Available at
WWW.MATHEMATICSWEB.ORG
POWERED BY SCIENCE @ DIRECT•

Journal of Approximation Theory 121 (2003) 1–12

JOURNAL OF
Approximation
Theory

http://www.elsevier.com/locate/jat

Duality for Bregman projections onto translated cones and affine subspaces

Heinz H. Bauschke*

Department of Mathematics and Statistics, University of Guelph, Guelph, Ont., Canada N1G 2W1

Received 11 February 2002; revised 18 October 2002
Communicated by Frank Deutsch

Abstract

In 2001, Della Pietra, Della Pietra, and Lafferty suggested a dual characterization of the Bregman projection onto linear constraints, which has already been applied by Collins, Schapire, and Singer to boosting algorithms and maximum likelihood logistic regression. The proof provided by Della Pietra et al. is fairly complicated, and their statement features a curious nonconvex component.

In this note, the Della Pietra et al. characterization is proved differently, using the powerful framework of convex analysis. Assuming a standard constraint qualification, the proof presented here is not only much shorter and cleaner, but it also reveals the strange nonconvex component as a reformulation of a *convex* (dual) optimization problem. Furthermore, the setting is extended from an affine subspace to a translated cone, and the convex function inducing the Bregman distance is only required to be Legendre. Various remarks are made on limitations and possible extensions.

Keywords: Affine subspace; Bregman distance; Bregman projection; Convex cone; Convex duality; Legendre function; Orthogonal complement

1. Introduction

Throughout this paper, we assume that

$$X \text{ is some Euclidean space } \mathbf{R}^J, \text{ with inner product } \langle x, y \rangle = \sum_j x_j y_j,$$

*Supported by the Natural Sciences and Engineering Research Council of Canada.
E-mail address: hbauschk@uoguelph.ca.

0021-9045/02/$ - see front matter © 2002 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0021-9045(02)00040-0

and that (Definition 2.1)

$$f : X \to]-\infty, +\infty] \text{ is a convex function of Legendre type.}$$

The function f induces a so-called *Bregman distance* D_f between two points $x \in X$ and $y \in \operatorname{int}\operatorname{dom} f$, defined by

$$D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle.$$

Suppose now that R is a closed convex set in X with $R \cap \operatorname{int}\operatorname{dom} f \neq \emptyset$, and let

$$y_0 \in \operatorname{int}\operatorname{dom} f.$$

The *Bregman projection* of y_0 onto R is

$$P_R^f(y_0) = \operatorname*{arg\,min}_{r \in R} D_f(r, y_0).$$

Because f is Legendre, the argmin is a single point belonging to the interior of the domain of f (Fact 2.6).

We will assume throughout that R is a translated cone:

$$R = K + x_0, \text{for some closed convex cone } K \text{ in } X \text{ and } x_0 \in \operatorname{int}\operatorname{dom} f.$$

This is flexible enough to cover Della Pietra et al.'s setting [10], where R is a set of linear constraints.

The objective of this paper is to find equivalent, potentially more useful descriptions of $P_R^f(y_0)$.

Our main result states that $P_R^f(y_0)$ can also be found from a certain *dual projection*. Let us outline the major steps in deriving this; details will be provided in later sections. Consider the set $T = K^{\oplus} + \nabla f(y_0)$, where $K^{\oplus} := \{x^* \in X^* : \inf \langle x^*, K \rangle \geqslant 0\}$ is the *positive dual cone of K*. Denote the classical conjugate function of f by f^*. The point $\nabla f(x_0)$ belongs to $\operatorname{int}\operatorname{dom} f^*$, and it has a unique projection onto T, namely $P_T^{f^*}(\nabla f(x_0))$. Then $P_R^f(y_0)$ and $P_T^{f^*}(\nabla f(x_0))$ are linked by the nice equation

$$(\nabla f)(P_R^f(y_0)) = P_T^{f^*}(\nabla f(x_0)). \tag{1}$$

Put differently,

$$\nabla f(P_R^f(y_0)) = \operatorname*{arg\,min}_{z^* \in T \cap \operatorname{int}\operatorname{dom} f^*} D_{f^*}(z^*, \nabla f(x_0)).$$

On the other hand, for arbitrary points x^*, y^* in $\operatorname{int}\operatorname{dom} f^*$, one has an identity connecting the Bregman distances induced by f and f^*: $D_{f^*}(x^*, y^*) = D_f(\nabla f^*(y^*), \nabla f^*(x^*))$. Altogether,

$$\nabla f(P_R^f(y_0)) = \operatorname*{arg\,min}_{z^* \in T \cap \operatorname{int}\operatorname{dom} f^*} D_f(x_0, \nabla f^*(z^*));$$

equivalently,

$$P_R^f(y_0) = \operatorname*{arg\,min}_{s \in \nabla f^*(T \cap \operatorname{int}\operatorname{dom} f^*)} D_f(x_0, s) = \operatorname*{arg\,min}_{s \in S} D_f(x_0, s), \tag{2}$$

where

$$S = \nabla f^*(T \cap \operatorname{int}\operatorname{dom} f^*).$$

Unless f is the energy $x \mapsto \frac{1}{2}\|x\|^2$, the set S is generally *nonconvex*. Eq. (2) is the unusual "dual" characterization involving a nonconvex set suggested by Della Pietra et al. for the case when R is an affine subspace!

In essence, this explains how the convex characterization (1) leads to the nonconvex characterization (2).

This nonconvex characterization is crucial in the proofs of convergence results on the algorithms discussed in [9,10].

The aim of this paper is to provide new and useful characterizations of the projection $P_R^f(y_0)$, from within the powerful framework of convex analysis. We extend Della Pietra et al.'s result from affine subspaces to translated cones, and also discuss limitations and possible extensions of our approach.

The notation employed is standard; see [4,12], for instance.

The paper is organized as follows. Section 2 reviews known results that will be useful later in the paper. The conical duality is derived in Section 3. In Section 4, we specialize this duality to affine subspaces which allows some stronger results.

2. A tool box

2.1. Legendre function

The notion of a convex function of Legendre type, due to Rockafellar [12, Section 26], is key to our analysis.

Definition 2.1 (Legendre function). Suppose g is a lower semicontinuous convex proper function from X to $]-\infty, +\infty]$. Then g is *Legendre*, if g is both essentially smooth and essentially strictly convex; equivalently, g is differentiable and strictly convex on $\operatorname{int}\operatorname{dom} g \neq \emptyset$; and $\lim_{t \to 0^+} \langle \nabla g(x + t(y-x)), y - x \rangle = -\infty$, for all $x \in \operatorname{bdry}(\operatorname{dom} g)$, $y \in \operatorname{int}\operatorname{dom} g$.

The class of Legendre function is rather large and encompasses many important functions from convex optimization, see [1, Section 6]. We now give three examples, including the perhaps two most important Legendre functions—the energy and the negative entropy:

Example 2.2. Each of the following functions is Legendre:

- energy $f(x) = \sum_j \frac{1}{2}|x_j|^2$;
- negative entropy $f(x) = \sum_j (x_j \ln(x_j) - x_j)$;
- Burg entropy $f(x) = -\sum_j \ln(x_j)$.

Fact 2.3. *Suppose $g : X \to]-\infty, +\infty]$ is lower semicontinuous, convex, and proper. Then g is Legendre if and only if its conjugate function*

$$g^* : X^* \to]-\infty, +\infty] : x^* \mapsto \sup_{x \in X} (\langle x^*, x\rangle - g(x))$$

is. Moreover, the gradient map

$$\nabla g : \ \operatorname{int} \operatorname{dom} g \to \operatorname{int} \operatorname{dom} g^*$$

is a topological isomorphism with inverse mapping $(\nabla g)^{-1} = \nabla g^*$.

In particular, f^* is Legendre.

2.2. Bregman distance and projection

Definition 2.4 (Bregman distance). Suppose $g : X \to]-\infty, +\infty]$ is convex, lower semicontinuous, and proper. Let g be differentiable on $\operatorname{int} \operatorname{dom} g \neq \emptyset$. Then the *Bregman distance* is defined by

$$D_g : X \times \operatorname{int} \operatorname{dom} g \to [0, +\infty] : (x, y) \mapsto g(x) - g(y) - \langle \nabla g(y), x - y\rangle.$$

This distance-like measure was first employed by Bregman [5] in 1965. The notion was coined and further developed by Censor and Lent [7]. Bregman distances lie at the heart of numerous applications—see the many corresponding references in the recent monographs [6,8].

The following result will come handy later.

Fact 2.5. *Suppose $g : X \to]-\infty, +\infty]$ is Legendre. Then*:

(i) $(\forall x \in X)(\forall y \in \operatorname{int} \operatorname{dom} g)$ $D_g(x, y) = g(x) + g^*(\nabla g(y)) - \langle \nabla g(y), x\rangle$.
(ii) $(\forall x \in \operatorname{int} \operatorname{dom} g)(\forall y \in \operatorname{int} \operatorname{dom} g)$ $D_g(x, y) = D_{g^*}(\nabla g(y), \nabla g(x))$.

Proof. (i) [1, Proposition 3.2(i)]. (ii) [1, Theorem 3.7(v)]. □

The Bregman distance between two points induces the distance between a point and a set, which in turn prompts the notion of a projection:

Fact 2.6 (Legendre function and Bregman projection). *Suppose $g : X \to]-\infty, +\infty]$ is Legendre, C is a closed convex set in X with $C \cap \operatorname{int} \operatorname{dom} g \neq \emptyset$ and $y \in \operatorname{int} \operatorname{dom} f$. Then the approximation problem*

$$\inf_{x \in C} D_g(x, y)$$

has a unique solution denoted $P^g_C(y)$ and called the (Bregman) projection of y onto C. Moreover, $P^g_C(y)$ is contained in $\operatorname{int} \operatorname{dom} g$.

For further properties and examples, see [1,6,8].

3. Duality for a translated cone

Recall the standing assumptions:

- $f : X \to]-\infty, +\infty]$ is Legendre;
- K is a closed convex cone in X;
- $\{x_0, y_0\} \subseteq \operatorname{int}\operatorname{dom} f$;
- $R = K + x_0$, $S = \nabla f^*(T \cap \operatorname{int}\operatorname{dom} f^*)$, and $T = K^{\oplus} + \nabla f(y_0)$.

Theorem 3.1 (Duality for a translated cone). *Each of the following conditions on a point $\bar{x} \in X$ provides a characterization of the projection $P^f_R(y_0)$:*

(i) $\bar{x} \in R \cap \operatorname{int}\operatorname{dom} f$ *and* $\sup_{r \in R} \langle r - \bar{x}, \nabla f(y_0) - \nabla f(\bar{x}) \rangle \leqslant 0$.
(ii) $\bar{x} \in R \cap \operatorname{int}\operatorname{dom} f$ *and* $D_f(r, \bar{x}) + D_f(\bar{x}, y_0) \leqslant D_f(r, y_0)$, *for every* $r \in R$.
(iii) $\nabla f(\bar{x}) = P^{f^*}_T(\nabla f(x_0))$.
(iv) $\bar{x} = \arg\min_{s \in S} D_f(x_0, s)$.

Moreover,

$$D_f(x_0, y_0) = \min D_f(x_0, S) + \min D_f(R, y_0),$$

and both minima are uniquely attained at $P^f_R(y_0)$.

Proof. (i) This is [1, Proposition 3.16].

(ii) Is equivalent to (i), since

$$D_f(r, \bar{x}) + D_f(\bar{x}, y_0) - D_f(r, y_0) = \langle r - \bar{x}, \nabla f(y_0) - \nabla f(\bar{x}) \rangle.$$

We now proceed to prove the remaining conclusions. Consider the primal optimization problem

$$p := \inf_{x \in R} D(x, y_0) = \inf_{x \in X} (D_f(x, y_0) + \iota_K(x - x_0)). \tag{3}$$

Of course, we know (by Fact 2.6) that (3) has a *unique* solution

$$\bar{x} := P^f_R(y_0) \in \operatorname{int}\operatorname{dom} f. \tag{4}$$

Abbreviate $\varphi(x) := D(x, y_0)$ and $\psi(x) := \iota_K(x - x_0)$ so that the primal problem (3) becomes $p = \inf_{x \in X}(\varphi(x) + \psi(x))$. Using Fact 2.5(i), we readily verify that the conjugate functions of φ and ψ are

$$\varphi^*(x^*) = f^*(\nabla f(y_0) + x^*) - f^*(\nabla f(y_0)) \quad \text{and} \quad \psi^*(x^*) = \langle x_0, x^* \rangle + \iota^*_K(x^*).$$

Since K is a cone, the conjugate function ι^*_K is simply the indicator function $\iota_{K^{\ominus}}$, where $K^{\ominus} := \{x^* \in X^*:\ \sup\langle x^*, K \rangle \leqslant 0\} = -K^{\oplus}$ denotes the *negative dual cone* of K.

In the sense of convex optimization, the problem dual to (3) is

$$d := -\inf_{x^* \in X^*} (\varphi^*(x^*) + \psi^*(-x^*)). \tag{5}$$

Using the definition of D_{f^*} and Fact 2.5(ii), we re-write (5) as

$$\begin{aligned} d &= -\inf_{x^*\in X^*}(f^*(x^*+\nabla f(y_0)) - f^*(\nabla f(y_0)) + \iota_{K^\ominus}(-x^*) - \langle x^*, x_0\rangle) \\ &= D_{f^*}(\nabla f(y_0), \nabla f(x_0)) - \inf_{x^*\in K^\oplus} D_{f^*}(x^*+\nabla f(y_0), \nabla f(x_0)) \\ &= D_f(x_0, y_0) - \inf_{x^*\in K^\oplus} D_{f^*}(x^*+\nabla f(y_0), \nabla f(x_0)). \end{aligned} \tag{6}$$

The last infimum corresponds to finding the Bregman projection (with respect to f^*) of $\nabla f(x_0)$ onto $K^\oplus + \nabla f(y_0) = T$. Clearly, T is closed, convex, and $\nabla f(y_0) \in T \cap \operatorname{int}\operatorname{dom} f^* \neq \emptyset$. By Fact 2.6, $P_T^{f^*}(\nabla f(x_0))$ exists uniquely in $\operatorname{int}\operatorname{dom} f^*$. In particular, the dual problem (5) has a unique solution

$$\bar{x}^* := P_T^{f^*}(\nabla f(x_0)) - \nabla f(y_0). \tag{7}$$

For the pair of optimization problems ((3) and (5)), one always has *weak duality*, i.e., $p \geqslant d$. Since $0 \in K$ and $x_0 \in \operatorname{int}\operatorname{dom} f$—equivalently, $x_0 \in \operatorname{dom}\psi \cap \operatorname{int}\operatorname{dom}\varphi$—we actually have (see [4] or [12]) *strong duality* $p = d$; equivalently, using (6),

$$D_f(x_0, y_0) = \min D_f(R, y_0) + \min D_{f^*}(T, \nabla f(x_0)). \tag{8}$$

Convex duality yields even more: in fact, the primal solution $\bar{x}$ and the dual solution $\bar{x}^*$ are related via the *optimality conditions* $\bar{x}^* \in \partial\varphi(\bar{x})$ and $-\bar{x}^* \in \partial\psi(\bar{x})$. Translating this to the notation of the original problem, this becomes

$$\bar{x}^* = \nabla f(\bar{x}) - \nabla f(y_0) \quad \text{and} \quad -\bar{x}^* \in N_R(\bar{x}). \tag{9}$$

Combining (7) and the equation in (9) yields $P_T^{f^*}(\nabla f(x_0)) = \nabla f(\bar{x})$. Hence item (iii) is verified. But now (iii) and Fact 2.5(ii) yields

$$\nabla f(\bar{x}) = P_T^{f^*}(\nabla f(x_0)) = \operatorname*{arg\,min}_{z^*\in T} D_{f^*}(z^*, \nabla f(x_0)) = \operatorname*{arg\,min}_{z^*\in T} D_f(x_0, \nabla f^*(z^*)).$$

Since ∇f is a topological isomorphism (Fact 2.3), we can "change variables" and rephrase this simply as

$$\bar{x} = \operatorname*{arg\,min}_{s\in S} D_f(x_0, s).$$

This establishes item (iv) and also (use (8)!) the "Moreover" part. The entire theorem is proven. □

Remark 3.2 (Formal duality). Consider Theorem 3.1 and its notation. If we identify $\bar{x}$ with the triple $(f, K + x_0, y_0)$ and agree upon that starring such a triple amounts to computing $\nabla f(\bar{x})$, then we can concisely rephrase Theorem 3.1(iii) as

$$(f, K + x_0, y_0)^* = (f^*, K^\oplus + \nabla f(y_0), \nabla f(x_0)).$$

Consequently, $(f, K + x_0, y_0)^{**} = (f, K + x_0, y_0)$.

Remark 3.3 (Classical orthogonal setting). Suppose $f = \frac{1}{2}\|\cdot\|^2$ is the energy so that the Bregman projections reduce to the classical orthogonal projections. Theorem

3.1(iii) then states

$$P_{K+x_0}(y_0) = P_{K^{\oplus}+y_0}(x_0).$$

(Here and in Remark 4.2, Bregman projections without superscripts correspond to orthogonal projections.)

Although this identity does not appear to be known explicitly, it can be pieced together from known results on orthogonal projections: Frank Deutsch kindly pointed out that the identity follows by combining Theorem 2.7(ii) and (iv), and Theorem 5.6(2) from his recent monograph [11].

Remark 3.4 (Translated cone). Theorem 3.1 is formulated for the projection onto $R = K + x_0$, the translate of the cone K. Does our proof of Theorem 3.1 generalize to a more general closed convex nonempty set K? The answer is negative: in order to relate the dual problem to another projection, the function ι_K^* must be the indicator function of some closed convex set, say $\tilde{K}$. As the conjugate of an indicator function, $\iota_K^* = \iota_{\tilde{K}}$ is sublinear. Thus $\tilde{K}$ is a closed convex cone. Since $\iota_K^{**} = \iota_K$, this implies that K is a closed convex cone, namely the negative dual cone of $\tilde{K}$. (We note in passing that items (i) and (ii), however, are valid for every closed convex nonempty set R with $R \cap \operatorname{int} \operatorname{dom} f \neq \emptyset$.)

Remark 3.5 (Forward projection). Theorem 3.1(iv) appears quite surprising at first glance, since neither is $D_f(x_0, \cdot)$ generally a convex function nor is S a convex set. However, we have revealed this apparently nonconvex problem as a reformulation of a well-behaved convex problem.

In [2,3], we discuss Legendre functions for which the induced Bregman distance is *jointly convex* and for which the new notion of a *forward projection*—where the first argument of the Bregman distance is fixed and the second one is varied over a closed convex set—can be well defined. While smaller than the class of Legendre functions, this particular subclass does include both the energy and the negative entropy.

Remark 3.6 (Strong minimizer). By Theorem 3.1,

$$\bar{x} := \operatorname*{arg\,min}_{r \in R} D_f(r, y_0) = \operatorname*{arg\,min}_{s \in S} D_f(x_0, s).$$

We now sketch a proof of the fact that $\bar{x}$ is a *strong minimizer* for both minimization problems.

We first establish that $\bar{x}$ is a strong minimizer for $\min_{r \in R} D_f(r, y_0)$. So pick a sequence (r_n) in R with $D_f(r_n, y_0) \to D_f(\bar{x}, y_0)$. We need to show that $r_n \to \bar{x}$. By Bauschke and Borwein [1, Theorem 3.7 (iii)], $D_f(\cdot, y_0)$ is coercive, hence (r_n) is bounded. Let $\bar{r}$ be a cluster point of (r_n). Since D_f is lower semicontinuous, we have $D_f(\bar{r}, y_0) \leqslant D_f(\bar{x}, y_0)$. On the other hand, $\bar{r} \in R$, since R is closed. By uniqueness of the projection (Fact 2.6), $\bar{r} = \bar{x}$. Thus, the entire sequence converges to $\bar{x}$.

Next, we show that $\bar{x}$ is a strong minimizer for $\min_{s\in S} D_f(x_0,s)$. Fix a sequence (s_n) in S with $D_f(x_0,s_n)\to D_f(x_0,\bar{x})$. Using Fact 2.5(ii), this is equivalent to $D_{f^*}(\nabla f(s_n),\nabla f(x_0))\to D_{f^*}(\nabla f(\bar{x}),\nabla f(x_0))$. By Theorem 3.1(iii) and the previous case, $\nabla f(\bar{x})$ is a strong minimizer for the minimization problem $\min_{z^*\in T} D_{f^*}(z^*,\nabla f(x_0))$. It follows that $\nabla f(s_n)\to\nabla f(\bar{x})$. By Fact 2.3, $s_n\to\bar{x}$, as required.

Remark 3.7 ($R\cap S$ may not be a singleton). If X is the Euclidean plane and R is the translation of the nonnegative orthant $K=K^{\oplus}$, then $R\cap S$ is never a singleton. On the other hand, as we will see in Section 4, $R\cap S$ is always a singleton provided that R is an affine subspace.

4. Duality for an affine subspace

We continue to work with the standing assumptions listed at the beginning of Section 3; in addition, we assume that

- $K=L$, where L is some (closed) linear subspace of X.

Hence $R=L+x_0$ is an affine subspace, which allows a refinement of items (i) and (ii) of Theorem 3.1, as well as two new conditions:

Theorem 4.1 (Duality for an affine subspace). *Each of the following conditions on a point $\bar{x}\in X$ provides a characterization of the projection $P^f_R(y_0)$:*

(i) $\bar{x}\in R\cap \operatorname{int}\operatorname{dom} f$ *and* $\nabla f(y_0)-\nabla f(\bar{x})\in L^{\perp}$.

(ii) $\bar{x}\in R\cap \operatorname{int}\operatorname{dom} f$ *and* $D_f(r,\bar{x})+D_f(\bar{x},y_0)=D_f(r,y_0)$, *for every* $r\in R$.

(iii) $\bar{x}\in R\cap S$.

(iv) $D_f(r,s)=D_f(r,\bar{x})+D_f(\bar{x},s)$, *for all* $r\in R$, $s\in S$.

Proof.

(i) This follows from Theorem 3.1(i) since L is a linear subspace.

(ii) Analogously to the proof of Theorem 3.1(ii).

(iii) and (iv) require some preparation. Let $\bar{x}:=P^f_R(y_0)$. Pick $s\in S$. On the one hand, $\bar{x}\in R$. On the other hand, $\nabla f(s)-\nabla f(\bar{x})=(\nabla f(s)-\nabla f(y_0))+(\nabla f(y_0)-\nabla f(\bar{x}))$. The first difference lies in $L^{\perp}$ (by definition of S), and so does the second (by item (i)). Hence $\nabla f(s)-\nabla f(\bar{x})\in L^{\perp}$. Altogether, using once again the characterization in item (i),

$$(\forall s\in S)\quad P^f_R(s)=\bar{x}=P^f_R(y_0). \tag{10}$$

Next, fix $x\in R\cap S$. Since $x\in S$, (10) results in $P^f_R(x)=\bar{x}$. On the other hand, as $x\in R$, we have $x=P^f_R(x)$. Altogether, $x=\bar{x}$. Combining with Theorem 3.1(iv) yields $R\cap S=\{\bar{x}\}$. Thus item (iii) is proved.

To tackle (iv), note first that the "Moreover" part of Theorem 3.1 yields

$$D_f(x_0, y_0) = D_f(x_0, \bar{x}) + D_f(\bar{x}, y_0). \tag{11}$$

Because R is an affine subspace, we have $R = L + x_0 = L + r$, for every $r \in R$. Put differently, in Eq. (11), we can and do replace x_0 replaced by an arbitrary $r \in R$ to obtain

$$(\forall r \in R) \quad D_f(r, y_0) = D_f(r, \bar{x}) + D_f(\bar{x}, y_0). \tag{12}$$

Moreover, because of (10), we may interchange y_0 with an arbitrary $s \in S$ in (12) and conclude that

$$(\forall r \in R)(\forall s \in S) \quad D_f(r, s) = D_f(r, \bar{x}) + D_f(\bar{x}, s). \tag{13}$$

To complete the proof of item (iv), we only need to show that $\bar{x}$ is the only point in X satisfying (13). So suppose $\tilde{x}$ is such that $(\forall r \in R)(\forall s \in S)$ $D_f(r, s) = D_f(r, \tilde{x}) + D_f(\tilde{x}, s)$. Since $\bar{x} \in R \cap S$ (see item (iii)), we have $0 = D_f(\bar{x}, \bar{x}) = D_f(\bar{x}, \tilde{x}) + D_f(\tilde{x}, \bar{x})$. Hence $\tilde{x} = \bar{x}$, and the entire theorem is proved. □

Remark 4.2. One of the striking differences between Theorems 3.1 and 4.1 is that $R \cap S$ is always a singleton in the affine case. (See Remark 3.7 for a conical example where $R \cap S$ is not a singleton.) Moreover, if f is the energy, then $\bar{x} = P_R(y_0)$ can be determined in closed form as follows: $\bar{x} - x_0 \in L \Leftrightarrow P_{L^\perp}(\bar{x} - x_0) = 0 \Leftrightarrow P_{L^\perp}\bar{x} = P_{L^\perp}x_0$. Analogously, $\bar{x} - y_0 \in L^\perp \Leftrightarrow P_L\bar{x} = P_L y_0$. Altogether, $\bar{x} = P_L\bar{x} + P_{L^\perp}\bar{x} = P_L y_0 + P_{L^\perp}(x_0)$.

We now discuss Della Pietra et al.'s main result [10] in our setting. It is noteworthy that item (iv) is crucial in their analysis of an algorithmic scheme.

Corollary 4.3 (Della Pietra et al.). *Suppose $Y := \mathbb{R}^M$ and $A : X \to Y$ is linear. Assume f is also co-finite*, i.e., $\operatorname{dom} f^* = X^*$. *Let $\tilde{R} := \{x \in \operatorname{dom} f\colon Ax = Ax_0\}$ and $\tilde{S} := \nabla f^*(\nabla f(y_0) + \operatorname{ran} A^*)$. Then each of the following conditions for a point $\bar{x} \in X$ characterize $P^f_R(y_0)$:*

(i) $\bar{x} \in \tilde{R} \cap \tilde{S}$.
(ii) $(\forall r \in \tilde{R})(\forall s \in \tilde{S}) \quad D_f(r, s) = D_f(r, \bar{x}) + D_f(\bar{x}, s)$.
(iii) $\bar{x} = \operatorname*{arg\,min}_{r \in \tilde{R}} D_f(r, y_0)$.
(iv) $\bar{x} = \operatorname*{arg\,min}_{s \in \tilde{S}} D_f(x_0, s)$.

Proof. We let $L := \ker A := \{x \in X\colon Ax = 0\}$ and $R := L + x_0$. Then $\tilde{R} = R \cap \operatorname{dom} f$. Fact 2.6 now shows that the unique point $\bar{x}$ satisfying item (iii) is $P^f_R(y_0)$. Next, $L^\perp = (\ker A)^\perp = \operatorname{ran} A^*$. Hence $\tilde{S} = \nabla f^*(\nabla f(y_0) + \operatorname{ran} A^*) = \nabla f^*(\nabla f(y_0) +$

$L^\perp) = S \subseteq \operatorname{dom} f$. Item (i) now follows from Theorem 4.1(iii). Also, item (ii) is implied by Theorem 4.1(iv). Finally, Theorem 3.1(iv) results in item (iv). □

Remark 4.4 (Impossibility to extend D_f continuously). Della Pietra et al. originally considered a more general situation than the present one—they did not require any constraint qualification, i.e., neither x_0 nor y_0 is assumed to belong to the interior of the domain of f. To tackle this case, they proposed to work with D_f *extended continuously* to $\operatorname{dom} f \times \operatorname{dom} f$. Unfortunately, it is impossible to carry out this approach because of the following result, applicable in particular to the negative entropy:

If $\operatorname{dom} f \subsetneqq X$, then the lower semicontinuous hull of D_f is never continuous on $\operatorname{cl} \operatorname{dom} f \times \operatorname{cl} \operatorname{dom} f$.

Proof (Sketch). Denote the lower semicontinuous hull of D_f by $\bar{D}$. Note that $\bar{D}(x,y) \geqslant 0$, for all x, y in $\operatorname{cl} \operatorname{dom} f$. Fix $\bar{y} \in \operatorname{bdry} \operatorname{dom} f$ and let (y_n) be an arbitrary sequence in $\operatorname{int} \operatorname{dom} f$ converging to $\bar{y}$. Now $D_f(y_n, y_n) \equiv 0$, hence

$$\bar{D}(\bar{y}, \bar{y}) = 0. \tag{14}$$

On the other hand, fix $x \in \operatorname{int} \operatorname{dom} f$ and an arbitrary sequence (x_n) in $\operatorname{int} \operatorname{dom} f$ converging to x. Continuity of f on $\operatorname{int} \operatorname{dom} f$ and the proof of [1, Theorem 3.8(i)] show that $D_f(x_n, y_n) \to +\infty$. Hence

$$(\forall x \in \operatorname{int} \operatorname{dom} f) \quad \bar{D}(x, \bar{y}) = +\infty. \tag{15}$$

Altogether, (14) and (15) imply that $\bar{D}$ is not continuous.

It is desirable to obtain duality results without assuming constraint qualifications, as this occurs quite naturally in some applications. It appears, however, that neither Della Pietra et al.'s nor the present approach extends to the general setting.

Finally, we point out that Della Pietra et al. considered the closure of $\tilde{S}$ in Corollary 4.3(iv)—in our setting, this is not necessary: on the one hand, by Bauschke and Borwein [1, Theorem 3.8(i)] and since $x_0 \in \operatorname{int} \operatorname{dom} f$, we have $D_f(x_0, s_n) \to +\infty$ for every sequence (s_n) in $\operatorname{int} \operatorname{dom} f$ converging to a boundary point of $\operatorname{dom} f$. On the other hand, $(\operatorname{cl}\tilde{S}) \backslash \tilde{S} \subseteq \operatorname{bdry} \operatorname{dom} f$, because $\tilde{S} = \nabla f^*(\nabla f(y_0) + \operatorname{ran} A^*) = (\nabla f)^{-1}(\nabla f(y_0) + \operatorname{ran} A^*)$ is closed in $\operatorname{int} \operatorname{dom} f$. Altogether, $\operatorname*{arg\,min}_{s \in \tilde{S}} D_f(x_0, s) = \operatorname*{arg\,min}_{s \in \operatorname{cl}\tilde{S}} D_f(x_0, s)$.

We conclude by providing a concrete example of Remark 4.4.

Example 4.5 (Impossibility to extend D_f continuously). Let $X = \mathbb{R}$ and f be the *negative entropy* given by

$$f(x) = \begin{Bmatrix} x\ln(x) - x & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ +\infty & \text{if } x < 0. \end{Bmatrix}$$

Then $\operatorname{dom} f = [0, +\infty[$ is closed, and

$$D_f(x, y) = x\ln(x) - x\ln(y) - x + y$$

for $(x, y) \in \operatorname{dom} D_f = [0, +\infty[\times]0, +\infty[$.

Now assume to the contrary it were possible to extend D_f continuously.

For $x > 0$ and $\alpha \geqslant 1$, let $y_\alpha(x) := \exp(-1/x^\alpha)$. Then for every $\alpha \geqslant 1$ and $x > 0$, we have $(x, y_\alpha(x)) \in]0, +\infty[^2$,

$$D_f(x, y_\alpha(x)) = x\ln(x) + x^{1-\alpha} - x + y_\alpha(x), \tag{16}$$

and

$$\lim_{x\to 0^+} (x, y_\alpha(x)) = (0, 0). \tag{17}$$

By (16), $\lim_{x\to 0^+} D_f(x, y_1(x)) = 1$, while $\lim_{x\to 0^+} D_f(x, y_2(x)) = +\infty$; consequently, in view of (17), it is impossible to extend D_f continuously at $(0, 0)$. Hoping for a separately continuous extension of D_f is also fruitless: let $\tilde{D} : [0, +\infty[^2 \to [0, +\infty]$ be such that $\widetilde{D} = D_f$ on $]0, +\infty[^2$ and $\widetilde{D}(0, 0) = 0$. Fix $x > 0$. Since $\lim_{y\to 0^+} D_f(x, y) = +\infty$, separate continuity results in $\widetilde{D}(x, 0) = +\infty$. Thus $\widetilde{D}(\cdot, 0)$ cannot be continuous at 0.

Acknowledgments

I thank John Lafferty for sending me the very stimulating paper [10], and Frank Deutsch for his helpful comments on an earlier version of this manuscript.

References

[1] H.H. Bauschke, J.M. Borwein, Legendre functions and the method of random Bregman projections, J. Convex Anal. 4 (1) (1997) 27–67.

[2] H.H. Bauschke, J.M. Borwein, Joint and separate convexity of the Bregman distance, in: D. Butnariu, Y. Censor, S. Reich (Eds.), Inherently Parallel Algorithms in Feasibility and Optimization and their Applications, Vol. 8, Studies in Computational Mathematics Elsevier, Amsterdam, 2001, pp. 23–36.

[3] H.H. Bauschke, D. Noll, The method of forward projections, J. Convex Anal. 3 (2) (2002) 191–205.

[4] J.M. Borwein, A.S. Lewis, Convex Analysis and Nonlinear Optimization, Springer, Berlin, 2000.

[5] L.M. Bregman, The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, USSR Comput. Math. Math. Phys. 7 (3) (1967) 200–217.

[6] D. Butnariu, A.N. Iusem, Totally Convex Functions for Fixed Point Computation and Infinite Dimensional Optimization, Kluwer Academic Publishers, Dordrecht, 2000.

[7] Y. Censor, A. Lent, An iterative row-action method for interval convex programming, J. Optim. Theory Appl. 34 (3) (1981) 321–353.

[8] Y. Censor, S.A. Zenios, Parallel Optimization, Oxford University Press, Oxford, 1997.

[9] M. Collins, R.E. Schapire, Y. Singer, Logistic Regression, AdaBoost and Bregman distances, Machine Learning 48 (2002). Special Issue on New Methods for Model Selection and Model Combination.

[10] S. Della Pietra, V. Della Pietra, J. Lafferty, Duality and Auxiliary Functions for Bregman Distances, Technical Report CMU-CS-01-109, School of Computer Science, Carnegie Mellon University, October 2001.

[11] F. Deutsch, Best Approximation in Inner Product Spaces, Springer, Berlin, 2001.

[12] R.T. Rockafellar, Convex Analysis, Princeton University Press, Princeton, NJ, 1970.