On Accelerated Proximal Gradient Methods for Convex-Concave Optimization[1]

May 21, 2008

Paul Tseng[2]

Abstract

Recently there has been active interest in accelerated proximal gradient methods for large-scale convex-concave optimization, as studied by Nesterov, Nemirovski, and others. We present a unified treatment of these methods, including new variants that perform either one or two projections per iteration, and give simple analyses of their iteration complexity. These methods are compared on a matrix game example.

Key words. Convex-concave optimization, monotone variational inequality, proximal point, gradient projection, interpolation, Bregman function, iteration complexity.

1 Introduction

Let $\mathcal{E}$ be a real linear space endowed with a norm $\|\cdot\|$. Let $\mathcal{E}^*$ be the vector space of continuous linear functionals on $\mathcal{E}$, endowed with the dual norm $\|x^*\|_* = \sup_{\|x\|\leq 1}\langle x^*, x\rangle$, where $\langle x^*, x\rangle$ denotes the value of $x^* \in \mathcal{E}^*$ at $x \in \mathcal{E}$. Consider the nonsmooth convex optimization problem

$$\min_x f^P(x) := f(x) + P(x), \tag{1}$$

where $P : \mathcal{E} \to (-\infty, \infty]$ and $f : \mathcal{E} \to (-\infty, \infty]$ are proper, lower semicontinuous (lsc), convex [38, 39]. We assume that dom$P$ is closed, f is differentiable on an open set containing dom$P$, and ∇f is Lipschitz continuous on dom$P$, i.e.,

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \quad \forall x, y \in \mathrm{dom}P, \tag{2}$$

for some $L > 0$. This class of problems was studied in [2, 17] and by others; see [45] and references therein. A well-known special case is smooth constrained convex optimization, for which P is the indicator function for a nonempty closed convex set $X \subseteq \mathcal{E}$, i.e.,

$$P(x) = \begin{cases} 0 & \text{if } x \in X; \\ \infty & \text{else.} \end{cases} \tag{3}$$

A second special case that has received much attention lately is ℓ_1-regularization, for which P is the 1-norm, as in lasso and basis pursuit and sparse covariance selection; see [1, 12, 16, 22, 43, 45, 47] and references therein. A third special case is the group lasso for regression, for which P is a sum of weighted 2-norms, i.e.,

$$P(x) = c_1\|x_1\|_2 + \cdots + c_N\|x_N\|_2, \tag{4}$$

where $x_1, \ldots, x_N$ denote disjoint subvectors of x and $c_j > 0$ for all j; see [25] and references therein. While (1) can be transformed into a smooth constrained convex optimization problem

$$\min_{x,\zeta}\{f(x) + \zeta \mid P(x) \leq \zeta\}, \tag{5}$$

we will see that this has undesirable consequences for computation.

[1] This work is supported by National Science Foundation grant DMS-0511283.

[2] Department of Mathematics, University of Washington, Seattle, WA 98195, U.S.A. (tseng@math.washington.edu)

For any $y \in \mathrm{dom}P$, consider the approximation of f^P by replacing f with its linear approximation at y:

$$\ell_f(x;y) := f(y) + \langle \nabla f(y), x - y \rangle + P(x).$$

The convexity of f and (2) imply that

$$f^P(x) \geq \ell_f(x;y) \geq f^P(x) - \frac{L}{2}\|x-y\|^2 \quad \forall x, y \in \mathrm{dom}P. \tag{6}$$

Choose a strictly convex function $h : \mathcal{E} \to (-\infty, \infty]$ that is differentiable on an open set containing $\mathrm{dom}P$,[3] and consider the corresponding "distance/proximity" function

$$D(x,y) := h(x) - h(y) - \langle \nabla h(y), x - y \rangle \quad \forall y \in \mathrm{dom}P,\ \forall x \in \mathcal{E},$$

which was studied by Bregman [10] and many others; see [4, 5, 18, 42] and references therein. By scaling h if necessary, we assume that

$$D(x,y) \geq \frac{1}{2}\|x-y\|^2 \qquad \forall\ x, y \in \mathrm{dom}P. \tag{7}$$

Then, for any $y, z \in \mathrm{dom}P$ and any $\alpha > 0$, the function $x \mapsto \ell_f(x;y) + \alpha D(x,z)$ has a unique minimum point and it lies in $\mathrm{dom}P$.

The classical gradient-projection method of Goldstein and Levitin, Polyak [9] naturally generalizes to solve (1), with constant stepsize $1/L$ and with D used in the nearest-point projection:

$$x_{k+1} = \arg\min_x \left\{ \ell_f(x;x_k) + LD(x,x_k) \right\}, \quad k = 0, 1, \ldots, \tag{8}$$

where $x_0 \in \mathrm{dom}P$. In the case of a quadratic kernel $h(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, this method was studied early on by Fukushima and Mine [17]. If P is the 1-norm or has the block-separable form (4), the new point x_{k+1} can be found in closed form, which is one key advantage of this method for large-scale optimization; see [45, 47] and references therein. This contrasts with applying the same method to the transformed problem (5), which requires projecting onto $\mathrm{epi}P = \{(x,\zeta) \mid P(x) \leq \zeta\}$ and is nontrivial even in the case of P being the 1-norm. In the smooth constrained case (3), the method (8) is closely related to the mirror descent method of Nemirovski and Yudin [27], as is discussed in [4, 6]. When X is the unit simplex, x_{k+1} can be found in closed form by taking $h(x)$ to be the $x \ln x$-entropy function [6, Section 5], [32, Lemma 4]; see Section 7 on dealing with $x \ln x$ being nondifferentiable at 0. Moreover, the corresponding $D(\cdot,\cdot)$ satisfies (7) with $\|\cdot\|$ being the 1-norm [6, Proposition 5.1], [32, Lemma 3]. It can be shown that

$$f^P(x_k) - \inf f^P \leq O\left(\frac{L}{k}\right) \quad \forall k,$$

and hence $O(L/\epsilon)$ iterations suffice to come within $\epsilon > 0$ of the optimal value; see, e.g., [31, Theorem 2.1.14], [36, page 166], [46]; also see [13] for a related analysis under certain local growth conditions on f.

In a series of work [28, 29, 32] (also see [36, page 171]), Nesterov proposed three methods for solving the smooth constrained case (3) that, at each iteration, use either one or two projection steps together with interpolation to accelerate convergence. These methods generate points x_k that achieve remarkably

$$f^P(x_k) - \inf f^P \leq O\left(\frac{L}{k^2}\right) \quad \forall k,$$

so that $O(\sqrt{L/\epsilon})$ iterations suffice to come within $\epsilon > 0$ of the optimal value. In [32], it is shown that various large convex-concave optimization problems can be efficiently solved by applying these methods to a smooth approximation with Lipschitz constant $L = O(1/\epsilon)$. Nesterov's second method [29] was further studied in [4, Section 5] using Bregman functions (also see [31, Section 2.2]), and his third method [32] was

[3] This assumption can be further relaxed as is discussed in Section 7.

applied in [1, 22, 23, 30] to sparse covariance selection, rank reduction in multivariate linear regression, and eigenvalue optimization; also see [37, Section 2.3]. It was further shown by Nesterov [32, Theorem 3] and generalized by Lu [22, Theorem 2.2] that if $\mathcal{E}$ is finite-dimensional, X in (3) is bounded, and f has the form

$$f(x) = \max_{v \in V} \phi(x, v), \tag{9}$$

then x_{k+1} generated by Nesterov's third method satisfies $f^P(x_{k+1}) - q^P(\bar{v}_k) = O(L/k^2)$, where q^P is the dual function

$$q^P(v) := \min_x \{\phi(x, v) + P(x)\}, \tag{10}$$

and $\bar{v}_k$ is a weighted sum of dual vectors associated with $x_0, x_1, \ldots, x_k$; also see Corollary 3(c). Here, V is a compact convex set in a finite-dimensional real vector space $\mathcal{F}$, $\phi : \mathcal{E} \times V \to (-\infty, \infty]$ is continuous on dom$P \times V$, $\phi(\cdot, v)$ is convex and differentiable on an open set containing dom$P$ for every $v \in V$, and $\phi(x, \cdot)$ is strictly concave for every $x \in$ dom$P$. The above duality gap provides an effective termination criterion [22, 23, 32]. The saddle structure (9) is further exploited in the primal-dual method of [33], which combines gradient projection with dynamically adjusted primal-dual smoothing. A variant of Nesterov's second method [29] was recently proposed by Lan, Lu, and Monteiro [21], with comparable iteration complexity and improved performance reported on random generated LP and SDP problems. These gradient methods can be significantly faster than interior-point methods on large-scale problems. Very recently, Nesterov extended his third method [32] with $D(x, y) = \frac{1}{2}\|x - y\|^2$ to solve (1) [35], while Beck and Teboulle extended Nesterov's first method to solve (1) [7]. Promising numerical results on ℓ_1-regularized least square problems are reported. While the methods in [4, 7, 21, 28, 29, 32], [31, Section 2.2] are remarkably simple, their analyses can be surprisingly intricate and lacking a unified framework.

Motivated by the above work, we propose a unified framework and simpler analysis of the $O(\sqrt{L/\epsilon})$ methods in [4, 7, 21, 28, 29, 31, 32], extended to solve (1). As a byproduct, we derive possibly new variants and refinements of $O(\sqrt{L/\epsilon})$ methods that use either one or two projections per iteration; see Corollaries 1, 2, 3. One variant uses a weighted sum of previous gradients as in [32], but uses one projection instead of two at each iteration. Recently, Nemirovski [26] proved the $O(L/\epsilon)$ iteration complexity for a prox-type method applied to monotone variational inequality and convex-concave optimization; also see [3, Section 4] for a closely related method using general proximity function, [34] for a related dual extrapolation method, and [24] for an application to large-scale SDP. We extend this method to solve a more general problem analogous to (1) (see (41)) and give a simple $O(L/\epsilon)$ iteration complexity proof; see Proposition 4 and the subsequent remarks. Key to our analyes are two basic properties of Bregman distance (Properties 1 and 2) and re-interpretations of the methods in [4, 7, 21, 26, 28, 29, 31, 32]. The aforementioned methods are compared on a matrix game example in Section 6. Extensions of these results are discussed in Section 7.

2 Basic properties of Bregman functions

We have the following basic properties of h and D. Property 1 ([11, Lemma 3.2], [21, Lemma 6]) will be used to prove Propositions 1 and 4, and Lemma 1. Property 2 will be used to prove Proposition 3.

Property 1 *For any proper lsc convex function $\psi : \mathcal{E} \to (-\infty, \infty]$ and any $z \in$ dom$P$, if*

$$z_+ = \arg\min_x \{\psi(x) + D(x, z)\}$$

and h is differentiable at z_+, then

$$\psi(x) + D(x, z) \geq \psi(z_+) + D(z_+, z) + D(x, z_+) \quad \forall x \in \text{dom}P.^4$$

[4] This follows from the optimality condition for z_+:

$$\psi(x) + \langle \nabla_x D(z_+, z), x - z_+ \rangle \geq \psi(z_+) \quad \forall x$$

Property 2 *For any proper lsc convex function* $\psi : \mathcal{E} \to (-\infty, \infty]$, *if*

$$z = \arg\min_x \{\psi(x) + h(x)\}$$

and h *is differentiable at* z, *then*

$$\psi(x) + h(x) \geq \psi(z) + h(z) + D(x, z) \quad \forall x \in \mathrm{dom} P.^5$$

3 1-memory $O(\sqrt{L/\epsilon})$ methods for convex optimization

In this section we present a unified framework and analysis of Nesterov's second $O(\sqrt{L/\epsilon})$ method [29, 31] and its variants in [4, 7, 21], extended to solve (1). A similar analyis of Nesterov's first method [28] and its extension to solve (1) [7] is presented in the remainder of this section.

Algorithm 1 *Choose* $\theta_0 \in (0, 1]$, $x_0, z_0 \in \mathrm{dom}P$. $k \leftarrow 0$. *Go to 1.*

1. *Choose a nonempty closed convex set* $X_k \subseteq \mathcal{E}$ *with* $X_k \cap \mathrm{dom}P \neq \emptyset$. *Let*

$$\begin{aligned} y_k &= (1-\theta_k)x_k + \theta_k z_k, && (11)\\ z_{k+1} &= \arg\min_{x \in X_k} \{\ell_f(x; y_k) + \theta_k L D(x, z_k)\}, && (12)\\ \hat{x}_{k+1} &= (1-\theta_k)x_k + \theta_k z_{k+1}. && (13) \end{aligned}$$

Choose x_{k+1} *to be no worse than* $\hat{x}_{k+1}$ *in* $\ell_f(\cdot; y_k) + \frac{L}{2}\|\cdot - y_k\|^2$ *value, i.e.,*

$$\ell_f(x_{k+1}; y_k) + \frac{L}{2}\|x_{k+1} - y_k\|^2 \leq \ell_f(\hat{x}_{k+1}; y_k) + \frac{L}{2}\|\hat{x}_{k+1} - y_k\|^2. \qquad (14)$$

Choose $\theta_{k+1} \in (0, 1]$ *satisfying*

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^2} \leq \frac{1}{\theta_k^2}. \qquad (15)$$

$k \leftarrow k+1$, *and go to 1.*

The set X_k should be chosen to contain a desired solution of (1); see Proposition 1. The simplest choice is $X_k = \mathcal{E}$, but it may be desirable to use a smaller X_k to accelerate convergence, at the expense of more computation to solve (12). By (6), $\epsilon + f^P(w) \geq f^P(x) \geq \ell_f(x; w)$ for any $w \in \mathrm{dom}P$ and any ϵ-minimum point x of f^P (i.e., $f^P(x) \leq \inf f^P + \epsilon$) with $\epsilon \geq 0$. Thus, the choice

$$X_k = \left\{ x \,\middle|\, \sum_{i \in I_{k,j}} \alpha_{k,i}(\ell_f(x; w_{k,i}) - f^P(w_{k,i})) \leq \epsilon,\ j = 1, \ldots, n_k \right\}, \qquad (16)$$

with $w_{k,i} \in \mathrm{dom}P$ (e.g., $w_{k,i} \in \{x_0, x_1, \cdots, x_k, z_0, z_1, \ldots, z_k\}$), $\alpha_{k,i} \geq 0$, $\sum_{i \in I_{k,j}} \alpha_{k,i} = 1$, $I_{k,j} \subseteq \{1, 2, \ldots\}$, and $n_k \geq 0$, contains all ϵ-minimum points of f^P. If $\inf f^P$ is attained, then we can replace "ϵ" by "0" in (16), in which case X_k contains all minimum points of f^P. In the smooth constrained case of (3), using (16) adds n_k cutting planes to (12) compared to using $X_k = \mathcal{E}$.

One choice for x_{k+1} is

$$x_{k+1} = \arg\min_x \left\{ \ell_f(x; y_k) + \frac{L}{2}\|x - y_k\|^2 \right\}. \qquad (17)$$

and $\nabla_x D(z_+, z) = \nabla h(z_+) - \nabla h(z)$. Rearranging terms yields

$$\psi(x) - \langle \nabla h(z), x - z \rangle \geq \psi(z_+) - \langle \nabla h(z), z_+ - z \rangle - \langle \nabla h(z_+), x - z_+ \rangle.$$

Add $h(x) - h(z)$ to both sides.

[5] This follows from the optimality condition for z: $\psi(x) + \langle \nabla h(z), x - z \rangle \geq \psi(z) \quad \forall x$.

In the case of (3), Algorithm 1 with $X_k = \mathcal{E}$ and this choice of x_{k+1} reduces to the Lan-Lu-Monteiro variant [21, Section 3] of Nesterov's method [32]. This choice requires two projection per iteration, but seems to be efficient in practice, according to the numerical results in [21].

A second choice for x_{k+1} is

$$x_{k+1} = \hat{x}_{k+1}. \tag{18}$$

In the case of (3), Algorithm 1 with $X_k = \mathcal{E}$ and this choice of x_{k+1} reduces to the Auslender-Teboulle extension [4, Section 5] of Nesterov's method from 1988 [29] (also see [31, page 90]), where the quadratic proximal term is replaced by the Bregman function D. This method requires only one projection per iteration. In his recent work [32, Section 3], Nesterov proposes an alternative method that uses a weighted sum of past gradients and two projections per iteration, as will be discussed in Section 4.

To gain some intuition for the improved efficiency of Algorithm 1 over classical gradient projection (8), suppose for simplicity that $P \equiv 0$, $h(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, $X_k = \mathcal{E}$, and x_{k+1} is given by (18). Then (11)–(13) simplify to

$$x_{k+1} = x_k + \theta_k(z_k - x_k) - \frac{1}{L}\nabla f\left(x_k + \theta_k(z_k - x_k)\right),$$

which is identical to (8) but for a key momentum term $\theta_k(z_k - x_k)$ added to x_k.

Note 1: Since h is coercive, z_{k+1} exists and belongs to dom$P$. Thus, h is differentiable at z_{k+1}. If h is separable quadratic, and P is the 1-norm or the indicator function for a box, then (12) and (17) have closed-form solutions.

Note 2: The condition (15) allows $\{\theta_k\}$ to decrease, but not too fast. For fastest convergence, $\{\theta_k\}$ should decrease as fast as possible, as Proposition 1 below suggests. The choice

$$\theta_k = \frac{2}{k+2} \tag{19}$$

satisfies (15). We can alternatively solve (15) with "$\leq$" replaced by "=," yielding

$$\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}, \tag{20}$$

which tends to zero somewhat faster.

Note 3: Algorithm 1 assumes L is known, but this can be relaxed by making an initial guess of L and decreasing L by a constant factor and repeating the iteration whenever the condition (23) below is violated. Under (2), the number of such decreases is finite; see [35] for bounds on this number.

Below we use Property 1 to give a simple proof of the $O(\sqrt{L/\epsilon})$ iteration complexity for Algorithm 1. Our proof is motivated by the proof of Theorem 5 in [21] for the case of (3), $X_k = \mathcal{E}$, and x_{k+1} given by (17), but with some simplification and generalization. To simplify notation, let

$$\Delta_f(x;y) := f^P(x) - \ell_f(x;y) = f(x) - f(y) - \langle \nabla f(y), x - y\rangle \quad \forall x, y \in \mathrm{dom}P. \tag{21}$$

Proposition 1 *Let $\{(x_k, y_k, z_k, \theta_k, X_k)\}$ be generated by Algorithm 1. For any $k = 0, 1, \ldots$ and any $x \in X_k \cap \mathrm{dom}P$, if $f^P(x) \leq f^P(x_{k+1})$ or (20) holds, then*

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^2}(f^P(x_{k+1}) - f^P(x)) + LD(x, z_{k+1}) \leq \frac{1-\theta_k}{\theta_k^2}(f^P(x_k) - f^P(x)) + LD(x, z_k) - \frac{\Delta_f(x; y_k)}{\theta_k}. \tag{22}$$

Proof. Fix any $k \in \{0, 1, \ldots\}$ and any $x \in X_k \cap \mathrm{dom}P$. By the second inequality in (6),

$$\begin{aligned} f^P(x_{k+1}) &\leq \ell_f(x_{k+1}; y_k) + \frac{L}{2}\|x_{k+1} - y_k\|^2 \\ &\leq \ell_f((1-\theta_k)x_k + \theta_k z_{k+1}; y_k) + \frac{L}{2}\|(1-\theta_k)x_k + \theta_k z_{k+1} - y_k\|^2 \end{aligned}$$

$$
\begin{aligned}
&\leq \; (1-\theta_k)\ell_f(x_k;y_k) + \theta_k \ell_f(z_{k+1};y_k) + \frac{L}{2}\theta_k^2\|z_{k+1}-z_k\|^2 \\
&\leq \; (1-\theta_k)\ell_f(x_k;y_k) + \theta_k \left(\ell_f(z_{k+1};y_k) + \theta_k L D(z_{k+1},z_k)\right) \qquad (23) \\
&\leq \; (1-\theta_k)\ell_f(x_k;y_k) + \theta_k \left(\ell_f(x;y_k) + \theta_k L D(x,z_k) - \theta_k L D(x,z_{k+1})\right) \\
&\leq \; (1-\theta_k)f^P(x_k) + \theta_k \left(f^P(x) - \Delta_f(x;y_k) + \theta_k L D(x,z_k) - \theta_k L D(x,z_{k+1})\right),
\end{aligned}
$$

where the second inequality uses (13), (14), the third inequality uses (11) and the convexity of $\ell_f(\cdot;y_k)$, the fourth inequality uses (7), the fifth inequality uses (12) and Property 1 with $\psi(x) = \ell_f(x;y_k)/(\theta_k L) + \delta_{X_k}(x)$, and the last inequality uses the first inequality in (6).

Subtracting $f^P(x)$ from both sides and then dividing both sides by θ_k^2 yields

$$\frac{1}{\theta_k^2}(f^P(x_{k+1}) - f^P(x)) \leq \frac{1-\theta_k}{\theta_k^2}(f^P(x_k) - f^P(x)) - \frac{\Delta_f(x;y_k)}{\theta_k} + LD(x;z_k) - LD(x;z_{k+1}). \qquad (24)$$

If $f^P(x) \leq f^P(x_{k+1})$, then this together with (15) proves (22). If (20) holds, then the "$\leq$" in (15) holds with equality and this again proves (22). ■

Corollary 1 *Let $\{(x_k, y_k, z_k, \theta_k, X_k)\}$ be generated by Algorithm 1 with $\theta_0 = 1$.*

(a) *Fix any $\epsilon > 0$. Suppose $\theta_k \leq \frac{2}{k+2}$ and X_k is given by (16) for all k. Then for any $x \in \mathrm{dom}P$ with $f^P(x) \leq \inf f^P + \epsilon$, we have*

$$\min_{i=0,1,\ldots,k+1}\{f^P(x_i)\} \;\leq\; f^P(x) + \epsilon \quad \text{whenever} \quad k \geq \sqrt{\frac{4LD(x,z_0)}{\epsilon}} - 2.$$

(b) *Suppose $\mathcal{E}$ is finite-dimensional,* $\mathrm{dom}P$ *is bounded, f has the form (9), θ_k is given by (20), and $X_k = \mathcal{E}$. Then*

$$0 \;\leq\; f^P(x_{k+1}) - q^P(\bar{v}_k) \;\leq\; \theta_k^2 L \max_{x\in\mathrm{dom}P} D(x,z_0), \quad k = 0,1,\ldots,$$

where q^P is given by (10), and we let $v_k = \arg\max_v \phi(y_k, v)$, and

$$\bar{v}_k \;=\; \left(\sum_{i=0}^{k}\frac{v_i}{\theta_i}\right) / \left(\sum_{i=0}^{k}\frac{1}{\theta_i}\right) \;=\; (1-\theta_k)\bar{v}_{k-1} + \theta_k v_k \qquad (\bar{v}_{-1} = 0). \qquad (25)$$

Proof. (a) For each k, either $\min_{i=0,1,\ldots,k+1}\{f^P(x_i)\} < f^P(x)$ or else $f^P(x_i) \geq f^P(x)$, $i = 0,1,\ldots,k+1$, in which case (22), (24), $D(x, z_{k+1}) \geq 0$, $\theta_0 = 1$, and $x \in X_k \cap \mathrm{dom}P$ yield

$$\frac{1}{\theta_k^2}(f^P(x_{k+1}) - f^P(x)) + \sum_{i=0}^{k}\frac{\Delta_f(x;y_i)}{\theta_i} \leq LD(x,z_0). \qquad (26)$$

Since $\Delta_f(x;y_i) \geq 0$, and $\theta_k \leq 2/(k+2)$, this implies $f^P(x_{k+1}) - f^P(x) \leq 4LD(x,z_0)/(k+2)^2$, whose right-hand side is below ϵ whenever $k+2 \geq \sqrt{4LD(x,z_0)/\epsilon}$.

(b) Since (20) holds and $X_k = \mathcal{E}$, by Proposition 1, (22) holds for all k and all $x \in \mathrm{dom}P$. Then, analogous to the proof of (a), we have that (26) holds for all k and all $x \in \mathrm{dom}P$. Since $\mathcal{E}$ is finite-dimensional and f has the form (9), by Danskin's theorem [9, Proposition B.25], $\nabla f(y_i) = \nabla_x\phi(y_i, v_i)$ and the convexity of $\phi(\cdot, v_i)$ yields

$$\Delta_f(x;y_i) = f(x) - \phi(y_i,v_i) - \langle\nabla_x\phi(y_i,v_i), x - y_i\rangle \geq f(x) - \phi(x,v_i), \quad i = 0,1,\ldots,k.$$

Dividing both sides by θ_i and summing over $i = 0,1,\ldots,k$ yields

$$\sum_{i=0}^{k}\frac{\Delta_f(x;y_i)}{\theta_i} \geq \sum_{i=0}^{k}\frac{f(x)}{\theta_i} - \sum_{i=0}^{k}\frac{\phi(x,v_i)}{\theta_i} \geq \sum_{i=0}^{k}\frac{1}{\theta_i}(f(x) - \phi(x,\bar{v}_k)) = \frac{1}{\theta_k^2}(f(x) - \phi(x,\bar{v}_k)),$$

where the second inequality uses the concavity of $\phi(x,\cdot)$ and the first equality in (25), and the equality follows from "$\le$" in (15) holding with equality, so that $\frac{1}{\theta_{k+1}^2} - \frac{1}{\theta_k^2} = \frac{1}{\theta_{k+1}}$, and an induction argument yields $\frac{1}{\theta_k^2} = \sum_{i=0}^k \frac{1}{\theta_i}$. The second equality in (25) follows similarly. Combining this with (26) yields, for each k,

$$\frac{1}{\theta_k^2}(f^P(x_{k+1}) - f^P(x)) + \frac{1}{\theta_k^2}(f(x) - \phi(x, \bar{v}_k)) \le LD(x, z_0) \quad \forall x \in \mathrm{dom}P.$$

Now take the maximum of both sides over $x \in \mathrm{dom}P$ and use (10). ■

Corollary 1(a) generalizes and slightly refines the complexity estimates in [21, Theorem 5], [4, Theorem 5.2], and [31, Theorem 2.2.3] for the special case of (3), $X_k = \mathcal{E}$, and x_{k+1} given by (17) or (18).

In the remainder of this section we further assume that f is differentiable on $\mathcal{E}$, and $\mathcal{E}$ is a Hilbert space, i.e., $\mathcal{E}^* = \mathcal{E}$, $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$. The following method is an extension of Nesterov's first $O(\sqrt{L/\epsilon})$ method [28] (also see [7]), using $h(\cdot) = \frac{1}{2}\|\cdot\|^2$ (so that $D(x,y) = \frac{1}{2}\|x-y\|^2$), to solve (1).

Algorithm 2 *Choose $\theta_0 = \theta_{-1} \in (0,1]$, $x_0 = x_{-1} \in \mathrm{dom}P$. $k \leftarrow 0$. Go to 1.*

1. *Choose a nonempty closed convex set $X_k \subseteq \mathcal{E}$ with $X_k \cap \mathrm{dom}P \ne \emptyset$. Let*

$$y_k = x_k + \theta_k(\theta_{k-1}^{-1} - 1)(x_k - x_{k-1}), \tag{27}$$

$$x_{k+1} = \arg\min_{x \in X_k} \left\{ \ell_f(x; y_k) + \frac{L}{2}\|x - y_k\|^2 \right\}. \tag{28}$$

Choose $\theta_{k+1} \in (0,1]$ satisfying (15). $k \leftarrow k+1$, and go to 1.

The set X_k and θ_k can be chosen as in Algorithm 1; see Note 2. Note that y_k may be outside of $\mathrm{dom}P$, and hence we need f to be differentiable outside of $\mathrm{dom}P$. Algorithm 2 assumes L is known, but this can be relaxed by increasing L and repeating the iteration whenever (29) below is violated; also see [7, FISTA with backtracking]. Algorithm 2 with $X_k = \mathcal{E}$ and θ_k chosen by (20) is essentially Nesterov's method in [28], as extended by Beck and Teboulle [7] to solve (1). This method is simpler than Algorithms 1 and 3 but is more limited in applicability.

Below we use Property 1 to give a simple proof of the $O(\sqrt{L/\epsilon})$ iteration complexity for Algorithm 2, analogous to that given in Proposition 1 for Algorithm 1. The proof uses the homogeneity property of $\|\cdot\|$.

Proposition 2 *Assume f is differentiable on $\mathcal{E}$, and $\mathcal{E}^* = \mathcal{E}$, $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$. Let $\{(x_k, y_k, \theta_k, X_k)\}$ be generated by Algorithm 2. For any $k = 0, 1, \ldots$ and any $x \in X_k \cap \mathrm{dom}P$, if $f^P(x) \le f^P(x_{k+1})$ or (20) holds, then (22) holds with $D(x,z) = \frac{1}{2}\|x - z\|^2$ and $z_k = x_{k-1} + \theta_{k-1}^{-1}(x_k - x_{k-1})$.*

Proof. Fix any $k \in \{0, 1, \ldots\}$ and any $x \in X_k \cap \mathrm{dom}P$. Let $y = (1-\theta_k)x_k + \theta_k x$, so that $y \in \mathrm{dom}P$. By the second inequality in (6),

$$\begin{aligned}
f^P(x_{k+1}) &\le \ell_f(x_{k+1}; y_k) + \frac{L}{2}\|x_{k+1} - y_k\|^2 \qquad (29)\\
&\le \ell_f(y; y_k) + \frac{L}{2}\|y - y_k\|^2 - \frac{L}{2}\|y - x_{k+1}\|^2 \\
&= \ell_f((1-\theta_k)x_k + \theta_k x; y_k) + \frac{L}{2}\|(1-\theta_k)x_k + \theta_k x - y_k\|^2 - \frac{L}{2}\|(1-\theta_k)x_k + \theta_k x - x_{k+1}\|^2 \\
&= \ell_f((1-\theta_k)x_k + \theta_k x; y_k) + \theta_k^2\frac{L}{2}\|x + \theta_k^{-1}(x_k - y_k) - x_k\|^2 - \theta_k^2\frac{L}{2}\|x + \theta_k^{-1}(x_k - x_{k+1}) - x_k\|^2 \\
&= \ell_f((1-\theta_k)x_k + \theta_k x; y_k) + \theta_k^2\frac{L}{2}\|x - z_k\|^2 - \theta_k^2\frac{L}{2}\|x - z_{k+1}\|^2 \\
&\le (1-\theta_k)\ell_f(x_k; y_k) + \theta_k\ell_f(x; y_k) + \theta_k^2\frac{L}{2}\|x - z_k\|^2 - \theta_k^2\frac{L}{2}\|x - z_{k+1}\|^2 \\
&\le (1-\theta_k)f^P(x_k) + \theta_k(f^P(x) - \Delta_f(x; y_k)) + \theta_k^2\frac{L}{2}\|x - z_k\|^2 - \theta_k^2\frac{L}{2}\|x - z_{k+1}\|^2,
\end{aligned}$$

where the second inequality uses (28) and Property 1 with $\psi(x) = \ell_f(x; y_k)/L + \delta_{X_k}(x)$, $D(x,z) = \frac{1}{2}\|x - z\|^2$, the third equality uses (27) (in fact, (27) is chosen to make this equality hold), the third inequality uses the convexity of $\ell_f(\cdot; y_k)$, and last inequality uses (6) and (21).

Subtracting $f^P(x)$ from both sides and then dividing both sides by θ_k^2 and using $D(x,z) = \frac{1}{2}\|x-z\|^2$ yields (24). If $f^P(x) \le f^P(x_{k+1})$, then this together with (15) proves (22). If (20) holds, then the "$\le$" in (15) holds with equality and this again proves (22). ■

Corollary 2 *Assume f is differentiable on $\mathcal{E}$, and $\mathcal{E}^* = \mathcal{E}$, $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$. Let $\{(x_k, y_k, \theta_k, X_k)\}$ be generated by Algorithm 2 with $\theta_0 = 1$. Then assertion (a) in Corollary 1 holds with $z_0 = x_0$. Assertion (b) in Corollary 1 holds with $z_0 = x_0$ if in addition ϕ is continuous on $\mathcal{E} \times V$, $\phi(\cdot, v)$ is convex and differentiable on $\mathcal{E}$ for every $v \in V$, and $\phi(x, \cdot)$ is strictly concave for every $x \in \mathcal{E}$.*

Proof. The proof is very similar to that of Corollary 1, with Proposition 2 replacing Proposition 1 and noting that $z_0 = x_0$. Since y_k may not lie in dom$P$, for (b) we need ϕ to have the desired properties on $\mathcal{E} \times V$ instead of dom$P \times V$. ■

Corollary 2(a) refines [28, Theorem 2] and [7, Theorem 4.1] for the case of $X_k = \mathcal{E}$ and (1) having a minimum point.

4 ∞-memory $O(\sqrt{L/\epsilon})$ methods for convex optimization

Nesterov's $O(\sqrt{L/\epsilon})$ method in [32] is remarkably different from those in [4, 21, 28, 29, 31] in that it uses, at each iteration, two projections and a weighted sum of previous gradients. In this section we present a general framework and analysis of Nesterov's method, extended to solve (1). Our framework and analysis parallel that of Algorithms 1 and 2, thus showing a close connection among all $O(\sqrt{L/\epsilon})$ methods. Moreover, our analysis shows that one projection per iteration suffices to achieve the same complexity for Nesterov's method.

Algorithm 3 *Choose $0 < \theta_0 \le 1, \vartheta_0 \ge \theta_0$, $x_0 \in \mathrm{dom}P$. Let $z_0 = \arg\min_{x \in \mathrm{dom}P} h(x)$, $X_{-1} = \mathcal{E}$. $k \leftarrow 0$. Go to 1.*

1. *Choose a nonempty closed convex set $X_k \subseteq X_{k-1}$ with $X_k \cap \mathrm{dom}P \ne \emptyset$. Let y_k be given by (11), let*

$$\psi_{k+1}(x) \;:=\; \sum_{i=0}^{k} \frac{\ell_f(x; y_i)}{\vartheta_i} \quad \forall x, \tag{30}$$

$$z_{k+1} \;=\; \arg\min_{x \in X_k} \{\psi_{k+1}(x) + Lh(x)\}, \tag{31}$$

and let x_{k+1} be given by (13), (14). Choose $0 < \theta_{k+1} \le 1, \vartheta_{k+1} \ge \theta_{k+1}$ satisfying

$$\frac{1 - \theta_{k+1}}{\theta_{k+1}\vartheta_{k+1}} = \frac{1}{\theta_k \vartheta_k}. \tag{32}$$

$k \leftarrow k+1$, and go to 1.

We can choose $X_k = \mathcal{E}$ or by (16), provided that $X_k \subseteq X_{k-1}$. We do not know if the latter restriction, which is needed in the proof of Proposition 3 below, can be relaxed. Nesterov's method in [32, Section 3] corresponds to Algorithm 3 with $X_k = \mathcal{E}$, x_{k+1} given by (17), and

$$\theta_k = \frac{2}{k+2}, \qquad \vartheta_k = \frac{2}{k+1}, \tag{33}$$

which satisfy (32). A better choice is

$$\theta_k \text{ given by (20) with } \theta_0 = 1, \qquad \vartheta_k = \theta_k, \tag{34}$$

which also satisfy (32) and tend to zero faster. Nesterov's recent method for solving (1) [35, Section 4] is similar to Algorithm 3 but with $D(x,y) = \frac{1}{2}\|x-y\|^2$, $X_k = \mathcal{E}$, and x_{k+1} given by (17). In [32, Section 5.2] is described a modified method that replaces (17) by

$$\hat{z}_{k+1} = \arg\min_{x \in X_k} \left\{ \ell_f(x; y_k) + \theta_k L D(x; z_k) \right\}, \tag{35}$$

$$x_{k+1} = (1-\theta_k)x_k + \theta_k \hat{z}_{k+1}. \tag{36}$$

This modified method may be viewed as a hybrid of Algorithms 1 and 3 (compare (35), (36) with (12), (13)). It uses the same kernel function h in both minimizations (31) and (35), which may be advantageous when h can be chosen to simplify the minimizations (e.g., (3) with X being the unit simplex and h being the $x \ln x$-entropy function). We can alternatively choose x_{k+1} by (18), so that only one projection is needed per iteration. This new variant may be advantageous when the projection is expensive.

Algorithm 3 resembles Algorithm 1, except that z_{k+1} is given by (31) instead of (12). (The differences between (15) and (32) are negligible.) In particular, by using (30), we can rewrite (31) as

$$z_{k+1} = \arg\min_{x \in X_k} \left\{ \frac{\ell_f(x; y_k)}{\vartheta_k} + \psi_k(x) + Lh(x) \right\},$$

with $\psi_0 = \delta_{\mathrm{dom}P}$. As we shall see in the analysis below, the term $\psi_k(x) + Lh(x)$ plays the role of $LD(x, z_k)$ in (12).

Note 4: By the same argument as in Note 1, z_{k+1} exists, belongs to dom$P$, and h is differentiable at z_{k+1}. Algorithm 3 assumes L is known, but this can be relaxed as described in Note 3, i.e., increase L whenever (23) is violated.

Note 5: If $\theta_0 = 1$, then (32) implies

$$\sum_{i=0}^{k} \frac{1}{\vartheta_i} = \frac{1}{\theta_k \vartheta_k}. \tag{37}$$

This clearly holds for $k = 0$. Suppose it holds for some $k \geq 0$. Then, by (32),

$$\sum_{i=0}^{k+1} \frac{1}{\vartheta_i} = \frac{1}{\theta_k \vartheta_k} + \frac{1}{\vartheta_{k+1}} = \frac{1-\theta_{k+1}}{\theta_{k+1}\vartheta_{k+1}} + \frac{1}{\vartheta_{k+1}} = \frac{1}{\theta_{k+1}\vartheta_{k+1}}.$$

Below we use Property 2 and (37) to give a simple proof of the $O(\sqrt{L/\epsilon})$ iteration complexity for Algorithm 3, paralleling that for Algorithms 1 and 2 (compare Propositions 1, 2, and 3).

Proposition 3 *Let* $\{(x_k, y_k, z_k, \theta_k, \vartheta_k, X_k)\}$ *be generated by Algorithm 3 or its modification where* (13), (14) *are replaced by* (35), (36). *Let* $\psi_0 = \delta_{\mathrm{dom}P}$. *For any* $k = 0, 1, \ldots$, *we have*

$$\frac{1-\theta_{k+1}}{\theta_{k+1}\vartheta_{k+1}} f^P(x_{k+1}) - (\psi_{k+1} + Lh)(z_{k+1}) \leq \frac{1-\theta_k}{\theta_k \vartheta_k} f^P(x_k) - (\psi_k + Lh)(z_k). \tag{38}$$

Proof. The definitions of z_0, ψ_0, and $X_{-1} = \mathcal{E}$ imply that (31) holds for $k = -1$. Fix any $k \in \{0, 1, \ldots\}$. If x_{k+1} is given by (13) and (14), then as in the proof of Proposition 1, we have that (23) holds. If x_{k+1} is given by (35) and (36), then as in the proof of Proposition 1, we have that (23) holds but with z_{k+1} replaced by $\hat{z}_{k+1}$. Combining this with $z_{k+1} \in X_k$ and (35), so that

$$\ell_f(\hat{z}_{k+1}; y_k) + \theta_k L D(\hat{z}_{k+1}; z_k) \leq \ell_f(z_{k+1}; y_k) + \theta_k L D(z_{k+1}; z_k),$$

we see that (23) holds. Using (23), we have

$$\begin{aligned}
f^P(x_{k+1}) &\le (1-\theta_k)\ell_f(x_k;y_k) + \theta_k\left(\ell_f(z_{k+1};y_k) + \theta_k LD(z_{k+1},z_k)\right) \\
&\le (1-\theta_k)f^P(x_k) + \theta_k\left(\ell_f(z_{k+1};y_k) + \vartheta_k LD(z_{k+1},z_k)\right) \\
&= (1-\theta_k)f^P(x_k) + \theta_k\vartheta_k\left(\frac{\ell_f(z_{k+1};y_k)}{\vartheta_k} + LD(z_{k+1},z_k)\right) \\
&\le (1-\theta_k)f^P(x_k) + \theta_k\vartheta_k\left(\frac{\ell_f(z_{k+1};y_k)}{\vartheta_k} + \psi_k(z_{k+1}) + Lh(z_{k+1}) - \psi_k(z_k) - Lh(z_k)\right), \\
&= (1-\theta_k)f^P(x_k) + \theta_k\vartheta_k\left(\psi_{k+1}(z_{k+1}) + Lh(z_{k+1}) - \psi_k(z_k) - Lh(z_k)\right),
\end{aligned}$$

where the second inequality uses the first inequality in (6) and $\theta_k \le \vartheta_k$, the last inequality uses $z_{k+1} \in X_k \subseteq X_{k-1}$, Property 2 with $\psi(x) = \psi_k(x)/L + \delta_{X_{k-1}}(x)$, and the fact that (31) holds when "$k+1$" is replaced by "k", the last equality uses (30) and $\psi_0 = \delta_{\mathrm{dom}P}$.

Dividing both sides by $\theta_k\vartheta_k$ and using (32) yields

$$\frac{1-\theta_{k+1}}{\theta_{k+1}\vartheta_{k+1}}f^P(x_{k+1}) \le \frac{1-\theta_k}{\theta_k\vartheta_k}f^P(x_k) + (\psi_{k+1}+Lh)(z_{k+1}) - (\psi_k+Lh)(z_k).$$

Rearranging terms yields (38). ■

Corollary 3 *Let $\{(x_k, y_k, z_k, \theta_k, \vartheta_k, X_k)\}$ be generated by Algorithm 3 or its modification described in Proposition 3, with $\theta_0 = 1$.*

(a) *For any $k \ge 0$ and any $x \in X_k \cap \mathrm{dom}P$, we have*

$$\frac{1}{\theta_k\vartheta_k}(f^P(x_{k+1}) - f^P(x)) + \sum_{i=0}^{k}\frac{\Delta_f(x;y_i)}{\vartheta_i} \;\le\; L(h(x) - h(z_0)). \tag{39}$$

(b) *Fix any $\epsilon > 0$. Suppose $\vartheta_k \le \frac{2}{k+1}$ and X_k is given by (16) for all k. Then for any $x \in \mathrm{dom}P$ with $f^P(x) \le \inf f^P + \epsilon$, we have*

$$f^P(x_{k+1}) \;\le\; f^P(x) + \epsilon \quad \text{whenever} \quad k \ge \sqrt{\frac{4L(h(x)-h(z_0))}{\epsilon}} - 1.$$

(c) *Suppose $\mathcal{E}$ is finite-dimensional, $\mathrm{dom}P$ is bounded, f has the form (9), and $X_k = \mathcal{E}$. Then*

$$0 \;\le\; f^P(x_{k+1}) - q^P(\bar{v}_k) \;\le\; \theta_k\vartheta_k L \max_{x\in\mathrm{dom}P}(h(x) - h(z_0)), \quad k = 0, 1, \ldots,$$

where q^P is given by (10), and we let $v_k = \arg\max_v \phi(y_k, v)$ and

$$\bar{v}_k \;=\; \left(\sum_{i=0}^{k}\frac{v_i}{\vartheta_i}\right) \Big/ \left(\sum_{i=0}^{k}\frac{1}{\vartheta_i}\right) \;=\; (1-\theta_k)\bar{v}_{k-1} + \theta_k v_k \qquad (\bar{v}_{-1} = 0).$$

Proof. (a) By applying (38) recursively and then using (32) and $\theta_0 = 1$, we have that, for each $k \ge 0$,

$$\frac{f^P(x_{k+1})}{\theta_k\vartheta_k} - (\psi_{k+1}+Lh)(z_{k+1}) \le -(\psi_0+Lh)(z_0).$$

Since $\psi_0 = \delta_{\mathrm{dom}P}$ so that $\psi_0(z_0) = 0$, this yields for any $x \in X_k \cap \mathrm{dom}P$ that

$$\frac{f^P(x_{k+1})}{\theta_k\vartheta_k} \;\le\; \psi_{k+1}(z_{k+1}) + Lh(z_{k+1}) - Lh(z_0)$$

$$
\begin{aligned}
&\le \psi_{k+1}(x) + Lh(x) - Lh(z_0) \\
&= \sum_{i=0}^{k} \frac{\ell_f(x;y_i)}{\vartheta_i} + Lh(x) - Lh(z_0) \\
&= \sum_{i=0}^{k} \frac{f^P(x)}{\vartheta_i} - \sum_{i=0}^{k} \frac{\Delta_f(x;y_i)}{\vartheta_i} + Lh(x) - Lh(z_0) \\
&= \frac{f^P(x)}{\theta_k\vartheta_k} - \sum_{i=0}^{k} \frac{\Delta_f(x;y_i)}{\vartheta_i} + Lh(x) - Lh(z_0),
\end{aligned}
$$

where the second inequality uses (31) and $x \in X_k$, and the three equalities use, respectively, (30), (21), and (37). Rearranging terms yields (39).

The proof of (b) and (c) is very similar to that of Corollary 1, with (39) replacing (26) and using $\theta_k \le \vartheta_k$, (32), (37). ∎

In the special case of (3), $X_k = \mathcal{E}$, x_{k+1} given by (17), and ϑ_k, θ_k given by (33), Corollary 3 reduces to [32, Theorems 2 and 3]; also see [22, Theorem 2.2]. In fact, Corollaries 1(b), 2(b) and 3(c) are motivated by [32, Theorem 3] and [22, Theorem 2.2]. We obtain slightly sharper complexity bound by replacing (33) with (34). For $k = 0, 1, \ldots, 5$, this yields $\theta_k\vartheta_k = 1, .38, .20, .13, .09, .06$ as compared to $\theta_k\vartheta_k = 2, .66, .33, .2, .13, .09$ for (33). In the special case of $X_k = \mathcal{E}$ and x_{k+1} given by (17), Corollary 3(b) is similar to [35, Eq. (4.17)]. Notice that Propositions 1, 2, 3 and Corollaries 1(a), 2(a), 3(b) do not assume (1) has a minimum point. If (1) has a minimum point x, then it can be used therein.

5 $O(L/\epsilon)$ method for convex-concave optimization and monotone variational inequality

Suppose $F : \mathrm{dom}P \to \mathcal{E}^*$ is monotone and Lipschitz continuous on $\mathrm{dom}P$, i.e.,

$$
\langle F(x) - F(y), x - y \rangle \ge 0 \quad \text{and} \quad \|F(x) - F(y)\|_* \le L\|x - y\| \quad \forall x, y \in \mathrm{dom}P. \tag{40}
$$

Consider the variational inequality problem of finding an $\bar{x} \in \mathrm{dom}P$ satisfying

$$
\langle F(\bar{x}), x - \bar{x} \rangle + P(x) - P(\bar{x}) \ge 0 \quad \forall x \quad \Longleftrightarrow \quad \bar{x} = \arg\min_x \ell_F(x;\bar{x}), \tag{41}
$$

where we define

$$
\ell_F(x;y) := \langle F(y), x \rangle + P(x). \tag{42}
$$

In the case of (3), this problem has been well studied; see [15] and references therein. A special case of interest is the min-max problem

$$
\min_u \max_v \{\phi(u,v) + \pi(u) - \varpi(v)\}, \tag{43}
$$

where $\pi : \Re^n \to (-\infty, \infty]$, $\varpi : \Re^m \to (-\infty, \infty]$ are proper, lsc, convex, and ϕ is convex-concave and differentiable on an open set containing $\mathrm{dom}\pi \times \mathrm{dom}\varpi$ (with Lipschitz continuous gradient). This generalizes (1). Moreover, an $\bar{x} \in \mathrm{dom}P$ solves (43) if and only if it satisfies (41) with

$$
F(u,v) = \begin{bmatrix} \nabla_u \phi(u,v) \\ -\nabla_v \phi(u,v) \end{bmatrix}, \qquad P(u,v) = \pi(u) + \varpi(v). \tag{44}
$$

In the case of (3), Nemirovski [26] proposed an $O(L/\epsilon)$ prox-type method to solve (41) and (43). His analysis shows, as a byproduct, that Korpelevich's extragradient method [20] achieves $O(L/\epsilon)$ iteration complexity in an ergodic sense; also see [3, Proposition 4.1] for a similar result with a general proximity function. This improves on analogous results for the mirror descent method [27, Section 6.2]. Below we extend Nemirovski's method to solve (41) and (43); compare (45) with [26, Eqs. (3.3), (3.4)].

Algorithm 4 *Choose $x_0 \in \mathrm{dom}P$. $k \leftarrow 0$. Go to 1.*

1. *Choose a nonempty closed convex set $X_k \subseteq \mathcal{E}$ with $X_k \cap \mathrm{dom}P \neq \emptyset$. Choose $\gamma_k > 0$ and $y_k \in \mathrm{dom}P$ satisfying*

$$\min_{x\in X_k} \left\{ \ell_F(x;y_k) + \frac{L}{\gamma_k} D(x,x_k) \right\} \quad \geq \quad \ell_F(y_k;y_k). \tag{45}$$

Let x_{k+1} attain the minimum in (45)*. $k \leftarrow k+1$, and go to 1.*

Note 6: Algorithm 4 assumes L is known, but this can be relaxed by making an initial guess of L and decreasing L by a constant factor and repeating the iteration whenever (45) is violated. Under (40) and with y_k given by (48) and $\gamma_k \leq 1$, the proof of Lemma 1 below shows that the number of such decreases is finite.

The set X_k should be chosen to contain a desired solution of (41). The simplest choice is $X_k = \mathcal{E}$. For any $w \in \mathrm{dom}P$ and any solution x of (41), we have $\langle F(w) - F(x), w - x\rangle \geq 0$ and $\langle F(x), w - x\rangle + P(w) - P(x) \geq 0$. Summing them yields $\langle F(w), w - x\rangle + P(w) - P(x) \geq 0$ or, equivalently, $\ell_F(x;w) \leq \ell_F(w;w)$. Thus, the choice

$$X_k = \left\{ x \,\middle|\, \sum_{i\in I_{k,j}} \alpha_{k,i}(\ell_F(x;w_{k,i}) - \ell_F(w_{k,i};w_{k,i}))) \leq \varepsilon_k,\ j = 1,\ldots,n_k \right\}, \tag{46}$$

with $w_{k,i} \in \mathrm{dom}P$, $\alpha_{k,i} \geq 0$, $\sum_{i\in I_{k,j}} \alpha_{k,i} = 1$, $I_{k,j} \subseteq \{1,2,\ldots\}$, $n_k \geq 0$, and $\varepsilon_k \geq 0$, contains all solutions of (41). In the case of (3), using (46) adds n_k cutting planes to (45) compared to using $X_k = \mathcal{E}$.

One choice of y_k is the unique solution of the strongly monotone variational inequality, obtained by adding the proximal term $D(\cdot, x_k)$ to P in (41), i.e.,

$$y_k \quad = \quad \operatorname*{arg\,min}_{x\in X_k} \left\{ \ell_F(x;y_k) + \frac{L}{\gamma_k} D(x,x_k) \right\}. \tag{47}$$

Thus y_k attains the minimum in (45) and hence (45) holds. In this case, Algorithm 4 reduces to the proximal point method for solving (41) using a Bregman function; see [14, 40, 41] and references therein. However, this choice of y_k is generally too expensive to compute.

A second choice of y_k, proposed in [26], is to approximate F by the constant mapping $F(x_k)$, i.e.,

$$y_k \quad = \quad \operatorname*{arg\,min}_{x\in X_k} \left\{ \ell_F(x;x_k) + \frac{L}{\gamma_k} D(x,x_k) \right\}. \tag{48}$$

In this case, Algorithm 4 with $X_k = \mathcal{E}$ reduces to Korpelevich's extragradient method for solving (41) using a Bregman function; see [3, 20, 26, 44] and references therein. As is noted in [26], (48) and (45) are equivalent to two fixed-point iterations of the contractive mapping $y \mapsto \arg\min_x \left\{ \ell_F(x;y) + \frac{L}{\gamma_k} D(x,x_k) \right\}$ whose fixed point satisfies (47). The lemma below, an extension of [26, Theorem 3.1], shows that this choice satisfies (45) provided $\gamma_k \leq 1$. We give a simple proof using Property 1.

Lemma 1 *For each $k \in \{0,1,\ldots\}$, if $\gamma_k \leq 1$, then y_k given by* (48) *satisfies* (45)*.*

Proof. By using (48) and Property 1 with $\psi(x) = \frac{\gamma_k}{L}\ell_F(x;x_k) + \delta_{X_k}(x)$, we have

$$\frac{\gamma_k}{L}\ell_F(x;x_k) + D(x,x_k) \quad \geq \quad \frac{\gamma_k}{L}\ell_F(y_k;x_k) + D(y_k,x_k) + D(x,y_k) \quad \forall x \in X_k \cap \mathrm{dom}P. \tag{49}$$

Then (42) yields, for all $x \in X_k \cap \mathrm{dom}P$,

$$\begin{aligned} \frac{\gamma_k}{L}\ell_F(x;y_k) + D(x,x_k) - \frac{\gamma_k}{L}\ell_F(y_k;y_k) \quad &\geq \quad \frac{\gamma_k}{L}\langle F(y_k) - F(x_k), x - y_k\rangle + D(y_k,x_k) + D(x,y_k) \\ &\geq \quad -\|y_k - x_k\|\|x - y_k\| + \frac{1}{2}\|y_k - x_k\|^2 + \frac{1}{2}\|x - y_k\|^2 \\ &\geq \quad 0, \end{aligned}$$

where the second inequality follows from $-\langle F(y_k) - F(x_k), x - y_k\rangle \leq \|F(y_k) - F(x_k)\|_* \|x - y_k\| \leq L\|y_k - x_k\|\|x - y_k\|$, $\gamma_k \leq 1$, and (7). Minimizing the left-hand side with respect to $x \in X_k$ yields (45). ■

Termination for Algorithm 4 will be based on the following error function

$$e(y) \quad := \quad \max_{x \in \mathrm{dom} P} \{E(x, y) - P(x) + P(y)\}\,, \tag{50}$$

where $E : \mathrm{dom}P \times \mathrm{dom}P \to \Re$ satisfies

$$E(x, \cdot) \text{ is convex}, \quad E(x, x) = 0, \quad \text{and} \quad E(x, y) \leq \langle F(y), y - x\rangle \quad \forall\, x, y \in \mathrm{dom}P. \tag{51}$$

Thus $e(\cdot)$ is convex and $e(y) \geq 0$ for all $y \in \mathrm{dom}P$. The following lemma relates $e(\cdot)$ to (41).

Lemma 2 *Let $E(x, \bar{x}) = \langle F(x), \bar{x} - x\rangle$ or, when F and P are given by (44), $E(x, \bar{x}) = \phi(\bar{u}, v) - \phi(u, \bar{v})$. Then E satisfies* (51). *Moreover, for any $\bar{x} \in \mathrm{dom}P$,* (41) *is equivalent to $e(\bar{x}) = 0$.*

Proof. Suppose $E(x, y) = \langle F(x), y - x\rangle$, which is convex in y. If $\bar{x}$ satisfies (41), then by using the monotonicity of F, e.g., $\langle F(x) - F(\bar{x}), x - \bar{x}\rangle \geq 0$, we obtain

$$-E(x, \bar{x}) + P(x) - P(\bar{x}) \geq 0 \quad \forall x \in \mathrm{dom}P.$$

Thus (50) yields $e(\bar{x}) = 0$. Conversely, if $e(\bar{x}) = 0$, then the above inequalities hold. We now argue similarly as in the case of (3); see [15, page 159]. Since $\mathrm{dom}P$ is convex, this implies

$$-E(\theta x + (1-\theta)\bar{x}, \bar{x}) + P(\theta x + (1-\theta)\bar{x}) - P(\bar{x}) \geq 0 \quad \forall x \in \mathrm{dom}P,\ \forall 0 < \theta < 1.$$

Using the convexity of P, i.e., $P(\theta x + (1-\theta)\bar{x}) \leq \theta P(x) + (1-\theta)P(\bar{x})$, we obtain that

$$\langle F(\theta x + (1-\theta)\bar{x}), x - \bar{x}\rangle + P(x) - P(\bar{x}) \geq 0 \quad \forall x \in \mathrm{dom}P,\ \forall 0 < \theta < 1.$$

Letting $\theta \to 0$ and using the continuity of F yields (41). Moreover, the monotonicity of F implies $\langle F(y), y - x\rangle \geq \langle F(x), y - x\rangle = E(x, y)$, where $x, y \in \mathrm{dom}P$. Thus (51) holds.

Suppose F and P are given by (44) and $E(x, \bar{x}) = \phi(\bar{u}, v) - \phi(u, \bar{v})$, which is convex in $\bar{x} = (\bar{u}, \bar{v})$. It is straightforward to verify that $e(\bar{x}) = 0$ is equivalent to $\bar{x}$ being a solution of (43). Moreover, the convex-concavity of ϕ implies

$$\begin{aligned}
\langle F(y), y - x\rangle &= \langle \nabla_u \phi(w, z), w - u\rangle - \langle \nabla_v \phi(w, z), z - v\rangle \\
&\geq \phi(w, z) - \phi(u, z) - (\phi(w, z) - \phi(w, v)) \\
&= \phi(w, v) - \phi(u, z) \;=\; E(x, y),
\end{aligned}$$

where $x = (u, v), y = (w, z) \in \mathrm{dom}P$. Thus (51) holds. ■

In the case of (3), $e(\cdot)$ with E given by Lemma 2 reduces to $\epsilon(\cdot)$ and $\epsilon_f(\cdot)$ in [26]. The former is known as the dual gap function (modulo negation); see [15, page 167]. Below we use Property 1 to give a simple proof of the $O(L/\epsilon)$ iteration complexity of Algorithm 4 for finding an $\bar{x}$ with $e(\bar{x}) \leq \epsilon$. This result shows that, for fast convergence, the stepsize γ_k should be as large as possible. For y_k given by (48), Lemma 1 suggests taking $\gamma_k = 1$. In fact, $\gamma_k > 1$ is allowable provided that (45) still holds.

Proposition 4 *Let $\{(x_k, y_k, \gamma_k, X_k)\}$ be generated by Algorithm 4.*

(a) *Suppose $X_k = \mathcal{E}$ for all k. For any integers $0 \leq s \leq t$, we have $\bar{x}_{s,t} \in \mathrm{dom}P$ and*

$$E(x, \bar{x}_{s,t}) - P(x) + P(\bar{x}_{s,t}) \;\leq\; \frac{L\, D(x, x_s)}{\gamma_{s,t}} \quad \forall x \in \mathrm{dom}P, \tag{52}$$

where

$$\gamma_{s,t} = \sum_{k=s}^{t} \gamma_k, \qquad \bar{x}_{s,t} = \frac{1}{\gamma_{s,t}} \sum_{k=s}^{t} \gamma_k y_k.$$

(b) *Suppose X_k is given by (46), X_k is nonempty, and there exist Lagrange multipliers $\lambda_{k,j} \geq 0$, $j = 1, \ldots, n_k$, for its constraints in the minimization of (45) for all k. For any integers $0 \leq s \leq t$, we have $\bar{x}_{s,t} \in \mathrm{dom} P$ and (52) holds, where*

$$\gamma_{s,t} = \sum_{k=s}^{t} \gamma_k \left(1 + \sum_{j=1}^{n_k} \sum_{i \in I_{k,j}} \lambda_{k,j} \alpha_{k,i} \right), \qquad \bar{x}_{s,t} = \frac{1}{\gamma_{s,t}} \sum_{k=s}^{t} \gamma_k \left(y_k + \sum_{j=1}^{n_k} \sum_{i \in I_{k,j}} \lambda_{k,j} \alpha_{k,i} w_{k,i} \right).$$

Proof. (a) Fix any $k \in \{0, 1, \ldots\}$ and any $x \in \mathrm{dom} P$. Since x_{k+1} attains the minimum in (45) and $X_k = \mathcal{E}$, Property 1 with $\psi(x) = \frac{\gamma_k}{L} \ell_F(x; y_k)$ yields

$$\frac{\gamma_k}{L} \ell_F(x; y_k) + D(x, x_k) \quad \geq \quad \frac{\gamma_k}{L} \ell_F(x_{k+1}; y_k) + D(x_{k+1}, x_k) + D(x, x_{k+1}) \quad \forall x \in \mathrm{dom} P. \tag{53}$$

Moreover, $x_{k+1} \in \mathrm{dom} P$ and setting $x = x_{k+1}$ in (45) yields

$$\ell_F(x_{k+1}; y_k) + \frac{L}{\gamma_k} D(x_{k+1}, x_k) \quad \geq \quad \ell_F(y_k; y_k).$$

Multiplying the above inequality by $\frac{\gamma_k}{L}$ and adding it to (53) yield

$$\frac{\gamma_k}{L} \left(\ell_F(x; y_k) - \ell_F(y_k; y_k) \right) + D(x, x_k) \quad \geq \quad D(x, x_{k+1}) \quad \forall x \in \mathrm{dom} P.$$

By using (42) and (51) to bound the left-hand side, we have

$$\frac{\gamma_k}{L} (-E(x, y_k) + P(x) - P(y_k)) + D(x, x_k) \quad \geq \quad D(x, x_{k+1}) \quad \forall x \in \mathrm{dom} P.$$

For each $x \in \mathrm{dom} P$ and any integers $0 \leq s \leq t$, summing this over $k = s, s+1, \ldots, t$ and dividing by $\gamma_{s,t} = \sum_{k=s}^{t} \gamma_k$ yields

$$-\sum_{k=s}^{t} \omega_k E(x, y_k) + P(x) - \sum_{k=s}^{t} \omega_k P(y_k) + L \frac{D(x, x_s) - D(x, x_{t+1})}{\gamma_{s,t}} \quad \geq \quad 0.$$

where we let $\omega_k = \gamma_k / \gamma_{s,t}$. Since $E(x, \cdot)$ and P are convex, and $D(x, x_{t+1}) \geq 0$, this and the definition of $\bar{x}_{s,t}$, $\gamma_{s,t}$ yields $\bar{x}_{s,t} \in \mathrm{dom} P$ and

$$-E(x, \bar{x}_{s,t}) + P(x) - P(\bar{x}_{s,t}) + \frac{L D(x, x_s)}{\gamma_{s,t}} \quad \geq \quad 0.$$

(b) For any integer $k \geq 0$, since $\lambda_{k,j} \geq 0$, $j = 1, \ldots, n_k$, are Lagrange multipliers for the constraints in (46), the left-hand side of (45) equals

$$\min_x \left\{ \ell_F(x; y_k) + \sum_{j=1}^{n_k} \sum_{i \in I_{k,j}} \lambda_{k,j} \alpha_{k,i} \left(\ell_F(x; w_{k,i}) - \ell_F(w_{k,i}; w_{k,i}) \right) + \frac{L}{\gamma_k} D(x, x_k) \right\}$$

and x_{k+1} attains the minimum. This is equivalent to (45) with $X_k = \mathcal{E}$ and with a nonnegative weighted sum of $\ell_F(x; w_{k,i}) - \ell_F(w_{k,i}; w_{k,i})$, $i \in \cup_{j=1}^{n_k} I_{k,j}$, added to $\ell_F(x; y_k)$. Thus we can repeat the argument for (a), but using this modification to $\ell_F(x; y_k)$. ■

If $\mathrm{dom} P$ is bounded and we take, say, $\gamma_k = 1$ for all k in Algorithm 4, then (52) shows that, after t iterations, we obtain an $\bar{x} \in \mathrm{dom} P$ satisfying

$$e(\bar{x}) = \max_{x \in \mathrm{dom} P} (E(x, \bar{x}) - P(x) + P(\bar{x})) \leq \frac{L R_s}{t - s + 1}, \qquad \textit{with} \qquad R_s = \max_{x \in \mathrm{dom} P} D(x, x_s)$$

for any $0 \le s \le t$. When F and P are given by (44) and $E(x,\bar{x}) = \phi(\bar{u},v) - \phi(u,\bar{v})$, this implies $\bar{x} = (\bar{u},\bar{v})$ is an ϵ-solution of the saddle-point problem (43) in the sense that

$$\max_v\{\phi(\bar{u},v) + \pi(\bar{u}) - \varpi(v)\} - \epsilon \le \phi(\bar{u},\bar{v}) + \pi(\bar{u}) - \varpi(\bar{v}) \le \min_u\{\phi(u,\bar{v}) + \pi(u) - \varpi(\bar{v})\} + \epsilon$$

whenever $t - s + 1 \ge LR_s/\epsilon$. In the case of (3) and $s = \lfloor t/2 \rfloor$, this result recovers those in [26, Sections 2 and 3]. Proposition 4(b) generalizes the "bundle" algorithm in [26, Section 4]. In the case of $\mathcal{E} = \Re^n$, X_k given by (46) is nonempty whenever (41) has a solution, and Lagrange multipliers $\lambda_{k,j}$ exist whenever (i) P is polyhedral or (ii) P is given by (3) and $X_k \cap \mathrm{ri}(X) \ne \emptyset$ (as is assumed in [26, Section 4]) or (iii) a Slater condition hold; see [38, Theorem 28.2]. Our proof using $D(\cdot,\cdot)$, $E(\cdot,\cdot)$, and Property 1 seems simpler. In the case of (3), y_k given by (48), $E(x,y) = \langle F(x), y - x\rangle$, and $s = 0$, Proposition 4(a) is similar to [3, Proposition 4.1] specialized to Bregman functions.

Suppose X_k is given by (46) and y_k is given by (48). For any solution $\bar{x}$ of (41), we have

$$\ell_F(x;\bar{x}) \quad \ge \quad \ell_F(\bar{x};\bar{x}) \quad \forall x. \tag{54}$$

Also, (49) and (53) hold for all $x \in X_k$. Setting $x = x_{k+1}$, $x = \bar{x}$, and $x = y_k$ in (49), (53), and (54), respectively, and summing, we obtain (also using (42))

$$\begin{aligned}
& D(\bar{x},x_k) - D(\bar{x},x_{k+1}) \\
\ge\ & \frac{\gamma_k}{L}\langle F(x_k) - F(y_k), y_k - x_{k+1}\rangle + \frac{\gamma_k}{L}\langle F(y_k) - F(\bar{x}), y_k - \bar{x}\rangle + D(x_{k+1},y_k) + D(y_k,x_k) \\
\ge\ & -\gamma_k\|x_k - y_k\|\|y_k - x_{k+1}\| + D(x_{k+1},y_k) + D(y_k,x_k) \\
\ge\ & (1-\gamma_k)\left(\|x_{k+1} - y_k\|^2 + \|y_k - x_k\|^2\right)/2,
\end{aligned}$$

where the second inequality uses (40), and the last inequality uses (7). Thus, $\{D(\bar{x},x_k)\}\downarrow$ for any solution $\bar{x}$ of (41) and $\{\|x_k\|\}$ is bounded. If in addition $\limsup_k \gamma_k < 1$, then $\{\|x_{k+1} - y_k\|^2 + \|y_k - x_k\|^2\} \to 0$, and it can be shown that $\{x_k\}$ converges weakly to a solution of (41) when $\mathcal{E}$ is finite-dimensional or a Hilbert space; see, e.g., [3, proof of Theorem 4.1(a)], [44, proof of Theorem 3.4(b)].

6 A Numerical Example

How do Algorithms 1, 3, 4 compare with each other? We apply them to solve the matrix game problem, also solved in [26, 32]:

$$\min_{u\in U}\max_{v\in V}\langle v, Au\rangle, \tag{55}$$

where U and V are the unit simplices in $\Re^n$ and $\Re^m$, $A \in \Re^{m\times n}$, and $\langle\cdot,\cdot\rangle$ is the usual inner product. Each entry of A is generated independently and uniformly in the interval $[-1,1]$ with probability p and otherwise is set to 0.

As in [32, Section 6], we consider a smooth approximation of (55):

$$\min_{x\in U} f_\mu(x) := \max_{v\in V}\left(\langle v, Ax\rangle - \mu\left(\ln m + \sum_{i=1}^m v_i \ln v_i\right)\right). \tag{56}$$

Then f_μ has the form (9) and $f_\mu(x) \le f_0(x) = \max_i(Ax)_i \le f_\mu(x) + \mu\ln m$ for all $x \in U$. By setting $\mu = \frac{\epsilon}{2\ln m}$, its optimal value is within $\epsilon/2$ of (55), and ∇f_μ is Lipschitz continuous on U with constant

$$L_\mu = \frac{1}{\mu} = \frac{2\ln m}{\epsilon}$$

with respect to the 1-norm; see [32, Eq. (4.8)]. Thus (56) is a special case of (1) with $\mathcal{E} = \Re^n$, $\|\cdot\|$ being the 1-norm, and $P = \delta_U$. We apply Algorithms 1 and 3 to (56) with $X_k = \Re^n$, $h(x) = \sum_{j=1}^n x_j \ln x_j$ (so (7) holds for $z \in U \cap (0,\infty)^n$), x_{k+1} given by either (18) or (35), (36) (so x_{k+1} has closed form and is computable in $O(n)$ time), θ_k, ϑ_k given by (34), and $x_0 = z_0 = (\frac{1}{n}, \ldots, \frac{1}{n})$. We do not consider (17) nor Algorithm 2 since they entail solving a quadratic knapsack problem per iteration, which requires a more complex algorithm to solve in $O(n)$ time; see [19] and references therein. To accelerate convergence, we initialize L to $L_\mu/8$ and increase L by a factor of 2 and repeat the iteration whenever $L < L_\mu$ and (23) is violated. Propositions 1 and 3 still hold with this modification. We have $f_0(x_{k+1}) \leq \inf_U f_0 + \epsilon$ whenever $f_\mu(x_{k+1}) \leq \inf_U f_\mu + \frac{\epsilon}{2}$ which, by Corollaries 1 and 3 (and using $\max_{x\in U} D(x, z_0) = \max_{x\in U} h(x) - h(z_0) = \ln n$), occurs whenever

$$k \geq \sqrt{\frac{4L_\mu \ln n}{\epsilon/2}} - 1 = \frac{4\sqrt{\ln m \ln n}}{\epsilon} - 1. \tag{57}$$

Thus, we can terminate the methods at iteration k when the bound in (57) is reached. To accelerate termination as is suggested in [22, 23, 32] and by Corollaries 1(b) and 3(c), we also check every 5 iterations whether the duality gap at x_{k+1} and $\bar{v}_k$ given by the second formula in (25), with $v_k = \left(\frac{e^{(Ay_k)_i/\mu}}{\sum_i e^{(Ay_k)_i/\mu}}\right)_{i=1,\ldots,m}$, is below ϵ, i.e.,

$$\max_i (Ax_{k+1})_i - \min_j (A^* \bar{v}_k)_j \leq \epsilon, \tag{58}$$

and terminate if yes. This check takes only $O(n)$ time since $A^* v_k$ is available as a byproduct of evaluating ∇f_μ at y_k and Ax_{k+1} is available from evaluating $f_\mu(x_{k+1})$ to check (23). If (23) has not been checked at iteration k, then we use Ay_k in place of Ax_{k+1}. The results obtained are reported below.

		Alg1	**Alg3a**	**Alg3b**	**Alg4**
$n/m/p$	ϵ	k/%/cpu time	k/%/cpu time	k/%/cpu time	t/%/cpu time
1000/100/.01	.001	3325/14/5	10510/46/9	9790/43/13	2400/20/5
	.0001	20635/9/23	61865/27/45	60215/26/71	1150/10/3
1000/100/.1	.001	4265/18/8	4265/18/8	4265/18/10	1150/10/3
	.0001	42470/18/87	70895/31/103	70850/31/136	11085/9/38
1000/1000/.01	.001	4760/17/12	4760/17/11	4760/17/14	1565/11/7
	.0001	50820/18/126	50820/18/121	50820/18/146	18485/13/90
1000/1000/.1	.001	3900/14/33	3900/14/33	3900/14/34	1050/7/32
	.0001	38605/14/333	49645/18/412	49275/17/436	9915/7/318
10000/100/.01	.001	10005/38/142	10005/38/128	10005/38/171	10005/72/187
10000/100/.1	.001	10005/38/201	10005/38/185	10005/38/238	10005/72/456
10000/1000/.01	.001	10005/31/202	10005/31/191	10005/31/238	10005/62/457
10000/1000/.1	.001	10005/31/706	10005/31/695	10005/31/743	10005/62/2977

Table 1: Comparing Algorithms 1, 3, 4 for different problem dimension, sparsity, and termination tolerance. (Algorithms 3a and 3b choose x_{k+1} by, respectively, (18) and (35), (36).)

The problem (55) is a special case of (43) with $\phi(u, v) = \langle v, Au \rangle$, $\pi = \delta_U$, $\varpi = \delta_V$. This corresponds to (41) with $\mathcal{E} = \Re^{n+m}$, and

$$F(u, v) = \begin{bmatrix} A^* v \\ -Au \end{bmatrix}, \qquad P(u, v) = \delta_{U\times V}(u, v).$$

We endow $\mathcal{E}$ with the 1-norm, so F is Lipschitz continuous on $\mathcal{E}$ with constant 1 (since $A \in [-1, 1]^{m\times n}$). Similar to [26, Section 6], we apply Algorithm 4 with $X_k = \Re^{n+m}$, $h(u, v) = \sum_{j=1}^n u_j \ln u_j + \sum_{i=1}^m v_i \ln v_i$ (so (7) holds for $z \in (U \times V) \cap (0, \infty)^{n+m}$), y_k given by (48) (so y_k has closed form and is computable in $O(n)$ time), $\gamma_k = 1$, $u_0 = (\frac{1}{n}, \ldots, \frac{1}{n})$, $v_0 = (\frac{1}{m}, \ldots, \frac{1}{m})$. We can choose $L = 1$. To accelerate convergence, we initialize L to $1/8$ and increase L by a factor of 2 and repeat the iteration whenever $L < 1$ and (45) is violated. We check every 5 iterations whethere the duality gap at $\bar{x}_{0,t} = (\bar{u}_{0,t}, \bar{v}_{0,t})$, defined as in (50) and Lemma 2, is below ϵ, i.e.,

$$e(\bar{x}_{0,t}) = \max_i (A\bar{u}_{0,t})_i - \min_j (A^* \bar{v}_{0,t})_j \leq \epsilon,$$

and terminate if yes. By the remarks following Proposition 4 (and using $\max_{x\in U\times V} D(x,x_0) = \ln n + \ln m$), this occurs whenever $\frac{\ln n+\ln m}{t+1} \le \epsilon$ or

$$t \ge \frac{\ln n + \ln m}{\epsilon} - 1. \tag{59}$$

The iteration bound (57) for Algorithms 1 and 3 is about a factor of 2 more than the bound (59) for Algorithm 4. On the other hand, Algorithms 1 and 3 require only two or three matrix-vector multiplications per iteration by A and A^* (depending on whether (23) is checked), whereas Algorithm 4 requires four such multiplications per iteration. The remaining computations per iteration are $O(n)$ for these methods.

All methods are coded in Matlab, with A stored in sparse format. All runs are performed on an HP DL360 workstation, under Matlab 7.2.0. We report in Table 1 the number of iterations, also expressed as a percentage of the bounds in (57) and (59), and the cpu time (in seconds) for Algorithms 1, 3 ("a" for (18) and "b" for (35), (36)), and Algorithm 4, for different m, n, p, ϵ. Thus Alg1, Alg3b, Alg4 are similar to the methods in [4, Section 5], [32, Section 5.3], [26], but with a dynamically adjusted L. As can be seen from Table 1, Alg3b is slower than the other three methods in terms of cpu time (though it uses the same or fewer iterations than Alg3a). Alg4 is fastest on problems with $n = 1000$ while Alg3a is fastest on problems with $n = 10000$. For Alg1, Alg3a, and Alg3b, termination occurs several times faster using the average dual vector $\bar{v}_k$ in (58) than using v_k. The number of iterations seems independent of the sparsity of A. These results corroborate those reported in [32, Tables 2 and 3] for a method that is essentially Algorithm 3 with $L = L_\mu$ and (17), and in [26, Table 6.1] for a method that is essentially Algorithm 4 with y_k given by (48), $L = 1$, h being a regularized entropy, and γ_k dynamically adjusted. The Matlab code can be downloaded from

`http://www.math.washington.edu/~tseng/papers.html`

7 Discussions and extensions

We assumed for simplicity that h is differentiable on an open set containing dom$P$. This assumption can be relaxed to

(a) dom$P \subseteq$ dom$h$ and dom∇h = int(dom$h$).

(b) For any linear function $\ell : \mathcal{E} \to \Re$ and $\alpha > 0$, $\arg\min_x\{\ell(x) + P(x) + \alpha h(x)\} \in$ dom∇h.

Then by taking $z_0 \in$ dom$P \cap$ dom∇h, Algorithms 1 and 3 with $X_k = \mathcal{E}$ would maintain $z_k, y_k \in$ dom$P \cap$ dom∇h for all k. Similarly, by taking $x_0 \in$ dom$P \cap$ dom∇h, Algorithm 4 with $X_k = \mathcal{E}$ and (48) would maintain $x_k, y_k \in$ dom$P \cap$ dom∇h for all k. The remaining proofs are unchanged. The above relaxed assumption is satisfied by (3) with X being a Cartesian product of simplices and spectahedra and h being the $x\ln(x)$-entropy function; see [3, Definition 2.1 and Proposition 2.1]. What if X_k is given by (16) or (46)? In the case of (3) with

$$\mathcal{E} = \Re^n, \qquad X = \{x \mid Ax = b,\ x \ge 0\} \text{ bounded}, \qquad h(x) = \sum_{j=1}^n h_j(x_j),$$

and each $h_j : [0,\infty) \to \Re$ being continuous and twice differentiable on $(0,\infty)$ with $h''(t) > 0$, $\lim_{t\to 0^+} h_j'(t) = -\infty$ (e.g., $h_j(t) = t\ln t$), it can be seen that if $\arg\min_{x\in X_k}\{\ell(x) + h(x)\}$ has zero components, with $\ell : \mathcal{E} \to \Re$ being any linear function, then those component must be zero for all $x \in X_k$ and hence can be eliminated from the computation. An alternative approach, suggested in [8, 26], is to use the regularized entropy $(t + \delta/n)\ln(t + \delta/n)$ with small $\delta > 0$ (e.g., $\delta = 10^{-16}$).

Given the similarities between Algorithms 1 and 3 and Propositions 1 and 3, we may ask whether we can switch between $D(\cdot, z_k)$ and $h(\cdot)$ in (12) and (31) and accordingly switch between Properties 1 and

2 in the proofs. The answer seems to be 'no.' Intuitively, if we use $D(\cdot, z_k)$, then we are limited to using Property 1 with $z = z_k$ and $z_+ = z_{k+1}$. If we use $h(\cdot)$, then we are limited to using Property 2 with $z = z_k$ or $z = z_{k+1}$. This in turn limits the choice of ψ for obtaining a recursion like (22) and (38).

Can the interpolation techniques and their analysis be extended to other gradient methods, such as incremental gradient methods [9, Section 1.5] and coordinate gradient descent methods [45]? Can the primal-dual method in [33] and the dual extrapolation method in [34] be treated in a similarly unified way? Can other proximity functions, such as those studied in [3, 4], be used?

Acknowledgement. The author thanks Dimitri Bertsekas for discussions and encouragements that stimulated this work. He also thanks Alfred Auslender, Zhaosong Lu, Renato Monteiro, Yurii Nesterov, and Marc Teboulle for comments on the first draft of this paper and, in particular, for bringing to his attention [3, 7, 11, 23, 34, 35]. A further thanks to Zhaosong and Renato for noting the dual averaging scheme in [32, Theorem 3] and [22, Theorem 2.2], and their unpublished results similar to parts of Sections 3 and 4.

References

[1] A. d'Aspremont, O. Banerjee, O., and L. E. Ghaoui, First-order methods for sparse covariance selection, EECS Department, University of California, Berkeley, 2007; to appear in SIAM J. Matrix Anal. Appl.

[2] A. Auslender, Minimisation de fonctions localement lipschitziennes: applications à la programmation mi-convexe, mi-différentiable, in O. L. Mangasarian, R. R. Meyer and S. M. Robinson, editors, Nonlinear Programming, 3, Academic Press, New York (1978), 429-460.

[3] A. Auslender and M. Teboulle, Interior projection-like methods for monotone variational inequalities, Math. Program. 104 (2005), 39-68.

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

[5] H. H. Bauschke, J. M. Borwein, and P. L. Combettes, Bregman monotone optimization algorithms, SIAM J. Control Optim. 42 (2003), 596-636.

[6] A. Beck and M. Teboulle, Mirror descent and nonlinear projected subgradient methods for convex optimization, Oper. Res. Letters 31 (2003), 167-175.

[7] A. Beck and M. Teboulle, A fast iterative shrinkage-threshold algorithm for linear inverse problems, Report, Department of Industrial Engineering and Management, Technion, Haifa, Israel, 2008.

[8] A. Ben-Tal and A. Nemirovski, Non-Euclidean restricted memory level method for large-scale convex optimization, Math. Program. 102 (2005), 407-456.

[9] D. P. Bertsekas, Nonlinear Programming, 2nd edition, Athena Scientific, Belmont, 1999.

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

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

[12] S. Chen, D. Donoho, D., and M. Saunders, Atomic decomposition by basis pursuit, SIAM Rev. 43 (2001), 129-159.

[13] J. C. Dunn, Global and asymptotic convergence rate estimates for a class of projected gradient processes, SIAM J. Control Optim. 19 (1981), 368-400.

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

[15] F. Facchinei and J.-S. Pang, Finite-Dimensional Variational Inequalities and Complementarity Problems, Vol. I, Springer-Verlag, New York, 2003.

[16] M. P. Friedlander and P. Tseng, Exact regularization of convex programs, SIAM J. Optim. 18 (2007), 1326-1350.

[17] M. Fukushima and H. Mine, A generalized proximal point algorithm for certain non-convex minimization problems, Int. J. Systems Sci. 12 (1981), 989-1000.

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

[19] K. C. Kiwiel, On linear time algorithms for the continuous quadratic knapsack problem, report, Systems Research Institute, Warsaw, Poland, 2006; to appear in J. Optim. Theory Appl.

[20] G.M. Korpelevich, The extragradient method for finding saddle points and other problems, Matecon 12 (1976), 747-756.

[21] G. Lan, Z. Lu, and R. D. C. Monteiro, Primal-dual first-order methods with $\mathcal{O}(1/\epsilon)$ iteration-complexity for cone programming, Manuscript, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, December 2006.

[22] Z. Lu, Smooth optimization approach for sparse covariance selection, Report, Department of Mathematics, Simon Fraser University, Burnaby, January 2008; submitted to SIAM J. Optim.

[23] Z. Lu, R. D. C. Monteiro, and M. Yuan, Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression, Manuscript, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, January 2008.

[24] Z. Lu, A. S. Nemirovski and R. D. C. Monteiro, Large-scale semidefinite programming via saddle point mirror-prox algorithm, Math. Program. 109 (2007), 211-237.

[25] L. Meier, S. van de Geer, and P. Bühlmann, The group Lasso for logistic regression, Report, Seminar für Statistik, ETH Zürich, Zürich, March 2006; to appear in J. Royal Statist. Soc. B.

[26] A. Nemirovski, Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems, SIAM J. Optim. 15 (2005), 229-251.

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

[28] Y. Nesterov, A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$, Doklady AN SSSR 269 (1983), 543-547; translated as Soviet Math. Dokl.

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

[30] Y. Nesterov, Smoothing technique and its applications in semidefinite optimization, Report, CORE, Catholic University of Louvain, Louvain-la-Neuve, Belgium, October 2004.

[31] Y. Nesterov, Introductory Lectures on Convex Optimization, Kluwer Academic Publisher, Dordrecht, The Netherlands, 2004.

[32] Y. Nesterov, Smooth minimization of nonsmooth functions, Math. Program. 103 (2005), 127-152.

[33] Y. Nesterov, Excessive gap technique in nonsmooth convex minimization, SIAM J. Optim. 16 (2005), 235-249.

[34] Y. Nesterov, Dual extrapolation and its applications to solving variational inequalities and related problems, Math. Program. 109 (2007), 319-344.

[35] Y. Nesterov, Gradient methods for minimizing composite objective function, Report, CORE, Catholic University of Louvain, Louvain-la-Neuve, Belgium, September 2007.

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

[37] P. Richtárik, Some algorithms for large-scale convex and linear minimization in relative scale, Ph.D. thesis, Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 2007.

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

[39] R. T. Rockafellar, Conjugate Duality and Optimization, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1974.

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

[41] P. J. Silva and J. Eckstein, Double-regularization proximal methods, with complementarity applications, Comp. Optim. Appl. 33 (2006), 115-156.

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

[43] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Royal Statist. Soc. B. 58 (1996), 267-288.

[44] P. Tseng, A modified forward-backward splitting method for maximal monotone mappings, SIAM J. Control Optim. 38 (2000), 431-446.

[45] P. Tseng and S. Yun, A coordinate gradient descent method for nonsmooth separable minimization, Report, Department of Mathematics, University of Washington, Seattle, June 2006 (revised Feb 2007); to appear in Math. Program. B.

[46] P. Tseng and S. Yun, A block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization, Report, Department of Mathematics, University of Washington, Seattle, January 2008; to appear in J. Optim. Theory Appl.

[47] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, Sparse reconstruction by separable approximation, Report, Computer Sciences Department, University of Wisconsin, Madison, October 2007.