$ | Set of possible values for a specific feature $f \in F$ |
| $F(i)$ | Feature vector of agent $i$ |
| $n_{f,v}$ | Number of agents in the population with value $v$ of feature $f$ |
| $\ell_{f,v}, u_{f,v}$ | Lower and upper quotas for every feature-value pair |
| $q_i$ | Probability that agent $i \in N$ enters the pool, conditioned on being invited |
| $q^*$ | Minimum value of $q_i$ over all agents ($q^* := \min_{i \in N} q_i$) |
| $\alpha$ | Parameter defined as $\alpha := q^* r/k$ |

## A.2 Supplementary Material for section 2.3

### A.2.1 Discussion of Theorem Preconditions

We show that pools are good with high probability under two preconditions: that each feature-value group constitutes at least $1/k$ fraction of the population (so $n_{f,v}/n \geq 1/k$ for all $f, v$), and that the number of recipients is sufficiently high relative to the participation probabilities and the panel size ($\alpha = q^* r/k \to \infty$).

The first condition is natural because if a group should proportionally receive less than one seat on the panel, any positive lower bound on selection probabilities for agents in groups would violate proportionality.

The second condition enforces that the number of agents invited $r$ is large enough relative to the minimum participation probability $q^*$ and the size of the panel. Without this condition, there can be a constant probability that the pool will feature zero agents with a certain feature-value: Suppose that $\alpha$ is an arbitrary positive constant, set all $q_i := \alpha\, k/r$, and consider a feature-value pair $f, v$ with $n_{f,v} = n/k$ agents. In expectation, there will be $(r/n)\,(n/k) = r/k$ agents with feature-value $f, v$ among the recipients. If $r \in \omega(k)$, there are at most $2\,r/k$ such recipients with high probability. Then, the probability that the pool contains no agent with $f, v$ is at least

$$(1 - \alpha\, k/r)^{2\,r/k} = (1 - q_i)^{2\,\alpha/q_i} = \Big( \underbrace{(1 - q_i)^{1/q_i}}_{\to\, 1/e \text{ as } q_i \to 0} \Big)^{2\,\alpha} \to e^{-2\,\alpha} > 0.$$

### A.2.2 Discussion of Ties to Discrepancy Theory

In rounding agents' marginal selection probabilities to select a panel, we round fractional variables to 0 or 1 such that the sum of certain sets of variables changed only by a small amount. This problem is closely connected to *combinatorial discrepancy* [77, 253], which can be summarized in the same words, by additionally assuming that the initial fractional values are 1/2. In fact, the original Beck-Fiala theorem arises in the context of discrepancy, showing that, if each variable appears in a bounded number $t$ of sets, discrepancy $\Theta(t)$ can be achieved (where in our setting, $t$ corresponds to $|F|$, the number of features). Beck and Fiala [40] conjectured that it is actually possible to achieve discrepancy in $O(\sqrt{t})$. Should this conjecture be true, similar ideas might translate to our setting to guarantee the satisfaction of quotas closer to exact proportionality. To this day, however, the best known bound in $t$ is still in $\Theta(t)$ [64]. In accordance with this result, we guarantee a relaxation of $|F|$ from proportional representation of groups.

We note that there do exist other discrepancy results that give sub-linear dependencies on $|F|$, but at the cost of introducing dependencies on other parameters. One such result is Theorem 5.3 in [34], which guarantees discrepancy a square-root dependency on $|F|$. However, subject to our requirement that the per-person marginal probability must deviate from $k/n$ by only $\pm\delta k/n$ where $\delta \in o(1)$, Bansal's result guarantees a discrepancy bound of $O(\sqrt{|F| \log(kn/\delta)})$, which grows in $n$, making it unfavorable in our setting.

### A.2.3 Proof of lemma 2.3.2

The results in this section allow $k \geq 1$ and $r \geq 1$ to vary arbitrarily in $n$; they just require that $\alpha := q^* r/k \to \infty$ as $n \to \infty$ (without requiring $\alpha$ to grow at a specific minimum rate relative to $n$). All convergences are relative to $n$ going to infinity.

**Lemma A.2.2.** *Suppose that $\alpha \to \infty$ and $n_{f,v} \geq n/k$ for all $f, v$. Then, for all agents $i \in Population$, $\mathbb{P}[Pool \text{ is good} \mid i \in Pool] \to 1$.*

In the following proofs, it is convenient to refer to $1/q^*$, the largest possible value of $a_i$, as $a^*$. Note that $a^* = \frac{r}{\alpha k}$. We will refer to the random set of recipients with a certain feature-value pair $f, v$ as $Recipients_{f,v} := \{i \in Recipients \mid f(i) = v\}$.

We begin by showing in lemmas A.2.1 and A.2.3 that, conditioned on $i$ being in the pool, the following three events occur with high probability:

**A.** $k\, a^* \leq \sum_{j \in Pool} a_j$

**B.** $\sum_{j \in Pool} a_j \in [(1 - \alpha^{-.492})\, r, (1 + \alpha^{-.492})\, r]$

**C.** $\sum_{j \in Pool : f(j) = v} a_j \in [(1 - \alpha^{-.492}) \frac{n_{f,v}}{n} r, (1 + \alpha^{-.492}) \frac{n_{f,v}}{n} r] \quad \forall f, v$

We then show in lemma A.2.4 that, when these events occur on some pool, the pool must be good, which concludes the proof of lemma 2.3.2.

**Lemma A.2.1.** *Under the assumptions of lemma 2.3.2, $\mathbb{P}$ [**Event A** $\wedge$ **Event B** $\mid i \in Pool] \to 1$.*

*Proof.* Fix the set of recipients $R$ (including $i$). With respect to the randomness in the pool self-selection, the random variables $a_j \cdot \mathbb{1}\{j \in Pool\}$ across all $j \in R \setminus \{i\}$ are independent, bounded in $[0, a^*]$, and have expected value $a_j\, q_j = 1$. Thus, by a Chernoff bound, and using that $a^* = r/(\alpha\, k)$,

$$\begin{aligned}
\mathbb{P}\left[\left|\sum_{j \in Pool \setminus \{i\}} a_j - (r-1)\right| \geq \alpha^{-.495}\,(r-1)\right] &\leq 2\, e^{-\alpha^{-.99} \frac{r-1}{a^*}/3} \\
&= 2\, e^{-\alpha^{-.99} \frac{r-1}{r} \alpha\, k/3} \\
&\leq 2\, e^{-\Omega(\alpha^{.01})} \to 0,
\end{aligned}$$

where the last inequality uses the fact that $r \geq 2$ for large enough $n$[1] and that $k \geq 1$.

Conditioning on this high-probability event, it follows that, for large enough $n$,

$$\sum_{j \in Pool} a_j \geq 1 + \sum_{j \in Pool \setminus \{i\}} a_j \geq 1 + (1 - \alpha^{-.495})\,(r-1) \geq (1 - \alpha^{-.492})\, r,$$

[1] Since $r = \alpha\, k/q^* \geq \alpha/q^* \geq \alpha \to \infty$.

which shows the lower bound in Event B. For the upper bound,

$$\sum_{j \in Pool} a_j \le a^* + \sum_{j \in Pool \setminus \{i\}} a_j \le a^* + (1 + \alpha^{-.495})\,(r-1) \le r/(\alpha\, k) + (1 + \alpha^{-.495})\, r$$
$$\le (1 + \alpha^{-.495} + 1/\alpha)\, r \le (1 + \alpha^{-.492})\, r \le 1/(1 - \alpha^{-.492})\, r.$$

This establishes Event B.

For large enough $n$, the lower bound on $\sum_{j \in Pool} a_j$ can be extended as

$$\sum_{j \in Pool} a_j \ge (1 - \alpha^{-.492})\, r \ge r/\alpha \ge k\, a^*,$$

which shows Event A. □

For Event C, we need to show that $\sum_{j \in Pool: f(j)=v} a_j$ is concentrated for a feature-value pair $f, v$. As an intermediate step, we first show that the *number* of pool members ("$\sum_{j \in Pool: f(j)=v} 1$") with this feature-value pair is concentrated:

**Lemma A.2.2.** *Under the assumptions of lemma 2.3.2, for each $f, v$,*

$$\mathbb{P}\left[(1 - \alpha^{-.495})\, \frac{n_{f,v}}{n}\, r \le \left|Recipients_{f,v}\right| \le (1 + \alpha^{-.495})\, \frac{n_{f,v}}{n}\, r \,\middle|\, i \in Pool\right] \to 1.$$

*Proof.* Conditioned on $i \in Pool \subseteq Recipients$, $Recipients \setminus \{i\}$ is distributed as if $r-1$ members of $Population \setminus \{i\}$ were drawn with equal probability and without replacement. Thus,

$$\mathbb{E}\left[\left|Recipients_{f,v}\right| \,\middle|\, i \in Pool\right] = \begin{cases} n_{f,v}\, \frac{r-1}{n-1} & \text{if } f(i) \ne v \\ 1 + (n_{f,v} - 1)\, \frac{r-1}{n-1} & \text{if } f(i) = v. \end{cases}$$

In both cases, we show that $\mathbb{E}\left[\left|Recipients_{f,v}\right| \,\middle|\, i \in Pool\right] \in [(1 - k/r)\, n_{f,v}\, \frac{r}{n}, (1 + k/r)\, n_{f,v}\, \frac{r}{n}]$. Indeed, for the upper bound,

$$\mathbb{E}\left[\left|Recipients_{f,v}\right| \,\middle|\, i \in Pool\right] \le 1 + (n_{f,v} - 1)\, \frac{r-1}{n-1} \le 1 + n_{f,v}\, \frac{r}{n} = \left(1 + \frac{n}{n_{f,v}}/r\right)\, n_{f,v}\, \frac{r}{n}$$
$$\le (1 + k/r)\, n_{f,v}\, \frac{r}{n} \le (1 + 1/\alpha)\, n_{f,v}\, \frac{r}{n}.$$

For the lower bound,

$$\mathbb{E}\left[\left|Recipients_{f,v}\right| \,\middle|\, i \in Pool\right] \ge n_{f,v}\, \frac{r-1}{n-1} = \frac{r-1}{r}\, n_{f,v}\, \frac{r}{n} = (1 - 1/r)\, n_{f,v}\, \frac{r}{n}$$
$$\ge (1 - k/r)\, n_{f,v}\, \frac{r}{n} \ge (1 - 1/\alpha)\, n_{f,v}\, \frac{r}{n}.$$

As the (independent) union of the deterministic set $\{i\}$ and indicator variables for sampling without replacement, the variables $\mathbb{1}\{j \in Recipients\}$ satisfy negative association and therefore Chernoff inequalities [273]. Thus, for the upper tail bound,

$$\mathbb{P}\left[\left|Recipients_{f,v}\right| \geq (1+\alpha^{-.497})\,(1+1/\alpha)\,n_{f,v}\,\frac{r}{n}\middle| i \in Pool\right] \leq e^{-\alpha^{-.994}\,(1+1/\alpha)\,n_{f,v}\,\frac{r}{n}/3}$$
$$\leq e^{-\alpha^{-.994}\,n_{f,v}\,\frac{r}{n}/3} \leq e^{-\alpha^{-.994}\,\frac{r}{k}/3} \leq e^{-\alpha^{-.994}\,\alpha/3} \leq e^{-\alpha^{.006}/3} \to 0.$$

Similarly, for the lower tail bound,

$$\mathbb{P}\left[\left|Recipients_{f,v}\right| \leq (1-\alpha^{-.497})\,(1-1/\alpha)\,n_{f,v}\,\frac{r}{n}\middle| i \in Pool\right] \leq e^{-\alpha^{-.994}\,(1-1/\alpha)\,n_{f,v}\,\frac{r}{n}/2}$$
$$\overset{(\alpha\geq 3)}{\leq} e^{-\alpha^{-.994}\,n_{f,v}\,\frac{r}{n}/3} \leq e^{-\alpha^{.006}/3} \to 0.$$

The claim follows from observing that, for $r/k$ large enough,

$$(1-\alpha^{-.497})\,(1-1/\alpha) \geq 1-\alpha^{-.497}-\alpha^{-1} \geq 1-\alpha^{-.495}$$

and

$$(1+\alpha^{-.497})\,(1+1/\alpha) = 1+\alpha^{-.497}+\alpha^{-1}+\alpha^{-1.497} \leq 1-\alpha^{-.495}.$$ □

**Lemma A.2.3.** *Under the assumptions of lemma 2.3.2,* $\mathbb{P}\left[\textbf{\textit{Event C}} \mid i \in Pool\right] \to 1$.

*Proof.* Fix a single feature-value pair $f, v$. By lemma A.2.2, with high probability, the number of recipients $r_{f,v}$ with feature-value pair $f, v$ is in

$$\left[(1-\alpha^{-.495})\,\frac{n_{f,v}}{n}\,r, (1+\alpha^{-.495})\,\frac{n_{f,v}}{n}\,r\right].$$

Going forward, we will fix a set of recipients $R$, and we assume that $r_{f,v}$ indeed falls in this range. For large enough $n$, this implies that $r_{f,v}$ is positive. For ease of notation, we will implicitly condition on $i \in Pool$ and these high-probability events.

The self-selection process of agents with feature-value pair $f, v$ might look a bit different depending on whether $f(i) = v$. If $f(i) \neq v$, the self selection of agents with feature-value pair $f, v$ is independent from our knowledge about $i$ being in the pool. Thus, the random variable $\sum_{\substack{j\in Pool,\\ f(j)=v}} a_j$ is the sum of independent random variables $a_j\,\mathbb{1}\{j \in Pool\}$ for each $j \in R, f(j) = v$, where each variable is bounded in $[0, a^*]$ and has expectation 1. In particular, $\mathbb{E}\left[\sum_{\substack{j\in Pool,\\ f(j)=v}} a_j\right] = r_{f,v}$.

Else, if $f(i) \neq v$, $\sum_{\substack{j\in Pool,\\ f(j)=v}} a_j$ is still the sum of independent random variables $a_j\,\mathbb{1}\{j \in Pool\}$ and each variable is bounded in $[0, a^*]$. However, the specific variable $a_i\,\mathbb{1}\{i \in Pool\}$ is deterministi-

cally $a_i$ (all other variables still have expectation 1). Thus, $\mathbb{E}\left[\sum_{\substack{j\in Pool,\\ f(j)=v}} a_j\right] = r_{f,v} - 1 + a_i$.

$$\begin{aligned} r_{f,v} - 1 + a_i &= \left(1 + \frac{a_i - 1}{r_{f,v}}\right) r_{f,v} \le \left(1 + \frac{a^*}{r_{f,v}}\right) r_{f,v} \le \left(1 + \frac{r/(\alpha\, k)}{(1-\alpha^{-.495})\, r\, n_{f,v}/n}\right) r_{f,v} \\ &\le \left(1 + \frac{r/(\alpha\, k)}{(1-\alpha^{-.495})\, r/k}\right) r_{f,v} = \left(1 + \frac{1}{(1-\alpha^{-.495})\, \alpha}\right) r_{f,v} \\ &\le (1 + 2/\alpha)\, r_{f,v}. \qquad \text{(for } \alpha^{.495} \ge 2\text{)} \end{aligned}$$

Thus, across both cases, the expectation $\mathbb{E}\left[\sum_{\substack{j\in Pool,\\ f(j)=v}} a_j\right]$ is at least $r_{f,v} \ge (1-\alpha^{-.495})\, \frac{n_{f,v}}{n}\, r$ and at most $(1+2/\alpha)\, r_{f,v} \le (1+2/\alpha)\,(1+\alpha^{-.495})\, \frac{n_{f,v}}{n}\, r \le (1+\alpha^{-.493})\, \frac{n_{f,v}}{n}\, r$ for large $n$, and we can use Chernoff bounds.

For bounding the lower tail,

$$\begin{aligned} &\mathbb{P}\left[\sum_{\substack{j\in Pool,\\ f(j)=v}} a_j \le (1-\alpha^{-.495})\,(1-\alpha^{-.495})\, \frac{n_{f,v}}{n}\, r\right] \le e^{-\alpha^{-.99}\,(1-\alpha^{-.495})\, \frac{n_{f,v}}{n}\, r/(2\,a^*)} \\ &\overset{(\alpha^{.495}\ge 3)}{\le} e^{-\alpha^{-.99}\, \frac{n_{f,v}}{n}\, r/(3\,a^*)} = e^{-\alpha^{-.99}\, \frac{n_{f,v}}{n}\, r/(3\, r/(\alpha\, k))} \le e^{-\alpha^{-.99}\, \frac{n_{f,v}}{n}\, \alpha\, k/3} \\ &\le e^{-\alpha^{-.99}\, \alpha/3} \\ &\le e^{-\alpha^{.01}/3} \to 0. \end{aligned}$$

For bounding the upper tail,

$$\begin{aligned} &\mathbb{P}\left[\sum_{\substack{j\in Pool,\\ f(j)=v}} a_j \ge (1+\alpha^{-.495})\,(1+\alpha^{-.493})\, \frac{n_{f,v}}{n}\, r\right] \le e^{-\alpha^{-.99}\,(1+\alpha^{-.493})\, \frac{n_{f,v}}{n}\, r/(3\,a^*)} \\ &\le e^{-\alpha^{-.99}\, \frac{n_{f,v}}{n}\, r/(3\,a^*)} = e^{-\alpha^{-.99}\, \frac{n_{f,v}}{n}\, \alpha\, k/3} \le e^{-\alpha^{-.99}\, \alpha/3} \le e^{-\alpha^{.01}/3} \to 0. \end{aligned}$$

Note that, for large $n$, $(1-\alpha^{-.495})\,(1-\alpha^{-.495}) \ge 1 - 2\,\alpha^{-.495} \ge 1 - \alpha^{-.492}$. Similarly, $(1+\alpha^{-.495})\,(1+\alpha^{-.493}) \in 1 + O(\alpha^{-.493})) \le 1 + \alpha^{-.492}$.

This shows that, for each $f, v$, $(1-\alpha^{-.492})\, \frac{n_{f,v}}{n}\, r \le \sum_{\substack{j\in Pool,\\ f(j)=v}} a_j \le (1+\alpha^{-.492})\, \frac{n_{f,v}}{n}\, r$ with high probability. The claim follows by a union bound over all (finitely many) feature-value pairs. □

**Lemma A.2.4.** *For large enough n, if Events A, B, and C occur for a pool P, P is good.*

*Proof.* Suppose that Events A, B, and C occur in a pool $P$.

Condition (2.1): $\forall j \in P.\ 0 \leq \pi_{j,P} \leq 1$. Clearly, $\pi_{j,P}$ is nonnegative, and Event A implies that $\pi_{j,P} = k\, a_j / \sum_{j' \in P} a_{j'} \leq k\, a^* / \sum_{j' \in P} a_{j'} \leq 1$.

Condition (2.2): $\forall f, v.\ (1 - \alpha^{-.49})\, k\, n_{f,v}/n \leq \sum_{j \in P: f(j)=v} \pi_{j,P} \leq (1 + \alpha^{-.49})\, k\, n_{f,v}/n$. Fix any feature-value pair $f, v$. Recall that, by Event B,

$$\sum_{j \in P} a_j \in [(1 - \alpha^{-.492})\, r, (1 + \alpha^{-.492})\, r],$$

and, by Event C,

$$\sum_{j \in P: f(j)=v} a_j \in [(1 - \alpha^{-.492})\, \frac{n_{f,v}}{n}\, r, (1 + \alpha^{-.492})\, \frac{n_{f,v}}{n}\, r].$$

Observe that, for any $x \in [0, 1/3]$,

$$\frac{1+x}{1-x} \leq \frac{1 + x + x\,(1 - 3\,x)}{1-x} = \frac{1 + 2\,x - 3\,x^2}{1-x} = 1 + 3\,x.$$

Then, if $n$ is large enough such that $\alpha^{-.492} \leq 1/3$, it follows that

$$\begin{aligned}\sum_{j \in P: f(j)=v} \pi_{j,P} = k\, \frac{\sum_{j \in P: f(j)=v} a_j}{\sum_{j \in P} a_j} &\leq k\, \frac{(1 + \alpha^{-.492})\, \frac{n_{f,v}}{n}\, r}{(1 - \alpha^{-.492})\, r} \leq (1 + 3\,\alpha^{-.492})\, k\, \frac{n_{f,v}}{n} \\ &\leq (1 + \alpha^{-.49})\, k\, \frac{n_{f,v}}{n}.\end{aligned}$$

Next, observe that, for any $x$,

$$\frac{1-x}{1+x} \geq \frac{1 - x - 2\,x^2}{1+x} = 1 - 2\,x.$$

Thus,

$$\begin{aligned}\sum_{j \in P: f(j)=v} \pi_{j,P} = k\, \frac{\sum_{j \in P: f(j)=v} a_j}{\sum_{j \in P} a_j} &\geq k\, \frac{(1 - \alpha^{-.492})\, \frac{n_{f,v}}{n}\, r}{(1 + \alpha^{-.492})\, r} \geq (1 - 2\,\alpha^{-.492})\, k\, \frac{n_{f,v}}{n} \\ &\geq (1 - \alpha^{-.49})\, k\, \frac{n_{f,v}}{n}.\end{aligned}$$

Condition (2.3): $\sum_{i \in P} a_i \leq r/(1 - \alpha^{-.49})$. This follows from Event B since $\sum_{j \in P} a_j \leq (1 + \alpha^{-.492})\, r \leq (1 + \alpha^{-.49})\, r = \frac{1 - \alpha^{-.98}}{1 - \alpha^{-.49}}\, r \leq r/(1 - \alpha^{-.49})$ for large enough $n$. □

### A.2.4 Proof of lemma 2.3.3

#### Rounding the Linear Program Using Discrepancy Methods

In Part II of the algorithm, we need to implement the marginal probabilities $\pi_{i,P}$ from Part I by randomizing over panels of size $k$. Additionally, the panels produced by this procedure should guarantee that the number of panel members of a feature-value pair $(f, v)$ lies in a narrow interval around the proportional number of panel members $k\, n_{f,v}/n$. Technically, this corresponds to randomly rounding the fractional solution $x_i := \pi_{i,P}$ of an LP, such that afterwards all variables are 0 or 1, i.e., indicator variables for membership in a random panel.

Formally, we prove the following lemma:

**Lemma A.2.3.** *There is a polynomial-time sampling algorithm that, given a good pool $P$, produces a random panel $Panel$ such that (1) $\mathbb{P}[i \in Panel] = \pi_{i,P}$ for all $i \in P$, (2) $|Panel| = k$, and (3) $\sum_{i:f(i)=v} \pi_{i,P} - |F| \leq |\{i \in Panel \mid f(i) = v\}| \leq \sum_{i:f(i)=v} \pi_{i,P} + |F|$.*

To round the linear program, we use an iterative rounding procedure based on the famous Beck-Fiala theorem [40]. For ease of exposition, we first describe an algorithm for deterministic rounding and describe in the subsequent subsection how to turn it into a randomized rounding procedure. From here on, we drop the index "$P$" from the marginal probabilities $\pi_{i,P}$, both for ease of notation and to emphasize that the lemma applies to any set of marginal probabilities adding up to $k$ (such other marginals might arise, say, from clipping and rescaling the $\pi_{i,P}$ if some of them are greater than 1).

**Lemma A.2.5.** *For a pool $P$, let $(\pi_i)_{i\in P}$ be any collection of variables in $[0, 1]$ such that $\sum_{i\in P} \pi_i = k$. Then, we can efficiently compute a deterministic 0/1 rounding $(x_i)_{i\in P}$ such that $\sum_{i\in P} x_i = k$ and such that, for each feature-value pair $f, v$,*

$$\sum_{i\in P:f(i)=v} \pi_i - |F| \leq \sum_{i\in P:f(i)=v} x_i \leq \sum_{i\in P:f(i)=v} \pi_i + |F|.$$

*Proof.* We initialize $x_i \leftarrow \pi_{i,P}$, and the following inequalities are therefore satisfied:

$$\sum_{i\in P} x_i = k \tag{A.1}$$

$$\sum_{i\in P:f(i)=v} x_i = \sum_{i\in P:f(i)=v} \pi_{i,P} \qquad \forall f, v. \tag{A.2}$$

We then iteratively update the $x_i$ and maintain a set of equations that starts as the equations in eqs. (A.1) and (A.2), but from which we will iteratively drop some equations of type (A.2). Throughout this process, we maintain that the $x_i$ satisfy all remaining (i.e., not dropped) equations and that $x_i \in [0, 1]$ for all $i$. We call $x_i \in (0, 1)$ *active*; once an $x_i$ stops being active, it stays at its value 0 or 1 to the end of the rounding. We continue our iterative process until no more active variables remain, at which point we return our 0/1 rounding.

Whenever the number of remaining equalities is lower than the number of active agents, the values $x_i$ for the active variables must be underdetermined by the equalities. More precisely, after considering all inactive $x_i$ as constants, the space of remaining $x_i$ that satisfies the remaining equalities forms an affine subspace of non-zero dimension. Since this subspace must intersect the boundary of the unit hypercube, there is a way of updating the $x_i$ such that all equalities are preserved, such that no inactive variable gets changed, and such that at least one additional variable becomes inactive (progress).[1]

Else, we know that the number of active agents $n'$ is at most the number of remaining equalities $m$. If $m = 1$, i.e., if eq. (A.1) is the only remaining equation, there cannot be any active agents since eq. (A.1) can only be satisfied if no $x_i$ or at least two $x_i$ are non-integer. Thus, in the following, $m \geq 2$. For any remaining equality of type (A.2) corresponding to some feature-value pair $f, v$, say that it *ranges over* $t$ many active variables if there are $t$ many active variables $x_i$ such that $f(i) = v$. Should any of the remaining constraints range over all $n'$ many active variables, then this constraint must be implied by constraint (A.1) and the values of the inactive variables. We can thus drop the redundant constraint without consequences (progress), and repeat the iterative process.

If none of these steps apply, we show that some constraint of type (A.2) ranges over at most $|F|$ active variables: Clearly, this is the case if $n' \leq |F|$, and furthermore if $n' = |F| + 1$ because we removed constraints of type (A.2) ranging over all active variables. If $n' > |F| + 1$, note that every active agent appears in at most $|F|$ many equations of type (A.2), at most one per feature. It follows that the total number of active agents summed up over all remaining equalities of this type is at most $n' |F| < n' |F| - (|F| + 1) + n' = (n' - 1)(|F| + 1) \leq (m - 1)(|F| + 1)$, which implies that one of the $m - 1$ equalities of type (A.2) ranges over less than $|F| + 1$ active variables. Drop all such equalities (progress) and repeat.

Since $n' + m$ decreases in every iteration, this algorithm will produce a deterministic panel in polynomial time. Since constraint (A.1) is never dropped, the panel size must be exactly $k$. By how much might the equations of type (A.2) for a feature-value pair $f, v$ be violated in the result? Clearly, they are maintained exactly up to the point where they are dropped.[2] From this point on, however, only $|F|$ many active variables could still change the value of $\sum_{i \in P: f(i)=v} x_i$. Since each of these variables remains in its range $[0, 1]$ throughout the rounding process, the final $x_i$ must satisfy

$$\sum_{i \in P: f(i)=v} \pi_i - |F| \leq \sum_{i \in P: f(i)=v} x_i \leq \sum_{i \in P: f(i)=v} \pi_i + |F|. \qquad \square$$

### Randomizing the Beck-Fiala rounding

We give two methods of transforming the previous deterministic rounding algorithm into a randomized rounding algorithm. To prove lemma 2.3.3, we can directly apply a result by Bansal [34]

[1] This step can be implemented in polynomial time by solving systems of linear equations.

[2] We do not count if the equality was dropped because it was implied by constraint (A.1), in which case it is preserved exactly throughout the rounding.

to our deterministic rounding procedure:

**Lemma A.2.3.** *There is a polynomial-time sampling algorithm that, given a good pool $P$, produces a random panel $Panel$ such that (1) $\mathbb{P}[i \in Panel] = \pi_{i,P}$ for all $i \in P$, (2) $|Panel| = k$, and (3) $\sum_{i:f(i)=v} \pi_{i,P} - |F| \leq |\{i \in Panel \mid f(i) = v\}| \leq \sum_{i:f(i)=v} \pi_{i,P} + |F|$.*

*Proof.* We apply Theorem 1.2 by Bansal [34] to the deterministic rounding procedure of lemma A.2.5. To apply the theorem, we need to give a $\delta > 0$ such that, when there are $n'$ many active variables left, the number of remaining equalities in the next iteration is at most $(1 - \delta)\, n'$ constraints. In lemma A.2.5, we showed that $m$ is always set to a value of at most $n' - 1$. Thus, for $\delta := 1/n$, we get that $m \leq n' - 1 = (1 - 1/n')\, n' \leq (1 - 1/n)\, n'$ and can apply the theorem. □

While the previous algorithm runs in polynomial time, we found an alternative way of randomizing the rounding to be more efficient in practice. This technique is based on naïve column generation, which is not guaranteed to run in polynomial time, but has the following advantages:

- it uses linear programs rather than semi-definite programs,
- instead of a single random panel, the column generation (deterministically) generates a *distribution* over panels, which allows us to analyze the distribution after a single run, and
- there is a continuous progress measure that allows us to stop the optimization process once we implement the $\pi_i$ with sufficient accuracy.

We describe this algorithm in the proof of the following version of lemma 2.3.3, which does not require polynomially-bounded runtime:

**Lemma A.2.6.** *There is a sampling algorithm that, given a good pool $P$, produces a random panel $Panel$ such that (1) $\mathbb{P}[i \in Panel] = \pi_{i,P}$ for all $i \in P$, (2) $|Panel| = k$, and (3) $\sum_{i:f(i)=v} \pi_{i,P} - |F| \leq |\{i \in Panel \mid f(i) = v\}| \leq \sum_{i:f(i)=v} \pi_{i,P} + |F|$.*

*Proof.* First, note that we can strengthen lemma A.2.5 slightly by giving it an arbitrary vector $\vec{c} \in \mathbb{R}^{|P|}$ as part of its input and additionally requiring that $\langle \vec{c}, \vec{x} \rangle \geq \langle \vec{c}, \vec{\pi} \rangle$, where $\vec{x}$ is the vector of $x_i$ and $\vec{\pi}$ the vector of $\pi_i$. This stronger statement follows from the same proof if we require every update of the $x_i$ to additionally maintain that $\langle \vec{c}, \vec{x} \rangle \geq \langle \vec{c}, \vec{\pi} \rangle$. Since this intersects the non-zero dimensional affine subspace formed by the constraints with a half space that contains at least the current point $\vec{x}$, the resulting intersection is still unbounded, which means that we can find an intersection with the boundary of the hypercube. We refer to this procedure as the "modified lemma A.2.5."

Now, let $\mathfrak{B} \neq \emptyset$ be any set of panels satisfying the constraints of the lemma, possibly exponentially

many. Consider the following linear program and its (simplified) dual:

PRIMAL($\mathfrak{B}$):

$$\begin{aligned} &\textit{minimize } \delta \\ &\textit{s.t.} \left| \pi_i - \sum_{B \in \mathfrak{B}: i \in B} \lambda_B \right| \leq \delta \quad \forall i \in P \\ &\sum_{B \in \mathfrak{B}} \lambda_B = 1 \\ &\delta \geq 0, \lambda_B \geq 0 \quad \forall B \in \mathfrak{B} \end{aligned}$$

DUAL($\mathfrak{B}$):

$$\begin{aligned} &\textit{maximize } \left( \sum_{i \in P} \pi_i \, z_i \right) - \hat{z} \\ &\textit{s.t.} \sum_{i \in B} z_i \leq \hat{z} \quad \forall B \in \mathfrak{B} \\ &|z_i| \leq 1 \quad \forall i \in P \end{aligned}$$

The primal LP searches for a distribution over the panels $\mathfrak{B}$ such that the largest absolute deviation between the marginal $\sum_{B \in \mathfrak{B}: i \in B} \lambda_B$ and the target value $\pi_i$ of any $i \in P$ is as small as possible. Let $\overline{\mathfrak{B}}$ denote the set of panels that can be returned by the modified lemma A.2.5, for any vector $\vec{c}$ in its input.

**Observation 1: For any $\mathfrak{B} \neq \emptyset$, the LP has an objective value $obj(\mathfrak{B}) \geq 0$.** Indeed, in the primal, the objective value is clearly bounded below by 0, and the LP is feasible for any distribution over $\mathfrak{B}$ and large enough $\delta$. By strong duality, the dual LP must have the same objective value.

**Observation 2: $obj(\overline{\mathfrak{B}}) = 0$.** For the sake of contradiction, suppose that the objective value was strictly positive, i.e., that $\vec{\pi}$ does not lie in the convex hull of $\overline{\mathfrak{B}}$. Then, there must be a plane separating $\vec{\pi}$ from this convex hull, and an orthogonal vector $\vec{c}$ such that $\langle \vec{c}, \vec{\pi} \rangle > \langle \vec{c}, \vec{x} \rangle$ for any $\vec{x}$ corresponding to a panel in $\overline{\mathfrak{B}}$. Applying the modified lemma A.2.5 with this vector $\vec{c}$ would lead to a contradiction.

Consider algorithm 1, which iteratively generates a subset $\mathfrak{B} \subseteq \overline{B}$ by column generation.

**Algorithm 1** Column generation

$\mathfrak{B} \leftarrow$ {result of running modified lemma A.2.5 with arbitrary $\vec{c}$} **while** $obj(\mathfrak{B}) > 0$ **do**
  fix optimal values $z_i, \hat{z}$ for DUAL($\mathfrak{B}$) $B \leftarrow$ result of running modified lemma A.2.5 with $\vec{c}$ as the vector of $z_i$
  $\mathfrak{B} \leftarrow \mathfrak{B} \cup \{B\}$
**return** $\mathfrak{B}$

**Observation 3: algorithm 1 terminates.** It suffices to show that, in appendix A.2.4, the generated panel $B$ is not yet contained in $\mathfrak{B}$ since, then, the size of $\mathfrak{B}$ grows in every iteration and is always upper-bounded by the finite cardinality of $\overline{\mathfrak{B}}$. By the definition of the modified lemma A.2.5, $B$ always satisfies $\sum_{i \in B} z_i \geq \sum_{i \in P} \pi_i \, z_i$. However, since the objective value is positive, any $B' \in \mathfrak{B}$ satisfies $\sum_{i \in P} \pi_i \, z_i > \hat{z} \geq \sum_{i \in B'} z_i$, which shows that $B \notin \mathfrak{B}$.

Once algorithm 1 terminates with a set $\mathfrak{B}$, we know that $obj(\mathfrak{B}) = 0$, which means that, by solving PRIMAL($\mathfrak{B}$), we obtain a distribution over valid panels that implements the marginals $\pi_i$, which concludes the proof. □

In practice, it makes sense to exit the while loop in appendix A.2.4 already when $obj(\mathfrak{B})$ is smaller than some small positive constant, which guarantees a close approximation to the marginal probabilities while reducing running time and preventing issues due to rounding errors.

### A.2.5 Proof of theorem 2.3.1

**Theorem A.2.1.** *Suppose that $\alpha \to \infty$ and $n_{f,v} \geq n/k$ for all feature-value pairs $f, v$. Consider a sampling algorithm that, on a good pool, selects a random panel, $Panel$, via the randomized version of lemma 2.3.3, and else does not return a panel. This process satisfies, for all $i$ in the population, that*

$$\mathbb{P}[i \in Panel] \geq (1 - o(1))\, k/n.$$

*All panels produced by this process satisfy the quotas $\ell_{f,v} := (1 - \alpha^{-.49})\, k\, n_{f,v}/n - |F|$ and $u_{f,v} := (1 + \alpha^{-.49})\, k\, n_{f,v}/n + |F|$ for all feature-value pairs $f, v$.*

*Proof.* The claim about the quotas immediately follows from lemma 2.3.3 and the definition of a good pool. Concerning the selection probabilities,

$$\mathbb{P}[i \in Panel] = \sum_{\substack{\text{good pools } P \\ i \in P}} \mathbb{P}[i \in Panel \mid Pool = P]\, \mathbb{P}[Pool = P] = \sum_{\substack{\text{good pools } P \\ i \in P}} \frac{k\, a_i}{\sum_{j \in P} a_j}\, \mathbb{P}[Pool = P].$$

Since $\sum_{j \in P} a_j \leq r/(1 - \alpha^{-.49})$ for good pools, we continue

$$\geq (1 - \alpha^{-.49})\, k/(r\, q_i) \sum_{\substack{\text{good pools } P \\ i \in P}} \mathbb{P}[Pool = P] = (1 - \alpha^{-.49})\, \frac{k}{r\, q_i}\, \mathbb{P}[i \in Pool \wedge Pool \text{ is good}]$$

$$= (1 - \alpha^{-.49})\, \frac{k}{r\, q_i}\, \underbrace{\mathbb{P}[Pool \text{ is good} \mid i \in Pool]}_{\in 1 - o(1) \text{ by lemma 2.3.2}}\, \underbrace{\mathbb{P}[i \in Pool]}_{= q_i\, r/n} \in (1 - o(1))\, \frac{k}{n}. \qquad \square$$

## A.3 Supplementary Material for section 2.4

**Participation Model** Let $y_i = 1$ for agents who would join the pool if invited, and $y_i = 0$ for agents who would not. We want to predict $q_i = \mathbb{P}[y_i = 1]$ for all agents in the pool. To do so, we learn the following parametric model, which describes the relationship between an agent's feature vector $F(i)$ and value of $q_i$.

$$q_i = \beta_0 \prod_{f \in F} \beta_{f,f(i)}$$

This type of generative model describes a decision process known as *simple independent action* [119, as cited in [280]]. To express this model in a more standard form, let $x_i$ be a vector describing agent $i$'s values for all features in $F$, where each index $j$ of $x_i$ corresponds to a feature-value $f, v$ and contains a binary indicator of whether agent $i$ has value $v$ for feature $f$. Let $M$ be

the length of $x_i$, where $M = 1 + \#feature\text{-}values$. We then reshape parameters $\beta_0$, $\beta_{f,v}$ for all $f, v$ into a parameter vector $\boldsymbol{\beta}$ of length $M$, and correspondingly, $x_i$ must have value 1 at its first index for all agents $i$, corresponding to the parameter $\beta_0$. We can then write an equivalent version of our model in more standard form. Note that $q_i$ is technically a function of $x_i, \boldsymbol{\beta}$, but we omit this notation for simplicity.

$$q_i = \prod_{j \in [M]} \boldsymbol{\beta}_j^{x_{i,j}}$$

Maximum Likelihood Estimation with Contaminated Controls To estimate the parameters $\boldsymbol{\beta}$ of this model on fixed pool $P$ and fixed background sample $B$, we apply the estimation methods in Section 3 of Lancaster and Imbens [185]. We use the objective function in Equation 3.3, which is designed to perform maximum-likelihood estimation (MLE) in the setting of contaminated controls. Let $z_i$ be an indicator such that $z_i = 1$ for $i \in P$ and $z_i = 0$ for $i \in B$. Let $w_i$ be the weight of agent $i \in B$ (for details on these weights, see Appendix A.4). Recall that $\overline{q}$ is the average participation probability in the underlying population. Then, the likelihood function $L(\boldsymbol{\beta})$ that we would maximize to directly learn our model is

$$L(\boldsymbol{\beta}) = \sum_{i \in B \cup P} \left( z_i \sum_{j \in [M]} \left( x_{i,j} \log \boldsymbol{\beta}_j \right) - w_i \log \left( \overline{q}\, |B|/|P| + \prod_{j \in [M]} \boldsymbol{\beta}_j^{x_{i,j}} \right) \right)$$

Unfortunately, $L(\boldsymbol{\beta})$ is not obviously concave in $\boldsymbol{\beta}$. To get around this, we re-parameterize our model such that we can instead learn the *logarithms* of our parameters. Defining a new parameter vector $\theta$ such that $\theta_j = \log(\boldsymbol{\beta}_j)$ for all $j \in [M]$, we can rewrite our model equivalently as the exponential model.

$$q_i = \prod_{j \in [M]} \boldsymbol{\beta}_j^{x_{i,j}} = \exp\left( \log \left( \prod_{j \in [M]} \boldsymbol{\beta}_j^{x_{i,j}} \right) \right) = \exp\left( \sum_{j \in [M]} x_{i,j} \log(\boldsymbol{\beta_j}) \right) = e^{\theta x_i}$$

By Equation 3.3 in Lancaster and Imbens [185], the likelihood function $L'(\theta)$ we maximize is now the following. By Theorem A.3.1, this objective function is concave, so it can therefore be maximized efficiently (under the constraint that $\theta \leq 0$).

$$L'(\theta) = \sum_i \left( z_i \theta x_i - w_i \log \left( \overline{q}\, |B|/|P| + e^{\theta x_i} \right) \right) \tag{A.3}$$

**Theorem A.3.1.** *The log-likelihood function for the simple independent action model under contaminated controls is concave in the model parameters.*

*Proof.* The first term of the sum is linear, so both concave and convex. The second term is concave by Lemma A.3.2, □

**Lemma A.3.2.** *Let function $f(\theta) = -\log(c + e^{\theta X})$, where $c > 0$ is a constant. $f$ is concave.*

*Proof.* The $i, j$th term of the Hessian matrix $H$ of $f$ can be written as

$$H_{i,j} = -X_i X_j \frac{ce^{\theta X}}{\left(c + e^{\theta X}\right)^2}$$

Now, let $\psi = \frac{\sqrt{ce^{\theta X}}}{c+e^{\theta X}}$. Noting that $X$ is considered a column vector, we can then rewrite the Hessian in terms of $\psi$ as $H = -(\psi X)(\psi X)^T$. In words, the negative Hessian can be written as the outer product of the vector $\psi X$ with itself. Therefore, the negative Hessian is positive semi-definite, and the Hessian is negative semi-definite, implying that $f$ is concave. □

Discussion of Methods The reader may note that we treat $\overline{q}$ as a known constant in our estimation, but the objective function we use from Lancaster and Imbens is designed for the setting in which $\overline{q}$ is a variable. There is precedent in the literature for doing so [277]. As Lancaster and Imbens discuss, using $\overline{q}$ as a constant rather than a variable when maximizing Equation 3.3 introduces issues of over-parameterization, because it is not enforced that the average $q_i$ over the population be $\overline{q}$. While we cannot estimate $q_i$ values for the entire population for lack of data, it would be a worrying sign if the average $q_i$ over the *background sample*, a uniform sample from the population, was far from our assumed $\overline{q}$. However, we find that the average of our estimated $q_i$ values over the background sample is 2.9%, which matches $\overline{q} = 2.9\%$.

## A.4 Supplementary Material for section 2.5

For estimation, we use two datasets. For our positively-labeled data, we use the set of pool members from the UK Climate Assembly (for details, see Appendix A.4.1). For our background sample, we use the European Social Survey (ESS), which serves as an unlabeled uniform sample of the population.

### A.4.1 Climate Assembly UK Details & Pool Dataset

Our pool dataset contains the agents from the pool of the *Climate Assembly UK*, a national-level sortition panel on climate change held in the UK in 2020. We use "panel" to refer to the group of people who deliberate, and "assembly" to refer to the actual deliberation step. The panel for this assembly as selected by the *Sortition Foundation*, a UK-based nonprofit that selects sortition panels. A document by the Sortition Foundation gives the following description of this assembly:[1]

> *This Citizens' Assembly will meet across four weekends in early 2020 to consider how the UK can meet the Governments legally binding target to reduce greenhouse gas emissions to net zero by 2050. The outcomes will be presented to six select committees of the UK parliament, who will form detailed plans on how to implement the assembly's recommendations. These plans will be debated in the House of Commons.*

[1] https://docs.google.com/spreadsheets/d/1kwgOpxMX4pwR3Myu4pXku4gjcnOS53bPOKwOGjZNxyI/edit#gid=0

In the formation of the panel for this assembly, 30 000 letters were sent out inviting people to participate. Of these letter recipients, 1 727 people entered the pool, and 110 people were selected for the panel. The features and corresponding sets of values used for this panel are described in Table A.1.

| Feature ($f \in F$) | Values ($V_f$) |
|---|---|
| Gender | Male, Female, ~~Other~~ |
| Age | 16-29, 30-44, 45-59, 60+ |
| Region | North East, North West, Yorkshire and the Humber, East Midlands, West Midlands, East of England, London, South East, South West, Wales, Scotland, Northern Ireland |
| Education Level | No Qualifications/Level 1, Level 2/Level 3/Apprenticeship/Other, Level 4 and above |
| ~~Climate Concern Level~~ | ~~Very concerned, Fairly concerned, Not very concerned, Not at all concerned, Other~~ |
| Ethnicity | White, Black or ethnic minority (BAME) |
| Urban / Rural | Urban, Rural |

Table A.1: Climate Assembly UK features and values.

Those with value *Other* for gender were dropped from the pool data because an equivalent value could not be constructed in the ESS data. This resulted in us dropping 12 people out of the original 1727, for a pool dataset of final size 1715. Note that dropping these people did not affect our estimate of $\overline{q}$ — before and after dropping these agents, it was 2.9%. The Climate Concern Level feature was dropped altogether from the set of features used for analysis because there were too few people in the pool with value *Not at all concerned* to give these agents proportional representation on the panel.

Due to privacy agreements between the Sortition Foundation and the pool members, we are unable to share this dataset.

### A.4.2 Background Data

We define the size of the ESS dataset to be the sum of the weights of the agents within it.[1] For details on weights, see the *Re-weighting* paragraph of this section. In order to use this data as our background sample, we construct feature vectors for each person in the ESS data that correspond to those used in Climate Assembly UK, as defined in Table A.1.

In this section, we describe how we constructed the variables corresponding to the features and their values as specified by the Sortition Foundation. We dropped 44 people out of the original

[1]This sum should ideally be equal to the number of people in the ESS data, but because we drop a few people, the sum of weights no longer exactly equals the number of people.

1959 people in the ESS dataset, and we briefly discuss this decision and its implications. Finally, we describe how we re-weighted the ESS data to correct for sampling and non-response bias to approximate the scenario in which the surveyed individuals were uniformly sampled from the population. This step is important because, in our $q_i$ estimation procedure, we assume that our background sample is uniformly sampled.

Variable construction Fortunately, the ESS data contained variables and categories that either exactly or very closely corresponded to the features and values specified by the Sortition Foundation. Essentially the only modification to the ESS data we made to construct valid feature vectors was the aggregation over categories in the *Education Level* and *Urban/Rural* ESS variables, which were broken down into more fine-grained categories than those specified in Table A.1. In general, for features with values containing the value "other", missing data was assigned the value "other". Below is a table showing which variables and values from the ESS data were used to construct each feature from the Climate Assembly UK. Exact details on how these variables were used is documented in the code (see Appendix A.4.3 for reference to readme).

| Feature (Climate Assembly UK) | Variable (ESS raw data) |
|---|---|
| Gender | gndr |
| Age | agea |
| Region | region |
| Education Level | edulvlb |
| Climate Concern Level | wrclmch |
| Ethnicity | blgetmg |
| Urban/Rural | domicil |

Dropping people As described in Table A.1, the Climate Assembly's youngest valid age category was 16-29. We therefore dropped all four people in the ESS data who were under 16 years old. Dropping people who fall outside our demographic ranges of interest is not a problem for weights, because the weights of all people of interest (who we want to be fair to) will remain the same relative to each other, and we care only about the composition of this relevant population. There were an additional 40 people who may have been within our demographic range of interest, but who were missing age, race, or urban/rural data. Among these 40 people, 33, 6, and 4 people did not have data for variables corresponding to the features *age*, *ethnicity*, and *urban / rural*, respectively. While dropping these people could affect the weighting scheme, the distribution of weights of those dropped is strongly right-skewed, meaning that those who we dropped belong to groups that tended to be oversampled in the ESS data. These people are therefore likely more numerous in the ESS data overall, and dropping some of them will have a smaller proportional effect.

Finally, the ESS did not permit people to answer "other" for gender, a category permitted on the Sortition Panel. Without any way to construct the *gender = other* feature-value in the ESS data, we dropped the members of the Climate Assembly pool with this feature-value.

Re-weighting The ESS recommends re-weighting their data to correct for bias, and they provide multiple sets of possible weighting schemes for doing so[1]. Of the provided options, we elected to apply the Post-Stratification Weights, because these weights account for not only sampling bias, but also non-response bias, by incorporating auxiliary information from other demographic surveys. By this weighting scheme, each person in the ESS data is given a weight $w_i$, representing how much that person should count in the analysis of the ESS data, where the weights are normalized to 1. This weight is encoded in the ESS data as 'pspwght'.

Estimation of $\overline{q}$ We bolster the identification of our model with an estimate of $\overline{q}$, the rate of true positives in the population. In our setting, this is the number of people who would ultimately enter the pool if invited. We estimate $\overline{q}$ in Climate Assembly UK data roughly as the fraction of people who joined the pool (1 715) out of those who were invited (30 000). These numbers seem to imply that the $\overline{q} \approx 1\,715/30\,000 = 5.7\%$. However, there is a complication: each letter is sent to a *household*, rather than an individual, and any eligible member of an invited household may join the pool. Using the ESS data, we compute (see below) the average number of eligible panel participants per household to be 2.00, implying that in reality, 60 000 eligible people were invited to participate in the pool. As a result, we estimate $\overline{q}$ to be $\overline{q} = 1\,715/60\,000 \approx 2.9\%$.

Let *ESS* be the set of agents in the cleaned ESS data. Computing the average number of eligible panel participants per household from the ESS data is not entirely trivial, because sampling *people* uniformly (or in the case of the ESS, approximating uniform sampling by re-weighting) is biased toward larger households. To account for this, for each person $i \in ESS$, we scale their weight $w_i$ by the inverse of the number of eligible people in their household, $householdsize_i$. Then,

$$\text{average number of eligible people per household} = \frac{\sum_{i \in ESS} \left( \frac{w_i}{householdsize_i} \right) \cdot householdsize_i}{\sum_{i \in ESS} \left( \frac{w_i}{householdsize_i} \right)}$$

We compute $householdsize_i$ for each person $i \in ESS$ using the weighted ESS data. Age is the only feature from the UK Climate Assembly for which the ESS data may contain values rendering a person ineligible (specifically, the ESS data surveys people down to age 15, while the climate assembly accepted only those over 16). To count the number of people in each household who are eligible, we use variables 'agea', 'pspwght', and 'yrbrn2-12', which describe the ages of person $i$'s household members (up to 12 household members).

### A.4.3 Implementation Details

Our experiments were implemented in Python, using PyTorch for the MLE estimation and Gurobi for solving the linear programs in the column generation. Our code is contained in the supplementary material and will be made available as open source when published. The file "README.md" in the code gives detailed instructions for reproducibility.

[1] https://www.europeansocialsurvey.org/docs/methodology/ESS_weighting_data_1.pdf

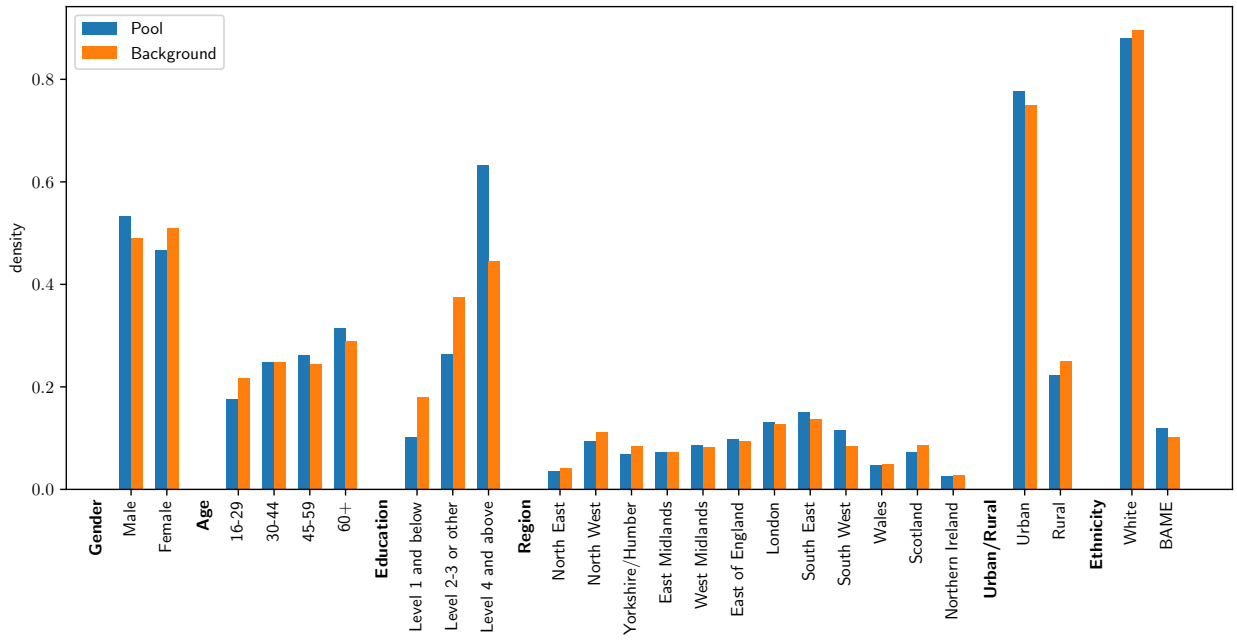
We found the log-likelihood presented in eq. (A.3) to be easy to maximize. For accuracy, we chose a small step size of $10^{-5}$ and a large number $10^5$ of optimization steps. The final objective was 4157.32345, and objective changes between iterations 20 000 and 100 000 were less than $3 \times 10^{-6}$.

Our experiments were run on a 13-inch MacBook Pro (2017) with a 3.1 GHz Dual-Core i5 processor. Optimizing the log-likelihood took 46 seconds. Running the column generation took 38 minutes to reach the desired accuracy of $10^{-6}$, which is much smaller than the smallest $\pi_{i,P}$ at around 2%. For the version including climate concern, MLE estimation took 37 seconds reaching a log-likelihood of 4601.01427, and column generation took 26 minutes.

Sampling 100 000 pools each and simulating our algorithm for the end-to-end experiments took 30 minutes for $r = 10\,000$, 55 minutes for $r = 11\,000$, 61 minutes for $r = 12\,000$, 76 minutes for $r = 15\,000$, and 95 minutes for $r = 60\,000$. All running times should be seen as upper bounds since other processes were running simultaneously. Sampling the same number of pools for the case including the climate concern feature took around 410 minutes for $r = 600\,000$. The equivalent experiments with the greedy algorithm took around 19 hours (floor and ceiling quotas) and around 12 hours (no quotas).

### A.4.4 Results and Validation of $\boldsymbol{\beta}, \boldsymbol{q_i}$ Estimation

**Pool and Background Data Composition** First, we examine the frequency at which each feature-value occurs in the pool and the background data. As shown in the figure below, those with the most education are highly over-represented in the Climate Assembly UK pool compared to the background sample, and people with low education are under-represented. Similarly, we see men are slightly over-represented in the pool, and increasing age also seems to increase likelihood of entering the pool.

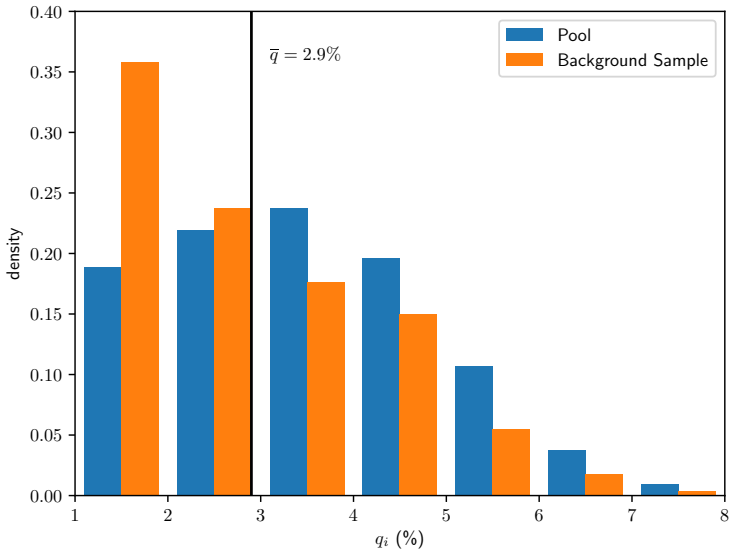


Estimates of $\boldsymbol{\beta}$ We find that $\beta_0 = 8.8\%$, meaning that all agents participate with a baseline probability of 8.8%. In the figure below are estimates of $\beta_{f,v}$ for all feature-values $f, v$. Recall that $1 - \beta_{f,v}$ can be interpreted as the probability of not participating due to having value $v$ for feature $f$; in other words if $\beta_{f,v}$ is 1, then feature-value $f, v$ has no adverse effect on whether a person participates.

Notably, these $\beta$ estimates are consistent with the composition of the pool compared to the background data. For example, people of increasing age were increasingly over-represented in the pool compared to the background data, and we see here that $\beta$ associated with age increase with increasing age. Similarly, we see that having low education greatly diminishes a person's likelihood of participation, corresponding to the observation that the pool contained a disproportionately low number of people with the two lower levels of education. In fact, one can confirm that across all feature-values, $\beta$ values correspond with the composition of the pool data compared to the background data, indicating that the $\beta$ values learned with our model are a good fit to the data used to learn them.


Estimates of $\boldsymbol{q_i}$ We compute our $q_i$ estimates based on $\beta$ estimates according to the model in Appendix A.3. We get the following distributions of $q_i$ values in the pool and background datasets.

The data shown in this plot is limited to density of $q_i$ values between 1% and 8%, because bins outside this range contain fewer than 7 people, and are withheld to avoid potential privacy issues. Less than 0.3% of agents in either dataset are excluded for this reason.




Not very surprisingly, we find that the pool overrepresents agents with higher participation probability with respect to their share in the background sample.

Test for Calibration of $q_i$ Estimates To validate whether our model fits the data well, we form a *hypothetical pool* by imagining that the weighted background sample was selected as the set of recipients and that the members of this set participate with our estimated probability $q_i$. For some attributes that agents might have or not have, the expected number of agents in the hypothetical pool with this attribute is

$$\sum_{i \in B : i \text{ has attribute}} q_i.$$ [1]

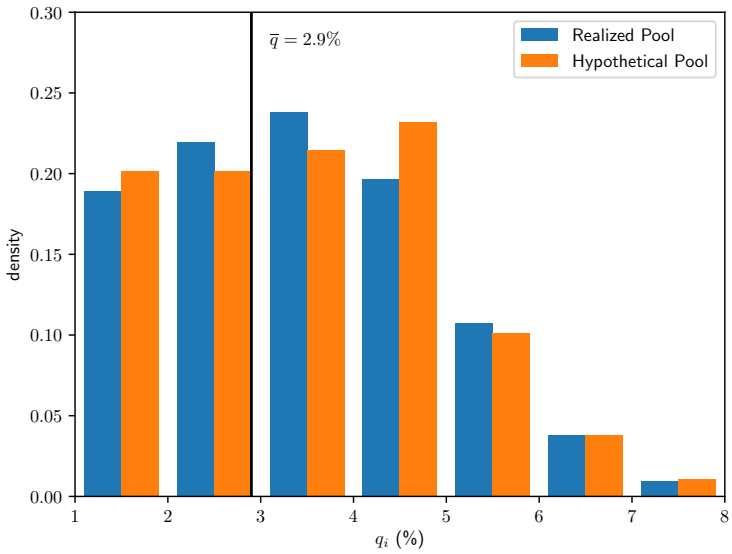
Since the set of invitation recipients to the Climate Assembly and the background sample are both assumed to be representative samples of the population, we would expect the above sum to be (close to) proportional to the fraction of pool members with this attribute — at least if the model fits the data well.

For instance, this idea allows us to re-examine the previous plot of $q_i$ values by letting the orange bars not denote the (scaled) *number* of members in the background sample with $q_i$ in the right range, but instead the (scaled) *sum of* $q_i$ *values* of members in the background sample with $q_i$ in this range.

The fact that these distributions align fairly well can be seen as our $q_i$ passing a sort of calibration test — of those agents with a certain $q_i$ value, roughly a $q_i$ proportion would participate when invited. Relative to our background sample, the Climate Assembly pool does not seem to untypically skew towards agents with low or high values of $q_i$.

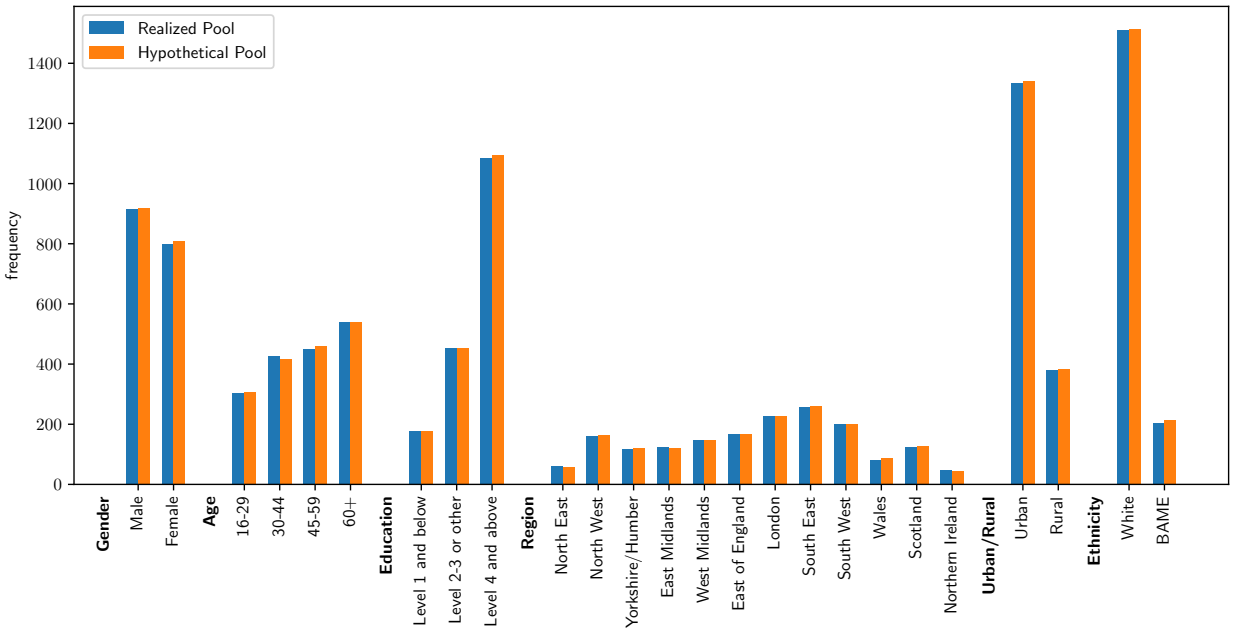
[1] Of course, all operations on the background sample respect the weights, which we ignore here for the sake of clarity.

Once again, for privacy reasons we display frequencies of $q_i$ values only between 1% and 8%. Once again, less than 0.3% of agents in either dataset are excluded for this reason.

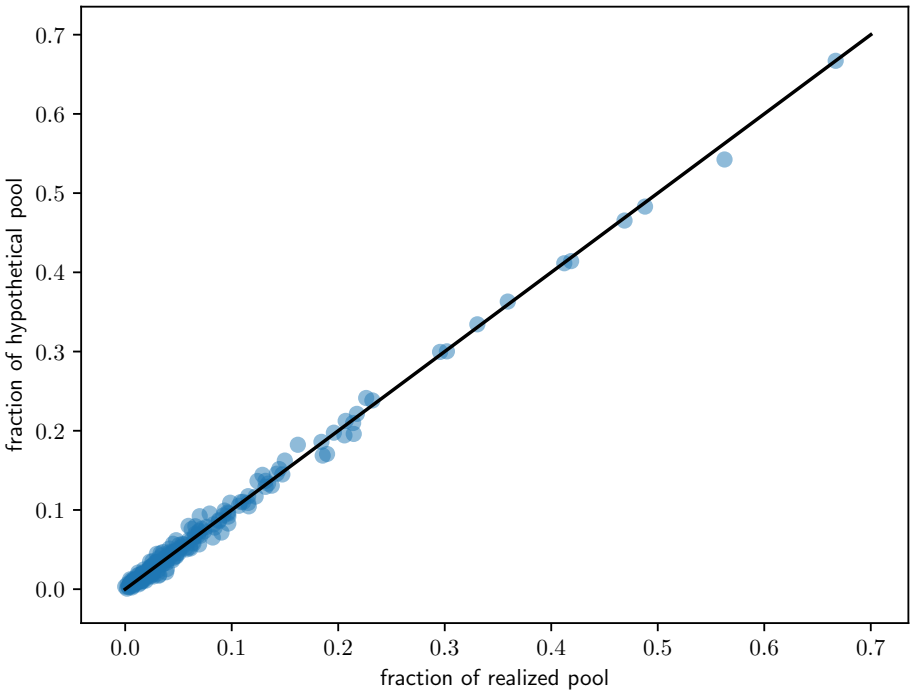



Comparison of Realized Pool Composition and Hypothetical Pool Composition We now plot the same comparison between the Climate Assembly pool and the hypothetical pool but for the prevalence of each feature-value pair.

The figure below shows that if our $\beta$ estimates and the $q_i$ estimates they yield are true for members of the population, then if we sampled the underlying population as was done to form the Climate Assembly UK pool, we would get in expectation a pool that looks almost identical to realized pool. This illustrates in another way that our $\beta$ estimates are a good fit to the data we provided.




Testing Model Capture of 2-Correlations Our model assumes that each feature-value affects people's probability of participating independently of all other feature-values. This analysis tests whether this causes our model to severely misjudge the participation probability for some group defined by the intersection of *two* feature-value pairs, again comparing the prevalence of these groups in the Climate Assembly UK pool vs. the hypothetical pool that would be drawn from a population with the same composition as the background sample. On the plot below, each point represents an intersection of two feature-values. Each point's $x$ and $y$ coordinates are the fraction of people with that intersection in the Climate Assembly UK pool and the fraction of the hypothetical pool, respectively. We would hope for this relationship to be exactly linear, illustrating that each pair of feature-values occurs at the same rate in the real vs. hypothetical pool.




### A.4.5 Details on End-To-End Experiment

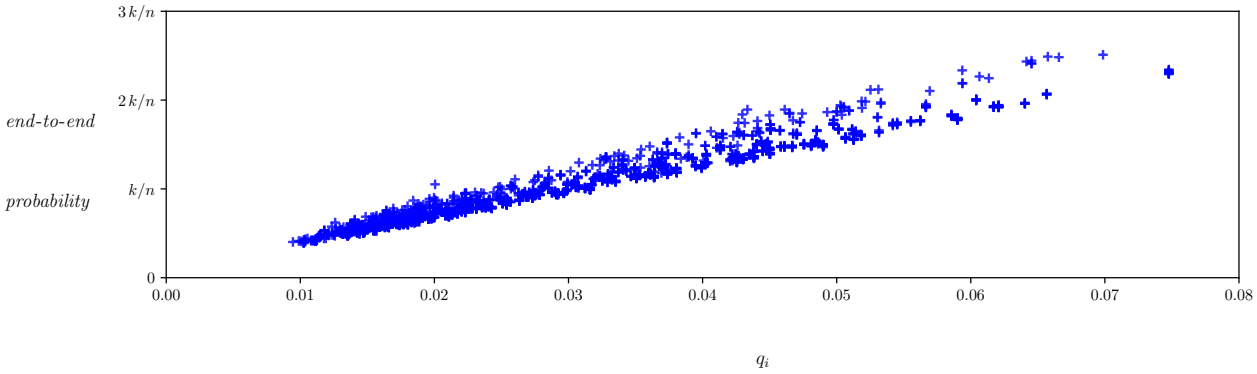
As described in the body of the paper, we generate a synthetic population by scaling up the ESS participants to a population of 60 million individuals. The number of copies of a participant is proportional to their weight in the ESS, and is rounded to an integer using the Hamilton apportionment method. 100 000 times per experiment, we select a set of letter recipients of size $r$ uniformly from the population, and flip a biased coin with probability $q_i$ for each letter recipient to determine whether she joins the pool. For each pool, we then obtain the selection probabilities of the pool members conditioned on this being the pool (or an unbiased estimate of these probabilities):

- For our algorithm, we check whether the pool $P$ is good. If the pool is not good, we (conservatively) assume that no panel is returned, and that pool members have zero probability of being selected. Else, we return the selection probabilities $\pi_{i,P}$.
- We use the implementation of the greedy algorithm developed by the Sortition Foundation and available at https://github.com/sortitionfoundation/stratification-app/tree/4a957359b708a327aad0103ab2a59d061aeaeeb4. Since we do not have a closed form for individual selection probabilities, we run the greedy algorithm 10 times and report the average time that each pool member was selected. While these estimates of selection probabilities are noisy, they are unbiased estimates of the end-to-end probability and independent between pools. Thus, the noise largely averages out over the 100 000 random pools. In no case did the greedy algorithm fail to satisfy the quotas.

Each point in the diagrams corresponds to one agent in the ESS sample and indicates this agents'

$q_i$ as well as the average selection probability of its copies, averaged over the different pools and the different copies. Since both our algorithm and the greedy algorithm treat agents with equal feature vector symmetrically, averaging over the copies of an ESS participant is a valid way to estimate the end-to-end probability of any single copy, which greatly reduces sample variance.

In the body of the paper, we mention the behavior of the greedy algorithm without any quotas. In this case, the panel members seem to be sampled with near-equal probability from the pool, which leads to end-to-end probabilities that are roughly proportional to $q_i$:




### A.4.6 Additional Results for Section 2.5

End-to-End Fairness Results for Varied $r$ Values This plot shows the end-to-end probabilities for all agents in the synthetically-generated population over varied values of $r$. To recall, we copied the agents in the background sample (in proportion to their weight) to obtain a synthetic population of size 60 million (the order of magnitude of eligible participants for the Climate Assembly).

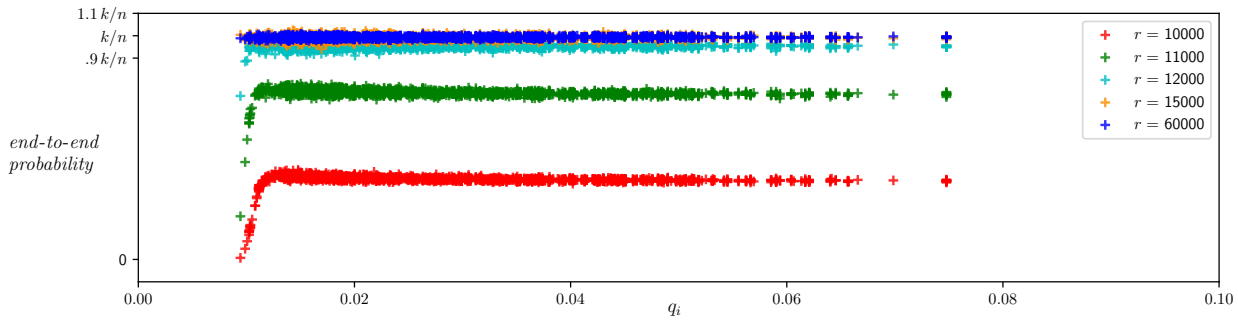
We display these end-to-end probabilities for $r$ values 11 000, 12 000, 13 000, and 60 000, where 60 000 is the $r$ value used to form the real-life Climate Assembly UK pool. Every point in the scatter plot corresponds to an original member of the background sample, and the point's y-value is the mean selection probabilities averaged over 100 000 sampled pools and over all copies of this background agent.[1]

An important question is what we do when a bad pool occurs. In the corresponding figure in the body of the text (examining only $r = 60\,000$), we did not credit any selection probability to any agent when bad pools occurred. When we take this approach for multiple $r$ values, the result shows a sharp discontinuity between $r = 11\,000$ (when everyone's end-to-end probability is essentially zero) and $r = 12\,000$ (when it is around 95%). As it turns out, the property that makes nearly all pools bad when $r = 11\,000$ is eq. (2.3). Note that this property is the least consequential of the three defining properties of a good pool: if we proceed with Part II of the algorithm on a pool that satisfies only eqs. (2.1) and (2.2), we still satisfy the quotas but just can't bound the end-to-end probabilities. Since the end-to-end probabilities are what we are measuring here anyway,

[1] Averaging over the copies of an agent makes use of the fact that the selection process treats copies of the same agent symmetrically, which makes the empirical means converge faster.

we will in the following graph count bad pools as good pools if they only violate eq. (2.3).

As shown in the figure below, we see a smooth transition towards the end-to-end guarantee, where higher values of $r$ give better guarantees. The agents with the lowest selection probabilities are suffering most from low values of $r$, with their end-to-end probability trailing that of the majority of other agents. From $r = 15\,000$ upwards, however, all agents in the population receive an end-to-end probability that is very close to $k/n$. This threshold roughly coincides with the point at which $\alpha$ becomes larger than one.

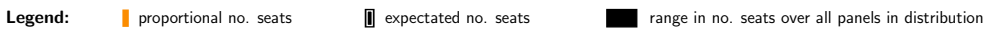


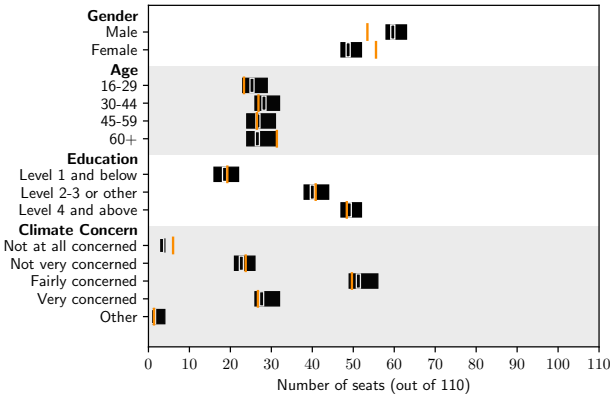

### A.4.7 Validation and Results including Climate Concern Feature

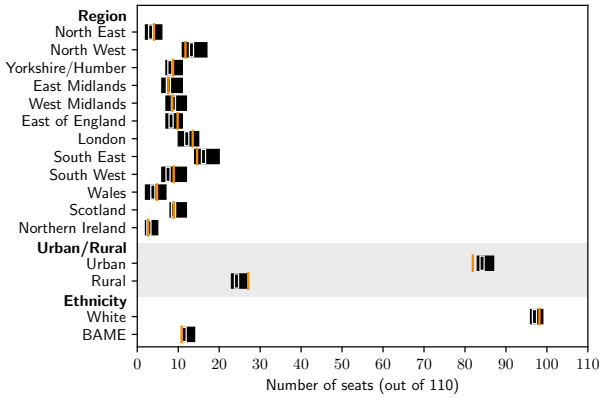
This section includes all the analysis in this paper and appendices, re-done with the climate concern level feature included. Figures in this section are provided in the same order as they were presented in the body of the paper, Appendix A.4.4, and Appendix A.4.6.

***(Figures from Paper Body)***

We omit the figure showing end-to-end probabilities at $r = 60,000$, because when the *Climate Concern Level* feature is included, good pools are so rare at this value of $r$ that all end-to-end probabilities are 0. Similarly, for the greedy algorithm, the floor and ceiling quotas are often not satisfiable. In 754 out of 1 000 random pools, this is because fewer pool members are "not at all concerned" about climate change than the lower quota for this feature, which is 6. In 86 out of the remaining pools, the greedy algorithm fails to identify a valid panel within the first 100 restarts. Only in the remaining 160 pools did the greedy algorithm find a valid panel in fewer than 100 iterations.

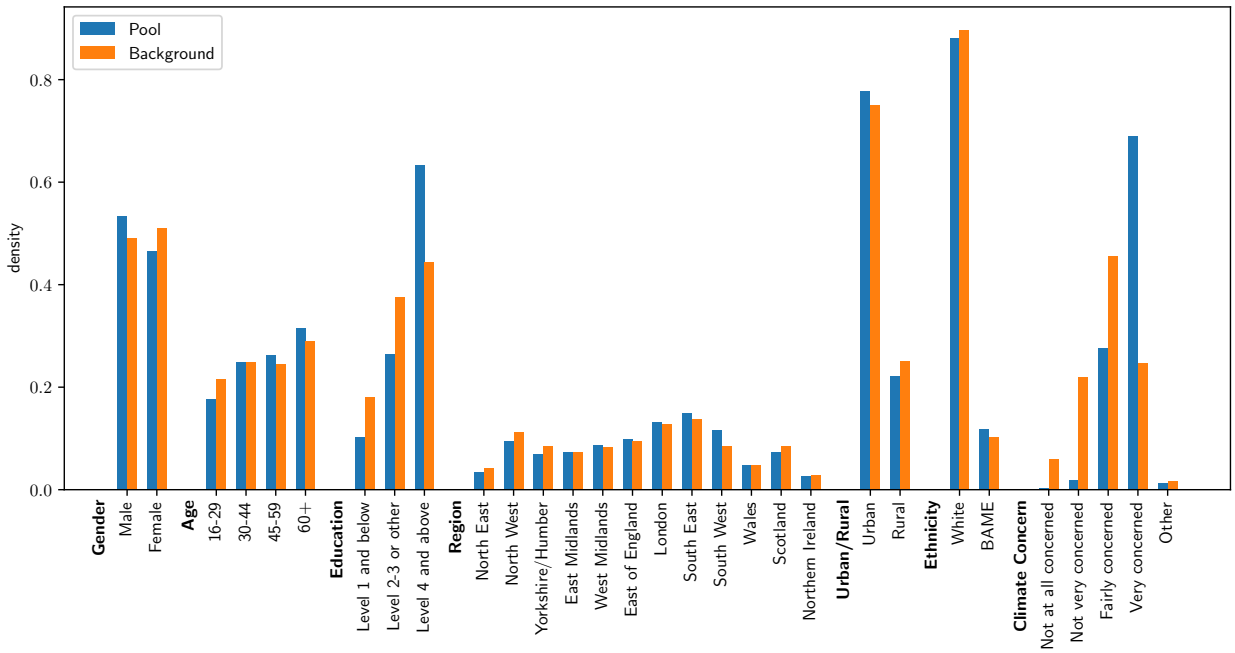



Gender
Male
Female
Age
16-29
30-44
45-59
60+
Education
Level 1 and below
Level 2-3 or other
Level 4 and above
Climate Concern
Not at all concerned
Not very concerned
Fairly concerned
Very concerned
Other
0 10 20 30 40 50 60 70 80 90 100 110
Number of seats (out of 110)

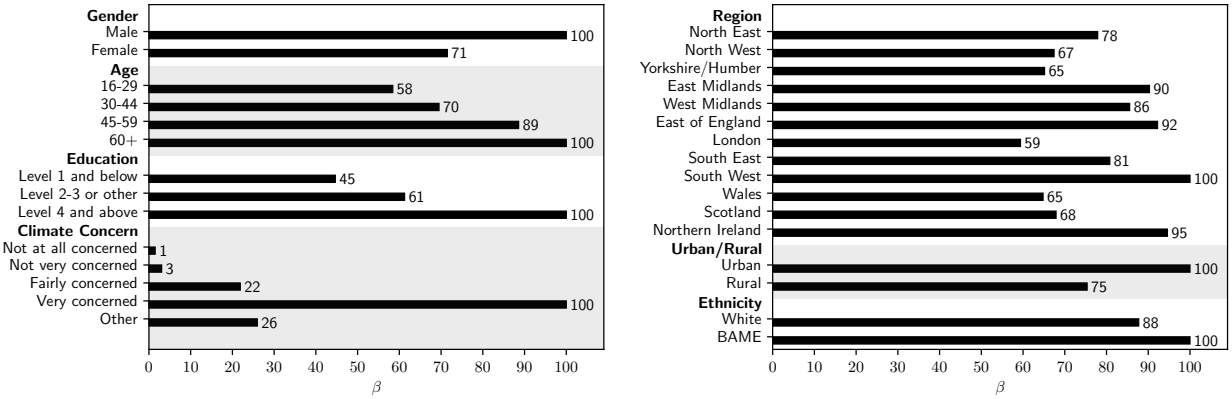

Region
North East
North West
Yorkshire/Humber
East Midlands
West Midlands
East of England
London
South East
South West
Wales
Scotland
Northern Ireland
Urban/Rural
Urban
Rural
Ethnicity
White
BAME
0 10 20 30 40 50 60 70 80 90 100 110
Number of seats (out of 110)

*(Figures from Appendix A.4.4)*

### Pool and Background Data Composition

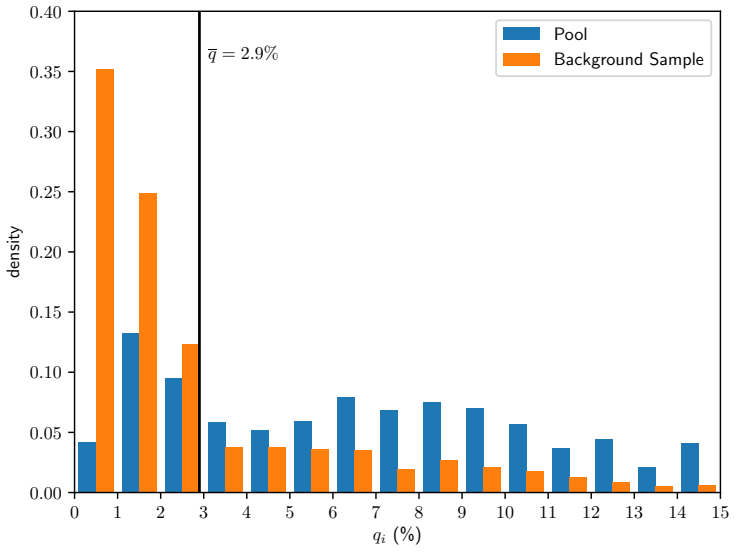



### Estimates of $\beta$

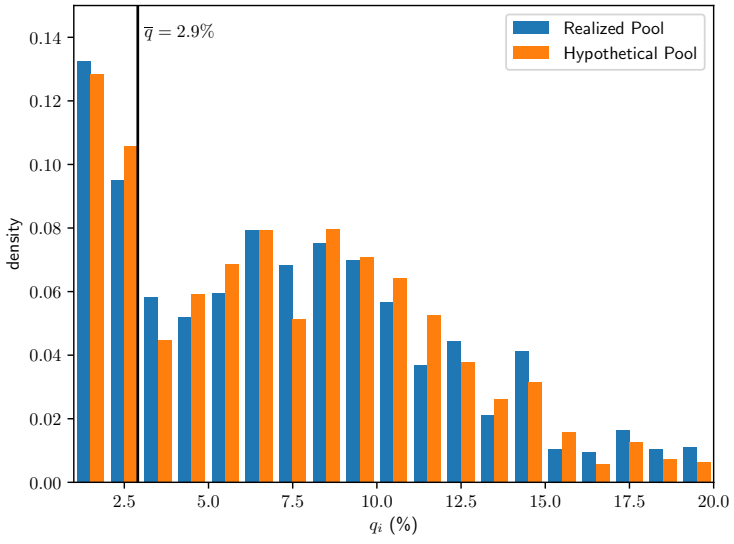



### Estimates of $q_i$

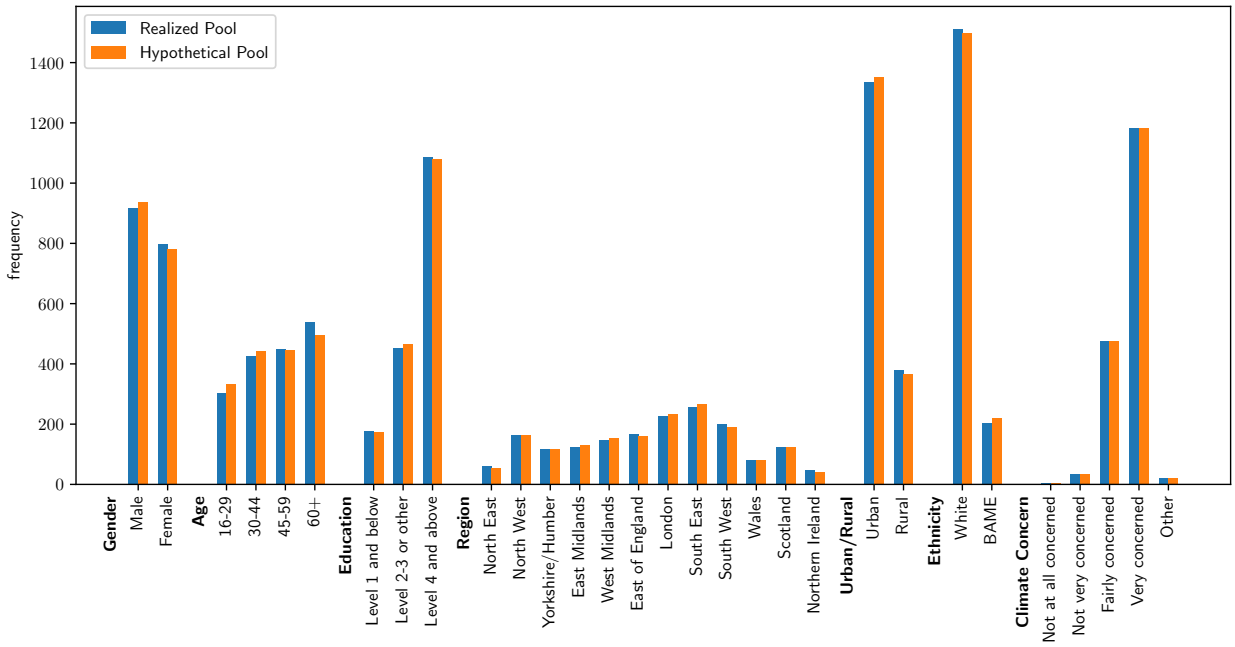
$\beta_0 = 24.3\%$. Frequencies of $q_i$ values above 15% are not shown due to privacy concerns. 6.8%, 1% of agents in pool, background datasets respectively are not presented for this reason.



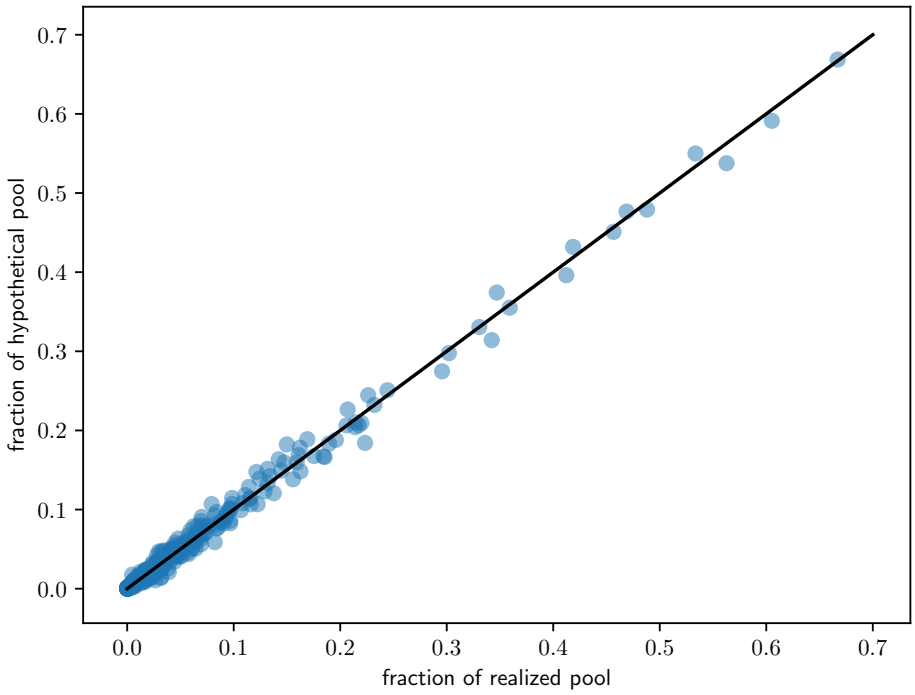

Test for calibration of $q_i$ estimates Frequencies of $q_i$ values above 20% are not shown due to privacy concerns. Less than 0.4% of agents in either dataset are not presented for this reason.




Comparison of Realized Pool Composition and Hypothetical Pool Composition

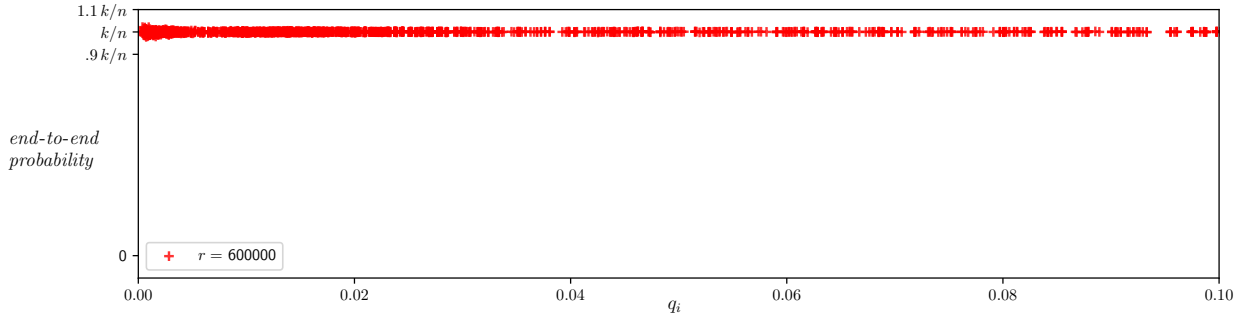



Testing model capture of 2-correlations




*(Figures from Appendix A.4.6)*

End-to-End Fairness Results for Varied $r$ Values This figure demonstrates that, for large enough $r$, we can get $k/n$ end-to-end probability for all agents in the synthetic population when we include the Climate Concern Level feature. We only include analysis for only one $r$ value because the $r$ values must be extremely large to give any end-to-end guarantees when the Climate Concern Feature is included, and running the analysis with such large $r$ costs substantial computational time.



# B

# Chapter 3 Appendix

# Contents

### B.1 Illustration of Definitions with Examples

Here, we introduce the definitions and concepts used in this paper through an example *instance*, which is composed of a pool, information about quotas, and a panel size $k$.

Example instance. Suppose we want to select a panel of size $k = 3$. Let the *features* on which we want to impose *quotas* be female, male, young, and old; and let the lower and upper quotas for each feature be as specified below:

| | female | male | young | old |
|---|---|---|---|---|
| lower quota | 1 | 1 | 2 | 1 |
| upper quota | 2 | 2 | 2 | 1 |

Finally, suppose that the *pool* of the instance contains $n = 5$ pool members, which are given with their features:

| name | features |
|---|---|
| Alice | young, female |
| Bob | old, male |
| Ciara | young, female |
| Dan | young, male |
| Ella | old, female |

Panels for the example instance. A *panel* for this instance is any set of 3 pool members in which 1 or 2 are female, 1 or 2 are male, exactly 1 is old, and exactly 2 are young. Therefore, the complete set of panels in this instance is:

$$\widehat{\mathcal{P}} = \{\{\text{Alice, Bob, Ciara}\}, \{\text{Alice, Bob, Dan}\}, \{\text{Ciara, Bob, Dan}\}, \{\text{Alice, Dan, Ella}\}, \{\text{Ciara, Dan, Ella}\}\}$$

Selection algorithms on this instance. In general, a *selection algorithm* takes in an arbitrary instance and must (randomly) return a panel for that instance. Thus, when a selection algorithm receives our example instance as its input, it must produce one of the panels in $\widehat{\mathcal{P}}$. Now, we compare the behavior of two selection algorithms, Legacy and Leximin , on this instance. (These algorithms are formally defined in appendices B.10 and B.11, but no knowledge of the algorithms is necessary to follow this example.)

Legacy[1] and Leximin each have a different *output distribution* on our instance, both of which are displayed on the left-hand side of the two tables below. While both algorithms return the same

[1]For one specific way of breaking ties between features (male > female > old > young), which is left unspecified by the algorithm (see appendix B.11).

set of panels, they differ in how likely each panel is to be selected; for example, Legacy selects the panel {Alice, Bob, Ciara} with probability 1/6 whereas Leximin selects that panel with probability 1/3.

Each algorithm's output distribution determines the *selection probability* of each pool member. For example, the probability that Legacy selects a panel containing Ella can be calculated by summing up the output probabilities of both panels that include her: Since Legacy selects {Alice, Dan, Ella} and {Ciara, Dan, Ella} each with probability 1/6, Ella's selection probability is 1/3. We refer to agents' collective selection probabilities as a *probability allocation*. The probability allocations of the two algorithms are given on the right-hand side of the two tables below.

*Fairness measures* evaluate the fairness of different probability allocations, which allows us to evaluate whether Legacy or Leximin is fairer on our instance. One important fairness measure ("egalitarian social welfare"; see appendix B.9) measures the fairness of a probability allocation by its minimum selection probability. Using this fairness measure, the fairness of Legacy's probability allocation is 1/3 whereas the fairness of Leximin 's probability allocation is 1/2. Since the latter value is higher, the fairness measure judges Leximin to be fairer on the example instance than Legacy.

In this paper, we develop maximally fair selection algorithms. As it turns out, Leximin is one such algorithm for the fairness measure above, in the sense that, for all instances, and for all other selection algorithms, the minimum selection probability of Leximin will be at least as large as the minimum selection probability of the other algorithm.

Legacy

| *Output Distribution* | *Probability Allocation* |
|---|---|
| $\mathbb{P}[\{\text{Alice, Bob, Ciara}\} \text{ selected}] = \frac{1}{6}$ | Alice: $\frac{1}{6} + \frac{1}{4} + \frac{1}{6} = \frac{7}{12}$ |
| $\mathbb{P}[\{\text{Alice, Bob, Dan}\} \text{ selected}] = \frac{1}{4}$ | Bob: $\frac{1}{6} + \frac{1}{4} + \frac{1}{4} = \frac{2}{3}$ |
| $\mathbb{P}[\{\text{Ciara, Bob, Dan}\} \text{ selected}] = \frac{1}{4}$ | Ciara: $\frac{1}{6} + \frac{1}{4} + \frac{1}{6} = \frac{7}{12}$ |
| $\mathbb{P}[\{\text{Alice, Dan, Ella}\} \text{ selected}] = \frac{1}{6}$ | Dan: $\frac{1}{4} + \frac{1}{4} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$ |
| $\mathbb{P}[\{\text{Ciara, Dan, Ella}\} \text{ selected}] = \frac{1}{6}$ | Ella: $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$ |

Leximin

| *Output Distribution* | *Probability Allocation* |
|---|---|
| $\mathbb{P}[\{\text{Alice, Bob, Ciara}\} \text{ selected}] = \frac{1}{3}$ | Alice: $\frac{1}{3} + \frac{1}{12} + \frac{1}{4} = \frac{2}{3}$ |
| $\mathbb{P}[\{\text{Alice, Bob, Dan}\} \text{ selected}] = \frac{1}{12}$ | Bob: $\frac{1}{3} + \frac{1}{12} + \frac{1}{12} = \frac{1}{2}$ |
| $\mathbb{P}[\{\text{Ciara, Bob, Dan}\} \text{ selected}] = \frac{1}{12}$ | Ciara: $\frac{1}{3} + \frac{1}{12} + \frac{1}{4} = \frac{2}{3}$ |
| $\mathbb{P}[\{\text{Alice, Dan, Ella}\} \text{ selected}] = \frac{1}{4}$ | Dan: $\frac{1}{12} + \frac{1}{12} + \frac{1}{4} + \frac{1}{4} = \frac{2}{3}$ |
| $\mathbb{P}[\{\text{Ciara, Dan, Ella}\} \text{ selected}] = \frac{1}{4}$ | Ella: $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ |

### B.2 Model

An *instance* consists of a set of agents $N = \{1, \ldots, n\}$, a desired panel size $k$, and a finite set of *features*. Examples of such features could be "female" or "older than 65". Let $N_f$ be the set of agents with feature $f$. Each feature $f$ is furthermore associated with a lower quota $\ell_f$ and an upper quota $u_f$, which specify lower and upper limits on the number of panel seats to be filled by agents in $N_f$. In a given instance, a *panel* $P$ is any subset of $N$ such that the following integer linear program (ILP) is satisfied by the set of 0–1 indicators $x_i$ that specify whether agent $i$ is in panel $P$:

$$\sum_{i \in N} x_i = k \qquad\qquad (P \text{ contains } k \text{ agents})$$

$$\ell_f \leq \sum_{i \in N_f} x_i \leq u_f \qquad \forall \text{ features } f \qquad (P \text{ satisfies all lower and upper quotas})$$

$$x_i \in \{0, 1\} \qquad \forall i \in N \qquad (\text{the } x_i \text{ are binary indicators}).$$

In the context of our column-generation framework, we call a set of panels within the same instance a *portfolio*.

To avoid issues of well-definedness, we formally restrict our definition of an instance to include only those in which there exists at least one panel. (In practice, this restriction is unproblematic, since the existence of a panel can be confirmed by checking the satisfiability of the ILP above with an ILP solver before applying a selection algorithm.)

A *selection algorithm* receives an instance as its input and must randomly choose a panel to return. We call the distribution describing the probability with which each panel is returned the selection algorithm's *output distribution* for this instance. If, for a given selection algorithm and input instance, we let the random variable $P$ denote the panel returned by the selection algorithm (its distribution then being the output distribution), the *selection probability* $p_i$ of an agent $i$ is defined as $\mathbb{P}[i \in P]$, and a *probability allocation* is a function mapping each agent $i \in N$ to their selection probability $p_i$.

Finally, a *fairness measure* for a specific instance is a function $F : [0, 1]^n \to (\mathbb{R} \cup \{-\infty\})$ mapping the probability allocations of that instance to a score, where larger scores denote preferable levels of fairness. To avoid artificially reducing the generality of our results, this definition of a fairness measure is specific to one instance. Where we speak of "fairness measures" in the body of the paper and in appendix B.9 (e.g., "Nash welfare" or "Gini coefficient"), we are formally referring to families of fairness measures, where each family contains one fairness measure for each possible instance.

### B.3 Stratified Sampling

One procedure for selecting random panels that is often discussed is *stratified sampling*. A stratified-sampling procedure is defined by what we will call a *stratification*: a partition of the population into disjoint subgroups (e.g., women, men, people of nonbinary gender), where each subgroup is associated with the number of panel seats they will receive (say, 19, 19, and 2 seats). Then, from each stratum, the procedure uniformly samples the specified number of panel members. Stratified sampling and our selection algorithms similarly strive to ensure descriptive representation. However, our algorithms accept a more flexible range of quotas for expressing constraints on descriptive representation, making them more widely applicable than stratified sampling. For instance, the quota constraints imposed in all ten citizens' assemblies analyzed in this paper cannot be expressed as stratifications.

To understand why the quotas imposed in practice are more general than those imposed by stratified sampling, we first note that the constraints expressed by a stratification can directly be expressed as a system of quotas. This is done by turning each stratum into a feature, and then setting both the feature's lower and upper quota to the desired number of panel seats. By contrast, not every system of quotas can be expressed as a stratification. This is for two reasons: first, whereas practitioners often permit a bit of tolerance between a feature's upper and lower quota, stratified sampling requires specifying the *exact number* of people to be chosen from each stratum. Second, and more fundamentally, quotas are often imposed on overlapping groups (e.g., the groups women and young people, where individuals can belong to both groups at once), whereas all strata must be disjoint.

To see why this restriction limits the generality of stratified sampling, consider an example in which we have overlapping categories gender and age, and want to impose quotas on women, men, people of non-binary gender, young people, and old people. In stratified sampling, one would define six disjoint strata: young women, young men, young people of nonbinary gender, old women, old men, and old people of nonbinary gender. One would then have to specify some exact number of people from each stratum; by contrast, the constraints expressed by quotas on the feature can be much more flexible since they, for example, do not directly constrain the age composition within the group of women.

As illustrated in the above example, one can implement quotas in practical settings by defining the strata to be all intersectional groups. However, this strategy does not extend practicably to the number of feature categories on which quotas are imposed in practice (in our instances, between 4 and 8). This is because imposing quotas on many orthogonal features (e.g, gender, age, region, and education level) would require setting aside a number of seats for exponentially many combinations of these features (e.g., "female, 18–25 years old, London, no diploma"), which would quickly exceed the number of panel seats.

## B.4 Desiderata for Sortition in the Political Science Literature

In this paper, we approach the problem of panel selection from a pragmatic angle. We ask: taking as given the overall panel selection process (sending out invitations uniformly at random, and then using quotas to enforce representativeness), what is the best selection algorithm for practitioners to use?

To identify desirable properties of a selection algorithm, it is natural to take inspiration from political theory, where advantages and disadvantages of sortition have been discussed in detail[86, 112, 124, 250, 258]. However, one should not expect the political theory literature to give concrete instructions for a practical selection algorithm, since the literature focuses on an idealized sortition process that ignores the complications of the real-world settings in which panels must be selected. In particular, the literature assumes that panels can be selected by sampling directly from the population, whereby each member of the population is selected with equal probability and will agree to participate if invited [71, 224, 257]. We refer to this procedure as *idealized sortition*. Usually, in practice, a large majority of people decline to participate when invited [250].

Though this literature does not immediately prescribe a practical selection algorithm, it informs our approach by identifying the values that should be pursued when designing selection algorithms. In this section, we outline several prominently advocated properties of idealized sortition, discuss how they are or are not conducive to algorithmic implementation, and describe how these properties complement or contradict one another. Ultimately, our approach of making selection probabilities as equal as possible strives for *promotion of equality*, while guaranteeing the achievement of *representativeness* as implemented by practitioners via quotas.

### B.4.1 Properties of Idealized Sortition

Following a model developed by Engelstad [112] and elaborated upon by others [71**? ?** ], sortition should simultaneously (1) *promote equality*, (2) *ensure representativeness*, (3) *maximize efficiency*, and (4) *protect against conflict and domination*.

#### Equality

According to Engelstad, "The strongest normative argument in favour of sortition is linked to the idea of social equality and individual welfare", which stems from the fact that every constituent has an equal selection probability. [112] Subsequent work in political theory has reaffirmed the importance of equal selection probabilities, even if different authors deduce this importance from slightly different ideals: Some [123, 124, 224, 258] see the equal selection probabilities of idealized sortition as an embodiment of *democratic equality*, the ideal that a democratic decision-making process should give equal consideration to all of its constituents' preferences. Other authors [71, 224] stress equal probabilities as the hallmark of (prospect-regarding [234]) *equality of opportunity*. A related argument is made by Stone [257, 258]. Rather than seeing equality as the goal in its own right, he views random allocation with equal probability as the only way to

satisfy *allocative justice* in the distribution of public offices among constituents who all have equal claims to authority.

As we discuss in the introduction, perfect equality of selection probabilities is not attainable within the constraints of practical sortition. In this paper, we handle this impossibility by proposing a more gradual version of this goal: Subject to achieving descriptive representation, one should make selection probabilities as equal as possible. The view of political office as a good, and of sortition as a means to allocative justice [258], is a natural foundation for the approach of treating panel selection as a problem of fair division (see appendix B.9).

### Representativeness

Another important benefit of ideal sortition is that, with high probability, the composition of the panel will resemble the population along all dimensions of interest [257]. Descriptive representation is a crucial assumption in Fishkin's argument that the result of a deliberative minipublic can reveal the likely outcome of the whole population deliberating [123, 124]. In addition to its contribution to the quality of deliberation, descriptive representation is particularly valuable in contexts of mistrust and marginalization [198].

As stated above, the statistical properties of idealized sortition imply that *any possible* division of the population is likely to be represented close to proportionally on the panel, provided that the panel size is sufficiently large. By contrast, no such guarantee can be provided in the realistic setting where constituents decline to participate, which forces practitioners to select specific features for which they want to enforce descriptive representation using quotas. Whereas our approach focuses on making selection probabilities close to equal, we do not sacrifice descriptive representation for this goal. Rather, organizing bodies can still set quotas to ensure a desired level of descriptive representation, and our methods only use the remaining freedom within these constraints to promote equality. In this way, our method allows an assembly organizer to trade off representation and equality by tightening or loosening the quotas.

### Efficiency

In comparison to selecting representatives by election, some authors argue that sortition is more efficient because it requires fewer resources [71, 112]. For instance, campaigning and organizing elections are not necessary. Arguably, this argument is more specific to the benchmark of elections than to sortition, and subsequent works have put little emphasis on this point [257].

When considering the design of the selection algorithm, the only major resource one might seek to use efficiently is time — namely, the time the algorithm takes to run. Given that the selection of the panel from the pool is only a minor task in organizing and convening a citizens' assembly, as organizers spend much more time recruiting the pool and organizing the deliberation. For this reason, reducing the running time of the algorithm seems a frivolous efficiency. As we show in Table 1, our algorithm Leximin runs in seconds for most instances and an hour at most. This is significantly longer than the running time of the benchmark algorithm Legacy, but much faster than the process of executing other selection algorithms using dice and spreadsheets, as practiced

by some organizations. We take this as an indicator that hours versus minutes of running time is not a significant consideration in terms of efficiency.

Existing algorithms often confront practitioners with a hard trade-off between representation and computational efficiency, since more numerous and tighter quotas may drastically increase the running time of these algorithms. While such a concern cannot be theoretically ruled out for any known algorithm (appendix B.6), our algorithms delegate the task of finding panels to a state-of-the-art ILP solver, a mature technology routinely used to solve much harder tasks [148] than all panel-selection subtasks we have encountered. Therefore, we expect our algorithm to allow for much more complex quotas without substantial increases in running time; the fundamental trade-offs between representativeness and equality, of course, persist. Our algorithms also have an advantage in the (undesirable) situation where no panel formed from the pool can satisfy the quotas. Whereas existing algorithms enter an infinite loop in this situation until the user gives up, our algorithms' first call to the ILP solver will immediately reveal that the quotas are infeasible; in these situations, our implementation solves a second ILP to suggest a minimal relaxation of the quotas that can be satisfied.

#### Protection against Conflict and Domination

A final family of arguments stresses that, if the members of a panel are chosen via idealized sortition, this procedure prevents interested parties from swaying the selection for their benefit [71, 97, 112]. Stone summarizes these arguments as follows:

> "First, [sortition] can prevent wrongful action on the part of the agent who must select officials. […] Second, it can prevent wrongful action on the part of the officials selected. If the method of selection is in any way predictable, outside interests might bribe or threaten officials into conformity with their wishes. If the method is unpredictable, then such wishes cannot be expressed at least until the results of the lottery become known. […] Finally, competing elites unable to stack the political process in their favor have less to fight about." [257]

In the practical setting of sortition, the additional stages of the selection process (as compared to idealized sortition) inherently create opportunities for dishonest agents to influence the composition and the decisions of the panel in ways that cannot be remedied by a change of selection algorithm. First, with respect to concerns about wrongful action on the part of the officials, the panel organizers wield a lot of influence in sending out the invitations, setting the quotas, and handling the process of selecting the panel from the pool.

More fundamentally, when any selection algorithm enforcing descriptive representation is used, a dishonest pool member can significantly increase their chances of selection by misrepresenting their features. For example, this pool member might pretend to have a different political orientation because they know that people with this orientation are unlikely to participate, and thus are likely to be underrepresented in the pool. Since, on average, the selection algorithm must choose pool members from this group with higher probability, reporting this feature will likely increase the agent's probability of being selected for the panel. So long as practitioners seek to enforce

descriptive representation in the presence of unequal rates of participation across subgroups, this type of manipulation seems unavoidable.

If, despite these challenges, one wanted to design a selection algorithm to discourage manipulation, one would have to target a specific kind of manipulation. For instance, for reducing the effect of bribing or intimidating pool members before they are selected, the algorithm within our framework minimizing the largest selection probabilities might be appropriate. Such an algorithm would increase the cost to the manipulator since any bribed pool member would have a substantial chance of not being selected to the panel, rendering the bribe futile. For other threat models, it would be natural for the selection algorithm to maximize not only the uncertainty of each agent being selected for the panel individually but the uncertainty about the composition of the whole panel. A selection algorithm maximizing this objective of *maximum entropy* could, in principle, be implemented by uniformly drawing sets of $k$ pool members, repeating this process until one set satisfies all quotas. Whether this selection algorithm can be sped up to the degree of being practically relevant is an interesting question for future work.

### B.4.2 Beyond Idealized Sortition, and the Objective of Maximal Fairness

As we have described, a large body of political theory literature characterizes the desiderata and benefits of *idealized* sortition. However, there is also research that engages, as we do in this work, with sortition beyond the idealized assumption that everyone is willing to participate. Such work often mentions *stratified sampling* [71, 188, 224, 250, 256] as a sampling method that can be used to reestablish descriptive representation despite differing response rates across subpopulations. For details on stratified sampling and how it relates to our work, see appendix B.3. In the political theory literature touching on stratified sampling, several authors point out that the benefits of idealized sortition do not perfectly extend to stratified sampling [96, 224, 250, 257]. To our knowledge, however, the literature stops short of proposing more gradual ideals, such as the maximal fairness objective we propose to approximate equality.

## B.5 Related Work on Panel Selection

The algorithmic problem of selecting panels for citizens' assemblies has motivated two previous papers. Both previous papers consider different models of sortition than does this work, and their results are not directly applicable to the practical setting we consider here.

In the first paper, Benadè, Gölz, and Procaccia [46] study a setting closely resembling what we call *idealized sortition* in appendix B.4 — that is, Benadè et al. assume that the panel-selection procedure can choose any constituent to participate (they assume it has full knowledge of the population) without taking into account that some constituents might not agree to serve on the panel. In this setting, *uniform sampling without replacement* is the most natural selection procedure, and it provides two important benefits: perfect equality of selection probabilities and probabilistic guarantees on the descriptive representation of any arbitrary group in the population. If one wants *deterministic* guarantees on descriptive representation along one specific category of attributes (say, gender), stratified sampling (appendix B.3) will give such guarantees. Benadè et al.

show that such deterministic guarantees can be imposed for certain groups with only marginal deterioration in the representation of other groups. Unfortunately, these results do not extend to the practical setting explored in this paper because, in addition to their unrealistic assumption that all constituents will participate, the set of quotas that can be imposed via stratified sampling is much more restrictive than those imposed in practice (see appendix B.3 for details).

The second paper, by Flanigan, Gölz, Gupta, and Procaccia [127], also develops a panel selection procedure, and, unlike Benadè et al., it accounts for the possibility that people invited to the panel may decline to join. Flanigan et al. consider the same general panel-selection pipeline as does this paper, with a uniform sample of the population being invited to participate, invitation recipients self-selecting into a pool of volunteers, and then a selection algorithm choosing the panel from the pool.

The main differences between the paper by Flanigan et al. and ours lies in the level of idealization of the models of sortition, and in the handling of quotas. On both of these counts, this paper engages more directly with the practical setting than does Flanigan et al.: In the present paper, we directly address the problem faced by practitioners when they sample their panel, which means taking as already decided the set of agents who opted into the pool and the quotas imposed by practitioners. As we described in the introduction, with these attributes of the problem already decided, equal selection probabilities are generally not attainable, which is why we focus on achieving equality to the maximum degree possible. By contrast, Flanigan et al. attempt to recover a notion of equal probabilities in an idealized probabilistic model of the panel-selection pipeline. Specifically, in their model, whether an invited agent joins the pool is decided by a biased coin flip, where the success probability of each agent's coin, the agent's *participation probability*, is known to the selection algorithm. Furthermore, quotas are not externally given, but are determined by what the selection algorithm can ensure for the given citizens' assembly. Under these assumptions and further assuming that all participation probabilities lie above a certain minimum bound, Flanigan et al. design a selection algorithm that achieves near-equal *end-to-end probabilities*, i.e., ensures that each agent reaches the panel from the population with similar probability. To do so, it prioritizes selecting those pool members who had the lowest probability of accepting their invitation, essentially canceling out the self-selection bias.

Note that Flanigan et al. and our paper pursue different notions of equality: Their paper aims to equalize the probability of each agent going *from population to panel* (calculated across all possible pools), whereas our paper aims for equality between the selection probabilities of members of a *single* pool. While their notion of equality is conceptually appealing, it is well-defined only relative to their modeling assumption that people decide to join the pool randomly. If one nevertheless wanted to apply their selection algorithm in practice, the agents' "participation probabilities" would have to be estimated using machine learning. Since, depending on these estimates, the selection algorithm might select an individual with much higher or lower selection probability, determining this number based on inherently imprecise techniques raises concerns about algorithmic bias and transparency. Finally, while their selection algorithm ensures some quotas, these guarantees only hold in the limit of very large pools and, even then, the gap between upper and lower quotas remains much looser than the gap between upper and lower quotas typically

imposed by practitioners.

### B.6 Computational Hardness

Here we show that, under standard complexity assumptions, there does not exist a selection algorithm (even an unfair one) that runs in polynomial time. At its core, this impossibility is a consequence of the following hardness result:

**Theorem B.6.1.** *For a given set of agents, panel size, and set of features with associated quotas, it is* NP*-hard to decide whether there exists a panel.*

*Proof.* By reduction from the NP-complete problem Exact Cover by 3-Sets (X3C) [142]. Fix an X3C instance consisting of a ground set $X$ with $|X| = 3q$ and of a collection $C$ of 3-element subsets of $X$. From this instance, construct an instance of the panel-selection problem as follows: Identify the pool members $N$ with the 3-sets $C$, create one feature $f_x$ per $x \in X$, and set the panel size $k$ to $q$. For every feature $f_x$, we impose quotas $\ell_{f_x} = u_{f_x} = 1$, and we set $N_{f_x}$ to the set of agents whose corresponding 3-set contains $x$.

It remains to show that there exists a panel iff there exists an exact cover for the X3C instance:
***m*⇒**: Suppose that there is a quota-compliant panel $P \subseteq N$. By the definition of the quotas, all features $f_x$ apply to exactly one agent in $P$. Thus, all elements $x \in X$ occur in exactly one of the three-sets corresponding to $P$, which means that this collection of 3-sets is an exact cover.
***m*⇐**: Let $C' \subseteq C$ be an exact cover for the X3C instance. Note that $|C'| = q = k$ because every set in $C'$ has exactly 3 elements and must cover a universe of size $|X| = 3q$. Set the panel $P$ to $C'$. Since $C'$ covers every element $x \in X$ exactly once, each feature $f_x$ applies to a single agent in $P$. This shows that the quotas are satisfied. □

Formally, the hardness of this decision problem does not *immediately* contradict the existence of polynomial-time selection algorithms, since our definition of a selection algorithm only allows for *instances* in the input of the algorithm, and instances are required to have at least one panel (appendix B.2). Nonetheless, the non-existence of polynomial-time algorithms follows as a simple corollary: if a selection algorithm produced a panel in polynomial time *with probability 1*, this would imply P = NP (corollary B.6.2 below), and, even if a selection algorithm succeeded at producing a panel in polynomial time only with constant probability, this would imply NP = RP (corollary B.6.3 below). The latter consequence would in turn imply NP = RP ⊆ P/poly [18] and thus that the polynomial-time hierarchy collapses [174], both of which are widely assumed to be false.

Since polynomial-time selection algorithms are unlikely to exist, this paper studies algorithms that are efficient in practice but whose worst-case running time might scale exponentially.

**Corollary B.6.2.** *Unless* P = NP*, there is no selection algorithm that finds a panel in polynomial time (with probability 1).*

*Proof.* By contrapositive. Suppose that there was a selection algorithm that would return a panel within $n^c$ computation steps for some constant $c$. Since our definition of instances assumes that all instances possess panels, this hypothetical algorithm may behave arbitrarily when provided with an input for which no panel exists. Still, this selection algorithm would allow to decide the NP-hard problem from theorem B.6.1 in polynomial time: Given a set of agents, a panel size, and a set of features, simply simulate the selection algorithm for $n^c$ steps and check whether a quota-compliant panel was returned. Since this polynomial-time algorithm decides an NP-hard problem, the existence of a polynomial-time selection algorithm would imply P = NP. □

**Corollary B.6.3.** *Unless* RP = NP*, there is no selection algorithm that, with constant probability, finds a panel in polynomial time.*

*Proof.* By contrapositive. Suppose that there was a selection algorithm that, for each instance, would succeed at returning a panel in $n^c$ computation steps (for some constant $c$) with constant probability. By again simulating this selection algorithm for $n^c$ steps and checking whether a quota-compliant panel was returned, one defines an RP-acceptor for the NP-hard language defined in theorem B.6.1, implying RP = NP. □

## B.7 Small Optimal Portfolios Exist

**Proposition B.7.1.** *Fix an arbitrary instance and a fairness measure $F$ for this instance. If there exists any maximally fair distribution over panels for $F$, there exists a maximally fair output distribution whose support includes at most $n + 1$ panels.*

*Proof.* Consider the hypercube $[0, 1]^n$, and associate each dimension with one agent. A panel $P$ can be embedded into this space by its characteristic vector $\vec{v}_P \in \{0, 1\}^n$, whose $i$th component is one exactly if $i \in P$.

Fix a maximally fair panel distribution, let $\mathcal{P}$ denote its support, and let $\{\lambda_P\}_{P\in\mathcal{P}}$ denote its probability mass function. Note that

$$\vec{p} \coloneqq \sum_{P\in\mathcal{P}} \lambda_P \vec{v}_P$$

is a probability allocation maximizing $F$, and that it is a convex combination of the $\{\vec{v}_P\}_{P\in\mathcal{P}}$. By Carathéodory's theorem, there is a subset $\mathcal{P}' \subseteq \mathcal{P}$ of size at most $n+1$ such that $\vec{p}$ still lies in the convex hull of this smaller set. Thus, there are nonnegative real numbers $\{\lambda'_P\}_{P\in\mathcal{P}'}$ adding up to one such that

$$\vec{p} = \sum_{P\in\mathcal{P}'} \lambda'_P \vec{v}_P.$$

These $\lambda'_P$ form the probability mass function of a distribution over at most $n + 1$ panels, which has the same probability allocation $\vec{p}$ as the original maximally fair distribution, which implies that the new distribution is also maximally fair for $F$. □

## B.8 Algorithmic Framework

In this section, we first summarize the high-level design of our algorithmic framework, how it is situated among existing algorithms and techniques, and how the framework applies to settings other than sortition. We then introduce the notion of a distribution-optimizer family, which encapsulates the information that the framework needs to optimize a fairness measure, and we formally describe the steps of the framework. Finally, we prove the correctness of the framework.

### B.8.1 Algorithmic Framework Overview and Context

At the highest level, each algorithm in our framework maximizes a concave function (the fairness measure). The approach our algorithms take to optimizing these concave functions generalizes a form of *column generation*, an algorithmic technique that is commonly used for solving linear programs with many variables and few constraints. [54] The existing column generation approach for solving such linear programs proceeds as follows: We first consider a version of the linear program in which all but a portfolio consisting of some $K$ of the variables are assumed to be non-basic and set to zero. This restricted version of the program then has only $K$ variables (and the same few constraints as in the original program), so its optimal primal and dual variables can be found efficiently. This primal solution (with zeros for the remaining variables) is then checked for optimality in the entire original program. This is done by looking for a column with negative reduced cost, i.e., a primal variable not currently in the portfolio such that slightly increasing its value from the current value of zero would lead to an increase in the objective. If such a column exists, it is then added to our portfolio of possibly basic variables, and the process is repeated for this slightly larger linear program. Once no such column exists, the solution for the restricted program is already optimal for the entire program.

Our column-generation algorithm applies the same general approach to convex programs satisfying strong duality. We are not aware of many previous papers applying column generation to convex optimization, and the papers we know of use column generation to refine linear approximations of convex functions, rather than directly optimizing the convex function over restricted sets of variables [54, 156]. One reason that column generation has not been applied to convex programs themselves might be that general convex programs may not have optimal solutions with few nonzero variables, and thus, column generation might not be faster than direct optimization of the full convex program. As we discuss below, however, the optimization problems considered in this paper have a special structure that ensures the existence of optimal solutions with few nonzero variables, which makes column generation a promising approach.

The convex program we solve, stated in its most general form, is as follows: Let $N$ be a finite set of *entities* (in our case: pool members), and let $\widehat{\mathcal{P}}$ be an implicitly defined (i.e., not explicitly given) family of subsets of $N$ (in our case: quota-compliant panels). Then, we consider a convex program of the following shape:

$$\begin{aligned} &\text{maximize } h(\vec{p}, \vec{x}) \\ &\text{subject to } g_r(\vec{p}, \vec{x}) \leq 0 && \forall 1 \leq r \leq m \\ &\qquad \vec{p} \in \mathsf{PossibleMarginals}(\widehat{\mathcal{P}}) && . \end{aligned} \tag{B.1}$$

Without the constraint in the last row, this would just be a general convex program, with a concave objective function $h$, $m$ many constraints defined by convex functions $g_r$, an arbitrary vector of variables $\vec{x}$, and a vector of special variables $\vec{p}$, one per entity. What makes this convex program special is the constraint "$\vec{p} \in \mathsf{PossibleMarginals}(\widehat{\mathcal{P}})$", which expresses that there exists some probability distribution over $\widehat{\mathcal{P}}$ such that the $p_1, \ldots, p_n$ in $\vec{p}$ are the entities' *marginals* induced by that distribution (where an entity's marginal is the probability that a set containing them is drawn from that distribution over $\widehat{\mathcal{P}}$). This last constraint could be easily expanded into additional linear constraints and exponentially many auxiliary variables $\lambda_P$, one for the probability mass of each set $P$ in $\widehat{\mathcal{P}}$, but this would require enumerating exponentially many sets in $\widehat{\mathcal{P}}$ and drastically increasing the size of the convex program. As we show in appendix B.7, Carathéodory's theorem implies that an optimal solution of this expanded program (if one exists) can set all but $|N| + 1$ of the $\lambda_P$ variables to zero.

Thus, our framework applies column generation to these $\lambda_P$ variables, repeatedly solving the expanded convex program under the restriction that all $\lambda_P$ except those in a small portfolio are non-basic and set to zero. Given some additional assumptions (guaranteeing that these restricted programs are solvable and satisfy strong duality), we can define the reduced cost of a set $P$ in $\widehat{\mathcal{P}}$ as a sum of Karush-Kuhn-Tucker (KKT) multipliers corresponding to the set's elements. Thus, our framework reduces optimizing the convex program with the special constraint "$\vec{p} \in \mathsf{PossibleMarginals}(\widehat{\mathcal{P}})$" to the problem of optimizing a linear objective over $\widehat{\mathcal{P}}$ (for finding the column with minimum reduced cost in each iteration of the column generation). When, as in this paper, $\widehat{\mathcal{P}}$ is implicitly defined by an ILP, the framework directly defines an algorithm by using an ILP solver for these subtasks.

### B.8.2 Applications of Framework to Other Problems

Solving convex programs of the form (B.1) identified above has immediate applications outside of sortition and to combinatorial structures other than quota-compliant panels: For example, Kurokawa, Procaccia, and Shah [182] study the problem of assigning classrooms to charter schools, where the implicit sets in $\widehat{\mathcal{P}}$ correspond to sets of schools that can simultaneously be matched in a bipartite matching with knapsack constraints. While Kurokawa et al. give an algorithm optimizing the leximin criterion in this domain, our framework immediately allows to optimize other fairness measures such as Nash welfare.

A second application lies in kidney exchange, where Roth, Sönmez and Ünver [242] again propose an algorithm for finding the leximin-optimal distribution over matchings, where each edge in the matching connects two donor–patient pairs matched for a 2-way exchange of kidneys. Not only does our framework allow the optimization of fairness measures other than leximin, but it also

extends to the more complex forms of kidney exchange encountered in practice, including longer cyclical exchanges and donation chains initiated by altruistic donors. The literature proposes multiple ILP formulations [93**?** ] that can be used for this purpose.

While both previous examples optimize individual fairness as their objective, our techniques apply to other convex optimization objectives as well. In appendix B.14.3, we give an example of an objective that optimizes the descriptive representation of groups rather than aiming for equal selection probabilities between individuals.

### B.8.3 Conditions for Applying the Framework

We now specify conditions that allow a convex program to be solved using our framework. Putting the outline in appendix B.8.1 into the language of panel selection, the column generation repeatedly (i) optimizes the convex program with the added restriction that the output probabilities of all panels not included in the current portfolio of panels $\mathcal{P}$ are set to zero, and then (ii) uses the KKT multipliers and an ILP solver to identify the panel to add to $\mathcal{P}$ that will allow the greatest marginal increase in fairness, until, eventually, the solution found in (i) is optimal for the unrestricted convex program. We will refer to the restricted convex program for a portfolio $\mathcal{P}$ as $C_{\mathcal{P}}$.

For the column generation to work, all programs $C_{\mathcal{P}}$ it optimizes should have an optimal solution and the KKT conditions should be necessary and sufficient. In particular, having an optimal solution implies that the portfolio must be non-empty from the start (since the output probabilities must add up to one, meaning that they cannot all be zero). We formalize these assumptions in a structure called a *distribution-optimizer family*:

**Definition B.8.1** (distribution-optimizer family)**.** *A distribution-optimizer family (DOF) $C$ for an instance is a family of convex programs that is fully specified by the tuple $(\mathcal{P}_{init}, t, h, \{g_r\}_r)$, where the four elements of this tuple are as follows:*

- *$\mathcal{P}_{init}$ is a non-empty portfolio of panels of the instance,*
- *$t \in \mathbb{N}_0$ is the number of auxiliary variables in each convex program,*
- *$h : ([0,1]^n \times \mathbb{R}^t) \to \mathbb{R}$ is a differentiable concave function (the objective of the convex programs), and*
- *the $g_r : ([0,1]^n \times \mathbb{R}^t) \to \mathbb{R}$ for $1 \leq r \leq m$ are some number $m \in \mathbb{N}_0$ of affine functions (defining auxiliary constraints in the convex programs).*[1]

*This tuple defines a family of convex programs $C = \{C_{\mathcal{P}}\}_{\mathcal{P} \supseteq \mathcal{P}_{init}}$, which includes one program $C_{\mathcal{P}}$ for each portfolio $\mathcal{P}$ in the instance such that $\mathcal{P} \supseteq \mathcal{P}_{init}$. Each such convex program $\mathcal{P}$ has variables $\{\lambda_P\}_{P \in \mathcal{P}}$ (representing the output probabilities of panels $P$), $\vec{p} = \{p_i\}_{i \in N}$ (representing the selection probabilities of agents $i$), and $\vec{x}$ (a $t$-dimensional vector of real-valued auxiliary variables), and the*

[1]The functions $g_r$ can be differentiable convex rather than affine as long as the strong duality of all convex problems $C_{\mathcal{P}}$ below is still ensured, for instance by Slater's condition.

*convex program is defined as follows:*

$$
\begin{aligned}
&\textit{maximize } h(\vec{p},\vec{x}) \\
&\textit{subject to } \sum_{P\in\mathcal{P}} \lambda_P = 1 && \textit{(output probabilities add to 1)} \\
&\qquad p_i = \sum_{\substack{P\in\mathcal{P}\\ i\in P}} \lambda_P && \forall i \in N \quad \textit{(marginals are sums of output probabilities)} \\
&\qquad g_r(\vec{p},\vec{x}) \le 0 && \forall 1 \le r \le m \quad \textit{(auxiliary constraints)} \\
&\qquad \lambda_P \ge 0 && \forall P \in \mathcal{P} \quad \textit{(output probabilities are nonnegative).}
\end{aligned}
$$

*For $C$ to be a DOF for the instance, in addition to being defined by a tuple as specified above, it must hold that all convex programs $C_{\mathcal{P}}$ for $\mathcal{P} \supseteq \mathcal{P}_{init}$ are solvable (i.e., they are feasible and the optimal value is attained).*

The algorithmic framework takes as input a specific instance and a DOF $C$ for this instance, and the framework then uses column generation to decide which convex programs from $C$ to run in what order to find the maximally fair distribution. Therefore, to use the framework to optimize a specific fairness measure $F$ on a given instance, one simply needs to find a DOF for that instance that optimizes $F$ (if one exists). The following definition formally connects a fairness measure with a DOF that optimizes it:

**Definition B.8.2** (implementation of a fairness measure by a DOF)**.** *For a specific instance, a fairness measure $F$ for the instance is implemented by a DOF $C = \{C_{\mathcal{P}}\}_{\mathcal{P}\supseteq\mathcal{P}_{init}}$ if, for any portfolio $\mathcal{P} \supseteq \mathcal{P}_{init}$, each optimal solution to $C_{\mathcal{P}}$ yields the probability mass function $\{\lambda_P^*\}_{P\in\mathcal{P}}$ of a distribution that is maximally fair according to $F$ among all distributions over the support $\mathcal{P}$.*

As we show below, for each DOF $C$ of an instance, it is easy to construct a fairness measure $F$ for that instance that is implemented by the DOF, by setting $F(\vec{p}) := \sup\{h(\vec{p},\vec{x}) \mid \vec{x} \in \mathbb{R}^t, \forall 1 \le r \le m.\ g_r(\vec{p},\vec{x}) \le 0\}$, with the convention that $\sup\emptyset = -\infty$. However, $C$ simultaneously implements other fairness measures whose optimization leads to the same optima (for example, the same DOF might implement the product of probabilities and the sum of their logarithms).

**Proposition B.8.3.** *For a fixed instance, a DOF $C = \{C_{\mathcal{P}}\}_{\mathcal{P}\supseteq\mathcal{P}_{init}}$ for this instance implements the fairness measure $F$ specified by*

$$F(\vec{p}) := \sup\{h(\vec{p},\vec{x}) \mid \vec{x} \in \mathbb{R}^t, \forall 1 \le r \le m.\ g_r(\vec{p},\vec{x}) \le 0\}.$$

*Proof.* Fix an instance and fix a portfolio $\mathcal{P} \supseteq \mathcal{P}_{init}$. Denote the optimal objective value of $C_{\mathcal{P}}$ by $obj^*$, and note that, by the definition of a DOF, this optimal value is attained.

We must show that, for any optimal solution of $C_{\mathcal{P}}$, the $\lambda_p^*$ are the probability mass function of a distribution that is maximally fair according to $F$ among distributions over the support $\mathcal{P}$, i.e., that the $\vec{p}^*$ optimize $F$. We will show this in two steps: In step (1), we show that, if $\vec{p}$ is the probability allocation corresponding to an optimal solution of $C_{\mathcal{P}}$, then $F(\vec{p}) = obj^*$. In step (2), we show

that, for each probability allocation $\vec{p}$ that can be obtained by a distribution over $\mathcal{P}$, it holds that $F(\vec{p}) \leq obj^*$. Together, these steps imply that a probability allocation $\vec{p}$ is optimal according to $F$ (among probability allocations of distributions over $\mathcal{P}$) iff $F(\vec{p}) = obj^*$, and that this is the case for the probability allocation of each panel distribution given by an optimal solution of $C_{\mathcal{P}}$.

**Step (1).** Consider an optimal solution $\vec{\lambda}^*, \vec{p}^*, \vec{x}^*$ to $C_{\mathcal{P}}$. Note that its objective value must be $obj^*$. Furthermore, note that if we added constraints fixing each selection probability $p_i$ to $p_i^*$ and each panel probability $\lambda_P$ to $\lambda_P^*$ to the convex program $C_{\mathcal{P}}$, the optimal objective value of the restricted problem would still be $obj^*$ and would still be attained. Since $F(\vec{p})$ is defined as the optimal objective value of this restricted problem, $F(\vec{p}) = obj^*$.

**Step (2).** Now, consider any probability allocation $\vec{p}^*$ that is the result of a distribution $\mathcal{D}$ over $\mathcal{P}$. By fixing $\vec{p}$ in $C_{\mathcal{P}}$ to $\vec{p}^*$ and by fixing $\vec{\lambda}$ to the probability mass function of $\mathcal{D}$, $C_{\mathcal{P}}$ simplifies to the optimization problem defining $F(\vec{p})$, which means that the optimal objective value $obj^*$ of the full convex program $C_{\mathcal{P}}$ is at least $F(\vec{p})$. □

### B.8.4 Definition of Framework

As described above, the algorithmic framework is an algorithm that takes as input an instance and a DOF of that instance. The framework then computes a distribution over panels that is maximally fair with respect to the fairness measure implemented by the DOF, and then samples this distribution to select the final panel. The full algorithm is specified below:

**Algorithm 2** Framework

**Input:** an instance and a corresponding DOF $C = \{C_{\mathcal{P}}\}_{\mathcal{P} \supseteq \mathcal{P}_{init}}$
**Output:** a randomly chosen panel for the instance

$\mathcal{P} \leftarrow \mathcal{P}_{init}$ **while** *true* **do**
  let $\vec{\lambda}^*, \vec{p}^*, \vec{x}^*$ denote an optimal solution for $C_{\mathcal{P}}$, and let $\mu_r^*$ be the dual value for each constraint $g_r(\vec{p}, \vec{x}) \leq 0$ at this optimum **for** $i \in N$ **do**
    $\eta_i^* \leftarrow \frac{\partial}{\partial p_i} h(\vec{p}^*, \vec{x}^*) - \sum_{r=1}^{m} \mu_r^* \frac{\partial}{\partial p_i} g_r(\vec{p}^*, \vec{x}^*)$
  $P_{new} \leftarrow$ panel $P$ maximizing $\sum_{i \in P} \eta_i^*$, found by ILP ($P$ need not be in $\mathcal{P}$) $P_{old} \leftarrow$ some panel $P \in \mathcal{P}$ such that $\lambda_P^* > 0$ **if** $\sum_{i \in P_{old}} \eta_i^* \geq \sum_{i \in P_{new}} \eta_i^*$ **then**
    $\mathcal{D} \leftarrow$ distribution over $\mathcal{P}$ with probability mass function $\vec{\lambda}^*$ **return** panel drawn from $\mathcal{D}$
  **else**
    $\mathcal{P} \leftarrow \mathcal{P} \cup \{P_{new}\}$

### B.8.5 Termination and Correctness of Framework

It remains to show that the above algorithm always terminates (theorem B.8.4) and that it selects panels in a maximally fair way (theorem B.8.5). In the proofs of these theorems, we will

extensively use the Karush-Kuhn-Tucker (KKT) conditions for the convex optimization problems $C_{\mathcal{P}}$. Consider a specific instance and a specific DOF $C = \{C_{\mathcal{P}}\}_{\mathcal{P} \supseteq \mathcal{P}_{init}}$ for this instance. Then, we denote

- the dual variable of the constraint $\sum_{P \in \mathcal{P}} \lambda_P = 1$ by $\eta_0$,
- the dual variables of the constraints $p_i = \sum_{P \in \mathcal{P} : i \in P} \lambda_P$ by $\eta_i$,
- the dual variables of the constraints $g_r(\vec{p}, \vec{x}) \leq 0$ by $\mu_r$, and
- the dual variables of the constraints $\lambda_P \geq 0$ by $\nu_P$.

Since $C_{\mathcal{P}}$ satisfies strong duality, the following KKT conditions are necessary and sufficient for optimality:

$$\sum_{P \in \mathcal{P}} \lambda_P = 1 \tag{B.2}$$

$$p_i = \sum_{\substack{P \in \mathcal{P} \\ i \in P}} \lambda_P \qquad \forall i \in N \tag{B.3}$$

$$g_r(\vec{p}, \vec{x}) \leq 0 \qquad \forall 1 \leq r \leq m \tag{B.4}$$

$$\lambda_P \geq 0 \qquad \forall P \in \mathcal{P} \tag{B.5}$$

$$\mu_r \geq 0 \qquad \forall 1 \leq r \leq m \tag{B.6}$$

$$\nu_P \geq 0 \qquad \forall P \in \mathcal{P} \tag{B.7}$$

$$\mu_r \, g_r(\vec{p}, \vec{x}) = 0 \qquad \forall 1 \leq r \leq m \tag{B.8}$$

$$\nu_P \, \lambda_P = 0 \qquad \forall P \in \mathcal{P} \tag{B.9}$$

$$\left( \sum_{i \in P} \eta_i \right) + \nu_P = \eta_0 \qquad \forall P \in \mathcal{P} \tag{B.10}$$

$$\eta_i = \frac{\partial}{\partial p_i} h(\vec{p}, \vec{x}) - \sum_{r=1}^{m} \mu_r \, \frac{\partial}{\partial p_i} g_r(\vec{p}, \vec{x}) \qquad \forall i \in N \tag{B.11}$$

$$\nabla_{\vec{x}} \, h(\vec{p}, \vec{x}) = \sum_{r=1}^{m} \mu_r \, \nabla_{\vec{x}} \, g_r(\vec{p}, \vec{x}) \tag{B.12}$$

In the following proofs, we will denote the set of all panels of the instance by $\widehat{\mathcal{P}}$.

**Theorem B.8.4.** *algorithm 2 terminates.*

*Proof.* Fix the input instance and the DOF $C = \{C_{\mathcal{P}}\}_{\mathcal{P} \supseteq \mathcal{P}_{init}}$. It suffices to show that $\mathcal{P}$ grows in every iteration since it is always a subset of the finite set $\widehat{\mathcal{P}}$ of all panels of the instance. More specifically, we need to show that, whenever the if branch in line 4 is not taken, $P_{new}$ was not yet in $\mathcal{P}$.

Note that, in line 3 of algorithm 2, the $\eta_i^*$ are set equal the dual variables $\eta_i$ at the optimum of $C_{\mathcal{P}}$

by eq. (B.11).[1] From complementary slackness (B.9) and the precondition $\lambda_{P_{old}} > 0$ (line 4), we know that $v_{P_{old}} = 0$, and thus, by eq. (B.10), that

$$\sum_{i \in P_{old}} \eta_i^* = \eta_0^* = \left( \sum_{i \in P'} \eta_i^* \right) + v_{P'}^* \geq \sum_{i \in P'} \eta_i^*$$

for all $P' \in \mathcal{P}$, where the last step uses eq. (B.7). Since, by assumption, the if branch in line 4 was not taken, we know that $\sum_{i \in P_{new}} \eta_i^* > \sum_{i \in P_{old}} \eta_i^* \geq \sum_{i \in P'} \eta_i^*$ for all $P' \in \mathcal{P}$, which shows that $P_{new}$ was not yet in $\mathcal{P}$. □

**Theorem B.8.5.** *Fix any instance, and let a DOF $C = \{C_{\mathcal{P}}\}_{\mathcal{P} \supseteq \mathcal{P}_{init}}$ for this instance implement a fairness measure $F$. Then, when algorithm 2 is called with the instance and $C$, its output distribution is maximally fair according to $F$.*

*Proof.* Consider the point in the execution of algorithm 2 just before returning, when the algorithm defines the distribution $\mathcal{D}$ in line 5. Since all computation steps so far are deterministic, and since the algorithm subsequently just returns a panel drawn from $\mathcal{D}$, $\mathcal{D}$ is the output distribution of the algorithm when given these inputs. It remains to show that $\mathcal{D}$ is maximally fair according to $F$.

Since $C_{\mathcal{P}}$ (for the value of $\mathcal{P}$ when the algorithm is in line 5) satisfies strong duality, we know that the variables $\vec{\lambda}^*, \vec{p}^*, \vec{x}^*, \vec{\mu}^*, \vec{\eta}^*$ can be extended by variables $(v_P^*)_{P \in \mathcal{P}}$ and $\eta_0^*$ to satisfy the KKT conditions of $C_{\mathcal{P}}$.

We will extend these variables for $C_{\mathcal{P}}$ to variables satisfying the KKT conditions for the larger convex program $C_{\widehat{\mathcal{P}}}$. In this extension, we preserve the values of all variables already present from $C_{\mathcal{P}}$, and set $\lambda_P^* := 0$ and $v_P^* := \eta_0^* - \sum_{i \in P} \eta_i^*$ for all $P \in \widehat{\mathcal{P}} \setminus \mathcal{P}$.

Next, we show that this assignment satisfies the KKT conditions for $C_{\widehat{\mathcal{P}}}$. Most of the conditions directly follow from the assumption that the KKT conditions hold for $C_{\mathcal{P}}$ because all variables in the equation remained the same (eqs. (B.4), (B.6), (B.8), (B.11) and (B.12); and eqs. (B.5), (B.7), (B.9) and (B.10) for all $P \in \mathcal{P}$). The first two conditions (eqs. (B.2) and (B.3)) are preserved because all newly introduced $\lambda_P^*$ are zero. Clearly, all $\lambda_P^*$ are nonnegative (eq. (B.5)). Similarly, the added $v_P^*$ for $P \in \widehat{\mathcal{P}} \setminus \mathcal{P}$ are nonnegative (eq. (B.7)) because the algorithm took the if branch in line 4, which means that

$$\sum_{i \in P} \eta_i^* \leq \sum_{i \in P_{new}} \eta_i^* \leq \sum_{i \in P_{old}} \eta_i^* \leq \left( \sum_{i \in P_{old}} \eta_i^* \right) + v_{P_{old}}^* = \eta_0^*.$$

Complementary slackness (eq. (B.9)) is satisfied because the added $\lambda_P^*$ are zero, and condition (B.10) holds by the definition of the new $v_P^*$. This shows that all KKT conditions for $C_{\widehat{\mathcal{P}}}$ are satisfied, implying the constructed assignment is optimal.

[1] Thus, the algorithm could alternatively have been written as taking the $\eta_i^*$ directly as the optimal dual variable values of the $\eta_i$. We do not do so to avoid ambiguity in the sign of $\eta_i^*$ and to stress that $\sum_{i \in P} \eta_i^*$ can be understood as a reduced cost of the column $\lambda_P$, based on the gradient of the convex function.

Since $C$ implements the fairness measure $F$, the distribution whose probability mass function is given by the constructed $\lambda_p^*$ is maximally fair among distributions over the support $\widehat{\mathcal{P}}$, and therefore maximally fair among all output distributions. Since, in extending the assignment, we only added $\lambda_p^*$ variables with value 0, $\mathcal{D}$ is equal to this maximally fair distribution. □

### B.9 Fairness Measures

In different sub-areas of fair division, researchers have developed metrics measuring how fairly utility is distributed over individuals by a given allocation of a resource [110, 206]. By casting the problem of panel selection as a fair-division problem below, we demonstrate how these metrics can be used to quantify the fairness of probability allocations produced by selection algorithms:

> *Consider each quota-compliant panel in a given instance to be a distinct public good, and suppose that society can select exactly one of these goods, possibly through a random lottery. Each agent in the pool has value 1 for any panel on which they are featured, and value 0 for any panel on which they are not featured; and an agent's utility for a lottery over panels is their expected value for the drawn panel.*

In this setup, each pool member's utility is exactly their selection probability, which is determined by the selected lottery over panels. Therefore, metrics for measuring the fairness of a utility profile in the fair division literature can be applied to measure the fairness of a distribution over panels by giving them a probability allocation as their input rather than a vector of utilities.

Now, we describe multiple metrics from the fair-division literature that can be used as fairness measures in the panel-selection setting. In the subsections below, we show how each of these fairness measures can be maximized using our framework.

**Egalitarian social welfare [111]:** Maximize the lowest selection probability, $\min_{i \in N} p_i$.

**Gini coefficient [110, 187]:** Minimize half of the relative mean absolute difference,

$$\frac{\sum_{i \in N} \sum_{j \in N} |p_i - p_j|}{2\, n \sum_{i \in N} p_i}.$$

**Atkinson indices [110, 247]:** For a given parameter $\epsilon \in (0, 1)$, minimize

$$1 - \frac{n}{\sum_{i \in N} p_i} \left( \frac{\sum_{i \in N} p_i^{1-\epsilon}}{n} \right)^{1/(1-\epsilon)}.$$[1]

**Nash social welfare [206]:** Maximize the product of selection probabilities, $\prod_{i \in N} p_i$.

Recall that our definition of a fairness measure (appendix B.2) assumes that higher values indicate higher levels of fairness. Thus, the sign of the Gini coefficient and the Atkinson indices needs to be inverted to obtain a fairness measure according to our formal definition.

[1] Note that, in our setting, minimizing the Atkinson index for $\epsilon = 1$ coincides with maximizing Nash welfare.

Given that Nash social welfare and egalitarian social welfare are listed as fairness measures above, one might expect utilitarian social welfare (i.e., the sum of selection probabilities) to also appear. However, since the sum of selection probabilities is equal to $k$ for all probability allocations, utilitarian welfare is a constant function in our setting, which can hardly be considered a measurement of fairness.

Another important formalization of fairness from the fair-division literature is the *leximin criterion* [206], which we implement in our algorithm Leximin . Recall that the leximin objective not only maximizes the lowest selection probability (as does egalitarian welfare), but then breaks ties in favor of the second-lowest selection probability, the third-lowest selection probability and so on. Since this objective cannot be represented as the maximization of a single real-valued score [206], leximin cannot formally be expressed as a fairness measure according to our definition (appendix B.2). Nevertheless, the leximin criterion defines a weak ordering of probability allocations, which is enough to define a maximally fair probability allocation. Specifically, to compare two probability allocations $\{p_i\}_{i\in N}$ and $\{q_i\}_{i\in N}$, one represents each by a vector of probability values sorted in non-decreasing order and compares these vectors using the lexicographic order.

### B.9.1 Maximizing Egalitarian Welfare

For any instance, the egalitarian-welfare fairness measure is defined by

$$F_{egal}(\vec{p}) = \min_{i\in N} p_i.$$

Let $P_o$ be an arbitrary panel for the instance, which can be found by ILP. We will show that the DOF $C_{egal} = \{C_{\mathcal{P}}\}_{\mathcal{P}}$ defined by the tuple

$$\langle \{P_o\}, 1, (\vec{p}, x) \mapsto x, \{(\vec{p}, x) \mapsto x - p_i\}_{i\in N} \rangle$$

implements $F_{egal}$. Since $t$, $h$, and the $g_r$ can be read from the convex optimization problem, it is more convenient to implicitly specify them via the parametric convex program $C_{\mathcal{P}}$:

$$\begin{aligned}
&\text{maximize } x \\
&\text{such that } \sum_{P\in\mathcal{P}} \lambda_P = 1 \\
&\qquad p_i = \sum_{\substack{P\in\mathcal{P} \\ i\in P}} \lambda_P && \forall i \in N \\
&\qquad x - p_i \le 0 && \forall i \in N \\
&\qquad \lambda_P \ge 0 && \forall P \in \mathcal{P}.
\end{aligned}$$

**Proposition B.9.1.** *For each instance, $C_{egal}$ is a DOF.*

*Proof.* We must show that, for each $\mathcal{P} \supseteq \mathcal{P}_{init} = \{P_o\}$, the optimal value of $C_{\mathcal{P}}$ is attained. Since $C_{\mathcal{P}}$ is a linear program, this reduces to showing that the program is feasible and bounded.

For any $\mathcal{P} \supseteq \{P_o\}$, $C_{\mathcal{P}}$ is feasible by setting $\lambda_{P_o} \coloneqq 1$, $\lambda_P \coloneqq 0$ for all other $P \in \mathcal{P}$, by setting the $p_i$ according to their functional dependency on the $\lambda_P$, and by setting $x \coloneqq 0$. Furthermore, the optimal value is bounded from above since, in any valid assignment, fixing an arbitrary agent $i \in N$,

$$x \leq p_i = \sum_{\substack{P \in \mathcal{P} \\ i \in P}} \lambda_P \leq \sum_{P \in \mathcal{P}} \lambda_P = 1. \qquad \square$$

**Proposition B.9.2.** *For each instance, the fairness measure $F_{egal}$ for this instance is implemented by the DOF $C_{egal}$ for this instance.*

*Proof.* By proposition B.8.3, $C_{egal}$ implements the fairness measure $F$ given by

$$\begin{aligned} F(\vec{p}) &= \sup\{x \mid x \in \mathbb{R}, \forall i \in N.\ x - p_i \leq 0\} \\ &= \sup\{x \mid x \in \mathbb{R}, \forall i \in N.\ x \leq p_i\} \\ &= min_{i \in N} p_i. \end{aligned} \qquad \square$$

### B.9.2 Minimizing the Gini Coefficient

For any instance, the Gini-coefficient fairness measure is defined by

$$F_{gini}(\vec{p}) = -\frac{\sum_{i \in N} \sum_{j \in N} |p_i - p_j|}{2\, n \sum_{i \in N} p_i}.$$

Again, let $P_o$ be an arbitrary panel of the instance, found by ILP. We will show that the DOF $C_{gini} = \{C_{\mathcal{P}}\}_{\mathcal{P} \supseteq \mathcal{P}_{init}}$ implements $F_{gini}$, where $C_{gini}$ is defined by setting $\mathcal{P}_{init} \coloneqq \{P_o\}$ and by implicitly defining $t$, $h$, and the $g_r$ through the following convex program $C_{\mathcal{P}}$:

$$\begin{aligned} \text{maximize } & -\sum_{i<j \in N} x_{i,j} \\ \text{such that } & \sum_{P \in \mathcal{P}} \lambda_P = 1 \\ & p_i = \sum_{\substack{P \in \mathcal{P} \\ i \in P}} \lambda_P && \forall i \in N \\ & -x_{i,j} + p_i - p_j \leq 0 && \forall i < j \in N \\ & -x_{i,j} - p_i + p_j \leq 0 && \forall i < j \in N \\ & \lambda_P \geq 0 && \forall P \in \mathcal{P}, \end{aligned}$$

where "$i < j \in N$" is short-hand for requiring that $i, j \in N$ and that $i$ precedes $j$ in a canonical ordering over agents.

**Proposition B.9.3.** *For each instance, $C_{gini}$ is a DOF.*

*Proof.* We must show that, for each $\mathcal{P} \supseteq \mathcal{P}_{init} = \{P_o\}$, the optimal value of $C_{\mathcal{P}}$ is attained. Since $C_{\mathcal{P}}$ is a linear program, it suffices to show that the program is feasible and bounded.

For any $\mathcal{P} \supseteq \{P_o\}$, $C_{\mathcal{P}}$ is feasible by setting $\lambda_{P_o} \coloneqq 1$, $\lambda_P \coloneqq 0$ for all other $P \in \mathcal{P}$, by setting the $p_i$ according to their functional dependency on the $\lambda_P$, and by setting all $x_{i,j}$ to 1 (since then, e.g., $-x_{i,j} + p_i - p_j \leq -1 + p_i \leq 0$). Furthermore, the optimal value is bounded from above since, in any valid assignment, the $x_{i,j}$ are constrained to be at least $p_i - p_j$ and at least $-p_i + p_j = -(p_i - p_j)$, which means that all $x_{i,j}$ are nonnegative and, thus, that $-\sum_{i<j\in N} x_{i,j}$ cannot be positive. □

**Proposition B.9.4.** *For each instance, the fairness measure $F_{gini}$ for this instance is implemented by the DOF $C_{gini}$ for this instance.*

*Proof.* By proposition B.8.3, $C_{gini}$ implements the fairness measure $F$ given by

$$\begin{aligned}
F(\vec{p}) &= \sup \left\{ -\textstyle\sum_{i<j\in N} x_{i,j} \;\middle|\; \begin{array}{l} \{x_{i,j}\}_{i<j\in N} \in \mathbb{R}^{\binom{n}{2}}, \\ \forall i, j \in N.\ x_{i,j} \geq p_i - p_j \text{ and } x_{i,j} \geq p_j - p_i \end{array} \right\} \\
&= \sup \left\{ -\textstyle\sum_{i<j\in N} x_{i,j} \;\middle|\; \begin{array}{l} \{x_{i,j}\}_{i<j\in N} \in \mathbb{R}^{\binom{n}{2}}, \\ \forall i, j \in N.\ x_{i,j} \geq |p_i - p_j| \end{array} \right\} \\
&= -\textstyle\sum_{i<j\in N} |p_i - p_j| \\
&= -\dfrac{\sum_{i\in N} \sum_{j\in N} |p_i - p_j|}{2} \\
&= F_{gini}(\vec{p})\, n \textstyle\sum_{i\in N} p_i \\
&= F_{gini}(\vec{p})\, n\, k.
\end{aligned}$$

Thus, $C_{gini}$ implements a fairness measure that is just $F_{gini}$ times the positive constant $n\,k$. Since multiplying a fairness measure by a positive constant does not change which probability allocations maximize the fairness measure, $C_{gini}$ also implements $F_{gini}$. □

### B.9.3 Minimizing the Atkinson Indices for $0 < \epsilon < 1$

For a fixed instance, and a fixed constant $\epsilon \in (0, 1)$, the Atkinson-index fairness measure is defined by

$$F_{atkinson}(\vec{p}) = \frac{n}{\sum_{i\in N} p_i} \left( \frac{\sum_{i\in N} p_i^{1-\epsilon}}{n} \right)^{1/(1-\epsilon)} - 1.$$

Again, let $P_o$ be an arbitrary panel of the instance, found by ILP. We will show that the DOF $C_{atkinson} = \{C_{\mathcal{P}}\}_{\mathcal{P}\supseteq\mathcal{P}_{init}}$ implements $F_{atkinson}$, where $C_{atkinson}$ is defined by setting $\mathcal{P}_{init} \coloneqq \{P_o\}$ and by implicitly defining $t$, $h$, and the $g_r$ through the following convex program $C_{\mathcal{P}}$:

$$
\begin{aligned}
\text{maximize } & \sum_{i \in N} p_i^{1-\epsilon} \\
\text{such that } & \sum_{P \in \mathcal{P}} \lambda_P = 1 \\
& p_i = \sum_{\substack{P \in \mathcal{P} \\ i \in P}} \lambda_P && \forall i \in N \\
& \lambda_P \geq 0 && \forall P \in \mathcal{P}.
\end{aligned}
$$

**Proposition B.9.5.** *For each instance, $C_{gini}$ is a DOF.*

*Proof.* We must show that, for each $\mathcal{P} \supseteq \mathcal{P}_{init} = \{P_o\}$, the optimal value of $C_{\mathcal{P}}$ is attained. Since there are no auxiliary constraints, feasibility is trivial given that $\mathcal{P}$ is nonempty. Since there are no auxiliary variables, all variables are naturally bounded in $[0, 1]$. Since the domain of valid assignments for $\vec{\lambda}$ and $\vec{p}$ is bounded and closed, thus compact, the continuous function $h$ attains its maximum on this domain. □

**Proposition B.9.6.** *For each instance, the fairness measure $F_{atkinson}$ for this instance is implemented by the DOF $C_{atkinson}$ for this instance.*

*Proof.* By proposition B.8.3, $C_{atkinson}$ implements the fairness measure $F$ given by

$$
\begin{aligned}
F(\vec{p}) &= \sup\{\textstyle\sum_{i \in N} p_i^{1-\epsilon}\} \\
&= \textstyle\sum_{i \in N} p_i^{1-\epsilon} \\
&= n \left(k/n \left(F_{atkinson}(\vec{p}) + 1\right)\right)^{1-\epsilon}.
\end{aligned}
$$

Since $F$ can be obtained by composing $F_{atkinson}$ with a strictly monotone function, it has the same maximally fair probability allocations. This shows that $C_{atkinson}$ also implements $F_{atkinson}$. □

### B.9.4 Maximizing Nash Social Welfare

For a fixed instance, and a fixed constant $\epsilon \in (0, 1)$, the Nash-welfare fairness measure is defined by

$$
F_{nash}(\vec{p}) = \prod_{i \in N} p_i.
$$

Using an ILP solver, one can determine all agents $i \in N$ who appear on any panel. If any agent $i$ does not appear on a panel, their selection probability must be 0, which means that $F_{nash}$ is constant on all probability allocations and can be maximized by deterministically returning any

panel.[1] Thus, without loss of generality, we assume that each agent $i \in N$ is contained in a panel $P_i$, which can be found by $n$ ILP calls.

Consider the family of concave programs $C_{nash} = \{C_{\mathcal{P}}\}_{\mathcal{P} \supseteq \mathcal{P}_{init}}$ where $\mathcal{P}_{init} = \{P_i \mid i \in N\}$ and the convex program $C_{\mathcal{P}}$ is given as

$$\begin{aligned}
\text{maximize} \quad & \sum_{i \in N} \log p_i \\
\text{such that} \quad & \sum_{P \in \mathcal{P}} \lambda_P = 1 \\
& p_i = \sum_{\substack{P \in \mathcal{P} \\ i \in P}} \lambda_P && \forall i \in N \\
& \lambda_P \geq 0 && \forall P \in \mathcal{P}.
\end{aligned}$$

We will show that, by inserting this family of concave programs into our framework, the framework optimizes $F_{nash}$. A formal complication is that the objective function $h$ defined above is not real-valued for all probability allocations, since it is $-\infty$ whenever one selection probability is zero. Thus, this family does not *quite* fit into our definition of a DOF. However, the proof of optimality of the framework still goes through given that the $C_P$ can be optimized by a convex-program solver and that the optimal values of all $C_{\mathcal{P}}$ are real-valued:

**Proposition B.9.7.** *For each $C_{\mathcal{P}}$ for some $\mathcal{P} \supseteq \mathcal{P}_{init}$, the optimal objective value is real-valued and attained.*

*Proof.* Fix some $\mathcal{P} \supseteq \mathcal{P}_{init}$. We will first show that the optimal objective value is not $-\infty$. Indeed, consider the distribution obtained by selecting each panel $P_i$ with probability $1/|\mathcal{P}_{init}|$. Since, by construction, each agent is contained in at least one panel in $\mathcal{P}_{init}$, each selection probability $p_i$ is at least $1/|\mathcal{P}_{init}| \geq 1/n$. This means that an objective value of $n \log(1/n) > -\infty$ can be attained and that the constraints are feasible. Furthermore, it shows that any probability allocation that selects some agent $i$ with probability strictly less than $1/n^n$ cannot be optimal, because its objective value $\sum_{j \in N} \log p_j \leq \log p_i < n \log(1/n)$ is lower than the previous value.

It remains to show that the optimal objective value can be attained. Consider the space of all valid assignments $\vec{\lambda}, \vec{p}$, which is bounded and closed. By the argument above, we do not change the optimal objective value of $C_{\mathcal{P}}$ by further restricting the program with the constraints $p_i \geq 1/n^n$ for all $i$, and the space of assignments for $\vec{\lambda}, \vec{p}$ still stays compact in this operation. Since $h(\vec{p}) = \sum_{i \in N} \log p_i$ is real-valued and continuous on this space, its maximum is attained. □

**Proposition B.9.8.** *For each instance, plugging $C_{nash}$ into the framework yields an output distribution that is maximally fair according to $F_{nash}$.*

[1] In practice, one would instead remove all agents from the pool who are not contained in any panel, and optimize Nash social welfare for the resulting instance with fewer agents.

*Proof.* Following the reasoning of the proof of theorem B.8.5, one shows that the probability mass function of the output distribution is optimal according to $C_{\vec{p}}$ in $C_{nash}$. By the reasoning of proposition B.8.3, this yields a probability allocation that maximizes the fairness measure $F$ given by

$$\begin{aligned} F(\vec{p}) &= \sup\{\textstyle\sum_{i \in N} \log p_i\} \\ &= \textstyle\sum_{i \in N} \log p_i \\ &= \log(F_{nash}(\vec{p})). \end{aligned}$$

Since this is a strictly monotone transformation of $F_{nash}$, the output distribution must also be maximally fair for $F_{nash}$. □

## B.10 Description of Leximin

### B.10.1 Overview

As we discussed in appendix B.9, leximin is not formally a fairness measure according to our definition, which means that it cannot be optimized with a single application of our framework. Instead, we repeatedly invoke the framework for different auxiliary DOFs as follows: In the first application of the framework, we maximize the minimum probability. Subject to fixing the selection probability of a specific set of agents at this value (we discuss below how these agents are chosen), we then maximize the minimum selection probability among all other agents in a second application of the framework. We continue by fixing the selection probabilities of more and more agents to their value in the leximin allocation until all probabilities are fixed.

The crucial step in the algorithm is knowing which agents' probabilities to fix in each iteration. For example, the first invocation of the framework, which maximizes the minimum selection probability, might result in a probability allocation in which multiple agents have this minimum selection probability. In this case, not all of these agents must have this minimum selection probability in the leximin-optimal distribution, so it is not obvious whose selection probability should be fixed. As in previous work [209], complementary slackness allows us to identify at least one agent in each iteration whose selection probability must be minimal across *all* distributions optimizing the current iteration's DOF. Since all leximin-optimal distributions are optimal for the current DOF, we can fix these agents' selection probabilities.

In the following, we first define the auxiliary DOFs and the Leximin algorithm. Then, we prove the correctness of the algorithm.

### B.10.2 Definition of Leximin

To define the algorithm, we must first specify the auxiliary DOFs used by it. Each auxiliary DOF is a family $C_{aux}(R, \rho, \mathcal{P}_{init})$ parametrized by a set $R \subsetneq N$ of agents and by a function $\rho : R \to [0, 1]$,

which together represent that the selection probability of each agent $i \in R$ has been fixed to $\rho(i)$; and by an initial portfolio.

For a set of agents $R \subsetneq N$, a function $\rho\colon R \to [0,1]$, and a non-empty portfolio $\mathcal{P}_{init}$, the DOF $C_{aux}(R, \rho, \mathcal{P}_{init}) = \{C_{\mathcal{P}}\}_{\mathcal{P} \supseteq \mathcal{P}_{init}}$ for an instance is defined via the initial portfolio $\mathcal{P}_{init}$ and the following optimization problem $C_{\mathcal{P}}$:

$$
\begin{aligned}
&\text{maximize } x \\
&\text{such that } \sum_{P \in \mathcal{P}} \lambda_P = 1 \\
&\qquad p_i = \sum_{\substack{P \in \mathcal{P} \\ i \in P}} \lambda_P && \forall i \in N \\
&\qquad x - p_i \leq 0 && \forall i \in N \setminus R \\
&\qquad p_i - \rho(i) \leq 0 && \forall i \in R \\
&\qquad \rho(i) - p_i \leq 0 && \forall i \in R \\
&\qquad \lambda_P \geq 0 && \forall P \in \mathcal{P}.
\end{aligned}
$$

We will show in lemma B.10.1 below that, whenever Leximin applies the framework to such a $C_{aux}(R, \rho, \mathcal{P}_{init})$, it indeed defines a DOF. Furthermore, we show in lemma B.10.2 that this DOF maximizes $\min_{i \in N \setminus R} p_i$ among all probability allocations that select each $i \in R$ with probability exactly $\rho(i)$.

We now define the Leximin algorithm:

**Algorithm 3** Leximin

**Input:** an instance
**Output:** a randomly chosen panel for the instance

8 $\mathcal{P}_{lexi} \leftarrow$ {arbitrary panel $P_o$ found by ILP} $R \leftarrow \emptyset$ initialize empty function $\rho : R \to [0,1]$ $\mathcal{D} \leftarrow$ deterministic distribution with value $P_o$ (for analysis only) **while** $R \subsetneq N$ **do**
9  execute algorithm 2 up to line 5 with the instance and the DOF $C_{aux}(R, \rho, \mathcal{P}_{lexi})$ as input; set $\vec{p}^*, \vec{\mu}^*, \mathcal{D}$ to their final values inside the subprocedure call; and set $\mathcal{P}_{lexi}$ to the final value of $\mathcal{P}$ in the call **for** $i \in N \setminus R$ **do**
10  **if** $\mu_r^* > 0$ *for* $r$ *corresponding to constraint* $x - p_i \leq 0$ **then**
11  $R \leftarrow R \cup \{i\}$ $\rho(i) \leftarrow p_i^*$
12 **return** panel drawn from $\mathcal{D}$

Note that, since $N \neq \emptyset$, the loop is executed at least once and the initialization of $\mathcal{D}$ in line 8 will never be used. However, this initialization will be convenient in the proof of correctness. In

theorems B.10.3 and B.10.4, we prove that the selection algorithm terminates and that it is indeed maximally fair according to the leximin criterion.

Our practical implementation of Leximin deviates from the formal specification of algorithm 3 by the following modifications, which speed up the practical runtime while preserving optimality: (i) implementing lines 2 to 3 of algorithm 2 purely in terms of the dual linear program, by (ii) solving these linear programs using interior-point barrier methods (which typically allow to fix more probabilities per iteration) and by (iii) initializing $\mathcal{P}_{lexi}$ in line 8 with multiple panels found through a multiplicative-weight heuristic.

### B.10.3 Proofs

**Lemma B.10.1.** *Whenever algorithm 3 applies the framework with an instance and $C_{aux}(R, \rho, \mathcal{P}_{lexi})$, the latter is a DOF for the instance.*

*Proof.* Fix any $\mathcal{P} \supseteq \mathcal{P}_{init} = \mathcal{P}_{lexi}$. We must show that the optimal value of $C_{\mathcal{P}}$ is attained. Because $C_{\mathcal{P}}$ is a linear program, it suffices to show that it is feasible and bounded.

Since $R \subsetneq N$, the objective value $x$ is clearly bounded from above since, for any $i \in N \setminus R$,

$$x \leq p_i = \sum_{\substack{P \in \mathcal{P} \\ i \in P}} \lambda_P \leq \sum_{P \in \mathcal{P}} \lambda_P = 1.$$

It remains to show that $C_{\mathcal{P}}$ is feasible. Indeed, in the very first application of the framework, $\mathcal{P}_{init}$ is chosen to contain any arbitrary panel $P_o$. Since $R = \emptyset$, $C_{aux}(\emptyset, \rho, \{P_o\})$ is equal to $C_{egal}$ as defined in appendix B.9.1 and a DOF by proposition B.9.1.

In subsequent applications, $\mathcal{P}_{init}$ is chosen to be the portfolio $\mathcal{P}_{lexi}$ produced by the previous iteration. In this case, $R$ and $\rho$ were updated such that the final values $\vec{\lambda}^*$ and $\vec{p}^*$ of the previous application of the framework are a feasible solution to the optimization problem of the current application (setting $\lambda_P$ of all $P \notin \mathcal{P}_{init}$ to zero). □

**Lemma B.10.2.** *Whenever algorithm 3 applies the framework with an instance and the DOF $C_{aux}(R, \rho, \mathcal{P}_{lexi})$, the DOF implements the fairness measure $F$ given by*

$$F(\vec{p}) = \begin{cases} \min_{i \in N \setminus R} p_i & \text{if } \forall i \in R.\ p_i = \rho(i) \\ -\infty & \text{otherwise.} \end{cases}$$

*Proof.* By proposition B.8.3, the DOF implements the fairness measure $F'$ given by

$$F'(\vec{p}) = \sup\{x \mid x \in \mathbb{R}, \forall i \in N \setminus R.\ p_i \geq x, \forall i \in R.\ p_i = \rho(i)\}.$$

We will show that $F' = F$, by fixing some $\vec{p}$ and showing that $F'(\vec{p}) = F(\vec{p})$. If $\forall i \in R.\ p_i = \rho(i)$, then

$$F'(\vec{p}) = \sup\{x \mid x \in \mathbb{R}, \forall i \in N \setminus R.\ p_i \geq x\} = \min_{i \in N \setminus R} p_i.$$

Else, i.e., if $p_i \neq \rho(i)$ for some $i \in R$, then $F'(\vec{p}) = \sup \emptyset = -\infty$. □

**Theorem B.10.3.** *Algorithm 3 terminates.*

*Proof.* It is enough to show that the size of $R \subseteq N$ grows in each iteration of the while loop.

Recall that the KKT stationarity condition on $\vec{x}$ (B.12) states that

$$\nabla_{\vec{x}} h(\vec{p}, \vec{x}) = \sum_{r=1}^{m} \mu_r \nabla_{\vec{x}} g_r(\vec{p}, \vec{x}).$$

Note that $\frac{\partial}{\partial x}(x - p_i) = 1$, that $\frac{\partial}{\partial x}(p_i - \rho(i)) = \frac{\partial}{\partial x}(\rho(i) - p_i) = 0$, and that $\frac{\partial}{\partial x} h(\vec{p}, x) = \frac{\partial}{\partial x} x = 1$. Thus, the stationarity condition simplifies to

$$1 = \sum_{r \text{ constraint of shape } x - p_i \leq 0} \mu_r.$$

This shows that at least one of the optimal dual variables $\mu_r^*$ for a constraint $x \leq p_i$ must be positive, and that the size of $R$ increases in line 11. □

**Theorem B.10.4.** *For any instance, the output distribution of algorithm 3 on this instance is maximally fair according to the leximin criterion.*

*Proof.* We will prove the following invariant for the while loop in line 8 of algorithm 3: (1) for all agents $i \in R$, $\rho(i)$ is this agent's selection probability in the leximin-optimal probability allocation,[1] and (2) $\mathcal{D}$ is a distribution over $\mathcal{P}_{lexi}$ giving each $i \in R$ selection probability exactly $\rho(i)$.

Before proving the loop invariant, we show that it implies the correctness of the algorithm. Indeed, when the while loop exits, $R = N$, which means that $\rho$ specifies the whole leximin-probability allocation by part (1) of the invariant. By part (2) of the invariant, the distribution $\mathcal{D}$, which is the output distribution of the algorithm, implements the best possible probability allocation according to the leximin criterion and is therefore itself maximally fair.

It is easy to see that the loop invariant holds when we enter the loop for the first time since it is nearly vacuous for $R = \emptyset$. It remains to show that each iteration of the loop preserves the loop invariant.

It follows from the definition of the leximin criterion and part (1) of the invariant that the leximin-optimal probability allocation maximizes $x = \min_{i \in N \setminus R} p_i$ among all possible probability allocations guaranteeing $p_i = \rho(i)$ for all $i \in R$. By lemma B.10.2 and theorem B.8.5, the output distribution of algorithm 2 with the arguments as provided in line 9 also is a solution to this maximization problem. Fix $p_i^*$, $\mu_r^*$, $\mathcal{D}$, and $\mathcal{P}_{lexi}$ as in line 9, and call the optimal objective value $x^* = \min_{i \in N \setminus R} p_i^*$.

[1] The leximin-optimal probability allocation is uniquely determined as shown for example in Theorem 3.7 by Kurokawa et al. [182].

To re-establish part (1) of the invariant, we must look at the agents $i \in N \setminus R$ whose selection probability gets fixed to $p_i^*$ in line 11. Note that the dual variable $\mu_r^*$ is positive, and, as shown in the proof of theorem B.8.5, that this is also an optimal assignment for the dual variable in the problem $C_{\widehat{\mathcal{P}}}$ in $C_{aux}(R, \rho, \mathcal{P}_{lexi})$, ranging over all panels. By complementary slackness (B.8), the positivity of $\mu_r^*$ implies that the constraint $x \leq p_i$ is tight, meaning that $\rho(i)$ is set to $p_i = x^*$. While it follows from the application of our framework that *some* agent in $N \setminus R$ must have probability $x^*$ in the leximin-optimal probability allocation, it is not immediately clear that this must be the case for the specific agent $i$. However, $\mu_r^* > 0$ furthermore implies that the constraint $x \leq p_i$ is tight in *all* optimal solutions to $C_{\widehat{\mathcal{P}}}$ (see p. 95 of Schrijver [248]), and all the leximin-optimal distributions are such optimal solutions. This shows that agent $i$'s selection probability is fixed to the probability $x^*$ the agent receives in the leximin-optimal probability allocation, as claimed. Part (2) of the loop invariant follows from the fact that the distribution returned by the call to algorithm 2 satisfies all fixed probabilities and has support $\mathcal{P}_{lexi}$. □

### B.11 Description of Legacy

The Legacy algorithm proceeds in $k$ rounds, adding one pool member to the panel per round. Each round begins by calculating the *need* of each feature $f$ remaining in the pool, which is defined as

$$need_f := \frac{\ell_f - (\#\text{ panel members already selected with feature } f)}{\#\text{ remaining pool members with feature } f}.$$

Note that $need_f$ may be negative. After calculating $need_f$ for all features, the algorithm chooses a feature $f_{max}$ with maximal need and draws the next panel member uniformly from the remaining pool members with feature $f_{max}$. The selected panel member is then removed from the pool.

After adding this person to the panel, the panel might, for one or more features $f$, now contain $u_f$ many people with feature $f$. In this case, all remaining pool members with feature $f$ are removed from the pool. If this procedure produces a quota-compliant panel after the $k$th round, this panel is returned. Else, i.e., if the pool becomes empty in an earlier round or if the final panel violates some quotas, the algorithm is restarted from the beginning.

For intuition, note that the panel resulting from this procedure can violate quotas for several different reasons: it could happen that the $k$th person is selected but not all the lower quotas are satisfied yet, or the algorithm could run out of people of a certain type before fulfilling a lower quota if some of these agents were previously removed when an upper quota was reached.

The selection algorithms developed by other practitioner organizations generally follow the same structure of selecting panel members one by one, determining which agents to choose next based on myopic heuristics. We describe these algorithms in the following section.

## B.12 Description of Other Existing Algorithms

All existing algorithms we have heard about are listed below, and all select panel members one-by-one, backtracking or restarting if they encounter a quota violation. In most cases, a fully specified algorithmic description was not available, but we did obtain a high-level sketch of how each of these algorithms selects the next panel member. We list these algorithms by organization below, and describe their basic functionality:

**G1000:** G1000's algorithm works similarly to Legacy, except that it calculates the need of a feature as a difference rather than as a ratio.

**IFOK:** IFOK's algorithm is also generally similar to Legacy, but, rather than choosing only the next panel member from the feature with greatest need and then recalculating need, the entire lower quota of the feature with highest need is filled at once.

**Nexus:** The algorithm used by Nexus focuses less on features but rather selects uniformly from the pool, removing people from the pool once any of their features has reached its upper quota.

**MASS LBP:** MASS LBP typically uses tight lower and upper quotas on all their features. Their algorithm uses one bin for each feature category (e.g., gender, ethnicity, …), each initially filled with $k$ balls labeled with the correct distribution of features of this category (e.g., $k/2$ women and $k/2$ men). In every round, one ball is drawn from each bin. If a member of the pool has exactly this set of features, the pool member is chosen as the next panel member. Since this will often not be possible, MASS LBP employs elaborate (and not fully formalized) procedures of redrawing balls and backtracking on earlier picks. [201]

## B.13 Instances where Legacy is Unfair

In this section, we define a family of instances on which Legacy selects one individual much more rarely than the others, even though it would be possible to select all agents with equal probability. For illustration, we present one specific instance before defining the family:

Say that we want to select an assembly of $k = 200$ people that includes at least 99 of each category: women, men, liberals, and conservatives. Let the pool consist of 1,000 conservative men, 999 liberal women, and 1 conservative woman. Note that the algorithm that selects 100 uniformly drawn women and 100 uniformly drawn men satisfies the quotas and selects each pool member with equal probability 10%. By contrast, one can verify that the Legacy algorithm alternates between seeing liberals and men as the categories with highest need, skipping the conservative woman in each of the first 198 draws. Depending on how ties are broken for the last two panel selections (when all lower quotas are met), the conservative woman might even be chosen with probability 0, but with at most probability 0.2%.

Definition B.13.1 below generalizes this example to a wide range of agent numbers and panel sizes. In all these instances, it is possible to select all agents with equal probability $k/n$. At the

same time, depending on tie breaking, LEGACY might select the conservative woman with probability probability as low as zero (proposition B.13.4) or up to a selection probability in $O(1/n)$ (proposition B.13.5). Note that the ratio of this latter probability and the probability of equal selection $k/n$ can be made arbitrarily small by scaling up the size of the instance (corollary B.13.6).

**Definition B.13.1.** *Let $n$ and $k$ be even, positive integers, such that $n \geq 2k$. Define the instance $Alternate(n, k)$ as follows:*

- *Set the panel size to $k$.*
- *Let there be four features: female ($f$), male ($m$), liberal ($\ell$), and conservative ($c$). Let each feature have a lower quota of $k/2 - 1$ and an upper quota of $k$ (i.e., there are effectively no upper quotas).*
- *Let the pool consist of $n/2$ conservative men, $n/2 - 1$ liberal women, and one conservative woman.*

**Proposition B.13.2.** *For any instance $Alternate(n, k)$, it is possible to select each agent with equal probability $k/n$.*

*Proof.* Consider the selection algorithm that chooses $k/2$ women and $k/2$ men, each uniformly at random without replacement. It is easy to verify that this procedure will select each woman and each man with probability $\frac{k/2}{n/2} = k/n$. Moreover, this procedure will always select exactly $k/2$ women, exactly $k/2$ men, between $k/2$ and $k/2 + 1$ conservatives and between $k/2 - 1$ and $k/2$ liberals; which means that all panels produced by the procedure satisfy the quotas. □

**Lemma B.13.3.** *When LEGACY is called on $Alternate(n, k)$,*

- *all picks numbered $1, 3, 5, \ldots, k - 3$ are liberal women, and*
- *all picks numbered $2, 4, 6, \ldots, k - 2$ are conservative men.*

*Proof.* By strong induction on the number $i = 0, 1, \ldots, k - 3$ of panel members picked so far.

Suppose that $i$ is even. We will show that the next pick (the $i + 1$th) is a liberal woman. By the induction hypothesis, $s_\ell = i/2$ liberal women and $s_m = i/2$ conservative men have been selected so far. The need for each of the four features is

$$\begin{aligned} need_f &= (k/2 - 1 - s_\ell)/(n/2 - s_\ell) \\ need_m &= (k/2 - 1 - s_m)/(n/2 - s_m) \\ need_\ell &= (k/2 - 1 - s_\ell)/(n/2 - 1 - s_\ell) \\ need_c &= (k/2 - 1 - s_m)/(n/2 + 1 - s_m). \end{aligned}$$

Note that all the numerators are positive and equal, and that all the denominators are positive. Thus, the feature with highest need is the feature with lowest denominator, which is $\ell$. Thus, the algorithm selects a liberal, which can only be a woman.

Now, suppose that $i$ is odd. We will show that the next pick (the $i+1$th) is a conservative man. By the induction hypothesis, $s_\ell = \lceil i/2 \rceil$ liberal women and $s_m = \lfloor i/2 \rfloor$ conservative men have been selected so far. The need for each of the four features is

$$\begin{aligned}
need_f &= (k/2 - 1 - s_\ell)/(n/2 - s_\ell) \\
need_m &= (k/2 - 1 - s_m)/(n/2 - s_m) \\
need_\ell &= (k/2 - 1 - s_\ell)/(n/2 - 1 - s_\ell) \\
need_c &= (k/2 - 1 - s_m)/(n/2 + 1 - s_m).
\end{aligned}$$

It is easy to see that $need_m > need_c$ and that $need_\ell > need_f$. Furthermore,

$$\begin{aligned}
\frac{need_m}{need_\ell} &= \frac{(n/2 - 1 - s_\ell)/(n/2 - s_m)}{(k/2 - 1 - s_\ell)/(k/2 - 1 - s_m)} \\
&= \frac{(n/2 - 2 - s_m)/(n/2 - s_m)}{(k/2 - 2 - s_m)/(k/2 - 1 - s_m)} \\
&= \frac{1 - 2/(n/2 - s_m)}{1 - 1/(k/2 - 1 - s_m)} \\
&= \frac{1 - 2/(n/2 - s_m)}{1 - 2/(k - 2 - 2\,s_m)} \\
&\geq \frac{1 - 2/(k - s_m)}{1 - 2/(k - 2 - 2\,s_m)} && (k \leq n/2) \\
&> 1.
\end{aligned}$$

This shows that the feature with highest need is male ($m$), which implies that the next pick must be a conservative man. □

**Proposition B.13.4.** *If* Legacy *breaks ties between features with equal need in a worst-case way, the conservative woman in* $Alternate(n, k)$ *is selected with zero probability.*

*Proof.* By lemma B.13.3, the conservative woman is never among the first $k - 2$ picks. For the $k - 1$th pick, all features are exactly at their lower quota and therefore have a need of 0. The implementation breaks ties in the order in which the features are specified, so might break the tie in favor of liberals ($\ell$), which would mean that another liberal woman is selected. Then, in the last pick, the categories liberal and female have negative need because they exceed their lower quota, whereas the categories male and conservative still have a need of 0. If the tie is broken in favor of male, the last selection is a conservative man. Since all quotas are satisfied, the algorithm does not restart but returns this panel. Assuming the above tie-breaking decisions, the conservative woman will never be selected. □

**Proposition B.13.5.** *No matter how* Legacy *breaks ties between features with equal need, the conservative woman in* $Alternate(n, k)$ *is selected with probability at most* $8/n$.

*Proof.* Again, lemma B.13.3 shows that the conservative woman is never among the first $k-2$ picks. At the time of $k-1$th pick, there are $n/2-(k/2-1)$ women left in the pool and $n/2+1-(k/2-1)$ conservatives. At the time of the $k$th pick, these numbers are at still least $n/2-k/2$ and $n/2+1-k/2$. Since all quotas are already satisfied by the first $k-2$ picks, the algorithm does not restart. Thus, by a union bound over the last two picks, the selection probability of the conservative woman is at most

$$\frac{1}{n/2-(k/2-1)}+\frac{1}{n/2-k/2}\leq\frac{2}{n/2-k/2}\leq\frac{2}{n/2-n/4}=\frac{8}{n}.$$ □

**Corollary B.13.6.** *Even assuming best-case tie breaking between features with equal need, for every $\epsilon>0$, there is an instance where it is possible to select agents with equal probability $k/n$, but where* Legacy *selects some agent with probability at most $\epsilon\,k/n$.*

*Proof.* Let $k$ be an even integer larger than $8/\epsilon$, and let $n=2\,k$. By proposition B.13.2, it is possible to select each agent with equal probability $k/n$. By proposition B.13.5, the selection probability of the conservative woman is at most $8/n\leq\epsilon\,k/n$. □

## B.14 Comparing Legacy and Leximin on Intersectional Representation

While most of the paper is concerned with representation guarantees to *individuals*, in this section, we consider how the selection algorithms Legacy and Leximin impact the representation of *groups*. Note that both selection algorithms must satisfy quotas, and thus both algorithms will proportionally represent the groups delineated by the features. Therefore, we direct our focus to groups defined by the *intersection* of multiple features (e.g., "young woman", where "young" and "woman" are the features being intersected). Throughout this section, we study each group's *panel share*, which is the expected value of the fraction of the pool filled with that group's members (i.e., the sum of selection probabilities of all of its members divided by $k$). Ideally, to provide perfectly accurate descriptive representation, each intersectional group's panel share would be equal to its share in the population.

A priori, we would expect neither Leximin nor Legacy to accurately represent intersectional groups in proportion to their population share, since neither of these algorithms has precise information about the population shares of these groups, and they do not explicitly try to give these groups accurate representation. Instead, the panel share of an intersectional group will likely arise incidentally from the algorithms' efforts to ensure the satisfaction of quotas. The panel shares given by Leximin may additionally be impacted by its effort to equalize the selection probabilities between pool members, which could result in groups' panel shares being closer to their representation levels in the pool.

In this section, we investigate how accurately each algorithm represents intersectional groups in one real-world instance, *sf(e)*. We find that the algorithms give similar levels of intersectional representation overall, and in fact, the level of representation given to each *specific group* is similar across the two algorithms. We then find evidence suggesting an explanation for this similarity: for

both algorithms, it seems that the panel shares of intersectional groups mainly reflect the quotas, rather than the frequency of groups in the pool. We conclude by suggesting two ways in which our framework can be used for explicitly promoting the accurate representation of intersectional groups.

We perform this analysis on only a single dataset because the analysis requires knowledge of the population shares of all intersectional groups. Effectively, this requires a separate survey dataset, conducted on the exact population underlying the panel and including all features protected by the assembly's quotas. For the instance *sf(e)*, a nation-wide panel in the UK, we make use of the 2016 European Social Survey (ESS) [216].[1] We restrict our analysis to combinations defined by two features ("2-intersections") because, for intersections of three or more features, many intersectional groups are so small that we do not expect the ESS to represent their true population shares.

### B.14.1 Level of Intersectional Representation in Legacy versus Leximin

ED Figure 4 compares the deviation from proportional representation given to each individual 2-intersection by each respective algorithm. The histograms on the margins of the plot show that these deviations are concentrated around zero, indicating that both algorithms give fairly accurate representation to most intersectional groups. Nonetheless, a few 2-intersections are misrepresented by more than 15 percentage points, i.e., their true and proportional panel shares differ by more than 0.15. We compare the relative performance of Legacy and LexiMin using the *mean squared error*, i.e., the mean (calculated over all 2-intersections) of the squared difference between the population share and the panel share. Smaller mean squared errors indicate more accurate descriptive representation. We find that this error value is essentially the same for both algorithms, indicating that they achieve essentially the same level of representation for these intersectional groups: Legacy gives a mean squared error of $1.40 \cdot 10^{-3}$, and Leximin one of $1.36 \cdot 10^{-3}$.

### B.14.2 Explanation for Intersectional Representation in Legacy and Leximin

As the scatter plot in the center of ED Figure 4 shows, the points track closely with a line of slope equal to 1, indicating that not only do Leximin and Legacy achieve similar overall levels of intersectional representation, but that they over- and underrepresent the same groups by similar amounts. Indeed, the mean squared error between a group's panel share for Legacy and a group's panel share for Leximin is $1.99 \cdot 10^{-4}$, implying that the panel shares of a given group by the two algorithms are more closely related to each other than to the population share. This suggests that another property associated with the 2-intersections might determine the group's panel share more accurately than the population share, across both selection algorithms.

One property of intersectional groups that might influence their panel shares across both algorithms is their share in the pool. This is particularly relevant — and of potential concern — for

[1] The ESS data is preprocessed as described in Appendix D.2 of Flanigan et al. [127], and the population shares of intersectional groups computed from this data are included in our code repository.

Leximin , whose efforts to equalize individuals' selection probabilities might push it to overrepresent groups that are overrepresented in the pool. Our findings do not substantiate these concerns: as measured by the mean squared error, the panel share given by either algorithm is less closely related to the pool share (Legacy: $2.60 \cdot 10^{-3}$, Leximin : $2.37 \cdot 10^{-3}$) than to the population share, and, while this distance is smaller for Leximin than for Legacy, the difference is small.

In contrast to the pool share, we find that a group's panel share as naïvely extrapolated from the quotas *does* closely mirror the panel shares we observe resulting from either algorithm. We extrapolate from the quotas to predicted panel shares by defining the *quota share* (related to the ratio product defined in the methods section "Individuals Rarely Selected by Legacy") of the intersection of features $f_1$ and $f_2$ as

$$\frac{\ell_{f_1} + u_{f_1}}{2\,k} \cdot \frac{\ell_{f_2} + u_{f_2}}{2\,k}.$$

This quota share can be understood as a naïve estimation of the population share of the 2-intersection, assuming that features $f_1$ and $f_2$ are uncorrelated. We find that the mean squared error between the 2-intersections' panel shares and their quota shares (Legacy: $1.69 \cdot 10^{-4}$, Leximin : $1.76 \cdot 10^{-4}$) are substantially smaller than the error between panel and population shares, and on the same scale as the distance between the panel shares of both algorithms. These findings suggest that the descriptive representation of an intersectional group is more directly determined by the quotas of its constituent features rather than its share in the population or the pool. These results also suggest that the panel produced by both selection algorithms do not automatically replicate the correlation of features found in the population, but rather tends towards a composition in which features are closer to uncorrelated. If this phenomenon generalizes across citizens' assemblies, this would be an argument in favor of explicitly promoting intersectional representation, as we do in the following subsection.

### B.14.3 Achieving Proportional Representation for Intersections with Our Framework

In the above, we observed that neither selection algorithm happens to represent intersectional groups at a high level of accuracy. This suggests that, if the accurate representation of intersectional groups is an important consideration, one should attempt to incorporate this goal (and the data about population shares) explicitly into the algorithm. Below, we present two ways of using our framework to make the expected representation of intersectional groups closer to proportional:

First, one could enforce hard constraints on the representation of these intersectional groups by imposing lower and upper quotas on them, just as is traditionally done for single-feature groups. In fact, practitioners already do this on occasion for intersectional groups of particular interest. The downside of this approach is that it poorly scales to large numbers of intersections, because it is difficult to estimate how tight these quotas can be before quota-compliant panels cease to exist. Moreover, the number and tightness of these quotas trade off against the goal of equalizing selection probabilities in ways that can be difficult to predict.

A method that side-steps these downsides is to promote the proportional representation of intersectional groups as a soft constraint, by incorporating it into the fairness measure. Specifically, if one has a collection of groups $g$, each of which is associated with a set of pool members $N_g$ and a population share $q_g \in [0, 1]$, maximizing the concave expression

$$-\sum_{\text{groups } g}\left(q_g - \sum_{i \in N_g} p_i/k\right)^2$$

minimizes the mean square error between the panel shares given by the algorithm and the population shares. This term can either be turned into a distribution-optimizer family (definition B.8.1) that minimizes this error without consideration for individual selection probabilities, or it can be added to the objective function of another DOF, and the user can then optimize a linear combination of the chosen fairness measure and this mean squared error term. In defining this objective, the user can choose how strongly they want to prioritize intersectional representation over individual fairness by modifying the coefficients of the linear combination.

## B.15 Axiomatic Analysis

In searching for fair selection algorithms, we found the approach of optimizing quantitative measures of fairness more useful than the axiomatic method. The main reason for this is that a range of standard axioms of fair division are either trivially satisfied by all selection algorithms or impossible to satisfy by any selection algorithm, making them useless for delineating "good" algorithms. For example, no selection algorithm can guarantee *envy freeness* [206] on all instances, since the quotas of most instances preclude selecting every agent with equal probability $k/n$. *Pareto efficiency* [266], on the other hand, is trivially satisfied by all selection algorithms, since the sum of selection probabilities is always $k$. In appendices B.15.1 and B.15.2 below, we show that the relational axioms *population monotonicity* [266] and *committee monotonicity* [109] are also impossible to guarantee.

Two classical axioms that *are* meaningful in comparing selection algorithms are *equal treatment of equals* [206] and a form of *proportionality* [85]. In appendices B.15.3 and B.15.4, respectively, we show, via standard arguments, that Leximin satisfies both of these axioms.

### B.15.1 Population Monotonicity

**Definition B.15.1** (population monotonicity)**.** *A selection algorithm guarantees population monotonicity if, when additional agents are added to an instance, the selection probability of all previously existing agents weakly decreases.*

**Theorem B.15.2.** *No selection algorithm can guarantee population monotonicity.*

*Proof.* Fix a selection algorithm $A$, and consider an instance with six agents, $k = 3$, and four features. We indicate an agent's feature membership as a four-element Boolean vector, where

the $i$th entry of the vector indicates whether the agent exhibits feature $i$. Using this convention, let the agents' features be given as as agent 1: $(1, 0, 0, 0)$, agent 2: $(0, 1, 0, 0)$, agent 3: $(1, 1, 0, 0)$, agent 4: $(0, 0, 1, 0)$, agent 5: $(0, 0, 0, 1)$, and agent 6: $(0, 0, 1, 1)$. For each feature $f$, set the lower quota $\ell_f$ to 1 and the upper quota to 3 (i.e., there is effectively no upper quota). This instance has quota-compliant panels, for example the panel $\{\text{agent } 1, \text{agent } 2, \text{agent } 6\}$. Consider the probability allocation of $A$ on this instance. Since $k = 3$, agents 1, 2, 4, and 5 cannot all simultaneously have zero selection probability. W.l.o.g., assume that agent 1 has positive selection probability.

Now, consider a modified instance in which agent 6 is removed. In this instance, one verifies that the only quota-compliant panel is {agent 3, agent 4, agent 5}, which means that $A$ must select agent 1 with zero probability. This violates population monotonicity since adding back agent 6 would strictly increase the selection probability of agent 1. □

### B.15.2 Committee Monotonicity

**Definition B.15.3** (committee monotonicity)**.** *A selection algorithm guarantees committee monotonicity if, when an instance is modified by increasing $k$ (and remains an instance), the selection probability of all agents weakly increase.*

**Proposition B.15.4.** *No selection algorithm can guarantee committee monotonicity.*

*Proof.* Consider an instance with three agents and two features. Define the features of the agents using the vector notation from the proof of theorem B.15.2 as agent 1: $(1, 0)$, agent 2: $(0, 1)$, and agent 3: $(1, 1)$. If the lower and upper quotas for both features are set to 1, the only panel for $k = 1$ is $\{\text{agent } 3\}$, and the only panel for $k = 2$ is $\{\text{agent } 1, \text{agent } 2\}$. Thus, any selection algorithm must strictly decrease agent 3's selection probability when going from $k = 1$ to $k = 2$. □

### B.15.3 Equal Treatment of Equals

**Definition B.15.5** (equal treatment of equals)**.** *A selection algorithm guarantees equal treatment of equals if, for every instance and for every pair of agents $i_1, i_2$ that have exactly the same set of features, $i_1$ and $i_2$ are selected with equal probability.*

**Theorem B.15.6.** *Leximin guarantees equal treatment of equals.*

*Proof.* Fix an instance and two agents $i_1, i_2$ with equal features. Let $\mathcal{D}$ denote the output distribution of Leximin on this instance. For the sake of contradiction, assume that $i_1$ is selected with a probability $p_1$ strictly higher than the selection probability $p_2$ of $i_2$ in $\mathcal{D}$. We will show that there exists another distribution $\mathcal{D}'$ over panels whose probability allocation is leximin-fairer than the probability allocation of $\mathcal{D}$, which will contradict the optimality of Leximin .

Let $d$ denote the probability mass function of $\mathcal{D}$, mapping each possible panel of the instance to the probability with which it is returned in $\mathcal{D}$. Furthermore, define for each panel $P$ a second panel $swap(P)$, in which $i_1$ is exchanged for $i_2$ and vice versa:

$$swap(P) \coloneqq \begin{cases} P \setminus \{i_1\} \cup \{i_2\} & \text{if } i_1 \in P \text{ and } i_2 \notin P \\ P \setminus \{i_2\} \cup \{i_1\} & \text{if } i_2 \in P \text{ and } i_1 \notin P \\ P & \text{otherwise.} \end{cases}$$

Since $i_1$ and $i_2$ have exactly the same features, $swap(P)$ is also a quota-compliant panel.

Now, define $\mathcal{D}_{swap}$ by the probability mass function $d_{swap}$ with values

$$d_{swap}(P) \coloneqq d(swap(P)).$$

For each agent $i \notin \{i_1, i_2\}$, their selection probability is equal in $\mathcal{D}$ and $\mathcal{D}_{swap}$, because the agent is included in a panel $P$ iff they are included in $swap(P)$. Also, the selection probability of $i_1$ in $\mathcal{D}_{swap}$ is $p_2$ and that of $i_2$ is $p_1$.

Now define the symmetrization $\mathcal{D}'$ of $\mathcal{D}$ over $i_1$ and $i_2$ as the mixture of distributions $\frac{1}{2}\mathcal{D} + \frac{1}{2}\mathcal{D}_{swap}$. In this distribution, each agent $i \notin \{i_1, i_2\}$ is selected with the same probability as in $\mathcal{D}$, but $i_1$ and $i_2$ are both selected with probability $(p_1 + p_2)/2$. This probability allocation is leximin-fairer than that of $\mathcal{D}$, contradiction. □

### B.15.4 Proportionality

If a selection algorithm satisfies proportionality, each agent $i$ should, on every instance, receive at least a $1/n$ fraction of the selection probability they would receive under their most preferred probability allocation for this instance (i.e., the probability allocation chosen if $i$ was a dictator [85]). Note that, if $i$ is contained in some panel $P$, the panel distribution that deterministically outputs $P$ gives rise to a probability allocation in which $i$ is chosen with probability 1. Thus, proportionality requires that $i$ is selected with probability at least $1/n$. Else, if $i$ is not contained in any panel, no probability allocation gives them positive selection probability, and proportionality does not guarantee them any minimum selection probability. Consequently, proportionality in the panel-selection setting can be defined as follows:

**Definition B.15.7** (proportionality)**.** *A selection algorithm guarantees proportionality if, on all instances, each agent $i$ has a selection probability of at least $1/n$ unless they are not contained in any possible panel.*

**Theorem B.15.8.** *Leximin guarantees proportionality.*

*Proof.* Fix an arbitrary instance. Partition the agents $N$ into two sets: the agents $N^+$ that are contained in at least one panel and the agents $N^-$ that are not contained in any panel. Since at least one panel must exist, $N^+ \neq \emptyset$.

First, consider the leximin-optimal probability allocation $\vec{p}_{lex}$. Assume for the sake of contradiction that Leximin violates proportionality on this instance, i.e., that some agent in $N^+$ is selected with probability $p < 1/n$.

Under this assumption, we will construct another panel distribution with a probability allocation $\vec{p}_{alt}$ that is strictly leximin-fairer than $\vec{p}_{lex}$, which will contradict the optimality of $\vec{p}_{lex}$. For each $i \in N^+$, let $P_i$ be a panel such that $i \in P_i$. Then, consider the distribution over panels resulting from choosing an agent $i \in N^+$ uniformly at random and returning $P_i$. Call the corresponding probability allocation $\vec{p}_{alt}$. Note that each $i \in N^+$ will be contained in the panel selected in this way with probability at least $1/|N^+| \geq 1/n$.

Clearly, each agent in $N^-$ must receive selection probability 0 in both $\vec{p}_{lex}$ and $\vec{p}_{alt}$. Since the next-lower selection probability of $\vec{p}_{alt}$ is at least $1/n$, and since the next-lower selection probability of $\vec{p}_{lex}$ is $p < 1/n$, $\vec{p}_{alt}$ would be leximin-fairer than $\vec{p}_{lex}$, contradiction. □

# C

# Chapter 4 Appendix

## C.1 Panel Selection Datasets

We examine data from the following 11 real-world sortition panel selection instances, generously provided to us by several groups that specialize in organizing citizens' assemblies. Appendix D.4.1 shows the instance short-names we use throughout the paper, and which organization was responsible for each panel. The final two columns compare the values of our theoretical upper bounds on the marginal discrepancy, illustrating that in all instances except "obf", the bound from Section 4.3.2 is tighter. Finally, we give some metadata about each instance, which is required for calculating the values of our theoretical upper bounds.

In particular, $n$ = number of pool members, $k$ = number panel members, $C$ = set of distinct realized feature-vectors in the pool. Precise constants used for computing exact the upper bounds are derived in appendix C.2: the Section 4.3.1 bound is exactly $k/m$, the Section 4.3.2 bound is exactly

$$\frac{\sqrt{\frac{1}{2}(1+\frac{\ln 2}{\ln|C|})}\cdot\sqrt{|C|\ln(|C|)}+1}{m},$$

and the Theorem C.2.8 bound is exactly $\frac{2k/n_{min}+1}{m}$. In all instances, $n_{min}=1$.

Table C.1: Instance parameters and resulting theoretical bounds

| Instance | Organization | $n$ | $k$ | $\lvert C\rvert$ | Thm 4.3.1 | Thm 4.3.2 | Thm C.2.8 |
|---|---|---|---|---|---|---|---|
| sf(a) | Sortition Foundation | 312 | 35 | 182 | $35/m$ | $24.2/m$ | $71/m$ |
| sf(b) | Sortition Foundation | 250 | 20 | 92 | $20/m$ | $16.5/m$ | $41/m$ |
| sf(c) | Sortition Foundation | 161 | 44 | 92 | $44/m$ | $16.5/m$ | $89/m$ |
| sf(d) | Sortition Foundation | 404 | 40 | 108 | $40/m$ | $18.0/m$ | $81/m$ |
| sf(e) | Sortition Foundation | 1727 | 110 | 762 | $110/m$ | $53.8/m$ | $221/m$ |
| cca | Center for Climate Assemblies | 825 | 75 | 554 | $75/m$ | $45.1/m$ | $151/m$ |
| hd | Healthy Democracy | 239 | 30 | 202 | $30/m$ | $25.6/m$ | $61/m$ |
| mass | MASS LBP | 70 | 24 | 25 | $24/m$ | $8.0/m$ | $49/m$ |
| nexus | Nexus | 342 | 170 | 242 | $170/m$ | $28.4/m$ | $341/m$ |
| obf | Of By For | 321 | 30 | 294 | $30/m$ | $31.6/m$ | $61/m$ |
| ndem | New Democracy | 398 | 40 | 173 | $40/m$ | $23.5/m$ | $81/m$ |

## C.2 Omitted Proofs and Additional Beyond-Worst-Case Upper Bounds from Section 4.3

### C.2.1 General Rounding Procedure

Throughout this section, we repeatedly face the task of rounding the entries of some distribution $p$ to some vector $\bar{p}$ that must also be a valid distribution (i.e., have entries in $[0,1]$ such that $\|\bar{p}\|_1 = 1$), and have entries that are integer multiples of $1/m$. However, many of the standard rounding procedures we apply, such as randomized rounding and discrepancy-based dependent rounding, only give guarantees for rounding probabilities to 0/1 vectors, rather than to multiples

of $1/m$. Thus, in several proofs (Section 4.3.1, Section 4.3.1, Section 4.3.2, Theorem C.2.8), we apply these canonical rounding methods to a *modified version* of our original vector $p$, called $x'$. After constructing $x'$, we round it to a 0/1 vector $\bar{x}'$, from which we finally compute $\bar{p}$. We more precisely define this general rounding procedure, and characterize some of its useful properties, below.

**Definition C.2.1** (Procedure for using 0/1 rounding procedure to round $p$ to $\bar{p}$)**.** *Let $p$ be a distribution, represented as a vector. Let $x$ be the vector $p$ with entries scaled by $m$, so that $x_j := m \cdot p_j$. Then, define the vector $\lfloor x \rfloor$, which we can think of as the "integer components" of each entry of $x$, i.e., $\lfloor x \rfloor_j := \lfloor m \cdot p_j \rfloor$. Finally, we define $x'$ as the "decimal components" of the entries of $x$, so that $x' := x - \lfloor x \rfloor$. We will round $x'$ to a 0/1 vector.*

*Then, construct $\bar{p}$ from $p$ as follows:*
*1. Construct the vector $x'$ as above.*
*2. Round $x'$ to some 0/1 vector $\bar{x}'$ via a given rounding procedure such that $\|\bar{x}\|_1 = \|x'\|_1$.*
*3. Set $\bar{p}$ such that*

$$\bar{p} := \frac{\lfloor x \rfloor + \bar{x}'}{m}.$$

At a high level, this rounding procedure can be thought of as scaling up the vector we want to round by $m$, holding this scaled vector's integer components aside and rounding its decimal components, and then adding the integer components back in and scaling back down by $m$.

Now, we show that this rounding procedure produces a $\bar{p}$ with the properties we want—(a) it has entries that are multiples of $1/m$ and (b) it is a valid distribution—as well as an additional property (c), which helps translate guarantees on existing roundinig schemes to guarantees in our setting.

**Lemma C.2.2.** *Suppose we are given a 0/1 rounding scheme which, given $x' \in [0,1]^{|\mathsf{K}|}$ and constraint matrix $M$, produces some $\bar{x}'$ which satisfies*

- $\bar{x}' \in \{0,1\}^{|\mathsf{K}|}$,
- $\|\bar{x}'\|_1 = \|x'\|_1$, *and*
- $|(M(x' - \bar{x}'))_i| \leq g(i)$ *for each row $i$.*

*Then given some distribution $p \in \mathsf{R}_+^{|\mathsf{K}|}$ and $m \in \mathbb{N}$, the procedure in Definition C.2.1, using such a 0/1 rounding scheme, produces $\bar{p}$ such that*

*(a)* $\bar{p} \in (\mathbb{Z}_+/m)^{|\mathsf{K}|}$,

*(b)* $\bar{p}$ *is a distribution, and*

*(c)* $|(M(p - \bar{p}))_i| \leq \frac{g(i)}{m}$ *for each row $i$.*

*Proof.* We prove each property separately:

(a) holds: $\bar{p}$ contains multiples of $1/m$, since in the general procedure (Definition C.2.1), its entries are set to the sum of two integers divided by $m$.

(b) holds: $\bar{p}$ is a valid distribution: all entries of $\bar{p}$ must be non-negative, and we have that $\|\bar{p}\|_1 = \|p\|_1 = 1$, as shown below.

$$\|\bar{p}\|_1 = \left\|\frac{\lfloor x \rfloor + \bar{x}'}{m}\right\|_1 = \left\|\frac{\lfloor x \rfloor}{m}\right\|_1 + \left\|\frac{\bar{x}'}{m}\right\|_1 = \left\|\frac{\lfloor x \rfloor}{m}\right\|_1 + \left\|\frac{x'}{m}\right\|_1 = \|p\|_1$$

(c) holds: Fix some $i$ and the corresponding row of $(M(p - \bar{p}))$, referred to as $(M(p - \bar{p}))_i$. Then,

$$|(M(p - \bar{p}))_i| = \left|\left(M\left(\frac{\lfloor x \rfloor + x'}{m} - \frac{\lfloor x \rfloor + \bar{x}'}{m}\right)\right)_i\right| = \frac{|(M(x' - \bar{x}'))_i|}{m} \leq \frac{g(i)}{m}$$

□

### C.2.2 Omitted Proofs

We will make repeated use of the following generalization of Hoeffding's inequality (see e.g. Proposition 5 of [101]):

**Lemma C.2.3.** *If $\{\xi_j\}$ are negatively associated random variables with $\xi_j \in [a_j, b_j]$ and $\xi = \sum_j \xi_j$, then*

$$\mathbf{Pr}\left[|\mathsf{E}[\xi] - \xi| \geq t\right] \leq 2\exp\left\{-\frac{2t^2}{\sum_j (b_j - a_j)^2}\right\}.$$

Here is our first use:

For any realizable $\pi$, we may efficiently randomly generate $\bar{p}$ such that its marginals $\bar{\pi}$ satisfy

$$\|\pi - \bar{\pi}\|_\infty = O\left(\frac{\sqrt{n \log n}}{m}\right).$$

*Proof of Section 4.3.1.* Given a vector of marginals $\pi$, let $p$ be a basic solution to $Mp = \pi$, where $M$ is the individual-feasible panel membership matrix, so that $|supp(p)| \leq n$.

Then, we will construct $\bar{p}$ from $p$ by constructing $x'$, rounding it to $\bar{x}' \in \{0,1\}^{|\mathcal{K}|}$, and then reconstructing $\bar{p}$ as described in Definition C.2.1. To do this 0/1 rounding, here we use any randomized rounding procedure that satisfies the following properties: preservation of adding up constraint $\|\bar{x}'\|_1 = \|x'\|_1$, preservation of marginals $E[\bar{x}'_j] = x'_j$, and that $\bar{x}'_j$ are *negatively associated*, as defined in [56, 101]. These properties are satisfied via any number of randomized rounding algorithms [56]. Note as in Definition C.2.1, $\|\bar{x}'\|_1 = \|x'\|_1$ implies that $\bar{p} \in \mathcal{D}$.

Now it remains to analyze the marginal $\bar{\pi}_i$ provided to any given individual $i$ by $\bar{p}$. Consider the collection of $\bar{x}'_j$ for which $i$ is contained in panel $j$. Then, using the negative association of these $\bar{x}'_j$s, we have that for any $t \geq 0$,

$$\mathbf{Pr}[|Mx' - M\bar{x}'| \geq t] = \mathbf{Pr}\left[\left|\mathsf{E}\left[\sum_{j \ni i} \bar{x}'_j\right] - \sum_{j \ni i} \bar{x}'_j\right| \geq t\right], \tag{C.1}$$

by the definition of $\bar{x}'_j$. Then by Hoeffding (Lemma C.2.3),

$$\leq 2\exp\left(\frac{-2t^2}{|\{j : i \in j\}|}\right) \tag{C.2}$$

$$\leq 2\exp\left(\frac{-2t^2}{n}\right), \tag{C.3}$$

where here we use that $|supp(p)| \leq n$. Then taking $t = \sqrt{\frac{1+\epsilon}{2} n \log n}$,

$$\leq \frac{2}{n^{1+\epsilon}}. \tag{C.4}$$

Taking a union bound over all $n$ rows $i$ then gives

$$\mathbf{Pr}\left[\|Mx' - M\bar{x}'\|_\infty \geq \sqrt{\frac{(1+\epsilon)}{2}} \cdot \sqrt{n \log n}\right] \leq \frac{2}{n^\epsilon} < 1.$$

By lemma C.2.2, we therefore have

$$\mathbf{Pr}\left[\|\pi - \bar{\pi}\|_\infty \leq \sqrt{\frac{1+\epsilon}{2}} \cdot \frac{\sqrt{n \log n}}{m}\right] \geq 1 - \frac{2}{n^\epsilon} > 0. \qquad \square$$

Note: if we are additionally guaranteed that all of the $\pi_i = \Omega(k/n)$, then a multiplicative form of Chernoff yields

$$\|\pi - \bar{\pi}\|_\infty = O\left(\sqrt{\frac{k \log n}{mn}}\right)$$

with constant probability.

For any realizable $\pi$, we may efficiently construct $\bar{p}$ such that its marginals $\bar{\pi}$ satisfy

$$\|\pi - \bar{\pi}\|_\infty \leq k/m.$$

*Proof of Section 4.3.1.* Here, we apply the rounding algorithm used by Flanigan et al. [128] (Lemma 9, Appendix B.4.1), which builds on a notable theorem by Beck and Fiala [40]. Since this rounding algorithm does 0/1 rounding, we apply their algorithm to round $x'$, as in Definition C.2.1, to some 0/1 vector $\bar{x}'$, from which we construct $\bar{p}$. By Lemma 9 in Appendix B.4.1 in [128], this algorithm ensures the preservation of the "adding up" constraint, that is, that $\|\bar{x}'\|_1 = \|x'\|_1$. Thus, by results (a) and (b) of Lemma C.2.2, $\bar{p} \in \bar{D}$.

Now, it remains to show that $\|\pi - \bar{\pi}\|_\infty = \|M(p - \bar{p})\|_\infty \leq k/m$. Fortunately, as they prove, the rounding procedure of Flanigan et al. [128] guarantees that when rounding $x'$ to $\bar{x}'$, for a constraint matrix $M$ with column sparsity $k$, $\|M(x' - \bar{x}')\|_\infty \leq k$. By Lemma C.2.2 result (c), this immediately implies that $\|\pi - \bar{\pi}\|_\infty \leq k/m$. $\square$

If $\pi$ is anonymous and realizable, then we may efficiently construct $\bar{p}$ such that its marginals $\bar{\pi}$ satisfy

$$\|\pi - \bar{\pi}\|_\infty = O\left(\frac{\sqrt{|C|\log|C|}}{m}\right).$$

*Proof of Section 4.3.2.* We begin with anonymous marginals $\pi$ witnessed by some distribution $p$ over $\mathsf{K}$. The first order of business is to project $p$ into "type space," in order to derive a distribution over panel types. Overloading $F$, we let $F(P) = \mathfrak{P}$ denote the panel type of a given panel $P$, defined as the multiset $F(P) = \{F(i) : i \in P\}$. Then we define the distribution over panel types induced by $p$ as $\mathfrak{p}$, where the probability of drawing panel type $\mathfrak{P}$ from $\mathfrak{p}$ is defined as $\mathfrak{p}_{\mathfrak{P}} := \sum_{P \in \mathsf{K}: F(P) = \mathfrak{P}} p_P$.

This $\mathfrak{p}$ satisfies the Panel Type LP in eq. (4.3). As an aside, note that this $\mathfrak{p}$ has support $supp(\mathfrak{p}) = \{F(P) : P \in supp(p)\}$. We will assume without loss of generality that $\mathfrak{p}$ is a basic solution to (4.3), so that it has at most $|\mathsf{C}|$ nonzero entries, where $\mathsf{C}$ is the set of all feature-vectors appearing in the pool, i.e., $supp(p) \leq |C|$. Since $|supp(p)| \leq n$ without loss of generality, $|supp(\mathfrak{p})| \leq n$ also, and so this basic $\mathfrak{p}$ may be found efficiently.

Given this distribution $\mathfrak{p}$ over panel types, we will round it to a uniform lottery $\bar{\mathfrak{p}}$ of size $m$ over panel types $\mathfrak{R}$. Finally, we will lift this distribution over panel types $\bar{\mathfrak{p}}$ back to a distribution $\bar{p}$ over panels with the desired guarantee, and argue that this lift can be performed when the original marginals $\pi$ are anonymous.

We generate $\bar{\mathfrak{p}}$, a distribution with all probabilities multiples of $1/m$, from $\mathfrak{p}$ via randomized rounding, as in section 4.3.1. To produce $\bar{\mathfrak{p}}$ via a 0/1 rounding algorithm, we follow the procedure given in Definition C.2.1, where here, $\mathfrak{p}, \bar{\mathfrak{p}}$ correspond to the $p, \bar{p}$ given in the definition. Via this definition, we construct $x, \lfloor x \rfloor, x', \bar{x}'$ analogously, so that $x = m\mathfrak{p}$, etc. By choosing a randomized rounding procedure that preserves $\|\bar{x}'\|_1 = \|x'\|_1$, by Lemma C.2.2 we have that $\bar{\mathfrak{p}}$ is a valid distribution containing multiples of $1/m$. We again assume this rounding procedure samples $\bar{x}'_j$ which are negatively associated, and preserves that $\mathsf{E}[\bar{x}'_j] = x'_j$ for all panel types $j$.

Recall that type marginals $\tau_c, \bar{\tau}_c$ represent the expected number of panel spots allocated to each feature vector $c$ by $\mathfrak{p}, \bar{\mathfrak{p}}$, respectively, and are given by $\tau = Q\mathfrak{p}$ and $\bar{\tau} = Q\bar{\mathfrak{p}}$. (Recall that $Q$, as described in Section 4.3, encodes the number of copies of each feature vector on each panel type.) We will next analyze the proximity of the rounded type marginals $\bar{\tau}_c$ to the original type marginals $\tau_c$.

Proceeding via an analysis similar to that of Section 4.3.1, we consider the collection of random variables $\bar{x}'_j$ for which feature vector $c$ appears on panel type $j$ (i.e., $Q_{cj} > 0$). We note that these $\bar{x}'_j$ are again negatively associated, and thus all $Q_{cj}\bar{x}'_j$ are negatively associated, since for a fixed instance all $Q_{cj}$ are constant.

Then for any $t \geq 0$,

$$\mathbf{Pr}[|(Qx' - Q\bar{x}')_c| \geq t] = \mathbf{Pr}\left[\left|\mathsf{E}\left[\sum_j Q_{cj}\bar{x}'_j\right] - \sum_j Q_{cj}\bar{x}'_j\right| \geq t\right], \tag{C.5}$$

by the definition of $x_j$ and $\tilde{x}_j$. Then by Hoeffding (Lemma C.2.3) with $\xi_j = Q_{cj}\tilde{x}_j$,

$$\leq 2\exp\left(\frac{-2t^2}{\sum_j Q_{cj}^2}\right) \tag{C.6}$$

$$\leq 2\exp\left(\frac{-2t^2}{|\mathsf{C}|m_c^2}\right), \tag{C.7}$$

where $m_c := \max_j Q_{cj}$, and (C.7) uses that for all $c$, $\sum_j Q_{cj}^2 \leq \sum_j m_c^2 \leq |supp(\mathfrak{p})|m_c^2 \leq |\mathsf{C}|m_c^2$. Thus, taking $t_c = \alpha \cdot m_c \cdot \sqrt{|\mathsf{C}|\log|\mathsf{C}|}$,

$$\leq \frac{2}{|\mathsf{C}|^{2\alpha^2}}. \tag{C.8}$$

Taking $\alpha > \sqrt{\frac{1}{2}(1+\frac{\log 2}{\log|\mathsf{C}|})}$ and union bounding over all $|C|$ feature vectors, we may therefore guarantee that with positive probability,

$$|(Qx' - Q\bar{x}')_c| \leq \alpha \cdot m_c\sqrt{|\mathsf{C}|\log|\mathsf{C}|}$$

for all $c$ simultaneously. By lemma C.2.2, the derived $\bar{\mathfrak{p}}$ and $\bar{\tau}$ and therefore satisfy

$$|\tau_c - \bar{\tau}_c| \leq \alpha \cdot m_c\frac{\sqrt{|\mathsf{C}|\log|\mathsf{C}|}}{m} \tag{C.9}$$

for all $c$ simultaneously.

Given such a $\bar{\mathfrak{p}}, \bar{\tau}$ over panel types, it remains to construct some uniform lottery $\bar{p}, \bar{\pi}$ over the panels in $\mathsf{K}$ which is consistent with $\bar{\tau}$ and satisfies the desired guarantees on $\bar{\pi}$, which are:

1. each individual appears on each panel in $\bar{p}$ at most once,[1]
2. $0 \leq \bar{\pi}_i \leq 1$ for all $i$, and
3. $|\pi_i - \bar{\pi}_i|$ is small for all $i$.

We will describe a procedure for forming $\bar{p}$ and $supp(\bar{p})$ from $\bar{\mathfrak{p}}$, and then argue that it satisfies all three of these criteria, as well as implies a valid distribution $\bar{p}$ for which all probabilities are multiples of $1/m$. At a high level, this algorithm starts with the panel types $\mathfrak{P}_j$ which form the support of $\mathfrak{p}$, and for each $c$ in turn allocates spots in these panel types $\mathfrak{P}_j$ with feature vector $c$ to individuals in $N_c := \{i \in [n] : F(i) = c\}$, the $n_c$ individuals with feature vector $c$. Given the type marginals $\bar{\tau} = Q\bar{\mathfrak{p}}$ output by our rounding procedure, it first calculates the "ideal" number of spots $\bar{s}_i$ to allocate to each individual $i \in N_c$ across all of $\bar{p}$. It then performs the allocation in such a way that the guarantees above are satisfied. Since $\bar{\mathfrak{p}} \in (\mathbb{Z}_+/m)^{|\hat{\mathfrak{K}}|}$ and this algorithm populates each $\mathfrak{P}_j$ in the support to create some $P_j \in \mathsf{K}$, it follows that the $\bar{p}$ which it ultimately produces is $\bar{p} \in (\mathbb{Z}_+/m)^{|\mathsf{K}|}$ also.

[1]We note that this is a concern because we will not simply be choosing known panels from collection $\mathcal{K}$, as we don't see the entire collection *a priori*; we will instead be *constructing* panels that must turn out to be feasible.

**Algorithm 4** PanelPacker

**Input:** $\bar{\mathfrak{p}} \in (\mathbb{Z}_+/m)^{|\mathfrak{K}|}$ a distribution over feasible panel types, $N$
**Output:** $\bar{p} \in (\mathbb{Z}_+/m)^{|\mathfrak{K}|}$ a distribution over feasible panels

**for** $j \in [m]$ **do**
  Initialize $P_j \leftarrow \emptyset$ for each $\mathfrak{P}_j \in \text{supp}(\bar{\mathfrak{p}})$
**for** $c \in C$ **do**
  Initialize spots $\bar{s}_i \in \{\lfloor m \cdot \bar{\tau}_c/n_c \rfloor, \lceil m \cdot \bar{\tau}_c/n_c \rceil\}$ for $i \in N_c$ such that $\sum_{i \in N_c} \bar{s}_i = m \cdot \bar{\tau}_c$
  Initialize $d_i^1 \leftarrow \bar{s}_i$ for $i \in N_c$
  **for** $j \in [m]$ **do**
    Let $I_{cj}$ be the first $Q_{cj}$ many $i \in N_c$ with largest $d_i^j$
    Update $P_j \leftarrow P_j \cup I_{jc}$
    Update $d_i^{j+1} \leftarrow d_i^j - \mathbb{1}\{i \in I_{cj}\}$ for all $i \in N_c$
**return** $\bar{p}$ *the uniform distribution over* $P_j$

For each panel type $\mathfrak{P}_j$ in the support of $\bar{\mathfrak{p}}$, algorithm 4 forms one panel in the support of $\bar{p}$ by, for each $c \in \mathsf{C}$, allocating each of panel type $\mathfrak{P}_j$'s $Q_{cj}$ "spots" to individuals $i \in N_c$. It populates each panel type $\mathfrak{P}_j$ with individuals for each $c$ independently. If algorithm 4 succeeds at step (20) for all $c \in \mathsf{C}$, then it produces a panel $P_j \in supp(\bar{p})$. We first argue that algorithm 4 succeeds in producing feasible panels.

*Proof that algorithm 4 succeeds.* In particular, we will argue that algorithm 4 succeeds for every iteration of step (20). Since $\sum_{i \in N_c} \bar{s}_i = \sum_{\mathfrak{P}_j \in \bar{\mathfrak{p}}} Q_{cj}$, this is equivalent to showing that it assigns all individuals $i \in N_c$ such that $d_i^{m+1} = 0$ for all $i$ and no individual appears on any panel more than once.

In each round we have

$$d_i^j := m \cdot \bar{\pi}_i - \sum_{j' < j} \mathbb{1}\{i \in P_{j'}\}$$

the number of spots in $\bar{\mathfrak{p}}$ of type $c$ on which $i$ still needs to be placed at the beginning of round $j$ in order to reach their allocation of $\bar{s}_i$ spots. (This $d_i^j$ can be viewed as the "unsatisfied demand" of individual $i$ at round $j$, according to the promised number of spots $m\bar{\pi}_i$.)

Because the $\bar{\pi}_i$ are all either $\frac{\lfloor m \cdot \bar{\tau}_c/n_c \rfloor}{m}$ or $\frac{\lceil m \cdot \bar{\tau}_c/n_c \rceil}{m}$, the initial values of $d_i^0$ for $i \in N_c$ are all within 1 of one another. Note that step (20) preserves this property that $d_i^j$ remain within 1 of one another for all rounds, since at each step $j$ it decreases some collection of maximal $d_i^j$ by 1.

Suppose for the sake of contradiction that for some $c$, algorithm 4 reaches some first step $j$ for which a $c$ position on panel $P_j$ cannot be allocated to any $i \in N_c$; then there are not enough individuals with remaining "unmet demand", so $Q_{cj} > |\{i : d_i^j > 0\}|$. Since $Q_{cj} \leq m_c \leq n_c$, it must be the case that some $i \in N_c$ have already been fully assigned by this step $j$ (meaning that for these $i$ it is the case that $d_i^j = 0$), and so all $d_i^j \in \{0, 1\}$ because the $d_i^j$ are within 1 of one

another. But $\sum_j Q_{cj} = \sum_i d_i^0 = m \cdot \bar{\tau}_c$, while at this point

$$\sum_{j' \geq j} Q_{cj'} \geq Q_{cj} > |\{i : d_i^j > 0\}| = \sum_i d_i^j,$$

meaning that the number of unallocated positions of type $c$ remaining at step $j$ exceeds the remaining unmet demand of the $i \in N_c$. This implies that strictly more than $Q_{cj'}$ individuals $i$ were given spots on panel $j'$ at step (20) for some earlier $j' < j$. But this is impossible by the definition of algorithm 4. Therefore algorithm 4 must succeed in feasibly assigning individuals of each type $c$ to panels.

Since algorithm 4 succeeds on step (20), it successfully puts $Q_{cj}$ individuals in $N_c$ onto panel $P_j$ for each $j$ and each $c$. By the feasibility of $\mathfrak{P}_j$ we therefore have that $|P_j| = k$ and $P_j$ is quota feasible, since $\mathfrak{P}_j$ is quota feasible and $P_j$ has the exact same numbers of individuals with each feature vector as $\mathfrak{P}_j$.

Therefore algorithm 4 terminates with a collection of quota-feasible panels, with no individual appearing on any panel more than once. □

We conclude by arguing that the output of algorithm 4 satisfies the desired guarantees.

First, it is clear that each individual $i$ appears on each panel $P_j \in supp(p)$ at most once. This is because for each individual $i \in N_c$ for some $c$, $i$ is assigned a position on $P_j$ if and only if $i \in I_{cj}$ at step (20), and $I_{cj}$ contains each $i$ at most once by definition. Therefore condition (1) is satisfied.

We next show that these output $\bar{\pi}_i$ satisfy condition (2). For each $i$, its value of $\bar{\pi}_i$ in the distribution $\bar{p}$ output by algorithm 4 is precisely $\bar{s}_i/m$.

Therefore clearly $\bar{\pi}_i \geq 0$, and since condition (1) holds we have $\sum_j \mathbb{1}\{i \in P_j\} \leq m$, and so $\bar{\pi}_i \leq 1$ also. For a more explicit proof that $\bar{\pi}_i \leq 1$, observe that since $\mathfrak{p}$ is a distribution,

$$\bar{\tau}_c = \sum_j \bar{\mathfrak{p}}_j Q_{cj} \leq \max_j Q_{cj} = m_c \leq n_c,$$

where the last inequality follows because all $\mathfrak{P}$ are feasible panel types, so they cannot contain more individuals $i \in N_c$ than exist in the pool. By algorithm 4 we have $\bar{s}_i \in \{\lfloor m \cdot \bar{\tau}_c/n_c \rfloor, \lceil m \cdot \bar{\tau}_c/n_c \rceil\}$. Dividing by $n_c$ and multiplying by $m$ yields $\bar{s}_i \leq m$, and so $\bar{\pi}_i = \bar{s}_i/m \leq 1$. Thus (2) is satisfied.

Finally, we confirm condition (3), that the individual marginals are close. By the anonymity of $\pi$, for all $i$ with $F(i) = c$ we have $\pi_i = \tau_c/n_c$, and by its choice of $\bar{s}_i$ and the fact that it succeeds, algorithm 4 guarantees that $\bar{\pi}_i = \bar{s}_i/m \in (\bar{\tau}_c/n_c - 1/m, \bar{\tau}_c/n_c + 1/m)$. Since $m_c \leq n_c$, therefore (C.9) implies

$$|\pi_i - \bar{\pi}_i| \leq \frac{m_c}{n_c} \cdot \alpha \cdot \frac{\sqrt{|\mathsf{C}| \log |\mathsf{C}|}}{m} + \frac{1}{m} = O\left(\frac{\sqrt{|\mathsf{C}| \log |\mathsf{C}|}}{m}\right),$$

for all $i$, satisfying condition (3) and showing the claim. □

There exist $p, \pi$ for which for all uniform lotteries $\bar{p}, \bar{\pi}$,

$$\min_{\bar{p} \in \overline{\mathcal{D}}} \|\pi - \bar{\pi}\|_\infty = \Omega\left(\frac{\sqrt{k}}{m}\right).$$

We will make use of the following lemma:

**Lemma C.2.4.** *Any $k$-uniform hypergraph on $[n]$ is realizable via quotas as the set of feasible panels for an instance of the panel selection problem with pool $[n]$.*

When individual membership in feasible panels is represented as $M \in \{0,1\}^{n \times |\mathsf{K}|}$, this lemma claims that any $M$ with uniform column norms is *realizable* by an instance of the panel selection problem, meaning that there exists an instance of the panel selection problem $(N, k, F, l, u)$ for which $M$ is precisely the individual-panel membership matrix for the set of feasible panels.

*Proof.* Given a set system $\mathcal{S} \subseteq \binom{[n]}{k}$, we may construct a set of upper quotas such that the collection of feasible panels is exactly $\mathcal{S}$.

To do this, construct a binary feature $f_T$ for each $T \notin \mathcal{S}$. For each $i$ in $[n]$, let $f_T(i) = 1$ if and only if $i \in T$; otherwise let $f_T(i) = 0$. Finally, enforce the upper quota that for all feasible panels $P \subset [n]$,

$$\sum_{i \in P} f_T(i) \leq k - 1,$$

for all $T \notin \mathcal{S}$—that is, no feasible panel has more than $k-1$ members belonging to any $T$. Clearly no $T \notin \mathcal{S}$ is a feasible panel. For $S \in \mathcal{S}$, observe that $|S| = k$, and so for all $T \notin \mathcal{S}$, we have $|S \cap T| \leq k - 1$. Therefore all $S \in \mathcal{S}$ are feasible.

Finally, it bears noting that this is also possible to execute using lower quotas: taking $f'_T(i) = 1 - f_T(i)$, we could instead enforce for each $T \notin \mathcal{S}$ that

$$\sum_{i \in P} f'_T(i) \geq 1.$$

□

*Proof of Section 4.3.3.* Using lemma C.2.4, our aim is to identify and deploy some matrix $M \in \{0,1\}^{n \times |\mathsf{K}|}$ for which

$$\min_{\bar{x} \in \bar{\Delta}} \|M\bar{x}\|_\infty = \Omega\left(\sqrt{k}\right),$$

where $\bar{\Delta} := \{x \in \{\ldots, -3, -1, 1, 3, \ldots\}^n : \sum_i x_i = 0\}$ and all columns of $M$ sum to $k$. Translating and scaling appropriately and applying lemma C.2.4, this will provide our desired $\Omega\left(\frac{\sqrt{k}}{m}\right)$ lower bound.

The common instances which provide lower bounds of $\Omega(\sqrt{k})$ for the Beck-Fiala problem are insufficient for our purposes in two respects. First, while they are column-sparse, they are generally

not uniform in column norm. Second, they are incomparable in terms of the $\bar{x}$ which they quantify over: the Beck-Fiala problem considers minimizing $\|M\bar{x}\|_\infty$ in the more restrictive rounding setting where $\bar{x} \in \{-1, 1\}^n$, while we are concerned with $\bar{x} \in \bar{\Delta}$.

We overcome these barriers by first modifying the Walsh matrices — a family of Hadamard matrices — in order to guarantee uniform column norms, and then modifying the Beck-Fiala lower bound proof of [253, Theorem 19] for arbitrary Hadamard matrices to apply to our matrices for all $\bar{x} \in (2\mathbb{Z} + 1)^n$.

To begin, let $H_t$ be the $2^t \times 2^t$ Walsh matrix, defined recursively by $H_0 = 1$ and

$$H_{t+1} = \begin{bmatrix} H_t & H_t \\ H_t & -H_t \end{bmatrix}.$$

Let $N := 2^t$ denote its dimension.[1] It is a fact that all rows (and columns) besides the first have an equal number of 1 and $-1$ entries. Therefore we take $H'_t$ to be the submatrix derived by dropping the first two columns of $H_t$. (We remove the first column so that all remaining columns have equal sum; we remove the second so that $\bar{\Delta}$ is nonempty). Additionally, let $h_i$ denote the rows of $H'_t$, and $h^j$ denote its columns. Then $H'_t$ has the property that $\sum_i h_i^j = 0$, and in particular all columns $h^j$ have $N/2$ 1-entries.

We have the following lemma:

**Lemma C.2.5.**

$$\min_{x \in \bar{\Delta}} \|H'_t\, x\|_\infty \geq \frac{N-2}{\sqrt{N}},$$

*where* $\bar{\Delta} := \{x \in \{\ldots, -3, -1, 1, 3, \ldots\}^{N-2}\}$.

*Proof.* This right-hand side is $H'_t\, x = (h_1 x, \ldots, h_N x)^T$. We aim to show that there is some $i$ for which $|h_i x|$ is large. Writing $\|H'_t\, x\|_2^2$ two ways, we have that

$$\begin{aligned} \sum_i (h_i x)^2 &= \|x_1 h^1 + \ldots + x_{N-2} h^{N-2}\|_2^2 \\ &= \sum_j x_j^2 \|h^j\|_2^2 + \sum_{j \neq k} x_j x_k (h^j \cdot h^k). \end{aligned}$$

The entries of $H_t$ are all $\pm 1$, and $h^j \cdot h^k = 0$ for $j \neq k$ (since the columns of $H_t$ and therefore $H'_t$ are orthogonal), so this becomes

$$\begin{aligned} &= (N-2) \sum_j x_j^2 \\ &\geq (N-2)^2, \end{aligned}$$

[1] Note that this $N$ is a variable used only in this proof, and it is unrelated to the pool $N$ and its magnitude $n$ as used in the paper body.

since $x_i^2 \geq 1$ by assumption. Therefore by averaging there is some $i$ for which $(h_i x)^2 \geq \frac{(N-2)^2}{N}$, and so $|h_i x| \geq \frac{N-2}{\sqrt{N}}$), as desired. □

Next we translate $H_t'$ into an instance of the panel selection problem and argue it has the desired properties. Take $M := \frac{1}{2}(H_t + 1^{N\times(N-2)})$ to be the $\{0,1\}$ matrix derived from $H_t'$.

The fact that $M$ has uniform column norm $k = N/2$ directly follows from a property of Walsh matrices. Therefore we may apply lemma C.2.4 to argue that $M$ is realizable as the individual-panel membership matrix for some instance of the panel selection problem, with $n = N$, $|\mathsf{K}| = N-2$, and $k = N/2$.

To conclude, consider the uniform $p = \left(\frac{1}{N-2}, \ldots, \frac{1}{N-2}\right)$, with $m = a(N-2) + (N-2)/2$ for any $a \in \mathbb{Z}_+$. In this case, each coordinate of $p$ falls evenly between multiples of $1/m$ and must be rounded to multiples of $1/m$. Letting $x := p - \lfloor mp \rfloor / m = (1/2m, \ldots, 1/2m)$ be this vector of remainders, we must replace it with some $\bar{x} \in (\mathbb{Z}/m)^{N-2}$, while maintaining that $\sum_j \bar{x}_j = \sum_j x_j = (N-2)/2m$, so that the resulting $\bar{p} = \lfloor mp \rfloor / m + \bar{x}$ remains a distribution over panels. (Note that here negative $\bar{x}_j$ signify that the distribution mass on panel $j$ decreases from $p$ to $\bar{p}$.)

Explicitly, we then have

$$\begin{aligned} \|\pi - \bar{\pi}\|_\infty &= \|Mp - M\bar{p}\|_\infty && \text{(C.10)} \\ &= \|M(x - \bar{x})\|_\infty && \text{(C.11)} \\ &= \frac{1}{2m}\|My\|_\infty, && \text{(C.12)} \end{aligned}$$

where $y := 2m(\bar{x} - x)$.

$$\begin{aligned} &= \frac{1}{2m}\|\frac{1}{2}H_t' y + \frac{1}{2}1^{N\times(N-2)} y\|_\infty && \text{(C.13)} \\ &= \frac{1}{4m}\|H_t' y\|_\infty, && \text{(C.14)} \end{aligned}$$

where $\sum_i y_i = 0$ because we require that $\bar{p}$ remain a distribution. Then since $y \in (2\mathbb{Z}+1)^{N-2}$, by lemma C.2.5 we have

$$\begin{aligned} &\geq \frac{N-2}{4m\sqrt{N}} && \text{(C.15)} \\ &= \Omega\left(\frac{\sqrt{k}}{m}\right), && \text{(C.16)} \end{aligned}$$

since $k = N/2$.

This holds for all $y \in (2\mathbb{Z}+1)^{N-2}$. Recall that $\overline{\mathcal{D}} := \{\bar{p} \in (\mathbb{Z}_+/m)^{|\mathsf{K}|} : \|\bar{p}\|_1 = 1\}$, and so

$$\overline{\mathcal{D}} \subseteq \{p + \bar{\Delta}/2m\}.$$

Therefore (C.16) implies that

$$\min_{\tilde{p} \in \overline{\mathcal{D}}} \|\pi - \bar{\pi}\|_\infty = \Omega\left(\frac{\sqrt{k}}{m}\right),$$

as desired. □

### C.2.3 Additional Beyond-Worst-Case Upper Bounds

Since some of our beyond-worst-case upper bounds apply to anonymous realizable $\pi$, it is reasonable to ask how prevalent anonymous realizable $\pi$ are, for arbitrary instances of sortition. Fortunately, we have the following claim:

**Claim C.2.6.** *For any instance of the panel selection problem and any realizable $\pi$, let $\pi'$ be the "anonymized" marginals obtained by setting $\pi'_i$ to the average $\pi_{i'}$ across all $i'$ with the same feature vector as $i$. Then $\pi'$ is realizable also.*

*Proof of Claim C.2.6.* Let $\pi^*$ denote the "anonymization" of $\pi$, and take

$$\Pi := \left\{ \pi' : \text{ realizable, and for all } c, \sum_{i:F(i)=c} \pi'_i = \sum_{i:F(i)=c} \pi_i \right\}.$$

We will show that $\pi^* \in \Pi$.

We argue by way of contradiction. Let $\hat{\pi}$ denote the "most anonymized" $\pi' \in \Pi$, in the sense that

$$\hat{\pi} = \arg\min_{\pi' \in \Pi} \max_c \left( \max_{i:F(i)=c} \pi'_i - \min_{i:F(i)=c} \pi'_i \right).$$

Let $i$ and $i'$ be some pair of individuals with $F(i) = F(i')$ witnessing this maximum diameter, and let $p$ be a distribution with marginals $\hat{\pi}$. For each such pair, we will argue that $p$ may be modified so that $\hat{\pi}_i = \hat{\pi}_{i'}$ while leaving all other marginals unchanged. By iteratively applying this to all such pairs, we will contradict the minimality of $\hat{\pi}$.

To start, observe that by assumption $\hat{\pi}_i > \hat{\pi}_{i'}$. Let $p'$ be the distribution over feasible panels which is the same as $p$, except that $i$ and $i'$ switch places in any panel on which either of them appear. All such panel replacements yield feasible panels, since they have the same feature vector $c$. Finally take $p_{new} = (p + p')/2$. As promised, this distribution has the property that $\pi_i = \pi_{i'}$ and all other marginals are unchanged. □

As a belated warm-up to the beyond-worst-case guarantees, we address the case when there is only one feature of interest, so that $F = \{f\}$. It turns out that we can obtain strong guarantees for this special case without using the machinery deployed in the proof of Section 4.3.2. We place no constraints on the size of the set of feature values $\Omega$, nor do we require that $\pi$ is anonymous.

**Theorem C.2.7.** *If $\pi$ is realizable and $|F| = 1$, then we may efficiently identify $\bar{p}$ such that its marginals $\bar{\pi}$ satisfy*

$$\|\pi - \bar{\pi}\|_\infty < \frac{2}{m}.$$

*Proof of Theorem C.2.7.* Given marginals $\pi$, let $p$ be a distribution over feasible panels $\mathsf{K}$ which witnesses $\pi$. The first step of this rounding is to consider the marginals $\tau_v$ of each feature value $v$: $\tau_v = \sum_{i:f(i)=v} \pi_i$. Note that $\sum_v \tau_v = \sum_i \pi_i = k$. Since there is only one feature, all feasible panels $P$ satisfy

$$l_v \le |\{i \in P : f(i) = v\}| \le u_v, \tag{C.17}$$

and taking the expectation of this over $p$ gives

$$l_v \le \mathsf{E}_p[|\{i \in P : f(i) = v\}|] \le u_v \tag{C.18}$$

$$l_v \le \tau_v \le u_v. \tag{C.19}$$

Therefore $l_v \le \lfloor \tau_v \rfloor$ and $u_v \ge \lceil \tau_v \rceil$. We will construct a new distribution $\bar{p}$ over panels $P$ which satisfy $\lfloor \tau_v \rfloor \le |\{i \in P : f(i) = v\}| \le \lceil \tau_v \rceil$ for all features $v$, and are therefore guaranteed to be feasible.

We will construct feasible panels via the following scheme. Consider the interval $[0, km] \subset \mathbb{R}$ as representing the $km$ spots to be allocated across the $m$ panels which will comprise our lottery, and let $s_t := [t-1, t)$ denote spot $t$. Next observe that $m \sum_i \pi_i = km$, and so $m\pi_i$ may be viewed as the expected number of spots which $p$ would give to $i$.

First group the $\pi_i$ by feature value to form $\tau_v = \sum_{i:f(i)=v} \pi_i$, and then pack them into $[0, km]$, so that individuals with common feature values have contiguous sections; let $S_i$ denote the portion of $[0, km]$ allocated to $i$, so that $|S_i| = \pi_i$. We will choose an individual $I(t)$ for each spot $s_t$, and then assemble the $m$ panels that comprise $\bar{p}$ by taking

$$P_r := \{I(t) : t = wm + r \text{ for } w \in \{0, \dots, k-1\}\}, \tag{C.20}$$

for $r \in \{1, \dots, m\}$.

How to choose which individual will get the spot $t$ for each $t$? If $S_i \supseteq s_t$ then $I(t) = i$. Otherwise, $s_t$ is split between two or more individuals, possibly with different feature values, in which case we call it *contested*. Observe that no matter how these contested $s_t$ are allocated (no matter the choice of $I(t)$ for split $t$), it will be the case that $|\pi_i - \bar{\pi}_i| < 2/m$, since there is at most one contested $s_t$ at each endpoint of the interval $S_i$.

It remains to argue that the panels chosen in (C.20) are feasible; in particular that $\lfloor \tau_v \rfloor \le \bar{\tau}_v \le \lceil \tau_v \rceil$ for all $v$. By construction, each panel $P_r$ has some number of spots which will necessarily be allocated to an individual with feature value $v$, and some number of spots which are contested and may or may not be allocated to an individual with feature vector $v$. For each value $v$, there are at most two spots in all of $[0, km]$ which are *type contested* in this way. If some panel $P_r$

contains at most one type-contested spot for type $v$, then no matter which way it is allocated, $|\{i \in P : f(i) = v\}| - \tau_v| < 1$, and so $P_r$ is feasible with respect to $v$. In the worst case, for some given $v$ both of the spots which are type-contested by $v$ appear on the same panel $P_r$. In order to ensure that $|\{i \in P : f(i) = v\}| - \tau_v| < 1$, it must be the case that exactly one of these two spots is allocated to some $i$ for which $f(i) = v$. Fortunately this constraint is easily satisfiable, even in the case when a given panel $P_r$ contains both of the type-contested spots for multiple features $v$.

Therefore the $\bar{p}$ as constructed by (C.20) is supported by panels which are not only feasible but respect quotas which are maximally tight, given that the input $p$, $\pi$ was realizable. Finally since each $i$ contests at most two spots, we have that

$$\|\pi - \bar{\pi}\|_\infty < \frac{2}{m}.$$ □

**Theorem C.2.8.** *Given realizable anonymous $\pi$, we may efficiently identify $\bar{p}, \bar{\pi}$ such that*

$$\|\pi - \bar{\pi}\|_\infty = O\left(\frac{1}{m} \max\left\{\frac{k}{n_{min}}, 1\right\}\right),$$

*where $n_{min} := \min_c n_c$ is the minimum number of individuals in the pool which share any one feature vector.*

*Proof.* We proceed as in the proof of Section 4.3.2, but apply a different rounding to the panel type LP to obtain $\bar{\mathfrak{p}}$. To begin, $p, \pi$ projects to some $\mathfrak{p}, \tau$. Without loss of generality assume that it is a basic solution to the Type LP (4.4).

We will construct $\bar{\mathfrak{p}}$ from $\mathfrak{p}$ by applying 0/1 rounding as in definition C.2.1.

Note that the constraint matrix $Q$ in (4.3) has the property that for all columns $q^j$, $\|q^j\|_1 = k$. As a special case of [94, Theorem 6], applied to $x'$ and the panel type LP, there exists an $\bar{x}' \in \{0,1\}^{|\mathfrak{K}|}$ such that

$$\|Q(x' - \bar{x}')\|_\infty < 2k.$$

and for which $\|\bar{x}'\|_1 = \|x\|_1$. (This follows from a generalization of the Beck-Fiala algorithm which both respects hard constraints and applies to arbitary matrices $Q$ with bounded column norms, and is therefore also algorithmic.)

Applying lemma C.2.2, we then have

$$\|\tau - \bar{\tau}\|_\infty < \frac{2k}{m}.$$

Given that such a $\bar{\mathfrak{p}}, \bar{\tau}$ exists, it remains to generate $\bar{p}$ and $\bar{\pi}$ in such a way as to give the desired bound on the discrepany in individual marginals. We proceed in a manner identical to the proof of Section 4.3.2.

Again we have that $\bar{\tau} \geq 0$ and $\bar{\tau} = \sum_j Q_{cj}\bar{p}_j \leq m_c \leq n_c$, where $m_c = \max_j Q_{cj}$ and $n_c$ is the number of individuals $i$ for which $F(i) = c$, since $\bar{\mathfrak{p}}$ is a distribution over feasible panel types $j$. Therefore dividing $\bar{\tau}$ amongst the $\bar{\pi}_i$ as equally as possible for each $c$ gives $\bar{\pi}_i \in [0, 1]$.

By the anonymity of $\pi$, for all $i$ with $F(i) = c$, $\pi_i = \tau_c/n_c$, and dividing the spots in $\bar{\mathfrak{p}}$ for feature vector $c$ as equally as possible amongst the $n_c$ individuals gives $\bar{\pi}_i \in \{\bar{\tau}_c/n_c \pm \frac{1}{m}\}$. This equal division of spots in order to form $\bar{p}$ from $\bar{\mathfrak{p}}$ is feasible by the same algorithm 4 as in the proof of Section 4.3.2. Therefore the resulting $\bar{p}, \bar{\pi}$ satisfies

$$\begin{aligned}\|\pi - \bar{\pi}\|_\infty &= \max_c |\tau_c/n_c - \bar{\pi}| \\ &< \frac{1}{n_c}\|\tau - \bar{\tau}\|_\infty + \frac{1}{m} \\ &< \frac{2k}{n_{min} \cdot m} + \frac{1}{m}.\end{aligned}$$

□

## C.3 Omitted Proofs from Section 4.4

There exists a Maximin-optimal $p^*$ such that, for all uniform lotteries $\bar{p}$,

$$\text{Maximin}(p^*) - \text{Maximin}(\bar{p}) = \Omega\left(\frac{\sqrt{k}}{m}\right).$$

*Proof of Section 4.4.1.* We will follow the proof of Section 4.3.3: first we use the Walsh matrices to construct a matrix with the desired properties, prove a modified version of Lemma C.2.5 for it, and then appeal to lemma C.2.4 to argue that it corresponds to a realizable instance of the panel selection problem.

In contrast to the construction in section 4.3.3, where we need only demonstrate that *some* $\bar{\pi}_i$ deviates from $\pi_i$, we must construct an instance for which (essentially) the minimum $\pi_i$ necessarily *decreases*. We accomplish this by first modifying the Walsh matrices to have uniform row norm, so that $\pi$ is uniform and all $\pi_i$ are minimal. We then introduce a second set of "twin" individuals, each $i'$ of which is a member of the panels which their twin $i$ is not. This ensures that any discrepancy in $\bar{\pi} - \pi$ is witnessed in the downward direction.

To begin, again let $H_t$ be the $2^t \times 2^t$ Walsh matrix, with $N := 2^t$ its dimension. This time we take $H_t^*$ to be the submatrix derived by dropping the first row of $H_t$. By properties of Walsh matrices, all remaining rows in $H_t^*$ have an equal number of 1 and $-1$ entries, (though this is no longer true of the columns).

Again letting $h_i$ denote the rows of $H_t^*$, and $h^j$ denote its columns, we have the following new version of lemma C.2.5, which requires the additional assumption that $\sum_j x_j = 0$:

**Lemma C.3.1.**

$$\min_{x \in \Delta^*} \|H_t^* x\|_\infty \geq \sqrt{N},$$

*where* $\Delta^* := \{x \in \{\ldots, -3, -1, 1, 3, \ldots\}^N : \sum_j x_j = 0\}$.

*Proof.* This right-hand side is $H_t' x = (h_1 x, \ldots, h_{N-1} x)^T$. We aim to show that there is some $i$ for which $|h_i x|$ is large. Writing $\|H_t' x\|_2^2$ two ways, we have that

$$\begin{aligned}\sum_i (h_i x)^2 &= \|x_1 h^1 + \ldots + x_N h^N\|_2^2 \\ &= \sum_j x_j^2 \|h^j\|_2^2 + \sum_{j \neq k} x_j x_k (h^j \cdot h^k)\end{aligned}$$

the entries of $H_t^*$ are all $\pm 1$, and $h^j \cdot h^k = -1$ for $j \neq k$ (since the columns of $H_t$ were orthogonal), so this becomes

$$\begin{aligned}&= (N-1) \sum_j x_j^2 - \sum_{j \neq k} x_j x_k \\ &= N \sum_j x_j^2 - \sum_j \sum_k x_j x_k \\ &= N \sum_j x_j^2 \\ &\geq N^2,\end{aligned}$$

since $x_i^2 \geq 1$ by assumption. Therefore by averaging, there is some $i$ for which $(h_i x)^2 \geq \frac{N^2}{N-1}$, and so $|h_i x| \geq \sqrt{N}$, as desired. □

As constructed, all rows of $H_t^*$ have the same number of 1s, so when we transform it into some $M$ for some instance of the panel selection problem, it will yield that the marginals $\pi$ of uniform $p$ are uniform. However we cannot yet apply lemma C.2.4, since the columns of the resulting $M$ do not have constant norm; in particular, the first column will be all 1s.

In order to simultaneously correct for this and translate from $\ell_\infty$ to Maximin lower bounds, we introduce "twins" for each $i$. Letting $M^* = \frac{1}{2}(H_t^* + 1^{(N-1)\times N})$ be this $\{0,1\}$ matrix, define $\bar{M}^* := 1^{(N-1)\times N} - M^*$ to be its complement, so that $M_{ij}^* = 1 - \bar{M}_{ij}^*$ for all $i, j$. Finally take

$$M = \begin{bmatrix} M^* \\ \bar{M}^* \end{bmatrix}$$

and observe that this $M \in \{0,1\}^{(2N-2)\times N}$ has uniform column norm $N-1$ because of $\bar{M}^*$. We may therefore apply lemma C.2.4 to claim that it is the individual-panel membership matrix of some instance of the panel selection problem.

The remainder of the argument proceeds similarly to that of lemma C.2.5, with additional step of showing that the lower bound holds for the maximin objective. We include the full argument for completeness.

Similarly take $p = \left(\frac{1}{N}, \ldots, \frac{1}{N}\right)^T$, with $m = aN + N/2$ for any $a \in \mathbb{Z}_+$, $n = 2N - 2$ (the number of individuals), and $k = N - 1$. This $p$ gives equal marginals: here $\pi_i = (Mp)_i = \frac{N-1}{2N-2} = \frac{k}{n}$ for all $i$. Again each coordinate of $p$ falls evenly between multiples of $1/m$ and must be rounded to multiples of $1/m$. Letting $x := p - \lfloor mp \rfloor / m = (1/2m, \ldots, 1/2m)^T$ be this vector of remainders, we must replace it with some $\bar{x} \in (\mathbb{Z}/m)^N$, while maintaining that $\sum_j \bar{x}_j = \sum_j x_j = N/2m$, so that the resulting $\bar{p} = \lfloor mp \rfloor / m + \bar{x}$ remains a distribution over panels.

Explicitly, we then have

$$\|\pi - \bar{\pi}\|_\infty = \|Mp - M\bar{p}\|_\infty \tag{C.21}$$
$$= \|M(x - \bar{x})\|_\infty \tag{C.22}$$
$$= \frac{1}{2m} \| \begin{bmatrix} M^* \\ \bar{M}^* \end{bmatrix} y \|_\infty, \tag{C.23}$$

where $y := 2m(\bar{x} - x)$. Because $\ell_\infty$ is a maximum, this is

$$\geq \frac{1}{2m} \|M^* y\|_\infty \tag{C.24}$$
$$= \frac{1}{2m} \| \frac{1}{2} H_t^* y + \frac{1}{2} 1^{(N-1)\times N} y \|_\infty \tag{C.25}$$
$$= \frac{1}{4m} \|H_t^* y\|_\infty, \tag{C.26}$$

where $\sum_i y_i = 0$ because we require that $\bar{p}$ remain a distribution. Then since $y \in (2\mathbb{Z} + 1)^{N-2}$, by lemma C.2.5 we have

$$\geq \frac{\sqrt{N}}{4m} \tag{C.27}$$
$$= \Omega\left(\frac{\sqrt{k}}{m}\right), \tag{C.28}$$

since $k = N - 1$. Again since $\overline{\mathcal{D}} \subseteq \{p + \bar{\Delta}/2m\}$, we then have

$$\min_{\bar{p} \in \overline{\mathcal{D}}} \|\pi - \bar{\pi}\|_\infty = \Omega\left(\frac{\sqrt{k}}{m}\right).$$

Since $\pi$ is uniform by construction (and so these $p$ and $\pi$ are optimal with respect ot Maximin), this is a lower bound on the discrepancy of each marginal which was minimal *before deviation*. It finally remains to show that this deviation happens in the downward direction, so that the minimum marginal decreases by at least this amount. Observe that by the construction of $\bar{M}^*$, for all $\bar{p}$ we have $(M^*\bar{p})_i = -(\bar{M}^*\bar{p})_i$. Therefore for any given $\bar{p}$, whichever coordinate $i$ satisfies $|(\pi - \bar{\pi})_i| = \Omega(\sqrt{k}/m)$, there is a coordinate $i'$ for which $(\pi - \bar{\pi})_{i'} = \Omega(\sqrt{k}/m)$. Therefore in this instance

$$\text{Maximin}(p^*) - \max_{\bar{p} \in \overline{\mathcal{D}}} \text{Maximin}(\bar{p}) = \Omega\left(\frac{\sqrt{k}}{m}\right),$$

as desired. □

For NW-optimal $p^*$ over a support of panels $supp(p^*)$, there exists a constant $\lambda \in \mathsf{R}^+$ such that, for all $P \in supp(p^*)$, $\sum_{i \in P} 1/\pi_i^* = \lambda$.

*Proof of Section 4.4.2.* We can write the problem of finding the NW optimizing distribution over a fixed panel support $\mathcal{P} \subseteq \mathcal{K}$ as below on the left, where $NW^n(p)$ is equal to the product of the $\pi_i$, the marginals implied by the panel distribution $p$ (in contrast, in Section 4.2, we let $NW(p)$ be the geometric mean—here we we take the $n^{th}$ power). On the right, we've rewritten the program in standard form, where we set $f(p) = -NW^n(p)$, $h(p) = p_1 + p_2 + \cdots + p_{|\mathcal{P}|} - 1$, and $g_j(p) = -p_j$. Observe that, $\forall j \in [|\mathcal{P}|]$, $\nabla h(p) = \mathbf{1}$ and $\nabla g_j(p) = -e_j$, where $e_j$ is the vector of 0s with a 1 at index $j$.

$$\begin{aligned} &\max_p NW^n(p) && \min_p f(p) \\ &\|p\|_1 = 1 && h(p) = 0 \\ &p_j \geq 0 \ \forall j \in [|\mathcal{P}|] && g_j(p) \leq 0 \ \forall j \in [|\mathcal{P}|] \end{aligned}$$

Now, let $p^*$ be an optimal solution to this program, and $supp(p^*)$ be its support, i.e., the set of panels to which $p^*$ assigns nonzero probability. Then, since the objective and constraints of the above program are continuously differentiable over their entire support (and thus at $p^*$), by the KKT condition Stationarity, there exist some constants $\lambda$ and $\mu_j$ for all $j \in [|supp(p^*)|]$ (where $\mathbf{0}$ is the zero vector) such that

$$\nabla f(p^*) + \lambda \nabla h(p^*) + \sum_{j \in [|supp(p^*)|]} \mu_j \nabla g(p^*) = \mathbf{0} \implies (\nabla f(p^*))_j = \mu_j - \lambda$$

By dual feasibility and primal feasibility respectively, we have that $\mu_j, p_j \geq 0$ for all $j \in [|supp(p^*)|]$; by complementary slackness, we have that $\sum_{j \in [|supp(p^*)|]} \mu_j p_j^* = 0$. Thus, for all $j$, either $p_j^* = 0$, or $p_j^* > 0$ and $\mu_j = 0$. We have restricted $supp(p^*)$ to panels $j$ in which $p_j^* > 0$, so we conclude that $\mu_j = 0$. It follows that

$$\frac{\partial NW^n(p^*)}{\partial p_j^*} = -\left(\nabla f(p^*)\right)_j = -(\mu_j - \lambda) = \lambda \quad \forall j \in supp(p^*)$$

Finally, we can conclude the proof by expressing this partial derivative for fixed $p_j$ (which as shown, has a constant value across all $j$ in the support) in terms of the marginals $\pi$. We obtain that for all $j$ in $supp(p^*)$,

$$\lambda = \frac{\partial NW^n(p^*)}{\partial p_j^*} = \sum_{i \in N} \frac{NW^n(p^*)}{\pi_i^*} \frac{\partial \pi_i^*}{\partial p_j^*} = \sum_{i \in P_j} \frac{NW^n(p^*)}{\pi_i^*} = NW^n(p^*) \left( \sum_{i \in P_j} \frac{1}{\pi_i^*} \right)$$

where $P_j$ is the $j^{th}$ panel in $supp(p^*)$. The second equality is by the product rule for derivatives, where each term of the resulting sum is equal to the derivative of $\pi_i^*$ with respect to $p_j^*$ multiplied by $NW/\pi_i^*$, the NW holding out the marginal of individual $i$. The third equality is by the fact that if $i \in P_j$, then $\partial \pi_i^* / \partial p_j^* = 1$; otherwise $\partial \pi_i^* / \partial p_j^* = 0$. □

For NW-optimal $p^*, \pi^*$, we have that $\pi_i^* \geq 1/n$ for all $i \in N$.

*Proof of Section 4.4.2.* Let $X[P \ni i]$ be the indicator that a panel $P$ contains individual $i$. Then,

$$\mathbb{E}_{P\sim p^*}\left[\sum_{i\in P}\frac{1}{\pi_i^*}\right] = \mathbb{E}_{P\sim p^*}\left[\sum_{i\in N}\frac{X[P\ni i]}{\pi_i^*}\right] = \sum_{i\in N}\frac{\mathbb{E}_{P\sim p^*}[X[P\ni i]]}{\pi_i^*} = \sum_{i\in N}\frac{\pi_i^*}{\pi_i^*} = n$$

By Section 4.4.2, we also have that $\mathsf{E}\left[\sum_{i\in P}\frac{1}{\pi_i^*}\right] = \lambda/NW^n(p^*)$, and thus $\lambda/NW^n(p^*) = n$. It follows that for all panels $P$, $\sum_{i\in P}\frac{1}{\pi_i^*} = \lambda/NW^n(p^*) = n$ and therefore $\pi_i^* \geq 1/n\,\forall i \in N$; otherwise, we would have some panel $P$ for which $\sum_{i\in P}\frac{1}{\pi_i} > n$, a contradiction. □

For NW-optimal $p^*, \pi^*$, there exists a uniform lottery $\bar{p}, \bar{\pi}$ that satisfies $\mathrm{NW}(p^*) - \mathrm{NW}(\bar{p}) \leq k\,\|\pi^* - \bar{\pi}\|_\infty$.

*Proof of Section 4.4.2.* Let $\pi_{min}^*$ be the smallest marginal of any individual implied by the Nash-optimal distribution over panels $p^*$, i.e., $\pi_{min}^* = \min_{i\in N}\pi_i^*$. Then, to upper-bound the loss in NW, we assume an unattainable worst case that between $p^*, \pi^*$ and a given uniform lottery $\bar{p}, \bar{\pi}$, *all* individuals probabilities suffer the largest loss of any marginal, $\|\pi^* - \bar{\pi}\|_\infty$, and that this loss manifests multiplicatively as badly as if all agents had original marginal probability $\pi_{min}^*$. This first gives the multiplicative bound:

$$NW(\overline{p}^*) \geq NW(p^*)\left(\frac{\pi_{min}^* - \|\pi^* - \bar{\pi}\|_\infty}{\pi_{min}^*}\right) = NW(p^*)\left(1 - \frac{\|\pi^* - \bar{\pi}\|_\infty}{\pi_{min}^*}\right).$$

Rearranging the above conclusion and then applying the facts that $NW(p^*) \leq k/n$ (trivially) and $\pi_{min}^* \geq 1/n$ (Section 4.4.2), we get the desired additive bound:

$$NW(p^*) - NW(\overline{p}) \leq NW(p^*)\cdot\frac{\|\pi^* - \bar{\pi}\|_\infty}{\pi_{min}^*} \leq \frac{k}{n}\cdot\frac{\|\pi^* - \bar{\pi}\|_\infty}{1/n} \leq k\,\|\pi^* - \bar{\pi}\|_\infty$$ □

## C.4 Omitted Materials from Section 4.5

### C.4.1 Algorithm Descriptions

**Algorithms for calculating optimal panel distributions.**
In this paper, we calculate optimal panel distributions across instances with respect to Maximin, NW, and Leximin objectives. To do this, we build on publicly-available code [163], which implements the column generation techniques from [130].

**Rounding algorithms.**
At a high level, the task solved by the Pipage and Beck-Fiala rounding algorithms in Section 4.5 can be thought of as rounding an input panel distribution $p$ to some uniform lottery $\bar{p}$ by rounding the Standard LP described in Section 4.3. However, neither of these rounding methods are used

to directly round $p$; rather, they are used to round a modified version $p'$, which transforms the task from rounding entries of $p$ to multiples of $1/m$ to the task of rounding entries of $p'$ to 0/1. The details of this transformation are described in the proof of Section 4.3.1 in Appendix C.2.

PIPAGE

We round $p'$ exactly according to the Pipage Rounding algorithm specified in Gandhi *et al* [141]. We note that their algorithm is specified for the task of rounding bipartite graphs; we apply their methods by formulating our rounding problem as a star graph, where each of the $|\mathcal{K}|$ vertices surrounding the central vertex corresponds to a feasible panel $P$. Each edge from the central vertex $i$ to a surrounding vertex $P$ has a weight (which will ultimately be rounded to 0/1) equal to $x_{i,P} = p'_P$, the probability of drawing panel $P$ from the modified version of the initial distribution $p'$. Gandhi *et al*'s degree preservation property guarantees the satisfaction of our adding up constraint $\|p'\| = \|\bar{p}'\|$.

BECK-FIALA

Our Beck-Fiala implementation is identical to the deterministic implementation specified in the proof of Lemma 9, Appendix B.4.1 of [128]. For details on the mapping of their setting to ours, see the proof of Section 4.3.1 in Appendix C.2.

**Integer Programs.**

IP-MAXIMIN

The below integer program computes a lottery $\bar{p} \in (\mathbb{Z}^+/m)^{|\mathcal{K}|}$, where the variables are $y$, the lower bound on any marginal probability; $\bar{p}$, the uniform lottery; and $\bar{\pi}$, the implied vector of marginals. The first constraint, along with the objective, result in the maximization of the minimum marginal. The second constraint imposes the relationship between the panel distribution $\bar{p}$ and the marginals $\bar{\pi}$. The third constraint imposes that the resulting panel distribution $x$ will be a uniform lottery. The fourth and fifth constraints impose that $\bar{p}$ is a valid distribution.

$$
\begin{aligned}
&\text{Maximize } y \\
&\quad \text{s.t. } \bar{\pi}_i \geq y && \forall i \in N \\
&\qquad \sum_{\substack{P \in \mathcal{K}, \\ P \ni i}} \bar{p}_P = \bar{\pi}_i && \forall i \in N \\
&\qquad m\, \bar{p}_P \in \mathbb{Z}^+ && \forall P \in \mathcal{K} \\
&\qquad \sum_{P \in \mathcal{K}} \bar{p}_P = 1 \\
&\qquad \bar{p}_P \geq 0 && \forall P \in \mathcal{K}
\end{aligned}
$$

IP-NW

This integer program is essentially the same as IP-MAXIMIN, except that instead of maximizing the lower bound on the marginals, it maximizes the geometric mean of the marginals by equivalently

maximizing the sum of their logarithms.

$$\begin{aligned}
\text{Maximize } & \sum_{i \in N} \log(\bar{\pi}_i) & \\
\text{s.t. } & \sum_{\substack{P \in \mathcal{K}, \\ P \ni i}} \bar{p}_P = \bar{\pi}_i & \forall i \in N \\
& m\, \bar{p}_P \in \mathbb{Z}^+ & \forall P \in \mathcal{K} \\
& \sum_{P \in \mathcal{K}} \bar{p}_P = 1 & \\
& \bar{p}_P \geq 0 & \forall P \in \mathcal{K}
\end{aligned}$$

IP-Marginals

This IP takes as input some panel distribution $p, \pi$ to be rounded, and minimizes the largest discrepancy of any resulting $\bar{\pi}_i$ from the corresponding $\pi_i$. Again, several of the constraints and variables are common with IP-Maximin.

$$\begin{aligned}
\text{Minimize } & z & \\
\text{s.t. } & |\pi_i - \bar{\pi}_i| \leq z & \forall i \in N \\
& \sum_{\substack{P \in \mathcal{K}, \\ P \ni i}} \bar{p}_P = \bar{\pi}_i & \forall i \in N \\
& m\, \bar{p}_P \in \mathbb{Z}^+ & \forall P \in \mathcal{K} \\
& \sum_{P \in \mathcal{K}} \bar{p}_P = 1 & \\
& \bar{p}_P \geq 0 & \forall P \in \mathcal{K}
\end{aligned}$$

### C.4.2 Implementation Notes and Algorithm Runtimes

Our experiments were implemented in Python and run on a 13-inch MacBook Air (2018) with a 1.6 GHz Intel Core i5 processor.

Runtimes of Pipage, Beck-Fiala, and IP-NW on rounding an unconstrained distribution are given in the table below. We optimized IP-NW with Gurobi using its built-in piecewise linear approximation of logarithms (given that IP-NW is nonlinear) with the parameter controlling the error in the piecewise approximation set to FuncPieceError=0.0001. This worked quite well in most instances, getting within $1/m$ of optimal fairness on 10 out of 11 instances.

IP-Maximin and IP-Marginals were run in Gurobi and struggled to converge completely (even after many hours), but showed good performance after a short time. The results in the paper show their solutions after 30 minutes of run-time.

* indicates capped at 7200s (2 hours). Time is measured in seconds. All times given (except those that timed out) represent the average over 3 runs.

Table C.2: Run-times for Pipage, Beck-Fiala, and IP-NW

| Instance | Pipage | Beck-Fiala | IP-NW |
|---|---|---|---|
| sf(a) | 1.5 | 1.6 | 17.1 |
| sf(b) | 1.3 | 1.3 | 27.8 |
| sf(c) | 1.0 | 1.1 | 33.1 |
| sf(d) | 2.1 | 2.3 | 40.6 |
| sf(e) | 17.0 | 28.3 | 7245* |
| cca | 4.4 | 6.4 | 7207* |
| hd | 1.5 | 1.7 | 120.1 |
| mass | 0.4 | 0.4 | 3.4 |
| nexus | 2.8 | 3.2 | 21.1 |
| obf | 2.3 | 2.4 | 22.3 |
| ndem | 2.2 | 2.6 | 34.8 |

### C.4.3 Analysis of Nash Welfare Fairness Preservation (Figure corresponding to Figure 4.2)

Here we give the corresponding analysis from Figure 4.2 for NW. We see, first that there is some algorithm in every instance that achieves within $0.1/m$ of $NW(p^*)$, where $p^*$ is the NW optimizing unconstrained distribution. This indicates that the cost of transparency to NW in practice is essentially 0. We note that in a few instances, IP-NW, which should theoretically dominate all other algorithms, is outperformed by either Pipage or Beck-Fiala. As we discuss in Appendix C.4.2, this is due to small errors in the integer optimization errors.

We find that our theoretical upper bounds on NW loss are less useful than those on the Maximin loss, because they are multiplied by an additional factor of $k$, while the value of the NW objective falls within a similar range to the Maximin objective. We note, however, that these bounds would be useful for larger $m$: currently, the maximum possible losses implied by the bounds fall between $191/m = 0.191$ and $5922/m = 5.922$. If we increased $m$ by a factor of 100 to $m = 100,000$ (this would mean drawing 5 lottery balls instead of 3), then our bounds would be nearly tight to optimal in multiple instances (e.g., in "sf(a)", this would yield a loss of 0.008), and would be meaningful in all instances.

### C.4.4 Analysis of Leximin Preservation (Figures corresponding to Figure 4.3)

Here we give the corresponding analysis from Figure 4.3 for all other instances. In all instances, the conclusions we draw are essentially the same as those drawn from Figure 4.3: in all instances, all algorithms almost exactly preserve the Leximin-optimal marginals. Our theoretical bounds are meaningful, but we consistently outperform them in practice.

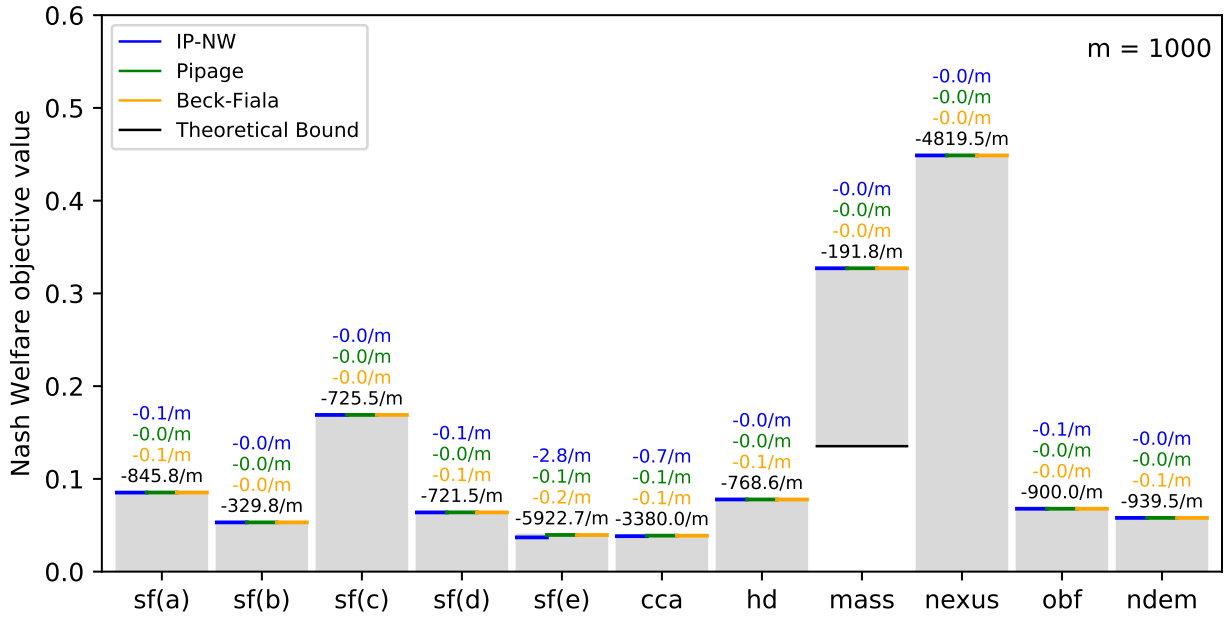



Figure C.1: $m = 1000$. Shaded regions extend from $NW(p^*)$, the fairness of the optimal unconstrained distribution, down to the minimum fairness implied by the tightest theoretical upper bound in that instance (in all instances but "obf" Section 4.3.2 is tightest). Each algorithm or bound's loss relative to $NW(p^*)$ is written above in the corresponding color. We show a representative run of Pipage, a randomized algorithm.

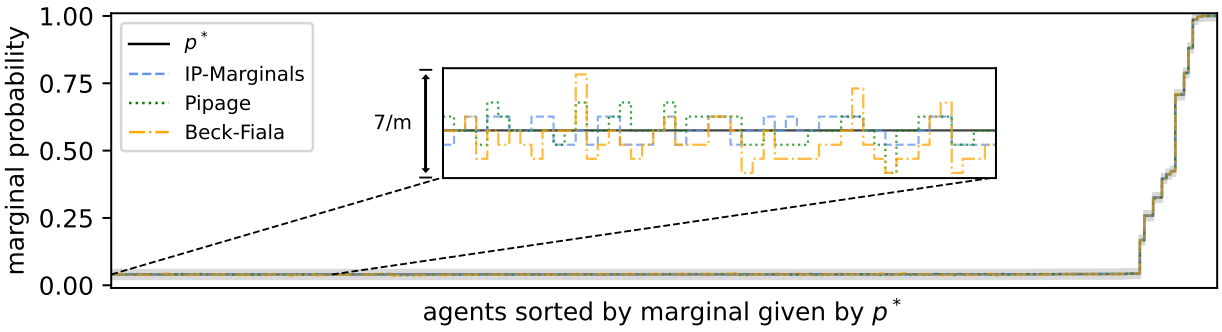



Figure C.2: sf(b)

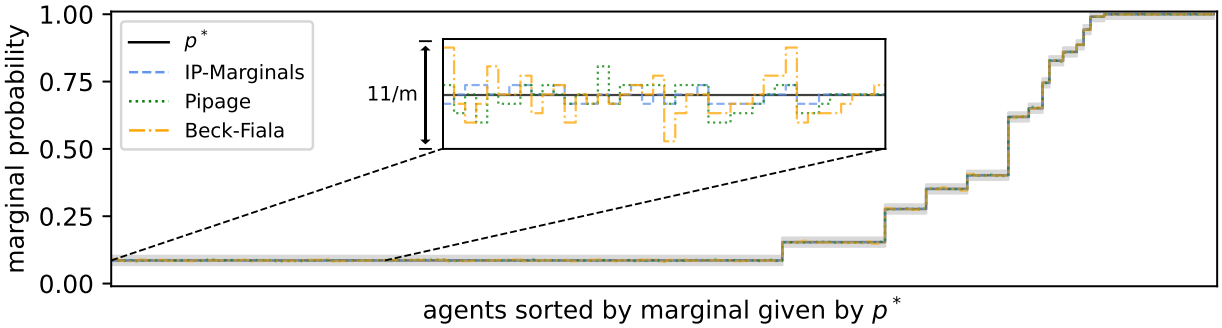



Figure C.3: sf(c)

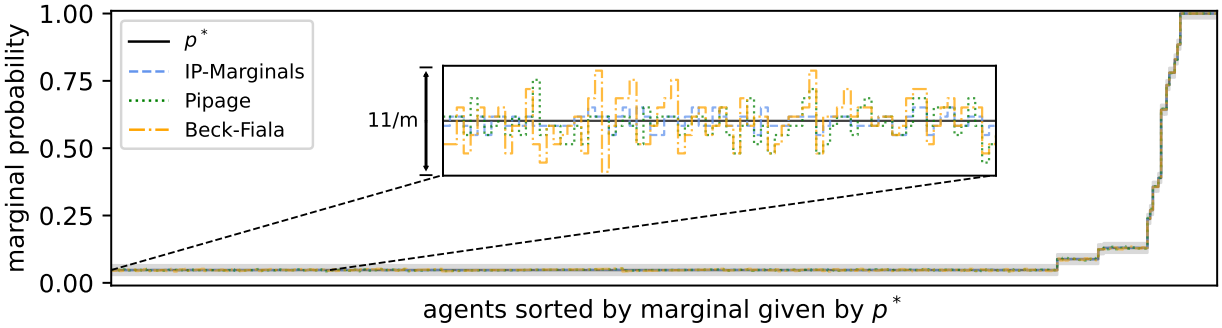



Figure C.4: sf(d)

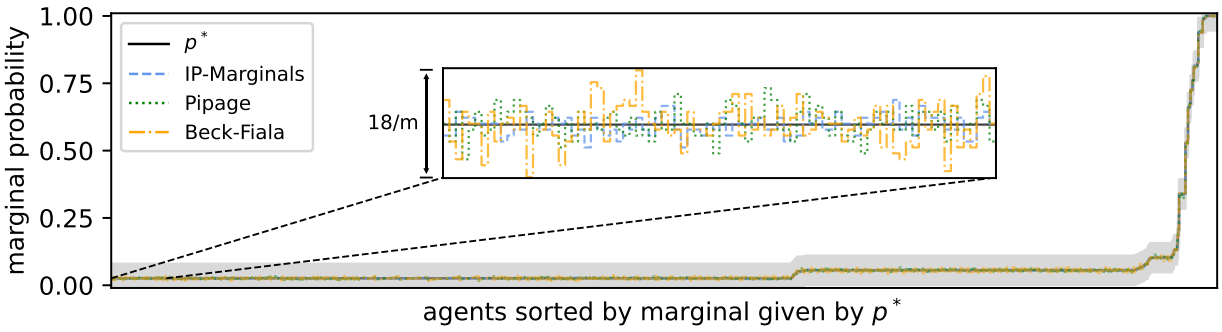



Figure C.5: sf(e)

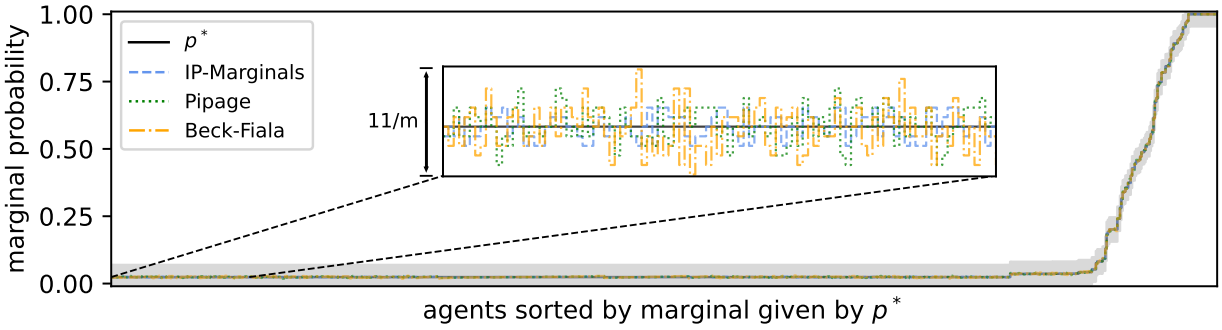



Figure C.6: cca

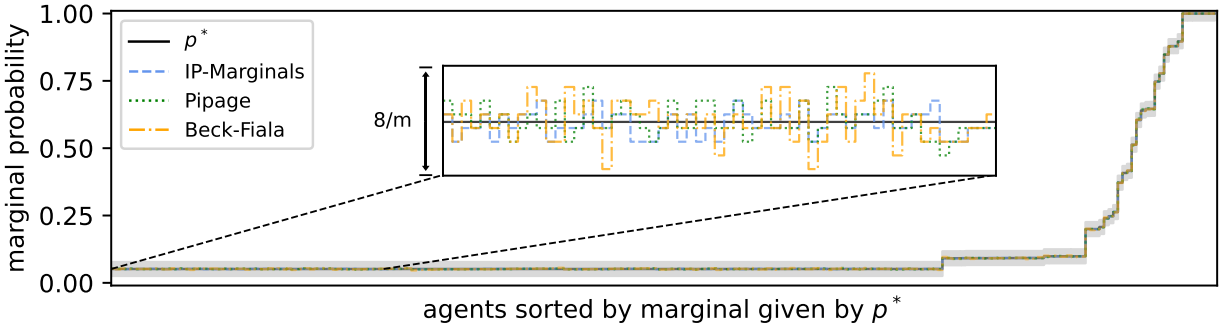



Figure C.7: hd

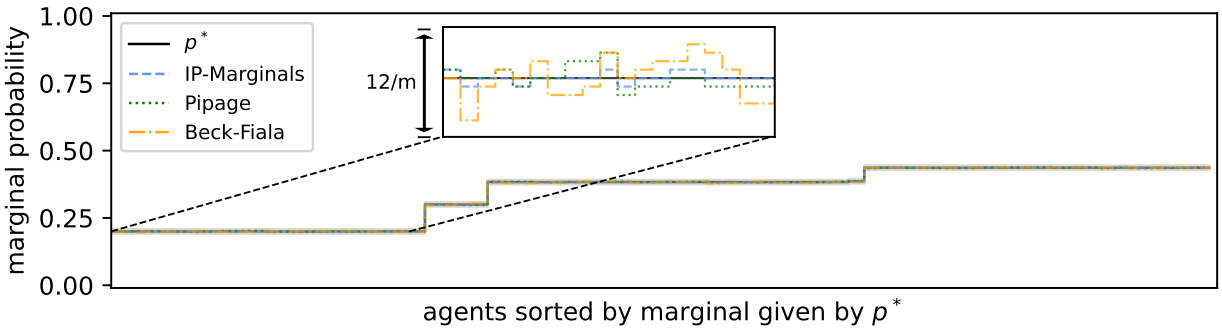



Figure C.8: mass

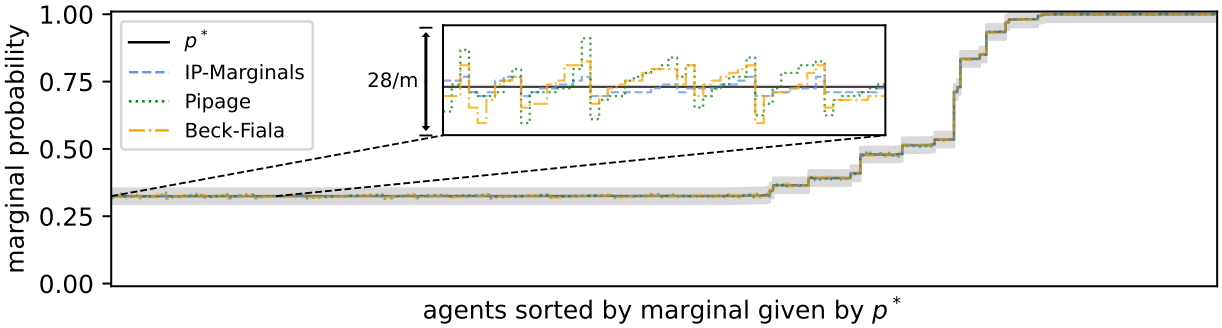



Figure C.9: nexus

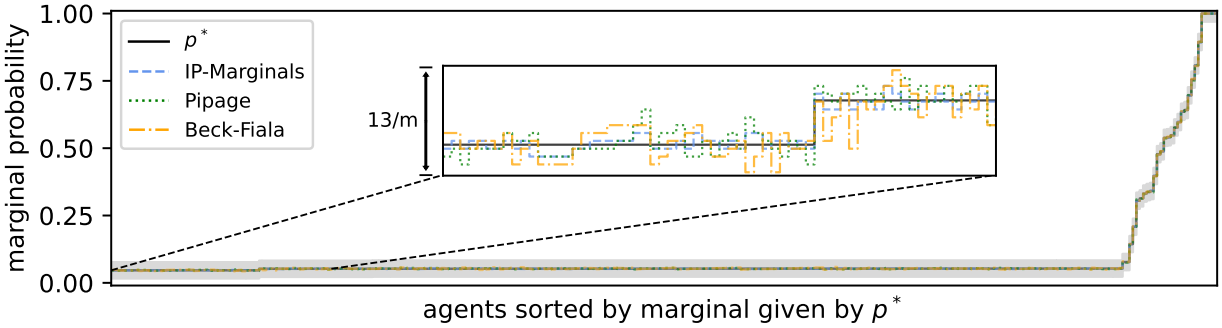



Figure C.10: obf

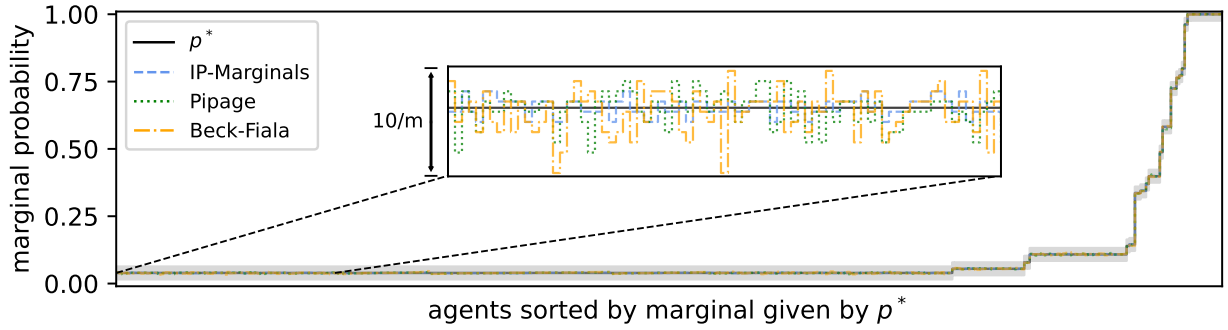



Figure C.11: ndem

# D

# Chapter 5 Appendix

## D.1 Supplemental materials from Section 5.2

### D.1.1 Details of *randomized rounding* step of rounding-based algorithms

**Inputs:** The randomized rounding task takes two inputs:

- a vector of marginal probabilities, $\pi \in [0,1]^n$ such that $\sum_{i\in[n]} \pi_i = k$, and
- an $|FV| + 1 \times n$ matrix $H$, which can be seen as the binary matrix defining the adding up constraints in OPT-PROB. That is, each column of $H$ corresponds to an agent, and each row (except the last) corresponds to a different feature-value $(f, v) \in FV$. The $i$th column has 1s in rows corresponding to feature-values possessed by $i$, and 0s elsewhere. The last row corresponds to the adding up constraints, and so contains a 1 in every column.

**Task:** The goal is to round the entries of $\pi$ into a vector $\tilde{\pi} \in \{0,1\}^n$ such that the three criteria below are satisfied. This rounding procedure will be randomized, so $\tilde{\pi}$ is a random variable. Conceptually, $\tilde{\pi}$ will encode the selected panel $K$, where $\tilde{\pi}_i = 1 \iff i \in K$.

- The adding up constraint is deterministically preserved:

$$\sum_{i\in[n]} \tilde{\pi}_i = k \text{ with probability 1}$$

- The representation constraints satisfied by the original selection probabilities are deterministically satisfied within a relaxation of $|F|$:

$$\sum_{i:f(i)=v} \tilde{\pi}_i \in \sum_{i:f(i)=v} \pi_i \pm |F| \text{ for all } (f,v) \in FV \text{ with probability 1}$$

- The selection probabilities in $\pi$ are preserved:

$$\mathsf{E}[\tilde{\pi}_i] = \pi_i \text{ for all } i \in [n]$$

**Algorithm** (Randomized-Round): This rounding algorithm is exactly the algorithm used to prove Lemma 3 in Flanigan et al. [128]. We outline their key arguments here, rephrased in our notation.

**Lemma D.1.1** (Lemma 9 in [128])**.** *Let $(\pi_i)_{i\in[n]}$ be any collection of variables in $[0,1]$ such that $\sum_{i\in[n]} \pi_i = k$. Then, we can efficiently compute a deterministic 0/1 rounding $(\tilde{\pi}_i)_{i\in[n]}$ such that $\sum_{i\in[n]} \tilde{\pi}_i = k$ and such that, for each feature-value pair $(f,v)$,*

$$\sum_{i:f(i)=v} \tilde{\pi}_i \in \sum_{i:f(i)=v} \pi_i \pm |F|.$$

The proof of this lemma is based on discrepancy theorem by Beck and Fiala [41], and crucially relies on the fact the underlying matrix $H$ is relatively sparse: that is, each agent $i$ has only $|F|+1$ 1s in their column of the $H$ matrix.

The above lemma implies a *deterministic* rounding procedure satisfying only criteria 1 and 2 above. To transform this deterministic rounding procedure into a randomized rounding procedure satisfying criteria 3, as do Flanigan et al., we can apply Theorem 1.2 from [34], which does exactly the needed transformation. We outsource the (relatively straightforward) details of applying this theorem in our setting to Lemma 3 in Flanigan et al. [128].

### D.1.2 Relationship between rounding-based and quota-based algorithms

Conceptually, OPT-PROB is equivalent to the *relaxation* of quota-based algorithms in which all agents are treated as divisible (i.e., a panel can contain fractional agents). We formalize this here now, defining all relevant programs and relaxations, and then proving the equivalence in Proposition D.1.2.

**Definition of quota-based algorithms: OPT-QUOTA.** To begin, we first formally specify the optimization solved by a quota-based algorithm, as used in practice and studied in Flanigan et al. [130]. At a high level, these algorithms differ from Equation (OPT-PROB) by requiring representation in a different way: they impose upper and lower *quotas* on all $(f, v)$, which impose some tolerance of error around each $p_{(f,v)} \cdot k$ that must satisfied deterministically by the chosen panel $K$. Formally, for all $(f, v) \in FV$, a *lower quota* is $\ell_{(f,v)}$, an upper quota is $u_{(f,v)}$, and the chosen panel $K$ is sampled from the *panel distribution* $\rho$ resulting from the optimization program below. We let $\mathcal{K}$ be the collection of all *feasible panels*, that is, all subsets of $[n]$ satisfying the following two constraints:

$$\mathcal{K} := \left\{K : |K| = k \quad \wedge \sum_{i:f(i)=v} \mathbf{1}(i \in K) \in \left[\ell_{(f,v)}, u_{(f,v)}\right] \text{ for all } (f,v) \in FV\right\}.$$

Implicitly, the panel distribution $\rho$ implies selection probabilities $\pi$: $\pi_i$ is equal to the probability of choosing any panel containing $i$, as defined by $p$. We encode this constraint in the optimization problem below:

$$\min_{\pi \in [0,1]^n,\ p \in [0,1]^{|\mathcal{K}|}} g(\pi) \quad \text{s.t.} \quad \sum_{K \in \mathcal{K}} \rho_K = 1 \quad \wedge \quad \pi_i = \sum_{K \in \mathcal{K}} \rho_K \cdot \mathbf{1}(i \in K) \text{ for all } i \in [n] \tag{OPT-QUOTA}$$

***Continuous relaxation* of quota-based algorithms: OPT-QUOTA-CONTINUOUS.**
Now, we define a version of OPT-QUOTA in which individuals are treated as divisible. Then, a panel $X \in [0,1]^n$ is a vector of length $n$, whose $i$-th entry $x_i$ specifies the fraction of agent $i$ included in panel $X$. Then, the set of feasible panels is the following, uncountable infinite set:

$$\mathcal{X} := \left\{X : \sum_{i \in [n]} x_i = k \quad \wedge \sum_{i:f(i)=v} x_i \in \left[\ell_{(f,v)}, u_{(f,v)}\right] \text{ for all } (f,v) \in FV\right\}.$$

Then, for variables $\pi$ and *panel density function* $\rho$, we optimize

$$\min_{\pi \in [0,1]^n,\ \rho} g(\pi) \quad \text{s.t.} \quad \int_{X \in \mathcal{X}} \rho_X dX = 1 \quad \wedge \quad \pi_i = \int_{X \in \mathcal{X}} \rho_X \, x_i dX \text{ for all } i \in [n] \tag{OPT-QUOTA-CONTINUOUS}$$

**Generalized version of rounding-based algorithms: OPT-PROB-RANGE.**
Here, we define a slightly generalized version of OPT-PROB, in which the representation targets are replaced with ranges (where we should think of this range encompassing the exact representation target in OPT-PROB, so $kp_{(f,v)} \in [\ell_{(f,v)}, u_{(f,v)}]$).

$$\min_{\pi \in [0,1]^n} g(\pi) \quad \text{s.t.} \quad \sum_{i:f(i)=v} \pi_i \in [\ell_{(f,v)}, u_{(f,v)}] \quad \text{for all } (f,v) \in FV \quad \wedge \quad \sum_{i \in [n]} \pi_i = k \tag{OPT-PROB-RANGE}$$

FORMAL EQUIVALENCE OF OPT-PROB-RANGE AND OPT-QUOTA-CONTINUOUS. Conceptually, what this shows is that for any quotas $\ell_{(f,v),u_{(f,v)}}$ imposed in quota-based algorithms, maximizing our fairness objective while treating people as *divisible* is equivalent — from the perspective of selection probabilities — to solving our rounding-based optimization problem with the same representation target ranges. To realize exactly OPT-PROB, one could run a quota-based algorithm with divisible agents and $\ell_{(f,v)} = u_{(f,v)} = kp_{(f,v)}$.

**Proposition D.1.2.** *$\pi$ is feasible in OPT-PROB-RANGE $\iff$ there exists a panel density function $\rho$ over $\mathcal{X}$ which realizes $\pi$ in OPT-QUOTA-CONTINUOUS.*

*Proof.* (Forward direction): *$\pi$ is feasible in OPT-PROB $\implies$ there exists a panel density function $\rho$ over $\mathcal{X}$ which realizes $\pi$:*

Fix a feasible $\pi$. Define a panel $X^*$ such that $x^*_i = \pi_i$ for all $i \in [n]$. By the fact that $\pi$ satisfies the constraints in OPT-PROB-RANGE, it follows immediately that $X^* \in \mathcal{X}$. Place all the mass in $\rho$ on $X^*$, so $\rho_{X^*} = 1$ and $\rho_X = 0$ for all $X \in \mathcal{X} \setminus \{X^*\}$. Then, by definition, $\pi$ is realized by this $\rho$, because for all $i \in [n]$,

$$\pi_i = \int_{X \in \mathcal{X}} \rho_X x_i dX = x_i.$$

(Reverse direction): *$\rho$ is a valid density function over $\mathcal{X}$ and implies $\pi$ $\implies$ $\pi$ is feasible in OPT-PROB-RANGE.*

Fix a valid $\rho$ and let it imply $\pi$. Then, we can confirm that $\pi$ satisfies the constraints of OPT-PROB:

$$\sum_{i \in [n]} \pi_i = \sum_{i \in [n]} \int_{X \in \mathcal{X}} \rho_X x_i dX = \int_{X \in \mathcal{X}} \rho_X \sum_{i \in [n]} x_i dX = \int_{X \in \mathcal{X}} \rho_X k dX = k.$$

For any $(f,v) \in FV$, let $\sum_{i:f(i)=v} x_i = r_{(f,v)}$.

$$\sum_{i:f(i)=v} \pi_i = \sum_{i:f(i)=v} \int_{X \in \mathcal{X}} \rho_X x_i dX = \int_{X \in \mathcal{X}} \rho_X \sum_{i:f(i)=v} x_i dX = \int_{X \in \mathcal{X}} \rho_X r_{(f,v)} dX$$
$$\in \left[\ell_{(f,v)}, u_{(f,v)}\right]. \qquad \square$$

## D.2 Supplemental materials from Section 5.3

### D.2.1 Proof of Theorem 5.3.1

*Proof.* This result is most naturally proven using feature vector-indexed analogs of our standard agent-indexed objects, so we define them now: we use $\nu_w(N) := |\{i : i \in [n], w(i) = w\}|/|N|$ to denote the fraction of the pool $N$ containing feature vector $w$, with $\boldsymbol{\nu}(N) = (\nu_w | w \in \mathcal{W})$. Noting that all reasonable objectives (including those considered here) will give all agents with the same vector the same selection probability, we will use $q_w(N)$ to denote the selection probability given to each individual agent with vector $w$, i.e., for all $i \in N : w(i) = w$, $q_w(N) = \pi_i$. We summarize these selection probabilities in the vector $\boldsymbol{q}(N)$.

Reformulation of optimization problem. We reformulate our optimization problem in terms of the variables $q_w$ here, for both objectives. First our feasible set of values of $q_w | w \in \mathcal{W}$, call it $\mathcal{Q}$, is defined by the following constraints, analogs of those defining $\mathcal{R}$: $\boldsymbol{q} \in \mathcal{Q} \iff \boldsymbol{q}$ satisfies

$$\sum_{w : w_f = v} \nu_w(N) q_w(N) = p_{(f,v)} \cdot k/n \text{ for all } (f, v) \in FV \quad \wedge \quad \sum_{w} \nu_w(N) q_w(N) = k/n \quad \wedge \quad \boldsymbol{q}(N) \in [0,1]^{|\mathcal{W}|}. \tag{D.1}$$

Now, to defining our full optimization problems: first, recalling that Leximin is just a refinement of maximin,

$$\text{maximin}(p, k, N) : \qquad \max_{q \in [0,1]^{\mathcal{W}_N}} \min_{w \in \mathcal{W}_N} q_w(N) \quad \text{s.t.} \quad \boldsymbol{q} \in \mathcal{Q}.$$

For Nash , we equivalently analyze the log of the geometric mean, whose optimizer is the same as that of the geometric mean:

$$\text{Nash}(p, k, N) : \qquad \min_{q \in [0,1]^{\mathcal{W}_N}} \sum_{w \in \mathcal{W}_N} -\nu_w(N) \log(q_w(N)) \quad \text{s.t.} \quad \boldsymbol{q} \in \mathcal{Q}.$$

**Instance.** Fix a $\delta \in [1, k/2)$. All four claims will be proven via the same class of instances (parameterized by $\delta$), which has two features $F = \{f_1, f_2\}$ with binary values in $\{0, 1\}$, and as such, $\mathcal{W}$ contains the feature vectors 00, 01, 10, 11. Now, to define this instance $p, k, N$: let the population rates be $p_{f_1,0} = p_{f_2,0} = 1/2$. Let $k \geq 2$. Fix a pool $N$ of size $n \geq k^2$ where $n$ is a multiple of both $\delta/(2k)$ and $1 - \delta/k$, with the following composition: $\nu_{00} = \nu_{11} = \nu^* = \delta/(2k)$, $\nu_{10} = 1 - 2\nu^*$, and $\nu_{01} = 0$. We note that in order for there to be enough people to fill the panel, $\nu^* \geq k/(2n)$.

**Optimal selection probabilities in instance (with true pool).** Now, we characterize the Leximin and Nash -optimal selection probabilities in this instance with the true pool. They are simple, because they are essentially determined by the constraints: notice that for any $n$, the constraints require $q_{10}(N) = 0$, because any probability mass added to $\nu_{(10)}$ will induce imbalance in the amount of probability given to $f_1 = 0$ versus $f_2 = 0$, a violation of the constraints that $\rho_{f_1=0} = \rho_{f_2=0} = 1/2$ that cannot be counteracted because the complementary vector (01) does

not exist in the pool. Also, because vectors 00 and 11 are completely symmetric in the instance, they must receive identical selection probabilities. Thus, $q_{00}(N) = q_{11}(N)$; by the adding up constraint, we have that $v^* q_{00}(N) + v^* q_{11}(N) = k/n$, implying that $q_{00}(N) = q_{11}(N) = 1/(2v^*) \cdot k/n$. To recap, for any $n$,

$$q_{00}^{\text{LEXIMIN}}(N) = q_{00}^{\text{NASH}}(N) = q_{11}^{\text{LEXIMIN}}(N) = q_{11}^{\text{NASH}}(N) = 1/(2v^*) \cdot k/n, \quad q_{10}^{\text{LEXIMIN}}(N) = q_{10}^{\text{NASH}}(N) = 0. \tag{D.2}$$

**Defining the manipulated pool.** Now, define the following manipulating coalition $C$ of size $c = k/2 - \delta$ such that $w(i) = 10$ for all $i \in C$ — that is, all agents in the coalition will have true vector 10. They will also all misreport the same vector 01, so $\tilde{w}(i) = 01$ for all $i \in C$. We define the resulting manipulated pool as $\tilde{N} := N_{-C} \cup (\tilde{w}(i) | i \in C)$. Then, in the corresponding $\tilde{v}$, we have that $\tilde{v}_{00} = \tilde{v}_{11} = v^*$, $\tilde{v}_{10} = 1 - 2v^* - c/n$, and $\tilde{v}_{01} = c/n$.

**Optimal selection probabilities in the manipulated pool.** Before analyzing any specific objective, we reduce the constraints to be in terms of a single selection probability $q_{01}$. Beginning with the raw constraints (where all probabilities $q_v$ here are implicitly $q_v(\tilde{N})$, the probabilities in the manipulated pool):

$$\begin{aligned} v^*\, q_{00} + c/n\, q_{01} &= 1/2 \cdot k/n \\ v^*\, q_{00} + (1 - 2v^* - c/n)\, q_{10} &= 1/2 \cdot k/n \\ v^*\, q_{00} + c/n\, q_{01} + (1 - 2v^* - c/n)\, q_{10} + v^*\, q_{11} &= k/n \end{aligned}$$

This system of 3 linear equations and 4 unknowns simplifies to the following expressions, where all the selection probabilities are in terms of $q_{01}$:

$$q_{00} = q_{11} = \frac{1/2 \cdot k/n - c/n q_{01}}{v^*} \qquad \text{and} \qquad q_{10} = \frac{1/2 \cdot k/n - (1/2 \cdot k/n - c/n q_{01})}{1 - 2v^* - c/n} = \frac{c/n \cdot q_{01}}{1 - 2v^* - c/n}.$$

**Handling box constraints.** Above, we expressed all agents' selection probabilities in terms of $q_{01}$. Now, we will show that for all $q_{01} \in [0, 1]$, all agents' selection probabilities fall between $[0, 1]$ for the parameter settings above. First, this is trivially true for $q_{01}$. For $q_{00} = q_{11}$, we have that

$$q_{00} = q_{11} = \frac{1/2 \cdot k/n - c/n q_{01}}{v^*} = \frac{k - 2(k/2 - \delta) q_{01}}{2n \cdot \delta/(2k)} = \frac{k(k - (k - 2\delta) q_{01})}{n\delta}$$

Bounding this above and below for all $q_{01} \in [0, 1]$:

$$0 \leq \frac{2k}{n} = \frac{k(k - (k - 2\delta))}{n\delta} \leq \frac{k(k - (k - 2\delta) q_{01})}{n\delta} \leq \frac{k^2}{n\delta} \leq 1.$$

Finally, for $q_{10}$,

$$\frac{c/n}{1 - 2v^* - c/n} \cdot q_{01} = \frac{k/2 - \delta}{n(1 - 2\delta/(2k)) - (k/2 - \delta)} = \frac{k/2 - \delta}{n(1 - \delta/k) - (k/2 - \delta)}$$

Bounding this above and below for all $q_{01} \in [0,1]$ (and assuming $k \geq 2$, as is always the case in real panels):

$$0 \leq \frac{k/2-\delta}{n(1-\delta/k)} \leq \frac{k/2-\delta}{n(1-\delta/k)-(k/2-\delta)} \leq \frac{k/2}{n(1-1/2)-k/2} = \frac{k}{n-k} \leq \frac{k}{k^2-k} = \frac{1}{k-1} \leq 1.$$

Now, we've shown that in this instance, the constraints $q_{00} \in [0,1]$, $q_{11} \in [0,1]$, and $q_{10} \in [0,1]$ in Equation (D.1) will never bind. This means that we have reduced the problem to a single-variable problem of the following form:

$$\min_{q_{01}} g(q_{01}) \quad \text{such that } q_{01} \in [0,1].$$

We now compute the optimizer of this program below for both $g = \textsc{Leximin}$ and $g = \textsc{Nash}$, showing that in either case, the optimizer sets $q_{01} = 1$.

**Analysis of Leximin.** Leximin has only one degree of freedom $q_{01}$, so it will maximize the minimum selection probability, i.e., it will set $q_{01}$ to maximize the following expression:

$$\min\left\{q_{01}, \frac{c/n \cdot q_{01}}{1-2v^*-c/n}, \frac{1/2 \cdot k/n - c/nq_{01}}{v^*}\right\} \tag{D.3}$$

We will show that the second term in this minimum is the smallest over the entire domain of $q_{01}$. First, comparing the second term to the first term in (D.3), we use that $c \leq k/2$ to show that

$$q_{01} \geq \frac{c/n \cdot q_{01}}{1-2v^*-c/n} \iff 1-2v^*-c/n \geq c/n \impliedby 1-2v^* \geq k/n.$$

Plugging in our parameters, we deduce that $1-2v^* = 1-2\delta/(2k) \geq 1-1/2 = 1/2 \geq k/n$, as needed.

Next, comparing the second term to the third term in (D.3), we deduce that

$$\begin{aligned}
\frac{1/2 \cdot k/n - c/nq_{01}}{v^*} \geq \frac{c/n \cdot q_{01}}{1-2v^*-c/n} &\iff (1/2 \cdot k/n - c/n \cdot q_{01})(1-2v^*-c/n) \geq c/nq_{01} \cdot v^* \\
&\iff k \geq \frac{2c \cdot q_{01}(v^*+1-2v^*-c/n)}{1-2v^*-c/n} \\
&\iff k \geq \frac{2c \cdot q_{01}(1-v^*-c/n)}{1-2v^*-c/n} \qquad \text{(b)}
\end{aligned}$$

Observe that if (b) holds for $q_{01} = 1$, it holds for all $q_{01} \in [0,1]$. Thus, setting $q_{01} = 1$, we deduce

the bound in reverse:

$$
\begin{aligned}
k \geq \frac{2c(1-v^*-c/n)}{1-2v^*-c/n} &\iff k \geq \frac{2(k/2-\delta)(1-\delta/(2k)-(k/2-\delta)/n)}{1-2\delta/(2k)-(k/2-\delta)/n} \\
&\iff k(1-\delta/k-(k-2\delta)/(2n)) \geq (k-2\delta)(1-\delta/(2k)-(k-2\delta)/(2n)) \\
&\iff k-\delta-\frac{k(k-2\delta)}{2n} \geq k-\delta/2-\frac{k(k-2\delta)}{2n}-2\delta+\delta^2/k+2\delta\frac{k-2\delta}{2n} \\
&\iff -\delta \geq \delta/2-2\delta+\delta^2/k+\delta\frac{k-2\delta}{n} \\
&\iff \delta \leq -\delta/2+2\delta-\delta^2/k-\delta\frac{k-2\delta}{n}
\end{aligned}
$$

Using that $k-2\delta > 0$ and $n \geq 2k$,

$$
\begin{aligned}
&\impliedby \delta \leq -\delta/2+2\delta-\delta^2/k-\delta\frac{k-2\delta}{2k} \\
&\iff \delta \leq \delta-\delta^2/k+\delta^2/k \\
&\iff \delta \leq \delta.
\end{aligned}
$$

Then, we have that (b) is true for all $q_{01} \in [0,1]$.

We have shown that the second term of the minimum in (D.3) is the smallest term over the entire support $q_{01} \in [0,1]$. Because this term is increasing in $q_{01}$, the Leximin optimal solution will maximize this term by setting $q_{01} = 1$.

We conclude that in this instance, $q_{01}^{\text{Leximin}}(\tilde{N}) = 1$. That is, on the manipulated pool, Leximin will give all agents in the manipulating coalition probability 1. Given that by Equation (D.2), $q_{10}^{\text{Leximin}}(N) = 0$ and $w(i) = 10$ for all $i \in C$, it follows that for any $i \in C$, $\pi_i^{\text{Leximin}}(\tilde{N}) - \pi_i^{\text{Leximin}}(N) = 1-0 = 1$.

Moreover, we've shown this for any size coalition $c \in [1, k/2)$. Setting $c = 1$ (corresponding setting to $\delta = k/2-(k/2-1)$), this implies that $\text{manip}_{int}(N, \text{Leximin}, 1) = 1$. For generic $\delta$, we conclude that $\text{manip}_{comp}(N, \text{Leximin}, k/2-\delta) = k/2-\delta$.

**Analysis of Nash.** Repeating the same analysis for Nash, the function Nash maximizes in this instance is

$$
\sum_w v_w \log(q_w) = 2v^* \log\left(\frac{1/2 \cdot k/n - c/n q_{01}}{v^*}\right) + (1-2v^*-c/n)\log\left(\frac{c/n}{1-2v^*-c/n} \cdot q_{01}\right) + c/n \log(q_{01})
$$

This function is concave in $q_{01}$, so it has a unique maximizer that can be found by the first-order condition: Thus, taking the derivative with respect to $q_{01}$ and setting it to zero, we get that this function is maximized when

$$
2v^* \cdot \frac{v^*}{1/2 \cdot k/n - c/n \cdot q_{01}} \cdot \frac{-c}{nv^*} + (1-2v^*-c/n) \cdot \frac{1-2v^*-c/n}{c/n \cdot q_{01}} \cdot \frac{c/n}{1-2v^*-c/n} + \frac{c}{n \cdot q_{01}} = 0.
$$

Dividing both sides by $c/n$ and making cancellations,

$$\begin{aligned}
&\iff \frac{-2v^*}{1/2 \cdot k/n - c/n \cdot q_{01}} + \frac{(1 - 2v^* - c/n)/(c/n)}{q_{01}} + \frac{1}{q_{01}} = 0 \\
&\iff \frac{-2v^*}{1/2 \cdot k/n - c/n \cdot q_{01}} + \frac{1 - 2v^*}{c/n \cdot q_{01}} = 0 \\
&\iff q_{01} = \frac{k(1 - 2v^*)}{2c}
\end{aligned}$$

Plugging in our values for $v^*, c$,

$$\begin{aligned}
&\iff q_{01} = \frac{k(1 - 2\delta/(2k))}{2(k/2 - \delta)} \\
&\iff q_{01} = \frac{k - \delta}{k - 2\delta} > 1
\end{aligned}$$

Of course, we have deduced that the unconstrained optimizer places $q_{01} > 0$. By the concavity of the objective, we know that the optimizer is then at $q_{01} = 1$, at the edge of the box constraint.

We conclude that in this instance, $q_{01}^{\text{NASH}}(\tilde{N}) = 1$ — that is, on the manipulated pool, NASH will give all agents in the manipulating coalition probability 1. Given that by Equation (D.2), $q_{10}^{\text{NASH}}(N) = 0$ and $w(i) = 10$ for all $i \in C$, for all $i \in C$, $\pi_i^{\text{NASH}}(\tilde{N}) - \pi_i^{\text{NASH}}(N) = 1 - 0 = 1$.

Moreover, we've shown this for any size coalition $c \in [1, k/2)$. Setting $c = 1$ (corresponding setting to $\delta = k/2 - (k/2 - 1)$), this implies that $\text{manip}_{int}(N, \text{NASH}, 1) = 1$. For generic $\delta$, we conclude that $\text{manip}_{comp}(N, \text{NASH}, k/2 - \delta) = k/2 - \delta$. □

### D.2.2 Proof of Proposition D.2.1

Let $k/n\mathbf{1}$ be the $n$-length vector whose entries are all $k/n$.

**Proposition D.2.1.** *Let $g$ be any strongly convex (with parameter $m$) that, when unconstrained, is minimized at at $k/n\mathbf{1}$, the point where all agents' selection probabilities are equalized. Let $\pi$ be a set of marginal probabilities with $q = \max_i \pi_i$. Then,*

$$g(\pi) \geq m/2|q - k/n|^2.$$

*Proof.* Then, by the definition of strong convexity,

$$\begin{aligned}
g(\pi) - g(k/n\mathbf{1}) + \nabla g(k/n\mathbf{1})^T(\pi - k/n\mathbf{1}) \geq m/2\|\pi - k/n\|^2 &= m/2 \sum_i (\pi_i - k/n)^2 \\
&\geq m/2|q - k/n|^2
\end{aligned}$$

Noting that $\nabla g(k/n\mathbf{1})^T = 0$,

$$\begin{aligned}
&\iff g(\pi) - g(k/n\mathbf{1}) \geq m/2|q - k/n|^2. \\
&\implies g(\pi) \geq m/2|q - k/n|^2.
\end{aligned}$$

□

## D.3 Supplemental materials from Section 5.4

### D.3.1 Practical justification of Assumption 5.4.1

What we need is a a set $\mathcal{W}^*$ of feature vectors within the pool such that each group $w \in \mathcal{W}^*$ grows *linearly in n* (up to the size of the total population) and that this set of vectors is sufficient to permit a feasible solution on their own. We cannot test this assumption directly in our data, since we only see one realized value of $n$. Thus, we base our discussion here on the statistical properties of the random pool recruitment process. Examining this process, we actually expect something stronger to be true, at least in expectation: *every vector group* to grow linearly in $n$, up to variance, which we will discuss at the end. This (expected) linear growth is due to how the pool is sampled: invitation recipients are uniformly selected from the population, so at least the *expected* pool composition, over the randomness of the invitation process, should be roughly constant in $n$ (i.e., *all groups* grow linearly in $n$). We formalize this intuition below with a simple model of the pool formation process, which will also help us more precisely discuss the role of variance.

To model the pool formation process with minimal assumptions, let $Y$ be the entire underlying population, Let $\mathcal{W}_Y$ be the set of all unique feature vectors in the population, $\gamma_w$ be the fraction of the population with feature vector $w$, and $q_i$ be the probability that each $i \in Y$ decides to participate conditional on being invited. Let $\bar{q}_w : \frac{1}{|\{i:w_i=w\}|} \sum_{i:w_i=w} q_i$ be the average rate of participation among population members with vector $w$. Then, in the process of sampling the pool $N$ (with corresponding $\boldsymbol{\nu}(N)$), there are two stages of randomness: that of inviting recipients, and their decision of whether to participate. Regardless of the size of the pool $N$, $\mathbb{E}[\nu_w(N)] = \gamma_w \bar{q}_w$ for all $w \in \mathcal{W}_Y$ — that is, in expectation, *all vector groups* in the pool are growing linearly in $n$ (and moreover, the randomness in this process consists of Bernoulli draws, so the pool composition should be concentrating around its expectation as $n$ gets large). Variance in this process could be in the $q_i$ values of agents with vector $w$ relative to $\bar{q}_w$; variance in the sampling of who receives letters; and variance in the Bernoulli draws by which people decide whether to participate. Based on this process, variance will mainly be a problem for ensuring linear growth among very small groups, particularly when $n$ is small.

The potential effects of variance in small groups, especially at practical sample sizes, is precisely the motivation for proving our results under Assumption 5.4.1 — a much weaker requirement than the assumption that *all* groups are growing in $n$. Under this assumption, we need only that *there some set of vectors yielding a feasible solution* growing linearly in $n$, rather than *all* vector groups in the pool. For our assumption to be violated, there would need to be *no such set of feature vectors*, corresponding to the unlikely case that *any possible set of feature vectors* supporting a feasible solution contains a group composing only a sliver of the population.

### D.3.2 Proof of

**Proposition D.3.1.** *The example used to prove Theorem 5.3.1 satisfies Assumption 5.4.1.*

*Proof.* The example in Theorem 5.3.1 satisfies Assumption 5.4.1 by setting $\mathcal{W}^* = \{00, 11\}$ and

$\kappa^* = v^* - k/n$, where $v^* > 0$ is a constant that we choose, and $k/n \to 0$ as $n$ grows large. For part (2) of Assumption 5.4.1, we can easily see that a feasible solution exists over $\mathcal{W}^* = \{00, 11\}$: spread all the probability equally among the agents with these vectors. For part (1) of Assumption 5.4.1, we need to verify that $\kappa^* = v^* - k/n > 0$. The only tricky part is that, as the coalition size approaches $k/2$, we need $v^*$ to get smaller. This is not a problem, though; we just need $n$ to be larger for the assumption to be satisfied, which is fine because Theorem 5.3.1 needs to hold just for very large $n$.

□

### D.3.3 Proof of Theorem 5.4.3

*Proof.* Fix an instance with one binary feature with values 0 and 1; let $p_{f,v_1} = 1/2$; then we know that the total probability given to agents with vector 0 and 1 is $1/2k$. Now, suppose $v_0 = 7/8$ and $v_1 = 1/8$. The probabilities are then

$$\pi_0 = \frac{1/2k}{7n/8} = \frac{4}{7}k/n, \qquad \pi_1 = \frac{1/2k}{n/8} = 4k/n.$$

We will deal with the largest coalition size $c = 5n/64$ only; the argument for all smaller coalition sizes follows in the same way. We assume that this coalition defects from vector 0 to vector 1. Then, the resulting probability for vector 1 is

$$\tilde{\pi}_1 = \frac{1/2k}{n/8 + c} = \frac{4k}{n + 8c} = 4k/n - \frac{4k * 8c}{n(n + 8c)} \geq 4k/n - \frac{32kc}{n^2}$$

Now, we characterize the three types of manipulability. Within the coalition, all members receive probability $\tilde{\pi}_1$ when before they received $\pi_0$, so

$$\mathsf{manip}_{int}(N, \mathcal{A}, c) \geq \tilde{\pi}_1 - \pi_0 \geq 4k/n - \frac{32kc}{n^2} - \frac{4}{7}k/n = \frac{k}{n}\left(\frac{24}{7} - \frac{32c}{n}\right) \geq \frac{k}{n}(3 - 2.5) = 1/2 \cdot k/n.$$

By joining group 1, the coalition decreases the existing members' probabilities by making their group more numerous:

$$\begin{aligned}\mathsf{manip}_{ext}(N, \mathcal{A}, c) \geq \pi_1 - \tilde{\pi}_1 = 4k/n - \left(4k/n - \frac{4k * 8c}{n(n + 8c)}\right) = \frac{4k * 8c}{n(n + 8c)} \geq \frac{32kc}{8n(n + c)} &= \frac{4k \cdot 5n/64}{n(n + 5n/64)} \\ &= \frac{5/16 \cdot k}{n(1 + 5/64)} \\ &= 20/69 \cdot k/n.\end{aligned}$$

Then, because there are $c$ true 0s impersonating 1s, the true seats given to 0 is, in expectation, $k/2$ (the number of seats that must be given to them in expectation, based on the perceived pool), plus however many seats 1-impersonators get in expectation:

$$\mathsf{manip}_{comp}(N, \mathcal{A}, c) \geq (k/2 + c \cdot \tilde{\pi}_1) - k/2 = c \cdot \tilde{\pi}_1 = c\left(4k/n - \frac{4k * 8c}{n(n + 8c)}\right) \geq 3ck/n.$$

Set $\eta = k \cdot 20/69$, and the proof is complete. □

## D.4 Supplemental materials from Section 5.5

### D.4.1 Panel selection instances

| Instance | Organization | $n$ | $k$ | # unique vectors | # features | $\Delta$ |
|---|---|---|---|---|---|---|
| sf(a) | Sortition Foundation | 312 | 35 | 182 | 6 | 6.08 |
| sf(b) | Sortition Foundation | 250 | 20 | 92 | 6 | 11.78 |
| sf(c) | Sortition Foundation | 161 | 44 | 92 | 7 | 3.18 |
| sf(d) | Sortition Foundation | 404 | 40 | 108 | 6 | 8.02 |
| sf(e) | Sortition Foundation | 1727 | 110 | 762 | 7 | 15.28 |
| cca | Center for Blue Democracy | 825 | 75 | 554 | 4 | 10.56 |
| hd | Healthy Democracy | 239 | 30 | 202 | 7 | 3.54 |
| newd | New Democracy | 398 | 40 | 173 | 6 | 4.16 |

Table D.1: Overview of real-world instances. $\Delta$ is a measure of the self-selection bias in the instance, as defined as Section 5.5.1.

### D.4.2 Additional instances for Figure 5.1

Below in Figure D.1 we present plots for all 6 other instances, corresponding to those in Figure 5.1. An interesting aspect of these results results: For instances cca and sf(d), the strategy *MU* is harmful for nearly all agents in the pool under all three algorithms. This is promising for practitioners; although deviating to the feature vector with the most underrepresented feature values is a strategy that is most likely to be used in practice, cca and sf(d) serve as counterexamples where Leximin and Nash are not arbitrarily manipulable against *MU*.

### D.4.3 Self-selection bias experiments: methods

Here, we describe the details of the experiments used to produce plots Figure 5.2(b) and Figure 5.2(c), and correspondingly, those in Appendix D.4.4. For both SSB by *interpolating* and SSB by *feature dropping*, we define the precise sequence of instances we test, and then prove that over the sequences of instances induced by either approach, $\Delta_{p,k,N}$ is decreasing (Claim D.4.1 for interpolation, and Claim D.4.2.

SSB by *interpolating* (corresponding to Figure 5.2(b)) Here, we studied how our selection algorithms performed against the OPT-1 strategy over a sequence of pools with decreasing SSB. The pools in this sequence are different convex combinations of two pools: $N$ (the original pool) and pool $N'$, defined as the solution of the convex program below, which finds the pool "closest" (by Euclidean distance) to $N$ that has SSB $\Delta_{p,k,N} = 0$.

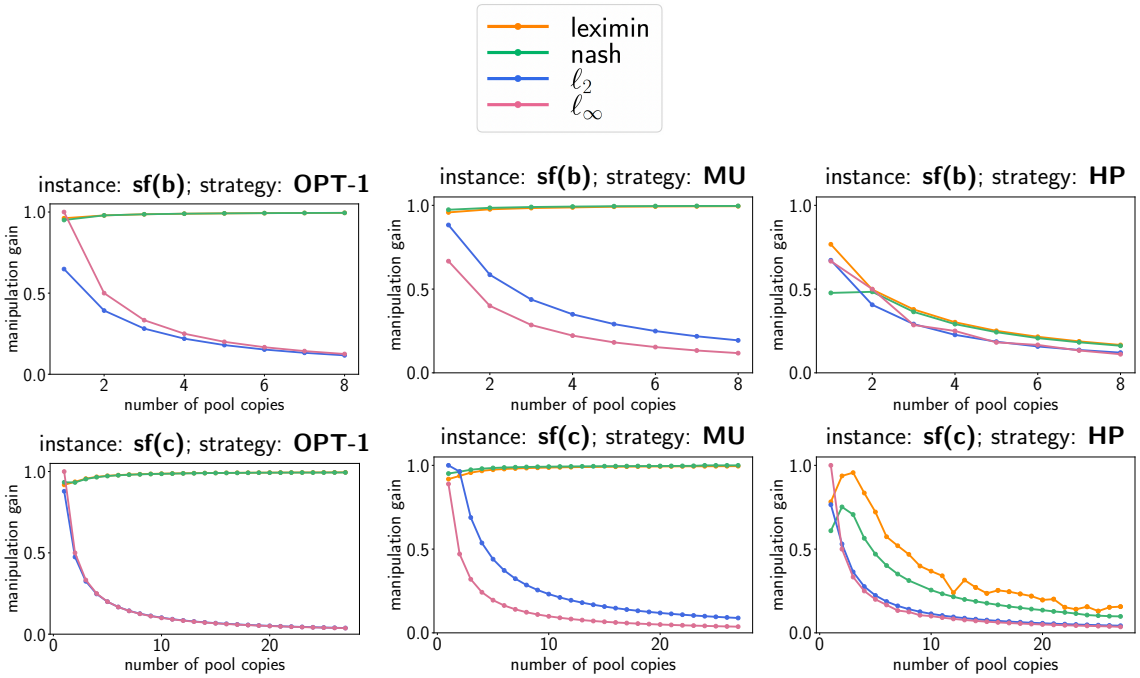
$$N' := \arg\max_{N'':|N''|=n} \|\boldsymbol{\nu}(N) - \boldsymbol{\nu}(N'')\|_2 \quad \text{s.t.} \quad \sum_{w:w_f=v} \nu_w(N'') = p_{(f,v)} \;\text{ for all } (f,v) \in FV.$$

Now, define the sequence of pools $N_0, N_1, \ldots, N_{10}$ in which $N_\ell$ is defined such that $|N_\ell| = n$ and

$$\nu_w(N_\ell) = (1 - \ell/10) \cdot \nu_w(N) + \ell/10 \cdot \nu_w(N') \qquad \text{for all } w \in \mathcal{W}.$$

In Figure 5.2(b), the $\alpha$ on the $x$ axis is then the interpolation weight, ranging over $\alpha = (\ell/10)_{\ell\in[10]}$. More formally, across the $x$ axis, we're testing the sequence of instances $p, k, N_0,\ p, k, N_1, \ldots, p, k, N_{10}$.

**Claim D.4.1.** *$\Delta_{p,k,N}$ is weakly decreasing over the sequence of instances $p, k, N_0, \ldots, p, k, N_{10}$.*



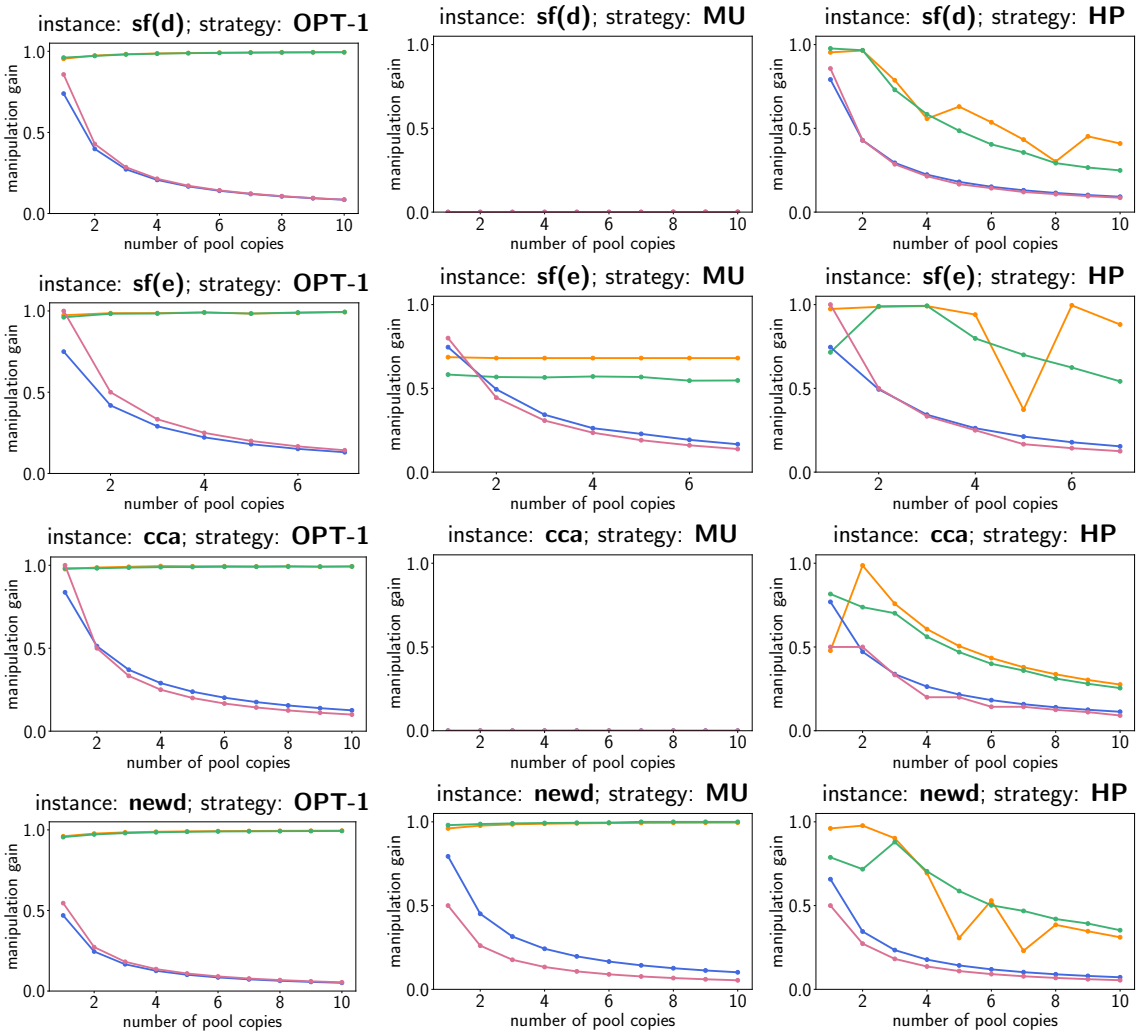



Figure D.1: Figures for remaining instances from analysis in Figure 5.1

.

*Proof.* Partition the feature-values $FV$ into three exhaustive subsets:

$$FV^{under}(N) := \left\{(f,v) : \frac{p_{(f,v)}}{\eta_{(f,v)}(N)} > 1\right\},\ FV^{over}(N) := \left\{(f,v) : \frac{p_{(f,v)}}{\eta_{(f,v)}(N)} < 1\right\},$$

$$\text{and } FV^{exact}(N) := \left\{(f,v) : \frac{p_{(f,v)}}{\eta_{(f,v)}(N)} = 1\right\}.$$

Observe that for all $(f,v)$, by the constraints defining $N'$, $\eta_{f,v}(N') = p_{(f,v)}$. Then we have that

$$\eta_{f,v}(N_\ell) = (1 - \ell/10)\,\eta_{(f,v)}(N) + \ell/10\eta_{(f,v)}(N') = (1 - \ell/10)\,\eta_{(f,v)}(N) + \ell/10 p_{(f,v)}$$

We can see from this expression that $(f,v) \in FV^{under}(N) \implies (f,v) \in FV^{under}(N_\ell)$ for all $\ell < 10$, and likewise for $FV^{under}, FV^{exact}$.

Now, let $\ell' > \ell$ for $\ell \in 0 \ldots 9$. We have that for all $(f, v) \in FV^{over}(N)$,

$$\frac{p_{f,v}}{\eta_{(f,v)}(N_\ell)} > \frac{p_{f,v}}{\eta_{(f,v)}(N_{\ell'})}$$

This is seen by the fact that for all $(f, v) \in FV$, the following quantity is decreasing And similarly, for all $(f, v) \in FV^{under}(N)$,

$$\frac{p_{f,v}}{\eta_{(f,v)}(N_\ell)} < \frac{p_{f,v}}{\eta_{(f,v)}(N_{\ell'})}.$$

Observing that if $FV^{under}(N)$ is non-empty (and thus $FV^{over}(N)$ is also non-empty) the feature values that yield the max and min terms in $\Delta_{p,k,N}$ must come from $FV^{under}(N)$ and $FV^{over}(N)$, respectively. Therefore, the difference between the max and the min must be decreasing, and $\Delta_{p,k,N_0}, \Delta_{p,k,N_1}, \ldots, \Delta_{p,k,N_{10}}$ is decreasing. □

SSB BY *FEATURE DROPPING.* In a fixed instance, we define the self-selection bias of a single feature according to $\Delta_{p,k,N}$ *restricted to the values of* $f$, or formally, as

$$\Delta^f_{p,k,N} := \max_{v \in V_f} p_{(f,v)} \,/\, \eta_{(f,v)}(N) - \min_{v \in V_f} p_{(f,v)} \,/\, \eta_{(f,v)}(N).$$

Now, let the features be ordered in decreasing order of their self-selection bias, so $\Delta^{f_1}_{p,k,N} \geq \Delta^{f_2}_{p,k,N} \geq \cdots \geq \Delta^{f_{|F|}}_{p,k,N}$. We will decrease the self-selection bias by successively drop features from the problem in this order.

When we "drop" a feature $f$ out of the problem, we are formally dropping constraints $\sum_{i:f(i)=v} \pi_i = k p_{(f,v)}$ for all $v \in V_f$ from Equation (OPT-PROB) Accordingly, dropping features corresponds to changing the instance $p, k, N$ by dropping entries of $p$. Formally, express $p = (p_{(f,v)})_{f \in F, v \in V_f}$. Now, we define a sequence of $p_1, \ldots, p_{|F|}$ where $p_\ell := (p_{(f,v)})_{f \in \{f_\ell \ldots f_{|F|}\}, v \in V_f}$. Then, across the $x$ axis of Figure 5.2(c) (and all corresponding figures for other instances), we are testing how our selection algorithms perform against the OPT-1 strategy over the sequence instances $p_1, k, N,\ p_2, k, N,\ \ldots p_{|F|}, k, N$.

**Claim D.4.2.** *$\Delta_{p,k,N}$ is weakly decreasing over the sequence of instances $p_1, k, N, \ldots p_{|F|}, k, N$.*

*Proof.* Dropping constraints $f, v$ out of $FV$ can only decrease $\max_{(f,v) \in FV} \frac{p_{(f,v)}}{\eta_{(f,v)}(N)}$ and increase $\min_{(f,v) \in FV} \frac{p_{(f,v)}}{\eta_{(f,v)}(N)}$. □

### D.4.4 Self-selection bias experiments: supplemental empirical results

Additional instances for Figure 5.2(b). Figure D.2 shows tests decreasing self-selection bias by *interpolation* for all remaining instances.

Additional instances for Figure 5.2(c). Figure D.3 shows tests decreasing self-selection bias by *feature dropping* for all remaining instances.

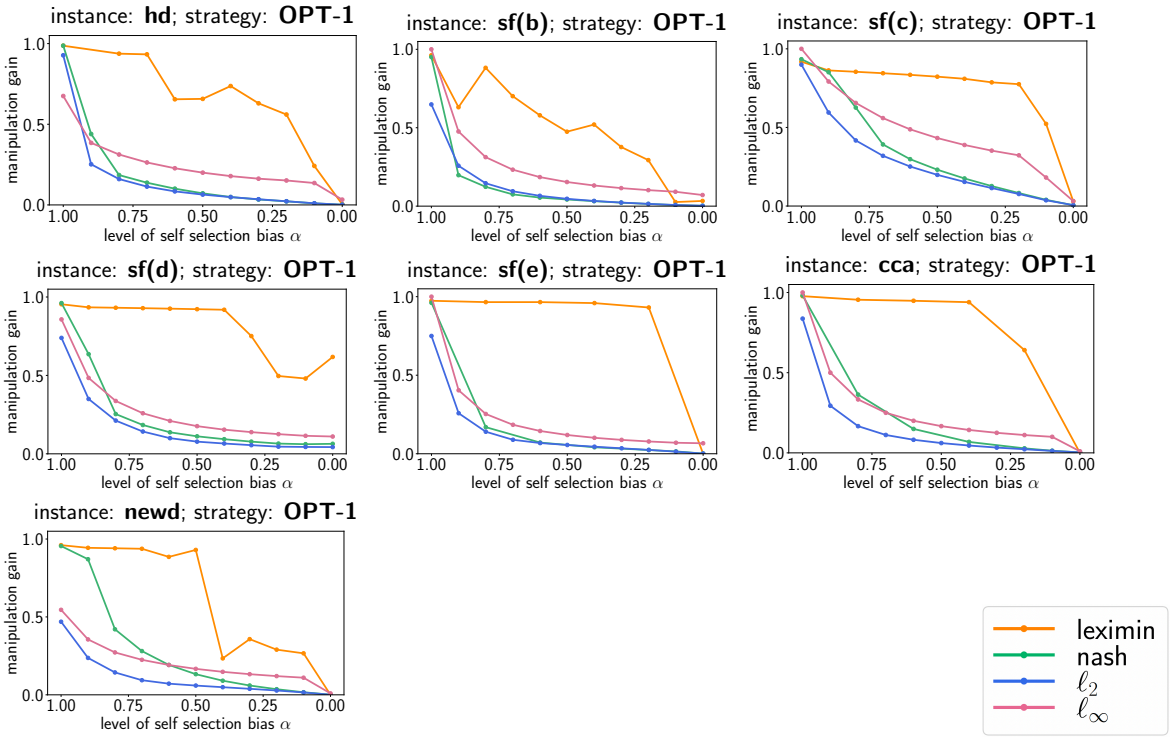



Figure D.2: Figures for remaining instances from analysis in Figure 5.2(b)

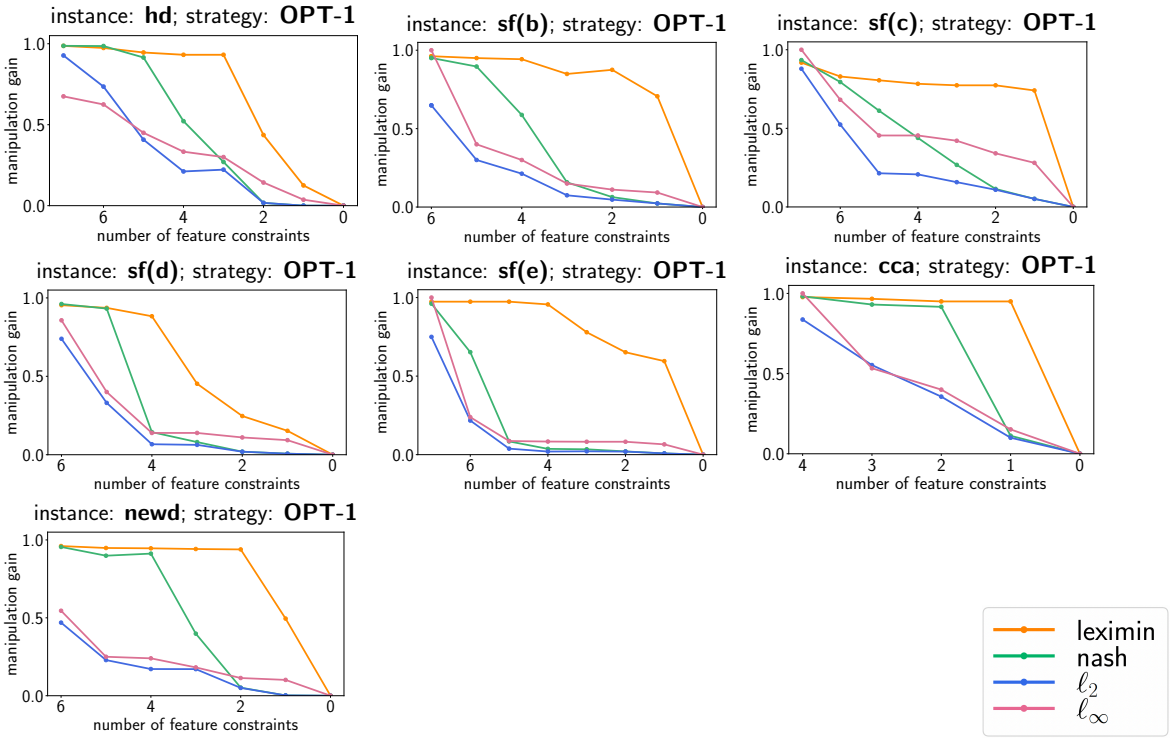



Figure D.3: Figures for remaining instances from analysis in Figure 5.2(c)

### D.4.5 Empirical minimum selection probabilities given by norms

| Minimum probability | sf(a) | sf(b) | sf(c) | sf(d) | sf(e) | sf(hd) | sf(newd) | sf(cca) |
|---|---|---|---|---|---|---|---|---|
| $\ell_2$ | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| $\ell_\infty$ | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

Table D.2: Minimum selection probability given to any agent by $\ell_2, \ell_\infty$ across instances

# E

# Chapter 6 Appendix

## E.1 Supplemental Materials for Section 6.2

Below, we show that all of our stated equality objectives in Section 6.2 are convex and satisfy conditional equitability and anonymity.

**Proposition E.1.1.** *Maximin, Minimax, Nash, Linear$_\gamma$, and Goldilocks$_\gamma$ are all convex.*

*Proof.* The convexity of *Maximin* and *Minimax* follows immediately from their definition: min is concave, so $-\min$ is convex, and max is a convex function. Geometric mean is known to be concave, and as we define the *Nash* objective to be the negative geometric mean, it is convex. *Linear*$_\gamma$ is the sum of two convex functions: max and $-\gamma \min$ for some $\gamma \geq 0$, hence it is also convex. Finally for *Goldilocks*$_\gamma$, we can rewrite the second term as $\frac{\gamma \max(1/\boldsymbol{\pi})}{n/k}$. $1/\boldsymbol{\pi}$ is convex as all entries of $\boldsymbol{\pi}$ are nonnegative, and max is convex and increasing. Hence the composition of these two functions is convex. Therefore, *Goldilocks*$_\gamma$ is the sum of two convex functions, and is itself convex. □

**Proposition E.1.2.** *Maximin, Minimax, Nash, Linear$_\gamma$, and Goldilocks$_\gamma$ are all conditionally equitable.*

*Proof.* We will simply lower bound each objective function for any $\boldsymbol{\pi} \in \Pi(\mathcal{I})$ and then show that $k/n\mathbf{1}^n$ achieves this bound. This will imply that $k/n\mathbf{1}^n \in \Pi^{\mathcal{E}}(\mathcal{I})$. Fix any solution $\pi \in \Pi(\mathcal{I})$. We know that $\max(\pi) \geq k/n$ and $\min(\pi) \leq k/n$— otherwise $\sum_{i\in[n]} \pi_i \neq k$. Hence, for any feasible solution:

$$\begin{aligned}
\mathit{Maximin}(\boldsymbol{\pi}) &\geq -k/n \\
\mathit{Minimax}(\boldsymbol{\pi}) &\geq k/n \\
\mathit{Goldilocks}_\gamma(\boldsymbol{\pi}) &\geq n/k \cdot k/n + \frac{\gamma}{n/k \cdot k/n} = 1 + \gamma \\
\mathit{Linear}_\gamma(\boldsymbol{\pi}) &\geq k/n - \gamma \cdot k/n = k/n \cdot (1-\gamma)
\end{aligned}$$

Each of these lower bounds are realized by the solution $k/n\mathbf{1}$. For *Nash* we use the AM-GM inequality as follows:

$$\mathit{Nash}(\boldsymbol{\pi}) = -\left(\Pi_{i\in[n]}\pi_i\right)^{1/n} \geq \frac{-1}{n}\sum_{i\in[n]}\pi_i = \frac{-k}{n}$$

Again, this lower bound is realized by the solution $k/n\mathbf{1}$. □

We transfer the following claim about anonymity from Flanigan et al. [131] as it is relevant to the structure of our final proposition proof and is of independent use in later proofs in the appendix.

**Claim E.1.3** ([131] Claim B.6)**.** *For any instance $\mathcal{I}$ and any realizable $\boldsymbol{\pi}$, let $\boldsymbol{\pi}'$ be the "anonymized" marginals obtained by setting $\pi'_i$ to the average $\boldsymbol{\pi}_j$ across all $j$ such that $w(j) = w(i)$. Then $\boldsymbol{\pi}'$ is realizable as well.*

**Proposition E.1.4** (Adapted from [131] Claim B.6)**.** *Maximin, Minimax, Nash, Linear$_\gamma$, and Goldilocks$_\gamma$ are all anonymous.*

*Proof.* Fix some instance $\mathcal{I}$, and $\mathcal{E} \in \{Maximin, Minimax, Nash, Linear_\gamma, \text{ and } Goldilocks_\gamma\}$. By Assumption 6.2.3, we have that $\mathcal{I}$ is feasible – hence $\Pi^{\mathcal{E}}(\mathcal{I})$ is nonempty. Now we will use a similar proof as to that of Claim E.1.3, but will pay attention to the impact of incrementally anonymizing the panel distribution on the equality objective.

Assume for sake of contradiction that there is no anonymous $\boldsymbol{\pi}$ such that $\boldsymbol{\pi} \in \Pi^{\mathcal{E}}(\mathcal{I})$. Let $\boldsymbol{\pi}$ be the most anonymized optimal vector of marginals, and $\boldsymbol{d}$ be the corresponding panel distribution inducing it. Formally:

$$\boldsymbol{\pi} = \arg \min_{\boldsymbol{\pi} \in \Pi^{\mathcal{E}}(\mathcal{I})} \max_{w \in \mathcal{W}_N} \left( \max_{i \in [n]\,:\, w(i)=w} \pi_i - \min_{i \in [n]\,:\, w(i)=w} \pi_i \right)$$

There must be a finite number of pairs of marginals that are maximizing this gap. We argue that we can equalize these pairs one-by-one without affecting other marginals, while never increasing $\mathcal{E}$. Let $i, j \in [n]$ be such that $w(i) = w(j)$ and they have the maximum gap between any marginals of the same feature-vector in $\boldsymbol{\pi}$. Without loss of generality, assume $\boldsymbol{\pi}_i > \boldsymbol{\pi}_j$. We construct a new panel distribution $\boldsymbol{d'}$ as follows: $\hat{\boldsymbol{d'}}$ is identical to $\hat{\boldsymbol{d}}$ except that it swaps $i$ for $j$ and vice versa on all panels. Then define $\boldsymbol{d''} = (\boldsymbol{d} + \boldsymbol{d'})/2$. Let $\boldsymbol{\pi''}$ be the marginals resulting from $\boldsymbol{d''}$. We have that $\pi''_i = \pi''_j$ and all other marginals remain the same. Now we consider how our equality objective, $\mathcal{E}$ might be impacted.

Notice that $\min(\boldsymbol{\pi}) \leq \pi_j < \pi''_j = \pi''_i < \pi_i \leq \max(\boldsymbol{\pi})$. Therefore, as all other marginals remain unchanged, we have that $\min(\boldsymbol{\pi}) \leq \min(\boldsymbol{\pi''})$ and $\max(\boldsymbol{\pi}) \geq \max(\boldsymbol{\pi''})$. Therefore:

$$\begin{aligned}
Maximin(\boldsymbol{\pi''}) &= -\min(\boldsymbol{\pi''}) \leq -\min(\boldsymbol{\pi}) \leq Maximin(\boldsymbol{\pi}) \\
Minimax(\boldsymbol{\pi''}) &= \max(\boldsymbol{\pi''}) \leq \max(\boldsymbol{\pi}) \leq Minimax(\boldsymbol{\pi}) \\
Linear_\gamma(\boldsymbol{\pi''}) &= \max(\boldsymbol{\pi''}) - \gamma \min(\boldsymbol{\pi''}) \leq \max(\boldsymbol{\pi}) - \gamma \min(\boldsymbol{\pi}) \leq Linear_\gamma(\boldsymbol{\pi}) \\
Goldilocks_\gamma(\boldsymbol{\pi''}) &= n/k \max(\boldsymbol{\pi''}) - \gamma \cdot \frac{1}{n/k \min(\boldsymbol{\pi''})} \leq n/k \max(\boldsymbol{\pi}) - \gamma \cdot \frac{1}{n/k \min(\boldsymbol{\pi})} \leq Goldilocks_\gamma(\boldsymbol{\pi})
\end{aligned}$$

Finally, we just consider the case of $Nash$:

$$Nash(\boldsymbol{\pi''}) = -\left(\Pi_{a \in [n]} \pi''_a\right)^{1/n} = -\left(\Pi_{a \in [n]} \pi_a\right)^{1/n} \cdot \left(\frac{\pi''_i \pi''_j}{\pi_i \pi_j}\right)^{1/n}$$

We have that $\pi''_i \pi''_j = \left(\frac{\pi_i + \pi_j}{2}\right)^2 = \frac{\pi_i^2 + 2\pi_i\pi_j + \pi_j^2}{4}$. We know that $(\pi_i - \pi_j)^2 \geq 0$ which implies that $\pi_i^2 + \pi_j^2 \geq 2\pi_i\pi_j$. So this gives us that $\pi''_i \pi''_j \geq \frac{4\pi_i\pi_j}{4} = \pi_i\pi_j$. Returning to our analysis of $Nash$, we see that we are multiplying the negative geometric mean of $\boldsymbol{\pi}$ by a value greater than or equal to 1. So we have that:

$$Nash(\boldsymbol{\pi''}) \leq -\left(\Pi_{a \in [n]} \pi_a\right)^{1/n} \leq Nash(\boldsymbol{\pi})$$

□

Then we have shown that for all equality objectives we're considering, $\mathcal{E}(\boldsymbol{\pi''}) \leq \mathcal{E}(\boldsymbol{\pi})$. Therefore, after repeating this adjustment for all of the finitely many pairs enforcing this maximum gap, we will have arrived at a *more* anonymized vector of marginals that has objective value at most $\boldsymbol{\pi}$. This is a contradiction to $\boldsymbol{\pi}$ being the most anonymized vector of marginals in $\Pi^{\mathcal{E}}(\mathcal{I})$. Therefore, we have arrived at a contradiction and can conclude that there exists an anonymous $\pi \in \Pi^{\mathcal{E}}(\mathcal{I})$.

## E.2 Supplemental Materials for Section 6.3

### E.2.1 Proof of Theorem 6.3.1

*Proof.* Our original instance is $\mathcal{I}^{=}$, meaning that by Observation 6.3.2, $\pi_i^{\mathrm{E}}(\mathcal{I}^{=}) = k/n$ for all $\mathrm{E} \in \{\textsc{Maximin}, \textsc{Leximin}, \textsc{Nash}\}$. After the coalition $C$ of size $c$, as constructed in the body, misreports, our resulting instance is $\mathcal{I}_c^*$ with associated valid panels $\mathcal{K}_c^*$. By Observation 6.3.3, for any $\mathbf{d} \in \Delta(\mathcal{K}_c^*)$ with corresponding probabilities on our two panel types $d_1$, $d_2$,

$$\mathsf{p}_{00} = \mathsf{p}_{11} = d_1 \frac{k/2}{(n-c)/2} + d_2 \frac{k/2-1}{(n-c)/2}, \qquad \mathsf{p}_{10} = d_2 \frac{1}{c-1}, \qquad \mathsf{p}_{01} = d_2. \tag{E.1}$$

Noting that $d_1 = (1 - d_2)$ and simplifying, we get that

$$\mathsf{p}_{00} = \mathsf{p}_{11} = \frac{k/2-d_2}{(n-c)/2}, \qquad \mathsf{p}_{10} = d_2 \frac{1}{c-1}, \qquad \mathsf{p}_{01} = d_2. \tag{E.2}$$

Now, we will argue that the optimal panel distributions for Maximin (and Leximin) and Nash will place at least probability $(c-1)k/n$ probability exclusively on panels of type 2, resulting in giving $i^*$ selection probability 1. Note that this will prove the requisite lower bounds on $\mathsf{manip}_{int}$ and $\mathsf{manip}_{comp}$: before misreporting, $i^*$ received selection probability $k/n$ after misreporting they will have gained selection probability $(c-1)k/n$, which can be driven up to 1 for a linear-size coalition. It will follow that $\mathsf{manip}_{int}$ is at least 1 for all algorithms considered.

Maximin (and Leximin). $\mathbf{d}^{\textsc{Maximin}}$ maximizes the minimum selection probability subject to the constraints in Equation (E.2) and that $d_2 \in [0,1]$. We now rewrite the objective by plugging in the constraints:

$$\min\left\{\mathsf{p}_{00}, \mathsf{p}_{11}, \mathsf{p}_{10}, \mathsf{p}_{01}\right\} = \min\left\{\frac{k/2-d_2}{(n-c)/2},\ d_2\frac{1}{c-1},\ d_2\right\}.$$

Clearly, $\mathsf{p}_{01}$ cannot be the minimum probability for any $d_2 \in [0,1]$. Thus, the minimum term in this objective must be between the first two. By arithmetic, we can derive that

$$\frac{k/2-d_2}{(n-c)/2} \geq d_2\frac{1}{c-1} \iff d_2 = \frac{k(c-1)}{n+c-2}.$$

At $d_2 = \frac{k(c-1)}{n+c-2}$, $\mathsf{p}_{00} = \mathsf{p}_{11} = \mathsf{p}_{10} = \frac{k}{n+c-2}$. $\mathsf{p}_{00} = \mathsf{p}_{11}$ are decreasing in $d_2$, so we if we increase $d_2$ from here, $\mathsf{p}_{00}$ will decrease and the minimum will decrease, making the objective value worse; $\mathsf{p}_{10}$ is increasing in $d_2$, so if we decrease $d_2$ from here, $\mathsf{p}_{10}$ will decrease, making the objective value worse. We conclude that the maximin-optimal solution sets $d_2 = \frac{k(c-1)}{n+c-2}$, meaning

$$\mathsf{p}_{01} = \frac{k(c-1)}{n+c-2} \geq \frac{k(c-1)}{2n}.$$

For a coalition of size $2n/k + 1$, $\mathsf{p}_{01} = 1$ (we can set $k$ relative to $n$ so that this coalition is at most size $n/2$, which is sufficient).

By misreporting their vector as part of coalition $C$, $i^*$ went from probability $k/n$ to probability 1, meaning that they gained probability at least $1 - k/n$, giving us the requisite lower bound on $\mathsf{manip}_{int}$ for Maximin. Because Leximin will first optimize Maximin and never decrease a selection probability in its later steps, this lower bound also applies to Leximin.

Nash. We prove this claim in much the same way. We will instead equivalently optimize the logarithm of Nash, defined as $\sum_{i\in[n]} \log(\pi_i)$. The same constraints hold, but we are instead optimizing the *product* of probabilities, so $\mathbf{d}^{\text{Nash}}$ optimizes the following objective, noting the symmetry between $\mathsf{p}_{00}$ and $\mathsf{p}_{11}$:

$$(n-c)\log(\mathsf{p}_{00}) + (c-1)\log(\mathsf{p}_{10}) + \log(\mathsf{p}_{01})$$

which, by the constraints is equal to

$$(n-c)\log\left(\frac{k/2-d_2}{(n-c)/2}\right) + (c-1)\log\left(d_2\frac{1}{c-1}\right) + \log(d_2). \tag{E.3}$$

Differentiating Equation (E.3) with respect to $d_2$, we get

$$\frac{\partial \text{Equation (E.3)}}{\partial d_2} = -\frac{2d_2n - ck}{d_2(k-2d_2)}$$

This derivative is 0 at $d_2 = ck/(2n)$. By the concavity of the original objective function, this value of $d_2$ must be its unique global optimizer. It follows that at the Nash-optimal solution,

$$\mathsf{p}_{01} = \min\left\{\frac{ck}{2n}, 1\right\}.$$

For a coalition of size $2n/k + 1$, this selection probability is 1. By misreporting their vector, $i^*$ has gained probability $1 - k/n$, giving us the requisite lower bound on $\mathsf{manip}_{int}$ for Nash. □

## E.3 Supplemental Materials for Section 6.3

### E.3.1 Proof of Theorem 6.3.4 2

**Truthful instance.** All truthful instances we consider in this proof will have two binary features: $F = (f_1, f_2)$ with $V_{f_1} = V_{f_2} = \{0, 1\}$, so $\mathcal{W} = \{00, 11, 01, 10\}$. Our truthful pool $N$ will have the following composition: $N_{00} = N_{11} = n/3$, $N_{01} = N_{10} = n/6$. Our truthful instance will have quotas that depend on the case:

- $0 \leq \gamma < 1$: $\mathcal{I}$ will have quotas $\ell_{f_1,0} = u_{f_1,0} = 2k/3$ and $\ell_{f_2,0} = u_{f_2,0} = k/3$.
- $1 \leq \gamma < n/3-1$ and $\gamma \geq n/3-1$: $\mathcal{I}'$ will have quotas $\ell_{f_1,0} = u_{f_1,0} = k/2$ and $\ell_{f_2,0} = u_{f_2,0} = k/2$.

**Coalitions.** What coalitions deviate from our truthful instance also depends on the case.

- $0 \leq \gamma < 1$: There is no coalition; in this case, we directly analyze $\mathcal{I}$, because we are analyzing the outcome of fairness, which is measured in the absence of any manipulation.
- $1 \leq \gamma < n/3 - 1$ and $\gamma < n/3 - 1$: We let the pool in $\mathcal{I}'$ be $N'$, and let $C \subseteq N'$ be of size $n/6$, where for all $i \in C, w(i) = 01$. Let $n/6 - 1$ agents misreport $\tilde{w}(i) = 10$, and let one agent $i^*$ still report vector $\tilde{w}(i^*) = 01$.

**Manipulated Instances.** The pools resulting from these coalitional manipulations are

- $0 \leq \gamma < 1$: Not applicable
- $1 \leq \gamma < n/3 - 1$ and $\gamma < n/3 - 1$: Call the manipulated pool $\tilde{N}'$ and the manipulated instance $\tilde{\mathcal{I}}'$. $\tilde{N}'_{00} = \tilde{N}'_{11} = n/3$, $\tilde{N}'_{10} = n/3 - 1$, and $\tilde{N}'_{01} = 1$.

Now, we handle each case separately. For convenience, we will first analyze the Case 1 instance in the relaxation of our setting studied in Flanigan et al. [135], where they study the same panel selection task but permit agents to be *divisible*. We call this setting the *continuous setting*. Formally speaking, in instance $\mathcal{I} = (N, k, \boldsymbol{\ell}, \boldsymbol{u})$ such that $\boldsymbol{\ell}, \boldsymbol{u})$, the set of feasible selection probability assignments over which $\mathcal{E}(\boldsymbol{\pi})$ could be optimized was

$$P(\mathcal{I}) = \left\{ \boldsymbol{\pi} : \boldsymbol{\pi} \in [0,1]^n \ \wedge \ \sum_{i \in [n]} \pi_i = k \ \wedge \ \sum_{i \in [n]: f(i) = v} \pi_i = k u_{f,v} \ \forall f, v \in FV \right\}.$$

We analogously define $P^{\mathcal{E}}(\mathcal{I}) := \arg\inf_{\pi \in P(\mathcal{I})} \mathcal{E}(\mathcal{I})$ as the set of $\mathcal{E}$-optimal selection probability assignments over the feasible space $P(\mathcal{I})$. Now we proceed with giving constructions in the continuous setting. At the end, we will prove a general method for translating lower bounds in the continuous setting to our setting.

**Case 1:** $0 \leq \gamma < 1$.

**Claim E.3.1.** *$\mathcal{I}$ satisfies Assumption 6.2.4.*

*Proof.* In $\mathcal{I}$, we have that $\ell_{f_1,0} = u_{f_1,0} = 2k/3$ and $\ell_{f_2,0} = u_{f_2,0} = k/3$ and $N_{00} = N_{11} = n/3$, $N_{01} = N_{10} = n/6$. First, observe that all groups are growing in $n$, meaning that we can set $n, k$ so that for some $\kappa^* > 0$, for all $w$, $N_w \geq \kappa^* n + k$.

Next, observe that we can place all vectors on a valid panel: consider all panels containing $k/2$ agents with vector 01, and $k/6$ agents of each remaining vector. Panels of this composition satisfy the quotas, and all agents can be contained on such a panel. □

**Claim E.3.2.** *For all* $\gamma \in [0, 1)$, $\min(\boldsymbol{\pi}^{\text{LINEAR}_\gamma}) = 0$.

*Proof.* Note that we only need one quota constraint per feature because the constraint for one value implies the constraint for the other. Our constraints are then:

$$\mathrm{p}_{01} \cdot \frac{n}{6} + \mathrm{p}_{10} \cdot \frac{n}{6} + \mathrm{p}_{00} \cdot \frac{n}{3} + \mathrm{p}_{11} \cdot \frac{n}{3} = k \tag{E.4}$$

$$\mathrm{p}_{01} \cdot \frac{n}{6} + \mathrm{p}_{00} \cdot \frac{n}{3} = \frac{2k}{3} \tag{E.5}$$

$$\mathrm{p}_{01} \cdot \frac{n}{6} + \mathrm{p}_{11} \cdot \frac{n}{3} = \frac{2k}{3} \tag{E.6}$$

**Showing that $\mathbf{p}_{10} = 0$ in the optimizer.** By constraints E.5 and E.6 we see that $\mathrm{p}_{00} = \mathrm{p}_{11} = k/n - \mathrm{p}_{10}/2$. Plugging this back into constraint E.4, we can solve for $\mathrm{p}_{01}$ and get that

$$\mathrm{p}_{01} = \frac{2k}{n} + \mathrm{p}_{10} \qquad \text{and} \qquad \mathrm{p}_{00} = \mathrm{p}_{11} = \frac{k}{n} - \frac{\mathrm{p}_{10}}{2}.$$

*Reducing the box constraints to constraints on $p_{10}$*: if $\mathrm{p}_{10}$ is close enough to 0 (or 0), clearly all probabilities are in $[0, 1]$.

Now, observe that $\mathrm{p}_{01}$ must be larger than both $\mathrm{p}_{10}$ and $\mathrm{p}_{00}$, so it is the maximum marginal. Then, our objective is the following:

$$\max\left\{\frac{2k}{n} + \mathrm{p}_{10} - \gamma \mathrm{p}_{10},\ \frac{2k}{n} + \mathrm{p}_{10} - \gamma\left(\frac{k}{n} - \frac{\mathrm{p}_{10}}{2}\right)\right\}.$$

where the two terms in the maximum account for either $\mathrm{p}_{10}$ or $\mathrm{p}_{00} = \mathrm{p}_{11}$ being the minimum marginal probability. By the fact that $0 \leq \gamma < 1$, both of these terms are increasing in $\mathrm{p}_{10}$. Thus, this is minimized when $\mathrm{p}_{10} = 0$.

We claim that this instance can be translated to our panel distribution setting. Fix the same $\mathcal{I}$ and require without loss of generality that that $n$ is divisible by 6 and $k$ is divisible by 3. We take the same definition of $N, \ell$, and $u$. First we observe that any $\pi \in \Pi(\mathcal{I})$ is also in $P(\mathcal{I})$. Intuitively this is because $P(\mathcal{I})$ is defined by a relaxation of the constraints defining $\Pi(\mathcal{I})$. This was shown formally in Appendix A.2 of Flanigan et al. [135]. From above, we know that $\boldsymbol{\pi}^* \in P^{\mathcal{E}}(I)$ is of the following form: $\mathrm{p}_{10}(\boldsymbol{\pi}^*) = 0, \mathrm{p}_{01}(\boldsymbol{\pi}^*) = \frac{2k}{n}, \mathrm{p}_{00}(\boldsymbol{\pi}^*) = \mathrm{p}_{11}(\boldsymbol{\pi}^*) = \frac{k}{n}$. We first construct a panel distribution, $\boldsymbol{d}$, to assign the same total probability to each feature vector group as $\boldsymbol{\pi}^*$. Let $K$ be a

panel populated with $k/3$ agents of type 01, $k/3$ agents of type 11 and, $k/3$ agents of type 00, and let $\boldsymbol{d}_K = 1$. By Claim E.1.3, we know that we can construct a new *anonymous* panel distribution $\boldsymbol{d'}$ with the same total probability assigned to each feature vector. Let $\boldsymbol{\pi'} = \boldsymbol{\pi}(\boldsymbol{d'})$. Therefore, we have that $\mathrm{p}_w(\boldsymbol{\pi'}) = \frac{\sum_{i:\, w(i)=w} \pi_i(\boldsymbol{d})}{N_{w(i)}}$ for all $i \in [n]$. Solving this gives us:

$$\mathrm{p}_{10}(\boldsymbol{\pi'}) = \frac{\sum_{i:\, w(i)=10} \pi_i(\boldsymbol{d})}{N_{w(i)}} = 0 \qquad \mathrm{p}_{01}(\boldsymbol{\pi'}) = \frac{\sum_{i:\, w(i)=01} \pi_i(\boldsymbol{d})}{N_{w(i)}} = \frac{k/3}{n/6} = \frac{2k}{n}$$

$$\mathrm{p}_{00}(\boldsymbol{\pi'}) = \frac{\sum_{i:\, w(i)=00} \pi_i(\boldsymbol{d})}{N_{w(i)}} = \frac{k/3}{n/3} = \frac{k}{n} \qquad \mathrm{p}_{11}(\boldsymbol{\pi'}) = \frac{\sum_{i:\, w(i)=11} \pi_i(\boldsymbol{d})}{N_{w(i)}} = \frac{k/3}{n/3} = \frac{k}{n}$$

Therefore, we can observe that $\mathrm{p}(\boldsymbol{\pi'}) = \mathrm{p}(\boldsymbol{\pi^*})$. Hence, because both $\boldsymbol{\pi'}$ and $\boldsymbol{\pi^*}$ are anonymous and on the same instance, we get that $\boldsymbol{\pi'} = \boldsymbol{\pi^*}$. As $\Pi(\mathcal{I}) \subseteq P(\mathcal{I})$, and $P^{\mathcal{E}}(\mathcal{I}) = \{\boldsymbol{\pi^*}\}$, we know that $\Pi^{\mathcal{E}}(\mathcal{I}) = \{\boldsymbol{\pi^*}\}$ as well — there cannot be any other optimal marginals, otherwise they would be in $P^{\mathcal{E}}(\mathcal{I})$ as well. Therefore, we also know that in the panel setting, optimizing $\textsc{Linear}_\gamma$ will set $\mathrm{p}_{10} = 0$. □

**Cases 2 and 3:** $1 \le \gamma < n/3 - 1$ and $\gamma \ge n/3 - 1$.

**Claim E.3.3.** *$\mathcal{I}'$ satisfies Assumption 6.2.4.*

*Proof.* In $\mathcal{I}''$, we have that $\ell_{f_1,0} = u_{f_1,0} = k/2$ and $\ell_{f_2,0} = u_{f_2,0} = k/2$ and $N_{00} = N_{11} = n/3$, $N_{01} = N_{10} = n/6$. First, observe that all groups are growing in $n$, meaning that we can set $n, k$ so that for some $\kappa^* > 0$, for all $w$, $N_w \ge \kappa^* n + k$.

Next, observe that we can place all vectors on a valid panel: consider all panels containing $k/3$ agents with 00, $k/3$ agents with 11, $k/6$ agents with 01, and $k/6$ agents with 10. Panels of this composition satisfy the quotas, and all agents can be contained on such a panel. □

**Claim E.3.4.** *For all $\gamma \in [1, n/3 - 1)$,*

$$p_{10}\left(\boldsymbol{\pi}^{\textsc{Linear}_\gamma}(\tilde{\mathcal{I}}')\right) = \frac{9k}{2(n^2 - 9)}$$

*and for all $\gamma \ge n/3 - 1$,*

$$p_{01}\left(\boldsymbol{\pi}^{\textsc{Linear}_\gamma}(\tilde{\mathcal{I}}')\right) = 1.$$

*Proof.* Take the instance $\tilde{\mathcal{I}}'$. Following Observation 6.3.3, notice that this is instance $\mathcal{I}^*_{n/3}$. Fix any $\mathbf{d} \in \Delta(\mathcal{K}^*_{n/3})$, and let $d_1, d_2$ represent the total probability $\mathbf{d}$ places on panels of Types 1 and 2, respectively. Then, by simply dividing the expected panel seats given to agents with each vector $w$ divided by the total number of pool members with vector $w$, the resulting selection probabilities (assumed to be anonymous) are:

$$\mathrm{p}_{00} = \mathrm{p}_{11} = d_1 \frac{k/2}{(n - n/3)/2} + d_2 \frac{k/2 - 1}{(n - n/3)/2}, \qquad \mathrm{p}_{10} = d_2 \frac{1}{n/3 - 1}, \qquad \mathrm{p}_{01} = d_2. \tag{E.7}$$

Using that $d_1 + d_2 = 1$ and simplifying, we get that

$$\mathsf{p}_{00} = \mathsf{p}_{11} = \frac{k/2 - d_2}{n/3}, \qquad \mathsf{p}_{10} = d_2 \frac{1}{n/3 - 1}, \qquad \mathsf{p}_{01} = d_2. \tag{E.8}$$

Now, we make some observations:

- $\mathsf{p}_{01} \geq \mathsf{p}_{10}$ for all $d_2 \in [0, 1]$
- $\mathsf{p}_{01} \geq \mathsf{p}_{00} \iff d_2 \geq \frac{3k}{6+2n}$
- $\mathsf{p}_{00} \geq \mathsf{p}_{10} \iff d_2 \leq \frac{k(n-3)}{4n-6)}$

$\textsc{Linear}_\gamma$ is in terms of the maximum and minimum. This gives us three cases for the values of the minimum and maximum probability:

**Case 1:** $\mathsf{p}_{00} \geq \mathsf{p}_{01} \geq \mathsf{p}_{10}$, which occurs when $d_2 \leq \frac{3k}{6+2n}$. Here,

$$\textsc{Linear}_\gamma = \mathsf{p}_{00} - \mathsf{p}_{10} = \frac{k/2 - d_2}{n/3} - \gamma \frac{d_2}{n/3 - 1} = \frac{3k}{2n} - d_2 \left(1 + \frac{\gamma}{n/3 - 1}\right).$$

For all $\gamma$, this objective is decreasing in $d_2$, meaning it is optimized when $d_2$ is maximized over the relevant domain. Then, the optimal solution over this domain is $d_2 = \frac{3k}{6+2n}$, at which point $\mathsf{p}_{00} = \mathsf{p}_{01}$, which means that this case at the optimizer over this domain is interchangeable with Case 3.

**Case 2:** $\mathsf{p}_{01} \geq \mathsf{p}_{10} \geq \mathsf{p}_{00}$, which occurs when $d_2 \geq \frac{k(n-3)}{4n-6}$. Here,

$$\textsc{Linear}_\gamma = \mathsf{p}_{01} - \mathsf{p}_{00} = d_2 - \gamma \frac{k/2 - d_2}{n/3} = d_2 \left(\frac{\gamma}{n/3} + 1\right) - \frac{\gamma 3k}{2n}.$$

For all $\gamma$, this objective is increasing in $d_2$, meaning it is optimized when $d_2$ is minimized over the relevant domain. Then, $d_2 = \frac{k(n-3)}{4n-6}$. Again, at at this point $\mathsf{p}_{00} = \mathsf{p}_{10}$, which means that this case at the optimizer over this domain is interchangeable with Case 3.

**Case 3:** $\mathsf{p}_{01} \geq \mathsf{p}_{00} \geq \mathsf{p}_{10}$, which occurs when $\frac{3k}{6+2n} \leq d_2 \leq \frac{k(n-3)}{4n-6)}$. Here,

$$\textsc{Linear}_\gamma = \mathsf{p}_{01} - \mathsf{p}_{10} = d_2 - \gamma \frac{d_2}{n/3 - 1} = d_2 \left(1 - \frac{\gamma}{n/3 - 1}\right).$$

When $\gamma < n/3 - 1$, this objective function is increasing in $d_2$, meaning that it is minimized by minimizing $d_2$ over the relevant domain; therefore, at the $\textsc{Linear}_\gamma$ optimizer, $d_2 = \frac{3k}{6+2n}$. It follows that

$$\mathsf{p}_{10} = \frac{3k}{(6 + 2n)(n/3 - 1)} = \frac{9k}{2(n^2 - 9)}.$$

When $\gamma \geq n/3 - 1$, this objective function is (weakly) decreasing in $d_2$, meaning that it is minimized by maximizing $d_2$ over the relevant domain; therefore, $d_2 = \min\{\frac{k(n-3)}{4n-6)}, 1\}$. It follows that

$$\mathsf{p}_{01} = \min\left\{\frac{k(n-3)}{2(n-3)}, 1\right\} = \min\{k/2, 1\}. \qquad \square$$

## E.4 Supplemental Materials for Section 6.4

### E.4.1 Proof of Lemma 6.4.2 for general $\gamma$

**Lemma E.4.1.** *Fix an instance $\mathcal{I}$, and suppose there exists a feasible solution $\pi \in \Pi(\mathcal{I})$ with associated $\delta_{below}, \delta_{above}$. Then, for all $i \in [n]$,*

$$\pi^{\textsc{Goldilocks}_\gamma}(\mathcal{I})_i \in \left[(2\max\{\delta_{below}(\pi), \delta_{above}(\pi)/\gamma\})^{-1}\, k/n,\ 2\max\{\gamma\, \delta_{below}(\pi), \delta_{above}(\pi)\}\, k/n\right].$$

*Proof.* Fix instance $\mathcal{I}$ and corresponding feasible solution $\pi$ with associated $\delta_{below}(\pi)$, $\delta_{above}(\pi)$, as given in the statement. We will use shorthand $\pi^* = \pi^{\textsc{Goldilocks}_\gamma}(\mathcal{I})$ and $\delta_{below}(\pi) = \delta_{below}$,$\delta_{above}(\pi) = \delta_{above}$.

First, we upper bound the optimal objective value using our feasible solution $\pi$:

$$\begin{aligned}\textsc{Goldilocks}_\gamma(\pi^*) \le n/k\ \max(\pi) + \frac{\gamma}{n/k\ \min(\pi)} &= \delta_{above} + \gamma\, \delta_{below} \\ &\le 2\max\{\gamma\, \delta_{below}, \delta_{above}\}. \qquad \text{(E.9)}\end{aligned}$$

Now, suppose there exists $i \in [n]$ such that $\pi^*_i > k/n \cdot 2\max\{\gamma\delta_{below}, \delta_{above}\}$. Then,

$$\textsc{Goldilocks}_\gamma(\pi^*) > n/k \cdot k/n \cdot 2\max\{\gamma\delta_{below}, \delta_{above}\} + 0 = 2\max\{\gamma\delta_{below}, \delta_{above}\},$$

which is a contradiction to (E.9). We conclude that $\pi^*_i \le k/n \cdot 2\max\{\gamma\delta_{below}, \delta_{above}\}$ for all $i \in [n]$.

Likewise, suppose that there exists $i \in [n]$ such that $\pi^*_i < k/n \cdot (2\max\{\delta_{below}, \delta_{above}/\gamma\})^{-1}$. Then,

$$\textsc{Goldilocks}_\gamma(\pi^*) > 0 + \frac{\gamma}{n/k \cdot k/n \cdot (2\max\{\delta_{below}, \delta_{above}/\gamma\})^{-1}} = 2\max\{\gamma\delta_{below}, \delta_{above}\}.$$

This is again a contradiction to (E.9), and we conclude that $\pi^*_i \ge k/n \cdot (2\max\{\delta_{below}, \delta_{above}/\gamma\})^{-1}$ for all $i \in [n]$, concluding the claim. □

### E.4.2 Notation and preliminaries for proofs of Lemma 6.4.3 and Lemma 6.4.4

In this proof, we will work exclusively with feature-vector indexed objects, which treat individuals with the same feature vector as interchangeable (this is without loss of generality because, by Proposition E.1.4, all objectives we consider are anonymous). To begin, we will define these objects, which collapse all individuals of the same feature vector.

***Pool and panel compositions***: For panel $K$, we let its panel *composition* $\mathsf{K}(K) \in [0,1]^{|\mathcal{W}|}$ describe the frequencies of each feature vector on a panel, with $w$-th entry

$$\mathsf{K}_w(K) = \frac{|\{i : i \in K \wedge w(i) = w\}|}{|K|} \quad \text{and} \quad \mathsf{K}(K) := (\mathsf{K}_w(K) | w \in \mathcal{W}).$$

We say that $\mathsf{K}$ *contains* vector $w$ iff $\mathsf{K}_w > 0$.

We define a ***pool composition*** $\mathsf{N}(N) \in [0,1]^{|\mathcal{W}|}$ analogously, so the pool composition of $N$ is given by

$$\mathsf{N}(N) := (\mathsf{N}_w(N) | w \in \mathcal{W}) \text{ where } \mathsf{N}_w(N) = \frac{|\{i : i \in N \wedge w(i) = w\}|}{|N|}.$$

When $N$ or $K$ is clear from context, or when referring to an arbitrary pool or panel composition, we will simply use $\mathsf{K}$ or $\mathsf{N}$ respectively.

Let the **set of *valid* panel compositions** be

$$\mathfrak{K}(\mathcal{K}) := \{\mathsf{K}(K) | K \in \mathcal{K}\}.$$

When $\mathcal{K}$ is clear, we will shorten this to $\mathfrak{K}$.

Then, a ***panel composition distribution*** is then any distribution over the set of valid panel compositions; that is, $\mathsf{d} \in \Delta(\mathfrak{K})$.

***Vector-indexed total probabilities*:** Finally, for a given panel composition distribution $\mathsf{d}$ we define the *total probabilities* given to each vector $\mathsf{t}(\mathsf{d}) \in [0,k]^{|\mathcal{W}|}$ as

$$\mathsf{t}_w(\mathsf{d}) := \sum_{\mathsf{K} \in \mathfrak{K}} \mathsf{d}_\mathsf{K} \cdot k \cdot \mathsf{K}_w \qquad \text{and} \qquad \mathsf{t}(\mathsf{d}) = (\mathsf{t}_w(\mathsf{d}) | w \in \mathcal{W}).$$

Notice that we can just as easily define these totals for $\mathsf{p}$ as $\mathsf{t}_w(\mathsf{p}) = N_w \cdot \mathsf{p}_w$ —abusing notation, we will allow this.

Before proceeding, we prove the following two lemmas, which show how to reconstruct a panel distribution from a panel composition distribution while preserving the vector-indexed total probabilities and vice versa.

**Lemma E.4.2.** *Fix a panel composition distribution $d$. We will now show how to construct a corresponding panel distribution $\mathbf{d}$ such that $\boldsymbol{\pi}(\mathbf{d})$ is anonymous with*

$$\pi_i(\mathbf{d}) = \frac{t_w(d)}{N_w} \quad \textit{for all } i : w(i) = w, \textit{ all } w \in \mathcal{W}.$$

*Proof.* Fix $\mathsf{d}$. We will construct $\mathbf{d}$ via the following algorithm.

Initialize $\mathbf{d} \leftarrow \mathbf{0}^{|\mathcal{K}|}$.

For all panel compositions $\mathsf{K} \in \mathfrak{K}$ such that $\mathsf{d}_\mathsf{K} > 0$, do the following:

> Let $W_\mathsf{K} := \{w : \mathsf{K}_w > 0\}$ be the set of all feature vectors contained by $\mathsf{K}$. Then, let $L$ be the least common multiple of $N_w | w \in W_\mathsf{K}$, i.e., the number of people in the pool with each such vector $w$. Now create $L$ panels $K_1^{(\mathsf{K})} \dots K_L^{(\mathsf{K})}$, where all these panels

contain $k \cdot \mathsf{K}_w$ seats reserved for people of vector $w$, for each $w \in W_{\mathsf{K}}$. Populate the seats reserved for vector $w$ on each panel with individuals with vector $w$ round-robin style until all panels of individuals are constructed. Because $L$ is a multiple of $N_w$ for all $w$, each $i$ with vector $w$ will be placed on the same number of panels, and will be placed on a total of $L \cdot k \cdot \mathsf{K}_w / N_w$ panels. Also, note that because $\mathsf{K}$ was a valid panel composition, $K_1^{(\mathsf{K})} \dots K_L^{(\mathsf{K})}$ must be valid panels.

Now, for each panel $K_j \in \{K_1^{(\mathsf{K})} \dots K_L^{(\mathsf{K})}\}$, $d_{K_j^{(\mathsf{K})}} \leftarrow d_{K_j^{(\mathsf{K})}} + \mathsf{d}_{\mathsf{K}}/L$.

Now, it just remains to prove that for all $i$ with $w(i) = w$, we have that $\pi_i(\mathbf{d}) = t_w(\mathsf{d})/N_w$, for all $w \in \mathcal{W}$. Fix such a $w$ and corresponding $i$ with $w(i)$. Then, based on the algorithm above,

$$\pi_i(\mathbf{d}) = \sum_{\mathsf{K}:\mathsf{d}_{\mathsf{K}}>0} L \cdot k \cdot \mathsf{K}_w/N_w \cdot \mathsf{d}_{\mathsf{K}}/L = \sum_{\mathsf{K}\in\mathfrak{K}} k \cdot \mathsf{K}_w \cdot \mathsf{d}_{\mathsf{K}}/N_w = \frac{\mathsf{t}_w(\mathsf{d})}{N_w}. \qquad \square$$

**Lemma E.4.3.** *Given a panel distribution* $\mathbf{d}$*, we will show how to construct a corresponding panel composition distribution* $d$ *such that*

$$t_w(d) = \sum_{i\in[n]\colon w(i)=w} \pi_i(\mathbf{d}) \quad \textit{for all } w \in \mathcal{W}.$$

*Proof.* Fix our panel distribution $\mathbf{d}$. We will essentially just abstract it into a panel composition distribution. Initialize $\mathsf{d} \leftarrow \mathbf{0}^{|\mathfrak{K}|}$.

For all panels $K \in \mathcal{K}$ such that $\mathbf{d}_K > 0$, update $\mathsf{d}$ as follows: $\mathsf{d}_{\mathsf{K}(K)} \leftarrow \mathsf{d}_{\mathsf{K}(K)} + \mathbf{d}_K$. This is clearly a valid distribution because all entries are non-negative and sum to 1 because we simply distribute the probability mass of $\mathbf{d}$ across panel compositions.

Fix some $w \in \mathcal{W}$. Based on the algorithm above, we have that:

$$\begin{aligned} t_w(\mathsf{d}) &= \sum_{\mathsf{K}\in\mathfrak{K}} \mathsf{d}_{\mathsf{K}} \cdot k \cdot \mathsf{K}_w = \sum_{\mathsf{K}\in\mathfrak{K}} \sum_{K\in\mathcal{K}\colon \mathsf{K}(K)=\mathsf{K}} \mathbf{d}_K \cdot k \cdot \mathsf{K}_w = \sum_{K\in\mathcal{K}} \mathbf{d}_K \cdot |\{i\colon i \in K \wedge w(i) = w\}| \\ &= \sum_{K\in\mathcal{K}} \sum_{i\colon i\in K\wedge w(i)=w} \mathbf{d}_K = \sum_{i\in[n]\colon w(i)=w} \sum_{K\in\mathcal{K}\colon i\in K} \mathbf{d}_K = \sum_{i\in[n]\colon w(i)=w} \pi_i(\mathbf{d}) \end{aligned}$$

$\square$

### E.4.3 Proof of Lemma 6.4.3

*Proof.* Fix an instance $\mathcal{I}$ whose pool $N$ satisfies Assumption 6.2.4 with constant $\kappa^*$. We will construct $\mathbf{d}$ by constructing a panel *composition distribution* $\mathsf{d}$, and then transforming it into a panel distribution via Lemma E.4.2. Recall that $\mathcal{W}_N$ represents the unique set of feature vectors in $N$. Note that because $N_w \geq n\kappa^* + k$ for all $w \in \mathcal{W}_N$ (Assumption 6.2.4), there must only exist a constant number of unique vectors in the pool:

$$|\mathcal{W}_N| \leq \frac{n}{n\kappa^* + k} \leq \frac{1}{\kappa^*}. \tag{E.10}$$

For each $w \in \mathcal{W}_N$, by Assumption 6.2.4, there must exist some panel composition $\mathsf{K} \in \mathfrak{K}$ such that $\mathsf{K}_w > 0$. Let $\mathsf{K}^{(w)}$ be this identified panel for each vector $w \in \mathcal{W}_N$ (these panel compositions need not be unique). By Equation (E.10), there are only at most $1/\kappa^*$ of these panels. Define our panel composition distribution $\mathrm{d}$ to as the uniform distribution over these panel compositions; that is;

$$\mathrm{d}_{\mathsf{K}} = \begin{cases} \frac{1}{\left|\bigcup_{w \in \mathcal{W}_N} \{\mathsf{K}^{(w)}\}\right|} & \text{if } \mathsf{K} \in \bigcup_{w \in \mathcal{W}_N} \{\mathsf{K}^{(w)}\} \\ 0 & \text{else.} \end{cases}$$

Then, the total probability given to each feature vector $w \in \mathcal{W}_N$ is bounded as follows, using that $\left|\bigcup_{w \in \mathcal{W}_N} \{\mathsf{K}^{(w)}\}\right| \leq 1/\kappa^*$ (Equation (E.10))

$$k\kappa^* \mathsf{K}_w^{(w)} \leq \frac{k\mathsf{K}_w^{(w)}}{\left|\bigcup_{w \in \mathcal{W}_N} \{\mathsf{K}^{(w)}\}\right|} \leq \mathrm{t}_w(\mathrm{d}) \leq k,$$

Finally, use Lemma E.4.2 to transform $\mathrm{d}$ into a panel distribution $\mathbf{d}$. Using that $\kappa^* n + k \leq N_w \leq n$ for all $w$, we get that

$$\frac{k\kappa^* \mathsf{K}_w^{(w)}}{n} \leq \pi_i(\mathbf{d}) \leq \frac{k}{\kappa^* n + k} \leq \frac{k}{\kappa^* n} \qquad \text{for all } i \in [n].$$

Because we want bounds that reflect the scaling in terms of $k$ and $n$, we just for now treat $k$ as an asymptotic parameter, whose composition remains constant as $k$ scales. Thus, we we treat $\mathsf{K}_w^{(w)}$ as constant in $k$ and $n$, and we get that

$$\pi_i(\mathbf{d}) \in \left[\Omega\left(\frac{k\kappa^*}{n}\right), O\left(\frac{k}{\kappa^* n}\right)\right] \implies \delta(\mathcal{I}) \leq O(1)$$

concluding the proof. □

### E.4.4 Formalization of observation that $\delta(\mathcal{I}_c^*) \geq \sqrt{c-1}$

**Theorem E.4.4** (Lower Bound)**.** *There exists $\mathcal{I}$ satisfying Assumption 6.2.4 such that $k/n\mathbf{1}^n \in \Pi(\mathcal{I})$ (so $\delta(\mathcal{I}) = 1$), but there exists a coalition of size $c$ which, after misreporting $\tilde{\mathbf{w}} \in \mathcal{W}^c$, can create a new instance $\tilde{\mathcal{I}}$ such that $\delta(\tilde{\mathcal{I}}) \geq \sqrt{c-1}$.*

*Proof.* Let our truthful instance be $\mathcal{I}^=$, and recall the observation Observation 6.3.2 that $k/n\mathbf{1} \in \Pi(\mathcal{I})$, meaning that $\delta(\tilde{I}) = 1$. Now, define a coalition $C \subseteq N$ exactly as in the proof of Theorem 6.3.1, so the resulting post-manipulation instance is exactly $\mathcal{I}_c^*$. It follows by Observation 6.3.3 that

$$\mathrm{p}_{10}(\boldsymbol{\pi}) \cdot (c-1) = \mathrm{p}_{01}(\boldsymbol{\pi}) \qquad \text{for all } \boldsymbol{\pi} \in \Pi(\tilde{\mathcal{I}}). \tag{E.11}$$

For any given $\boldsymbol{\pi} \in \Pi(\tilde{\mathcal{I}})$, there are then two options: either $\mathrm{p}_{10}(\boldsymbol{\pi}) = \mathrm{p}_{01}(\boldsymbol{\pi}) = 0$, or $\mathrm{p}_{10}(\boldsymbol{\pi}) > 0$ and $\mathrm{p}_{01}(\boldsymbol{\pi}) > 0$. For any $\boldsymbol{\pi}$ in the former category, $\delta_{below}(\boldsymbol{\pi}) = \infty$, because $\frac{k/n}{0} = \infty$. For any $\boldsymbol{\pi}$ in the latter category, it must be the case that $\max\{\delta_{below}(\boldsymbol{\pi}), \delta_{above}(\boldsymbol{\pi})\} \geq \sqrt{c-1}$. If not, then it would have to be the case that $\max(\boldsymbol{\pi}) < k/n\sqrt{c-1}$ *and* $\min(\boldsymbol{\pi}) > \frac{k/n}{\sqrt{c-1}}$, and all probabilities in $\boldsymbol{\pi}$ would be bounded within less than a $c-1$ factor of each other, a contradiction of Equation (E.11). We conclude that $\delta(\tilde{\mathcal{I}}) \geq \sqrt{c-1}$. □

### E.4.5 Proof of Lemma 6.4.4

*Proof.* Fix an instance $\mathcal{I} = (N, k, \boldsymbol{\ell}, \boldsymbol{u})$ such that $N$ satisfies Assumption 6.2.4 for constant $\kappa^*$. Fix the panel distribution $\mathbf{d} \in \Delta(\mathcal{K})$ that implies selection probability assignment $\boldsymbol{\pi} \in \Pi(\mathcal{I})$ giving all agents selection probability in $\Theta(k/n)$, which we know to exist by Lemma 6.4.3. Let $\kappa \in (0, \kappa^*)$ be a constant, let $C \subset N$ be an arbitrary coalition of size $|C| = c = \kappa\, n/\sqrt{k}$, and let this coalition misreport feature vectors $\tilde{\boldsymbol{w}} \in \mathcal{W}^c$. Let $\tilde{N} := N \setminus C \cup \tilde{\boldsymbol{w}}$ be the pool created by $C$ misreporting their vectors as $\tilde{\boldsymbol{w}}$, and let $\tilde{\mathcal{I}} = (\tilde{N}, k, \boldsymbol{\ell}, \boldsymbol{u})$ be the resulting instance. Let $\mathcal{K}$ be the set of valid panels in $\mathcal{I}$ and $\tilde{\mathcal{K}}$ be the set of valid panels in $\tilde{\mathcal{I}}$; likewise, let $\mathfrak{K}$ be the set of valid panel *compositions* in $\mathcal{I}$, and let $\tilde{\mathfrak{K}}$ be the set of valid panel *compositions* in $\tilde{\mathcal{I}}$.

Our approach will be to construct a panel distribution $\tilde{\mathbf{d}} \in \Delta(\tilde{K})$ with the desired properties from our original panel distribution $\mathbf{d}$. We will do this construction in panel composition space. We begin with Claim 1, which characterizes the space of valid panel compositions in instance $\mathcal{I}$ versus $\tilde{\mathcal{I}}$.

*Claim 1:* $\mathfrak{K} \subseteq \tilde{\mathfrak{K}}$ (the set of valid compositions only grows after $C$ misreports).

*Proof of Claim 1.* Recall that $\mathcal{W}_N$ describes the unique feature vectors in pool $N$. By assumption, we know that $N_w \geq \kappa^* n + k$ for all $w \in \mathcal{W}_N$. Because our coalition is of size $c \leq \kappa n/\sqrt{k}$, we conclude the following lower bound on $\tilde{N}_w$ for each $w \in \mathcal{W}_N$:

$$\tilde{N}_w \geq n\kappa^* + k - c \geq n(\kappa^* - \kappa/\sqrt{k}) + k \geq \max\{k, n(\kappa^* - \kappa/\sqrt{k})\} \qquad \text{for all } w \in \mathcal{W}_N \tag{E.12}$$

By Equation (E.12), $\tilde{N}$ still contains at least $k$ people of each feature vector present in $N$, meaning that

$$\mathsf{K} \in \mathfrak{K} \implies \mathsf{K} \in \tilde{\mathfrak{K}}.$$

*(End proof of Claim 1).*

While the set of feasible panel compositions could not have *shrunk* due to $C$ misreporting, it could certainly have grown, as members of $C$ may have reported vectors not present in $N$. We now partition $C$ into three mutually exclusive and exhaustive subsets:

- $C_1 \subseteq C$ contains all $i$ whose feature vector $\tilde{w}(i)$ is contained on some panel composition $\mathsf{K} \in \mathfrak{K}$
- $C_2 \subseteq C$ contains all $i$ whose feature vector $\tilde{w}(i)$ is not contained on some panel composition $\mathsf{K} \in \mathfrak{K}$, but is contained on some panel composition $\mathsf{K} \in \tilde{\mathfrak{K}} \setminus \mathfrak{K}$
- $C_3 \subseteq C$ contains all $i$ whose feature vector $\tilde{w}(i)$ is not contained on any panel composition in $\tilde{\mathfrak{K}}$.

By Assumption 6.2.3, we know that given instance $\tilde{\mathcal{I}}$, the selection algorithm will ignore agents in $C_3$, meaning that our effective pool size in $\tilde{\mathcal{I}}$ is $\tilde{n} := n - |C_3|$. We correspondingly redefine $\tilde{N} := \tilde{N} \setminus C_3$. Note that the lower bound from Equation (E.12) on $\tilde{N}_w$ for all $w \in \mathcal{W}_N$ remains unchanged (note that we do indeed mean $\mathcal{W}_N$ here, rather than $\mathcal{W}_{\tilde{N}}$. Then, this inequality still

holds because the number of people with any vector $w \in \mathcal{W}_N$ will not be changed by dropping people of vectors that cannot be included, which must be $w \notin \mathcal{W}_N$).

Now, for each $i \in C_2$, identify a panel $K^{(i)} \in \tilde{\mathcal{K}}$ such that $i \in K^{(i)}$ (these panels needs not be unique). Let $Z$ represent the maximum total number seats reserved for any single vector across these panels (counting duplicates with their multiplicity):

$$Z := \max_{w \in W_{\tilde{N}}} \sum_{i \in C_2} k \cdot \mathsf{K}_w(K^{(i)}).$$

Let $g \colon \bigcup_{i \in C_2} \{\mathsf{K}(K^{(i)})\} \to \mathbb{N}$ map a given panel composition to the number of agents in $C_2$ whose chosen panel have that composition. Formally, it is defined as $g(\mathsf{K}) = |i \in C_2 \colon \mathsf{K}(K^{(i)}) = \mathsf{K}|$. Note that grouping these representative panels by panel composition is a partition of $|C_2|$, so $\sum_{\mathsf{K} \in \bigcup_{i \in C_2} \{\mathsf{K}(K^{(i)})\}} g(\mathsf{K}) = |C_2|$, and additionally $g(\mathsf{K}) \leq |C_2|$ for all $\mathsf{K}$.

Note that $Z \geq 1$, because each agent $i \in C_2$ is given at least one seat on one panel $K^{(i)}$. Also note that $Z \leq k|C_2|$, as we sum over $|C_2|$ panels that can allot at most $k$ seats to any vector. We have that $Z \geq g(\mathsf{K})$ for all $\mathsf{K}$ because if there are $g(\mathsf{K})$ many copies of the same panel composition in the representative panels then there are at least $g(\mathsf{K})$ seats reserved for any given vector on this panel composition.

We will now construct a new panel composition distribution $\tilde{\mathrm{d}}$ from $\mathbf{d}$ transformed into a panel composition distribution $\mathrm{d}$ as given by Lemma E.4.3. By Lemma 6.4.3, we know that

$$\kappa^* \leq \mathrm{t}_w(\mathrm{d}) \leq k \qquad \text{for all } w \in \mathcal{W}_N. \tag{E.13}$$

Colloquially, in this construction we will add the necessary newly feasible panels to the support and redistribute some probability mass over them. Define panel composition distribution $\tilde{\mathrm{d}}$ as follows:

$$\tilde{\mathrm{d}}_{\mathsf{K}} := \begin{cases} \mathrm{d}_{\mathsf{K}} \cdot \left(1 - \frac{\sqrt{k}|C_2|}{\sqrt{Z}\tilde{n}}\right) & \text{if } \mathsf{K} \in \mathfrak{K} \\ \frac{g(\mathsf{K})\sqrt{k}}{\sqrt{Z}\tilde{n}} & \text{if } \mathsf{K} \in \bigcup_{i \in C_2} \{\mathsf{K}(K^{(i)})\} \\ 0 & \text{else} \end{cases} \qquad \text{for all } \mathsf{K} \in \tilde{\mathfrak{K}}.$$

*Claim 2:* $\tilde{\mathrm{d}}$ is a well-defined distribution.

*Proof of Claim 2.* First, note that for every $\mathsf{K} \in \tilde{\mathfrak{K}}$, $\tilde{\mathrm{d}}_{\mathsf{K}}$ is set to a single value. This is because the cases are by definition mutually exclusive: if $\mathsf{K} \in \mathfrak{K}$, it cannot be among the panels compositions $\mathsf{K}(K^{(i)}) | i \in C_2$ by definition. First we show that $\tilde{\mathrm{d}}_{\mathsf{K}} \leq 1$ for all $\mathsf{K} \in \tilde{\mathfrak{K}}$ using that $Z \geq 1$ and $g(\mathsf{K}) \leq |C_2| \leq c - |C_3| = \kappa n/\sqrt{k} - |C_3|$:

$$\frac{g(\mathsf{K})\sqrt{k}}{\sqrt{Z}\tilde{n}} \leq \frac{|C_2|\sqrt{k}}{\sqrt{Z}\tilde{n}} \leq \frac{\kappa n - \sqrt{k}|C_3|}{n - |C_3|} \leq \frac{\kappa n - \kappa|C_3|}{n - |C_3|} \leq \kappa \leq 1$$

To show that $\tilde{\mathsf{d}}_{\mathsf{K}} \geq 0$ for all $\mathsf{K} \in \tilde{\mathfrak{K}}$, we can directly reuse our analysis from above:

$$1 - \frac{\sqrt{k}|C_2|}{\sqrt{Z}\tilde{n}} \geq 1 - \kappa \geq 0. \tag{E.14}$$

Finally, we see that all probabilities in this distribution sum to 1:

$$\begin{aligned}
\sum_{\mathsf{K} \in \tilde{\mathfrak{K}}} \tilde{\mathsf{d}}_{\mathsf{K}} &= \sum_{\mathsf{K} \in \bigcup_{i \in C_2} \mathsf{K}(K^{(i)})} \frac{g(\mathsf{K})\sqrt{k}}{\sqrt{Z}\tilde{n}} + \sum_{\mathsf{K} \in \tilde{\mathfrak{K}} \setminus \bigcup_{i \in C_2} \mathsf{K}(K^{(i)})} \mathsf{d}_{\mathsf{K}} \cdot \left(1 - \frac{\sqrt{k}|C_2|}{\sqrt{Z}\tilde{n}}\right) \\
&= \frac{\sqrt{k}}{\sqrt{Z}\tilde{n}} \sum_{\mathsf{K} \in \bigcup_{i \in C_2} \mathsf{K}(K^{(i)})} g(\mathsf{K}) + \left(1 - \frac{\sqrt{k}|C_2|}{\sqrt{Z}\tilde{n}}\right) \sum_{\mathsf{K} \in \tilde{\mathfrak{K}} \setminus \bigcup_{i \in C_2} \mathsf{K}(K^{(i)})} \mathsf{d}_{\mathsf{K}} \\
&= \frac{\sqrt{k}|C_2|}{\sqrt{Z}\tilde{n}} + \left(1 - \frac{\sqrt{k}|C_2|}{\sqrt{Z}\tilde{n}}\right) \cdot 1 \\
&= 1
\end{aligned}$$

*(End proof of Claim 2).*

In the next part, our goal will be to lower and upper bound $\mathrm{t}_w(\tilde{\mathsf{d}})$ for all $w \in \mathcal{W}_N \cup \mathcal{W}_{C_2}$. We begin by looking at $w \in \mathcal{W}_{C_2}$. We first make some observations about the $g(\mathsf{K})$, and in particular their relationship to sets of agents:

1. $\tilde{N}_w \leq \sum_{\mathsf{K} \in \bigcup_{i \in C_2} \{\mathsf{K}(K^{(i)}) \wedge \mathsf{K}_w > 0} g(\mathsf{K})$ for all $w \in \mathcal{W}_{C_2}$. To see this, note that we can partition agents in $C_2$ according to the panel composition of $K^{(i)}$, the panel we identified to include them. $g(\mathsf{K})$ is then exactly the number of agents who chose $\mathsf{K}$. Adding up over all panel compositions including vector $w$ will necessarily add 1 per person with vector $w$, since for each such person there is at least one panel composition containing them whose composition group they belong to.
2. $g(\mathsf{K})\frac{\sqrt{k}}{\sqrt{Z}\tilde{n}} = \sum_{i \in C_2 : \mathsf{K}(K^{(i)}) = \mathsf{K}} \frac{\sqrt{k}}{\sqrt{Z}\tilde{n}}$; this is by definition of $g(\mathsf{K})$.

Now, we lower bound $t_w(\tilde{\mathsf{d}})$ for all $w \in \mathcal{W}_{C_2}$:

$$t_w(\tilde{\mathsf{d}}) = \sum_{\mathsf{K} \in \bigcup_{i \in C_2} \{\mathsf{K}(K^{(i)})\}} \frac{g(\mathsf{K})\sqrt{k}}{\sqrt{Z}\tilde{n}} \cdot k \cdot \mathsf{K}_w \geq \frac{\sqrt{k}}{\sqrt{Z}\tilde{n}} \tilde{N}_w \geq \frac{1}{\sqrt{cn}} \tilde{N}_w.$$

The first inequality comes from applying observation (1) above, noting that when $\mathsf{K}_w > 0$, $k\mathsf{K}_w \geq 1$. The the final step uses that $Z \leq k|C_2|$, implying that $Z \leq k(c - |C|_3) \leq kc$.

Now, to upper bound $\mathrm{t}_w(\tilde{\mathsf{d}})$ for all $w \in \mathcal{W}_{C_2}$, we apply observation (2) above.

$$t_w(\tilde{\mathsf{d}}) = \sum_{\mathsf{K} \in \bigcup_{i \in C_2} \{\mathsf{K}(K^{(i)})\}} \frac{g(\mathsf{K})\sqrt{k}}{\sqrt{Z}\tilde{n}} \cdot k \cdot \mathsf{K}_w = \frac{\sqrt{k}}{\sqrt{Z}\tilde{n}} \sum_{\mathsf{K} \in \bigcup_{i \in C_2} \{\mathsf{K}(K^{(i)})\}} \sum_{i \in C_2 : \mathsf{K}(K^{(i)}) = \mathsf{K}} k \cdot \mathsf{K}_w$$

We can condense the sums: the first is over all panel compositions, and the second is over all $i \in C_2$ whose $K^{(i)}$ fits that composition; therefore, this is just

$$= \frac{\sqrt{k}}{\sqrt{Z}\tilde{n}} \sum_{i \in C_2} k \cdot \mathsf{K}_w(K^{(i)})$$

This sum is by definition $Z$:

$$= \frac{\sqrt{k}}{\sqrt{Z}\tilde{n}} Z$$
$$\leq \frac{\sqrt{k}}{\sqrt{Z}\tilde{n}} Z$$

$$\frac{1}{\sqrt{c}\tilde{n}} \tilde{N}_w \leq \frac{\sqrt{k}}{\sqrt{Z}\tilde{n}} \cdot \tilde{N}_w \leq \mathrm{t}_w(\tilde{\mathrm{d}}) \leq \frac{\sqrt{k}}{\sqrt{Z}\tilde{n}} Z \leq \frac{k\sqrt{c}}{\tilde{n}}.$$

We conclude that for all $w \in \mathcal{W}_{C_2}$,

$$\mathrm{t}_w(\tilde{\mathrm{d}}) \in \left[ \frac{1}{\sqrt{c}\tilde{n}} \tilde{N}_w,\ \frac{k\sqrt{c}}{\tilde{n}} \right]. \tag{E.15}$$

For all other $w \in \mathcal{W}_{\tilde{N}} \setminus \mathcal{W}_{C_2}$, we deduce the following bounds on $\mathrm{t}_w(\tilde{\mathrm{d}})$, where the lower bound corresponds to the case where $w$ occurs on no panel compositions in $\bigcup_{\tilde{w} \in \mathcal{W}_{C_2}} \{\mathsf{K}^{(\tilde{w})}\}$, and the upper bound corresponds to the case where this vector occurs on all of them to an extent captured in $Z$. First, by Equations (E.13) and (E.14),

$$\mathrm{t}_w(\tilde{\mathrm{d}}) \geq (1-\kappa)\mathrm{t}_w(\mathrm{d}) \geq \kappa^*(1-\kappa)$$

Next, by Equation (E.13),

$$\mathrm{t}_w(\tilde{\mathrm{d}}) \leq \mathrm{t}_w(\mathrm{d}) + \frac{\sqrt{k}}{\sqrt{Z}\tilde{n}} Z \cdot \tilde{N}_w \leq k + \frac{\sqrt{k}}{\sqrt{Z}\tilde{n}} Z \cdot \tilde{N}_w = k + k\sqrt{c}$$

We conclude that for all $w \in \mathcal{W}_{\tilde{N}}$,

$$\mathrm{t}_w(\tilde{\mathrm{d}}) \in \left[ \min\left\{ \frac{1}{\sqrt{c}\tilde{n}} \tilde{N}_w,\ \kappa^*(1-\kappa) \right\},\ k + k\sqrt{c} \right]. \tag{E.16}$$

Now, we apply Lemma E.4.2 to transform our panel composition distribution $\tilde{\mathrm{d}}$ into a corresponding panel distribution $\tilde{\mathbf{d}}$ such that for the $\tilde{\boldsymbol{\pi}}$ implied by $\tilde{\mathbf{d}}$, it holds that

$$\tilde{\pi}_i = \frac{\mathrm{t}_w(\tilde{\mathrm{d}})}{\tilde{N}_w} \qquad \text{for all } i : \tilde{w}(i) = w,\ w \in \mathcal{W}.$$

First, we bound the probabilities for those whose reported vector $\tilde{w} \in \mathcal{W}_{C_2}$. Here, we will use that $\tilde{n} = n - |C_3| \in [n-c, n] \in [n(1-\kappa^*/\sqrt{k}), n]$ and $\tilde{N}_w \geq 1$. By Equation (E.15), we get that

$$\tilde{\pi}_i \in \left[\frac{1}{\sqrt{c}n},\ \frac{k\sqrt{c}}{n(1-\kappa^*/\sqrt{k})}\right] \qquad \text{for all } i : \tilde{w}(i) = w, w \in C_2.$$

Similarly, we bound the probabilities for those whose reported vector $\tilde{w} \in \mathcal{W}_N$. Here, we will use the fact that $\tilde{N}_w \geq (\kappa^* - \kappa)n$ by Equation (E.12), and $\tilde{N}_w \leq \tilde{n} \leq n$. By Equation (E.16), it then follows that

$$\tilde{\pi}_i \in \left[\min\left\{\frac{1}{\sqrt{c}n},\ \frac{\kappa^*(1-\kappa)}{n}\right\},\ \frac{k+k\sqrt{c}}{(\kappa^*-\kappa)n}\right] \qquad \text{for all } i : \tilde{w}(i) \in \mathcal{W}_N.$$

Taking the union of both these ranges to bound the probabilities of all agents $i \in \tilde{N}$, we conclude the following, where we use that $c \in \Omega(1)$ to eliminate the second term of the minimum above:

$$\tilde{\pi}_i \in \left[\Omega\left(\frac{1}{\sqrt{c}n}\right),\ O\left(\frac{k+k\sqrt{c}}{n}\right)\right] = \left[\Omega\left(\frac{k}{k\sqrt{c}n}\right),\ O\left(\frac{k(1+\sqrt{c})}{n}\right)\right] \qquad \text{for all } i \in \tilde{N}.$$

So we get that $\delta_{below}(\tilde{\boldsymbol{\pi}}) = O(k\sqrt{c})$, and $\delta_{above}(\tilde{\boldsymbol{\pi}}) = O(\sqrt{c})$.

We conclude that $\delta(\tilde{\mathcal{I}}) \leq O(k\sqrt{c})$. □

### E.4.6 Proof of Theorem 6.4.5

*Proof.* Let our truthful instance be $\mathcal{I}^=$, and let our coalition so that when they misreport, they create instance $\mathcal{I}_c^*$. By Observation 6.3.2, $\boldsymbol{\pi}^{\text{Goldilocks}_1}(\mathcal{I}^=) = k/n\mathbf{1}$. Now, applying Observation 6.3.3 to analyze $\mathcal{I}_c^*$, we apply that $d_1 + d_2 = 1$ to get that all selection probabilities realizable in this instance must satisfy

$$\mathsf{p}_{00} = \mathsf{p}_{11} = \frac{k/2 - d_2}{(n-c)/2}, \qquad \mathsf{p}_{10} = d_2\frac{1}{c-1}, \qquad \mathsf{p}_{01} = d_2. \tag{E.17}$$

Now, we establish for which domains of $d_2$ each possible relative ordering of these terms holds, noting that $\mathsf{p}_{00}(\boldsymbol{\pi})$ and $\mathsf{p}_{11}(\boldsymbol{\pi})$ must be symmetric:

- $\mathsf{p}_{01}(\boldsymbol{\pi}) \geq \mathsf{p}_{00}(\boldsymbol{\pi}) \iff d_2 \geq \frac{k}{n-c+2}$.
- $\mathsf{p}_{00}(\boldsymbol{\pi}) \geq \mathsf{p}_{10}(\boldsymbol{\pi}) \iff d_2 \leq \frac{k(c-1)}{n+c-2}$
- $\mathsf{p}_{01}(\boldsymbol{\pi}) \geq \mathsf{p}_{10}(\boldsymbol{\pi}) \qquad \forall d_2 \in [0,1]$.

Then, there can be three possible orderings of the selection probabilities in $\boldsymbol{\pi}^{\text{Goldilocks}_1}$ with the following corresponding domains, which we handle separately:

**Case 1:** $\mathrm{p}_{01}(\boldsymbol{\pi}) \geq \mathrm{p}_{10}(\boldsymbol{\pi}) \geq \mathrm{p}_{00}(\boldsymbol{\pi})$ when $d_2 \geq \frac{k(c-1)}{n+c-2}$. In this case, $\textsc{Goldilocks}_1(\boldsymbol{\pi})$ is equal to

$$n/k \cdot \mathrm{p}_{01}(\boldsymbol{\pi}) + \frac{1}{n/k \cdot \mathrm{p}_{00}(\boldsymbol{\pi})} = n/k \cdot d_2 + \frac{(n-c)/2}{n/k(k/2 - d_2)}.$$

This is increasing in $d_2$ for all $d_2 \in [0, 1]$, so this objective is minimized when $d_2$ is minimized over the relevant domain. Therefore the optimal solution over this domain is $d_2 = \frac{k(c-1)}{n+c-2}$, in which case $\mathrm{p}_{00}(\boldsymbol{\pi}) = \mathrm{p}_{10}(\boldsymbol{\pi})$, and this is interchangeable with case 3.

**Case 2:** $\mathrm{p}_{00}(\boldsymbol{\pi}) \geq \mathrm{p}_{01}(\boldsymbol{\pi}) \geq \mathrm{p}_{10}(\boldsymbol{\pi})$ when $d_2 \leq \frac{k}{n-c+2}$. In this case, $\textsc{Goldilocks}_1(\boldsymbol{\pi})$ is equal to

$$n/k \cdot \mathrm{p}_{00}(\boldsymbol{\pi}) + \frac{1}{n/k \cdot \mathrm{p}_{10}(\boldsymbol{\pi})} = n/k \cdot \frac{k/2 - d_2}{(n-c)/2} + \frac{c-1}{n/k \cdot d_2}.$$

This is decreasing in $d_2$ for all $d_2 \in [0, 1]$, so this objective is minimized when $d_2$ is maximized over the relevant domain. Therefore the optimal solution over this domain is $d_2 = \frac{k}{n-c+2}$, in which case $\mathrm{p}_{00}(\boldsymbol{\pi}) = \mathrm{p}_{01}(\boldsymbol{\pi})$, and this is interchangeable with case 3.

**Case 3:** $\mathrm{p}_{01}(\boldsymbol{\pi}) \geq \mathrm{p}_{00}(\boldsymbol{\pi}) \geq \mathrm{p}_{10}(\boldsymbol{\pi})$ when $\frac{k}{n-c+2} \leq d_2 \leq \frac{k(c-1)}{n+c-2}$. In this case, $\textsc{Goldilocks}_1(\boldsymbol{\pi})$ is equal to

$$n/k \cdot \mathrm{p}_{01}(\boldsymbol{\pi}) + \frac{1}{n/k \cdot \mathrm{p}_{10}(\boldsymbol{\pi})} = n/k d_2 + \frac{c-1}{n/k \cdot d_2}.$$

Taking the derivative of this function with respect to $d_2$ and setting it equal to 0, we get that

$$n/k - \frac{c-1}{n/k \cdot d_2^2} = 0 \iff (k/n)^2 = \frac{d_2^2}{c-1} \iff d_2 = k/n \cdot \sqrt{c-1}.$$

Applying Equation (E.17), this setting of $d$ gives us that $\mathrm{p}_{01} = k/n \cdot \sqrt{c-1}$. It follows that

$$\begin{aligned}
\mathrm{p}_{00}(\boldsymbol{\pi}^{\textsc{Goldilocks}_1}(\mathcal{I}_c^*)) &= \mathrm{p}_{11} = \frac{k/2 - d_2}{(n-c)/2}, \\
\mathrm{p}_{10}(\boldsymbol{\pi}^{\textsc{Goldilocks}_1}(\mathcal{I}_c^*)) &= k/n \cdot \frac{1}{\sqrt{c-1}}, \\
\mathrm{p}_{01}(\boldsymbol{\pi}^{\textsc{Goldilocks}_1}(\mathcal{I}_c^*)) &= k/n \cdot \sqrt{c-1}.
\end{aligned}$$

Now, let $i^* \in C$ be the single agent who misreported $\tilde{w}(i^*) = 01$. $\pi_{i^*}^{\textsc{Goldilocks}_1}(\mathcal{I}^=) = k/n$ and $\pi_{i^*}^{\textsc{Goldilocks}_1}(\mathcal{I}_c^*) = k/n \cdot \sqrt{c-1}$, meaning that

$$\mathsf{manip}_{int}(\mathcal{I}^=, \textsc{Goldilocks}_1, c) = k/n(\sqrt{c-1} - 1).$$

Moreover, given that $\mathrm{p}_{10}(\boldsymbol{\pi}^{\textsc{Goldilocks}_1}(\mathcal{I}_c^*)) = k/n \cdot \frac{1}{\sqrt{c-1}}$, we immediately have that

$$\mathsf{manip\text{-}fairness}(\mathcal{I}^=, \textsc{Goldilocks}_1, c) = k/n \cdot \frac{1}{\sqrt{c-1}}.$$ □

## E.5 Supplemental Materials for Section 6.6

### E.5.1 Investigation of alternative $\gamma$ values

We determine $\gamma$ values according to three different methods; the first is generic across instances, and the second two are instance-wise, aiming to respond to how quotas and self-selection bias in a given instance necessitate practically significant probability gaps.

While static $\gamma$ will behave well as $n$ grows large, for practical $n$ this may not be ideal. This is because these instances display some necessary deviation from $k/n$ in selection probabilities due to the quotas, sometimes to the extent that some people *must* receive very higih or very low probability. While in our theoretical analysis these constant gaps diminish in $n$, in these real-world instances, $n$ is small and these constants matter. This means we might want to tune $\gamma$ on an instance-by-instance basis: for example, if the quotas require someone to receive probability 1, we are better off setting $\gamma$ to be extremely large and prioritizing only low probabilities, since we cannot gain anything on the high end.

**$\gamma_1$: minimax/maximin-balanced.** While we don't know the best solution in any given instance, we can try to approximately balance the terms using our knowledge of $\min(\boldsymbol{\pi}^{\text{Maximin}}(\mathcal{I}))$, the maximal minimum probability, and $\max(\boldsymbol{\pi}^{\text{Minimax}}(\mathcal{I}))$, the minimal maximum probability. The bounds given by our algorithm depend on $\max\{\gamma d, d'\}$, where $k/(dn)$ is the minimum probability and $d'k/n$ is the maximum probability in the feasible instance. To roughly balance these terms relative to one another, we can set

$$k/(dn) = \min(\boldsymbol{\pi}^{\text{Maximin}}(\mathcal{I})) \iff d = \frac{k}{n} \cdot \frac{1}{\min(\boldsymbol{\pi}^{\text{Maximin}}(\mathcal{I}))}$$

and

$$d'k/n = \max(\boldsymbol{\pi}^{\text{Minimax}}(\mathcal{I})) \iff d' = \frac{n}{k} \max(\boldsymbol{\pi}^{\text{Minimax}}(\mathcal{I})),$$

thereby optimistically proceeding as though there exists an instance where we can achieve the maximal minimum probability and the minimal maximum probability simultaneously. Given that our bounds depend on $\max\{\gamma d, d'\}$ by Lemma E.4.1 (the $\gamma$-general version of Lemma 6.4.2), we set $\gamma$ so that $\gamma d$ and $d'$ are balanced:

$$\begin{aligned} \gamma d = d' &\iff \gamma = \frac{k}{n} \cdot \frac{1}{\min(\boldsymbol{\pi}^{\text{Maximin}}(\mathcal{I}))} = \frac{n}{k} \max(\boldsymbol{\pi}^{\text{Minimax}}(\mathcal{I})) \\ &\implies \gamma = \frac{n^2}{k^2} \cdot \max(\boldsymbol{\pi}^{\text{Minimax}}(\mathcal{I})) \cdot \min(\boldsymbol{\pi}^{\text{Maximin}}(\mathcal{I})). \end{aligned}$$

Some observations about this method: As we approach the ability to perfectly equalize, $\gamma \to 1$. As $\max(\boldsymbol{\pi}^{\text{Minimax}}(\mathcal{I})) \to 1$ but $\min(\boldsymbol{\pi}^{\text{Maximin}}(\mathcal{I}))$ is around $k/n$, this gets large, approaching order $n/k$ and prompting us to prioritize low probabilities, as desired. Likewise, if $\max(\boldsymbol{\pi}^{\text{Minimax}}(\mathcal{I}))$ is around $k/n$ but $\min(\boldsymbol{\pi}^{\text{Maximin}}(\mathcal{I})) \to 0$, this approaches 0, prompting us to prioritize only the higher probabilities, as desired.

$\gamma_2$ **selection bias-balanced.** The weakness of method 2 is that we have to optimize minimax and maximin before we can optimize GOLDILOCKS. We can maybe get around this by getting a coarse-grained approximation to the above approach, which estimates how much gap must exist in the selection probabilities to satisfy individual constraints. Building on Flanigan et al. [135]'s measure $\Delta_{p,k,N}$, we set $\eta_{f,v}(N) := |\{i|f(i) = v\}|/|N|$ and let

$$k/(nd) = \min_{(f,v)\in FV} \frac{(\ell_{f,v} + u_{f,v})/2}{\eta_{f,v}(N)} \cdot k/n \qquad \text{and} \qquad d'k/n = \max_{(f,v)\in FV} \frac{(\ell_{f,v} + u_{f,v})/2}{\eta_{f,v}(N)} \cdot k/n.$$

Computing $\gamma$ to balance terms as we did in Method 2, we get that

$$\gamma d = d' \iff \gamma = \min_{f,v\in FV} \frac{(\ell_{f,v} + u_{f,v})/2}{\eta_{f,v}(N)} \cdot \max_{(f,v)\in FV} \frac{(\ell_{f,v} + u_{f,v})/2}{\eta_{f,v}(N)}.$$

We now show that this has the same desirable behavior as Method 2: first, notice as the self-selection bias goes away, both these terms approach 1 and we get $\gamma = 1$. If the self-selection bias requires very high probabilities for some feature-vector, making the max term very large, this will make $\gamma$ larger, prompting us to prioritize low probabilities. If the self-selection bias requires very low probabilities for some feature-vector, this will make $\gamma$ term smaller, prompting us to prioritize high probabilities. If they depart equally from 1 (multiplicatively), then the terms will cancel and $\gamma = 1$.

| **Instances** | MINIMAX | LEXIMIN | GOLDILOCKS(1) | GOLDILOCKS($\gamma_1$) | GOLDILOCKS($\gamma_2$) |
|---|---|---|---|---|---|
| 1 | (0.0, 1.0) | (1.0, 2.0) | (0.73, 1.14) | (0.76, 1.19) | (0.77, 1.2) |
| 2 | (0.0, 1.0) | (1.0, 1.33) | (0.9, 1.0) | (0.94, 1.03) | (0.94, 1.03) |
| 3 | (0.0, 1.0) | (1.0, 1.0) | (0.99, 1.0) | (0.99, 1.0) | (0.99, 1.0) |
| 4 | (0.0, 1.0) | (1.0, 1.0) | (0.97, 1.0) | (0.98, 1.0) | (0.98, 1.0) |
| 5 | (0.0, 1.0) | (1.0, 1.17) | (0.92, 1.0) | (0.93, 1.01) | (0.93, 1.01) |
| 6 | (0.25, 1.0) | (1.0, 1.11) | (1.0, 1.09) | (1.0, 1.09) | (1.0, 1.09) |
| 7 | (0.0, 1.0) | (1.0, 3.5) | (0.7, 1.45) | (0.74, 1.51) | (0.79, 1.63) |
| 8 | (0.0, 1.0) | (0.98, 1.0) | (0.96, 1.0) | (0.99, 1.0) | (0.99, 1.0) |
| 9 | (0.0, 1.0) | (1.0, 1.0) | (0.94, 1.0) | (0.97, 1.0) | (0.98, 1.0) |

Table E.1: We compare the performance of the two instance-specific gamma values described above against MINIMAX, LEXIMIN, and GOLDILOCKS with a gamma value of 1.

### E.5.2 INSTANCES

Table E.3 gives the values of $n$, $k$, and $|\mathcal{W}_N|$ associated with our 9 instances.

### E.5.3 DESCRIPTION OF LEGACY

The *Legacy* algorithm is a greedy heuristic that populates the panel person by person, in each of its $k$ steps uniformly randomizing over all remaining pool members (not yet placed on the panel) who have value $v'$ for feature $f'$, where this feature-value is defined by the following ratio:

| | Instances | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| $n$ | 239 | 312 | 161 | 250 | 404 | 70 | 321 | 1727 | 825 |
| $k$ | 30 | 35 | 44 | 20 | 40 | 24 | 30 | 110 | 75 |
| $\lvert\mathcal{W}_N\rvert$ | 202 | 182 | 92 | 92 | 108 | 25 | 294 | 762 | 554 |

Table E.2: $k$, $n$, and $|\mathcal{W}_N|$ values across all 9 instances we analyze.

$$f', v' := \arg\max_{f,v \in FV} \frac{\ell_{f,v} - \#\text{ people already selected for the panel with } f, v}{\#\text{ people left in the pool with } f, v}.$$

Intuitively, this is computing how desperate we are for quota $f, v$: the top is how many more people we need to fill the quota, and the bottom is how many we have left. If this is large, then the quota is more desperate. The algorithm proceeds this way until either a valid panel is created, or it is impossible to satisfy the quotas with the remaining pool members, at which case it starts over. A more detailed description of how this algorithm handles corner cases can be found in Appendix 11 of [130]; these details are not pertinent to our results.

### E.5.4 Implementation of algorithmic framework

We (will) provide our code for implementing the framework for all $\mathcal{E}$ we optimize. We give approximate runtimes for optimizing the objectives we study below. These runtimes were obtained on a 13-inch MacBook Pro (2020) with an Apple M1 chip. For clarity, we select a representative run of a smaller instance (Instance 1) as well as a representative run of a large instance (Instance 8) with all of our equality objectives.

| Instance | Minimax | Maximin | Leximin | Nash | Goldilocks |
|---|---|---|---|---|---|
| 1 | 10.05 | 10.68 | 32.35 | 28.59 | 47.55 |
| 8 | 35.67 | 33.16 | 358.66 | 857.19 | 8473.9 |

Table E.3: Times (seconds) of a representative run of all of the various objective-optimizing algorithms on Instances 1 and 8.

To calculate optimal distributions under various equality objectives, we used pre-existing implementations of Maximin, Leximin, Nash, and Legacy from publicly available code [130, 131]. We implemented Minimax and Goldilocks using the algorithmic framework provided by [130]. Implementing Minimax is straightforward, as its implementation almost exactly the same as Maximin. To implement Goldilocks, we needed it to be differentiable; thus, we instead optimized the function definition below:

$$n/k \cdot \left(\sum_{i \in [n]} \pi_i^p\right)^{1/p} + 1/n/k \cdot \left(\sum_{i \in [n]} 1/\pi_i^p\right)^{1/p}.$$

We characterize the relationship of this objective and our standard GOLDILOCKS$_1$ function in the propositions below. With this objective in hand, we implement [130]'s framework in the following way. In our experiments, we use an auxiliary constraint (as permitted within the framework) to enforce that selection probabilities are anonymous within a tolerance of 0.01.

GOLDILOCKS Primal program:

$$\begin{aligned}
\max \quad & -\frac{n}{k}\left(\sum_{i\in[n]}\pi_i^p\right)^{1/p} - \frac{k\gamma}{n}\left(\sum_{i\in[n]}\frac{1}{\pi_i^p}\right)^{1/p} \\
\text{s.t.} \quad & \pi_i = \sum_{P\in\mathcal{K}:i\in P} q_P \quad \text{for all } i\in[n] \\
& |\pi_i - \pi_j| \le 0.01 \quad \text{for all } i,j\in[n]: w(i)=w(j) \\
& \sum_{P\in\mathcal{K}} q_P = 1 \\
& q_P \ge 0 \quad \text{for all } P\in\mathcal{K}
\end{aligned}$$

With our stopping condition defined in terms of

$$\eta_i = \frac{\partial h(\vec{\pi})}{\partial \pi_i} = \frac{-n}{kp}\left(\sum_{i'\in[n]}\pi_{i'}^p\right)^{1/p-1}\cdot p\pi_i^{p-1} + \frac{k\gamma}{n}\left(\sum_{i'\in[n]}\frac{1}{\pi_{i'}^p}\right)^{1/p-1}\frac{1}{\pi_i^{p+1}}$$

as

$$\sum_{i\in P'}\eta_i \le \sum_{i\in P^*}\eta_i + \varepsilon_{GL}$$

where $P'$ is the maximizing panel not currently in the support of the panel distribution for the sum of $\eta_i : i\in P'$, and $P^*$ is the maximizing panel currently in the support of the panel distribution for the corresponding sum. The stopping condition as defined in the framework has $\varepsilon_{GL}=0$, but for computational constraints we set $\varepsilon_{GL}=1$. We experimented with thresholds down to $\varepsilon_{G:}=0.1$ and found the degradation of the solution with $\varepsilon_{GL}=1$ to be relatively insignificant.

**Proposition E.5.1.** *Fix $\boldsymbol{\pi}$ and let* $\max := \max_{i\in[n]}\pi_i$ *and likewise,* $\min := \min_{i\in[n]}\pi_i$.

$$\lim_{p\to\infty}\left(\sum_{i\in[n]}\pi_i^p\right)^{1/p} + \gamma\left(\sum_{i\in[n]}\frac{1}{\pi_i^p}\right)^{1/p} = \max + \gamma\frac{1}{\min}$$

*Proof.* We'll analyze the two terms separately, both in much the same way:

$$\max = (\max{}^p)^{1/p} \le \left(\sum_{i\in[n]}\pi_i^p\right)^{1/p} \le \left(\sum_{i\in[n]}\max{}^p\right)^{1/p} = n^{1/p}\max \xrightarrow{p\to\infty} \max$$

$$\frac{1}{\min} = \left(\frac{1}{\min}^{p}\right)^{1/p} \leq \left(\sum_{i\in[n]} \frac{1}{\pi_i}^{p}\right)^{1/p} \leq \left(\sum_{i\in[n]} \frac{1}{\min}^{p}\right)^{1/p} = n^{1/p}\frac{1}{\min} \xrightarrow{p\to\infty} \frac{1}{\min} \quad \square$$

**Lemma E.5.2.** *Suppose there exists a feasible solution $\pi$ in which $\pi_i \in \left[\frac{1}{d(n)\cdot n}, \frac{d(n)'}{n}\right]$ for all $i \in [n]$, where $d(n), d'(n) \in \Omega(1)$. Then, let $\pi^*$ be the optimizer of the following objective in this instance:*

$$f_{gold}(\pi) = \left(\sum_{i\in[n]} \pi_i^p\right)^{1/p} + \frac{1}{n^2}\left(\sum_{i\in[n]} \frac{1}{\pi_i^p}\right)^{1/p}.$$

*Then, for all $i \in [n]$, we have that*

$$\pi_i^* \in \left[\frac{1}{2n^{1+1/p}} \cdot \frac{1}{\max\{d(n), d'(n)\}},\ \frac{2}{n^{1-1/p}} \cdot \max\{d(n), d'(n)\}\right].$$

*Proof.* First, we start with our feasible solution and upper bound the objective:

$$\begin{aligned} f_{gold}(\pi) = \left(\sum_{i\in[n]} \pi_i^p\right)^{1/p} + \frac{1}{n^2}\left(\sum_{i\in[n]} \frac{1}{\pi_i^p}\right)^{1/p} &\leq \left(\sum_{i\in[n]} \left(\frac{d(n)'}{n}\right)^p\right)^{1/p} + \frac{1}{n^2}\left(\sum_{i\in[n]} \left(\frac{1}{d(n)\cdot n}\right)^p\right)^{1/p} \\ &= \frac{d'(n)}{n^{1-1/p}} + \frac{d(n)}{n^{1-1/p}} \\ &\leq \frac{2}{n^{1-1/p}} \max\{d(n), d'(n)\} \end{aligned}$$

Now, assume for the sake of contradiction that in the optimizer $\pi^*$, someone receives higher-order probability than $\frac{2}{n^{1-1/p}} \cdot \max\{d(n), d'(n)\}$. Then,

$$f_{gold}(\pi^*) > \frac{2}{n^{1-1/p}} \cdot \max\{d(n), d'(n)\}$$

which is a contradiction to the optimality of this solution. Likewise, assume for the sake of contradiction that in the optimizer $\pi^*$, someone receives lower-order probability than $\frac{1}{2n^{1+1/p}} \cdot \frac{1}{\max\{d(n),d'(n)\}}$. Then,

$$f_{gold}(\pi^*) > \frac{2n^{1+1/p} \cdot \max\{d(n), d'(n)\}}{n^2} = \frac{2}{n^{1-n^{1-1/p}}} \cdot \max\{d(n), d'(n)\}$$

Again, encountering a contradiction. We conclude the claim. $\square$

### E.5.5 Feature Dropping Methods & Results for Additional Instances

In $\mathcal{I}$, we define the selection bias of feature $f$ exactly as in [135]:

$$\Delta^{f}_{N,k,\boldsymbol{\ell},\boldsymbol{u}} := \max_{v \in V_f} \frac{(\ell_{f,v} + u_{f,v})/2}{\eta_{f,v}(N)} - \min_{v \in V_f} \frac{(\ell_{f,v} + u_{f,v})/2}{\eta_{f,v}(N)}$$

where $\eta_{f,v}(N)$ represents the fraction of people in the pool $N$ with value $v$ for feature $f$.

Then, we order the features in decreasing order of $\Delta^{f}_{N,k,\boldsymbol{\ell},\boldsymbol{u}}$ as follows

$$\Delta^{f_1}_{N,k,\boldsymbol{\ell},\boldsymbol{u}} \geq \Delta^{f_2}_{N,k,\boldsymbol{\ell},\boldsymbol{u}} \geq \cdots \geq \Delta^{f_{|F|}}_{N,k,\boldsymbol{\ell},\boldsymbol{u}}$$

And in Figure 6.1, we drop features $f_1$ (1 feature dropped), then $f_1$ and $f_2$ (2 features dropped), then $f_1, f_2, f_3$ (3 features dropped), and so forth. Dropping a feature, formally speaking, means that we are dropping their associated quota constraints; so after we have dropped $y$ features, we are imposing quotas

$$\boldsymbol{\ell}' := (\ell_{f,v} | v \in V_f, f \in F \setminus \{f_1, \ldots, f_y\}), \quad \text{and} \quad \boldsymbol{u}' := (u_{f,v} | v \in V_f, f \in F \setminus \{f_1, \ldots, f_y\}).$$

**Results for additional instances.** In Figure E.1, we provide the analog to Figure 6.1 for the remaining 6 instances omitted from the body.

### E.5.6 Manipulation Robustness Experimental Methods

In our manipulability experiments, we used the high level structure implemented in [135], but modified it to be in the *panel* distribution setting as opposed to the continuous setting. We now formally describe several aspects of our experimental design.

**Simulating the growth of the pool via *pool copies*.** On the horizontal axis of our plots, we vary the number of *pool copies*. In an instance $\mathcal{I} = (N, k, \boldsymbol{\ell}, \boldsymbol{u})$, 1 pool copy means the pool is $N$; 2 pool copies means that the pool is $N \cup N$ (that is, we duplicate each agent in the original pool once, and leave all else about the instance the same).

**The *Most Underrepresented (MU)* strategy** Fix an $\mathcal{I} = (N, k, \boldsymbol{\ell}, \boldsymbol{u})$. For every feature $f$, let the most underrepresented value be $v^*_f$, defined as

$$v^*_f := \arg\max_{v \in V_f} \frac{(\ell_{f,v} + u_{f,v})/2}{\eta_{f,v}(N)},$$

with $\eta_{f,v}(N)$ defined the same way as above. Then, when an agent $i \in N$ employs the *MU* strategy, they misreport the vector

$$w^{MU} := (v^*_f | f \in F).$$

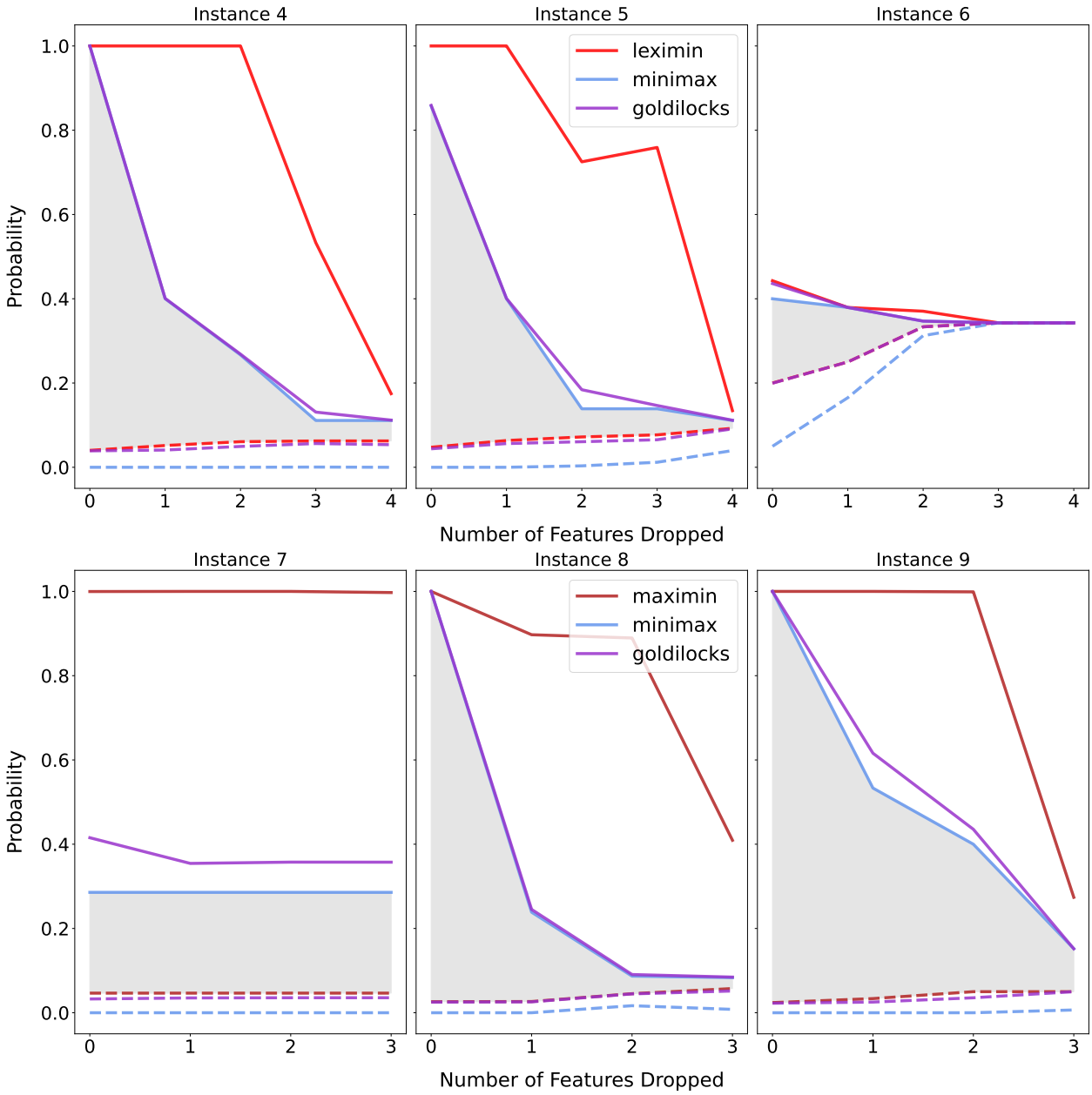



Figure E.1: In instances 7-9, we use MAXIMIN instead of LEXIMIN to indicate the optimal minimum marginal probability because of computational costs due to the size of these instances. We additionally drop only 3 features instead of 4 because instance 9 only has 4 features.

**Computing the worst-case *MU* manipulator.** Fix an $\mathcal{I} = (N, k, \boldsymbol{\ell}, \boldsymbol{u})$ and a maximally equal algorithm E. As above, let $\tilde{\mathcal{I}}_i = (N_{-i} \cup w^{MU}, k, \boldsymbol{\ell}, \boldsymbol{u})$ be the instance in which $i$ has employed the $MU$ manipulation strategy, and all other agents are truthful. Then, we run the following algorithm (pseudocode here) to compute the most any $MU$ manipulator in the instance can gain.

- max-gain $\leftarrow 0$
- compute $\boldsymbol{\pi}^{\mathrm{E}}(\mathcal{I})$
- for all $i \in N$:
  - compute $\boldsymbol{\pi}^{\mathrm{E}}(\tilde{\mathcal{I}}_i)$
  - if $\pi_i^{\mathrm{E}}(\tilde{\mathcal{I}}_i) - \pi_i^{\mathrm{E}}(\mathcal{I}) >$max-gain, set max-gain to this larger difference.
- return max-gain

#### E.5.7 Transparency Experimental Methods

We model our transparency experiments after the experiments done by Flanigan et al. [131]. The two rounding procedures that we utilize in this paper are *ILP* and *Pipage*.

**Theoretical Bounds.** In order to get theoretical upper bounds on the change in any individual's marginal probabilities as a result of rounding, we utilize the results from Flanigan et al. [131]. Theorem 3.2 gives us an upper bound of $b_1 := k/m$, while Theorem 3.3 gives a bound of:

$$b_2 := \frac{\sqrt{\frac{1}{2}(1 + \frac{\ln 2}{\ln |\mathcal{W}_N|})} \cdot \sqrt{|\mathcal{W}_N| \ln(|\mathcal{W}_N|)} + 1}{m}$$

You can find instance-specific values of $n, k$ and $|\mathcal{W}_N|$ in Appendix E.5.4. For our experiments, we set the number of panels $m$ to 1000.

Then, for a given instance $\mathcal{I}$, we derived our theoretical bound on the minimum probability as $\min(\boldsymbol{\pi}^{\textsc{Goldilocks}_1}(\mathcal{I})) - \min(b_1, b_2)$ and the theoretical bound on the maximum probability as $\max(\boldsymbol{\pi}^{\textsc{Goldilocks}_1}(\mathcal{I})) + \min(b_1, b_2)$.

**ILP Rounding.** For the ILP Rounding algorithm, we implemented a new ILP based off of the Nash-optimizing ILP in [131] that optimized the *Goldilocks* objective.

Due to constraints of the IP-solver, we were unable to implement the continuous version of *Goldilocks* used in our optimal distribution algorithm for sufficiently high values of $p$. Instead, we implemented the objective of the form $Goldilocks_1(\boldsymbol{\pi}) = n/k \cdot \max(\boldsymbol{\pi}) + \frac{1}{n/k \cdot \min(\boldsymbol{\pi})}$.

IP-GOLDILOCKS

$$
\begin{aligned}
\max \quad & n/k\|\pi\|_\infty + k/n\|1/\pi\|_\infty \\
\text{s.t.} \quad & \pi_i = \sum_{P \in \mathcal{K}: i \in P} q_P \quad \text{for all } i \in [n] \\
& mq_P \in \mathbb{Z}^+ \\
& \sum_{P \in \mathcal{K}} q_P = 1 \\
& q_P \geq 0 \quad \text{for all } P \in \mathcal{K}
\end{aligned}
$$

It is worth noting that IP-GOLDILOCKS does not give formal guarantees on both maximum and minimum probabilities simultaneously. In IP-MAXIMIN, for example, Flanigan et al. [131] could give guarantees on the minimum probabilities outputted by the rounding algorithm. This is because IP-MAXIMIN optimizes solely for maximizing the minimum probability, and we have a theoretical bound on how much the former minimum probability will be reduced when the panel distribution is rounded. In contrast, *Goldilocks* optimizes for both maximum and minimum probabilities simultaneously, so it is not clear whether it will stay within the theoretical bounds for one or the other on a given instance.

**Pipage.** We ran the pipage algorithm implemented by Flanigan et al. [131] for 1000 independent repetitions for each of our instances. We stored the minimum and maximum marginals from each repetitions and computed the average minimum and average maximum marginal. Additionally, we computed the standard deviation of minimum and maximum marginals across these repetitions. We ultimately found that the spread of the data was very low — standard deviation of minimum and maximum marginals across repetitions did not exceed 0.0015 across all of our instances, and was typically much lower.

# F

# Chapter 9 Appendix

## F.1 Supplemental Materials from Section 9.1.2

### F.1.1 Connections to existing results

We connect our results to those in three papers. The first two study distortion under the $s$-unit stakes assumption, and the third assumes utility queries.

**Caragiannis et al. [70] ($s$-unit stakes, deterministic rules)**
This paper assumes sum-unit-stakes. Although this paper proves distortion bounds for both deterministic and randomized rules, we do not discuss their analysis of randomized rules, as such bounds are more directly addressed in later work, described next. Another similarity between our bounds: $\beta_f$ that our bounds depend on is similar to the dependency of their analysis on alternatives' plurality score.

*Upper bounds (deterministic rules):* Theorem 1 of their paper proves an $O(m^2)$ upper bound on the distortion of Plurality under sum-unit-stakes (i.e., unit-sum utilities). We can recover this bound via our Theorem 9.3.4: First observe that $\kappa$-upper(sum) $= m$ and $\kappa$-lower(sum) $= 1$ (given by utility vectors $\mathbf{1}$ and $\mathbf{1}_1\mathbf{0}_{m-1}$, respectively). Recall also that $\beta_{\text{Plurality}} = 1/m$. By Theorem 9.3.4, it follows that $\text{dist}^{\text{sum}}(\text{Plurality}) \leq m^2$.

*Lower bounds (deterministic rules):* Theorem 1 of their paper proves an $\Omega(m^2)$ lower bound on the distortion of any voting rule under sum-unit stakes (unit-sum utilities). We do show a lower-bound on the distortion of all deterministic voting rules, but due to its is general across *any stakes function* (not just sum), our (tight) lower bound is $\Omega(m)$ (Theorem 9.3.1). However, we can recover the $\Omega(m^2)$ bound specifically for $s = \text{sum}$ for *most* voting rules by combining two of our results. First, by Appendix F.2.6, many voting rules have unbounded distortion under $s$-unit stakes with respect to *any* $s$, including sum. Among the remaining rules with $\beta_f > 0$, we can recover a lower bound of $\Omega(m^2)$ (with tighter constants) for Plurality via Proposition F.2.3: We have $\kappa$-upper(sum) from above, and $\tilde{\kappa}$-lower(sum) $= 2$, given by $\mathbf{1}_2\mathbf{0}_{m-2}$. Then, by Proposition F.2.3, $\text{dist}^{\text{sum}}(\text{Plurality}) \geq (m-1)m/2$. Our bounds here are tighter, improving upon the gap from a factor of 8 to a factor of 2.

**Ebadian et al. [105] ($s$-unit-stakes, randomized rules)**
This paper studies only randomized rules, under both the sum-unit stakes assumption and the max-unit stakes assumption (which they call "range").

*Upper bounds (randomized rules):* As discussed in Section 9.3.2, we use our reduction in Appendix F.2.8 to directly apply their upper bounds on the distortion of Stable Lottery under the sum- and max-unit stakes assumptions (their Theorem 3.4) to prove our upper bound in Theorem 9.3.11.

*Lower bounds (randomized rules):* In Theorem 3.7 of their paper, they show a lower bound of $\Omega(\sqrt{m})$ on the distortion of any randomized rule under max-unit-stakes. This complements a

previously-known bound by **?** ] of $\Omega(\sqrt{m})$ on the distortion of any randomized rule under sum-unit stakes. Our lower bound in Theorem 9.3.11 is weaker than these bounds by a $\log(m)$ factor, but it applies to $s$-unit stakes for *any 1-homogeneous* $s$ (which includes both sum and max). We suspect our lower bound can be tightened to $\Omega(\sqrt{m})$, which would make it a strict generalization of the existing bound.

**Amanatidis et al. [22] (utility queries, deterministic rules)** This paper considers deterministic voting rules with access to one of two kinds of queries: *value* queries, where the voting mechanism can directly ask agents about any one of their utilities; and *comparison* queries, where the voting mechanism can ask agents: "for alternatives $a$ and $b$, is your utility for $a$ at least $d$ times your utility for $b$?" Stakes information according to an arbitrary $s$ can be recovered by some number of either type of these queries (trivially, $m$ value queries, but in many cases, far less). Determining the optimal set of queries of these types to recover a given stakes function is outside the scope of this appendix. Thus, when thinking about upper bounds, we will restrict our consideration here to the stakes functions max, which can be recovered by 1 value query. Similar reasoning applies for range. Finally, we remark that their permission of *noisy* queries (i.e., queries within a constant factor of the truth) are related to our robustness results in Theorem 9.4.1, though due to the generality of our class of stakes functions (and the fact that errors are occurring on *stakes functions' output* rather than utilities directly) requires us to handle additional technicalities.

*Upper bounds (deterministic rules):* Their upper bound in Theorem 1 shows that their mechanism 1-PRV — equivalent to Plurality under stakes-proportionality with respect to max — gives distortion $O(m)$. This result corresponds to our upper bound on the distortion of plurality under max-proportionality, proven via Theorem 9.3.4.

*Lower bounds (deterministic rules):* Their Theorem 7 shows that any single-value query can enable at best $\Omega(m)$ distortion. Our lower bound in Theorem 9.3.1 generalizes this lower bound, showing that *any system of queries yielding the value of a scalar-valued stakes function*, when paired with a deterministic voting rule, can achieve at best $\Omega(m)$ distortion.[1] Of course, this is not to say that we generalize *all* their lower bounds — they prove several other lower bounds for their setting, which are incomparable to ours.

[1] A necessary step in showing that our lower bound subsumes theirs is arguing that our lower bound actually applies to *any stakes-dependent electoral recomposition*, not just proportional recomposition, which we do in the body of the paper.

## F.2 Supplemental Materials from Section 9.3

### F.2.1 All deterministic rules have unbounded distortion

**Proposition F.2.1.** *For all deterministic rules $f$, $dist(f) = \infty$.*

*Proof.* Fix an arbitrary deterministic voting rule $f$. Consider an election with $n$ voters and $m = 2$ alternatives. For $\epsilon > 0$, let $n/2$ voters have utility vector $(\epsilon, 0)$ and let the other half have $(0, 1)$. Then, half of voters will vote for $a$ and the other half for $b$. In this example, $a$ or $b$ are indistinguishable to $f$; suppose it chooses $a$. Then, the distortion in this instance is $\frac{n/2}{\epsilon n/2} \to \infty$ as $\epsilon \to 0$. □

### F.2.2 Proof of Theorem 9.3.1

**Theorem 9.3.1 (lower bound).** *For all $s$ and deterministic $f$,*

$$\mathrm{dist}^s(f) \geq m - 1.$$

*Proof.* We will define two instances, $U$ and $U'$, and show that all $f$ must have at least $m - 1$ distortion in one of these two instances. We will construct $U, U'$ in the following way: first, set aside one alternative $a'$, and let the remaining alternatives be $A_\ell = \{a_j | j \in [m] \setminus \{\ell\}\}$. For all $\ell$, when we write $A_\ell$ in a ranking it represents a ranking over all the alternatives within it, *in increasing order of index.* Divide voters in into $m - 1$ groups, and consider a voter $i$ in group $\ell$: we will assign utility vectors to these voters so that their ranking $\pi_i = a_\ell \succ a' \succ A_\ell$. We display $i$'s utility vectors $\mathbf{u}_i$ and $\mathbf{u}'_i$, as given by $U$ and $U'$ respectively, in *sorted* order, to emphasize how their utilities correspond to their resulting ranking:

| alternative: | $a_\ell$ | $\succ$ | $a'$ | $\succ$ | $A_\ell$ |
|---|---|---|---|---|---|
| sorted $\mathbf{u}_i$ for $i \in$ group $\ell$: | 1 | | 1 | | $0 \dots 0$ |
| sorted $\mathbf{u}'_i$ for $i \in$ group $\ell$: | 1 | | 0 | | $0 \dots 0$ |

We now make three observations:

1. $\mathsf{hist}(U) \equiv \mathsf{hist}(U')$ — that is, the utility matrices induce the same preference histogram. This is true because for every $\ell$, voters in the $\ell$-th group of $U$ and $U'$ have the same ranking.
2. $\mathsf{hist}^s(U) \equiv \mathsf{hist}(U)$ and $\mathsf{hist}^s(U') \equiv \mathsf{hist}(U')$ — that is, the $s$-proportional profiles are identical to the standard profiles for both utility matrices. This is because within each utility matrix, all voters have the same ordered utility vector and thus have the same stakes.
3. $\mathrm{sw}(a', U) = n$ while $\mathrm{sw}(a', U') = 0$. Moreover, $\mathrm{sw}(a_\ell, U) = \mathrm{sw}(a_\ell, U') = n/(m-1)$ for all $\ell \in [m-1]$.

We distinguish between two cases, depending on whether $f(\mathsf{hist}(U)) = a'$ or $f(\mathsf{hist}(U)) \neq a'$. If $f(\mathsf{hist}(U)) = a'$, by (1), we also have that $f(\mathsf{hist}(U')) = a'$. Then, since $\mathrm{sw}(a', U') = 0$,

$$\mathrm{dist}^s_{U'}(f) \overset{(2)}{=} \mathrm{dist}_{U'}(f) = \frac{\mathrm{sw}(a_1, U')}{\mathrm{sw}(a', U')} \overset{(3)}{=} \frac{n/(m-1)}{0} = \infty.$$

If $f(\mathsf{hist}(U)) \neq a'$, then there must exist some $\ell \in [m-1]$ such that $f(\mathsf{hist}(\mathbf{u})) = a_\ell$. Then, fixing this $\ell$,

$$\mathsf{dist}^{\mathrm{s}}_U(f) \overset{(2)}{=} \mathsf{dist}_U(f) = \frac{\mathsf{sw}(a', U)}{\mathsf{sw}(a_\ell, U)} \overset{(3)}{=} \frac{1}{1/(m-1)} = m - 1. \qquad \square$$

### F.2.3 Theorem 9.3.4 holds when $\kappa$'s are defined with range instead of max

**Observation F.2.2.** *The bound in Theorem 9.3.4 remains true also for a slightly different definition of the coefficients* $\kappa$-lower$(s)$, $\kappa$-upper$(s)$ *where* max$(\cdot)$ *is replaced by* range$(\cdot)$,

$$\kappa\text{-upper}(s) := \sup_{\mathbf{u} \in \mathbb{R}^m_{\geq 0}} \frac{s(\mathbf{u})}{\mathit{range}(\mathbf{u})}, \quad \textit{and} \quad \kappa\text{-lower}(s) := \inf_{\mathbf{u} \in \mathbb{R}^m_{\geq 0}} \frac{s(\mathbf{u})}{\mathit{range}(\mathbf{u})}.$$

*Proof.* Let $U \in \mathbb{R}^{m \times n}_{\geq 0}$ be any utility matrix. Then, let $\tilde{U}$ denote the utility matrix in which each agent $i$'s utility vector $\mathbf{u}_i$ is altered by

$$\tilde{u}_i(a) = u_i(a) - \min_{a \in [m]} u_i(a),$$

i.e., the utilities are shifted down such that each voter's minimum utility is 0. Then, letting $c := \sum_{i \in [N]} \min_a u_i(a)$, we obtain that

$$\frac{\mathsf{sw}(a^*, U)}{\mathsf{sw}(a', U)} \leq \frac{\mathsf{sw}(a^*, U) - c}{\mathsf{sw}(a', U) - c} = \frac{\mathsf{sw}(a^*, \tilde{U})}{\mathsf{sw}(a, \tilde{U})}.$$

Then, we may restrict the arguments in the proof of Theorem 9.3.4 to utility vectors with zero minimum entry. This leads to a bound where we may use, instead of $\kappa$-upper$(s)$ and $\kappa$-lower$(s)$

$$\sup_{\mathbf{u} \in \mathsf{R}^m_{\geq 0} : \min_{a \in [m]} u(a) = 0} \frac{s(\mathbf{u})}{\mathsf{max}(\mathbf{u})} \quad \text{and} \quad \inf_{\mathbf{u} \in \mathsf{R}^m_{\geq 0} : \min_{a \in [m]} u(a) = 0} \frac{s(\mathbf{u})}{\mathsf{max}(\mathbf{u})}$$

in place of $\kappa$-upper$(s)$ and $\kappa$-lower$(s)$. We may further upper and lower bound these last two quantities, respectively, by

$$\sup_{\mathbf{u} \in \mathsf{R}^m_{\geq 0} : \min_a u(a) = 0} \frac{s(\mathbf{u})}{\mathsf{max}(\mathbf{u})} = \sup_{\mathbf{u} \in \mathsf{R}^m_{\geq 0} : \min_a u(a) = 0} \frac{s(\mathbf{u})}{\mathsf{range}(\mathbf{u})} \leq \sup_{\mathbf{u} \in \mathsf{R}^m_{\geq 0}} \frac{s(\mathbf{u})}{\mathsf{range}(\mathbf{u})},$$
$$\inf_{\mathbf{u} \in \mathsf{R}^m_{\geq 0} : \min_a u(a) = 0} \frac{s(\mathbf{u})}{\mathsf{max}(\mathbf{u})} = \inf_{\mathbf{u} \in \mathsf{R}^m_{\geq 0} : \min_a u(a) = 0} \frac{s(\mathbf{u})}{\mathsf{range}(\mathbf{u})} \geq \inf_{\mathbf{u} \in \mathsf{R}^m_{\geq 0}} \frac{s(\mathbf{u})}{\mathsf{range}(\mathbf{u})},$$

and we then in particular obtain a distortion upper bound with the two expressions on the right hand side in place of $\kappa$-upper$(s)$ and $\kappa$-lower$(s)$. $\square$

### F.2.4 Proof of Lemma 9.3.6

*Proof.* Fix any deterministic voting rule $f$, and define the quantity

$$\kappa_f = \min_{\mathbf{h}\in\Delta(S_m)} \min_{a\neq f(\mathbf{h})} \sum_{\pi\in S_m} h_\pi \mathbb{I}(f(\mathbf{h}) \succ_\pi a),$$

which captures the minimum fraction of people by whom the winner $f(\boldsymbol{\pi})$ ranked ahead of any other given alternative $a$. In [134], it is shown that for any voting rule $f$, we have that

$$\kappa_f \leq \kappa_{\text{Minimax}} = 1/m,$$

where Minimax is the voting rule which chooses the alternative $a$ that suffers the least severe worst pairwise defeat; see [134] for details. Moreover, we have that for any histogram profile $\mathbf{h}$ and any alternative $a \neq f(\mathbf{h})$,

$$\sum_{\pi\in S_m} h_\pi \mathbb{I}(\pi^{-1}(f(\mathbf{h})) = 1) \leq \sum_{\pi\in S_m} h_\pi \mathbb{I}(f(\mathbf{h}) \succ_\pi a)$$

It follows that $\beta_f \leq \kappa_f \leq 1/m$, which proves the first part of the claim.

Now, for the second part of the claim: the fact that $\beta_{\text{Plurality}} \geq 1/m$ follows immediately its definition: there always exists an alternative which is first-ranked in at least a $1/m$ fraction of the population – therefore, the Plurality winner also has to rank first at least in a $1/m$ fraction of the population. □

### F.2.5 Proof of Proposition F.2.3

**Proposition F.2.3.** *For all $s$,*

$$\text{dist}^{\mathbf{s}}(\textit{Plurality}) \geq (m-1) \cdot \kappa\text{-upper}(s)/\tilde{\kappa}\text{-lower}(s).$$

*Proof.* Formally, we define $\tilde{\kappa}^{\text{lower}}$ as

$$\tilde{\kappa}^{\text{lower}} = \inf_{u\in\mathcal{U}} \frac{s(u)}{\max \mathbf{u}}, \qquad \mathcal{U} := \left\{u \in \mathsf{R}^m_{\geq 0} : u_1 = u_2 \geq \cdots \geq u_m = 0\right\}. \tag{F.1}$$

We will construct an instance which exhibits distortion of the desired order.

**Step 1: Designing the ordered utilities.** There are two population groups: one *high-stake* population group which we call $G_1$ and on *low-stake* population group which we call $G_2$. We denote the proportional group size of $G_1$ by $p = |G_1|/n \in (0,1)$, $1-p = |G_2|/n$. The exact value of $p$ will be determined later in Step 3 of this proof.

Since we are considering proportional recomposition, we may assume without loss of generality that across agents, their maximal utility is equal to 1. Suppose that $u^{\text{upper}}$ is an ordered utility

vector which maximizes the supremum in $\kappa$-upper, such that $\max_{a\in[m]} u^{\text{upper}}(a) = 1$. Similarly, let $u^{\text{lower}}$ denote the utility vector in $\mathcal{U}$ that minimizes the infimum in (F.1). Now, we assign to $G_1$ the ordered utility vector $u^{\text{upper}}$, and to $G_2$ the ordered utility vector $u^{\text{lower}}$. Then, agents in these two population groups have respective stakes of

$$s(u^{\text{upper}}) = \kappa\text{-upper}, \qquad s(u^{\text{lower}}) = \tilde{\kappa}^{\text{lower}}.$$

**Step 2: Designing the rankings.**

- In group $G_1$, we first-rank an alternative $a'$ – this alternative, by appropriate choice of $p$, will later turn out to be the winner of the plurality election. The second to last ranked alternatives in group $G_1$ can be chosen arbitrarily.
- In group $G_2$, the first-rank positions are divided up equally between the remaining $m-1$ alternatives in $[m]\setminus\{a'\}$. Out of those $m-1$ alternatives, we choose an arbitrary alternative which we will make the highest-welfare alternative, called $a^*$. This alternative $a^*$ is ranked second throughout the group $G_2$, whenever it does not rank first.
- Finally, we also specify that the alternative $a'$ is ranked last throughout group $G_2$. The remaining places in $G_2$'s preference profile may be filled arbitrarily.

**Step 3: Specifying the group size $p$.** It remains to calculate $p$. Since $G_1$ has stakes $\kappa$-upper and $G_2$ has stakes $\tilde{\kappa}^{\text{lower}}$, the stakes-weighted plurality score obtained by $a'$ is $p\kappa$-upper. Any other alternative $a \neq a'$ obtains a stakes-weighted plurality score of $(1-p)\tilde{\kappa}^{\text{lower}}/(m-1)$. Thus, $a'$ winning the election amounts to the inequality

$$p\kappa\text{-upper} \geq \frac{1-p}{m-1}\tilde{\kappa}^{\text{lower}} \iff p(\kappa\text{-upper}+\frac{\tilde{\kappa}^{\text{lower}}}{m-1}) \geq \frac{\tilde{\kappa}^{\text{lower}}}{m-1} \iff p \geq \frac{\tilde{\kappa}^{\text{lower}}}{\tilde{\kappa}^{\text{lower}} + (m-1)\kappa\text{-upper}}.$$

Thus, let us set $p$ to be equal to the last expression, i.e.

$$p = \frac{|G_1|}{n} = \frac{\tilde{\kappa}^{\text{lower}}}{\tilde{\kappa}^{\text{lower}} + (m-1)\kappa\text{-upper}}.$$

With this choice of $p$, we notice that

$$\frac{\text{sw}(a', U)}{n} = p, \qquad \text{and} \qquad \frac{\text{sw}(a^*, U)}{n} \geq \frac{1}{n}\sum_{i\in G_2} u_i(a^*) = 1-p,$$

since agents in $G_2$ have utility 1 for $a^*$, and agents in $G_1$ *may* have positive utility for $a^*$. In conclusion, the distortion in this instance is lower bounded by

$$\frac{\text{sw}(a^*, U)}{\text{sw}(a', U)} \geq \frac{1-p}{p} = \frac{\frac{(m-1)\kappa\text{-upper}}{\tilde{\kappa}^{\text{lower}}+(m-1)\kappa\text{-upper}}}{\frac{\tilde{\kappa}^{\text{lower}}}{\tilde{\kappa}^{\text{lower}}+(m-1)\kappa\text{-upper}}} = \frac{(m-1)\kappa\text{-upper}}{\tilde{\kappa}^{\text{lower}}}.$$

□

### F.2.6 Proof: $\beta_f = 0$ for many established voting rules

**Observation F.2.4.** *$\beta_f = 0$ for many established voting rules.*

This observation is shown via a simple instance. Before presenting this instance, we define the voting rules we will address.

**Voting rules.** Borda Count and Veto are positional scoring rules, which are rules defined by a scoring vector $w \in [0,1]^m$ with $j$-th entry $w_j$. In these scoring rules, an alternative receives $w_j$ points for each voter who ranks it $j$-th, and the winner in a given profile is the alternative with the most points. Borda Count is defined by the linearly-decreasing scoring vector $\mathbf{w} = (1, (m-2)/(m-1), \ldots, 1/(m-1), 0)$, and Veto is defined by $\mathbf{w} = \mathbf{1}_{m-1}\mathbf{0}_1$. We also consider the entire class of *Condorcet-consistent rules*. To define this class, we say that *$a$ pairwise-dominates $a'$* in $\mathbf{h}$ if $a$ is ranked ahead of $a'$ in at least half of the electorate. We say that $\mathbf{h}$ has a *Condorcet winner $a$* if $a$ pairwise-dominates all other alternatives. A Condorcet-consistent rule is one which $f(\mathbf{h})$ will be the Condorcet winner on all profiles $\mathbf{h}$ in which a Condorcet winner exists. We will consider this large class of voting rules as a whole, but will not consider any specific rule in this class.

**Instance.** Indeed, consider the following instance with 4 alternatives, $a, b, c, d$:

- 1 voter has $c \succ a \succ d \succ b$
- $n/3 - 1$ voters have $c \succ a \succ b \succ d$
- $n/3$ voters have $b \succ a \succ c \succ d$
- $n/3$ voters have $d \succ a \succ b \succ c$

Then, $a$ is ranked ahead of any other alternative by $2/3$ of voters, and is the Condorcet winner; it will also be the Borda winner, and the Veto winner. Yet, it is never ranked first.

### F.2.7 Proof of Theorem 9.3.11

**Theorem 9.3.11 (lower bound).** *For all 1-homogeneous $s$, randomized $f$,* $\mathrm{dist}^s(f) \geq \frac{\sqrt{m}}{10+3\log m}$.

*Proof.* Define the vector $\mathbf{1}_z\mathbf{0}_{z'}$ to be the vector consisting of $z$ ones followed by $z'$ zeroes.

Case 1: Suppose that there exists some $z \leq (\log m) - 1$ such that $s(\mathbf{1}_{z+1}\mathbf{0}_{m-z-1})/s(\mathbf{1}_z\mathbf{0}_{m-z}) \leq e$. Fix this $z$. We now design a utility instance and associated preference histogram which exhibits a distortion of the order $\sqrt{m/\log m}$.

**Step 1: Designing the rankings.** We begin by designing the preference histogram. We divide the population into $m/\log m$ groups

$$G_1, \ldots G_{m/\log m}.$$

Let alternatives $1, ..., m/\log m$ occupy the first positions in each of the groups $G_1, ...G_{m/\log m}$, respectively. Similarly, we occupy the second to $z$-th rank of those groups by following alternatives:

$$\begin{array}{rccccc} \text{Rank:} & 1 & 2 & \dots & z \\ \text{Group } G_1\text{:} & 1 & m/\log m + 1 & \dots & (z-1)m/\log m + 1 \\ & \vdots & & & \vdots \\ \text{Group } G_{m/\log m}\text{:} & m/\log m & 2m/\log m & \dots & zm/\log m. \end{array}$$

Next, we also divide the population into $\sqrt{m}$ parts $H_1, ..., H_{\sqrt{m}}$ of equal size, based on which alternatives occupy the $(z+1)$-the position. We may design this partition in a way such that

$$\forall k \in [\sqrt{m}] : \ \left|\{l \in [m/\log m] : H_k \cap G_l \neq \emptyset\}\right| \leq \frac{\sqrt{m}}{\log m} + 2.$$

Intuitively, this is because the groups $H_k$ are *larger* by a factor of $\sqrt{m}/\log m$ than the groups $G_l$. We may thus pick the partition into $H_k$ such that each $H_k$ overlaps with at most $\sqrt{m}/\log m + 2$ many groups $G_l$. For each $k \in [\sqrt{m}]$, we assign the $(z+1)$-th position in group $H_k$ to be occupied by the alternative $zm/\log m + k$. Finally, we fill the rest of the positions in the preference histogram – i.e. the $(z+2)$-th to last ranks – arbitrarily.

**Step 2: Designing the utilities.** Amongst the $\sqrt{m}$ alternatives which are ranked in the $(z+1)$-th position, there must exist one alternative which we call $\bar{a}$ which is chosen by the voting rule $f$ with probability at most $1/\sqrt{m}$. That is, if $\mathbf{h}$ denotes the preference histogram constructed in Step 1, then

$$f_{\bar{a}}(\mathbf{h}) \leq 1/\sqrt{m}.$$

Let $H_{\bar{k}}$ be the unique group which ranks $\bar{a}$ in the $(z+1)$-th position. Now, we assign utilities as follows. Define the following ratio of stakes:

$$c_z := \frac{s(\mathbf{1}_{z+1}\mathbf{0}_{m-z-1})}{s(\mathbf{1}_z\mathbf{0}_{m-z})} \leq e.$$

- **Group $H_{\bar{k}}$.** We assign to agents in $H_k$ the ranked utilities $s(\mathbf{1}_{z+1}\mathbf{0}_{m-z-1})$.
- **Remainder.** In the remaining population $H_k^c$, we assign the ranked utilities $c_z \cdot s(\mathbf{1}_z\mathbf{0}_{m-z})$.

These ordered utilities, together with the rankings designed in Step 1, determine a utility matrix which we call $U$.

1. The alternative $\bar{a}$ has average utility $\mathrm{sw}(\bar{a}, U) = 1/\sqrt{m}$.
2. All other alternatives $a \neq \bar{a}$ have average utility at most $\mathrm{sw}(a, U) = c_z \log m/m \leq e \log m/m$.
3. By the homogeneity of the stakes function $s(\cdot)$, all voters have equal stakes. Therefore, we have that $\mathsf{hist}^s(U) = \mathsf{hist}(U) = \mathbf{h}$, and thus also

   $$f(\mathsf{hist}^s(U)) = f(\mathbf{h}).$$

   In particular, $\bar{a}$ is chosen by the voting rule with probability at most $1/\sqrt{m}$ in $f(\mathsf{hist}^s(U))$.

Together, these observations yield that

$$\mathsf{E}\big[\mathsf{sw}(f(\mathsf{hist}^s(U)))\big] \le \frac{e\log m}{m} + \frac{1}{\sqrt{m}}\frac{1}{\sqrt{m}} = \frac{e\log m + 1}{m},$$

and thus the $f$ in Case 1 is at least

$$\frac{\max_a \mathsf{sw}(a,U)}{\mathsf{E}\big[\mathsf{sw}(f(\mathsf{hist}^s(U)))\big]} \ge \frac{1/\sqrt{m}}{(1+e\log m)/m} = \frac{\sqrt{m}}{1+e\log m}.$$

Case 2: It remains to treat the case when the premise of Case 1 is not fulfilled, that is, for every $z \le \log m - 1$, it holds that $s(\mathbf{1}_{z+1}\mathbf{0}_{m-z-1})/s(\mathbf{1}_z\mathbf{0}_{m-z}) \ge e$. By multiplying this equality for all $z = 2,\dots,\log m - 1$, it follows that

$$\frac{s\big(\mathbf{1}_{\log(m)-1}\mathbf{0}_{m-\log(m)+1}\big)}{s(\mathbf{1}_1\mathbf{0}_{m-1})} \ge 2^{\log m - 2} \ge \frac{m}{e^2}. \tag{F.2}$$

Now let us consider a histogram profile where the population is divided in $\sqrt{m}$ many equal sizes groups, which first-rank alternatives $1,\dots\sqrt{m}$, respectively. We fill up the remaining positions in the histogram arbitrarily. Denote this histogram by $\mathbf{h}$.

We now assign utilities to induce $\mathbf{h}$. There must exist one alternative among the $\sqrt{m}$ first-ranked alternatives that receives $\le 1/\sqrt{m}$ probability of selection by $f(\mathbf{h})$. Let us call this alternative $a^*$, and let us call the group which ranks $a^*$ first $G$.

- **Group** $G$**.** In this group, we assign the ordered utility vector $\mathbf{1}_1\mathbf{0}_{m-1}$.
- **Group** $G^c$**.** In the remainder of the population, we assign the ordered utility vector

$$\frac{s(\mathbf{1}_1\mathbf{0}_{m-1})}{s(\mathbf{1}_{\log(m)-1}\mathbf{0}_{m-\log(m)+1})} \cdot \mathbf{1}_{\log(m)-1}\mathbf{0}_{m-\log(m)+1}$$

Let us denote the resulting utility matrix by $U$. We observe the following.

1. The average utility of $a^*$ is at least $\mathsf{sw}(a^*,U)/n \ge 1/\sqrt{m}$.
2. By equation (F.2), the average utility of any other alternative $a \ne a^*$ is at most
$$\frac{\mathsf{sw}(a,U)}{n} \le \frac{e^2}{m}.$$
3. All voters have equal stakes. Therefore $f(\mathbf{h}) = f(\mathsf{hist}(U)) = f(\mathsf{hist}^s(U))$ and we may estimate
$$\mathsf{E}[\mathsf{sw}(f(\mathsf{hist}^s(U)),U)] \le \frac{1}{\sqrt{m}}\frac{1}{\sqrt{m}} + \frac{e^2}{m^2} \le \frac{10}{m}.$$

We obtain an overall distortion of at least

$$\mathsf{dist}^s_U(f) \ge \frac{\sqrt{m}}{10},$$

and the proof is complete. □

### F.2.8 Proof of Lemma 9.3.13

Theorem F.2.5: reduction for rational-valued histograms

Here, we state and prove the reduction assuming $\mathsf{hist}^s(U)$ has only rational entries, which we ensure by restricting to rational utility matrices $U \in \mathbb{Q}_{\geq 0}^{n\times m}$ and *rationality-preserving* stakes functions $s$ (i.e., $s(\mathbf{u}) \in \mathbb{Q}$ whenever $\mathbf{u} \in \mathbb{Q}_{\geq 0}^m$).

**Theorem F.2.5.** *Let $f$ be a voting rule, $s$ a rationality-preserving and 1-homogeneous stakes function, and let $\mathcal{U}_s$ be the set of all rational utility matrices satisfying the $s$-unit-stakes assumption. Then,*

$$\sup_{n\geq 1}\sup_{U\in\mathcal{U}_s} \mathit{dist}_U(f) = \sup_{n\geq 1}\sup_{U\in\mathbb{Q}_{\geq 0}^{n\times m}} \mathit{dist}_U^s(f).$$

*Proof.* We show the claimed equality by separately proving the directions '$\leq$' and '$\geq$'. In order to see the direction '$\leq$', we note that for any unit-stakes utility matrix $U \in \mathcal{U}_s$, $\mathsf{hist}(U) = \mathsf{hist}^s(U)$: the standard and stakes-proportional histograms are the same. Therefore, $\mathsf{dist}_U(f) = \mathsf{dist}_U^s(f)$. Taking suprema over $n \geq 1$ and $U \in \mathcal{U}_s$, we obtain the '$\leq$' direction.

It remains to show '$\geq$'. In order to prove this direction, we fix any utility matrix $U \in \mathbb{Q}_{\geq 0}^{n\times m}$, and construct a unit-stakes utility matrix $\tilde{U}$ such that $\mathsf{dist}_{\tilde{U}}(f) = \mathsf{dist}_U^s(f)$. We let

$$\bar{s}_i = \frac{s(\mathbf{u}_i)}{\sum_{i\in[n]} s(\mathbf{u}_i)}, \quad i \in [n]$$

be the weights with which voter $i$ is represented in the stakes-recomposed election. Since $\bar{s}_i \in \mathbb{Q}$, there exists some $\tilde{n}$ such that $\bar{s}_i\tilde{n}$ is again an integer for each $i \in [n]$. We fix such an $\tilde{n}$ and now construct a utility matrix $\tilde{U} \in \mathbb{Q}_{\geq 0}^{\tilde{n}\times m}$ for which $f$ (without taking into account stakes) exhibits the same distortion as $U$ (while accounting for stakes).

- We divide the electorate of $\tilde{n}$ into $n$ groups, each of them of size $\bar{s}_i\tilde{n}$. Call these groups $G_1, \ldots G_n$.
- Within each group $G_i$, voters have the same ranking $\pi_i(U)$ as voter $i$ in $U$. However, they possess *scaled* utilities $\mathbf{u}_i/s(\mathbf{u}_i)$.

Then we notice that by definition, $\mathsf{hist}(\tilde{U}) = \mathsf{hist}^s(U)$, and therefore also $f(\mathsf{hist}(\tilde{U})) = f(\mathsf{hist}^s(U))$. Moreover, since $s$ is 1-homogeneous, it holds that for all $i$,

$$s\Big(\frac{\mathbf{u}_i}{s(\mathbf{u}_i)}\Big) = \frac{1}{s(\mathbf{u}_i)} s(\mathbf{u}_i) = 1,$$

which yields that $\mathcal{U}_s$ satisfies the unit-stakes property. Moreover, for all alternatives $a \in [m]$, it holds that

$$\frac{\mathsf{sw}(a,U)}{n} = \sum_{i\in[n]} u_i(a) = \frac{\sum_i s(\mathbf{u}_i)}{n} \sum_{i\in[n]} \bar{s}_i \frac{u_i(a)}{s(\mathbf{u}_i)} = \sum_{i\in[n]} s(\mathbf{u}_i) \cdot \frac{\mathsf{sw}(a,\tilde{U})}{n} = \sum_{i\in[n]} s(\mathbf{u}_i)\frac{\tilde{n}}{n} \cdot \frac{\mathsf{sw}(a,\tilde{U})}{\tilde{n}}.$$

Since $\sum_{i\in[n]} s(\mathbf{u}_i)\frac{\tilde{n}}{n}$ is a fixed constant independent of $i$ and $a$, it follows that the average utilities in $U$ and $\tilde{U}$ are equal up to multiplication with a fixed constant — thus distortion is preserved. □

Theorem F.2.7: Extension of Theorem F.2.5 to real-valued histograms

Under an additional *very mild* restrictions on the voting rule $f$, it is possible to prove the correspondence between stakes-based procedures and unit-stakes assumptions from Theorem F.2.5 not just for rational utilities, but for all real-valued utility functions. We term this assumption for $f$ to be *rationally approximable*, which amount to the outcome of $f(\mathbf{h})$ for any preference histogram being *well-approximated* by some preference histogram $\tilde{\mathbf{h}}$ with only rational entries.

**Definition F.2.6** (Rationally approximable rules)**.** *We say that a (deterministic or randomized) voting rule $f : \Delta(S_m) \to \Delta([m])$ is 'rationally approximable' if for every $\mathbf{h} \in \Delta(S_m)$ and every $\epsilon > 0$ there exists another histogram $\tilde{\mathbf{h}} \in \mathbb{Q}_{\geq 0}^{n \times m}$ with only rational entries such that*

$$\sup_{\pi \in S_m} |h_\pi - \tilde{h}_\pi| \leq \epsilon \quad \text{and} \quad \sup_{a \in [m]} |f_a(\mathbf{h}) - f_a(\tilde{\mathbf{h}})| \leq \epsilon,$$

*where $f_a(\mathbf{h})$ denotes the win probability of $a$ in $f(\mathbf{h})$.*

**Theorem F.2.7.** *For any 1-homogeneous stakes function $s$ and any voting rule $f : \Delta([m!]) \to \Delta([m])$, we have that*

$$\sup_{n \geq 1} \sup_{U \in \mathcal{U}_s} \mathit{dist}_U(f) \leq \mathit{dist}^s(f).$$

*If additionally $s$ is 1-homogeneous and $f$ is either (i) weakly locally constant or (ii) continuous, then the reverse inequality is also true,*

$$\sup_{n \geq 1} \sup_{U \in \mathcal{U}_s} \mathit{dist}_U(f) \geq \mathit{dist}^s(f).$$

*Proof of Theorem F.2.7.* The first inequality is immediately implied by the fact that for any $U \in \mathcal{U}_s$, the stakes-recomposed electorate is identical to the original electorate. Indeed, in this case stakes-based election yields the same outcome as the non-stakes-based election, $f(\mathsf{hist}(U)) = f(\mathsf{hist}^s(U))$, so that $\mathrm{dist}_U(f) = \mathrm{dist}_U^s(f)$. It thus only remains to prove the reverse inequality.

Let us fix an arbitrary $n \geq 1$ and utility matrix $U \in \mathsf{R}^{n \times m}$, and let $\mathsf{hist}^s(U) \in \Delta(S_m)$ denote the stakes-recomposed profile corresponding to $U$. Without loss of generality, we may assume both $\mathrm{sw}(a^*, U) > 0$ (since otherwise $U = 0$) and

$$\mathbb{E}\big[\mathrm{sw}(f(\mathsf{hist}^s(U)), U)\big] > 0,$$

since otherwise $\mathrm{dist}_U^s(f) = \infty$ and there remains nothing to prove. By Proposition F.2.8, given *any* $\rho > 0$ we may choose a unit-stakes utility matrix $\tilde{U} \in \mathsf{R}_{\geq 0}^{\tilde{n} \times m}$ such that

$$\sup_{a \in [m]} |f_a(\mathsf{hist}^s(U)) - f_a(\mathsf{hist}(\tilde{U}))| \leq \rho \qquad \text{and} \qquad \sup_{a \in [m]} \left| \frac{\mathrm{sw}(a, U)}{n} - \frac{\mathrm{sw}(a, \tilde{U})}{\tilde{n}} \right| \leq \rho.$$

These two properties, taken together, imply the convergence

$$\left| \mathsf{E}\Big[\frac{\mathrm{sw}(f(\mathsf{hist}^s(U)), U)}{n}\Big] - \mathsf{E}\Big[\frac{\mathrm{sw}(f(\mathsf{hist}(\tilde{U})), \tilde{U})}{\tilde{n}}\Big] \right| \xrightarrow{\rho \to 0} 0,$$

as well as the convergence

$$\Big|\max_{a\in[m]}\frac{\mathrm{sw}(a,U)}{n}-\max_{a\in[m]}\frac{\mathrm{sw}(a,\tilde{U})}{\tilde{n}}\Big|\xrightarrow{\rho\to 0}0.$$

Taken together, this implies that

$$|\mathrm{dist}^s_U(f)-\mathrm{dist}_{\tilde{U}}(f)|\xrightarrow{\rho\to 0}0,$$

which proves the claim. □

**Proposition F.2.8** (Approximation of social welfares)**.** *Suppose $f$ is a rationally approximable voting rule. Let $U\in\mathsf{R}^{n\times m}$ be any non-zero utility matrix. Then, for any $\rho>0$ there exists some large enough $\tilde{n}$ and a unit-stakes utility matrix $\tilde{U}\in\mathsf{R}^{\tilde{n}\times m}$ such that*

- *The election outcomes are close,*

$$\sup_{a\in[m]}|f_a(\mathit{hist}^s(U))-f_a(\mathit{hist}(\tilde{U}))|\le\rho.$$

- *For all $a\in[m]$, the average utilities in $U$ and $\tilde{U}$ are close,*

$$\Big|\frac{\mathit{sw}(a,U)}{n}-\frac{\mathit{sw}(a,\tilde{U})}{\tilde{n}}\Big|\le\rho.$$

*Proof.* Let $\epsilon>0$ be arbitrary and fix any $U$. By Definition F.2.6, we can choose some $\tilde{\mathbf{h}}\in\mathbb{Q}^{S_m}_{\ge 0}$ with rational coefficients such that

$$\sup_{\pi\in S_m}|\mathsf{hist}^s_\pi(U)-\tilde{h}_\pi|\le\epsilon\quad\text{and}\quad\sup_{a\in[m]}|f\big(\mathsf{hist}^s_\pi(U)\big)-f_a(\tilde{h})|\le\epsilon,$$

**Step 1: Construction of utility matrix which induces $\tilde{\mathbf{h}}$.** Since $\tilde{\mathbf{h}}\in\mathbb{Q}^{S_m}_{\ge 0}$ only has rational coefficients, there exists some electorate with $\tilde{n}$ many voters and preferences $(\tilde{\pi}_i : i\le\tilde{n})$ such that for each $\pi\in S_m$, exactly a $\tilde{h}_\pi$ fraction of the voters have ranking $\pi$. Now, we construct a unit-stakes utility matrix $\tilde{U}\in\mathcal{U}_s\cap\mathsf{R}^{\tilde{n}\times m}$ which induces those rankings to the $\tilde{n}$ voters, and which in turn will induce the profile $\tilde{\mathbf{h}}$, $\mathsf{hist}(\tilde{U})=\tilde{\mathbf{h}}$. To this end, let

$$\bar{s}_i:=\frac{s(u_i)}{\sum_{i\in[n]}s(u_i)},\qquad\sum_{i\in[n]}\bar{s}_i=1,$$

denote the weights corresponding to each voter $i$'s preferences in the stakes-recomposed electorate. Since $s$ is 1-homogeneous, we may assume without loss of generality that $\sum_{i\in[n]}s(u_i)=n$, by simply scaling the utilities (note that this leaves $\mathsf{hist}^s(U)$ and also $\mathrm{dist}^s_U(f)$ unchanged). We partition in the new 'unit-stakes electorate' (which consists of $\tilde{n}$ voters) into $n+1$ parts, which we denote by $G_1,\ldots,G_{n+1}$. Within each of those groups, voters share the same ordered utility vector.

**Groups $G_1, \ldots G_n$.** The first $n$ groups $G_1, \ldots, G_n$ are specified as follows. Voters in group $i$ have the utilities $\frac{u_i}{s(u_i)}$, i.e., the same utilities as voter $i$ in the original electorate, but *scaled* to unit-stakes. In particular, voters in group $G_i$ will inherit the same ranking $\pi_i$ as the $i-th$ voter from the original electorate. Let the (fraction) size of the $i$-th group be denoted by $g_i$, i.e., $g_i = |G_i|/\tilde{n}$. We now determine those sizes. Since

$$\sup_{\pi \in S_m} \left| \tilde{\mathbf{h}}_\pi - \mathsf{hist}^s_\pi(U) \right| \leq \epsilon,$$

we can now choose the $(g_i : i \in n)$ in such a way such that simultaneously, the following properties are satisfied. First, $g_i \in [\bar{s}_i - \epsilon, \bar{s}_i]$, and second, for every $\pi \in S_m$,

$$\sum_{i \in n} g_i \mathbb{I}\big(\pi_i = \pi\big) \leq \tilde{\mathbf{h}}_\pi. \tag{F.3}$$

The first property states that the group size $G_i$ does not exceed the amount of representation of voter $i$ in the stakes-recomposed electorate $\bar{s}_i$. The second property states that by assigning group sizes $g_i$, compared to the histogram $\tilde{\mathbf{h}}$, none of the rankings is *overrepresented*. Note that

$$\sum_i g_i \leq \sum_i \bar{s}_i \leq 1, \quad \text{and} \quad \sum_i g_i \geq \sum_i \bar{s}_i - \epsilon \geq 1 - n\epsilon.$$

**Group $G_{n+1}$.** This group constitutes the remainder of the population. Within this group, everyone has the same *ordered utility vector*, but not the same rankings of alternatives. In this group, we assign the ordered utility vector $(x, 0, \ldots, 0)$, where $x$ is given by $x = s((1, 0, \ldots, 0))^{-1} > 0$. Note that $x$ is the (unique) constant such that $s((x, 0, \ldots, 0)) = 1$. In terms of the orderings of alternatives in group $G_{n+1}$, we assign the exact rankings which are needed to complete the correct histogram $\tilde{\mathbf{h}}$ which we aim to realize. Since from Groups $G_1, \ldots, G_n$, none of the rankings $\pi \in S_m$ was overrepresented compared to $\tilde{\mathbf{h}}$ – see equation (F.3) – this is possible. The group $G_{n+1}$ has size at most $n\epsilon$.

Let us denote the utility matrix which arises from this construction by $\tilde{U} \in \mathsf{R}^{\tilde{n} \times m}_{\geq 0}$.

**Step 2: Approximation of social welfares.** It remains to check that the distortion $\mathrm{dist}_{\tilde{U}}(f)$ induced by $\tilde{U}$ approximates the distortion $\mathrm{dist}^s_U(f)$ for the stakes-based election. To this end, we upper and lower bound the difference in average utilities induced by $U$ and $\tilde{U}$, respectively. First, recalling that $\sum_i s(u_i) = n$, we have the lower bound

$$\begin{aligned}
\frac{\mathsf{sw}(a, \tilde{U})}{\tilde{n}} - \frac{\mathsf{sw}(a, U)}{n} &\geq \sum_{i=1}^n g_i \frac{u_i(a)}{s(u_i)} - \frac{1}{n} \sum_{i=1}^n u_i(a) \\
&\geq \sum_{i=1}^n (\bar{s}_i - \epsilon) \frac{u_i(a)}{s(u_i)} - \frac{1}{n} \sum_{i=1}^n u_i(a) \\
&\geq \sum_{i=1}^n \frac{s(u_i)}{\sum_{j \in [n]} s(u_j)} \frac{u_i(a)}{s(u_i)} - \frac{1}{n} \sum_{i=1}^n u_i(a) - \sum_{i=1}^n \epsilon \frac{u_i(a)}{s(u_i)} \\
&= -\epsilon \sum_{i=1}^n \frac{u_i(a)}{s(u_i)}.
\end{aligned}$$

Similarly, we may derive an upper bound, recalling the constant $x = s((1,0,...,0))^{-1}$:

$$\frac{\mathrm{sw}(a,\tilde{U})}{\tilde{n}} - \frac{\mathrm{sw}(a,U)}{n} \leq \sum_{i=1}^{n} g_i \frac{u_i(a)}{s(u_i)} + n\epsilon x - \frac{1}{n}\sum_{i=1}^{n} u_i(a) \leq \sum_{i=1}^{n} \bar{s}_i \frac{u_i(a)}{s(u_i)} + n\epsilon x - \frac{1}{n}\sum_{i=1}^{n} u_i(a) = n\epsilon \cdot x.$$

Since $\epsilon > 0$ was arbitrary, and since both of the latter two bounds tend to 0 as $\epsilon \to 0$, we can now choose $\epsilon > 0$ small enough to fulfill all of the inequalities in the Proposition F.2.8 for any prescribed threshold $\rho > 0$. This proves the claim. □

Our result shows that, from the perspective of worst-case distortion, using a stakes-based recomposition is equivalent to assuming across the population that every voter has equal stakes.

### F.2.9 Formalisms about the Stable Lottery Rule

We now define the Stable lottery rule,following [105]. Since only the case of a stable lottery of size $\sqrt{m}$ is relevant to us, we shall restrict our definition to this special case. Let $\mathcal{P}_{\sqrt{m}}([m])$ be the set of all subsets (or 'committees') of $[m]$, of size $\sqrt{m}$, and let $\Delta(\mathcal{P}_{\sqrt{m}}([m]))$ be the set all of all distributions on $\mathcal{P}_{\sqrt{m}}([m])$. Given a subset $A \subseteq [m]$ of alternatives, an alternative $a \in [m]$ and a histogram profile $\mathbf{h} \in \Delta(S_m)$, let us denote the fraction of voters who rank $a$ ahead of all of $A$ by

$$Freq_{a \succ A}(\mathbf{h}) = \sum_{\pi \in S_m} h_\pi \mathbb{I}(a \succ_\pi A).$$

If $a \in A$, then we set $Freq_{a \succ A}(\mathbf{h}) = 0$ for all $\mathbf{h}$.

**Definition F.2.9** (Stable lottery)**.** *Given a preference histogram* $\mathbf{h}$*, a stable lottery (of size* $\sqrt{m}$*) is a probability distribution* $P(\mathbf{h}) \in \Delta(\mathcal{P}_{\sqrt{m}}([m]))$ *(i.e., a random selection of a committee of size* $\sqrt{m}$*) such that for all* $\mathbf{h}$*,*

$$\max_{a \in [m]} \mathsf{E}_{A \sim P(\mathbf{h})}\left[Freq_{a \succ A}(\mathbf{h})\right] < \frac{1}{\sqrt{m}}.$$

It is well-known that a stable lottery always exists, see, e.g. [105]. Building on this definition, we define the Stable Lottery Rule in terms of histograms.

**Definition F.2.10** (Stable Lottery Rule)**.** *Given a histogram* $\mathbf{h}$*, let* $P(\mathbf{h})$ *be a stable lottery. With probability* $1/2$*, sample a committee* $A$ *of size* $\sqrt{m}$ *from* $P(\mathbf{h})$*, and then choose an alternative uniformly at random from* $A$*. Else, with the remaining probability* $1/2$*, simply choose an alternative uniformly at random from* $[m]$*.*

*Proof of Theorem 9.3.12.*

**Theorem 9.3.12 (upper bound).** *For* $s \in \{sum, max, range\}$, $\mathrm{dist}^s(\text{Stable Lottery}) \in O(\sqrt{m})$.

First, assume that $s \in \{\max,\ \text{sum}\}$, and let $f = $ Stable Lottery Rule. Then, by a well-established result from Ebadian *et al* [105], we know that both for $s = \text{sum}$ and $s = \max$, the worst-case distortion over unit-stakes instances is of the order $O(\sqrt{m})$,

$$\sup_{n\geq 1}\ \sup_{U\in\mathcal{U}_s} \text{dist}_U(\text{Stable Lottery Rule}) \in O(\sqrt{m}),$$

where we recall the notation $\mathcal{U}_s$ for the set of utility matrices $U$ where each voter has unit stakes, $s(\mathbf{u}_i) = 1$. Our goal is to use Theorem F.2.5 to conclude that the stakes-proportional procedure also has distortion of the order at most $O(\sqrt{m})$. For this, we need to confirm that the Stable Lottery Rule is rationally approximable in the sense of Definition F.2.6. Indeed, this is seen as follows. Let $\mathbf{h}$ be an arbitrary preference histogram. In [105], it is proven not just that a stable lottery always exists for $\mathbf{h}$; indeed, a slightly stronger requirement is validated, namely, that the lottery satisfies

$$\max_{a\in[m]} \mathsf{E}_{A\sim P(\mathbf{h})}\left[Freq_{a\succ A}(\mathbf{h})\right] \leq \frac{1}{\sqrt{m}+1}.$$

Now, let $\epsilon > 0$. Suppose that $\tilde{\mathbf{h}}$ is another histogram profile with rational entries such that

$$\sup_{\pi\in S_m} |h_\pi - \tilde{h}_\pi| \leq \epsilon.$$

We may also choose $\tilde{\mathbf{h}}$ such that the difference $\left|Freq_{a\succ A}(\mathbf{h}) - Freq_{a\succ A}(\tilde{\mathbf{h}})\right| \leq \epsilon$ for any $a$. Choosing $\epsilon$ small enough, $P(\mathbf{h})$ is a permissible stable lottery also for $\tilde{h}$. Using this stable lottery, we have that $f(\mathbf{h}) = f(\tilde{h})$; thus $f$ is rationally approximable; the statement follows for $s \in \{\max, \text{sum}\}$.

It remains to show the claim for $s = \text{range}$. Here, we argue along the same lines as Observation F.2.2: The worst-case distortion both for $s = \text{range}$ and for $s = \max$ can be realized while only considering utility matrices in which each voter has minimum utility 0. Let this set of utilities be denoted by $\mathcal{V}$. Then,

$$\sup_{U\in\mathsf{R}^{n\times m}_{\geq 0}} \text{dist}^{\text{range}}_U(f) = \sup_{U\in\mathcal{V}} \text{dist}^{\text{range}}_U(f) = \sup_{U\in\mathcal{V}} \text{dist}^{\max}_U(f) = \sup_{U\in\mathsf{R}^{n\times m}_{\geq 0}} \text{dist}^{\max}_U(f).$$

□

### F.2.10 Folklore: all randomized rules have at least $m$ distortion.

**Fact F.2.11.** *For all voting rules $f$, dist$(f) \geq m$.*

*Proof.* Consider a histogram in which each of the $m$ alternatives occupies a $1/m$ fraction of the first positions and the second to last positions are occupied arbitrarily. There exists some alternative $a$ which will be chosen by the randomized rule with probability at most $1/m$. Let $G$ denote the group in which $a$ is ranked first. In this group, let us assign the ordered utility vector $(1, 0, ..., 0)$. In the remainder of the population $G^c$, we assign the zero utility vector. Let us denote this utility matrix by $U$. Then, since $f$ selects $a$ with probability at most $1/m$, denoting the winner of the election by $a'$, we obtain $\mathsf{E}[\text{sw}(a', U)/n] \leq 1/m^2$, while the maximum welfare alternative has average utility $\text{sw}(a, U)/n = 1/m$; thus the distortion of $f$ is at least $m$. □

# G

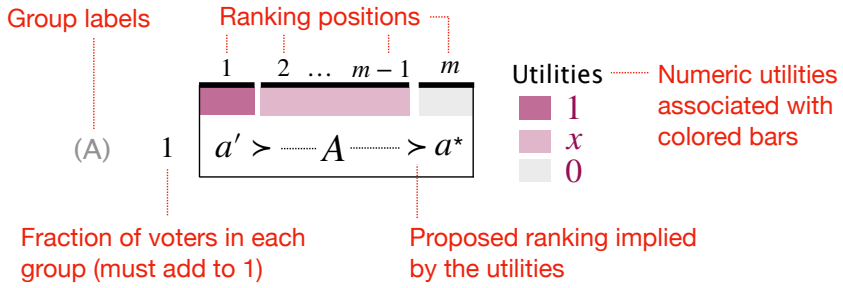
# Chapter 12 Appendix

## G.1 Supplemental Materials from Section 12.3

By convention, throughout the appendices $a'$ denotes the winner of the election, i.e. $a' = f(\boldsymbol{\pi})$, and $a^*$ denotes the highest-welfare alternative.

### G.1.1 Explanation of instance diagrams

In this section of the appendix, we will present the utility matrices of counterexample instances (usually for proving lower bounds) via diagrams. Below, we show the anatomy of such a diagram:




Voters. Most diagrams will have multiple rows, but this one has just a single row, reflecting the fact that this utility matrix has only one *group* of voters, labeled as group (A) on the left. All members of a given group have the same utilities for all alternatives, and thus the same ranking over alternatives. On the left of the box is the number 1, indicating that all voters (a 1-fraction)

belong to group (A).

Alternatives. The alternatives are listed in the white region of the box. In this instance, there are $m$ alternatives: $a'$, $a^*$, and all alternatives in $A$, which represents a bloc of alternatives that are interchangeable in the instance, i.e., treated identically by all voters.

Utilities. We encode voters' utilities for alternatives with colored bars corresponding to the alternative below them, where darker colors correspond to higher utilities. The utility value associated with each color is on the right hand side of the diagram in the key labeled 'Utilities'. Sometimes, this key will contain variables like $x$, which we will set carefully in the proof, as they are functions of $\gamma$. For example, in the diagram above, every voter in group (A) has utility 1 for alternative $a'$, $x$ for all $a \in A$, and 0 for $a^*$. In these examples, we will occasionally set utilities to be larger than 1 to make the math clearer because the scaling is more convenient.

Rankings. Finally, these diagrams encode the rankings that we propose are implied by the utilities. These rankings are denoted by the list of alternatives in the box, separated by $\succ$ symbols to denote that they are ordered. For instance, the ranking proposed in the above instance is $a' \succ A \succ a^*$, i.e., all voters in group (A) rank $a'$ first, $a^*$ is last, and all other alternatives in between. Of course, the fact that these rankings are realized by the given utilities requires proving, which we will do when we prove our lower bounds. The ranking positions are given above the box.

Regarding the rankings of alternatives in blocs like $A$, we will make various assumptions about how the alternatives within $A$ are ranked, via arbitrarily small perturbations of the utilities of those alternatives.[1]

### G.1.2 Proof of Proposition 12.3.5

**Proposition G.1.1.** *$\kappa_{\textsc{Plurality}} = 1/m$, so for all $\boldsymbol{\gamma}$ with $\gamma_{min} > 0$, $dist(\textsc{Plurality}, \boldsymbol{\gamma}) \leq m\frac{1-\gamma_{min}}{\gamma_{min}} + 1$.*

*Proof.* In light of Corollary 12.3.4, proving the claim amounts to proving $\kappa_{\text{Plurality}}(m) \geq 1/m$. Let $f = \text{Plurality}$. For the sake of contradiction, suppose there exists a profile $\boldsymbol{\pi}$ and an alternative $a$ such that $|\{i | f(\boldsymbol{\pi}) \succ_{\pi_i} a\}|/n < 1/m$. For shorthand, let $a' = f(\boldsymbol{\pi})$. Then, $a'$ must be ranked first by less than a $1/m$ fraction of the voters in $\boldsymbol{\pi}$, meaning $a'$ receives strictly less than $n/m$ points. There are $n$ total points awarded across alternatives, so by averaging, there must be an alternative $a \neq a'$ that receives strictly more than $1/m$ points, implying that $f(\boldsymbol{\pi}) \neq a'$, a contradiction. □

### G.1.3 Proof of Proposition 12.3.6

**Proposition G.1.2.** *$\kappa_{\textsc{Borda}} = 1/m$, so for all $\boldsymbol{\gamma}$ with $\gamma_{min} > 0$, $dist(\textsc{Borda}, \boldsymbol{\gamma}) \leq m\frac{1-\gamma_{min}}{\gamma_{min}} + 1$.*

[1] We will usually assume that the alternatives in $A$ are cycled symmetrically across voters' rankings (using arbitrarily small epsilons to tie-break), but sometimes we will instead assume that these alternatives are always ranked consistently. Either way, we can do this tie-breaking without affecting the distortion.

*Proof.* Proving this claim amounts to proving that $\kappa_{\text{BORDA}}(m) \geq 1/m$. Let $f = \text{BORDA}$. For the sake of contradiction, suppose there exists a profile $\boldsymbol{\pi}$ and an alternative $a$ such that $|\{i : f(\boldsymbol{\pi}) \succ_{\pi_i} a\}|/n < 1/m$. For shorthand, let $a' = f(\boldsymbol{\pi})$.

Now, divide voters into two groups: those who rank $a' \succ a$ (an $x < 1/m$ fraction of the voters), and those who rank $a \succ a'$ (the remaining $1 - x$ fraction of voters). Among all voters in the first group, the point gap between $a'$ and $a$ is at most 1, corresponding to $a'$ ranked first and $a$ last. For all voters in the second group, the point gap between $a'$ and $a$ is at most $-1/(m-1)$, i.e., $a$ receives at least $1/(m-1)$ more points than $a'$ from each of these voters' rankings. Then, denoting the respective point totals by $P(a')$ and $P(a)$,

$$P(a') - P(a) \leq x \cdot 1 + (1-x) \cdot \frac{-1}{m-1} < \frac{1}{m} + \left(1 - \frac{1}{m}\right) \cdot \frac{-1}{m-1} = 0.$$

Therefore, $a'$ must receive less points than $a$ and cannot be the BORDA winner, a contradiction. □

### G.1.4 Proof of Proposition 12.3.7

**Proposition G.1.3.** *For all $\boldsymbol{\gamma}$ with (fixed) $\gamma_{min} > 0$, dist(PIECEWISE, $\boldsymbol{\gamma}$) $\in O(m^{2/3})$.*

*Proof.* Let $U \in \mathsf{R}_{\geq 0}^{n \times m}$, fix arbitrary $\boldsymbol{\gamma}$ with $\gamma_{min} > 0$ (as in the hypothesis), and let $\boldsymbol{\pi} \in \Pi_{V(\boldsymbol{\gamma},U)}$. Let $a' = \text{PIECEWISE}(\boldsymbol{\pi})$ and $a^*$ denote the winner and the highest-welfare alternative, respectively. Without loss of generality, let us assume that the average utility of $a^*$ is $\text{sw}(a^*)/n = 1$. We treat separately the scenarios where the lower bound on the social welfare of $a'$ comes from $a'$ *having to beat* $a^*$ (Case 1), and when it comes from *having to beat some other alternative* (Case 2).

**Case 1:** Suppose at least half of voters rank $a^*$ in the first $m^{2/3}$ positions. Let us call this subset of voters

$$N^* = \{i : (\pi_i)^{-1}(a^*) \leq m^{2/3}\},$$

satisfying $|N^*| \geq n/2$. If $a'$ ranks ahead of $a^*$ in more than half of $N^*$, then Lemma 12.3.1 immediately gives a *constant* distortion bound. If on the other hand $a'$ ranks behind $a^*$ in more than half of $N^*$, and since $a^*$ is located in the first $m^{2/3}$ positions where the spacing between consecutive positions is $s_t - s_{t-1} = m^{-2/3}$, $a'$ amasses a point deficit of at least

$$m^{-2/3} \cdot \frac{|N^*|}{2} \geq m^{-2/3} \cdot \frac{n}{4},$$

relative to $a^*$. Thus, in order to beat $a^*$ overall, $a'$ must rank ahead of $a^*$ at least $m^{-2/3} \cdot n/4$ times. Therefore, using Lemma 12.3.1, we obtain a distortion bound of the order $O(m^{2/3})$:

$$\frac{\text{sw}(a^*, U)}{\text{sw}(a', U)} \leq \frac{1 - \gamma_{min}}{\gamma_{min}} \cdot 4m^{2/3} + 1.$$

**Case 2:** Now suppose $a^*$ is ranked in the first $m^{2/3}$ positions by *less* than $1/2$ of the voters. Again let $N^*$ be again the voters where $a^*$ ranks in the first $m^{2/3}$ positions; we have that $|(N^*)^c| \geq n/2$

(using $(\cdot)^c$ to denote the complement). Now, for each alternative $a$, define the frequency with which $a$ occurs in the first $m^{2/3}$ positions amongst $(N^*)^c$ by

$$F_a = \frac{|\{i \in (N^c)^* : (\pi_i)^{-1}(a) \leq m^{2/3}\}|}{n}.$$

Since $|(N^*)^c| \geq n/2$, the *average* frequency of occurrence in the first $m^{2/3}$ positions must satisfy

$$\frac{1}{m}\sum_{a\in[m]} F_a \geq \frac{nm^{2/3}}{2mn} = \frac{m^{-1/3}}{2}. \tag{G.1}$$

Now, we need a further case distinction, based on *how many alternatives* have, roughly speaking, above-average frequency of occurrence in the first $m^{2/3}$ positions. To this end, let $\bar{A}$ be the set of alternatives that have $F_a \geq m^{-1/3}/4$:

$$\bar{A} := \{a \in A : F_a \geq m^{-1/3}/4\}.$$

**Case 2a:** Suppose $|\bar{A}| > m^{2/3}$. Let us now lower bound the average utility of alternatives $a \in \bar{A}$. First, since agents $i \in (N^*)^c$ rank $a^*$ in a lower position than $m^{2/3}$, the set featuring in the definition of $F_a$ is contained as follows

$$\{i : (N^c)^* : (\pi_i)^{-1}(a) \leq m^{2/3}\} \subseteq \{i : a \succ_{\pi_i} a^*\}.$$

Therefore, we may use Lemma 12.3.1 to estimate

$$\frac{n}{\mathrm{sw}(a,U)} \leq \frac{1-\gamma_{min}}{\gamma_{min}}\frac{n}{\{i : a \succ_{\pi_i} a^*\}} + 1 \leq \frac{1-\gamma_{min}}{\gamma_{min}}\frac{1}{F_a} + 1 \leq \frac{1-\gamma_{min}}{\gamma_{min}}4m^{1/3} + 1, \tag{G.2}$$

which leads to the lower bound

$$\frac{\mathrm{sw}(a,U)}{n} \geq \left(\frac{1-\gamma_{min}}{\gamma_{min}}4m^{1/3} + 1\right)^{-1} =: \bar{w}, \quad \text{satisfying } \bar{w} = \Omega(m^{-1/3}).$$

Next, we deduce from this a lower bound on the social welfare of $a'$. Since there are in total $nm^{2/3}/2$ points awarded in the election, $a'$ has to score at least $m^{-1/3}/2$ points per voter (on average) to win. Thus, $a'$ has to rank in the first $m^{2/3}$ positions at least $n/(2m^{1/3})$ many times – denote this set of voters by

$$N' := \{i : (\pi_i)^{-1}(a') \leq m^{2/3}\}, \quad \text{satisfying } |N'|/n \geq m^{-1/3}/2.$$

Since $|\bar{A}| > m^{2/3}$, every time that $a'$ ranks in the first $m^{2/3}$ positions, it has to rank ahead of an alternative $a \in \bar{A}$, whose average utility is lower bounded by $\bar{w}$. Therefore, arguing as in Lemma 12.3.1, we obtain that

$$\frac{\bar{w}}{\mathrm{sw}(a',U)/n} \leq \frac{1-\gamma_{min}}{\gamma_{min}}\frac{n}{|N'|} + 1,$$

which implies

$$\frac{\mathrm{sw}(a', U)}{n} \geq \bar{w}\Big(\frac{1-\gamma_{min}}{\gamma_{min}}\frac{n}{|N'|}+1\Big)^{-1} \geq \bar{w}\Big(\frac{1-\gamma_{min}}{\gamma_{min}}2m^{1/3}+1\Big)^{-1} = \Omega(m^{-2/3}).$$

**Case 2b:** Now suppose $|\bar{A}| \leq m^{2/3}$. Using (G.1), we then obtain that

$$\begin{aligned}\frac{m^{-1/3}}{2} &\leq \frac{1}{m}\Big(\sum_{a\in\bar{A}} F_a + \sum_{a\notin\bar{A}} F_a\Big)\\ &= \frac{|\bar{A}|}{m}\cdot\frac{1}{|\bar{A}|}\sum_{a\in\bar{A}} F_a + \frac{|\bar{A}^c|}{m}\cdot\frac{1}{|\bar{A}^c|}\sum_{a\notin\bar{A}}\frac{m^{-1/3}}{4}\\ &\leq \frac{|\bar{A}|}{m}\cdot\frac{1}{|\bar{A}|}\sum_{a\in\bar{A}} F_a + \frac{m^{-1/3}}{4}.\end{aligned}$$

Rearranging and using that $|\bar{A}| \leq m^{2/3}$, we obtain that

$$\begin{aligned}\frac{m^{-1/3}}{4} \leq \frac{|\bar{A}|}{m}\cdot\frac{1}{|\bar{A}|}\sum_{a\in\bar{A}} F_a \leq \frac{m^{2/3}}{m}\cdot\frac{1}{|\bar{A}|}\sum_{a\in\bar{A}} F_a = m^{-1/3}\frac{1}{|\bar{A}|}\sum_{a\in\bar{A}} F_a\\ \implies \frac{1}{|\bar{A}|}\sum_{a\in\bar{A}} F_a \geq \frac{1}{4}.\end{aligned}$$

It follows that there must exist at least one alternative $\bar{a} \in \bar{A}$ such that $F_{\bar{a}} \geq 1/4$. Since at least $n/4$ voters rank $\bar{a}$ ahead of $a^*$, Lemma 12.3.1 implies that

$$\frac{n}{\mathrm{sw}(\bar{a}, U)} \leq 4\frac{1-\gamma_{min}}{\gamma_{min}} + 1,$$

and thus the average utility of $\bar{a}$ is lower bounded by a constant, $\mathrm{sw}(\bar{a}, U)/n = \Omega(1)$. We may now complete the proof by arguing as in Case 1: Indeed, each time $a'$ ranks behind $\bar{a}$, it incurs a scoring deficit of $m^{-2/3}$. It thus must rank ahead of $\bar{a}$ at least $\Omega(nm^{-2/3})$ times, which, via Lemma 12.3.1, gives the desired lower bound $\mathrm{sw}(a', U)/n = \Omega(m^{-2/3})$. □

### G.1.5 Proof of Proposition 12.3.8

The goal of this section is to show the following upper bound for the distortion of Maximin.

**Proposition G.1.4.** *$\kappa_{\textsc{Maximin}} = 1/m$, so for all $\boldsymbol{\gamma}$ with $\gamma_{min} > 0$, $dist(\textsc{Maximin}, \boldsymbol{\gamma}) \leq m\frac{1-\gamma_{min}}{\gamma_{min}} + 1$.*

Our high-level proof strategy is to show that $\kappa_{\text{Maximin}} = 1/m$. Then, the proposition follows immediately from an application of Corollary 12.3.4. Since every voting rule satisfies $\kappa_f \leq 1/m$, we only have to show that $\kappa_{\text{Maximin}} \geq 1/m$, which is directly implied by the following lemma.

**Lemma G.1.5.** *For every $\boldsymbol{\pi}$ preference profile, there exists some alternative $\bar{a} \in [m]$ such that*

$$\min_{a \neq \bar{a}} \{i : \bar{a} \succ_{\boldsymbol{\pi}_i} a\} \geq n/m.$$

*In particular, the* Maximin *winner $a'$ (which is the alternative with the smallest maximum pairwise loss) must also satisfy*

$$\min_{a \neq a'} |\{i : a' \succ_{\pi_i} a\}| \geq n/m.$$

*Consequently, it also holds that $\kappa_{\textsc{Maximin}} \geq 1/m$.*

*Proof.* We define a sequence of alternatives $(a_j : j \geq 1)$ as follows. Start with an arbitrary alternative $a_1$. Given $a_j$, we let $a_{j+1}$ be the alternative which pairwise-dominates $a_j$ by the most,

$$a_{j+1} := \arg \max_{a \in [m] \setminus \{a_j\}} |\{i : a \succ_{\pi_i} a_j\}|.$$

In this process, if we encounter an alternative that has previously been part of the sequence, i.e. $a_{j+1} = a_k$ for some $k \leq j$, then we exit the recursive procedure, and draw a cycle $(a_k, \ldots, a_{j+1})$. Then, the longest such cycle we can create is of length $m + 1$. Since $a_1$ was arbitrary, we may without loss of generality assume that the constructed cycle starts at $k = 1$, and has length $L$, i.e. the cycle is $(a_1, \ldots, a_L)$ with $a_1 = a_L$. Now, let $N_j \subseteq [n]$ denote the set of voters who rank $a_{j+1} \succ a_j$, i.e., who contribute to $a_j$'s worst pairwise defeat. We now make the following claim.
**Claim:** There exists some $j^* \in [L]$ such that $|N_{j^*}| \leq \frac{L-2}{L-1} n$.

To prove the claim, we first note that there cannot exist any voter $i$ such that

$$a_1 \succ_{\pi_i} \ldots \succ_{\pi_i} a_L \succ_{\pi_i} a_1,$$

since this ranking would be cyclical. It follows that

$$\bigcap_{j=1}^{L-1} N_j = \emptyset.$$

Now, assume for the sake of contradiction that for all $j = 1, \ldots, L$ it holds that $|N_j| > \frac{L-2}{L-1} n$. Then, this implies that.

$$\Big|\bigcap_{j=1}^{J} N_j\Big| > \frac{L-1-J}{L-1} n, \quad \text{for all } J = 1, \ldots, L-1.$$

Intuitively, we are saying that if all $N_j$ individually comprise nearly the entire set of voters, their intersection must be somewhat large. Now, looking in particular at the case where $J = L - 1$, the above inequality implies that $|\bigcap_{j=1}^{L-1} N_j| > 0$, which contradicts that the intersection of all $N_j : j \in [L-1]$ must be empty, as above. We conclude that the claim is true.

Since $L - 1 \leq m$, the preceding claim implies that there exists some $a_{j^*}$ whose worst defeat is by less than $\frac{L-1}{L} \leq \frac{m-1}{m}$ fraction of voters, i.e.,

$$\max_{a \neq a_{j^*}} \frac{|\{i : a \succ_{\pi_i} a_{j^*}\}|}{n} \leq \frac{m-1}{m}.$$

This proves the first assertion of the proposition, that is, by setting $\bar{a} = a_{j^*}$, we obtain the desired alternative for which at least $n/m$ voters must rank $\bar{a} \succ a$.

Since $a'$ is the Maximin winner, we further obtain that

$$\min_{a \neq a'} |\{i : a' \succ_{\pi_i} a\}| \geq \min_{a \neq a_{j^*}} |\{i : a \succ_{\pi_i} a_{j^*}\}| \geq \frac{n}{m}.$$

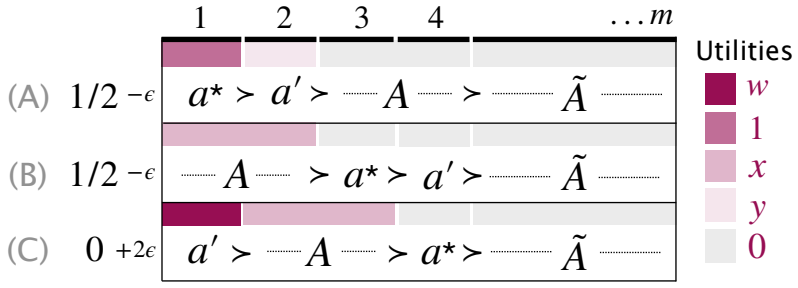
□

### G.1.6 Proof of Proposition 12.3.9

**Proposition G.1.6.** *For all uniform* $\boldsymbol{\gamma} = \gamma \mathbf{1}$, $\gamma \in [0, 1]$, *dist*(Copeland, $\boldsymbol{\gamma}$) $\geq \left(\frac{2(1-\gamma)}{\gamma} + 1\right)^2$.

*Proof.* The claim is true when $\gamma = 1$ trivially, so we will consider $\gamma < 1$. Let $U$ be the utility matrix described by the diagram (see Appendix G.1.1 for a primer on reading these diagrams), where $\epsilon > 0$ and $W$ is some sufficiently large value that depends on $\gamma$ (but not $\epsilon$).

$$w = W, \quad x = \frac{\gamma/2}{1 - \gamma/2}, \quad y = \left(\frac{\gamma/2}{1 - \gamma/2}\right)^2.$$




Observe that $1 > x > y > 0$, and the average utilities of alternatives are the following, where here and throughout this analysis, we will gray out $\epsilon$ terms, as they can be made arbitrarily small.

Now, establishing the average utilities: $\text{sw}(a^*, U)/n = 1/2 - \epsilon$; $\text{sw}(a', U)/n = y/2 + \epsilon(2w - y/2)$; for all $a \in A$, $\text{sw}(a, U)/n = x(1/2 + \epsilon) = x/2 + x\epsilon$; and for all $a \in \tilde{A}$, $\text{sw}(a, U)/n = 0$.

***Claim 1.*** *The utilities imply the claimed rankings.* First, observe that by virtue of having zero social welfare, the alternatives in $\tilde{A}$ are always ranked last. We will consider only the other relative rankings throughout this analysis. We confirm each group's ranking left to right by comparing the values of $v_i(a, \boldsymbol{\gamma}, U)$ (Equation (12.2)), derived below.

Let $i \in$ Group (A) and $a \in A$. Then, $a^* \succ_{\pi_i} a' \succ_{\pi_i} a$:

$$v_i(a^*, \boldsymbol{\gamma}, U) = (1-\gamma) + \gamma(1/2 - \epsilon) = 1 - \gamma/2 - \gamma\epsilon$$
$$v_i(a', \boldsymbol{\gamma}, U) = (1-\gamma)y + \gamma(y(1/2-\epsilon) + 2\epsilon W) = y(1-\gamma/2) + \epsilon\gamma(2W - y)$$
$$v_i(a, \boldsymbol{\gamma}, U) = \gamma x(1/2+\epsilon) = \gamma/2 \cdot \left(\frac{\gamma/2}{1-\gamma/2}\right) + \epsilon\gamma x = y(1-\gamma/2) + \epsilon\gamma x$$

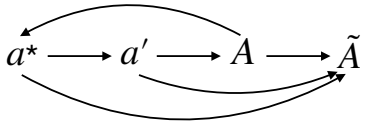
Let $i \in$ Group (B) and $a \in A$. Then, $a \succ_{\pi_i} a^* \succ_{\pi_i} a'$:

$$v_i(a, \boldsymbol{\gamma}, U) = x(1 - \gamma + \gamma(1/2+\epsilon)) = x(1-\gamma/2) + \gamma x\epsilon = \gamma/2 + \gamma x\epsilon$$
$$v_i(a^*, \boldsymbol{\gamma}, U) = \gamma(1/2 - \epsilon) = \gamma/2 - \gamma\epsilon$$
$$v_i(a', \boldsymbol{\gamma}, U) = \gamma(y(1/2-\epsilon) + 2\epsilon W) = \gamma y/2 + \epsilon\gamma(2W - y)$$

Let $i$ be in Group (C), and $a \in A$. Then, $a' \succ_{\pi_i} a \succ_{\pi_i} a^*$:

$$v_i(a', \boldsymbol{\gamma}, U) = (1-\gamma)W + \gamma(y(1/2-\epsilon) + 2W\epsilon) = (1-\gamma)W + \gamma y/2 + \epsilon\gamma(2W - y)$$
$$v_i(a, \boldsymbol{\gamma}, U) = x(1 - \gamma + \gamma(1/2+\epsilon)) = \left(\frac{\gamma/2}{1-\gamma/2}\right)(1-\gamma/2) + \epsilon\gamma x = \gamma/2 + \epsilon\gamma x$$
$$v_i(a^*, \boldsymbol{\gamma}, U) = \gamma(1/2 - \epsilon) = \gamma/2 - \epsilon\gamma$$

***Claim 2.*** *$a'$ is the* Copeland *winner.* To do this analysis quickly, we draw the pairwise majority graph for this instance, where an arrow $a \to \tilde{a}$ indicates that $a$ pairwise-dominates $\tilde{a}$:




Because we assume that items are symmetrically within $A$, and similarly within $\tilde{A}$, $a'$ is the unique Copeland winner:[1]

- $a'$ gets $m-2$ points by strictly pairwise defeating 2 items in $A$ and $m-4$ items in $\tilde{A}$.
- $a^*$ gets $m-3$ points by strictly pairwise defeating $a'$ and $m-4$ items in $\tilde{A}$.
- all $a \in A$ get $m-3$ points by strictly pairwise defeating $a^*$ and $m-4$ items in $\tilde{A}$.
- all $a \in \tilde{A}$ get 0 points.

[1]Here, we additionally assume that $n$ is even (a similar instance, with a third identical alternative added to the set $A$ to form a Condorcet cycle within $A$, would work for odd $n$, see Appendix G.2.2 for a similar construction.).

Distortion. It follows that the distortion in this instance, provided the proposed rankings are realized, approaches the following quantity as $\epsilon \to 0$:

$$\frac{\mathrm{sw}(a^*, U)}{\mathrm{sw}(a', U)} \xrightarrow{\epsilon \to 0} \frac{1/2}{y/2} = \left(\frac{1-\gamma/2}{\gamma/2}\right)^2 = \left(\frac{2-\gamma}{\gamma}\right)^2 = \left(\frac{2(1-\gamma)}{\gamma} + 1\right)^2$$

□

### G.1.7 Proof of Proposition 12.3.10

**Proposition G.1.7.** *For all uniform $\boldsymbol{\gamma} = \gamma\mathbf{1}$, $\gamma \in [0,1]$, dist(Slater, $\boldsymbol{\gamma}$) $\geq \left(\frac{2(1-\gamma)}{\gamma} + 1\right)^2$.*

*Proof.* We can lower-bound Slater's distortion identically to Copeland's, as in Proposition 12.3.9, via the same instance (with slightly different treatment of the alternatives in $A, \tilde{A}$). In particular, where before we cycled alternatives symmetrically in these set, now assume that items are always ordered the same way within $A$, and similarly within $\tilde{A}$. In particular, let $\pi_A, \pi_{\tilde{A}}$ be these consistent sub-rankings. Fix this instance $\boldsymbol{\gamma}, U$. Then, $a'$ is the unique Slater winner, by the argument below. Note that this is all we need to prove identical distortion to Proposition 12.3.9, because we have already confirmed that the rankings in this instance are realized by the utilities, as well as the distortion itself, in the proof of Proposition 12.3.9.

First, we will pare down the possible slater rankings. Observe that because items within $A$, $\tilde{A}$ are always ranked as $\pi_A, \pi_{\tilde{A}}$ in $\boldsymbol{\pi}^{\boldsymbol{\gamma},U}$, the slater ranking must also rank them in this order to minimize pairwise disagreements. Similarly, the slater ranking will always rank everything in $\tilde{A}$ in the last $m-4$ slots, as those items are always in those slots in $\boldsymbol{\pi}^{\boldsymbol{\gamma},U}$.

That leaves us with the possible slater rankings listed below, using $\pi_A, \pi_{\tilde{A}}$ to denote all alternatives in those sets in their fixed ordering. Note that $A$ contains 2 alternatives and $\tilde{A}$ contains $m-4$ alternatives. For each ranking, we tally its disagreements with the pairwise majority graph.

- $a' \succ \pi_A \succ a^* \succ \pi_{\tilde{A}}$ disagrees with 1
- $a' \succ a^* \succ \pi_A \succ \pi_{\tilde{A}}$ disagrees with 3
- $a^* \succ a' \succ \pi_A \succ \pi_{\tilde{A}}$ disagrees with 2
- $a^* \succ \pi_A \succ a' \succ \pi_{\tilde{A}}$ disagrees with 4
- $\pi_A \succ a^* \succ a' \succ \pi_{\tilde{A}}$ disagrees with 2
- $\pi_A \succ a' \succ a^* \succ \pi_{\tilde{A}}$ disagrees with 3

The slater ranking is the first one, so the winner is $a'$. □

### G.1.8 Proof of Theorem 12.3.11

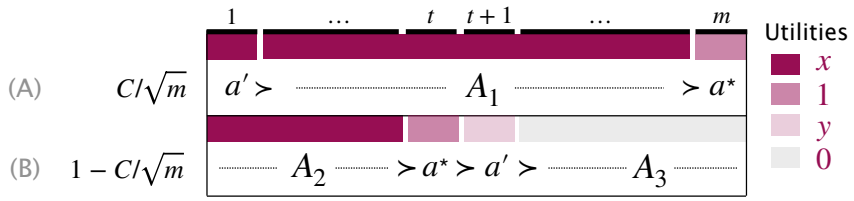
**Theorem 12.3.11.** *For all positional scoring rules $f$ and uniform $\boldsymbol{\gamma} = \gamma\mathbf{1}$ with (fixed) $\gamma \in [0,1)$,*

$$dist(f, \boldsymbol{\gamma}) \in \Omega(\sqrt{m}).$$

*Proof.* Let $\mathbf{s} = (s_1, \dots, s_m)$ denote the (decreasing) scoring vector of $f$, and recall that $s_1 = 1$, $s_m = 0$. Then, there must exist some position $t \in \{1, \dots, \sqrt{m}\}$ such that $s_t - s_{t+1} \leq 1/\sqrt{m}$. We then construct a utility matrix $U$ as pictured in the diagram below (see Appendix G.1.1 for a primer on reading these diagrams), where

$$x = \frac{1}{1-\gamma} \quad \text{and} \quad y = C'/\sqrt{m}$$

and $C, C'$ are constant to be chosen later.




For the ranking of group (A), we assume that $A_1$ contains alternatives $1, \dots, m-2$ occupy the ranks in *cyclically*, i.e. that any given alternative $a = 1, \dots, m-2$ occupies any rank $r \in \{2, \dots, m-1\}$ in a $1/(m-2)$ fraction of group (A) (this is permitted since $a = 1, \dots, m-2$ are treated symmetrically, so we may choose the preference orderings between them arbitrarily when the PS-values are tied.) Similarly, in group (B) we may assume that the alternatives $1, \dots, m-2$ are cycled through the $m-2$ occupied by $A_2 \cup A_3$ – this way their welfares are equal, $\mathrm{sw}(1, U) = \cdots = \mathrm{sw}(m-2, U)$, and the PS-values are hence always tied between the alternatives within positions $A_2$, and within positions $A_3$.

We now argue that the above utilities induce the the rankings profile shown in the diagram. To verify the rankings in group (A), we first note that

$$v_i(a^*, \boldsymbol{\gamma}, U) = 1 \leq (1-\gamma) u_i(a') \leq v_i(a', \boldsymbol{\gamma}, U).$$

Moreover, for any $1 \leq a \leq m-2$, since $t \leq \sqrt{m}$ we have

$$\frac{\mathrm{sw}(a, U)}{n} \leq \frac{C}{(1-\gamma)\sqrt{m}} + \frac{t-1}{(1-\gamma)(m-2)} \leq \frac{C+2}{(1-\gamma)\sqrt{m}}.$$

while for $a'$ we have, for any $m$ large enough such that $C/\sqrt{m} \leq 1/2$,

$$\frac{\mathrm{sw}(a', U)}{n} = \frac{C}{(1-\gamma)\sqrt{m}} + \left(1 - \frac{C}{\sqrt{m}}\right)\frac{C'}{\sqrt{m}} \geq \frac{C}{(1-\gamma)\sqrt{m}} + \frac{C'}{2\sqrt{m}}.$$

Thus, for $C'$ chosen large enough (depending on $\gamma, C$), we obtain that for $a = 1, \dots, m-2$, $\mathrm{sw}(a', U) \geq \mathrm{sw}(a, U)$. It follows that also $v_i(a', \boldsymbol{\gamma}, U) \geq v_i(a, \boldsymbol{\gamma}, U)$, and the rankings of group (A) are confirmed.

We now verify the rankings in group (B). For alternatives $a$ in positions $A_2$, we have

$$v_i(a^*, \boldsymbol{\gamma}, U) = 1 \leq (1-\gamma) u_i(a) \leq v_i(a, \boldsymbol{\gamma}, U),$$

so that they indeed rank ahead of $a^*$. Since $C'$ was chosen above such that $\mathrm{sw}(a', U) \geq \mathrm{sw}(a, U)$ (for all $a = 1, \ldots, m-2$), $a'$ is indeed ranked ahead of $A_3$, and $\mathrm{sw}(a', U) = O(1/\sqrt{m}) = o(\mathrm{sw}(a^*, U))$, we conclude that $a'$ is indeed ranked in the $t+1$-st position. Thus the positions in group (B) are confirmed, too.

It remains to verify that in the ranking profile from the diagram, $a'$ is indeed the positional scoring rule winner. For $i \in [n], a \in [m]$, let $\pi_i^{-1}(a)$ denote the position that voter $i$ ranks alternative $a$ in. Then, we may write the point totals as

$$P(a) := \sum_{i \in [n]} s_{\pi_i^{-1}(a)}, \quad a \in [m]$$

Firstly, $a'$ beats $a^*$, since

$$\frac{1}{n}(P(a') - P(a^*)) = \frac{C}{\sqrt{m}} - \left(1 - \frac{C}{\sqrt{m}}\right)(s_t - s_{t+1}) \geq \frac{C-1}{\sqrt{m}} > 0,$$

as long as we choose $C > 1$. Secondly, to see that $a'$ beats $1, \ldots, m-2$, we prove that $P(a') > P(1)$ (which suffices because $P(1) = \cdots = P(m-2)$). Note that the fraction of times alternative 1 occupies any position $l \in \{1, \ldots, t+1\}$ is bounded by

$$\frac{|\{i : \pi_i^{-1}(1) \leq t+1\}|}{n} = \frac{C}{\sqrt{m}} \frac{t}{m-2} + \big(1 - \frac{C}{\sqrt{m}}\big) \frac{t-1}{m} \leq \frac{t}{m-2} \lesssim \frac{1}{\sqrt{m}},$$

where we again used that $t \leq \sqrt{m}$. Since $a'$ ranks first a $C/\sqrt{m}$ fraction of times, and otherwise occupies the $(t+1)$-th place, we may enforce that $P(a') > P(1)$, by choosing $C > 0$ large enough. □

### G.1.9 Proof of Lemma 12.3.12

**Lemma G.1.8.** *For all positional scoring rules $f$ and uniform $\boldsymbol{\gamma} = \gamma \mathbf{1}, \gamma \in [0, 1]$, $dist(f, \boldsymbol{\gamma}) \geq \frac{1-\gamma}{\gamma \Delta_f} + 1$.*

*Proof.* The claim is true when $\gamma = 1$, because given that all positional scoring rules $f$ are unanimous, $\mathrm{dist}_{\mathbf{1}}(f) = 1$. For the remainder of the proof, we will thus consider $\gamma < 1$.

Fix an arbitrary positional scoring rule $f$ with gap $\Delta_f$, defined as the gap between the scores given to the first two positions (i.e., $s_1 - s_2$). Fix some $\gamma \in [0, 1)$, and let $\boldsymbol{\gamma} = \gamma \mathbf{1}$. Now, consider the instance $(\boldsymbol{\gamma}, U)$ depicted in the diagram below, where $U$ is as shown in the following diagram with $\epsilon > 0$ and

$$x = \frac{\gamma(1 - \Delta_f)}{1 - \gamma + \gamma \Delta_f} \iff x(1 - \gamma + \gamma \Delta_f) = \gamma(1 - \Delta_f)$$

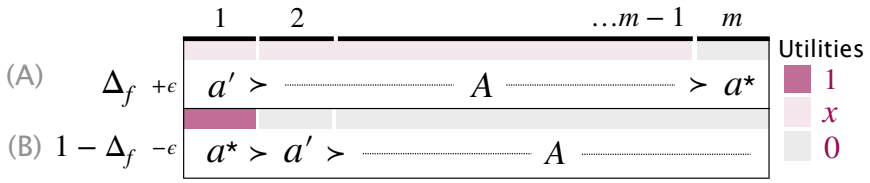



Figure G.1: $A$ contains all alternatives other than $a', a^*$, cycled symmetrically over rankings, and all $\pm\epsilon$ are used for tie-breaking only.

First, we prove two necessary claims, and then analyze the distortion given that $a'$ is the winner by $f$.

***Claim 1:*** *the utilities imply the proposed rankings.* Since $a'$ always has PS-values which are always greater or equal than that of any alternative in $A$, we may always rank $a'$ ahead of all alternatives in $A$, and thus the relative rankings of $a', A$ are correct. Now, we verify the relative orderings of $a^*$ and all other alternatives in both groups:

$$\begin{aligned} x = \frac{\gamma(1-\Delta_f)}{1-\gamma+\gamma\Delta_f} &\iff x(1-\gamma+\gamma\Delta_f) = \gamma(1-\Delta_f) \\ &\implies (1-\gamma+\gamma(\Delta_f+\epsilon))x > \gamma(1-\Delta_f-\epsilon) \\ &\iff v_i(a,\boldsymbol{\gamma},U) > v_i(a^*,\boldsymbol{\gamma},U) \text{ for all } a \neq a^*, i \in \text{group (A)}. \end{aligned}$$

We now analyze group (B)'s ranking. Since $u_i(a') = u_i(a)$ for all $i \in [n]$ and $a \in A$, it suffices to check that

$$\gamma \mathrm{sw}(a', U) = v_i(a', \boldsymbol{\gamma}, U) \leq v_i(a^*, \boldsymbol{\gamma}, U) \text{ for all } i \in \text{group (B)}.$$

Since $u_i(a') = 0$ in group (B), it suffices to verify that $\mathrm{sw}(a', U) \leq \mathrm{sw}(a^*, U)$:

$$\begin{aligned} \mathrm{sw}(a', U) \leq \mathrm{sw}(a^*, U) &\iff x(\Delta_f+\epsilon) \leq (1-\Delta_f-\epsilon) \\ &\iff \frac{\gamma(\Delta_f+\epsilon)}{1-\gamma+\gamma\Delta_f}(1-\Delta_f) \leq (1-\Delta_f-\epsilon) \\ &\iff \gamma(1-\Delta_f)(\Delta_f+\epsilon) \leq \gamma(1-\Delta_f-\epsilon)(\Delta_f+1/\gamma-1) \end{aligned}$$

Since we assumed that $\gamma < 1$, clearly we may choose $\epsilon > 0$ small enough such that the inequality in the last line holds true. This confirms the rankings in group (B).

***Claim 2:*** *$a'$ is the winner per the proposed rankings.* $a'$ is always ranked ahead of all $a \in A$, so $a'$ must receive a higher score than all these alternatives. $a'$ also receives more points than $a^*$: $a'$ receives $\Delta_f + \epsilon + (1-\Delta_f-\epsilon)(1-\Delta_f) > 1-\Delta_f$ points, which is larger than the $1-\Delta_f-\epsilon$ points received by $a^*$.

Now, to analyze the distortion we let $\epsilon \to 0$:

$$\mathrm{dist}_{\boldsymbol{\gamma}}(f) \geq \frac{\mathrm{sw}(a^*, U)}{\mathrm{sw}(a', U)} \xrightarrow{\epsilon \to 0} \frac{1 - \Delta_f}{\Delta_f x} = \frac{1 - \gamma}{\gamma \Delta_f} + 1.$$

□

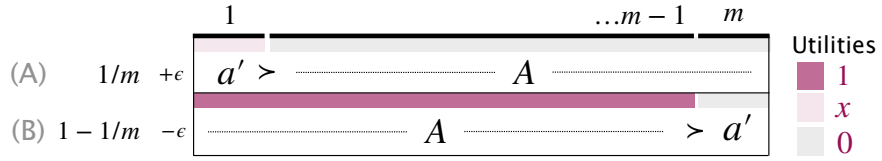
### G.1.10 Proof of Proposition 12.3.15

**Proposition G.1.9.** *For all $s$,*

$$\mathrm{dist}^{\mathrm{s}}(\textsc{Plurality}) \geq (m-1) \cdot \kappa\text{-upper}(s)/\tilde{\kappa}\text{-lower}(s).$$

*Proof.* Fix an arbitrary uniform $\boldsymbol{\gamma} = \mathbf{1}\gamma$ and let $U$ be the utility matrix depicted in the following diagram, where all alternatives in $A$ are cycled symmetrically, and

$$x = \frac{\gamma(m-1)/m}{1 - \gamma + \gamma/m}$$




The average utilities of the alternatives are then the following: $\mathrm{sw}(a', U)/n = x(1/m + \epsilon) = x/m + x\epsilon$, and for all $a \in A$, $\mathrm{sw}(a, U)/n = (m-1)/m - \epsilon$.

***Claim 1.*** *The proposed rankings are realized by the utilities.* We confirm each ranking left to right by comparing voters' PS-values, per Equation (12.2).

Let $i \in$ group (A) and $a \in A$. Then,

$$\begin{aligned} v_i(a', \boldsymbol{\gamma}, U) = x(1 - \gamma + \gamma(1/m + \epsilon)) = x(1 - \gamma + \gamma/m) + x\epsilon &= \left(\frac{\gamma(m-1)/m}{1 - \gamma + \gamma/m}\right)(1 - \gamma + \gamma/m) + x\epsilon \\ &= \gamma(m-1)/m + x\epsilon \\ v_i(a, \boldsymbol{\gamma}, U) = \gamma(1 - 1/m - \epsilon) &= \gamma(m-1)/m - \gamma\epsilon \end{aligned}$$

Let $i \in$ group (B) and $a \in A$. Then,

$$\begin{aligned} v_i(a, \boldsymbol{\gamma}, U) = 1 - \gamma + \gamma(1 - 1/m - \epsilon) &= 1 - \gamma/m - \gamma\epsilon \\ v_i(a', \boldsymbol{\gamma}, U) = \gamma(1/m + \epsilon)x = \gamma x/m + \gamma x\epsilon &= \gamma/m \cdot \left(\frac{\gamma(m-1)/m}{1 - \gamma + \gamma/m}\right) + \gamma x\epsilon \\ &= \gamma/m \cdot \left(\frac{1}{1 - \gamma + \gamma/m} - 1\right) + \gamma x\epsilon \\ &= \frac{\gamma/m}{1 - \gamma + \gamma/m} - \gamma/m + \gamma x\epsilon \\ &< 1 - \gamma/m - \gamma\epsilon \end{aligned}$$

Where the last step holds for sufficiently small $\epsilon$, and $\gamma/m \leq 1 - \gamma/m$ holds when $m \geq 2$.

***Claim 2.*** *$a'$ is the winner.* $a'$ is the Plurality winner because it is ranked first a $1/m + \epsilon$ fraction of the time, while all other alternatives $a \in A$ are ranked first a $1/m - \epsilon/(m-1)$ fraction of the time.

By Claims 1 and 2, the distortion in this instance approaches the following as $\epsilon \to 0$ (where $a$ is an arbitrary alternative in $A$):

$$\frac{\mathrm{sw}(a, U)}{\mathrm{sw}(a', U)} = \frac{(m-1)/m}{x/m} = (m-1) \cdot \frac{1 - \gamma + \gamma/m}{\gamma(m-1)/m} = \frac{1 - \gamma}{\gamma} m + 1.$$

□

#### G.1.11 Proof of Proposition 12.3.16

**Proposition G.1.10.** *For all uniform $\boldsymbol{\gamma} = \gamma\mathbf{1}$, dist(Maximin, $\boldsymbol{\gamma}$) $\geq (m-1) \cdot \frac{1-\gamma}{\gamma} + 1$.*

*Proof.* We first specify a preference profile $\boldsymbol{\pi}$ with $m$ alternatives in which $a'$ is the winner, i.e., $\textsc{Maximin}(\boldsymbol{\pi}) = a'$; we will later show that $\boldsymbol{\pi}$ can be realized by suitable utilities.

We split the population into two groups, A and B:

- **Group A** is of size $n/(m-1)$, and voters $i$ in group A rank

$$a' \succ_i \text{all other } m-1 \text{ alternatives.}$$

- **Group B** contains the rest of the voters, i.e. is of size $n(m-2)/(m-1)$. In this group, voters $i$ have ranking of the form

$$\text{all other } m-1 \text{ alternatives} \succ_{\pi_i} a'.$$

In these rankings, we assume that the $m-1$ non-winning alternatives (call them $1 \dots m-1$) are ranked cyclically – that is, each group is further divided into $m-1$ subgroups of equal size, where voters $i$ in the respective $k$-th subgroups rank

$$k \succ_{\pi_i} k+1 \succ_{\pi_i} \cdots \succ_{\pi_i} m-1 \succ_{\pi_i} 1 \succ_{\pi_i} \cdots \succ_{\pi_i} k-1.$$

We now verify that indeed $a' = \textsc{Maximin}(\boldsymbol{\pi})$. Firstly, $a'$ performs equally well in all comparisons with other alternatives, i.e.

$$\max_{a \neq a'} |\{i : a \succ_{\pi_i} a'\}| = |\{i : 1 \succ_{\pi_i} a'\}| = n - \frac{n}{m-1} = n\frac{m-2}{m-1}.$$

On the other hand, for each of the remaining alternatives $k = 1, \dots, m-1$, their worst defeat comes from the preceding alternative $k-1$ (for $k=1$, this alternative is $m-1$) – in particular, the cyclical rankings in both Group 1 and 2 immediately imply that

$$\max_{a \neq k} |\{i : a \succ_{\pi_i} k\}| = |\{i : k-1 \succ_{\pi_i} k\}| = n\frac{m-2}{m-1} \geq \max_{a \neq a'} |\{i : a \succ_{\pi_i} a'\}|,$$

confirming that $a'$ wins the election.

We now specify the utilities as follows.

- In **Group A**, voters $i$ have $u_i(a') = \frac{\gamma(m-2)}{(1-\gamma)(m-1)+\gamma}$ and $u_i(a) = 0$ for all remaining alternatives.
- In **Group B**, voters $i$ have $u_i(a') = 0$ and $u_i(a) = 1$ for all remaining alternatives $a = 1, \dots, m-1$.

The cyclical rankings amongst $a = 1, \dots, m-1$ can be realized since we have treated those alternatives symmetrically, so that $v_i(a, \boldsymbol{\gamma}, U)$ are tied for all $i \in [n]$ and $a = 1, \dots, m-1$. The ranking of voters in group $B$ is confirmed by comparison of social welfares and utilities. $a'$ is ranked ahead of all other $a$ for all $i$ in Group A by the following reasoning:

$$\begin{aligned} v_i(a', \boldsymbol{\gamma}, U) &= (1-\gamma)u_i(a') + \mathrm{sw}(a', U) \\ &= (1-\gamma)\frac{\gamma(m-2)}{(1-\gamma)(m-1)+\gamma} + \frac{\gamma}{m-1}\frac{\gamma(m-2)}{(1-\gamma)(m-1)+\gamma} \\ &= (1-\gamma+\frac{\gamma}{m-1})\frac{\gamma(m-2)}{(m-1)[(1-\gamma)+\gamma/(m-1)]} \\ &= \gamma\frac{m-2}{m-1} = v_i(a, \boldsymbol{\gamma}, U). \end{aligned}$$

Since we may assume worst-case tie breaking, we may rank $a'$ ahead of $a$. Note that in this profile, all the alternatives $a = 1, \dots, m-1$ have equal social welfare. Fixing any such $a$, the distortion in this instance is

$$\frac{\mathrm{sw}(a, U)}{\mathrm{sw}(a', U)} = \frac{m-2}{m-1} \cdot \frac{m-1}{\frac{\gamma(m-2)}{(1-\gamma)(m-1)+\gamma}} = \frac{(1-\gamma)(m-1)+\gamma}{\gamma}.$$

□

## G.2 Supplemental Material for Section 12.4

In this appendix, we will often apply the following lemma:

**Lemma G.2.1.** *For any utility matrix, decreasing a voter $i$'s level of public spirit cannot result in them promoting a higher-welfare alternative over a lower-welfare alternative in their ranking.*

*Proof.* Fix an arbitrary utility matrix $U$, arbitrary voter $i$, and $\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}$ which differ in that $\tilde{\gamma}_i < \gamma_i$ (they may also differ in other ways — it is irrelevant to this proof). Fix corresponding profiles $\boldsymbol{\pi} \in \Pi_{V(\boldsymbol{\gamma},U)}$ and $\tilde{\boldsymbol{\pi}} \in \Pi_{V(\tilde{\boldsymbol{\gamma}},U)}$. Let $a, a'$ be an arbitrary pair of alternatives such that $a' \succ_{\pi_i} a$ and $\mathrm{sw}(a,U) \geq \mathrm{sw}(a',U)$ (if no such pair exists, because $i$'s alternatives are ranked in decreasing order of welfare, and thus we are done because $i$ cannot promote a higher-welfare alternative over a lower-welfare alternative). We will show that $a$ cannot be promoted over $a'$ from $\pi_i$ to $\tilde{\pi}_i$—that is, $a' \succ_{\tilde{\pi}_i} a$, thereby showing the claim.

First, observe that because $a$ has greater social welfare than $a'$, $i$ must have higher utility for $a'$ than $a$ to create their relative ranking in $\pi_i$:

$$a' \succ_{\pi_i} a \implies u_i(a') > u_i(a).$$

Then, by $\tilde{\gamma}_i < \gamma_i$, $\mathrm{sw}(a',U) - \mathrm{sw}(a,U) < 0$ and $u_i(a') - u_i(a) > 0$,

$$\begin{aligned} v_i(a', \tilde{\boldsymbol{\gamma}}, U) - v_i(a, \tilde{\boldsymbol{\gamma}}, U) &= (1-\tilde{\gamma}_i)(u_i(a') - u_i(a)) + \tilde{\gamma}_i(\mathrm{sw}(a',U) - \mathrm{sw}(a,U)) \\ &> (1-\gamma_i)(u_i(a') - u_i(a)) + \gamma_i(\mathrm{sw}(a',U) - \mathrm{sw}(a,U)) \\ &= v_i(a', \boldsymbol{\gamma}, U) - v_i(a, \boldsymbol{\gamma}, U) > 0. \end{aligned}$$

The inequality deduced above concludes the proof: $v_i(a', \tilde{\boldsymbol{\gamma}}, U) - v_i(a, \tilde{\boldsymbol{\gamma}}, U) > 0 \implies a' \succ_{\tilde{\pi}_i} a$. □

### G.2.1 Proof of Proposition 12.4.5

**Proposition G.2.2.** *If $m \leq 3$, then all voting rules exhibit nonuniform monotonicity.*

We prove this for $m = 2$ and $m = 3$ separately, though the arguments use the same overall strategy. We present the proof of the $m = 2$ case more gently as a warm-up, to illustrate the high-level approach; the proof of $m = 3$ requires more careful handling of additional technicalities.

**Proposition 12.4.5(a).** *When $m = 2$, all voting rules exhibit nonuniform monotonicity.*

*Proof.* Fix an arbitrary resolute voting rule $f$, and suppose our two alternatives are $a, b$. To show the claim, it suffices to show that, starting with an instance $\boldsymbol{\gamma}, U$ and given a $\tilde{\boldsymbol{\gamma}}$ which only differs from $\boldsymbol{\gamma}$ in that $\tilde{\gamma}_1 < \gamma_1$ (i.e., only a single voter's public spirit is decreased), we can find some $\tilde{U}$ with the following two properties:

- *Property 1:* $\mathrm{sw}(a,U) = \mathrm{sw}(a,\tilde{U})$ and $\mathrm{sw}(b,U) = \mathrm{sw}(b,\tilde{U})$
- *Property 2:* $\Pi_{V(\boldsymbol{\gamma},U)} \subseteq \Pi_{V(\tilde{\boldsymbol{\gamma}},\tilde{U})}$.

Together these properties imply that $\mathrm{dist}(f, \boldsymbol{\gamma}, U) \leq \mathrm{dist}(f, \tilde{\boldsymbol{\gamma}}, \tilde{U})$.

Construction of $\tilde{U}$. Note that for all $i > 1$, we immediately have that $v_i(a, \boldsymbol{\gamma}, U) = v_i(a, \tilde{\boldsymbol{\gamma}}, U)$ and $v_i(b, \boldsymbol{\gamma}, U) = v_i(b, \tilde{\boldsymbol{\gamma}}, U)$. Then, if it is already the case voter 1's values under $U$ match ordinally across $\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}$ —that is $v_1(a, \boldsymbol{\gamma}, U) \geq v_1(b, \boldsymbol{\gamma}, U)$ and $v_1(a, \tilde{\boldsymbol{\gamma}}, U) \geq v_1(b, \tilde{\boldsymbol{\gamma}}, U)$, or $v_1(b, \boldsymbol{\gamma}, U) \geq v_1(a, \boldsymbol{\gamma}, U)$ and $v_1(b, \tilde{\boldsymbol{\gamma}}, U) \geq v_1(a, \tilde{\boldsymbol{\gamma}}, U)$ — then we are done: set $\tilde{U} = U$, and we automatically get properties 1 and 2.

Else, we have that $v_1(a, \boldsymbol{\gamma}, U) \geq v_1(b, \boldsymbol{\gamma}, U)$ and $v_1(b, \tilde{\boldsymbol{\gamma}}, U) \geq v_1(a, \tilde{\boldsymbol{\gamma}}, U)$, where moreover, one of these inequalities is strict. Then, we have the following facts:

**Fact G.2.3.** *By Lemma G.2.1 and $\tilde{\gamma}_1 < \gamma_1$, $sw(b, U) < sw(a, U)$.*

**Fact G.2.4.** *By the fact that $v_1(b, \tilde{\boldsymbol{\gamma}}, U) > v_1(a, \tilde{\boldsymbol{\gamma}}, U)$ and Fact G.2.3, $u_1(b) > u_1(a)$.*

Let $N'$ be the set of all voters $i$ for whom $u_i(a) > u_i(b)$. Note that by Fact G.2.3, $v_i(a, \boldsymbol{\gamma}, U) > v_i(b, \boldsymbol{\gamma}, U)$ for all $i \in N'$. We now show Equation (G.3), which states that in order for $\mathrm{sw}(a, U) \geq \mathrm{sw}(b, U)$, the gap between voters' utilities for $a$ and $b$ in $N'$ must at least compensate for the gap between voter 1's utilities for $b$ and $a$:

$$\begin{aligned} 0 \leq \mathrm{sw}(a, U) - \mathrm{sw}(b, U) &= -(u_1(b) - u_1(a)) + \sum_{i \in N \setminus \{1\}} (u_i(a) - u_i(b)) \\ &\leq -(u_1(b) - u_1(a)) + \sum_{i \in N'} (u_i(a) - u_i(b)). \end{aligned}$$

and we conclude

$$\sum_{i \in N'} (u_i(a) - u_i(b)) \geq u_1(b) - u_1(a). \tag{G.3}$$

Then, by Equation (G.3), there must exist some vector of non-negative real numbers $\boldsymbol{\delta} = (\delta_i : i \in N')$ such that

$$0 \leq \delta_i \leq u_i(a) - u_i(b) \text{ for all } i \in N' \qquad \text{and} \qquad \sum_{i \in N'} \delta_i \geq u_1(b) - u_1(a).$$

Fix this vector $\boldsymbol{\delta}$, and use it to construct $\tilde{U}$ in the following way: first, for all voters $i$, set $\tilde{u}_i(b) = u_i(b)$. Then, set voters' utilities for $a$ as follows:

- $\tilde{u}_1(a) = u_1(a) + \sum_{i \in N'} \delta_i$,
- for all $i \in N'$, $\tilde{u}_i(a) = u_i(a) - \delta_i$, and
- for all other $i$, $\tilde{u}_i(a) = u_i(a)$.

By inspection, per this construction we have *Property 1*: that $\mathrm{sw}(a, U) = \mathrm{sw}(a, \tilde{U})$ and $\mathrm{sw}(b, U) = \mathrm{sw}(b, \tilde{U})$.

Finally, we show *Property 2*, that $\Pi_{V(\boldsymbol{\gamma}, U)} \subseteq \Pi_{V(\tilde{\boldsymbol{\gamma}}, \tilde{U})}$. First for all voters $i \in N' \cup \{1\}$, we have by the construction above that $\tilde{u}_i(a) \geq \tilde{u}_i(b)$; By Property 1, we also have that $\mathrm{sw}(a, \tilde{U}) > \mathrm{sw}(b, \tilde{U})$.

Thus, $v_i(a, \tilde{\boldsymbol{\gamma}}, \tilde{U}) > v_i(b, \tilde{\boldsymbol{\gamma}}, \tilde{U})$. This is consistent with the fact that $v_i(a, \boldsymbol{\gamma}, U) > v_i(b, \boldsymbol{\gamma}, U)$ for all $i \in N' \cup \{1\}$, as fixed earlier in the proof. For all remaining voters $i \notin N' \cup \{1\}$, we did not change their utilities from $U$ to $\tilde{U}$, so we have that $v_i(a, \tilde{\boldsymbol{\gamma}}, \tilde{U}) = v_i(a, \boldsymbol{\gamma}, U)$ and $v_i(b, \tilde{\boldsymbol{\gamma}}, \tilde{U}) = v_i(b, \boldsymbol{\gamma}, U)$. We conclude that all PS-values are ordinally consistent for all voters across $V(\boldsymbol{\gamma}, U)$ and $V(\tilde{\boldsymbol{\gamma}}, \tilde{U})$, and thus $\Pi_{V(\boldsymbol{\gamma},U)} \subseteq \Pi_{V(\tilde{\boldsymbol{\gamma}},\tilde{U})}$, concluding the proof. □

**Proposition 12.4.5(b).** *When $m = 3$, all voting rules exhibit nonuniform monotonicity.*

*Proof.* Fix an arbitrary $f$. Fix $U$ and $\tilde{\boldsymbol{\gamma}} < \boldsymbol{\gamma}$ where $\tilde{\gamma}_1 < \gamma_1$ and $\tilde{\gamma}_i = \gamma_i$ for all $i > 1$. We will prove the claim by showing that we can find some other utility matrix $\tilde{U}$ so that $\mathsf{dist}(f, \boldsymbol{\gamma}, U) \leq \mathsf{dist}(f, \tilde{\boldsymbol{\gamma}}, \tilde{U})$.

For notational convenience, for any instance $(\boldsymbol{\gamma}, U)$ we will write $\boldsymbol{\pi}^{\boldsymbol{\gamma},U}$ to denote a profile compatible with $(\boldsymbol{\gamma}, U)$. Fix an arbitrary $\boldsymbol{\pi}^{\boldsymbol{\gamma},U}$, and fix another profile $\boldsymbol{\pi}^{\tilde{\boldsymbol{\gamma}},U}$ with the same tie-breaking when PS-values are equal. Note that these two profiles may differ only in voter 1's ranking (and if they don't, we can set $\tilde{U} = U$ and we are done). This proof will be conceptually similar to that of Proposition 12.4.5(a), except that instead of correcting one pairwise ranking, we must correct multiple in succession.

Define the $\mathsf{swap}(\pi, a, b)$ function as one that intakes a ranking and two alternatives that are ranked adjacently in $\pi$, and outputs the ranking in which they are swapped; e.g., $\mathsf{swap}(b \succ a, b, a) = a \succ b$. Now, define a sequence of unique pairwise swaps of alternatives adjacent in $\pi_1^{\tilde{\boldsymbol{\gamma}},U}$ such that, if made, would transform $\pi_1^{\tilde{\boldsymbol{\gamma}},U}$ into $\pi_1^{\boldsymbol{\gamma},U}$. Let this sequence be $(a_1, b_1), (a_2, b_2), \ldots, (a_T, b_T)$ where, by convention, we are swapping $b_t \succ a_t \to a_t \succ b_t$. That is, if we apply swap successively to $\pi_1^{\tilde{\boldsymbol{\gamma}},U}$ for alternatives $a_1, b_1 \ldots a_T, b_T$, we will get $\pi_1^{\boldsymbol{\gamma},U}$.

By Lemma G.2.5 (below), we can define a sequence of utility matrices $\tilde{U}_1, \tilde{U}_2, \ldots \tilde{U}_T$ such that

- **Property 1:** $\pi_i^{\tilde{U}_t,\tilde{\boldsymbol{\gamma}}} = \pi_i^{\boldsymbol{\gamma},U}$ for all $i \neq 1, t \in [T]$
  (the rankings of all voters other than 1 are preserved in step $t$)
- **Property 2** $\pi_1^{\tilde{U}_{t+1},\tilde{\boldsymbol{\gamma}}} = \mathsf{swap}(\pi_1^{\tilde{U}_t,\tilde{\boldsymbol{\gamma}}}, b_{t+1}, a_{t+1})$ for all $t \in [T-1]$
  (so 1's pairwise mis-ordering of $a_{t+1}$ and $b_{t+1}$ is corrected in the $t+1$-st step)
- **Property 3**: $\mathsf{sw}(a, U) = \mathsf{sw}(a, \tilde{U}_t)$ for all $a \in [m], t \in [T]$
  (the welfares are preserved in step $t$)

It follows that $\boldsymbol{\pi}^{\tilde{U}_T,\tilde{\boldsymbol{\gamma}}} = \boldsymbol{\pi}^{\boldsymbol{\gamma},U}$ and $\mathsf{sw}(a, U) = \mathsf{sw}(a, \tilde{U}_T)$, together implying that

$$\mathsf{dist}(f, \boldsymbol{\gamma}, U) = \mathsf{dist}(f, \tilde{\boldsymbol{\gamma}}, \tilde{U}_T),$$

concluding the proof. □

**Lemma G.2.5.** *Let $m = 3$. Fix arbitrary $U$ and $\boldsymbol{\gamma} > \tilde{\boldsymbol{\gamma}}$, where $\tilde{\gamma}_1 < \gamma_1$ but all other voters' entries are identical. Let alternatives $a, b$ be such that $a \succ_{\pi_1^{\boldsymbol{\gamma},U}} b$ and $b \succ_{\pi_1^{\tilde{\boldsymbol{\gamma}},U}} a$, where $a$ and $b$ are ranked adjacently in $\pi_1^{\tilde{\boldsymbol{\gamma}},U}$. Then, there exists a $\tilde{U}$ such that:*

- ***Property 1*** $\pi_i^{\tilde{\gamma},\tilde{U}} = \pi_i^{\gamma,U}$ *for all* $i \neq 1$
  *(the rankings of all voters other than 1 are preserved)*
- ***Property 2*** $\pi_1^{\tilde{\gamma},\tilde{U}} = swap(\pi_1^{\tilde{\gamma},U}, b, a)$
  *(so 1's pairwise mis-ordering of* $a$ *and* $b$ *is corrected)*
- ***Property 3****:* $sw(a, U) = sw(a, \tilde{U})$ *for all* $a \in [m]$
  *(the welfares are preserved)*

*Proof.* We begin by establishing a series of facts:

**Fact G.2.6.** *By the fact that* $a, b$ *were in the list of pairwise swaps,* $a \succ_{\pi_1^{\gamma,U}} b$ *and* $b \succ_{\pi_1^{\tilde{\gamma},U}} a$.

**Fact G.2.7.** *By Lemma G.2.1, the fact that* $b \succ_{\pi_1^{\tilde{\gamma},U}} a$, *and* $\tilde{\gamma}_1 < \gamma_1$,

$$sw(b, U) \leq sw(a, U).$$

**Fact G.2.8.** *By Fact G.2.6 and Fact G.2.7,*[1] *we have that*

$$u_1(b) > u_1(a).$$

Now, define our set of voters $N'$ as in the $m = 2$ proof, i.e., as the set of all voters $i \in [n]$ such that $u_i(a) > u_i(b)$ (and thus, given Fact G.2.7, $a \succ_{\pi_i^{U,\tilde{\gamma}}} b$). Then, we know that by the same argument as before, using Facts G.2.7 and G.2.8, that

$$\sum_{i \in N'} (u_i(a) - u_i(b)) > u_1(b) - u_1(a). \tag{G.4}$$

Now, we have that for all $i \in N'$, we have that $a \succ_{\pi_i^{U,\tilde{\gamma}}} b$. Let $c$ be the third, remaining alternative that is not equal to $a$ or $b$. Then, a voter $i \in N'$ can have one of three possible rankings in $\boldsymbol{\pi}^{U,\tilde{\gamma}}$:

$$(1)\ a \succ b \succ c, \quad (2)\ c \succ a \succ b, \text{ or } \quad (3)\ a \succ c \succ b.$$

We will now prove three claims, one per ranking, which will lay the foundations for our later construction of $\tilde{U}$. We use $N'_{(1)}$ to mean the set of voters in $N'$ with ranking (1), and likewise for (2) and (3).

***Claim 1:*** For all voters $i \in N'_{(1)}$ and for all $\delta_i^1 \in [0, u_i(a) - u_i(b)]$,

$$v_i(a, \tilde{\boldsymbol{\gamma}}, U) \geq (1 - \tilde{\gamma}_i)(u_i(b) + \delta_i) + \tilde{\gamma}_i \mathrm{sw}(b, U) \geq v_i(c, \tilde{\boldsymbol{\gamma}}, U).$$

[1] This is strict for the same reason as in the $m = 2$ case.

The first inequality holds by Fact G.2.7 combined with $\delta_i$ being defined in $[u_i(b), u_i(a)]$. The second inequality is implied by the fact that $v_i(b, \tilde{\boldsymbol{\gamma}}, U) \geq v_i(c, \tilde{\boldsymbol{\gamma}}, U)$, inferred from the fact that $b \succ_{\pi_i^{U,\tilde{\gamma}}} c$ (i.e., $i$ ranks $b$ ahead of $c$).

***Claim 2:*** For all voters $i \in N'_{(2)}$, and for all $\delta_i \in [0, u_i(a) - u_i(b)]$,

$$v_i(c, \tilde{\boldsymbol{\gamma}}, U) \geq (1 - \tilde{\gamma}_i)(u_i(a) - \delta_i) + \tilde{\gamma}_i \mathrm{sw}(a, U) \geq v_i(b, \tilde{\boldsymbol{\gamma}}, U).$$

*Proof of Claim 2:* The proof is essentially the same as that of Claim 1: The first inequality is implied by the fact that $v_i(c, \tilde{\boldsymbol{\gamma}}, U) \geq v_i(a, \tilde{\boldsymbol{\gamma}}, U)$, inferred from the fact that $i$ ranks $c$ ahead of $a$, and the second inequality holds by Fact G.2.7 combined with $\delta_i$ being defined in $[u_i(b), u_i(a)]$.

***Claim 3:*** For all voters $i \in N'_{(3)}$, there exists some $u^*$ in the following interval

$$\left[ u_i(c) + \frac{\tilde{\gamma}_i(\mathrm{sw}(c, U) - \mathrm{sw}(a, U))}{(1 - \tilde{\gamma}_i)n},\ u_i(c) + \frac{\tilde{\gamma}_i(\mathrm{sw}(c, U) - \mathrm{sw}(b, U))}{(1 - \tilde{\gamma}_i)n} \right],$$

such that $u^*$ is also in the interval $[u_i(b), u_i(a)]$ and satisfies

$$(1 - \tilde{\gamma}_i)u^* + \tilde{\gamma}_i \mathrm{sw}(a)/n \geq u_i^{\tilde{\gamma}}(c) \geq (1 - \tilde{\gamma}_i)u^* + \tilde{\gamma}_i \mathrm{sw}(b)/n. \tag{G.5}$$

*Proof of Claim 3:* By Fact G.2.7, the upper end of the interval is indeed at least the lower end, so there can exist a $u^*$, as this is a non-empty region of the real line. Second, fixing any $u^*$ in this interval, the chain of inequalities in (G.5) is proven by simply rearranging the given fact that $u^*$ is in the provided interval. Finally, the given interval must overlap the interval $[u_i(b), u_i(a)]$, so we can choose some $u^*$ within both intervals. We show in two steps. First, the upper end of the interval is weakly larger than $u_i(b)$:

$$\begin{aligned} v_i(b, \tilde{\boldsymbol{\gamma}}, U) \leq v_i(c, \tilde{\boldsymbol{\gamma}}, U) &\iff (1 - \tilde{\gamma}_i)u_i(b) + \tilde{\gamma}_i \mathrm{sw}(b, U)/n \leq (1 - \tilde{\gamma}_i)u_i(c) + \tilde{\gamma}_i \mathrm{sw}(c, U)/n \\ &\iff u_i(c) + \frac{\tilde{\gamma}_i(\mathrm{sw}(c, U) - \mathrm{sw}(b, U))}{(1 - \tilde{\gamma}_i)n} \geq u_i(b). \end{aligned}$$

And the lower end of the interval is at most $u_i(a)$:

$$\begin{aligned} v_i(a, \tilde{\boldsymbol{\gamma}}, U) \leq v_i(c, \tilde{\boldsymbol{\gamma}}, U) &\iff (1 - \tilde{\gamma}_i)u_i(a) + \tilde{\gamma}_i \mathrm{sw}(a, U)/n \geq (1 - \tilde{\gamma}_i)u_i(c) + \tilde{\gamma}_i \mathrm{sw}(c, U)/n \\ &\iff u_i(c) + \frac{\tilde{\gamma}_i(\mathrm{sw}(c, U) - \mathrm{sw}(b, U))}{(1 - \tilde{\gamma}_i)n} \leq u_i(a). \end{aligned}$$

*End of proof of Claim 3.*

**Claim 4 (Corollary of Claim 3).** For arbitrary $u^*$ satisfying the conditions of Claim 3, for all $\delta_i^{3,a} \in [0, u_i(a) - u^*], i \in N'_{(3)}$ and all $\delta_i^{3,b} \in [0, u^* - u_i(b)], i \in N'_{(3)}$, we have that

$$\begin{aligned}(1-\tilde{\gamma}_i)(u_i(a) - \delta_i^{3,a}) + \tilde{\gamma}_i \mathrm{sw}(a)/n &\geq (1-\tilde{\gamma}_i)u^* + \tilde{\gamma}_i \mathrm{sw}(a)/n && (G.5)\\ &\geq v_i(c, \tilde{\boldsymbol{\gamma}}, U)\\ &\geq (1-\tilde{\gamma}_i)u^* + \tilde{\gamma}_i \mathrm{sw}(b)/n && (G.5)\\ &\geq (1-\tilde{\gamma}_i)(u_i(b) + \delta^{3,b}) + \tilde{\gamma}_i \mathrm{sw}(a)/n.\end{aligned}$$

Choosing the $\boldsymbol{\delta}$s. Taking the $\delta_i^{1,a}$, $\delta_i^{2,b}$, $\delta_i^{3,a}$ and $\delta_i^{3,b}$ and their domains from Claims 1, 2, and 4, we have that

$$\begin{aligned}0 &\leq \sum_{i \in N'_{(1)}} \delta_i^{1,a} + \sum_{i \in N'_{(2)}} \delta_i^{2,b} + \sum_{i \in N'_{(3)}} (\delta_i^{3,a} + \delta_i^{3,b})\\ &\leq \sum_{i \in N'_{(1)} \cup N'_{(2)}} (u_i(a) - u_i(b)) + \sum_{i \in N'_{(3)}} (u_i(a) - u^*) + (u^* - u_i(b))\\ &= \sum_{i \in N'} u_i(a) - u_i(b)\\ &> u_1(b) - u_1(a). && (G.4)\end{aligned}$$

Thus, for any constant $t \in [0, u_1(b) - u_1(a)]$, there must exist settings of these deltas so that their sum over $i \in N'$ is equal to $t$. We will choose $\delta^*$ values $\delta_i^{*1,a}$ for all $i \in N_{(1)}$, $\delta_i^{*2,b}$ for all $i \in N_{(2)}$, $\delta_i^{*3,a}$ and $\delta_i^{*3,b}$ for all $i \in N_{(3)}$, so that they add up to

$$t^* = u_1(b) - u_1(a) - \frac{\tilde{\gamma}}{1-\tilde{\gamma}}(\mathrm{sw}(a, U) - \mathrm{sw}(b, U))/n \qquad (G.6)$$

Note that this value falls in the permitted range as it is clearly at most $u_1(b) - u_1(a)$, and it is at least 0 by a simple rearrangement of the known inequality $v_i(b, \tilde{\boldsymbol{\gamma}}, U) \geq v_i(a, \tilde{\boldsymbol{\gamma}}, U)$.

Construction of $\tilde{U}$.

- For all $i \notin N' \cup \{1\}$, set $i$'s utilities in $\tilde{U}$ as in $U$, i.e., $\tilde{u}_i(a) = u_i(a)$ and likewise for $b$ and $c$.
- For all $i \in N'$ set $\tilde{u}_i(c) = u_i(c)$, and
  - for $i \in N'_{(1)}$, set
$$\begin{aligned}\tilde{u}_i(a) &= u_i(a) - \delta_i^{*1,a},\\ \tilde{u}_i(b) &= u_i(b)\end{aligned}$$
  - for $i \in N'_{(2)}$, set
$$\begin{aligned}\tilde{u}_i(a) &= u_i(a)\\ \tilde{u}_i(b) &= u_i(b) + \delta_i^{*2,b}\end{aligned}$$

– for $i \in N'_{(3)}$, set

$$\tilde{u}_i(a) = u_i(a) - \delta_i^{*3,a}$$
$$\tilde{u}_i(b) = u_i(b) + \delta_i^{*3,b}.$$

- For voter 1, set $\tilde{u}_1(c) = u_1(c)$, and set

$$\tilde{u}_1(a) = u_1(a) + \sum_{i \in N'_{(1)}} \delta_i^{*1,a} + \sum_{i \in N'_{(3)}} \delta_i^{*3,a}$$
$$\tilde{u}_1(b) = u_1(b) - \sum_{i \in N'_{(2)}} \delta_i^{*2,b} - \sum_{i \in N'_{(3)}} \delta_i^{*3,b}.$$

By construction, all utilities are nonnegative.

$\tilde{U}$ satisfies Property 3. We want to show that $\mathrm{sw}(a, \tilde{U}) = \mathrm{sw}(a, U)$, and likewise for alternatives $b$ and $c$. This is true for $c$ by inspection, as for all $i \in [n]$, $\tilde{u}_i(c) = u_i(c)$. For $a$ and $b$, the argument is also by inspection, noting that the utility added or subtracted among the $N'$ group for either alternative is exactly compensated by the change to voter 1's utility for that alternative.

$\tilde{U}$ satisfies Property 1. We need to conclude that by our construction, all voters' other than 1's rankings were preserved, i.e., $\pi_i^{\tilde{\gamma},U} = \pi_i^{\tilde{\gamma},\tilde{U}}$ for all $i \neq 1$. We will confirm this by group:

- For all $i \notin N' \cup \{1\}$, this holds simply by the fact that $\tilde{U}$ satisfies Property 3, $\tilde{\gamma}_i = \gamma_i$, and $\tilde{u}_i(a) = u_i(a)$, $\tilde{u}_i(b) = u_i(b)$, and $\tilde{u}_i(c) = u_i(c)$.
- For all $i \in N'$, this follows from claims 1, 2, and 3 and the fact that we set the $\delta$s as specified according to the conditions of those claims.

$\tilde{U}$ satisfies Property 2. This is implied by the fact that $v_1(a, \tilde{\boldsymbol{\gamma}}, \tilde{U}) = v_1(b, \tilde{\boldsymbol{\gamma}}, \tilde{U})$, which we will prove now. First, we will the following equality using (G.6):

$$\tilde{u}_1(b) - \tilde{u}_1(a) = u_1(b) - u_1(a) - t^* = \frac{\tilde{\gamma}}{1 - \tilde{\gamma}}(\mathrm{sw}(a, U) - \mathrm{sw}(b, U))/n.$$

Then, applying this equality,

$$\begin{aligned} v_1(b, \tilde{\boldsymbol{\gamma}}, \tilde{U}) - v_1(a, \tilde{\boldsymbol{\gamma}}, \tilde{U}) &= (1 - \tilde{\gamma}_1)(\tilde{u}_1(b) - \tilde{u}_1(a)) + \tilde{\gamma}_1(\mathrm{sw}(b), U) - \mathrm{sw}(a, U))/n \\ &= (1 - \tilde{\gamma}_1) \cdot \frac{\tilde{\gamma}_1}{1 - \tilde{\gamma}_1}(\mathrm{sw}(a, U) - \mathrm{sw}(b, U)/n + \tilde{\gamma}_1(\mathrm{sw}(b), U) - \mathrm{sw}(a, U))/n \\ &= 0. \end{aligned}$$

□

### G.2.2 Proof of Proposition 12.4.6

**Proposition G.2.9.** *Copeland is nonuniform PS-monotonic.*

*Proof.* Let $f = \text{Copeland}$. Since the case $m \leq 3$ is covered by Proposition 12.4.5, we may assume here that $m \geq 4$. For notational convenience, for any instance $(\boldsymbol{\gamma}, U)$ we will write $\boldsymbol{\pi}^{\boldsymbol{\gamma},U}$ to denote a profile compatible with $(\boldsymbol{\gamma}, U)$.

It suffices to show that when a single voter's public spirit level is decreased, the worst-case distortion weakly increases. Suppose this voter is voter 1, and that their public spirit is decreased from $\gamma_1$ to $\tilde{\gamma}_1$, corresponding to a change from PS-vector $\boldsymbol{\gamma}$ to $\tilde{\boldsymbol{\gamma}}$ (all else kept the same). To prove monotonicity, it suffices to prove that for an arbitrary utility matrix $U$, we can find a utility matrix $\tilde{U}$ such that the winner $a'$ remains the same (i.e., $a' = \text{Copeland}(\boldsymbol{\pi}^{\boldsymbol{\gamma},U}) = \text{Copeland}(\boldsymbol{\pi}^{\tilde{\boldsymbol{\gamma}},\tilde{U}})$), and such that $\text{sw}(a', U) = \text{sw}(a', \tilde{U})$, $\text{sw}(a^*, U) = \text{sw}(a^*, \tilde{U})$. We make a case distinction now on whether $a'$ pairwise-dominates $a^*$ in $\boldsymbol{\pi}^{\boldsymbol{\gamma},U}$.

Case 1: If $a'$ strictly pairwise-dominates $a^*$ in $\boldsymbol{\pi}^{\boldsymbol{\gamma},U}$, then define $\tilde{U}$ such that for all $i \in [n]$,

- $\tilde{u}_i(a) = 0$ for all $a \notin \{a', a^*\}$
- $\tilde{u}_i(a) = u_i(a)$ for all $a \in \{a', a^*\}$

Now, we argue that $\text{dist}(\text{Copeland}, \boldsymbol{\gamma}, U) = \text{dist}(\text{Copeland}, \tilde{\boldsymbol{\gamma}}, \tilde{U})$:

Observation 1. The welfares of $a', a^*$ are preserved across $U, \tilde{U}$, i.e., $\text{sw}(a', U) = \text{sw}(a', \tilde{U})$, $\text{sw}(a^*, U) = \text{sw}(a^*, \tilde{U})$.

Observation 2. for all voters $i \neq 1$, $i$ has the same relative ordering of $a'$,$a^*$ in $\boldsymbol{\pi}^{\boldsymbol{\gamma},U}$ and $\boldsymbol{\pi}^{\tilde{\boldsymbol{\gamma}},\tilde{U}}$. This is because from $\boldsymbol{\gamma}, U$ to $\tilde{\boldsymbol{\gamma}}, \tilde{U}$, $a', a^*$'s average utilities don't change, $i$'s utilities for $a', a^*$ don't change, and $\gamma_i$ doesn't change, meaning that $v_i(a', \boldsymbol{\gamma}, U) = v_i(a', \tilde{\boldsymbol{\gamma}}, \tilde{U})$ and $v_i(a^*, \boldsymbol{\gamma}, U) = v_i(a^*, \tilde{\boldsymbol{\gamma}}, \tilde{U})$.

Observation 3. In $\boldsymbol{\pi}^{\tilde{\boldsymbol{\gamma}},\tilde{U}}$, $a'$ and $a^*$ pairwise-dominate all $a \notin \{a', a^*\}$. This is because all voters must rank $a', a^*$ in the first two positions and all the other alternatives in positions $3 \ldots m$, by virtue of the fact that we can wlog assume that some voter has nonzero utility for $a^*$ (else the distortion will be 0), and thus some voter has nonzero utility for $a'$ (since it is sometimes ranked ahead of $a'$). In contrast, all other alternatives have average utility 0, and thus must be ranked behind $a', a^*$.

Observation 4. $a'$ pairwise-dominates $a^*$ in $\boldsymbol{\pi}^{\tilde{\boldsymbol{\gamma}},\tilde{U}}$. If $a^* \succ_{\pi_1^{\boldsymbol{\gamma},U}} a'$, then either voter 1's ranking is preserved, or $a' \succ_{\pi_i^{\tilde{\boldsymbol{\gamma}},\tilde{U}}} a^*$, which can only strengthen $a'$'s pairwise domination of $a^*$. Conversely, if $a' \succ_{\boldsymbol{\pi}^{\boldsymbol{\gamma},U}} a^*$, $a^*$ cannot overtake $a'$ by Lemma G.2.1.

These four observations, taken together, imply that in $\boldsymbol{\pi}^{\tilde{U},\tilde{\boldsymbol{\gamma}}}$, $a'$ still pairwise-dominates $a^*$, and moreover, both $a'$ and $a^*$ pairwise-dominate everything else. We conclude that the uncovered set is $\{a'\}$, and thus $a'$ is the unique winner. By Observation 1, this directly implies that the distortion is preserved across $(\boldsymbol{\gamma}, U)$ and $(\tilde{\boldsymbol{\gamma}}, \tilde{U})$.

Case 2: Now, suppose $a'$ does not strictly dominate $a^*$. We may without loss of generality assume that $a^*$ is not a Copeland winner — indeed, if it were, then for this $U$ we would have $\text{dist}(\text{Copeland}, \boldsymbol{\gamma}, U) = 1$, in which case the distortion can only increase when voter 1's PS-level is dropped.

When $a^*$ is not a Copeland winner, it has a strictly lower Copeland score than $a'$, and thus there must exist some alternative $b$ such that $a'$ *strictly* pairwise-dominates $b$ and $b$ weakly pairwise dominates $a^*$. We now construct $\tilde{U}$ from $U$ in three steps. In the first step, for all alternatives $a \notin \{a', a^*, b\}$ and all voters $i \in [n]$, we set $\tilde{u}_i(a) = 0$. For all voters $i \neq 1$, we set their utilities in $\tilde{U}$ for $a', a^*, b$ to be the same as in $U$.

In the second step, we set the utilities for $a^*, a', b$ for voter $i = 1$, depending on the following case distinction.

- Suppose $\text{sw}(b, U) > \text{sw}(a', U)$.
  - In this case, the social welfares are ordered $\text{sw}(a^*, U) \geq \text{sw}(b, U) > \text{sw}(a', U)$, while the above pairwise wins are

    $$a' \xrightarrow{\text{strictly}} b \xrightarrow{\text{weakly}} a^*.$$

    Since dropping $\gamma_1$ can only promote lower-welfare alternatives, we keep the same utilities for voter 1, and these pairwise wins will continue to hold.
  - Now, suppose $\text{sw}(b, U) \leq \text{sw}(a', U)$.
    - In this case, we know that dropping $\gamma_1$ can lead to the following promotions in 1's ranking: $b$ over $a'$, $b$ over $a^*$, or $a'$ over $a^*$. The last one doesn't concern us, as the promotion of $a'$ only helps $a'$ win, and the second-last one does not concern us because it will just strengthen the existing pairwise win of $b$ versus $a^*$. Thus, as long as the first promotion doesn't occur, we keep the same utilities as before.
    - If $b$ is promoted over $a'$, we drop its utility to $\tilde{u}_1(b) = 0$ for voter 1 (then, it will not be promoted, leaving only the option of promoting $a'$ over $a^*$). Then, if there exists someone who ranks $b$ ahead of $a'$, we add this utility to someone who ranks $b$ ahead of $a'$. If the person ranks $b$ ahead of $a^*$, this preserves their exact ranking; if they rank $b$ behind $a^*$, this may result in a strengthening of the pairwise defeat of $a^*$ by $b$, which does not change the Copeland winner. Else, if there is no one who ranks $b$ ahead of $a'$, then $b$ dominating $a'$ pairwise is not possible by changing any single person's ranking, so add this utility arbitrarily.

Finally, in the third step, we add identical copies of the 'intermediate' alternative $b$, to make $a'$ the unique Copeland winner. Again, we need a case distinction.

- $n$ **is even.** We take an 'empty' alternative $\bar{b} \in [m] \setminus \{a', a^*, b\}$ for which we previously set the utilities to 0, and re-set its utilities to be identical to $b$. We moreover choose the preference profile where any individual's preference between $a', a^*, \bar{b}$ is identical to the preference between $a', a^*, \bar{b}$ (i.e. $b, \bar{b}$ are always neighbouring in any $\pi_i$), and that $b, \bar{b}$ are in a tie (i.e. $|\{i : b \succ b'\}| = n/2$). In this constellation, $a'$ at least pairwise beats $b$ and $\bar{b}$ ($\geq 2$ points), $b$ and $\bar{b}$ at best pairwise beat $a^*$ ($\leq 1$ point), and $a^*$ at best beats $a'$ ($\leq 1$ point), so $a'$ is the winner.
- $n$ **is odd and** $m \geq 5$**.** Since we are unable to create pairwise ties when $n$ is odd, we have to treat this case separately. Let us assume first that $m \geq 5$. Then, we have at least two 'empty' alternatives for which we previously set the utilities to 0; let us call these $\bar{b}, \tilde{b} \in [m] \setminus \{a', a^*, b\}$. We then re-set the utilities for $\bar{b}, \tilde{b}$ to be identical to $b$, such that they are ranked relative to $a', a^*$ the same as $b$ by any individual. We moreover order $b, \bar{b}, \tilde{b}$ in so that they form a Condorcet cycle, and

$$b \xrightarrow{\text{strictly}} \bar{b} \xrightarrow{\text{strictly}} \tilde{b} \xrightarrow{\text{strictly}} b.$$

  Note that we may do so freely, since all three alternatives are identical.

  In this scenario, the Copeland scores are

  - $a'$ gets 3 points (for beating $b, \bar{b}, \tilde{b}$),
  - $a^*$ gets 1 point (for beating $a'$),
  - $b, \bar{b}, \tilde{b}$ get 2 points,

  whence $a'$ wins.
- $n$ **is odd and** $m = 4$**.** The previous arguments held for the Copeland rule with arbitrary tie-breaking between alternatives with identical Copeland score. In the specific case of $n$ being odd and $m = 4$, we need to make a slight refinement to our definition of distortion, namely that the distortion is a supremum over the whole Copeland set $CS(\boldsymbol{\pi}^{\boldsymbol{\gamma},U})$ for any $(\boldsymbol{\gamma}, U)$-compatible profile $\boldsymbol{\pi}^{\boldsymbol{\gamma},U}$.

$$\text{dist}(\text{Copeland}, \boldsymbol{\gamma}, U) = \sup_{a \in CS(\boldsymbol{\pi}^{\boldsymbol{\gamma},U})} \frac{\text{sw}(a^*, U)}{\text{sw}(a, U)}.$$

  It then suffices to ensure that $a'$ is *one* of the Copeland winners under $(\tilde{U}, \tilde{\gamma})$, not the unique one. Let the four alternatives be called $a', a^*, b, \bar{b}$. Since (i) $n$ is odd, (ii) we assumed that $a'$ does not strictly pairwise dominate $a^*$ and since we assumed that $a^*$ is not a Copeland winner, we can deduce that

  - $a'$ has exactly Copeland score 2 (for beating $b, \bar{b}$.)

- $a^*$ has exactly Copeland score 1 (for beating $a'$.)
- There exist exactly two elements in the Copeland set (alternatives with score 2), suppose that $b$ is this element.
- Note that this $b$ is an admissible choice in the second step, We assume that it was chosen in the second step.

After the second step, we may here create a $\bar{b}$ identical to $b$, and suppose that $b$ pairwise beats $\bar{b}$. Then the Copeland set will again consist of the same alternatives $\{a', b\}$. Since the welfare of $b$ was preserved in the second step, the proof is now complete.

□

### G.2.3 Proof of Proposition 12.4.7

**Proposition G.2.10.** *Plurality is nonuniform PS-monotonic.*

*Proof.* For notational convenience, for any instance $(\boldsymbol{\gamma}, U)$ we will write $\boldsymbol{\pi}^{\boldsymbol{\gamma},U}$ to denote any profile compatible with $(\boldsymbol{\gamma}, U)$.

It suffices to prove that when a single voter's public spirit level is decreased, the worst-case distortion increases. Suppose this voter is voter $i = 1$, and that $\gamma_1$ is changed from some value $\gamma_1 = \rho$ (Scenario 1) is changed to some lower value $\gamma_1 = \tilde{\rho} < \rho$ (Scenario 2). Let us denote by $\boldsymbol{\gamma}$ the original PS-vector (with $\gamma_1 = \rho$), and by $\tilde{\boldsymbol{\gamma}}$ the one which arises from lowering $\gamma_1$ to $\tilde{\rho}$. To prove monotonicity, it suffices to prove that for any utility matrix $U \in \mathsf{R}^{n\times m}$, we can find a utility matrix $\tilde{U}$ such that

1. the winner remains the same, $a' = f(\boldsymbol{\pi}^{\boldsymbol{\gamma},U}) = f(\boldsymbol{\pi}^{\tilde{\boldsymbol{\gamma}},\tilde{U}})$,
2. the social welfares of $a', a^*$ are preserved, i.e.

$$\mathsf{sw}(a', U) = \mathsf{sw}(a', \tilde{U}),\ \mathsf{sw}(a^*, U) = \mathsf{sw}(a^*, \tilde{U}).$$

If voter 1's first-ranked alternative remains unchanged, there is nothing to prove, so let us assume the first-ranked alternative does change – let us denote by $\pi_1^{\boldsymbol{\gamma},U}(1) = a$ the alternatives which receive voter 1's vote in scenario 1 such that voter 1's rankings are of the form

- Scenario 1: $a \succ$ all other alternatives,
- Scenario 2: alternatives $A_1 \succ a \succ$ alternatives $A_2$.

Since the second ranking arises from the first ranking by lowering $\tilde{\gamma}$, only *alternatives with lower welfare* can be promoted over $a$, i.e. $A_1$ consists of alternatives with welfare below $\mathsf{sw}(a, U)$.

We construct $\tilde{U}$ from $U$ in two steps. First, we set voter 1's utility for all alternatives in $A_1$ to zero. Since all those alternatives have lower welfare than $a$, this will restore $a$ as voter 1's first-ranked alternative. Since the highest-welfare alternative $a^*$ cannot have not been promoted over $a$, i.e. $a^* \in A_2$, its welfare remains unchanged.

This second step is to restore *some* of the welfares of alternatives in $A_1$ which were affected by the previous step. Specifically, let $\bar{a} \in A_1$. If there is a non-empty set $N_{\bar{a}} \subseteq [n]$, $|N_{\bar{a}}| \geq 1$ of voters (in Scenario 1) who rank $\bar{a}$ first, we add an $u_1(\bar{a})/|N_{\bar{j}}|$ amount of utility to all the voters in $N_{\bar{a}}$,

$$\tilde{u}_i(\bar{a}) = u_i(\bar{a}) + \frac{u_1(\bar{a})}{|N_{\bar{a}}|}, \quad \forall i \in N_{\bar{a}}.$$

If on the other hand $\bar{a}$ is ranked first by no voter, we do not intervene.

We claim that these two steps combined restore the first-ranked alternatives of all voters, and thus the winner of the election. To see this, we notice the following.

- **Welfares.** For any $\bar{a} \in A_1$ with $N_{\bar{a}} \neq \emptyset$, $\mathrm{sw}(\bar{a}, U) = \mathrm{sw}(\bar{a}, \tilde{U})$. The other alternatives $\bar{a} \in A_1$ with $N_{\bar{a}} = \emptyset$ may have lower welfare $\mathrm{sw}(\bar{a}, \tilde{U}) \leq \mathrm{sw}(\bar{a}, U)$. The welfares of alternatives in $A_2$, in particular of $a^* \in A_2$, remain unchanged.
- **Voters with first-choice in $A_1$.** If a voter first-ranks some alternative $\bar{a} \in \tilde{A}_1$ in Scenario 1 $(\boldsymbol{\gamma}, U)$, then they still do so in Scenario 2 $(\tilde{\boldsymbol{\gamma}}, \tilde{U})$, since they have added utility for $\bar{a}$ while the welfares of all other alternatives are either the same or lower.
- **Voters with first-choice in $\{a\} \cup A_2$.** Suppose a voter first-ranks some $\bar{a} \in \{a\} \cup A_2$ under $(\boldsymbol{\gamma}, U)$. Then, since both their utility and welfare for $\bar{a}$ are the same under $(\tilde{\boldsymbol{\gamma}}, \tilde{U})$ while the welfares of other alternatives can only have decreased, they continue to first-rank $\bar{a}$ under $(\tilde{\boldsymbol{\gamma}}, U)$.

This concludes the proof. □

### G.2.4 Proof of Lemma 12.4.13

**Lemma G.2.11.** *If $f$ is weakly unanimous and instance-wise PS-monotonic, then it is monotonic.*

*Proof.* Suppose that $f$ is weakly unanimous but not monotonic; we will show that it is not instance-wise PS-monotonic. Fix a pair of profiles $\boldsymbol{\pi}$, $\boldsymbol{\pi}'$ in which monotonicity is violated, i.e., where there exists some voter $i^* \in [n]$ such that $a$ is promoted via an adjacent swap in $\pi'_{i^*}$ compared to $\pi_{i^*}$, but $f(\boldsymbol{\pi}) = a$ and $f(\boldsymbol{\pi}') = b$. Let $\tilde{a}$ be the alternative over which $a$ is promoted from $\pi_{i^*}$ to $\pi'_{i^*}$.

Given that $f(\boldsymbol{\pi}) = a$ and the fact that $f$ is weakly unanimous, for every $c \neq a$, there must exist some voter $i_c$ such that $a \succ_{\pi_{i_c}} c$. Arbitrarily choose one such voter per $c$ and denote them $i_c$, for all $c \neq a$. Note that it is possible that some such $i_c = i^*$; we will handle this in the proof.

Now, we will construct a pair of instances $\boldsymbol{\gamma}, U$ and $\boldsymbol{\gamma}, U'$ such that $\boldsymbol{\gamma}'$ differs from $\boldsymbol{\gamma}$ only in that $\gamma'_{i^*} \succ \gamma_{i^*}$, and that three claims hold: *Claim (1):* $\mathrm{dist}(f, \boldsymbol{\gamma}', U) > \mathrm{dist}(f, \boldsymbol{\gamma}, U)$, *Claim (2):* $\boldsymbol{\pi} \in \Pi_{V(\boldsymbol{\gamma}, U)}$, *Claim (3):* $\boldsymbol{\pi}' \in \Pi_{V(\boldsymbol{\gamma}, U')}$. Together, these claims constitute a violation of instance-wise PS-monotonicity.

**Construction of $\boldsymbol{\gamma}$, $\boldsymbol{\gamma}'$:** Let $\boldsymbol{\gamma} = \mathbf{0}$ (i.e., all voters have public spirit level 0). Let $\boldsymbol{\gamma}'$ be defined such that $\gamma'_i = \gamma_i = 0$ for all $i \neq i^*$, and let $\gamma'_{i^*} = \epsilon$, where $\epsilon > 0$ is set to some number smaller than $1/2m^2$.

**Construction of $U$:**

- Group 1: For all voters $i \neq i^*$ and $i \notin \{i_c | c \in [m] \setminus \{a\}\}$, let $i$ have 0 utility for all alternatives.
- Group 2: For all voters $i \neq i^*$ and $i \in \{i_c | c \in [m] \setminus \{a\}\}$, let $i$ have utility 1 for $a$ and all alternatives ranked ahead of $a$ in $\pi_i$, and 0 for all other alternatives.
- For $i^*$: starting at the first-ranked alternative in $\pi_{i^*}$, assign utilities starting at 1 and let them descend at intervals of $1/m^2$ until we reach alternative $a$. Then, assign $u_{i^*}(a)$ so that $u_{i^*}(\tilde{a}) - u_{i^*}(a) = \epsilon^2/n$. Now, continuing in order of the $\pi_{i^*}$ after $a$, continue assigning alternatives utilities descending at intervals of $1/m^2$.

**Proof of Claims (1), (2), and (3):**

*Claim (1):* We prove this by proving that $a$ has strictly higher social welfare than any other alternative. Then, the winner changing from $a$ to $b$ from $\boldsymbol{\pi}$ to $\boldsymbol{\pi}'$ must increase the distortion, i.e., $\text{dist}(f, \boldsymbol{\gamma}', U) > \text{dist}(f, \boldsymbol{\gamma}, U)$.

First, if there is no $c$ such that $i_c = i^*$, then for all $c \neq a$, we have that $\sum_{i \in \text{ Group } 1}(u_i(a) - u_i(c)) = 0$, $\sum_{i \in \text{ Group } 2}(u_i(a) - u_i(c)) \geq 1$, and $u_{i^*}(a) - u_{i^*}(c) \geq -1/m$. Thus, $\sum_{i \in [n]}(u_i(a) - u_i(c)) > 0$, equivalent to $\text{sw}(a, U) > \text{sw}(c, U)$.

If there exists $c^*$ such that $i_{c^*} = i^*$, then the previous case holds for all $c \neq c^*$. For $c^*$, we repeat the above analysis: $\sum_{i \in \text{ Group } 1}(u_i(a) - u_i(c^*)) = 0$, $\sum_{i \in \text{ Group } 2}(u_i(a) - u_i(c^*)) \geq 0$, and $u_{i^*}(a) - u_{i^*}(c^*) \geq 1/m^2$, the final inequality by the fact that $a \succ_{\pi_{i^*}} c^*$. Thus, again $\sum_{i \in [n]}(u_i(a) - u_i(c^*)) > 0$, equivalent to $\text{sw}(a, U) > \text{sw}(c^*, U)$.

*Claim (2):* We have assigned voters' utilities in weakly decreasing order according to $\pi_i$ for all $i$, and $\boldsymbol{\gamma} = \mathbf{0}$, meaning that voters' individual utilities fully determine their rankings: thus, $\boldsymbol{\pi} \in \Pi_{V(\boldsymbol{\gamma}, U)}$.

*Claim (3):* The high level proof of this claim is the following: First, for all voters $i \neq i^*$, their $\gamma_i = \gamma'_i$, so their rankings implied by $U\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}, U'$ are the same, as is consistent with $\pi_i = \pi'_i$. For voter $i^*$, the separation between the utilities for all pairs of alternatives other than $\tilde{a}, a$ are too large for an $\epsilon$ increase in public spirit to flip them; however, the separation between the utilities of $\tilde{a}, a$ are small enough for this increase to flip them, realizing the transformation from $\pi_{i^*} \to \pi'_{i^*}$.

Building on the notation of $\boldsymbol{\pi} \in \Pi_{V(\boldsymbol{\gamma}, U)}$ (meaning that the profile $\boldsymbol{\pi}$ is consistent with the instance $U, \boldsymbol{\gamma}$), we use $\pi_i \in \Pi_{V_i(\boldsymbol{\gamma}, U)}$ to mean a voter $i$'s ranking $\pi_i$ is consistent with the vector of PS-values implied by the $i$th row of the matrix $V(\boldsymbol{\gamma}, U)$.

For all voters $i \neq i^*$, by construction of $\boldsymbol{\pi}, \boldsymbol{\pi}'$ we have that $\pi_i = \pi'_i$. Moreover, $\gamma_i = \gamma'_i$ implies that $\Pi_{V_i(\boldsymbol{\gamma}, U)} = \Pi_{V_i(\boldsymbol{\gamma}, U')}$. By these two equalities, $\pi_i \in \Pi_{V_i(\boldsymbol{\gamma}, U)}$ (as shown in Claim (2)) implies $\pi'_i \in \Pi_{V_i(\boldsymbol{\gamma}, U')}$.

Now, it only remains to show that $\pi'_{i^*} \in \Pi_{V_{i^*}(\boldsymbol{\gamma},U')}$. First, we observe that for all alternatives $c$,

$$|v_{i^*}(c,\boldsymbol{\gamma},U) - v_{i^*}(c,\boldsymbol{\gamma}',U)| = |u_{i^*}(c) - (1-\epsilon)u_{i^*}(c) - \epsilon \mathrm{sw}(c,U)/n| = \epsilon |u_{i^*}(c) - \mathrm{sw}(c,U)/n| \leq \epsilon, \quad \text{(G.7)}$$

where the final step holds because all utilities in $U$ are bounded between 0 and 1.

Next, we observe that for all pairs of alternatives $(c,c') \neq (\tilde{a},a)$, we have that

$$|v_{i^*}(c,\boldsymbol{\gamma},U) - v_{i^*}(c',\boldsymbol{\gamma},U)| = |u_{i^*}(c) - u_{i^*}(c)| \geq 1/m^2 > 2\epsilon. \quad \text{(G.8)}$$

Now, fix an arbitrary pair of alternatives $(c,c') \neq (\tilde{a},a)$ such that $c \succ_{\pi_{i^*}} c'$, and thus $v_{i^*}(c,\boldsymbol{\gamma},U) \geq v_{i^*}(c',\boldsymbol{\gamma},U)$. Then, by Equations (G.7) and (G.8) we have that $v_{i^*}(c,\boldsymbol{\gamma}',U) \geq v_{i^*}(c',\boldsymbol{\gamma}',U)$:

$$\begin{aligned}
0 &< v_{i^*}(c,\boldsymbol{\gamma},U) - v_{i^*}(c',\boldsymbol{\gamma},U) - 2\epsilon && \text{by (G.8)}\\
&\leq v_{i^*}(c,\boldsymbol{\gamma},U) - v_{i^*}(c',\boldsymbol{\gamma},U) - |v_{i^*}(c,\boldsymbol{\gamma},U) - v_{i^*}(c,\boldsymbol{\gamma}',U)| - |v_{i^*}(c',\boldsymbol{\gamma}',U) - v_{i^*}(c',\boldsymbol{\gamma},U)| && \text{by (G.7)}\\
&\leq v_{i^*}(c,\boldsymbol{\gamma},U) - v_{i^*}(c',\boldsymbol{\gamma},U) - (v_{i^*}(c,\boldsymbol{\gamma},U) - v_{i^*}(c,\boldsymbol{\gamma}',U)) - (v_{i^*}(c',\boldsymbol{\gamma}',U) - v_{i^*}(c',\boldsymbol{\gamma},U))\\
&= v_{i^*}(c,\boldsymbol{\gamma}',U) - v_{i^*}(c',\boldsymbol{\gamma}',U)
\end{aligned}$$

We conclude that for all such pairs $(c,c')$,

$$v_{i^*}(c,\boldsymbol{\gamma},U) \geq v_{i^*}(c',\boldsymbol{\gamma},U) \implies v_{i^*}(c,\boldsymbol{\gamma}',U) \geq v_{i^*}(c',\boldsymbol{\gamma}',U). \quad \text{(G.9)}$$

Next, we consider the remaining pair $(a,\tilde{a})$. First, we observe that

$$\mathrm{sw}(a,U) - \mathrm{sw}(\tilde{a},U) > \epsilon,$$

by the fact that $\sum_{i \in \text{Group } 1}(u_i(a) - u_i(\tilde{a})) = 0$, $\sum_{i \in \text{Group } 2}(u_i(a) - u_i(\tilde{a})) \geq 1$ (note that it cannot be that, given the existence of a $c^* : i_{c^*} = i^*$, $c^* = \tilde{a}$, because we know that $\tilde{a} \succ_{\pi_{i^*}} a$), and $u_{i^*}(a) - u_{i^*}(\tilde{a}) = -\epsilon^2/n$. Adding up over voters, these inequalities imply that $\mathrm{sw}(a,U) - \mathrm{sw}(\tilde{a},U) \geq 1 - \epsilon^2/n > \epsilon$.

Then, we show the the inequality

$$v_{i^*}(a,\boldsymbol{\gamma}',U) > v_{i^*}(\tilde{a},\boldsymbol{\gamma}',U) \quad \text{(G.10)}$$

via the following deduction, where the first inequality uses that $u_{i^*}(a) - u_{i^*}(\tilde{a}) = -\epsilon^2/n < 0$, and the second inequality uses that $\mathrm{sw}(a,U) - \mathrm{sw}(\tilde{a},U) \geq \epsilon$:

$$\begin{aligned}
v_{i^*}(a,\boldsymbol{\gamma}',U) - v_{i^*}(\tilde{a},\boldsymbol{\gamma}',U) &= (1-\epsilon)(u_{i^*}(a) - u_{i^*}(\tilde{a})) + \epsilon(\mathrm{sw}(a,U)/n - \mathrm{sw}(\tilde{a},U)/n)\\
&> -\epsilon^2/n + \epsilon(\mathrm{sw}(a,U) - \mathrm{sw}(\tilde{a},U))/n\\
&\geq -\epsilon^2/n + \epsilon \cdot \epsilon/n\\
&= 0.
\end{aligned}$$

By Equations (G.9) and (G.10), we have that any ranking $\pi$ with the following two properties must be consistent with $\Pi_{V_{i^*}(\boldsymbol{\gamma},U')}$: First, for all pairs of alternatives $(c,c') \neq (a,\tilde{a})$, $c \succ_{\pi_{i^*}} c' \implies c \succ_{\pi} c'$, and second, $a \succ_{\pi} \tilde{a}$. $\pi'_{i^*}$ satisfies these criteria by construction, and thus $\pi'_{i^*} \in \Pi_{V_{i^*}(\boldsymbol{\gamma},U')}$, as needed, concluding the proof. □

### G.2.5 Proof of Lemma 12.4.15

**Lemma G.2.12.** *If $f$ weakly unanimous and monotonic, then if $f$ is instance-wise PS-monotonic, it must also be swap-invariant.*

*Proof.* We will prove the contrapositive. Suppose $f$ is not swap-invariant. Then, there exists two profiles $\boldsymbol{\pi}, \boldsymbol{\pi}'$ that differ only in that for some voter $i^*$, $b$ and $c$ are adjacently swapped in their ranking, and $f(\boldsymbol{\pi}) = a$ but $f(\boldsymbol{\pi}') = b$. By the monotonicity of $f$, we know that $c \succ_{\pi_{i^*}} b$: otherwise, going from $\boldsymbol{\pi}' \to \boldsymbol{\pi}$, $b$ would be promoted over $c$ but lose the winning spot, violating monotonicity. Now, we will break into cases depending on the nature of $\boldsymbol{\pi}$, and in either case, show that PS-monotonicity is violated.

**Case 1: $\boldsymbol{\pi}$** contains at least one voter $i$ who ranks $b \succ_{\pi_i} c$.

Now, we will construct $\boldsymbol{\gamma}, U, \boldsymbol{\gamma}'$ such that the following claims hold: *Claim (1):* $\text{sw}(a, U) > \text{sw}(b, U) > \text{sw}(c, U)$; *Claim (2):* $\boldsymbol{\pi} \in \Pi_{V(\boldsymbol{\gamma}, U)}$; and *Claim (3):* $\boldsymbol{\pi}' \in \Pi_{V(\boldsymbol{\gamma}, U')}$. If these claims are true, then by the construction of our example, we have found $\boldsymbol{\gamma} \leq \boldsymbol{\gamma}'$ such that by increasing the public spirit from $\boldsymbol{\gamma}$ to $\boldsymbol{\gamma}'$, we can change the winner from $a$ to $b$, thereby increasing the distortion, a violation of instance-wise PS-monotonicity.

**Construction of $\boldsymbol{\gamma}, \boldsymbol{\gamma}'$.** Let $\boldsymbol{\gamma} = \mathbf{0}, \boldsymbol{\gamma}'$ such that $\gamma'_i = \gamma_i = 0$ for all $i \neq i^*$, and $\gamma'_{i^*} = \epsilon$ for some small $\epsilon > 0$ where $\epsilon < 1/(16m)$.

**Construction of $U$.** We set the utilities according to three cases (where latter cases apply only if earlier cases do not hold):

A. If there exists an $i$ who ranks $a \succ_{\pi_i} b \succ_{\pi_i} c$, set $i$'s utilities in weakly decreasing order of $\pi_i$ such that $a$ (and everything before it) gets utility 1, $b$ (and everything after $a$ and before $b$) gets utility 1/2, and $c$ (and everything after) gets utility 0.

   Give all remaining voters besides $i^*$ utility 0 for all alternatives.

B. Else if there exists an $i$ where $b \succ_{\pi_i} a \succ_{\pi_i} c$, set $i$'s utilities in weakly decreasing order of $\pi_i$: give $a$ and everything ranked before it (including $b$) utility 1, and everything after $a$ (including $c$) 0 utility.

   Then, by weak unanimity of $f$, there must be another voter $i'$ where $a \succ_{\pi_{i'}} b$, whose utilities we assign based on two cases:

   B1. If $i' \neq i^*$, set $i'$'s utilities according to $\pi_{i'}$: give all alternatives ranked ahead of $b$ utility 1/2 (this must include $a$ and $c$), and utility 0 to $b$ and all alternatives ranked after.

   B2. If $i' = i^*$, note that $i'$ must have ranking $a \succ_{\pi_{i'}} c \succ_{\pi_{i'}} b$, because $c$ and $b$ must be ranked adjacently. Then, give utility 1 to $a$, 1/2 to $c$, $1/2 - \epsilon^2/n$ to $b$, and set the rest of the alternatives' utilities so they are decreasing at intervals of at least $1/(4m)$.

   Give all other voters except $i^*$ with thus far unset utilities 0 utility for all alternatives.

C. Else, by the falseness of cases A and B and our assumption that there is some $i$ for which $b \succ_{\pi_i} c$, there must exist some voter $i$ who ranks $b \succ_{\pi_i} c \succ_{\pi_i} a$. Set $i$'s utilities in weakly decreasing order of $\pi_i$: Give $b$ and all alternatives before it utility $1/2 + \epsilon^2/n$ (the $+\epsilon^2/n$ is for convenience of arguments later), and all alternatives after it (including $c$ and $a$) utility 0.

Then, by the weak unanimity of $f$, there must exist one voter $i'$ where $a \succ_{\pi_{i'}} b$ and $a \succ_{\pi_{i'}} c$, in which case, by the falseness of cases A and B they must have ranking $a \succ_{\pi_{i'}} c \succ_{\pi_{i'}} b$.[1] Set $i'$'s utilities in weakly decreasing order of $\pi'_i$, based on two cases:

C1. If $i' \neq i^*$: let $i'$ have utility $1 + \epsilon^2/n$ for $a$ (the $+\epsilon^2/n$ is for convenience of arguments later) and all alternatives ranked before it and utility 0 for all alternatives ranked after it (including $c$ and $b$).

C2. If $i' = i^*$, set $i'$'s utilities similar to how we did in B2: give utility $3/2$ to $a$, $1/2$ to $c$, $1/2 - \epsilon^2/n$ to $b$, and set the rest of the alternatives' utilities so they are decreasing at intervals of at least $1/(4m)$.

Give all other voters with unset utilities except $i^*$ 0 utility for all alternatives.

If we have not already set $i^*$'s utilities in cases B or C (we cannot set them in case A), set $i^*$'s utilities in weakly decreasing order of $\pi_{i^*}$: give the alternatives ahead of (and including) $c$ utilities starting at $1/4$ and dropping by additive gaps of $1/(4m)$. Then, set $u_{i^*}(b)$ such that $u_{i^*}(c) - u_{i^*}(b) = \epsilon^2/n$. Then, for alternatives ranked after $b$, continue assigning utilities decreasing by additive gaps of $1/(4m)$.

**Proofs of Claims (1), (2), and (3).**

*Claim (1):* Let $N_A = [n] \setminus \{i\}$, $N_{B1} = [n] \setminus \{i\}$ $N_{B2} = [n]$, $N_{C1} = [n] \setminus \{i\}$, $N_{C2} = [n]$, denote the sets of voters whose utilities are set within cases $A$, $B$ and $C$, depending on which cases are invoked.

Now, we will show that for any $N \in \{N_A, N_{B1}, N_{B2}, N_{C1}, N_{C2}\}$, we have that

$$\sum_{i \in N} u_i(a) > \sum_{i \in N} u_i(b) > \sum_{i \in N} u_i(c),$$

and moreover, that these inequalities hold by a margin of at least $1/2$.

(Case A): letting $N = N_A$, we have $\sum_{i \in N} u_i(a) = 1$, $\sum_{i \in N} u_i(b) = 1/2$, $\sum_{i \in N} u_i(c) = 0$.

(Case B1): letting $N = N_{B1}$, we have $\sum_{i \in N} u_i(a) = 3/2$, $\sum_{i \in N} u_i(b) = 1$, $\sum_{i \in N} u_i(c) = 1/2$.

(Case B2): letting $N = N_{B2}$, we have $\sum_{i \in N} u_i(a) = 2$, $\sum_{i \in N} u_i(b) = 3/2 - \epsilon^2/n$, $\sum_{i \in N} u_i(c) = 1/2$.

(Case C1): letting $N = N_{C1}$, we have $\sum_{i \in N} u_i(a) = 1 + \epsilon^2/n$, $\sum_{i \in N} u_i(b) = 1/2 + \epsilon^2/n$, $\sum_{i \in N} u_i(c) = 0$.

(Case C2): letting $N = N_C$, we have $\sum_{i \in N} u_i(a) = 3/2$, $\sum_{i \in N} u_i(b) = 1$, $\sum_{i \in N} u_i(c) = 1/2$.

[1] The alternative would be that there would have to exist two voters, the first for whom $c \succ a \succ b$, and the second for whom $b \succ a \succ c$, which is not possible by the falseness of case B.

If cases B2 or C2 was the binding case — that is, we set $i^*$ while within the three cases — , then we have already concluded the claim, and $\text{sw}(a, U) - \text{sw}(b, U) \geq 1/2$ and $\text{sw}(b, U) - \text{sw}(c, U) \geq 1/2$. Otherwise, we note that for any pair of alternatives $d, e$, $|u_{i^*}(d) - u_{i^*}(e)| \leq 1/4$; therefore, these social welfare gaps cannot be closed by more than $1/4$, and we conclude that $\text{sw}(a, U) - \text{sw}(b, U) \geq 1/4$ and $\text{sw}(b, U) - \text{sw}(c, U) \geq 1/4$. We will use this lower bound on these gaps later, in Claim (3).

*Claim (2):* We have assigned voters' utilities in weakly decreasing order according to $\pi_i$ for all $i$, and $\boldsymbol{\gamma} = \mathbf{0}$, meaning that voters' individual utilities fully determine their rankings: thus, $\boldsymbol{\pi} \in \Pi_{V(\boldsymbol{\gamma}, U)}$.

*Claim (3)*: The proof of this claim follows the same structure as that of Claim (3) in the proof of Lemma 12.4.13, so we will be slightly more brief here, and invoke parts of that argument when useful. We again use the notation $\pi_i \in \Pi_{V_i(\boldsymbol{\gamma}, U)}$ to mean a voter $i$'s ranking $\pi_i$ is consistent with the vector of PS-values implied by the $i$th row of the matrix $V(\boldsymbol{\gamma}, U)$.

First, for all voters $i \neq i^*$, by construction of $\boldsymbol{\pi}, \boldsymbol{\pi}'$ we have that $\pi_i = \pi'_i$. Moreover, $\gamma_i = \gamma'_i$ implies that $\Pi_{V_i(\boldsymbol{\gamma}, U)} = \Pi_{V_i(\boldsymbol{\gamma}, U')}$. By these two equalities, $\pi_i \in \Pi_{V_i(\boldsymbol{\gamma}, U)}$ (as shown in Claim (2)) implies $\pi'_i \in \Pi_{V_i(\boldsymbol{\gamma}, U')}$.

Now considering voter $i^*$, we want to show that $\pi'_{i^*} \in \Pi_{V_{i^*}(\boldsymbol{\gamma}, U')}$. To show this, first fix a pair of alternatives $(d, d') \neq (b, c)$. By the same type of reasoning as in Lemma 12.4.13, we have that $|v_{i^*}(d, \boldsymbol{\gamma}, U) - v_{i^*}(d', \boldsymbol{\gamma}, U)| \geq 1/(4m) > 4\epsilon\}$, and also that $|v_{i^*}(d, \boldsymbol{\gamma}, U) - v_{i^*}(d, \boldsymbol{\gamma}', U)| \leq 2\epsilon$ and $|v_{i^*}(d', \boldsymbol{\gamma}, U) - v_{i^*}(d', \boldsymbol{\gamma}', U)| \leq 2\epsilon$, by the fact that all utilities in $U$ are bounded between 0 and 2. Putting these facts together, we get that for all such pairs $d, d'$,

$$v_{i^*}(d, \boldsymbol{\gamma}, U) \geq v_{i^*}(d', \boldsymbol{\gamma}, U) \implies v_{i^*}(d, \boldsymbol{\gamma}', U) \geq v_{i^*}(d', \boldsymbol{\gamma}', U). \tag{G.11}$$

Now, finally considering the pair $b, c$, we have the following, using that $\text{sw}(c, U) - \text{sw}(b, U) \geq 1/4$, as shown in the proof of Claim (1):

$$\begin{aligned} v_{i^*}(b, \boldsymbol{\gamma}', U) - v_{i^*}(c, \boldsymbol{\gamma}', U) &= (1 - \epsilon)(u_{i^*}(b) - u_{i^*}(c)) + \epsilon(\text{sw}(b, U)/n - \text{sw}(c, U)/n) \\ &> -\epsilon^2/n + \epsilon(\text{sw}(b, U) - \text{sw}(b, U))/n \\ &\geq -\epsilon^2/n + \epsilon/(4n) \\ &\geq 0. \end{aligned}$$

We conclude that

$$v_{i^*}(b, \boldsymbol{\gamma}', U) - v_{i^*}(c, \boldsymbol{\gamma}', U) > 0. \tag{G.12}$$

By Equations (G.11) and (G.12), we have that any ranking $\pi$ with the following two properties must be consistent with $\Pi_{V_{i^*}(\boldsymbol{\gamma}, U')}$: First, for all pairs of alternatives $(d, d') \neq (b, c)$, $d \succ_{\pi_{i^*}} d' \implies d \succ_{\pi} d'$, and second, $b \succ_{\pi} c$. $\pi'_{i^*}$ satisfies these criteria by construction, and thus $\pi'_{i^*} \in \Pi_{V_{i^*}(\boldsymbol{\gamma}, U')}$, as needed, concluding the proof of Case 1.

**Case 2: $\boldsymbol{\pi}$** does not contain a voter $i$ who ranks $b \succ_{\pi_i} c$.

First, observe that $c \succ_{\pi_i} b$ for all $i$ implies that $\boldsymbol{\pi}$ contains at least one voter in which $b \succ_{\pi_i} a$. To see this, first observe that $f(\boldsymbol{\pi}') = b$ implies that $b$ cannot always be ranked behind $a$ in $\boldsymbol{\pi}'$ by weak unanimity; thus there must be a voter $i'$ such that $b \succ_{\pi'_{i'}} a$. Next, observe that swapping $b$ and $c$ from $\boldsymbol{\pi} \to \boldsymbol{\pi}'$ cannot change the relative ordering of either of these alternatives with $a$, so it must also be the case that $b \succ_{\pi_{i'}} a$ (i.e., there exists such a voter in $\boldsymbol{\pi}$). We let $i'$ be this voter throughout this case.

Now, we will construct $\boldsymbol{\gamma}, U, \boldsymbol{\gamma}'$ such that three claims are true: *Claim (1):* $\text{sw}(c, U) > \text{sw}(b, U) > \text{sw}(a, U)$, *Claim (2):* $\boldsymbol{\pi}' \in \Pi_{V(\boldsymbol{\gamma}, U')}$, and *Claim (3):* $\boldsymbol{\pi} \in \Pi_{V(\boldsymbol{\gamma}, U)}$. If these claims hold, then we will have *decreased* voters' public spirit from $\boldsymbol{\gamma} \to \boldsymbol{\gamma}'$, which realizes the transformation from $\boldsymbol{\pi} \to \boldsymbol{\pi}'$. This transformation changed the winner from $a$ to $b$ — an *increase* in the social welfare and a violation of PS-monotonicity.

**Construction of $\boldsymbol{\gamma}$,$\boldsymbol{\gamma}'$.** We let $\boldsymbol{\gamma}$ such that $\gamma_i = 0$ for all $i \neq i^*$, and $\gamma_{i^*} = \epsilon$ for some small $0 < \epsilon < 1/m^4$, and let $\boldsymbol{\gamma}' = \mathbf{0}$.

**Construction of $U$.** First, for $i^*$, let their utility for the first-ranked alternative in $\pi'_{i^*}$ be $1/m$, and then, in order of $\pi_{i^*}$, assign the alternatives utilities descending at intervals of $1/m^2$ until we reach $b$. Then set $u_{i^*}(b)$ so that $u_{i^*}(b) - u_{i^*}(c) = \epsilon^2/n$. Then, starting after $c$, continue down $\pi'_{i^*}$ assigning alternatives decreasing utilities at intervals of $1/m^2$. For the remaining voters, we break into cases:

- If $i' \neq i^*$, we assign $i'$'s utilities according to $\pi'_{i'}$: let $i'$'s utilities be 1 for $c$ and all alternatives $i'$ ranks ahead of $c$; $1/2$ for $b$ and all alternatives $i'$ ranks between $c$ and $b$, and 0 for all alternatives $i'$ ranks after $b$ (note that this includes $a$, by selection of $i'$). Give all other voters besides $i^*$ and $i'$ 0 utility for all alternatives.
- If $i' = i^*$, then pick another arbitrary voter $i''$ for whom $c \succ_{\pi'_{i''}} b$. We assign $i''$'s utilities according to their ranking $\pi'_{i''}$: give $c$ and all alternatives ranked ahead of $c$ utility $1/m^3$, and give 0 utility to all alternatives they rank after $c$. Give all other voters besides $i^*$ and $i''$ 0 utility for all alternatives.

**Proofs of Claims (1), (2), and (3).**

*Claim (1):* If $i^* \neq i'$, then the only voters with any nonzero utilities are $i^*$ and $i'$; by their utilities, $\text{sw}(c) - \text{sw}(b) = 1/2 - \epsilon^2/n$ and $\text{sw}(b) - \text{sw}(a) \geq 1/2 - 1/m$ (where the $-1/m$ is the maximum possible gap between $u_{i^*}(a)$ and $u_{i^*}(b)$). If $i^* = i'$, then the only voters with any nonzero utilities are $i^*$ and $i''$; by their utilities, $\text{sw}(c) - \text{sw}(b) = 1/m^3 - \epsilon^2/n$, and $\text{sw}(b) - \text{sw}(a) \geq 1/m^2 - 1/m^3$.

*Claim (2):* We have assigned voters' utilities in weakly decreasing order according to $\pi'_i$ for all $i$, and $\boldsymbol{\gamma}' = \mathbf{0}$, meaning that voters' individual utilities fully determine their rankings: thus, $\boldsymbol{\pi}' \in \Pi_{V(\boldsymbol{\gamma}', U)}$.

*Claim (3):* The proof of this claim follows the same structure as that of Claim (3) in Case 1, so we will be slightly more brief here, and invoke parts of that argument when useful. We again use the notation $\pi_i \in \Pi_{V_i(\boldsymbol{\gamma}, U)}$ to mean a voter $i$'s ranking $\pi_i$ is consistent with the vector of PS-values implied by the $i$th row of the matrix $V(\boldsymbol{\gamma}, U)$.

First, for all voters $i \neq i^*$, by construction of $\boldsymbol{\pi}, \boldsymbol{\pi}'$ we have that $\pi_i = \pi'_i$. Moreover, $\gamma_i = \gamma'_i$ implies that $\Pi_{V_i(\boldsymbol{\gamma},U)} = \Pi_{V_i(\boldsymbol{\gamma},U')}$. By these two equalities, $\pi'_i \in \Pi_{V_i(\boldsymbol{\gamma},U')}$ (as shown in Claim (2)) implies $\pi_i \in \Pi_{V_i(\boldsymbol{\gamma},U)}$.

Now considering voter $i^*$, we want to show that $\pi_{i^*} \in \Pi_{V_{i^*}(\boldsymbol{\gamma},U)}$. To show this, first fix a pair of alternatives $(d, d') \neq (b, c)$. By the same type of reasoning as in Case 1, we have that $|v_{i^*}(d, \boldsymbol{\gamma}', U) - v_{i^*}(d', \boldsymbol{\gamma}', U)| \geq 1/m^2 > 2\epsilon\}$, and also that $|v_{i^*}(d, \boldsymbol{\gamma}, U) - v_{i^*}(d, \boldsymbol{\gamma}', U)| \leq \epsilon$ and $|v_{i^*}(d', \boldsymbol{\gamma}, U) - v_{i^*}(d', \boldsymbol{\gamma}', U)| \leq \epsilon$, by the fact that all utilities in $U$ are bounded between 0 and 1. Putting these facts together, we get that for all such pairs $d, d'$,

$$v_{i^*}(d, \boldsymbol{\gamma}', U) \geq v_{i^*}(d', \boldsymbol{\gamma}', U) \implies v_{i^*}(d, \boldsymbol{\gamma}, U) \geq v_{i^*}(d', \boldsymbol{\gamma}, U). \tag{G.13}$$

Now, finally considering the pair $b, c$, we have the following, using that $\text{sw}(c, U) - \text{sw}(b, U) \geq 1/m^4 > \epsilon$, as shown in the proof of Claim (1):

$$\begin{aligned} v_{i^*}(c, \boldsymbol{\gamma}, U) - v_{i^*}(b, \boldsymbol{\gamma}, U) &= (1 - \epsilon)(u_{i^*}(c) - u_{i^*}(b)) + \epsilon(\text{sw}(c, U)/n - \text{sw}(b, U)/n) \\ &> -\epsilon^2/n + \epsilon(\text{sw}(c, U) - \text{sw}(b, U))/n \\ &\geq -\epsilon^2/n + \epsilon^2/n \\ &= 0. \end{aligned}$$

We conclude that

$$v_{i^*}(b, \boldsymbol{\gamma}', U) - v_{i^*}(c, \boldsymbol{\gamma}', U) > 0. \tag{G.14}$$

By Equations (G.13) and (G.14), we have that any ranking $\pi$ with the following two properties must be consistent with $\Pi_{V_{i^*}(\boldsymbol{\gamma},U')}$: First, for all pairs of alternatives $(d, d') \neq (b, c)$, $d \succ_{\pi'_{i^*}} d' \implies d \succ_{\pi} d'$, and second, $c \succ_{\pi} b$. $\pi_{i^*}$ satisfies these criteria by construction, and thus $\pi_{i^*} \in \Pi_{V_{i^*}(\boldsymbol{\gamma},U)}$, as needed, concluding the proof of Case 2.

□

### G.2.6 Proof of Lemma 12.4.17

**Lemma G.2.13.** *If $f$ is monotonic and swap-invariant, then it is Maskin-monotonic.*

*Proof.* Fix a monotonic and swap-invariant voting rule $f$, and fix a profile $\boldsymbol{\pi}$ such that $f(\boldsymbol{\pi}) = a$. Let $\boldsymbol{\pi}'$ be an arbitrary other profile such that such that $a \succ_{\pi'_i} b$ whenever $a \succ_{\pi_i} b$ for every voter $i$ and for all $b \neq a$. Now, we will show that we can construct $\boldsymbol{\pi}'$ from $\boldsymbol{\pi}$ by promoting $a$ and/or swapping $b$ with alternatives other than $a$. By monotonicity and swap-invariance, this will preserve the winner thus it will hold that $f(\boldsymbol{\pi}') = a$, thereby proving the Maskin monotonicity of $f$.

Fix an $i$, and consider $\pi_i$, from which we must construct $\pi'_i$. First, let $A_1$ be the set of all alternatives ranked ahead of $a$ in $\pi_i$ but behind $a$ in $\pi_i$. Swap the alternatives in $A_1$ with other alternatives ahead of $a$ in $\pi_i$ so that all these alternatives are ranked just ahead of $a$. These swaps didn't change the $f$ winner by the swap invariance of $f$. Then, swap $a$ ahead of all alternatives in $A_1$–this does

not change the $f$ winner by the monotonicity of $f$. Finally, swap alternatives other than $a$ to make the relative ordering of all alternatives ahead of and behind $a$, respectively, match their relative ordering in $\pi'_i$; by swap invariance of $f$, this again does not change the $f$ winner. We can do this procedure to the rankings if all $i$, and thereby construct $\boldsymbol{\pi}'$ from $\boldsymbol{\pi}$ while preserving $a$ as the winner. □

# H

# Chapter 13 Appendix

Appendix

## H.1 Rankings by Value for Money

In the ballot format *rankings by value for money* (vfm), $\mathcal{L}_{\mathsf{vfm}}$ is still the set of all rankings over alternatives, but now each voter $i$ submits a ranking $\rho_i$ of the alternatives by their PS-value divided by cost, i.e., such that for every $a, b \in A$, $v_i(a)/c(a) > v_i(b)/c(b)$ implies $a \succ_{\rho_i} b$; the voter can break ties arbitrarily.

### H.1.1 Deterministic Rules

Benadè et al. [45] show that no deterministic rule for rankings by value for money can achieve bounded distortion, even under the unit-sum assumption. Moreover, in their construction, all voters submit the same ranking. Adding any amount of public spirit would therefore leave the rankings and their analysis unchanged, implying that the distortion remains unbounded even with public spirit. We formalize this in Theorem H.1.1.

**Theorem H.1.1** (lower bound)**.** *For rankings by value for money, every deterministic rule $f$ has unbounded distortion: $\mathrm{dist}_{\mathsf{vfm}}(f) = \infty$.*

*Proof.* We use the exact same construction used by Benadè et al. [45]. Fix $a, b \in A$, and let $c_a = \epsilon > 0$ and $c_x = 1$ for all $x \in A \setminus \{a\}$. Construct an input profile $\vec{\rho}$ where each voter has alternatives $a$ and $b$ in positions 1 and 2, and let $f$ be some deterministic aggregation rule.

If $f(\vec{\rho}, c) \neq a$, then construct a utility profile where $u_i(a) = 1$ and $u_i(x) = 0$ for all $x \in A \setminus \{a\}$. Then the distortion is infinite.

If $f(\vec{\rho}, c) = a$, then construct a utility profile where $u_i(a) = \epsilon$, $u_i(b) = 1$ and $u_i(x) = 0$ for $x \in A \setminus \{a, b\}$. Then,

$$\frac{v_i(a)}{c_a} = \frac{(1-\gamma_i)\epsilon + \gamma_i \frac{(n\epsilon)}{n}}{\epsilon} = \frac{(1-\gamma_i) + \gamma_i}{1} = \frac{v_i(b)}{c_b},$$

and so the ranking of each voter is consistent with this utility profile. But, the distortion is:

$$\frac{n}{n\epsilon} = \frac{1}{\epsilon},$$

which as $\epsilon \to 0$ tends to infinity. □

### H.1.2 Randomized Rules

For randomized rules, we show the same upper bound (up to a constant) for rankings by value for money as for rankings by value. The result uses a similar construction, too: First, we bucket alternatives as in Lemma 13.4.6, so that the alternatives in each bucket differ in cost by a factor of at most 2. Due to these similar costs, a ranking by value for money of the alternatives within any is a good approximation of their ranking by value, allowing us to apply our reductions from PB to committee selection to single-winner selection, except we lose an additional factor of 2.

**Theorem H.1.2** (upper bound)**.** *For rankings by value for money, there exists a randomized rule $f$ with distortion*

$$dist_{\mathsf{vfm}}(f) \leq 8\left(\left\lceil \log_2(m) \right\rceil + 1\right)\left(2\gamma_{\min}^{-1} - 1\right).$$

**Lemma H.1.3.** *For rankings by value for money, there exists a $k$-committee-selection voting rule $f$ such that on all sets of alternatives with costs in $[2^{-\ell}, 2^{1-\ell}]$ for some $\ell \geq 0$, $f$ has distortion $4(2\gamma_{\min}^{-1} - 1)$.*

*Proof.* Notice that if $a$ beats $b$, then $v_i(a)/c_a \geq v_i(b)/c_b$ at least $n/2$ times. Since the costs differ by at most a factor of 2, $2v_i(a) \geq v_i(b)$.

We can use the exact same rule as in Theorem 13.3.5. Indeed, everything is the same, except that when $b$ beats $a^*$ in a pairwise election (i.e. at least $n/2$ times), we get the following distortion by Lemma 13.2.1:

$$\frac{\mathsf{sw}(a^*)}{\mathsf{sw}(b)} \leq 2\left(2\frac{1-\gamma_{\min}}{\gamma_{\min}} + 1\right).$$

Then, the distortion of our rule is, by the same analysis in Theorem 13.3.5:

$$8\frac{1-\gamma_{\min}}{\gamma_{\min}} + 4.$$

From here, we can convert this single winner rule into a committee selection rule with the same distortion by using Lemma 13.4.7. □

Having proved this lemma, we utilise an argument similar to Lemma 13.4.6.

*Proof of Theorem H.1.2.* Let $g$ be the rule in Lemma H.1.3, and let the distortion it achieves, $\left(4\frac{1-\gamma_{\min}}{\gamma_{\min}}+2\right)$, be $d$. By the same mechanism in Lemma 13.4.6, we will convert $g$ to a ranking by value per cost rule.

Indeed, divide the alternatives into buckets $A_0, A_1, \ldots, A_{\lceil \log_2(m) \rceil}$, where for $i \neq 0$:

$$A_i = \left\{ a \in A : \frac{2^{i-1}}{m} < c_a \leq \frac{2^i}{m} \right\},$$

and

$$A_0 = \{a \in A : c_a \leq 1/m\}.$$

Recall the mechanism used:

1. Pick the bucket $A_j$ uniformly at random.
2. Consider the restricted election with only the alternatives in $A_j$.
3. Use $g$ to pick the top $\left\lfloor \frac{m}{2^j} \right\rfloor$ alternatives in the restricted election.

Consider any PB instance. Split the alternatives into buckets $A_0, A_1, \ldots, A_{\lceil \log_2(m) \rceil}$, where for $i \neq 0$:

$$A_i = \left\{ a \in A : 2^{i-1}/m < c_a \leq 2^i/m \right\},$$

and

$$A_0 = \{a \in A : c_a \leq 1/m\}.$$

The randomized PB rule $f$ is as follows:

1. Pick $j \in \{ 0, 1, \ldots, \lceil \log_2(m) \rceil \}$ uniformly at random.
2. Consider the restricted instance with only the alternatives in $A_j$.
3. With $m' = |A_j|$ and $k = \min(m', \left\lfloor \frac{m}{2^j} \right\rfloor)$, use the $k$-committee selection rule $f_{m',k}$ on this restricted instance to pick a set of $k$ alternatives and return it.

Let $A^*$ be the optimal budget-feasible subset of the alternatives, $L_j^*$ be the optimal $\left\lfloor \frac{m}{2^j} \right\rfloor$-committee of $A_j$, and $L_j$ be the one selected by the $k$-committee rule. For $j \neq 0$, $A^* \cap A_j$ is of size at most $\frac{m}{2^{j-1}}$. That means $\mathrm{sw}(A^* \cap A_j) \leq 2\mathrm{sw}(L_j^*)$ for any $j \neq 0$.

In addition for $j = 0$, $L_0^* = A_0$ which implies $\mathrm{sw}(A^* \cap A_j) \leq \mathrm{sw}(L_j^*)$. Since the $k$-committee selection rule has distortion of $d$ for any $j$ we have $\mathrm{sw}(L_j^*) \leq d\mathrm{sw}(L_j)$ which gives us $\mathrm{sw}(A^* \cap$

$A_j) \leq 2d\mathrm{sw}(L_j)$. Let $\delta$ be the distribution of the output of the mechanism, we have:

$$\begin{aligned}
\mathsf{E}_{L\sim\delta}[\mathrm{sw}(L)] &= \frac{1}{\lceil \log_2(m)\rceil + 1} \sum_{j=0}^{\lceil \log_2(m)\rceil} \mathrm{sw}(L_j) \\
&\geq \frac{1}{\lceil \log_2(m)\rceil + 1} \sum_{j=0}^{\lceil \log_2(m)\rceil} \frac{\mathrm{sw}(A^* \cap A_j)}{2d} \\
&\geq \frac{\mathrm{sw}(A^*)}{2d(\lceil \log_2(m)\rceil + 1)},
\end{aligned}$$

which gives us the desired distortion bound. □

Whether this is (asymptotically) the best distortion that randomized rules for rankings by value for money can achieve remains an open question.

### H.2 Threshold Approval Votes

Finally, we investigate the distortion under the ballot format of *threshold approval votes*. Under this ballot format with threshold $\tau > 0$ ($\tau$-th), each voter $i$ reports the subset of alternatives for which her PS-value is at least a $\tau$ fraction of her total PS-value for all alternatives in $A$, i.e., $\rho_i = \{a \in A : v_i(a) \geq \tau \cdot \sum_{b\in A} v_i(b)\}$. Thus, $\mathcal{L}_{\tau\text{-th}} = 2^A$, as with knapsack votes. Benadè et al. [45] introduce this ballot format for unit-sum utilities and our definition extends it to arbitrary utilities.[1]

It is easy to see that without a unit sum assumption, the distortion of any deterministic rule is unbounded, even with public-spirited voters.

**Proposition H.2.1.** *The distortion associated with deterministic fixed thresholds (using the same definition as in [45]) is unbounded for any choice of threshold.*

*Proof.* Suppose we use a threshold of $t$. Then, consider an input profile where no voter approves any alternative. Suppose that $f$ picks $a^* \in A$. Then, consider a preference profile where $u_i(a^*) = 0$ and $u_i(b) = t/2$ for all $i \in N$ and all $b \neq a^*$.

Then, $v_i(a^*) = (1-\gamma_i)\cdot 0 + \gamma_i \cdot \frac{0}{n} = 0 < t$ and $v_i(b) = (1-\gamma_i)\cdot t/2 + \gamma_i \cdot \frac{nt/2}{n} = t/2 < t$, meaning the utility profile is consistent with the input, but the distortion is infinite. □

#### H.2.1 Deterministic Rules

By setting $\tau = 1/m$, we can achieve the following distortion upper bound.

[1] One could also conceive of using an *absolute* threshold (i.e., voter $i$ asked to approve all $a$ with $v_i(a) \geq \tau$), instead of making it relative to the total value. But in Proposition H.2.1, we show that this leads to the worst possible distortion: unbounded for deterministic rules and $m$ for randomized rules.

**Theorem H.2.2** (upper bound)**.** *For threshold approval votes with threshold $\tau = 1/m$, there exists a deterministic rule $f$ with distortion*

$$dist_{(1/m)\text{-th}}(f) \leq m\left(m\gamma_{\min}^{-1} - m + 1\right).$$

*Proof.* We can use the voting rule that simply picks the plurality winner: the alternative with most approvals. Let $a$ be the plurality winner.

Let $S^*$ be the optimal feasible subset of alternatives. Then, if voter $i$ approves alternative $a$:

$$\frac{v_i(a)}{\sum_{b\in A} v_i(b)} \geq 1/m,$$

and so:

$$mv_i(a) \geq v_i(A).$$

Notice that every voter must approve at least one alternative, as at least one alternative must have value at least the average: $\frac{\sum_{a\in A} v_i(a)}{m}$. Therefore, by the pigeonhole principle, the plurality winner must appear at least $n/m$ times, and so $mv_i(a) \geq v_i(A)$ for at least $n/m$ voters $i$.

By Lemma 13.2.1,

$$\frac{\mathrm{sw}(A)}{\mathrm{sw}(a)} \leq m\left(\frac{1-\gamma_{\min}}{\gamma_{\min}}m + 1\right).$$

as claimed. □

As with rankings by value, it turns out that linear distortion is unavoidable, even when voters exhibit perfect public spirit and submit the same vote.

**Theorem H.2.3** (lower bound)**.** *For all deterministic $f$ and all threshold values $\tau > 0$,*

$$dist_{\tau\text{-th}}(f) \geq m - 1.$$

*Proof.* Let $t > 0$ be the threshold.

Consider the case where alternative $a$ costs 1, and alternatives $b_1, \ldots, b_{m-1}$ cost $\frac{1}{m-1}$.

Suppose all voters approve only $a$. Then, we have two cases. If the voting rule $f$ doesn't pick alternative $a$, then we incur infinite distortion when the utility of $a$ is 1, and the utility of $b_1, \ldots, b_{m-1}$ is 0 for all voters.

If $f$ does pick $a$, then it cannot pick anything else as the budget is exhausted. Let the utility of $a$ be $t + \epsilon$ and the utility of $b_j$ be $t - \epsilon$ for all voters, and any small $\epsilon > 0$.

Then, we could have gotten a utility of $(m-1)(t-\epsilon)$, but instead get $t + \epsilon$. As $\epsilon \to 0$, the distortion goes to $m - 1$. □

### H.2.2 Randomized Rules

Turning to randomized rules for threshold approval votes with threshold $\tau$, we get the same results under public-spirited behavior with arbitrary utilities as Benadè et al. [45] get under the unit-sum assumption.

**Theorem H.2.4** (lower bound)**.** *For threshold approval votes with any threshold $\tau > 0$, every randomized rule $f$ has distortion*

$$dist_{\tau\text{-th}}(f) \geq \frac{1}{2}\left(\left\lfloor \frac{\sqrt{m}}{2} \right\rfloor + 1\right).$$

*Proof.* We are borrowing the construction from Theorem 3.4 in Benadè et al. [45]. Consider the case where each alternative has cost 1. We consider two cases. First suppose that $\tau \leq 1/\lfloor\sqrt{m}\rfloor$. Fix a set $S$ of $\lfloor\sqrt{m}/2\rfloor + 1$ alternatives. Construct the input profile $\vec{\rho}$ where $\rho_i = S$ for all $i \in N$. There must exist $a^* \in S$ where $\mathbf{Pr}[a^*] \leq 1/|S|$. Consider the utility matrix $U$ where for all $i \in N$, $u_i(a^*) = 1/2$ and for $a \in S \setminus \{a^*\}$, $u_i(a) = 2/\lfloor\sqrt{m}/2\rfloor$ and $u_i(a) = 0$ for $a \in A \setminus S$. Note that since voters have identical utilities, we have $u_i(a) = v_i(a)$ for any alternative $a \in A$. We have $\mathrm{sw}(a^*) = n/2$ and for $a \in A \setminus \{a^*\}$, $\mathrm{sw}(a) \leq n/\sqrt{m}$. That gives us

$$\begin{aligned}
\mathrm{dist}_\tau\text{-th}(f) &\geq \frac{\mathrm{sw}(a^*)}{\mathsf{E}_{a\sim f(\vec{\rho},c)}[\mathrm{sw}(a)]} \\
&\geq \frac{\frac{n}{2}}{\frac{1}{\lfloor\sqrt{m}/2\rfloor+1}\frac{n}{2} + \frac{\lfloor\sqrt{m}/2\rfloor}{\lfloor\sqrt{m}/2\rfloor+1}\frac{n}{\sqrt{m}}} \\
&\geq \frac{}{\frac{1}{\lfloor\sqrt{m}/2\rfloor+1} + \frac{1}{\lfloor\sqrt{m}/2\rfloor+1}} \geq \frac{1}{2}\left(\left\lfloor \frac{\sqrt{m}}{2} \right\rfloor + 1\right).
\end{aligned}$$

On the other hand if $\tau > 1/\lfloor\sqrt{m}\rfloor$, construct the input profile $\vec{\rho}$ where $\rho_i = \emptyset$ for $i \in N$. In this case there exists $a^* \in A$ where $Pr[a^*] \leq 1/m$. Consider the utility matrix $U$ where for every voter $u_i(a^*) = 1/\lfloor\sqrt{m}\rfloor$ and for $a \in A \setminus \{a^*\}$, $u_i(a) = (1 - 1/\lfloor\sqrt{m}\rfloor)/(m-1)$. We have $\mathrm{sw}(a^*) = n/\lfloor\sqrt{m}\rfloor$, and $sw(a) = n(1 - 1/\lfloor\sqrt{m}\rfloor)/(m-1)$ for $a \in A \setminus \{a^*\}$. That gives us:

$$\begin{aligned}
\mathrm{dist}_\tau\text{-th}(f) &\geq \frac{\mathrm{sw}(a^*)}{\mathsf{E}_{a\sim f(\vec{\rho},c)}[\mathrm{sw}(a)]} \\
&\geq \frac{\frac{n}{\lfloor\sqrt{m}\rfloor}}{\frac{1}{m}\frac{n}{\lfloor\sqrt{m}\rfloor} + \frac{m-1}{m}\frac{n\left(1-\frac{1}{\lfloor\sqrt{m}\rfloor}\right)}{m-1}} \geq \frac{m}{\lfloor\sqrt{m}\rfloor} \geq \lfloor\sqrt{m}\rfloor,
\end{aligned}$$

which gives us the desired bound. □

Benadè et al. [45] consider an additional source of randomness, whereby the designer samples a threshold $\tau$ from a distribution $R$ over support $[0, 1]$, and then all voters are asked to submit

their threshold approval votes using this value of $\tau$ (same for all voters). We refer to this ballot format as *randomized threshold approval votes* with threshold distribution $D$ ($D$-rth). Note that $\mathcal{L}_{D\text{-rth}} = \mathcal{L}_{\tau\text{-th}} = 2^A$. Since randomness is already introduced, it makes sense to also allow the aggregation rule $f$ to be randomized in this case. When defining the distortion of a randomized rule $f$, we take expectation over the sampling of threshold $\tau$ (before taking any worst case).

**Theorem H.2.5** (lower bound)**.** *For randomized threshold approval votes with the threshold sampled from any distribution $D$, every randomized rule $f$ has distortion*

$$dist_{D\text{-rth}}(f) \geq \frac{1}{2}\left\lceil \frac{\log_2(m)}{\log_2(2\lceil \log_2(m)\rceil)} \right\rceil.$$

*Proof.* We are borrowing the construction directly from Theorem 3.6 in Benadè et al. [45]. Consider the case where $c_a = 1$ for all $a \in A$, and let $f$ be an arbitrary rule that both returns a threshold and a set of alternatives randomly.

Split up the $(1/m, 1]$ interval into $\lceil \log_2(m)/\log_2(2\log_2(m))\rceil$ parts $I_j$ defined such that

$$I_j = \left( \frac{(2\log_2(m))^{j-1}}{m}, \min\left\{ \frac{(2\log_2(m))^{j}}{m}, 1 \right\} \right].$$

Define $u_j$ and $\ell_j$ to be the largest and smallest points in $I_j$ respectively. By construction, $u_j \leq 2\log_2(m)\ell_j$ for all $j$.

The key idea is to give utilities to alternatives within the interval that the threshold with least probability is contained in, so that with high probability, the alternatives are either all approved or all disapproved.

Indeed, let $k$ be a value such that

$$\mathbf{Pr}(t \in I_k) \leq \lceil \log_2(m)/\log_2(2\log_2(m))\rceil^{-1},$$

which must exist by the pigeonhole principle.

Fix a subset $S \subseteq A$ of size $\lceil \log_2(m)\rceil$, and let $V = u_k/2 + (\lceil \log_2(m)\rceil - 1)\ell_k$.

We will give each voter the same utilities, so that $u(a) := u_i(a) = v_i(a)$ for all $i \in N, a \in A$. For all $a \in S$, assign utilities so that $\sum_{a\in S} u(a) = V$, for all $a \notin S$, let $u(a) = (1-V)/(m - \lceil \log_2(m)\rceil)$.

We can verify that $\ell_k \leq \frac{1}{2\log_2(m)} u_k$ for all $k$. We can then see that the utilities sum to one, and are all positive as:

$$V = \frac{u_k}{2} + (\lceil \log_2(m)\rceil - 1)\ell_k \leq \frac{1}{2} + \frac{\lceil \log_2(m)\rceil - 1}{2\log_2(m)} \leq 1.$$

We construct this so that all alternatives in $S$ have utilities contained in $I_k$. Thus, when $t \notin I_k$, all voters either approve $S$ or disapprove $S$. Therefore, there must exist some $a^* \in S$ such that

$$\mathbf{Pr}(a^* \text{is returned} \mid t \notin I_k) \leq 1/\lceil \log_2(m)\rceil.$$

Now, we can assign $u(a^*) = u_k/2$ and $u(a) = \ell_k$ for $a \in S \setminus \{a^*\}$. Then, the optimal choice is $a^*$ with social welfare $nu_k/2$, but instead, since $\ell_k > (1-V)/(m - \log_2(m))$, we pick with high probability an alternative with at most $n\ell_k$ utility.

Indeed, the expected social welfare of $f$ is:

$$\begin{aligned}
&\mathbf{Pr}(t \in I_k) \cdot \frac{nu_k}{2} + \mathbf{Pr}(t \notin I_k) \left( \frac{1}{\lceil \log_2(m) \rceil} \cdot \frac{nu_k}{2} + \frac{\lceil \log_2(m) \rceil - 1}{\lceil \log_2(m) \rceil} \cdot n\ell_k \right) \\
&\quad \le \left( \left\lceil \log_2(m)/\log_2(2\log_2(m)) \right\rceil^{-1} + \frac{1}{\lceil \log_2(m) \rceil} + \frac{\lceil \log_2(m) \rceil - 1}{\lceil \log_2(m) \rceil} \cdot \frac{1}{\log_2(m)} \right) \frac{nu_k}{2} \\
&\quad \le \left( \left\lceil \log_2(m)/\log_2(2\log_2(m)) \right\rceil^{-1} \right) nu_k.
\end{aligned}$$

The maximum social welfare that we can get is $nu_k/2$, so the distortion is:

$$\mathrm{dist}_{D\text{-rth}}(f) \ge \frac{\frac{nu_k}{2}}{nu_k \left\lceil \frac{\log_2(m)}{\log_2(2\log_2(m))} \right\rceil^{-1}} = \frac{1}{2} \left\lceil \frac{\log_2(m)}{\log_2(2\lceil \log_2(m) \rceil)} \right\rceil. \qquad \square$$

Theorems H.2.4 and H.2.5 are corollaries of Theorems 3.4 and 3.6 of Benadè et al. [45], respectively. Their lower bound, derived under the unit-sum assumption, carries over to our more general setup. While they do not allow public-spirited behavior, in their construction the utility of each alternative is the same across all voters, ensuring that any level of public-spirited behavior does not affect their construction. The only reason we provide full proofs is that Benadè et al. [45] derive only an asymptotic lower bound by making several simplifying assumptions, which we carefully remove to derive an exact lower bound.

## H.3 Proofs from Section 13.2 (Preliminaries)

### H.3.1 Proof of Lemma 13.2.1

**Lemma H.3.1.** *Let $A_1, A_2 \subseteq A$ be two arbitrary subsets of alternatives. Fix any $\alpha \ge 0$ and define $N_{A_1 \succ A_2} = \{i \in N : \alpha \cdot v_i(A_1) \ge v_i(A_2)\}$. Then:*

$$\frac{sw(A_2)}{sw(A_1)} \le \alpha \cdot \left( \frac{1-\gamma_{\min}}{\gamma_{\min}} \frac{n}{|N_{A_1 \succ A_2}|} + 1 \right).$$

*Proof.* The proof is the same as the proof of Lemma 3.1 by Flanigan et al. [134]. Indeed, for each voter $i \in N_{A_1 \succ A_2}$, we know that $\alpha v_i(A_1) \ge v_i(A_2)$, and so:

$$\alpha \left( (1-\gamma_i) u_i(A_1) + \gamma_i \frac{\mathrm{sw}(A_1)}{n} \right) \ge (1-\gamma_i) u_i(A_2) + \gamma_i \frac{\mathrm{sw}(A_2)}{n} \ge \gamma_i \frac{\mathrm{sw}(A_2)}{n}.$$

Dividing by $\gamma_i$ and using the fact that $\frac{1-\gamma_i}{\gamma_i}$ is decreasing in $\gamma_i$ we have:

$$\alpha\left(\frac{1-\gamma_{\min}}{\gamma_{\min}}\cdot u_i(A)+\frac{\mathsf{sw}(A_1)}{n}\right)\geq\frac{\mathsf{sw}(A_2)}{n}.$$

Summing over all voters in $N_{A_1\succ A_2}$,

$$\alpha\left(\frac{1-\gamma_{\min}}{\gamma_{\min}}\sum_{i\in N_{A_1\succ A_2}}u_i(A_1)+\frac{\mathsf{sw}(A_1)|N_{A_1\succ A_2}|}{n}\right)\geq\frac{\mathsf{sw}(A_2)|N_{A_1\succ A_2}|}{n}.$$

Using the fact that $\sum_{i\in N_{A_1\succ A_2}}u_i(A_1)\leq\sum_{i\in N}u_i(A_1)=\mathsf{sw}(A_1)$,

$$\alpha\left(\frac{1-\gamma_{\min}}{\gamma_{\min}}\mathsf{sw}(A_1)+\frac{\mathsf{sw}(A_1)|N_{A_1\succ A_2}|}{n}\right)\geq\frac{\mathsf{sw}(A_2)|N_{A_1\succ A_2}|}{n},$$

and, after some simplification, we finally get the desired upper bound:

$$\frac{\mathsf{sw}(A_2)}{\mathsf{sw}(A_1)}\leq\alpha\left(\frac{1-\gamma_{\min}}{\gamma_{\min}}\frac{n}{|N_{A_1\succ A_2}|}+1\right).$$ □

### H.3.2 Distortion Without Public Spirit

In this section, we consider the distortion that can be achieved under various ballot formats without an assumption of public-spirited voters, or equivalently, when $\gamma_i = 0$ for every voter $i \in N$. This serves as a benchmark and motivates the need for cultivating public spirit among voters. It is also interesting to note that without any public spirit, the information in the ballots is useless as rules that ignore the ballots altogether turn out to be worst-case optimal. In contrast, the worst-case optimal rules in the presence of even a little bit of public spirit are both qualitatively and quantitatively fairer.

**Proposition H.3.2.** *For any ballot format* $\mathsf{X} \in \{\mathsf{rbv}, \mathsf{vfm}, \mathit{knap}, \tau\text{-}\mathsf{th}, D\text{-}\mathsf{rth}\}$ *(with any threshold* $\tau$ *and threshold distribution* $D$*), every deterministic rule has unbounded distortion when* $\gamma_i = 0$ *for all* $i \in N$*.*

*Proof.* First, consider the ballot formats other than randomized threshold approval votes. For deterministic threshold approval votes, pick any threshold $\tau \in [0, 1]$. Let $n$ be even.

Consider an instance as follows. The cost of each alternative is 1, i.e., $c(a) = 1$ for each $a \in A$. Pick any two alternatives $a_1, a_2 \in A$, and let the input profile be as follows. Partition the voters into two equal-sized groups $N_1, N_2$.

- Under $\mathsf{X} \in \{\mathsf{rbv}, \mathsf{vfm}\}$, each voter in $N_1$ ranks $a_1$ at the top, $a_2$ next, and the remaining alternatives afterwards (arbitrarily); and each voter in $N_2$ ranks $a_2$ at the top, $a_1$ next, and the remaining alternatives afterwards (arbitrarily).
- Under $\mathsf{X} \in \{\mathsf{knap}, \tau\text{-}\mathsf{th}\}$ (where $\tau \neq 0$), each voter in $N_1$ submits $\{ a_1 \}$ and each voter in $N_2$ submits $\{ a_2 \}$.

- Under $\mathsf{X} = \tau$-th with $\tau = 0$, every voter approves all the alternatives.

Fix any of the above ballot formats $\mathsf{X}$ and consider any deterministic rule $f_{\mathsf{X}}$. Suppose it picks alternative $a$. Note that at least one of $a_1$ and $a_2$ is not picked by $f_{\mathsf{X}}$. Due to the symmetry, assume without loss of generality that it is $a_1$. Then, for an arbitrarily chosen $\epsilon \in (0, 1)$, consider the following consistent utility matrix $U$.

- Each voter in $N_1$ has utility 1 for $a_1$ and 0 for all other alternatives.
- Each voter in $N_2$ has utility $\epsilon$ for $a_2$ and 0 for all other alternatives.

Then, the distortion of $f_{\mathsf{X}}$ is at least

$$\frac{\mathsf{sw}(a_1, U)}{\mathsf{sw}(a, U)} = \frac{n/2}{\epsilon \cdot n/2} = \frac{1}{\epsilon}.$$

Because $\epsilon \in (0, 1)$ was chosen arbitrarily, we can take the worst case by letting $\epsilon \to 0$, which establishes unbounded distortion.

For randomized threshold approval votes with any threshold distribution $D$, we cannot fix the input profile upfront as it depends on the threshold $\tau$ sampled from $D$. However, we can assume that for each $\tau$ the rule sees the profile described above for $\tau$-th. The proof continues to work because the consistent utility matrix $U$ described above is independent of the value of $\tau$ (and hence, can be set upfront without knowing the value of $\tau$). □

**Proposition H.3.3.** *For any ballot format $\mathsf{X} \in \{\mathsf{rbv}, \mathsf{vfm}, \mathit{knap}, \tau\text{-}\mathsf{th}, D\text{-}\mathsf{rth}\}$ (with any threshold $\tau$ and threshold distribution $D$), every randomized rule has distortion at least $m$ when $\gamma_i = 0$ for all $i \in N$ and this is tight.*

*Proof.* For the upper bound under all ballot formats, it suffices to show that the trivial randomized rule $f$, which does not take any ballots as input and simply returns a single alternative chosen uniformly at random, achieves distortion at most $m$. Fix any instance $U$ and let $A^*$ be an optimal budget-feasible set of alternatives. Then, the expected social welfare under $f$ is

$$\frac{1}{m} \sum_{a \in A} \mathsf{sw}(a, U) \geq \frac{1}{m} \mathsf{sw}(A^*, U),$$

which implies the desired upper bound of $m$ on the distortion of $f$.

For the lower bound, we simply extend the argument from the proof of Proposition H.3.2. Define an instance with $m$ alternatives $a_1, a_2, \ldots, a_m$, all with cost 1 (i.e., $c(a_j) = 1$ for all $j \in [m]$). Fix any randomized rule $f_{\mathsf{X}}$ for each ballot $\mathsf{X}$ in the statement of the proposition.

Let us first consider ballot formats other than randomized threshold approval votes. Consider the following symmetric profiles for each ballot format. Suppose $n$ divides $m$ and voters are partitioned into $m$ equal-size groups $N_1, \ldots, N_m$. Then:

- for $\mathsf{X} \in \{\mathsf{rbv}, \mathsf{vfm}\}$, for each $j \in [m]$, every voter in $N_j$ submits the ranking $a_j \succ a_{j+1} \succ \cdots \succ a_m \succ a_1 \succ \cdots \succ a_{j-1}$, and

- for $\mathsf{X} = \{\, \mathsf{knap}, \tau\text{-}\mathsf{th} \,\}$ (for any $\tau$), for each $j \in [m]$, every voter in $N_j$ submits the set of alternatives $\{a_j\}$.

For $\tau$-threshold approval votes, there is an edge case where this profile may not be feasible with $\tau = 0$, in which case we can set the profile to have every voter approving all alternatives. The utility matrix defined below would still remain consistent in this case.

For each ballot format $\mathsf{X}$, the corresponding rule must pick at least one alternative with probability $p_{\mathsf{X}} \leq 1/m$. Due to the symmetry, we can assume without loss of generality that this alternative is $a_1$.

Fix any $\epsilon \in (0, 1)$. We define a consistent utility matrix $U$ that works for all of the above ballot formats:

- Every voter in $N_1$ has utility 1 for $a_1$ and 0 for all other alternatives.
- For each $j \in \{\, 2, \ldots, m \,\}$, every voter in $N_j$ has utility $\epsilon$ for $a_j$ and 0 for all other alternatives.

Finally, notice that the maximum possible social welfare is $\mathrm{sw}(a_1, U) = 1$, whereas the expected social welfare under the rule $f_{\mathsf{X}}$ is $p_{\mathsf{X}} \cdot 1 + (1 - p_{\mathsf{X}}) \cdot \epsilon \leq 1/m + (1 - 1/m) \cdot \epsilon$. Thus, the distortion of $f_{\mathsf{X}}$ is at least $\frac{1}{1/m+(1-1/m)\cdot\epsilon}$. Since $\epsilon \in (0, 1)$ was chosen arbitrarily, we can take the worst case by letting $\epsilon \to 0$, in which case we get that the distortion must be at least $m$.

For randomized threshold approval votes with threshold distribution $D$, we cannot fix the input profile as the input profile depends on the threshold $\tau$ sampled from $D$. However, we can assume that the rule sees the generic input profile described above (where each voter approves only her most favorite alternative) for any $\tau \neq 0$ and the edge-case input profile (where every voter approves all the alternatives). Due to the symmetry, the rest of the argument goes through as the final utility matrix $U$ constructed above is consistent with these input profiles for all $\tau$. □

## H.4 Proofs from Section 13.3 (Single Winner)

### H.4.1 Proof of Theorem 13.3.1

**Theorem 13.3.1** (Lower Bound - Deterministic)**.** *Any deterministic single-winner voting rules $f$ with ranked preferences has distortion*

$$dist_{rbv}(f) \geq 1 + 2\frac{1 - \gamma_{\min}}{\gamma_{\min}} \cdot \frac{m^2}{2\gamma_{\min} + \gamma_{\min} m^2 + (2 - 3\gamma_{\min})m} \in \Omega\left(\frac{1}{\gamma_{\min}} \cdot \min\left\{\, m, \frac{1}{\gamma_{\min}} \,\right\}\right).$$

*Proof.* Suppose we have $m$ alternatives $a_1, \ldots, a_m$ and $n$ voters each with the same PS-value of $\gamma = \gamma_{\min}$. For ease of exposition, let $n$ be divisible by $m$. Our construction consists of $m$ types of voters, equally distributed with $n/m$ voters of each type. Let $N_k$ be the set of voters of type $k$. Suppose each voter type votes as follows,

$$
\begin{array}{cccccccccc}
N_1 & : & a_1 & \succ & a_2 & \succ & \dots & \succ & a_{m-1} & \succ & a_m \\
N_2 & : & a_2 & \succ & a_3 & \succ & \dots & \succ & a_m & \succ & a_1 \\
\vdots & & & & & & & & & & \\
N_{m-1} & : & a_{m-1} & \succ & a_m & \succ & \dots & \succ & a_{m-3} & \succ & a_{m-2} \\
N_m & : & a_m & \succ & a_1 & \succ & \dots & \succ & a_{m-2} & \succ & a_{m-1}
\end{array}
$$

so that $N_i$ prefers alternative $a_i$ most, and cycles through the rest.

Without the loss of generality, suppose the voting rule picks $a_1$. We will set the utilities so that $\mathrm{sw}(a_m) > \mathrm{sw}(a_{m-1}) > \cdots > \mathrm{sw}(a_2) > \mathrm{sw}(a_1)$. To do so, set for all voters $i$,

$$
u_i(a_m) = \begin{cases} 1 & \text{if } i \in N_m \\ 0 & \text{if } i \in N_1 \\ u_i(a_1) & \text{otherwise} \end{cases}.
$$

For all $k$ from 1 to $m-1$ and for all $i \in N_1$,

$$
u_i(a_k) = \frac{\gamma}{1-\gamma}\left(\frac{\mathrm{sw}(a_m) - \mathrm{sw}(a_k)}{n}\right),
$$

and for all $j$ from 2 to $m$, for all $i \in N_j$, for $k$ from 1 to $m-1$, when $k < j-1$:

$$
u_i(a_k) = \frac{\gamma}{1-\gamma}\left(\frac{\mathrm{sw}(a_{j-1}) - \mathrm{sw}(a_k)}{n}\right),
$$

and when $k \geq j$:

$$
u_i(a_k) = \frac{\gamma}{1-\gamma}\left(\frac{\mathrm{sw}(a_m) - \mathrm{sw}(a_k)}{n} + \frac{\mathrm{sw}(a_{j-1}) - \mathrm{sw}(a_1)}{n}\right),
$$

and $u_i(a_{j-1}) = 0$.

Then, for $k$ from 1 to $m-1$,

$$
\begin{aligned}
\mathrm{sw}(a_k) &= \sum_{j=1}^{m} \sum_{i \in N_j} u_i(a_k) \\
&= \frac{\gamma}{1-\gamma} \cdot \frac{1}{n}\Bigg( \sum_{i \in N_1} \Big(\mathrm{sw}(a_m) - \mathrm{sw}(a_k)\Big) + \sum_{j=2}^{k} \sum_{i \in N_j} \Big(\mathrm{sw}(a_m) - \mathrm{sw}(a_k) + \mathrm{sw}(a_{j-1}) - \mathrm{sw}(a_1)\Big) + 0 \\
&\qquad + \sum_{j=k+2}^{m} \sum_{i \in N_j} \Big(\mathrm{sw}(a_{j-1}) - \mathrm{sw}(a_k)\Big)\Bigg) \\
&= \frac{\gamma}{1-\gamma} \cdot \frac{1}{n} \cdot \frac{n}{m}\left((k-1)(\mathrm{sw}(a_m) - \mathrm{sw}(a_1)) - (m-1)\mathrm{sw}(a_k) + \sum_{j=1, j \neq k}^{m} \mathrm{sw}(a_j)\right) \\
&= \frac{\gamma}{1-\gamma} \cdot \frac{1}{m}\left((k-1)(\mathrm{sw}(a_m) - \mathrm{sw}(a_1)) - m \cdot \mathrm{sw}(a_k) + \sum_{j=1}^{m} \mathrm{sw}(a_j)\right).
\end{aligned}
$$

Let $S = \sum_{j=1}^{m} \mathrm{sw}(a_j)$. Adding $\frac{\gamma}{1-\gamma}\mathrm{sw}(a_k)$ to both sides of the above and rearranging, we get:

$$\mathrm{sw}(a_k) = \frac{\gamma}{m}\left((k-1)(\mathrm{sw}(a_m) - \mathrm{sw}(a_1)) + S\right).$$

In particular, $\mathrm{sw}(a_1) = \frac{\gamma}{m}S$, so

$$\mathrm{sw}(a_k) = \frac{\gamma}{m}\left((k-1)\mathrm{sw}(a_m) + S \cdot \frac{m-(k-1)\gamma}{m}\right).$$

Via the same reasoning,

$$\begin{aligned}
\mathrm{sw}(a_m) &= \sum_{j=1}^{m}\sum_{i\in N_j} u_i(a_m) \\
&= \frac{\gamma}{1-\gamma}\cdot\frac{1}{n}\left(\sum_{j=2}^{m-1}\sum_{i\in N_j}\left(\mathrm{sw}(a_{j-1}) - \mathrm{sw}(a_1)\right)\right) + \frac{n}{m} \\
&= \frac{\gamma}{1-\gamma}\cdot\frac{1}{m}\left(\sum_{j=2}^{m-1}\left(\mathrm{sw}(a_{j-1}) - \mathrm{sw}(a_1)\right)\right) + \frac{n}{m} \\
&= \frac{\gamma}{1-\gamma}\cdot\frac{1}{m}\left(S - (m-2)\mathrm{sw}(a_1) - \mathrm{sw}(a_m) - \mathrm{sw}(a_{m-1})\right) + \frac{n}{m} \\
&= \frac{\gamma}{1-\gamma}\cdot\frac{1}{m}\left(S - \frac{\gamma(m-2)}{m}S - \mathrm{sw}(a_m) - \frac{\gamma}{m}\left((m-2)\mathrm{sw}(a_m) + S\cdot\frac{m-(m-2)\gamma}{m}\right)\right) + \frac{n}{m} \\
&= \frac{\gamma}{1-\gamma}\cdot\frac{1}{m}\left(\frac{m-(m-2)\gamma}{m}\cdot\frac{m-\gamma}{m}S - \frac{m+\gamma(m-2)}{m}\mathrm{sw}(a_m)\right) + \frac{n}{m} \\
&= \frac{\gamma}{1-\gamma}\cdot\frac{1}{m}\left(\frac{m-(m-2)\gamma}{m}\cdot\frac{m-\gamma}{m}S\right) + \frac{n}{m} - \frac{\gamma(m+\gamma(m-2))}{(1-\gamma)m^2}\mathrm{sw}(a_m).
\end{aligned}$$

Adding $\frac{\gamma(m+\gamma(m-2))}{(1-\gamma)m^2}\mathrm{sw}(a_m)$ to both sides and rearranging:

$$\begin{aligned}
\mathrm{sw}(a_m) &= \frac{(1-\gamma)m^2}{(1-\gamma)m^2+\gamma(m+\gamma(m-2))}\left(\frac{\gamma}{1-\gamma}\cdot\frac{1}{m}\left(\frac{m-(m-2)\gamma}{m}\cdot\frac{m-\gamma}{m}S\right) + \frac{n}{m}\right) \\
&= \frac{\gamma m}{(1-\gamma)m^2+\gamma(m+\gamma(m-2))}\left(\frac{m-(m-2)\gamma}{m}\cdot\frac{m-\gamma}{m}S\right) + \frac{(1-\gamma)mn}{(1-\gamma)m^2+\gamma(m+\gamma(m-2))} \\
&= \frac{\gamma(m-(m-2)\gamma)}{(1-\gamma)m^2+\gamma(m+\gamma(m-2))}\cdot\frac{m-\gamma}{m}S + \frac{(1-\gamma)nm}{(1-\gamma)m^2+\gamma(m+\gamma(m-2))}.
\end{aligned}$$

Now, we can finally solve for $S$:

$$\begin{aligned}
S &= \sum_{k=1}^{m} \mathrm{sw}(a_k) \\
&= \mathrm{sw}(a_m) + \frac{\gamma}{m} \sum_{k=1}^{m-1} \left( (k-1)\mathrm{sw}(a_m) + S\frac{m-(k-1)\gamma}{m} \right) \\
&= \mathrm{sw}(a_m) + \frac{\gamma(m-1)(m-2)}{2m}\mathrm{sw}(a_m) + \frac{\gamma}{m^2} S \sum_{k=1}^{m-1} (m-(k-1)\gamma) \\
&= \frac{2m+\gamma(m-1)(m-2)}{2m}\mathrm{sw}(a_m) + \frac{\gamma}{m^2} S \cdot \frac{(m-1)(2\gamma+m(2-\gamma))}{2} \\
&= \frac{2m+\gamma(m-1)(m-2)}{2m} \left( \frac{\gamma(m-(m-2)\gamma)}{(1-\gamma)m^2+\gamma(m+\gamma(m-2))} \cdot \frac{m-\gamma}{m} S + \frac{(1-\gamma)nm}{(1-\gamma)m^2+\gamma(m+\gamma(m-2))} \right) \\
&\qquad + S \cdot \frac{\gamma(m-1)(2\gamma+m(2-\gamma))}{2m^2}.
\end{aligned}$$

After simplifying this, we get:

$$S = n\frac{2\gamma+\gamma m^2+(2-3\gamma)m}{2(1-\gamma)m^2+2\gamma(\gamma+1)m-4\gamma^2}.$$

This then implies that

$$\mathrm{sw}(a_m) = \frac{n}{m} \cdot \frac{2m^2(1-\gamma)+\left(m(2-3\gamma)+2\gamma+m^2\gamma\right)\gamma}{2(1-\gamma)m^2+2\gamma(\gamma+1)m-4\gamma^2},$$

and so we ultimately get the following social welfare for each alternative, for $k$ from 1 to $m-1$:

$$\mathrm{sw}(a_k) = \frac{n}{m} \cdot \frac{\gamma\left(2(1-\gamma)km+\gamma\left(m^2-m+2\right)\right)}{2(1-\gamma)m^2+2\gamma(\gamma+1)m-4\gamma^2}.$$

The chain of inequalities $\mathrm{sw}(a_m) > \cdots > \mathrm{sw}(a_1)$ does indeed hold, and knowing this, we can verify that the above utilities are non-negative.

This gives distortion, after simplifying:

$$\frac{\mathrm{sw}(a_m)}{\mathrm{sw}(a_1)} = 1 + \frac{2(1-\gamma)m^2}{\gamma\left(2\gamma+\gamma m^2+(2-3\gamma)m\right)}.$$

To show that this is asymptotically as desired, we can write this as:

$$1 + \frac{2(1-\gamma)}{\gamma}\left(\frac{2\gamma+\gamma m^2+(2-3\gamma)m}{m^2}\right)^{-1}.$$

Since, for any positive $a, b$, we have that $(a+b)^{-1} \geq \frac{1}{2}\min\{a^{-1}, b^{-1}\}$, this expression is in:

$$\Omega\left(1 + \frac{1-\gamma}{\gamma}\min\left\{\frac{m^2}{\gamma(m^2+2)}, \frac{m^2}{m(2-3\gamma)}\right\}\right) = \Omega\left(1 + \frac{1-\gamma}{\gamma}\min\left\{\frac{1}{\gamma}, m\right\}\right),$$

which in the $\gamma \to 0$ regime is asymptotic in $\Omega\left(\frac{\min\{1/\gamma, m\}}{\gamma}\right)$. □

### H.4.2 Proof of Theorem 13.3.2

**Theorem 13.3.2** (Lower Bound - Randomized)**.** *Any randomized single-winner voting rules $f$ with ranked preferences has distortion*

$$dist_{rbv}(f) \in \Omega\left(\min\left\{m, \frac{1}{\gamma_{\min}}\right\}\right).$$

*Proof.* Use the same input profile $\vec{\rho}$ as in the proof of Theorem 13.3.1. Let $p(a_i)$ be the probability that $a_i$ is picked by rule $f$ and without the loss of generality, suppose that $a_{\min} = \arg\min_a p(a)$.

Then, for any $j$, $1 = \sum_i p(a_i) \geq p(a_j) + (m-1)p(a_{\min})$, so $p(a_j) \leq 1 - (m-1)p(a_j)$

By the proof of Theorem 13.3.1, $\mathrm{sw}(a_1) \leq \mathrm{sw}(a_2) \leq \cdots \leq \mathrm{sw}(a_m)$, and so we can maximize social welfare by picking $a_m$.

The expected social welfare of $f$ is at most:

$$\begin{aligned}
\mathsf{E}_{a\sim f(\vec{\rho})}[\mathrm{sw}(a)] &= \frac{1}{m}\mathrm{sw}(a_m) + \frac{m-1}{m}\max_{k=1}^{m-1}\mathrm{sw}(a_k) \\
&= \frac{n}{m(2(1-\gamma)m^2 + 2\gamma(\gamma+1)m - 4\gamma^2)} \cdot \left(\frac{2m^2(1-\gamma) + \left(m(2-3\gamma) + 2\gamma + m^2\gamma\right)\gamma}{m}\right. \\
&\qquad \left. + \frac{m-1}{m} \cdot \left(\gamma\left(2(1-\gamma)(m-1)m + \gamma\left(m^2 - m + 2\right)\right)\right)\right) \\
&= \frac{n}{m} \cdot \frac{\gamma(\gamma-2)(m-2)(m-1) - 2m}{2((1-\gamma)m + 2\gamma)(m-\gamma)}.
\end{aligned}$$

So, the distortion is:

$$\begin{aligned}
\frac{\mathrm{sw}(a_m)}{\mathsf{E}_{a\sim f(\vec{\rho})}[\mathrm{sw}(a)]} &= \frac{n}{m} \cdot \frac{2m^2(1-\gamma) + \left(m(2-3\gamma) + 2\gamma + m^2\gamma\right)\gamma}{2(1-\gamma)m^2 + 2\gamma(\gamma+1)m - 4\gamma^2} \\
&\qquad \cdot \left(\frac{n}{m} \cdot \frac{\gamma(\gamma-2)(m-2)(m-1) - 2m}{2((1-\gamma)m + 2\gamma)(m-\gamma)}\right)^{-1} \\
&= 1 + \frac{2(1-\gamma)(m-1)((1-\gamma)m + 2\gamma)}{\gamma(2-\gamma)(m-2)(m-1) + 2m} \\
&\geq 1 + \frac{2(1-\gamma)^2(m-1)m}{\gamma(2-\gamma)(m-2)(m-1) + 2m}.
\end{aligned}$$

Since, for any positive $a, b$, we have that $(a+b)^{-1} \geq \frac{1}{2}\min\{a^{-1}, b^{-1}\}$:

$$\frac{\mathrm{sw}(a_m)}{\mathsf{E}_{a\sim f(\vec{\rho})}[\mathrm{sw}(a)]} \in \Omega\left((1-\gamma)^2 \min\left\{\frac{2(m-1)m}{\gamma(2-\gamma)(m-2)(m-1)}, \frac{2(m-1)m}{2m}\right\}\right)$$
$$\in \Omega\left((1-\gamma)^2 \min\left\{\frac{1}{\gamma}, m\right\}\right),$$

which in the $\gamma \to 0$ regime, is $\Omega\left(\min\left\{1/\gamma, m\right\}\right)$. □

## H.5 Proofs from Section 13.4 (Rankings by Value)

### H.5.1 Proof of Theorem 13.4.1

**Theorem 13.4.1** (lower bound)**.** *For rankings by value, every deterministic rule $f$ has distortion*

$$dist_{\mathsf{rbv}}(f) \geq \frac{m-1}{\gamma_{\min}} \in \Omega\left(\frac{m}{\gamma_{\min}}\right).$$

*Proof.* Consider an instance with $A = \{a, b_1, \ldots b_{m-1}\}$, where $a$ costs 1 and every other alternative costs $1/(m-1)$. Define $p = \frac{1-\gamma_{\min}}{1-\gamma_{\min}+m^2}$. Let $N_1$ be a set of $n(1-p)$ voters and $N_2 = N \setminus N_1$. Suppose that members of $N_1$ submit ranking $(a \succ b_1 \succ \ldots \succ b_{m-1})$ and members of $N_2$ vote $(b_1 \succ \ldots \succ b_{m-1} \succ a)$.

Now consider two cases.

Case 1: If the aggregation rule selects $a$, consider utility matrix $U$ where members of $N_1$ have utility of $\frac{\gamma_{\min}p}{1-p\gamma_{\min}}$ for $a$ and 0 for the rest, while members of $N_2$ have utility of 0 for $a$ and 1 for the rest of the alternatives. This means $\mathrm{sw}(a) = n(1-p)\frac{\gamma_{\min}p}{1-\gamma_{\min}p}$, and $\mathrm{sw}(b) = np$ for $b \in A \setminus \{a\}$. Alongside with the PS-vector $\vec{\gamma} = [\gamma_{\min}]^n$ we have value matrix $V_{\vec{\gamma},U}$ first of all we have to make sure that this is consistent with the input profile. For $i \in N_1$,

$$v_i(a) = (1-\gamma_{\min})\frac{\gamma_{\min}p}{1-\gamma_{\min}p} + \gamma_{\min}(1-p)\frac{\gamma_{\min}p}{1-\gamma_{\min}p}$$
$$= (1-\gamma_{\min}p)\frac{\gamma_{\min}p}{1-\gamma_{\min}p} = \gamma_{\min}p,$$

and $v_i(b_j) = (1-\gamma_{\min}) \times 0 + \gamma_{\min}p = \gamma_{\min}p$. Therefore, the value matrix is consistent with the ranking of the members of $N_1$. On the other hand for $i \in N_2$ we have, $v_i(a) = \gamma_{\min}(1-p)\frac{\gamma_{\min}p}{1-\gamma_{\min}p}$, and $v_i(b_j) = 1-\gamma_{\min}+\gamma_{\min}p$, where for $p = \frac{1-\gamma_{\min}}{1-\gamma_{\min}+m^2}$ we have:

$$v_i(a) = \frac{\gamma_{\min}^2 m^2 (1-\gamma_{\min})}{(m^2+1-\gamma_{\min})(m^2+(1-\gamma_{\min})^2)},$$
$$v_i(b_j) = \frac{(m^2+1)(1-\gamma_{\min})}{m^2+1-\gamma_{\min}}.$$

This gives us:

$$\frac{v_i(a)}{v_i(b_j)} = \frac{\gamma_{\min}^2 m^2}{(m^2+1)\,(m^2+(1-\gamma_{\min})^2)} \leq 1$$
$$\implies v_i(b_j) \geq v_i(a),$$

and therefore the votes of voters in $N_2$ are consistent with the value matrix $V_{\vec{\gamma},U}$.

By picking budget-feasible set $\{b_1, \ldots, b_{m-1}\}$ we can get a social welfare of $n(m-1)p$, while instead we got $n(1-p)\frac{\gamma_{\min} p}{1-p\gamma_{\min}}$ by choosing $a$. This leaves us with a distortion of

$$\frac{(m-1)(1-p\gamma_{\min})}{(1-p)\gamma_{\min}}.$$

Since $p \geq 0$ and $\gamma_{\min} \leq 1$, $p \geq p\gamma_{\min}$, and so $1-p\gamma_{\min} \geq 1-p$. Therefore, we get the desired distortion:

$$\frac{(m-1)(1-p\gamma_{\min})}{(1-p)\gamma_{\min}} \geq \frac{m-1}{\gamma_{\min}}.$$

Case 2: If the aggregation rule does not select $a$, consider the utility matrix $U$ where members of $N_1$ have utility of 1 for $a$ and 0 for the rest, while members of $N_2$ have utility of 0 for $a$ and $\frac{\gamma_{\min}(1-p)}{1-\gamma_{\min}(1-p)}$ for the rest of the alternatives. This gives us $\mathrm{sw}(a) = n(1-p)$, and $\mathrm{sw}(b) = np\frac{\gamma_{\min}(1-p)}{1-\gamma_{\min}(1-p)}$ for $b \in A \setminus \{a\}$. Again we have to check that the value matrix $V_{\vec{\gamma},U}$ is consistent with the input profile. For $i \in N_1$ we have: $v_i(a) = 1-\gamma_{\min}+\gamma_{\min}(1-p) = 1-\gamma_{\min}p$, and $v_i(b_j) = \gamma_{\min}p\frac{\gamma_{\min}(1-p)}{1-\gamma_{\min}(1-p)}$.

The value matrix is consistent with the ranking of the members of $N_1$, i.e. $v_i(a) \geq v_i(b_j)$, as:

$$\gamma_{\min} \leq 1 \implies 0 \leq \gamma_{\min}p \leq 1-\gamma_{\min}(1-p)$$
$$\implies \gamma_{\min}p\frac{1}{1-\gamma_{\min}(1-p)} \leq 1$$
$$\implies \gamma_{\min}p\frac{\gamma_{\min}(1-p)}{1-\gamma_{\min}(1-p)} \leq 1-\gamma_{\min}p.$$

Moreover, for $i \in N_2$ we have: $v_i(a) = \gamma_{\min}(1-p)$, and

$$v_i(b_j) = (1-\gamma_{\min})\frac{\gamma_{\min}(1-p)}{1-\gamma_{\min}(1-p)} + \gamma_{\min}p\frac{\gamma_{\min}(1-p)}{1-\gamma_{\min}(1-p)}$$
$$= (1-\gamma_{\min}(1-p))\frac{\gamma_{\min}(1-p)}{1-\gamma_{\min}(1-p)} = \gamma_{\min}(1-p).$$

So we have $v_i(a) = v_i(b_j)$ which means that the value matrix is consistent with the ranking of the members of $N_2$ as well.

Since $a$ is not picked by the aggregation rule, we get a maximum social welfare of $(m-1)np\frac{\gamma_{\min}(1-p)}{1-\gamma_{\min}(1-p)}$ when we could have gotten a social welfare of $np$ from $a$ meaning a distortion of:

$$\mathsf{dist}_{\mathsf{rbv}}(f) \geq \frac{1-\gamma_{\min}(1-p)}{\gamma_{\min}p(m-1)} \geq \frac{m-1}{\gamma_{\min}}.$$

All the conditions above hold for $m \geq 2$, so we have a distortions of at least: $\frac{m-1}{\gamma_{\min}}$. □

### H.5.2 Proof of Lemma 13.4.7

**Lemma H.5.1** (Single-Winner → Committee)**.** *Fix any $k \in [m]$ and $d \geq 1$. If there exists a single-winner rule with distortion at most $d$ for each $m' \leq m$, then there exists a $k$-committee selection rule with distortion at most $d$. The committee selection rule is deterministic if the underlying rule is deterministic, and it is randomized if the underlying rule is randomized.*

*Proof.* Let $A^* = \{a_1^*, \ldots, a_k^*\}$ be the optimal budget-feasible set, sorted from highest social welfare to the lowest so that $i < j \implies \mathsf{sw}(a_i^*) \geq \mathsf{sw}(a_j^*)$. Let $S$ denote the set of alternatives that our algorithm picks.

Consider the $i$th iteration of the procedure. Let $a^+{}_i$ be the alternative with the highest social welfare among the remaining alternatives, and $a_i$ be the random alternative picked by the single-winner voting rule in this round. We know that $\mathsf{sw}(a_i^+) \geq \mathsf{sw}(a_i^*)$ and since the single-winner rule has expected distortion of $d$, we have $\mathsf{E}[\mathsf{sw}(a_i)] \geq \frac{\mathsf{sw}(a_i^+)}{d}$ which implies $\mathsf{E}[\mathsf{sw}(a_i)] \geq \frac{\mathsf{sw}(a_i^*)}{d}$. Summing this over all iterations and using linearity of expectation, we get that

$$\sum_{i=0}^{k} \mathsf{E}[\mathsf{sw}(a_i)] \geq \sum_{i=0}^{k} \mathsf{sw}(a_i^*) \,/\, d$$
$$\implies \mathsf{sw}(A^*) \,/\, \mathbb{E}[\mathsf{sw}(S)] \leq d. \qquad \square$$

## H.6 Proofs from Section 13.5.1 ($k$-Approvals)

### H.6.1 Proof of Proposition 13.5.3

**Proposition H.6.1** (LB, 1-app, Deterministic)**.** *For 1-approval ballot format, every deterministic rule $f$ has distortion*

$$\mathit{dist}_{\text{1-app}}(f) \in \Omega\left(\frac{m^2}{\gamma_{\min}}\right).$$

*Proof.* We take $m$ to be sufficiently large. Consider an instance with $\frac{m}{2}$ alternatives $a_1, \ldots, a_{m/2}$ of cost 1 and $\frac{m}{2}$ alternatives $b_1, \ldots, b_{m/2}$ of cost $\frac{2}{m}$, and all the voters have the same PS-value of $\gamma = \gamma_{\min}$. Suppose $\frac{2n}{m}$ voters vote for each $a_i$.

If a PB rule picks the bundle $b_1, \ldots, b_{m/2}$, then consider the instance where every voter assigns a value of 1 to each $a_i$ and a value of 0 to each $b_i$. This is consistent with the input, and results in infinite distortion.

Instead, suppose the PB rule, without the loss of generality, picks $a_{m/2}$. Then, suppose that every voter who votes for $a_{m/2}$ gives it a value of $\gamma\frac{m-2}{m-2\gamma_{\min}}$, and everything else a value of 0, and suppose that all other voters give their top choice a value of 1, the $b_i$ a value of $\frac{m-\gamma(m-2)}{m-2\gamma}$, and everything else a value of zero.

Then, $\mathrm{sw}(b_i) = \frac{m-\gamma(m-2)}{m-2\gamma} \cdot \frac{m-2}{m} \cdot n$ for all $i$ from 1 to $\frac{m}{2}$, and $\mathrm{sw}(a_i) = \frac{2n}{m}$ for $i \neq \frac{m}{2}$ with $\mathrm{sw}(a_{m/2}) = \frac{2n}{m} \cdot \gamma\frac{m-2}{m-2\gamma}$.

Then, the utilities for voters $i$ who vote for $a_{m/2}$ are consistent as

$$\begin{aligned} v_i(a_{m/2}) &= (1-\gamma)\frac{m-2}{m-2\gamma} + \gamma\frac{m-2}{m-2\gamma}\frac{2}{m} \\ &= \frac{m-2}{m-2\gamma}\left(1-\gamma\frac{m-2}{m}\right) \\ &= \frac{m-2}{m-2\gamma}\frac{m-\gamma(m-2)}{m} \\ &\geq \gamma\frac{m-\gamma(m-2)}{m-2\gamma}\frac{m-2}{m} = v_i(b_j) \end{aligned}$$

for all $b_j$, where the last inequality holds because $m \geq m - 2\gamma$. Similarly,

$$\begin{aligned} v_i(a_{m/2}) &= (1-\gamma)\frac{m-2}{m-2\gamma} + \gamma\frac{m-2}{m-2\gamma}\frac{2}{m} \\ &= \frac{m-2}{m-2\gamma}\frac{m-\gamma(m-2)}{m} \\ &\geq \gamma\frac{2}{m} = v_i(a_j) \end{aligned}$$

for all $a_j \neq a_{m/2}$, where the last inequality holds for sufficiently large $m$, so $a_{m/2}$ is indeed the alternative of highest value.

The utilities of voters $i$ who vote for $a_j \neq a_{m/2}$ is consistent as:

$$\begin{aligned} v_i(b_i) &= (1-\gamma)\frac{m-\gamma(m-2)}{m-2\gamma} + \gamma\frac{m-\gamma(m-2)}{m-2\gamma} \cdot \frac{m-2}{m} \\ &= \frac{m-\gamma(m-2)}{m-2\gamma}\left(1-\gamma+\gamma\frac{m-2}{m}\right) \\ &= \frac{m-\gamma(m-2)}{m} \\ &= (1-\gamma) + \gamma \cdot \frac{2}{m} = v_i(a_j) \end{aligned}$$

for all $b_i$. And $v_i(a_j) \geq v_i(a_k)$ for all $k \neq j$ as $\mathrm{sw}(a_k) \leq \mathrm{sw}(a_j)$ and voter $i$ gives $a_k$ zero utility. So, $a_j$ is indeed the highest ranking alternative.

But, the distortion we get is:

$$\begin{aligned}\frac{\sum_i \mathsf{sw}(b_i)}{\mathsf{sw}(a_{m/2})} &= \frac{m}{2} \cdot \frac{m-\gamma(m-2)}{m-2\gamma} \cdot n \cdot \left(\frac{2n}{m} \cdot \gamma \frac{m-2}{m-2\gamma}\right)^{-1} \\ &= \frac{m^2}{4} \cdot \frac{m-\gamma(m-2)}{\gamma(m-2)} \\ &= \frac{m^2}{4} \cdot \left(\frac{1}{\gamma} \cdot \frac{m}{m-2} - 1\right) \\ &\geq \frac{m^2}{4} \cdot \frac{1-\gamma}{\gamma},\end{aligned}$$

as claimed. □

## H.7 Proofs from Section 13.5.2 (Knapsack)

### H.7.1 Proof of Theorem 13.5.6

**Theorem 13.5.6** (LB, knap, Randomized)**.** *For knapsack ballot format, every randomized rules $f$ has distortion*

$$dist_{knap}(f) \geq m(1-\gamma_{\min}) + \gamma_{\min}.$$

*Proof.* Formally, consider a case where $n$ is divisible by $m$, all the voters have the same PS-value of $\gamma = \gamma_{\min}$, and every alternative $a \in A$ has a cost of $c_a = 1$. In this case, each vote consists of exactly one alternative. For any alternative $a \in A$, let $N_a$ be the set of voters who vote for alternative $a$. Choose the input profile $\vec{\rho}$ so that $n/m$ voters vote for each alternative so that $|N_a| = \frac{n}{m}$ for all $a \in A$. Our randomized voting rule $f$ must pick some alternative $a^*$ with probability at most $1/m$.

Suppose that all voters in $N_{a^*}$ have a utility of $\frac{m(1-\gamma)+\gamma}{\gamma}$ for $a^*$ and utility zero for everything else. Moreover, voters in $N_a$ with $a \neq a^*$ have utility 1 for $a$ and zero utility for the rest of the alternatives. We can see that the social welfare of $a^*$ is $\frac{m(1-\gamma)+\gamma}{\gamma} \cdot \frac{n}{m}$, and the social welfare of any other alternative is $\frac{n}{m}$.

First of all, we have to make sure that this utility matrix and PS-vector yield a value matrix consistent with the input profile. For any $a \neq a^*$ and $i \in N_a$ we have:

$$\begin{aligned}v_i(a^*) &= \gamma \frac{m(1-\gamma)+\gamma}{\gamma} \cdot \frac{1}{m} \\ &= \frac{m(1-\gamma)+\gamma}{m} = (1-\gamma) + \frac{\gamma}{m} \\ &= v_i(a).\end{aligned}$$

Furthermore, for voter $i \in N_{a^*}$ and any $a \neq a*$ as:

$$\begin{aligned} v_i(a^*) &= (1-\gamma)\frac{m(1-\gamma)+\gamma}{\gamma} + \gamma\frac{m(1-\gamma)+\gamma}{\gamma} \cdot \frac{1}{m} \\ &= \left(1 - \gamma\frac{m-1}{m}\right)\frac{m(1-\gamma)+\gamma}{\gamma} \\ &= \frac{m-\gamma(m-1)}{m} \cdot \frac{m(m-\gamma)+\gamma}{\gamma} \\ &= \frac{\gamma}{m} \cdot \frac{(1-\gamma)m+\gamma}{\gamma} \cdot \frac{m(m-\gamma)+\gamma}{\gamma} \\ &\geq \frac{\gamma}{m} = v_i(a), \end{aligned}$$

where the last inequality follows from the fact that $\gamma \leq 1$. That means the value matrix is consistent with the input profile for all the voters.

After that, we can see the distortion that the rule incurs. We could have gotten a utility of $\frac{n}{m} \cdot \frac{m(1-\gamma)+\gamma}{\gamma}$ by choosing $a^*$, but instead, we got the expected utility of the following

$$\begin{aligned} \mathsf{E}_{a\sim f(\vec{\rho},c)}[\mathsf{sw}(a)] \leq& \frac{1}{m}\mathsf{sw}(a^*) + \frac{m-1}{m} \cdot \frac{n}{m} \\ &= \frac{1}{m} \cdot \frac{n}{m} \cdot \frac{m(1-\gamma)+\gamma}{\gamma} + \frac{m-1}{m} \cdot \frac{n}{m} \\ &= n\left(\frac{m(1-\gamma)+\gamma+(m-1)\gamma}{m^2\gamma}\right) \\ &= \frac{n}{\gamma m}, \end{aligned}$$

and so the distortion is at least:

$$\begin{aligned} \mathsf{dist}_{\mathsf{knap}}(f,\vec{\rho},c) &= \frac{\mathsf{sw}(a^*)}{\mathsf{E}_{a\sim f(\vec{\rho},c)}[\mathsf{sw}(a)]} \\ &\geq \frac{\frac{n}{m} \cdot \frac{m(1-\gamma_{\min})+\gamma_{\min}}{\gamma_{\min}}}{\frac{n}{\gamma_{\min} m}} \\ &= m(1-\gamma_{\min}) + \gamma_{\min}. \end{aligned}$$

□

### H.7.2 Knapsack for Committee Selection

We can improve the analysis of the knapsack voting when all alternatives have the same cost.

**Theorem H.7.1.** *We can get a distortion of* $1 + \frac{m}{2} + \frac{1-\gamma_{\min}}{\gamma_{\min}}m^2$ *in the deterministic knapsack setting for* $m/2$*-multiwinner elections (or equivalently when* $c_a = \frac{2}{m}$ *for all* $a \in A$*).*

*Proof.* Recall the notation used in the proof of Theorem 13.5.8. For any subset of alternatives $S \subseteq A$, let $n_S := \sum_{i \in N} \mathbb{I}(S \subseteq \rho_i)$ be the number of voters whose knapsack set contains $S$. We use shorthand $n_a := n_{\{a\}}$ and $n_{a,b} := n_{\{a,b\}}$ for all $a, b \in A$. Then, informally, $n_{a,b}$ is the number of voters who vote for both $a$ and $b$.

The voting rule we will use is as follows: assign a plurality score to each alternative, where the score is simply the number of times each alternative appears.

Pick the $m/2$ alternatives with the largest plurality score, $A$. Indeed, every alternative can appear at most $n$ times, as every voter can vote for them only once. Therefore, in the worst case, if the top $m/2-1$ alternatives appear $n$ times there must remain $nm/2 - n(m/2-1) = n$ appearances of other alternatives. By the pigeonhole principle from here, the remaining plurality winner must be chosen $n/(m/2+1) > n/m$ times. Thus, the minimum number of times a plurality winner can appear is $n/m$.

Moreover, because $n_a > n_b$ for all $a \in A$ and $b \notin A$, and $\sum_{a \in A} n_a + \sum_{b \notin A} n_b = mn/2$, we get that $2\sum_{a \in A} n_a \geq mn/2$ and so $\sum_{a \in A} n_a \geq mn/4$.

Then, let $A^*$ be the optimal set of alternatives. Note then that:

$$\begin{aligned}
\frac{\mathrm{sw}(A^*, U)}{\mathrm{sw}(A, U)} &= \frac{\sum_{a^* \in A^*} \mathrm{sw}(a^*, U)}{\sum_{a \in A} \mathrm{sw}(a, U)} \\
&= \frac{\sum_{a^* \in A^* \cap A} \mathrm{sw}(a^*, U)}{\sum_{a \in A} \mathrm{sw}(a, U)} + \frac{\sum_{a^* \in A^* \setminus A} \mathrm{sw}(a^*, U)}{\sum_{a \in A} \mathrm{sw}(a, U)} \\
&\leq 1 + \sum_{a^* \in A^* \setminus A} \frac{\mathrm{sw}(a^*, U)}{\sum_{a \in A} \mathrm{sw}(a, U)}. \qquad \text{(H.1)}
\end{aligned}$$

We will show that for all $a^* \in A^* \setminus A$, there exists some $a \in A$ such that:

$$\frac{\mathrm{sw}(a^*)}{\mathrm{sw}(a)} \leq 2\frac{1-\gamma_{\min}}{\gamma_{\min}} m + 1,$$

by considering two cases:

1. Suppose that for all $a^* \in A^* \setminus A$, there exists some $a \in A$ such that $n_{a,a^*}/n_a \leq 1/2$. Then, $n_a - n_{a,a^*} \geq n_a/2 \geq n/2m$. Therefore, by Lemma 13.2.1:

$$\frac{\mathrm{sw}(a^*)}{\mathrm{sw}(a)} \leq 2\frac{1-\gamma_{\min}}{\gamma_{\min}} m + 1.$$

2. Suppose that for some $a^* \in A^* \setminus A$, and for all $a \in A$, $n_{a,a^*}/n_a > 1/2$. Let $a_{\max} = \arg\max_{a \in A} n_a$

and $a_{\min} = \arg\min_{a \in A} n_a$. Then, in particular,

$$\begin{aligned} n_{a_{\max}} &< 2n_{a_{\max},a^*} \\ &\leq 2n_{a^*} \\ &\leq 2n_{a_{\min}}, \end{aligned}$$

where the last equality holds because $a_{\min}$ is a plurality winner, and $a^*$ isn't

Since $(m/2)n_{a_{\max}} \geq \sum_{a \in A} n_a \geq nm/4$, $n_{a_{\max}} \geq n/2$ and so $n_{a_{\min}} \geq n/4$. Therefore, we can improve the lower bound for plurality winners: for all $a \in A$, $n_a \geq n/4$.

By Lemma H.7.2 below, we know that for all $a^* \in A^* \setminus A$, there exists some $a \in A$ such that $n_{a,a^*}/n_a \leq (m-2)/m$. Therefore, $n_a - n_{a,a^*} \geq 2n_a/m \geq n/2m$. Thus, by Lemma 13.2.1 in [134]:

$$\frac{\mathrm{sw}(a^*)}{\mathrm{sw}(a)} \leq 2\frac{1-\gamma_{\min}}{\gamma_{\min}}m + 1.$$

From here we can prove an $m^2$ bound easily by taking $a^*_{\max} = \mathrm{argmax}_{a^* \in A^*} \mathrm{sw}(a^*, U)$. Then, continuing off of (H.1), and using the fact that there exists some $\hat{a} \in A$ such that $\frac{\mathrm{sw}(a^*_{\max},U)}{\mathrm{sw}(\hat{a},U)} \leq 2\frac{1-\gamma_{\min}}{\gamma_{\min}}m + 1$:

$$\begin{aligned} \frac{\mathrm{sw}(A^*,U)}{\mathrm{sw}(A,U)} &\leq 1 + \frac{m}{2} \cdot \frac{\mathrm{sw}(a^*_{\max},U)}{\sum_{a \in A} \mathrm{sw}(a,U)} \\ &\leq 1 + \frac{m}{2} \cdot \frac{\mathrm{sw}(a^*_{\max},U)}{\mathrm{sw}(\hat{a},U)} \\ &\leq 1 + \frac{1-\gamma_{\min}}{\gamma_{\min}}m^2 + \frac{m}{2}, \end{aligned}$$

as claimed! □

**Lemma H.7.2.** *When $A^*$ is the optimal subset and $A$ is the subset chosen by the repeated plurality rule, for all $a^* \in A^* \setminus A$, there exists some $a \in A$ such that:*

$$\frac{N(a,a^*)}{N(a)} \leq (m-2)/m.$$

*Proof.* Note that $\sum_{a \in A} N(a,a^*)$ is the number of times a voter votes for some alternative and $a^*$. Each voter can vote for at most $m/2$ alternatives. Since there are then at most $m/2-1$ alternatives in $A$ that any voter who votes for $a^*$ could have voted for:

$$\sum_{a \in A} N(a,a^*) \leq N(a^*)(m/2-1) \leq N(a^*) \cdot \frac{m-2}{2}.$$

From here, let $a_{\min} = \operatorname{argmin}_{a \in A} N(a, a^*)$. Then, substituting this into the inequality above, and using that $|A| = \frac{m}{2}$:

$$\frac{m}{2} N(a_{\min}, a^*) \leq N(a^*) \cdot \frac{m-2}{2}.$$

Since $N(a^*) \leq N(a_{\min})$ as $a^*$ is not in $A$ and therefore must occur at most as many times as any plurality winner,

$$\frac{m}{2} N(a_{\min}, a^*) \leq N(a_{\min}) \cdot \frac{m-2}{2},$$

and so finally

$$\frac{N(a_{\min}, a^*)}{N(a_{\min})} \leq \frac{m-2}{m},$$

as desired! □