Duality and Auxiliary Functions for Bregman Distances

Stephen Della Pietra, Vincent Della Pietra, and John Lafferty

October 8, 2001

CMU-CS-01-109

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

We formulate and prove a convex duality theorem for Bregman distances and present a technique based on auxiliary functions for deriving and proving convergence of iterative algorithms to minimize Bregman distance subject to linear constraints.

This research was partially supported by the Advanced Research and Development Activity in Information Technology (ARDA), contract number MDA904-00-C-2106, and by the National Science Foundation (NSF), grant CCR-9805366.

Keywords: Bregman distance, convex duality, Legendre functions, auxiliary functions

I. Introduction

Convexity plays a central role in a wide variety of machine learning and statistical inference problems. A standard paradigm is to distinguish a preferred member from a set of candidates based upon a convex impurity measure or loss function tailored to the specific problem to be solved. Examples include least squares regression, decision trees, boosting, online learning, maximum likelihood for exponential models, logistic regression, maximum entropy, and support vector machines. Such problems can often be naturally cast as convex optimization problems involving a Bregman distance, which can lead to new algorithms, analytical tools, and insights derived from the powerful methods of convex analysis.

In this paper we formulate and prove a convex duality theorem for minimizing a general class of Bregman distances subject to linear constraints. The duality result is then used to derive iterative algorithms for solving the associated optimization problem. Our presentation is motivated by the recent work of Collins, Schapire, and Singer (2001), who showed how certain boosting algorithms and maximum likelihood logistic regression can be unified within the framework of Bregman distances. In particular, specific instances of the results given here are used by Collins et al. (2001) to show the convergence of a family iterative algorithms for minimizing the exponential or logistic loss.

While invoking methods from convex analysis can unify and clarify the relationship between different methods, the higher level of abstraction often comes at a price, since there can be considerable technicalities. For example, in some treatments the assumptions on the convex functions that can be used to define Bregman distances are very technical and difficult to verify. Here we trade off generality for relative simplicity by working with a restricted class of Bregman distances, which however includes many of the examples that arise in machine learning. Our treatment of duality and auxiliary functions for Bregman distances closely parallels the results presented by Della Pietra et al. (1997) for the Kullback-Leibler divergence. In particular, the statement and proof of the duality theorem given in (Della Pietra et al., 1997) carries over with only a few changes to the class of Bregman distances we consider.

Our approach differs from much of the literature in convex analysis in several ways. First, we work primarily with the *argument* at which a convex conjugate takes on its value, rather than the value of the function itself. The reason for this is that the argument corresponds to a statistical model, which is the main object of interest in statistical or machine learning applications, while the value corresponds to a likelihood or loss function. Second, while Bregman distances are typically defined only on the interior of the domain of the underlying convex function, we assume that there is a continuous extension to the entire domain. This makes it possible to formulate a very natural duality theorem that also includes many cases required in practice, when the desired model may lie on the boundary of the domain.

The following section recalls the standard definitions from convex analysis that will be required, and presents the technical assumptions made on the class of Bregman distances that we work with. We also introduce some new terminology, using the terms Legendre-Bregman conjugate and Legendre-Bregman projection to extend the classical notion of the Legendre conjugate and transform to Bregman distances. Section 3 contains the statement and proof of the duality theorem that connects the primal problem with its dual, showing that the solution is characterized in geometrical terms by a Pythagorean equality. Section 4 defines the notion of an auxiliary function, which is used to construct iterative algorithms for solving constrained optimization problems. This section shows how convexity can be used to derive an auxiliary function for Bregman distances based on separable functions. The last section summarizes the main results of the paper.

II. Bregman Distances and Legendre-Bregman Projections

In this section we begin by establishing our notation and recalling the relevant notions from convex analysis that we require; the classic text (Rockafellar, 1970) remains one of the best references for this material. We then define Bregman distances and their associated conjugate functions and projections, and derive various relations between these that will be important in proving the duality theorem. Next we state our assumptions on the underlying convex function that enable us to derive these properties for the continous extension of the Bregman distance to the entire domain.

A. Notation and Basic Definitions

We will use notation that is suggestive of our main applications: rather than $\phi(x)$ we will write $\phi(p)$ or $\phi(q)$, having in mind probability distributions p or q. A convex function $\phi : S \subset \mathbb{R}^m \longrightarrow [-\infty, +\infty]$ is *proper* if there is no $q \in S$ with $\phi(q) = -\infty$ and there is some q with $\phi(q) \neq \infty$. The *effective domain of* ϕ, denoted Δ_ϕ, is the set of points where ϕ is finite: $\Delta_\phi = \{q \in S \mid \phi(q) < \infty\}$; for brevity we usually refer to Δ_ϕ as simply the *domain* of ϕ. A proper convex function is *closed* if it is lower semi-continuous. The *conjugate* ϕ^* of ϕ is given by

$$\phi^*(v) \quad = \quad \sup_{q \in S} \left(\langle q, v \rangle - \phi(q) \right) \tag{2.1}$$

A proper convex function ϕ is said to be *essentially smooth* or *steep* if it is differentiable on the interior of its domain $\text{int}(\Delta_\phi) \neq \emptyset$, and if $\lim_{n \to \infty} |\nabla \phi(q_n)| = +\infty$ whenever q_n is a sequence in $\text{int}(\Delta_\phi)$ converging to a point on the boundary of $\text{int}(\Delta_\phi)$. The function ϕ is said to be *coercive* in case the level set $\{q \in S \mid \phi(q) \leq c\}$ is bounded for every $c \in \mathbb{R}$.

Definition 2.1. *Let ϕ be a closed, convex and proper function defined on $S \subset \mathbb{R}^m$, such that ϕ is differentiable on $\text{int}(\Delta_\phi) \neq \emptyset$. The Bregman distance $D_\phi : \Delta_\phi \times \text{int}(\Delta_\phi) \longrightarrow [0, \infty)$ is defined by*

$$D_\phi(p, q) \quad = \quad \phi(p) - \phi(q) - \langle \nabla \phi(q), p - q \rangle \tag{2.2}$$

Bregman distance can be interpreted as a measure of the convexity of ϕ. This is easy to visualize in the one-dimensional case: drawing a tangent line to the graph of ϕ at q, the Bregman distance $D_\phi(p, q)$ is seen as the vertical distance between this line and the point $\phi(p)$.

Legendre functions are a very well behaved family of convex functions that will make working with Bregman distances much easier.

Definition 2.2. *A closed convex function ϕ is Legendre, or a convex function of Legendre type, in case $\text{int}(\Delta_\phi)$ is convex and ϕ is essentially smooth and strictly convex on $\text{int}(\Delta_\phi)$.*

The primary properties that make working with Legendre functions convenient are summarized in the following results quoted from (Rockafellar, 1970).

Proposition 2.3. (Rockafellar, 1970; Theorem 26.5) *If ϕ is a convex function of Legendre type then $\nabla\phi : \mathrm{int}(\Delta_\phi) \longrightarrow \mathrm{int}(\Delta_{\phi^*})$ is a bijection, continuous in both directions, and $\nabla\phi^* = (\nabla\phi)^{-1}$.*

Proposition 2.4. *Suppose that ϕ is Legendre, and ψ is a proper closed, convex function that is also essentially smooth. Then $\phi + \psi$ is Legendre.*

In particular, since for fixed $q \in \mathrm{int}(\Delta_\phi)$ the mapping $p \longmapsto \langle \nabla\phi(q), p - q\rangle + \phi(q)$ is affine linear, the function $p \longmapsto D_\phi(p, q)$ is Legendre with domain Δ_ϕ, and with conjugate domain $\Delta_{\phi^*} - \nabla\phi(q)$.

Definition 2.5. *For ϕ a convex function of Legendre type we define the Legendre-Bregman conjugate $\ell_\phi : \mathrm{int}(\Delta_\phi) \times \mathbb{R}^m \longrightarrow \mathbb{R} \cup \{\infty\}$ as*

$$\ell_\phi(q, v) = \sup_{p \in \Delta_\phi} \left(\langle v, p\rangle - D_\phi(p, q)\right) \tag{2.3}$$

We define the Legendre-Bregman projection $\mathcal{L}_\phi : \mathrm{int}(\Delta_\phi) \times \mathbb{R}^m \longrightarrow \Delta_\phi$ as

$$\mathcal{L}_\phi(q, v) = \arg\max_{p \in \Delta_\phi} \left(\langle v, p\rangle - D_\phi(p, q)\right) \tag{2.4}$$

whenever this is well defined.

Let us explain our choice of terminology in the above definition, which is nonstandard. When h is a convex function of Legendre type, the Legendre conjugate, as defined in (Rockafellar 1970; Chapter 26), corresponds to the convex conjugate h^*. For fixed q, our definition of the Legendre-Bregman conjugate is simply the classical Legendre conjugate for the convex function $h(p) = D_\phi(p, q)$. Note that Rockafellar defines the *Legendre transform* as the mapping from the original convex function (and domain) to its Legendre conjugate (and associated domain). The Legendre-Bregman projection $\mathcal{L}_\phi(q, v)$ is the actual argument at which the maximum is attained. As shown by the following result, our use of the term "projection" accords with the standard terminology of Bregman projections.

Proposition 2.6. *Let ϕ be Legendre. Then for $q \in \mathrm{int}(\Delta_\phi)$ and $v \in \mathrm{int}(\Delta_{\phi^*}) - \nabla\phi(q)$, the Legendre-Bregman projection is given explicitly by*

$$\mathcal{L}_\phi(q, v) = (\nabla\phi^*)\left(\nabla\phi(q) + v\right) \tag{2.5}$$

Moreover, it can be written as a Bregman projection

$$\mathcal{L}_\phi(q, v) = \arg\min_{p \in \Delta_\phi \cap H} D_\phi(p, q) \tag{2.6}$$

for the hyperplane $H = \{p \in \mathbb{R}^m \mid \langle p, v\rangle = b\}$ with $b = \langle \mathcal{L}_\phi(q, v), v\rangle$.

Proof. To prove the first statement, note that a stationary point p^* of $\langle p, v \rangle - D_\phi(p, q)$ must satisfy:

$$\nabla\phi(p^*) = \nabla\phi(q) + v \tag{2.7}$$

Since ϕ is Legendre, for $\nabla\phi(q) + v \in \operatorname{int}(\Delta_{\phi^*})$, we then have that

$$\begin{aligned} p^* &= (\nabla\phi)^{-1}(\nabla\phi(q) + v) &(2.8)\\ &= (\nabla\phi^*)(\nabla\phi(q) + v) &(2.9) \end{aligned}$$

using the fact that $(\nabla\phi)^{-1}$ is well defined and equal to $\nabla\phi^*$ from Proposition 2.3. The second statement follows from, for example, the results in Section 2.2 of Censor and Zenios (1997) on projections onto hyperplanes, noting that every Legendre function is zone consistent. □

In our formulation of the duality theorem, it is the Legendre-Bregman *projection* $\mathcal{L}_\phi(q, v)$ rather than the conjugate function $\ell_\phi(q, v)$ that plays a central role, leading to a natural and simple statement of convex duality for Bregman distances. This projection corresponds to a statistical model, which is the primary object of interest for machine learning problems.

B. Basic Relations

We now derive some basic algebraic relations between D_ϕ, $\mathcal{L}_\phi$, and ℓ_ϕ. These relations will be important in establishing the geometrical aspects of the duality theorem, as well as for deriving auxiliary functions. In order to free us from having to specify the domain of ℓ_ϕ and $\mathcal{L}_\phi$, we will in the following assume that $\Delta_{\phi^*} = \mathbb{R}^m$.

Proposition 2.7. *Let ϕ be Legendre, with $\Delta_{\phi^*} = \mathbb{R}^m$. For fixed $p \in \Delta_\phi$, $D_\phi(p, \mathcal{L}_\phi(q, v))$ is continuous in $q \in \operatorname{int}(\Delta_\phi)$ and convex in v. Together, the Legendre-Bregman conjugate and projection satisfy*

$$\begin{aligned} D_\phi(p, q) - D_\phi(p, \mathcal{L}_\phi(q, v)) &= \langle v, p \rangle - \ell_\phi(q, v) &(2.10)\\ &= D(\mathcal{L}_\phi(q, v), q) + \langle v, p - \mathcal{L}_\phi(q, v) \rangle &(2.11) \end{aligned}$$

for all $p \in \Delta_\phi$, $q \in \operatorname{int}(\Delta_\phi)$ and $v \in \mathbb{R}^m$.

Proof. Using the definition of Bregman distance and Proposition 2.6, we have for $q \in \operatorname{int}(\Delta_\phi)$ that

$$\begin{aligned} &D_\phi(p, q) - D_\phi(p, \mathcal{L}_\phi(q, v))\\ &\quad = \phi(\mathcal{L}_\phi(q, v)) - \phi(q) + \langle \nabla\phi(\mathcal{L}_\phi(q, v)), p - \mathcal{L}_\phi(q, v) \rangle - \langle \nabla\phi(q), p - q \rangle &(2.12)\\ &\quad = \phi(\mathcal{L}_\phi(q, v)) - \phi(q) + \langle \nabla\phi(q) + v, p - \mathcal{L}_\phi(q, v) \rangle - \langle \nabla\phi(q), p - q \rangle &(2.13)\\ &\quad = \langle v, p \rangle - \langle v, \mathcal{L}_\phi(q, v) \rangle + \phi(\mathcal{L}_\phi(q, v)) - \phi(q) - \langle \nabla\phi(q), \mathcal{L}_\phi(q, v) - q \rangle &(2.14)\\ &\quad = \langle v, p \rangle - \langle v, \mathcal{L}_\phi(q, v) \rangle + D_\phi(\mathcal{L}_\phi(q, v), q) &(2.15)\\ &\quad = \langle v, p \rangle - \ell_\phi(q, v) &(2.16) \end{aligned}$$

Therefore (2.10) holds for $p \in \Delta_\phi$ and $q \in \mathrm{int}(\Delta_\phi)$. From the definition of the Legendre-Bregman conjugate we have for $q \in \mathrm{int}(\Delta_\phi)$, that

$$\ell_\phi(q, v) \quad = \quad \langle v, \mathcal{L}_\phi(q, v) \rangle - D_\phi(\mathcal{L}_\phi(q, v), q) \tag{2.17}$$

Equation (2.24) results from combining this with (2.10). The convexity of $D_\phi(p, \mathcal{L}_\phi(q, v))$ in v follows from (2.10), which expresses $D_\phi(p, \mathcal{L}_\phi(q, v))$ as a sum of the convex functions $\ell_\phi(q, v)$ and $D_\phi(p, q) - \langle v, p \rangle$. $\square$

The next result shows how $D_\phi(p, \mathcal{L}_\phi(q, v))$ varies with v, and will be an important ingredient in the duality result of the next section.

Proposition 2.8. *Let ϕ be Legendre, with $p \in \Delta_\phi$ and $q \in \mathrm{int}(\Delta_\phi)$. Then for $v \in \mathbb{R}^m$, the mapping $t \mapsto D_\phi(p, \mathcal{L}_\phi(q, tv))$ is differentiable at $t = 0$, with derivative*

$$\left.\frac{d}{dt}\right|_{t=0} D_\phi(p, \mathcal{L}_\phi(q, tv)) = \langle v, q \rangle - \langle v, p \rangle \ . \tag{2.18}$$

Proof. Since ϕ is Legendre, if $q \in \mathrm{int}(\Delta_\phi)$ then $\mathcal{L}_\phi(q, tv) \in \mathrm{int}(\Delta_\phi)$; see for example Theorem 3.12 of (Bauschke & Borwein, 1997). From Proposition 2.7 we have that

$$\begin{aligned}
\left.\frac{d}{dt}\right|_{t=0} D_\phi(p, \mathcal{L}_\phi(q, tv)) \quad &= \quad \left.\frac{d}{dt}\right|_{t=0} \left(\langle tv, \mathcal{L}_\phi(q, tv) \rangle - \langle tv, p \rangle + D_\phi(\mathcal{L}_\phi(q, tv), q) \right) &(2.19)\\
&= \quad \langle v, q \rangle - \langle v, p \rangle + \left.\frac{d}{dt}\right|_{t=0} \phi(\mathcal{L}_\phi(q, tv)) - \langle \nabla\phi(q), \left.\frac{d}{dt}\right|_{t=0} \mathcal{L}_\phi(q, tv) \rangle &(2.20)\\
&= \quad \langle v, q \rangle - \langle v, p \rangle &(2.21)
\end{aligned}$$

which proves the result for $p \in \Delta_\phi$ and $q \in \mathrm{int}(\Delta_\phi)$. $\square$

C. The Continuous Extension

The results above are given in terms of the Bregman distance using its standard definition as a function on $\Delta_\phi \times \mathrm{int}(\Delta_\phi)$. We now make assumptions that allow us to work with D_ϕ as an extended real-valued function on $\Delta_\phi \times \Delta_\phi$. This enables us to formlate a very natural and general duality result, presented in the following section.

Informally, we assume that D_ϕ extends continously from $\Delta_\phi \times \mathrm{int}(\Delta_\phi)$ to $\Delta_\phi \times \Delta_\phi$, and that $\mathcal{L}_\phi$ extends continously from $\mathrm{int}(\Delta_\phi) \times \mathbb{R}^m$ to $\Delta_\phi \times \mathbb{R}^m$. In addition, we require a form of compactness to guarantee the existence of certain minimizers. As before, in order to simplify the presentation we assume that the range of $\nabla\phi$ is all of $\mathbb{R}^m$.

Thus, we make the following assumptions on ϕ:

A1. ϕ is of Legendre type;

A2. $\Delta_{\phi^*} = \mathbb{R}^m$;

A3. D_ϕ extends to a function $D_\phi : \Delta_\phi \times \Delta_\phi \longrightarrow [0, \infty]$ such that $D_\phi(p, q)$ is jointly continuous in p and q, and satisfies $D_\phi(p, q) = 0$ if and only if $p = q$.

A4. $\mathcal{L}_\phi$ extends to a function $\mathcal{L}_\phi : \Delta_\phi \times \mathbb{R}^m \to \Delta_\phi$ such that $\mathcal{L}_\phi(q, v)$ is jointly continuous in q and v, and satisfies $\mathcal{L}_\phi(q, 0) = q$.

A5. $D_\phi(p, \cdot)$ is coercive for every $p \in \Delta_\phi \backslash \mathrm{int}(\Delta_\phi)$;

Note that since for a Legendre function $\nabla\phi$ is continous on $\mathrm{int}(\Delta_\phi)$ (Proposition 2.3), it follows from Definition 2.1 that $D_\phi(p, q)$ is jointly continous on $\Delta_\phi \times \mathrm{int}(\Delta_\phi)$ and from Proposition 2.6 that $\mathcal{L}_\phi(q, v)$ is jointly continous on $\mathrm{int}(\Delta_\phi) \times \mathbb{R}^m$. We also note that since we assume $\Delta_{\phi^*} = \mathbb{R}^m$, $D_\phi(p, \cdot)$ is automatically coercive for $p \in \mathrm{int}(\Delta_\phi)$. Together, conditions A1–A5 imply that ϕ is a *Bregman-Legendre function* as defined by Bauschke and Borwein (1997).

Now, from the definition of the Legendre-Bregman conjugate we have

$$\ell_\phi(q, v) \quad = \quad \langle v, \mathcal{L}_\phi(q, v) \rangle - D_\phi(\mathcal{L}_\phi(q, v), q) \tag{2.22}$$

for $q \in \mathrm{int}(\Delta_\phi)$. Properties A4 and A5 allow us to define $\ell_\phi : \Delta_\phi \times \mathbb{R}^m \longrightarrow \mathbb{R}$ as the continuous extension of $\ell_\phi : \mathrm{int}(\Delta_\phi) \times \mathbb{R}^m \longrightarrow \mathbb{R}$, satisfying the same identity. Thus, the Legendre-Bregman conjugate $\ell_\phi(q, v)$ is continuous in q, continuous and convex in v, and satisfies $\ell_\phi(q, 0) = 0$.

Proposition 2.7 now generalizes to the continuous extension.

Proposition 2.9. *Let ϕ satisfy A1–A4. For fixed $p \in \Delta_\phi$, $D_\phi(p, \mathcal{L}_\phi(q, v))$ is continuous in q and convex in v. Together, the Legendre-Bregman conjugate and projection satisfy*

$$\begin{aligned} D_\phi(p, q) - D_\phi(p, \mathcal{L}_\phi(q, v)) \quad &= \quad \langle v, p \rangle - \ell_\phi(q, v) & (2.23) \\ &= \quad D(\mathcal{L}_\phi(q, v), q) + \langle v, p - \mathcal{L}_\phi(q, v) \rangle & (2.24) \end{aligned}$$

for all $p, q \in \Delta_\phi$ and $v \in \mathbb{R}^m$.

Proof. This follows directly from the joint continuity of D_ϕ, $\mathcal{L}_\phi$, and ℓ_ϕ. $\square$

The differential identity in Proposition 2.8 also extends.

Proposition 2.10. *Let ϕ satisfy A1–A4, and let $p, q \in \Delta_\phi$ with $D_\phi(p, q) < \infty$. Then for $v \in \mathbb{R}^m$, the mapping $t \mapsto D_\phi(p, \mathcal{L}_\phi(q, tv))$ is differentiable at $t = 0$, with derivative*

$$\left.\frac{d}{dt}\right|_{t=0} D_\phi(p, \mathcal{L}_\phi(q, tv)) = \langle v, q \rangle - \langle v, p \rangle \ . \tag{2.25}$$

Proof. Let $q \in \text{int}(\Delta_\phi)$. From Proposition 2.8 we know that

$$\left.\frac{d}{dt}\right|_{t=0} D_\phi(p, \mathcal{L}_\phi(q, tv)) \;\; = \;\; \langle v, q \rangle - \langle v, p \rangle \tag{2.26}$$

Thus, since $\mathcal{L}_\phi(q, (t+s)v) = \mathcal{L}_\phi(\mathcal{L}_\phi(q, tv), sv)$, we also have that

$$\frac{d}{dt} D(p, \mathcal{L}_\phi(q, tv)) \;\; = \;\; \langle v, \mathcal{L}_\phi(q, tv) \rangle - \langle v, p \rangle \tag{2.27}$$

To show that the result holds when $q \in \Delta_\phi \backslash \text{int}(\Delta_\phi)$, we'll use a fact from elementary analysis: if $f_n \to f$, and f_n' is continuous with $f_n'(t) \to g(t)$ uniformly for $t \in [a, b]$, then g is continous and $f'(t) = g(t)$. First, let $q \in \text{int}(\Delta_\phi)$ and $p \notin \text{int}(\Delta_\phi)$. Suppose $p_n \in \text{int}(\Delta_\phi)$ with $p_n \to p$. Since $\mathcal{L}_\phi(q, tv) \in \text{int}(\Delta_\phi)$, we have from the above calculation that

$$\frac{d}{dt} D(p_n, \mathcal{L}_\phi(q, tv)) \;\; = \;\; \langle v, \mathcal{L}_\phi(q, tv) \rangle - \langle v, p_n \rangle \;\longrightarrow\; \langle v, \mathcal{L}_\phi(q, tv) \rangle - \langle v, p \rangle \tag{2.28}$$

where the convergence is uniform on every interval $[a, b]$ around zero; property (2.25) follows.

Now suppose that $p \in \Delta_\phi$, $q \in \Delta_\phi \backslash \text{int}(\Delta_\phi)$, and $q_n \in \text{int}(\Delta_\phi)$ with $q_n \to q$. Because $\mathcal{L}_\phi(q, v)$ is jointly continous in q and v, it is uniformly continuous on every compact set of (q, v). In particular, $\langle v, \mathcal{L}_\phi(q_n, tv) \rangle - \langle v, p \rangle$ converges uniformly in t to $\langle v, \mathcal{L}_\phi(q, tv) \rangle - \langle v, p \rangle$ on every interval $[a, b]$. Thus property (2.25) holds for all $p, q \in \Delta_\phi$. $\square$

Proposition 2.9 and 2.10 are the main computations that we will require in the following section.

III. Duality

In this section we derive the main duality result. The setup is that we have a set of *features* $f^{(j)} \in \mathbb{R}^m$, $j = 1, 2, \ldots, n$ and denote by F the $m \times n$ matrix with columns given by the $f^{(j)}$. These features correspond to the "weak learners" in boosting, or to the sufficient statistics in an exponential model. The primal problem constrains the values $\langle p, f^{(j)} \rangle$, and these constraints carry over to Lagrange multipliers in a family of Legendre-Bregman projections $\mathcal{L}(q, F\lambda)$ in the dual problem.

Definition 3.1. *For a given element $p_0 \in \Delta_\phi$, the feasible set for p_0 and F is defined by*

$$\mathcal{P}(p_0, F) \;\; = \;\; \left\{ p \in \Delta_\phi \;\mid\; \langle p, f^{(j)} \rangle = \langle p_0, f^{(j)} \rangle,\; j = 1, \ldots n \right\} \tag{3.1}$$

For a given $q_0 \in \Delta_\phi$, the Legendre-Bregman projection family for q_0 and F is defined by

$$\mathcal{Q}(q_0, F) \;\; = \;\; \{ q \in \Delta_\phi \;\mid\; q = \mathcal{L}_\phi(q_0, F\lambda) \text{ for some } \lambda \in \mathbb{R}^n \} \tag{3.2}$$

Trivially, both sets are non-empty since $p_0 \in \mathcal{P}(p_0, F)$ and $q_0 \in \mathcal{Q}(q_0, F)$. Since p_0, q_0, and F will be fixed, we will use abbreviated notation and refer to these sets as $\mathcal{P}$ and $\mathcal{Q}$. We use $\overline{\mathcal{Q}}$ to denote the closure of $\mathcal{Q}(q_0, F)$ as a subset of $\mathbb{R}^m$. Duality relates the projection onto $\mathcal{P}$ to the projection onto $\overline{\mathcal{Q}}$.

Proposition 3.2. *Let ϕ satisfy A1–A5, and suppose that $p_0, q_0 \in \Delta_\phi$ with $D_\phi(p_0, q_0) < \infty$. Then there exists a unique $q_\star \in \Delta_\phi$ satisfying the following four properties:*

(1) $q_\star \in \mathcal{P} \cap \overline{\mathcal{Q}}$

(2) $D_\phi(p,q) = D_\phi(p,q_\star) + D_\phi(q_\star,q)$ *for any* $p \in \mathcal{P}$ *and* $q \in \overline{\mathcal{Q}}$

(3) $q_\star = \arg\min_{p \in \mathcal{P}} D_\phi(p, q_0)$

(4) $q_\star = \arg\min_{q \in \overline{\mathcal{Q}}} D_\phi(p_0, q)$

Moreover, any one of these four properties determines $q_\star$ uniquely.

In order to prove Proposition 3.2, we first prove two lemmas. The first shows that there is at least one member in common between $\mathcal{P}$ and $\overline{\mathcal{Q}}$; the second shows that the *Pythagorean equality* (2) holds for any such member.

Lemma 3.3. *If $D_\phi(p_0, q_0) < \infty$ then $\mathcal{P} \cap \overline{\mathcal{Q}}$ is nonempty.*

Proof. Note that since $D_\phi(p_0, q_0) < \infty$, $D_\phi(p_0, q)$ is not identically ∞ on $\overline{\mathcal{Q}}$. Also, the mapping $\lambda \mapsto D_\phi(p_0, \mathcal{L}_\phi(q_0, F\lambda))$ is continuous and convex. Let $\mathcal{R}$ be the level set

$$\mathcal{R} \;=\; \{q \in \Delta_\phi \mid D_\phi(p_0, q) \le D(p_0, q_0)\} \tag{3.3}$$

We know from Assumption A5 that $\mathcal{R}$ is bounded. Thus $D_\phi(p_0, q)$ attains its minimum at a (not necessarily unique) point $q_\star \in \overline{\mathcal{Q} \cap \mathcal{R}} \subset \overline{\mathcal{Q}}$. We will show that $q_\star$ is also in $\mathcal{P}$.

Let $\overline{q} \in \overline{\mathcal{Q}}$, and let $\mu_j \in \mathbb{R}^n$ be such that $\overline{q} = \lim_{j\to\infty} \mathcal{L}_\phi(q_0, F\mu_j)$. Then by the continuity of $\mathcal{L}_\phi(\cdot,\cdot)$,

$$\mathcal{L}_\phi(\overline{q}, F\lambda) \;=\; \lim_{j\to\infty} \mathcal{L}_\phi(\mathcal{L}_\phi(q_0, F\mu_j), F\lambda) \tag{3.4}$$

$$\;=\; \lim_{j\to\infty} \mathcal{L}_\phi(q_0, F(\mu_j + \lambda)) \;\in\; \overline{\mathcal{Q}} \tag{3.5}$$

Thus $\overline{\mathcal{Q}}$ is closed under the mapping $q \mapsto \mathcal{L}_\phi(q, F\lambda)$ for $\lambda \in \mathbb{R}^m$, and $\mathcal{L}_\phi(q_\star, F\lambda)$ is in $\overline{\mathcal{Q}}$ for any λ. By the definition of $q_\star$, it follows that $\lambda = 0$ is a minimum of the function $\lambda \mapsto D_\phi(p_0, \mathcal{L}_\phi(q_\star, F\lambda))$. Taking derivatives with respect to λ and using Proposition 2.10 we conclude that $\langle q_\star, f\rangle = \langle p_0, f\rangle$; thus $q_\star \in \mathcal{P}$. □

Lemma 3.4. *If $q_\star \in \mathcal{P} \cap \overline{\mathcal{Q}}$ then the Pythagorean equality $D_\phi(p,q) = D_\phi(p,q_\star) + D_\phi(q_\star,q)$ holds for any $p \in \mathcal{P}$ and $q \in \overline{\mathcal{Q}}$.*

Proof. Suppose that $p_1, p_2, q_1, q_2 \in \Delta_\phi$ with $q_2 = \mathcal{L}_\phi(q_1, F\lambda)$. From Proposition 2.9 we have that

$$D_\phi(p_1, q_1) - D_\phi(p_1, q_2) \;=\; \langle p_1, F\lambda\rangle - \ell_\phi(p_1, F\lambda) \tag{3.6}$$

and similarly

$$D_\phi(p_2, q_1) - D_\phi(p_2, q_2) \;=\; \langle p_2, F\lambda\rangle - \ell_\phi(p_2, F\lambda) \tag{3.7}$$

Therefore,

$$\begin{aligned} D_\phi(p_1, q_1) - D_\phi(p_1, q_2) - D_\phi(p_2, q_1) + D_\phi(p_2, q_2) &= \langle p_1, F\lambda\rangle - \langle p_2, F\lambda\rangle \\ &= \sum_{j=1}^{n} \lambda_j \left(\langle p_1, f^{(j)}\rangle - \langle p_2, f^{(j)}\rangle \right) \end{aligned} \tag{3.8}$$

It follows from this identity and the continuity of D_ϕ that

$$D_\phi(p_1, q_1) - D_\phi(p_1, q_2) - D_\phi(p_2, q_1) + D_\phi(p_2, q_2) = 0 \tag{3.9}$$

if $p_1, p_2 \in \mathcal{P}$ and $q_1, q_2 \in \overline{\mathcal{Q}}$. The lemma follows by taking $p_1 = q_1 = q_\star$. □

Proof of Proposition 3.2. Choose $q_\star$ to be any point in $\mathcal{P} \cap \overline{\mathcal{Q}}$. Such a $q_\star$ exists by Lemma 3.3. It satisfies property (1) by definition, and it satisfies property (2) by Lemma 3.4. As a consequence of property (2), it also satisfies properties (3) and (4). To check property (3), note that if q is any point in $\overline{\mathcal{Q}}$, then $D_\phi(p_0, q) = D_\phi(p_0, q_\star) + D_\phi(q_\star, q) \geq D_\phi(p_0, q_\star)$. Similarly, property (4) must hold since if p is any point in $\mathcal{P}$, then $D_\phi(p, q_0) = D_\phi(p, q_\star) + D_\phi(q_\star, q_0) \geq D(q_\star, q_0)$.

It remains to prove that each of the four properties (1)–(4) determines $q_\star$ uniquely. In other words, we need to show that if m is a point in Δ_ϕ satisfying any of the four properties (1)–(4), then $m = q_\star$. Suppose that m satisfies property (1). Then property (2) with $p = q = m$ implies that $D_\phi(m, m) = D_\phi(m, q_\star) + D_\phi(q_\star, m)$. Since $D_\phi(m, m) = 0$, it follows that $D_\phi(m, q_\star) = 0$ so $m = q_\star$. If m satisfies property (2), then the same argument with $q_\star$ and m reversed proves that $m = q_\star$. Suppose that m satisfies property (3). Then

$$D_\phi(p_0, q_\star) \geq D_\phi(p_0, m) = D_\phi(p_0, q_\star) + D_\phi(q_\star, m) \tag{3.10}$$

where the second equality follows from property (2) for $q_\star$. Thus $D_\phi(q_\star, m) \leq 0$ so $m = q_\star$. If m satisfies property (4), then

$$D_\phi(q_\star, q_0) \geq D_\phi(m, q_0) = D_\phi(m, q_\star) + D_\phi(q_\star, q_0) \tag{3.11}$$

showing once again that $m = q_\star$. □

In the following section we outline the auxilary function method for building iterative algorithms to compute $q_\star$, and show how to use convexity to derive an auxiliary function for separable Bregman distances.

IV. Auxiliary Functions

The auxiliary function approach is conceptually simple: bound the change in Bregman distance from below using a function that is easy to compute and that decouples the constraints. Maximizing this auxiliary function we obtain new parameters $\lambda' = \lambda + \Delta\lambda$ and a new model $q_{\lambda+\Delta\lambda}$ given by

$$\begin{aligned} q_{\lambda+\Delta\lambda} &= \mathcal{L}_\phi(q_\lambda, F\Delta\lambda) &(4.1)\\ &= \mathcal{L}_\phi(\mathcal{L}_\phi(q_0, F\lambda), F\Delta\lambda) &(4.2)\\ &= \mathcal{L}_\phi(q_0, F(\lambda + \Delta\lambda)) &(4.3)\end{aligned}$$

We then use the duality theorem to show that when $\Delta\lambda = 0$, we must have that $q = q_\star$.

The strategy is very similar to EM. In an EM algorithm, the Q-function $Q(\lambda', \lambda)$ is computed as a lower bound to the change in log-likelihood:

$$\begin{aligned} \sum_x p_0(x) \log \frac{q(x \mid \lambda')}{q(x \mid \lambda)} &= \sum_x p_0(x) \log \frac{\sum_h q(x, h \mid \lambda')}{q(x \mid \lambda)} &(4.4)\\ &= \sum_x p_0(x) \log \sum_h q(h \mid x, \lambda) \frac{q(x, h \mid \lambda')}{q(x, h \mid \lambda)} &(4.5)\\ &\geq \sum_h q(h \mid x, \lambda) \log \frac{q(x, h \mid \lambda')}{q(x, h \mid \lambda)} &(4.6)\\ &\stackrel{\text{def}}{=} \mathcal{Q}(\lambda', \lambda) &(4.7)\end{aligned}$$

where (x, h) is the complete data, x is the incomplete data, and the inequality follows from the concavity of the logarithm. After computing Q in the E-step, it is then maximized over λ' in the M-step.

In the same way, for Bregman distances the aim is to derive an auxiliary function $\mathcal{A}(\lambda', \lambda)$ whose calculation can be carried out efficiently in something like an E-step, and such that it can be easily maximized over λ' in an M-step. However, just as for EM, this is a general strategy more than it is a precise algorithm. A particular Bregman distance problem may require some ingenuity in order to come up with an appropriate auxiliary function.

This is the general motivation behind the following two definitions.

Definition 4.1. *A function $\mathcal{A} : \Delta_\phi \times \mathbb{R}^n \longrightarrow \mathbb{R}$ is called an auxiliary function for p_0 and F in case*

1. *$\mathcal{A}(q, \lambda)$ is continuous in q and $\mathcal{A}(q, 0) = 0$*
2. *$D_\phi(p_0, q) - D_\phi(p_0, \mathcal{L}_\phi(q, F\lambda)) \geq \mathcal{A}(q, \lambda)$*
3. *If $\lambda = 0$ is a maximum of $\mathcal{A}(q, \lambda)$, then $\langle q, f^{(j)} \rangle = \langle p_0, f^{(j)} \rangle$ for $j = 1, \ldots, n$.*

Definition 4.2. *Let $\mathcal{A}$ be an auxiliary function and $q_0 \in \Delta_\phi$. The update sequence for q_0 with respect to $\mathcal{A}$ is defined by $q^{(0)} = q_0$ and*

$$q^{(t+1)} \stackrel{\text{def}}{=} \mathcal{L}_\phi(q^{(t)}, F\lambda^{(t)}) \;\text{ where } \lambda^{(t)} = \arg\max_\lambda \mathcal{A}(q^{(t)}, \lambda) \tag{4.8}$$

The reason for defining auxiliary functions in this way is the following result, which can be proved in a similar way to Proposition 5 in (Della Pietra et al., 1997) or Lemma 2 in (Collins et al., 2001). The compactness assumption will in general follow from coercivity in Assumption A5.

Proposition 4.3. *Suppose that the sequence $q^{(t)}$ lies in a compact set. Then*

$$\lim_{t \to \infty} q^{(t)} \quad = \quad \arg\min_{q \in \overline{\mathcal{Q}}} D_\phi(p_0, q) \tag{4.9}$$

As we now explain, auxiliary functions can be conveniently constructed by using the relation

$$D_\phi(p, q) - D_\phi(p, \mathcal{L}_\phi(q, v)) = \langle p, v \rangle - \ell_\phi(q, v) \tag{4.10}$$

from Proposition 2.9 and exploiting the convexity of $\ell_\phi(q, v)$. For simplicity, we will assume that ϕ is *separable*, so that $\phi(p) = \sum_i \phi_i(p_i)$ with each $\phi_i : \mathbb{R} \to \mathbb{R}$ satisfying properties A1–A5 (with $m = 1$). Auxiliary functions in the general case can be derived using similar arguments. For the separable case, clearly

$$D_\phi(p, q) \quad = \quad \sum_{i=1}^{m} D_{\phi_i}(p_i, q_i) \tag{4.11}$$

$$\ell_\phi(q, v) \quad = \quad \sum_{i=1}^{m} \ell_{\phi_i}(q_i, v_i) \tag{4.12}$$

where $\ell_{\phi_i}(q, v) = \sup_{p \in \Delta_{\phi_i}} (pv - D_{\phi_i}(p, q))$ is the Legendre-Bregman conjugate of ϕ_i.

Proposition 4.4. *For each $i = 1, \ldots, m$, select N_i so that $\sum_{j=1}^{n} |f_i^{(j)}| \leq N_i$, and set $s_{ij} = \operatorname{sign}(f_i^{(j)})$. Then*

$$\mathcal{A}(q, \lambda) \stackrel{\text{def}}{=} \sum_{j=1}^{n} \lambda_j \left\langle p_0, f^{(j)} \right\rangle - \sum_{i=1}^{m} \frac{1}{N_i} \sum_{j=1}^{n} |f_i^{(j)}| \, \ell_{\phi_i}(q_i, s_{ij} N_i \lambda_j) \tag{4.13}$$

is an auxiliary function for p_0 and F, and the corresponding update scheme is given by

$$q^{(t+1)} = \mathcal{L}_\phi(q^{(t)}, F\lambda^{(t)}) \tag{4.14}$$

where $\lambda_j^{(t)}$ satisfies

$$\sum_{i=1}^{m} f_i^{(j)} \, \mathcal{L}_{\phi_i}(q_i^{(t)}, s_{ij} N_i \lambda_j) = \left\langle p_0, f^{(j)} \right\rangle \tag{4.15}$$

The idea of factoring out the signs s_{ij} is taken from Collins et al. (2001). If we take $N_i = 1$ for all i we obtain the algorithms presented in that paper. If we take $N_i = \sum_j |f_i^{(j)}|$ then we obtain algorithms analogous to the IIS algorithm of Della Pietra et al. (1997).

Proof. We verify that the function defined in (4.13) satifies the three properties of Definition 4.1. Property (1) of the definition holds since $\ell_{\phi_i}(q_i, 0) = 0$. Property (2) follows from the convexity of ℓ_{ϕ_i}. In particular, we have the inequality

$$\ell_\phi(q, F\lambda) \quad = \quad \sum_{i=1}^{m} \ell_{\phi_i}(q_i, (F\lambda)_i) \tag{4.16}$$

$$= \sum_{i=1}^{m} \ell_{\phi_i}(q_i, \sum_{j=1}^{n} s_{ij}|F_{ij}|\lambda_j) \tag{4.17}$$

$$\leq \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \frac{|F_{ij}|}{N_i} \ell_{\phi_i}(q_i, s_{ij}N_i\lambda_j) + \left(1 - \sum_{j=1}^{n} \frac{|F_{ij}|}{N_i}\right) \ell_{\phi_i}(q_i, 0) \right) \tag{4.18}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{|F_{ij}|}{N_i} \ell_{\phi_i}(q_i, s_{ij}N_i\lambda_j) \tag{4.19}$$

which together with Proposition 2.9 says that

$$D_\phi(p,q) - D_\phi(p, \mathcal{L}_\phi(q, F\lambda)) \geq \sum_{j=1}^{n} \lambda_j \left\langle p, f^{(j)} \right\rangle - \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{|F_{ij}|}{N_i} \ell_{\phi_i}(q_i, s_{ij}N_i\lambda_j) \tag{4.20}$$

Now, using Propositions 2.9 and 2.10, it can be shown that

$$\frac{\partial}{\partial v} \ell_{\phi_i}(q_i, v) = \mathcal{L}_{\phi_i}(q_i, v) \tag{4.21}$$

which shows that (4.15) is the correct update. Therefore, at a maximum λ^* of $\mathcal{A}(q, \lambda)$ we have that

$$\left\langle p_0, f^{(j)} \right\rangle = \sum_{i=1}^{m} f_i^{(j)} \mathcal{L}_{\phi_i}(q_i, s_{ij}N_i\lambda_j^*) \ \text{ for each } j \tag{4.22}$$

If $\lambda^* = 0$, then

$$\left\langle p_0, f^{(j)} \right\rangle = \sum_{i=1}^{m} f_i^{(j)} \mathcal{L}_{\phi_i}(q_i, 0) = \left\langle q, f^{(j)} \right\rangle \tag{4.23}$$

showing that Property (3) holds. Thus (4.13) defines an auxiliary function. □

While the auxiliary function (4.13) looks a bit messy, both the "E-step" and "M-step" for this type of auxiliary function are generally quite practical and easy to implement.

V. Conclusion

This paper has presented a convex duality theorem for constrained optimization using Bregman distances. The main result, Proposition 3.2, differs from results presented in the convex analysis literature in that the Bregman distance is defined on the entire essential domain, rather than only on the interior. This generality is needed in many applications. Though the assumptions A1–A5 that we make on the underlying convex function are fairly restrictive, it may well be possible to relax these assumptions to cover a broader class of examples. In particular, assumption A2, which states that the conjugate domain is all of $\mathbb{R}^m$, may not be essential in our approach.

The auxiliary function technique presented in Section 4 is a general and practical method for deriving algorithms for solving the dual problem. Although the specific auxiliary function we derive assumes the Bregman distance is separable, similar arguments can be used for non-separable

Bregman distances. The analysis given here makes clear the role of convexity, as the bounds are derived using only the properties of the underlying Legendre-Bregman conjugate.

Acknowledgments

We are grateful to Rob Schapire for encouraging us to write this paper, and for helpful questions and suggestions. We also thank Jonathan Borwein and Heinz Bauschke for comments on the paper.

References

Bauschke, H., & Borwein, J. (1997). Legendre functions and the method of random Bregman projections. *Journal of Convex Analysis*, *4*, 27–67.

Censor, Y., & Zenios, S. A. (1997). *Parallel optimization: Theory, algorithms and applications.* Oxford University Press.

Collins, M., Schapire, R., & Singer, Y. (2001). Logistic regression, AdaBoost, and Bregman distances. *Machine Learning.* In press.

Della Pietra, S., Della Pietra, V., & Lafferty, J. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*, 380–393.

Rockafellar, R. T. (1970). *Convex analysis.* Princeton University Press.