# Calibration with Privacy in Peer Review: A Theoretical Study

**Wenxin Ding**

CMU-CS-21-127

August 2021

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

**Thesis Committee:**
Nihar B. Shah (Co-Chair)
Weina Wang (Co-Chair)
Giulia Fanti

*Submitted in partial fulfillment of the requirements
for the degree of Master of Science.*

This work was supported by a National Science Foundation CAREER Award 1942124, NSF Grant CIF 1763734, and a Google Research Scholar Award.

**Keywords:** Peer-review, calibration, utility privacy tradeoff, Pareto frontier

# Abstract

Reviewers in peer review are often miscalibrated: they may be strict, lenient, extreme, moderate, etc. Various attempts have been made to calibrate reviews in conference peer review, but they are hampered by the critical bottleneck of a small number of samples (reviews) per reviewer. To increase the sample sizes, we consider using exogenously obtained information about reviewers' calibration, such as data from past conferences. The problem with this approach is that it may compromise the privacy of which reviewer reviewed which paper. We formulate this problem as that of calibrating reviews while ensuring privacy. We undertake a theoretical study of this problem under a simplified yet challenging model involving two reviewers, two papers, and a MAP-computing adversary. Our main results establish the Pareto frontier of the tradeoff between privacy and utility (accepting the better papers), and design computationally-efficient algorithms that are Pareto optimal. Our work provides a foundation for future research to address the important problem of miscalibration on a larger scale.

This thesis contains joint work with Nihar B. Shah, Weina Wang, and Gautam Kamath.

## Acknowledgments

I would like to thank Prof. Nihar B. Shah and Prof. Weina Wang for being amazing advisors and guiding me through my way in scientific research. They led my way to research in the field of computer science and always inspired me with great ideas and helpful advice. I am grateful for their valuable feedback that will keep echoing in my future life and study.

I would like to thank Prof. Nihar B. Shah, Prof. Weina Wang, and Prof. Gautam Kamath for working with me on this project. They brought great inspirations and insights to this project and taught me important skills in scientific research. I would like to thank Prof. Giulia Fanti for being on my thesis committee and provided valuable feedback.

Thanks to my family and friends for supporting and encouraging me at all times. Thank you for helping me grow over the years and being witnesses of my growth.

# Contents

# List of Figures

# Chapter 1

# Introduction

It is well known that ratings provided by people are frequently miscalibrated. In the application of peer review, reviewers may be strict, lenient, extreme, moderate, or have other forms of miscalibration. This leads to unfairness in peer review, for instance, disadvantaging papers that happen to go to strict reviewers [28]: *"the existence of disparate categories of reviewers creates the potential for unfair treatment of authors. Those whose papers are sent by chance to assassins/demoters are at an unfair disadvantage, while zealots/pushovers give authors an unfair advantage."*

A number of algorithms are proposed in the literature to address the problem of miscalibration (detailed in Section 2). Program chairs of conferences have tried to use some algorithms to calibrate reviewers' scores, but have found the outcomes to be unsatisfactory. For instance, John Langford, the program chair of ICML 2012 says that *"We experimented with reviewer normalization and generally found it significantly harmful"* [17].

A critical challenge in addressing miscalibration is the tiny sample size. Many conferences have each reviewer reviewing just a handful papers (typically 1 to 6 papers). As a consequence, it is often hard to decipher the miscalibration of any reviewer, particularly since human miscalibration can be quite complex [4]. Our approach to address this challenge is to use exogenous information about the miscalibration of reviewers, e.g., reviewers' calibration information from other conferences where they have reviewed.

In peer review, the identity of which reviewer reviews which paper is confidential. A naïve attempt at calibration can compromise this confidentiality. As an example, consider an adversary trying to guess the reviewer of a paper between two possibilities – reviewer X or reviewer Y. The review for the paper is lukewarm, and for simplicity suppose this is the only review. We assume the "OpenReview" model where all submitted papers, reviews, and final decisions are public (but reviewer identities are not). Also suppose it is known that reviewer X is quite strict but reviewer Y is not. Then the paper will not be accepted unless the conference performs a calibration using this information *and* the reviewer is X. The acceptance of the paper will provide the adversary the necessary information to infer the reviewer as X.

The compromise in confidentiality due to calibration can occur in the conference at hand (as in the example above), or in previous conferences if data from them used. In the conference at hand, such compromise could occur when calibrating using exogenous information (as in the example above) or using information only from within the conference (as done in past literature).

This problem of privacy in calibration that we identify is quite challenging in full generality, and in this paper, we analyze a specific setting of privacy in the conference at hand when calibrating using exogenous information. We consider a simplistic–yet challenging–model with two reviewers and two papers and where an adversary attempts to guess the reviewer assignment based on maximum a posteriori (MAP) computation. Our contributions are summarized as follows:

- We identify the problem of privacy in calibration, and we initiate a theoretical study with the formulation of a specific problem that incorporates various key challenges of the more general setting.
- We provide an algorithm for calibration with privacy that optimally trades off the error of the conference (in terms of accepting the better paper) and the error of the adversary (in terms of guessing the reviewer).
- We establish the structure of the Pareto optimal curve between the two aforementioned desiderata. We observe that there is a linear tradeoff between the two errors up to a certain point, after which the error of the adversary does not decrease even if the conference adds more randomness in its protocols.

Our work aims to found a building block for more research on this important problem of miscalibration (possibly using exogenous information about calibration) while ensuring privacy of reviewers.

# Chapter 2

# Related Work

Peer review is extensively used for evaluating scientific papers and grant proposals, although it is known to suffer from various challenges such as miscalibration, biases, subjectivity, dishonest behavior and others [27].

The problem of miscalibration is well recognized in the literature. A common approach to design calibration algorithms is to assume a certain model of miscalibration, and under the assumed model, estimate the calibrated data (or the model parameters) from the data. The papers [2, 9, 12, 19, 24, 25, 26] in this line of literature assume affine models: it assumes that each paper has some "true" real-valued quality and that the score provided by any reviewer is some affine transform (plus noise) of this true quality. The affine transform captures the reviewer's miscalibration. In our formulation (detailed subsequently in Section 3) we also assume papers have true qualities, and a part of our work also assumes affine miscalibrations.

A second line of literature [1, 10, 23] recognizes the problem of miscalibration, and takes the approach of using only the ranking of papers induced by the ratings given by any individual reviewer, or alternatively, asking each reviewer to only provide a ranking of the papers they are reviewing. Using rankings alone thus gets rid of any miscalibrations, but on the downside, can lose some information contained in ratings. Moreover, a recent work [35] showed that under certain settings, ratings can yield more information than rankings even if the miscalibration is adversarial.

Notably, these works consider addressing miscalibration using data from within the conference at hand, and moreover do not consider the issue of compromise of privacy.

We assume an "open review" model where all submitted papers and all reviews are available publicly (see `openreview.net`). This model is followed in the ICLR conference as well as other venues. In a survey [29] at the ICLR 2013 conference, researchers felt that this open review model leads to benefits of more accountability of authors (in terms of not submitting below-par papers) as well as reviewers (in terms of giving high-quality reviews). The publicly available data has resulted in another benefit: it has yielded a rich dataset for research on peer review [3, 15, 20, 34, 36, 37]. A downside of the open review approach is that if a rejected paper is resubmitted elsewhere, the (publicly available) knowledge of previous rejection may bias the reviewer [31].

Our work considers explicitly randomized assignments and decisions. In practice, the assignments and decision protocols are typically deterministic (although some variations naturally arise due to human involvement in various parts of the peer review process). The assignment of reviewers to papers is done by solving a certain optimization problem [5, 11, 13, 16, 30, 32] involving similarities computed between each pair of reviewer and paper [5, 8, 21, 22]. Decisions are arrived at after discussions between the reviewers. That said, there are notable instances where randomization has been explicitly used in practice in peer review: randomization can help mitigate dishonest behavior [14] and can help make more fair decisions for borderline papers or grants [6, 18]. Finally, the algorithms in the theoretical work [35] also employ randomization.

Issues of privacy in peer review also arise when releasing data to researchers. The program chairs of the WSDM 2017 conference performed a remarkable controlled experiment to test for biases in peer review, and in their paper [33] they point out privacy-related concerns in releasing data: *"We would prefer to make available the raw data used in our study, but after some effort we have not been able to devise an anonymization scheme that will simultaneously protect the identities of the parties involved and allow accurate aggregate statistical analysis. We are familiar with the literature around privacy preserving dissemination of data for statistical analysis and feel that releasing our data is not possible using current state-of-the-art techniques."* We are aware of two past works which deal with privacy in peer review [7, 14]. In particular, both papers consider privacy-preserving release of peer-review data. The paper [7] provides an algorithm to optimize utility when releasing histograms pertaining to the reviews, miscalibration or subjectivity in a privacy-preserving manner. The paper [14] uses randomized assignments to guarantee privacy of the reviewer-paper assignment when data pertaining to similarities between reviewer-paper pairs is released.

# Chapter 3

# Problem Formulation and Preliminaries

In this section, we present the formal problem specification.

**Papers and reviewers.** We consider a setting with two reviewers and two papers. Each paper $i \in \{1,2\}$ has some true quality $\theta_i \in \mathbb{R}$. We assume that the qualities $\theta_1$ and $\theta_2$ are drawn i.i.d. according to the standard Normal distribution.

**Reviewer assignment.** Each reviewer reviews one paper and each paper is reviewed by one reviewer. There are thus two possible assignments: we let $A_1$ denote the assignment of reviewer 1 to paper 1 and reviewer 2 to paper 2, and $A_2$ denote the assignment of reviewer 1 to paper 2 and reviewer 2 to paper 1. We assume that the assignment is chosen uniformly at from these two possibilities. We use $\mathcal{A}$ to denote the random variable for the assignment.

**Miscalibration and reviewer scores.** Following [35], we assume that each reviewer $j \in \{1,2\}$ has a function $\beta_j : \mathbb{R} \to \mathbb{R}$ which captures their miscalibration. If reviewer $j \in \{1,2\}$ reviews paper $i \in \{1,2\}$, we assume that the reviewer provides a score

$$\beta_j(\theta_i) + \epsilon_j,$$

where $\epsilon_j$ is a Gaussian random variable with mean zero independent of everything else. We call $\beta_j$ the *miscalibration function* of reviewer $j$. We assume that the functions $\beta_1$ and $\beta_2$ are increasing and invertible. In one part of our work, we further make an assumption that the miscalibration functions are affine, and we detail this subsequently in the associated section.

We let $s_1$ denote the score received by paper 1 and $s_2$ denote the score received by paper 2.

The realizations of these scores are $S = [s_1, s_2]$ where .

We use $s$ to denote the final score given by a reviewer. Such final score $s$ also has a distribution and we denote the probability density function of final score given by reviewer $j$ as $f_j$.

**Conference.** The goal of the conference is to accept high-quality papers. In the scenario with two papers, the conference wants to choose the paper with higher quality between the two. The conference has full information on identities and scores of papers, identities, miscalibration functions, noise distributions of reviewers, and the true assignment. Conference does calibration by using some strategy to improve its probability of accepting higher-quality papers.

**Adversary.** To measure the privacy leakage of calibration, we assume an adversary in the process. The goal of the adversary is to guess the true assignment. The adversary has information on identities and scores of papers, identities, miscalibration functions, and noise distributions of reviewers, the mechanism used by the conference to do calibration and the decision made (i.e., the paper being accepted) by the conference. However, since the adversary does not know the true assignment, it does not know the output of $h$ in the realization. The adversary predicts the assignment that maximizes the posterior probability given all its observations and knowledge.

**Settings and errors.** We study two settings, noiseless and noisy. If both reviewers' noises are constant 0, then the reviewer is noiseless and it is the noiseless setting. Otherwise, the reviewer is noisy and it is the noisy setting. There are two types of error in the analysis, per-instance error and average-case error. Per-instance error refers to the error with respect to some specific scores. In this case, the conference observes the scores and does calibration, we then look at the error probability for both the conference and adversary under the calibration strategy used by the conference. On the other hand, average-case error is computed across the distributions of scores. The conference has a strategy before seeing the sores and we study the error of the strategy under the distribution of scores for both the conference and the adversary.

**Definition 3.0.1.** Per-instance error of the conference given $S = [s_1, s_2]$, denoted as $\mathcal{E}_C([s_1, s_2])$, is computed as $\Pr($conference accepts lower-quality paper $|S = [s_1, s_2])$.

**Definition 3.0.2.** The average-case error of the conference is the average per-instance error, where the averaging is done over the distribution of the scores $s_1$ and $s_2$. Specifically, $\int_{s_1} \int_{s_2} \Pr($conference accepts lower-quality paper $|S = [s_1, s_2])f_S([s_1, s_2])$ where $f_S$ is the p.d.f of the joint distribution of $[s_1, s_2]$.

**Definition 3.0.3.** Per-instance error of the adversary given $S = [s_1, s_2]$, denoted as $\mathcal{E}_A([s_1, s_2])$, is computed as $\Pr($adversary guesses wrong assignment $|S = [s_1, s_2])$.

**Definition 3.0.4.** The average-case error of the adversary is the average per-instance error, where the averaging is done over the distribution of the scores $s_1$ and $s_2$. Specifically $\int_{s_1} \int_{s_2} \Pr($adversary guesses wrong assignment $|S = [s_1, s_2])f_S([s_1, s_2])$ where $f_S$ is the p.d.f of the joint distribution of $[s_1, s_2]$.

**Pareto frontier.** A calibration strategy is Pareto efficient if for any given maximum error of the conference, it maximized the error of the adversary with the smallest error of the conference under the threshold. Therefore, we define the Pareto frontier as follows:

**Definition 3.0.5.** A Pareto frontier of the error of the adversary against the error of the conference contains all points such that the error of the adversary cannot be increased without increasing the error of the conference.

**Our goal.** Our goal is to design a strategy for the conference to calibrate the scores of the papers while protect the privacy of the assignments. For given desired error of the conference, we would like to maximize the error of the adversary. In the analysis of the two-paper two-reviewer scenario, we find the Pareto frontier of the errors and conclude an optimal strategy for the conference to calibrate.

**Calibration.** Calibration is a function that takes the information known by the conference and outputs the paper to be accepted. The most general strategy for conference calibration is to accept each paper with a certain probability. The conference observes the scores $S = [s_1, s_2]$ and the assignment $A$ and the most general strategy the conference can use is to have $g : S \times A \to [0, 1]$. The function $g$ decides the probability that the conference accepts paper 1 and we call function $g$ a probability function. The conference accepts paper 1 with probability $g(S, A)$ and accepts paper 2 with probability $1 - g(S, A)$. The conference then decides the paper to accepts using the probabilities.

If the conference uses the probability function $g$ to do calibration, the error of the conference is computed as $((1 - g(S, A_1)) \Pr(\mathcal{A} = A_1|\theta_1 > \theta_2, S) + (1 - g(S, A_2)) \Pr(\mathcal{A} = A_2|\theta_1 > \theta_2, S)) \Pr(\theta_1 > \theta_2|S) + (g(S, A_1) \Pr(\mathcal{A} = A_1|\theta_1 < \theta_2, S) + g(S, A_2) \Pr(\mathcal{A} = A_2|\theta_1 < \theta_2, S)) \Pr(\theta_1 < \theta_2|S)$.

Another way of doing calibration to minimize error of the conference is to compute the MAP of the quality of the paper. When the conference calibrates in the two paper scenario, it accepts the paper with higher probability to be higher in quality between the two using the known information. That is to say, the conference computes $\Pr(\theta_1 > \theta_2|S, A)$ and if the value is greater than $\frac{1}{2}$ then it accepts paper 1. If the value is less than $\frac{1}{2}$, the conference accepts paper 2. In the case of the value equal to $\frac{1}{2}$, we assume the conference accepts a paper uniformly at random. The error of the conference $\mathcal{E}_C$ is the probability that the conference accepts the paper with lower quality.

To simplify our analysis, we consider a mechanism for the conference to do calibration with MAP. Observing the scores $S = [s_1, s_2]$ and the assignment $A$, the conference uses a function $h : S \times A \to [0, 1]$ to decide the probability of calibrating the scores under the true assignment. We call function $h$ a calibration function. After seeing the result of $h$, the conference first decides the assignment it calibrates under. It calibrates under the true assignment with probability $h(S, A)$ and under the wrong assignment with probability $1 - h(S, A)$. Then the conference calibrates under the chosen assignment and makes decision.

The following lemma states that switching from the most general mechanism of probability function $g$ to the mechanism using calibration function $h$ does not reduce optimality of the conference. So the conference can use calibration function to make decisions of paper acceptance and our analysis is done under this mechanism.

**Lemma 3.0.6.** *Without loss of optimality, the conference can calibrate use the mechanism of calibration function $h$ instead of the mechanism of probability function $g$.*

# Chapter 4

# Main Results

We study a peer review process with two papers and two reviewers. We analyze the tradeoff between utility and privacy of the process using the Pareto frontier of error of the conference against error of the adversary. We study the Pareto frontier and find optimal strategy for the conference to do calibration.

## 4.1 Noiseless Setting

We first study the noiseless setting where both reviewers' noises are constant 0. In the noiseless case, the conference always accepts the higher-quality paper when it calibrates under the correct assignment by finding the quality of papers using inverse functions of miscalibration functions and the scores. Lemma 3.0.6 indicates that if calibrating under both assignments makes the conference accept the same paper, then there is no need to have non-zero probability to accept the other paper. Otherwise, the error of the conference is increased but the error of the adversary remains the same.

### 4.1.1 Pareto Frontier

We first find the Pareto frontier of the error of the adversary against the error of the conference in the noiseless setting. For the given scores, if the same paper has higher quality under both assignments, then the output of the calibration function $h$ does not make a difference to the conference decision. Therefore, for scores in such range, the Pareto efficient situation is where the conference has zero error. For the rest of the scores, the conference shall accept different papers by calibrating under different assignments. Then the calibration function needs to be carefully designed.

Given scores $S = [s_1, s_2]$ and knowing the miscalibration functions, the conference can calculate the true qualities of papers under each assignment. Under $A_1$ we have $\theta_1 = \beta_1^{-1}(s_1)$ and $\theta_2 = \beta_2^{-1}(s_2)$. Under $A_2$ we have $\theta_1 = \beta_2^{-1}(s_1)$ and $\theta_2 = \beta_1^{-1}(s_2)$. If $s_1 > \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$, then $\theta_1 > \theta_2$ under both assignments and paper 1 should be accepted. Similarly, if $s_1 < \min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$, paper 2 should be accepted. We will study the Pareto frontier for scores where $\min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\} \leq s_1 \leq \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$.

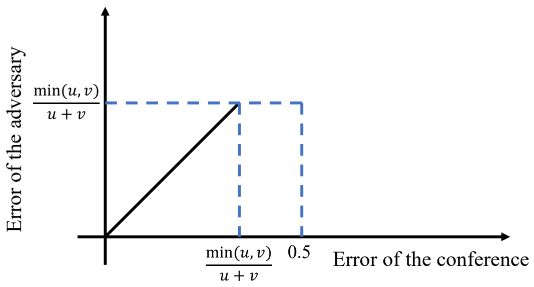


Figure 4.1: Pareto frontier in the noiseless setting with $u = f_1(s_1)f_2(s_2), v = f_2(s_1)f_1(s_2)$.

In this range of scores, the error of the adversary does not surpass the error of the conference. The Pareto frontier is a line with slope 1 as stated in the following theorem:

**Theorem 4.1.1.** *In the noiseless setting where the scores satisfy* $\min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\} < s_1 < \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$*, the Pareto frontier of error of the adversary against error of the conference is a line of slope* 1 *starting from the origin as shown in Figure 4.1.*

### 4.1.2 Per-instance Error

We analyze the per-instance error with a realization of the scores $S = [s_1, s_2]$. From Lemma 3.0.6, we know that if paper 1 has higher quality under both assignments, the conference should accept paper 1 and the same argument holds for paper 2. For the rest range of scores, the conference adopts the calibration mechanism with function $h$.

Since $S$ is a fixed realization in the analysis, we simplify the mechanism for the conference as

$$q_1 = h(S, A_1)$$
$$q_2 = h(S, A_2).$$

With the simplification, $q_1$ is the probability for the conference to calibrate under the true assignment when the true assignment is $A_1$ and $q_2$ is such probability when the true assignment is $A_2$. Therefore, given a value for the maximum error of the conference $\mathcal{E}_C$, our goal is to find values of $q_1$ and $q_2$ that is Pareto efficient.

For a Pareto efficient calibration mechanism, the error of the conference and the adversary should stay on the Pareto frontier as in Figure 4.1. If the error of the conference is less than $\frac{\min(u,v)}{u+v}$ then the mechanism has error of the adversary being the same. Otherwise, error of the conference and the adversary are both $\frac{\min(u,v)}{u+v}$.

**Theorem 4.1.2.** *Algorithm 1 describes the optimal strategy for conference calibration in the noiseless setting with per-instance error* $\mathcal{E}_C([s_1, s_2])$.

### 4.1.3 Average-case Error

We analyze the average-case error with respect to the distributions of scores $S = [s_1, s_2]$.

**Algorithm 1:** Conference calibration with per-instance error in the noiseless setting

```
Input: scores $S = [s_1, s_2]$, maximum per-instance error of the conference $\mathcal{E}_C([s_1, s_2])$
if $s_1 > \max\{\beta_1(\beta_2^{-1}(s_2)), \beta_2(\beta_1^{-1}(s_2))\}$ then
    accept paper 1
else if $s_1 < \min\{\beta_1(\beta_2^{-1}(s_2)), \beta_2(\beta_1^{-1}(s_2))\}$ then
    accept paper 2
else if $\mathcal{E}_C([s_1, s_2]) \geq \frac{\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ then
    choose $q_1, q_2 \in [0, 1]$ such that
    $\min\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}\, q_2 = \max\{f_1(s_1)f_2(s_2), f_2(s_1)f_1(s_2)\}\,(1 - q_1)$
else
    choose $q_1, q_2 \in [0, 1]$ such that $\mathcal{E}_C([s_1, s_2]) = 1 - \frac{f_1(s_1)f_2(s_2)q_1+f_2(s_1)f_1(s_2)q_2}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$
end if
```

**Algorithm 2:** A strategy that always operates at the first change point

```
Input: scores $S = [s_1, s_2]$
if $s_1 > \beta_2(\beta_1^{-1}(s_2))$ then
    accept paper 1
else if $s_1 < \beta_1(\beta_2^{-1}(s_2))$ then
    accept paper 2
else if $f_1(s_1)f_2(s_2) > f_2(s_1)f_1(s_2)$ then
    choose $q_1, q_2 \in [0, 1]$ such that $f_1(s_1)f_2(s_2)q_1 = f_1(s_1)f_2(s_2) - f_2(s_1)f_1(s_2)q_2$
else if $f_1(s_1)f_2(s_2) \leq f_2(s_1)f_1(s_2)$ then
    choose $q_1, q_2 \in [0, 1]$ such that $f_2(s_1)f_1(s_2)q_1 = f_2(s_1)f_1(s_2) - f_1(s_1)f_2(s_2)q_2$
end if
```

Knowing the miscalibration functions and the distribution of quality of papers, the conference can find distributions of the scores. We first consider the strategy of always operating at the first change point within the interesting intervals. Algorithm 2 describes the strategy. We use $\zeta$ to denote the average error of the conference by using Algorithm 2.

Then we have an optimal strategy for the conference to guarantee an average-case error of $\mathcal{E}_C$.

**Theorem 4.1.3.** *Algorithm 3 describes the optimal strategy for conference calibration in the noiseless setting with average-case error $\mathcal{E}_C$.*

This strategy yields no error outside the interesting intervals for the conference and error of the conference equals error of the adversary within the interesting intervals. Thus, it is Pareto efficient in all intervals and is optimal for the conference.

## 4.2 Noisy Setting

We now study the noisy setting. We consider both miscalibration functions $\beta_1$ and $\beta_2$ to be affine and both reviewers' noises $\epsilon_1$ and $\epsilon_2$ to be Gaussian. Furthermore, the distributions of the noise

**Algorithm 3:** Conference calibration with average-case error

```
Input: maximum average-case error of the conference $\mathcal{E}_C$
Let $\zeta$ = error of the conference for always adopting Algorithm 2
if $\mathcal{E}_C > \zeta$ then
    the desired conference error is Pareto inefficient and operate at $\mathcal{E}_C = \zeta$
else if $\mathcal{E}_C = \zeta$ then
    always adopt Algorithm 2
else if $\mathcal{E}_C < \zeta$ then
    toss a coin that has probability of appearing head $\frac{\mathcal{E}_C}{\zeta}$
    if coin appears head then
        always adopt Algorithm 2
    else
        always calibrate under correct assignment
    end if
end if
```

are the same for both reviewers with mean zero and some known variance $\sigma^2$.

$$\begin{aligned}
&\beta_1(\theta) = a_1 \cdot \theta + b_1 \\
&\beta_2(\theta) = a_2 \cdot \theta + b_2 \\
&\epsilon_1 \sim N(0, \sigma^2) \\
&\epsilon_2 \sim N(0, \sigma^2)
\end{aligned}$$

In the noisy case, the conference does calibration by accepting the paper that is more likely to be higher-quality given the scores, miscalibration functions and an assignment. Lemma 3.0.6 still indicates that if calibrating under both assignments makes the conference accept the same paper, then there is no need to have non-zero probability to accept the other paper.

### 4.2.1 Pareto Frontier

We first find the Pareto frontier of the error of the adversary against the error of the conference in the noisy setting. For the given scores, if the same paper has higher quality under both assignments, then the output of the calibration function $h$ does not make a difference to the conference decision. Therefore, for scores in such range, the Pareto efficient situation is where the conference always accepts the paper that is more likely to be higher-quality. For the rest of the scores, the conference shall accept different papers by calibrating under different assignments. Then the calibration function needs to be carefully designed.

Given scores $S = [s_1, s_2]$ and knowing the miscalibration functions, the conference can calculate the probability that paper 1 has higher quality under each assignment. Under $A_1$ we have $\Pr(\theta_1 > \theta_2 | \mathcal{A} = A_1, S = [s_1, s_2]) = 1 - \Phi\left(\frac{a_2(a_1^2+\sigma^2)(s_2-b_2) - a_1(a_2^2+\sigma^2)(s_1-b_1)}{\sqrt{\sigma^2(a_1^2+a_2^2+2\sigma^2)(a_1^2+\sigma^2)(a_2^2+\sigma^2)}}\right)$. Under $A_2$

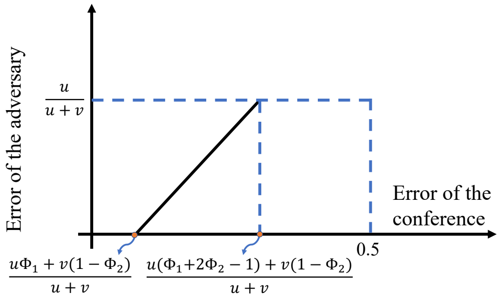


Figure 4.2: A Pareto frontier in the noisy setting with assumptions: $u < v$, $\Phi_1 = \frac{1}{2} - \varphi_1$ and $\Phi_2 = \frac{1}{2} + \varphi_2$ with $0 < \varphi_2 < \varphi_1$ where $u = f_1(s_1)f_2(s_2)$ and $v = f_2(s_1)f_1(s_2)$. Definitions of $\Phi_1$ and $\Phi_2$ are in 4.2.1 and 4.2.2.

we have $\Pr(\theta_1 > \theta_2|\mathcal{A} = A_2, S = [s_1, s_2]) = 1 - \Phi\left(\frac{a_1(a_2^2+\sigma^2)(s_2-b_1)-a_2(a_1^2+\sigma^2)(s_1-b_2)}{\sqrt{\sigma^2(a_1^2+a_2^2+2\sigma^2)(a_1^2+\sigma^2)(a_2^2+\sigma^2)}}\right)$ where $\Phi$ is the cumulative distribution function of the standard normal distribution. For simplicity, we let

$$\Phi_1 = \Phi\left(\frac{a_2(a_1^2+\sigma^2)(s_2-b_2) - a_1(a_2^2+\sigma^2)(s_1-b_1)}{\sqrt{\sigma^2(a_1^2+a_2^2+2\sigma^2)(a_1^2+\sigma^2)(a_2^2+\sigma^2)}}\right) \tag{4.2.1}$$

$$\Phi_2 = \Phi\left(\frac{a_1(a_2^2+\sigma^2)(s_2-b_1) - a_2(a_1^2+\sigma^2)(s_1-b_2)}{\sqrt{\sigma^2(a_1^2+a_2^2+2\sigma^2)(a_1^2+\sigma^2)(a_2^2+\sigma^2)}}\right). \tag{4.2.2}$$

Therefore, if $\Phi_1$ and $\Phi_2$ are both less than $\frac{1}{2}$, the conference should accept paper 1. Similarly, if $\Phi_1$ and $\Phi_2$ are both greater than $\frac{1}{2}$, the conference should accept paper 2. We will study the Pareto frontier for scores where $\Phi_1 - \frac{1}{2}$ and $\Phi_2 - \frac{1}{2}$ have opposite signs. It corresponds to the region where the scores satisfy $\min\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\} \leq s_1 \leq \max\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$. In this range of scores, the Pareto frontier is an increasing line:

**Theorem 4.2.1.** *In the noisy setting where* $\min\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\} \leq s_1 \leq \max\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$*, the Pareto frontier of error of the adversary against error of the conference is an increasing line.*

We will show the Pareto frontier with some additional assumptions in Figure 4.2. Note that removing the assumptions does not affect the shape of the Pareto frontier but the coordinates of the plot. The relationship between $u$ and $v$ combining with the values of $\Phi_1$ and $\Phi_2$ and their distance to $\frac{1}{2}$, we have eight different combinations of these values. In all eight cases, the Pareto frontier is an increasing line and maximum error of the adversary is $\frac{\min(u,v)}{u+v}$.

**Algorithm 4:** Conference calibration with per-instance error in the noisy setting

```
Input:scores $S = [s_1, s_2]$, maximum per-instance error of the conference $\mathcal{E}_C([s_1, s_2])$
if $s_1 > \max\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$ then
    accept paper 1
else if $s_1 < \min\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$ then
    accept paper 2
else if $\mathcal{E}_C([s_1, s_2]) < \frac{u\Phi_1+v(1-\Phi_2)}{u+v}$ then
    error conference of cannot be achieved
else if $\mathcal{E}_C([s_1, s_2]) \geq \frac{u(\Phi_1+2\Phi_2-1)+v(1-\Phi_2)}{u+v}$ then
    choose $q_1 = 1, q_2 = \frac{(v-u)(1-2\Phi_2)}{u+v}$
else
    choose $q_1 = 1, q_2 = \frac{T(\mathcal{E}_C([s_1,s_2]))-(2\Phi_1-1)u}{(1-2\Phi_2)v}$
end if
```

### 4.2.2 Per-instance Error

We analyze the per-instance error with respect to a realization of the scores $S = [s_1, s_2]$. Since $S$ is fixed in the analysis, we simplify the mechanism for the conference as

$$\begin{aligned} q_1 &= h(S, A_1) \\ q_2 &= h(S, A_2). \end{aligned}$$

Therefore, given a value for the maximum error of the conference $\mathcal{E}_C$, our goal is to find values of $q_1$ and $q_2$ that maximize the error of the adversary $\mathcal{E}_A$. We carry the notations from Section 4.2.1 and Figure 4.2. In addition, we let $T(\mathcal{E}_C) = \mathcal{E}_C(u + v) - u \cdot (1 - \Phi_1) - v \cdot \Phi_2$ to be a function that takes the error of the conference $\mathcal{E}_C$ as input.

Along with Figure 4.2, we present the algorithm with the assumptions that $u < v$, $\Phi_1 = \frac{1}{2} - \varphi_1$ and $\Phi_2 = \frac{1}{2} + \varphi_2$ with $0 < \varphi_2 < \varphi_1$. For a Pareto efficient calibration mechanism, the error of the conference and the adversary should stay on the Pareto frontier as in Figure 4.2. If the desired error of the conference is less than $\frac{u\Phi_1+v(1-\Phi_2)}{u+v}$, there is no feasible calibration mechanism that satisfies this error due to the noise in the scores given by the reviewers.

**Theorem 4.2.2.** *Algorithm 4 describes the optimal strategy for conference calibration in the noisy setting with per-instance error $\mathcal{E}_C([s_1, s_2])$.*

# Chapter 5

# Discussion

Our work is only a starting point towards addressing the important problem of calibration with privacy in its full generality. A number of challenges must be addressed in future work in order to design practical algorithms with guarantees for calibration using exogenous information in peer review.

One, relax certain assumptions made in this paper such as affine calibration in the noisy case, heterogeneity and knowledge of the noise variance, and the privacy criterion.

Two, consider a general number of reviewers and papers. This will require carefully defining the metric for the conference's utility as well as the adversary's goals.

Three, instead of assuming precise exogenous knowledge of the reviewers' miscalibration functions, consider having access to data from other conferences. One may assume that the entire data can be pooled, which may allow for using algorithms from past literature [2, 9, 12, 19, 24, 25, 26], but will still require ensuring privacy on top of it.

Four, our work considered ensuring privacy in the conference at hand. If using data from previous conferences, one would need to also ensure the privacy of data from those conferences.

# Chapter 6

# Appendix

In the appendix, we present complete proofs of the results claimed in the main text.

## 6.1 Proof of Lemma 3.0.6

Calibrating using the mechanism of calibration function $h$ differs from Calibrating using the mechanism of probability function $g$ only when the same paper has higher quality under both assignments by the MAP. Since other wise, by adjusting the output of $h(S, A_1)$ and $h(S, A_2)$, either paper can have arbitrary non-zero probability of being accepted (their probabilities sum to 1), and it is the same mechanism as using the probability function $g$.

Note that the adversary makes its guess using the MAP $\operatorname{argmax}_{A \in \{A_1, A_2\}} \Pr(\mathcal{A} = A | \boldsymbol{D} = P, S = [s_1, s_2])$ where $\boldsymbol{D}$ is the random variable for the decision made by the conference (acceptance of paper) and $P$ is the paper being accepted. By expanding the probability expression, we have that

$$\begin{aligned}
&\underset{A \in \{A_1, A_2\}}{\operatorname{argmax}} \Pr(\mathcal{A} = A | \boldsymbol{D} = P, S = [s_1, s_2]) \\
&= \underset{A \in \{A_1, A_2\}}{\operatorname{argmax}} \frac{\Pr(\boldsymbol{D} = P | \mathcal{A} = A, S = [s_1, s_2]) \Pr(\mathcal{A} = A | S = [s_1, s_2])}{\Pr(\boldsymbol{D} = P | S = [s_1, s_2])} \\
&= \underset{A \in \{A_1, A_2\}}{\operatorname{argmax}} \Pr(\boldsymbol{D} = P | \mathcal{A} = A, S = [s_1, s_2]) \Pr(\mathcal{A} = A | S = [s_1, s_2]).
\end{aligned}$$

If the same paper has higher-quality under both assignments, and the conference accepts the believed higher-quality paper, then the adversary guesses the assignment based on the scores only. Because the adversary knows the mechanism used by the conference, if $P$ is the paper that has higher-quality under both assignments, then $\Pr(\boldsymbol{D} = P | \mathcal{A} = A, S = [s_1, s_2]) = 1$ for both $\mathcal{A} = A_1$ and $\mathcal{A} = A_2$. Therefore, the conference does not have extra privacy leak by accepting $P$ since the adversary is making its guess based on the information that is already public (the scores). In addition, if the conference has non-zero probability of accepting the other paper, its utility decreases because it would have higher probability of accepting the lower-quality paper. However, the error of the adversary remains unchanged as it can use the scores to guess the assignment without being affected by the conference decision. Thus, there is no need for the conference to have non-zero probability for accepting the paper that has lower-quality under both assignments.

In conclusion, calibrating using the mechanism of calibration function $h$ instead of the mechanism of probability function $g$ does not reduce the optimally of the conference. Therefore, we consider the calibration mechanism with calibration function $h$ in our analysis.

## 6.2 Proof of Theorem 4.1.1

To find the Pareto frontier of per-instance error of the adversary against per-instance error of the conference, in the noiseless setting where the scores satisfy $\min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\} < s_1 < \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$, we first find the maximum per-instance error of the adversary given per-instance error of the conference in this range. We will show the proof with the assumptions that $\beta_2(\beta_1^{-1}(s_2)) > \beta_1(\beta_2^{-1}(s_2))$ and $f_1(s_1)f_2(s_2) > f_2(s_1)f_1(s_2)$. The analysis is of the same procedure for different assumptions on the values $\beta_2(\beta_1^{-1}(s_2))$, $\beta_1(\beta_2^{-1}(s_2))$, $f_1(s_1)f_2(s_2)$, and $f_2(s_1)f_1(s_2)$.

In the noiseless setting, the conference uses the reverse functions of miscalibration functions and the scores to exactly compute the quality of the papers. In the interesting region, the conference always accepts higher-quality paper if it calibrates under the correct assignment. And the conference always accepts lower-quality paper if it calibrates assuming the wrong assignment. We use $\mathcal{A}$ to denote the random variable for the assignment, $\boldsymbol{D}$ to denote the random variable for the conference decision and $S$ is the scores. In addition, we use $\boldsymbol{C}$ to denote the calibration status. If the conference calibrates under the correct assignment then $\boldsymbol{C} = T$. Otherwise, $\boldsymbol{C} = F$.

$$\begin{aligned}
&\Pr(\text{conference accepts lower-quality paper}|S = [s_1, s_2]) \\
=&\Pr(\boldsymbol{C} = F, \mathcal{A} = A_1|S = [s_1, s_2]) + \Pr(\boldsymbol{C} = F, \mathcal{A} = A_2|S = [s_1, s_2]) \\
=&\Pr(\boldsymbol{C} = F|\mathcal{A} = A_1, S = [s_1, s_2]) \Pr(\mathcal{A} = A_1|S = [s_1, s_2]) \\
&+ \Pr(\boldsymbol{C} = F|\mathcal{A} = A_2, S = [s_1, s_2])P(\mathcal{A} = A_2|S = [s_1, s_2]) \\
=&(1 - q_1) \cdot \frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2) + f_2(s_1)f_1(s_2)} + (1 - q_2) \cdot \frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2) + f_2(s_1)f_1(s_2)} \\
=&1 - \frac{f_1(s_1)f_2(s_2)q_1 + f_2(s_1)f_1(s_2)q_2}{f_1(s_1)f_2(s_2) + f_2(s_1)f_1(s_2)}
\end{aligned}$$

The adversary uses MAP to guess the assignment. If the two assignments have the same a posteriori probability, then the adversary makes a random guess between the assignments where either assignment has probability $\frac{1}{2}$ of being guessed. When making a guess, the adversary observes the scores and the conference decision. So the adversary finds $\operatorname{argmax}_{A \in \{A_1, A_2\}} \Pr(\mathcal{A} = A|\boldsymbol{D} = P, S = [s_1, s_2])$ where $P$ is the paper being accepted.

$$\begin{aligned}
&\operatorname*{argmax}_{A\in\{A_1,A_2\}} \Pr(\mathcal{A}=A|\boldsymbol{D}=P,S=[s_1,s_2])\\
&=\operatorname*{argmax}_{A\in\{A_1,A_2\}} \frac{\Pr(\boldsymbol{D}=P|\mathcal{A}=A,S=[s_1,s_2])\Pr(\mathcal{A}=A|S=[s_1,s_2])}{\Pr(\boldsymbol{D}=P|S=[s_1,s_2])}\\
&=\operatorname*{argmax}_{A\in\{A_1,A_2\}} \Pr(\boldsymbol{D}=P|\mathcal{A}=A,S=[s_1,s_2])\Pr(\mathcal{A}=A|S=[s_1,s_2])\\
&=\operatorname*{argmax}_{A\in\{A_1,A_2\}} (\Pr(\boldsymbol{D}=P|\mathcal{A}=A,S=[s_1,s_2],\boldsymbol{C}=T)\Pr(\boldsymbol{C}=T|\mathcal{A}=A,S=[s_1,s_2])\\
&\quad+\Pr(\boldsymbol{D}=P|\mathcal{A}=A,S,\boldsymbol{C}=F)\Pr(\boldsymbol{C}=F|\mathcal{A}=A,S=[s_1,s_2]))\cdot\Pr(\mathcal{A}=A|S=[s_1,s_2])\\
&=\operatorname*{argmax}_{A\in\{A_1,A_2\}} (\Pr(\boldsymbol{D}=P|\mathcal{A}=A,S=[s_1,s_2],\boldsymbol{C}=T)\Pr(h(\mathcal{A}=A,S=[s_1,s_2])=1)\\
&\quad+\Pr(\boldsymbol{D}=P|\mathcal{A}=A,S,\boldsymbol{C}=F)\Pr(h(\mathcal{A}=A,S=[s_1,s_2])=0))\cdot\Pr(\mathcal{A}=A|S=[s_1,s_2])
\end{aligned}$$

Since both the conference and the adversary know the miscalibration functions and there is no noise, the quality of papers can be computed exactly using the inverse functions of reviewer functions. Therefore, $\Pr(\boldsymbol{D}=P|\mathcal{A}=A,S=[s_1,s_2])$ is either 0 or 1. Note that if paper 1 has higher quality under both assignments $A_1$ and $A_2$, then the adversary guesses the assignment based on the scores only because in this case $\operatorname{argmax}_{A\in\{A_1,A_2\}}\Pr(\boldsymbol{D}=P|\mathcal{A}=A,S=[s_1,s_2])\Pr(\mathcal{A}=A|S=[s_1,s_2])=\operatorname{argmax}_{A\in\{A_1,A_2\}}\Pr(\mathcal{A}=A|S=[s_1,s_2])$. So there is no need for the conference to accept paper 2 in this case. Same argument applies when paper 2 has higher quality under both assignments $A_1$ and $A_2$, then the conference should accept paper 2.

We then look into the region of the scores where $\min\{\beta_2(\beta_1^{-1}(s_2)),\beta_1(\beta_2^{-1}(s_2))\}<s_1<\max\{\beta_2(\beta_1^{-1}(s_2)),\beta_1(\beta_2^{-1}(s_2))\}$. If paper 1 is accepted, then if $f_1(s_1)f_2(s_2)q_1>f_2(s_1)f_1(s_2)(1-q_2)$, the adversary guesses $\mathcal{A}=A_1$. If $f_1(s_1)f_2(s_2)q_1<f_2(s_1)f_1(s_2)(1-q_2)$, the adversary guesses $\mathcal{A}=A_2$. If $f_1(s_1)f_2(s_2)q_1=f_2(s_1)f_1(s_2)(1-q_2)$, the adversary guesses makes a guess of the assignment with probability $\frac{1}{2}$ for either assignment. Similarly, if paper 2 is accepted, the adversary compares $f_1(s_1)f_2(s_2)(1-q_1)$ and $f_2(s_1)f_1(s_2)q_2$. There are 2 papers and 2 possible assignments, so we have 4 scenarios combining decisions and assignments.

1. If $\mathcal{A}=A_1$ and $\boldsymbol{D}=P_1$, then the adversary guesses wrong if $f_1(s_1)f_2(s_2)q_1<f_2(s_1)f_1(s_2)(1-q_2)$. This scenario happens with probability $\Pr(\mathcal{A}=A_1,\boldsymbol{D}=P_1|S=[s_1,s_2])=\frac{f_1(s_1)f_2(s_2)q_1}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
2. If $\mathcal{A}=A_1$ and $\boldsymbol{D}=P_2$, then the adversary guesses wrong if $f_1(s_1)f_2(s_2)(1-q_1)<f_2(s_1)f_1(s_2)q_2$. This scenario happens with probability $\Pr(\mathcal{A}=A_1,\boldsymbol{D}=P_2|S=[s_1,s_2])=\frac{f_1(s_1)f_2(s_2)(1-q_1)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
3. If $\mathcal{A}=A_2$ and $\boldsymbol{D}=P_1$, then the adversary guesses wrong if $f_1(s_1)f_2(s_2)q_1>f_2(s_1)f_1(s_2)(1-q_2)$. This scenario happens with probability $\Pr(\mathcal{A}=A_1,\boldsymbol{D}=P_1|S=[s_1,s_2])=\frac{f_2(s_1)f_1(s_2)(1-q_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
4. If $\mathcal{A}=A_2$ and $\boldsymbol{D}=P_2$, then the adversary guesses wrong if $f_1(s_1)f_2(s_2)(1-q_1)>f_2(s_1)f_1(s_2)q_2$. This scenario happens with probability $\Pr(\mathcal{A}=A_1,\boldsymbol{D}=P_2|S=[s_1,s_2])=\frac{f_2(s_1)f_1(s_2)q_2}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.

To compute the error of the adversary, we need to compare $f_1(s_1)f_2(s_2)$ and $f_2(s_1)f_1(s_2)$. So we suppose $f_1(s_1)f_2(s_2) > f_2(s_1)f_1(s_2)$. Then we consider 5 cases of the value $f_1(s_1)f_2(s_2)q_1$ that result in different error of the adversary. We refer to the 4 scenarios of $(\mathcal{A}, \boldsymbol{D})$ above.

- If $f_1(s_1)f_2(s_2)q_1 < f_2(s_1)f_1(s_2) - f_2(s_1)f_1(s_2)q_2$, the adversary guesses wrong in scenarios 1 and 4.

  Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{f_1(s_1)f_2(s_2)q_1 + f_2(s_1)f_1(s_2)q_2}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, which is the opposite of the error of the conference $\mathcal{E}_C([s_1, s_2])$. Error of the adversary ranges from 0 to $\frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. The relation between error of the adversary and error of the conference is $\mathcal{E}_A([s_1, s_2]) = 1 - \mathcal{E}_C([s_1, s_2])$ for $\mathcal{E}_C([s_1, s_2]) \in (\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}, 1]$.

- If $f_1(s_1)f_2(s_2)q_1 = f_2(s_1)f_1(s_2) - f_2(s_1)f_1(s_2)q_2$, the adversary makes random guess in scenarios 1 and 3 and guesses wrong in scenario 4.

  Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ and error of the conference $\mathcal{E}_C([s_1, s_2])$ is $\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.

- If $f_2(s_1)f_1(s_2) - f_2(s_1)f_1(s_2)q_2 < f_1(s_1)f_2(s_2)q_1 < f_1(s_1)f_2(s_2) - f_2(s_1)f_1(s_2)q_2$, the adversary guesses wrong in scenarios 3 and 4.

  Error of the the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, which is constant. The relation between error of the adversary and error of the conference is $\mathcal{E}_A([s_1, s_2]) = \frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. Error of the adversary stays constant for error of the conference $\mathcal{E}_C([s_1, s_2]) \in (\frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}, \frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)})$.

- If $f_1(s_1)f_2(s_2)q_1 = f_1(s_1)f_2(s_2) - f_2(s_1)f_1(s_2)q_2$, the adversary makes random guess in scenarios 2 and 4 and guesses wrong in scenario 3.

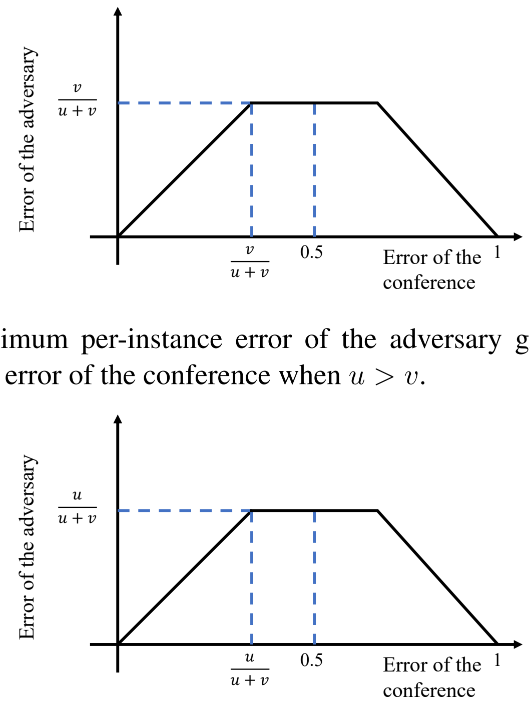
  Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ and error of the conference $\mathcal{E}_C([s_1, s_2])$ is also $\frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.

- If $f_1(s_1)f_2(s_2)q_1 > f_1(s_1)f_2(s_2) - f_2(s_1)f_1(s_2)q_2$, the adversary guesses wrong in scenarios 2 and 3.

  Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $1 - \frac{f_1(s_1)f_2(s_2)q_1 + f_2(s_1)f_1(s_2)q_2}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, which is the same as the error of the conference $\mathcal{E}_C([s_1, s_2])$. Error of the adversary $\mathcal{E}_A([s_1, s_2])$ ranges from 0 to $\frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. The relation between error of the adversary and error of the conference is $\mathcal{E}_A([s_1, s_2]) = \mathcal{E}_C([s_1, s_2])$ for $\mathcal{E}_C([s_1, s_2]) \in [0, \frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)})$.

Therefore, the relation between error of the adversary and error of the conference when $f_1(s_1)f_2(s_2) > f_2(s_1)f_1(s_2)$ is of the shape of a trapezoid in $[0, 1]$ with the three line segments of the slope +1, 0, and -1 as in Figure 6.1a. Note that the relation between the per-instance errors does not change with the relation between values of $f_1(s_1)f_2(s_2)$ and $f_2(s_1)f_1(s_2)$. So Figure 6.1a is the relation between the errors when $u > v$. Similarly, Figure 6.1b is the relation between the errors when $u \leq v$.

From Figure 6.1 we see that the conference should keep its per-instance error less than $\frac{\min(u,v)}{u+v}$ to stay optimal. Because if error of the conference is greater than $\frac{\min(u,v)}{u+v}$, increasing its error does not increase error the adversary and thus is not optimal. Thus, the Pareto frontier of per-instance error of the adversary against error of the conference is the first line segment



(a) Maximum per-instance error of the adversary given per-instance error of the conference when $u > v$.

(b) Maximum per-instance error of the adversary given per-instance error of the conference when $u \leq v$

Figure 6.1: Relation between error of the adversary and error of the conference with $u = f_1(s_1)f_2(s_2)$ and $v = f_2(s_1)f_1(s_2)$.

with slope 1 in both Figure 6.1a and Figure 6.1b when $\min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\} < s_1 < \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$.

## 6.3 Proof of Theorem 4.1.2

We prove that Algorithm 1 is optimal for each instance of scores $S = [s_1, s_2]$ with desired error of the conference $\mathcal{E}_C([s_1, s_2])$ in the noiseless setting.

From Lemma 3.0.6 we know that if a paper has higher quality under both assignments, the conference should accept the paper. This is the optimal strategy for the conference.

Otherwise when $\min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\} < s_1 < \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$, we use the Pareto frontiers analyze the optimality of our algorithm. Theorem 4.1.1 shows that the Pareto frontier in the noiseless setting within this region. Suppose $f_1(s_1)f_2(s_2) \leq f_2(s_1)f_1(s_2)$, then the endpoint on the Pareto frontier has both error of the conference and error of the adversary being $\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$. If $\mathcal{E}_C([s_1, s_2]) < \frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, we maximized the error of the adversary by operating on the Pareto frontier. If $\mathcal{E}_C([s_1, s_2]) \geq \frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, we operate at the endpoint where error of the adversary is maximum and error of the conference is no larger than the desired $\mathcal{E}_C([s_1, s_2])$. The endpoint is the point with minimum error of the conference such that error of the adversary is maximum. Therefore, it is optimal for the conference.

Similarly, if $f_1(s_1)f_2(s_2) > f_2(s_1)f_1(s_2)$, the algorithm is also optimal by maximizing error of the adversary under desired error of the conference following the Pareto frontier. Algorithm 1 follows the procedure by choosing the corresponding $q_1$ and $q_2$ for each point on the Pareto frontier and thus is optimal for the conference.

## 6.4 Proof of Theorem 4.1.3

Algorithm 2 is an algorithm that operates on the endpoint of the Pareto frontier when teh scores satisfy $\min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\} < s_1 < \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$. We use $\zeta$ to denote the error of adopting Algorithm 2 across all scores. Then we have Algorithm 3 that has a desired overall error of the conference $\mathcal{E}_C$ as input.

If $\mathcal{E}_C \geq \zeta$, we operate at $\mathcal{E}_C = \zeta$ by adopting Algorithm 2. error of the adversary is maximized because when $\min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\} < s_1 < \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$, error of the adversary is maximized by operating at the first changing point. Any other point on the Pareto frontier has the error of the adversary less than or equal the error at the first changing point. Outside the region, error of the adversary is fixed because the adversary's guess is based on the scores only. The conference has zero error outside the region so it is optimal.

If $\mathcal{E}_C < \zeta$, when $\min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\} < s_1 < \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$, error of the adversary is the same as error of the conference. Note that the slope of the Pareto frontier is 1 when $\min\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\} < s_1 < \max\{\beta_2(\beta_1^{-1}(s_2)), \beta_1(\beta_2^{-1}(s_2))\}$, therefore error of the adversary cannot exceed error of the conference with the maximum being equal to error of the conference. So the algorithm is optimal within this region. Otherwise, error of the adversary is fixed and is optimal for the conference.

## 6.5 Proof of Theorem 4.2.1

To find the Pareto frontier of per-instance error of the adversary against error of the conference in the noisy setting where the scores satisfy $\min\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\} \leq s_1 \leq \max\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$, we first find the maximum per-instance error of the adversary given per-instance error of the conference in this range. We will show the proof with the assumptions that $f_1(s_1)f_2(s_2) < f_2(s_1)f_1(s_2)$ and $\Phi_1 = \frac{1}{2} - \varphi_1$ and $\Phi_2 = \frac{1}{2} + \varphi_2$ with $0 < \varphi_2 < \varphi_1$. The analysis is of the same procedure for different assumptions on the values of $f_1(s_1)f_2(s_2)$, $f_2(s_1)f_1(s_2)$, $\Phi_1$ and $\Phi_2$ with $\Phi_1 - \frac{1}{2}$ and $\Phi_2 - \frac{1}{2}$ having different signs. The notations are of the same meaning as in Section 6.3. In the noisy setting, even if the conference calibrates under the true assignment, there is still possibility to accept the lower-quality paper due to the noise in the scores given by the reviewers. Note that with the assumptions and when $\min\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\} \leq s_1 \leq \max\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$, the conference accepts paper 1 if calibrates under $A_1$ and accepts paper 2 if calibrates under $A_2$. So we have

$$\begin{aligned}
&\Pr(\text{conference accepts lower-quality paper}|S=[s_1,s_2])\\
=&\Pr(\text{conference accepts } P_1,\theta_1<\theta_2|S=[s_1,s_2])+\Pr(\text{conference accepts } P_2,\theta_1>\theta_2|S=[s_1,s_2])\\
=&\Pr(\text{conference accepts } P_1|\theta_1<\theta_2,S=[s_1,s_2])\cdot\Pr(\theta_1<\theta_2|S=[s_1,s_2])\\
&+\Pr(\text{conference accepts } P_2|\theta_1>\theta_2,S=[s_1,s_2])\cdot\Pr(\theta_1>\theta_2|S=[s_1,s_2]).
\end{aligned}$$

We then expand each of the two terms.

$$\begin{aligned}
&\Pr(\text{conference accepts } P_1|\theta_1<\theta_2,S=[s_1,s_2])\\
=&\Pr(\text{conference accepts } P_1,\mathcal{A}=A_1|\theta_1<\theta_2,S=[s_1,s_2])\\
&+\Pr(\text{conference accepts } P_1,\mathcal{A}=A_2|\theta_1<\theta_2,S=[s_1,s_2])\\
=&\Pr(\text{conference accepts } P_1|\mathcal{A}=A_1,\theta_1<\theta_2,S=[s_1,s_2])\cdot P(\mathcal{A}=A_1|\theta_1<\theta_2,S=[s_1,s_2])\\
&+\Pr(\text{conference accepts } P_1|\mathcal{A}=A_2,\theta_1<\theta_2,S=[s_1,s_2])\cdot\Pr(\mathcal{A}=A_2|\theta_1<\theta_2,S=[s_1,s_2])\\
=&\Pr(\boldsymbol{C}=T|\mathcal{A}=A_1,\theta_1<\theta_2,S=[s_1,s_2])\Pr(\mathcal{A}=A_1|\theta_1<\theta_2,S=[s_1,s_2])\\
&+\Pr(\boldsymbol{C}=F|\mathcal{A}=A_2,\theta_1<\theta_2,S=[s_1,s_2])\Pr(\mathcal{A}=A_2|\theta_1<\theta_2,S=[s_1,s_2])\\
=&q_1\Pr(\mathcal{A}=A_1|\theta_1<\theta_2,S=[s_1,s_2])+(1-q_2)\Pr(\mathcal{A}=A_2|\theta_1<\theta_2,S=[s_1,s_2])\\
=&q_1\frac{\Pr(\theta_1<\theta_2|\mathcal{A}=A_1,S=[s_1,s_2])\cdot\Pr(\mathcal{A}=A_1|S=[s_1,s_2])}{\Pr(\theta_1<\theta_2|S=[s_1,s_2])}\\
&+(1-q_2)\frac{\Pr(\theta_1<\theta_2|\mathcal{A}=A_2,S=[s_1,s_2])\cdot\Pr(\mathcal{A}=A_2|S=[s_1,s_2])}{\Pr(\theta_1<\theta_2|S=[s_1,s_2])}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\Pr(\text{conference accepts } P_2|\theta_1>\theta_2,S=[s_1,s_2])\\
=&(1-q_1)\frac{\Pr(\theta_1>\theta_2|\mathcal{A}=A_1,S=[s_1,s_2])\cdot\Pr(\mathcal{A}=A_1|S=[s_1,s_2])}{\Pr(\theta_1>\theta_2|S=[s_1,s_2])}\\
&+q_2\frac{\Pr(\theta_1>\theta_2|\mathcal{A}=A_2,S=[s_1,s_2])\cdot\Pr(\mathcal{A}=A_2|S=[s_1,s_2])}{\Pr(\theta_1>\theta_2|S=[s_1,s_2])}
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\Pr(\text{conference accepts lower-quality paper}|S=[s_1,s_2])\\
=&q_1\Pr(\theta_1<\theta_2|\mathcal{A}=A_1,S=[s_1,s_2])\cdot\Pr(\mathcal{A}=A_1|S=[s_1,s_2])\\
&+(1-q_2)\Pr(\theta_1<\theta_2|\mathcal{A}=A_2,S=[s_1,s_2])\cdot\Pr(\mathcal{A}=A_2|S=[s_1,s_2])\\
&+(1-q_1)\Pr(\theta_1>\theta_2|\mathcal{A}=A_1,S=[s_1,s_2])\cdot\Pr(\mathcal{A}=A_1|S=[s_1,s_2])\\
&+q_2\Pr(\theta_1>\theta_2|\mathcal{A}=A_2,S=[s_1,s_2])\cdot\Pr(\mathcal{A}=A_2|S=[s_1,s_2])\\
=&\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}\cdot(q_1\Phi_1+(1-q_1)(1-\Phi_1)\\
&+\frac{f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}\cdot((1-q_2)\Phi_2+q_2(1-\Phi_2)
\end{aligned}$$

Under the assumptions that $\Phi_1 = \frac{1}{2} - \varphi_1$ and $\Phi_2 = \frac{1}{2} + \varphi_2$ where $0 < \varphi_2 < \varphi_1$ and $f_1(s_1)f_2(s_2) < f_2(s_1)f_1(s_2)$, we analyze the per-instance error of the adversary similar to the procedure in Section 6.2. There are 4 scenarios combining the decision and the true assignment.

1. If $\mathcal{A} = A_1$ and $\boldsymbol{D} = P_1$, then the adversary guesses wrong if $q_1 f_1(s_1)f_2(s_2) < (1 - q_2)f_2(s_1)f_1(s_2)$. This scenario happens with probability $\Pr(\mathcal{A} = A_1, \boldsymbol{D} = P_1 | S = [s_1, s_2]) = \frac{f_1(s_1)f_2(s_2)q_1}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
2. If $\mathcal{A} = A_1$ and $\boldsymbol{D} = P_2$, then the adversary guesses wrong if $(1 - q_1)f_1(s_1)f_2(s_2) < q_2 f_2(s_1)f_1(s_2)$. This scenario happens with probability $\Pr(\mathcal{A} = A_1, \boldsymbol{D} = P_1 | S = [s_1, s_2]) = \frac{f_1(s_1)f_2(s_2)(1-q_1)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
3. If $\mathcal{A} = A_2$ and $\boldsymbol{D} = P_1$, then the adversary guesses wrong if $q_1 f_1(s_1)f_2(s_2) > (1 - q_2)f_2(s_1)f_1(s_2)$. This scenario happens with probability $\Pr(\mathcal{A} = A_1, \boldsymbol{D} = P_1 | S = [s_1, s_2]) = \frac{f_2(s_1)f_1(s_2)(1-q_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
4. If $\mathcal{A} = A_2$ and $\boldsymbol{D} = P_2$, then the adversary guesses wrong if $(1 - q_1)f_1(s_1)f_2(s_2) > q_2 f_2(s_1)f_1(s_2)$. This scenario happens with probability $\Pr(\mathcal{A} = A_1, \boldsymbol{D} = P_1 | S = [s_1, s_2]) = \frac{f_2(s_1)f_1(s_2)q_2}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.

To compute the error of the adversary, we need to compare $f_1(s_1)f_2(s_2)$ and $f_2(s_1)f_1(s_2)$. So we suppose $f_1(s_1)f_2(s_2) < f_2(s_1)f_1(s_2)$. Then we consider 5 cases of the value $f_1(s_1)f_2(s_2)q_1$ that result in different error of the adversary. We refer to the 4 scenarios of $(\mathcal{A}, \boldsymbol{D})$ above.

- If $q_1 f_1(s_1)f_2(s_2) < f_1(s_1)f_2(s_2) - q_2 f_2(s_1)f_1(s_2)$, the adversary guesses wrong in scenarios 1 and 4. Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{q_1 f_1(s_1)f_2(s_2)+q_2 f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
- If $q_1 f_1(s_1)f_2(s_2) = f_1(s_1)f_2(s_2) - q_2 f_2(s_1)f_1(s_2)$, the adversary makes random guess in scenarios 2 and 4 and guesses wrong in scenario 1. Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{q_1 f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)} + \frac{1}{2}\left(\frac{(1-q_1)f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)} + \frac{q_2 f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}\right) = \frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
- If $f_1(s_1)f_2(s_2) - q_2 f_2(s_1)f_1(s_2) < q_1 f_1(s_1)f_2(s_2) < f_2(s_1)f_1(s_2) - q_2 f_2(s_1)f_1(s_2)$, the adversary guesses wrong in scenarios 1 and 2. Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
- If $q_1 f_1(s_1)f_2(s_2) = f_2(s_1)f_1(s_2) - q_2 f_2(s_1)f_1(s_2)$, the adversary makes random guess in scenarios 1 and 3 and guesses wrong in scenario 2. Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{(1-q_1)f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)} + \frac{1}{2}\left(\frac{q_1 f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)} + \frac{(1-q_2)f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}\right) = \frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
- If $q_1 f_1(s_1)f_2(s_2) > f_2(s_1)f_1(s_2) - q_2 f_2(s_1)f_1(s_2)$, the adversary guesses wrong in scenarios 2 and 3. Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $1 - \frac{q_1 f_1(s_1)f_2(s_2)+q_2 f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.

To find the maximum error of the adversary given error of the conference, we solve an optimization problem. In order to formulate the optimization problem, we can combine the 5 cases above into 3 cases for simplicity.

- If $q_1 f_1(s_1)f_2(s_2) \leq f_1(s_1)f_2(s_2) - q_2 f_2(s_1)f_1(s_2)$, error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{q_1 f_1(s_1)f_2(s_2)+q_2 f_2(s_1)f_1(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.
- If $f_1(s_1)f_2(s_2) - q_2 f_2(s_1)f_1(s_2) \leq q_1 f_1(s_1)f_2(s_2) \leq f_2(s_1)f_1(s_2) - q_2 f_2(s_1)f_1(s_2)$, error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.

- If $q_1 f_1(s_1) f_2(s_2) \geq f_2(s_1) f_1(s_2) - q_2 f_2(s_1) f_1(s_2)$, error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $1 - \frac{q_1 f_1(s_1) f_2(s_2) + q_2 f_2(s_1) f_1(s_2)}{f_1(s_1) f_2(s_2) + f_2(s_1) f_1(s_2)}$.

We let $T(\mathcal{E}_C) = \mathcal{E}_C(u + v) - u \cdot (1 - \Phi_1) - v \cdot \Phi_2$ to be a function that takes the error of the conference as input.

- Maximize $\frac{q_1 f_1(s_1) f_2(s_2) + q_2 f_2(s_1) f_1(s_2)}{f_1(s_1) f_2(s_2) + f_2(s_1) f_1(s_2)}$ subject to $\mathcal{E}_C([s_1, s_2])(f_1(s_1) f_2(s_2) + f_2(s_1) f_1(s_2)) - f_1(s_1) f_2(s_2) \cdot (1 - \Phi_1) - f_2(s_1) f_1(s_2) \cdot \Phi_2 = f_1(s_1) f_2(s_2)(2\Phi_1 - 1) q_1 + f_2(s_1) f_1(s_2) \cdot (1 - 2\Phi_2) q_2$ and $q_1 f_1(s_1) f_2(s_2) \leq f_1(s_1) f_2(s_2) - q_2 f_2(s_1) f_1(s_2)$.

  The maximum occurs at $q_1 f_1(s_1) f_2(s_2) = f_1(s_1) f_2(s_2) - q_2 f_2(s_1) f_1(s_2)$. Then the intersection of the two lines is $q_1 = 1 - \frac{(2\Phi_1 - 1)u - T(\mathcal{E}_C([s_1, s_2]))}{(2\Phi_1 + 2\Phi_2 - 2)u}$ and $q_2 = \frac{(2\Phi_1 - 1)u - T(\mathcal{E}_C([s_1, s_2]))}{(2\Phi_1 + 2\Phi_2 - 2)v}$.

  - If the intersection point can be reached, $q_1, q_2 \in [0, 1]$, $(2\Phi_1 - 1)u \leq T(\mathcal{E}_C([s_1, s_2])) \leq (1 - 2\Phi_2)u$, then error of the conference $\mathcal{E}_C([s_1, s_2])$ ranges from $\frac{f_1(s_1) f_2(s_2) \Phi_1}{f_1(s_1) f_2(s_2) + f_2(s_1) f_1(s_2)} + \frac{f_2(s_1) f_1(s_2) \Phi_2}{f_1(s_1) f_2(s_2) + f_2(s_1) f_1(s_2)}$ to $\frac{f_1(s_1) f_2(s_2)(2 - \Phi_1 - 2\Phi_2)}{f_1(s_1) f_2(s_2) + f_2(s_1) f_1(s_2)} + \frac{f_2(s_1) f_1(s_2) \Phi_2}{f_1(s_1) f_2(s_2) + f_2(s_1) f_1(s_2)}$.

    Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{f_1(s_1) f_2(s_2)}{f_1(s_1) f_2(s_2) + f_2(s_1) f_1(s_2)}$.
  - If the intersection point can not be reached and $T(\mathcal{E}_C([s_1, s_2])) < (2\Phi_1 - 1)u$, then no $q_1, q_2$ are qualified for the constraints.
  - If the intersection point can not be reached and $T(\mathcal{E}_C([s_1, s_2])) > (1 - 2\Phi_2)u$.
    - If $(1 - 2\Phi_2)u < T(\mathcal{E}_C([s_1, s_2])) \leq 0$ then the maximum is reached when $q_1 = 0$ and $q_2 = \frac{T(\mathcal{E}_C([s_1, s_2]))}{(1 - 2\Phi_2)v}$.

      Error of the conference $\mathcal{E}_C([s_1, s_2])$ ranges from $\frac{(2 - \Phi_1 - 2\Phi_2)u + \Phi_2 v}{u + v}$ (when $T(\mathcal{E}_C([s_1, s_2])) = (1 - 2\Phi_2)u$) to $\frac{(1 - \Phi_1)u + \Phi_2 v}{u + v}$ (when $T(\mathcal{E}_C([s_1, s_2])) = 0$).

      Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{T(\mathcal{E}_C([s_1, s_2]))}{(1 - 2\Phi_2)(u + v)}$, ranges from $\frac{u}{u + v}$ (when $T(\mathcal{E}_C([s_1, s_2])) = (1 - 2\Phi_2)u$) to 0 (when $T(\mathcal{E}_C([s_1, s_2])) = 0$).
    - If $T(\mathcal{E}_C([s_1, s_2])) > 0$ then no $q_1, q_2$ are qualified for the constraints.

- Error of the adversary $\mathcal{E}_A([s_1, s_2])$ is $\frac{f_1(s_1) f_2(s_2)}{f_1(s_1) f_2(s_2) + f_2(s_1) f_1(s_2)}$ subject to $f_1(s_1) f_2(s_2) - q_2 f_2(s_1) f_1(s_2) \leq q_1 f_1(s_1) f_2(s_2) \leq f_2(s_1) f_1(s_2) - q_2 f_2(s_1) f_1(s_2)$.

  From Figure 6.2 we can see that error of the conference $\mathcal{E}_C([s_1, s_2])$ has its extremes at $q_1 = 0, q_2 = \frac{u}{v}$ and $q_1 = 1, q_2 = 1 - \frac{u}{v}$. Therefore, error of the conference ranges from $\frac{(2 - \Phi_1 - 2\Phi_2)u + \Phi_2 v}{u + v}$ to $\frac{(\Phi_1 + 2\Phi_2 - 1)u + (1 - \Phi_2)v}{u + v}$.

- Maximize $1 - \frac{q_1 f_1(s_1) f_2(s_2) + q_2 f_2(s_1) f_1(s_2)}{f_1(s_1) f_2(s_2) + f_2(s_1) f_1(s_2)}$ subject to $\mathcal{E}_C([s_1, s_2])(f_1(s_1) f_2(s_2) + f_2(s_1) f_1(s_2)) - f_1(s_1) f_2(s_2) \cdot (1 - \Phi_1) - f_2(s_1) f_1(s_2) \cdot \Phi_2 = f_1(s_1) f_2(s_2)(2\Phi_1 - 1) q_1 + f_2(s_1) f_1(s_2) \cdot (1 - 2\Phi_2) q_2$ and $q_1 f_1(s_1) f_2(s_2) \geq f_2(s_1) f_1(s_2) - q_2 f_2(s_1) f_1(s_2)$.

  The maximum occurs at $q_1 f_1(s_1) f_2(s_2) = f_2(s_1) f_1(s_2) - q_2 f_2(s_1) f_1(s_2)$. Then the intersection of the two lines is $q_1 = \frac{(1 - 2\Phi_2)v - T(\mathcal{E}_C([s_1, s_2]))}{(2 - 2\Phi_1 - 2\Phi_2)u}$ and $q_2 = \frac{T(\mathcal{E}_C([s_1, s_2])) - (2\Phi_1 - 1)v}{(2 - 2\Phi_1 - 2\Phi_2)v}$.

  - If the intersection point can be reached, $q_1, q_2 \in [0, 1]$, $(1 - 2\Phi_2)v - (2 - 2\Phi_1 - 2\Phi_2)u \leq T(\mathcal{E}_C([s_1, s_2])) \leq (1 - 2\Phi_2)v$, then error of the conference $\mathcal{E}_C([s_1, s_2])$ ranges from $\frac{f_1(s_1) f_2(s_2)(1 - \Phi_1)}{f_1(s_1) f_2(s_2) + f_2(s_1) f_1(s_2)} + \frac{f_2(s_1) f_1(s_2)(1 - \Phi_2)}{f_1(s_1) f_2(s_2) + f_2(s_1) f_1(s_2)}$ (when $T(\mathcal{E}_C([s_1, s_2])) =$

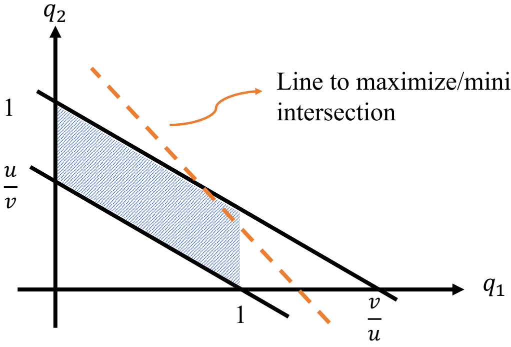


Figure 6.2: A diagram illustrates the optimization problem in this case.

$(1-2\Phi_2)v$) to $\frac{f_1(s_1)f_2(s_2)(\Phi_1+2\Phi_2-1)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)} + \frac{f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ (when $T(\mathcal{E}_C([s_1,s_2])) = (1-2\Phi_2)v - (2-2\Phi_1-2\Phi_2)u$).

Error of the adversary $\mathcal{E}_A([s_1,s_2])$ is $\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$.

- If the intersection point can not be reached and $T(\mathcal{E}_C([s_1,s_2])) > (1-2\Phi_2)v$, then no $q_1, q_2$ are qualified for the constraints.
- If the intersection point can not be reached and $T(\mathcal{E}_C([s_1,s_2])) < (1-2\Phi_2)v - (2-2\Phi_1-2\Phi_2)u$.
  - If $(2\Phi_1-1)u + (1-2\Phi_2)v \leq T(\mathcal{E}_C([s_1,s_2])) < (1-2\Phi_2)v - (2-2\Phi_1-2\Phi_2)u$ then the maximum is reached when $q_1 = 1$ and $q_2 = \frac{T(\mathcal{E}_C([s_1,s_2]))-(2\Phi_1-1)u}{(1-2\Phi_2)v}$.

    Error of the conference $\mathcal{E}_C([s_1,s_2])$ ranges from $\frac{(\Phi_1+2\Phi_2-1)u+(1-\Phi_2)v}{u+v}$ (when $T(\mathcal{E}_C([s_1,s_2])) = (1-2\Phi_2)v - (2-2\Phi_1-2\Phi_2)u$) to $\frac{\Phi_1 u+(1-\Phi_2)v}{u+v}$ (when $T(\mathcal{E}_C([s_1,s_2])) = (2\Phi_1-1)u + (1-2\Phi_2)v$).

    Error of the adversary $\mathcal{E}_A([s_1,s_2])$ is $1 - \frac{T(\mathcal{E}_C([s_1,s_2]))+(2-2\Phi_1-2\Phi_2)u}{(1-2\Phi_2)(u+v)}$, ranges from $\frac{u}{u+v}$ (when $T(\mathcal{E}_C([s_1,s_2])) = (1-2\Phi_2)v - (2-2\Phi_1-2\Phi_2)u$) to 0 (when $T(\mathcal{E}_C([s_1,s_2])) = (2\Phi_1-1)u + (1-2\Phi_2)v$).
  - If $T(\mathcal{E}_C([s_1,s_2])) < (2\Phi_1-1)u + (1-2\Phi_2)v$ then no $q_1, q_2$ are qualified for the constraints.

Therefore, the relation between error of the adversary and error of the conference when $\Phi_1 = \frac{1}{2} - \varphi_1$ and $\Phi_2 = \frac{1}{2} + \varphi_2$ where $0 < \varphi_2 < \varphi_1$ and $f_1(s_1)f_2(s_2) < f_2(s_1)f_1(s_2)$ is of the shape of a trapezoid in $[0,1]$ as in Figure 6.3. Note that the relation between the per-instance errors does not change with the relation between values of $f_1(s_1)f_2(s_2)$ and $f_2(s_1)f_1(s_2)$ or with the values of $\Phi_1$ and $\Phi_2$.

From Figure 6.3 we see that the conference should keep its per-instance error between $\frac{u\Phi_1+v(1-\Phi_2)}{u+v}$ and $\frac{u(\Phi_1+2\Phi_2-1)+v(1-\Phi_2)}{u+v}$ to stay optimal. The conference cannot have its error less than $\frac{u\Phi_1+v(1-\Phi_2)}{u+v}$ due to the reviewers' noise. If error of the conference is greater than $\frac{u(\Phi_1+2\Phi_2-1)+v(1-\Phi_2)}{u+v}$, increasing its error does not increase error the adversary and thus is not optimal. Thus, the Pareto

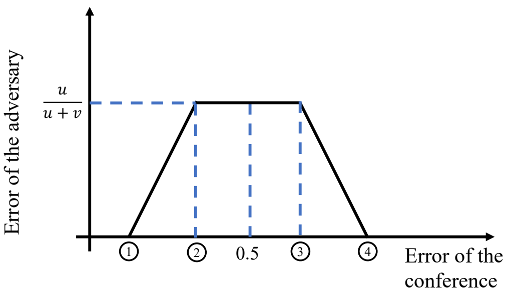


Figure 6.3: Maximum per-instance error of the adversary given per-instance error of the conference when $u < v$, $\Phi_1 = \frac{1}{2} - \varphi_1$ and $\Phi_2 = \frac{1}{2} + \varphi_2$ with $0 < \varphi_2 < \varphi_1$. The coordinates in the plot are: ① = $\frac{u\Phi_1+v(1-\Phi_2)}{u+v}$, ② = $\frac{u(\Phi_1+2\Phi_2-1)+v(1-\Phi_2)}{u+v}$, ③ = $\frac{u(2-\Phi_1-2\Phi_2)+v\Phi_2}{u+v}$, ④ = $\frac{u(1-\Phi_1)+v\Phi_2}{u+v}$.

frontier of per-instance error of the adversary against error of the conference is the first line segment with positive slope in Figure 6.3 when $\min\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\} \leq s_1 \leq \max\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$.

## 6.6 Proof of Theorem 4.2.2

We prove that Algorithm 4 is optimal for each instance of scores $S = [s_1, s_2]$ with desired error of the conference $\mathcal{E}_C([s_1, s_2])$ in the noisy setting. We carry the assumptions from Section 6.5 that $\Phi_1 = \frac{1}{2} - \varphi_1$ and $\Phi_2 = \frac{1}{2} + \varphi_2$ where $0 < \varphi_2 < \varphi_1$ and $f_1(s_1)f_2(s_2) < f_2(s_1)f_1(s_2)$.

From Lemma 3.0.6 we know that if a paper has higher quality under both assignments, the conference should accept the paper. This is the optimal strategy for the conference.

Otherwise when the scores are in the region $\min\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\} \leq s_1 \leq \max\left\{\frac{a_2(a_1^2+\sigma^2)(s_2-b_2)}{a_1(a_2^2+\sigma^2)} + b_1, \frac{a_1(a_2^2+\sigma^2)(s_2-b_1)}{a_2(a_1^2+\sigma^2)} + b_2\right\}$, we use the Pareto frontiers analyze the optimality of our algorithm. Theorem 4.2.1 shows that the Pareto frontier in the noiseless setting within this region. The analysis is similar to the one in the noiseless setting in Section 6.3.

Suppose $f_1(s_1)f_2(s_2) < f_2(s_1)f_1(s_2)$, then the endpoint on the Pareto frontier has error of the adversary being $\frac{f_1(s_1)f_2(s_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$ and error of the conference being $\frac{f_1(s_1)f_2(\Phi_1+2\Phi_2-1)+f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2+f_2(s_1)f_1(s_2)}$. If $\frac{f_1(s_1)f_2(s_2)\Phi_1+f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)} \leq \mathcal{E}_C([s_1, s_2]) < \frac{f_1(s_1)f_2(s_2)(\Phi_1+2\Phi_2-1)+f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, we maximized the error of the adversary by operating on the Pareto frontier. If error of the conference $\mathcal{E}_C([s_1, s_2]) \geq \frac{f_1(s_1)f_2(s_2)(\Phi_1+2\Phi_2-1)+f_2(s_1)f_1(s_2)(1-\Phi_2)}{f_1(s_1)f_2(s_2)+f_2(s_1)f_1(s_2)}$, we operate at the endpoint where error of the adversary is maximum and error of the conference is no larger than the desired $\mathcal{E}_C([s_1, s_2])$. The endpoint is the point with minimum error of the conference such that error of the adversary is maximum. Therefore, it is optimal for the conference.

Algorithm 4 follows the procedure by choosing the corresponding $q_1$ and $q_2$ for each point on the Pareto frontier and thus is optimal for the conference.

# Bibliography

[1] Ammar Ammar and Devavrat Shah. Efficient rank aggregation using partial data. In *SIGMETRICS*, 2012. 2

[2] Yukino Baba and Hisashi Kashima. Statistical quality estimation for general crowdsourcing tasks. In *KDD*, 2013. 2, 5

[3] Homanga Bharadhwaj, Dylan Turpin, Animesh Garg, and Ashton Anderson. De-anonymization of authors through arxiv submissions during double-blind review. *arXiv preprint arXiv:2007.00177*, 2020. 2

[4] Lyle Brenner, Dale Griffin, and Derek J Koehler. Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. *Organizational Behavior and Human Decision Processes*, 97(1):64–81, 2005. 1

[5] L. Charlin and R. S. Zemel. The Toronto Paper Matching System: An automated paper-reviewer assignment system. In *ICML Workshop on Peer Reviewing and Publishing Models*, 2013. 2

[6] Dalmeet Singh Chawla. Swiss funder draws lots to make grant decisions. *Nature*, 2021. 2

[7] Wenxin Ding, Nihar B. Shah, and Weina Wang. On the privacy-utility tradeoff in peer-review data analysis. In *AAAI Privacy-Preserving Artificial Intelligence (PPAI-21) workshop*, 2020. 2

[8] T Fiez, N Shah, and L Ratliff. A SUPER* algorithm to optimize paper bidding in peer review. In *Conference on Uncertainty in Artificial Intelligence*, 2020. 2

[9] Peter Flach, Sebastian Spiegler, Bruno Golénia, Simon Price, John Guiver, Ralf Herbrich, Thore Graepel, and Mohammed Zaki. Novel tools to streamline the conference review process: Experiences from SIGKDD'09. *SIGKDD Explor. Newsl.*, 11(2):63–67, May 2010. ISSN 1931-0145. 2, 5

[10] Yoav Freund, Raj D. Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003. URL `http://www.jmlr.org/papers/v4/freund03a.html`. 2

[11] N. Garg, T. Kavitha, A. Kumar, K. Mehlhorn, and J. Mestre. Assigning papers to referees. *Algorithmica*, 58(1):119–136, Sep 2010. 2

[12] Hong Ge, Max Welling, and Zoubin Ghahramani. A Bayesian model for calibrating conference review scores, 2013. URL `http://mlg.eng.cam.ac.uk/hong/nipsrevcal.pdf`. 2, 5

[13] Judy Goldsmith and Robert H Sloan. The ai conference paper assignment problem. In *Proc. AAAI Workshop on Preference Handling for Artificial Intelligence, Vancouver*, pages 53–57, 2007. 2

[14] Steven Jecmen, Hanrui Zhang, Ryan Liu, Nihar B. Shah, Vincent Conitzer, and Fei Fang. Mitigating manipulation in peer review via randomized reviewer assignments. In *NeurIPS*, 2020. 2

[15] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (peerread): Collection, insights and nlp applications. *arXiv preprint arXiv:1804.09635*, 2018. 2

[16] Ari Kobren, Barna Saha, and Andrew McCallum. Paper matching with local fairness constraints. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019. 2

[17] John Langford. ICML acceptance statistics, 2012. `http://hunch.net/?p=2517` [Online; accessed 14-May-2021]. 1

[18] Mengyao Liu, Vernon Choy, Philip Clarke, Adrian Barnett, Tony Blakely, and Lucy Pomeroy. The acceptability of using a lottery to allocate research funding: a survey of applicants. *Research integrity and peer review*, 5(1):1–7, 2020. 2

[19] R. S. MacKay, R. Kenna, R. J. Low, and S. Parker. Calibration with confidence: a principled method for panel assessment. *Royal Society Open Science*, 4(2), 2017. doi: 10.1098/rsos.160760. 2, 5

[20] Emaad Manzoor and Nihar B Shah. Uncovering latent biases in text: Method and application to peer review. In *AAAI*, 2021. 2

[21] Reshef Meir, Jérôme Lang, Julien Lesca, Natan Kaminsky, and Nicholas Mattei. A market-inspired bidding scheme for peer review paper assignment. In *Games, Agents, and Incentives Workshop at AAMAS*, 2020. 2

[22] David Mimno and Andrew McCallum. Expertise modeling for matching papers with reviewers. In *KDD*, 2007. 2

[23] Ioannis Mitliagkas, Aditya Gopalan, Constantine Caramanis, and Sriram Vishwanath. User rankings from comparisons: Learning permutations in high dimensions. In *Allerton Conference*, 2011. 2

[24] S. R. Paul. Bayesian methods for calibration of examiners. *British Journal of Mathematical and Statistical Psychology*, 34(2):213–223, 1981. 2, 5

[25] Magnus Roos, Jörg Rothe, and Björn Scheuermann. How to calibrate the scores of biased reviewers by quadratic programming. In *AAAI Conference on Artificial Intelligence*, 2011. 2, 5

[26] Magnus Roos, Jörg Rothe, Joachim Rudolph, Björn Scheuermann, and Dietrich Stoyan. A statistical approach to calibrating the scores of biased reviewers: The linear vs. the nonlinear model. In *Multidisciplinary Workshop on Advances in Preference Handling*, 2012. 2, 5

[27] Nihar B Shah. Systemic challenges and solutions on bias and unfairness in

peer review. Preprint `http://www.cs.cmu.edu/~nihars/preprints/Shah_Survey_PeerReview.pdf`, July 2021. 2

[28] Stanley S Siegelman. Assassins and zealots: variations in peer review. *Radiology*, 178(3): 637–642, 1991. 1

[29] David Soergel, Adam Saunders, and Andrew McCallum. Open scholarship and peer review: a time for experimentation. 2013. 2

[30] Ivan Stelmakh, Nihar Shah, and Aarti Singh. PeerReview4All: Fair and accurate reviewer assignment in peer review. *JMLR*, 2021. 2

[31] Ivan Stelmakh, Nihar Shah, Aarti Singh, and Hal Daumé III. Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. In *CSCW*, 2021. 2

[32] Camillo J Taylor. On the optimal assignment of conference papers to reviewers. 2008. 2

[33] Andrew Tomkins, Min Zhang, and William D Heavlin. Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48): 12708–12713, 2017. 2

[34] David Tran, Alex Valtchanov, Keshav Ganapathy, Raymond Feng, Eric Slud, Micah Goldblum, and Tom Goldstein. An open review of openreview: A critical analysis of the machine learning conference review process. *arXiv preprint arXiv:2010.05137*, 2020. 2

[35] Jingyan Wang and Nihar B Shah. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *AAMAS*, 2019. 2, 3

[36] Yichong Xu, Han Zhao, Xiaofei Shi, and Nihar Shah. On strategyproof conference review. In *IJCAI*, 2019. 2

[37] Weizhe Yuan, Pengfei Liu, and Graham Neubig. Can we automate scientific reviewing? *arXiv preprint arXiv:2102.00176*, 2021. 2