An accelerated non-Euclidean hybrid proximal extragradient-type Algorithm for convex-concave saddle-point Problems

O. Kolossoski * R.D.C. Monteiro †

September 18, 2015

Abstract

This paper describes an accelerated HPE-type method based on general Bregman distances for solving monotone saddle-point (SP) problems. The algorithm is a special instance of a non-Euclidean hybrid proximal extragradient framework introduced by Svaiter and Solodov [28] where the prox sub-inclusions are solved using an accelerated gradient method. It generalizes the accelerated HPE algorithm presented in [13] in two ways, namely: a) it deals with general monotone SP problems instead of bilinear structured SPs; and b) it is based on general Bregman distances instead of the Euclidean one. Similar to the algorithm of [13], it has the advantage that it works for any constant choice of proximal stepsize. Moreover, a suitable choice of the stepsize yields a method with the best known iteration-complexity for solving monotone SP problems. Computational results show that the new method is superior to Nesterov's smoothing scheme [23].

2010 Mathematics Subject Classification: 90C25, 90C30, 47H05.

Key words: convex programming, complexity, ergodic convergence, maximal monotone operator, hybrid proximal extragradient method, accelerated gradient method, inexact proximal method, saddle point problem, Bregman distances.

1 Introduction

Given nonempty closed convex sets $X \subset \mathcal{X}$ and $Y \subset \mathcal{Y}$ where $\mathcal{X}$ and $\mathcal{Y}$ are two inner product spaces, and a convex-concave map $\hat{\Phi} : X \times Y \to \mathbb{R}$, our goal in this paper is to develop

*Departamento de Matemática, Universidade Federal do Paraná, Curitiba, PR, 81531-990. (email: oliver.kolossoski@ufpr.br). The work of this author was partially supported by Capes under Grant 99999.003842/2014-02.

†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.

algorithms for finding (approximate) saddle-points of $\hat{\Phi}$, i.e., pairs $(\bar{x}, \bar{y}) \in X \times Y$ such that

$$\hat{\Phi}(\bar{x}, y) \leq \hat{\Phi}(\bar{x}, \bar{y}) \leq \hat{\Phi}(x, \bar{y}) \quad \forall (x, y) \in X \times Y, \tag{1}$$

or equivalently, a zero of the operator $T : \mathcal{X} \times \mathcal{Y} \rightrightarrows \mathcal{X} \times \mathcal{Y}$ defined as

$$T(x, y) = \begin{cases} \partial[\hat{\Phi}(\cdot, y) - \hat{\Phi}(x, \cdot)](x, y), & \text{if} (x, y) \in X \times Y; \\ \emptyset, & \text{otherwise.} \end{cases} \tag{2}$$

Under mild assumptions on Φ, the operator T is maximal monotone, and hence an approximate zero of T can be computed by using an inexact proximal-type algorithm such as one of the algorithms presented in [13, 17, 18, 20, 21, 22, 26, 27, 28, 29, 30].

In particular, He and Monteiro [13] presented an inexact proximal-point method for solving the special case of the saddle-point problem in which $\hat{\Phi}$ is of the form

$$\hat{\Phi}(x, y) = f(x) + \langle Ax, y \rangle + g_1(x) - g_2(y) \tag{3}$$

where $A : \mathcal{X} \to \mathcal{Y}$ is a linear operator, g_1 and g_2 are proper closed convex functions, f is a differentiable convex function with Lipschitz continuous gradient and $X \times Y = \operatorname{dom} g_1 \times \operatorname{dom} g_2$. The method is a special instance of the hybrid proximal extragradient (HPE) framework introduced in [26]. Any instance of the HPE framework is essentially an inexact proximal point method which allows for a relative error in the prox sub-inclusions. More specifically, consider the monotone inclusion problem

$$0 \in T(z), \tag{4}$$

where T is a maximal monotone operator. Recall that for a given pair (z_-, λ), the exact proximal point method computes the next iterate z as $z = (\lambda T + I)^{-1}(z_-)$, or equivalently, as the (unique) solution of

$$\frac{z_- - z}{\lambda} \in T(z). \tag{5}$$

An instance of the HPE framework on the other hand allows an approximate solution of (4) satisfying the (relative) HPE error criterion, namely, for some tolerance $\sigma \in [0, 1]$, a triple $(\tilde{z}, z, \varepsilon)$ is computed in such a way as to satisfy

$$\frac{z_- - z}{\lambda} \in T^{\varepsilon}(\tilde{z}), \qquad \frac{1}{2}\|\tilde{z} - z\|^2 + \lambda\varepsilon \leq \frac{1}{2}\sigma^2\|\tilde{z} - z_-\|^2 \tag{6}$$

where T^{ε} denotes the ε-enlargement [2] of T. (It has the property that $T^{\varepsilon}(u) \supset T(u)$ for each u with equality holding when $\varepsilon = 0$.) Clearly, when $\sigma = 0$ in (6), then $z = \tilde{z}$ and $\varepsilon = 0$, and the inclusion in (6) reduces to (5). As opposed to other HPE-type methods in the literature (see for instance [12, 21]) which have to choose λ relatively small, the HPE method of [13] for solving (4) with T as in (2) can choose an arbitrarily sized $\lambda > 0$ and computes the triple $(\tilde{z}, z, \varepsilon)$ satisfying the HPE error condition (6) with the aid of an accelerated gradient method (e.g., see [23, 31]) applied to a certain regularized convex-concave min-max problem related to $\hat{\Phi}$ in (3).

The main goal of this paper is to develop a non-Euclidean HPE (NE-HPE) method which extends the one of [13] in two relevant ways. First, it solves saddle-point problems with general

convex-concave functions $\hat{\Phi}$ such that $\nabla_x \hat{\Phi}$ is Lipschitz continuous instead of those with $\hat{\Phi}$ given by (3). Second, the method is a special instance of a more general non-Euclidean HPE framework which is based on a general Bregman distance instead of the specific Euclidean one. More specifically, let $d_z(z') = w(z') - w(z) - \langle \nabla w(z), z' - z \rangle$ for every z, z' where w is a differentiable convex function. Then, the Euclidean distance is obtained by setting $w(\cdot) = \|\cdot\|^2/2$ in which case $d_z(z') = \|z' - z\|^2/2$ for all z', z and (6) can be written as

$$\frac{1}{\lambda} \nabla (dw)_z(z_-) \in T^{\varepsilon}(\tilde{z}), \quad (dw)_z(\tilde{z}) + \lambda \varepsilon \leq \sigma (dw)_{z_-}(\tilde{z}). \tag{7}$$

The non-Euclidean HPE framework generalizes the HPE one in that it allows an approximate solution of (4) satisfying the more general NE-HPE error condition (7) where w is an arbitrary convex function. As an important step towards analyzing the new NE-HPE method, we establish the ergodic iteration-complexity of the NE-HPE framework for solving inclusion (4) where T is a maximal monotone operator with bounded domain. Similar to the method in [13], the new NE-HPE method chooses an arbitrary $\lambda > 0$ and computes a triple $(\tilde{z}, z, \varepsilon)$ satisfying the HPE error condition (7) with the aid of an accelerated gradient method applied to a certain regularized convex-concave min-max problem. Under the assumption that the feasible set of the saddle-point problem is bounded, an ergodic iteration-complexity bound is developed for the total number of inner (accelerated gradient) iterations performed by the new NE-HPE method. Finally, it is shown that if the stepsize λ and Bregman distance are properly chosen, then the derived ergodic iteration-complexity reduces to the one obtained in [23] for Nesterov's smoothing scheme which finds approximate solutions of a bilinear structured convex-concave saddle-point problem. Such complexity bound is known to be optimal (see for example the discussion in paragraph (1) of Subsection 1.1 of [6]).

Our paper is organized as follows: Section 2 contains two subsections which provide the necessary background material for our presentation. Subsection 2.1 introduces some notation, presents basic definitions and properties of point-to-set operators and convex functions, and discusses the saddle-point problem and some of its basic properties. Subsection 2.2 reviews an accelerated gradient method for solving composite convex optimization problems. Section 3 contains two subsections. Subsection 3.1 reviews the notion of Bregman distances, then presents the NE-HPE framework for solving (4) and establishes its ergodic iteration-complexity. Subsection 3.2 presents a sufficient condition which ensures that the HPE error condition (7) holds for a special class of monotone operators (which includes the one given by (2)). Section 4 describes the new accelerated NE-HPE method for solving the saddle-point problem, i.e., inclusion (4) with the operator given by (2). It contains three subsections as follows. Subsection 4.1 presents a scheme based on the accelerated gradient method of Subsection 2.2 for finding an approximate solution of the prox sub-inclusion according to the NE-HPE error criterion (7) and states its iteration-complexity result. Subsection 4.2 completely describes the new accelerated NE-HPE method for solving the saddle-point problem and establishes its overall ergodic inner iteration-complexity. It also discusses a way of choosing the prox stepsize λ so that the overall ergodic inner iteration-complexity bound reduces to the one obtained for Nesterov's smoothing scheme [23]. Subsection 4.3 provides the proof of the iteration-complexity result stated in Subsection 4.1. Finally, numerical results are presented in Section 5 showing that the new method

outperforms the scheme of [23] on three classes of convex-concave saddle-point problems.

1.1 Previous related works

In the context of variational inequalities, Nemirovski [22] established the ergodic iteration-complexity of an extension of Korpelevich's method [17], namely, the mirror-prox algorithm, under the assumption that the feasible set of the problem is bounded. More recently, Dang and Lan [8] established the iteration-complexity of a class of non-Euclidean extragradient methods for solving variational inequalities when the operators are not necessarily monotone. Also, Lan et al. [7] established the iteration-complexity of an accelerated mirror-prox method which finds weak solutions of a class of variational inequalities. They obtained optimal complexity for the case where the feasible set of the problem is bounded.

Nesterov [23] developed a smoothing scheme for solving bilinear structured saddle-point problems under the assumption that X and Y are compact convex sets. It consists of first approximating the objective function of the associated convex-concave saddle-point problem by a convex differentiable function with Lipschitz continuous gradient and then applying an accelerated gradient-type method (see e.g. [23, 31]) to the resulting approximation problem.

The HPE framework and its convergence results are studied in [26] and its iteration-complexity is established in [20] (see also [18, 21]). The complexity results in [20] 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 [20] and in the subsequent papers [18, 21]. More specifically, by viewing Korpelevich's method [17] as well as Tseng's modified forward-backward splitting (MF-BS) method [30] as special cases of the HPE framework, the authors have established in [18, 20] 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.

Solodov and Svaiter [28] has studied a more specialized version of the NE-HPE framework which allows approximate solutions of (4) according to (7) but with $\varepsilon = 0$. Finally, extensions of the proximal method to the context of Bregman distances have been studied in [4, 5, 10, 11, 15, 16]. However, none of the works cited in this paragraph deal with iteration-complexity results.

2 Background material

This section provides background material necessary for the paper presentation. The first one presents the notation and basic definitions that will be used in the paper. The second subsection reviews a variant of Nesterov's accelerated method for the composite convex optimization problem.

2.1 Basic notation, definitions and results

This subsection establishes notation and gives basic results that will be used throughout the paper.

The set of real numbers is denoted by $\mathbb{R}$. The set of non-negative real numbers and the set of positive real numbers are denoted respectively as $\mathbb{R}_+$ and $\mathbb{R}_{++}$. Let $\lceil z \rceil$ denote the smallest integer not less than $z \in \mathbb{R}$.

2.1.1 Convex functions, monotone operators and their enlargements

Let $\mathcal{Z}$ denote a finite dimensional inner product space with inner product denoted by $\langle \cdot, \cdot \rangle$. For a set $Z \subset \mathcal{Z}$, its relative interior is denoted by $\mathrm{ri}\, Z$ and its closure as $\mathrm{cl} Z$. 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

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

The domain of T is defined as

$$\mathrm{Dom}\, T := \{z \in \mathcal{Z} : T(z) \neq \emptyset\}.$$

The domain of definition of a point-to-point map F is also denoted by $\mathrm{Dom}\, F$. An operator $T : \mathcal{Z} \rightrightarrows \mathcal{Z}$ is *monotone* if

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

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

$$T^\varepsilon(z) := \{v \in \mathcal{Z} \ : \langle z - \tilde{z}, v - \tilde{v} \rangle \geq -\varepsilon, \quad \forall \tilde{z} \in \mathcal{Z},\ \forall \tilde{v} \in T(\tilde{z})\} \quad \forall z \in \mathcal{Z}. \tag{8}$$

The following result gives some useful properties of ε-enlargements.

Proposition 2.1. *Let $T, T' : \mathcal{Z} \rightrightarrows \mathcal{Z}$ be given. Then, the following statement holds: $T^{\varepsilon_1}(z) + (T')^{\varepsilon_2}(z) \subset (T + T')^{\varepsilon_1 + \varepsilon_2}(z)$ for every $z \in \mathcal{Z}$ and $\varepsilon_1, \varepsilon_2 \in \mathbb{R}$.*

Proof. The proof follows directly from definition (8). ∎

Let $f : \mathcal{Z} \to [-\infty, \infty]$ be given. The effective domain of f is defined as

$$\mathrm{dom}\, f := \{z \in \mathcal{Z} : f(z) < \infty\}.$$

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

$$\partial_\varepsilon f(z) := \{v : f(\tilde{z}) \geq f(z) + \langle \tilde{z} - z, v \rangle - \varepsilon,\ \forall \tilde{z} \in \mathcal{Z}\} \quad \forall z \in \mathcal{Z}. \tag{9}$$

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 [25].

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

$$\mathcal{I}_\Omega(z) := \begin{cases} 0, & z \in \Omega, \\ \infty, & z \notin \Omega. \end{cases} \tag{10}$$

The following simple but useful result shows that adding a maximal monotone operator T to the subdifferential of the indicator function of a convex set containing $\mathrm{Dom}\, T$ does not change T.

Proposition 2.2. *Assume that T is a maximal monotone operator and $\Omega \subset \mathcal{Z}$ is a convex set containing $\mathrm{Dom}\, T$. Then, $T + \partial\mathcal{I}_\Omega = T$.*

Proof. Clearly, $\partial\mathcal{I}_\Omega$ is monotone since, by assumption, the set Ω, and hence the indicator function $\mathcal{I}_\Omega$, is convex. Since T is also monotone by assumption, it follows that $T + \partial\mathcal{I}_\Omega$ is monotone in view of Proposition 6.1.1(b) of [1]). Clearly, $T \subset T + \partial\mathcal{I}_\Omega$ due to the assumption that $\mathrm{Dom}\, T \subset \Omega$ and the fact that $0 \in \partial\mathcal{I}_\Omega(x)$ for every $x \in \Omega$. The conclusion of the proposition now follows from the above two observations and the assumption that T is maximal monotone. ∎

2.1.2 The saddle-point problem

Let $\mathcal{X}$ and $\mathcal{Y}$ be finite dimensional inner product spaces with inner products denoted respectively by $\langle\cdot,\cdot\rangle_\mathcal{X}$ and $\langle\cdot,\cdot\rangle_\mathcal{Y}$ and endow the product space $\mathcal{X}\times\mathcal{Y}$ with the canonical inner product defined as

$$\langle (x,y),(x',y')\rangle = \langle x,x'\rangle_\mathcal{X} + \langle y,y'\rangle_\mathcal{Y} \quad \forall (x,y),(x',y') \in \mathcal{X}\times\mathcal{Y}. \tag{11}$$

Let $X \subseteq \mathcal{X}$ and $Y \subseteq \mathcal{Y}$ be nonempty sets and consider a function $\hat{\Phi} : \mathrm{Dom}\,\hat{\Phi} \to \mathbb{R}$ such that $\mathrm{Dom}\,\hat{\Phi} \supset X \times Y$. A pair $(\bar{x},\bar{y}) \in X \times Y$ is called a saddle-point of $\hat{\Phi}$ with respect to $X \times Y$ if it satisfies (1). The problem of determining such a pair is called the saddle-point problem determined by $\hat{\Phi}$ and $Z := X \times Y$ and is denoted by $\mathrm{SP}(\hat{\Phi}; Z)$.

Setting $\mathcal{Z} := \mathcal{X}\times\mathcal{Y}$, define $T_{\hat{\Phi}} : \mathcal{Z} \rightrightarrows \mathcal{Z}$ as

$$T_{\hat{\Phi}}(z) := \begin{cases} \partial(\hat{\Phi}_z)(z), & \text{if } z \in Z; \\ \emptyset, & \text{otherwise} \end{cases} \tag{12}$$

where, for every $z = (x,y) \in Z$, the function $\hat{\Phi}_z : \mathcal{Z} \to (-\infty, +\infty]$ is defined as

$$\hat{\Phi}_z(x',y') = \begin{cases} \hat{\Phi}(x',y) - \hat{\Phi}(x,y'), & \forall (x',y') \in Z; \\ +\infty, & \text{otherwise.} \end{cases} \tag{13}$$

Clearly, $z = (x,y)$ is a saddle-point of $\hat{\Phi}$ with respect to Z if and only if z is a solution of the inclusion

$$0 \in T_{\hat{\Phi}}(z). \tag{14}$$

The operator $T_{\hat{\Phi}}$ admits the ε-enlargement as in (8). It also admits an ε-saddle-point enlargement which exploits its natural saddle-point nature, namely, $\partial_\varepsilon(\hat{\Phi}_z)(z)$ for $z \in Z$. The following result whose proof can be found for example in Lemma 3.2 of [13] follows straightforwardly from definitions (8) and (9).

Proposition 2.3. *For every $z \in Z$ and $\varepsilon \geq 0$, the inclusion $\partial_\varepsilon(\hat{\Phi}_z)(z) \subset [T_{\hat{\Phi}}]^\varepsilon(z)$ holds.*

The following result (see for example Theorem 6.3.2 in [1]) gives sufficient conditions for the operator $T_{\hat{\Phi}}$ in (12) to be maximal monotone.

Proposition 2.4. *The following statements hold:*

(a) $T_{\hat{\Phi}}$ is monotone;

(b) if the function $\hat{\Phi}_z : \mathcal{Z} \to (-\infty, +\infty]$ is closed convex for every $z \in Z$, then $T_{\hat{\Phi}}$ is maximal monotone.

Note that, due to the definition of T^ε, the verification of the inclusion $v \in T^\varepsilon(x)$ requires checking an infinite number of inequalities. This verification is feasible only for specially-structured instances of operators T. However, it is possible to compute points in the graph of T^ε using the following *weak transportation formula* [3].

Proposition 2.5. *Suppose that $T : \mathcal{Z} \rightrightarrows \mathcal{Z}$ is maximal monotone. Moreover, for $i = 1, \ldots, k$, let $z_i, r_i \in \mathcal{Z}$ and $\varepsilon_i, \alpha_i \in \mathbb{R}_+$ satisfying $\sum_{i=1}^k \alpha_i = 1$ be given and define*

$$z^a = \sum_{i=1}^k \alpha_i z_i, \quad r^a = \sum_{i=1}^k \alpha_i r_i, \quad \varepsilon^a = \sum_{i=1}^k \alpha_i \left[\varepsilon_i + \langle z_i - z^a, r_i - r^a \rangle\right].$$

Then, the following statements hold:

(a) if $r_i \in T^{\varepsilon_i}(z_i)$ for every $i = 1, \ldots, k$, then $\varepsilon^a \geq 0$ and $r^a \in T^{\varepsilon^a}(z^a)$;

(b) if $T = T_{\hat{\Phi}}$ where $\hat{\Phi}$ is a saddle function satisfying the assumptions of Proposition 2.4 and the stronger inclusion $r_i \in \partial_{\varepsilon_i}(\hat{\Phi}_{z_i})(z_i)$ holds for every $i = 1, \ldots, k$, then

$$r^a \in \partial_{\varepsilon^a}(\hat{\Phi}_{z^a})(z^a).$$

2.2 Accelerated method for composite convex optimization

This subsection reviews a variant of Nesterov's accelerated first-order method [23, 31] for minimizing a (possibly strongly) convex composite function. In what follows, we refer to convex functions as 0-strongly convex functions. This terminology has the benefit of allowing us to treat both the convex and strongly convex case simultaneously.

Let $\mathcal{X}$ denote a finite dimensional inner product space with an inner product denoted by $\langle \cdot, \cdot \rangle_{\mathcal{X}}$ and a norm denoted by $\| \cdot \|_{\mathcal{X}}$ which is not necessarily the one induced by the inner product. Consider the following composite optimization problem

$$\inf f(x) + g(x) \tag{15}$$

where $f : X \subset \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{X} \to [-\infty, +\infty]$ satisfy the following conditions:

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

A.2) X is a convex set such that $X \supset \operatorname{dom} g$;

A.3) there exist constant $L > 0$ and function $\nabla f : X \to X$ such that for every $x, x' \in X$,

$$f(x') + \langle \nabla f(x'), x - x' \rangle_{\mathcal{X}} \leq f(x) \leq f(x') + \langle \nabla f(x'), x - x' \rangle_{\mathcal{X}} + \frac{L}{2}\|x - x'\|_{\mathcal{X}}^2.$$

Even though the map ∇f is not necessarily the gradient of f, it plays a similar role to it and hence our notation.

The accelerated method for solving problem (15) stated below requires the specification of a point $x_0 \in \operatorname{dom} g$ and a function $h : \mathcal{X} \to (-\infty, \infty]$ satisfying

A.4) h is a proper closed convex function such that $\operatorname{dom} h \supset \operatorname{dom} g$;

A.5) h is 1-strongly convex on $\operatorname{dom} g$;

A.6) $x_0 = \operatorname{argmin}\{h(x) : x \in \operatorname{dom} g\}$.

Clearly, if $\operatorname{dom} g$ is closed then the above optimization problem always has a unique global minimum which can be taken to be the point x_0. The special case of the method below with $\mu = 0$ is the same as the accelerated variant stated in Algorithm 3 of [31]. Its proof for $\mu > 0$ is not given in [31] but follows along the same line as the one for Algorithm 3 of [31] (see also Subsection 2.2 of [13] for the proof of the case where $\mu > 0$, X is closed and $h(\cdot) = \|\cdot - u_0\|^2/2$ for some $u_0 \in \mathcal{X}$).

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

0) Set $A_0 := 0$, $\Lambda_0 :\equiv 0$, $k = 1$ and $\tilde{x}_0 = x_0$ where x_0 is as in A.6;

1) compute

$$A_k := A_{k-1} + \frac{(1 + \mu A_{k-1}) + \sqrt{(1 + \mu A_{k-1})^2 + 4L(1 + \mu A_{k-1})A_{k-1}}}{2L}, \tag{16}$$

$$\breve{x}_k := \frac{A_{k-1}}{A_k}\tilde{x}_{k-1} + \frac{A_k - A_{k-1}}{A_k}x_{k-1}, \tag{17}$$

$$\Lambda_k := \frac{A_{k-1}}{A_k}\Lambda_{k-1} + \frac{A_k - A_{k-1}}{A_k}\left[f(\breve{x}_k) + \langle \nabla f(\breve{x}_k), \cdot - \breve{x}_k \rangle_{\mathcal{X}}\right]; \tag{18}$$

2) iterate x_k and $\tilde{x}_k$ as

$$x_k := \operatorname{argmin}\left\{\Lambda_k(x) + g(x) + \frac{1}{A_k}h(x)\right\}, \tag{19}$$

$$\tilde{x}_k := \frac{A_{k-1}}{A_k}\tilde{x}_{k-1} + \frac{A_k - A_{k-1}}{A_k}x_k; \tag{20}$$

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

end

The main technical result which yields the convergence rate of the above accelerated method is as follows.

Proposition 2.6. *The sequences $\{A_k\}$, $\{\tilde{x}_k\}$ and $\{\Lambda_k\}$ generated by Algorithm 1 satisfy the following inequalities for any $k \geq 1$:*

$$A_k \geq \frac{1}{L}\max\left\{\frac{k^2}{4}, \left(1+\sqrt{\frac{\mu}{4L}}\right)^{2(k-1)}\right\}, \tag{21}$$

$$\Lambda_k \leq f, \quad (f+g)(\tilde{x}_k) \leq \Lambda_k(x) + g(x) + \frac{1}{A_k}\left[h(x) - h(x_0)\right] \quad \forall x \in \operatorname{dom} g. \tag{22}$$

3 The non-Euclidean hybrid proximal extragradient framework

This section describes the non-Euclidean hybrid proximal extragradient framework for solving monotone inclusion problems. Such framework as well as its convergence properties are stated on Subsection 3.1. A sufficient condition for obtaining a HPE error condition is given on Subsection 3.2.

It is assumed throughout this section that $\mathcal{Z}$ is an inner product space with inner product $\langle\cdot,\cdot\rangle$, and that $\|\cdot\|$ is a (general) norm in $\mathcal{Z}$, which is not necessarily the inner product induced norm.

3.1 The non-Euclidean hybrid proximal extragradient framework for monotone inclusion problem

This subsection establishes the non-Euclidean hybrid proximal extragradient framework introduced in [26] for finding an approximate solution of the monotone inclusion problem (4) where $T : \mathcal{Z} \rightrightarrows \mathcal{Z}$ is a maximal monotone operator such that $\operatorname{Dom} T$ is bounded.

Before presenting the framework, we review the notions of distance generating functions and Bregman distances.

Definition 3.1. *A proper closed convex function $w : \mathcal{Z} \to [-\infty, \infty]$ is called a distance generating function if it satisfies the following conditions:*

(i) $W := \operatorname{dom} w$ is closed and $W^0 := int(W) = \{z \in \mathcal{Z} : \partial w(z) \neq \emptyset\}$;

(ii) w restricted to W is continuous and w is continuously differentiable on W^0.

Moreover, w induces the map $dw : \mathcal{Z} \times W^0 \to \mathbb{R}$ defined as

$$(dw)(z'; z) := w(z') - w(z) - \langle \nabla w(z), z' - z\rangle \quad \forall z' \in \mathcal{Z}, \forall z \in W^0, \tag{23}$$

referred to as the Bregman distance over W induced by w, or simply, a Bregman distance over W when w does not need to be emphasized.

For simplicity, for every $z \in W^0$, the function $(dw)(\cdot; z)$ will be denoted by $(dw)_z$ so that

$$(dw)_z(z) = (dw)(z'; z) \quad \forall z \in W^0, \forall z' \in \mathcal{Z}.$$

The following useful identities follow straightforwardly from (23):

$$\nabla(dw)_z(z') = -\nabla(dw)_{z'}(z) = \nabla w(z') - \nabla w(z) \quad \forall z, z' \in W^0, \tag{24}$$

$$(dw)_v(z') - (dw)_v(z) = \langle \nabla(dw)_v(z), z' - z \rangle + (dw)_z(z') \quad \forall z' \in \mathcal{Z}, \forall v, z \in W^0. \tag{25}$$

The description of the non-Euclidean hybrid proximal extragradient framework is based on a distance generating function w which also satisfies the following condition:

B.1) $\mathrm{Dom}\, T \subset W$.

From now on, denote $Z := \mathrm{cl}(\mathrm{Dom}\, T)$. Clearly, due to B.1 and the fact that W is closed we have that $Z \subset W$.

We are now ready to state the non-Euclidean hybrid proximal extragradient framework.

Non-Euclidean Hybrid Proximal Extragradient (NE-HPE) Framework:

0) Let $z_0 \in W^0$ be given and set $j = 1$;

1) choose $\sigma_j \in [0, 1]$, and find $\lambda_j > 0$ and $(\tilde{z}_j, z_j, \varepsilon_j) \in W \times W^0 \times \mathbb{R}_+$ such that

$$r_j := \frac{1}{\lambda_j} \nabla(dw)_{z_j}(z_{j-1}) \in T^{\varepsilon_j}(\tilde{z}_j), \tag{26}$$

$$(dw)_{z_j}(\tilde{z}_j) + \lambda_j \varepsilon_j \leq \sigma_j (dw)_{z_{j-1}}(\tilde{z}_j); \tag{27}$$

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

end

We now make several remarks about the NE-HPE framework. First, the NE-HPE framework does not specify how to find λ_j and $(\tilde{z}_j, z_j, \varepsilon_j)$ satisfying (26) and (27). The particular scheme for computing λ_j and $(\tilde{z}_j, z_j, \varepsilon_j)$ will depend on the instance of the framework under consideration and the properties of the operator T. Second, if $\sigma_j = 0$, then (27) implies that $\varepsilon_j = 0$ and $z_j = \tilde{z}_j$, and hence that $r_j \in T(z_j)$ in view of (26). Therefore, the HPE error conditions (26)-(27) can be viewed as a relaxation of an iteration of the exact non-Euclidean proximal point method, namely,

$$\frac{1}{\lambda_j} \nabla(dw)_{z_j}(z_{j-1}) \in T(z_j).$$

Third, if dw satisfies some additional conditions (i.e., the one in Definition 3.6), then it is shown in Proposition A.2 of Appendix A that the above inclusion has a unique solution z_j, from which we conclude that, for any given $\lambda_j > 0$, it is always possible to obtain a triple $(\tilde{z}_j, z_j, \varepsilon_j)$ of the form $(z_j, z_j, 0)$ satisfying (26)-(27) with $\sigma_j = 0$. Clearly, computing the triple in this (exact)

manner is expensive, and hence (inexact) triples $(\tilde{z}_j, z_j, \varepsilon_j)$ satisfying the HPE (relative) error conditions with $\sigma_j > 0$ are computationally more appealing.

It is possible to show that some well-known methods such as the ones in [17, 30] can be viewed as special instances of the NE-HPE framework. Section 4 presents another instance of the above framework in the context of the saddle-point problem (and hence with $T = T_{\hat{\Phi}}$) in which λ_j is chosen in a interval of the form $[\tau\lambda, \lambda]$ for some fixed $\lambda > 0$ and $\tau \in (0, 1)$ and the triples as in step 1 are obtained by means of the accelerated gradient method of Subsection 2.2 applied to a certain optimization problem.

In the remaining part of this subsection, we focus our attention on establishing ergodic convergence rate bounds for the NE-HPE framework. We start by deriving some preliminary technical results.

Lemma 3.2. *For every $j \geq 1$, the following statements hold:*

(a) for every $z \in W$, we have

$$(dw)_{z_{j-1}}(z) - (dw)_{z_j}(z) = (dw)_{z_{j-1}}(\tilde{z}_j) - (dw)_{z_j}(\tilde{z}_j) + \lambda_j \langle r_j, \tilde{z}_j - z\rangle;$$

(b) for every $z \in W$, we have

$$(dw)_{z_{j-1}}(z) - (dw)_{z_j}(z) \geq (1-\sigma_j)(dw)_{z_{j-1}}(\tilde{z}_j) + \lambda_j(\langle r_j, \tilde{z}_j - z\rangle + \varepsilon_j).$$

Proof. (a) Using (25) twice and using the definition of r_j given by (26), we have that

$$\begin{aligned}
&(dw)_{z_{j-1}}(z) - (dw)_{z_j}(z) = (dw)_{z_{j-1}}(z_j) + \langle \nabla(dw)_{z_{j-1}}(z_j), z - z_j\rangle \\
&= (dw)_{z_{j-1}}(z_j) + \langle \nabla(dw)_{z_{j-1}}(z_j), \tilde{z}_j - z_j\rangle + \langle \nabla(dw)_{z_{j-1}}(z_j), z - \tilde{z}_j\rangle \\
&= (dw)_{z_{j-1}}(\tilde{z}_j) - (dw)_{z_j}(\tilde{z}_j) + \langle \nabla(dw)_{z_{j-1}}(z_j), z - \tilde{z}_j\rangle \\
&= (dw)_{z_{j-1}}(\tilde{z}_j) - (dw)_{z_j}(\tilde{z}_j) + \lambda_j\langle r_j, \tilde{z}_j - z\rangle.
\end{aligned}$$

(b) This statement follows as an immediate consequence of (a) and (27). ∎

The following result follows as an immediate consequence of Lemma 3.2(b).

Lemma 3.3. *For every $j \geq 1$ and $z \in W$, we have that*

$$(dw)_{z_0}(z) - (dw)_{z_j}(z) \geq \sum_{i=1}^{j}(1-\sigma_i)(dw)_{z_{i-1}}(\tilde{z}_i) + \sum_{i=1}^{j}\lambda_i\left[\varepsilon_i + \langle r_i, \tilde{z}_i - z\rangle\right].$$

Proof. The lemma follows by adding the inequality in Lemma 3.2(b) from 1 to j. ∎

Lemma 3.4. *For every $j \geq 1$, define $\Lambda_j := \sum_{i=1}^{j}\lambda_i$,*

$$\tilde{z}_j^a := \frac{1}{\Lambda_j}\sum_{i=1}^{j}\lambda_i\tilde{z}_i, \quad r_j^a := \frac{1}{\Lambda_j}\sum_{i=1}^{j}\lambda_i r_i, \quad \varepsilon_j^a := \frac{1}{\Lambda_j}\sum_{i=1}^{j}\lambda_i\left[\varepsilon_i + \langle r_i, \tilde{z}_i - \tilde{z}_j^a\rangle\right].$$

Then, we have

$$\varepsilon_j^a \geq 0, \quad r_j^a \in T^{\varepsilon_j^a}(\tilde{z}_j^a), \tag{28}$$

$$\varepsilon_j^a + \langle r_j^a, \tilde{z}_j^a - z\rangle \leq \frac{(dw)_{z_0}(z)}{\Lambda_j} \quad \forall z \in Z. \tag{29}$$

Proof. The relations on (28) follow from (26) and Proposition 2.5(a). Moreover, Lemma 3.3, the assumption that $\sigma_j \in [0,1]$, the fact that $Z \subset W$, and the definitions of ε_j^a and r_j^a, imply that for every $z \in Z$,

$$(dw)_{z_0}(z) - (dw)_{z_j}(z) \geq \sum_{i=1}^{j} \lambda_i \left[\varepsilon_i + \langle r_i, \tilde{z}_i - z\rangle\right] =$$

$$\sum_{i=1}^{j} \lambda_i \left[\varepsilon_i + \langle r_i, \tilde{z}_i - \tilde{z}_j^a\rangle + \langle r_i, \tilde{z}_j^a - z\rangle\right] = \Lambda_j \left[\varepsilon_j^a + \langle r_j^a, \tilde{z}_j^a - z\rangle\right],$$

and hence that (29) holds. ∎

To state the main result of this subsection, which establishes an ergodic convergence rate bound for the NE-HPE framework, define the quantity

$$R = R(z_0; Z) := \sup\left\{(dw)_{z_0}(z) : z \in Z\right\} < \infty \tag{30}$$

and observe that R is finite due to the boundedness assumption on $\mathrm{Dom}\, T$ and the facts that $Z \subset W$ and $(dw)_{z_0}(\cdot)$ is a continuous function on W (see Definition 3.1(ii)).

Theorem 3.5. *For every $j \geq 1$, define Λ_j, $\tilde{z}_j^a$, r_j^a and ε_j^a as in Lemma 3.4, and also*

$$\tilde{\varepsilon}_j := \varepsilon_j^a + \max\left\{\langle r_j^a, \tilde{z}_j^a - z\rangle : z \in Z\right\}. \tag{31}$$

Then, the following statements hold:

(a) for every $j \geq 1$, it holds

$$0 \in T^{\tilde{\varepsilon}_j}(\tilde{z}_j^a), \quad \tilde{\varepsilon}_j \leq \frac{R}{\Lambda_j}; \tag{32}$$

(b) if $T = T_{\hat{\Phi}}$ where $\hat{\Phi}$ is a function satisfying the assumptions of Proposition 2.4, and the stronger inclusion

$$r_j \in \partial_{\varepsilon_j}(\hat{\Phi}_{\tilde{z}_j})(\tilde{z}_j) \tag{33}$$

holds for every $j \geq 1$, then

$$0 \in \partial_{\tilde{\varepsilon}_j}(\hat{\Phi}_{\tilde{z}_j^a})(\tilde{z}_j^a) \quad \forall j \geq 1.$$

Proof. (a) Inequality (29), the definition of R in (30) and the definition of $\tilde{\varepsilon}_j$ in (31) clearly imply the inequality in (32). Now, let $\delta_j := \tilde{\varepsilon}_j - \varepsilon_j^a$ and note that (31) and the definitions of the ε-subdiffential and the indicator function in (9) and (10), respectively, imply that $-r_j^a \in \partial_{\delta_j}(\mathcal{I}_Z)(\tilde{z}_j^a)$. This inclusion, the inclusion in (28) and Propositions 2.1 and 2.2, then imply that

$$0 \in T^{\varepsilon_j^a}(\tilde{z}_j^a) + (\partial \mathcal{I}_Z)^{\delta_j}(\tilde{z}_j^a) \subset (T + \partial \mathcal{I}_Z)^{\varepsilon_j^a + \delta_j}(\tilde{z}_j^a) = T^{\varepsilon_j^a + \delta_j}(\tilde{z}_j^a) = T^{\tilde{\varepsilon}_j}(\tilde{z}_j^a)$$

where the last equality is due to the definition of δ_j.

(b) This statement follows by using similar arguments as the ones used in the inclusions of Lemma 3.4 and statement (a) except that Proposition 2.5(b) is used in place of Proposition 2.5(a). ∎

Note that $\tilde{\varepsilon}_j$ in (31) can be easily computed for those instances of (31) for which the minimization of a linear function on Z can be trivially performed. Note also that if Λ_j grows to ∞, relation (32) implies that any limit point of $\tilde{z}_j^a$ is a solution of (4). The inequality in this relation implies that the convergence rate of $\tilde{z}_j^a$, measured in terms of the size of $\tilde{\varepsilon}_j$, is on the order of $\mathcal{O}(1/\Lambda_j)$. Clearly, this convergence rate reduces to $\mathcal{O}(1/j)$ for the case in which the sequence of stepsizes $\{\lambda_j\}$ is constant.

3.2 A sufficient condition for the HPE error conditions

In this subsection, a sufficient condition which ensures the HPE error conditions (26)-(27) is derived for special classes of maximal monotone operators and Bregman distances.

Observe that given $z_- = z_{j-1} \in W^0$ and $\sigma = \sigma_j \in [0,1]$, each iteration of the NE-HPE framework involves the computation of a stepsize $\lambda = \lambda_j > 0$ and a triple $(\tilde{z}, z, \varepsilon) = (\tilde{z}_j, z_j, \varepsilon_j) \in W \times W^0 \times \mathbb{R}_+$ satisfying the HPE error conditions:

$$\frac{1}{\lambda}\nabla(dw)_z(z_-) \in T^\varepsilon(\tilde{z}), \tag{34}$$

$$(dw)_z(\tilde{z}) + \lambda\varepsilon \leq \sigma(dw)_{z_-}(\tilde{z}). \tag{35}$$

Our goal in this subsection is to derive a sufficient condition for obtaining such a triple. Our discussion in this subsection applies to maximal monotone operators satisfying the following condition:

C.1) for every $\tilde{z} \in \mathrm{Dom}\, T$, there exists a proper closed convex function $f_{\tilde{z}} : \mathcal{Z} \to [-\infty, \infty]$ such that $\tilde{z} \in \mathrm{dom}(f_{\tilde{z}})$ and $\partial_\varepsilon(f_{\tilde{z}})(\tilde{z}) \subset T^\varepsilon(\tilde{z})$ for every $\varepsilon \geq 0$.

Note that Proposition 2.3 implies that the operator $T_{\hat{\Phi}}$ defined in (12) satisfies condition C.1 with $f_{\tilde{z}} = \hat{\Phi}_{\tilde{z}}$ for every $\tilde{z} \in \mathrm{Dom}\, T_{\hat{\Phi}}$ where $\hat{\Phi}_{\tilde{z}}$ is defined in (13).

To state the main result of this subsection, we need to introduce a special class of Bregman distances that will also be used in other parts of the paper.

Definition 3.6. *For a given scalar $\mu > 0$ and a nonempty convex set $\Omega \subset \mathcal{Z}$, a Bregman distance dw over W is called a (μ, Ω)-Bregman distance over W if $\Omega \cap W^0 \neq \emptyset$ and*

$$(dw)_z(z') \geq \frac{\mu}{2}\|z' - z\|^2 \quad \forall z, z' \in \Omega \cap W^0. \tag{36}$$

We now make some remarks about the above definition. First, for every $z \in \Omega \cap W^0$, the inequality in (36) holds for every $z' \in \Omega \cap W$ due to the continuity of $(dw)_z(\cdot)$. Second, (36) is equivalent to the distance generating function w being μ-strongly convex on $\Omega \cap W$.

We now present the aforementioned sufficient condition for obtaining a triple satisfying (34)-(35).

Proposition 3.7. *Suppose that T is a maximal monotone operator satisfying condition C.1 and that dw is a $(\mu, \operatorname{Dom} T)$-Bregman distance over W for some $\mu > 0$. Let $\lambda > 0$, $z_- \in W^0$ and $\tilde{z} \in \operatorname{Dom} T \cap W^0$ be given. Assume that there exists a proper closed convex function such that $\Gamma_{\tilde{z}} \leq f_{\tilde{z}}$ and define the quantities*

$$z := \operatorname{argmin}_u \left\{ \Gamma_{\tilde{z}}(u) + \frac{1}{\lambda}(dw)_{z_-}(u) \right\} \tag{37}$$

$$\varepsilon := f_{\tilde{z}}(\tilde{z}) - \Gamma_{\tilde{z}}(z) - \langle r, \tilde{z} - z \rangle \tag{38}$$

where

$$r := \frac{1}{\lambda} \nabla (dw)_z(z_-). \tag{39}$$

Then, the following statements hold:

(a) z is well-defined, $\varepsilon \in [0, \infty)$ and the following inclusion (which is stronger than (34) due to C.1) holds:

$$r \in \partial_\varepsilon (f_{\tilde{z}})(\tilde{z}); \tag{40}$$

(b) if, in addition, for a given scalar $\sigma \geq 0$, we have

$$f_{\tilde{z}}(\tilde{z}) + \frac{1-\sigma}{\lambda}(dw)_{z_-}(\tilde{z}) \leq \inf \left\{ \Gamma_{\tilde{z}}(u) + \frac{1}{\lambda}(dw)_{z_-}(u) : u \in \mathcal{Z} \right\} \tag{41}$$

then (35) holds.

Proof. (a) The assumptions that $\tilde{z} \in \operatorname{Dom} T \cap W^0$ and $\Gamma_{\tilde{z}} \leq f_{\tilde{z}}$ together with condition C.1 imply that $\tilde{z} \in \operatorname{dom} f_{\tilde{z}} \cap W^0 \subset \operatorname{dom} \Gamma_{\tilde{z}} \cap W^0$. Since dw is a $(\mu, \operatorname{Dom} T)$-Bregman distance over W, it follows from (39) and Proposition A.1 of Appendix A with $\psi = \lambda \Gamma_{\tilde{z}}$ that z is well-defined and satisfies

$$r \in \partial \Gamma_{\tilde{z}}(z). \tag{42}$$

Clearly, the latter conclusion and the fact that $\tilde{z} \in \operatorname{dom} f_{\tilde{z}}$ imply that $\varepsilon < \infty$. Using the assumption that $\Gamma_{\tilde{z}} \leq f_{\tilde{z}}$, and relations (38) and (42), we conclude that

$$f_{\tilde{z}}(u) \geq \Gamma_{\tilde{z}}(u) \geq \Gamma_{\tilde{z}}(z) + \langle r, u - z \rangle = f_{\tilde{z}}(\tilde{z}) + \langle r, u - \tilde{z} \rangle - \varepsilon \quad \forall u \in \mathcal{Z}, \tag{43}$$

and hence that the first inclusion in (40) holds. Note that the second inclusion in (40) is due to condition C.1. Clearly, (43) with $u = \tilde{z}$ implies that $\varepsilon \geq 0$.

(b) Note that (41) and (37) clearly imply that

$$f_{\tilde{z}}(\tilde{z}) + \frac{1-\sigma}{\lambda}(dw)_{z_-}(\tilde{z}) \leq \Gamma_{\tilde{z}}(z) + \frac{1}{\lambda}(dw)_{z_-}(z). \tag{44}$$

Moreover, relations (24), (25) and (39) imply that

$$(dw)_{z_-}(\tilde{z}) - (dw)_{z_-}(z) = (dw)_z(\tilde{z}) + \langle \nabla (dw)_{z_-}(z), \tilde{z} - z \rangle = (dw)_z(\tilde{z}) - \lambda \langle r, \tilde{z} - z \rangle. \tag{45}$$

Now, using (38), (44) and (45), we conclude that

$$\begin{aligned} (dw)_z(\tilde{z}) + \lambda \varepsilon &= (dw)_z(\tilde{z}) + \lambda [f_{\tilde{z}}(\tilde{z}) - \Gamma_{\tilde{z}}(z) - \langle r, \tilde{z} - z \rangle] \\ &= (dw)_{z_-}(\tilde{z}) - (dw)_{z_-}(z) + \lambda [f_{\tilde{z}}(\tilde{z}) - \Gamma_{\tilde{z}}(z)] \leq \sigma (dw)_{z_-}(\tilde{z}), \end{aligned}$$

and hence that (35) holds. ∎

4 An accelerated instance of the NE-HPE framework

This section presents and establishes the (inner) iteration-complexity of a particular instance of the NE-HPE framework for solving the saddle-point problem where the triple $(\tilde{z}_j, z_j, \varepsilon_j)$ in step 1 of the framework is computed with the aid of the accelerated gradient method of Subsection 2.2.

Throughout this section, we assume that $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$, X, Y, Z, $\langle\cdot,\cdot\rangle_{\mathcal{X}}$, $\langle\cdot,\cdot\rangle_{\mathcal{Y}}$, $\langle\cdot,\cdot\rangle$ and $\hat{\Phi}$ are as in Subsubsection 2.1.2. Moreover, let $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$ be norms in $\mathcal{X}$ and $\mathcal{Y}$, respectively, which are not necessarily the ones induced by their corresponding inner products. Our problem of interest is the saddle-point problem $\mathrm{SP}(\hat{\Phi}; Z)$ endowed with a certain composite structure on the space $\mathcal{X}$ which consists of the existence of a proper closed convex function $\phi : \mathcal{X} \to (-\infty, +\infty]$ and a function $\Phi : \mathrm{Dom}\,\Phi \supset Z \to \mathbb{R}$ satisfying

$$\mathrm{dom}\,\phi = X, \tag{46}$$

$$\hat{\Phi}(x,y) = \Phi(x,y) + \phi(x) \quad \forall (x,y) \in Z, \tag{47}$$

and the following additional conditions:

D.1) Z is a nonempty bounded convex set;

D.2) for every $z \in Z$, the function $\hat{\Phi}_z$ given by (13) is closed and convex;

D.3) for every $y \in Y$, the function $\Phi(\cdot, y)$ is differentiable on X and there exist nonnegative constants L_{xx} and L_{xy} such that

$$\|\nabla_x \Phi(x',y') - \nabla_x \Phi(x,y)\|_{\mathcal{X}}^* \le L_{xx}\|x - x'\|_{\mathcal{X}} + L_{xy}\|y - y'\|_{\mathcal{Y}} \quad \forall (x,y),(x',y') \in X \times Y,$$

where $\|\cdot\|_{\mathcal{X}}^*$ denotes the dual norm of $\|\cdot\|_{\mathcal{X}}$ defined as

$$\|x\|_{\mathcal{X}}^* := \max_{\|x'\|_{\mathcal{X}}=1} \{\langle x, x'\rangle_{\mathcal{X}} : x' \in \mathcal{X}\} \quad \forall x \in \mathcal{X}.$$

Observe that D.2 and Proposition 2.4(b) imply that the operator $T_{\hat{\Phi}}$ given by (12) is maximal monotone.

Our goal in this section is to develop an accelerated instance of the NE-HPE framework for (approximately) solving the saddle-point problem $\mathrm{SP}(\hat{\Phi}; Z)$, or equivalently, the inclusion 14, under the above assumptions. The following definition describes the notion of approximate solution considered in our analysis.

Definition 4.1. *Given $\bar{\varepsilon} \ge 0$, a pair $(z, \varepsilon) \in Z \times \mathbb{R}_+$ satisfying $0 \in \partial_\varepsilon(\hat{\Phi}_z)(z)$ is called an $\bar{\varepsilon}$-saddle-point of $\hat{\Phi}$ with respect to Z if $\varepsilon \le \bar{\varepsilon}$, where $\hat{\Phi}_z$ is given by (13).*

We now describe the Bregman distance used by our instance. Let dw^X (resp., dw^Y) be a (η_X, X)-Bregman (resp., (η_Y, Y)-Bregman) distance over $W_X \subset \mathcal{X}$ (resp., $W_Y \subset \mathcal{Y}$). Letting $W = W_X \times W_Y$ and $W^0 = \mathrm{int}(W)$, the function dw defined as

$$(dw)_z(z') := (dw^X)_x(x') + (dw^Y)_y(y') \quad \forall z = (x,y) \in W^0,\ \forall z' = (x',y') \in W \tag{48}$$

is a Bregman distance over W.

It is assumed that $Z \subset W$ in order to ensure that the operator $T = T_{\hat{\Phi}}$ given by (12) satisfies condition B.1, and hence the results of Subsection 3.1 carry over to the present context.

To describe our instance, it suffices to explain how step 1 of the NE-HPE framework is implemented. This will be the subject of Subsection 4.1 below which describes a scheme for implementing this step based on the acceleration gradient method of Subsection 2.2. For now, we just mention that the stepsize λ_j is not chosen to be constant but rather is computed within an interval of the form $[\tau\lambda, \lambda]$ where $\lambda > 0$ and $\tau \in (0, 1)$ are fixed throughout our instance. In addition, the scheme of Subsection 4.1 also describes how to compute a triple $(\tilde{z}_j, z_j, \varepsilon_j)$ satisfying condition (27) with dw given by (48), and the stronger inclusion (33).

More specifically, Subsection 4.1 describes a scheme for solving the following problem.

(P1) *Given a pair $z_- = (x_-, y_-) \in W^0$, and scalars $\sigma \in (0, 1]$, $\lambda > 0$ and $\tau \in (0, 1)$, the problem is to find $\tilde{\lambda} \in [\tau\lambda, \lambda]$ and a triple $(\tilde{z}, z, \varepsilon) \in W \times W^0 \times \mathbb{R}_+$ such that*

$$r := \frac{1}{\tilde{\lambda}} \nabla(dw)_z(z_-) \in \partial_\varepsilon(\hat{\Phi}_{\tilde{z}})(\tilde{z}), \tag{49}$$

$$(dw)_z(\tilde{z}) + \tilde{\lambda}\varepsilon \leq \sigma(dw)_{z_-}(\tilde{z}). \tag{50}$$

with $\hat{\Phi}_{\tilde{z}}$ given by (13).

In addition to Subsection 4.1, this section contains two other subsections. Subsection 4.2 completely describes the accelerated instance of the NE-HPE framework for solving SP$(\hat{\Phi}; Z)$ and its corresponding iteration-complexity result. It also discusses optimal ways of choosing the prox stepsize in order to minimize the overall inner iteration-complexity of the instance. The proof of the main complexity result of Subsection 4.1 is only given in Subsection 4.3.

4.1 An accelerated scheme for solving (P1)

This subsection presents a scheme for finding a solution of problem (P1) based on the accelerated gradient method of Subsection 2.2 applied to a certain regularized convex-concave min-max problem.

With the above goal in mind, consider the regularized convex-concave min-max problem

$$\min_{x \in X} \max_{y \in Y} \hat{\Phi}(x, y) + \frac{1}{\lambda}(dw^X)_{x_-}(x) - \frac{1}{\lambda}(dw^Y)_{y_-}(y). \tag{51}$$

It is easy to see that the exact solution of (51) determines a solution of (P1) with $\sigma = 0$ in which $\tilde{\lambda} = \lambda$. Letting

$$f_\lambda(x) := \max_{y \in Y} \left\{ \Phi(x, y) - \frac{1}{\lambda}(dw^Y)_{y_-}(y) \right\} \quad \forall x \in X, \tag{52}$$

$$g_\lambda(x) := \frac{1}{\lambda}(dw^X)_{x_-}(x) + \phi(x) \quad \forall x \in \mathcal{X}, \tag{53}$$

it follows from (47), (52) and (53) that (51) is equivalent to (15) with $(f, g) = (f_\lambda, g_\lambda)$. Moreover, conditions A.1 and A.2 are satisfied with $\mu = \eta_X/\lambda$ due to (52) and the fact that dw^X is an (η_X, X)-Bregman distance over W_X. Also, the following result establishes the validity of A.3.

Proposition 4.2. *The constant $L = L_\lambda$ and function $\nabla f = \nabla f_\lambda : X \to X$ defined as*

$$L_\lambda := 2\left(L_{xx} + \frac{\lambda}{\eta_Y} L_{xy}^2\right), \quad \nabla f_\lambda(x) := \nabla_x \Phi(x, y_\lambda(x)) \quad \forall x \in X, \tag{54}$$

respectively, where $y_\lambda(x)$ is defined as

$$y_\lambda(x) := \operatorname{argmax}_{y \in Y} \left\{\Phi(x, y) - \frac{1}{\lambda}(dw^Y)_{y_-}(y)\right\} \quad \forall x \in X, \tag{55}$$

satisfy condition A.3 with $f = f_\lambda$.

Proof. The result follows from Proposition 4.1 of [19] with the function Ψ given by

$$\Psi(x, y) = \Phi(x, y) - \frac{1}{\lambda}(dw^Y)_{y_-}(y) \quad \forall (x, y) \in X \times Y,$$

and with $\eta = 0$ and $\beta = \eta_Y/\lambda$. ∎

Next we present a scheme for solving (P1) under the assumption that the input z_- lies in $W^0 \cap Z$. The scheme consists on applying the accelerated method of Subsection 2.2 to problem (15) with $(f, g) = (f_\lambda, g_\lambda)$ where f_λ and g_λ are as in (52) and (53), respectively.

[Algorithm 2] Accelerated scheme for solving (P1).

Input: $\sigma \in (0, 1]$, $\lambda > 0$, $\tau \in (0, 1)$ and $z_- = (x_-, y_-) \in W^0 \cap Z$.

0) Set $A_0 = 0$, $k = 1$, $\tilde{\Lambda}_0 \equiv 0$, $\tilde{y}_0 = 0$, L_λ as in (54), and $x_0 = \tilde{x}_0 := x_-$;

1) compute A_k as in (16) with $\mu = \eta_X/\lambda$, iterate $\breve{x}_k$ as in (17), compute $y_\lambda(\breve{x}_k)$ according to (55), and the affine function $\tilde{\Lambda}_k$ as

$$\tilde{\Lambda}_k := \frac{A_{k-1}}{A_k}\tilde{\Lambda}_{k-1} + \frac{A_k - A_{k-1}}{A_k}\left[\Phi(\breve{x}_k, y_\lambda(\breve{x}_k)) + \langle \nabla\Phi(\breve{x}_k, y_\lambda(\breve{x}_k)), \cdot - \breve{x}_k\rangle_{\mathcal{X}}\right] \tag{56}$$

2) set

$$\lambda_k = \left(\frac{1}{\lambda} + \frac{1}{\eta_X A_k}\right)^{-1}, \tag{57}$$

and compute iterates x_k and $\tilde{y}_k$ as

$$x_k = \operatorname{argmin}\left\{\tilde{\Lambda}_k(x) + \phi(x) + \frac{1}{\lambda_k}(dw^X)_{x_-}(x)\right\}, \tag{58}$$

$$\tilde{y}_k = \frac{A_{k-1}}{A_k}\tilde{y}_{k-1} + \frac{A_k - A_{k-1}}{A_k} y_\lambda(\breve{x}_k), \tag{59}$$

and $\tilde{x}_k$ as in (20);

3) if $\lambda_k \geq \max\{1-\sigma, \tau\}\lambda$, then compute $y_k := y_{\lambda_k}(\tilde{x}_k)$ according to (55), set $\tilde{\lambda} = \lambda_k$, $\tilde{z} = \tilde{z}_k := (\tilde{x}_k, \tilde{y}_k)$, $z = z_k := (x_k, y_k)$ and

$$\varepsilon = \varepsilon_k := \hat{\Phi}(\tilde{x}_k, y_k) - \tilde{\Lambda}_k(x_k) - \phi(x_k) - \frac{1}{\lambda_k} \langle \nabla(dw)_z(z_-), \tilde{z} - z \rangle ,$$

output $\tilde{\lambda}$ and the triple $(\tilde{z}, z, \varepsilon)$, and terminate; otherwise, set $k \leftarrow k+1$ and go to step 1.

end

We now make several remarks about Algorithm 2. First, due to the stopping criterion and (57), Algorithm 2 outputs $\tilde{\lambda} \in [\tau\lambda, \lambda]$. Second, due to Proposition A.1 and relations (46), (55) and (58), the output z lies in $W^0 \cap Z$. Third, steps 1 and 2 of Algorithm 2 are specializations of steps 1 and 2 of Algorithm 1 to the instance of (15) in which (f, g) is given by (f_λ, g_λ) with f_λ and g_λ as in (52) and (53), respectively. The only difference is the extra computation of $\tilde{y}_k$ in (59) which is used to compute the component $\tilde{z}$ of the output. Fourth, even though the affine function $\tilde{\Lambda}_k$ given in (56) and the affine function Λ_k given by (18) with $f = f_\lambda$ are not the same, they both have the same gradient due to (54), and hence the subproblems (58) and (19) are equivalent. Fifth, each iteration of Algorithm 2 before the last one requires solving two subproblems, namely, (58) and one of the form (55), while the last one requires one additional subproblem of the form (55) in step 3. Sixth, when the termination criterion in step 3 is met, this extra step computes the output $\tilde{\lambda}$ and $(\tilde{z}, z, \varepsilon)$ which solve (P1) (see Proposition 4.3 below). Seventh, another possible way to terminate Algorithm 2 would be to compute the triple $(\tilde{z}, z, \varepsilon) = (\tilde{z}_k, z_k, \varepsilon_k)$ as described in its step 3 at every iteration and check whether $\tilde{\lambda} = \lambda_k$ and this triple satisfy the HPE error criterion (50). (They always satisfy (49) due to Proposition 4.3(a).) The drawback of this stopping criterion is that it requires solving an additional subproblem of the form (55) at every iteration. Our computational benchmark presented in Section 5 is based on the stopping criterion of Algorithm 2.

The following result establishes the correctness and iteration-complexity of Algorithm 2. Its proof is given in Subsection 4.3.

Proposition 4.3. *For every $k \geq 1$ the following statements hold:*

(a) the scalar $\tilde{\lambda} = \lambda_k$ and the triple $(\tilde{z}, z, \varepsilon) = (\tilde{z}_k, z_k, \varepsilon_k)$ satisfy inclusion (49)*;*

(b) If $\lambda_k \geq (1-\sigma)\lambda$, then $\tilde{\lambda} = \lambda_k$ and the triple $(\tilde{z}, z, \varepsilon) = (\tilde{z}_k, z_k, \varepsilon_k)$ satisfy the condition (50)*.*

Also, Algorithm 2 solves problem (P1) in at most

$$\mathcal{O}\left(\left\lceil \sqrt{\frac{\lambda\left(L_{xx} + \frac{\lambda}{\eta_Y} L_{xy}^2\right)}{\eta_X}} \right\rceil\right) \tag{60}$$

iterations.

4.2 An accelerated NE-HPE instance for solving SP$(\hat{\Phi}; Z)$

This subsection describes an accelerated instance of NE-HPE framework for solving the saddle-point problem SP$(\hat{\Phi}; Z)$ and its corresponding iteration-complexity result. It also discusses optimal ways of choosing the prox stepsize in order to minimize the overall inner iteration-complexity of the instance.

We start by stating an accelerated instance of the NE-HPE framework for solving SP$(\hat{\Phi}; Z)$ which computes the required stepsize λ_j and triple $(\tilde{z}_j, z_j, \varepsilon_j)$ in its step 1 with the aid of Algorithm 2.

Accelerated NE-HPE method for the saddle-point problem

0) Let $z_0 \in W^0$, $\lambda > 0$, $\sigma \in (0, 1]$ and $\tau \in (0, 1)$ be given and set $j = 1$;

1) invoke Algorithm 2 with input σ, λ, τ and $z_- = z_{j-1}$ to obtain a stepsize λ_j and a triple $(\tilde{z}_j, z_j, \varepsilon_j)$ satisfying (33) and (27);

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

end

In view of Proposition 4.3, the accelerated NE-HPE method satisfies the error conditions (33) and (27) of step 1 of the NE-HPE framework. Therefore, the accelerated NE-HPE method is clearly a special case of the NE-HPE framework. It follows that the ergodic (outer) convergence rate bound for the accelerated NE-HPE method is as described in Theorem 3.5.

Theorem 4.4. *let $R = R(z_0; Z)$ be given by (30). Consider the sequences $\{\tilde{z}_j\}$, $\{z_j\}$ and $\{\varepsilon_j\}$ generated by the accelerated NE-HPE method and the respective ergodic sequences $\{\tilde{z}_j^a\}$, $\{z_j^a\}$ and $\{\varepsilon_j^a\}$ as in Lemma 3.4. Then, the following statements hold:*

(a) for every positive scalar $\bar{\varepsilon}$, there exists

$$j_0 = \mathcal{O}\left(\left\lceil \frac{R}{\lambda \bar{\varepsilon}} \right\rceil\right)$$

such that for every $j \geq j_0$, $(\tilde{z}_j^a, \varepsilon_j^a)$ is a $\bar{\varepsilon}$-saddle-point of $\hat{\Phi}$ with respect to Z;

(b) each iteration of the accelerated NE-HPE method performs at most

$$\mathcal{O}\left(\left\lceil \sqrt{\frac{\lambda\left(L_{xx} + \frac{\lambda}{\eta_Y} L_{xy}^2\right)}{\eta_X}} \right\rceil\right)$$

inner iterations.

As a consequence, the accelerated NE-HPE method finds an $\bar{\varepsilon}$-saddle-point of $\hat{\Phi}$ with respect to Z by performing no more than

$$\mathcal{O}\left(\left\lceil\sqrt{\frac{\lambda\left(L_{xx}+\frac{\lambda}{\eta_Y}L_{xy}^2\right)}{\eta_X}}\right\rceil\left\lceil\frac{R}{\lambda\bar{\varepsilon}}\right\rceil\right) \tag{61}$$

inner iterations.

Proof. Since the accelerated NE-HPE method is a special instance of the NE-HPE framework, (a) follows from Theorem 3.5 (a) and from the fact that $\lambda_j \geq \tau\lambda$ for every $j \geq 1$. Statement (b) follows from Proposition 4.3. The last assertion of the theorem follows immediately from (a) and (b). ∎

We end this subsection by making a remark about the complexity bound (61) in light of the one obtained in relation (4.4) of [23]. Clearly, when $\lambda = R/\bar{\varepsilon}$, the complexity bound (61) reduces to

$$\mathcal{O}\left(1+\frac{RL_{xy}}{\bar{\varepsilon}\sqrt{\eta_X\eta_Y}}+\sqrt{\frac{RL_{xx}}{\bar{\varepsilon}\eta_X}}\right). \tag{62}$$

It turns out that, for suitably chosen scaled Bregman distances with respect to X and Y, this bound reduces to

$$\mathcal{O}\left(1+\frac{\sqrt{R_XR_Y}L_{xy}}{\bar{\varepsilon}\sqrt{\eta_X\eta_Y}}+\sqrt{\frac{R_XL_{xx}}{\bar{\varepsilon}\eta_X}}\right), \tag{63}$$

where

$$R_X := \max\{(dw^X)_{x_0}(x) : x \in X\}, \quad R_Y := \max\{(dw^Y)_{y_0}(y) : y \in Y\}.$$

The latter bound generalizes the one in relation (4.4) of [23] which is valid only for a special bilinear structured case of SP($\hat{\Phi}; Z$).

To obtain the bound (63), consider the Bregman distances defined as

$$dw^{X,\theta} := \theta dw^X, \quad dw^{Y,\theta} := \theta^{-1}dw^Y$$

where $\theta > 0$ is a fixed parameter. Clearly, $dw^{X,\theta}$ (resp., $dw^{Y,\theta}$) is a $(\theta\eta_X, X)$-Bregman distance (resp., $(\theta^{-1}\eta_Y, Y)$-Bregman distance) over W_X (resp., over W_Y). In this case, R becomes

$$R = R_\theta := \theta R_X + \theta^{-1}R_Y.$$

Hence, chosing $\theta = (R_Y/R_X)^{1/2}$, the quantities R, η_X and η_Y in this case reduce to

$$R = 2\sqrt{R_XR_Y}, \quad \eta_X = \sqrt{\frac{R_Y}{R_X}}\eta_X, \quad \eta_Y = \sqrt{\frac{R_X}{R_Y}}\eta_Y,$$

and hence (62) reduces to (63).

4.3 Proof of Proposition 4.3

This subsection proves Proposition 4.3.

We start by establishing the following technical result.

Lemma 4.5. *For every $k \geq 1$, the affine function $\tilde{\Lambda}_k$ given by (56) satisfies the following statements:*

(a) $\tilde{\Lambda}_k(x) \leq \Phi(x, \tilde{y}_k)$ *for every* $x \in X$*;*

(b) setting $\tilde{z}_k := (\tilde{x}_k, \tilde{y}_k)$*, the function* $\Gamma_{\tilde{z}_k} : \mathcal{Z} \to [-\infty, \infty]$ *defined as*

$$\Gamma_{\tilde{z}_k}(z) := \begin{cases} \tilde{\Lambda}_k(x) + \phi(x) - \hat{\Phi}(\tilde{x}_k, y), & \forall z = (x, y) \in Z; \\ +\infty, & \text{otherwise.} \end{cases} \tag{64}$$

minorizes the function $\hat{\Phi}_{\tilde{z}_k}$ *defined in* (13)*;*

(c) if $\lambda_k \geq (1 - \sigma)\lambda$ *then*

$$\frac{(1-\sigma)}{\lambda_k}(dw)_{(z_-)}(\tilde{z}_k) \leq \inf_{u \in \mathcal{Z}} \left\{ \Gamma_{\tilde{z}_k}(u) + \frac{1}{\lambda_k}(dw)_{(z_-)}(u) \right\}. \tag{65}$$

Proof. (a) Using the definitions of $\tilde{\Lambda}_k$ and $\tilde{y}_k$ given in (56) and (58) as well as the fact that $\Phi(\cdot, y) - \Phi(x, \cdot)$ is convex for every $(x, y) \in Z$, we see that for every $x \in X$

$$\begin{aligned} \tilde{\Lambda}_k(x) &= \sum_{i=1}^{k} \frac{A_i - A_{i-1}}{A_k} \left[\Phi(\breve{x}_i, y_\lambda(\breve{x}_i)) + \langle \nabla \Phi(\breve{x}_i, y_\lambda(\breve{x}_i)), x - \breve{x}_i \rangle_{\mathcal{X}} \right] \\ &\leq \sum_{i=1}^{k} \frac{A_i - A_{i-1}}{A_k} \left[\Phi(x, y_\lambda(\breve{x}_i)) \right] \leq \Phi\left(x, \sum_{i=1}^{k} \frac{A_i - A_{i-1}}{A_k} y_\lambda(\breve{x}_i) \right) = \Phi(x, \tilde{y}_k), \end{aligned}$$

which proves (a).

(b) By (13), (47) and (64), we see that $\Gamma_{\tilde{z}_k}$ minorizes $\hat{\Phi}_{\tilde{z}_k}$ if and only if $\tilde{\Lambda}_k \leq \Phi(\cdot, \tilde{y}_k)$, and hence (b) follows.

(c) Assume that $\lambda_k \geq (1 - \sigma)\lambda$. Observe that due to (52), (54), (55), (56), (59) and the convexity of $(dw^Y)_{y_-}(\cdot)$, we have that

$$\begin{aligned} &\tilde{\Lambda}_k(x) - \frac{1}{\lambda}(dw^Y)_{y_-}(\tilde{y}_k) \\ &\geq \sum_{i=1}^{k} \frac{A_i - A_{i-1}}{A_i} \left[\Phi(\breve{x}_i, y_\lambda(\breve{x}_i)) + \langle \nabla \Phi(\breve{x}_i, y_\lambda(\breve{x}_i)), x - \breve{x}_i \rangle_{\mathcal{X}} - \frac{1}{\lambda}(dw^Y)_{y_-}(y_\lambda(\breve{x}_i)) \right] \\ &= \sum_{i=1}^{k} \frac{A_i - A_{i-1}}{A_i} \left[f_\lambda(\breve{x}_i) + \langle \nabla f_\lambda(\breve{x}_i), x - \breve{x}_i \rangle_{\mathcal{X}} \right]. \end{aligned} \tag{66}$$

Now, letting $x_0 = x_-$, $g = g_\lambda$ and $h = (1/\eta_X)dw^X_{x_-}$ where g_λ is as in (53), and using the fact that dw^X is an (η_X, X)-Bregman distance over W, we easily see that x_0, g and h satisfy conditions A.4-A.6. Since the set X and the functional pair $(f, g) = (f_\lambda, g_\lambda)$ satisfy conditions A.1-A.3 where f_λ is as in (52), and Algorithm 2 corresponds to Algorithm 1 applied to (15) with $(f, g) = (f_\lambda, g_\lambda)$ and h as above, it follows from (22), (52), (53) (57) and (66) that

$$\begin{aligned}
&\tilde{\Lambda}_k(x) + \phi(x) - \frac{1}{\lambda}(dw^Y)_{y_-}(\tilde{y}_k) + \frac{1}{\lambda_k}(dw^X)_{x_-}(x) \\
&= \tilde{\Lambda}_k(x) + \phi(x) - \frac{1}{\lambda}(dw^Y)_{y_-}(\tilde{y}_k) + \left(\frac{1}{\lambda} + \frac{1}{\eta_X A_k}\right)(dw^X)_{x_-}(x) \\
&\geq \sum_{i=1}^{k} \frac{A_i - A_{i-1}}{A_i}\left[f_\lambda(\breve{x}_i) + \langle \nabla f_\lambda(\breve{x}_i), x - \breve{x}_i\rangle_{\mathcal{X}}\right] + g_\lambda(x) + \frac{1}{A_k}h(x) \\
&\geq (f_\lambda + g_\lambda)(\tilde{x}_k) \geq \hat{\Phi}(\tilde{x}_k, y) - \frac{1}{\lambda}(dw^Y)_{y_-}(y) + \frac{1}{\lambda}(dw^X)_{x_-}(\tilde{x}_k) \quad \forall (x, y) \in Z.
\end{aligned}$$

Now, rearranging the last inequality and using the definitions of dw and $\Gamma_{\tilde{z}_k}$ in (48) and (64), respectively, and the fact that $(1-\sigma)\lambda \leq \lambda_k \leq \lambda$ where the second inequality is due to (57), we easily see that (65) holds. ∎

We are now ready to prove Proposition 4.3.

Proof. [**Proof of Proposition 4.3**] Let $k \geq 1$ be given. The proofs of the two statements are based on Proposition 3.7 specialized to $\lambda = \lambda_k$, $\tilde{z} = \tilde{z}_k$, $\varepsilon = \varepsilon_k$ and operator $T = T_{\hat{\Phi}}$ given by (12), which satisfies condition C.1 with $f_{\tilde{z}} = \hat{\Phi}_{\tilde{z}}$ given by (13) for every $\tilde{z} \in Z$ (see the remark preceding Proposition 3.7).

(a) Lemma 4.5(b) implies that $\Gamma_{\tilde{z}_k}$ defined in (64) minorizes $\hat{\Phi}_{\tilde{z}_k}$. Moreover, using the fact that $f_{\tilde{z}}(\tilde{z}) = \hat{\Phi}_{\tilde{z}}(\tilde{z}) = 0$ for every $\tilde{z} \in Z$ due to definition (13), it is easy to see that the quantities z and ε computed according to (37) and (38), respectively, with $\tilde{z} = \tilde{z}_k$ is equivalent to the way z_k and ε_k are computed in Algorithm 2. Hence, (a) follows from Proposition 3.7(a).

(b) This statement follows from the same arguments above, Lemma 4.5(c) and Proposition 3.7(b).

We now establish the last conclusion of the proposition. When the termination of Algorithm 2 holds (see Step 3), it easily follows from (57) that the output $\tilde{\lambda} = \lambda_k$ satisfies $\tilde{\lambda} \in [\tau\lambda, \lambda]$. Hence, in view of statements (a) and (b), we conclude that Algorithm 2 outputs $\tilde{\lambda}$ and $(\tilde{z}, z, \varepsilon)$ which solves (P1). Finally, using the the estimate $A_k \geq k^2/4L_\lambda$ given in (21), and the definitions of L_λ and λ_k given in (54) and (57), respectively, it is easy to verify that the number of iterations until the stopping criterion of Algorithm 2 occurs is bounded by (60) (when τ and σ are viewed as universal constants such that $\max\{1-\sigma, \tau\}$ is not close to either zero or one). ∎

5 Numerical experiments

This section presents computational results showing the numerical performance of the accelerated NE-HPE method on a collection of saddle-point problems. All the computational results

were obtained using MATLAB R2014a on a Windows 64 bit machine with processor Intel 2.16 GHz with 4 GB memory.

The accelerated NE-HPE method (referred to as ACC-HPE) is compared with Nesterov's smoothing scheme [23] (referred to as NEST). We have implemented both algorithms based on the Euclidean distance and the Bregman distance induced by the Kulback-Leibler divergence, namely, $dw_{z^2}(z^1) = \sum_i z_i^1 \log(z_i^1/z_i^2) + z_i^1 - z_i^2$. Our computational results then consider four variants, namely, E-ACC-HPE, L-ACC-HPE, E-NEST and L-NEST, where the ones starting with E- (resp., L-) are the ones based on the Euclidean (resp., Kulback-Leibler log distance). To improve the performance of the L-variants, we have used the adaptive scheme for choosing the parameter L given in [31], i.e., the initial value of L is set to a fraction of the true Lipschitz constant value and is increased by a factor of 2 whenever it fails to satisfy a certain convergence criterion (see equations (23) and (45) of [31]). The fraction $1/2^9$ was used in our experiments. The same scheme was not used for the E-variants since we have observed that it does not improve their performance. The value of L at the last iteration divided by the true Lipschitz constant varied between $1/64$ and 1 in our experiments. More specifically, this ratio was $1/64$ for one instance, $1/32$ for three instances, $1/8$ for one instance, $1/4$ for four instances, $1/2$ for thirteen instances and 1 for the remaining instances.

The following three subsections report computational results on the following classes of problems: (a) zero-sum matrix game; (b) vector-matrix saddle-point; and (c) quadratic game. The results are reported in tables and in performance profiles (see [9]). We recall the following definition of a performance profile. For a given instance, a method A is said to be at most x times slower than method B, if the time taken by method A is at most x times the time taken by method B. A point (x, y) is in the performance profile curve of a method if it can solve exactly 100% of all the tested instances x times slower than any other competing method.

For all problem classes, the stopping criterion used to terminate all methods at the k-th iteration is

$$\max_{y\in Y} \hat{\Phi}(\tilde{x}_k, y) - \min_{x\in X} \hat{\Phi}(x, \tilde{y}_k) \leq \bar{\varepsilon}.$$

The use of this criterion for the second and third problem classes is not the best strategy from the computational point of view, since the computation of the dual function involves solving a quadratic programming problem over the unit simplex. Note that our method has the ability to compute at every iteration a pair $((\tilde{x}_k, \tilde{y}_k), \varepsilon_k)$ such that the above inequality holds with $\bar{\varepsilon} = \varepsilon_k$ and hence the above termination criterion will be satisfied whenever $\varepsilon_k \leq \bar{\varepsilon}$. Since the usual description of Nesterov's smoothing scheme generates $(\tilde{x}_k, \tilde{y}_k)$ but not ε_k, we have opted for the gap criterion but adopted the convention of excluding the effort to evaluate the dual functions from the reported cpu times.

We let $\mathbb{R}^n$ denote the n-dimensional Euclidean space and $\mathcal{S}^n$ denote the linear space of $n \times n$ real symmetric matrices. The unit simplex in $\mathbb{R}^n$ is defined as

$$\Delta_n := \left\{ x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1, x \geq 0 \right\}.$$

5.1 Zero-sum matrix game problem

This subsection compares the performance of the four variants on instances of the zero-sum matrix game problem

$$\min_{x\in\Delta_n}\max_{y\in\Delta_m}\langle x, Ay\rangle$$

where A is a real $n \times m$ matrix. The matrices were generated so that its elements are non-zero with probability p and the nonzero ones are randomly generated in the interval $[-1, 1]$. We have tested the methods for a set of problems with different sizes of matrices and different values of p. The tolerance used here was $\bar{\varepsilon} = 10^{-3}$.

Table 1 reports the results of the four variants applied to several instances of this problem with different sizes of matrices and different values of p.

[Table 1 near here]

Figure 1 gives the performance profile for the same set of instances. Overall, it shows that the accelerated NE-HPE variants perform better than NEST variants on this set of zero-sum games instances.

[Figure 1 near here]

5.2 Vector-matrix saddle-point problem

This subsection compares the four variants on instances of the vector-matrix saddle-point problem. Given $c \in \mathbb{R}^n$, a real $n \times n$ matrix B and a linear operator $\mathcal{A} : \mathbb{R}^n \to \mathcal{S}^m$, the vector-matrix saddle-point problem is

$$\min_{x\in\Delta_n}\frac{1}{2}\|Bx + c\|^2 + \theta_{\max}(\mathcal{A}(x))$$

where $\theta_{\max}(\mathcal{A}(x))$ denotes the largest eigenvalue of $\mathcal{A}(x)$. Such problem is equivalent to the saddle-point problem

$$\min_{x\in\Delta_n}\max_{y\in\Omega}\frac{1}{2}\|Bx + c\|^2 + \langle\mathcal{A}(x), y\rangle,$$

where $\Omega := \{y \in \mathcal{S}^m : \mathrm{tr}(y) = 1, y \text{ is positive definite}\}$. We have tested the four variants on a set of problems where the matrices B and $\mathcal{A}_i := \mathcal{A}(e_i)$, $i = 1, \ldots, n$, were generated so that its elements are non-zero with probability 0.1 and the non-zero ones are randomly generated in the interval $[-1, 1]$. (Here, e_i denotes the i-th unit n dimensional vector.) The tolerance used was $\bar{\varepsilon} = 10^{-2}$.

Table 2 reports the results of the four variants applied to several instances of this problem with different sizes of matrices.

[Table 2 near here]

Figure 2 gives the performance profile for the same set of instances. It also shows that the accelerated NE-HPE variants perform better than NEST variants on this set of vector-matrix saddle-point instances.

[Figure 2 near here]

5.3 Quadratic game problem

This subsection compares the four variants on instances of the quadratic game problem

$$\min_{x\in\Delta_n}\max_{y\in\Delta_m}\frac{1}{2}\|Bx\|^2+\langle x,Ay\rangle$$

for different sizes of matrices and different values of p. The matrices were generated in the same way as in the zero-sum matrix game problem (see Subsection 5.1). The tolerance used was $\bar{\varepsilon}=10^{-4}$.

Table 3 reports the results of the four variants applied to several instances of this problem with different sizes of matrices and different values of p.

[Table 3 near here]

Figure 3 gives the performance profile for the same set of instances. It shows that the accelerated NE-HPE variant based on the Euclidean (resp., Kulback-Leibler log) distance performs better than the NEST variant based on the Euclidean (resp., Kulback-Leibler log) distance on this set of quadratic game instances.

[Figure 3 near here]

5.4 Concluding Remarks

In this subsection, we make some final remarks about the computational results described in this section. We have shown in Subsection 4.2 that the accelerated NE-HPE method has the same complexity as the Nesterov's smoothing technique of [23]. The experiment results of this section involving three problem sets have shown that the accelerated NE-HPE variants outperform the variants of Nesterov's smoothing scheme. The experiments have also shown that the accelerated NE-HPE variant based on the Euclidean distance performs better than the accelerated NE-HPE variant based on the Kulback-Leibler log distance.

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] 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.

[3] BURACHIK, R. S., SAGASTIZÁBAL, C. A., AND SVAITER, B. F. ϵ-enlargements of maximal monotone operators: theory and applications. In *Reformulation: nonsmooth, piecewise smooth, semismooth and smoothing methods (Lausanne, 1997)*, vol. 22 of *Appl. Optim.* Kluwer Acad. Publ., Dordrecht, 1999, pp. 25–43.

[4] CENSOR, Y., AND ZENIOS, S. A. Proximal minimization algorithm with D-functions. *Journal of Optimization Theory and Applications 73* (1992).

[5] CHEN, G., AND TEBOULLE, M. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization 3* (1993), 538–543.

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

[7] CHEN, Y., LAN, G., AND OUYANG, Y. Accelerated schemes for a class of variational inequalities. *submitted to Mathematical Programming* (2014).

[8] DANG, C. D., AND LAN, G. On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators. *Computational Optimization and applications 60*, 2 (2015), 277–310.

[9] DOLAN, E. D., AND MORÉ, J. J. Benchmarking optimization software with performance profiles. *Mathematical Programming 91*, 2 (2002), 201–213.

[10] ECKSTEIN, J. Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. *Mathematics of Operations Research 18*, 1 (1993).

[11] EGGERMONT, P. P. B. Multiplicative iterative algorithms for convex programming. *Linear Algebra and Applications 130* (1990), 25–32.

[12] 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).

[13] HE, Y., AND MONTEIRO, R. D. C. An accelerated HPE-type algorithm for a class of composite convex-concave saddle-point problems. *working paper* (2014).

[14] HIRRIART-URRUTY, J., AND LEMARECHAL, C. Convex analysys and minimization algorithms i. *Comprehensive Study in Mathematics 305* (1993).

[15] IUSEM, A. N., AND SOLODOV, M. V. Newton-type methods with generalized distances for constrained optimization. *Optimization 41* (1997), 257 278.

[16] KIWIEL, K. Proximal minimization methods with generalized Bregman functions. *SIAM Journal on Control Optimization 35* (1997), 11421168.

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

[18] 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.

[19] MONTEIRO, R. D. C., AND SVAITER, B. F. Convergence rate of inexact proximal point methods with relative error criteria for convex optimization. *submitted to SIAM Journal on Optimization* (2010).

[20] 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.

[21] 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.

[22] 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.

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

[24] ROCKAFELLAR, R. T. Convex analysys. *Princeton Math. Series 28* (1970).

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

[26] 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.

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

[28] 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.

[29] 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.

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

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

A Appendix

This appendix presents two existence/uniqueness results about solutions of certain regularized convex minimization and/or monotone inclusion problems.

We begin by stating without proof a well-known result about regularized convex minimization problems.

Proposition A.1. *Let $\psi : \mathcal{Z} \to [-\infty, \infty]$ be a proper closed convex function and, for some $\mu > 0$, assume that dw is a $(\mu, \operatorname{dom}\psi)$-Bregman distance over W (see Definition 3.6). Then, for any $z_- \in W^0 \cap \operatorname{dom}\psi$, the problem*

$$\inf\{\psi(u) + (dw)_{z_-}(u) : u \in \mathcal{Z}\}$$

has an unique optimal solution z, which necessarily lies in $W^0 \cap \operatorname{dom}\psi$. Moreover, z is the unique zero of the inclusion $\nabla w(z_-) \in (\partial\psi + \partial w)(z)$.

The next result generalizes Proposition A.1 to the context of regularized monotone operators.

Proposition A.2. *Let $T : \mathcal{Z} \rightrightarrows \mathcal{Z}$ be a maximal monotone operator and, for some $\mu > 0$, assume that dw is a $(\mu, \operatorname{Dom} T)$-Bregman distance over W (see Definition 3.6). Then, for every $z' \in \mathcal{Z}$, the inclusion*

$$z' \in (T + \partial w)(z)$$

has an unique solution z (which necessarily lies on $W^0 \cap \operatorname{Dom} T$ due to Definition 3.1(i)).

Proof. Define $\bar{w} := w + \mathcal{I}_Z$ where $Z := \operatorname{cl}(\operatorname{Dom} T)$. We claim that $\bar{w}$ is a proper closed μ-strongly convex function such that

$$\operatorname{Dom} \partial\bar{w} = Z \cap W^0, \quad T + \partial w = T + \partial\bar{w}.$$

Indeed, in view of Proposition 6.4.1 of [1], the set Z, and hence the indicator function $\mathcal{I}_Z$, is closed convex. This conclusion with Definition 3.6 imply that $\bar{w}$ is a proper closed μ-strongly convex function. Since the assumption that dw is a $(\mu, \operatorname{Dom} T)$-Bregman distance over W implies that $W^0 \cap \operatorname{Dom} T \neq \emptyset$ and W^0 is open, it is straightforward to see that $W^0 \cap \operatorname{ri} Z = W^0 \cap \operatorname{ri}(\operatorname{Dom} T) \neq \emptyset$, and hence that the relative interiors of the domains of the convex functions w and $\mathcal{I}_Z$ intersect. This conclusion together with Theorem 23.8 of [24] then imply that $\partial\bar{w} = \partial(w + \mathcal{I}_Z) = \partial w + \partial\mathcal{I}_Z$, and hence that $\operatorname{Dom} \partial\bar{w} = \operatorname{Dom} \partial\mathcal{I}_Z \cap \operatorname{Dom} \partial w = Z \cap W^0$. Now using the assumption that T is maximal, Proposition 2.2 and the latter conclusion, we conclude that

$$T + \partial\bar{w} = T + (\partial w + \partial\mathcal{I}_Z) = (T + \partial\mathcal{I}_Z) + \partial w = T + \partial w,$$

and hence that the claim holds.

We next establish the conclusion of the Lemma. By changing μ if necessary, we may assume without any loss of generality that $\|\cdot\|$ is the norm associated with the inner product $\langle \cdot, \cdot \rangle$. Since $\bar{w}$ is a proper closed μ-strongly convex function, it follows from the above claim and Proposition 1.12 in Chapter IV of [14] that $\bar{w}_0 := \bar{w} - \mu\|\cdot\|^2/2$ is a proper closed convex function. Now, define $\bar{T} := T + \partial\bar{w}_0$ and note that

$$T + \partial w = T + \partial\bar{w} = \bar{T} + \mu I,$$

due to the above claim and the fact that $\partial\bar{w} = \partial\bar{w}_0 + \mu I$. Hence, the conclusion of the Lemma will follow from Minty's theorem (e.g., see Theorem 6.2.2 of [1]) if we show that $\bar{T}$ is maximal monotone. Indeed, first note that the above claim implies that

$$\operatorname{ri}(\operatorname{Dom} \partial\bar{w}) = \operatorname{ri}(Z \cap W^0) = \operatorname{ri} Z \cap W^0$$

due to the fact that the latter set is nonempty. Since both T and $\partial\bar{w}_0$ are maximal monotone (the second one due to Theorem 6.3.1 of [1]) and the intersection of the relative interior of their domains is clearly equal to $\operatorname{ri} Z \cap W^0 \neq \emptyset$, it follows from Theorem 6.5.6 of [1] that $\bar{T}$ is maximal monotone. ∎

A.1 List of Tables

Table 1 – Test results for the zero-sum matrix game problem

Size			E-ACC-HPE		E-NEST		L-ACC-HPE		L-NEST	
n	m	p	time	iter.	time	iter.	time	iter.	time	iter.
1000	100	0.01	0.2721	196	1.8915	1806	6.0765	4364	6.0991	4979
1000	100	0.1	0.6587	480	14.5410	12738	8.88550	5734	7.7157	5808
1000	1000	0.01	0.4562	224	3.9617	2560	7.1988	1245	30.0177	2287
1000	1000	0.1	0.7927	213	47.3204	14602	12.2948	2723	17.1655	4603
1000	10000	0.01	0.7082	100	28.0016	4213	12.9047	1799	83.6426	5098
1000	10000	0.1	3.5575	196	698.2147	38410	67.6110	2541	259.4104	5306
10000	100	0.01	1.8140	461	33.8041	9606	36.3373	7010	33.5731	7245
10000	100	0.1	11.5663	1381	745.7268	100384	48.0043	7671	47.5234	7740
10000	1000	0.01	0.7976	121	34.0278	5038	62.5932	6572	67.1911	7608
10000	1000	0.1	8.1079	287	1566.3	56744	165.6846	7449	161.6192	7816

Table 2 – Test results for the vector-matrix saddle-point game problem

Size		E-ACC-HPE		E-NEST		L-ACC-HPE		L-NEST	
n	m	time	iter.	time	iter.	time	iter.	time	iter.
50	50	1.5987	266	3.9912	750	1.2079	240	1.7442	300
50	100	10.4205	456	44.2678	2335	3.7254	235	7.9121	410
50	200	56.4287	765	510.9588	7560	16.7225	265	24.5651	310
100	50	0.6970	113	4.4376	725	2.2108	315	3.1361	460
100	100	3.6108	218	32.5960	1855	5.9332	395	8.2586	480
100	200	22.6149	316	370.1234	5145	26.1798	330	43.8539	350
200	50	2.2427	214	4.3349	605	2.1502	305	3.2779	450
200	100	3.9892	281	30.1342	1530	7.4743	345	8.7948	475
200	200	22.9626	312	342.4116	4335	23.9602	355	31.2051	495

Table 3 – Test results for the quadratic game problem

Size			E-ACC-HPE		E-Nest		L-ACC-HPE		L-Nest	
n	m	p	time	iter.	time	iter.	time	iter.	time	iter.
100	100	0.01	2.9931	3325	3.4325	3865	15.5991	16485	24.5904	29115
100	1000	0.01	1.5389	1375	22.0383	19595	33.8117	30060	42.1352	39040
1000	100	0.01	6.1160	4535	4.3215	2675	21.3346	12010	34.0820	20285
1000	1000	0.01	1.1044	595	8.8350	4960	22.6147	9230	53.5389	23255
100	100	0.1	4.4928	4985	8.6743	10345	10.9572	13125	12.0119	15415
100	1000	0.1	6.7362	5040	39.7344	31185	15.3777	12915	18.1534	15010
1000	100	0.1	42.2032	15105	41.8742	15310	47.2913	16540	52.9498	18810
1000	1000	0.1	116.7657	22820	118.2056	22850	50.3541	9450	84.0200	16820
100	100	0.2	7.7375	8585	14.9389	17815	18.8705	22605	20.6869	26545
100	1000	0.2	11.6011	8675	68.4306	39355	26.4835	22240	31.2638	25845
1000	100	0.2	28.8739	26015	72.1152	26415	81.4451	16275	91.1901	28970
1000	1000	0.2	72.6806	39305	203.5757	53710	86.7198	28490	144.6992	32390

A.2 List of Figures

Figure 1 – Performance profile for the zero-sum matrix problem

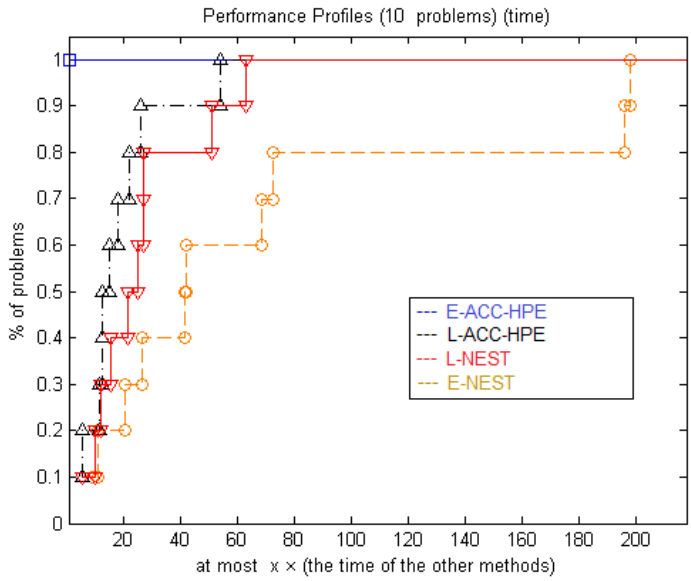

Figure 2 – Performance profile for the vector-matrix saddle-point problem

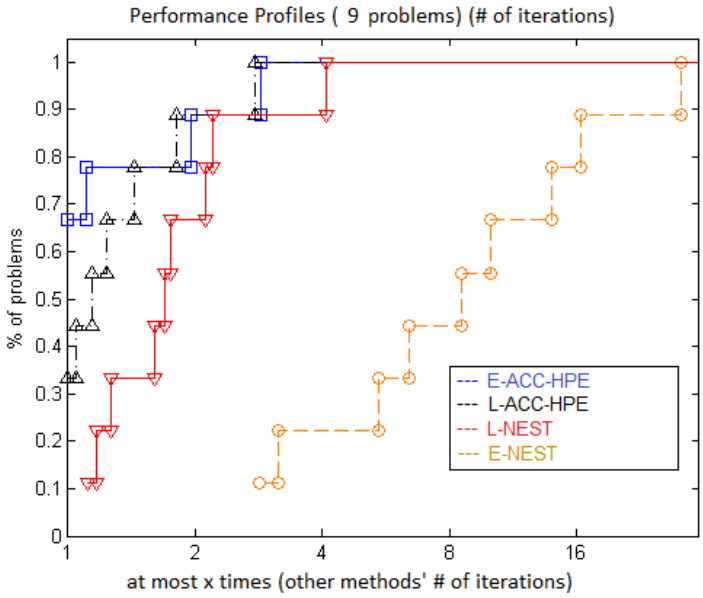

Figure 3 – Performance profile for the quadratic game problem