Regular self-proximal distances are Bregman

Felipe Alvarez*, Rafael Correa†& Matthieu Marćhal‡

Abstract

Bregman distances play a key role in generalized versions of the proximal algorithm. This paper proposes a new characterization of Bregman distances in terms of their gradient and Hessian matrix. Thanks to this characterization, we obtain two results: all the so called self-proximal distances are Bregman, and all the induced proximal distances, under some regularity assumptions, are Bregman functions.

1 Introduction

Given an open convex $C \subset \mathbb{R}^n$, we consider the optimization problem

$$f_* = \inf_{x \in \overline{C}} f(x), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper, closed and convex function, and $\overline{C}$ stands for the closure of C in $\mathbb{R}^n$. Auslender and Teboulle in 2006 [2] developed a complete study of the convergence of the *Interior Proximal Algorithm* (IPA), which consists of generating a sequence $(x^k)_k$ satisfying, for all $k \in \mathbb{N}$,

$$x^{k+1} \in \text{Argmin}\{\lambda_k f(x) + d(x, x^k) \mid \ x \in \overline{C}\}, \ \ k = 0, 1, 2, \dots, \tag{2}$$

Here, $(\lambda_k)_k$ is a sequence of positive reals and $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+ \cup \{+\infty\}$, with $\mathbb{R}_+ := [0, +\infty)$, is a distance-like function which is called

*Centro de Modelamiento Matemático (CNRS UMI 2807), Departamento de Ingeniería Matemática, Universidad de Chile, Beauchef 851, Santiago, Chile. Partially supported by the Institute on Complex Engineering Systems (ICM: P-05-004-F, CONICYT: FBO16) and FONDECYT 1130176.

†Centro de Modelamiento Matemático, Departamento de Ingenería Matemática, U de Chile, Partially Supported by ICM; P-05-004-F; CONICYT: FB016, PFP03, FB0003; FONDECYT 1110019

‡Instituto de ciencias basicas, Facultad de Ingenería, Universidad Diego Portales, Ejército 441, Santiago, Chile, Partially supported by Centro de Modelamiento Matematico, U de Chile and FONDECYT 3130596.

a *proximal distance with respect to* C. The proximal distance d is required to have good properties, which force the iterates to stay in C, the interior of the feasible set given by $\overline{C}$ (cf. [2, Definition 2.1] and Definition 2.1 below).

The convergence analysis of this algorithm relies on the existence of an *induced proximal distance*, that is a function H satisfying

$$\forall a, b, c \in C \ , \ \langle c - b, \nabla_1 d(b, a) \rangle \leq H(c, a) - H(c, b), \tag{3}$$

where ∇_1 stands for the gradient with respect to the first variable of $d(\cdot, \cdot)$. An important class of proximal distances is the family of Bregman distances [3, 4, 5, 6, 9, 10]. The Bregman distances are *self-proximal*, that is, the previous inequality holds with $H = d$. Another popular class of proximal distances are the φ-divergence distances, and their regularized versions [1, 6, 7, 8].

Our work is motivated by the observation that in the literature, the only known self-proximal distances are Bregman distances. Therefore we investigated whether all the self-proximal distances are Bregman. To this end, we obtained a new characterization of the Bregman distances.

The paper is organized as follows. In section 2 we recall the definition of proximal distance and motivate the notions of induced proximal and self-proximal distances. Then we obtain that under assumptions of regularity, the induced proximal distance is uniquely determined by the proximal distance. Section 3 is devoted to Bregman distances; we obtain a characterization of these distances, and deduce that all self-proximal distances are Bregman. Moreover, under an assumption of suitable regularity, we prove that all the induced distances are Bregman.

2 Induced proximal distances

Given a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ we define the (effective) domain of f by $\operatorname{dom} f := \{x \in \mathbb{R}^n \mid f(x) < +\infty\}$. We say that f is proper if $\operatorname{dom} f \neq \emptyset$, and f is lower semi continuous (lsc) on $\mathbb{R}^n$ if for all $\lambda \in \mathbb{R}$ the set $S_\lambda = \{x \in \mathbb{R}^n \mid f(x) \leq \lambda\}$ is closed. Given $\epsilon \geq 0$, an ϵ-subgradient of f at $x \in \operatorname{dom} f$ is an element $x^* \in \mathbb{R}^n$ verifying

$$f(x') + \epsilon \geq f(x) + \langle x^*, x' - x \rangle, \quad \forall x' \in \mathbb{R}^n,$$

where $\langle \cdot, \cdot \rangle$ stands for the classical inner product of $\mathbb{R}^n$. The set of all ϵ-subgradients of f at x is the ϵ-subdifferential of f at x, it is denoted by $\partial_\epsilon f(x)$, with the convention that $\partial_\epsilon f(x) = \emptyset$ when $x \notin \operatorname{dom} f$. When $\epsilon = 0$, we simply write $\partial f(x)$.

Definition 2.1 [2, Definition 2.1] *A function* $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+ \cup \{+\infty\}$ *is called a* ***proximal distance*** *with respect to an open nonempty convex set* $C \subset \mathbb{R}^n$ *if for each* $y \in C$ *it satisfies the following properties:*

(P_1) $d(\cdot, y)$ *is proper, lsc, convex on* $\mathbb{R}^n$ *and* C^1 *on* C*;*

(P_2) $\operatorname{dom} d(\cdot, y) \subset \overline{C}$ *and* $\operatorname{dom} \partial d(\cdot, y) \subset C$*.*

(P_3) $d(\cdot, y)$ *is level bounded in* $\mathbb{R}^n$*, i.e.,* $\lim_{\|u\| \to \infty} d(u, y) = +\infty$*;*

(P_4) $d(y, y) = 0$*.*

We denote by $\mathcal{D}(C)$ *the set of functions* d *satisfying these properties.*

As we mentioned in the introduction, the IPA solves (1) by generating a sequence $(x^k)_k$ according to the iterative scheme (2), where $d \in \mathcal{D}(C)$. Due to (P_2), under suitable assumptions on f and C, the iterates generated by (2) stay in C, the interior of the constraints set $\overline{C}$, as the following proposition shows.

Proposition 2.1 [2, Proposition 2.1] *Suppose that* $f_* > -\infty$ *and* $\operatorname{dom} f \cap C \neq \emptyset$*, with* f_* *defined by* (1)*. Let* $d \in \mathcal{D}(C)$*, and for all* $v \in C$ *consider the optimization problem*

$$(P(v)) \qquad f_*(v) = \inf_{u \in \overline{C}} f(u) + d(u, v).$$

Then the optimal set of $P(v)$ *is nonempty and compact. For each* $\epsilon \geq 0$ *there exist* $u(v) \in C$*,* $g \in \partial_\epsilon f(u(v))$ *such that* $g + \nabla_1 d(u(v), v) = 0$*. For such* $u(v) \in C$*, we have* $f(u(v)) + d(u(v), v) \leq f_*(v) + \epsilon$*.*

We now turn to the definition of the *induced proximal distance.*

Definition 2.2 *We say that a function* $H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+ \cup \{+\infty\}$ *is an* ***induced proximal distance*** *to* $d \in \mathcal{D}(C)$ *if it satisfies the following properties:*

1. H *is finite-valued on* $C \times C$*.*
2. $\forall a \in C$*,* $H(a, a) = 0$
3. $\forall a, b, c \in C$*,* $\langle c - b, \nabla_1 d(b, a) \rangle \leq H(c, a) - H(c, b)$*.*

We denote by $\mathcal{F}(C)$ *the set of all pairs* (d, H) *with* $d \in \mathcal{D}(C)$ *and* H *an induced proximal distance to* d*.*

We say that d is **self-proximal** when $(d, d) \in \mathcal{F}(C)$, that is when

$$\forall a, b, c \in C, \quad \langle c - b, \nabla_1 d(b, a) \rangle \leq d(c, a) - d(c, b). \tag{4}$$

Let us motivate the introduction of the induced proximal distance H for the analysis of IPA. Let x^{k+1} be generated by the iterative scheme given by (2). By Proposition 2.1, there exists $g^{k+1} \in \partial f(x^{k+1})$ such that $\lambda_k g^{k+1} + \nabla_1 d(x^{k+1}, x^k) = 0$. If $(d, H) \in \mathcal{F}(C)$, we get

$$\begin{aligned}\lambda_k(f(x^{k+1}) - f(x)) &\leq \langle -\lambda_k g^{k+1}, x - x^{k+1}\rangle \\ &= \langle \nabla_1 d(x^{k+1}, x^k), x - x^{k+1}\rangle \\ &\leq H(x, x^k) - H(x, x^{k+1}),\end{aligned}$$

for all $x \in C$. Next, suppose that the inequality can be extended to every $x \in \overline{C}$. Set $S := \text{Argmin}_{\overline{C}} f$, which is supposed to be nonempty, and take any $x^* \in S$. Then $\lambda_k(f(x^{k+1}) - f(x^*)) \leq H(x^*, x^k) - H(x^*, x^{k+1})$. Since $f(x^{k+1}) \geq f^* = f(x^*)$, we deduce that

$$H(x^*, x^{k+1}) \leq H(x^*, x^k), \quad \forall x^* \in S.$$

Thus the sequence $(x^k)_k$ is *H-Fejer monotone* with respect to the optimal set S. This is a key property for the asymptotic convergence of $(x^k)_k$ towards an optimal solution of (1). See [2] for all details.

The previous argument motivates the introduction of the set of pairs $\mathcal{F}(\overline{C})$ defined as follows: a pair (d, H) belongs to $\mathcal{F}(\overline{C})$ if $(d, H) \in \mathcal{F}(C)$ and satisfies the following properties:

1. H is finite-valued on $\overline{C} \times C$.
2. $\forall a, b \in C$, $\forall c \in \overline{C}$, $\langle c - b, \nabla_1 d(b, a)\rangle \leq H(c, a) - H(c, b)$.
3. $\forall c \in \overline{C}$, the function $H(c, \cdot)$ is level-bounded on $\mathbb{R}^n$, i.e. for all $\lambda \in \mathbb{R}$, the set $\{y \in \mathbb{R}^n \mid H(c, y) \leq \lambda\}$ is bounded.

Our first result is the uniqueness of the induced proximal distance satisfying nice regularity assumptions.

Theorem 2.1 *Let $(d, H) \in \mathcal{F}(C)$. Suppose that for every $x \in C$, $H(x, \cdot)$ is C^1 on C. Then:*

(a) If for each $y \in C$, $d(\cdot, y)$ is C^2 on C, then

1. $\forall x, y \in C$, $\nabla_2 H(x, y) = -\nabla^2_{1,1} d(y, y)(x - y)$,

2. $H(x, y) = \int_0^1 t\langle \nabla^2_{1,1} d(z_t, z_t)(y - x), y - x\rangle dt$,

where $z_t = x + t(y - x)$.

(b) If for each $x \in C$, $\nabla_1 d(x, \cdot)$ is C^1 on C, then

3. $\forall x, y \in C$, $\nabla_2 H(x, y) = \nabla^2_{1,2} d(y, y)(x - y)$,

4. $H(x, y) = -\int_0^1 t\langle \nabla^2_{1,2} d(z_t, z_t)(y - x), y - x\rangle dt$, *where $z_t = x + t(y - x)$.*

Thus, under any of the above assumptions on d, there exists a unique induced proximal distance H to d such that for all $x \in C$, $H(x, \cdot)$ is C^1 on C.

Proof. We prove *(a)*. We suppose that for all $y \in C$, $d(\cdot, y)$ is C^2 on C and for all $x \in C$, $H(x, \cdot)$ is C^1 on C. Let $x, y \in C$, and define $\Phi : C \to \mathbb{R}$ by $\Phi(z) := H(x, y) - H(x, z) - \langle x - z, \nabla_1 d(z, y)\rangle$. Since H is an induced proximal distance to d, for all $z \in C$, we have $\Phi(z) \geq 0$. Since $d(\cdot, y) \geq 0$ on C and $d(y, y) = 0$, we have $\nabla_1 d(y, y) = 0$. This yields $\Phi(y) = 0$, then y minimizes Φ on C. We deduce that $\nabla\Phi(y) = 0$ (because C is an open set), and computing this gradient, we have, for all $x, y \in C$, $\nabla_2 H(x, y) = -\nabla^2_{1,1} d(y, y)(x - y)$. Let $x, y \in C$. Since $H(x, x) = 0$, we have:

$$\begin{aligned} H(x,y) &= \int_0^1 \langle \nabla_2 H(x, x + t(y-x)), y - x\rangle dt \\ &= -\int_0^1 \langle \nabla^2_{1,1} d(z_t, z_t)(x - z_t), y - x\rangle dt \\ &= \int_0^1 t\langle \nabla^2_{1,1} d(z_t, z_t))(y - x), y - x\rangle dt, \end{aligned}$$

where $z_t = x + t(y - x)$. Therefore H is uniquely determined by d, so there exists a unique proximal distance H to d such that for all $x \in C$, $H(x, \cdot)$ is C^1 on C. We have thus proven items 1. and 2.; items 3. and 4. are obtained by similar arguments. □

As a direct consequence of Theorem 2.1, we get the following necessary conditions on the proximal distance d to have that the corresponding induced proximal distance is given by the Euclidean norm, an interesting special case for practical computations.

Corollary 2.1 *Let $d \in \mathcal{D}(C)$ be a proximal distance such that there exists a real $\eta > 0$ satisfying $(d, (x, y) \to \eta\|y - x\|^2) \in \mathcal{F}(C)$, that is*

$$\forall a, b, c \in C,\ \langle c - b, \nabla_1 d(b, a)\rangle \leq \eta(\|c - a\|^2 - \|c - b\|^2). \tag{5}$$

The following assertions hold:

(a) if for all $y \in C$, $d(\cdot, y)$ is C^2 on C, then $\nabla^2_{1,1} d(y, y) = 2\eta I$,

(b) if for all $x \in C$, $\nabla_1 d(x, \cdot)$ is C^1 on C, then $\nabla^2_{1,2} d(y, y) = -2\eta I$.

The notation I stands for the identity matrix on $\mathbb{R}^n$.

3 Bregman distances

Let us recall the definition of Bregman distance.

Definition 3.1 *Let $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be proper, lsc and convex on $\mathbb{R}^n$, with dom $h \subset \overline{C}$ and dom $\nabla h = C$. Suppose that h is strictly convex and continuous on dom h, and C^1 on int(dom h) $= C$. Set*

$$\begin{aligned} D_h(x,y) :=& \ \ h(x) - h(y) - \langle \nabla h(y), x-y\rangle, \ x \in \mathbb{R}^n, \ \ y \in C \\ :=& \ \ +\infty \ \ \textit{otherwise.} \end{aligned} \tag{6}$$

The function D_h is the ***Bregman distance*** *with kernel h.*

It is easy to see that if the kernel h is level bounded, then the corresponding D_h is a proximal distance in the sense of Definition 2.1.

Remark 3.1 *Every Bregman distance D_h satisfies the following property, which is referred to as the* three point identity *(see [5]): $\forall a, b \in C$, $\forall c \in$ dom h*

$$D_h(c,a) = D_h(c,b) + D_h(b,a) + \langle \nabla_1 D_h(b,a), c-b\rangle. \tag{7}$$

It is straightforward that (7) implies (4) with $d = D_h$, so that every Bregman distance is self-proximal. We will show that under fairly general conditions, self-proximal distances are indeed Bregman distances. To this end, we introduce new characterizations of Bregman functions.

Remark 3.2 *When h is C^2 on C, then $\nabla^2_{1,1} D_h(x,y) = \nabla^2 h(x)$ and $\nabla_2 D_h(x,y) = -\nabla^2 h(y)(x-y)$, therefore D_h enjoys the following property:*

$$\forall x, y \in C \ , \ \nabla_2 D_h(x,y) = -\nabla^2_{1,1} D_h(y,y)(x-y). \tag{8}$$

Thus (8) is a necessary condition for a proximal distance to be Bregman. An interesting question is to know whether this is also sufficient for a proximal distance to be a Bregman distance.

Remark 3.3 *If $d = D_h$ is a Bregman distance, and $\tilde{h} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a convex function, then $D_h = D_{\tilde{h}}$ if and only if dom h = dom $\tilde{h}$ and $h - \tilde{h}$ is an affine function on dom h. Therefore, we can easily establish that for all arbitrary $y_0 \in C$, the function $\tilde{h}(x) := d(x, y_0)$ verifies the following properties: dom $\tilde{h}$ = dom h and $h - \tilde{h}$ is an affine function on dom h. Thus we have $d = D_{d(\cdot, y_0)}$. Moreover, h is strictly convex if and only if there exists an element $y \in C$ such that $d(\cdot, y)$ is strictly convex.*

Since the three point identity (7) is stronger than inequality (4), we first verify that every proximal distance satisfying the three point identity is indeed Bregman.

Proposition 3.1 *Let $d \in \mathcal{D}(C)$ be a proximal distance satisfying the three point identity, that is,*

$$\forall a, b \in C,\ \forall c \in \mathbb{R}^n, \quad d(c,a) = d(c,b) + d(b,a) + \langle c-b, \nabla_1 d(b,a)\rangle. \tag{9}$$

Then there exists a function $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ lsc, proper and convex on $\mathbb{R}^n$, C^1 on C, with dom $\nabla h = C$, $C \subset$ dom $h \subset \overline{C}$ and for all $(x,y) \in \mathbb{R}^n \times C$, $d(x,y) = D_h(x,y)$.

Proof. Let $y_0 \in C$ and $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be defined by $h(x) := d(x, y_0)$. Since d is a proximal distance, the function h is lsc, proper and convex on $\mathbb{R}^n$, C^1 on C with dom $\nabla h = C$ and $C \subset$ dom $h \subset \overline{C}$. Let $x \in \mathbb{R}^n$ and $y \in C$. Applying the equality (9), we obtain

$$d(x, y_0) = d(x,y) + d(y,y_0) + \langle x-y, \nabla_1 d(y,y_0)\rangle,$$

which implies that

$$h(x) = d(x,y) + h(y) + \langle x-y, \nabla h(y)\rangle.$$

Therefore, for all $x \in \mathbb{R}^n$ and $y \in C$, we have $d(x,y) = D_h(x,y)$. □

The following theorem gives us a necessary and sufficient condition for a function defined on $C \times C$ to be written as a Bregman distance.

Theorem 3.1 *Let a function $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+ \cup \{+\infty\}$ be such that $C \times C \subset$ dom F and for all $x \in C$, $F(x,x) = 0$. We suppose that F is C^1 on $C \times C$, and for all $y \in C$, $F(\cdot, y)$ is C^2 on C. Then there exists a function $h : C \to \mathbb{R}$ convex and C^2 on C, satisfying*

$$\begin{aligned} \forall x,y \in C, \quad F(x,y) &= h(x) - [h(y) + \langle \nabla h(y), x-y\rangle] \\ &= D_h(x,y), \end{aligned} \tag{10}$$

if and only if

$$\forall x,y \in C,\ \nabla_2 F(x,y) = -\nabla^2_{1,1} F(y,y)(x-y), \tag{11}$$

and in this case, for any $y_0 \in C$, the function $h : C \to \mathbb{R}$ defined by $h(x) = F(x, y_0)$ satisfies the equality (10).
Additionally, if for all $y \in C$, $F(\cdot, y)$ is finite-valued and continuous on $\overline{C}$, then any C^1 function $h : C \to \mathbb{R}$ satisfying the equality (10) can be extended in a continuous function $\bar{h}$ on $\overline{C}$, so that the equality (10) holds with $\bar{h}$ for all $(x,y) \in \overline{C} \times C$.

Remark 3.4 *The equality (10) mixed with the convexity of h on C tells us that $F(\cdot, y)$ must be convex for all $y \in C$, so the above theorem shows us that the first assumptions about F (namely $F(\cdot,\cdot) \geq 0$ on $C \times C$ and $F(x,x) = 0$ for all $x \in C$) mixed with the assumption (11) imply that $F(\cdot, y)$ is convex on C for all $y \in C$.*

Remark 3.5 *The second part of the theorem shows that if for all $y \in C$, the function $F(\cdot, y)$ is finite-valued and continuous on $\overline{C}$, then (11) implies the equality $F = D_{\bar{h}}$ on $\overline{C} \times C$ with $\bar{h}$ a function which is C^1 on C and continuous on $\overline{C}$. That implies the continuity of F on $\overline{C} \times C$.*

Proof of Theorem 3.1. From (8), equality (10) implies equality (11).

We now prove the converse, more precisely we prove that equality (11) implies that $F = D_{F(\cdot, y_0)}$ on $C \times C$ for any arbitrary $y_0 \in C$. We suppose that (11) holds. Let $y_1, y_2 \in C$. We show that the function $F(\cdot, y_2) - F(\cdot, y_1)$ is an affine function on C: $\forall x \in C$,

$$\begin{aligned} F(x,y_2) - F(x,y_1) &= \int_0^1 \langle \nabla_2 F(x, y_1 + t(y_2 - y_1)), y_2 - y_1 \rangle dt \\ &= -\int_0^1 \langle \nabla^2_{1,1} F(y_t, y_t)(x - y_t), y_2 - y_1 \rangle dt \text{ by (11)} \\ &= -\int_0^1 \langle x - y_t, \nabla^2_{1,1} F(y_t, y_t)(y_2 - y_1) \rangle dt, \end{aligned} \tag{12}$$

because the Hessian matrix $\nabla^2_{1,1} F(y_t, y_t)$ is a symmetric matrix with $y_t := y_1 + t(y_2 - y_1)$. The last expression is an affine function with respect to x, therefore, since the Hessian matrix of an affine function is equal to zero, we deduce that

$$\forall x, y_1, y_2 \in C\ ,\ \nabla^2_{1,1} F(x, y_1) = \nabla^2_{1,1} F(x, y_2). \tag{13}$$

We choose an element $y_0 \in C$ and define $h : C \to \mathbb{R}$ by

$$\forall x \in C\ ,\ h(x) := F(x, y_0). \tag{14}$$

By assumption, h is C^2 on C. We now define $H : C \times C \to \mathbb{R}$ by

$$\begin{aligned} \forall x, y \in C \quad H(x,y) &:= h(x) - [h(y) + \langle \nabla h(y), x - y \rangle] \\ &= D_h(x,y). \end{aligned} \tag{15}$$

We will show that $F = H$ on $C \times C$. Actually, we will show that $\nabla F = \nabla H$ on $C \times C$. We first compute ∇H:

$$\nabla_1 H(x,y) = \nabla h(x) - \nabla h(y), \qquad \nabla_2 H(x,y) = -\nabla^2 h(y)(x - y). \tag{16}$$

We now compute $\nabla F(x,y)$. For $y \in C$ and $t \in [0,1]$, we set $y_t := y_0 + t(y-y_0)$, and obtain

$$\begin{aligned}
\forall x,y \in C\ ,\ F(x,y) &= F(x,y_0) - \int_0^1 \langle x - y_t, \nabla^2_{1,1}F(y_t,y_t)(y-y_0)\rangle dt \text{ by (12)}\\
&= h(x) - \int_0^1 \langle x, \nabla^2_{1,1}F(y_t,y_t)(y-y_0)\rangle dt\\
&+ \int_0^1 \langle y_t, \nabla^2_{1,1}F(y_t,y_t)(y_-y_0)\rangle dt\\
&= h(x) - \left\langle x, \int_0^1 \nabla^2_{1,1}F(y_t,y_t)(y-y_0)dt \right\rangle\\
&+ \int_0^1 \langle y_t, \nabla^2_{1,1}F(y_t,y_t)(y-y_0)\rangle dt
\end{aligned}$$

with $\int_0^1 \nabla^2_{1,1}F(y_t,y_t)(y-y_0)dt$ a vectorial integral, that is an integral component by component. Therefore we obtain:

$$\begin{aligned}
\forall x,y \in C\ ,\ \nabla_1 F(x,y) &= \nabla h(x) - \int_0^1 \nabla^2_{1,1}F(y_t,y_t)(y-y_0)dt\\
&= \nabla h(x) - \int_0^1 \nabla^2_{1,1}F(y_t,y_0)(y-y_0)dt \text{ by (13)}\\
&= \nabla h(x) - (\nabla_1 F(y,y_0) - \nabla_1 F(y_0,y_0))\\
&= \nabla h(x) - \nabla h(y).
\end{aligned} \tag{17}$$

In the last line we use the equality $\nabla_1 F(y_0,y_0) = 0$ (this is due to $F(\cdot,y_0) \geq 0$ on C and $F(y_0,y_0) = 0$, so y_0 is a minimizer of $F(\cdot,y_0)$ on the open set C).

We now compute $\nabla_2 F(x,y)$:

$$\begin{aligned}
\forall x,y \in C\ ,\ \nabla_2 F(x,y) &= -\nabla^2_{1,1}F(y,y)(x-y) \text{ by (11)}\\
&= -\nabla^2_{1,1}F(y,y_0)(x-y) \text{ by (13)}\\
&= -\nabla^2 h(y)(x-y).
\end{aligned} \tag{18}$$

From the equalities (16), (17) and (18), we deduce that $\nabla F = \nabla H$ on $C \times C$, thus $F = H + c$ on $C \times C$, with c being a constant. Since $F(x,x) = H(x,x) = 0$, for all $x \in C$, we deduce that $c = 0$, then $F = H$ on $C \times C$. We conclude that for all $(x,y) \in C \times C$, $F(x,y) = h(x) - h(y) - \langle \nabla h(y), x-y\rangle$.

By the non-negativity of F, we have $h(x) \geq h(y) + \langle \nabla h(y), x-y\rangle$, for all $(x,y) \in C$, which implies the convexity of h on C.

We suppose that for all $y \in C$, $F(\cdot,y)$ is finite-valued and continuous on $\overline{C}$. Consider a C^1 function $h : C \to \mathbb{R}$ such that $F = D_h$ on $C \times C$.

We choose an element $y_1 \in C$ and define $\tilde{F} := F(\cdot, y_1)$. By assumption $\tilde{F}$ is finite-valued and continuous on $\overline{C}$. An easy computation gives:

$$\begin{aligned}\tilde{F}(x) &= F(x, y_1)\\ &= D_h(x, y_1)\\ &= h(x) - h(y_1) - \langle \nabla h(y_1), x - y_1\rangle.\end{aligned}$$

Therefore $\tilde{F} - h = -[h(y_1) + \langle \nabla h(y_1), \cdot - y_1\rangle]$ is an affine function on C, $\tilde{F}$ is continuous on $\overline{C}$, so we deduce that h can be extended in a continuous function on $\overline{C}$. Denoting by $\bar{h}$ its continuous extension on $\overline{C}$, we obtain that $D_{\bar{h}}$ is continuous on $\overline{C} \times C$.

Let $(x, y) \in \overline{C} \times C$. Since $F(\cdot, y) = D_{\bar{h}}(\cdot, y)$ on C and both functions are continuous on $\overline{C}$, we deduce that $F(\cdot, y) = D_{\bar{h}}(\cdot, y)$ on $\overline{C}$, thus $F(x, y) = D_{\bar{h}}(x, y)$. □

From this theorem we deduce a characterization of Bregman distances.

Proposition 3.2 *Let $d \in \mathcal{D}(C)$. We suppose that d is C^1 on $C \times C$, and for all $y \in C$, $d(\cdot, y)$ is C^2 on C. Then there exists a function $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ lsc, proper and convex on $\mathbb{R}^n$, C^2 on C, with $\mathrm{dom}\, \nabla h = C$, $\mathrm{dom}\, h \subset \overline{C}$ and*

$$\begin{aligned}\forall x, y \in C \quad d(x, y) &= h(x) - [h(y) + \langle \nabla h(y), x - y\rangle]\\ &= D_h(x, y),\end{aligned} \tag{19}$$

if and only if

$$\forall x, y \in C\ ,\ \nabla_2 d(x, y) = -\nabla^2_{1,1} d(y, y)(x - y). \tag{20}$$

In this case, if for all $y \in C$, $d(\cdot, y)$ is finite-valued and continuous on $\overline{C}$, then any function $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ lsc, proper and convex on $\mathbb{R}^n$, C^1 on C, with $\mathrm{dom}\, \nabla h = C$, $\mathrm{dom}\, h \subset \overline{C}$ and $d = D_h$ on $C \times C$ satisfies the following items:

1. *$\mathrm{dom}\, h = \overline{C}$ and h is continuous on $\overline{C}$.*
2. *$\forall x \in \mathbb{R}^n$, $\forall y \in C$, $d(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y\rangle$.*

Moreover, if there exists an element $y \in C$ such that $d(\cdot, y)$ is strictly convex on $\overline{C}$, then h is also strictly convex on $\overline{C}$.

Remark 3.6 *In the second part of the proposition, the strict convexity and the continuity of h on its domain mixed with the equality $d = D_h$ on $\mathbb{R}^n \times C$ ensure that d is a Bregman distance in the sense of Definition 3.1. Similar to Theorem 3.1, Proposition 3.2 shows that under assumption (20), if for all $y \in C$, $d(\cdot, y)$ is finite-valued and continuous on $\overline{C}$, then d is continuous on $\overline{C} \times C$.*

Proof of Proposition 3.2. Suppose that the equality (20) holds. Take an element $y_0 \in C$ and define $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ by $h(x) = d(x, y_0)$. By Theorem 3.1 we have $d = D_h$ on $C \times C$. By definition of a proximal distance, the function h is lsc, proper and convex on $\mathbb{R}^n$, $\operatorname{dom} \nabla h = C$ and $\operatorname{dom} h \subset \overline{C}$. Moreover h is C^2 on C by assumption on d.

Suppose that for all $y \in C$, $d(\cdot, y)$ is finite-valued and continuous on $\overline{C}$. Consider an arbitrary function $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ lsc and convex on $\mathbb{R}^n$, C^1 on C with $\operatorname{dom} \nabla h = C$, $\operatorname{dom} h \subset \overline{C}$ and $d = D_h$ on $C \times C$. Let $(\bar{x}, y) \in \overline{C} \times C$. Since $d(\cdot, y)$ is continuous on $\overline{C}$ and h is lsc on $\mathbb{R}^n$, we have:

$$\begin{aligned} d(\bar{x}, y) &= \lim_{x \xrightarrow{x \in C} \bar{x}} d(x, y) \\ &= \lim_{x \xrightarrow{x \in C} \bar{x}} D_h(x, y) \\ &= \lim_{x \xrightarrow{x \in C} \bar{x}} h(x) - [h(y) + \langle \nabla h(y), x - y \rangle] \\ &\geq h(\bar{x}) - [h(y) + \langle \nabla h(y), \bar{x} - y \rangle]. \end{aligned}$$

This inequality implies that $h(\bar{x}) < +\infty$, moreover $\bar{x}$ is an arbitrary element of $\overline{C}$, so h is finite-valued on $\overline{C}$. By assumption $\operatorname{dom} h \subset \overline{C}$, then $\operatorname{dom} h = \overline{C}$. We now prove that $d(\bar{x}, y) \leq h(\bar{x}) - [h(y) + \langle \nabla h(y), \bar{x} - y \rangle]$. Take an arbitrary element $x \in C$, since h is convex we have

$$\forall t \in (0, 1) \ , \ h((1 - t)\bar{x} + tx) \leq (1 - t)h(\bar{x}) + th(x). \tag{21}$$

Since C is a nonempty open convex set, $x \in C$ and $\bar{x} \in \overline{C}$, we have, for all $t \in (0, 1)$, $(1 - t)\bar{x} + tx \in C$. This implies that

$$d((1 - t)\bar{x} + tx, y) = D_h((1 - t)\bar{x} + tx, y),$$

thus

$$h((1 - t)\bar{x} + tx) = d((1 - t)\bar{x} + tx, y) + h(y) + \langle \nabla h(y), (1 - t)\bar{x} + tx - y \rangle. \tag{22}$$

Finally, combining (21) with (22), we obtain, for all $t \in (0, 1)$,

$$d((1-t)\bar{x}+tx, y)+h(y)+\langle \nabla h(y), (1-t)\bar{x}+tx-y \rangle \leq (1-t)h(\bar{x})+th(x).$$

Letting t tend to zero and using the continuity of $d(\cdot, y)$ at $\bar{x}$, we deduce that

$$d(\bar{x}, y) \leq h(\bar{x}) - [h(y) + \langle \nabla h(y), \bar{x} - y \rangle],$$

which proves that $d(\bar{x}, y) = h(\bar{x}) - [h(y) + \langle \nabla h(y), \bar{x} - y \rangle]$. We have proven that $d = D_h$ on $\overline{C} \times C$.

Let $x \in \mathbb{R}^n \setminus \overline{C}$ and $y \in C$. By definition of a proximal distance we have $d(x, y) = +\infty$, since $\operatorname{dom} h = \overline{C}$ we have $h(x) = +\infty$, so $D_h(x, y) = +\infty = d(x, y)$.

The continuity of h on $\overline{C}$ results from the equality, for an arbitrary $y_0 \in C$, $d(\cdot, y_0) = D_h(\cdot, y_0) = h - [h(y_0) + \langle \nabla h(y_0), \cdot - y_0 \rangle$ on $\overline{C}$. Since $d(\cdot, y_0)$ is continuous on $\overline{C}$, h is also continuous on $\overline{C}$.

We now suppose that there exists an element $y_1 \in C$ such that $d(\cdot, y_1)$ is strictly convex on $\overline{C}$. Since $d(x, y_1) = h(x) - [h(y_1) + \langle \nabla h(y_1), x - y_1 \rangle]$ for all $(x, y) \in \overline{C} \times C$, the function $h - d(\cdot, y_1)$ is an affine function on $\overline{C}$, thus the function h is strictly convex on $\overline{C}$ because $d(\cdot, y_1)$ is strictly convex on $\overline{C}$. □

The following corollary shows that the unique self-proximal distances satisfying a regularity assumption are the Bregman distances.

Corollary 3.1 *Let $d \in \mathcal{D}(C)$ be a self-proximal distance, that is d satisfies the inequality (4). Moreover, we suppose that d is C^1 on $C \times C$, and for all $y \in C$, $d(\cdot, y)$ is C^2 on C. Then there exists a function $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ lsc, proper and convex on $\mathbb{R}^n$, C^2 on C, with dom $\nabla h = C$, dom $h \subset \overline{C}$ and*

$$\forall x, y \in C, \quad \begin{aligned} d(x, y) &= h(x) - [h(y) + \langle \nabla h(y), x - y \rangle] \\ &= D_h(x, y). \end{aligned} \tag{23}$$

Additionally, if for all $y \in C$, $d(\cdot, y)$ is finite-valued and continuous on $\overline{C}$, then any function $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ lsc, proper and convex on $\mathbb{R}^n$, C^1 on C, with dom $\nabla h = C$, dom $h \subset \overline{C}$ and $d = D_h$ on $C \times C$ satisfies the following items:

1. $\operatorname{dom} h = \overline{C}$ *and h is continuous on $\overline{C}$.*
2. $\forall x \in \mathbb{R}^n$, $\forall y \in C$, $d(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle$.

Moreover, if there exists an element $y \in C$ such that $d(\cdot, y)$ is strictly convex on $\overline{C}$, then h is also strictly convex on $\overline{C}$.

Proof of Corollary 3.1. Applying Theorem 2.1 with $H = d$, we have

$$\forall x, y \in C \ , \ \nabla_2 d(x, y) = -\nabla^2_{1,1} d(y, y)(x - y).$$

Therefore, Corollary 3.1 is a direct consequence of Proposition 3.2. □

It is important to remark that the assumptions of regularity for the function d in the above theorems are stronger than in the definition of a proximal distance. However, the actual functions that are used in practice typically enjoy some regularity. The following theorem shows that under a stronger assumption of regularity on d and H, for any $(d, H) \in \mathcal{F}(C)$, H is a Bregman function on C.

Theorem 3.2 *Let $(d, H) \in \mathcal{F}(C)$. Suppose that d is C^4 on $C \times C$, and for all $x \in C$, $H(x, \cdot)$ is C^1 on C. Then there exists a function $h : C \to \mathbb{R}$ convex and C^2 on C, with*

$$\forall x, y \in C, \quad H(x, y) = h(x) - [h(y) + \langle \nabla h(y), x - y \rangle] = D_h(x, y). \tag{24}$$

Additionally, if for all $y \in C$, $H(\cdot, y)$ is finite-valued and continuous on $\overline{C}$, then any C^1 function $h : C \to \mathbb{R}$ satisfying the equality (24) can be extended in a continuous function $\bar{h}$ on $\overline{C}$, so that the equality (24) holds with $\bar{h}$ for all $(x, y) \in \overline{C} \times C$.

Remark 3.7 *We observe that according to Theorem 3.2, if d is C^4 on $C \times C$ and for all $x \in C$, $H(x, \cdot)$ is C^1 on C, then for all $y \in C$, $H(\cdot, y)$ is convex on C, while the convexity of $H(\cdot, y)$ was not needed in the definition of the induced proximal distance.*

Proof of Theorem 3.2. From Theorem 2.1 we have, for all $x, y \in C$,

$$\nabla_2 H(x, y) = -\nabla^2_{1,1} d(y, y)(x - y). \tag{25}$$

In order to apply Theorem 3.1, we show that for all $y \in C$,

$$\nabla^2_{1,1} d(y, y) = \nabla^2_{1,1} H(y, y).$$

We start with the following equality, proven in Proposition 2.1:

$$H(x, y) = \int_0^1 t \langle \nabla^2_{1,1} d(x + t(y - x), x + t(y - x))(y - x), y - x \rangle dt.$$

Let $y \in C$, and define $G_y : C \times [0, 1] \to \mathbb{R}$ by

$$G_y(x, t) = t \langle \nabla^2_{1,1} d(x + t(y - x), x + t(y - x))(y - x), y - x \rangle.$$

We see that

$$H(x, y) = \int_0^1 G_y(x, t) dt.$$

Since d is C^4 on $C \times C$, the function G_y is C^2 on $C \times [0, 1]$, and $[0, 1]$ is a compact set, therefore, by theorem of derivation of a parametric integral, $H(\cdot, y)$ is C^2 on C and we have

$$\nabla^2_{1,1} H(x, y) = \int_0^1 \nabla^2_{1,1} G_y(x, t) dt,$$

this integral of matrix is component by component.

We need now to compute the quantity $\nabla^2_{1,1} G_y(y,t)$. Define the function $F_y : C \times [0,1] \to \mathbb{R}^n$ by

$$F_y(x,t) = \nabla^2_{1,1} d(x + t(y-x), x + t(y-x))(x-y).$$

Since we have $G_y(x,t) = t\langle F_y(x,t), x-y\rangle$, we deduce that

$$\nabla_1 G_y(x,t) = tD_1F_y(x,t)(x-y) + tF_y(x,t).$$

Therefore we have

$$\nabla^2_{1,1} G_y(x,t) = tD^2_{1,1}F_y(x,t)(x-y) + 2tD_1F_y(x,t).$$

Then $\nabla^2_{1,1} G_y(y,t) = 2tD_1F_y(y,t)$. The last step consists of computing $D_1F_y(y,t)$. We fix an arbitrary $t \in [0,1]$. By continuity, we have

$$\lim_{x\to y} \nabla^2_{1,1} d(x + t(y-x), x + t(y-x)) = \nabla^2_{1,1} d(y,y),$$

thus, there exists a real $r > 0$ and a function $\varepsilon_{t,y} : B(0,r) \to \mathbb{R}^{n\times n}$ satisfying, for all $x \in B(y,r)$,

$$\nabla^2_{1,1} d(x + t(y-x), x + t(y-x)) = \nabla^2_{1,1} d(y,y) + \varepsilon_{t,y}(x-y)$$

with $\varepsilon_{t,y}(s) \to 0$ whenever $\|s\| \to 0$. It ensues that

$$\begin{aligned}
F_y(x,t) &= (\nabla^2_{1,1} d(y,y) + \varepsilon_{t,y}(x-y))(x-y) \\
&= \nabla^2_{1,1} d(y,y)(x-y) + \varepsilon_{t,y}(x-y)(x-y) \\
&= F_y(y,t) + \nabla^2_{1,1} d(y,y)(x-y) + \varepsilon_{t,y}(x-y)(x-y) \text{ since } F_y(y,t) = 0.
\end{aligned}$$

Therefore $D_1F_y(y,t) = \nabla^2_{1,1} d(y,y)$ then $\nabla^2_{1,1} G_y(y,t) = 2t\nabla^2_{1,1} d(y,y)$, which gives

$$\begin{aligned}
\nabla^2_{1,1} H(y,y) &= \int_0^1 \nabla^2_{1,1} G_y(y,t)dt \\
&= \int_0^1 2t\nabla^2_{1,1} d(y,y)dt = \nabla^2_{1,1} d(y,y).
\end{aligned}$$

This and (25) tell us that $\forall x, y \in C$, $\nabla_2 H(x,y) = -\nabla^2_{1,1} H(y,y)(x-y)$. Thus we can conclude by Theorem 3.1. Indeed, by Theorem 3.1, there exists a function $h : C \to \mathbb{R}$ convex and C^2 on C, with

$$\begin{aligned}
\forall x, y \in C \quad d(x,y) &= h(x) - [h(y) + \langle \nabla h(y), x-y\rangle] \\
&:= D_h(x,y).
\end{aligned}$$

Still by Theorem 3.1, if $\overline{C} \times C \subset \operatorname{dom} H$ and $H(\cdot, y)$ is continuous on $\overline{C}$ for all $y \in C$, then there exists a continuous extension of h satisfying the equality on $\overline{C} \times C$. □

It is worth noticing that Corollary 3.1 is not a consequence of Theorem 3.2, because Theorem 3.2 needs the function to be C^4, while Corollary 3.1 only needs d to be C^2.

References

[1] Auslender A, Teboulle M (2005). Interior projection-like methods for monotone variational inequalities. *Mathematical programming*, 104(1):39–68.

[2] Auslender A, Teboulle M (2006) Interior gradient and proximal methods for convex and conic optimization. *SIAM J. Optim.*, 16:697–725.

[3] Bauschke H, Borwein J, Combettes P (2003) Bregman monotone optimization algorithms. *SIAM J. Control Optim.*, 42:596–636.

[4] Burachick R, Iusem A (1998) A generalized proximal point algorithm for the variational inequality problem in a Hilbert Space. *SIAM J. Optim.*, 8:197–216.

[5] Chen G, Teboulle M (1993) Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM J. Optimization*, 3:538–543.

[6] Iusem A (1995) Some properties of generalized proximal points methods for quadratic and linear programming. *Journal of Optimization Theory and Applications*, 85:593–612.

[7] Iusem A, Teboulle M (1995). Convergence rate analysis of non-quadratic proximal methods for convex and linear programming. *Mathematics of Operations Research*, 20(3):657–677.

[8] Iusem A, Svaiter B, Teboulle M (1994). Entropy-like proximal methods in convex programming. *Mathematics of Operations Research*, 19(4):790–814.

[9] Kiwiel K (1997) Proximal minimization methods with generalized Bregman functions. *SIAM J. Control Optim.* 35:1142–1168.

[10] Solodov M, Svaiter B (2000) An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions. *Mathematics of Operations Research* 25:214–230.

[11] Teboulle M (1997). Convergence of proximal-like algorithms. *SIAM Journal on Optimization*, 7(4):1069–1083.

[12] Xu H, Ori R (2001). An implicit iteration process for nonexpansive mappings. *Numerical functional analysis and optimization*, 22(5):767–773.

The authors are grateful to the referee by numerous observatipns that significantly improved this work.