Journal of Computational Mathematics
Vol.xx, No.x, 2023, 1–21.
http://www.global-sci.org/jcm
doi:10.4208/jcm.2301-m2022-0099

SEMI-PROXIMAL POINT METHOD FOR NONSMOOTH CONVEX-CONCAVE MINIMAX OPTIMIZATION*

Yuhong Dai
LSEC, ICMSEC, AMSS, Chinese Academy of Sciences, Beijing 100190, China
School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
Email: dyh@lsec.cc.ac.cn
Jiani Wang[1)]
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
Email: wjiani@lsec.cc.ac.cn
Liwei Zhang
School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China
Email: lwzhang@dlut.edu.cn

Abstract

Minimax optimization problems are an important class of optimization problems arising from modern machine learning and traditional research areas. While there have been many numerical algorithms for solving smooth convex-concave minimax problems, numerical algorithms for nonsmooth convex-concave minimax problems are rare. This paper aims to develop an efficient numerical algorithm for a structured nonsmooth convex-concave minimax problem. A semi-proximal point method (SPP) is proposed, in which a quadratic convex-concave function is adopted for approximating the smooth part of the objective function and semi-proximal terms are added in each subproblem. This construction enables the subproblems at each iteration are solvable and even easily solved when the semiproximal terms are cleverly chosen. We prove the global convergence of our algorithm under mild assumptions, without requiring strong convexity-concavity condition. Under the locally metrical subregularity of the solution mapping, we prove that our algorithm has the linear rate of convergence. Preliminary numerical results are reported to verify the efficiency of our algorithm.

Mathematics subject classification: 90C30.
Key words: Minimax optimization, Convexity-concavity, Global convergence, Rate of convergence, Locally metrical subregularity.

1. Problem Setting

In this paper, we consider the following nonsmooth minimax optimization problem:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} L(x,y) := f(x) + K(x,y) - g(y), \tag{1.1}$$

where $K : \mathcal{X}\times\mathcal{Y} \to \Re$ is a continuously differentiable convex-concave function, and $f : \mathcal{X} \to \overline{\Re}$, $g : \mathcal{Y} \to \overline{\Re}$ are proper lower semi-continuous convex functions. $\mathcal{X}$ and $\mathcal{Y}$ be two finite-dimensional real Hilbert spaces equipped with a scalar product $\langle\cdot,\cdot\rangle$ and its induced norm $\|\cdot\|$.

The mathematical model (1.1) covers a lot of interesting convex-concave minimax problems appeared in the literature. We only list two examples here.

* Received April 20, 2022 / Revised version received July 26, 2022 / Accepted January 3, 2023 / Published online April 7, 2023 /
1) Corresponding author

Example 1.1. Let $X \subset \mathcal{X}$ and $Y \subset \mathcal{Y}$ be two closed convex sets. The following constrained minimax optimization problem are frequently studied in the literature:

$$\min_{x \in X} \max_{y \in Y} K(x, y). \tag{1.2}$$

Obviously, the constrained minimax optimization problem (1.2) can be written as the form of problem (1.1) if we set $f(x) = \delta_X(x)$ and $g(y) = \delta_Y(y)$.

Example 1.2. In Example 1.1, let $X \subset \mathcal{X}$ and $Y \subset \mathcal{Y}$ be two closed convex sets specified as

$$X = \{x \in \mathcal{X} : G(x) \in C\}, \quad Y = \{y \in \mathcal{Y} : H(y) \in D\},$$

where C and D are closed convex sets in some finite dimensional spaces and the set-valued mappings

$$x : \to \;\; G(x) - C, \quad y : \to \;\; H(y) - D$$

are graph-convex (under this assumption X and Y are convex sets). Then the constrained minimax optimization problem (1.2) can be written as the form of problem (1.1) if we set $f(x) = \delta_C(G(x))$ and $g(y) = \delta_D(H(y))$.

The study of algorithms for solving convex-concave minimax problems of the form (1.1) is active. For the case when K is a bilinear function, there are many publications about constructing and analyzing numerical algorithms for the minimax problem. The first work was due to Arrow *et al.* [1], where they proposed an alternating coordinate method, leaving the convergence unsolved. Nemirovski [14] considered the following minimax problem:

$$\min_{x} \max_{y \in Y} g(x) + x^T A y + h^T y,$$

where Y is a compact convex set and g is a $\mathcal{C}^{1,1}$ convex function. He proposed a mirror-prox algorithm which returns an approximate saddle point within the complexity of $\mathcal{O}(1/\varepsilon)$. Nesterov [15] developed a dual extrapolation algorithm for solving variational inequalities which owns the complexity bound $\mathcal{O}(1/\varepsilon)$ for Lipschitz continuous operators and applied the algorithm to bilinear matrix games. Chen *et al.* [7] presented a novel accelerated primal-dual (APD) method for solving this class of minimax problems, and showed that the APD method achieves the same optimal rate of convergence as Nesterov's smoothing technique. Chambolle and Pock [4] proposed a first-order primal-dual algorithm and established the convergence of the algorithm. Later, Chambolle and Pock provided the ergodic convergence rate [5] and explored the rate of convergence for accelerated primal-dual algorithms [6].

For smooth convex-concave minimax problems when K is not bilinear, many numerical algorithms are proposed such as the projection method [19], extragradient method [10], Tseng's accelerated proximal gradient algorithm [21], catalyst algorithm framework [24]. Recently, Mokhtari *et al.* [12] proposed algorithms admitting a unified analysis as approximations of the classical proximal point method for solving saddle point problems. Mokhtari *et al.* [13] proved that the optimistic gradient and extra-gradient methods achieve a convergence rate of $\mathcal{O}(1/k)$ for smooth convex-concave saddle point problems. Yoon and Ryu [25] combined extra-gradient steps with anchoring to reduce the squared gradient magnitude for smooth minimax problems which allows algorithms to obtain accelerated $\mathcal{O}(1/k^2)$ last-iterate rates.

Recently, Lin *et al.* [11] announced that they solved a longstanding open question pertaining to the design of near-optimal first-order algorithms for smooth and strongly-convex-strongly-concave minimax problems by presenting an algorithm with $\tilde{O}(\sqrt{\kappa_x \kappa_y})$ gradient complexity, matching the lower bound up to logarithmic factors [20]. In 2020, Wang and Li [23] proposed the proximal best response method, which improved over the best known upper bound by [11], and achieved linear convergence rate and tighter dependency on condition numbers.

For the nonsmooth convex-concave minimax problem (1.1), Valkonen [22] gave a modified primal-dual hybrid gradient method which is an extension of the primal-dual method in [4]. Furthermore, following the line of [4], Clason *et al.* [8] proposed a generalized primal-dual proximal splitting (GPDPS) method for solving the problem (1.1). However, the convergence of GPDPS depends on both Lipschitz gradients and bounded gradient of K. Recently, Hamedani [9] proposed a primal-dual algorithm for problem (1.1) and achieved an ergodic convergence rate of function value with $\mathcal{O}(1/k)$. To deal with the nonsmooth term in coupling function $K(x, y)$, Boţ *et al.* [3] designed an optimistic gradient ascent-proximal point algorithm and obtained a convergence rate of order $\mathcal{O}(1/K)$ for convex-concave saddle point problem. Distinct from the above research, in this paper, we build a semi-proximal alternating coordinate method and the convergence of iteration (x^k, y^k) only depends on Lipschitz gradients of K. Moreover, we establish the linear convergence rate of (x^k, y^k) provided with local metric subregularity, which, as far as we know, is not provided in other work for the general problem (1.1).

Averaging the weights of neural nets is a prohibitive approach in particular because the zero-sum game that is defined by training one deep net against another is not a convex-concave zero-sum game. Thus it seems essential to identify training algorithms that make the last iterate of the training be very close to the equilibrium, rather than only the average. However, the averaging technique is a popular tool for achieving good complexity, most of the mentioned papers (including [4, 5, 7, 13–15]) adopted that averaging technique. In this paper, we will not adopt the averaging technique in our algorithm.

Contribution. We propose a semi-proximal point method to solve nonsmooth convex-concave minimax problems with a more general coupling term K that is not bilinear. As K may have a complex structure, we use a convex-concave quadratic function to approximate function K. Then using the proximal point method to solve the sum of approximate quadratic function and nonsmooth convex function (with respect to x) and concave function (with respect to y), we design a numerical method for alternating iteration of x and y. Provided with some Lipschitz properties of the gradient of K, which are common in nonlinear optimization, (x^k, y^k) converges monotonically to a saddle point of general form of the problem (1.1).

In order to characterize the convergence rate, we define the local metric subregularity of the problem (1.1), which is an important definition in nonsmooth optimization and nonconvex optimization (see [2, 18]). If the local metric subregularity holds at all saddle points, (x^k, y^k) converges linearly to a saddle point.

This paper is organized as follows. In Section 2, we provide some technical results about minimizing the sum of a convex function and a quadratic proximal term, which will play an important role in the convergence analysis of SPP. In Section 3, we propose mild assumptions and prove the global convergence of SPP for solving problem (1.1). In Section 4, the linear rate of convergence of SPP is demonstrated under the assumption that the solution mapping is locally metrically subregular. In Section 5, we report some preliminary numerical results for linear regression problems with the saddle point formulation and for solving separate linearly constrained nonsmooth convex-concave minimax problems. Some discussions are made in the last section.

2. Preliminary

In this section, we give several results about properties of minimizing the sum of a convex function and a quadratic proximal term. The results will be used in the next section.

Lemma 2.1. *Let $f : \mathcal{Z} \to \Re$ be twice differentiable and $\mathcal{Z}$ be a finite-dimensional real Hilbert space with a scalar product $\langle \cdot, \cdot \rangle$ and its induced norm $\|\cdot\|$. Assume that there exists a self-adjoint operator $\Sigma_f : \mathcal{Z} \to \mathcal{Z}$ such that $\mathrm{D}^2 f(z) \succeq \Sigma_f$ for any $z \in \mathcal{Z}$ (which means that $\mathrm{D}^2 f(z) - \Sigma_f$ is a positive-definite self-adjoint operator). Define*

$$q(z', z) = f(z) - \left[f(z') + \mathrm{D}f(z')(z - z') + \frac{1}{2}\|z - z'\|^2_{\Sigma_f} \right]. \tag{2.1}$$

Then for any $z' \in \mathcal{Z}$, $q(z', z)$ is a convex function of z.

Proof. Note that

$$\mathrm{D}^2_z q(z', z) = \mathrm{D}^2 f(z) - \Sigma_f \succeq 0,$$

hence the conclusion is obtained from the definition of convex function in [16]. □

For a smooth convex optimization problem, the proximal point method often involves minimizing the sum of a smooth convex function, a nonsmooth convex function and a quadratic proximal term. An important inequality will be established in the following lemma.

Lemma 2.2. *Let the assumptions of Lemma 2.1 be satisfied and $T : \mathcal{Z} \to \mathcal{Z}$ be a positively semidefinite self-adjoint operator such that*

$$T + \Sigma_f \succ 0.$$

Then for any proper lower semi-continuous convex function $\psi : \mathcal{Z} \to \overline{\Re}$ and any $z^c \in \mathcal{Z}$, the problem

$$\min \psi(z) + f(z) + \frac{1}{2}\|z - z^c\|^2_T \tag{2.2}$$

has a unique solution, denoted by z^+. Moreover, for any $z \in \mathcal{Z}$,

$$\psi(z) + f(z) + \frac{1}{2}\|z - z^c\|^2_T - \frac{1}{2}\|z - z^+\|^2_{T+\Sigma_f} \geq \psi(z^+) + f(z^+) + \frac{1}{2}\|z^+ - z^c\|^2_T. \tag{2.3}$$

Proof. Define

$$\phi_c(z) = \psi(z) + f(z) + \frac{1}{2}\|z - z^c\|^2_T.$$

Then

$$\phi_c(z) = \left[f(z^+) + \mathrm{D}f(z^+)(z - z^+) + \frac{1}{2}\|z - z^+\|^2_{\Sigma_f} \right] + \frac{1}{2}\|z - z^c\|^2_T + q(z^+, z) + \psi(z).$$

Let

$$q_0(z) = \psi(z) + q(z^+, z) + \mathrm{D}f(z^+)(z - z^+) + \frac{1}{2}\|z - z^c\|^2_T - \frac{1}{2}\|z - z^+\|^2_T.$$

Then q_0 is convex and

$$\phi_c(z) = f(z^+) + \frac{1}{2}\|z - z^+\|^2_T + \frac{1}{2}\|z - z^+\|^2_{\Sigma_f} + q_0(z).$$

Since $0 \in \partial \phi_c(z^+)$, we have $0 \in \partial q_0(z^+)$ and q_0 arrives its minimum value at z^+. Then for every $z \in \mathcal{Z}$,

$$q_0(z) \geq q_0(z^+) = \psi(z^+) + \frac{1}{2}\|z^+ - z^c\|_T^2,$$

which is equivalent to

$$\psi(z) + f(z) + \frac{1}{2}\|z - z^c\|_T^2 - \frac{1}{2}\|z - z^+\|_{T+\Sigma_f}^2 \geq \psi(z^+) + f(z^+) + \frac{1}{2}\|z^+ - z^c\|_T^2.$$

The proof is complete. □

In the next section, we will find that the inequality (2.3) plays an important role in establishing the convergence of the proposed algorithm.

3. The Algorithm and Global Convergence

To begin with, we introduce some notation. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and denote $z = (x, y)$, $z' = (x', y'), z^k = (x^k, y^k)$ and $z^{k+1/2} = (x^{k+1/2}, y^{k+1/2})$. Let K be a smooth convex-concave function on an open set $\mathcal{O} \supset \operatorname{dom} f \times \operatorname{dom} g$; i.e., for each $(x, y) \in \mathcal{O}, K(\cdot, y)$ and $-K(x, \cdot)$ are smooth convex functions. Let $\widehat{\Sigma}_f$ and $\widehat{\Sigma}_g$ be self-adjoint and positive semidefinite linear operators. For any $(x, y), (x', y') \in \mathcal{O}$, we define

$$\begin{aligned}
\widehat{K}(x, y; x', y') &= K(x', y') + \langle \mathrm{D}_x K(x', y'), x - x'\rangle + \langle \mathrm{D}_y K(x', y'), y - y'\rangle \\
&\quad + \frac{1}{2}\|x - x'\|_{\widehat{\Sigma}_f}^2 - \frac{1}{2}\|y - y'\|_{\widehat{\Sigma}_g}^2, \\
\widehat{L}(x, y; x', y') &= f(x) + \widehat{K}(x, y; x', y') - g(y).
\end{aligned}$$

We propose a semi-proximal point algorithm (SPP) for solving problem (1.1) as below.

Algorithm 3.1: Semi-proximal Point Algorithm (SPP)

Step 0. Input $z^0 = (x^0, y^0) \in \operatorname{dom} f \times \operatorname{dom} g$. Set $k := 0$.
Step 1. Compute $z^{k+1/2} = (x^{k+1/2}, y^{k+1/2})$ and $z^{k+1} = (x^{k+1}, y^{k+1})$ by

$$\begin{cases}
x^{k+\frac{1}{2}} = \arg\min\limits_{x\in\mathcal{X}} \sigma\big[f(x) + \widehat{K}(x, y^k; z^k)\big] + \frac{1}{2}\|x - x^k\|_{\mathcal{S}}^2, \\
y^{k+\frac{1}{2}} = \arg\min\limits_{y\in\mathcal{Y}} \sigma\big[- \widehat{K}(x^k, y; z^k) + g(y)\big] + \frac{1}{2}\|y - y^k\|_{\mathcal{T}}^2, \\
x^{k+1} = \arg\min\limits_{x\in\mathcal{X}} \sigma\big[f(x) + \widehat{K}(x, y^{k+\frac{1}{2}}; z^{k+\frac{1}{2}})\big] + \frac{1}{2}\|x - x^k\|_{\mathcal{S}}^2, \\
y^{k+1} = \arg\min\limits_{y\in\mathcal{Y}} \sigma\big[- \widehat{K}(x^{k+\frac{1}{2}}, y; z^{k+\frac{1}{2}}) + g(y)\big] + \frac{1}{2}\|y - y^k\|_{\mathcal{T}}^2.
\end{cases}$$

Step 2. If a termination criterion is not met, set $k := k + 1$ and go to Step 1.

The motivation for proposing the above algorithm comes from the following observations:

(i) The reason for using $\widehat{K}(\cdot, \cdot; z)$ instead of $K(\cdot, \cdot)$ is that this makes the subproblems for determining $z^{k+1/2}$ and z^{k+1} to be easily solvable, especially when K is a complicated smooth convex-concave function. Furthermore, the subproblems for determining $z^{k+1/2}$ and z^{k+1} may have explicit solutions when f and g are simple convex functions.

(ii) The use of the semi-proximal terms (i.e., $\mathcal{S}$ and $\mathcal{T}$ are only required to be positively semidefinite) leaves the user a freedom to choose $\mathcal{S}$ and $\mathcal{T}$ so that the subproblems for determining $z^{k+1/2}$ and z^{k+1} are well-conditioned or are easily solved.

To analyze the global convergence of Algorithm 3.1, we need the following mild assumptions about functions in problem (1.1).

Assumption 3.1. *Let K be a smooth convex-concave function on an open set $\mathcal{O} \supset \operatorname{dom} f \times \operatorname{dom} g$; i.e., for each $(x, y) \in \mathcal{O}$, $K(\cdot, y)$ and $-K(x, \cdot)$ are smooth convex functions. Suppose there exist self-adjoint and positive semidefinite linear operators $\widehat{\Sigma}_f$ and $\widehat{\Sigma}_g$ such that for any $(x, y), (x', y') \in \mathcal{O}$,*

$$K(x, y) \geq K(x', y) + \langle \mathrm{D}_x K(x', y), x - x' \rangle + \frac{1}{2}\|x - x'\|^2_{\widehat{\Sigma}_f}, \tag{3.1}$$

$$-K(x, y) \geq -K(x, y') - \langle \mathrm{D}_y K(x, y'), y - y' \rangle + \frac{1}{2}\|y - y'\|^2_{\widehat{\Sigma}_g}. \tag{3.2}$$

For convenience, we introduce a linear operator $\widehat{\Sigma} : \mathcal{Z} \to \mathcal{Z}$ by

$$\widehat{\Sigma}(x, y) = \big(\widehat{\Sigma}_f x, \widehat{\Sigma}_g y\big). \tag{3.3}$$

Assumption 3.2. *Suppose that K is continuously differentiable on an open set $\mathcal{O} \supset \operatorname{dom} f \times \operatorname{dom} g$ and the derivative mapping $\mathrm{D}K$ is Lipschitz continuous with constant $\eta_0 > 0$, i.e.,*

$$\|\mathrm{D}K(x, y) - \mathrm{D}K(x', y')\| \leq \eta_0 \|(x, y) - (x', y')\|, \quad \forall (x', y), (x, y) \in \mathcal{O}.$$

Assumption 3.3. *The set of saddle points of L over $\operatorname{dom} f \times \operatorname{dom} g$ is nonempty, i.e., $\overline{\Omega} \neq \emptyset$, where $\overline{\Omega}$ is defined by*

$$\overline{\Omega} = \big\{(x, y) \in \mathcal{Z} : L(x, y') \leq L(x, y) \leq L(x', y),\ \forall (x', y') \in \operatorname{dom} f \times \operatorname{dom} g\big\}.$$

Define for $z = (x, y), z' = (x', y')$,

$$\Phi_K(z', z) = K(x', y) - K(x, y'), \quad \Phi_L(z', z) = L(x', y) - L(x, y').$$

Then we have for $z', z \in \operatorname{dom} f \times \operatorname{dom} g$ that

$$\Phi_K(z, z) = 0, \quad \Phi_L(z, z) = 0, \quad \Phi_K(z', z) + \Phi_K(z, z') = 0, \quad \Phi_L(z', z) + \Phi_L(z, z') = 0.$$

Furthermore, define for $z = (x, y), z' = (x', y')$ and $z'' = (x'', y'')$,

$$\widehat{\Phi}_K(z', z''; z) = \widehat{K}(x', y''; z) - \widehat{K}(x'', y'; z), \quad \widehat{\Phi}_L(z', z''; z) = \widehat{L}(x', y''; z) - \widehat{L}(x'', y'; z).$$

Then we have for $z', z'', z \in \operatorname{dom} f \times \operatorname{dom} g$ that

$$\begin{aligned} &\widehat{\Phi}_K(z', z'; z) = 0, && \widehat{\Phi}_L(z', z'; z) = 0, \\ &\widehat{\Phi}_K(z', z''; z) + \widehat{\Phi}_K(z'', z'; z) = 0, && \widehat{\Phi}_L(z', z''; z) + \widehat{\Phi}_L(z'', z'; z) = 0. \end{aligned}$$

Proposition 3.1. *Let Assumption 3.2 be satisfied. Then for $z \in \operatorname{dom} f \times \operatorname{dom} g$ and $h', h \in \mathcal{X} \times \mathcal{Y}$ such that $z + h + h', z + h \in \operatorname{dom} f \times \operatorname{dom} g$, one has that*

$$\begin{aligned} &\Big| \big[\widehat{\Phi}_K(z + h + h', z; z) - \widehat{\Phi}_K(z + h, z; z)\big] \\ &\quad - \big[\widehat{\Phi}_K(z + h + h', z + h; z + h) - \widehat{\Phi}_K(z + h, z + h; z + h)\big] \Big| \leq \widehat{\eta}_0 \|h\|\,\|h'\|, \end{aligned} \tag{3.4}$$

where

$$\widehat{\eta}_0 = \|\widehat{\Sigma}\| + \eta_0.$$

Proof. Let $h = (h_x, h_y) \in \mathcal{Z}$. Define an operation $\widetilde{\mathrm{D}}$ by

$$\widetilde{\mathrm{D}}K(z)h = \big(\mathrm{D}_x K(z)h_x, -\mathrm{D}_y K(z)h_y\big).$$

Since $\widehat{\Phi}_K(z+h, z+h; z+h) = 0$, we only need to consider the other three terms. For $h = (h_x, h_y)$ and $h' = (h'_x, h'_y)$,

$$\begin{aligned}
\widehat{\Phi}_K(z+h+h', z; z) &= \widehat{K}(x+h_x+h'_x, y; z) - \widehat{K}(x, y+h_y+h'_y; z) \\
&= \mathrm{D}_x K(z)(h_x+h'_x) - \mathrm{D}_y K(z)(h_y+h'_y) \\
&\quad + \frac{1}{2}\|h_x+h'_x\|^2_{\widehat{\Sigma}_f} + \frac{1}{2}\|h_y+h'_y\|^2_{\widehat{\Sigma}_g} \\
&= \widetilde{\mathrm{D}}K(z)(h+h') + \frac{1}{2}\|h+h'\|^2_{\widehat{\Sigma}}.
\end{aligned} \tag{3.5}$$

Similarly, we can get that

$$\widehat{\Phi}_K(z+h, z; z) = \widetilde{\mathrm{D}}K(z)(h) + \frac{1}{2}\|h\|^2_{\widehat{\Sigma}}, \tag{3.6}$$

$$\widehat{\Phi}_K(z+h+h', z+h; z+h) = \widetilde{\mathrm{D}}K(z+h)(h') + \frac{1}{2}\|h'\|^2_{\widehat{\Sigma}}. \tag{3.7}$$

Combing (3.5)-(3.7), we obtain

$$\begin{aligned}
&\Big|\big[\widehat{\Phi}_K(z+h+h', z; z) - \widehat{\Phi}_K(z+h, z; z)\big] \\
&\quad - \big[\widehat{\Phi}_K(z+h+h', z+h; z+h) - \widehat{\Phi}_K(z+h, z+h; z+h)\big]\Big| \\
&= \Big|\big[\widehat{\Phi}_K(z+h+h', z; z) - \widehat{\Phi}_K(z+h, z; z)\big] - \widehat{\Phi}_K(z+h+h', z+h; z+h)\Big| \\
&= \Big|\big[\widetilde{\mathrm{D}}K(z) - \widetilde{\mathrm{D}}K(z+h)\big](h') + \frac{1}{2}\|h+h'\|^2_{\widehat{\Sigma}} - \frac{1}{2}\|h\|^2_{\widehat{\Sigma}} - \frac{1}{2}\|h'\|^2_{\widehat{\Sigma}}\Big| \\
&= \big|\big[\widetilde{\mathrm{D}}K(z) - \widetilde{\mathrm{D}}K(z+h)\big](h') + \langle\widehat{\Sigma}h, h'\rangle\big| \\
&\le \big\|\big[\widetilde{\mathrm{D}}K(z) - \widetilde{\mathrm{D}}K(z+h)\big]\big\|\,\|h'\| + \|\widehat{\Sigma}\|\,\|h\|\,\|h'\| \\
&\le \big(\eta_0 + \|\widehat{\Sigma}\|\big)\|h\|\,\|h'\| = \widehat{\eta}_0\|h\|\,\|h'\|.
\end{aligned}$$

The proof is complete. $\square$

Remark 3.1. From the definition of $\widehat{\Phi}_K$ and $\widehat{\Phi}_L$, we have

$$\begin{aligned}
&\Big|\big[\widehat{\Phi}_L(z+h+h', z; z) - \widehat{\Phi}_L(z+h, z; z)\big] \\
&\quad - \big[\widehat{\Phi}_L(z+h+h', z+h; z+h) - \widehat{\Phi}_L(z+h, z+h; z+h)\big]\Big| \\
&= \Big|\big[\widehat{\Phi}_K(z+h+h', z; z) - \widehat{\Phi}_K(z+h, z; z)\big] \\
&\quad - \big[\widehat{\Phi}_K(z+h+h', z+h; z+h) - \widehat{\Phi}_K(z+h, z+h; z+h)\big]\Big|,
\end{aligned}$$

which yields from (3.4) that

$$\begin{aligned}
&\Big|\big[\widehat{\Phi}_L(z+h+h', z; z) - \widehat{\Phi}_L(z+h, z; z)\big] \\
&\quad - \big[\widehat{\Phi}_L(z+h+h', z+h; z+h) - \widehat{\Phi}_L(z+h, z+h; z+h)\big]\Big| \le \widehat{\eta}_0\|h\|\,\|h'\|.
\end{aligned}$$

Let $\sigma > 0$ be a given parameter. Let $\mathcal{S}$ and $\mathcal{T}$ be given self-adjoint linear operators satisfying

$$\mathcal{S} \succeq 0, \quad \mathcal{T} \succeq 0, \quad \sigma\widehat{\Sigma}_f + \mathcal{S} \succ 0, \quad \sigma\widehat{\Sigma}_g + \mathcal{T} \succ 0. \tag{3.8}$$

Define a linear operator $\Theta : \mathcal{Z} \to \mathcal{Z}$ by

$$\Theta(x, y) = (\mathcal{S}x, \mathcal{T}y). \tag{3.9}$$

Then the formulas for $(x^{k+1/2}, y^{k+1/2})$ and (x^{k+1}, y^{k+1}) in Step 1 of Algorithm 3.1 can be written as

$$\begin{aligned} z^{k+\frac{1}{2}} &= \arg\min_{z\in\mathcal{Z}} \left\{ \frac{1}{2}\|z - z^k\|_\Theta^2 + \sigma\widehat{\Phi}_L(z, z^k; z^k) \right\}, \\ z^{k+1} &= \arg\min_{z\in\mathcal{Z}} \left\{ \frac{1}{2}\|z - z^k\|_\Theta^2 + \sigma\widehat{\Phi}_L\big(z, z^{k+\frac{1}{2}}; z^{k+\frac{1}{2}}\big) \right\}. \end{aligned} \tag{3.10}$$

Noting that $\widehat{\Phi}_L(z, z^{k+1/2}; z^{k+1/2})$ and $\widehat{\Phi}_L(z, z^k; z^k)$ are convex and problems in (3.10) have the same structure as problem (2.2), we may use Lemma 2.2 to estimate $\|z - z^{k+1}\|_{\sigma\widehat{\Sigma}+\Theta}$ and $\|z - z^{k+1/2}\|_{\sigma\widehat{\Sigma}+\Theta}$. Specifically, by Lemma 2.2, we know from the definition of z^{k+1} that for any $z \in \operatorname{dom} f \times \operatorname{dom} g$,

$$\begin{aligned} &\frac{1}{2}\|z - z^k\|_\Theta^2 + \sigma\widehat{\Phi}_L\big(z, z^{k+\frac{1}{2}}; z^{k+\frac{1}{2}}\big) - \frac{1}{2}\|z - z^{k+1}\|_{\sigma\widehat{\Sigma}+\Theta}^2 \\ \geq\ & \frac{1}{2}\|z^{k+1} - z^k\|_\Theta^2 + \sigma\widehat{\Phi}_L\big(z^{k+1}, z^{k+\frac{1}{2}}; z^{k+\frac{1}{2}}\big). \end{aligned} \tag{3.11}$$

Similarly, we have from the definition of $z^{k+1/2}$ that for any $z \in \operatorname{dom} f \times \operatorname{dom} g$,

$$\begin{aligned} &\frac{1}{2}\|z - z^k\|_\Theta^2 + \sigma\widehat{\Phi}_L(z, z^k; z^k) - \frac{1}{2}\big\|z - z^{k+\frac{1}{2}}\big\|_{\sigma\widehat{\Sigma}+\Theta}^2 \\ \geq\ & \frac{1}{2}\big\|z^{k+\frac{1}{2}} - z^k\big\|_\Theta^2 + \sigma\widehat{\Phi}_L\big(z^{k+\frac{1}{2}}, z^k; z^k\big). \end{aligned} \tag{3.12}$$

We first establish the relation between $\|z^{k+1} - z^{k+1/2}\|_{\Theta+\sigma\widehat{\Sigma}}$ and $\|z^{k+1/2} - z^k\|_{\Theta+\sigma\widehat{\Sigma}}$ in the following proposition.

Proposition 3.2. *Let Assumptions* 3.1 *and* 3.2 *be satisfied,* $\mathcal{S}$ *and* $\mathcal{T}$ *satisfy* (3.8) *or equivalently* $\Theta \succeq 0$ *and* $\Theta + \sigma\widehat{\Sigma} \succ 0$. *Let* $\{z^k = (x^k, y^k)\}, \{z^{k+1/2} = (x^{k+1/2}, y^{k+1/2})\}$ *be generated by Algorithm* 3.1. *Then*

$$\big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|_{\Theta+\sigma\widehat{\Sigma}} \leq \vartheta(\sigma)\big\|z^{k+\frac{1}{2}} - z^k\big\|_{\Theta+\sigma\widehat{\Sigma}}, \tag{3.13}$$

where

$$\vartheta(\sigma) = \frac{\sigma\widehat{\eta}_0}{\lambda_{\min}(\Theta + \sigma\widehat{\Sigma})}.$$

Proof. Setting $z = z^{k+1/2}$ in (3.11), we obtain

$$\begin{aligned} &\frac{1}{2}\big\|z^{k+\frac{1}{2}} - z^k\big\|_\Theta^2 + \sigma\widehat{\Phi}_L\big(z^{k+\frac{1}{2}}, z^{k+\frac{1}{2}}; z^{k+\frac{1}{2}}\big) - \frac{1}{2}\big\|z^{k+\frac{1}{2}} - z^{k+1}\big\|_{\sigma\widehat{\Sigma}+\Theta}^2 \\ \geq\ & \frac{1}{2}\|z^{k+1} - z^k\|_\Theta^2 + \sigma\widehat{\Phi}_L\big(z^{k+1}, z^{k+\frac{1}{2}}; z^{k+\frac{1}{2}}\big). \end{aligned} \tag{3.14}$$

Setting $z = z^{k+1}$ in (3.12), we obtain

$$\begin{aligned}&\frac{1}{2}\|z^{k+1}-z^k\|_\Theta^2+\sigma\widehat{\Phi}_L(z^{k+1},z^k;z^k)-\frac{1}{2}\big\|z^{k+1}-z^{k+\frac{1}{2}}\big\|_{\sigma\widehat{\Sigma}+\Theta}^2\\ \geq\;&\frac{1}{2}\big\|z^{k+\frac{1}{2}}-z^k\big\|_\Theta^2+\sigma\widehat{\Phi}_L\big(z^{k+\frac{1}{2}},z^k;z^k\big).\end{aligned}\tag{3.15}$$

Summing (3.14) and (3.15), we get from Proposition 3.1 (with $z = z^k, h = z^{k+1/2} - z^k$, $h' = z^{k+1} - z^{k+1/2}$) that

$$\begin{aligned}\big\|z^{k+1}-z^{k+\frac{1}{2}}\big\|_{\sigma\widehat{\Sigma}+\Theta}^2 &\leq \sigma\Big\{\widehat{\Phi}_L\big(z^{k+\frac{1}{2}},z^{k+\frac{1}{2}};z^{k+\frac{1}{2}}\big)-\widehat{\Phi}_L\big(z^{k+1},z^{k+\frac{1}{2}};z^{k+\frac{1}{2}}\big)\\ &\qquad -\big[\widehat{\Phi}_L\big(z^{k+\frac{1}{2}},z^k;z^k\big)-\widehat{\Phi}_L(z^{k+1},z^k;z^k)\big]\Big\}\\ &=\sigma\Big\{\widehat{\Phi}_K\big(z^{k+\frac{1}{2}},z^{k+\frac{1}{2}};z^{k+\frac{1}{2}}\big)-\widehat{\Phi}_K\big(z^{k+1},z^{k+\frac{1}{2}};z^{k+\frac{1}{2}}\big)\\ &\qquad -\big[\widehat{\Phi}_K\big(z^{k+\frac{1}{2}},z^k;z^k\big)-\widehat{\Phi}_K(z^{k+1},z^k;z^k)\big]\Big\}\\ &\leq \sigma\widehat{\eta}_0\big\|z^{k+1}-z^{k+\frac{1}{2}}\big\|\,\big\|z^{k+\frac{1}{2}}-z^k\big\|.\end{aligned}\tag{3.16}$$

Therefore, in view of

$$\sqrt{\lambda_{\min}(\sigma\widehat{\Sigma}+\Theta)}\|z\|\leq\|z\|_{\sigma\widehat{\Sigma}+\Theta},\quad\forall z\in\mathcal{Z}$$

we obtain the desired result from the last expression of (3.16). The proof is complete. □

Now we provide the main result about the global convergence of Algorithm 3.1.

Theorem 3.1. *Let Assumptions* 3.1-3.3 *be satisfied and $\mathcal{S}$ and $\mathcal{T}$ satisfy* (3.8) *or equivalently $\Theta \succeq 0$ and $\Theta+\sigma\widehat{\Sigma} \succ 0$. Consider the sequences $\{z^k = (x^k, y^k)\}$ and $\{z^{k+1/2} = (x^{k+1/2}, y^{k+1/2})\}$ generated by Algorithm* 3.1*. Suppose that $\sigma, \mathcal{S}$ and $\mathcal{T}$ satisfy*

$$\Theta\succ\widehat{\eta}_0\sigma I.\tag{3.17}$$

Then $\{z^k = (x^k, y^k)\}$ converges monotonically with respect to some norm to an element of $\overline{\Omega}$.

Proof. Choose an element $z^* \in \overline{\Omega}$. Setting $z = z^*$ in (3.11), we obtain

$$\begin{aligned}&\frac{1}{2}\|z^*-z^k\|_\Theta^2+\sigma\widehat{\Phi}_L\big(z^*,z^{k+\frac{1}{2}};z^{k+\frac{1}{2}}\big)-\frac{1}{2}\|z^*-z^{k+1}\|_{\sigma\widehat{\Sigma}+\Theta}^2\\ \geq\;&\frac{1}{2}\|z^{k+1}-z^k\|_\Theta^2+\sigma\widehat{\Phi}_L\big(z^{k+1},z^{k+\frac{1}{2}};z^{k+\frac{1}{2}}\big).\end{aligned}\tag{3.18}$$

Setting $z = z^{k+1}$ in (3.12), we obtain

$$\begin{aligned}&\frac{1}{2}\|z^{k+1}-z^k\|_\Theta^2+\sigma\widehat{\Phi}_L(z^{k+1},z^k;z^k)-\frac{1}{2}\big\|z^{k+1}-z^{k+\frac{1}{2}}\big\|_{\sigma\widehat{\Sigma}+\Theta}^2\\ \geq\;&\frac{1}{2}\big\|z^{k+\frac{1}{2}}-z^k\big\|_\Theta^2+\sigma\widehat{\Phi}_L\big(z^{k+\frac{1}{2}},z^k;z^k\big).\end{aligned}\tag{3.19}$$

Summing (3.18) and (3.19), we have

$$\begin{aligned}\|z^k - z^*\|_\Theta^2 \geq{}& \|z^{k+1} - z^*\|^2_{\sigma\widehat{\Sigma}+\Theta} + \big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|^2_{\sigma\widehat{\Sigma}+\Theta} + \big\|z^{k+\frac{1}{2}} - z^k\big\|^2_\Theta \\ &+ 2\sigma\Big[\widehat{\Phi}_L\big(z^{k+1}, z^{k+\frac{1}{2}}; z^{k+\frac{1}{2}}\big) - \widehat{\Phi}_L\big(z^*, z^{k+\frac{1}{2}}; z^{k+\frac{1}{2}}\big) \\ &\quad + \widehat{\Phi}_L\big(z^{k+\frac{1}{2}}, z^k; z^k\big) - \widehat{\Phi}_L(z^{k+1}, z^k; z^k)\Big] \\ ={}& \|z^{k+1} - z^*\|^2_{\sigma\widehat{\Sigma}+\Theta} + \big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|^2_{\sigma\widehat{\Sigma}+\Theta} + \big\|z^{k+\frac{1}{2}} - z^k\big\|^2_\Theta \\ &+ 2\sigma\Big[\widehat{\Phi}_L\big(z^{k+1}, z^{k+\frac{1}{2}}; z^{k+\frac{1}{2}}\big) - \widehat{\Phi}_L\big(z^{k+\frac{1}{2}}, z^{k+\frac{1}{2}}; z^{k+\frac{1}{2}}\big) \\ &\quad + \widehat{\Phi}_L\big(z^{k+\frac{1}{2}}, z^k; z^k\big) - \widehat{\Phi}_L(z^{k+1}, z^k; z^k)\Big] \\ &+ 2\sigma\Big[\widehat{\Phi}_L\big(z^{k+\frac{1}{2}}, z^{k+\frac{1}{2}}; z^{k+\frac{1}{2}}\big) - \widehat{\Phi}_L\big(z^*, z^{k+\frac{1}{2}}; z^{k+\frac{1}{2}}\big)\Big]. \end{aligned} \tag{3.20}$$

Noting that

$$\widehat{\Phi}_L\big(z^{k+\frac{1}{2}}, z^{k+\frac{1}{2}}; z^{k+\frac{1}{2}}\big) = 0,$$

we need to estimate the term $-\widehat{\Phi}_L(z^*, z^{k+1/2}; z^{k+1/2})$. In fact, we have from Assumption 3.1 that

$$\begin{aligned} &- \widehat{\Phi}_L\big(z^*, z^{k+\frac{1}{2}}; z^{k+\frac{1}{2}}\big) = \widehat{L}\big(x^{k+\frac{1}{2}}, y^*; z^{k+\frac{1}{2}}\big) - \widehat{L}\big(x^*, y^{k+\frac{1}{2}}; z^{k+\frac{1}{2}}\big) \\ &= f\big(x^{k+\frac{1}{2}}\big) + \widehat{K}\big(x^{k+\frac{1}{2}}, y^*; z^{k+\frac{1}{2}}\big) - g(y^*) - \Big[f(x^*) + \widehat{K}\big(x^*, y^{k+\frac{1}{2}}; z^{k+\frac{1}{2}}\big) - g\big(y^{k+\frac{1}{2}}\big)\Big] \\ &= f\big(x^{k+\frac{1}{2}}\big) - g(y^*) + K\big(z^{k+\frac{1}{2}}\big) + \big\langle \mathrm{D}_y K\big(z^{k+\frac{1}{2}}\big), y^* - y^{k+\frac{1}{2}}\big\rangle - \frac{1}{2}\big\|y^* - y^{k+\frac{1}{2}}\big\|^2_{\widehat{\Sigma}_g} \\ &\quad - \Big[f(x^*) - g\big(y^{k+\frac{1}{2}}\big) + K\big(z^{k+\frac{1}{2}}\big) + \big\langle \mathrm{D}_x K\big(z^{k+\frac{1}{2}}\big), x^* - x^{k+\frac{1}{2}}\big\rangle\Big] - \frac{1}{2}\big\|x^* - x^{k+\frac{1}{2}}\big\|^2_{\widehat{\Sigma}_f} \\ &= f\big(x^{k+\frac{1}{2}}\big) - g(y^*) - f(x^*) + g\big(y^{k+\frac{1}{2}}\big) \\ &\quad - \Big[- K\big(z^{k+\frac{1}{2}}\big) - \big\langle \mathrm{D}_y K\big(z^{k+\frac{1}{2}}\big), y^* - y^{k+\frac{1}{2}}\big\rangle + \frac{1}{2}\big\|y^* - y^{k+\frac{1}{2}}\big\|^2_{\widehat{\Sigma}_g}\Big] \\ &\quad - \Big[K\big(z^{k+\frac{1}{2}}\big) + \big\langle \mathrm{D}_x K\big(z^{k+\frac{1}{2}}\big), x^* - x^{k+\frac{1}{2}} - x^*\big\rangle + \frac{1}{2}\big\|x^* - x^{k+\frac{1}{2}}\big\|^2_{\widehat{\Sigma}_f}\Big] \\ &\geq f\big(x^{k+\frac{1}{2}}\big) - g(y^*) - f(x^*) + g\big(y^{k+\frac{1}{2}}\big) + K\big(x^{k+\frac{1}{2}}, y^*\big) - K\big(x^*, y^{k+\frac{1}{2}}\big) \\ &= \Big[f\big(x^{k+\frac{1}{2}}\big) + K\big(x^{k+\frac{1}{2}}, y^*\big) - g(y^*) - L(x^*, y^*)\Big] \\ &\quad + \Big[L(x^*, y^*) - \Big(f(x^*) + K\big(x^*, y^{k+\frac{1}{2}}\big) - g\big(y^{k+\frac{1}{2}}\big)\Big)\Big] \geq 0, \end{aligned} \tag{3.21}$$

where the last inequality comes from $(x^*, y^*) \in \overline{\Omega}$.

Thus we can get from (3.20) and (3.21) that

$$\begin{aligned}\|z^k - z^*\|_\Theta^2 \geq{}& \|z^{k+1} - z^*\|^2_{\sigma\widehat{\Sigma}+\Theta} + \big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|^2_{\sigma\widehat{\Sigma}+\Theta} + \big\|z^{k+\frac{1}{2}} - z^k\big\|^2_\Theta \\ &+ 2\sigma\Big[\widehat{\Phi}_L\big(z^{k+1}, z^{k+\frac{1}{2}}; z^{k+\frac{1}{2}}\big) - \widehat{\Phi}_L\big(z^{k+\frac{1}{2}}, z^{k+\frac{1}{2}}; z^{k+\frac{1}{2}}\big) \\ &\quad + \widehat{\Phi}_L\big(z^{k+\frac{1}{2}}, z^k; z^k\big) - \widehat{\Phi}_L(z^{k+1}, z^k; z^k)\Big]. \end{aligned} \tag{3.22}$$

In view of Remark 3.1, we obtain from (3.22) that

$$\begin{aligned}\|z^k - z^*\|_\Theta^2 \geq\ & \|z^{k+1} - z^*\|^2_{\sigma\widehat{\Sigma}+\Theta} + \big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|^2_{\sigma\widehat{\Sigma}+\Theta} + \big\|z^{k+\frac{1}{2}} - z^k\big\|^2_\Theta \\ & - 2\sigma\widehat{\eta}_0\big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|\,\big\|z^{k+\frac{1}{2}} - z^k\big\| \\ \geq\ & \|z^{k+1} - z^*\|^2_{\sigma\widehat{\Sigma}+\Theta} + \big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|^2_{\sigma\widehat{\Sigma}+\Theta} + \big\|z^{k+\frac{1}{2}} - z^k\big\|^2_\Theta \\ & - \sigma\widehat{\eta}_0\Big[\big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|^2 + \big\|z^{k+\frac{1}{2}} - z^k\big\|^2\Big] \\ =\ & \|z^{k+1} - z^*\|^2_{\sigma\widehat{\Sigma}+\Theta} + \big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|^2_{\sigma\widehat{\Sigma}+\Theta-\sigma\widehat{\eta}_0 I} + \big\|z^{k+\frac{1}{2}} - z^k\big\|^2_{\Theta-\sigma\widehat{\eta}_0 I}. \end{aligned} \tag{3.23}$$

Define

$$G(\sigma) = \sigma\widehat{\Sigma} + \Theta, \quad N(\sigma) = \sigma\widehat{\Sigma} + \Theta - \widehat{\eta}_0\sigma I, \quad H(\sigma) = \Theta - \widehat{\eta}_0\sigma I.$$

The relation (3.23) implies that

$$\|z^k - z^*\|^2_{G(\sigma)} \geq \|z^{k+1} - z^*\|^2_{G(\sigma)} + \big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|^2_{N(\sigma)} + \big\|z^{k+\frac{1}{2}} - z^k\big\|^2_{H(\sigma)}. \tag{3.24}$$

From (3.17), we know that $H(\sigma)$ and $N(\sigma)$ are positively definite. Hence by (3.24), we get that

$$\|z^k - z^*\|^2_{G(\sigma)} \geq \|z^{k+1} - z^*\|^2_{G(\sigma)} + \big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|^2_{N(\sigma)} + \big\|z^{k+\frac{1}{2}} - z^k\big\|^2_{H(\sigma)}. \tag{3.25}$$

Summing the inequality (3.25) over k from 0 to N, we obtain

$$\|z^0 - z^*\|^2_{G(\sigma)} \geq \big\|z^{N+1} - z^*\big\|^2_{G(\sigma)} + \sum_{k=1}^{N} \big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|^2_{N(\sigma)} + \sum_{k=1}^{N} \big\|z^{k+\frac{1}{2}} - z^k\big\|^2_{H(\sigma)}. \tag{3.26}$$

Thus we obtain from (3.26) that

$$\begin{aligned} & \|z^0 - z^*\|^2_{G(\sigma)} \geq \big\|z^{N+1} - z^*\big\|^2_{G(\sigma)}, \\ & \sum_{k=1}^{\infty} \big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|^2_{N(\sigma)} < \infty, \\ & \sum_{k=1}^{\infty} \big\|z^{k+\frac{1}{2}} - z^k\big\|^2_{H(\sigma)} \leq \infty, \end{aligned}$$

implying that

$$\|z^{k+1} - z^k\| \ \to\ 0. \tag{3.27}$$

Since the sequence $\{z^k\}$ is bounded, there exist an element $\bar{z}$ and $\{k_i\} \subset \mathbf{N}$ such that $z^{k_i} \to \bar{z}$. It follows from (3.12) that, for $z \in \mathrm{dom} f \times \mathrm{dom} g$,

$$\begin{aligned} & \frac{1}{2}\|z - z^k\|^2_\Theta + \sigma\widehat{\Phi}_L(z, z^k; z^k) - \frac{1}{2}\big\|z - z^{k+\frac{1}{2}}\big\|^2_{\sigma\widehat{\Sigma}+\Theta} \\ \geq\ & \sigma\widehat{\Phi}_L\big(z^{k+\frac{1}{2}}, z^k; z^k\big) + \frac{1}{2}\big\|z^{k+\frac{1}{2}} - z^k\big\|^2_\Theta, \end{aligned}$$

which is equivalent to

$$\begin{aligned} & \frac{1}{2}\|z - z^k\|^2_{\sigma\widehat{\Sigma}+\Theta} - \frac{1}{2}\big\|z - z^{k+\frac{1}{2}}\big\|^2_{\sigma\widehat{\Sigma}+\Theta} \\ & \quad + \sigma\Big[f(x) + K(z^k) + \langle \mathrm{D}_x K(z^k), x - x^k\rangle + \frac{1}{2}\|x - x^k\|^2_{\widehat{\Sigma}_f} - g(y^k)\Big] \\ & \quad - \sigma\Big[f(x^k) + K(z^k) + \langle \mathrm{D}_y K(z^k), y - y^k\rangle - \frac{1}{2}\|y - y^k\|^2_{\widehat{\Sigma}_g} - g(y)\Big] \\ \geq\ & \sigma\widehat{\Phi}_L\big(z^{k+\frac{1}{2}}, z^k; z^k\big) + \frac{1}{2}\big\|z^{k+\frac{1}{2}} - z^k\big\|^2_\Theta. \end{aligned}$$

The above relation indicates that

$$\begin{aligned}
&\frac{1}{2}\|z-z^k\|^2_{\sigma\widehat{\Sigma}+\Theta}-\frac{1}{2}\|z-z^{k+\frac{1}{2}}\|^2_{\sigma\widehat{\Sigma}+\Theta}+\sigma\Big[f(x)+\langle \mathrm{D}_xK(z^k),x-x^k\rangle-g(y^k)\Big]\\
&\quad-\sigma\Big[f(x^k)+\langle \mathrm{D}_yK(z^k),y-y^k\rangle-g(y)\Big]\\
&\geq\frac{1}{2}\big\|z^{k+\frac{1}{2}}-z^k\big\|^2_{\Theta}+\sigma\Big[f\big(x^{k+\frac{1}{2}}\big)+\big\langle \mathrm{D}_xK(z^k),x^{k+\frac{1}{2}}-x^k\big\rangle+\frac{1}{2}\big\|x^{k+\frac{1}{2}}-x^k\big\|^2_{\widehat{\Sigma}_f}-g(y^k)\Big]\\
&\quad-\sigma\Big[f(x^k)+\big\langle \mathrm{D}_yK(z^k),y^{k+\frac{1}{2}}-y^k\big\rangle-\frac{1}{2}\big\|y^{k+\frac{1}{2}}-y^k\big\|^2_{\widehat{\Sigma}_g}-g\big(y^{k+\frac{1}{2}}\big)\Big].
\end{aligned}$$

Thus we can get that

$$\begin{aligned}
&\frac{1}{2}\|z-z^k\|^2_{\sigma\widehat{\Sigma}+\Theta}-\frac{1}{2}\|z-z^{k+\frac{1}{2}}\|^2_{\sigma\widehat{\Sigma}+\Theta}+\sigma\Big[f(x)+\langle \mathrm{D}_xK(z^k),x-x^k\rangle\Big]\\
&\quad-\sigma\Big[\langle \mathrm{D}_yK(z^k),y-y^k\rangle-g(y)\Big]\\
&\geq\frac{1}{2}\big\|z^{k+\frac{1}{2}}-z^k\big\|^2_{\Theta}+\sigma\Big[\big\langle \mathrm{D}_xK(z^k),x^{k+\frac{1}{2}}-x^k\big\rangle+\frac{1}{2}\big\|x^{k+\frac{1}{2}}-x^k\big\|^2_{\widehat{\Sigma}_f}\Big]\\
&\quad-\sigma\Big[\big\langle \mathrm{D}_yK(z^k),y^{k+\frac{1}{2}}-y^k\big\rangle-\frac{1}{2}\big\|y^{k+\frac{1}{2}}-y^k\big\|^2_{\widehat{\Sigma}_g}\Big]+\sigma f\big(x^{k+\frac{1}{2}}\big)+\sigma g\big(y^{k+\frac{1}{2}}\big).
\end{aligned}\tag{3.28}$$

Since f and g are lower semi-continuous and $\|z^{k_i+1/2}-z^{k_i}\|\to 0$ as $i\to\infty$, letting $k=k_i$ and taking the lower limit along $\{k_i\}$ on both sides of (3.28), we obtain

$$\big[f(x)+\langle \mathrm{D}_xK(\bar{z}),x-\bar{x}\rangle-g(\bar{y})\big]-\big[f(\bar{x})+\langle \mathrm{D}_yK(\bar{z}),y-\bar{y}\rangle-g(y)\big]\geq 0. \tag{3.29}$$

In view of the convexity-concavity of K, we have that

$$\begin{aligned}
\langle \mathrm{D}_xK(\bar{z}),x-\bar{x}\rangle&\leq K(x,\bar{y})-K(\bar{x},\bar{y}),\\
\langle \mathrm{D}_yK(\bar{z}),y-\bar{y}\rangle&\geq K(\bar{x},y)-K(\bar{x},\bar{y}).
\end{aligned}$$

Combining these with (3.29), we obtain

$$\big[f(x)+K(x,\bar{y})-g(\bar{y})\big]-\big[f(\bar{x})+K(\bar{x},y)-g(y)\big]\geq 0.$$

This implies that $\bar{z}\in\overline{\Omega}$.

Thus, any limit point of the sequence $\{z^k\}$ is a solution of the problem. For any limit point z^* of $\{z^k\}$, the relation (3.25) indicates that the quantity $\|z^k-z^*\|^2_{G(\sigma)}$ decreases monotonically. Combing these two facts, we know that the sequence $\{z^k\}$ can only have one limit point, that is, z^k converges monotonically with respect to $G(\sigma)$-norm to one of the solutions of the problem, i.e., $z^k\to z^*$. This proves the statement. □

4. The Rate of Convergence

Under Assumption 3.1, L is a convex-concave function and $(\bar{x},\bar{y})\in\overline{\Omega}$ if and only if $(\bar{x},\bar{y})$ satisfies

$$\begin{aligned}
0&\in\partial f(x)+\mathrm{D}_xK(x,y),\\
0&\in\partial g(y)-\mathrm{D}_yK(x,y).
\end{aligned}\tag{4.1}$$

This is a generalization for the optimality. We can also express the optimality as an equation

$$R(z) = R(x, y) = 0,$$

where

$$R(z) = \begin{bmatrix} x - \mathbf{P}_f\big(x - \mathrm{D}_x K(x, y)\big) \\ y - \mathbf{P}_g\big(y + \mathrm{D}_y K(x, y)\big) \end{bmatrix}, \tag{4.2}$$

and $\mathbf{P}_\psi$ is the proximal mapping. Here, for a convex function $\psi, \mathbf{P}_\psi$ is defined by

$$\mathbf{P}_\psi(w) = \arg\min_{w'} \left\{ \psi(w') + \frac{1}{2}\|w' - w\|^2 \right\}.$$

Then we can express the set of saddle points of L as

$$\overline{\Omega} = R^{-1}(0).$$

To develop the rate of convergence of Algorithm 3.1, we need the following metric subregularity of R at $(z^*, 0)$ which is also defined in [2, Definition 2.80].

Assumption 4.1. *Suppose that R is locally metrically subregular at $(z^*, 0)$, i.e., there exist $\varepsilon_0 > 0$ and $\kappa_0 > 0$ such that*

$$\mathrm{dist}\,(z, R^{-1}(0) = \overline{\Omega}) \le \kappa_0 \|R(z)\|, \quad \forall z \in \boldsymbol{B}(z^*, \varepsilon_0). \tag{4.3}$$

The metric subregularity of R at $(z^*, 0)$ is instrumental for estimating the distance to the optimal solution set of the minimax optimization problem. From [2, Theorem 2.81], the metric subregularity of R at $(z^*, 0)$ is equivalent to $0 \in \mathrm{int}\,(\mathrm{range}\, R)$. Hence, if the optimal solution $z^* = (x^*, y^*)$ of the minimax problem (1.1) satisfies $(x^*, y^*) \in \mathrm{int}\, L$, then the local metric subregularity of R at $(z^*, 0)$ holds.

Proposition 4.1. *Let Assumption* 3.2 *be satisfied. Let $\{z^k = (x^k, y^k)\}$ and $\{z^{k+1/2} = (x^{k+1/2}, y^{k+1/2})\}$ be generated by Algorithm* 3.1. *Then for $k \ge 0$,*

$$\big\|R(z^{k+1})\big\|^2 \le \big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|^2_{6\eta_0^2 I + 3(\widehat{\Sigma} + \sigma^{-1}\Theta)^*(\widehat{\Sigma} + \sigma^{-1}\Theta)} + \big\|z^{k+\frac{1}{2}} - z^k\big\|^2_{3\sigma^{-2}\Theta^*\Theta}. \tag{4.4}$$

Proof. From the definition of z^{k+1} in (3.10) and the optimality condition, we obtain

$$\begin{aligned} 0 &\in \sigma^{-1}\mathcal{S}(x^{k+1} - x^k) + \mathrm{D}_x \widehat{K}\big(x^{k+1}, y^{k+\frac{1}{2}}; z^{k+\frac{1}{2}}\big) + \partial f(x^{k+1}), \\ 0 &\in \sigma^{-1}\mathcal{T}(y^{k+1} - y^k) - \mathrm{D}_y \widehat{K}\big(x^{k+\frac{1}{2}}, y^{k+1}; z^{k+\frac{1}{2}}\big) + \partial g(y^{k+1}). \end{aligned}$$

Then by the definition of $\widehat{K}(z; z^{k+1/2})$, we know that

$$\begin{aligned} 0 &\in \sigma^{-1}\mathcal{S}(x^{k+1} - x^k) + \mathrm{D}_x K\big(z^{k+\frac{1}{2}}\big) + \widehat{\Sigma}_f\big(x^{k+1} - x^{k+\frac{1}{2}}\big) + \partial f(x^{k+1}), \\ 0 &\in \sigma^{-1}\mathcal{T}(y^{k+1} - y^k) - \mathrm{D}_y K\big(z^{k+\frac{1}{2}}\big) + \widehat{\Sigma}_g\big(y^{k+1} - y^{k+\frac{1}{2}}\big) + \partial g(y^{k+1}), \end{aligned} \tag{4.5}$$

and

$$\begin{aligned} x^{k+1} &= \mathbf{P}_f\Big(x^{k+1} - \mathrm{D}_x K\big(z^{k+\frac{1}{2}}\big) - \widehat{\Sigma}_f\big(x^{k+1} - x^{k+\frac{1}{2}}\big) - \sigma^{-1}\mathcal{S}(x^{k+1} - x^k)\Big), \\ y^{k+1} &= \mathbf{P}_g\Big(y^{k+1} + \mathrm{D}_y K\big(z^{k+\frac{1}{2}}\big) - \widehat{\Sigma}_g\big(y^{k+1} - y^{k+\frac{1}{2}}\big) - \sigma^{-1}\mathcal{T}(y^{k+1} - y^k)\Big). \end{aligned}$$

From the nonexpansiveness of proximal mapping, we have

$$\|\mathbf{P}_f(x) - \mathbf{P}_f(x')\| \le \|x - x'\|, \quad \|\mathbf{P}_g(y) - \mathbf{P}_g(y')\| \le \|y - y'\|.$$

Therefore we obtain

$$\begin{aligned}
&\|R(z^{k+1})\|^2 \\
\le\ & \big\|\mathrm{D}_x K(x^{k+1}, y^{k+1}) - \mathrm{D}_x K\big(z^{k+\frac{1}{2}}\big) - \widehat{\Sigma}_f\big(x^{k+1} - x^{k+\frac{1}{2}}\big) - \sigma^{-1}\mathcal{S}(x^{k+1} - x^k)\big\|^2 \\
& + \big\| - \mathrm{D}_y K(x^{k+1}, y^{k+1}) + \mathrm{D}_y K\big(z^{k+\frac{1}{2}}\big) - \widehat{\Sigma}_g\big(y^{k+1} - y^{k+\frac{1}{2}}\big) - \sigma^{-1}\mathcal{T}(y^{k+1} - y^k)\big\|^2 \\
=\ & \big\|\big[\mathrm{D}_x K(z^{k+1}) - \mathrm{D}_x K\big(z^{k+\frac{1}{2}}\big)\big] - (\sigma^{-1}\mathcal{S} + \widehat{\Sigma}_f)\big(x^{k+1} - x^{k+\frac{1}{2}}\big) - \sigma^{-1}\mathcal{S}\big(x^{k+\frac{1}{2}} - x^k\big)\big\|^2 \\
& + \big\|\big[\mathrm{D}_y K\big(z^{k+\frac{1}{2}}\big) - \mathrm{D}_y K(z^{k+1})\big] - (\sigma^{-1}\mathcal{T} + \widehat{\Sigma}_g)\big(y^{k+1} - y^{k+\frac{1}{2}}\big) - \sigma^{-1}\mathcal{T}\big(y^{k+\frac{1}{2}} - y^k\big)\big\|^2 \\
\le\ & 3\eta_0^2\big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|^2 + 3\big\|x^{k+1} - x^{k+\frac{1}{2}}\big\|^2_{(\sigma^{-1}\mathcal{S}+\widehat{\Sigma}_f)^*(\sigma^{-1}\mathcal{S}+\widehat{\Sigma}_f)} + 3\big\|x^{k+\frac{1}{2}} - x^k\big\|^2_{\sigma^{-2}\mathcal{S}^*\mathcal{S}} \\
& + 3\eta_0^2\big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|^2 + 3\big\|y^{k+1} - y^{k+\frac{1}{2}}\big\|^2_{(\sigma^{-1}\mathcal{T}+\widehat{\Sigma}_g)^*(\sigma^{-1}\mathcal{T}+\widehat{\Sigma}_g)} + 3\big\|y^{k+\frac{1}{2}} - y^k\big\|^2_{\sigma^{-2}\mathcal{T}^*\mathcal{T}} \\
=\ & \big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|^2_{6\eta_0^2 I + 3(\widehat{\Sigma}+\sigma^{-1}\Theta)^*(\widehat{\Sigma}+\sigma^{-1}\Theta)} + \big\|z^{k+\frac{1}{2}} - z^k\big\|^2_{3\sigma^{-2}\Theta^*\Theta},
\end{aligned}$$

which implies the truth of the statement. □

Now we define

$$\mathrm{dist}_{\sigma\widehat{\Sigma}+\Theta}\,(z, \overline{\Omega}) = \inf_{z' \in \overline{\Omega}} \big\{\|z' - z\|_{\sigma\widehat{\Sigma}+\Theta}\big\}.$$

With the help of Proposition 4.1, we can prove the linear rate of convergence under the locally metrical subregularity of R at $(z^*, 0)$.

Theorem 4.1. *Let Assumptions* 3.1-4.1 *be satisfied at every point* $z^* \in \overline{\Omega}$. *Let* $\{z^k = (x^k, y^k)\}$ *and* $\{z^{k+1/2} = (x^{k+1/2}, y^{k+1/2})\}$ *be generated by Algorithm* 3.1. *Suppose that* $\mathcal{S}$ *and* $\mathcal{T}$ *satisfy* (3.8) *or equivalently* $\Theta \succeq 0$ *and* $\Theta + \sigma\widehat{\Sigma} \succ 0$. *Suppose also that* $\sigma, \mathcal{S}$ *and* $\mathcal{T}$ *satisfy*

$$\Theta \succ \widehat{\eta}_0 \sigma I. \tag{4.6}$$

Then $\{z^k = (x^k, y^k)\}$ *converges linearly with respect to some norm to an element of* $\overline{\Omega}$.

Proof. From Theorem 3.1, $\{z^k\}$ converges to an element of $\overline{\Omega}$, say $\overline{z}$. This indicates that

$$z^k \in \mathbf{B}(\overline{z}, \varepsilon_0) \subset \mathbf{B}(\overline{\Omega}, \varepsilon_0)$$

for any $k > N$, where N is a large integer.

Recall from the proof of Theorem 3.1 that $G(\sigma), N(\sigma)$ and $H(\sigma)$ are positively definite, where

$$G(\sigma) = \sigma\widehat{\Sigma} + \Theta, \quad N(\sigma) = \sigma\widehat{\Sigma} + \Theta - \widehat{\eta}_0\sigma I, \quad H(\sigma) = \Theta - \widehat{\eta}_0\sigma I.$$

Also recall from (3.25) that we have the following relation:

$$\|z^k - z^*\|^2_{G(\sigma)} \ge \|z^{k+1} - z^*\|^2_{G(\sigma)} + \big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|^2_{N(\sigma)} + \big\|z^{k+\frac{1}{2}} - z^k\big\|^2_{H(\sigma)}. \tag{4.7}$$

Since both $N(\sigma)$ and $H(\sigma)$ are positively definite, there must exist a positive number $\mu > 0$ such that

$$6\eta_0^2 I + 3\big(\widehat{\Sigma} + \sigma^{-1}\Theta\big)^*\big(\widehat{\Sigma} + \sigma^{-1}\Theta\big) \prec \mu N(\sigma), \quad 3\sigma^{-2}\Theta^*\Theta \prec \mu H(\sigma).$$

Then for $k > N$, from Assumption 4.1, we have from Proposition 4.1 for any $z^* \in \overline{\Omega}$ that

$$\begin{aligned}
&\operatorname{dist}_{G(\sigma)}(z^{k+1}, \overline{\Omega})^2 \\
&\le \lambda_{\max}\big(G(\sigma)\big)\operatorname{dist}(z^{k+1}, \overline{\Omega})^2 \le \kappa_0^2 \lambda_{\max}\big(G(\sigma)\big)\|R(z^{k+1})\|^2 \\
&\le \kappa_0^2 \lambda_{\max}\big(G(\sigma)\big)\Big[\big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|^2_{6\eta_0^2 I + 3(\widehat{\Sigma}+\sigma^{-1}\Theta)^*(\widehat{\Sigma}+\sigma^{-1}\Theta)} + \big\|z^{k+\frac{1}{2}} - z^k\big\|^2_{3\sigma^{-2}\Theta^*\Theta}\Big] \\
&\le \kappa_0^2 \mu \lambda_{\max}\big(G(\sigma)\big)\Big[\big\|z^{k+1} - z^{k+\frac{1}{2}}\big\|^2_{N(\sigma)} + \big\|z^{k+\frac{1}{2}} - z^k\big\|^2_{H(\sigma)}\Big] \\
&\le \kappa_0^2 \mu \lambda_{\max}\big(G(\sigma)\big)\Big[\|z^k - z^*\|^2_{G(\sigma)} - \|z^{k+1} - z^*\|^2_{G(\sigma)}\Big],
\end{aligned} \tag{4.8}$$

where the last inequality comes from (4.7). Taking z^* in (4.8) as

$$z^* = \arg\min\left\{\|z - z^k\|_{G(\sigma)} : z \in \overline{\Omega}\right\},$$

and noting that

$$\operatorname{dist}_{G(\sigma)}(z^{k+1}, \overline{\Omega})^2 \le \|z^{k+1} - z^*\|^2_{G(\sigma)},$$

we get from (4.8) that

$$\operatorname{dist}_{G(\sigma)}(z^{k+1}, \overline{\Omega})^2 + \kappa_0^2 \mu \lambda_{\max}\big(G(\sigma)\big)\|z^{k+1} - z^*\|^2_{G(\sigma)} \le \kappa_0^2 \mu \lambda_{\max}\big(G(\sigma)\big)\operatorname{dist}_{G(\sigma)}(z^k, \overline{\Omega})^2.$$

The above relation indicates that

$$\operatorname{dist}_{G(\sigma)}(z^{k+1}, \overline{\Omega}) \le \frac{1}{\sqrt{1 + \big[\kappa_0^2 \mu \lambda_{\max}\big(G(\sigma)\big)\big]^{-1}}}\operatorname{dist}_{G(\sigma)}(z^k, \overline{\Omega}).$$

This means that z^k converges to an element of $\overline{\Omega}$ with linear rate of convergence with respect to $G(\sigma)$-norm. The proof is complete. □

5. Numerical Experiments

In this section, we present some preliminary numerical experiments to illustrate the performance of Algorithm 3.1. All numerical experiments are implemented by MATLAB R2019a on a laptop with Intel(R) Core(TM) i5-6200U 2.30 GHz and 8 GB memory. We discuss the application of the majorized semi-proximal alternating coordinate method in four different forms of minimax optimization problems. The advantages of SPP are shown for two cases when f and g are smooth or nonsmooth.

5.1. Smooth saddle point problems

When f and g are smooth, the minimax optimization problem can be abbreviated as the following form:

$$\min_{x \in \Re^n} \max_{y \in \Re^m} K(x, y). \tag{5.1}$$

Many methods can solve saddle point problems, such as the proximal point (PP) method (see [17]), the optimistic gradient descent ascent (OGDA) method and the extra-gradient (EG) method (see [12]). We compare the proposed method, SPP, with these methods for the linear regression problem. For solving problem (5.1), the updating formula of SPP can be simplified as

$$\begin{aligned}
z^{k+\frac{1}{2}} &= -\sigma(\sigma\widehat{\Sigma} + \Theta)^{-1}\widetilde{\mathrm{D}}K(z^k) + z^k, \\
z^{k+1} &= (\sigma\widehat{\Sigma} + \Theta)^{-1}\big[-\widetilde{\mathrm{D}}K\big(z^{k+\frac{1}{2}}\big) + \sigma\widehat{\Sigma}\big(z^{k+\frac{1}{2}}\big) + \Theta(z^k)\big],
\end{aligned} \tag{5.2}$$

where $z^k = (x^k, y^k)$ and $\widehat{\Sigma}$ and Θ are two linear operators defined in (3.3) and (3.9), respectively. The operator $\widetilde{\mathrm{D}}K(\cdot)$ is expressed as

$$\widetilde{\mathrm{D}}K(z) = \begin{pmatrix} \mathrm{D}_x K(x,y) \\ -\mathrm{D}_y K(x,y) \end{pmatrix}$$

for any $z = (x, y) \in \Re^n \times \Re^m$.

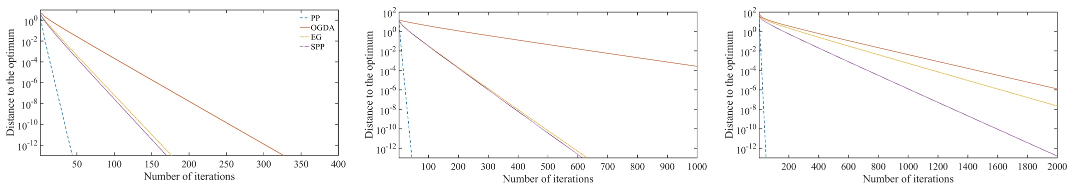
The saddle point reformulation of the linear regression is of the form

$$\min_{x\in\Re^n} \max_{y\in\Re^m} \frac{1}{m}\left[-\frac{1}{2}\|y\|^2 - b^T y + y^T Ax\right] + \frac{\lambda}{2}\|x\|^2. \tag{5.3}$$

We set $n = m$ and the rows of the matrix A are generated by a Gaussian distribution $\mathcal{N}(0, I_n)$. Let $b = 0$ and $\lambda = 1/m$. In PP, EG and OGDA, parameters and step sizes are selected for best performance. In SPP, we set $\sigma = 1$. The linear operators are chosen as $\mathcal{S} = \|A\|_2 I_n$, $\mathcal{T} = \|A\|_2 I_m$, $\widehat{\Sigma}_f = 0.5m\lambda I_n$ and $\widehat{\Sigma}_g = 0.5m\lambda I_m$.

In Fig. 5.1, we compare the performances of the four methods with respect to the number of iterations when the dimension varies from $n = 10, 100, 1000$. The same initial point x^0 is chosen. Generally speaking, all the four methods converge linearly to the optimal solution, and the proximal point (PP) method has the best performance. Our method, SPP, is the second best, which converges faster than EG and OGDA. It can be observed that for the low-dimensional strongly convex-strongly concave saddle point problem (5.3), the convergence rate of EG is very close to SPP. When the dimension n increases, the performance of SPP becomes much better then EG and OGDA.

It is reasonable to explain the best performance of PP because the proximal point method is asymptotically superlinear and it has an explicit solution for every subproblem when solving this simple problem.



(a) n=10 (b) n=100 (c) n=1000

Fig. 5.1. Compare SPP, PP, EG and OGDA in terms of number of iterations under different dimensions for the linear regression. Step sizes of EG and OGDA are tuned for best performance.

5.2. Convex-concave minimax optimization problems with ∞-norm

In this part, we focus on the minimax optimization problem of the form

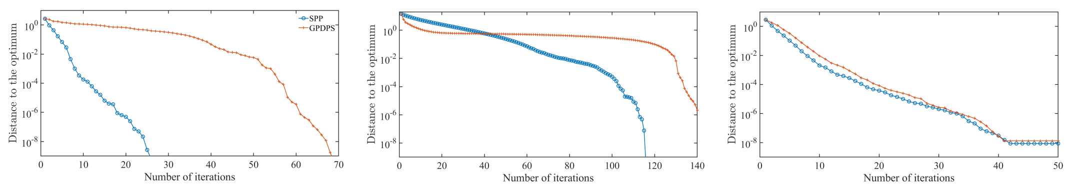
$$\min_{x\in X} \max_{y\in Y} \mu_x\|x\|_\infty + \frac{\lambda}{2}\|x\|^2 + \frac{1}{m}\left[-\frac{1}{2}\|y\|^2 - b^T y + y^T Ax\right] - \mu_y\|y\|_\infty, \tag{5.4}$$

where $X \in \Re^n$ and $Y \in \Re^m$ are two convex sets and $\|\cdot\|_\infty$ represents the infinite norm of a finite-dimensional vector space.

We can not use EG and OGDA to solve problem (5.4) as there is a nonsmooth term ∞-norm in the minimax optimization problem (5.4). Hence, we compare our algorithm and GPDPS in [8] for the problem (5.4). GPDPS was proposed for solving the nonsmooth convex minimax problem, which combined the proximal point method with linearization techniques. Different from GPDPS, our algorithm introduces a semi-proximal term during the computational iterations and the convergence advantages of the semi-proximal term will be shown in the next numerical experiments. As SPP and GPDPS can not provide the explicit solutions of the subproblems, we use Matlab code fminunc to solve the subproblems.

Unconstrained convex minimax optimization problem. We set $n = m, b = 0, \lambda = 1/m$ and $\mu_x = \mu_y = 1$. In this case, the linear operator $\mathcal{S} = \mathcal{T} = \|A\|_2 I_n$ and $\widehat{\Sigma}_f = \widehat{\Sigma}_g = 0.5m\lambda I_n$. The convergence of SPP is shown in Table 5.1, with respect to the number of iterations and CPU time under different dimensions, different condition numbers of matrix A, and different values of the parameter σ. It is easy to see that the convergence rate becomes significantly slower as the dimension n increases. However, the condition number of matrix A has very little effect on the convergence of SPP. On the other hand, within the same CPU time, the convergence of SPP becomes faster as σ increases.

The comparison between our algorithm and GPDPS for unconstrained minimax problem (5.4) is presented in Fig. 5.2. Same initial point (x^0, y^0) is chosen for them. Generally, GPDPS performs slower than SPP within the same iterations under different dimensions, which further illustrates that the semi-proximal terms used in SPP accelerates the convergence of iterations.



(a) n=50 (b) n=100 (b) n=100

Fig. 5.2. Compare SPP and GPDPS in terms of number of iterations under different dimensions for (5.4).

Convex minimax optimization problem with convex constraints. We tested three cases as follows:

Case 1. Linear equality constraints:

$$X = \{x \in \Re^n : B_x x = b_{ex}\}, \quad Y = \{y \in \Re^m : B_y y = b_{ey}\}.$$

Case 2. Linear inequality constraints:

$$X = \{x \in \Re^n : M_x x \leq b_{ix}\}, \quad Y = \{y \in \Re^m : M_y y \leq b_{iy}\}.$$

Case 3. Quadratic constraints:

$$X = \{x \in \Re^n : x^T Q_x x + c_x^T x + b_{qx} \leq 0\}, \quad Y = \{y \in \Re^m : y^T Q_y y + c_y^T y + b_{qy} \leq 0\},$$

where Q_x and Q_y are positive semi-definite matrices.

We rewrite problem (5.4) as

$$\min_{x \in \Re^n} \max_{y \in \Re^m} \delta_X(x) + \mu_x \|x\|_\infty + \frac{\lambda}{2}\|x\|^2 + \frac{1}{m}\left[-\frac{1}{2}\|y\|^2 - b^T y + y^T A x\right] - \mu_y \|y\|_\infty + \delta_Y(y). \quad (5.5)$$

Table 5.1: Numerical results of SPP for the minimax problem (5.4). $n = m = \dim x$; κ – condition number of A; σ – the parameter of SPP; ϵ – the relative error (i.e., $\|(x^k; y^k)\|_2/\|(x^0; y^0)\|_2$); T – number of iterations; t – CPU time.

n	κ	σ	$\epsilon = 10^{-1}$		$\epsilon = 10^{-3}$		$\epsilon = 10^{-5}$		$\epsilon = 10^{-7}$		$\epsilon = 10^{-9}$	
			T	t(s)	T	t(s)	T	t(s)	T	t(s)	T	t(sec)
10	10	1	4	1.43	7	2.01	7	2.01	8	2.15	8	2.15
		0.1	30	3.19	60	6.13	62	6.53	62	6.53	63	6.85
	50	1	4	2.44	7	3.47	8	3.80	8	3.80	9	4.11
		0.1	32	5.84	63	12.56	65	13.11	65	13.11	66	13.41
	$2 * 10^2$	1	4	1.30	6	1.87	7	2.20	7	2.20	8	2.52
		0.1	31	4.58	53	8.73	55	9.01	55	9.01	56	9.17
50	10^2	1	2	1.90	7	8.23	11	17.21	17	24.02	22	26.44
		0.1	19	6.47	38	37.23	49	49.10	54	52.22	59	54.62
	10^3	1	2	2.87	7	10.00	11	13.60	17	18.20	23	21.04
		0.1	8	6.33	37	54.97	47	74.96	52	83.85	56	87.36
	$5 * 10^3$	1	2	3.88	7	11.28	12	21.17	18	26.79	25	33.18
		0.1	8	5.82	37	48.94	47	70.12	52	79.52	57	83.12
10^2	10^2	1	4	2.40	58	133.88	99	304.28	107	331.25	114	346.78
		0.1	55	17.98	648	921.93	871	1858.38	889	1933.12	895	1944.91
	10^3	1	4	2.43	57	119.91	92	251.12	100	275.27	107	286.13
		0.1	54	14.28	635	741.39	867	1514.54	886	1559.01	893	1570.01
	10^4	1	4	3.80	58	120.22	97	250.40	110	282.20	115	290.32
		0.1	56	15.28	651	744.71	870	1446.34	894	1499.63	900	1508.61
200	10^2	1	2	11.73	8	82.27	18	192.83	29	287.45	40	340.61
		0.1	10	46.55	43	390.50	90	923.73	106	1050.30	113	1079.66
	10^3	1	3	21.17	8	76.97	17	173.58	28	285.79	35	325.52
		0.1	10	53.34	43	394.46	87	889.48	102	1026.98	109	1061.02
	10^5	1	3	25.85	8	82.84	18	190.43	34	335.50	42	370.18
		0.1	11	59.58	48	452.18	98	1015.31	113	1154.47	120	1181.84

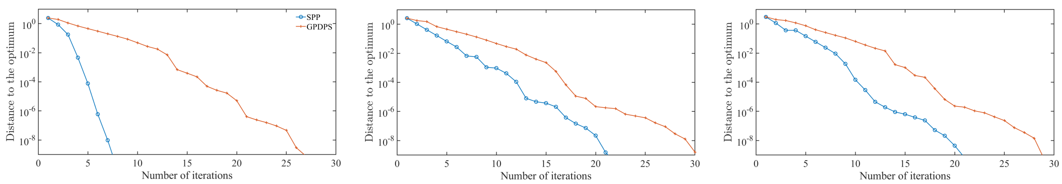
For simplicity, let $b = 0, \lambda = 1/m$ and $\mu_x = \mu_y = 1$. We consider the situation when $n \neq m$ in problem (5.5). The linear operators are selected as $\mathcal{S} = \|A\|_2 I_n, \mathcal{T} = \|A\|_2 I_m, \widehat{\Sigma}_f = 0.5m\lambda I_n$ and $\widehat{\Sigma}_g = 0.5m\lambda I_m$ and the parameter $\sigma = 1$.

The performance of SPP under three different constraints is shown in Table 5.2. We can see from Table 5.2 that, under each of the three different constraints, SPP converges to the optimal solution of the problem (5.5) within a few number of iterations. Furthermore, it can be observed that as the dimensions n and m increase, the CPU time for implementing SPP increases rapidly.

We next compare our algorithm and GPDPS under three different constraints. We choose same initial point (x^0, y^0) for them in minimax problem (5.5). In Fig. 5.3, SPP decreases more rapidly than GPDPS in general for all cases, which shows advantages of the semi-proximal terms in SPP.

Table 5.2: Numerical results of SPP for the minimax problem (5.5). $n = \dim x$, $m = \dim y$; ϵ – the relative error (i.e., $\|(x^k; y^k)\|_2/\|(x^0; y^0)\|_2$); T – number of iterations; t – CPU time.

$n,\ m$	Case	$\epsilon = 10^{-1}$		$\epsilon = 10^{-3}$		$\epsilon = 10^{-5}$		$\epsilon = 10^{-7}$		$\epsilon = 10^{-9}$	
		T	t(sec)	T	t(sec)	T	t(sec)	T	t(sec)	T	t(sec)
$n = 10$ $m = 10$	Case 1.	3	2.01	8	4.85	9	5.34	9	5.34	10	5.65
	Case 2.	3	2.44	11	5.92	14	6.81	16	7.18	17	7.30
	Case 3.	3	2.55	11	6.34	13	7.07	15	7.57	17	7.80
$n = 20$ $m = 50$	Case 1.	2	4.63	4	8.16	5	9.27	6	11.86	7	12.55
	Case 2.	3	3.63	7	7.43	13	14.44	17	18.80	21	22.49
	Case 3.	3	4.50	8	10.97	11	15.65	14	20.07	20	22.84
$n = 50$ $m = 20$	Case 1.	3	3.59	5	4.78	5	4.78	7	7.10	8	8.25
	Case 2.	3	5.15	8	11.47	12	19.30	18	26.97	23	32.37
	Case 3.	3	5.40	9	11.36	13	19.83	21	30.18	25	34.00
$n = 100$ $m = 100$	Case 1.	3	10.79	8	28.80	15	52.63	23	74.78	30	84.57
	Case 2.	3	9.59	8	28.35	14	53.23	23	75.96	31	86.40
	Case 3.	3	13.98	8	32.73	14	59.54	23	81.24	29	91.86
$n = 100$ $m = 300$	Case 1.	2	19.09	7	100.39	15	252.69	26	385.63	32	424.86
	Case 2.	2	18.41	8	124.50	17	281.16	28	439.69	34	493.03
	Case 3.	2	16.31	8	113.01	13	200.53	23	333.90	30	383.02
$n = 200$ $m = 200$	Case 1.	3	26.75	8	101.04	18	240.70	31	387.49	38	428.70
	Case 2.	3	24.78	8	90.06	20	245.65	33	398.40	41	439.93
	Case 3.	3	21.69	8	83.61	17	190.40	28	305.33	34	339.87
$n = 300$ $m = 100$	Case 1.	2	21.59	8	124.71	16	260.36	24	370.11	29	399.34
	Case 2.	2	18.24	7	101.87	14	227.72	24	384.40	31	431.59
	Case 3.	2	18.68	8	115.78	15	239.84	32	492.24	40	551.52



(a) Linear equality constraints (b) Linear inequality constraints (c) Quadratic constraints

Fig. 5.3. Compare SPP and GPDPS in terms of number of iterations under different constraints for (5.5).

6. Some Concluding Remarks

Nonsmooth convex-concave minimax optimization problems are an important class of optimization problems with many applications. However, there are few numerical algorithms for solving this type of problems when the smooth parts in the objective function are not bilinear. We developed a semi-proximal point method (SPP) for solving a nonsmooth convex-concave minimax problem of the form (1.1). We demonstrated the global convergence of the algorithm

SPP under mild assumptions without requiring strong convexity-concavity condition and the linear rate of convergence under the locally metrical subregularity of the solution mapping. Preliminary numerical results have been reported, which shows the efficiency of the proposed SPP method.

There are many interesting problems worth consideration. In this paper, we only tested SPP for two types of examples. How is the performance of SPP for other types of minimax optimization problems? Presently we only considered convex-concave minimax problems of the form (1.1); i.e., $K(\cdot,\cdot)$ is required to be smooth convex-concave. How to construct efficient numerical algorithms for solving the nonsmooth minimax problem (1.1) when $K(\cdot,\cdot)$ is not a smooth convex-concave function?

Acknowledgments. The authors are very grateful to Dr. W. Zhang for mentioning us the important reference [22].

This research was partially supported by the Natural Science Foundation of China (Grant Nos. 11991021, 11991020, 12021001, 11971372, 11971089, 11731013), by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDA27000000) and by the National Key R&D Program of China (Grant Nos. 2021YFA1000300, 2021YFA1000301).

References

[1] K.J. Arrow, L. Hurwicz, and H. Uzawa, *Studies in Linear and Non-linear Programming*, Stanford University Press, 1958.

[2] J.F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, 2000.

[3] R.I. Boţ, E.R. Csetnek, and M. Sedlmayer, An accelerated minimax algorithm for convex-concave saddle point problems with nonsmooth coupling function, *Comput. Optim. Appl.*, (2022), `https://doi.org/10.1007/s10589-022-00378-8`.

[4] A. Chambolle and T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, *J. Math. Imaging Vis.*, **40** (2011), 120–145.

[5] A. Chambolle and T. Pock, On the ergodic convergence rates of a first-order primal-dual algorithm, *Math. Program.*, **159** (2016), 253–287.

[6] A. Chambolle and T. Pock, An introduction to continuous optimization for imaging, *Acta Numer.*, **25** (2016), 161–319.

[7] Y. Chen, G. Lan, and Y. Ouyang, Optimal primal-dual methods for a class of saddle point problems, *SIAM J. Optim.*, **24**:4 (2014), 1779–1814.

[8] C. Clason, S. Mazurenko, and T. Valkonen, Primal-dual proximal splitting and generalized conjugation in non-smooth non-convex optimization, *Appl. Math. Optim.*, **84**:2 (2021), 1239–1284.

[9] E.Y. Hamedani and N.S. Aybat, A primal-dual algorithm with line search for general convex-concave saddle point problems, *SIAM J. Optim.*, **31**:2 (2021), 1299–1329.

[10] G. Korpelevich, The extragradient method for finding saddle points and other problems, *Matecon*, **12** (1976), 747–756.

[11] T. Lin, C. Jin, and M.I. Jordan, Near-optimal algorithms for minimax optimization, in: Proceedings of Thirty Third Conference on Learning Theory, *Proceedings of Machine Learning Research*, **125** (2020), 2738–2779.

[12] A. Mokhtari, A. Ozdaglar, and S. Pattathil, A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach, in: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, *Proceedings of Machine Learning Research*, **108** (2020), 1497–1507.

[13] A. Mokhtari, A.E. Ozdaglar, and S. Pattathil, Convergence rate of $O(1/k)$ for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems, *SIAM J. Optim.*, **30**:4 (2020), 3230–3251.

[14] A. Nemirovski, Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems, *SIAM J. Optim.*, **15**:1 (2004), 229–251.

[15] Y. Nesterov, Dual extrapolation and its applications to solving variational inequalities and related problems, *Math. Program.*, **109**:2-3 (2007), 319–344.

[16] R.T. Rockafellar, *Convex Analysis*, Vol. 18, Princeton University Press, 1970.

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

[18] R.T. Rockafellar and J.B. Wets, *Variational Analysis*, Springer Science and Business Media, 2004.

[19] M. Sibony, Méthodes itératives pour les équations et iné equations aux dérivées partielles non linéaires de type monotone, *Calcolo*, **7**:1-2 (1970), 65–183.

[20] P. Tseng, On linear convergence of iterative methods for the variational inequality problem, *J. Comput. Appl. Math.*, **60**:1-2 (1995), 237–252.

[21] P. Tseng, On accelerated proximal gradient methods for convex-concave optimization, (2008), `https://www.mit.edu/~dimitrib/PTseng/papers/apgm.pdf`.

[22] T. Valkonen, A primal-dual hybrid gradient method for nonlinear operators with applications to MRI, *Inverse Problems*, **30**:5 (2014), 055012.

[23] Y. Wang and J. Li, Improved algorithms for convex-concave minimax optimization, *Adv. Neural Inf. Process. Syst.*, **33** (2020), 4800–4810.

[24] J. Yang, S. Zhang, N. Kiyavash, and N. He, A catalyst framework for minimax optimization, *Adv. Neural Inf. Process. Syst.*, **33** (2020), 5667–5678.

[25] T. Yoon and E.K. Ryu, Accelerated algorithms for smooth convex-concave minimax problems with $O(1/k^2)$ rate on squared gradient norm, in: Proceedings of the 38th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, **139** (2021), 12098–12109.