An accelerated HPE-type algorithm for a class of composite convex-concave saddle-point problems

Yunlong He [*]   Renato D.C. Monteiro [†]

April 29, 2014

Abstract

This article proposes a new algorithm for solving a class of composite convex-concave saddle-point problems. The new algorithm is a special instance of the hybrid proximal extragradient framework in which a Nesterov's accelerated variant is used to approximately solve the prox subproblems. One of the advantages of the new method is that it works for any constant choice of proximal stepsize. Moreover, a suitable choice of the latter stepsize yields a method with the best known (accelerated inner) iteration complexity for the aforementioned class of saddle-point problems. In contrast to the smoothing technique of [12], our accelerated method does not assume that feasible set is bounded due to its proximal point nature. Experiment results on three problem sets show that the new method significantly outperforms Nesterov's smoothing technique of [12] as well as a recently proposed accelerated primal-dual method in [5].

Keywords: saddle-point problem, composite convex optimization, monotone inclusion problem, inexact proximal point method, hybrid proximal extragradient, accelerated method, complexity, smoothing.

AMS subject classifications: 90C60, 90C25, 90C30, 47H05, 47J20, 65K10, 65K05.

1   Introduction

A broad class of optimization, saddle-point (SP), equilibrium and variational inequality problems can be posed as the *monotone inclusion problem*, namely: finding x such that

$$0 \in T(x), \tag{1}$$

where T is a maximal monotone point-to-set operator. The proximal point method, proposed by Rockafellar [14], is a classical iterative scheme for solving the monotone inclusion problem which generates a sequence $\{z_k\}$ according to

$$\|z_k - (\lambda_k T + I)^{-1}(z_{k-1})\| \le e_k, \qquad \sum_{k=1}^{\infty} e_k < \infty.$$

---

[*]School of Mathematics, Georgia Institute of Technology, Atlanta, GA, 30332. (email: `heyunlong@gatech.edu`).

[†]School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0205. (email: `monteiro@isye.gatech.edu`). The work of this author was partially supported by NSF Grant CMMI-1300221.

This method has been used as a generic framework for the design and analysis of several implementable algorithms.

New inexact versions of the proximal point method which uses instead relative error criteria were proposed by Solodov and Svaiter [16, 17, 18, 19]. In this article, we will use one of these variants, namely, the hybrid proximal extragradient (HPE) framework studied in [16], to develop and analyze a new algorithm, and we now briefly discuss this framework. The *exact* proximal point iteration from z with stepsize $\lambda > 0$ is given by $z_+ = (\lambda T + I)^{-1}(z)$, which is equivalent to

$$r \in T(z_+), \qquad \lambda r + z_+ - z = 0. \tag{2}$$

In each step of the HPE, the above *proximal system* is solved inexactly with $(z, \lambda) = (z_{k-1}, \lambda_k)$ to obtain $z_k = z_+$ as follows. For a given constant $\sigma \in [0, 1)$, a triple $(\tilde{z}, \tilde{r}, \varepsilon) = (\tilde{z}_k, \tilde{r}_k, \varepsilon_k)$ satisfying the HPE error criteria

$$\tilde{r} \in T^\varepsilon(\tilde{z}), \qquad \|\lambda \tilde{r} + \tilde{z} - z\|^2 + 2\lambda \varepsilon \le \sigma^2 \|\tilde{z} - z\|^2 \tag{3}$$

is found, where T^ε denotes the ε-enlargement [3] of T. (It has the property that $T^\varepsilon(z) \supset T(z)$ for each z.) Note that this construction relaxes both the inclusion and the equation in (2). Finally, instead of choosing $\tilde{z}$ as the next iterate z_+, the HPE framework computes the next iterate z_+ by means of the *extragradient* step $z_+ = z - \lambda \tilde{r}$.

Iteration complexity results for the HPE framework were established in [9] and these results depend on the distance of the initial iterate to the solution set instead of the diameter of the feasible set. Applications of the HPE framework to the iteration complexity analysis of several zero-order (resp., first-order) methods for solving monotone variational inequalities and monotone inclusions (resp., saddle-point problems) are discussed in [9] and in the subsequent papers [8, 10]. More specifically, by viewing Korpelevich's method [7] as well as Tseng's modified forward-backward splitting (MF-BS) method [20] as special cases of the HPE framework, the authors have established in [8, 9] the pointwise and ergodic iteration complexities of these methods applied to either: monotone variational inequalities, monotone inclusions consisting of the sum of a Lipschitz continuous monotone map and a maximal monotone operator with an easily computable resolvent, and convex-concave saddle-point problems.

A framework of block-decomposition (BD) prox-type algorithms is introduced in [10] for solving the monotone inclusion problem consisting of the sum of a continuous monotone map and a point-to-set maximal monotone operator with a separable two-block structure. The above BD framework is a special case of the HPE framework which approximately solves the proximal subproblem corresponding to the two-block inclusion by (possibly, approximately) solving two smaller proximal single-block subproblems. When the stepsize is sufficiently small, the latter two subproblems can be approximately solved by performing a (single) step similar to the one performed by the gradient method in the case of composite convex optimization or by Korpelevich's and/or Tseng's MF-BS method in the more general context of variational inequality and maximal monotone inclusions. More recently, the authors have studied in [6] an accelerated BD prox-type algorithm for solving the saddle-point (and more generally Nash equilibrium) problem where the above two proximal subproblems (which in this case are composite convex optimization) are solved by an accelerated variant of Nesterov's optimal method. The accelerated BD method is generally able to take a much larger stepsize than the ones for the aforementioned BD methods and, as a consequence, performs significantly less number of outer iterations. Moreover, computational results have shown that the

accelerated BD method can substantially outperform the aforementioned ones on many relevant classes of saddle-point and Nash equilibrium instances.

Given proper closed convex (possibly nonsmooth) functions g_1 and g_2, this paper considers the class of composite convex-concave saddle-point problem

$$\min_{x \in X} \max_{y \in Y} \Psi(x, y) = f(x) + \langle Ax, y \rangle + g_1(x) - g_2(y) \tag{4}$$

where $X := \operatorname{dom} g_1$, $Y := \operatorname{dom} g_2$, A is a linear operator and f is a differentiable convex function whose gradient is L_f-Lipschitz continuous on X. It is assumed that g_1 and g_2 are simple functions in the sense that subproblems of the form

$$\min_{x \in X} \frac{1}{2} \|x - \tilde{x}\|^2 + \lambda g_1(x) \quad \text{and} \quad \min_{y \in Y} \frac{1}{2} \|y - \tilde{y}\|^2 + \lambda g_2(y) \tag{5}$$

are easy to solve for any $\tilde{x}$, $\tilde{y}$ and $\lambda > 0$. Since (4) is well-known to be equivalent to monotone inclusion problem (1) with T given by

$$T(x, y) = \partial(\Psi(\cdot, y) - \Psi(x, \cdot))(x, y), \tag{6}$$

any instance of the HPE method, including the ones already discussed above, can be used to solve it. In particular, by taking a sufficiently small stepsize λ, Korpelevich's (resp., Tseng's) method is able to approximately solve the current proximal subproblem (i.e., a triple satisfying (3)) by solving at most four (resp., two) subproblems of the form (5).

This paper presents an accelerated instance of the HPE framework which arbitrarily chooses the stepsize λ and solves (3) with T given by (6) by using a Nesterov's accelerated variant for smooth composite saddle-point problems. Both the outer (i.e., HPE) iteration complexity and the inner (i.e., accelerated variant) iteration complexity are derived for the method in terms of a general stepsize λ. As in Tseng's and Korpelevich's methods, just a few (namely, three) subproblems of the form (5) are solved within an inner iteration. Hence, choosing λ so as to minimize the overall number of inner iterations is the best strategy towards minimizing the overall complexity of the accelerated HPE method. An explicit formula in terms of $\|A\|$, L_f, the distance d_0 of the initial iterate to the set of saddle-points of (4) and the specified tolerances is then derived for such a stepsize. Clearly, since d_0 is not known a priori, the above stepsize can not be computed but an alternative stepsize λ depending only on $\|A\|$ and L_f is provided which is optimal for the most common saddle-point problems of the form (4). Moreover, when the feasible set $X \times Y$ is bounded, the expression for the above optimal stepsize with d_0 replaced by the diameter of $X \times Y$ yields another stepsize which implies (if an appropriate choice of inner product in the (x, y)-space is made) an overall complexity for the accelerated HPE method that is similar to that of Nesterov's smoothing technique (see [12]) for finding an ε-saddle-point of (4) (see equation (76) below). It is worth emphasizing that, in contrast to Nesterov's smoothing technique of [12], our accelerated method for solving (4) does not assume that $X \times Y$ is bounded due to its proximal point nature.

Our paper is organized as follows. Section 2 contains three subsections which provide the necessary background material for our presentation. More specifically, Subsection 2.1 presents the notation and basic definitions used in the paper. Subsection 2.2 reviews a Nesterov's accelerated variant for solving composite convex optimization problems. Subsection 2.3 discusses the HPE framework for the monotone inclusion problem. Section 3 reviews the definition of the saddle-point problem, its connection to the composite convex-concave min-max problem and the notion

of an approximate saddle-point. Moreover, this section specializes the HPE framework to the context of the saddle-point problem and states its convergence properties. Section 4 presents a scheme for finding a solution of (3) with T given by (4) and (6) (and w.l.o.g. $\lambda = 1$) based on the Nesterov's accelerated variant of Subsection 2.2 applied to an associated composite convex-concave min-max problem. Section 5 presents a special instance of the HPE framework based on the accelerated scheme of Section 4 for solving the composite convex-concave min-max problem (4) and derives its pointwise and ergodic outer, and corresponding overall inner, iteration complexities for finding approximate saddle-points. It also discusses optimal ways of choosing the stepsize so as to minimize the overall ergodic inner iteration complexity of the accelerated HPE method for solving (4). Finally, numerical results are presented in Section 6 showing that the new method significantly outperforms Nesterov's smooth approximation scheme [12] and the accelerated primal-dual method in [5] on three classes of composite convex-concave min-max problems of the form (4).

1.1 Previous most related works

In the context of variational inequalities, Nemirovski [11] has established the ergodic iteration complexity of an extension of Korpelevich's method [7], namely, the Mirror-prox algorithm, under the assumption that the feasible set of the problem is bounded.

Nesterov's smoothing scheme [12] solves problem (4) under the assumption that X and Y are compact convex sets and g_1 is the indicator function of X. It consists of first approximating the objective function of (4) by a convex differentiable function with Lipschitz continuous gradient and then applying an accelerated gradient-type method (see e.g. [12, 2, 21]) to the resulting approximation problem. It is shown that, if the approximation is properly chosen, the above scheme obtains an ε solution of (4) in at most

$$\mathcal{O}\left(\frac{\|A\|}{\varepsilon}D_X D_Y + \sqrt{\frac{L_f}{\varepsilon}}D_X\right)$$

iterations where D_X and D_Y are the diameters of X and Y. The latter bound is also known to be optimal (see for example the discussion in paragraph (1) of Subsection 1.1 of [5]).

Chambolle and Pock [4] have developed and established the convergence rate for a primal-dual method for solving problem (4) in the context of $f(x)$ being simple and g_1 being the indicator function of the feasible set X. A recent paper [5] considers problem (4) with g_1 being the indicator function of the feasible set X and proposed an accelerated primal-dual algorithm that achieved optimal convergence rate for both cases that the feasible set of the problem is bounded or unbounded.

2 Preliminaries

This section contains three subsections. The first one presents the notation and basic definitions that will be in the paper. The second subsection reviews a variant of Nesterov's accelerated method for composite convex optimization problem. The third subsection describes the HPE framework for the monotone inclusion problem.

2.1 Notation and basic definitions

We denote the sets of real numbers by $\Re$. For a matrix $W \in \Re^{m \times n}$, we denote its Frobenius norm by $\|W\|$. Let $\mathcal{S}^n$ represent the linear space of $n \times n$ real symmetric matrices. For a matrix $W \in \mathcal{S}^n$,

we denote its largest eigenvalue by $\theta_{\max}(W)$. Let $\lceil z \rceil$ denote the smallest integer not less than $z \in \Re$. The n-th unit simplex $\Delta_n \subseteq \Re^n$ is defined as

$$\Delta_n = \left\{ z \in \Re^n : \sum_{i=1}^n z_i = 1, z_i \geq 0, i = 1, \ldots, n \right\}. \tag{7}$$

Throughout this paper, we let $\mathcal{Z}$ denote a finite dimensional inner product space with associated inner product denoted by $\langle \cdot, \cdot \rangle$ and the induced norm denoted by $\| \cdot \|$. For a given set $\Omega \subset \mathcal{Z}$, the diameter D_Ω of Ω is defined as

$$D_\Omega := \sup\{\|z - z'\| : z, z' \in \Omega\} \tag{8}$$

and the indicator function $\mathcal{I}_\Omega : \mathcal{Z} \to (-\infty, \infty]$ of Ω is defined as

$$\mathcal{I}_\Omega(z) = \begin{cases} 0, & z \in \Omega, \\ \infty, & z \notin \Omega. \end{cases}$$

Also, if Ω is nonempty and convex, the *orthogonal projection* $P_\Omega : \mathcal{Z} \to \mathcal{Z}$ onto Ω is defined as

$$P_\Omega(z) = \operatorname{argmin}_{z' \in \Omega} \|z' - z\| \quad \forall z \in \mathcal{Z}.$$

A relation $T \subseteq \mathcal{Z} \times \mathcal{Z}$ can be identified with a point-to-set operator $T : \mathcal{Z} \rightrightarrows \mathcal{Z}$ in which

$$T(z) := \{v \in \mathcal{Z} : (z, v) \in T\}, \quad \forall z \in \mathcal{Z}.$$

Note that the relation T is then the same as the graph of the point-to-set operator T defined as

$$Gr(T) := \{(z, v) \in \mathcal{Z} \times \mathcal{Z} : v \in T(z)\}.$$

An operator $T : \mathcal{Z} \rightrightarrows \mathcal{Z}$ is *monotone* if

$$\langle v - \tilde{v}, z - \tilde{z} \rangle \geq 0, \qquad \forall (z, v), (\tilde{z}, \tilde{v}) \in Gr(T).$$

Moreover, T is *maximal monotone* if it is monotone and maximal in the family of monotone operators with respect to the partial order of inclusion, i.e., $S : \mathcal{Z} \rightrightarrows \mathcal{Z}$ monotone and $Gr(S) \supset Gr(T)$ implies that $S = T$. Given a scalar ε, the ε-enlargement of a point-to-set operator $T : \mathcal{Z} \rightrightarrows \mathcal{Z}$ is the point-to-set operator $T^\varepsilon : \mathcal{Z} \rightrightarrows \mathcal{Z}$ defined as

$$T^\varepsilon(z) = \{v \in \mathcal{Z} \mid \langle z - \tilde{z}, v - \tilde{v} \rangle \geq -\varepsilon, \quad \forall \tilde{z} \in \mathcal{Z}, \forall \tilde{v} \in T(\tilde{z})\}, \quad \forall z \in \mathcal{Z}. \tag{9}$$

The effective domain of a function $f : \mathcal{Z} \to [-\infty, \infty]$ is defined as $\operatorname{dom} f := \{z \in \mathcal{Z} : f(z) < \infty\}$. Moreover, if f is differentiable at point $\tilde{z}$ such that $f(\tilde{z}) \in \Re$, its first-order (affine) approximation at $\tilde{z}$ is defined as

$$l_f(z; \tilde{z}) := f(\tilde{z}) + \langle \nabla f(\tilde{z}), z - \tilde{z} \rangle \qquad \forall z \in \mathcal{Z}. \tag{10}$$

The conjugate f^* of f is the function $f^* : \mathcal{Z} \to [-\infty, \infty]$ defined as

$$f^*(v) = \sup_{z \in \mathcal{Z}} \langle v, z \rangle - f(z), \quad \forall v \in \mathcal{Z}.$$

Given a scalar $\varepsilon \geq 0$, the ε-subdifferential of a function $f : \mathcal{Z} \to [-\infty, +\infty]$ is the operator $\partial_\varepsilon f : \mathcal{Z} \rightrightarrows \mathcal{Z}$ defined as

$$\partial_\varepsilon f(z) = \{v \mid f(\tilde{z}) \geq f(z) + \langle \tilde{z} - z, v \rangle - \varepsilon, \ \forall \tilde{z} \in \mathcal{Z}\}, \quad \forall z \in \mathcal{Z}. \tag{11}$$

When $\varepsilon = 0$, the operator $\partial_\varepsilon f$ is simply denoted by ∂f and is referred to as the subdifferential of f. The operator ∂f is trivially monotone if f is proper. If f is a proper closed convex function, then ∂f is maximal monotone [13].

The following result lists some useful properties about the ε-subdifferential of a proper convex function.

Proposition 2.1. *Let $f : \mathcal{Z} \to [-\infty, +\infty]$ be a proper convex function. Then*

(a) $\partial_\varepsilon f(z) \subseteq (\partial f)^\varepsilon (z)$ *for any $\varepsilon \geq 0$ and $z \in \mathcal{Z}$;*

(b) *if $v \in \partial f(z)$ and $f(\tilde{z}) < \infty$, then $v \in \partial_\varepsilon f(\tilde{z})$, where $\varepsilon := f(\tilde{z}) - [f(z) + \langle \tilde{z} - z, v \rangle] \geq 0$;*

(c) *if, in addition, f is closed, then $v \in \partial f(z)$ is equivalent to $z \in \partial f^*(v)$.*

The domain of a point-to-point map F is denoted by $\mathrm{Dom}\, F$. For a constant $L \geq 0$, a map $F : \mathrm{Dom}\, F \subseteq \mathcal{Z} \to \mathcal{Z}$ is said to be L-Lipschitz continuous on $\Omega \subseteq \mathrm{Dom}\, F$ if

$$\|F(z) - F(\tilde{z})\| \leq L\|z - \tilde{z}\|, \quad \forall z, \tilde{z} \in \Omega;$$

moreover, if in addition $\Omega = \mathrm{Dom}\, F$, we will simply say that F is L-Lipschitz continuous.

The following result gives a characterization of a strongly convex function in terms of its conjugate.

Proposition 2.2. *For a scalar $\beta > 0$ and a proper closed convex function $f : \mathcal{Z} \to [-\infty, \infty]$, the following two properties are equivalent:*

(a) *f is strongly convex with modulus β;*

(b) *f^* is differentiable everywhere and ∇f^* is $1/\beta$-Lipschitz continuous.*

Proof. This proposition is equivalent to Proposition 12.60 of [15] in view of the well-known fact that $f = f^{**}$. $\qquad\square$

2.2   Accelerated method for composite convex optimization

This subsection reviews a variant of Nesterov's accelerated first-order method [12, 21] for solving the composite convex optimization problem.

Let $\mathcal{X}$ denote a finite dimensional inner product space with associated inner product and norm denoted by $\langle \cdot, \cdot \rangle_{\mathcal{X}}$ and $\| \cdot \|_{\mathcal{X}}$, respectively. Consider the following composite convex optimization problem

$$\inf p(u) := \psi(u) + g(u) \tag{12}$$

where the functions $\psi : \mathrm{dom}\, \psi \to \Re$ and $g : \mathcal{X} \to [-\infty, \infty]$ satisfy the following conditions:

A.1) g is a proper closed convex function;

A.2) ψ is convex and differentiable on a closed convex set $\Omega \supseteq X := \operatorname{dom} g$;

A.3) the gradient of the function ψ is L-Lipschitz continuous on Ω.

We now explicitly state a variant of Nesterov's accelerated method for solving problem (12), which is due to Tseng (see Algorithm 3 in [21]).

[Algorithm 1] A variant of Nesterov's accelerated algorithm:

0) Let $u_0 \in \mathcal{X}$ and $\lambda > 0$ and set $\Gamma_0 = 0$, $\tilde{u}_0 = w_0 = P_\Omega(u_0)$, $k = 1$;

1) compute $(u_k, w_k, \tilde{u}_k) \in \Omega \times X \times X$ as

$$\Gamma_k := \Gamma_{k-1} + \frac{1 + \sqrt{1 + 4L\Gamma_{k-1}}}{2L}, \tag{13}$$

$$u_k := \frac{\Gamma_{k-1}}{\Gamma_k}\tilde{u}_{k-1} + \frac{\Gamma_k - \Gamma_{k-1}}{\Gamma_k}w_{k-1}, \tag{14}$$

$$w_k := \operatorname{argmin} \sum_{i=1}^{k} \frac{\Gamma_i - \Gamma_{i-1}}{\Gamma_k}l_\psi(u; u_i) + g(u) + \frac{1}{2\Gamma_k}\|u - u_0\|_\mathcal{X}^2, \tag{15}$$

$$\tilde{u}_k := \frac{\Gamma_{k-1}}{\Gamma_k}\tilde{u}_{k-1} + \frac{\Gamma_k - \Gamma_{k-1}}{\Gamma_k}w_k; \tag{16}$$

2) set $k \leftarrow k + 1$ and go to step 1.

end

We now make a few remarks about the relationship between the above method and Algorithm 3 of [21]. First, the latter method computes w_k as in (15) but with the quadratic term $\|u - u_0\|_\mathcal{X}^2/2$ replaced by a general strongly convex function $h(u)$. Second, Algorithm 3 of [21] assumes that X is closed, $\Omega = X$ and $u_0 \in X$ so that $u_0 = \tilde{u}_0 = w_0$. On the other hand, Algorithm 1 can start from any point in $\mathcal{X}$ and can handle problems in which X is not necessarily closed. In fact, its ability to start from any point in $\mathcal{X}$ will be exploited later on in Section 4.

We now state the main technical result from which the convergence rate of the above variant of Nesterov's accelerated algorithm immediately follows. The proof of this convergence result is similar to the proof of Corollary 3(a) of [21]. For the sake of completeness, we provide its proof in the Appendix.

Proposition 2.3. *The sequences $\{\Gamma_k\}$, $\{\tilde{u}_k\}$ and $\{u_k\}$ generated by Algorithm 1 satisfy*

$$\Gamma_k \geq \frac{k^2}{4L}, \qquad \Gamma_k p(\tilde{u}_k) \leq \sum_{i=1}^{k}(\Gamma_i - \Gamma_{i-1})[l_\psi(u; u_i) + g(u)] + \frac{1}{2}\|u - u_0\|_\mathcal{X}^2 \quad \forall u \in X. \tag{17}$$

As a consequence, the sequence $\{l_{\psi,k}\}$ of affine functions defined as

$$l_{\psi,k}(u) := \sum_{i=1}^{k} \frac{\Gamma_i - \Gamma_{i-1}}{\Gamma_k}l_\psi(u; u_i) \quad \forall u \in \mathcal{X} \tag{18}$$

satisfies

$$l_{\psi,k} \leq \psi, \qquad p(\tilde{u}_k) \leq l_{\psi,k}(u) + g(u) + \frac{1}{2\Gamma_k}\|u - u_0\|_\mathcal{X}^2 \quad \forall u \in X. \tag{19}$$

7

2.3 HPE framework for the monotone inclusion problem

Let $T : \mathcal{Z} \rightrightarrows \mathcal{Z}$ be a maximal monotone operator. The monotone inclusion problem for T consists of finding $z \in \mathcal{Z}$ such that

$$0 \in T(z). \tag{20}$$

We also assume throughout this subsection that this problem has a solution, that is, $T^{-1}(0) \neq \emptyset$.

We next review the HPE framework introduced in [16] for solving the above problem and state the iteration complexity results obtained for it in [9].

[HPE] Hybrid Proximal Extragradient Framework:

0) Let $z_0 \in \mathcal{Z}$ and $0 \leq \sigma < 1$ be given and set $k = 1$;

1) choose $\lambda_k > 0$ and find $\tilde{z}_k, \tilde{r}_k \in \mathcal{Z}$, $\sigma_k \in [0, \sigma]$ and $\varepsilon_k \geq 0$ such that

$$\tilde{r}_k \in T^{\varepsilon_k}(\tilde{z}_k), \qquad \|\lambda_k \tilde{r}_k + \tilde{z}_k - z_{k-1}\|^2 + 2\lambda_k \varepsilon_k \leq \sigma_k^2 \|\tilde{z}_k - z_{k-1}\|^2; \tag{21}$$

2) set $z_k = z_{k-1} - \lambda_k \tilde{r}_k$, set $k \leftarrow k + 1$, and go to step 1.

end

We now make several remarks about the HPE framework. First, the HPE framework does not specify how to choose λ_k and how to find $\tilde{z}_k$, $\tilde{r}_k$ and ε_k as in (21). The particular choice of λ_k and the algorithm used to compute $\tilde{z}_k, \tilde{r}_k$ and ε_k will depend on the particular implementation of the method and the properties of the operator T. Second, if $\tilde{z} := (\lambda_k T + I)^{-1} z_{k-1}$ is the *exact* proximal point iterate, or equivalently

$$\tilde{r} \in T(\tilde{z}), \tag{22}$$

$$\lambda_k \tilde{r} + \tilde{z} - z_{k-1} = 0, \tag{23}$$

for some $\tilde{r} \in \mathcal{Z}$, then $(\tilde{z}_k, \tilde{r}_k) = (\tilde{z}, \tilde{r})$ and $\varepsilon_k = 0$ satisfies (21). Therefore, the error criterion (21) relaxes the inclusion (22) to $\tilde{r} \in T^\varepsilon(\tilde{z})$ and relaxes equation (23) by allowing a small error relative to $\|\tilde{z}_k - z_{k-1}\|$.

We define a sequence of ergodic means $\{\tilde{z}_k^a\}$ associated with $\{\tilde{z}_k\}$ as

$$\tilde{z}_k^a := \frac{1}{\Lambda_k} \sum_{i=1}^{k} \lambda_i \tilde{z}_i, \qquad \text{where} \qquad \Lambda_k := \sum_{i=1}^{k} \lambda_i, \tag{24}$$

and define the sequences of ergodic residuals $\{\tilde{r}_k^a\}$ and $\{\varepsilon_k^a\}$ as

$$\tilde{r}_k^a := \frac{1}{\Lambda_k} \sum_{i=1}^{k} \lambda_i \tilde{r}_i, \qquad \varepsilon_k^a := \frac{1}{\Lambda_k} \sum_{i=1}^{k} \lambda_i (\varepsilon_i + \langle \tilde{z}_i - \tilde{z}_k^a, \tilde{r}_i - \tilde{r}_k^a \rangle). \tag{25}$$

The following result describes the pointwise and ergodic convergence rate properties of the HPE framework. Its proof can be found in Theorem 4.4, Lemma 4.5 and Theorem 4.7 of [9].

Theorem 2.4. *Let d_0 denote the distance of z_0 to $T^{-1}(0)$. Then, for every $k \in \mathbb{N}$, the following statements hold:*

8

(a) (pointwise convergence rate) $\tilde{r}_k \in T^{\varepsilon_k}(\tilde{z}_k)$ and there exists an index $i \leq k$ such that

$$\|\tilde{r}_i\| \leq d_0 \sqrt{\frac{1+\sigma}{1-\sigma}\left(\frac{1}{\sum_{j=1}^{k}\lambda_j^2}\right)}, \qquad \varepsilon_i \leq \frac{\sigma^2 d_0^2 \lambda_i}{2(1-\sigma^2)\sum_{j=1}^{k}\lambda_j^2}.$$

(b) (ergodic convergence rate) $\tilde{r}_k^a \in T^{\varepsilon_k^a}(\tilde{z}_k^a)$ and

$$\|\tilde{r}_k^a\| \leq \frac{2d_0}{\Lambda_k}, \qquad 0 \leq \varepsilon_k^a \leq \frac{2d_0^2}{\Lambda_k}\left(1+\frac{\sigma}{\sqrt{(1-\sigma^2)}}\right).$$

3 HPE framework for saddle-point problem

The section reviews the definition of the saddle-point problem, its connection to the composite convex-concave min-max problem and the notion of an approximate saddle-point. Moreover, this section specializes the HPE framework to the context of the saddle-point problem and states its convergence properties.

Throughout this paper, we let $\mathcal{X}$ be the finite dimensional inner product space as described in Subsection 2.2 and $\mathcal{Y}$ denote a finite dimensional inner product space with associated inner product denoted by $\langle \cdot, \cdot \rangle_{\mathcal{Y}}$ and associated norm denoted by $\|\cdot\|_{\mathcal{Y}}$. We endow the product space $\mathcal{X} \times \mathcal{Y}$ with the canonical inner product defined as

$$\langle (x,y),(x',y') \rangle = \langle x,x' \rangle_{\mathcal{X}} + \langle y,y' \rangle_{\mathcal{Y}}, \quad \forall (x,y),(x',y') \in \mathcal{X} \times \mathcal{Y}. \tag{26}$$

The associated norm, denoted by $\|\cdot\|$ for shortness, is then given by

$$\|(x,y)\| = \sqrt{\|x\|_{\mathcal{X}}^2 + \|y\|_{\mathcal{Y}}^2}, \quad \forall (x,y) \in \mathcal{X} \times \mathcal{Y}.$$

We will now review the saddle-point problem and some of its basic properties. Given two nonempty convex sets $X \subseteq \mathcal{X}$ and $Y \subseteq \mathcal{Y}$, we consider throughout this section a function $\Psi : \mathcal{X} \times \mathcal{Y} \to [-\infty, +\infty]$ satisfying the following condition:

B.1) $\Psi(x,y)$ is finite-valued on $X \times Y$ and

$$\Psi(x,y) = \begin{cases} \infty, & x \notin X, \\ -\infty, & x \in X, y \notin Y. \end{cases} \tag{27}$$

The saddle-point problem determined by the triple $(\Psi; X, Y)$, denoted by $SP(\Psi; X, Y)$, consists of finding a pair $(x,y) \in X \times Y$ such that

$$\Psi(x,y') \leq \Psi(x,y) \leq \Psi(x',y), \quad \forall (x',y') \in X \times Y. \tag{28}$$

Clearly, (x,y) is a saddle-point of $SP(\Psi; X, Y)$ if and only if $(x,y) \in X \times Y$ and

$$(0,0) \in T(x,y) := \partial[\Psi(\cdot,y) - \Psi(x,\cdot)](x,y). \tag{29}$$

Define the primal and dual functions $p : X \to (-\infty, +\infty]$ and $d : Y \to [-\infty, +\infty)$, respectively, as

$$p(\tilde{x}) = \sup_{\tilde{y} \in Y} \Psi(\tilde{x}, \tilde{y}), \qquad d(\tilde{y}) = \inf_{\tilde{x} \in X} \Psi(\tilde{x}, \tilde{y}), \qquad \forall (\tilde{x}, \tilde{y}) \in X \times Y, \tag{30}$$

9

and consider the pair of optimization problems associated with $SP(\Psi; X, Y)$:

$$p_* := \inf_{\tilde{x} \in X} p(\tilde{x}) = \inf_{\tilde{x} \in X} \sup_{\tilde{y} \in Y} \Psi(\tilde{x}, \tilde{y}) \tag{31}$$

and

$$d_* := \sup_{\tilde{y} \in Y} d(\tilde{y}) = \sup_{\tilde{y} \in Y} \inf_{\tilde{x} \in X} \Psi(\tilde{x}, \tilde{y}) \tag{32}$$

Then, the weak duality inequality says that

$$p(\tilde{x}) \geq d(\tilde{y}), \quad \forall (\tilde{x}, \tilde{y}) \in X \times Y. \tag{33}$$

Moreover, it is well-known that (x, y) is a saddle-point if and only if $(x, y) \in X \times Y$ and $p(x) = d(y)$. In view of (33), the latter condition is equivalent to $x \in X$ and $y \in Y$ be optimal solutions of (31) and (32), respectively, and the optimal duality gap $p_* - d_*$ be equal to zero.

We now give a definition of an approximate saddle-point.

Definition 3.1. *Given $(\rho, \varepsilon) \in \Re_+ \times \Re_+$, $z = (x, y) \in X \times Y$, $r \in \mathcal{X} \times \mathcal{Y}$ and $\tilde{\varepsilon} \in \Re_+$, the triple $(z, r, \tilde{\varepsilon})$ is called a (ρ, ε)-saddle-point of Ψ if $\|r\| \leq \rho$, $\tilde{\varepsilon} \leq \varepsilon$ and*

$$r \in \partial_{\tilde{\varepsilon}}[\Psi(\cdot, y) - \Psi(x, \cdot)](x, y), \tag{34}$$

Moreover, the pair $(z, \tilde{\varepsilon})$ is a called a ε-saddle-point if $(z, 0, \tilde{\varepsilon})$ is a $(0, \varepsilon)$-saddle-point.

Before describing a special case of the HPE framework for solving the saddle-point problem, we introduce two more assumptions:

B.2) $\Psi(\cdot, y)$ and $-\Psi(x, \cdot)$ are proper closed convex functions for every $(x, y) \in X \times Y$;

B.3) the inclusion (29) has a solution, i.e., $T^{-1}(0) \neq \emptyset$.

A function $\Psi : \mathcal{X} \times \mathcal{Y} \to [-\infty, +\infty]$ satisfying conditions B.1 and B.2 for some nonempty convex sets X and Y is called a *closed convex-concave function on $X \times Y$*. It is well-known that its associated map T defined in (29) is maximal monotone (see for example Theorem 6.3.2 in [1]).

We are ready to state a special case of the HPE framework for solving the monotone inclusion problem (29), and hence the saddle-point problem $SP(\Psi; X, Y)$.

[SP-HPE] Hybrid proximal extragradient framework for solving $SP(\Psi; X, Y)$:

    0) Let $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$, $\lambda > 0$ and $0 \leq \sigma < 1$ be given and set $k = 1$;

    1) find $(\tilde{x}_k, \tilde{y}_k) \in \mathcal{X} \times \mathcal{Y}$, $\tilde{r}_k = (\tilde{r}_k^x, \tilde{r}_k^y) \in \mathcal{X} \times \mathcal{Y}$ and $\varepsilon_k \geq 0$ such that

$$(\tilde{r}_k^x, \tilde{r}_k^y) \in \partial_{\varepsilon_k}[\Psi(\cdot, \tilde{y}_k) - \Psi(\tilde{x}_k, \cdot)](\tilde{x}_k, \tilde{y}_k), \tag{35}$$
$$\|\lambda \tilde{r}_k^x + \tilde{x}_k - x_{k-1}\|_{\mathcal{X}}^2 + \|\lambda \tilde{r}_k^y + \tilde{y}_k - y_{k-1}\|_{\mathcal{Y}}^2 + 2\lambda\varepsilon_k \tag{36}$$
$$\leq \sigma^2 \left(\|\tilde{x}_k - x_{k-1}\|_{\mathcal{X}}^2 + \|\tilde{y}_k - y_{k-1}\|_{\mathcal{Y}}^2 \right);$$

    2) set $x_k = x_{k-1} - \lambda \tilde{r}_k^x$, $y_k = y_{k-1} - \lambda \tilde{r}_k^y$ and $k \leftarrow k + 1$, and go to step 1.

    end

We now make several remarks about the SP-HPE framework. First, due to Lemma 3.2 below, the SP-HPE framework is a special case of the HPE framework in which $\lambda_k := \lambda$. In fact, the SP-HPE framework could be stated in terms of a sequence of variable stepsizes $\{\lambda_k\}$, but we assume for simplicity $\lambda_k = \lambda$. Second, similar to the HPE framework, the SP-HPE framework does not specify how to find $(\tilde{x}_k, \tilde{y}_k)$, $\tilde{r}_k$ and ε_k satisfying the HPE error condition in (35) and (36). Section 5 describes a special instance of the SP-HPE framework in which $(\tilde{x}_k, \tilde{y}_k)$, $\tilde{r}_k$ and ε_k are obtained by a variant of Nesterov's accelerated method. Third, using the fact that the inclusion (35) is stronger than the inclusion in (21), we derive in Theorem 3.4 a finer version of Theorem 2.4 with $\lambda_k = \lambda$ specialized to the context of the saddle-point problem (28).

Before stating the pointwise and ergodic convergence rate results for the SP-HPE framework, we give two preliminary technical results.

Lemma 3.2. *For each $(x, y) \in X \times Y$ and $\varepsilon \geq 0$, we have*

$$\partial_\varepsilon(\Psi(\cdot, y) - \Psi(x, \cdot))(x, y) \subseteq T^\varepsilon(x, y),$$

where T is defined in (29).

Proof. Let $r \in \partial_\varepsilon(\Psi(\cdot, y) - \Psi(x, \cdot))(x, y)$ be given. This clearly implies that

$$\Psi(\tilde{x}, y) - \Psi(x, \tilde{y}) \geq \langle (\tilde{x} - x, \tilde{y} - y), r \rangle - \varepsilon \qquad \forall (\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y}.$$

On the other hand, it follows from the definition of T in (29) that any $\tilde{v} \in T(\tilde{x}, \tilde{y})$ satisfies

$$\Psi(x, \tilde{y}) - \Psi(\tilde{x}, y) \geq \langle (x - \tilde{x}, y - \tilde{y}), \tilde{v} \rangle.$$

Summing up the above two inequalities, we then conclude that

$$\langle (x - \tilde{x}, y - \tilde{y}), v - \tilde{v} \rangle \geq -\varepsilon \qquad \forall (\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y}, \; \forall \tilde{v} \in T(\tilde{x}, \tilde{y}),$$

and hence that $v \in T^\varepsilon(x, y)$ in view of the the definition of $T^\varepsilon(\cdot)$ in (9). $\qquad\square$

Lemma 3.3. *Let $X \subseteq \Re^n$ and $Y \subseteq \Re^m$ be given convex sets and let $\Gamma : X \times Y \to \Re$ be a function such that, for each pair $(x, y) \in X \times Y$, the function $\Gamma(\cdot, y) - \Gamma(x, \cdot) : X \times Y \to \Re$ is convex. Suppose that, for $i = 1, \ldots, k$, $(x_i, y_i) \in X \times Y$ and $(v_i, w_i) \in \Re^n \times \Re^m$ satisfies*

$$(v_i, w_i) \in \partial_{\varepsilon_i}\left(\Gamma(\cdot, y_i) - \Gamma(x_i, \cdot)\right)(x_i, y_i).$$

Let $\alpha_1, \ldots, \alpha_k \geq 0$ be such that $\sum_{i=1}^k \alpha_i = 1$ and define

$$(x^a, y^a) = \sum_{i=1}^k \alpha_i(x_i, y_i), \quad (v^a, w^a) = \sum_{i=1}^k \alpha_i(v_i, w_i),$$

$$\varepsilon^a := \sum_{i=1}^k \alpha_i[\varepsilon_i + \langle x_i - x^a, v_i \rangle + \langle y_i - y^a, w_i \rangle]$$

Then, $\varepsilon^a \geq 0$ and

$$(v^a, w^a) \in \partial_{\varepsilon^a}\left(\Gamma(\cdot, y^a) - \Gamma(x^a, \cdot)\right)(x^a, y^a). \tag{37}$$

11

The proof of Lemma 3.3 can be found in Proposition 5.1 of [8].

The following result describes the pointwise and ergodic convergence rate properties of the SP-HPE framework.

Theorem 3.4. *Consider the sequences* $\{(\tilde{x}_k, \tilde{y}_k)\}, \{(\tilde{r}_k^x, \tilde{r}_k^y)\}$ *and* $\{\varepsilon_k\}$ *generated by the SP-HPE framework and define for every* $k \in \mathbb{N}$,

$$(\tilde{x}_k^a, \tilde{y}_k^a) := \frac{1}{k}\sum_{i=1}^{k}(\tilde{x}_i, \tilde{y}_i), \quad \tilde{r}_k^a := \frac{1}{k}\sum_{i=1}^{k}(\tilde{r}_i^x, \tilde{r}_i^y) \tag{38}$$

and

$$\varepsilon_k^a := \frac{1}{k}\sum_{i=1}^{k}[\varepsilon_i + \langle(\tilde{x}_i - \tilde{x}_k^a, \tilde{y}_i - \tilde{y}_k^a), (\tilde{r}_i^x, \tilde{r}_i^y)\rangle]. \tag{39}$$

Let d_0 *denote the distance of* (x_0, y_0) *to the solution set of* $SP(\Psi; X, Y)$. *Then, for every* $k \in \mathbb{N}$, *the following statements hold:*

(a) *(pointwise convergence rate) the triple* $((\tilde{x}_k, \tilde{y}_k), \tilde{r}_k, \varepsilon_k)$ *is a* $(\|\tilde{r}_k\|, \varepsilon_k)$*-saddle-point of* Ψ, *or equivalently* (35) *holds, and there exists an index* $i \leq k$ *such that*

$$\|\tilde{r}_i\| \leq \frac{d_0}{\lambda}\sqrt{\frac{1+\sigma}{k(1-\sigma)}}, \quad \varepsilon_i \leq \frac{\sigma^2 d_0^2}{2k\lambda(1-\sigma^2)}; \tag{40}$$

(b) *(ergodic convergence rate) the triple* $((\tilde{x}_k^a, \tilde{y}_k^a), \tilde{r}_k^a, \varepsilon_k^a)$ *is a* $(\|\tilde{r}_k^a\|, \varepsilon_k^a)$*-saddle-point of* Ψ, *or equivalently*

$$\tilde{r}_k^a \in \partial_{\varepsilon_k^a}(\Psi(\cdot, \tilde{y}_k^a) - \Psi(\tilde{x}_k^a, \cdot))(\tilde{x}_k^a, \tilde{y}_k^a), \tag{41}$$

and

$$\|\tilde{r}_k^a\| \leq \frac{2d_0}{\lambda k}, \quad 0 \leq \varepsilon_k^a \leq \frac{2d_0^2}{\lambda k}\left(1 + \frac{\sigma}{\sqrt{(1-\sigma^2)}}\right). \tag{42}$$

Proof. The first claim in (a) is obvious. Since, by (35) and Lemma 3.2, we have $\tilde{r}_k \in T^{\varepsilon_k}(\tilde{x}_k, \tilde{y}_k)$ where T is defined in (29), we conclude that the SP-HPE framework is a special instance of the HPE framework applied to (29) where $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ is endowed with the inner product defined in (26). The second claim in (a) then follows Theorem 2.4(a). Moreover, inclusion (41) follows from (35) and Lemma 3.3, and the bounds in (42) follow from Theorem 2.4(b) with $\lambda_k = \lambda$. $\square$

4 Solving the HPE error condition

This section presents a scheme, together with its iteration-complexity analysis, for finding a solution of the HPE error condition (35)-(36) with Ψ given by (4) (and w.l.o.g. $\lambda = 1$). The scheme is based on the Nesterov's accelerated variant of Subsection 2.2 applied to an associated composite convex-concave min-max problem.

This section considers the following problem corresponding to the special case of step 1 of the SP-HPE framework in which $\lambda = 1$.

(P1) Given convex sets $X \subset \mathcal{X}$ and $Y \subset \mathcal{Y}$, a closed convex-concave function Ψ on $X \times Y$, a pair $(u_0, v_0) \in \mathcal{X} \times \mathcal{Y}$ and a scalar $\sigma > 0$, the problem is to find $(\tilde{u}, \tilde{v}) \in \mathcal{X} \times \mathcal{Y}, (\tilde{r}^u, \tilde{r}^v) \in \mathcal{X} \times \mathcal{Y}$ and $\tilde{\varepsilon} \geq 0$ such that

$$(\tilde{r}^u, \tilde{r}^v) \in \partial_{\tilde{\varepsilon}} \left[\Psi(\cdot, \tilde{v}) - \Psi(\tilde{u}, \cdot) \right] (\tilde{u}, \tilde{v}), \tag{43}$$

$$\|\tilde{r}^u + \tilde{u} - u_0\|_{\mathcal{X}}^2 + \|\tilde{r}^v + \tilde{v} - v_0\|_{\mathcal{Y}}^2 + 2\tilde{\varepsilon} \leq \sigma^2 \left(\|\tilde{u} - u_0\|_{\mathcal{X}}^2 + \|\tilde{v} - v_0\|_{\mathcal{Y}}^2 \right). \tag{44}$$

This section presents a scheme based on the Nesterov's accelerated variant of Subsection 2.2 for solving problem $(P1)$ where Ψ has the bilinear structure

$$\Psi(u, v) = f(u) + \langle Au, v \rangle + g_1(u) - g_2(v), \quad \forall (u, v) \in X \times Y \tag{45}$$

and the following conditions hold:

C.1) $A : \mathcal{X} \to \mathcal{Y}$ is a linear operator;

C.2) $g_1 : \mathcal{X} \to [-\infty, \infty]$ and $g_2 : \mathcal{Y} \to [-\infty, \infty]$ are proper closed convex functions such that $\operatorname{dom} g_1 = X$ and $\operatorname{dom} g_2 = Y$;

C.3) f is convex on a closed convex set $\Omega \supseteq X$;

C.4) f is differentiable on Ω and ∇f is L_f-Lipschitz continuous on Ω.

We now make two remarks about problem $(P1)$. First, finding the solution of the exact version of problem $(P1)$, i.e., the one in which $\sigma = 0$, is equivalent to finding the unique saddle-point of

$$\min_{u \in X} \max_{v \in Y} \Psi(u, v) + \frac{1}{2} \|u - u_0\|^2 - \frac{1}{2} \|v - v_0\|^2 \tag{46}$$

where Ψ is given by (45). More specifically, if $(\tilde{u}, \tilde{v})$ is the exact saddle-point of the above problem, then $(\tilde{u}, \tilde{v})$ and the quantities $(\tilde{r}^u, \tilde{r}^v) := (u_0 - \tilde{u}, v_0 - \tilde{v})$ and $\tilde{\varepsilon} := 0$ satisfy (43) and (44) with $\sigma = 0$. Second, although the above saddle-point problem has essentially the same structure as the one we are interested in solving, namely (4), its primal function (see (30)) has the key property that it is the composite sum of the easy convex nonsmooth function g_1 and a smooth convex function with Lipschitz continuous gradient. Hence, approximate solutions of (46) can be obtained by using a Nesterov's accelerated variant for composite convex optimization problems (e.g., the one in Subsection 2.2).

In view of the two observations above, it is reasonable to expect that approximate solutions of (46) yield solutions of problem $(P1)$ (with $\sigma > 0$). Rather than tackling the latter issue in an abstract setting, we instead propose a scheme based on the Nesterov's accelerated variant of Subsection 2.2 applied to (46) to obtain a solution of problem $(P1)$ and derive its corresponding iteration complexity.

We next discuss how the composite convex-concave min-max problem (46) can be viewed as a composite convex optimization problem (12) satisfying conditions A.1-A.3. Clearly, (46) is a special case of (12) in which

$$\psi(u) := f(u) + \tilde{\phi}(u), \qquad g(u) := g_1(u) + \frac{1}{2} \|u - u_0\|_{\mathcal{X}}^2 \tag{47}$$

13

and

$$\tilde{\phi}(u) := \max_v \left\{ \phi(u, v) := \langle Au, v \rangle - g_2(v) - \frac{1}{2}\|v - v_0\|_{\mathcal{Y}}^2 \right\}. \tag{48}$$

It is apparent that the above function g satisfies condition A.1. The following result implies that the above ψ satisfies conditions A.2 and A.3. Its proof for the case in which Y is compact is well-known (see for example [12]). Since we are not assuming that the latter condition, we include for sake of completeness a simple proof for the more general version given below. Its statement uses the following notion of the induced norm of a linear operator $A : \mathcal{X} \to \mathcal{Y}$ defined as

$$\|A\| := \max_y \{ \|Ax\|_{\mathcal{Y}} : \|x\|_{\mathcal{X}} \le 1 \}.$$

Proposition 4.1. *The following statements hold:*

(a) for every $u \in \mathcal{X}$, the maximization problem in (48) has a unique optimal solution $v(u)$, i.e.,

$$v(u) := \arg\max_v \langle Au, v \rangle - g_2(v) - \frac{1}{2}\|v - v_0\|_{\mathcal{Y}}^2; \tag{49}$$

(b) $\tilde{\phi}$ is convex, differentiable everywhere on $\mathcal{X}$, $\nabla\tilde{\phi}$ is $\|A\|^2$-Lipschitz continuous on $\mathcal{X}$ and

$$\nabla\tilde{\phi}(u) = A^* v(u) \quad \forall u \in \mathcal{X}; \tag{50}$$

(c) for every $u, \tilde{u} \in \mathcal{X}$,

$$l_{\tilde{\phi}}(u; \tilde{u}) = \phi(u, v(\tilde{u})). \tag{51}$$

Proof. (a) This statement follows immediately from the fact that the negative of the objective function of the max problem in (49) is proper, closed and strongly convex.

(b) Letting $\tilde{g}_2(v) := g_2(v) + \|v - v_0\|^2/2$ and using the definition of $\tilde{\phi}$ in (48), we easily see that

$$\tilde{\phi}(u) = \tilde{g}_2^*(Au) \quad \forall u \in \mathcal{X}. \tag{52}$$

Moreover, noting that $\tilde{g}_2$ is a proper closed strongly convex with modulus one, we conclude from Proposition 2.2 with $f = \tilde{g}_2$ that $\tilde{g}_2^*$ is differentiable everywhere on $\mathcal{Y}$ and $\nabla\tilde{g}_2^*$ is 1-Lipschitz continuous. The above two observations then easily imply that $\tilde{\phi}$ is convex, differentiable everywhere on $\mathcal{X}$ and $\nabla\tilde{\phi}$ is $\|A\|^2$-Lipschitz continuous on $\mathcal{X}$. Moreover, the optimality condition for (49) implies that $Au \in \partial\tilde{g}_2(v(u))$, and hence that $v(u) = \nabla\tilde{g}_2^*(Au)$ in view of Proposition 2.1(c). Now, (50) follows by differentiating (52) and using the latter conclusion.

(c) Using (50), and the definitions of $l_{\tilde{\phi}}(\cdot, \cdot)$, $\phi(\cdot, \cdot)$ and $v(u)$ in (10), (48) and (49), respectively, we easily see that

$$l_{\tilde{\phi}}(u; \tilde{u}) = \tilde{\phi}(\tilde{u}) + \langle \nabla\tilde{\phi}(\tilde{u}), u - \tilde{u} \rangle = \phi(\tilde{u}, v(\tilde{u})) + \langle A^* v(\tilde{u}), u - \tilde{u} \rangle = \phi(u, v(\tilde{u})).$$

$\square$

In view of the above result, we conclude that the function ψ defined in (47) satisfies conditions A.2 and A.3 of Subsection 2.2 with $L = L_f + \|A\|^2$. We can then use Algorithm 1 to approximately solve (46), and hence $(P1)$ as will be shown later in this section.

14

We now state our accelerated scheme for solving problem (P1). It is essentially Algorithm 1 applied to (12) with Ψ and g given by (47) and (48), respectively, endowed with two important refinements. The first one due to Nesterov (see (4.2) of [12] or Corollary 3(c) of [21]) computes a dual iterate $\tilde{v}_k$ as in (53), which together with the primal iterate $\tilde{u}_k$, provides the first candidate pair $(\tilde{u}, \tilde{v}) = (\tilde{u}_k, \tilde{v}_k)$ for (P1). The second one (see step 2 below) gives a recipe for computing the second candidate pair $(\tilde{r}^u, \tilde{r}^v) \in \mathcal{X} \times \mathcal{Y}$ and scalar $\tilde{\varepsilon} \geq 0$ which, together with the above pair $(\tilde{u}, \tilde{v})$, yield a candidate solution for (P1).

[Algorithm 2] Accelerated method for problem (P1):

Input: f, L_f, A, g_1 and g_2 as in conditions C.1-C.4, $(u_0, v_0) \in \mathcal{X} \times \mathcal{Y}$ and $\sigma \in (0, 1)$.

0) set $L = L_f + \|A\|^2$, $\Gamma_0 = 0$, $\tilde{u}_0 = w_0 = P_\Omega(u_0)$, $\tilde{v}_0 = 0$ and $k = 1$;

1) compute Γ_k, u_k and $v(u_k)$ as in (13), (14) and (49), respectively, $(\tilde{v}_k, w_k) \in Y \times X$ as

$$\tilde{v}_k := \frac{\Gamma_{k-1}}{\Gamma_k}\tilde{v}_{k-1} + \frac{\Gamma_k - \Gamma_{k-1}}{\Gamma_k}v(u_k), \tag{53}$$

$$w_k := \operatorname{argmin}_u l_{f,k}(u) + \langle A^*\tilde{v}_k, u\rangle + g_1(u) + \frac{c_k}{2}\|u - u_0\|_{\mathcal{X}}^2, \tag{54}$$

and $\tilde{u}_k$ as in (16), where

$$c_k := 1 + \frac{1}{\Gamma_k}, \qquad l_{f,k}(u) := \sum_{i=1}^{k} \frac{\Gamma_i - \Gamma_{i-1}}{\Gamma_k} l_f(u; u_i); \tag{55}$$

2) set

$$\tilde{\varepsilon}_k = \frac{1}{2\Gamma_k}\|\tilde{u}_k - u_0\|_{\mathcal{X}}^2, \quad \tilde{r}_k^u := c_k(u_0 - w_k), \quad \tilde{r}_k^v := v_0 - v(\tilde{u}_k); \tag{56}$$

3) if $\|\tilde{r}_k^u + \tilde{u}_k - u_0\|_{\mathcal{X}}^2 + \|\tilde{r}_k^v + \tilde{v}_k - v_0\|_{\mathcal{Y}}^2 + 2\tilde{\varepsilon}_k \leq \sigma^2\|\tilde{u}_k - u_0\|_{\mathcal{X}}^2 + \sigma^2\|\tilde{v}_k - v_0\|_{\mathcal{Y}}^2$, then terminate and go to **Output**; otherwise, set $k \leftarrow k + 1$ and go to step 1.

Output: ouput $(\tilde{u}, \tilde{v}) = (\tilde{u}_k, \tilde{v}_k)$, $(\tilde{r}^u, \tilde{r}^v) = (\tilde{r}_k^u, \tilde{r}_k^v)$ and $\tilde{\varepsilon} = \tilde{\varepsilon}_k$.

The following simple result shows that step 1 of Algorithm 2 corresponds to an iteration of Algorithm 1 applied to (15) with ψ and g defined according to (47) and (48).

Lemma 4.2. *Let ψ and g be defined according to (47) and (48). Then, the following statements hold for every $k \geq 1$:*

(a) *the function $l_{\psi,k}(u) - (l_{f,k}(u) + \langle A^*\tilde{v}_k, u\rangle)$ is constant;*

(b) *(54) is equivalent to (15).*

Proof. (a) Relation (53) and the fact that $\Gamma_0 = 0$ imply that

$$\tilde{v}_k = \sum_{i=1}^{k} \frac{\Gamma_i - \Gamma_{i-1}}{\Gamma_k}v(u_i). \tag{57}$$

15

Using the first identity in (47) and Proposition 4.1(b), we have that $\nabla\psi(u) = \nabla f(u) + A^*v(u)$, which together with definition (10) then imply that

$$l_\psi(u; u_i) = l_f(u; u_i) + [\tilde{\phi}(u_i) + \langle A^*v(u_i), u - u_i \rangle] \quad \forall i \geq 1.$$

Statement (a) now follows from the previous identity and relations (18), (55) and (57).

(b) This statement immediately follows from (a), the definition of g in (47) and the definition of c_k in (55). $\qquad \square$

Ignoring steps 2 and 3 of Algorithm 2 which are essentially computing $(\tilde{r}^u, \tilde{r}^v) = (\tilde{r}_k^u, \tilde{r}_k^v)$ and $\tilde{\varepsilon} = \tilde{\varepsilon}_k$ satisfying (43) and checking whether these entities, together with the primal-dual iterate $(\tilde{u}_k, \tilde{v}_k)$, satisfy (44), Lemma 4.2 immediately implies that Algorithm 2 is nothing more than Algorithm 1 applied to problem (12) with ψ and g given by (47).

The following technical result follows as a consequence of the latter observation and Proposition 2.3.

Lemma 4.3. *Consider the sequences $\{(\tilde{u}_k, \tilde{v}_k)\}$ generated by Algorithm 2 and define*

$$\tilde{\varepsilon}_k' := \frac{1}{2\Gamma_k}\|\tilde{u}_k - u_0\|_{\mathcal{X}}^2 + l_{f,k}(\tilde{u}_k) - f(\tilde{u}_k), \tag{58}$$

$$\Psi_k(u, v) := l_{f,k}(u) + \langle Au, v \rangle + g_1(u) - g_2(v), \tag{59}$$

$$q_k(u, v) := \frac{c_k}{2}\|u - u_0\|_{\mathcal{X}}^2 + \frac{1}{2}\|v - v_0\|_{\mathcal{Y}}^2, \tag{60}$$

where c_k and $l_{f,k}$ are defined in (55). Then,

$$0 \in \partial_{\tilde{\varepsilon}_k'} \left[\Psi_k(\cdot, \tilde{v}_k) - \Psi_k(\tilde{u}_k, \cdot) + q_k(\cdot, \cdot) \right] (\tilde{u}_k, \tilde{v}_k), \tag{61}$$

Proof. Consider the functions ψ, g and ϕ defined in (47) and (48). It follows from (47), (48) and Proposition 2.3 that

$$f(\tilde{u}_k) + \phi(\tilde{u}_k, v) + g_1(\tilde{u}_k) + \frac{1}{2}\|\tilde{u}_k - u_0\|_{\mathcal{X}}^2 \leq (\psi + g)(\tilde{u}_k)$$

$$\leq l_{\psi,k}(u) + g_1(u) + \frac{c_k}{2}\|u - u_0\|_{\mathcal{X}}^2 \quad \forall (u, v) \in \mathcal{X} \times \mathcal{Y}$$

where $l_{\psi,k}(\cdot)$ is defined in (18). Using the definitions of ψ and $\tilde{\phi}$ in (47) and (48), relation (10), definitions of $l_{\psi,k}(u)$ and $l_{f,k}(u)$ in (18) and (55), identities (51) and (57), and the fact $\phi(u, \cdot)$ is concave for any $u \in \mathcal{X}$, we conclude that

$$l_{\psi,k}(u) = \sum_{i=1}^{k} \frac{\Gamma_i - \Gamma_{i-1}}{\Gamma_k}\left(l_f(u; u_i) + l_{\tilde{\phi}}(u; u_i)\right) = l_{f,k}(u) + \sum_{i=1}^{k} \frac{\Gamma_i - \Gamma_{i-1}}{\Gamma_k}\phi\left(u, v(u_i)\right)$$

$$\leq l_{f,k}(u) + \phi\left(u, \sum_{i=1}^{k} \frac{\Gamma_i - \Gamma_{i-1}}{\Gamma_k}v(u_i)\right) = l_{f,k}(u) + \phi(u, \tilde{v}_k) \quad \forall u \in \mathcal{X}.$$

16

Combining the above two relations and using the definition of ϕ, Ψ_k and $\tilde{\varepsilon}_k'$ in (48), (59) and (58), respectively, we then conclude that

$$
\begin{aligned}
\Psi_k(\tilde{u}_k, v) - \frac{1}{2}\|v - v_0\|_{\mathcal{Y}}^2 &+ \frac{c_k}{2}\|\tilde{u}_k - u_0\|_{\mathcal{X}}^2 - \tilde{\varepsilon}_k' \\
&= l_{f,k}(\tilde{u}_k) + \phi(\tilde{u}_k, v) + g_1(\tilde{u}_k) + \frac{c_k}{2}\|\tilde{u}_k - u_0\|_{\mathcal{X}}^2 - \tilde{\varepsilon}_k' \\
&\leq l_{f,k}(u) + \phi(u, \tilde{v}_k) + g_1(u) + \frac{c_k}{2}\|u - u_0\|_{\mathcal{X}}^2 \\
&= \Psi_k(u, \tilde{v}_k) - \frac{1}{2}\|\tilde{v}_k - v_0\|_{\mathcal{Y}}^2 + \frac{c_k}{2}\|u - u_0\|_{\mathcal{X}}^2 \quad \forall (u, v) \in X \times Y. \quad (62)
\end{aligned}
$$

Now, using the definition of the ε-differential in (11) and the definition of $q_k(\cdot, \cdot)$ in (60), the above inequality can be easily seen to be equivalent to (61).

$\square$

The following result quantifies the quality of the entities $(\tilde{u}_k, \tilde{v}_k)$, $\tilde{\varepsilon}_k$ and $(\tilde{r}_k, \tilde{r}_k^v)$ generated at the k-th iteration of Algorithm 2 as a candidate solution for problem $(P1)$.

Lemma 4.4. *Consider the sequences $\{(\tilde{u}_k, \tilde{v}_k)\}$, $\{\tilde{\varepsilon}_k\}$ and $\{(\tilde{r}_k, \tilde{r}_k^v)\}$ generated by Algorithm 2. Then, for every $k \geq 1$,*

$$
(\tilde{r}_k^u, \tilde{r}_k^v) \in \partial_{\tilde{\varepsilon}_k}\left[\Psi(\cdot, \tilde{v}_k) - \Psi(\tilde{u}_k, \cdot)\right](\tilde{u}_k, \tilde{v}_k), \tag{63}
$$

$$
\|\tilde{r}_k^u + \tilde{u}_k - u_0\|_{\mathcal{X}}^2 + \|\tilde{r}_k^v + \tilde{v}_k - v_0\|_{\mathcal{Y}}^2 + 2\tilde{\varepsilon}_k \leq \left(\frac{3}{\Gamma_k} + \frac{4}{\Gamma_k^2}\right)\|\tilde{u}_k - u_0\|_{\mathcal{X}}^2, \tag{64}
$$

where $\Psi(\cdot)$ is defined in (45).

Proof. Equations (49) and (54) and the definitions of Ψ_k and q_k in (59) and (60) imply that

$$
(w_k, v(\tilde{u}_k)) = \arg\min_{(u,v)} \Psi_k(u, \tilde{v}_k) - \Psi_k(\tilde{u}_k, v) + q_k(u, v).
$$

In view of the optimality condition of the above minimization problem, the definitions of $\tilde{r}_k^u$ and $\tilde{r}_k^v$ in (56), and the definition of $q_k(\cdot, \cdot)$ in (60), we then conclude that

$$
(\tilde{r}_k^u, \tilde{r}_k^v) = -\nabla q_k(w_k, v(\tilde{u}_k)) \in \partial[\Psi_k(\cdot, \tilde{v}_k) - \Psi_k(\tilde{u}_k, \cdot)](w_k, v(\tilde{u}_k)).
$$

Hence, by Proposition 2.1(b) we have

$$
(\tilde{r}_k^u, \tilde{r}_k^v) = -\nabla q_k(w_k, v(\tilde{u}_k)) \in \partial_{\delta_k}[\Psi_k(\cdot, \tilde{v}_k) - \Psi_k(\tilde{u}_k, \cdot)](\tilde{u}_k, \tilde{v}_k). \tag{65}
$$

where

$$
\delta_k := -\left[\Psi_k(w_k, \tilde{v}_k) - \Psi_k(\tilde{u}_k, v(\tilde{u}_k))\right] - \langle -\nabla q_k(w_k, v(\tilde{u}_k)), (\tilde{u}_k, \tilde{v}_k) - (w_k, v(\tilde{u}_k))\rangle \geq 0.
$$

On the other hand, in view of Lemma 4.3, inclusion (61) holds, or equivalently inequality (62) holds. The latter inequality with $(u, v) = (w_k, v(\tilde{u}_k))$, together with the definitions of $\tilde{\varepsilon}_k$ and $\tilde{\varepsilon}_k'$ in (56) and (58), then implies that

$$
\begin{aligned}
\tilde{\varepsilon}_k \geq \tilde{\varepsilon}_k' &\geq -\Psi_k(w_k, \tilde{v}_k) + \Psi_k(\tilde{u}_k, v(\tilde{u}_k)) + q_k(\tilde{u}_k, \tilde{v}_k) - q_k(w_k, v(\tilde{u}_k)) \\
&= \delta_k + \frac{c_k}{2}\|\tilde{u}_k - w_k\|_{\mathcal{X}}^2 + \frac{1}{2}\|\tilde{v}_k - v(\tilde{u}_k)\|_{\mathcal{Y}}^2, \tag{66}
\end{aligned}
$$

17

where the last equality comes from the definition of δ_k and the fact that the second order Taylor expansion of the quadratic function q_k at an arbitrary point agrees with q_k itself. In view of (65) and (66), we then conclude that

$$(\tilde{r}_k^u, \tilde{r}_k^v) \in \partial_{\tilde{\varepsilon}_k'}[\Psi_k(\cdot, \tilde{v}_k) - \Psi_k(\tilde{u}_k, \cdot)](\tilde{u}_k, \tilde{v}_k).$$

Using the definition of ε-subdifferential in (11) and the definitions of $\tilde{\varepsilon}_k$, $\tilde{\varepsilon}_k'$, Ψ and Ψ_k in (56), (58), (45) and (59), respectively, and the fact that $\Psi_k(\cdot, \tilde{v}_k)$ is majorized $\Psi(\cdot, \tilde{v}_k)$, it is now easy to see that the above inclusion implies (63).

Moreover, inequality (66), the definitions of $\tilde{r}_k^u$, $\tilde{r}_k^v$ and $\tilde{\varepsilon}_k$ in (56), and the fact $c_k = 1 + 1/\Gamma_k$ imply that

$$
\begin{aligned}
&\|\tilde{r}_k^u + \tilde{u}_k - u_0\|_{\mathcal{X}}^2 + \|\tilde{r}_k^v + \tilde{v}_k - v_0\|_{\mathcal{Y}}^2 + 2\tilde{\varepsilon}_k \\
=& \|(u_0 - \tilde{u}_k)/\Gamma_k + c_k(\tilde{u}_k - w_k)\|_{\mathcal{X}}^2 + \|\tilde{v}_k - v(\tilde{u}_k)\|_{\mathcal{X}}^2 + 2\tilde{\varepsilon}_k \\
\leq& \frac{2}{\Gamma_k^2}\|\tilde{u}_k - u_0\|_{\mathcal{X}}^2 + 2c_k^2\|\tilde{u}_k - w_k\|_{\mathcal{X}}^2 + \|\tilde{v}_k - v(\tilde{u}_k)\|_{\mathcal{X}}^2 + 2\tilde{\varepsilon}_k \\
\leq& \frac{2}{\Gamma_k^2}\|\tilde{u}_k - u_0\|_{\mathcal{X}}^2 + (4c_k + 2)\tilde{\varepsilon}_k = \left(\frac{3}{\Gamma_k} + \frac{4}{\Gamma_k^2}\right)\|\tilde{u}_k - u_0\|_{\mathcal{X}}^2.
\end{aligned}
$$

$\square$

As an immediate consequence of Lemma 4.4, we can now derive the iteration-complexity for Algorithm 2 to solve problem ($P1$).

Proposition 4.5. *Algorithm 2 terminates in at most*

$$\mathcal{O}\left(\left\lceil\sqrt{(L_f + \|A\|^2)\lceil\sigma^{-2}\rceil}\right\rceil\right) \tag{67}$$

iterations with an output which solves problem ($P1$).

Proof. The inclusion (63) and the termination criterion in step 3 of Algorithm 2 show that the output of Algorithm 2 solves problem ($P1$). To show the corollary, it suffices to show that Algorithm 2 finishes in at most

$$k_0 := \left\lceil 4\sqrt{(L_f + \|A\|^2)\lceil\sigma^{-2}\rceil}\right\rceil = \left\lceil 4\sqrt{L\lceil\sigma^{-2}\rceil}\right\rceil. \tag{68}$$

iterations, where the second equality is due to the definition of L in step 0 of Algorithm 2. Indeed, assume for contradiction that Algorithm 2 has not terminated at an iteration $k > k_0$. The latter condition on k together with (17) and (68) can be easily seen to imply that

$$\Gamma_k > 4\max\{1, \sigma^{-2}\}.$$

Moreover, since Algorithm 3 has not terminated at the k-th iteration of Algorithm 2, it follows from the termination criterion on its step 3 and relation (64) that

$$\sigma^2\|\tilde{u}_k - u_0\|_{\mathcal{X}}^2 < \|\tilde{r}_k^u + \tilde{u}_k - u_0\|_{\mathcal{X}}^2 + \|\tilde{r}_k^v + \tilde{v}_k - v_0\|_{\mathcal{Y}}^2 + 2\tilde{\varepsilon}_k \leq \left(\frac{3}{\Gamma_k} + \frac{4}{\Gamma_k^2}\right)\|\tilde{u}_k - u_0\|_{\mathcal{X}}^2.$$

Hence, it follows from the above two conclusions that

$$\sigma^2 < \left(\frac{3}{\Gamma_k} + \frac{4}{\Gamma_k^2}\right) < \frac{4}{\Gamma_k} < \sigma^2.$$

$\square$

5   Accelerated SP-HPE method for problem (4)

This section presents a special instance of the SP-HPE framework introduced in Section 3, which we refer to as the Acc-SP-HPE method, for solving the class of composite convex-concave min-max problem (4), or equivalently, the saddle-point problem $SP(\Psi; X, Y)$ with Ψ defined in (45). Each (outer) iteration of the Acc-SP-HPE method, which is essentially a special iteration of the SP-HPE framework, invokes Algorithm 2 to obtain a solution of the inexact prox subproblem (35)-(36). A complexity bound on the total number of iterations performed by Algorithm 2 (called the inner iterations) performed by the Acc-SP-HPE method to find a (ρ, ε)-saddle-point is derived in section. Moreover, an inner-iteration complexity for the Acc-SP-HPE method to find an ε-saddle-point is also derived for the case when the feasible set $X \times Y$ is bounded.

We assume in this section that the solution set of the composite convex-concave min-max problem (4) is nonempty, and assumptions C.1-C.4 are satisfied. Initially, we do not assume boundedness of the feasible set $X \times Y$. The case where $X \times Y$ is assumed to be bounded will be discussed in Subsection 5.1.

Recall that in Section 4 we have motivated the introduction of problem $(P1)$ as a special case of the inexact prox subproblem (35)-(36) in which $\lambda = 1$. The following result shows in fact that problem $(P1)$ is as general as subproblem (35)-(36) for any value of $\lambda > 0$.

Proposition 5.1. *Let* $\lambda > 0$ *and* Ψ *be a closed convex-concave function and consider the k-th iteration of the SP-HPE framework. If* $(\tilde{u}, \tilde{v}) \in \mathcal{X} \times \mathcal{Y}$, $(\tilde{r}^u, \tilde{r}^v) \in \mathcal{X} \times \mathcal{Y}$ *and* $\tilde{\varepsilon} \geq 0$ *solve problem* $(P1)$ *with input* $\Psi = \lambda \Psi$, $(u_0, v_0) = (x_{k-1}, y_{k-1})$ *and* $\sigma > 0$, *then*

$$(\tilde{x}_k, \tilde{y}_k) := (\tilde{u}, \tilde{v}), \quad (\tilde{r}_k^x, \tilde{r}_k^y) := \frac{1}{\lambda}(\tilde{r}^u, \tilde{r}^v), \quad \varepsilon_k := \frac{1}{\lambda}\tilde{\varepsilon}$$

satisfy the conditions (35) *and* (36) *of step 1 of the SP-HPE framework.*

Proof. The conclusion follows immediately from the identity

$$\lambda \partial_\varepsilon \left[\Psi(\cdot, \tilde{v}) - \Psi(\tilde{u}, \cdot) \right] (\tilde{u}, \tilde{v}) = \partial_{\lambda \varepsilon} \left[\lambda \Psi(\cdot, \tilde{v}) - \lambda \Psi(\tilde{u}, \cdot) \right] (\tilde{u}, \tilde{v})$$

which holds for every $\varepsilon \geq 0$, $\lambda \geq 0$ and $(\tilde{u}, \tilde{v}) \in \mathcal{X} \times \mathcal{Y}$. $\square$

In view of the above result, we can use Algorithm 2 to solve the inexact prox subproblem (35)-(36). This is the key idea behind the following special case of the SP-HPE framework, referred to as the Acc-SP-HPE method, for solving the saddle-point problem (4).

[Acc-SP-HPE] Accelerated SP-HPE Method for solving problem (4):

0) Let $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$, $\lambda > 0$ and $0 < \sigma < 1$ be given and set $k = 1$;

1) invoke Algorithm 2 with input

$$f = \lambda f, \quad A = \lambda A, \quad g_1 = \lambda g_1, \quad g_2 = \lambda g_2, \quad (u_0, v_0) = (x_{k-1}, y_{k-1}), \quad L_f = \lambda L_f,$$

and set

$$(\tilde{x}_k, \tilde{y}_k) := (\tilde{u}, \tilde{v}), \quad \tilde{r}_k = (\tilde{r}_k^x, \tilde{r}_k^y) := \frac{1}{\lambda}(\tilde{r}^u, \tilde{r}^v), \quad \varepsilon_k := \frac{1}{\lambda}\tilde{\varepsilon}$$

where $(\tilde{u}, \tilde{v})$, $(\tilde{r}^u, \tilde{r}^v)$ and $\tilde{\varepsilon}$ are the output generated by Algorithm 2;

2) set $x_k = x_{k-1} - \lambda \tilde{r}_k^x$, $\quad y_k = y_{k-1} - \lambda \tilde{r}_k^y$, set $k \leftarrow k + 1$, and go to step 1.

end

Proposition 5.2. *Acc-SP-HPE method is a special case of the SP-HPE framework for solving the composite convex-concave min-max problem* (4).

Proof. In view of Proposition 5.1, the sequences $\{(\tilde{x}_k, \tilde{y}_k)\}$, $\{(\tilde{r}_k^x, \tilde{r}_k^y)\}$ and $\{\varepsilon_k\}$ generated by the Acc-SP-HPE method satisfy the conditions (35) and (36) of step 1 of the SP-HPE framework. Therefore, Acc-SP-HPE method is clearly a special case of the SP-HPE framework. $\square$

It follows as a consequence of Proposition 5.2 that the pointwise and ergodic (outer) convergence rate bounds for the Acc-SP-HPE method are as described in statements (a) and (b) of Theorem 3.4, respectively.

Theorem 5.3. *Assume that conditions C.1-C.4 hold, $\max\{\sigma^{-1}, (1-\sigma)^{-1}\} = \mathcal{O}(1)$ and the (convex) set of saddle-points of* (4) *is non-empty, and let d_0 denote the distance of the initial iterate (x_0, y_0) of the Acc-SP-HPE method with respect to this set. Consider the sequences $\{(\tilde{x}_k, \tilde{y}_k)\}$, $\{(\tilde{r}_k^x, \tilde{r}_k^y)\}$ and $\{\varepsilon_k\}$ generated by the Acc-SP-HPE method and the ergodic sequences $\{(\tilde{x}_k^a, \tilde{y}_k^a)\}$, $\{\tilde{r}_k^a\}$ and $\{\varepsilon_k^a\}$ defined in Theorem 3.4. Then, the following statements hold:*

(a) *for every pair of positive scalars (ρ, ε), there exists*

$$k_0 = \mathcal{O}\left(\max\left\{1, \frac{d_0}{\lambda\rho}, \frac{d_0^2}{\lambda\varepsilon}\right\}\right)$$

such that for every $k \geq k_0$, the triple $((\tilde{x}_k^a, \tilde{y}_k^a), \tilde{r}_k^a, \varepsilon_k^a)$ is a (ρ, ε)-saddle-point of (4);

(b) *each iteration of the Acc-SP-HPE method performs at most*

$$\mathcal{O}\left(\left\lceil \sqrt{\lambda L_f + \lambda^2 \|A\|^2} \right\rceil\right)$$

inner iterations (and hence resolvent evaluations of ∂g_1 and ∂g_2).

As a consequence, the Acc-SP-HPE method finds a (ρ, ε)-saddle-point of (4) *by performing no more than*

$$\mathcal{O}\left(\left\lceil \sqrt{(\lambda L_f + \lambda^2 \|A\|^2)} \right\rceil \max\left\{1, \frac{d_0}{\lambda\rho}, \frac{d_0^2}{\lambda\varepsilon}\right\}\right) \tag{69}$$

inner iterations (and hence resolvent evaluations of ∂g_1 and ∂g_2).

Proof. Since by Proposition 5.2 the Acc-SP-HPE method is a special instance of the SP-HPE framework, (a) follows immediately from Theorem 3.4(b). Statement (b) follows immediately from Proposition 4.5 with $L_f = \lambda L_f$ and $A = \lambda A$, and the fact that each iteration of Algorithm 2 performs one resolvent evaluation of ∂g_1 and two resolvent evaluations of ∂g_2. The last assertion of the theorem follows immediately from (a) and (b). □

We now make some remarks about possible values of λ which minimize the complexity bound (69) (up to an additive and multiplicative $\mathcal{O}(1)$ constant). Noting that (69) is equivalent to

$$\mathcal{O}\left(\max\left\{\frac{1}{\lambda}, \sqrt{\frac{L_f}{\lambda}}, \|A\|\right\} \max\left\{\lambda, \frac{d_0}{\rho}, \frac{d_0^2}{\varepsilon}\right\}\right)$$

then it is straightforward to see that the following claims hold depending on whether the condition

$$\lambda_1 := \max\left\{\frac{L_f}{\|A\|^2}, \frac{1}{\|A\|}\right\} \leq \max\left\{\frac{d_0}{\rho}, \frac{d_0^2}{\varepsilon}\right\} =: \lambda_2 \tag{70}$$

holds (case 1) or not (case 2):

1) if (70) holds then any $\lambda \in [\lambda_1, \lambda_2]$ minimizes (69) with minimum value equal to

$$\mathcal{O}\left(\|A\| \max\left\{\frac{d_0}{\rho}, \frac{d_0^2}{\varepsilon}\right\}\right);$$

2) otherwise, if $\lambda_1 > \lambda_2$, then $\lambda = \lambda_2$ minimizes (69) with minimum value equal to

$$\mathcal{O}\left(1 + \sqrt{L_f} \max\left\{\sqrt{\frac{d_0}{\rho}}, \frac{d_0}{\sqrt{\varepsilon}}\right\}\right).$$

Ideally, one should choose λ according to the above discussion in order to minimize the total number of resolvent evaluations of ∂g_1 and ∂g_2. But, since d_0 is usually not known a priori, we can not compute λ_2, and as a result choose $\lambda = \lambda_2$ as proposed in case 2 above. Note however that we can always choose $\lambda = \lambda_1$ since the latter is easily computable. Clearly, this choice is optimal when case 1 holds and, even though is not optimal when case 2 holds, we believe it might be a good practical choice in both cases due to the fact that case 2 is quite unlikely.

5.1 Scaling of Acc-SP-HPE method for bounded min-max problem

In this subsection, we consider the special case of problem (4) where the feasible set $X \times Y$ is bounded and derive a complexity bound on the number of inner iterations performed by the Acc-SP-HPE method to find an ε-saddle-point.

Corollary 5.4. *Suppose that the assumptions of Theorem 5.3 hold, $(x_0, y_0) \in X \times Y$ and the diameter D of the set $X \times Y$ defined in (8) is finite. Then, for any $\varepsilon > 0$, Acc-SP-HPE method finds an ε-saddle-point of (4) by performing no more than*

$$\mathcal{O}\left(\left\lceil\sqrt{(\lambda L_f + \lambda^2\|A\|^2)}\right\rceil \max\left\{1, \frac{d_0 D}{\lambda\varepsilon}\right\}\right) \leq \mathcal{O}\left(\left\lceil\sqrt{(\lambda L_f + \lambda^2\|A\|^2)}\right\rceil \max\left\{1, \frac{D^2}{\lambda\varepsilon}\right\}\right) \tag{71}$$

resolvent evaluations of ∂g_1 and ∂g_2.

Proof. Under the assumption that D is finite, it is straightforward to see from Definition 3.1 and the definition of subdifferential that a $(\varepsilon/2D, \varepsilon/2)$-saddle-point is always a ε-saddle-point. The first bound in (71) now follows immediately from the fact that $d_0 \leq D$ in view of the assumption that $(x_0, y_0) \in X \times Y$, and from the bound (69) in Theorem 5.3 with $(\rho, \varepsilon) = (\varepsilon/(2D), \varepsilon/2)$. Clearly, $d_0 \leq D$ also implies the second bound in (71). $\qquad\square$

We now make a few comments about choosing λ so as to minimize the right hand side of (71) (up to an additive and multiplicative $\mathcal{O}(1)$ constant). Similar to the discussion in the previous subsection, if

$$\widehat{\lambda}_1 := \max\left\{\frac{L_f}{\|A\|^2}, \frac{1}{\|A\|}\right\} \leq \frac{D^2}{\varepsilon} =: \widehat{\lambda}_2 \tag{72}$$

holds, then any $\lambda \in [\widehat{\lambda}_1, \widehat{\lambda}_2]$ minimizes the right hand side of (71) with minimum value equal to $\mathcal{O}(1 + D^2\|A\|/\varepsilon)$. Otherwise, if $\widehat{\lambda}_1 > \widehat{\lambda}_2$, then $\lambda = \widehat{\lambda}_2$ minimizes the right hand side of (71) with minimum value equal to $\mathcal{O}(1 + D\sqrt{L_f/\varepsilon})$. Observe that regardless of which case holds, the right hand side of (71) assume its minimum value

$$\mathcal{O}\left(1 + D^2\frac{\|A\|}{\varepsilon} + D\sqrt{\frac{L_f}{\varepsilon}}\right) \tag{73}$$

when $\lambda = \min\left\{\widehat{\lambda}_1, \widehat{\lambda}_2\right\}$.

Clearly, letting D_X and D_Y denote the diameter of X and Y, we have $D = (D_X^2 + D_Y^2)^{1/2}$. Hence, we have $D_X \leq D$ and $D_X D_Y \leq D^2/2$, and it is clearly possible that $D_X << D$ and/or $D_X D_Y << D^2/2$. The rest of this subsection shows that the Acc-SP-HPE method applied to problem (4) with $\mathcal{X}$ and $\mathcal{Y}$ endowed with suitable scaled inner products has a resolvent complexity similar to (73) but with D^2 in the first term replaced by $D_X D_Y$ and D in the second term replaced by D_X.

To achieve the above goal, we endow $\mathcal{X}$ and $\mathcal{Y}$ with new inner products

$$\langle \cdot, \cdot \rangle_{\mathcal{X},\theta} := \theta\langle \cdot, \cdot \rangle, \quad \langle \cdot, \cdot \rangle_{\mathcal{Y},\theta} := \theta^{-1}\langle \cdot, \cdot \rangle, \tag{74}$$

respectively, and the associated norms then become

$$\| \cdot \|_{\mathcal{X},\theta} := \theta^{1/2}\| \cdot \|_{\mathcal{X}}, \quad \| \cdot \|_{\mathcal{Y},\theta} := \theta^{-1/2}\| \cdot \|_{\mathcal{Y}}$$

and problem (4) becomes

$$\min_{x \in X} \max_{y \in Y} \Psi(x, y) = f(x) + \langle A_\theta x, y \rangle_{\mathcal{Y},\theta} + g_1(x) - g_2(y), \tag{75}$$

where $A_\theta := \theta A$. Moreover, $\|A_\theta\|_\theta = \|A\|$ where $\|C\|_\theta := \max_y\{\|Cx\|_{\mathcal{Y},\theta} : \|x\|_{\mathcal{X},\theta} \leq 1\}$ and the gradient of f with respect to $\langle \cdot, \cdot \rangle_{\mathcal{X},\theta}$ is $L_{f,\theta}$-Lipschitz continuous on Ω where $L_{f,\theta} = \theta^{-1}L_f$. Also, the diameter of the feasible set $X \times Y$ with the product space $\mathcal{X} \times \mathcal{Y}$ endowed with the Cartesian inner product $\langle \cdot, \cdot \rangle_{\mathcal{X},\theta} + \langle \cdot, \cdot \rangle_{\mathcal{Y},\theta}$ is

$$D_\theta^2 := \theta D_X^2 + \theta^{-1}D_Y^2.$$

Using the above observations, we immediately see that the Acc-SP-HPE method applied to problem (4) where θ and λ are chosen as

$$\theta = \frac{D_Y}{D_X}, \quad \lambda = \min\left\{\max\left\{\frac{L_f D_X}{\|A\|^2 D_Y}, \frac{1}{\|A\|}\right\}, \frac{2D_X D_Y}{\varepsilon}\right\},$$

and $\mathcal{X}$ and $\mathcal{Y}$ are endowed with the inner products (74), computes an ε-saddle-point of (4) by performing no more than

$$\mathcal{O}\left(1 + \frac{\|A\|}{\varepsilon}D_X D_Y + \sqrt{\frac{L_f}{\varepsilon}}D_X\right) \tag{76}$$

resolvent evaluations of ∂g_1 and ∂g_2.

It is worth noting that the above complexity is the same as the complexity of Nesterov's smoothing method (see (4.4) in [12]).

6   Numerical experiments

This section presents computational results showing the good performance of the Acc-SP-HPE method on a collection of convex optimization problems that are either in the form of or can be easily reformulated as (4). All the computational results were obtained using MATLAB R2013a on a quad-core Linux machine with 8GB memory.

Acc-SP-HPE method method is compared with the following two other methods: i) Nesterov's smooth approximation scheme [12] (referred to Nest-App) where the smooth approximation is solved by the Nesterov's accelerated variant introduced in Subsection 2.2, and ii) the accelerated primal-dual (referred to APD) method proposed in [5]. For the sake of a fair comparison, we have implemented the latter two methods with $\mathcal{X}$ and $\mathcal{Y}$ endowed with the Euclidean (or Frobenius) norm $\|\cdot\|_2$ and based on the distance generating function for $\|\cdot\|_2^2/2$.

The following three subsections report computational results on the following classes of convex optimization problems: a) zero-sum matrix games; b) quadratic game problem, and; c) vector-matrix saddle-point problem.

6.1   Zero-sum matrix game

This subsection compares Acc-SP-HPE with Nest-App and APD on a collection of instances of the zero-sum matrix game problem

$$\min_{x \in \Delta_m} \max_{y \in \Delta_n} \Psi(x, y) = \langle Ax, y \rangle. \tag{77}$$

where $A \in \Re^{n \times m}$.

In the numerical experiment, the matrix A in problem (77) is generated such that each entry is nonzero with probability p and each nonzero entry is generated independently and uniformly in the interval $[-1, 1]$. The methods are terminated whenever the duality gap at the iterate $(\tilde{x}_k, \tilde{y}_k)$ is less than a given tolerance ϵ, i.e.,

$$\max_i (A\tilde{x}_k)_i - \min_i (A^\top \tilde{y}_k)_i \leq \epsilon. \tag{78}$$

Table 1 reports the CPU time and the number of (inner) iterations for each method (Acc-SP-HPE). Table 1 shows that the method Acc-SP-HPE has the best performance on all instances of the zero-sum game problem.

Problem size			Acc-SP-HPE		Nest-App		APD	
m	n	p	time	#iter.	time	#iter.	time	#iter.
1000	100	0.01	**0.92**	376	6.25	5087	1.72	3368
1000	100	0.1	**0.90**	283	7.70	6751	3.63	7509
1000	1000	0.01	**1.66**	66	27.92	2278	18.15	3048
1000	1000	0.1	**3.33**	133	44.64	3573	33.88	5466
1000	10000	0.01	**15.48**	63	219.88	1987	130.35	2691
1000	10000	0.1	**35.10**	145	335.85	3057	420.42	8645
10000	100	0.01	**7.66**	274	64.20	4725	33.61	4772
10000	100	0.1	**22.00**	765	150.16	10700	125.67	17204
10000	1000	0.01	**15.41**	64	240.81	2170	133.00	2765
10000	1000	0.1	**38.57**	159	511.09	4649	420.21	8815
10000	10000	0.01	**42.58**	18	1302.7	1221	989.41	2148
10000	10000	0.001	**6.81**	3	337.61	320	235.59	509

Table 1: Computational results for the methods Acc-SP-HPE, Nest-App and APD on two-player zero-sum games with different sizes and sparsities. All methods are terminated using criterion (78) with $\epsilon = 10^{-3}$. CPU time in seconds and number of (inner) iterations are reported for each method.

6.2 Quadratic game problem

This subsection compares Acc-SP-HPE with Nest-App and APD for solving a collection of instances of the quadratic game problem

$$\min_{x \in \Delta_m} \max_{y \in \Delta_n} \frac{1}{2} \|Bx\|^2 + y^\top A x \qquad (79)$$

where $A \in \Re^{n \times m}$ and $B \in \Re^{m \times m}$. In our numerical experiments, the matrices A and B were randomly generated such that each component is nonzero with probability p and each nonzero component is generated independently and uniformly in the interval $[-1, 1]$.

Observe that computing the duality gap of the problem (79) requires solving a quadratic programming over the unit simplex, and hence its use as a termination criterion for the three benchmarked algorithms is not suitable. We instead use the notion of (ρ, ε)-saddle-points (see Definition 3.1) as termination criteria for these algorithms. Similar to the Acc-SP-HPE method, we can also show that Nest-App naturally generates a triple $(\tilde{z}_k, \tilde{r}_k, \varepsilon_k)$ at every iteration which can be checked for being a (ρ, ε)-saddle-point. For the sake of shortness, we leave out the details of this construction. Moreover, it is shown in Lemma 4.3 of [5] that APD generates a triple $(\tilde{z}_k, \tilde{r}_k, \varepsilon_k)$ at every iteration which can be checked for being a (ρ, ε)-saddle-point. For simplicity we set $\rho = \varepsilon$ and the three methods are terminated whenever

$$\max\{\|\tilde{r}_k\|, \varepsilon_k\} < \epsilon \qquad (80)$$

is satisfied. Table 2 reports the CPU time and the number of (inner) iterations for each method.

6.3 Vector-matrix saddle-point problem

This subsection compares Acc-SP-HPE with Nest-App and APD for solving a collection of instances of the minimization problem

$$\min_{x \in \Delta_m} \frac{1}{2} \|Cx - b\|^2 + \theta_{max}(A(x)), \qquad (81)$$

Problem size			Lip. const.		Acc-SP-HPE		Nest-App		APD	
m	n	p	$\|B\|^2$	$\|A\|$	time	#iter.	time	#iter.	time	#iter.
100	100	0.01	2.21	1.61	**0.39**	1010	4.41	12406	5.43	18669
100	1000	0.01	2.69	2.61	**15.45**	10077	47.17	39926	24.19	52391
1000	100	0.01	16.11	2.91	**9.1**	2877	39.70	13644	101.66	48650
1000	1000	0.01	16.04	3.99	**13.17**	519	290.05	14335	174.36	30474
100	100	0.1	13.86	3.50	**0.14**	361	7.23	21248	14.83	50520
100	1000	0.1	15.06	7.53	**0.86**	616	46.16	38736	40.95	89841
1000	100	0.1	131.85	7.57	**1.58**	526	83.77	30669	124.30	61229
1000	1000	0.1	133.87	11.59	**17.49**	678	593.56	39552	532.13	95249
100	100	0.2	28.47	5.24	**0.22**	524	11.26	33669	21.02	72126
100	1000	0.2	25.78	10.64	**1.93**	1154	69.28	54343	61.86	133663
1000	100	0.2	267.38	10.59	**2.21**	708	110.39	39761	194.20	93167
1000	1000	0.2	263.79	16.35	**21.43**	838	655.70	44198	797.53	139323

Table 2: Computational results for the methods Acc-SP-HPE, Nest-App and APD on two-player quadratic games with different sizes and sparsities. All methods are terminated using criterion (80) with $\epsilon = 10^{-4}$. CPU time in seconds and number of (inner) iterations are reported for each method.

where $C \in \Re^{m \times m}$, $b \in \Re^m$, $A_1, \cdots, A_m \in \mathcal{S}^n$ and $A(x) = \sum_{i=1}^m x_i A_i \in \mathcal{S}^{n \times n}$. It is easy to verify that the above problem is equivalent to the vector-matrix saddle-point problem

$$\min_{x \in \Delta_m} \max_{y \in \Omega} \ \Psi(x,y) = \frac{1}{2}\|Cx - b\|^2 + \langle A(x), y \rangle \tag{82}$$

where $\Omega = \{y \in \mathcal{S}^n : tr(y) = 1, y \succeq 0\}$. Hence, we can apply the above methods on the saddle-point problem (82). In our numerical experiments, the matrices $A_1, \cdots, A_m$ and C were randomly generated such that each component is nonzero with probability 0.1 and each nonzero component is generated independently and uniformly in the interval $[-1, 1]$ and $A_1, \cdots, A_m$ are then symmetrized. As discussed in the previous subsection, all three methods naturally generate a triple $(\tilde{z}_k, \tilde{r}_k, \varepsilon_k)$ at every iteration which can be checked for being a (ρ, ε)-saddle-point. We have terminated the three methods whenever (80) is satisfied. Table 3 reports the CPU time and the number of eigen-decompositions (resolvent evaluation of $\partial \mathcal{I}_\Omega$) for each method.

6.4 Concluding remarks

The experiment results based on three problem sets have shown that the new method Acc-SP-HPE significantly outperforms Nesterov's smoothing technique of [12] as well as the accelerated primal-dual method in [5]. However, we should note that both Nesterov's smoothing technique and the APD method can be implemented using the entropy distance-generating function and with $\mathcal{X}$ and $\mathcal{Y}$ endowed with the L_1-norm. In the future, we plan to design a variant of Acc-SP-HPE that can take advantage of entropy distance-generating function and compare it with the corresponding variants of Nesterov's method and the APD method.

Problem size		Lip. const.		Acc-SP-HPE		Nest-App		APD	
m	n	$\|C\|^2$	$\|A\|$	time	#eigen.	time	#eigen.	time	#eigen.
50	50	9.06	7.78	**1.77**	482	17.44	7954	12.49	5895
50	100	6.68	14.25	**13.00**	1108	129.25	14562	91.65	11197
50	200	6.19	26.95	**84.25**	1788	1096.60	27508	682.57	20844
100	50	14.65	8.37	**4.67**	934	28.83	7320	16.28	5935
100	100	15.01	14.82	**15.60**	898	183.27	12846	105.41	10011
100	200	13.6	27.6	**121.99**	1780	1347.50	23872	749.80	19186
200	50	27.58	9.01	**6.90**	1080	33.80	6712	22.41	5602
200	100	27.38	15.60	**42.51**	1436	267.54	11556	143.48	10161
200	200	27.2	28.4	**276.64**	2440	1886.70	21088	946.69	17593

Table 3: Computational results for the methods Acc-SP-HPE, Nest-App and APD on vector-matrix saddle-point problems (82) with different sizes. All methods are terminated using criterion (80) with $\epsilon = 10^{-2}$. CPU time in seconds and number of eigen-decompositions are reported for each method.

A   Proof of Proposition 2.3

To prove Proposition 2.3, we first prove an intermediate result in Lemma A.1.

Lemma A.1. *Define, for $k \geq 0$,*

$$\Lambda_k := \min_{u \in \Omega} \left\{ \sum_{i=1}^{k} (\Gamma_i - \Gamma_{i-1})[l_\psi(u; u_i) + g(u)] + \frac{1}{2}\|u - u_0\|_{\mathcal{X}}^2 \right\}. \tag{83}$$

Then, for every $k \geq 0$,

$$\Lambda_{k+1} - \Lambda_k \geq \Gamma_{k+1} p(\tilde{u}_{k+1}) - \Gamma_k p(\tilde{u}_k). \tag{84}$$

Proof. Since the function in the minimization problem (83) is strongly convex with modulus 1, we have

$$\Lambda_k + \frac{1}{2}\|w_k - w_{k+1}\|_{\mathcal{X}}^2 \leq \sum_{i=1}^{k} (\Gamma_i - \Gamma_{i-1})[l_\psi(w_{k+1}; u_i) + g(w_{k+1})] + \frac{1}{2}\|w_{k+1} - u_0\|_{\mathcal{X}}^2$$

$$= \Lambda_{k+1} - (\Gamma_{k+1} - \Gamma_k)[l_\psi(w_{k+1}; u_{k+1}) + g(w_{k+1})]. \tag{85}$$

Now, using the definition of $\tilde{u}_k$ in (16), the definitions (10) and (12) and the convexity of the function $l_\psi(\cdot; u_{k+1}) + g(\cdot)$, we have

$$\Gamma_{k+1}[l_\psi(\tilde{u}_{k+1}; u_{k+1}) + g(\tilde{u}_{k+1})] \leq (\Gamma_{k+1} - \Gamma_k)[l_\psi(w_{k+1}; u_{k+1}) + g(w_{k+1})] + \Gamma_k[l_\psi(\tilde{u}_k; u_{k+1}) + g(\tilde{u}_k)]$$

$$\leq (\Gamma_{k+1} - \Gamma_k)[l_\psi(w_{k+1}; u_{k+1}) + g(w_{k+1})] + \Gamma_k p(\tilde{u}_k). \tag{86}$$

Since the relation (13) implies $\Gamma_k = L(\Gamma_k - \Gamma_{k-1})^2$, using the definitions of u_k and $\tilde{u}_k$ in (14) and (16), we have

$$\|\tilde{u}_{k+1} - u_{k+1}\|^2 = \frac{(\Gamma_{k+1} - \Gamma_k)^2}{\Gamma_{k+1}^2}\|w_{k+1} - w_k\|^2 = \frac{1}{\Gamma_{k+1}L}\|w_{k+1} - w_k\|^2.$$

Therefore, the equality above and the inequalities (85) and (86) imply that

$$\Lambda_{k+1} - \Lambda_k \geq \Gamma_{k+1}[l_\psi(\tilde{u}_{k+1}; u_{k+1}) + g(\tilde{u}_{k+1})] + \frac{\Gamma_{k+1}L}{2}\|\tilde{u}_{k+1} - u_{k+1}\|^2 - \Gamma_k p(\tilde{u}_k).$$

Since ψ is L-Lipschitz continuous on Ω, we have

$$l_\psi(\tilde{u}_{k+1}; u_{k+1}) + \frac{L}{2}\|\tilde{u}_{k+1} - u_{k+1}\|^2 \geq \psi(\tilde{u}_{k+1}),$$

which, together with the above inequality and the definition (12), implies (84). $\square$

Proof of Proposition 2.3. It follows from (84) that the sequence $\{\Lambda_k - \Gamma_k p(\tilde{u}_k)\}$ is nondecreasing, which, together with the definition of Λ_k in (83) and the fact $\Gamma_0 = 0$, implies that

$$\Lambda_k - \Gamma_k p(\tilde{u}_k) \geq \Lambda_0 - \Gamma_0 p(\tilde{u}_0) = \min_{u \in \Omega} \frac{1}{2}\|u - u_0\|_{\mathcal{X}}^2 \geq 0,$$

and hence we obtained the second inequality in (17).

Moreover, since the relation (13) implies

$$\Gamma_k = L(\Gamma_k - \Gamma_{k-1})^2 = L(\Gamma_k^{1/2} - \Gamma_{k-1}^{1/2})^2(\Gamma_k^{1/2} + \Gamma_{k-1}^{1/2})^2 \leq 4L\Gamma_k(\Gamma_k^{1/2} - \Gamma_{k-1}^{1/2})^2,$$

we have

$$\Gamma_k \geq (\Gamma_{k-1}^{1/2} + \frac{1}{\sqrt{4L}})^2,$$

and hence we obtain the first inequality in (17) by induction. The inequalities in (19) follow immediately from the second inequality in (17) and the definitions (18) and (10). $\square$

References

[1] AUSLENDER, A., AND TEBOULLE, M. *Asymptotic cones and functions in optimization and variational inequalities.* Springer Monographs in Mathematics. Springer-Verlag, New York, 2003.

[2] BECK, A., AND TEBOULLE, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences 2*, 1 (2009), 183–202.

[3] BURACHIK, R. S., IUSEM, A. N., AND SVAITER, B. F. Enlargement of monotone operators with applications to variational inequalities. *Set-Valued Anal. 5*, 2 (1997), 159–180.

[4] CHAMBOLLE, A., AND POCK, T. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision 40*, 1 (2011), 120–145.

[5] CHEN, Y., LAN, G., AND OUYANG, Y. Optimal primal-dual methods for a class of saddle point problems. *arXiv preprint arXiv:1309.5548* (2013).

[6] HE, Y., AND MONTEIRO, R. D. C. Accelerating block-decomposition first-order methods for solving generalized saddle-point and Nash equilibrium problems. *Optimization-online preprint* (2013).

[7] KORPELEVIČ, G. M. An extragradient method for finding saddle points and for other problems. *Èkonom. i Mat. Metody 12*, 4 (1976), 747–756.

[8] MONTEIRO, R. D. C., AND SVAITER, B. Complexity of variants of Tseng's modified F-B splitting and korpelevich's methods for hemivariational inequalities with applications to saddle-point and convex optimization problems. *SIAM Journal on Optimization 21*, 4 (2011), 1688–1720.

[9] MONTEIRO, R. D. C., AND SVAITER, B. F. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization 20*, 6 (2010), 2755–2787.

[10] MONTEIRO, R. D. C., AND SVAITER, B. F. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM Journal on Optimization 23*, 1 (2013), 475–507.

[11] NEMIROVSKI, A. 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 Journal on Optimization 15*, 1 (2004), 229–251.

[12] NESTEROV, Y. Smooth minimization of non-smooth functions. *Mathematical Programming 103*, 1 (2005), 127–152.

[13] ROCKAFELLAR, R. T. On the maximal monotonicity of subdifferential mappings. *Pacific J. Math. 33* (1970), 209–216.

[14] ROCKAFELLAR, R. T. Monotone operators and the proximal point algorithm. *SIAM J. Control Optimization 14*, 5 (1976), 877–898.

[15] ROCKAFELLAR, R. T., AND WETS, R. J.-B. *Variational Analysis*, vol. 317. Springer, 1998.

[16] SOLODOV, M. V., AND SVAITER, B. F. A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Anal. 7*, 4 (1999), 323–345.

[17] SOLODOV, M. V., AND SVAITER, B. F. A hybrid projection-proximal point algorithm. *J. Convex Anal. 6*, 1 (1999), 59–70.

[18] SOLODOV, M. V., AND SVAITER, B. F. An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions. *Math. Oper. Res. 25*, 2 (2000), 214–230.

[19] SOLODOV, M. V., AND SVAITER, B. F. A unified framework for some inexact proximal point algorithms. *Numer. Funct. Anal. Optim. 22*, 7-8 (2001), 1013–1035.

[20] TSENG, P. A modified forward-backward splitting method for maximal monotone mappings. *SIAM J. Control Optim. 38*, 2 (2000), 431–446 (electronic).

[21] TSENG, P. On accelerated proximal gradient methods for convex-concave optimization. *submitted to Journal on Optimization* (2008).