SIAM J. OPTIM.
Vol. 16, No. 3, pp. 697–725

INTERIOR GRADIENT AND PROXIMAL METHODS FOR CONVEX AND CONIC OPTIMIZATION*

ALFRED AUSLENDER† AND MARC TEBOULLE‡

Abstract. Interior gradient (subgradient) and proximal methods for convex constrained minimization have been much studied, in particular for optimization problems over the nonnegative octant. These methods are using non-Euclidean projections and proximal distance functions to exploit the geometry of the constraints. In this paper, we identify a simple mechanism that allows us to derive global convergence results of the produced iterates as well as improved global rates of convergence estimates for a wide class of such methods, and with more general convex constraints. Our results are illustrated with many applications and examples, including some new explicit and simple algorithms for conic optimization problems. In particular, we derive a class of interior gradient algorithms which exhibits an $O(k^{-2})$ global convergence rate estimate.

Key words. convex optimization, interior gradient/subgradient algorithms, proximal distances, conic optimization, convergence and efficiency

AMS subject classifications. 90C25, 90C30, 90C22

DOI. 10.1137/S1052623403427823

1. Introduction.

Consider the following convex minimization problem:

$$\text{(P)} \qquad f_* = \inf\{f(x) \mid x \in \overline{C}\},$$

where $\overline{C}$ denotes the closure of C, a nonempty convex open set in $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper, lower semicontinuous (lsc) convex function. In this paper we study two closely related iterative schemes for solving (P). The first one is proximal based. Given some proximity measure d, it consists of generating a sequence $\{x^k\}$ via the iteration

$$x^k \in \operatorname{argmin}\{\lambda_k f(x) + d(x, x^{k-1}) \mid x \in \overline{C}\}, \quad k = 1, 2, \dots \quad (\lambda_k > 0). \tag{1.1}$$

The second iterative scheme is subgradient based (or explicit proximal) and produces a sequence $\{x^k\}$ via

$$x^k \in \operatorname{argmin}\{\lambda_k \langle g^{k-1}, x\rangle + d(x, x^{k-1}) \mid x \in \overline{C}\}, \quad k = 1, 2, \dots, \tag{1.2}$$

where $\langle \cdot, \cdot \rangle$ is an inner product on $\mathbb{R}^n$ and g^{k-1} is a subgradient of the function f at the point x^{k-1}. With the choice $d(x, y) = 2^{-1}\|x - y\|^2$, one recovers the proximal algorithm (PA) (see, e.g., Martinet [31] and Rockafellar [40]) and the projected subgradient method (see, e.g., [41]), respectively. In that case, the sequence $\{x^k\}$ produced by either one of the above algorithms does not necessarily belong to C. In this paper, the proximal term $d(x, y)$ will play the role of a distance-like function satisfying certain desirable properties (see section 2), which will force the iterates of the produced sequence to stay in C, and thus automatically eliminate the constraints.

[*] Received by the editors May 13, 2003; accepted for publication (in revised form) June 4, 2005; published electronically January 6, 2006.
http://www.siam.org/journals/siopt/16-3/42782.html

[†] Département de Mathématiques, University Lyon I, Lyon, France (aauslen@cegetel.net).

[‡] School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel (teboulle@math.tau.ac.il). This author's work was partially supported by the United States–Israel Binational Science Foundation, BSF grant 2002-2010.

1.1. Motivation and related works. The idea of replacing the quadratic proximal term by a proximal function $d(x, y)$ has been pursued in the literature in several works. In the context of proximal-based methods of the form (1.1), two popular choices for d include either a Bregman distance (see, e.g., [15, 16, 19, 29]) or a φ-divergence distance (see, e.g., [21, 43, 44]). More recent works have also proposed proximal methods based on second order homogeneous kernels; see, e.g., [4, 5, 10, 42, 45]. These works have concentrated on the ground set $\overline{C}$ being polyhedral and in particular when $\overline{C}$ is the nonnegative octant in $\mathbb{R}^n$. For semidefinite programming problems, two particular proximal distances were proposed by Doljansky and Teboulle [18]. Furthermore, applications of these algorithms to the dual of convex programs, leading to smooth Lagrangian multiplier methods as well as extension to variational inequalities over polyhedral constraints, have also been developed in many studies, e.g., [3, 27, 28, 36, 37]. More recent applications include continuous time models of proximal-based methods; see, e.g., [1, 2, 13] and references therein.

In the context of explicit proximal methods, namely subgradient projection-type algorithms of the form (1.2), a recent paper of Ben-Tal, Margalit, and Nemirovski [9] has shown that an algorithm based on the mirror descent algorithm of Nemirovski and Yudin introduced in [32] can be used to solve efficiently convex minimization problems over the unit simplex with millions of variables. In a more recent study, Beck and Teboulle [8] have shown that the mirror descent method can be viewed as a subgradient projection algorithm based on a Bregman distance and have proposed a specific variant for convex minimization over the unit simplex. Other interior gradient schemes can be found, for example, in [25, 26] and references therein. These two works study multiplicative interior gradient-type schemes for minimizing a continuously differentiable function over the nonnegative octant under various assumptions, the former being a scheme suggested by [21], and the latter being based on the φ-divergence distance. The revived interest in such gradient-type methods relies mainly on the following facts. They require only first order information (e.g., function and subgradient evaluation at each step), they often lead to simple iterative schemes for particular types of constraints (e.g., by picking the appropriate proximal distance), and they exhibit a nearly dimension independent computational complexity in terms of the problem's dimension; see, e.g., [9, 8]. One main disadvantage of gradient-based methods is that they often share a slow convergence rate for producing high accuracy solutions, typically an $O(k^{-1})$ global convergence rate estimate for function values, where k is the iteration counter; see, e.g., [32, 41]. In comparison, the theoretically more efficient polynomial interior point methods (IPM) can achieve high accuracy but require second order derivative information and an increase in computational effort that depends on and grows polynomially in the design dimension n of the problem. Thus, gradient-based methods can be a suitable practical alternative for solving large-scale applications where high accuracy is not needed, and where the IPM are not computationally tractable; see [9] and references therein.

1.2. Main contributions and summary of results. Motivated by all these works, this paper focuses on the following three main theoretical goals: (a) to uncover the main tools needed to analyze and design interior gradient and proximal methods, (b) to establish convergence and global efficiency estimates for the basic representative schemes of these methods, and (c) to devise some new methods with improved complexity. To achieve these goals, we first develop a general and simple principle, also capable of handling more general constraints than the one alluded to above. This is achieved by identifying the common mechanism underlying the analysis of interior proximal methods. This is developed in section 2, where we derive general results on the convergence of the sequence produced by proximal-type methods and establish global rates of convergence estimates in terms of function values. This development allows us to recover some of the well-known variants of such methods, and to derive and analyze new schemes. This is illustrated in section 3, which includes many examples and applications. In section 4, we continue along this line of analysis, to develop a simple and general framework for the interior subgradient/gradient methods, akin to the ones given in [9, 8]. The interior gradient methods we propose include both fixed and Armijo–Goldstein stepsize rules. Applications of these results to conic optimization and to convex minimization over the unit simplex are given. In particular we propose, for the first time to our knowledge, new explicit interior gradient methods for semidefinite and second order conic programs. Further motivated by the discussion and references outlined above on the potential usefulness of interior gradient methods, it is natural to ask if one can devise simple interior gradient algorithms with an improved computational complexity. Building on our previous results, and inspired by the work of Nesterov [34], we answer this question positively in the last section for one of the class of interior gradient algorithms discussed in section 4. The scheme we propose naturally extends the optimal classical gradient scheme given in [34], and it leads to a class of interior gradient algorithms for solving conic problems which exhibits the faster global convergence rate estimate $O(k^{-2})$.

1.3. Notation.

We adopt the standard notation of convex analysis [39]. For a proper convex and lsc function $F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, its effective domain is defined by $\operatorname{dom} F = \{x \mid F(x) < +\infty\}$, and for all $\epsilon \geq 0$ its ϵ-subdifferential at x is defined by $\partial_\epsilon F(x) = \{g \in \mathbb{R}^n \mid \forall z \in \mathbb{R}^n,\ F(z) + \epsilon \geq F(x) + \langle g, z - x\rangle\}$, which coincides with the usual subdifferential $\partial F \equiv \partial_0 F$ whenever $\epsilon = 0$. We set $\operatorname{dom} \partial F = \{x \in \mathbb{R}^n \mid \partial F(x) \neq \emptyset\}$. For any closed convex set $S \subset \mathbb{R}^n$, δ_S denotes the indicator function of S, $\operatorname{ri} S$ its relative interior, and $N_S(x) = \partial\delta_S(x) = \{\nu \in \mathbb{R}^n \mid \langle \nu, z - x\rangle \leq 0\ \forall z \in S\}$ the normal cone to S at $x \in S$. The set of n-vectors with nonnegative (positive) components is denoted by $\mathbb{R}^n_+$ ($\mathbb{R}^n_{++}$).

2. A general framework for interior proximal methods.

Let C be a nonempty convex open set in $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ a proper, lsc, and convex function. Consider the optimization problem

$$\text{(P)} \qquad f_* = \inf\{f(x) \mid x \in \overline{C}\},$$

where $\overline{C}$ denotes the closure of C. Unless otherwise specified, throughout this paper we make the following standing assumptions on (P):

(a) $\operatorname{dom} f \cap C \neq \emptyset$,

(b) $-\infty < f_*$.

We study the behavior of the following basic proximal iterative scheme to solve (P):

$$x^k \in \operatorname{argmin}\{\lambda_k f(x) + d(x, x^{k-1}) \mid x \in \overline{C}\}, \quad k = 1, 2, \ldots \ (\lambda_k > 0),$$

where d is some proximal distance. Our approach is motivated by and patterned after many of the studies mentioned in the introduction, and our objective is to develop a general framework to analyze the convergence of the resulting methods under various settings. Given the optimization problem (P), essentially the basic ingredients needed to achieve the aforementioned goals are

- to pick an appropriate proximal distance d which allows us to eliminate the constraints,

- given d, to find an induced proximal distance H, which will control the behavior of the resulting method.

We begin by defining an appropriate proximal distance d for problem (P).

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, and C^1 on C;*

(P_2) *$\operatorname{dom} d(\cdot, y) \subset \overline{C}$ and $\operatorname{dom} \partial_1 d(\cdot, y) = C$, where $\partial_1 d(\cdot, y)$ denotes the subgradient map of the function $d(\cdot, y)$ with respect to the first variable;*

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

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

We denote by $\mathcal{D}(C)$ the family of functions d satisfying Definition 2.1. Property (P_1) is needed to preserve convexity of $d(\cdot, y)$, (P_2) will force the iterate x^k to stay in C, and (P_3) is used to guarantee the existence of such an iterate. For each $y \in C$, let $\nabla_1 d(\cdot, y)$ denote the gradient map of the function $d(\cdot, y)$ with respect to the first variable. Note that by definition $d(\cdot, \cdot) \geq 0$, and from (P_4) the global minimum of $d(\cdot, y)$ is obtained at y, which shows that $\nabla_1 d(y, y) = 0$.

Proposition 2.1. *Let $d \in \mathcal{D}(C)$, and for all $y \in C$ consider the optimization problem*

$$P(y) \qquad f_*(y) = \inf\{f(u) + d(u, y) \mid u \in \mathbb{R}^n\}.$$

Then the optimal set $S(y)$ of $P(y)$ is nonempty and compact, and for each $\epsilon \geq 0$ there exist $u(y) \in C$, $g \in \partial_\epsilon f(u(y))$ such that

$$g + \nabla_1 d(u(y), y) = 0, \tag{2.1}$$

where $\partial_\epsilon f(u(y))$ denotes the ϵ-subdifferential of f at $u(y)$. For such a $u(y) \in C$ we have

$$f(u(y)) + d(u(y), y) \leq f_*(y) + \epsilon. \tag{2.2}$$

Proof. We set $t(u) = f(u) + d(u, y) + \delta_{\overline{C}}(u)$. Then by (P_2) we have $f_*(y) = \inf\{t(u) \mid u \in \mathbb{R}^n\}$. Furthermore, since f_* is finite, it follows by (P_3) that $t(\cdot)$ is level bounded. Therefore with $t(\cdot)$ being a proper, lsc convex function, it follows that $S(y)$ is nonempty and compact. From the optimality conditions, for each $u(y) \in S(y)$ we have $0 \in \partial t(u(y))$. Now, since $\operatorname{dom} f \cap C \neq \emptyset$ and C is open, we can apply [39, Theorem 23.8] so that

$$\partial t(u) = \partial f(u) + \nabla_1 d(u, y) + N_{\overline{C}}(u) \quad \forall u.$$

Since $\operatorname{dom} \partial_1 d(\cdot, y) = C$, it follows that $u(y) \in C$, and hence $N_{\overline{C}}(u(y)) = \{0\}$, and (2.1) holds for $\epsilon = 0$ with $g \in \partial f(u(y))$. For $\epsilon > 0$, (2.1) holds for such a pair $(u(y), g)$ since $\partial f(u(y)) \subset \partial_\epsilon f(u(y))$, and thus the first part of the proposition is proved. Finally, since for each $y \in C$ the function $d(\cdot, y)$ is convex, and since $g \in \partial_\epsilon f(u(y))$, we have

$$f(u) + d(u, y) \geq f(u(y)) + d(u(y), y) + \langle g + \nabla_1 d(u(y), y), u - u(y)\rangle - \epsilon$$

so that $f_*(y) = \inf\{f(u) + d(u, y) \mid u \in \overline{C}\} \geq f(u(y)) + d(u(y), y) - \epsilon$. □

Thanks to the above proposition, the following basic algorithm is well defined.

Interior proximal algorithm (IPA). Given $d \in \mathcal{D}(C)$, start with a point $x^0 \in C$, and for $k = 1, 2, \ldots$ with $\lambda_k > 0$, $\epsilon_k \geq 0$, generate a sequence

$$\{x^k\} \in C \text{ with } g^k \in \partial_{\epsilon_k} f(x^k) \tag{2.3}$$

such that

$$\lambda_k g^k + \nabla_1 d(x^k, x^{k-1}) = 0. \tag{2.4}$$

The IPA can be viewed as an approximate interior proximal method when $\epsilon_k > 0$ $\forall k \in \mathbb{N}$ (the set of natural numbers), which becomes exact for the special case $\epsilon_k = 0$ $\forall k \in \mathbb{N}$.

The next step is to associate with each given $d \in \mathcal{D}(C)$ a corresponding proximal distance satisfying some desirable properties needed to analyze the IPA.

Definition 2.2. *Given $C \subset \mathbb{R}^n$, open and convex, and $d \in \mathcal{D}(C)$, a function $H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+ \cup \{+\infty\}$ is called the induced proximal distance to d if H is finite valued on $C \times C$ and for each $a, b \in C$ satisfies*

$$H(a, a) = 0, \tag{2.5}$$

$$\langle c - b, \nabla_1 d(b, a)\rangle \leq H(c, a) - H(c, b) \quad \forall c \in C. \tag{2.6}$$

We write $(d, H) \in \mathcal{F}(C)$ to quantify the triple $[C, d, H]$ that satisfies the premises of Definition 2.2.

Likewise, we will write $(d, H) \in \mathcal{F}(\overline{C})$ for the triple $[\overline{C}, d, H]$ whenever there exists H which is finite valued on $\overline{C} \times C$, satisfies (2.5)–(2.6) for any $c \in \overline{C}$, and is such that $\forall c \in \overline{C}$ one has $H(c, \cdot)$ level bounded on C. Clearly, one has $\mathcal{F}(\overline{C}) \subset \mathcal{F}(C)$.

The motivation behind such a construction is not as mysterious as it might look at first sight. Indeed, for the moment, notice that the classical PA, which corresponds to the special case $C = \overline{C} = \mathbb{R}^n$, $d(x, y) = 2^{-1}\|x - y\|^2$ and the induced proximal distance H being exactly d, clearly satisfies (2.6), thanks to the well-known identity

$$\|z - x\|^2 = \|z - y\|^2 + \|y - x\|^2 + 2\langle z - y, y - x\rangle.$$

IPA with $d \equiv H$ will be called *self-proximal.* Several useful examples of more general self-proximal methods for various classes of constraint sets C will be given in the next section.

As we shall see below, the requested properties for the function H associated with d naturally emerge from the analysis of the classical PA as given in [24] and later extended for various specific classes of IPA in [16, 44, 5]. Building on these works, we can already easily obtain global rates of convergence estimates as well as convergence in limit points of the produced sequence by IPA. To derive the global convergence of the sequence $\{x^k\}$ to an optimal solution of (P), additional assumptions on the induced proximal distance H, akin to the properties of norms, will be required.

Before giving our convergence results, we recall the following well-known properties on nonnegative sequences, which will be useful to us throughout this work.

Lemma 2.1 (see [35]). *Let $\{v_k\}$, $\{\gamma_k\}$, and $\{\beta_k\}$ be nonnegative sequences of real numbers satisfying $v_{k+1} \leq (1+\gamma_k)v_k + \beta_k$ and such that $\sum_{k=1}^{\infty} \beta_k < \infty$, $\sum_{k=1}^{\infty} \gamma_k < \infty$. Then, the sequence $\{v_k\}$ converges.*

Lemma 2.2 (see [35]). *Let $\{\lambda_k\}$ be a sequence of positive numbers, $\{a_k\}$ a sequence of real numbers, and $b_n := \sigma_n^{-1} \sum_{k=1}^{n} \lambda_k a_k$, where $\sigma_n = \sum_{k=1}^{n} \lambda_k$. If $\sigma_n \to \infty$, one has*

(i) $\liminf a_n \leq \liminf b_n \leq \limsup b_n \leq \limsup a_n$,
(ii) $\lim b_n = a$ *whenever* $\lim a_n = a$.

THEOREM 2.1. *Let* $(d, H) \in \mathcal{F}(C)$ *and let* $\{x^k\}$ *be the sequence generated by IPA. Set* $\sigma_n = \sum_{k=1}^n \lambda_k$. *Then the following hold:*

(i) $f(x^n) - f(x) \leq \sigma_n^{-1} H(x, x^0) + \sigma_n^{-1} \sum_{k=1}^n \sigma_k \epsilon_k \ \forall x \in C$.

(ii) *If* $\lim_{n\to\infty} \sigma_n = +\infty$ *and* $\epsilon_k \to 0$, *then* $\liminf_{n\to\infty} f(x^n) = f_*$ *and the sequence* $\{f(x^k)\}$ *converges to* f_* *whenever* $\sum_{k=1}^\infty \epsilon_k < \infty$.

(iii) *Furthermore, suppose the optimal set* X_* *of problem* (P) *is nonempty, and consider the following cases:*

(a) X_* *is bounded,*

(b) $\sum_{k=1}^\infty \lambda_k \epsilon_k < \infty$ *and* $(d, H) \in \mathcal{F}(\overline{C})$.

Then, under either (a) *or* (b), *the sequence* $\{x^k\}$ *is bounded with all its limit points in* X_*.

Proof. (i) From (2.4), since $g^k \in \partial_{\epsilon_k} f(x^k)$ we have

$$\lambda_k(f(x^k) - f(x)) \leq \langle x - x^k, \nabla_1 d(x^k, x^{k-1})\rangle + \lambda_k \epsilon_k \quad \forall x \in C. \tag{2.7}$$

Using (2.6) at the points $c = x$, $a = x^{k-1}$, $b = x^k$, the above inequality implies that

$$\lambda_k(f(x^k) - f(x)) \leq H(x, x^{k-1}) - H(x, x^k) + \lambda_k \epsilon_k \quad \forall x \in C. \tag{2.8}$$

Summing over $k = 1, \dots, n$ we obtain

$$-\sigma_n f(x) + \sum_{k=1}^n \lambda_k f(x^k) \leq H(x, x^0) - H(x, x^n) + \sum_{k=1}^n \lambda_k \epsilon_k. \tag{2.9}$$

Now setting $x = x^{k-1}$ in (2.8), we obtain

$$f(x^k) - f(x^{k-1}) \leq \epsilon_k. \tag{2.10}$$

Multiplying the latter inequality by σ_{k-1} (with $\sigma_0 \equiv 0$) and summing over $k = 1, \dots, n$, we obtain, after some algebra,

$$\sigma_n f(x^n) - \sum_{k=1}^n \lambda_k f(x^k) \leq \sum_{k=1}^n \sigma_{k-1} \epsilon_k.$$

Adding this inequality to (2.9) and recalling that $\lambda_k + \sigma_{k-1} = \sigma_k$, it follows that

$$f(x^n) - f(x) \leq \sigma_n^{-1}[H(x, x^0) - H(x, x^n)] + \sigma_n^{-1} \sum_{k=1}^n \sigma_k \epsilon_k \quad \forall x \in C, \tag{2.11}$$

proving (i), since $H(\cdot, \cdot) \geq 0$.

(ii) If $\sigma_n \to +\infty$ and $\epsilon_k \to 0$, then dividing (2.9) by σ_n and invoking Lemma 2.2(i), we obtain from (2.9) that $\liminf_{n\to\infty} f(x^n) \leq \inf\{f(x) \mid x \in C\}$, which together with $f(x^n) \geq \inf\{f(x) \mid x \in \overline{C}\}$ implies that $\liminf_{n\to\infty} f(x^n) = \inf\{f(x) \mid x \in \overline{C}\} = f_*$. From (2.10) we have

$$0 \leq f(x^k) - f_* \leq f(x^{k-1}) - f_* + \epsilon_k.$$

Then using Lemma 2.1 it follows that the sequence $\{f(x^k)\}$ converges to f_* whenever $\sum_{k=1}^\infty \epsilon_k < \infty$.

(iii) Case (a): If X_* is bounded, then f is level bounded over $\overline{C}$, and since the sequence $\{f(x^k)\}$ converges to f_*, it follows that the sequence $\{x^k\}$ is bounded. Since f is lsc, passing to the limit, and recalling that $\{x^k\} \subset C$, it follows that each limit point is an optimal solution.

Case (b): Here, we suppose that $\sum_{k=1}^{\infty} \lambda_k \epsilon_k < \infty$ and that $(d, H) \in \mathcal{F}(\overline{C})$. Then (2.8) holds for each $x \in \overline{C}$, and in particular for $x \in X_*$, so that

$$H(x, x^k) \leq H(x, x^{k-1}) + \lambda_k \epsilon_k \quad \forall x \in X_*. \tag{2.12}$$

Summing over $k = 1, \ldots, n$, we obtain

$$H(x, x^n) \leq H(x, x^0) + \sum_{k=1}^{\infty} \lambda_k \epsilon_k.$$

But, since in this case $H(x, \cdot)$ is level bounded, the last inequality implies that the sequence $\{x^k\}$ is bounded, and thus as in Case (a) it follows that all its limit points are in X_*. □

An immediate byproduct of the above analysis yields the following global rate of convergence estimate for the exact version of IPA, i.e., with $\epsilon_k = 0 \; \forall k$.

Corollary 2.1. *Let $(d, H) \in \mathcal{F}(\overline{C})$, $X_* \neq \emptyset$, and $\{x^k\}$ be the sequence generated by IPA with $\epsilon_k = 0 \; \forall k$. Then, $f(x^n) - f_* = O(\sigma_n^{-1}) \; \forall x \in \overline{C}$.*

Proof. Under the given hypothesis, Theorem 2.1(i) holds for any $x \in \overline{C}$, and it follows that $f(x^n) - f_* \leq (\sigma_n)^{-1} H(x^*, x^0)$. □

To establish the global convergence of the sequence $\{x^k\}$ to an optimal solution of problem (P), we need to make further assumptions on the induced proximal distance H, mimicking the behavior of norms.

Let $(d, H) \in \mathcal{F}_+(\overline{C}) \subset \mathcal{F}(\overline{C})$ be such that the function H satisfies the following two additional properties:

(a_1) $\forall y \in \overline{C}$ and $\forall \{y_k\} \subset C$ bounded with $\lim_{k \to +\infty} H(y, y_k) = 0$, we have $\lim_{k \to +\infty} y_k = y$;

(a_2) $\forall y \in \overline{C}$ and $\forall \{y_k\} \subset C$ converging to y, we have $\lim_{k \to +\infty} H(y, y_k) = 0$.

With these additional hypotheses on H we immediately obtain that IPA globally converges to an optimal solution of (P).

Theorem 2.2. *Let $(d, H) \in \mathcal{F}_+(\overline{C})$ and let $\{x^k\}$ be the sequence generated by IPA. Suppose that the optimal set X_* of* (P) *is nonempty, $\sigma_n = \sum_{k=1}^{n} \lambda_k \to \infty$, $\sum_{k=1}^{\infty} \lambda_k \epsilon_k < \infty$, and $\sum_{k=1}^{\infty} \epsilon_k < \infty$. Then the sequence $\{x^k\}$ converges to an optimal solution of* (P).

Proof. Let $x \in X_*$. Then, since $(d, H) \in \mathcal{F}_+(\overline{C})$, from (2.12) with $\sum_{k=1}^{n} \lambda_k \epsilon_k < +\infty$ and Lemma 2.1 we obtain that the sequence $\{H(x, x^k)\}$ converges to some $a(x) \in \mathbb{R} \; \forall x \in X_*$. Let x_∞ be the limit of a subsequence $\{x^{k_l}\}$. Obviously, from Theorem 2.1, $x_\infty \in X_*$. Then by assumption (a_2) $\lim_{l \to \infty} H(x_\infty, x^{k_l}) = 0$, so that $\lim_{k \to \infty} H(x_\infty, x^k) = 0$, and by assumption ($a_1$) it follows that the sequence $\{x^k\}$ converges to x_∞. □

Note that we have separated the two types of convergence results to emphasize

- the differences and roles played by each of the three classes $\mathcal{F}_+(\overline{C}) \subset \mathcal{F}(\overline{C}) \subset \mathcal{F}(C)$,
- that the largest, and less demanding, class $\mathcal{F}(C)$ already provides reasonable convergence properties for IPA, with minimal assumptions on the problem's data.

These aspects are illustrated by several application examples in the next section.

Relations (2.3), (2.4) defining IPA can sometimes be difficult to implement, since at each step we have to find by some algorithm in a finite number of steps an ϵ_k-solution for the minimization of the function $\lambda_k f(\cdot) + d(\cdot, x^{k-1})$. To overcome this difficulty, we consider here (among others) a variant of the approximate rule proposed in [20] for self-proximal Bregman methods.

Interior proximal algorithm with approximation rule (IPA1). Let $(d, H) \in \mathcal{F}(C)$, $\lambda_* > 0$, and for each $k = 1, 2, \dots$, let $\lambda_k \geq \lambda_*$, $\eta_k > 0$, and $\epsilon_k > 0$ with $\sum_{k=1}^{\infty} \epsilon_k < \infty$, $\sum_{k=1}^{\infty} \eta_k < \infty$. Starting from a point $x^0 \in C$, for all $k \geq 1$ we generate the sequences $\{x^k\}_{k=1}^{\infty} \subset C$, $\{e^k\}_{k=1}^{\infty} \subset \mathbb{R}^n$ via

$$e^k = \lambda_k g^k + \nabla_1 d(x^k, x^{k-1}) \text{ with } g^k \in \partial f(x^k), \tag{2.13}$$

where the error sequence $\{e^k\}$ satisfies the conditions

$$\|e^k\| \leq \epsilon_k, \quad \|e^k\| \sup(\|x^k\|, \|x^{k-1}\|) \leq \eta_k. \tag{2.14}$$

Remark 2.1. From Proposition 2.1, a sequence $\{x^k\}$ given by relations (2.13), (2.14) always exists. Furthermore, if f is C^1 on C (C^2 on C with $d(\cdot, y) \in C^2$ on C for all $y \in C$), then any convergent gradient-type method (Newton-type method) will provide such an x^k in a finite number of steps.

Theorem 2.3. *Let $(d, H) \in \mathcal{F}(C)$, and let $\{x^k\}$ be a sequence generated by IPA1. Then we have the following:*

(i) *The sequence $\{f(x^k)\}$ converges to f_*.*

(ii) *Furthermore, suppose that the optimal set X_* is nonempty, and consider the following cases:*

  (a) *X_* is bounded;*

  (b) *$(d, H) \in \mathcal{F}(\overline{C})$;*

  (c) *$(d, H) \in \mathcal{F}_+(\overline{C})$.*

Then under (a) *or* (b)*, the sequence $\{x^k\}$ is bounded with all limit points in X_*, while under* (c) *the sequence $\{x^k\}$ converges to an optimal solution.*

Proof. Since $g^k \in \partial f(x^k)$, using (2.6) and the Cauchy–Schwarz inequality we get for any $x \in C$

$$\begin{aligned} \lambda_k(f(x^k) - f(x)) &\leq \langle x - x^k, \nabla_1 d(x^k, x^{k-1})\rangle + \langle e^k, x^k - x\rangle \\ &\leq H(x, x^{k-1}) - H(x, x^k) + \tilde{\epsilon_k}(x), \end{aligned} \tag{2.15}$$

with $\tilde{\epsilon_k}(x) := \|e^k\|\|x\| + \langle x^k, e^k\rangle$. Summing (2.15) over $k = 1, \dots, n$ and dividing by $\sigma_n = \sum_{i=1}^{n} \lambda_k$ we obtain

$$-f(x) + \sum_{k=1}^{n} \frac{\lambda_k f(x^k)}{\sigma_n} \leq \sigma_n^{-1} \left[H(x, x^0) - H(x, x^n) + \sum_{k=1}^{n} \tilde{\epsilon_k}(x) \right]. \tag{2.16}$$

Now setting $x = x^{k-1}$ in (2.15) and $\alpha_k := |\tilde{\epsilon_k}(x^{k-1})|\lambda_*^{-1}$, we obtain $(f(x^k) - f(x^{k-1})) \leq \alpha_k$. But using (2.14), one has $\sum_{k=1}^{\infty} \tilde{\epsilon_k}(x) < \infty$ and $\sum_{k=1}^{\infty} \alpha_k < \infty$. Therefore by passing to the limit in (2.16) and invoking Lemma 2.2(i) it follows that $\liminf_{n\to\infty} f(x^n) - f(x) \leq 0$ for each $x \in C$ so that $\liminf_{n\to\infty} f(x^n) \leq \inf\{f(x) \mid x \in C\}$. From here, the proof can be completed with the same arguments as in the proofs of Theorems 2.1 and 2.2. □

Theorem 2.3(c) recovers and extends [20, Theorem 1, p. 120] for the case of convex minimization, which was proved there only for the Bregman self-proximal method.

3. Proximal distances (d, H): Examples.

It turns out that in most situations, when constructing an IPA for solving the convex problem (P), the proximal distance H induced by d will be a Bregman proximal distance D_h generated by some convex kernel h. In the first part of this section we recall the special features of the Bregman proximal distance. In the second part we consider various types of constraint sets $\overline{C}$ for problem (P). We demonstrate through many examples for the pair (d, H) that many well-known proximal methods, as well as new ones, can be handled through our framework.

3.1. Bregman proximal distances.

Let $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper, lsc, and convex function with $\operatorname{dom} h \subset \overline{C}$ and $\operatorname{dom} \nabla h = C$, strictly convex and continuous on $\operatorname{dom} h$, C^1 on $\operatorname{int} \operatorname{dom} h = C$. Define

$$\begin{aligned} H(x,y) := D_h(x,y) &:= h(x) - [h(y) + \langle \nabla h(y), x - y\rangle] \quad \forall x \in \mathbb{R}^n,\ \forall y \in \operatorname{dom} \nabla h \\ &= +\infty \quad \text{otherwise.} \end{aligned} \tag{3.1}$$

The function D_h enjoys a remarkable three point identity [16, Lemma 3.1],

$$H(c,a) = H(c,b) + H(b,a) + \langle c - b, \nabla_1 H(b,a)\rangle \quad \forall a, b \in C,\ \forall c \in \operatorname{dom} h. \tag{3.2}$$

This identity plays a central role in the convergence analysis.

To handle the constraint cases C versus $\overline{C}$, we consider two types of kernels h. The first type consists of convex kernel functions h (often called a Bregman function with zone C; see, e.g., [15]) that satisfy the following conditions:

(B_1) $\operatorname{dom} h = \overline{C}$;

(B_2) (i) $\forall x \in \overline{C}$, $D_h(x, \cdot)$ is level bounded on $\operatorname{int}(\operatorname{dom} h)$;
(ii) $\forall y \in C$, $D_h(\cdot, y)$ is level bounded;

(B_3) $\forall y \in \operatorname{dom} h$, $\forall \{y_k\} \subset \operatorname{int}(\operatorname{dom} h)$ with $\lim_{k\to\infty} y_k = y$, one has $\lim_{k\to\infty} D_h(y, y_k) = 0$;

(B_4) if $\{y_k\}$ is a bounded sequence in $\operatorname{int}(\operatorname{dom} h)$ and $y \in \operatorname{dom} h$ such that $\lim_{k\to\infty} D_h(y, y_k) = 0$, then $y = \lim_{k\to\infty} y_k$.

Note that (B_4) is a direct consequence of the first three properties, a fact proved by Kiwiel in [29, Lemma 2.16].

Let $\mathcal{B}$ be the class of kernels h satisfying properties (B_1)–(B_4). More general Bregman proximal distances such as those introduced in [29] could also be candidates. For the sake of simplicity we consider here only the case $h \in \mathcal{B}$.

For the second type of kernels, we require the convex kernel h to satisfy two (weaker)[1] conditions:

(WB_1) $\operatorname{dom} h = C$;

(WB_2) (i) $\forall x \in C$, $D_h(x, \cdot)$ is level bounded on C;
(ii) $\forall y \in C$, $D_h(\cdot, y)$ is level bounded.

We denote by $\mathcal{WB}$ the set of such convex kernels h.

We give here some examples that underline the difference between the classes $\mathcal{B}$ and $\mathcal{WB}$.

Example 3.1. Let $C = \mathbb{R}^n_{++}$. Separable Bregman proximal distances are the most commonly used in the literature. Let $\theta : \mathbb{R} \to \mathbb{R} \cup +\infty$ be a proper convex and lsc function with $(0, +\infty) \subset \operatorname{dom} \theta \subset [0, +\infty)$ and such that $\theta \in C^2(0, +\infty)$, $\theta''(t) > 0$

[1] The terminology "weaker" is used here to indicate that "weaker type" of convergence results can be derived for this class. Indeed, with $h \in \mathcal{WB}$ one has $(d, D_h) \in \mathcal{F}(C)$ and only Theorem 2.1 (except (iii)(b)) can be applied.

$\forall t > 0$, and $\lim_{t\to 0^+} \theta'(t) = -\infty$. We denote this class by Θ_0 if $\theta(0) < +\infty$ and by Θ_+ whenever $\theta(0) = +\infty$ and θ is also assumed nonincreasing. Given θ in either class, define $h(x) = \sum_{j=1}^n \theta(x_j)$ so that D_h is separable. The first two examples are functions $\theta \in \Theta_0$, i.e., with $\operatorname{dom}\theta = [0, +\infty)$, and the last two are in Θ_+, i.e., with $\operatorname{dom}\theta = (0, +\infty)$:

- $\theta_1(t) = t\log t$ (Shannon entropy),
- $\theta_2(t) = (pt - t^p)/(1-p)$ with $p \in (0, 1)$,
- $\theta_3(t) = -\log t$ (Burg entropy),
- $\theta_4(t) = t^{-1}$.

More examples can be found in, e.g., [29, 43]. Then, the corresponding proximal distances $D_{h_1}, D_{h_2} \in \mathcal{B}$, while $D_{h_3}, D_{h_4} \in \mathcal{WB}$.

3.2. Self-proximal methods. The three point identity (3.2) plays a fundamental role in the convergence of Bregman-based self-proximal methods, namely those for which we take d itself as a Bregman proximal distance, that is, $d(x, y) = H(x, y) = D_h(x, y)$, with D_h as defined in (3.1). Whenever $h \in \mathcal{B}$, or in $\mathcal{WB}$, properties (P_1), (P_2), and (P_3) hold for $d = D_h$.

Clearly, $D_h(a, a) = 0\ \forall a \in C$, so that (P_4) holds, and since H is always nonnegative it follows from (3.2) that (2.6) holds. Therefore for $h \in \mathcal{WB}$ one has $(d, H) = (D_h, D_h) \in \mathcal{F}(C)$, while if $h \in \mathcal{B}$, then $(d, H) = (D_h, D_h) \in \mathcal{F}_+(\overline{C})$.

When $C = \mathbb{R}^n$, with $h(\cdot) = \|\cdot\|^2/2 \in \mathcal{B}$, then $D_h(x, y) = \|x - y\|^2/2$, and with $(d, H) = (D_h, D_h) \in \mathcal{F}_+(\mathbb{R}^n)$, the IPA is exactly the classical proximal method and Theorems 2.1 and 2.2 cover the usual convergence results, e.g., [24, 30, 31].

We now list several interesting special cases for the pair (d, H) leading to self-proximal schemes for various types of constraints.

Nonnegative constraints. Let $C = \mathbb{R}^n_{++}$ and $\overline{C} = \mathbb{R}^n_+$. For the examples given in Example 3.1, the resulting self-proximal algorithms, namely with $d = H = D_{h_i}$, yield $(d, D_{h_i}) \in \mathcal{F}_+(\overline{C})$ for $i = 1, 2$ and $(d, D_{h_i}) \in \mathcal{F}(C)$ for $i = 3, 4$.

Semidefinite constraints. We denote by S^n the linear space of symmetric real matrices equipped with the trace inner product $\langle x, y\rangle := \operatorname{tr}(xy)$ and $\|x\| = \sqrt{\operatorname{tr}(x^2)}$ $\forall x, y \in S^n$, where $\operatorname{tr}(x)$ is the trace of the matrix x and $\det x$ its determinant. The cone of $n \times n$ symmetric positive semidefinite (positive definite) matrices is denoted by S^n_+ (S^n_{++}). Let $C = S^n_{++}$ and $\overline{C} = S^n_+$. Let $h_1 : S^n_+ \to \mathbb{R}$, $h_1(x) = \operatorname{tr}(x\log x)$ and $h_3 : S^n_{++} \to \mathbb{R}$, $h_3(x) = -\operatorname{tr}(\log x) = -\log\det(x)$ (which corresponds to θ_1 and θ_3, respectively, of Example 3.1). For any $y \in S^n_{++}$, let

$$\begin{aligned} d_1(x, y) &= \operatorname{tr}(x\log x - x\log y + y - x) \text{ with } \operatorname{dom} d_1(\cdot, y) = S^n_+,\\ d_3(x, y) &= \operatorname{tr}(-\log x + \log y + xy^{-1}) - n\\ &\qquad = -\log\det(xy^{-1}) + \operatorname{tr}(xy^{-1}) - n \text{ with } \operatorname{dom} d_3(\cdot, y) = S^n_{++}. \end{aligned}$$

The proximal distances d_1, d_3 are Bregman type corresponding to h_1, h_3, respectively, and were proposed by Doljansky and Teboulle in [18], who derived convergence results for the associated IPA. From the results of [18] it is easy to see that $d_i \in \mathcal{D}(C)$, $i = 1, 3$, and with $H(x, y) = d_i(x, y)$ it follows that $(d_1, H) \in \mathcal{F}(S^n_+)$ and $(d_3, H) \in \mathcal{F}(S^n_{++})$ so that we recover the convergence results of [18] through Theorem 2.1. However, as noticed in a counterexample [18, Example 4.1], property (B_3) does not hold even for d_1, and therefore $(d_i, H) \notin \mathcal{F}_+(\overline{C})$, $i = 1, 3$. Consequently, Theorem 2.2 does not apply, i.e., global convergence to an optimal solution cannot be guaranteed. Similar results can be easily extended to the more general case with $C = \{x \in \mathbb{R}^m \mid B(x) \in S^n_{++}\}$ assumed nonempty, with $B(x) = \sum_{i=1}^m x_i B_i - B_0$, where

$B_i \in S^n$ $\forall i = 0, 1, \dots, m$, and the map $x \to \sum_{i=1}^m x_i B_i$ being onto, by considering the corresponding proximal distances,

$$D_1(x, y) = d_1(B(x), B(y)), \qquad D_3(x, y) = d_3(B(x), B(y)).$$

Convex programming. Let $f_i : \mathbb{R}^n \to \mathbb{R}$ be concave and C^1 on $\mathbb{R}^n$ for each $i \in [1, m]$. We suppose that Slater's condition holds, i.e., there exists some point $x_0 \in \mathbb{R}^n$ such that $f_i(x_0) > 0$ $\forall i \in [1, m]$ and that the open convex set C is described by

$$C = \{x \in \mathbb{R}^n \mid f_i(x) > 0 \ \forall i = 1, \dots, m\}$$

so that by Slater's assumption $C \neq \emptyset$ and $\overline{C} = \{x \in \mathbb{R}^n \mid f_i(x) \geq 0,\ i \in [1, m]\}$. Consider the class Θ_+ of functions defined in Example 3.1, and for each $\theta \in \Theta_+$ let

$$h(x) = \begin{cases} \sum_{i=1}^m \theta(f_i(x)) & \text{if } x \in C, \\ +\infty & \text{otherwise.} \end{cases} \tag{3.3}$$

Obviously h is a proper, lsc, and convex function. Now, consider the Bregman proximal distance associated with $h_\nu(x) := h(x) + \frac{\nu}{2}\|x\|^2$ with $\nu > 0$. Then, we take $d(x, y) = D_{h_\nu}(x, y)$, where D_{h_ν} is the Bregman distance associated with h_ν. Thanks to the condition $\nu > 0$, it follows that $h_\nu \in \mathcal{WB}$ and $(d, D_{h_\nu}) \in \mathcal{F}(C)$. An important and interesting case is obtained by choosing the Burg function, $\theta_3(t) = -\log t$. In this case we obtain the following:

$$d(x, y) = \sum_{i=1}^m -\log \frac{f_i(x)}{f_i(y)} + \frac{\langle \nabla f_i(y), x - y\rangle}{f_i(y)} + \frac{\nu}{2}\|x - y\|^2. \tag{3.4}$$

Note that in this case the function $d(\cdot, y)$ enjoys other interesting properties: for example, when the functions f_i are concave quadratic, then $d(\cdot, y)$ is self-concordant for each $y \in C$, a property which is very useful when minimizing the function with Newton-type methods [33]. When $\nu = 0$, i.e., with $d = D_h$, such proximal distance has been recently introduced by Alvarez, Bolte, and Brahic [1], in the context of dynamical systems to study interior gradient flows, but it requires a nondegeneracy condition, $\forall x \in C$: $\operatorname{span}\{\nabla f_i(x) \mid i = 1, \dots, m\} = \mathbb{R}^n$, which is satisfied mostly in the polyhedral case. Here, in the context of proximal methods, the addition of the regularized term in D_h precludes the use of such a condition.

Second order cone constraints. Let $C = L^n_{++} := \{x \in \mathbb{R}^n \mid x_n > (x_1^2 + \cdots + x_{n-1}^2)^{1/2}\}$ be the interior of the Lorentz cone, with closure denoted by L^n_+. Let J_n be a diagonal matrix with its first $(n-1)$ entries being -1 and the last being 1, and define $h : L^n_{++} \to \mathbb{R}$ by $h(x) = -\log(x^T J_n x)$. Then h is proper, lsc, and convex on $\operatorname{dom} h = L^n_{++}$. Let $h_\nu(x) = h(x) + \nu\|x\|^2/2$. Then thanks to $\nu > 0$, one has $h_\nu \in \mathcal{WB}$, and the Bregman proximal distance associated with h_ν is given by

$$D_{h_\nu}(x, y) = -\log \frac{x^T J_n x}{y^T J_n y} + \frac{2x^T J_n y}{y^T J_n y} - 2 + \frac{\nu}{2}\|x - y\|^2, \tag{3.5}$$

and we have $(D_{h_\nu}, D_{h_\nu}) \in \mathcal{F}(L^n_{++})$. As in the convex and semidefinite programming cases, one can easily handle the more general case with a nonempty $C = \{x \in \mathbb{R}^n \mid Ax - b \in L^m_{++}\}$, where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, by choosing $h(x) := -\log(Ax - b)^T J_m (Ax - b)$.

We will now show that, interestingly, even for IPA which are not self-proximal, the induced proximal distance H from the choice of d for various types of constraints will still be a Bregman proximal distance D_h with an appropriate convex kernel h in the class $\mathcal{B}$ or $\mathcal{WB}$.

3.3. Proximal functions based on φ-divergences.

φ-divergence kernels. Let $\varphi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be an lsc, convex, proper function such that $\operatorname{dom}\varphi \subset \mathbb{R}_+$ and $\operatorname{dom}\partial\varphi = \mathbb{R}_{++}$. We suppose in addition that φ is C^2, strictly convex, and nonnegative on $\mathbb{R}_{++}$ with $\varphi(1) = \varphi'(1) = 0$. We denote by Φ the class of such kernels and by Φ_1 the subclass of these kernels satisfying

$$\varphi''(1)\left(1 - \frac{1}{t}\right) \le \varphi'(t) \le \varphi''(1)\log t \quad \forall t > 0. \tag{3.6}$$

The other subclass of Φ of interest is denoted by Φ_2, where (3.6) is replaced by

$$\varphi''(1)\left(1 - \frac{1}{t}\right) \le \varphi'(t) \le \varphi''(1)(t-1) \quad \forall t > 0. \tag{3.7}$$

Examples of functions in Φ_1, Φ_2 are (see, e.g., [5, 44])

$$\begin{aligned}\varphi_1(t) &= t\log t - t + 1, \quad \operatorname{dom}\varphi = [0, +\infty),\\ \varphi_2(t) &= -\log t + t - 1, \quad \operatorname{dom}\varphi = (0, +\infty),\\ \varphi_3(t) &= 2(\sqrt{t} - 1)^2, \quad \operatorname{dom}\varphi = [0, +\infty).\end{aligned}$$

Corresponding to the classes Φ_r, with $r = 1, 2$, we define a φ-divergence proximal distance by

$$d_\varphi(x, y) = \sum_{i=1}^n y_i^r \varphi\left(\frac{x_i}{y_i}\right).$$

For any $\varphi \in \Phi$, since $\operatorname{argmin}\{\varphi(t) \mid t \in \mathbb{R}\} = \{1\}$, φ is coercive and thus it follows that $d_\varphi \in \mathcal{D}(C)$, with $C = \mathbb{R}^n_{++}$.

The use of φ-divergence proximal distances is particularly suitable for handling polyhedral constraints. Let $C = \{x \in \mathbb{R}^n \mid Ax < b\}$, where A is an (m, n) matrix of full rank m $(m \ge n)$. Particularly important cases include $C = \mathbb{R}^n_{++}$ or $C = \{x \in \mathbb{R}^n \mid a_i < x_i < b_i \ \forall i = 1, \dots, n\}$, with $a_i, b_i \in \mathbb{R}$. For the sake of simplicity we consider here only the case where $C = \mathbb{R}^n_{++}$. Indeed, as is already noted in several works (e.g., [3, 4, 44]), since these proximal distances are separable they can thus be extended without difficulty to the polyhedral case by redefining d in the form $d(x, y) := \sum_{i=1}^m d_i(b_i - \langle a_i, x\rangle, b_i - \langle a_i, y\rangle)$, where a_i are the rows of the matrix A and $d_i(u_i, v_i) = v_i^r \varphi(u_i v_i^{-1})$.

The class Φ_1. It turns out that the induced proximal distance H associated with d_φ is the Bregman proximal distance obtained from the kernel $h(x) = \sum_{j=1}^n x_j \log x_j$ (obtained from θ_1) and given by

$$D_h(x, y) := K(x, y) = \sum_{j=1}^n x_j \log\frac{x_j}{y_j} + y_j - x_j \quad \forall x \in \mathbb{R}^n_+,\ \forall y \in \mathbb{R}^n_{++}, \tag{3.8}$$

which is the Kullback–Liebler relative entropy. The fact that K plays a central role in the analysis of IPA based on $\varphi \in \Phi_1$ was already realized in [28] and later formalized in [44, Lemma 4.1(ii)], which shows that for any $\varphi \in \Phi_1$ one has

$$\langle c-b, \nabla_1 d_\varphi(b,a)\rangle \le \varphi''(1)[K(c,a)-K(c,b)]. \tag{3.9}$$

Therefore (2.6) is verified for $H = \varphi''(1)K$ and it follows that for any $\varphi \in \Phi_1$ one has $(d_\varphi, \varphi''(1)K) \in \mathcal{F}_+(\overline{C})$ and all the convergence results of section 2 apply for the corresponding IPA. We note parenthetically that the induced proximal distance K can also be obtained from the φ-divergence with the kernel φ_1. In fact this should not be surprising, since it can be verified that $d_\varphi = D_h$ if and only if $h(x) = \sum_{j=1}^n \varphi_1(x_j)$.

Regularized class $\mathbf{\Phi_1}$. Let $\varphi \in \Phi_1$ and define $d(x,y) = d_\varphi(x,y) + 2^{-1}\nu\|x-y\|^2$, with $\nu > 0$. This proximal distance was recently considered in [2] in the context of Lotka–Volterra dynamical systems with the choice $\varphi = \varphi_2$. As shown there, one can verify that with $H(x,y) = K(x,y) + \frac{\nu}{2}\|x-y\|^2$, one has $(d_{\varphi_2}, H) \in \mathcal{F}_+(\overline{C})$.

The class $\mathbf{\Phi_2}$: Second order homogeneous proximal distances [5, 10, 45]**.** Let $\varphi(t) = \mu p(t) + \frac{\nu}{2}(t-1)^2$ with $\nu \ge \mu > 0$, $p \in \Phi_2$, and let the associated proximal distance be defined by

$$d_\varphi(x,y) = \sum_{j=1}^n y_j^2 \varphi\left(\frac{x_j}{y_j}\right).$$

In particular, $p(t) = -\log t + t - 1$ gives the so-called logarithmic-quadratic proximal distance [5]. Obviously $d_\varphi \in \mathcal{D}(C)$, and from the key inequality [4, Lemma 3.4] one has

$$\langle c-b, \nabla_1 d(b,a)\rangle \le \eta(\|c-a\|^2 - \|c-b\|^2) \quad \forall a,b \in \mathbb{R}_{++},\ \forall c \in \mathbb{R}_+$$

with $\eta = 2^{-1}(\mu+\nu)$. Therefore with $H(x,y) = \eta\|x-y\|^2$ it follows that $(d_\varphi, H) \in \mathcal{F}_+(\overline{C})$.

4. Interior gradient methods.

When $\overline{C} = \mathbb{R}^n$, Correa and Lemarechal [17] and Robinson [38] have remarked that the PA can be viewed as an ϵ-subgradient descent method. This idea was recently extended by Auslender and Teboulle [7] for the logarithmic-quadratic proximal method which allows us to handle linear inequality constraints directly. Given the framework developed in section 2, we extend these results for more general constraints and with various classes of proximal distances.

We first give the main convergence result. We then present applications and examples which allow us to improve some known interior gradient–based methods as well as to derive new and simple convergent algorithms for conic optimization problems.

4.1. A general convergence theorem.

To solve problem (P) $\inf\{f(x) \mid x \in \overline{C}\}$ we consider the following general projected subgradient-based algorithm (PSA).

Take $d \in \mathcal{D}(C)$. Let $\lambda_k > 0$, $\epsilon_k \ge 0$, and $m \in (0,1]$, and for $k \ge 1$ generate the sequence $\{x^k, g^k\}$ such that

$$x^{k-1} \in C, \quad g^{k-1} \in \partial_{\varepsilon_k} f(x^{k-1}), \tag{4.1}$$

$$x^k \in \operatorname{argmin}\{\lambda_k\langle g^{k-1}, x\rangle + d(x, x^{k-1}) \mid x \in C\}, \tag{4.2}$$

$$f(x^k) \le f(x^{k-1}) + m(\langle g^{k-1}, x^k - x^{k-1}\rangle - \epsilon_k). \tag{4.3}$$

Let us briefly recall why the sequence $\{x^k\}$, constructed by the exact IPA ($\epsilon_k = 0$) in section 2 via (2.3) and (2.4), fits in PSA (see, e.g., [17, 7] for more details). Starting

IPA with $x^0 \in C$, one has $x^k \in C$, and it can be verified that $g^k \in \partial f(x^k)$ is equivalent to saying that

$$g^k \in \partial_{\epsilon_k^*} f(x^{k-1}) \quad \text{with } \epsilon_k^* = f(x^{k-1}) - f(x^k) + \langle g^k, x^k - x^{k-1}\rangle \geq 0.$$

Therefore (2.3) and (2.4) are nothing else but (4.1) and (4.2). Then, with $m = 1$ and with ϵ_k^* as defined above, inequality (4.3) holds as an equality, showing that the sequence $\{x^k\}$ generated by IPA satisfies (4.1), (4.2), and (4.3).

Building on the material developed in section 2 it is now possible to establish convergence results of PSA for various instances of the triple $[C, d, H]$, extending recent convergence results given in [7, Theorem 4.1]. Before doing so, we first note that by using the same arguments as in the proof of Proposition 2.1, it is easily seen that the existence of $x^k \in C$ is guaranteed.

THEOREM 4.1. *Let $\{x^k\}$ be a sequence generated by PSA with $(d, H) \in \mathcal{F}(C)$. Set $\sigma_n = \sum_{k=1}^n \lambda_k$ and $\alpha_k = \langle g^{k-1}, x^{k-1} - x^k\rangle$. Then,*

(i) $\sum_{k=1}^\infty \alpha_k < \infty$, $\sum_{k=1}^\infty \epsilon_k < \infty$, *and* $\alpha_k \geq \lambda_k^{-1} H(x^k, x^{k-1}) \geq 0$ $\forall k \in \mathbb{N}$.

(ii) $\forall z \in C$, $f(x^n) - f(z) \leq \sigma_n^{-1}[H(z, x^0) + \sum_{k=1}^n \lambda_k(\alpha_k + \varepsilon_k)]$.

(iii) *The sequence $\{f(x^k)\}$ is nonincreasing and converges to f_* as $\sigma_n \to \infty$.*

(iv) *Suppose that the optimal set X_* is nonempty and $\sigma_n \to \infty$. Then the sequence $\{x^k\}$ is bounded with all its limit points in X_* under either one of the following conditions:*

(a) *X_* is bounded.*

(b) *$(d, H) \in \mathcal{F}(\overline{C})$ and $\sum_{k=1}^\infty \lambda_k \epsilon_k < +\infty$ (which in particular is true if $\{\lambda_k\}$ is bounded above).*

In addition, if $(d, H) \in \mathcal{F}_+(\overline{C})$, then $\{x^k\}$ converges to an optimal solution of (P).

Proof. (i) From the optimality conditions, (4.2) is equivalent to

$$\lambda_k g^{k-1} + \nabla_1 d(x^k, x^{k-1}) = 0.$$

Since $H(\cdot,\cdot) \geq 0$ and $H(a, a) = 0$, from (2.6) with $c = a = x^{k-1}$, $b = x^k$ we then obtain

$$\lambda_k \alpha_k = \langle \nabla_1 d(x^k, x^{k-1}), x^k - x^{k-1}\rangle \geq H(x^{k-1}, x^k) \geq 0.$$

Furthermore, from (4.3) we obtain

$$m(\alpha_k + \epsilon_k) \leq f(x^{k-1}) - f(x^k), \tag{4.4}$$

which also shows that $\{f(x^k)\}$ is nonincreasing. Summing over $k = 1, \dots, n$ in the last inequality it follows that

$$m \sum_{k=1}^n (\alpha_k + \epsilon_k) \leq f(x^0) - f(x^n) \leq f(x^0) - f_*, \tag{4.5}$$

proving (i). Now since $\sigma_n = \sum_{k=1}^n \lambda_k$, using $\sigma_k = \lambda_k + \sigma_{k-1}$ (with $\sigma_0 = 0$), multiplying (4.4) by σ_{k-1}, and summing over $k = 1, \dots n$, we obtain

$$\sum_{k=1}^n [(\sigma_k - \lambda_k) f(x^k) - \sigma_{k-1} f(x^{k-1})] \leq 0,$$

which reduces to

$$\sigma_n f(x^n) - \sum_{k=1}^{n} \lambda_k f(x^k) \leq 0. \tag{4.6}$$

Now, since $g^{k-1} \in \partial_{\epsilon_k} f(x^{k-1})$, then for any $z \in C$ one has

$$\begin{aligned} f(z) - f(x^{k-1}) + \epsilon_k &\geq \langle g^{k-1}, z - x^{k-1} \rangle \\ &= \langle g^{k-1}, z - x^k \rangle + \langle g^{k-1}, x^k - x^{k-1} \rangle \\ &= -\frac{1}{\lambda_k} \langle z - x^k, \nabla_1 d(x^k, x^{k-1}) \rangle - \alpha_k \\ &\geq \frac{1}{\lambda_k} [H(z, x^k) - H(z, x^{k-1})] - \alpha_k, \end{aligned}$$

where the last inequality uses (2.6) with $b = x^k$, $a = x^{k-1}$. Since $f(x^k) \leq f(x^{k-1})$, it then follows that

$$\lambda_k (f(x^k) - f(z)) \leq H(z, x^{k-1}) - H(z, x^k) + \lambda_k (\alpha_k + \varepsilon_k).$$

Summing the above inequality over $k = 1, \ldots, n$, we obtain

$$-\sigma_n f(z) + \sum_{k=1}^{n} \lambda_k f(x^k) \leq H(z, x^0) - H(z, x^n) + \sum_{k=1}^{n} \lambda_k (\alpha_k + \epsilon_k).$$

Adding this inequality to (4.6) and dividing by σ_n one obtains

$$f(x^n) - f(z) \leq \frac{H(z, x^0)}{\sigma_n} + \sum_{k=1}^{n} \frac{\lambda_k (\alpha_k + \epsilon_k)}{\sigma_n} \quad \forall z \in C.$$

This proves (ii). Suppose $\sigma_n \to \infty$. Since the sequences $\{\alpha_k\}$ and $\{\epsilon_k\}$ converge to 0, invoking Lemma 2.2 and passing to the limit we obtain

$$\lim_{n \to \infty} f(x^n) = \limsup_{n \to \infty} f(x^n) \leq \inf\{f(x) \mid x \in C\} = f_*,$$

proving (iii). The rest of the proof is exactly the same as in the proof of Theorems 2.1 and 2.2. □

Using (ii) of Theorem 4.1 with (4.5), we obtain the following corollary.

COROLLARY 4.1. *Let $(d, H) \in \mathcal{F}(\overline{C})$ and let $\{x^k\}$ be the sequence produced by PSA. Suppose that $X_* \neq \emptyset$ and $0 < \lambda_* \leq \lambda_k \leq \lambda^*$. Then we have the global estimation $f(x^n) - f_* = O(n^{-1})$.*

4.2. Conic optimization: Interior projected gradient methods with strongly convex proximal distance.

4.2.1. Preliminaries.

We consider now the problem

$$\text{(M)} \qquad \inf\{f(x) \mid x \in \overline{C} \cap \mathcal{V}\},$$

where $\mathcal{V} = \{x : Ax = b\}$, with $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$, $n \geq m$, $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex and lsc, and we assume that $\exists x^0 \in \operatorname{dom} f \cap C : Ax^0 = b$.

When $\overline{C}$ is a convex cone, problem (M) is the standard conic optimization problem (see, e.g., [33]), while whenever $\mathcal{V} = \mathbb{R}^n$ it is just a pure conic optimization problem.

In the following subsection, we assume also that f is continuously differentiable with ∇f Lipschitz on $C \cap \mathcal{V}$ and Lipschitz constant L, i.e., $\exists L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall x, y \in C \cap \mathcal{V}. \tag{4.7}$$

We consider now $(d, H) \in \mathcal{F}(C)$ such that d satisfies the following properties:

(s1) $\exists \sigma > 0 : \forall y \in C \cap \mathcal{V}$, $d(\cdot, y)$ is σ-strongly convex over $C \cap \mathcal{V}$, i.e.,

$$\langle \nabla_1 d(x_1, y) - \nabla_1 d(x_2, y), x_1 - x_2\rangle \geq \sigma\|x_1 - x_2\|^2 \quad \forall x_1, x_2 \in C \cap \mathcal{V}, \tag{4.8}$$

for some norm $\|\cdot\|$ in $\mathbb{R}^n$.

(s2) $\forall y \in C \cap \mathcal{V}$, $d(\cdot, y)$ is C^2 on C with Hessian function denoted by $\nabla_1^2 d(\cdot, y)$.

Therefore with the same arguments as the ones given in the proof of Proposition 2.1, it follows that for each $x \in C \cap \mathcal{V}$, for each $v \in \mathbb{R}^n$ there exists a unique (by strong convexity) point $u(v, x) \in C \cap \mathcal{V}$ solving

$$u(v, x) = \operatorname{argmin}\{\langle v, z\rangle + d(z, x) \mid z \in \mathcal{V}\}. \tag{4.9}$$

Then from the optimality conditions for the convex problem (4.9) (see, e.g., [39, section 28]), $\exists \mu := \mu(v, x) \in \mathbb{R}^m$ such that[2]

$$v + A^t\mu + \nabla_1 d(u(v, x), x) = 0, \quad Au(v, x) = b. \tag{4.10}$$

Clearly, problem (M) can be equivalently formulated in the form of problem (P) as follows:

$$f_* = \min\{f_0(x) \mid x \in \overline{C}\} \quad \text{with} \quad f_0 = f + \delta_{\mathcal{V}}.$$

Define $\mathcal{V}_0 = \{x : Ax = 0\}$. Note that for any $w \in \mathcal{V}$ one has $f(w) = f_0(w)$ and

$$\gamma(\eta, x) := (\nabla f(x) + A^t\eta) \in \partial f_0(x) \quad \forall x \in C \cap \mathcal{V},\ \forall \eta \in \mathbb{R}^m. \tag{4.11}$$

Indeed, for any $z, x \in \mathcal{V}$ we have $z - x \in \mathcal{V}_0$ and thus, for any $\eta \in \mathbb{R}^m$,

$$\begin{aligned} f_0(z) = f(z) \geq f(x) + \langle \nabla f(x), z - x\rangle &= f_0(x) + \langle \nabla f(x) + A^t\eta, z - x\rangle \\ &= f_0(x) + \langle \gamma(\eta, x), z - x\rangle. \end{aligned}$$

Since for $z \notin \mathcal{V}$ this inequality obviously holds, (4.11) is verified.

4.2.2. Algorithms.

We can now propose for solving problem (M) the basic iteration of our algorithm. Given a step-size rule for choosing λ_k at each step k, starting from a point $x^0 \in C \cap \mathcal{V}$ we generate iteratively the sequence $x^k \in C \cap \mathcal{V}$ by the relation

$$x^k = u(\lambda_k \nabla f(x^{k-1}), x^{k-1}). \tag{4.12}$$

As a consequence of the above discussion relations (4.1) and (4.2) are satisfied with f replaced by f_0, $\epsilon_k = 0$, and

$$g^{k-1} = \gamma\left(\frac{\mu(\lambda_k \nabla f(x^{k-1}), x^{k-1})}{\lambda_k}, x^{k-1}\right) \in \partial f_0(x^{k-1}). \tag{4.13}$$

[2] Note that the first relation in the optimality condition can be rewritten equivalently as $v + \nabla_1 d(u(v, x), x) \in \mathcal{V}_0^\perp$, where $\mathcal{V}_0 = \{x : Ax = 0\}$.

We propose now two step-size rules, and for each rule we will show that inequality (4.3) holds and $\sum_{k=1}^{\infty} \lambda_k = \infty$. As a consequence we will be able to apply Theorem 4.1 and then to devise two convergent interior gradient projection algorithms which naturally extend the results of Auslender and Teboulle [7].

Algorithm 1 (constant step-size rule). Let $\epsilon \in]0,1[$ and set $\lambda^* := 2\epsilon\sigma L^{-1}$, $\lambda_* \in (0, \lambda^*)$. Start from a point $x^0 \in C \cap \mathcal{V}$ and generate the sequence $\{x^k\} \in C \cap \mathcal{V}$ as follows: if $\nabla f(x^{k-1}) \in \mathcal{V}_0^\perp$, stop. Otherwise, compute

$$x^k = x^k(\lambda_k) := u(\lambda_k \nabla f(x^{k-1}), x^{k-1}) \quad \text{with } \lambda_k \in (\lambda_*, \lambda^*]. \tag{4.14}$$

Theorem 4.2. *Let $\{x^k\}$ be the sequence produced by Algorithm* 1. *If at step k one has $\nabla f(x^{k-1}) \in \mathcal{V}_0^\perp$, then x^{k-1} is an optimal solution. Otherwise, the sequence $\{f(x^k)\}$ is nonincreasing and converges to f_*. Moreover, suppose that the optimal set X_* is nonempty; then*

(a) *if X_* is bounded, the sequence $\{x^k\}$ is bounded with all its limit points in X_*;*

(b) *if $(d, H) \in \mathcal{F}_+(\overline{C})$, the sequence $\{x^k\}$ converges to an optimal solution of* (P).

Proof. First, if $\nabla f(x^{k-1}) \in \mathcal{V}_0^\perp$, since $x^{k-1} \in C \cap \mathcal{V}$ then obviously, from the optimality conditions (4.10), it follows that x^{k-1} is also an optimal solution. Suppose now that $\nabla f(x^{k-1}) \notin \mathcal{V}_0^\perp$. Since $\lambda_k \geq \lambda_*$, then $\sigma_n = \sum_{k=1}^{n} \lambda_k \to \infty$. Thus, it remains to show (4.3), and our result would follow as a direct consequence of Theorem 4.1. Since ∇f is Lipschitz, by the well-known descent lemma (see, e.g., [12, p. 667]) one has

$$f(x^k) \leq f(x^{k-1}) + \langle \nabla f(x^{k-1}), x^k - x^{k-1} \rangle + \frac{L}{2}\|x^k - x^{k-1}\|^2. \tag{4.15}$$

Now, we first remark that

$$(x^k - x^{k-1}) \in \mathcal{V}_0. \tag{4.16}$$

Then using (4.8), with $x_1 = y = x^{k-1} \in C \cap \mathcal{V}$, $x_2 = u(v, x^{k-1}) \in C \cap \mathcal{V}$, and $v = \lambda_k \nabla f(x^{k-1})$; (4.10); and g^{k-1} as defined in (4.13) (recalling that $\nabla_1 d(y,y) = 0$), it follows that

$$\lambda_k \langle g^{k-1}, x^{k-1} - x^k \rangle = \lambda_k \langle \nabla f(x^{k-1}), x^{k-1} - x^k \rangle \geq \sigma \|x^k - x^{k-1}\|^2.$$

This combined with (4.15) yields

$$f(x^k) \leq f(x^{k-1}) + \langle x^k - x^{k-1}, g^{k-1} \rangle \left(1 - \frac{L\lambda_k}{2\sigma}\right),$$

so that with $f_0(x^k) = f(x^k)$, $f_0(x^{k-1}) = f(x^{k-1})$ we get

$$f_0(x^k) \leq f_0(x^{k-1}) + \langle x^k - x^{k-1}, g^{k-1} \rangle \left(1 - \frac{L\lambda_k}{2\sigma}\right).$$

Then with $\lambda^* = \frac{2\epsilon\sigma}{L}$, we get $f_0(x^k) \leq f_0(x^{k-1}) + \langle x^k - x^{k-1}, g^{k-1} \rangle (1 - \epsilon)$, showing that (4.3) holds with $m = 1 - \epsilon$. □

The second algorithm extends the method proposed in [7] and allows us to use a generalized step-size rule, reminiscent of the one used in the classical projected gradient method as studied by Bertsekas [11].

Algorithm 2 (Armijo–Goldstein step-size rule). Let $\beta \in (0,1)$, $m \in (0,1)$, and $s > 0$ be fixed chosen scalars. Start from a point $x^0 \in C \cap \mathcal{V}$ and generate the sequence $\{x^k\} \in C \cap \mathcal{V}$ as follows: if $\nabla f(x^{k-1}) \in \mathcal{V}_0^\perp$ stop. Otherwise, with $x^k(\lambda) = u(\lambda \nabla f(x^{k-1}), x^{k-1})$, set $\lambda_k = \beta^{j_k} s$, where j_k is the first nonnegative integer j such that

$$f(x^k(\beta^j s)) - f(x^{k-1}) \leq m\langle \nabla f(x^{k-1}), x^k(\beta^j s) - x^{k-1}\rangle. \tag{4.17}$$

Then set $x^k = x^k(\lambda_k)$.

In order to show that this step-size rule is well defined, we need the following proposition.

PROPOSITION 4.1. *For any $x \in C \cap \mathcal{V}$, any $v \in \mathbb{R}^n$, and $\lambda > 0$, the unique solution $u(\lambda v, x)$ defined by (4.9) satisfies $u(0,x) = x$ and the following properties hold:*

(i) $\sigma\|x - u(\lambda v, x)\|^2 \leq \lambda\langle x - u(\lambda v, x), v\rangle$,

(ii) $\frac{\|u(\lambda v,x)-x\|}{\lambda} \leq \sigma^{-1}\|v\|$,

(iii) $\lim_{\lambda\to 0^+} \frac{u(\lambda v,x)-x}{\lambda}$ *exists and is equal to $\rho(v,x) = u$, where $u \in \mathcal{V}_0$ satisfies*

$$Q(x)u + v \in \mathcal{V}_0^\perp \tag{4.18}$$

with $Q(x) = \nabla_1^2 d(x,x)$,

(iv) $\langle -\rho(v,x), v\rangle \geq \sigma\|\rho(v,x)\|^2$.

Proof. Fix any $x \in C \cap \mathcal{V}$. By (4.9), we have $u(0,x) = \operatorname{argmin}\{d(z,x) \mid z \in \mathcal{V}\}$, and thus by optimality conditions (4.10) with $\mu = 0$ it follows that $u(0,x) = x$. Furthermore, from (4.10) we have

$$\langle \lambda v + \nabla_1 d(u(\lambda v,x),x), x - u(\lambda v,x)\rangle = 0,$$

from which the inequality in (i) follows immediately by using the strong convexity inequality (4.8) at $y = x_1 = x$, $x_2 = u(\lambda v, x)$, and (ii) follows from (i) and the Cauchy–Schwarz inequality.

(iii) Since $d(\cdot, y)$ is strongly convex on $C \cap \mathcal{V}$ it follows from (4.8) that

$$\langle Q(x)h, h\rangle \geq \sigma\|h\|^2 \quad \forall h \in \mathcal{V}_0. \tag{4.19}$$

As a consequence of the Lax–Milgram theorem (see, for example, [14, Corollary 5.8]), (4.18) admits exactly one solution $\rho(v,x)$. Note that $\nabla_1 d(x,x) = 0$. Then, since by (4.10) we have

$$\lambda v + \nabla_1 d(u(\lambda v,x),x) + A^t\mu(\lambda v,x) = 0,$$

it follows that

$$\forall h \in \mathcal{V}_0 : \quad \lambda^{-1}\langle \nabla_1 d(u(\lambda v,x),x) - \nabla_1 d(x,x), h\rangle = \langle -v, h\rangle.$$

Denote $s(\lambda) := \frac{u(\lambda v,x)-x}{\lambda}$. Now, from (ii) the generalized sequence $\{s(\lambda)\}_{\lambda>0}$ is bounded. Since $u(\lambda v,x) = x + \lambda s(\lambda)$, taking the limit as $\lambda \to 0^+$ in the last equation and using the definition of the derivative (recall that here $d(\cdot,y) \in C^2$ for every $y \in C \cap \mathcal{V}$), it follows that any limit point u of the generalized sequence $\{s(\lambda)\}_{\lambda>0}$ satisfies $u \in \mathcal{V}_0$ such that

$$\langle \nabla_1^2 d(x,x)u, h\rangle = \langle Q(x)u, h\rangle = -\langle v, h\rangle \quad \forall h \in \mathcal{V}_0,$$

which is equivalent to (4.18). As a consequence $u = \rho(v,x)$ and $\lim_{\lambda\to 0^+} s(\lambda)$ exists and is equal to $\rho(v,x)$. To prove (iv), take $h = \rho(v,x) \in \mathcal{V}_0$ in the last equality and use (4.19). □

We can now prove the convergence of Algorithm 2.

THEOREM 4.3. *Let $\{x^k\}$ be the sequence generated by Algorithm 2. If at step k one has $\nabla f(x^{k-1}) \in \mathcal{V}_0^\perp$, then x^{k-1} is an optimal solution. Otherwise, the algorithm is well defined, i.e., there exists an integer j_k such that $\lambda_k = \beta^{j_k}$, and the sequence $\{\lambda_k\}$ is bounded below by $\lambda_* = \min(2\sigma\beta L^{-1}(1-m), s) > 0$. Furthermore, Theorem 4.2 holds for the sequence produced by Algorithm 2.*

Proof. We have only to prove that the algorithm is well defined and that $\lambda_k \geq \lambda_*$ (so that $\lim_{n\to\infty}\sigma_n = +\infty$). Indeed, if we set $\epsilon_k = 0$ and g^{k-1} as given in (4.13), then by definition of Algorithm 2 the sequence $\{x^k\}$ satisfies relations (4.1), (4.2), and (4.3). To simplify the notation set $x := x^{k-1}$, $v = \nabla f(x^{k-1})$, $x(\lambda) = u(\lambda v, x)$. First, if $v \in \mathcal{V}_0^\perp$, since $x \in C \cap \mathcal{V}$, then obviously, from optimality conditions (4.10), it follows that x is also an optimal solution. Suppose now that $v \notin \mathcal{V}_0^\perp$ and that (4.17) does not hold. That is,

$$f(x(\beta^j s)) - f(x) > m\langle x(\beta^j s) - x, v\rangle \quad \forall j \in \mathbb{N}. \tag{4.20}$$

Invoking the mean value theorem, $\exists z_j \in]x, x(\beta^j s)[$ such that

$$\left\langle \nabla f(z_j), \frac{x(\beta^j s) - x}{\beta^j s} \right\rangle > m \left\langle \frac{x(\beta^j s) - x}{\beta^j s}, v \right\rangle \quad \forall j \in \mathbb{N}.$$

But by Proposition 4.1(i) it follows that $\lim_{j\to\infty} z_j = x$. Moreover, passing to the limit in the last inequality and using (iii) and (iv) of the same proposition, we obtain

$$\sigma(1-m)\|\rho(v,x)\|^2 \leq (1-m)\langle -v, \rho(v,x)\rangle \leq 0,$$

which implies that $\rho(v,x) = 0$, and hence by (4.18) it follows that $v \in \mathcal{V}_0^\perp$, and we have reached a contradiction. Now let us prove that $\lambda_k \geq \lambda_*$. As for the case of the constant step-size rule, with the same arguments (using again the descent lemma; cf. (4.15)) we obtain

$$f_0(x^k(\lambda)) - f_0(x^{k-1}) \leq \langle g^{k-1}, x^k(\lambda) - x^{k-1}\rangle \left(1 - \frac{L\lambda}{2\sigma}\right) \quad \forall \lambda > 0,$$

where $x^k(\lambda) = u(\lambda\nabla f(x^{k-1}), x^{k-1})$ so that (4.17) holds for all $j \in \mathbb{N}$ with $\beta^j s \leq 2\sigma L^{-1}(1-m)$. But since by definition if $j_k \neq 0$, $\lambda_k\beta^{-1}$ does not satisfy (4.17), then $\lambda_k\beta^{-1} > 2\sigma L^{-1}(1-m)$, it follows that $\lambda_k \geq \lambda_*$ $\forall k$. From here, we can then proceed with the same statements and conclusions of Theorem 4.2 for Algorithm 2. $\square$

In general $u(v,x)$ is not given explicitly and has to be computed by an algorithm. However, there are important cases where the function d is such that $u(v,x)$ is given explicitly by an analytic formula, making these algorithms particularly attractive, which we now describe.

4.2.3. Application examples.

Consider the functions d with $(d, H) \in \mathcal{F}(C)$ which are regularized distances of the form

$$d(x,y) = p(x,y) + \frac{\sigma}{2}\|x-y\|^2, \tag{4.21}$$

with $p \in \mathcal{D}(C)$.

Note that the log-quad function belongs to this class. Such a class has been recently introduced and studied by Bolte and Teboulle [13] in the context of gradient-like continuous dynamical systems for constrained minimization.

We begin with two pure conic optimization problems, i.e., with $\mathcal{V} = \mathbb{R}^n$. We then consider the semidefinite and second order conic problems. To the best of our knowledge, this leads to the first explicit interior gradient methods for these problems with convergence results. The last application considers convex minimization over the unit simplex.

A. Convex minimization over $C = \mathbb{R}^n_{++}$. Let $\varphi \in \Phi_r$ with $r = 1, 2$ and let d be given by (4.21) with $p(z,x) = \mu \sum_{j=1}^n x_j^r \varphi(x_j^{-1} z_j)$, $\sigma \geq \mu > 0$ for $(z,x) \in C \times C$. Take for example $\varphi(t) = -\log t + t - 1$ and $r = 2$, namely the log-quad function. Then, (4.21) can be written as

$$d(z,x) = \sum_{j=1}^n x_j^2 \omega(x_j^{-1} z_j) \quad \text{with} \quad \omega(t) = \frac{\sigma}{2}(t-1)^2 + \mu(t - \log t - 1).$$

Solving (4.9), one easily obtains (see also [7, eq. (2.3), p. 4]) the following explicit formulas:

$$\forall i = j, \dots, n, \quad u_j(v,x) = x_j (\omega^*)'(-v_j x_j^{-1})$$
$$\text{with} \quad (\omega^*)'(s) = (2\sigma)^{-1}\{(\sigma - \mu) + s + \sqrt{((\sigma-\mu)+s)^2 + 4\mu\sigma}\}.$$

In the case $r = 1$, (4.9) reduces to solve the equation in $z \equiv u(v,x) > 0$ given by

$$v + \mu(1 - x_j z_j^{-1}) + \sigma(z_j - x_j) = 0, \quad j = 1, \dots, n.$$

A simple calculation then yields the unique positive solution of this quadratic equation:

$$u_j(v,x) = (2\sigma)^{-1}\left[\sigma x_j - \mu - v_j + \sqrt{(\sigma x_j - \mu - v_j)^2 + 4\sigma\mu x_j}\right] \quad \forall j = 1, \dots, n.$$

B. Semidefinite programming, $C = S^n_{++}$. Take (as in section 3.2)

$$\begin{aligned} p(x,y) &= \operatorname{tr}(-\log x + \log y + xy^{-1}) - n \quad \forall x, y \in S^n_{++} \\ &= +\infty \quad \text{otherwise,} \end{aligned}$$

which is obtained from the Bregman kernel $h : S^n_{++} \to \mathbb{R}$ defined by $h(x) = -\ln\det(x)$. Using the fact that $\nabla h(x) = -x^{-1}$, the optimality conditions for (4.9) allow us to solve for $z \equiv u(v,x)$ the matrix equation

$$\sigma z - z^{-1} = \rho \quad \text{with} \quad \rho := \sigma x - v - x^{-1}.$$

A direct calculation shows that the matrix

$$u(v,x) = (2\sigma)^{-1}(\rho + \sqrt{\rho^2 + 4\sigma I}) \quad \forall x \in S^n_{++}, \ \forall v \in S^n$$

(where I denotes the $n \times n$ identity matrix) is the unique solution of this equation, with $u(v,x) \in S^n_{++}$, since its eigenvalues are positive.

C. Second order cone programming, $C = L^n_{++}$. As in section 3, we take $h_\nu(x) = -\log(x^T J_n x) + \frac{\nu}{2}\|x\|^2$, with J_n a diagonal matrix with its first $(n-1)$ entries being -1 and the last entry being 1. Consider the associated Bregman distance $D \equiv D_{h_\nu}$ (as given by (3.5), with $\nu \equiv 2\sigma > 0$, the multiplication by 2 being just for computational convenience):

$$D(x,y) = -\log \frac{x^T J_n x}{y^T J_n y} + \frac{2x^T J_n y}{y^T J_n y} - 2 + \sigma\|x - y\|^2 \quad \forall x, y \in L^n_{++}.$$

Moreover, we use the following notation. For any $\xi \in \mathbb{R}^n$, we set $\tau(\xi) := \xi^T J_n \xi$ and we write $\xi := (\bar{\xi}, \xi_n) \in \mathbb{R}^{n-1} \times \mathbb{R}$. Writing the optimality conditions for (4.9), we have to find the unique solution $u(v,x) \equiv z \in L^n_{++}$ (namely with $\tau(z) > 0$) solving

$$v + \nabla h(z) - \nabla h(x) + 2\sigma(z - x) = 0. \tag{4.22}$$

Using $\nabla h(z) = -2\tau(z)^{-1} J_n z$ and defining $w := (\nabla h(x) + 2\sigma x - v)/2 := (\bar{w}, w_n) \in \mathbb{R}^{n-1} \times \mathbb{R}$, (4.22) reduces to

$$\sigma z - \tau(z)^{-1} J_n z = w. \tag{4.23}$$

Decomposing (4.23) in the product space $\mathbb{R}^{n-1} \times \mathbb{R}$ yields

$$\sigma \bar{z} + \tau(z)^{-1} \bar{z} = \bar{w}, \quad \sigma z_n - \tau(z)^{-1} z_n = w_n, \tag{4.24}$$

and by eliminating $\tau(z) > 0$ from these last two equations we obtain

$$(2\sigma z_n - w_n)\bar{z} = z_n \bar{w} \iff (2\sigma \bar{z} - \bar{w}) z_n = w_n \bar{z}. \tag{4.25}$$

Now, multiplying (4.23) by z, we obtain $\sigma\|z\|^2 - w^T z - 1 = 0$, which after completing the square can be rewritten as $\|2\sigma\bar{z} - \bar{w}\|^2 + (2\sigma z_n - w_n)^2 = \|w\|^2 + 4\sigma$. Using (4.25) and defining $\zeta := 2\sigma z_n - w_n$, the last equation reads

$$\frac{w_n^2 \|\bar{w}\|^2}{\zeta^2} + \zeta^2 = \|w\|^2 + 4\sigma. \tag{4.26}$$

Now, it is easy to verify that $\zeta > 0$. Indeed, since $z \in L^n_{++}$, then $z_n > 0$, and by (4.24) one also has $w_n < \sigma z_n$, and it follows that $\zeta = 2\sigma z_n - w_n > \sigma z_n - w_n > 0$. Out of the two remaining solutions of (4.26), a direct computation (using the fact that $(\|w\|^2 + 4\sigma)^2 - 4w_n^2\|\bar{w}\|^2 = (w_n^2 - \|\bar{w}\|^2 + 4\sigma)^2 + 16\sigma\|\bar{w}\|^2)$ shows that the unique positive solution of (4.26) that will warrant $\tau(z) > 0$ is given by the following:

$$\zeta = \left(\frac{\|w\|^2 + 4\sigma + \sqrt{(\|w\|^2 + 4\sigma)^2 - 4w_n^2\|\bar{w}\|^2}}{2} \right)^{1/2}.$$

Therefore using (4.25) it follows that the unique solution $u \equiv z \in L^n_{++}$ of (4.22) is given by $z = (\bar{z}, z_n)$ with

$$\bar{z} = \frac{z_n}{\zeta}\bar{w} = \frac{1}{2\sigma}\left(1 + \frac{w_n}{\zeta}\right)\bar{w}, \quad z_n = \frac{1}{2\sigma}(w_n + \zeta). \tag{4.27}$$

Remark 4.1. It is worthwhile to mention that an alternative derivation of (4.27) could also have been obtained by using properties and facts on Jordan algebra associated with the second order cone; see, e.g., [22, 23].

D. Convex minimization over the unit simplex. An interesting special case of a conic optimization, with $\mathcal{V} \neq \mathbb{R}^n$, where $u(v,x)$ can be explicitly given, and where all this theory applies, is when $\overline{C} = \mathbb{R}^n_+$ and $A = e^T$, $b = 1$, i.e., $\mathcal{V} = \{x \in \mathbb{R}^n \mid \sum_{j=1}^n x_j = 1\}$, so that problem (M) reduces to a convex minimization problem over the unit simplex $\Delta = \{x \in \mathbb{R}^n \mid \sum_{j=1}^n x_j = 1,\ x \geq 0\}$. This problem arises in important applications. In [9], Ben-Tal, Margalit, and Nemirovski demonstrated that an algorithm based on the mirror descent (MDA) can be successfully used to solve very large-scale instances of computerized tomography problems, modeled through (M). Recently, Beck and Teboulle [8] have shown that the MDA can be viewed as a projection subgradient algorithm with strongly convex Bregman proximal distances. As a result, to handle the simplex constraints Δ, they proposed to use a Bregman proximal distance based on the entropy kernel

$$\psi(x) = \begin{cases} \sum_{j=1}^{n} x_j \log x_j & \text{if } x \in \Delta, \\ +\infty & \text{otherwise} \end{cases} \tag{4.28}$$

to produce an entropic mirror descent algorithm (EMDA). It was shown in [8] that the EMDA preserved the same computational efficiency as the MDA (grows slowly with the dimension of the problem), but has the advantage of being given explicitly by a simple formula, since the problem

$$u(v,x) = \operatorname*{argmin}_{z\in\Delta}\{\langle v, z\rangle + D_\psi(z,x)\} \tag{4.29}$$

can be easily solved analytically and yields

$$u_j(v,x) = \frac{x_j \exp(-v_j)}{\sum_{i=1}^{n} x_i \exp(-v_i)}, \quad j = 1,\dots,n. \tag{4.30}$$

The resulting EMDA of [8] was then defined as follows: for each $j = 1,\dots,n$ with $v_j = \frac{\partial f}{\partial x_j}(x^{k-1})$,

$$x_j^k(\lambda_k) = u_j(\lambda_k v, x^{k-1}) = \frac{x_j^{k-1} \exp\left(-\lambda_k \frac{\partial f}{\partial x_j}(x^{k-1})\right)}{\sum_{i=1}^{n} x_i^{k-1} \exp\left(-\lambda_k \frac{\partial f}{\partial x_i}(x^{k-1})\right)}, \tag{4.31}$$

$$\lambda_k = \frac{\sqrt{2\log k}}{L_f\sqrt{k}}, \tag{4.32}$$

where the objective function was supposed to be Lipschitz on Δ and L_f is the Lipschitz constant.

We can modify the EMDA with an Armijo–Goldstein step-size rule. Such a version of the EMDA can be more practical, since we do not need to know/compute the constant L_f. Indeed, it is well known (see, e.g., [8]) that

$$\langle \nabla\psi(x) - \nabla\psi(y), x-y\rangle \ge \|x-y\|_1^2 \quad \forall x,y \in \Delta_+ = \left\{ x\in\mathbb{R}^n \;\middle|\; \sum_{j=1}^{n} x_j = 1,\ x>0 \right\},$$

namely ψ is 1-strongly convex with respect to the norm $\|\cdot\|_1$, and hence so is $d = H = D_\psi$. Therefore we can apply Theorem 4.3, proving that the sequence $\{x^k\}$ defined by (4.31), and with λ_k defined by the Armijo–Goldstein step-size rule (4.17), converges to an optimal solution of (M).

5. Interior gradient methods with improved efficiency.

In this section, we further analyze the global convergence rate of interior gradient methods, and we propose a new interior scheme which improves their efficiency. The classical gradient method for minimizing a continuously differentiable function over $\mathbb{R}^n$ with Lipschitz gradient is known to exhibit an $O(k^{-1})$ global convergence rate estimate for function values. In [34], Nesterov developed what he called an "optimal algorithm" for smooth convex minimization and was able to improve the efficiency of the gradient method by constructing a method that keeps the simplicity of the gradient method but with the faster rate $O(k^{-2})$. Inspired by this work, it is thus natural to ask if this kind of result can be extended to interior gradient methods. We answer this question positively for a class of interior gradient methods. We propose an algorithm that provides a natural extension of the results of [34] and leads to a simple "optimal" interior gradient method for convex conic problems.

Consider the conic optimization problem as described in section 4.2.1, i.e.,

$$\text{(M)} \qquad \inf\{f(x) : x \in \overline{C} \cap \mathcal{V}\},$$

where $\mathcal{V} := \{x \in \mathbb{R}^n \mid Ax = b\}$, with $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$, $n \geq m$, $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex and lsc, and we assume that $\exists x^0 \in \operatorname{dom} f \cap C : Ax^0 = b$. We assume also that f is continuously differentiable with ∇f Lipschitz on $C \cap \mathcal{V}$ and Lipschitz constant L, i.e., satisfying (4.7).

The basic idea is to generate a sequence of functions $\{q_k\}$ that approximate the function f in such a way that at each step $k \geq 0$ the difference $q_k(x) - f(x)$ is reduced by a fraction $(1 - \alpha_k)$, where $\alpha_k \in [0, 1)$, that is,

$$q_{k+1}(x) - f(x) \leq (1 - \alpha_k)(q_k(x) - f(x)) \quad \forall x \in \overline{C} \cap \mathcal{V}. \tag{5.1}$$

Whenever (5.1) holds, we then obtain

$$q_k(x) - f(x) \leq \gamma_k(q_0(x) - f(x)) \quad \forall x \in \overline{C} \cap \mathcal{V}, \tag{5.2}$$

where

$$\gamma_k := \prod_{l=0}^{k-1}(1 - \alpha_l). \tag{5.3}$$

Thus, if at step k we have a sequence $\{x^k\} \in C \cap \mathcal{V}$ such that $f(x^k) \leq \inf_{z\in\overline{C}\cap\mathcal{V}} q_k(z) := q_k^*$, assuming that the optimal solution set X_* of problem (P) is nonempty, we obtain from (5.2) the global convergence rate estimate

$$f(x^k) - f(x^*) \leq \gamma_k(q_0(x^*) - f(x^*)). \tag{5.4}$$

From the latter inequality it follows that if $\gamma_k \to 0$, then the sequence $\{x^k\}$ is a minimizing sequence for f and the convergence rate of $f(x^k)$ to $f(x^*)$ is measured by the magnitude of γ_k. Therefore to construct algorithms based on the above scheme which was proposed in [34] we need

- to generate an appropriate sequence of functions $\{q_k(\cdot)\}$,
- to guarantee that at each iteration k one can guarantee

$$f(x^k) \leq \min_{z\in\overline{C}\cap\mathcal{V}} q_k(z) := q_k^*.$$

We begin by constructing the sequence of functions $\{q_k(\cdot)\}$. For that purpose, we take here $d \equiv H \in \mathcal{D}(C)$, where H is a Bregman proximal distance (cf. (3.1)) with kernel h such that

(h1) $\operatorname{dom} h = \overline{C}$,

(h2) h is σ-strongly convex on $C \cap \mathcal{V}$.

For every $k \geq 0$ and for any $x \in \overline{C} \cap \mathcal{V}$, we construct the sequence $\{q_k(x)\}$ recursively via

$$q_0(x) = f(x^0) + cH(x, x^0), \tag{5.5}$$

$$q_{k+1}(x) = (1 - \alpha_k)q_k(x) + \alpha_k l_k(x, y^k), \tag{5.6}$$

$$l_k(x, y^k) = f(y^k) + \langle x - y^k, \nabla f(y^k)\rangle. \tag{5.7}$$

Here, $c > 0$ and $\alpha_k \in [0, 1)$. The point x^0 is chosen such that $x^0 \in C \cap \mathcal{V}$, while the point $y_k \in C$ is arbitrary and will be generated in a specific way later. We first show that the sequence of functions $\{q_k(\cdot)\}$ satisfies (5.1).

Lemma 5.1. *The sequence $\{q_k(x)\}$ defined by* (5.5)–(5.7) *satisfies*

$$q_{k+1}(x) - f(x) \leq (1 - \alpha_k)(q_k(x) - f(x)) \quad \forall x \in \overline{C} \cap \mathcal{V}.$$

Proof. Since f is convex, we have $f(x) \geq l_k(x, y^k)$ $\forall x \in \overline{C} \cap \mathcal{V}$, and together with (5.6) we thus obtain

$$q_{k+1}(x) \leq (1 - \alpha_k)q_k(x) + \alpha_k f(x) \quad \forall x \in \overline{C} \cap \mathcal{V},$$

from which the desired result follows. □

Using the notation of section 4, we recall that for each $z \in C \cap \mathcal{V}$, for each $v \in \mathbb{R}^n$ there exists a unique (by strong convexity of $H(\cdot, z)$) point $u(v, z) \in C \cap \mathcal{V}$ solving

$$u(v, z) = \operatorname{argmin}\{\langle v, x\rangle + H(x, z) \mid x \in \overline{C} \cap \mathcal{V}\}. \tag{5.8}$$

The next result is crucial and shows that the sequence $\{q_k(\cdot)\}$ admits a simple generic form.

Lemma 5.2. *For any $k \geq 0$, one has*

$$q_k(x) = q_k^* + c_k H(x, z^k) \quad \forall x \in \overline{C} \cap \mathcal{V} \tag{5.9}$$

with

$$z^k = \operatorname*{argmin}_{x \in \overline{C} \cap \mathcal{V}} q_k(x),\ q_k^* = q_k(z^k),\ c_0 = c,\ z^0 = x^0 \in C \cap \mathcal{V}. \tag{5.10}$$

Furthermore, the sequence $\{z^k\} \in C \cap \mathcal{V}$ is uniquely defined by

(5.11)

$$z^{k+1} = \operatorname{argmin}\left\{\left\langle x, \frac{\alpha_k}{c_{k+1}}\nabla f(y^k)\right\rangle + H(x, z^k) \;\middle|\; x \in \overline{C} \cap \mathcal{V}\right\} \equiv u\left(\frac{\alpha_k}{c_{k+1}}\nabla f(y^k), z^k\right),$$

where the positive sequence $\{c_k\}$ satisfies $c_{k+1} = (1 - \alpha_k)c_k$.

Proof. The proof is by induction and will use key identity (3.2). For $k = 0$, since $z^0 = x^0$ by (5.5), one has $q_0(x) = f(x^0) + cH(x, z^0)$. Then since $c\nabla_1 H(z^0, z^0) = 0$ (recall the properties of H), and since $z_0 \in C \cap \mathcal{V}$, the optimality conditions imply that $z^0 = \operatorname{argmin}_{x \in \overline{C} \cap \mathcal{V}} q_0(x)$. Now suppose that (5.9) holds for some k and let us prove that for any $x \in \overline{C} \cap \mathcal{V}$,

$$q_{k+1}(x) = q_{k+1}^* + c_{k+1}H(x, z^{k+1}). \tag{5.12}$$

Substituting (5.9) into (5.6) and using $c_{k+1} = (1 - \alpha_k)c_k$, one obtains

$$q_{k+1}(x) = (1 - \alpha_k)q_k^* + c_{k+1}H(x, z^k) + \alpha_k l_k(x, y^k). \tag{5.13}$$

Then by definition of z^{k+1} we have

$$z^{k+1} = \operatorname*{argmin}_{x \in \overline{C} \cap \mathcal{V}} q_{k+1}(x) = u\left(\frac{\alpha_k}{c_{k+1}} \nabla f(y^k), z^k\right)$$

with $z^{k+1} \in C \cap \mathcal{V}$, and

$$q^*_{k+1} = q_{k+1}(z^{k+1}) = (1-\alpha_k)q^*_k + c_{k+1}H(z^{k+1}, z^k) + \alpha_k l_k(z^{k+1}, y^k). \tag{5.14}$$

Subtracting (5.14) from (5.13), one obtains, using (5.7),

$$\begin{aligned} q_{k+1}(x) &= q^*_{k+1} + c_{k+1}[H(x,z^k) - H(z^{k+1},z^k)] + \alpha_k[l_k(x,y^k) - l_k(z^{k+1},y^k)] \\ &= q^*_{k+1} + c_{k+1}[H(x,z^k) - H(z^{k+1},z^k)] + \alpha_k\langle z^{k+1} - x, -\nabla f(y^k)\rangle. \end{aligned} \tag{5.15}$$

Now, since $z^{k+1} = \operatorname{argmin}_{x \in \overline{C} \cap \mathcal{V}} q_{k+1}(x)$, then writing the optimality conditions for (5.13) (recalling the properties of H) yields

$$c_{k+1}\langle \nabla_1 H(z^{k+1}, z^k), z^{k+1} - x\rangle = -\langle \alpha_k \nabla f(y^k), z^{k+1} - x\rangle \quad \forall x \in \overline{C} \cap \mathcal{V}. \tag{5.16}$$

Using (5.16) in (5.15), it follows that for any $x \in \overline{C} \cap \mathcal{V}$,

$$q_{k+1}(x) = q^*_{k+1} + c_{k+1}[H(x,z^k) - H(z^{k+1},z^k) + \langle z^{k+1} - x, \nabla_1 H(z^{k+1}, z^k)\rangle]. \tag{5.17}$$

Invoking the identity (3.2) at $c = x$, $b = z^{k+1}$, and $a = z^k$, the right-hand side of (5.17) reduces to $q_{k+1}(x) = q^*_{k+1} + c_{k+1}H(x, z^{k+1})$, and the lemma is proved. □

The next result is fundamental to determining the main steps of the algorithm, namely the formulas needed to update the sequence $\{x^k\}$ and to determine the choice of the intermediary point y^k.

THEOREM 5.1. *Let $\sigma > 0$, $L > 0$ be given. Suppose that for some $k \geq 0$ we have a point $x^k \in C \cap \mathcal{V}$ such that $f(x^k) \leq q^*_k = \min\{q_k(x) : x \in \overline{C} \cap \mathcal{V}\}$. Let $\alpha_k \in [0,1)$, $c_{k+1} = (1-\alpha_k)c_k$, and $C \cap \mathcal{V} \ni \{z^k\}$ be given by (5.11). Define*

$$y^k = (1-\alpha_k)x^k + \alpha_k z^k, \tag{5.18}$$

$$x^{k+1} = (1-\alpha_k)x^k + \alpha_k z^{k+1}. \tag{5.19}$$

Then, the following inequality holds:

$$q^*_{k+1} \geq f(x^{k+1}) + \frac{1}{2}\left(\frac{c_{k+1}\sigma}{\alpha_k^2} - L\right)\|x^{k+1} - y^k\|^2.$$

Proof. Let $x \in \overline{C} \cap \mathcal{V}$. Since $q_k(x) = q^*_k + c_k H(x, z^k)$, then by (5.6) and using $c_{k+1} = (1-\alpha_k)c_k$ one has

$$q_{k+1}(x) = (1-\alpha_k)q^*_k + c_{k+1}H(x,z^k) + \alpha_k l_k(x, y^k),$$

and with $z^{k+1} = \operatorname{argmin}_{x \in \overline{C} \cap \mathcal{V}} q_{k+1}(x)$ one obtains

$$q_{k+1}(z^{k+1}) = q^*_{k+1} = (1-\alpha_k)q^*_k + c_{k+1}H(z^{k+1}, z^k) + \alpha_k l_k(z^{k+1}, y^k). \tag{5.20}$$

Under our assumption, we have $q^*_k \geq f(x^k)$, and thus using the gradient inequality for f we have

$$q^*_k \geq f(x^k) \geq f(y^k) + \langle x^k - y^k, \nabla f(y^k)\rangle,$$

and it follows from (5.20) and (5.7) that

$$q^*_{k+1} \geq f(y^k) + c_{k+1} H(z^{k+1}, z^k) + \langle \nabla f(y^k), r^k \rangle, \tag{5.21}$$

where $r^k = \alpha_k(z^{k+1} - y^k) + (1 - \alpha_k)(x^k - y^k)$. Noting that r^k can be written as

$$r^k = (1 - \alpha_k)x^k + \alpha_k z^k - y^k + \alpha_k(z^{k+1} - z^k),$$

and since by definition one has $(1 - \alpha_k)x^k + \alpha_k z^k - y^k = 0$, then (5.21) reduces to

$$q^*_{k+1} \geq f(y^k) + c_{k+1} H(z^{k+1}, z^k) + \langle \alpha_k(z^{k+1} - z^k), \nabla f(y^k) \rangle. \tag{5.22}$$

Using the definition of $y^k, x^{k+1} \in C \cap \mathcal{V}$ given in (5.18)–(5.19), one has $x^{k+1} - y^k = \alpha_k(z^{k+1} - z^k)$. Since by hypothesis (h2) h is σ-strongly convex, it follows that $H(z^{k+1}, z^k) \geq \sigma/2\|z^{k+1} - z^k\|^2$, and then from (5.22) we have obtained

$$q^*_{k+1} \geq f(y^k) + \frac{1}{2}\frac{c_{k+1}\sigma}{\alpha_k^2}\|x^{k+1} - y^k\|^2 + \langle \nabla f(y^k), x^{k+1} - y^k \rangle. \tag{5.23}$$

Now, since we assumed that f in $C^{1,1}(C \cap \mathcal{V})$, then by the descent lemma (cf. (4.15)) we have

$$f(y^k) + \langle x^{k+1} - y^k, \nabla f(y^k) \rangle \geq f(x^{k+1}) - \frac{L}{2}\|x^{k+1} - y^k\|^2. \tag{5.24}$$

Combining the latter inequality with (5.23) we obtain

$$q^*_{k+1} \geq f(x^{k+1}) + \frac{1}{2}\left(\frac{c_{k+1}\sigma}{\alpha_k^2} - L\right)\|x^{k+1} - y^k\|^2. \qquad \square$$

Therefore by taking a sequence $\{\alpha_k\}$ with $\sigma c_{k+1} \geq L\alpha_k^2$ we can guarantee that $q^*_{k+1} \geq f(x^{k+1})$. In particular, we can choose $L\alpha_k^2 = \sigma c_k(1 - \alpha_k)$, and this leads to the following improved interior gradient algorithm.

Improved interior gradient algorithm (IGA).

Step 0. Choose a point $x^0 \in C \cap \mathcal{V}$ and a constant $c > 0$. Define $z^0 = x^0 = y^0$, $c_0 = c$, $\lambda = \sigma L^{-1}$.

Step *k*. For $k \geq 0$, compute the following:

$$\begin{aligned}
\alpha_k &= \frac{\sqrt{(c_k\lambda)^2 + 4c_k\lambda} - \lambda c_k}{2},\\
y^k &= (1 - \alpha_k)x^k + \alpha_k z^k,\\
c_{k+1} &= (1 - \alpha_k)c_k,\\
z^{k+1} &= \operatorname*{argmin}_{x \in \overline{C} \cap \mathcal{V}} \left\{ \left\langle x, \frac{\alpha_k}{c_{k+1}}\nabla f(y^k) \right\rangle + H(x, z^k) \right\} = u\left(\frac{\alpha_k}{c_{k+1}}\nabla f(y^k), z^k\right),\\
x^{k+1} &= (1 - \alpha_k)x^k + \alpha_k z^{k+1}.
\end{aligned}$$

Note that the computational work of this algorithm is exactly the same as that of the interior gradient method in section 4 via the computation of z^{k+1}, since the remaining steps involve trivial computations. To estimate the rate of convergence we need the following simple lemma on the sequence α_k; see [34] for a proof.

LEMMA 5.3. *Let $\lambda_k > 0$, $c_k > 0$ with $c_0 = c$, and let $\{\alpha_k\}$ be the sequence with $\alpha_k \in [0,1[$ defined by $\alpha_k^2 = \lambda_k c_k(1-\alpha_k)$ with $c_{k+1} = (1-\alpha_k)c_k$. Set $\gamma_k := \prod_{l=0}^{k-1}(1-\alpha_l)$. Then*

$$\gamma_k \le \left(1 + \frac{\sqrt{c}}{2}\sum_{l=0}^{k-1}\sqrt{\lambda_l}\right)^{-2}.$$

In particular, with $\lambda_l = \lambda\ \forall l$ we have $\gamma_k \le 4(k\sqrt{\lambda c}+2)^{-2}$.

We thus obtain a convergent interior gradient method with an improved convergence rate estimate.

THEOREM 5.2. *Let $\{x^k\}, \{y^k\}$ be the sequences generated by IGA and let x^* be an optimal solution of* (P). *Then for any $k \ge 0$ we have*

$$f(x^k) - f(x^*) \le \frac{4L}{\sigma k^2 c}C(x^*, x^0) = O\left(\frac{1}{k^2}\right),$$

where $C(x^*, x^0) = c_0 H(x^*, x^0) + f(x^0) - f(x^*)$ and the sequence $\{x^k\}$ is minimizing, i.e., $f(x^k) \to f(x^*)$.

Proof. By Lemma 5.1, the sequence of functions $\{q_k(\cdot)\}$ satisfies (5.1) and thus (5.4) holds; i.e., using (5.5) we have

$$f(x^k) - f(x^*) \le \gamma_k(q_0(x^*) - f(x^*)) = \gamma_k(f(x^0) + c_0H(x^*, x^0) - f(x^*)) = \gamma_k C(x^*, x^0).$$

Specializing Lemma 5.3 with $\lambda_k = \sigma L^{-1}$, we obtain

$$\gamma_k \le \frac{4L}{\left(k\sqrt{\sigma c} + 2\sqrt{L}\right)^2} \le \frac{4L}{\sigma c k^2},$$

from which the desired result follows. □

Thus, to solve (P) to accuracy $\varepsilon > 0$, one needs no more than $\lfloor O(1/\sqrt{\varepsilon})\rfloor$ iterations of IGA, which is a significant reduction (by a squared root factor) in comparison to the interior gradient method of section 4. In particular, we note that IGA can be used to solve convex minimization over the unit simplex with this improved global convergence rate estimate for the EMDA of section 4.

REFERENCES

[1] F. ALVAREZ, J. BOLTE, AND O. BRAHIC, *Hessian Riemannian gradient flows in convex programming*, SIAM J. Control Optim., 43 (2004), pp. 477–501.

[2] H. ATTOUCH AND M. TEBOULLE, *A regularized Lotka Volterra dynamical system as a continuous proximal-like method in optimization*, J. Optim. Theory Appl., 121 (2004), pp. 541–570.

[3] A. AUSLENDER AND M. HADDOU, *An interior proximal method for convex linearly constrained problems and its extension to variational inequalities*, Math. Program., 71 (1995), pp. 77–100.

[4] A. AUSLENDER, M. TEBOULLE, AND S. BEN-TIBA, *A logarithmic-quadratic proximal method for variational inequalities*, Comput. Optim. Appl., 12 (1999), pp. 31–40.

[5] A. AUSLENDER, M. TEBOULLE, AND S. BEN-TIBA, *Interior proximal and multiplier methods based on second order homogeneous kernels*, Math. Oper. Res., 24 (1999), pp. 645–668.

[6] A. AUSLENDER AND M. TEBOULLE, *Asymptotic Cones and Functions in Optimization and Variational Inequalities*, Springer Monogr. Math., Springer-Verlag, New York, 2003.

[7] A. AUSLENDER AND M. TEBOULLE, *Interior gradient and epsilon-subgradient methods for constrained convex minimization*, Math. Oper. Res., 29 (2004), pp. 1–26.

[8] A. BECK AND M. TEBOULLE, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Oper. Res. Lett., 31 (2003), pp. 167–175.

[9] A. Ben-Tal, T. Margalit, and A. Nemirovski, *The ordered subsets mirror descent optimization method with applications to tomography*, SIAM J. Optim., 12 (2001), pp. 79–108.

[10] A. Ben-Tal and M. Zibulevsky, *Penalty/barrier methods for convex programming problems*, SIAM J. Optim., 7 (1997), pp. 347–366.

[11] D. P. Bertsekas, *On the Goldstein-Levitin-Polyak gradient projection method*, IEEE Trans. Automat. Control, 21 (1976), pp. 174–183.

[12] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.

[13] J. Bolte and M. Teboulle, *Barrier operators and associated gradient-like dynamical systems for constrained minimization problems*, SIAM J. Control Optim., 42 (2003), pp. 1266–1292.

[14] H. Brezis, *Analyse Fonctionnelle: Theorie et applications*, Masson, Paris, 1987.

[15] Y. Censor and S. Zenios, *The proximal minimization algorithm with D-functions*, J. Optim. Theory Appl., 73 (1992), pp. 451–464.

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

[17] R. Correa and C. Lemarechal, *Convergence of some algorithm for convex programming*, Math. Program., 62 (1993), pp. 261–275.

[18] M. Doljansky and M. Teboulle, *An interior proximal algorithm and the exponential multiplier method for semidefinite programming*, SIAM J. Optim., 9 (1998), pp. 1–13.

[19] J. Eckstein, *Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming*, Math. Oper. Res., 18 (1993), pp. 202–226.

[20] J. Eckstein, *Approximate iterations in Bregman-function-based proximal algorithms*, Math. Program., 83 (1998), pp. 113–123.

[21] P. P. B. Eggermont, *Multiplicatively iterative algorithms for convex programming*, Linear Algebra Appl., 130 (1990), pp. 25–42.

[22] J. Faraut and A. Korányi, *Analysis on Symmetric Cones*, Oxford Math. Monogr., The Claredon Press, Oxford University Press, New York, 1994.

[23] M. Fukushima, Z.-Q. Luo, and P. Tseng, *Smoothing functions for second-order-cone complementarity problems*, SIAM J. Optim., 12 (2001), pp. 436–460.

[24] O. Güler, *On the convergence of the proximal point algorithm for convex minimization*, SIAM J. Control Optim., 29 (1991), pp. 403–419.

[25] A. N. Iusem, *Interior point multiplicative methods for optimization under positivity constraints*, Acta Appl. Math., 38 (1995), pp. 163–184.

[26] A. N. Iusem, B. F. Svaiter, and M. Teboulle, *Multiplicative interior gradient methods for minimization over the nonnegative orthant*, SIAM J. Control Optim., 34 (1996), pp. 389–406.

[27] A. N. Iusem, B. Svaiter, and M. Teboulle, *Entropy-like proximal methods in convex programming*, Math. Oper. Res., 19 (1994), pp. 790–814.

[28] A. N. Iusem and M. Teboulle, *Convergence rate analysis of nonquadratic proximal and augmented Lagrangian methods for convex and linear programming*, Math. Oper. Res., 20 (1995), pp. 657–677.

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

[30] B. Lemaire, *The proximal algorithm*, in New Methods in Optimization and Their Industrial Uses, Internat. Schriftenreihe Numer. Math. 87, J. P. Penot, ed., Birkhäuser, Basel, 1989, pp. 73–87.

[31] B. Martinet, *Regularisation d'inéquations variationnelles par approximations successives*, Rev. Française Informat. Recherche Opérationnelle, 4 (1970), pp. 154–158.

[32] A. Nemirovski and D. Yudin, *Problem Complexity and Method Efficiency in Optimization*, John Wiley, New York, 1983.

[33] Y. Nesterov and A. Nemirovskii, *Interior Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.

[34] Y. Nesterov, *On an approach to the construction of optimal methods of minimization of smooth convex functions*, Èkonom. i Mat. Metody, 24 (1988), pp. 509–517.

[35] B. T. Polyak, *Introduction to Optimization*, Optimization Software, New York, 1987.

[36] R. A. Polyak, *Nonlinear rescaling vs. smoothing technique in constrained optimization*, Math. Program., 92 (2002), pp. 197–235.

[37] R. A. Polyak and M. Teboulle, *Nonlinear rescaling and proximal-like methods in convex optimization*, Math. Program., 76 (1997), pp. 265–284.

[38] S. M. Robinson, *Linear convergence of epsilon subgradients methods for a class of convex functions*, Math. Program., 86 (1999), pp. 41–50.

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

[40] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.

[41] N. Z. Shor, *Minimization Methods for Nondifferentiable Functions*, Springer-Verlag, Berlin, 1985.

[42] P. J. da Silva e Silva, J. Eckstein, and C. Humes, Jr., *Rescaling and stepsize selection in proximal methods using separable generalized distances*, SIAM J. Optim., 12 (2001), pp. 238–261.

[43] M. Teboulle, *Entropic proximal mappings with applications to nonlinear programming*, Math. Oper. Res., 17 (1992), pp. 670–681.

[44] M. Teboulle, *Convergence of proximal-like algorithms*, SIAM J. Optim., 7 (1997), pp. 1069–1083.

[45] P. Tseng and D. P. Bertsekas, *On the convergence of the exponential multiplier method for convex programming*, Math. Program., 60 (1993), pp. 1–19.