AN ACCELERATED INEXACT PROXIMAL POINT METHOD FOR
SOLVING NONCONVEX-CONCAVE MIN-MAX PROBLEMS

WEIWEI KONG* AND RENATO D.C. MONTEIRO*

Abstract. This paper presents smoothing schemes for obtaining approximate stationary points
of unconstrained or linearly-constrained composite nonconvex-concave min-max (and hence non-
smooth) problems by applying well-known algorithms to composite smooth approximations of the
original problems. More specifically, in the unconstrained (resp. constrained) case, approximate
stationary points of the original problem are obtained by applying, to its composite smooth approx-
imation, an accelerated inexact proximal point (resp. quadratic penalty) method presented in a
previous paper by the authors. Iteration complexity bounds for both smoothing schemes are also
established. Finally, numerical results are given to demonstrate the efficiency of the unconstrained
smoothing scheme.

Key words. quadratic penalty method, composite nonconvex problem, iteration-complexity,
inexact proximal point method, first-order accelerated gradient method, minimax problem.

AMS subject classifications. 47J22, 90C26, 90C30, 90C47, 90C60, 65K10.

1. Introduction. The first goal of this paper is to present and study the com-
plexity of an accelerated inexact proximal point smoothing (AIPP-S) scheme for fin-
ding approximate stationary points of the (potentially nonsmooth) min-max compos-
ite nonconvex optimization (CNO) problem

\[ \min_{x \in X} \{ \hat{p}(x) := p(x) + h(x) \} \]

where \( h \) is a proper lower-semicontinuous convex function, \( X \) is a nonempty convex
set, and \( p \) is a max function given by

\[ p(x) := \max_{y \in Y} \Phi(x, y) \quad \forall x \in X, \]

for some nonempty compact convex set \( Y \) and function \( \Phi \) which, for some scalar
\( m > 0 \) and open set \( \Omega \supseteq X \), is such that: (i) \( \Phi \) is continuous on \( \Omega \times Y \); (ii) the
function \( -\Phi(x, \cdot) : Y \to \mathbb{R} \) is lower-semicontinuous and convex for every \( x \in X \); and
(ii) for every \( y \in Y \), the function \( \Phi(\cdot, y) + m\| \cdot \|_2^2 \) is convex, differentiable, and its
gradient is Lipschitz continuous on \( X \times Y \). Here, the objective function is the sum of a
convex function \( h \) and the pointwise supremum of (possibly nonconvex) differentiable
functions which is generally a (possibly nonconvex) nonsmooth function.

When \( Y \) is a singleton, the max term in (1.1) becomes smooth and (1.1) reduces
to a smooth CNO problem for which many algorithms have been developed in the
literature. In particular, accelerated inexact proximal points (AIPP) methods, i.e.
methods which use an accelerated composite gradient variant to approximately solve
the generated sequence of prox subproblems, have been developed for it (see, for
example, [5,16]). When \( Y \) is not a singleton, (1.1) can no longer be directly solved by
an AIPP method due to the nonsmoothness of the max term. The AIPP-S scheme
developed in this paper is instead based on a perturbed version of (1.1) in which the
max term in (1.1) is replaced by a smooth approximation and the resulting smooth
CNO problem is solved by an AIPP method.

*School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA,
30332-0205. (E-mails: wkong37@gatech.edu & monteiro@isye.gatech.edu). The works of these
authors were partially supported by ONR Grant N00014-18-1-2077.
Throughout our presentation, it is assumed that efficient oracles for evaluating the quantities $\Phi(x,y)$, $\nabla_x \Phi(x,y)$, and $h(x)$ and for obtaining exact solutions of the problems

$$
\min_{x \in X} \left\{ \lambda h(x) + \frac{1}{2} \|x - x_0\|^2 \right\}, \quad \max_{y \in Y} \left\{ \lambda \Phi(x_0,y) - \frac{1}{2} \|y - y_0\|^2 \right\}
$$

for any $(x_0,y_0)$ and $\lambda > 0$, are available. Throughout this paper, the terminology “oracle call” is used to refer to a collection of the above oracles of size $O(1)$ where each of them appears at least once. We refer to the computation of the solution of the first problem above as a $h$-resolvent evaluation. In this manner, the computation of the solution of the second one is a $[-\Phi(x_0,\cdot)]$-resolvent evaluation.

We first develop an AIPP-S scheme that obtains a stationary point based on a primal-dual formulation of (1.1). More specifically, given a tolerance pair $(\rho_x, \rho_y) \in \mathbb{R}_{++}$, it is shown that an instance of this scheme obtains a quadruple $(\tilde{u}, \tilde{v}, \tilde{x}, \tilde{y})$ such that

$$
(1.4) \quad \left( \begin{array}{l}
\tilde{u} \\
\tilde{v}
\end{array} \right) \in \left( \begin{array}{l}
\nabla_x \Phi(\tilde{x}, \tilde{y}) \\
0
\end{array} \right) + \left( \begin{array}{l}
\partial h(\tilde{x}) \\
\partial [-\Phi(\tilde{x}, \cdot)](\tilde{y})
\end{array} \right), \quad \|\tilde{u}\| \leq \rho_x, \quad \|\tilde{v}\| \leq \rho_y
$$

in $O(\rho_x^{-2} \rho_y^{-1/2})$ oracle calls, where $\partial \phi(z)$ is the subdifferential of a convex function $\phi$ at a point $z$ (see (1.9) with $\varepsilon = 0$). We then show that another instance of this scheme can obtain an approximate stationary point based on the directional derivative of $\hat{p}$. More specifically, given a tolerance pair $\delta > 0$, it is shown that this instance computes a point $x \in X$ such that

$$
(1.5) \quad \exists \hat{x} \in X \text{ s.t. } \inf_{\|d\| \leq 1} \hat{p}'(\hat{x}; d) \geq -\delta, \quad \|\hat{x} - x\| \leq \delta,
$$

in $O(\delta^{-3})$ oracle calls, where $\hat{p}'(x; d)$ is the directional derivative of $\hat{p}$ at the point $x$ along the direction $d$ (see (1.10)).

The second goal of this paper is to develop a quadratic penalty AIPP-S (QP-AIPP-S) scheme to obtain approximate stationary points of a linearly constrained version of (1.1), namely

$$
(1.6) \quad \min_{x \in X} \{ p(x) + h(x) : A x = b \}
$$

where $p$ is as in (1.2), $A$ is a linear operator, and $b$ is in the range of $A$. The scheme is a penalty-type method which approximately solves a sequence of penalty subproblems of the form

$$
(1.7) \quad \min_{x \in X} \left\{ p(x) + h(x) + \frac{c}{2} \|A x - b\|^2 \right\}
$$

for an increasing sequence of positive penalty parameters $c$. Similar to the approach used for the first goal of this paper, the method considers a perturbed variant of (1.7) in which the objective function is replaced by a smooth approximation and the resulting problem is solved by the quadratic-penalty AIPP (QP-AIPP) method proposed in [16]. For a given tolerance triple $(\rho_x, \rho_y, \eta) \in \mathbb{R}_{++}^3$, it is shown that the method computes a quintuple $(\tilde{u}, \tilde{v}, \tilde{x}, \tilde{y}, \tilde{r})$ satisfying

$$
(1.8) \quad \left( \begin{array}{l}
\tilde{u} \\
\tilde{v}
\end{array} \right) \in \left( \begin{array}{l}
\nabla_x \Phi(\tilde{x}, \tilde{y}) + A^* r \\
0
\end{array} \right) + \left( \begin{array}{l}
\partial h(\tilde{x}) \\
\partial [-\Phi(\tilde{x}, \cdot)](\tilde{y})
\end{array} \right), \quad \|\tilde{u}\| \leq \rho_x, \quad \|\tilde{v}\| \leq \rho_y, \quad \|A \tilde{x} - b\| \leq \eta.
$$
Finally, it is worth mentioning that all of the above complexities are obtained under the mild assumption that the optimal value in each of the respective optimization problems, namely (1.1) and (1.6) is bounded below. Moreover, it is neither assumed that $X$ be bounded nor that (1.1) or (1.6) has an optimal solution.

Related Works. Since the case when $\Phi(\cdot, \cdot)$ in (1.1) is convex-concave has been well-studied in the literature (see, for example, [1, 12, 14, 21, 22, 23, 27]), we will make no more mention of it here. Instead, we will focus on papers that consider (1.1) where $\Phi(\cdot, y)$ is differentiable and nonconvex for every $y \in Y$ and there are mild conditions on $\Phi(x, \cdot)$ for every $x \in X$.

Denoting $\delta_C$ to be the indicator function of a closed convex set $C \subseteq X$ (see Subsection 1.1) and $\delta = \min\{\rho_x, \rho_y\}$, we present Tables 1.1 and 1.2, which compare our contributions to past [24, 26] and subsequent [17, 25, 30] works. It is worth mentioning that the above works consider termination conditions that are slightly different than the ones in this paper. In Subsection 2.1, we show that they are equivalent to the ones in this paper up to a multiplicative constant that is independent of the tolerances, i.e., $\rho_x, \rho_y, \delta$.

### Table 1.1

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Oracle Complexity</th>
<th>Use Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGSF [24]</td>
<td>$O(\rho^{-3})$</td>
<td>$D_h = \infty$ $h \equiv 0$ $h \equiv \delta_C$ General $h$</td>
</tr>
<tr>
<td>Minimax-PPA [17]</td>
<td>$O(\rho^{-2} \log(\rho^{-1}))$</td>
<td>$\checkmark$ $\checkmark$ $\checkmark$ $\times$</td>
</tr>
<tr>
<td>FNE Search [25]</td>
<td>$O(\rho^{-2} \rho_y^{-1/2} \log(\rho^{-1}))$</td>
<td>$\checkmark$ $\checkmark$ $\checkmark$ $\times$</td>
</tr>
<tr>
<td>AIPP-S</td>
<td>$O(\rho^{-2} \rho_y^{-1/2})$</td>
<td>$\checkmark$ $\checkmark$ $\checkmark$ $\checkmark$</td>
</tr>
</tbody>
</table>

Comparison of iteration complexities and possible use cases under notions equivalent to (1.4) with $\rho := \min\{\rho_x, \rho_y\}$.

### Table 1.2

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Oracle Complexity</th>
<th>Use Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>PG-SVRG [26]</td>
<td>$O(\delta^{-6} \log(\delta^{-1}))$</td>
<td>$D_h = \infty$ $h \equiv 0$ $h \equiv \delta_C$ General $h$</td>
</tr>
<tr>
<td>Minimax-PPA [17]</td>
<td>$O(\delta^{-3} \log^2(\delta^{-1}))$</td>
<td>$\checkmark$ $\checkmark$ $\checkmark$ $\times$</td>
</tr>
<tr>
<td>Prox-DIAG [30]</td>
<td>$O(\delta^{-3} \log^2(\delta^{-1}))$</td>
<td>$\checkmark$ $\checkmark$ $\times$ $\times$</td>
</tr>
<tr>
<td>AIPP-S</td>
<td>$O(\delta^{-3})$</td>
<td>$\checkmark$ $\checkmark$ $\checkmark$ $\checkmark$</td>
</tr>
</tbody>
</table>

Comparison of iteration complexities and possible use cases under notions equivalent to (1.5).

To the best of our knowledge, this work is the first one to analyze the complexity of a smoothing scheme for finding approximate stationary points of (1.6).

**Organization of the paper.** Subsection 1.1 presents notation and some basic definitions that are used in this paper. Subsection 1.2 presents several motivating applications that are of the form in (1.1). Section 2 is divided into two subsections. The first one precisely states the assumptions underlying problem (1.1) and discusses four notions of stationary points. The second one presents a smooth approximation of the function $p$ in (1.1). Section 3 is divided into two subsections. The first one reviews the AIPP method in [16] and its iteration complexity. The second one presents the
AIPP-S scheme its iteration complexities for finding approximate stationary points as in (1.4) and (1.5). Section 4 is also divided into two subsections. The first one reviews the QP-AIPP method in [16] and its iteration complexity. The second one presents the QP-AIPP-S scheme its iteration complexity for finding points satisfying (1.8). Section 5 presents some computational results. Section 6 gives some concluding remarks. Finally, several appendices at the end of this paper contain proofs of technical results needed in our presentation.

1.1. Notation and basic definitions. This subsection provides some basic notation and definitions.

The set of real numbers is denoted by \( \mathbb{R} \). The set of non-negative real numbers and the set of positive real numbers is denoted by \( \mathbb{R}_+ \) and \( \mathbb{R}_{++} \) respectively. The set of natural numbers is denoted by \( \mathbb{N} \). For \( t > 0 \), define \( \log^+(t) := \max\{1, \log(t)\} \).

Let \( \mathbb{R}^n \) denote a real-valued \( n \)-dimensional Euclidean space with standard norm \( \| \cdot \| \).

Given a linear operator \( A : \mathbb{R}^n \to \mathbb{R}^p \), the operator norm of \( A \) is denoted by \( \|A\| := \sup\{\|Az\|/\|z\| : z \in \mathbb{R}^n, z \neq 0\} \).

The following notation and definitions are for a general complete inner product space \( Z \), whose inner product and its associated induced norm are denoted by \( \langle \cdot, \cdot \rangle \) and \( \| \cdot \| \) respectively. Let \( \psi : Z \to (-\infty, \infty] \) be given. The effective domain of \( \psi \) is denoted as \( \text{dom} \psi := \{z \in Z : \psi(z) < \infty\} \) and \( \psi \) is said to be proper if \( \text{dom} \psi \neq \emptyset \).

The set of proper, lower semi-continuous, convex functions \( \psi : Z \to (-\infty, \infty] \) is denoted by \( \text{Conv}(Z) \). Moreover, for a convex set \( Z \subseteq \mathbb{R}^n \), we denote \( \text{Conv}(Z) \) to be a set of functions in \( \text{Conv}(Z) \) whose effective domain is equal to \( Z \). For \( \varepsilon \geq 0 \), the \( \varepsilon \)-subdifferential of \( \psi \in \text{Conv}(Z) \) at \( z \in \text{dom} \psi \) is denoted by

\[
\partial_{\varepsilon} \psi(z) := \{w \in \mathbb{R}^n : \psi(z') \geq \psi(z) + \langle w, z' - z \rangle - \varepsilon, \forall z' \in Z\},
\]

and we denote \( \partial \psi \equiv \partial_0 \psi \). The directional derivative of \( \psi \) at \( z \in Z \) in the direction \( d \in Z \) is denoted by

\[
\psi'(z;d) := \lim_{t \to 0} \frac{\psi(z + td) - \psi(z)}{t}.
\]

It is well-known that if \( \psi \) is differentiable at \( z \in \text{dom} \psi \), then for a given direction \( d \in Z \) we have \( \psi'(z;d) = \langle \nabla \psi(z), d \rangle \).

For a given \( Z \subseteq \mathbb{R}^n \), the indicator function of \( Z \), denoted by \( \delta_Z \), is defined as \( \delta_Z(z) = 0 \) if \( z \in Z \) and \( \delta_Z(z) = \infty \) if \( z \notin Z \). Moreover, the closure, interior, and relative interior of \( Z \) are denoted by \( \text{cl} Z \), \( \text{int} Z \), and \( \text{ri} Z \), respectively. The support function of \( Z \) at a point \( z \) is denoted by \( \sigma_Z(z) := \sup_{z' \in Z} \langle z, z' \rangle \).

1.2. Motivating applications. This subsection lists motivating applications that are of the form in (1.1). In Section 5, we examine the performance of our proposed smoothing scheme on some special instances of these applications.

1.2.1. Maximum of a finite number of nonconvex functions. Given a family of functions \( \{f_i\}_{i=1}^k \) that are continuously differentiable everywhere with Lipschitz continuous gradients, define the (possibly) nonsmooth function \( f_{\max}(x) := \max_{1 \leq i \leq k} f_i(x) \) for every \( x \in \mathbb{R}^n \) and let \( C \subseteq \mathbb{R}^n \) be a closed convex set. The problem of interest is the minimization of \( f_{\max} \) over the set \( C \), i.e.,

\[
\min_{x \in \mathbb{R}^n} \{f_{\max}(x) : x \in C\},
\]
1.2.2. Robust regression. Given a set of observations \( \sigma := \{ \sigma_i \}_{i=1}^n \) and a compact convex set \( \Theta \in \mathbb{R}^k \), let \( \{ \ell_\theta(x|\sigma) \}_{\theta \in \Theta} \) be a family of nonconvex loss functions in which: (i) \( \ell_\theta(x|\sigma) \) is concave in \( \theta \) for every \( x \in \mathbb{R}^n \); and (ii) \( \ell_\theta(x|\sigma) \) is continuously differentiable in \( x \) with Lipschitz continuous gradient for every \( \theta \in \Theta \). The problem of interest is to minimize the worst-case loss in \( \Theta \), i.e.,

\[
\min_{x \in \mathbb{R}^n} \max_{\theta \in \Theta} \ell_\theta(x|\sigma),
\]

which is clearly an instance of (1.1) where

\[
Y = \Theta, \quad \Phi(x, y) = \ell_y(x|\sigma), \quad h(x) = 0.
\]

1.2.3. Min-max games with an adversary. Let \( \{ U_j(x_1, ..., x_k, y) \}_{j=1}^k \) be a set of utility functions in which: (i) \( U_j \) is nonconvex and continuously differentiable in its first \( k \) arguments, but concave in its last argument; (ii) \( \nabla_x U_j(x_1, ..., x_k, y) \) is Lipschitz continuous for every \( 1 \leq i \leq k \). Given input constraint sets \( \{ B_i \}_{i=1}^k \) and \( B_y \), the problem of interest is to maximize the total utility of the players (indices 1 to \( k \)) given that the adversary (index \( k + 1 \)) seeks to maximize his own utility, i.e.,

\[
\min_{x_1, ..., x_k} \max_y \left\{ -\sum_{i=1}^k U_j(x_1, ..., x_k, y) : x_i \in B_i, i = 0, ..., k \right\},
\]

which is clearly an instance of (1.1) where \( x = (x_1, ..., x_k) \) and

\[
Y = B_y, \quad \Phi(x, y) = -\sum_{i=1}^k U_j(x_1, ..., x_k, y), \quad h(x) = \delta_{B_1 \times ... \times B_k}(x).
\]

2. Preliminaries. This section presents some preliminary material and is divided into two subsections. The first one precisely describes the assumptions and various notions of stationary points for problem (1.1) and briefly compares two approaches for obtaining them. The second one presents a smooth approximation of the max function \( p \) in (1.1) and some of its properties.

2.1. Assumptions and notions of stationary points. This subsection describes the assumptions and four notions of stationary points for problem (1.1). It is worth mentioning that the complexities of the smoothing scheme of Section 3 are presented with respect to two of these notions. In order to understand how these results can be translated to the other two alternative notions, which have been used in a few papers dealing with problem (1.1), we present a few results discussing some useful relations between all these notions.

Throughout our presentation, we let \( \mathcal{X} \) and \( \mathcal{Y} \) be finite dimensional inner product spaces. We also make the following assumptions on problem (1.1):

(A0) \( X \subset \mathcal{X} \) and \( Y \subset \mathcal{Y} \) are nonempty convex sets, and \( Y \) is also compact;

(A1) there exists an open set \( \Omega \supseteq X \) such that \( \Phi(\cdot, \cdot) \) is finite and continuous on \( \Omega \times Y \); moreover, \( \nabla_x \Phi(x, y) \) exists and is continuous at every \((x, y) \in \Omega \times Y\);
(A2) \( h \in \text{Conv}(X) \) and \(-\Phi(x, \cdot) \in \text{Conv}(Y)\) for every \( x \in \Omega \);

(A3) there exist scalars \((L_x, L_y) \in \mathbb{R}^2_{++}\), and \( m \in (0, L_x]\) such that

\[
\Phi(x, y) - [\Phi(x', y) + \langle \nabla_x \Phi(x', y), x - x' \rangle] \geq \frac{m}{2} \| x - x' \|^2, \tag{2.1}
\]

\[
\| \nabla_x \Phi(x, y) - \nabla_x \Phi(x', y') \| \leq L_x \| x - x' \| + L_y \| y - y' \|, \tag{2.2}
\]

for every \( x, x' \in X \) and \( y, y' \in Y \);

(A4) \( \hat{p} := \inf_{x \in X} \hat{p}(x) \) is finite, where \( \hat{p} \) is as in (1.1);

We make three remarks about the above assumptions. First, it is well-known that condition (2.2) implies that

\[
\Phi(x', y) - [\Phi(x, y) + \langle \nabla_x \Phi(x, y), x' - x \rangle] \leq \frac{L_x}{2} \| x' - x \|^2, \tag{2.3}
\]

for every \((x', x, y) \in X \times X \times Y\). Second, functions satisfying the lower curvature condition in (2.1) are often referred to as weakly convex functions (see, for example, [6,7,8,9]). Third, the aforementioned weak convexity condition implies that, for any \( y \in Y \), the function \( \Phi(\cdot, y) + m \| \cdot \|^2/2 \) is convex, and hence \( p + m \| \cdot \|^2/2 \) is as well.

Note that while \( \hat{p} \) is generally nonconvex and nonsmooth, it has the nice property that \( \hat{p} + m \| \cdot \|^2/2 \) is convex.

We now discuss two stationarity conditions of (1.1) under assumptions (A0)–(A3).

First, denoting

\[
\hat{\Phi}(x, y) := \Phi(x, y) + h(x) \quad \forall (x, y) \in X \times Y,
\]

it is well-known that (1.1) is related to the saddle-point problem which consists of finding a pair \((x^*, y^*) \in X \times Y\) such that

\[
\hat{\Phi}(x^*, y) \leq \hat{\Phi}(x^*, y^*) \leq \hat{\Phi}(x, y^*), \tag{2.5}
\]

for every \((x, y) \in X \times Y\). More specifically, \((x^*, y^*)\) satisfies (2.5) if and only if \( x^* \) is an optimal solution of (1.1), \( y^* \) is an optimal solution of the dual of (1.1), and there is no duality gap between the two problems. Using the composite structure described above for \( \hat{\Phi} \), it can be shown that a necessary condition for (2.5) to hold is that \((x^*, y^*)\) satisfy the stationarity condition

\[
\nabla_x \Phi(x^*, y^*) + \begin{pmatrix} \frac{\partial h(x^*)}{\partial x} \\ \frac{\partial h(x^*)}{\partial y} \end{pmatrix} \leq 0, \tag{2.6}
\]

When \( m = 0 \), the above condition also becomes sufficient for (2.5) to hold. Second, it can be shown that \( \hat{p}'(x^*; d) \) is well-defined for every \( d \in X \) and that a necessary condition for \( x^* \in X \) to be a local minimum of (1.1) is that it satisfies the stationarity condition

\[
\inf_{\| d \| \leq 1} \hat{p}'(x^*; d) \geq 0. \tag{2.7}
\]

When \( m = 0 \), the above condition also becomes sufficient for \( x^* \) to be a global minimum of (1.1). Moreover, in view of Lemma 20 in Appendix D with \((\bar{u}, \bar{v}, \bar{x}, \bar{y}) = (0, 0, x^*, y^*)\), it follows that \( x^* \) satisfies (2.7) if and only if there exists \( y^* \in Y \) such that \((x^*, y^*)\) satisfies (2.6).

Note that finding points that satisfy (2.6) or (2.7) exactly is generally a difficult task. Hence, in this section and the next one, we only consider approximate versions of (2.6) or (2.7), which are (1.4) and (1.5), respectively. For ease of future reference, we say that:
(i) a quadruple \((\tilde{u}, \tilde{v}, \tilde{x}, \tilde{y})\) is a \((\rho_x, \rho_y)\)-primal-dual stationary point of (1.1) if it satisfies (1.4);

(ii) a point \(\tilde{x}\) is a \(\delta\)-directional stationary point of (1.1) if it satisfies the first inequality in (1.5).

It is worth mentioning that (1.5) is generally hard to verify for a given point \(x \in X\). This is primarily because the definition requires us to check an infinite number of directional derivatives for a (potentially) nonsmooth function at points \(\tilde{x}\) near \(\tilde{x}\). In contrast, the definition of an approximate primal-dual stationary point is generally easier to verify because the quantities \(\|\tilde{u}\|\) and \(\|\tilde{v}\|\) can be measured directly, and the inclusions in (1.4) are easy to verify when the prox oracles for \(h\) and \(\Phi(x, \cdot)\), for every \(x \in X\), are readily available.

The next result, whose proof is given in Appendix D, shows that a \((\rho_x, \rho_y)\)-primal-dual stationary point, for small enough \(\rho_x\) and \(\rho_y\), yields a point \(x\) satisfying (1.5). Its statement makes use of the diameter of \(Y\) defined as

\[
D_y := \inf_{y, y' \in Y} \|y - y'\|. 
\]

**PROPOSITION 1.** If the quadruple \((\tilde{u}, \tilde{v}, \tilde{x}, \tilde{y})\) is a \((\rho_x, \rho_y)\)-primal-dual stationary point of (1.1), then there exists a point \(\hat{x} \in X\) such that

\[
\inf_{\|d\| \leq 1} \|\tilde{d}\| \geq -\rho_x - 2\sqrt{2mD_y\rho_y}, \quad \|\tilde{x} - \hat{x}\| \leq \sqrt{\frac{2D_y\rho_y}{m}}.
\]

The iteration complexities in this paper (see Section 3) are stated with respect to the two notions of stationary points (1.4) and (1.5). However, it is worth discussing below two other notions of stationary points that are common in the literature as well as some results that relate all four notions.

Given \((\lambda, \varepsilon) \in \mathbb{R}^2_+\), a point \(x\) is said to be a \((\lambda, \varepsilon)\)-prox stationary point of (1.1) if the function \(\hat{p} + \|\cdot\|^2/(2\lambda)\) is strongly convex and

\[
\frac{1}{\lambda}\|x - x_\lambda\| \leq \varepsilon, \quad x_\lambda = \arg\min_{u \in X} \left\{ \hat{p}(u) := \hat{p}(u) + \frac{1}{2\lambda}\|u - x\|^2 \right\}.
\]

The above notion is considered, for example, in [17, 26, 30]. The result below, whose proof is given in Appendix D, shows how it is related to (1.5).

**PROPOSITION 2.** For any given \(\lambda \in (0, 1/m)\), the following statements hold:

(a) for any \(\varepsilon > 0\), if \(x \in X\) satisfies (1.5) and

\[
0 < \delta \leq \frac{\lambda^3 \varepsilon}{\lambda^2 + 2(1 - \lambda m)(1 + \lambda)}, 
\]

then \(x\) is a \((\lambda, \varepsilon)\)-prox stationary point;

(b) for any \(\delta > 0\), if \(x \in X\) is a \((\lambda, \varepsilon)\)-prox stationary point for some \(\varepsilon \leq \delta \cdot \min\{1, 1/\lambda\}\), then \(x\) satisfies (1.5) with \(\hat{x} = x_\lambda\), where \(x_\lambda\) is as in (2.9).

Note that for a fixed \(\lambda \in (0, 1/m)\) such that \(\max\{\lambda^{-1}, (1 - \lambda m)^{-1}\} \in \mathcal{O}(1)\), the largest \(\delta\) in part (a) is \(\mathcal{O}(\varepsilon)\). Similarly, for part (b), if \(\lambda^{-1} = \mathcal{O}(1)\) then largest \(\varepsilon\) in part (b) is \(\mathcal{O}(\delta)\). Combining these two observations, it follows that (2.9) and (1.5) are equivalent (up to a multiplicative factor) under the assumption that \(\delta = \Theta(\varepsilon)\).
Given \((\rho_x, \rho_y) \in \mathbb{R}_{++}^2\), a pair \((\bar{x}, \bar{y})\) is said to be a \((\rho_x, \rho_y)\)-first-order Nash equilibrium point of (1.1) if
\[
(2.11) \quad \inf_{\|d_x\| \leq 1} S'_y(\bar{x}; d_x) \geq -\rho_x, \quad \sup_{\|d_y\| \leq 1} S'_x(\bar{y}; d_y) \leq \rho_y,
\]
where \(S_y := \Phi(\cdot, \bar{y}) + h(\cdot)\) and \(S_x := \Phi(\bar{x}, \cdot)\). The above notion is considered, for example, in \([17, 24, 25]\). The next result, whose proof is given in Appendix D, shows that (2.11) is equivalent to (1.4).

**Proposition 3.** A pair \((\bar{x}, \bar{y})\) is a \((\rho_x, \rho_y)\)-first-order Nash equilibrium point if and only if there exists \((\bar{u}, \bar{v}) \in X \times Y\) such that \((\bar{u}, \bar{v}, \bar{x}, \bar{y})\) satisfies (1.4).

We now end this subsection by briefly discussing some approaches for finding approximate stationary points of (1.1). One approach is to apply a proximal descent type method directly to problem (1.1), but this would lead to subproblems with nonsmooth convex composite functions. A second approach is based on first applying a smoothing method to (1.1) and then using a prox-convexifying descent method such as the one in [16] to solve the perturbed unconstrained smooth problem. An advantage of the second approach, which is the one pursued in this paper, is that it generates subproblems with smooth convex composite objective functions. The next subsection describes one possible way to smooth the (generally) nonsmooth function \(p\) in (1.1).

### 2.2. Smooth approximation

This subsection presents a smooth approximation of the function \(p\) in (1.1).

For every \(\xi > 0\), consider the smoothed function \(p_\xi\) defined by
\[
(2.12) \quad p_\xi(x) := \max_{y \in Y} \\left\{ \Phi_\xi(x, y) := \Phi(x, y) - \frac{1}{2\xi} \|y - y_0\|^2 \right\} \quad \forall x \in X,
\]
for some \(y_0 \in Y\). The following proposition presents the key properties of \(p_\xi\) and its related quantities.

**Proposition 4.** Let \(\xi > 0\) be given and assume that the function \(\Phi\) satisfies conditions (A0)–(A3). Let \(p_\xi(\cdot)\) and \(\Phi_\xi(\cdot, \cdot)\) be as defined in (2.12) and define
\[
(2.13) \quad y_\xi(x) := \arg\max_{y' \in Y} \Phi_\xi(x, y') \quad \forall x \in X.
\]
Then, the following properties hold:
(a) \(y_\xi(\cdot)\) is \(Q_\xi\)-Lipschitz continuous on \(X\) where
\[
(2.14) \quad Q_\xi := \xi L_y + \sqrt{\xi (L_x + m)};
\]
(b) \(p_\xi(\cdot)\) is continuously differentiable on \(X\) and \(\nabla p_\xi(x) = \nabla_x \Phi(x, y_\xi(x))\) for every \(x \in X\);
(c) \(\nabla p_\xi(\cdot)\) is \(L_\xi\)-Lipschitz continuous on \(X\) where
\[
(2.15) \quad L_\xi := L_y Q_\xi + L_x \leq \left( L_y \sqrt{\xi} + \sqrt{L_x} \right)^2;
\]
(d) for every \(x, x' \in X\), we have
\[
(2.16) \quad p_\xi(x) - [p_\xi(x') + \langle \nabla p_\xi(x'), x - x' \rangle] \geq -\frac{m}{2} \|x - x'\|^2;
\]
Proof. The inequality in (2.15) follows from (a), the fact that \( m \leq L_x \), and the bound

\[
L \xi = L_y \left[ \xi L_y + \sqrt{\xi (L_x + m)} \right] + L_x \leq \xi L_y^2 + 2 \sqrt{\xi L_x} + L_x = \left( L_y \sqrt{\xi} + \sqrt{L_x} \right)^2.
\]

The other conclusions of (a)–(c) follow from Lemma 13 and Proposition 14 in Appendix B with \((\Psi, q, y) = (\Phi_\xi, p_\xi, y_\xi)\). We now show that the conclusion of (d) is true. Indeed, if we consider (2.1) at \((y, x') = (y_\xi(x'), x')\), the definition of \(\Phi_\xi\), and use the definition of \(\nabla p_\xi\) in (b), then

\[
\frac{m}{2} \|x - x'\|^2 \leq \Phi(x', y_\xi(x)) - \Phi(x, y_\xi(x)) + \langle \nabla_x \Phi(x, y_\xi(x)), x' - x \rangle
\]

\[
= \Phi_\xi(x', y_\xi(x)) - [p_\xi(x) + \langle \nabla p_\xi(x), x' - x \rangle] \leq p_\xi(x') - [p_\xi(x) + \langle \nabla p_\xi(x), x' - x \rangle],
\]

where the last inequality follows from the optimality of \(y\).

We now make two remarks about the above properties. First, the Lipschitz constants of \(y_\xi\) and \(\nabla p_\xi\) depend on the value of \(\xi\) while the weak convexity constant \(m\) in (2.16) does not. Second, as \(\xi \to \infty\), it holds that \(p_\xi \to p\) pointwise and \(Q_\xi, L_\xi \to \infty\). These remarks are made more precise in the next result.

Lemma 5. For every \(\xi > 0\), it holds that

\[-\infty < p(x) - \frac{D_y^2}{2\xi} \leq p_\xi(x) \leq p(x) \quad \forall x \in X,\]

where \(D_y\) is as in (2.8).

Proof. The fact that \(p(x) > -\infty\) follows immediately from assumption (A4). To show the other bounds, observe that for every \(y_\xi \in Y\), we have

\[
\Phi(x, y) + h(x) \geq \Phi(x, y) - \frac{1}{2\xi} \|y - y_\xi\|^2 + h(x) \geq \Phi(x, y) - \frac{D_y^2}{2\xi} + h(x)
\]

for every \((x, y) \in X \times Y\). Taking the supremum of the bounds over \(y \in Y\) and using the definitions of \(p\) and \(p_\xi\) yields the remaining bounds.

3. Unconstrained min-max optimization. This section presents our proposed AIPP-S scheme for solving the min-max CNO problem (1.1) and is divided into two subsections. The first one reviews an AIPP method for solving smooth CNO problems. The second one presents the AIPP-S scheme and its iteration complexity for finding stationary points as in (1.4) and (1.5).

Before proceeding, we briefly outline the idea of the AIPP-S scheme. The main idea is to apply the AIPP method described in the next subsection to the smooth CNO problem

\[(3.1) \quad \min_{x \in X} \{\tilde{p}_\xi(x) := p_\xi(x) + h(x)\},\]

where \(p_\xi\) is as in (2.12) and \(\xi\) is a positive scalar that will depend on the tolerances in (1.4) and (1.5). The above smoothing approximation scheme is similar to the one used in [23]; the approximation function \(\tilde{p}_\xi\) used in both schemes is smooth, but the one here is nonconvex while the one in [23] is convex. Moreover, while [23] uses an ACG variant to approximately solve (3.1), the AIPP-S scheme uses the AIPP method discussed below for this purpose.
3.1. AIPP method for smooth CNO problems. This subsection describes the AIPP method studied in [16], and its corresponding iteration complexity result, for solving a class of smooth CNO problems.

We first describe the problem that the AIPP method is intended to solve. Let \( \mathcal{X} \) be a finite-dimensional inner product and consider the smooth CNO problem

\[
(3.2) \quad \phi_* := \inf_{x \in \mathcal{X}} [\phi(x) := f(x) + h(x)]
\]

where \( h : \mathcal{X} \to (-\infty, \infty] \) and function \( f \) satisfy the following assumptions:

(P1) \( h \in \text{Conv}(\mathcal{X}) \) and \( f \) is differentiable on \( \text{dom} h \);

(P2) for some \( M \geq m > 0 \), the function \( f \) satisfies

\[
(3.3) \quad -\frac{m}{2} \|x' - x\|^2 \leq f(x') - [f(x) + \langle \nabla f(x), x' - x \rangle],
\]

\[
(3.4) \quad \|\nabla f(x') - \nabla f(x)\| \leq M \|x' - x\|,
\]

for any \( x, x' \in \text{dom} h \);

(P3) \( \phi_* \) defined in (3.2) is finite.

We now make four remarks about the above assumptions. First, it is well-known that a necessary condition for \( x^* \in \text{dom} h \) to be a local minimum of (3.2) is that \( x^* \) is a stationary point of \( \phi \), i.e. \( 0 \in \nabla f(x^*) + \partial h(x^*) \). Second, it is well-known that (3.4) implies that (3.3) holds for any \( m \in [-M, M] \). Third, it is easy to see from Proposition 4 that \( \rho_\xi \) in (2.12) satisfies assumption (P2) with \( (M, f) = (L_\xi, p_\xi) \) where \( L_\xi \) is as in (2.15). Fourth, it is also easy to see that the function \( \rho_\xi \) in (2.12) satisfies assumption (P3) with \( \phi_* = \inf_{x \in \mathcal{X}} \rho_\xi(x) \) in view of assumption (A4) and Lemma 5.

For the purpose of discussing the complexity results of this subsection, we consider the following notion of an approximate stationary point of (3.2): given a tolerance \( \bar{\rho} > 0 \), a pair \((\bar{x}, \bar{u})\) satisfying (3.5).

\[
(3.5) \quad \bar{u} \in \nabla f(\bar{x}) + \partial h(\bar{x}), \quad \|\bar{u}\| \leq \bar{\rho}.
\]

We now state the AIPP method for finding a pair \((\bar{x}, \bar{u})\) satisfying (3.5).

**AIPP method**

**Input:** a function pair \((f, h)\), a scalar pair \((m, M) \in \mathbb{R}^2_+\) satisfying (P2), scalars \( \lambda \in (0, 1/(2m)] \) and \( \sigma \in (0, 1) \), an initial point \( x_0 \in \text{dom} h \), and a tolerance \( \bar{\rho} > 0 \);

**Output:** a pair \((\bar{x}, \bar{u}) \in \text{dom} h \times \mathcal{X}\) satisfying (3.5);

(0) set \( k = 1 \) and define

\[
\bar{\rho} := \frac{\bar{\lambda}}{4}, \quad \bar{\epsilon} := \frac{\bar{\rho}^2}{32(M + \lambda^{-1})}, \quad M_\lambda := M + \lambda^{-1};
\]

(1) perform at least \([6\sqrt{2}\lambda M + 1]\) iterations of the accelerated composite gradient (ACG) method in Appendix A with inputs \( z_0, (\mu, L) \), and \( (\psi_s, \psi_n) \) given by

\[
z_0 = x_{k-1}, \quad (\mu, L) = (1/2, \lambda M + 1/2), \quad (\psi_s, \psi_n) = \left( \lambda f + \frac{1}{4} \| -x_{k-1} \|^2, \lambda h + \frac{1}{4} \| -x_{k-1} \|^2 \right).
\]
in order to obtain a triple \((x, u, \varepsilon) \in X \times X \times \mathbb{R}_+\) satisfying
\[(3.6)\quad u \in \partial \left( \lambda \phi + \frac{1}{2} \| -x_{k-1} \|^2 \right) (x), \quad \|u\|^2 + 2\varepsilon \leq \sigma \|x_{k-1} - x + u\|^2; \]

(2) if the residual
\[(3.7)\quad \|x_{k-1} - x + u\| \leq \frac{\lambda \rho}{5}, \]
then go to (3); otherwise set \((x_k, \tilde{u}_k, \tilde{\varepsilon}_k) = (x, u, \varepsilon)\), increment \(k = k + 1\) and
go to (1);
(3) restart the previous call to the ACG method in step 1 to find a triple \((\tilde{x}, \tilde{u}, \tilde{\varepsilon})\)
such that \(\tilde{\varepsilon} \leq \tilde{\varepsilon} \lambda\) and \((x, u, \varepsilon) = (\tilde{x}, \tilde{u}, \tilde{\varepsilon})\) satisfies (3.6);
(4) compute
\[(3.8)\quad \bar{x} := \arg\min_{x' \in X} \left\{ \langle \nabla f(x), x' - x \rangle + b(x') + \frac{M_\lambda}{2} \|x' - \bar{x}\|^2 \right\}, \]
\[(3.9)\quad \bar{u} := M_\lambda (x - \bar{x}) + \nabla f(\bar{x}) - \nabla f(x), \]
where \(M_\lambda\) is as in step 0, and output the pair \((\bar{x}, \bar{u})\).

We now make four remarks about the above AIPP method. First, at the \(k\)th
iteration of the method, its step 1 invokes an ACG method, whose description is given
in Appendix A, to approximately solve the strongly convex proximal subproblem
\[(3.10)\quad \min_{x \in X} \left\{ \lambda \phi(x) + \frac{1}{2} \|x - x_{k-1}\|^2 \right\}, \]
according to (3.6). Second, Lemma 12 shows that every ACG iterate \((z, u, \varepsilon)\) satisfies
the inclusion in (3.6), and hence, only the inequality in (3.6) needs to be verified.
Third, note that (3.4) implies that the gradient of the function \(\psi_s\) defined in step 1 of
the AIPP method is \((\lambda M + 1/2)\)-Lipschitz continuous. As a consequence, Lemma 12
with \(L = \lambda M + 1/2\) implies that the triple \((z, u, \varepsilon)\) in step 1 of any iteration of the
AIPP method can be obtained in \(O(\sqrt{\lambda M + 1/\sigma})\) ACG iterations.

Note that the above method differs slightly from the one presented in [16] in that
it adds step 4 in order to directly output a \(\tilde{\rho}\)-approximate stationary point as in (3.5).
The justification for the latter claim follows from [16, Lemma 12], [16, Theorem 13],
and [16, Corollary 14], which also imply the following complexity result.

**Proposition 6.** The AIPP method terminates with a \(\tilde{\rho}\)-approximate stationary
point of (3.2) in
\[(3.11)\quad \mathcal{O} \left( \sqrt{\lambda M + 1} \left[ \frac{R(\phi; \lambda)}{\sqrt{\sigma(1 - \sigma)^2 \lambda^2 \tilde{\rho}^2}} + \log_1^+ (\lambda M) \right] \right), \]
\(ACG\) iterations, where
\[(3.12)\quad R(\phi; \lambda) = \inf_{x'} \left\{ \frac{1}{2} \|x_0 - x'\|^2 + \lambda [\phi(x') - \phi] \right\}. \]
We now make two remarks about the quantity $R(\phi; \lambda)$ in (3.11). First, scaling $R(\phi; \lambda)$ by $1/\lambda$ and then shifting by $\phi_*$ results in the $\lambda$-Moreau envelope\footnote{See [28, Chapter 1.G] for an exact definition.} of $\phi$. Second, it admits the upper bound

$$\begin{align*}
R(\phi; \lambda) \leq \min \left\{ \frac{1}{2} d_0^2, \lambda \| \phi(x_0) - \phi_* \| \right\}
\end{align*}$$

where $d_0 := \inf \{ \| x_0 - x_* \| : x_* \text{ is an optimal solution of (3.2)} \}$.

\[ \tag{3.13} \]

### 3.2. AIPP-S scheme for min-max CNO problems

We are now ready to state the AIPP-S scheme for finding approximate stationary points of the unconstrained min-max CNO problem (1.1).

It is stated in a incomplete manner in the sense that it does not specify how the parameter $\xi$ and the tolerance $\rho$ used in its step 2 are chosen. Two invocations of this method, with different choices of $\xi$ and $\rho$, are considered in Propositions 8 and 9, which describe the iteration complexities for finding approximate stationary points as in (1.4) and (1.5), respectively.

**AIPP-S scheme**

\textbf{Input}: a triple $(m, L_x, L_y) \in \mathbb{R}_+^3$ satisfying (A3), a smoothing constant $\xi > 0$, an initial point $(x_0, y_0) \in X \times Y$, and a tolerance $\rho > 0$;

\textbf{Output}: a pair $(x, u) \in X \times X$;

(0) set $L_\xi$ as in (2.15), $\sigma = 1/2$, $\lambda = 1/(4m)$, and define $p_\xi$ as in (2.12);

(1) apply the AIPP method with inputs $(m, L_\xi), (p_\xi, h), \lambda, \sigma, x_0$, and $\rho$ to obtain a pair $(x, u)$ satisfying

$$\begin{align*}
u \in \nabla p_\xi(x) + \partial h(x), \quad \|u\| \leq \rho;
\end{align*}$$

\[ \tag{3.14} \]

(2) output the pair $(x, u)$.

We now give four remarks about the above method. First, the AIPP method invoked in step 2 terminates due to [16, Theorem 13] and the third and fourth remarks following assumptions (P1)–(P3). Second, since the AIPP-S scheme is a one-pass method (as opposed to an iterative method), the complexity of the AIPP-S scheme is essentially that of the AIPP method. Third, similar to the smoothing scheme of [23] which assumes $m = 0$, the AIPP-S scheme is also a smoothing scheme for the case in which $m > 0$. On the other hand, in contrast to the algorithm of [23] which uses an ACG variant, AIPP-S invokes the AIPP method to solve (3.1) due to its nonconvexity. Finally, while the AIPP method in step 2 is called with $(\sigma, \lambda) = (1/2, 1/(4m))$, it can also be called with any $\sigma \in (0, 1)$ and $\lambda \in (0, 1/(2m))$ to establish the desired termination of the AIPP-S scheme.

For the remainder of this subsection, our goal will be to show that a careful selection of the parameter $\xi$ and the tolerance $\rho$ will allow the AIPP-S method to generate approximate stationary points as in (1.5) and (1.4).

Before proceeding, we first present a bound on the quantity $R(\hat{p}_\xi; \lambda)$ in terms of the data in problem (1.1). Its importance derives from the fact that the AIPP method applied to the smoothed problem (3.1) yields the bound (3.11) with $\phi = \hat{p}_\xi$. 

\[ \footnote{See [28, Chapter 1.G] for an exact definition.} \]
Lemma 7. For every $\xi > 0$ and $\lambda \geq 0$, it holds that
\begin{equation}
R(\hat{p}_\xi; \lambda) \leq R(\hat{p}; \lambda) + \frac{\lambda D^2_y}{2\xi},
\end{equation}
where $R(\cdot, \cdot)$ and $D_y$ are as in (3.12) and (2.8), respectively.

Proof. Using Lemma 5 and the definitions of $\hat{p}$ and $\hat{p}_\xi$, it holds that
\begin{equation}
\hat{p}_\xi(x) - \inf_{x'} \hat{p}_\xi(x') \leq \hat{p}(x) - \inf_{x'} \hat{p}(x') + \frac{D^2_y}{2\xi}, \quad \forall x \in X.
\end{equation}
Multiplying the above expression by $(1 - \sigma)\lambda$ and adding the quantity $\|x_0 - x\|^2/2$ yields the inequality
\begin{equation}
\frac{1}{2}\|x_0 - x\|^2 + (1 - \sigma)\lambda \left[ \hat{p}_\xi(x) - \inf_{x'} \hat{p}_\xi(x') \right]
\leq \frac{1}{2}\|x_0 - x\|^2 + (1 - \sigma)\lambda \left[ \hat{p}(x) - \inf_{x'} \hat{p}(x') \right] + (1 - \sigma)\frac{\lambda D^2_y}{2\xi} \quad \forall x \in X,
\end{equation}
Taking the infimum of the above expression, and using the definition of $R(\cdot, \cdot)$ in (3.12) yields the desired conclusion.

We now show how the AIPP-S scheme generates a $(\rho_x, \rho_y)$–primal-dual stationary point, i.e. one satisfying (1.4). Recall the definition of “oracle call” in the paragraph containing (1.3).

Proposition 8. For a given tolerance pair $(\rho_x, \rho_y) \in \mathbb{R}^2_{++}$, let $(x, u)$ be the pair output by the AIPP-S scheme with input parameter $\xi$ and tolerance $\rho$ satisfying
\begin{equation}
\xi \geq \frac{D_y}{\rho_y}, \quad \rho = \rho_x.
\end{equation}
Moreover, define
\begin{equation}
(\bar{u}, \bar{v}) := \left( u, \frac{y_0 - y_\xi(x)}{\xi} \right), \quad (\bar{x}, \bar{y}) := (x, y_\xi(x)),
\end{equation}
where $y_\xi$ is as in (2.13). Then, the following statements hold:
(a) the AIPP-S scheme performs
\begin{equation}
\mathcal{O}\left( \Omega_\xi \left[ \frac{m^2 R(\rho_y; 1/4m)}{\rho_x^2} + \frac{m D^2_y}{\xi \rho_x^2} + \log^+_1(\Omega_\xi) \right] \right)
\end{equation}
oracle calls, where $R(\cdot, \cdot)$ and $D_y$ are as in (3.12) and (2.8), respectively, and
\begin{equation}
\Omega_\xi := 1 + \frac{\sqrt{\xi L_y + \sqrt{L_x}}}{\sqrt{m}};
\end{equation}
(b) the quadruple $(\bar{u}, \bar{v}, \bar{x}, \bar{y})$ is a $(\rho_x, \rho_y)$–primal-dual stationary point of (1.1).

Proof. (a) Using the inequality in (2.15), it holds that
\begin{equation}
\sqrt{\frac{L_y}{4m}} + 1 \leq 1 + \sqrt{\frac{L_y}{4m}} \leq 1 + \frac{\sqrt{\xi L_y + \sqrt{L_x}}}{2\sqrt{m}} = \Theta(\Omega_\xi).
\end{equation}
Moreover, using Proposition 6 with \( (\phi, M) = (\hat{\phi}_\xi, L_\xi) \), Lemma 7, and bound (3.22), it follows that the number of ACG iterations performed by the AIPP-S scheme is on the order given by (3.20). Since step 1 of the AIPP invokes once the ACG variant in Appendix A with a pair \((\psi_s, \psi_n)\) of the form

\[
\psi_s = \lambda p_\xi + \frac{1}{4} \| -\tilde{z} \|^2, \quad \psi_n = \lambda h + \frac{1}{4} \| -\tilde{z} \|^2
\]

for some \( \tilde{z} \) and each iteration of this ACG variant performs \( O(1) \) gradient evaluations of \( \psi_s \), \( O(1) \) function evaluations of \( \psi_s \) and \( \psi_n \), and \( O(1) \) \( \psi_n \)-resolvent evaluations, it follows from Proposition 4(b) and the definition of an “oracle call” in the paragraph containing (1.3) that each one of the above ACG iterations requires \( O(1) \) oracle calls.

Statement (a) now follows from the above two conclusions.

(b) It follows from the definitions of \( p_\xi \), tolerance \( \rho \), and \((\bar{y}, \bar{u})\) in (2.12), (3.18), and (3.19), respectively, Proposition 4(b), and the inclusion in (3.14) that \( \| \bar{u} \| \leq \rho_x \) and

\[
\bar{u} \in \nabla p_\xi(\bar{x}) + \partial h(\bar{x}) = \nabla_x \Phi(\bar{x}, y_\xi(\bar{x})) + \partial h(\bar{x}) = \nabla_x \Phi(\bar{x}, \bar{y}) + \partial h(\bar{x}).
\]

Hence, we conclude that the top inclusion and the upper bound on \( \| \bar{u} \| \) in (1.4) hold.

Next, the optimality condition of \( \bar{y} = y_\xi(\bar{x}) \) as a solution to (2.12) and the definition of \( \bar{v} \) in (2.12) give

\[
\tag{3.23}
0 \in \partial [\Phi(\bar{x}, \cdot)](\bar{y}) + \frac{\bar{y} - y_0}{\xi} = \partial [\Phi(\bar{x}, \cdot)](\bar{y}) - \bar{v}
\]

Moreover, the definition of \( \xi \) implies that

\[
\tag{3.24}
\| \bar{v} \| = \frac{\| \bar{y} - y_0 \|}{\xi} \leq \frac{D_y}{D_y/\rho_y} = \rho_y.
\]

Hence, combining (3.23) and (3.24), we conclude that the bottom inclusion and the upper bound on \( \| \bar{v} \| \) in (1.4) hold. \( \Box \)

We now make three remarks about Proposition 8. First, recall that \( R(\hat{\rho}; 1/(4m)) \) in the complexity (3.20) can be majorized by the rightmost quantity in (3.13) with \((\phi, \lambda) = (\hat{\rho}, 1/(4m))\). Second, under the assumption that (3.18) is satisfied as equality, the complexity of AIPP-S scheme reduces to

\[
\tag{3.25}
O \left( m^{3/2} \cdot R(\hat{\rho}; 1/(4m)) \cdot \frac{L_x^{1/2}}{\rho_x^2} + \frac{L_y D_y^{1/2}}{\rho_y^{3/2}} \right)
\]

under the reasonable assumption that the \( O(\rho_x^{-2} + \rho_x^{-2} \rho_y^{-1/2}) \) term in (3.20) dominates the other terms. Third, recall from the last remark following the previous proposition that when \( Y \) is a singleton, (1.1) becomes a special instance of (3.2) and the AIPP-S scheme becomes equivalent to the AIPP method of Subsection 3.1. It similarly follows that the complexity in (3.25) reduces to

\[
\tag{3.26}
O \left( \frac{m^{3/2} L_x^{1/2} R(\hat{\rho}; 1/(4m))}{\rho_x^2} \right)
\]

and, in view of this remark, the \( O(\rho_x^{-2} \rho_y^{-1/2}) \) term in (3.25) is attributed to the (possible) nonsmoothness in (1.1).
We next show how the AIPP-S scheme generates a point that is near a \( \delta \)-directional stationary point, i.e., one satisfying (1.5). Recall the definition of “oracle call” in the paragraph containing (1.3).

**Proposition 9.** Let a tolerance pair \( \delta > 0 \) be given and consider the AIPP-S scheme with input parameter \( \xi \) and tolerance \( \rho \) satisfying

\[
(3.27) \quad \xi \geq \frac{D_y}{\tau}, \quad \rho = \frac{\delta}{2}, \quad \tau \leq \min \left\{ \frac{m\delta^2}{2D_y}, \frac{\delta^2}{32mD_y} \right\}.
\]

Then, the following statements hold:

(a) the AIPP-S scheme performs

\[
(3.28) \quad \mathcal{O} \left( \Omega_{\xi} \left[ \frac{R(\hat{\rho}; \lambda)}{\lambda^2 \delta^2} + \frac{D_y^2}{\lambda^2 \xi^2} + \log^+(\Omega_{\xi}) \right] \right)
\]

oracle calls where \( \Omega_{\xi} \), \( R(\cdot; \cdot) \), and \( D_y \) are as in (3.21), (3.12), and (2.8), respectively;

(b) the first argument \( x \) in the pair output by the AIPP-S scheme satisfies (1.5).

*Proof.* (a) Using Proposition 8 with \( (\rho_x, \rho_y) = (\delta/2, \tau) \) and the bound on \( \tau \) in (3.27) it follows that the AIPP-S stops in a number of ACG iterations bounded above by (3.28).

(b) Let \( (x, u) \) be the \( \hat{\rho} \)-approximate stationary point of (3.1) generated by the AIPP-S scheme (see step 2) with \( \xi \) and \( \hat{\rho} \) satisfying (3.27). Defining \( (\tilde{v}, \tilde{y}) \) as in (3.19), it follows from Proposition 8 with \( (\rho_x, \rho_y) = (\delta/2, \tau) \) that \( (u, \tilde{v}, x, \tilde{y}) \) is a \( (\delta/2, \tau) \)-primal-dual stationary point of (1.1). As a consequence, it follows from Proposition 1 with \( (\rho_x, \rho_y) = (\delta/2, \tau) \) that there exists a point \( \hat{x} \) satisfying

\[
(3.29) \quad \|\hat{x} - x\| \leq \sqrt{\frac{2D_y\tau}{m}}, \quad \inf_{\|d\| \leq 1} \hat{\rho}'(\hat{x}; d) \geq -\frac{\delta}{2} - 2\sqrt{2mD_y\tau}.
\]

Combining the above bounds with the bound on \( \tau \) in (3.27) yields the desired conclusion in view of (1.5).

We now give four remarks about the above result. First, recall that \( R(\hat{\rho}; 1/(4m)) \) in the complexity (3.28) can be majorized by the rightmost quantity in (3.13) with \( (\phi, \lambda) = (\hat{\rho}, 1/(4m)) \). Second, Proposition 9(b) states that, while \( x \) not a stationary point itself, it is near a \( \delta \)-directional stationary point \( \hat{x} \). Third, under the assumption that (3.27) is satisfied as equality, the complexity of the AIPP-S scheme reduces to

\[
(3.30) \quad \mathcal{O} \left( m^{3/2} \cdot R(\hat{\rho}; 1/(4m)) \cdot \left[ \frac{L_x^{1/2} + L_y}{\delta^2} \right] \right)
\]

under the reasonable assumption that the \( \mathcal{O}(\delta^{-2} + \delta^{-3}) \) term in (3.28) dominates the other \( \mathcal{O}(\delta^{-1}) \) terms. Fourth, when \( Y \) is a singleton, it is easy to see that (1.1) becomes a special instance of (3.2), the AIPP-S scheme becomes equivalent to the AIPP method of Subsection 3.1, and the complexity in (3.30) reduces to

\[
(3.31) \quad \mathcal{O} \left( \frac{m^{3/2}L_x^{1/2} R(\hat{\rho}; 1/(4m))}{\delta^2} \right).
\]

In view of the last remark, the \( \mathcal{O}(\delta^{-3}) \) term in (3.30) is attributed to the (possible) nonsmoothness in (1.1).
4. Linearly-constrained min-max optimization. This section presents our proposed QP-AIPP-S scheme for solving the linearly constrained min-max CNO problem (1.6), and it is divided into two subsections. The first one reviews a QP-AIPP method for solving smooth linearly-constrained CNO problems. The second one presents the QP-AIPP-S scheme and its iteration complexity for finding stationary points as in (1.8). Throughout our presentation, we let $\mathcal{X}, \mathcal{Y}$, and $\mathcal{U}$ be finite dimensional inner product spaces.

Before proceeding, let us give the precise assumptions underlying the problem of interest and discuss the relevant notion of stationarity. For problem (1.6) suppose that assumptions (A0)–(A3) hold and that the linear operator $A : \mathcal{X} \mapsto \mathcal{U}$ and vector $b \in \mathcal{U}$ satisfy:

(A5) $A \not\equiv 0$ and $F := \{ x \in \mathcal{X} : Ax = b \} \not\equiv \emptyset$;

(A6) there exists $\hat{c} \geq 0$ such that

$$\inf_{x \in \mathcal{X}} \left\{ \hat{p}(x) + \frac{\hat{c}}{2} \| Ax - b \|^2 \right\} > -\infty.$$  

Note that (A4) in Subsection 2.1 is replaced by (A6) which is required by the QP-AIPP method of the next subsection.

Analogous to the first remark following (2.6), it is known that if $(x^*, y^*)$ satisfies (2.5) for every $(x, y) \in F \times \mathcal{Y}$ and $\hat{\Phi}$ as in (2.4), then there exists a multiplier $r^* \in \mathcal{U}$ such that

$$\left( \begin{array}{c} 0 \\ 0 \end{array} \right) \in \left( \begin{array}{c} \nabla_x \Phi(x^*, y^*) + A^* r^* \\ 0 \end{array} \right) + \left( \begin{array}{c} \partial h(x^*) \\ \partial [-\Phi(x^*, \cdot)](y^*) \end{array} \right),$$

holds. Hence, in view of the third remark in the paragraph following (2.7), we only consider the approximate version of (4.1) which is (1.8).

We now briefly outline the idea of the QP-AIPP-S scheme. The main idea is to apply the QP-AIPP method described in the next subsection to the smooth linearly-constrained CNO problem

$$\min_{x \in \mathcal{X}} \{ p_\xi(x) + h(x) : Ax = b \},$$

where $p_\xi$ is as in (1.2) and $\xi$ is a positive scalar that will depend on the tolerances in (1.8). This idea is similar to the one in Section 3 in that it applies an accelerated solver to a perturbed version of the problem of interest.

4.1. QP-AIPP method for constrained smooth CNO problems. This subsection describes the QP-AIPP method studied in [16], and its corresponding iteration complexity, for solving linearly-constrained smooth CNO problems.

We begin by describing the problem that the QP-AIPP method intends to solve. Consider the linearly-constrained smooth CNO problem

$$\hat{\phi}_\ast := \inf_{x \in \mathcal{X}} \{ \phi(x) := f(x) + h(x) : Ax = b \}$$

where $h : \mathcal{X} \mapsto (-\infty, \infty]$ and a function $f$ satisfy assumptions (P1)–(P3), the operator $A : \mathcal{X} \mapsto \mathcal{U}$ is linear, $b \in \mathcal{U}$, and the following additional assumptions hold:

(Q1) $A \not\equiv 0$ and $F := \{ x \in \text{dom} h : Ax = b \} \not\equiv \emptyset$;

(Q2) there exists $\hat{c} \geq 0$ such that $\hat{\phi}_\ast > -\infty$ where

$$\hat{\phi}_c := \inf_{x \in \mathcal{X}} \left\{ \phi_c(x) := \phi(x) + \frac{c}{2} \| Ax - b \|^2 \right\}, \quad \forall c \geq 0.$$
We now give some remarks about the above assumptions. First, similar to problem (3.2), it is well-known that a necessary condition for \( x^* \in \text{dom} \, h \) to be a local minimum of (4.3) is that \( x^* \) satisfies \( 0 \in \nabla f(x^*) + \partial h(x^*) + A^* r^* \) for some \( r^* \in \mathcal{U} \). Second, it is straightforward to verify that \( (p, h, A, b) \) in (1.6) satisfy (Q1)–(Q2) in view of assumptions (A5)–(A6). Third, since every feasible solution of (4.3) is also a feasible solution of (4.4), it follows from assumptions (Q2) that \( \hat{c} \leq \phi_c > -\infty \) (e.g., \( \text{dom} \, h \) is compact) then (Q2) holds with \( \hat{c} = 0 \).

Our interest in this subsection is in finding an approximate stationary point of (4.3) in the following sense: given a tolerance pair \((\bar{\rho}, \bar{\eta}) \in \mathbb{R}_+^2\), a triple \((\bar{x}, \bar{u}, \bar{r}) \in \text{dom} \, h \times \mathcal{X} \times \mathcal{U} \) is said to be a \((\bar{\rho}, \bar{\eta})\)-approximate stationary point of (4.3) if

\[
\bar{u} \in \nabla f(\bar{x}) + \partial h(\bar{x}) + A^* \bar{r}, \quad \|\bar{u}\| \leq \bar{\rho}, \quad \|A\bar{x} - b\| \leq \bar{\eta}.
\]

We now state the QP-AIPP method for finding \((\bar{x}, \bar{u}, \bar{r})\) satisfying (4.5).

**QP-AIPP method**

**Input:** a function pair \((f, h)\), a scalar pair \((m, M) \in \mathbb{R}_+^2\) satisfying (3.3), scalars \(\lambda \in (0, 1/(2m)]\) and \(\sigma \in (0, 1)\), a scalar \(\hat{c}\) satisfying assumption (Q2), an initial point \(x_0 \in \text{dom} \, h\), and a tolerance pair \((\tilde{\rho}, \tilde{\eta}) \in \mathbb{R}_+^2\);

**Output:** a triple \((\bar{x}, \bar{u}, \bar{r}) \in \text{dom} \, h \times \mathcal{X} \times \mathcal{U} \) satisfying (4.5);

1. set \(\tilde{c} = \hat{c} + M/\|A\|^2\);
2. define the quantities

\[
M_c := M + c\|A\|^2, \quad f_c := f + \frac{c}{2} \|A(. - b\|^2, \quad \phi_c = f_c + h,
\]

and apply the AIPP method with inputs \((m, M_c), (f_c, h), \lambda, \sigma, x_0, \) and \(\tilde{\rho}\) to obtain a \(\tilde{\rho}\)-approximate stationary point \((\bar{x}, \bar{u})\) of (3.2) with \(f = f_c\);
3. if \(\|A\bar{x} - b\| > \tilde{\eta}\) then set \(c = 2\tilde{c}\) and go to (1); otherwise, set \(\bar{r} = c(A\bar{x} - b)\) and output the triple \((\bar{x}, \bar{u}, \bar{r})\).

We now give two remarks about the above method. First, it straightforward to see that QP-AIPP method terminates due to the results in [16, Section 4]. Second, in view of Proposition 6 with \((\phi, M_\phi) = (\phi_c, M_c)\), it is easy to see that the number of ACG iterations executed in step 1 at any iteration of the method is

\[
O \left( \sqrt{\lambda M_c} + \frac{R(\phi_c; \lambda)}{\sqrt{\sigma(1 - \sigma)^2 \lambda^2 \tilde{\rho}^2}} + \log_\gamma^+ (\lambda M_c) \right)
\]

and that the pair \((\bar{x}, \bar{u})\) computed in step 1 satisfies the inclusion and the first inequality in (4.5).

We now focus on the iteration complexity of the QP-AIPP method. Before proceeding, we first define the useful quantity

\[
R_c(\phi; \lambda) := \inf_{x'} \left\{ \frac{1}{2} \|x_0 - x'\|^2 + \lambda \left[ \phi(x') - \phi_c \right] : x' \in \mathcal{F} \right\},
\]

for every \(c \geq \hat{c}\), where \(\phi_c\) is as defined in (4.4). The quantity in (4.8) plays an analogous role as (3.12) in (3.11) and, similar to the discussion following Proposition 6, it is a scaled and shifted \(\lambda\)-Moreau envelope of \(\phi + \delta \mathcal{F}\). Moreover, due to [16, Lemma 16], it
also admits the upper bound
\[
R_c(\phi; \lambda) \leq R_c(\phi; \lambda) \leq \min \left\{ \frac{1}{2} d_0^2, \lambda \left[ \hat{\phi}_s - \hat{\phi}_c \right] \right\}
\]
where \( \hat{\phi}_s \) is as defined in (4.3) and
\[
d_0 := \inf \{ \|x_0 - x_*\| : x_* \text{ is an optimal solution of (4.3)} \}.
\]

We now state the iteration complexity of the QP-AIPP method, whose proof may be adapted from [16, Lemma 12] and [16, Theorem 18].

**Proposition 10.** Let a constant \( \hat{\phi}_c \) as in assumption (Q2), scalar \( \sigma \in (0, 1) \), curvature pair \((m, M) \in \mathbb{R}^2_+ \), and a tolerance pair \((\hat{\rho}, \hat{\eta}) \in \mathbb{R}^2_+ \) be given. Moreover, define
\[
(4.10) \quad T_\eta := \frac{2R_c(\hat{\phi}; \lambda)}{\eta^2(1 - \sigma)\lambda} + \hat{\phi}_c, \quad \Theta_\eta := M + T_\eta \|A\|^2.
\]

Then, the QP-AIPP method outputs a triple \((\bar{u}, \bar{v}, \bar{r})\) satisfying (4.5) in
\[
(4.11) \quad \mathcal{O} \left( \sqrt{\lambda \Theta_\eta} \right) \left[ \frac{R_c(\hat{\phi}; \lambda)}{\sqrt{\sigma(1 - \sigma)^2 \lambda^2 \rho^2} + \log^+ (\lambda \Theta_\eta)} \right]
\]
ACG iterations.

### 4.2. QP-AIPP-S scheme for constrained min-max CNO problems.

We are now ready to state the QP-AIPP smoothing scheme for finding an approximate primal-dual stationary point of the linearly-constrained min-max CNO problem (1.6).

**QP-AIPP-S scheme**

**Input:** a triple \((m, L_x, L_y) \in \mathbb{R}^2_+\) satisfying assumption (A3), a scalar \( \hat{\phi}_c \) satisfying assumption (A6), a smoothing constant \( \xi \geq D_y/\rho_y \), an initial point \((x_0, y_0) \in X \times Y\), and a tolerance triple \((\rho_x, \rho_y, \eta) \in \mathbb{R}^3_+\);

**Output:** a triple \((\bar{u}, \bar{v}, \bar{x}, \bar{y}, \bar{r})\) satisfying (1.8);

1. set \( L_\xi \) as in (2.15), \( \sigma = 1/2, \lambda = 1/(4m) \), and define \( p_\xi \) as in (2.12);
2. apply the QP-AIPP method of Subsection 4.1 with inputs \((m, L_\xi), (p_\xi, h), \lambda, \sigma, \hat{\phi}_c, x_0, (\rho_x, \eta)\) to obtain a triple \((\bar{u}, \bar{x}, \bar{r})\) satisfying
\[
(4.12) \quad \bar{u} \in \nabla p_\xi(\bar{x}) + \partial h(\bar{x}) + A^* \bar{r}, \quad \|\bar{u}\| \leq \rho_x, \ \|A \bar{x} - b\| \leq \eta.
\]
3. define \((\bar{v}, \bar{y})\) as in (3.19) and output the quintuple \((\bar{u}, \bar{v}, \bar{x}, \bar{y}, \bar{r})\).

Some remarks about the above method are in order. First, the QP-AIPP method invoked in step 1 terminates due to the remarks following assumptions (Q1)–(Q2) and the results in Subsection 4.1. Second, since the QP-AIPP-S scheme is a one-pass algorithm (as opposed to an iterative algorithm), the complexity of the QP-AIPP-S scheme is essentially that of the QP-AIPP method. Finally, while the QP-AIPP method in step 2 is called with \((\sigma, \lambda) = (1/2, 1/(4m))\), it can also be called with any \(\sigma \in (0, 1)\) and \(\lambda \in (0, 1/(2m))\) to establish the desired termination of the QP-AIPP-S scheme.

We now show how the QP-AIPP-S scheme generates a point \((\bar{u}, \bar{v}, \bar{x}, \bar{y}, \bar{r})\) satisfying (1.8). Recall the definition of “oracle call” in the paragraph containing (1.3).
Proposition 11. Let a tolerance triple $(\rho_x, \rho_\eta, \eta) \in \mathbb{R}^3_+$ be given and let the quadruple $(\bar{u}, \bar{v}, \bar{x}, \bar{y}, \bar{r})$ be the output obtained by the QP-AIPP-S scheme. Then the following properties hold:

(a) the QP-AIPP-S scheme terminates in

\begin{equation}
(4.13) \quad \mathcal{O} \left( \Omega_{\xi, \eta} \left[ \frac{m^2 R_c(\hat{p}; 1/(4m))}{\rho_x^2} + \frac{m D_y^2}{\xi \rho_x^2} + \log_1^+ (\Omega_{\xi, \eta}) \right] \right)
\end{equation}

oracle calls, where

\begin{equation}
(4.14) \quad \Omega_{\xi, \eta} := \Omega_{\xi} + \left( R_c(\hat{p}; 1/(4m)) + \frac{D_y^2}{8m\xi} \right)^{1/2} \frac{\|A\|}{\eta}
\end{equation}

and $\Omega_{\xi}$, $R(\cdot, \cdot)$, and $D_y$ are as in (3.21), (3.12), and (2.8), respectively;

(b) the quintuple $(\bar{u}, \bar{v}, \bar{x}, \bar{y}, \bar{r})$ satisfies (1.8).

Proof. (a) Let $\Theta_i$ be as in (4.10) with $M = L_\xi$. Using the same arguments as in Lemma 7, it is easy to see that

\begin{equation}
(4.15) \quad R_c(\hat{p}_c; 1/(4m)) \leq R_c(\hat{p}; 1/(4m)) + \frac{D_y^2}{8m\xi}.
\end{equation}

and hence, using (3.22) and (4.15), we have

\begin{equation}
(4.16) \quad \sqrt{\frac{\Theta_i}{4m}} + 1 \leq 1 + \sqrt{\frac{L_\xi}{4m}} + \sqrt{\frac{4R_c(\hat{p}_c; 1/(4m))\|A\|^2}{\eta^2}}
\end{equation}

\begin{equation}
(4.17) \quad \leq 1 + \sqrt{\frac{\tilde{\xi} L_y + \sqrt{L_x}}{2\sqrt{m}}} + 2 \left( R_c(\hat{p}; 1/(4m)) + \frac{D_y^2}{8m\xi} \right)^{1/2} \frac{\|A\|}{\eta} = \Theta(\Omega_{\xi, \eta}).
\end{equation}

The complexity in (4.13) now follows from Proposition 10 with $(\phi, f, M) = (p, p_\xi, L_\xi)$, (4.16), and (4.15).

(b) The top inclusion and bounds involving $\|\bar{u}\|$ and $\|A \bar{x} - b\|$ in (1.8) follow from Proposition 4(b), the definition of $\bar{g}$ in step 2 of the algorithm, and Proposition 10 with $f = \hat{p}_c$. The bottom inclusion and bound involving $\|\bar{v}\|$ follow from similar arguments given for Proposition 8(b).

We now make three remarks about the above complexity bound. First, recall that $R_c(p; 1/(4m))$ in the complexity (11) can be majorized by the rightmost quantity in (4.9) with $\lambda = 1/(4m)$. Second, under the assumption that $\xi = D_y/\rho_y$, the complexity of the QP-AIPP-S scheme reduces to

\begin{equation}
(4.18) \quad \mathcal{O} \left( m^{3/2} \cdot R_c(\hat{p}; 1/(4m)) \cdot \left[ \frac{L_x^{1/2}}{\rho_x^2} + \frac{L_y D_y^{1/2}}{\rho_y^{1/2} \rho_x^2} + \frac{m^{1/2}\|A\| \|\hat{R}_c^{1/2}(p; 1/(4m))\|}{\eta \rho_x^2} \right] \right),
\end{equation}

under the reasonable assumption that the $\mathcal{O}(\rho_x^{-2} + \eta^{-1} \rho_x^{-2} + \rho_y^{-1} \rho_x^{-2})$ term in (4.13) dominates the other terms. Third, when $Y$ is a singleton, it is easy to see that (1.6) becomes a special instance of the linearly-constrained smooth CNO problem (4.3), the QP-AIPP-S of this subsection becomes equivalent to the QP-AIPP method of Subsection 4.1, and the complexity in (4.17) reduces to

\begin{equation}
(4.19) \quad \mathcal{O} \left( m^{3/2} \cdot R_c(\hat{p}; 1/(4m)) \cdot \left[ \frac{L_x^{1/2}}{\rho_x^2} + \frac{m^{1/2}\|A\| \|\hat{R}_c^{1/2}(p; 1/(4m))\|}{\eta \rho_x^2} \right] \right).
\end{equation}
In view of the last remark, the $O(\rho^{-2} \rho^{-1/2})$ term in (4.17) is attributed to the (possible) nonsmoothness in (1.6).

Let us now conclude this section with a remark about the penalty subproblem

\[ \min_{x \in X} \left\{ p_\xi(x) + h(x) + \frac{c}{2} \|Ax - b\|^2 \right\}, \tag{4.19} \]

which is what the AIPP method considers every time it is called in the QP-AIPP-S scheme (see step 1). First, observe that (1.6) can be equivalently reformulated as

\[ \min_{x \in X} \max_{y \in \mathcal{Y}, r \in \mathcal{U}} [\Psi(x, y, r) := \Phi(x, y) + h(x) + \langle r, Ax - b \rangle]. \tag{4.20} \]

Second, it is straightforward to verify that problem (4.19) is equivalent to

\[ \min_{x \in X} \{ \tilde{p}_{c, \xi}(x) := p_{c, \xi}(x) + h(x) \}, \tag{4.21} \]

where the function $p_{c, \xi} : X \mapsto \mathbb{R}$ is given by

\[ p_{c, \xi}(x) := \max_{y \in \mathcal{Y}, r \in \mathcal{U}} \left\{ \Psi(x, y, r) - \frac{1}{2c} \|r\|^2 - \frac{1}{2\xi} \|y - y_0\|^2 \right\} \quad \forall x \in X \tag{4.22} \]

with $\Psi$ as in (4.20). As a consequence, problem (4.21) is similar to (3.1) in that a smooth approximate is used in place of the nonsmooth component of the underlying saddle function $\Psi$. On the other hand, observe that we cannot directly apply the smoothing scheme developed in Subsection 3.2 to (4.21) as the set $\mathcal{U}$ is generally unbounded. One approach that avoids this problem is to invoke the AIPP method of Subsection 3.1 to solve a sequence subproblems of the form in (4.21) for increasing values of $c$. However, in view of the equivalence of (4.19) and (4.21), this is exactly the approach taken by the QP-AIPP-S scheme of this section.

5. Numerical experiments. This section presents numerical results that illustrate the computational efficiency of the our proposed smoothing scheme. It contains three subsections. Each subsection presents computational results for a specific unconstrained nonconvex min-max optimization problem class.

Each unconstrained problem considered in this section is of the form in (1.1) and is such that the computation of the function $y_\xi$ in (2.13) is easy. Moreover, for a given initial point $x_0 \in X$, three algorithms are run for each problem instance until a quadruple $(\bar{u}, \bar{v}, \bar{x}, \bar{y})$ satisfying

\[ \left( \begin{array}{c} \bar{u} \\ \bar{v} \end{array} \right) \in \left( \begin{array}{c} \nabla_x \Phi(\bar{x}, \bar{y}) \\ 0 \end{array} \right) + \left( \begin{array}{c} \partial h(\bar{x}) \\ \partial [-\Phi(\bar{x}, \cdot)](\bar{y}) \end{array} \right), \]

\[ \frac{\|\bar{u}\|}{\|\nabla p_\xi(z_0)\| + 1} \leq \rho_x, \quad \|\bar{v}\| \leq \rho_y, \tag{5.1} \]

is obtained, where $\xi = D_y / \rho_y$.

We now describe the three nonconvex-concave min-max methods that are being compared in this section, namely: (i) the R-AIPP-S method; (ii) the accelerated gradient smoothing (AG-S) scheme; and (iii) the projected gradient step framework (PGSF). Both the AG-S and R-AIPP-S schemes are modifications of the AIPP-S scheme which, instead of using the AIPP method in its step 1, use the AG method of [11] and R-AIPP method of [15], respectively. The PGSF is a simplified variant.

This manuscript is for review purposes only.
Note that, like the AIPP method, the R-AIPP similarly: (i) invokes at each of its
(outter) iterations an ACG method to inexactly solve the proximal subproblem (3.10);
and (ii) outputs a $\tilde{p}$-approximate stationary point of (3.2). However, the R-AIPP
method is more computationally efficient due to three key practical improvements
over the AIPP method, namely: (i) it allows the stepsizes $\lambda$ to be significantly larger
than the $1/(2m)$ upper bound in the AIPP method using adaptive estimates of $m$;
(ii) it uses a weaker ACG termination criterion compared to the one in (3.6); and (iii)
it does not prespecify the minimum number of ACG iterations as the AIPP method
does in its step 1.

The AG method is implemented as described in Algorithm 2 of [11]. We now
describe the details of our implementation of the R-AIPP method. We assume that
the reader is familiar with an iteration of the R-AIPP method of [15] and its various
parameters, e.g. $\lambda, \lambda_k, \theta, \tau$, etc. A single iteration of our R-AIPP implementation is
the same as an iteration of the R-AIPP method of [15] with $\theta = 4$, except that $\lambda_k$
and $\tau$ are updated differently at the end of the iteration. In order to describe the
update rule for $\lambda_k$, we first define an iteration $k \geq 1$ to be “good” if the intermediate
parameter $\lambda$ has not been halved in step 1 or 2 of the R-AIPP method for any iteration
$j \leq k$. The update rule for $\lambda_k$ at the end of the $k^{th}$ iteration is then

$$
\lambda_k = \begin{cases} 
\min \{2\lambda, 100/m\}, & \text{if iteration } k \text{ is "good"} \\
\lambda, & \text{otherwise,}
\end{cases}
$$

with $\lambda_0$ set to be $1/m$. The update rule for $\tau$ at the end of the $k^{th}$ iteration is

$$
(5.2) \quad \tau = \begin{cases} 
1.5\tau_{prev}/\pi_k, & \text{if } \pi_k > 1.5, \\
1.2\tau_{prev}/\pi_k, & \text{if } \pi_k < 1.2, \\
\tau_{prev}, & \text{otherwise,}
\end{cases}
$$

where $\tau_{prev}$ denotes the value of $\tau$ at the end of the $(k-1)^{th}$ iteration, $\pi_k := 
(\lambda_k||\bar{\theta}_k||)/||v_k + z_{k-1} - z_k||$, and $\tau_{prev}$ is set to be $10(\lambda_0 M + 1)$ at the first itera-
tion. Also, like the implementations of the R-AIPP method in [15], the R-AIPP
implementation in this section adaptively estimates the constant $\bar{M}$ — equivalently
$L$ (and hence $M$) in the ACG method of Appendix A — used in every iteration of
the R-ACG subroutine that is called in its step 1.

It worth noting the update rules for $\lambda_k$ and $\tau$ in the previous paragraph are
different from the implementations of the R-AIPP method in [15]. More specifically,
in [15] the iterates $\lambda_k$ are simply set to be $\lambda$ and $\tau$ is set to be $\tau_{prev}$ at every iteration
$k \geq 1$. Moreover, the parameter $\tau$ in the experiments of [15] was chosen according
to the problem class being examined and, hence, the update of $\tau$ in (5.2) has the
advantage that it dynamically adjusts to the geometry of a problem instance.

We also state some additional details about the numerical experiments. First,
each algorithm is run with a time limit of 4000 seconds. If an algorithm does not
terminate with a solution for a particular problem instance, we do not report any
details about its iteration count or function value at the point of termination and the
runtime for that instance is marked with a [*] symbol. Second, the iterations listed in
the tables of this section include even those that are extraneously performed due to an
improper choice of an input parameter, e.g., $\lambda_0$ and $\bar{M}$. Third, the bold numbers in
each of the computational tables in this section highlight the algorithm that performed
the most efficiently in terms of iteration count or total runtime. Moreover, each of tables contain a column labeled $\bar{p}_k(x)$ that contains the smallest obtained value of the smoothed function in (3.1), across all of the tested algorithms. Fourth, the description of $y_k$ and justification of the constants $m, L_x, \text{ and } L_y$ for each of the considered optimization problems are given in Appendix E. Fifth, $y_0$ is chosen to be 0 for all of the experiments. Finally, all algorithms described at the beginning of this section are implemented in MATLAB 2019a and are run on Linux 64-bit machines each containing Xeon E5520 processors and at least 8 GB of memory.

Before proceeding, it is worth mentioning that the code for generating the results of this section is available online\textsuperscript{2}.

5.1. Maximum of a finite number of nonconvex quadratic forms. This subsection presents computational results for a minmax quadratic vector problem, which is based on a similar problem in [15].

We first describe the problem. Given a dimension triple $(n,l,k) \in \mathbb{N}^3$, a set of parameters $\{(\alpha_i, \beta_i)\}_{i=1}^k \subseteq \mathbb{R}_{++}^2$, a set of vectors $\{d_i\}_{i=1}^k \subseteq \mathbb{R}^l$, a set of diagonal matrices $\{D_i\}_{i=1}^k \subseteq \mathbb{R}^{n \times n}$, and matrices $\{C_i\}_{i=1}^k \subseteq \mathbb{R}^{l \times n}$ and $\{B_i\}_{i=1}^k \subseteq \mathbb{R}^{n \times n}$, the problem of interest is the quadratic vector minmax (QVM) problem

$$
\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^k} \left\{ \delta_{\Delta^k}(x) + \sum_{i=1}^k y_i g_i(x) : y \in \Delta^k \right\},
$$

where, for every index $1 \leq i \leq k$, integer $p \in \mathbb{N}$, and $x \in \mathbb{R}^n$,

$$
g_i(x) := \frac{\alpha_i}{2} \|C_i x - d_i\|^2 - \frac{\beta_i}{2} \|D_i B_i x\|^2, \quad \Delta^p := \left\{ z \in \mathbb{R}^p_+ : \sum_{i=1}^p z_i = 1, z \geq 0 \right\}.
$$

We now describe the experiment parameters for the instances considered. First, the dimensions are set to be $(n,l,k) = (200, 10, 5)$ and only 5.0% of the entries of the submatrices $B_i$ and $C_i$ are nonzero. Second, the entries of $B_i, C_i$, and $d_i$ (resp., $D_i$) are generated by sampling from the uniform distribution $U[0, 1]$ (resp., $U[1, 1000]$). Third, the initial starting point is $z_0 = I_n/n$, where $I_n$ is the $n$-dimensional identity matrix. Fourth, with respect to the termination criterion (5.1), the inputs, for every $(x, y) \in \mathbb{R}^n \times \mathbb{R}^k$, are

$$
\Phi(x, y) = \sum_{i=1}^k y_i g_i(x), \quad h(x) = \delta_{\Delta^k}(x), \quad \rho_x = 10^{-2}, \quad \rho_y = 10^{-1}, \quad Y = \Delta^k.
$$

Fifth, each problem instance considered is based on a specific curvature pair $(m, M) \in \mathbb{R}^{2+}$ satisfying $m \leq M$, for which each scalar pair $(\alpha_i, \beta_i) \in \mathbb{R}^{2+}$ is selected so that

$$
M = \lambda_{\text{max}}(\nabla^2 g_i), \quad -m = \lambda_{\text{min}}(\nabla^2 g_i).
$$

Finally, the Lipschitz and curvature constants selected are

$$
m = m, \quad L_x = M, \quad L_y = M \sqrt{k} + \|P\|,
$$

where $P$ is an $n$–by–$k$ matrix whose $i^{th}$ column is equal to $\alpha_i C_i^T d_i$.

We now present the results in Table 5.1.

\textsuperscript{2}See the examples in ./examples/minmax/ from the GitHub repository https://github.com/wwkong/nc_opt/.

This manuscript is for review purposes only.
5.2. Truncated robust regression. This subsection presents computational results for the robust regression problem in [26]. It is worth mentioning that [26] also presents a min-max algorithm for obtaining a stationary point as in (5.1). However, its iteration complexity, which is $O(\rho^{-6})$ when $\rho = \rho_x = \rho_y$, is significantly worse than the other algorithms considered in this section and, hence, we choose not to include this algorithm in our benchmarks.

We now describe the problem. Given a dimension pair $(n, k) \in \mathbb{N}^2$, a set of $n$ data points $\{(a_j, b_j)\}_{j=1}^n \subseteq \mathbb{R}^k \times \{1, -1\}$ and a parameter $\alpha > 0$, the problem of interest is the truncated robust regression (TRR) problem

$$\min_{x \in \mathbb{R}^k} \max_{y \in \mathbb{R}^n} \left\{ \sum_{j=1}^n y_j (\phi_\alpha \circ \ell_j)(x) : y \in \Delta^n \right\},$$

where $\Delta^n$ is as in (5.3) with $p = n$ and, for every $(\alpha, t, x) \in \mathbb{R}_{++} \times \mathbb{R}_{++} \times \mathbb{R}^k$,

$$\phi_\alpha(t) := \alpha \log \left(1 + \frac{t}{\alpha}\right), \quad \ell_j(x) := \log \left(1 + e^{-b_j \langle a_j, x \rangle}\right).$$

We now describe the experiment parameters for the instances considered. First, $\alpha$ is set to 10 and the data points $\{(a_j, b_j)\}$ are taken from different datasets in the LIBSVM library\(^3\) for which each problem instance is based off of (see the “data name” column in the table below, which corresponds to a particular LIBSVM dataset). Second, the initial starting point is $z_0 = 0$. Third, with respect to the termination criterion (5.1), the inputs, for every $(x, y) \in \mathbb{R}^k \times \mathbb{R}^n$, are

$$\Phi(x, y) = \sum_{j=1}^n y_j (\phi_\alpha \circ \ell_j)(x), \quad h(x) = 0, \quad \rho_x = 10^{-5}, \quad \rho_y = 10^{-3}, \quad Y = \Delta^n.$$

Finally, the Lipschitz and curvature constants selected are

$$m = L_x = \frac{1}{\alpha} \max_{1 \leq j \leq n} \|a_j\|^2, \quad L_y = \sqrt{\sum_{j=1}^n \|a_j\|^2}.$$

We now present the results in Table 5.2.

5.3. Power control in the presence of a jammer. This subsection presents computational results for the power control problem in [18].

\(^3\)See https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html.
It is worth mentioning that [18] also presents a min-max algorithm for obtaining stationary points for the aforementioned problem. However, its termination criterion and notion of stationarity are significantly different than what is being considered in this paper and, hence, we choose not to include the algorithm of [18] in our benchmarks.

We now describe the problem. Given a dimension pair \((N, K)\) \(\in \mathbb{N}^2\), a pair of parameters \((\sigma, R)\) \(\in \mathbb{R}^2_{+}\), a 3D tensor \(A \in \mathbb{R}^{K \times K \times N}\), and a matrix \(B \in \mathbb{R}^{K \times N}\), the problem of interest is the power control (PC) problem

\[
\min_{X \in \mathbb{R}^{K \times N}} \max_{y \in \mathbb{R}^N} \left\{ \sum_{k=1}^{K} \sum_{n=1}^{N} f_{k,n}(X, y) : 0 \leq X \leq R, 0 \leq y \leq \frac{N}{2}, \right\},
\]

where, for every \((X, y) \in \mathbb{R}^{K \times N} \times \mathbb{R}^N\),

\[
f_{k,n}(X, y) := -\log \left( \frac{A_{k,k,n}X_{k,n}}{\sigma^2 + B_{k,n}y_n + \sum_{j=1, j \neq k}^{K} A_{j,k,n}X_{j,n}} \right).
\]

We now describe the experiment parameters for the instances considered. First, the scalar parameters are set to be \((\sigma, R) = (1/\sqrt{2}, K^{1/K})\) and the quantities \(A\) and \(B\) are set to be the squared moduli of the entries of two Gaussian sampled complex-valued matrices \(\mathcal{H} \in \mathbb{C}^{K \times K \times N}\) and \(P \in \mathbb{C}^{K \times N}\). More precisely, the entries of \(\mathcal{H}\) and \(P\) are sampled from the standard complex Gaussian distribution \(\mathcal{C}\mathcal{N}(0,1)\) and

\[
|A_{j,k,n}| = |\mathcal{H}_{j,k,n}|^2, \quad |B_{k,n}| = |P_{k,n}|^2 \quad \forall (j, k, n).
\]

Second, the initial starting point is \(z_0 = 0\). Third, with respect to the termination criterion (5.1), the inputs, for every \((X, y) \in \mathbb{R}^{K \times N} \times \mathbb{R}^N\), are

\[
\Phi(X, y) = \sum_{k=1}^{K} \sum_{n=1}^{N} f_{k,n}(X, y), \quad h(X) = \delta_{Q_{R}^{K \times N}}(X),
\]

\[
\rho_x = 10^{-1}, \quad \rho_y = 10^{-1}, \quad Y = Q_{N/2}^{N \times 1}.
\]

where \(Q_{U \times V}^T := \{ z \in \mathbb{R}^{p \times q} : 0 \leq z \leq T \} \) for every \(T > 0\) and \((U, V) \in \mathbb{N}^2\). Fourth, each problem instance considered is based on a specific dimension pair \((N, K)\). Finally, the Lipschitz and curvature constants selected are

\[
m = L_x = \frac{2}{\min\{\sigma^4, \sigma^6\}} \max_{1 \leq k \leq K} \sum_{1 \leq n \leq N} A_{k,j,n}^2, \quad L_y = \frac{2}{\min\{\sigma^4, \sigma^6\}} \max_{1 \leq n \leq N} \sum_{j=1}^{K} B_{j,n}A_{k,j,n}.
\]

We now present the results in Table 5.3.

<table>
<thead>
<tr>
<th>data name</th>
<th>(\hat{p}_x(\bar{x}))</th>
<th>Iteration Count</th>
<th>Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R-AIPP-S</td>
<td>AG-S</td>
</tr>
<tr>
<td>heart</td>
<td>6.70E-01</td>
<td>425</td>
<td>1747</td>
</tr>
<tr>
<td>diabetes</td>
<td>6.70E-01</td>
<td>852</td>
<td>1642</td>
</tr>
<tr>
<td>ionosphere</td>
<td>6.70E-01</td>
<td>1197</td>
<td>8328</td>
</tr>
<tr>
<td>sonar</td>
<td>6.70E-01</td>
<td>45350</td>
<td>96209</td>
</tr>
<tr>
<td>breast-cancer</td>
<td>1.11E-03</td>
<td>46097</td>
<td>-</td>
</tr>
</tbody>
</table>

*Table 5.2: Iteration counts and runtimes for TRR problems.*
### Table 5.3

<table>
<thead>
<tr>
<th>$N$</th>
<th>$K$</th>
<th>$\hat{p}_k(\bar{x})$</th>
<th>Iteration Count</th>
<th>Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>R-AIPP-S</td>
<td>AG-S</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>-3.64E+00</td>
<td>37</td>
<td>322832</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>-2.82E+00</td>
<td>54</td>
<td>33399</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
<td>-4.52E+00</td>
<td>183</td>
<td>-</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>-4.58E+00</td>
<td>566</td>
<td>-</td>
</tr>
</tbody>
</table>

### 6. Concluding Remarks

This section makes some concluding remarks.

We first make a final remark about the AIPP-S smoothing scheme. Recall that the main idea of AIPP-S is to call the AIPP method to obtain a pair satisfying (3.14), or equivalently

\[
\inf_{\|d\| \leq 1} (\hat{p}_\xi)'(x; d) \geq -\rho.
\]

Moreover, using Proposition 8 with $(\rho_x, \rho_y) = (\rho, D_y/\xi)$, it straightforward to see that the number of oracle calls, in terms of $(\xi, \rho)$, is $O(\rho^{-2}\xi^{1/2})$, which reduces to $O(\rho^{-2.5})$ if $\xi$ is chosen so as to satisfy $\xi = \Theta(\rho^{-1})$. The latter complexity bound improves upon the one obtained for an algorithm in [24] which obtains a point $x$ satisfying (6.1) with $\xi = \Theta(\rho^{-1})$ in $O(\rho^{-3})$ oracle calls.

We now discuss some possible extensions of this paper. First, it is worth investigating whether complexity results for the AIPP-S method can be derived for the case where $Y$ is unbounded. Second, it is worth investigating if the notions of stationary points in Subsection 2.1 are related to first-order stationary points of the related mathematical program with equilibrium constraints:

\[
\min_{(x,y) \in X \times Y} \{\Phi(x, y) + h(y) : 0 \in \partial[-\Phi(\cdot, y)](x)\}.
\]

Finally, it remains to be seen if a similar prox-type smoothing scheme can be developed for the case in which assumption (A2) is relaxed to the condition that there exists $m_y > 0$ such that $-\Phi(x, \cdot)$ is $m_y$-weakly convex for every $x \in X$.

### Appendix A

This appendix contains a description and a result about an ACG variant used in the analysis of [16].

Part of the input of the ACG variant, which is described below, consists of a pair of functions $(\psi_n, \psi_s)$ satisfying:

1. $\psi_n \in \text{Conv}(Z)$ is $\mu$-strongly convex for some $\mu \geq 0$;
2. $\psi_s$ is a convex differentiable function on $\text{dom} \psi_n$ whose gradient is $L$-Lipschitz continuous for some $L > 0$.

### ACG method

**Input:** a scalar pair $(\mu, L) \in \mathbb{R}^2_{++}$, a function pair $(\psi_n, \psi_s)$, and an initial point $z_0 \in \text{dom} \psi_n$;

1. set $y_0 = z_0$, $A_0 = 0$, $\Gamma_0 \equiv 0$ and $j = 0$;

---

4 See Lemma 16 with $f = \hat{p}_\xi$.

5 See, for example, [19, Chapter 3].
We now discuss some implementation details of the ACG method. First, a single iteration requires the evaluation of two distinct types of oracles, namely: (i) the evaluation of the functions $\psi_n, \psi_s, \nabla \psi_s$ at any point in $\text{dom} \psi_n$; and (ii) the computation of the exact solution of subproblems of the form

\[
\min_y \left\{ \psi_n(y) + \frac{1}{2\alpha} \|y - a\|^2 \right\}
\]

for any $a \in \mathcal{Z}$ and $\alpha > 0$. In particular, the latter is needed in the computation of $y_{j+1}$. Second, because $\Gamma_{j+1}$ is affine, an efficient way to store it is in terms of a normal vector and a scalar intercept that is updated recursively at every iteration. Indeed, if $\Gamma_j = \alpha_j + \langle \cdot, \beta_j \rangle$ for some $(\alpha_j, \beta_j) \in \mathbb{R} \times \mathcal{Z}$, then step 1 of the ACG method implies that $\Gamma_{j+1} = \alpha_{j+1} + \langle \cdot, \beta_{j+1} \rangle$ where

\[
\begin{align*}
\alpha_{j+1} &:= \frac{A_j}{A_{j+1}} \alpha_j + \frac{A_{j+1} - A_j}{A_{j+1}} [\psi_s(\tilde{z}_j) - \langle \nabla \psi_s(\tilde{z}_j), \tilde{z}_j \rangle], \\
\beta_{j+1} &:= \frac{A_j}{A_{j+1}} \beta_j + \frac{A_{j+1} - A_j}{A_{j+1}} [\nabla \psi_s(\tilde{z}_j)].
\end{align*}
\]

The following result, whose proof is given in [16, Lemma 9], is used to establish the iteration complexity of obtaining the triple $(z, u, \varepsilon)$ in step 1 of the AIPP method of Subsection 3.1.

**Lemma 12.** Let $\{(A_j, z_j, u_j, \varepsilon_j)\}$ be the sequence generated by the ACG method. Then, for any $\sigma > 0$, the ACG method obtains a triple $(z, u, \varepsilon)$ satisfying

\[
\begin{align*}
u &\in \partial_u (\psi_s + \psi_n)(z) \quad \|u\|^2 + 2\varepsilon \leq \sigma \|z_0 - z + u\|^2
\end{align*}
\]

in at most \(2\sqrt{2L}(1 + \sqrt{\sigma})/\sqrt{\sigma}\) iterations.
Appendix B. This appendix contains results about functions that can be described as the maximum of a family of differentiable functions.

The technical lemma below, which is a special case of [10, Theorem 10.2.1], presents a key property about max functions.

**Lemma 13.** Assume that the triple \((X,Y,\Psi)\) satisfies (A0)–(A1) in Subsection 2.1 with \(\Phi = \Psi\). Moreover, define

\[(B.1) \quad q(x) := \sup_{y \in Y} \Psi(x,y), \quad Y(x) := \\{y \in Y : \Psi(x,y) = q(x)\}, \quad \forall x \in X.\]

Then, for every \((x,d) \in X \times X\), it holds that

\[q'(x;d) = \max_{y \in Y(x)} \langle \nabla_x \Psi(x,y), d \rangle.\]

Moreover, if \(Y(x)\) reduces to a singleton, say \(Y(x) = \{y(x)\}\), then \(q\) is differentiable at \(x\) and \(\nabla q(x) = \nabla_x \Psi(x,y(x))\).

Under assumptions (A0)–(A3) in Subsection 2.1, the next result establishes Lipschitz continuity of the gradient of \(q\). It is worth mentioning that it generalizes related results in [2, Theorem 5.26] (which covers the case where \(\Psi\) is bilinear) and [20, Proposition 4.1] (which makes the stronger assumption that \(\Psi(\cdot,y)\) is convex for every \(y \in Y\)).

**Proposition 14.** If the triple \((X,Y,\Psi)\) satisfies (A0)–(A3) in Subsection 2.1 with \(\Phi = \Psi\) and, for some \(\mu > 0\), the function \(\Psi(x,\cdot)\) is \(\mu\)-strongly concave on \(Y\) for every \(x \in X\), then the following properties hold:

- (a) the function \(y(\cdot)\) given by
  \[y(x) := \arg\max_{y \in Y} \Psi(x,y) \quad \forall x \in X\]
  is \(Q_\mu\)-Lipschitz continuous on \(X\), where

  \[(B.2) \quad Q_\mu := \frac{L_y}{\mu} + \sqrt{\frac{L_x + m}{\mu}};\]

- (b) \(\nabla q(\cdot)\) is \(L_\mu\)-Lipschitz continuous on \(X\), where \(q\) is as in (B.1) and

  \[(B.3) \quad L_\mu := L_y Q_\mu + L_x.\]

**Proof.** (a) Let \(x, x' \in X\) be given and denote \((y,\tilde{y}) = (y(x), y(x'))\). Define

\[(B.4) \quad \alpha(u) := \Psi(u,y) - \Psi(u,\tilde{y}) \quad \forall u \in X.\]

and observe that the optimality conditions of \(y\) and \(\tilde{y}\) imply that

\[(B.5) \quad \alpha(x) \geq \frac{\mu}{2} \|y - \tilde{y}\|^2, \quad -\alpha(x') \geq \frac{\mu}{2} \|y - \tilde{y}\|^2.\]

Using (B.5), (2.1), (2.2), (2.3), and the Cauchy-Schwarz inequality, we conclude that

\[\mu \|y - \tilde{y}\|^2 \leq \alpha(x) - \alpha(x') \leq \langle \nabla_x \Psi(x,y) - \nabla_x \Psi(x,\tilde{y}), x - x' \rangle + \frac{L_x + m}{2} \|x - x'\|^2\]

\[\leq \|\nabla_x \Psi(x,y) - \nabla_x \Psi(x,\tilde{y})\| \cdot \|x - x'\| + \frac{L_x + m}{2} \|x - x'\|^2\]

This manuscript is for review purposes only.
\[ \leq L_y \| y - \hat{y} \| \cdot \| x - \bar{x} \| + \frac{L_x + m}{2} \| x - \bar{x} \|^2. \]

Considering the above as a quadratic inequality in \( \| y - \hat{y} \| \) yields the bound
\[
\| y - \hat{y} \| \leq \frac{1}{2\mu} \left[ L_y \| x - \bar{x} \| + \sqrt{L_y^2 \| x - \bar{x} \|^2 + 4\mu(L_x + m) \| x - \bar{x} \|^2} \right]
\]
\[
\leq \left[ \frac{L_y}{\mu} + \sqrt{\frac{L_x + m}{\mu}} \right] \| x - \bar{x} \| = Q_\mu \| x - \bar{x} \|
\]

which is the conclusion of (a).

(b) Let \( x, \bar{x} \in X \) be given and denote \( (y, \hat{y}) = (y(x), y(\bar{x})) \). Using part (a), Lemma 13, and (2.2) we have that
\[
\| \nabla q(x) - \nabla q(\bar{x}) \| = \| \nabla_x \Psi(x, y) - \nabla_x \Psi(\bar{x}, \hat{y}) \|
\]
\[
\leq \| \nabla_x \Psi(x, y) - \nabla_x \Psi(x, \bar{y}) \| + \| \nabla_x \Psi(x, \bar{y}) - \nabla_x \Psi(\bar{x}, \hat{y}) \|
\]
\[
\leq L_y \| y - \bar{y} \| + L_x \| x - \bar{x} \| \leq (L_y Q_\mu + L_x) \| x - \bar{x} \| = L_\mu \| x - \bar{x} \|,
\]

which is the conclusion of (b). □

**Appendix C.** The main goal of this appendix is to prove Propositions 18 and 19, which are used in the proofs of Propositions 1, 2, and 3 given in Appendix D. Several technical lemmas are stated and proved to accomplish the above goal. Some of these technical results (e.g., Lemmas 15(a) and 17) are stated without proof as they are broadly available in the convex analysis literature. Others (e.g., Lemmas 15(b) and 16) are given proofs because we could not find a suitable reference for them.

The first technical lemma presents some general results about proper convex functions and nonempty closed convex sets.

**Lemma 15.** Let \( \psi \) be a convex function and let \( C \subseteq X \) be a nonempty closed convex set. Then, the following statements hold:

(a) \( \inf_{\|d\| \leq 1} \sigma_C(d) = \lim_{\|u\| \to 0} \inf_C \| u \|; \)

(b) if \( C \cap \text{ri}(\text{dom} \psi) \neq \emptyset, \) then \( \inf_{x \in C} \text{cl} \psi(x) = \inf_{x \in C} \psi(x) < \infty. \)

**Proof.** (a) See, for example, the proof of [3, Lemma 5.1] with \( y = 0. \)

(b) Define \( \psi_* := \inf_{x \in C} \psi(x), \quad \psi_*^\circ := \inf_{x \in C} \text{cl} \psi(x). \) Then, note that the assumption of (b) implies that \( \psi_* < \infty. \) Now, assume for contradiction that the conclusion of (b) does not hold. Since \( \text{cl} \psi \leq \psi, \) and hence \( \psi_*^\circ \leq \psi_* \), we must have \( \psi_*^\circ < \psi_* \). Hence, due to a well-known infimum property, there exists \( \bar{x} \in C \) such that \( \text{cl} \psi(\bar{x}) < \psi_* < \infty. \) In particular, it follows that \( \psi_* \in \mathbb{R}, \) and hence that \( \psi(x) > -\infty \) for every \( x \in C, \) in view of the definition of \( \psi_* \). Now, by assumption, there exists \( x_0 \in C \cap \text{ri}(\text{dom} \psi) \) which, in view of the previous conclusion, satisfies \( \psi(x_0) > -\infty. \)

As \( x_0 \in \text{ri}(\text{dom} \psi), \) this implies that \( \psi \) is proper due to [27, Theorem 7.2]. Hence, in view of [27, Theorem 7.5] with \( f = \psi, \) we have
\[
\text{cl} \psi(\bar{x}) = \lim_{y \in (\bar{x}, x_0), y \to \bar{x}} \psi(y)
\]

where \( (\bar{x}, x_0) := \{ tx_0 + (1 - t) \bar{x} : t \in (0, 1]\}. \) On the other hand, as \( x_0, \bar{x} \in C \) and \( C \) is convex, we have \( (\bar{x}, x_0) \subseteq C. \) This inclusion and the definition of \( \psi_* \) then imply that the above limit, and hence \( \text{cl} \psi(\bar{x}), \) is greater than or equal to \( \psi_*, \) which contradicts the previously obtained inequality \( \psi_* > \text{cl} \psi(\bar{x}). \) □
The following technical lemma presents an important property about the directional derivative of a composite function \((f + h)\).

**Lemma 16.** Let \(h : \mathcal{X} \mapsto (-\infty, \infty] \) be a proper convex function and let \(f\) be a differentiable function on \(\text{dom } h\). Then, for any \(x \in \text{dom } h\), it holds that

\[
\inf_{\|d\| \leq 1} (f + h)'(x; d) = \inf_{\|d\| \leq 1} \left[ \langle \nabla f(x), d \rangle + \sigma_{\partial h(x)}(d) \right] = - \inf_{u \in \nabla f(x) + \partial h(x)} \|u\|.
\]

**Proof.** Let \(x \in \text{dom } h\) be fixed and define \(\tilde{h}(\cdot) := \langle \nabla f(x), \cdot \rangle + h(\cdot)\). We first claim that \(\inf_{\|d\| \leq 1} \tilde{h}'(x; d) = \inf_{\|d\| \leq 1} [\sigma_{\partial h(x)}(d)]\). Before showing this claim, let us show how it proves the desired conclusion. Since the definition of \(\tilde{h}\) implies that \((f + h)'(x; \cdot) = \tilde{h}'(x; \cdot)\) and \(\partial \tilde{h}(x) = \nabla f(x) + \partial h(x)\), it follows from our previous claim and [27, Theorem 23.2] with \(f = \tilde{h}\) that

\[
\inf_{\|d\| \leq 1} (f + h)'(x; d) = \inf_{\|d\| \leq 1} \tilde{h}'(x; d) = \inf_{\|d\| \leq 1} \sigma_{\partial h(x)}(d) = \inf_{\|d\| \leq 1} \left[ \langle \nabla f(x), d \rangle + \sigma_{\partial h(x)}(d) \right],
\]

which gives the first identity in (C.1). The second identity in (C.1) follows from Lemma 15 with \(C = \partial \tilde{h}(x)\) and the last identity in (C.2).

To complete the proof, we now justify the claim made in the previous paragraph. Define \(\mathcal{B} := \{d \in \mathcal{X} : \|d\| \leq 1\}\) and \(\psi(\cdot) := \tilde{h}'(x; \cdot)\). In view of Lemma 15 with \(C = \mathcal{B}\), it suffices to show that \(\mathcal{B} \cap \text{ri}(\text{dom } \psi) \neq \emptyset\). To show this, note that the convexity of \(\tilde{h}\) and the discussion following [27, Theorem 23.1] imply that \(\text{dom } \psi = \bigcup_{t > 0} (\text{dom } h - x)/t\), which is a nonempty convex cone. Hence, it follows from [27, Theorem 6.2] and the discussion in the second paragraph following [27, Corollary 6.8.1] that \(\text{ri}(\text{dom } \psi)\) is also a nonempty convex cone. This conclusion clearly implies that \(\mathcal{B} \cap \text{ri}(\text{dom } \psi) \neq \emptyset\).

It is worth mentioning that the result above is a generalization of the one given in [4, Lemma 5.1], which only considers the case where \((f + h)\) is real-valued and locally Lipschitz.

The next technical lemma, which can be found in [29, Corollary 3.3], presents a well-known min-max identity.

**Lemma 17.** Let a convex set \(D \subseteq \mathcal{X}\) and compact convex set \(Y \subseteq \mathcal{Y}\) be given. Moreover, let \(\psi : D \times Y \mapsto \mathbb{R}\) be a function in which \(\psi(\cdot, y)\) is convex lower semicontinuous for every \(y \in Y\) and \(\psi(d, \cdot)\) is concave upper semicontinuous for every \(d \in D\). Then,

\[
\inf_{d \in D} \sup_{y \in Y} \psi(d, y) = \sup_{y \in Y} \inf_{d \in D} \psi(d, y).
\]

The next result establishes an identity similar to Lemma 16 but for the case where \(f\) is a max function.

**Proposition 18.** Assume the quadruple \((\Psi, h, X, Y)\) satisfies assumptions (A0)–(A3) of Subsection 2.1 with \(\Phi = \Psi\). Moreover, suppose that \(\Psi(\cdot, y)\) is convex for every \(y \in Y\), and let \(q\) and \(Y(\cdot)\) be as in Lemma 13. Then, for every \(\bar{x} \in X\), it holds that

\[
\inf_{\|d\| \leq 1} (q + h)'(\bar{x}; d) = - \inf_{u \in Q(\bar{x})} \|u\|
\]

where

\[
Q(\bar{x}) := \partial h(\bar{x}) + \bigcup_{y \in Y(\bar{x})} \{\nabla \Psi(\bar{x}, y)\}.
\]
Moreover, if \( \partial h(\bar{x}) \) is nonempty, then the infimum on the right-hand side of (C.3) is achieved.

**Proof.** Let \( \bar{x} \in X \) and define
\[
(C.5) \quad \psi(d, y) := (\Psi_y + h)'(\bar{x}; d), \quad \forall (d, x, y) \in \mathcal{X} \times \Omega \times Y.
\]
We claim that \( \psi \) in (C.5) satisfies the assumptions on \( \psi \) in Lemma 17 with \( Y = Y(\bar{x}) \) and \( D \) given by
\[
D := \{ d \in Z : \|d\| \leq 1, d \in F_X(\bar{x}) \},
\]
where \( F_X(\bar{x}) := \{ t(x - \bar{x}) : x \in X, t \geq 0 \} \) is the set of feasible directions at \( \bar{x} \).

Before showing this claim, we use it to show that (C.3) holds. First observe that (A1) and Lemma 13 imply that
\[
\partial h(\bar{x}) = \sup_{y \in Y(\bar{x})} (\Psi_y)'(\bar{x}; d) \quad \forall d \in D.
\]
Before showing this claim, we use it to show that (C.3) holds. First observe that (A1) and Lemma 13 imply that
\[
\partial h(\bar{x}) = \sup_{y \in Y(\bar{x})} \min_{d \subseteq X} (\|d\|_1) = \sup_{y \in Y(\bar{x})} \min_{d \subseteq X} (\|d\|_1)
\]
and the previous observation, we have that
\[
(C.6) \quad \inf_{\|d\| \leq 1} (q + h)'(\bar{x}; d) = \inf_{d \in D} (q + h)'(\bar{x}; d) = \inf_{d \in D} \sup_{y \in Y(\bar{x})} (\Psi_y + h)'(\bar{x}; d)
\]
\[
= \inf_{d \in D} \sup_{y \in Y(\bar{x})} \psi(d, y) = \sup_{y \in Y(\bar{x})} \inf_{d \in D} (\psi(d, y) = \sup_{y \in Y(\bar{x})} \inf_{\|d\| \leq 1} (\Psi_y + h)'(\bar{x}; d)
\]
\[
(C.7) \quad \psi(d, y) = (\Psi_y)'(\bar{x}; d) = \frac{\Psi_y(\bar{x} + td) - q(\bar{x})}{t}, \quad \forall d \in X.
\]
Since assumption (A2) implies that \( \Psi(\bar{x}, \cdot) \) is upper semicontinuous and concave on \( Y \), it follows from (C.7), [27, Theorem 5.5], and [27, Theorem 9.4] that \( \psi(d, \cdot) \) is upper semicontinuous and concave on \( Y \) for every \( d \in \mathcal{X} \). On the other hand, since \( \Psi(\cdot, y) \) is assumed to be lower semicontinuous and convex on \( X \) for every \( y \in Y \), it follows from (C.7), the fact that \( \bar{x} \in \text{int} \Omega \), and [27, Theorem 23.4], that \( \psi(\cdot, y) \) is lower semicontinuous and convex on \( \mathcal{X} \), and hence \( D \subseteq \mathcal{X} \), for every \( y \in Y(\bar{x}) \).

To complete the proof, we now justify the above claim on \( \psi \). First, for any given \( y \in Y(\bar{x}) \), it follows from [27, Theorem 23.1] with \( f(\cdot) = \Psi_y(\cdot) \) and the definitions of \( q \) and \( Y(\bar{x}) \) that
\[
(C.8) \quad \psi(d, y) = (\Psi_y)'(\bar{x}; d) = \frac{\Psi_y(\bar{x} + td) - q(\bar{x})}{t}, \quad \forall d \in X.
\]
Since assumption (A2) implies that \( \Psi(\bar{x}, \cdot) \) is upper semicontinuous and concave on \( Y \), it follows from (C.7), [27, Theorem 5.5], and [27, Theorem 9.4] that \( \psi(d, \cdot) \) is upper semicontinuous and concave on \( Y \) for every \( d \in \mathcal{X} \). On the other hand, since \( \Psi(\cdot, y) \) is assumed to be lower semicontinuous and convex on \( X \) for every \( y \in Y \), it follows from (C.7), the fact that \( \bar{x} \in \text{int} \Omega \), and [27, Theorem 23.4], that \( \psi(\cdot, y) \) is lower semicontinuous and convex on \( \mathcal{X} \), and hence \( D \subseteq \mathcal{X} \), for every \( y \in Y(\bar{x}) \).

The last technical result is a specialization of the one given in [13, Theorem 4.2.1].

**PROPOSITION 19.** Let a proper closed function \( \phi : \mathcal{X} \mapsto (-\infty, \infty] \) and assume that \( \phi + \| \cdot \|^2 / 2\Lambda \) is \( \mu \)-strongly convex for some scalars \( \mu, \lambda > 0 \). If a quadruple \( (x^-, x, u, \varepsilon) \in \mathcal{X} \times \text{dom} \phi \times \mathcal{X} \times \mathbb{R}_+ \) together with \( \lambda \) satisfy
\[
(C.8) \quad u \in \partial_{\varepsilon} \left( \phi + \frac{1}{2\lambda} \| \cdot - x^- \|^2 \right)(x),
\]
then the point \( \hat{x} \in \text{dom} \phi \) given by
\[
(C.9) \quad \hat{x} := \arg \min_{x'} \left\{ \phi_\lambda(x') := \phi(x') + \frac{1}{2\lambda} \| x' - x^- \|^2 - \langle u, x' \rangle \right\}
\]
satisfies

\[
\inf_{\|d\| \leq 1} \phi'(\hat{x}; d) \geq -\frac{1}{\lambda} \|x^- - x + \lambda u\| - \sqrt{\frac{2\varepsilon}{\lambda^2 \mu}}, \quad \|\hat{x} - x\| \leq \sqrt{\frac{2\varepsilon}{\mu}}.
\]

Proof. We first observe that (C.8) implies that

\[
\phi_\lambda(x') \geq \phi_\lambda(x) - \varepsilon \quad \forall x' \in \mathcal{X}.
\]

Remark that (C.11) at \(x' = \hat{x}\), the optimality of \(\hat{x}\), and the \(\mu\)-strong convexity of \(\phi_\lambda\) imply that

\[
\frac{\mu}{2} \|\hat{x} - x\|^2 \leq \phi_\lambda(x) - \phi_\lambda(\hat{x}) \leq \varepsilon
\]

from which we conclude that \(\|\hat{x} - x\| \leq \sqrt{2\varepsilon/\mu}\), i.e., the second inequality in (C.10).

On the other hand, using the definition of \(\phi_\lambda\), the triangle inequality, and the previous observation, and the fact that \(\inf_{\Psi} \Psi(\bar{y})\) satisfies the assumptions on \(\phi_\lambda\) and let

\[
\inf_{\|d\| \leq 1} \phi'(\hat{x}; d) = \inf_{\|d\| \leq 1} \phi'(\hat{x}; d) - \frac{1}{\lambda} \langle d, \lambda u + x^- - \hat{x}\rangle
\]

\[
\leq \inf_{\|d\| \leq 1} \phi'(\hat{x}; d) + \frac{\|x^- - x + \lambda u\|}{\lambda} + \|x - \hat{x}\|
\]

\[
\leq \inf_{\|d\| \leq 1} \phi'(\hat{x}; d) + \frac{\|x^- - x + \lambda u\|}{\lambda} + \sqrt{\frac{2\varepsilon}{\lambda^2 \mu}}.
\]

which clearly implies the first inequality in (C.10).

Appendix D. This appendix presents the proofs of Propositions 1, 2, and 3.

The first technical result shows that an approximate primal-dual stationary point is equivalent to an approximate directional stationary point of a perturbed version of problem (1.1).

Lemma 20. Suppose the quadruple \((\Phi, h, X, Y)\) satisfies assumptions (A0)--(A3) of Subsection 2.1 and let \((\hat{x}, \bar{u}, \bar{v}) \in X \times \mathcal{X} \times \mathcal{Y}\) be given. Then, there exists \(\bar{y} \in Y\) such that the quadruple \((\bar{u}, \bar{v}, \hat{x}, \bar{y})\) satisfies the inclusion in (1.4) if and only if

\[
\inf_{\|d\| \leq 1} (p_{\bar{u}, \bar{v}} + h)'(\hat{x}; d) \geq 0,
\]

where the function \(p_{\bar{u}, \bar{v}}\) is given by

\[
p_{\bar{u}, \bar{v}}(x) := \max_{y \in Y} \left[ \Phi(x, y) + \langle \bar{v}, y \rangle - \langle \bar{u}, x \rangle \right] \quad \forall x \in \Omega.
\]

Proof. Let \((\bar{x}, \bar{u}, \bar{v}) \in X \times \mathcal{X} \times \mathcal{Y}\) be given, define

\[
\Psi(x, y) := \Phi(x, y) + \langle \bar{v}, y \rangle - \langle \bar{u}, x \rangle + m\|x - \bar{x}\|^2 \quad \forall (x, y) \in \Omega \times Y,
\]

and let \(q\) and \(Y(\cdot)\) be as in Lemma 13. It is easy to see that \(q = p_{\bar{u}, \bar{v}}\), the function \(\Psi\) satisfies the assumptions on \(\Psi\) in Proposition 18, and \(\bar{x}\) satisfies (D.1) if and only if \(\inf_{\|d\| \leq 1} (q + h)'(\hat{x}; d) \geq 0\). The desired conclusion follows from Proposition 18, the previous observation, and the fact that \(\bar{y} \in Y(\bar{x})\) if and only if \(\bar{v} \in \partial\mathcal{J}(\Phi(\bar{x}, \cdot))(\bar{y})\).

We are now ready to give the proof of Proposition 1.
Proof of Proposition 1. Suppose \( (\bar{u}, \bar{v}, \bar{x}, \bar{y}) \) is a \((\rho_x, \rho_y)\)-primal-dual stationary point of \((1.1)\). Moreover, let \( \Psi, q, \) and \( D_y \) be as in \((D.3), (B.1)\) and \((2.8)\), respectively, and define
\[
\hat{q}(x) := q(x) + h(x) \quad \forall x \in X.
\]
Using Lemma 20, we first observe that
\[
\inf_{\|d\| \leq 1} \hat{q}(\bar{x}; d) \geq 0.
\]
Since \( \hat{q} \) is convex from assumption \((A3)\), it follows from the previous bound and Lemma 16 with \((f, h) = (0, \hat{q})\), that \( \min_{u \in \partial \hat{q}(\bar{x})} \|u\| \leq 0 \), and hence, \( 0 \in \partial \hat{q}(\bar{x}) \). Moreover, using the Cauchy-Schwarz inequality, the second inequality in \((1.4)\), the previous inclusion, and the definition of \( q \) and \( \Psi \), it follows that for every \( x \in \mathcal{X} \),
\[
\hat{p}(x) + D_y \rho_y - \langle \bar{u}, x \rangle + m \|x - \bar{x}\|^2 \geq \hat{q}(x) \geq \hat{p}(x) - D_y \rho_y - \langle \bar{u}, x \rangle,
\]
and hence that \( \bar{u} \in \partial \varepsilon(\hat{p} + m \| \cdot - \bar{x} \|^2)(\bar{x}) \) where \( \varepsilon = 2D_y \rho_y \). Using now the first inequality in \((1.4)\), Proposition 19 with \((\phi, x, x^{-}, u) = (\hat{p}, \bar{x}, \bar{u}) \) and also \((\varepsilon, \lambda, \mu) = (D_y \rho_y, 1/(2m), m)\), we conclude that there exists \( \hat{x} \) such that \( \| \hat{x} - \bar{x} \| \leq \sqrt{2D_y \rho_y}/m \) and
\[
\inf_{\|d\| \leq 1} \hat{p}'(\hat{x}; d) \geq -\|\bar{u}\| - 2\sqrt{2mD_y \rho_y} \geq -\rho_x - 2\sqrt{2mD_y \rho_y}.
\]
We next give the proof of Proposition 2.

Proof of Proposition 2. (a) We first claim that \( \hat{P}_\lambda \) is \( \alpha \)-strongly convex, where \( \alpha = 1/\lambda - m \). To see this, note that \( \Phi(-, y) + m \| \cdot \|^2/2 \) is convex for every \( y \in \mathcal{Y} \) from assumption \((A3)\). The claim now follows from assumption \((A2)\), the fact that the supremum of a collection of convex functions is also convex, and the definition of \( \hat{p} \) in \((1.1)\).

Suppose the pair \((x, \delta)\) satisfies \((1.5)\) and \((2.10)\). If \( \hat{x} = x_\lambda \) in \((1.5)\), then clearly the second inequality in \((1.5)\), the fact that \( \lambda < 1/m \), and \((2.10)\) imply the inequality in \((2.9)\), and hence, that \( x \) is a \((\lambda, \varepsilon)\)-prox stationary point. Suppose now that \( \hat{x} \neq x_\lambda \).

Using the convexity of \( \hat{P}_\lambda \), we first have that
\[
\hat{P}_\lambda'(\hat{x}; d) = \inf_{t > 0} \left[ \hat{P}_\lambda(\hat{x} + td) - \hat{P}_\lambda(\hat{x}) \right]/t
\]
for every \( d \in \mathcal{X} \). Hence, using both inequalities in \((1.5)\) and the previous identity, it holds that
\[
\frac{\hat{P}_\lambda(x_\lambda) - \hat{P}_\lambda(\hat{x})}{\|x_\lambda - \hat{x}\|} \geq \hat{P}_\lambda'(\hat{x}; \frac{x_\lambda - \hat{x}}{\|x_\lambda - \hat{x}\|}) = \hat{p}'(\hat{x}; \frac{x_\lambda - \hat{x}}{\|x_\lambda - \hat{x}\|}) + \frac{1}{\lambda} \left( \frac{x_\lambda - \hat{x}}{\|x_\lambda - \hat{x}\|}, \hat{x} - x \right)
\]
\[
\geq -\delta - \frac{1}{\lambda} \|\hat{x} - x\| \geq -\delta \left( \frac{1 + \lambda}{\lambda} \right).
\]
Using the optimality of \( x_\lambda \), the \( \alpha \)-strong convexity of \( \hat{P}_\lambda \) (see our claim on \( \hat{p} \) in the first paragraph), and the above bound, we conclude that
\[
\frac{1}{2\alpha} \|\hat{x} - x_\lambda\|^2 \leq \hat{P}_\lambda(\hat{x}) - \hat{P}_\lambda(x_\lambda) \leq \delta \left( \frac{1 + \lambda}{\lambda} \right) \|\hat{x} - x_\lambda\|.
\]
Thus, \( \|\hat{x} - x_\lambda\| \leq 2\alpha \delta (1 + \lambda)/\lambda \). Using the previous bound, the second inequality in \((1.5)\), and \((2.10)\) yields
\[
\|x - x_\lambda\| \leq \|x - \hat{x}\| + \|\hat{x} - x_\lambda\| \leq \left( 1 + 2\alpha \left[ \frac{1 + \lambda}{\lambda} \right] \right) \delta \leq \lambda \varepsilon,
\]
which implies \((2.9)\), and hence, that \( x \) is a \((\lambda, \varepsilon)\)-prox stationary point.

This manuscript is for review purposes only.
Suppose that the point $x$ is a $(\lambda, \varepsilon)$-prox stationary point with $\varepsilon \leq \delta \cdot \min\{1, 1/\lambda\}$. Then the optimality of $x_\lambda$, the fact that $\hat{P}_\lambda$ is convex (see the beginning of part (a)), the inequality in (2.9), and the Cauchy-Schwarz inequality imply that

$$0 \leq \inf_{\|d\| \leq 1} \left[ \tilde{p}'(x_\lambda; d) + \frac{1}{\lambda} \langle d, x_\lambda - x \rangle \right] \leq \inf_{\|d\| \leq 1} \tilde{p}'(x_\lambda; d) + \varepsilon \leq \inf_{\|d\| \leq 1} \tilde{p}'(x_\lambda; d) + \delta,$$

which, together with the fact that $\lambda \varepsilon \leq \delta$, imply that $x$ satisfies (1.5) with $\hat{x} = x_\lambda$. \hfill \square

Finally, we give the proof of Proposition 3.

**Proof of Proposition 3.** This follows by using Lemma 16 with $(f, h) = (\Phi(\cdot, \bar{y}), h)$ and $(f, h) = (0, -\Phi(\bar{x}, \cdot))$. \hfill \square

**Appendix E.** This appendix presents the description of $y_\xi$ and justification for the constants $m$, $L_x$, and $L_y$ for each of the optimization problems in Section 5.

**Maximum of a finite number of nonconvex functions.** Since $Y = \Delta^k$, it is easy to verify that

$$y_\xi(x) = \arg\max_{y'} \{\|y' - \xi g_t(x)\| : y' \in \Delta^k\} \quad \forall x \in \mathbb{R}^n.$$

For the validity of the constants $m$, $L_x$, and $L_y$, we first define, for every $1 \leq i \leq k$, the quantities

$$P_i = \alpha_i C^T_i d_i, \quad Q_i = \alpha_i C^T_i C_i x - \beta_i B^T_i D_i B_i x \quad \forall x \in \mathbb{R}^n,$$

and observe that $\nabla_x \Phi(x, y) = \sum_{i=1}^k (Q_i + P_i) y_i$. Now, using the fact that $y \in \Delta^k$, (5.5), and defining $N_i := \alpha_i C^T_i C_i - \beta_i B^T_i D_i B_i$, we then have that

$$\lambda_{\max}(\nabla_{xx}^2 \Phi) \leq \sum_{i=1}^k y_i \lambda_{\max}(N_i) = \sum_{i=1}^k y_i \lambda_{\max}(\nabla^2 g_i) \leq M = L_x,$$

$$\lambda_{\min}(\nabla_{xx}^2 \Phi) \geq \sum_{i=1}^k y_i \lambda_{\min}(N_i) = \sum_{i=1}^k y_i \lambda_{\min}(\nabla^2 g_i) \geq -m \geq -L_x,$$

and hence we conclude that the choice of $m$ and $L_x$ in (5.6) is valid. On the other hand, using the fact that $\|x\| \leq 1$ for every $x \in \Delta^2$ and (5.5), we then have that for every $y, y' \in Y$,

$$\|\nabla_x \Phi(x, y) - \nabla_x \Phi(x, y')\| = \left\| \sum_{i=1}^k (Q_i + P_i)(y_i - y'_i) \right\|$$

$$\leq \sqrt{\sum_{i=1}^k M^2 \|x\|^2 + \|P\|} \|y - y'\| \leq L_y \|y - y'\|,$$

where $P$ is a an $n$–by–$k$ matrix whose $i^{th}$ column is $\alpha_i C^T_i d_i$, and hence we conclude that the choice of $L_y$ in (5.6) is valid.

---

This manuscript is for review purposes only.
Truncated robust regression. Since $Y = \Delta^n$, it is easy to verify that
\[
y^*_\xi(x) = \arg\max_{y'} \{||y' - \xi g_i(x)|| : y' \in \Delta^n\} \quad \forall x \in \mathbb{R}^k.
\]
For the validity of the constants $m$, $L_x$, and $L_y$, we first define for every $1 \leq i \leq k$ the function
\[
\tau_j(x) := \left[e^{-b_j(x,z)}\right] \left[1 + e^{-b_j(x,z)}\right]^{-1} [\alpha + \ell_j(x)]^{-1} \quad \forall x \in \mathbb{R}^k;
\]
and observe that $\nabla_x \Phi(x, y) = -\alpha \sum_{j=1}^n [y_j b_j \tau_j(x)] a_j$ and also that
\[
(E.1) \quad \sup_{x \in \mathbb{R}^k} |\tau_j(x)| \leq \alpha^{-1},
\]
for every $1 \leq j \leq n$. Now, using the fact that $y \in \Delta^n$, the bound (E.1), and the Mean Value Theorem applied to $\tau_j$, we have that for every $x, x' \in \mathbb{R}^k$,
\[
\|
\nabla_x \Phi(x, y) - \nabla_x \Phi(x', y)\| \leq \alpha \sum_{j=1}^n y_j |a_j| |\tau_j(x) - \tau_j(x')| \|
\leq \alpha \max_j \left(||a_j| |\tau_j(x) - \tau_j(x')| ||\right) = \alpha \max_{1 \leq j \leq n} \left(||a_j| |\tau_j(x) - \tau_j(x')| ||\right)
\leq \alpha \max_{1 \leq j \leq n} \left(||a_j| |\sup_{x \in \mathbb{R}^k} |\nabla \tau_j(x)|| ||x - x'|| ||\right)
\leq \alpha \max_{1 \leq j \leq n} \left(||a_j| |2 |\sup_{x \in \mathbb{R}^k} \left|\frac{\tau_j(z)}{\alpha + \ell_j(z)}\right|| ||x - x'|| ||\right)
\leq \frac{1}{\alpha} \max_{1 \leq j \leq n} ||a_j| |2 ||x - x'|| || = L_x ||x - x'||,
\]
and hence we conclude that the choice of $m = L_x$ in (5.7) is valid. On the other hand, using the bound (E.1), we have that for every $y, y' \in \mathbb{R}^n$,
\[
||\nabla_x \Phi(x, y) - \nabla_x \Phi(x', y')|| = \alpha \sum_{j=1}^n y_j |a_j| |\tau_j(x) - \tau_j(x')| || \leq L_y ||y - y'||,
\]
and hence we conclude that the choice of $L_y$ in (5.7) is valid.

Power control in the presence of a jammer. For every $1 \leq k \leq K$ and $1 \leq n \leq N$, we first define the quantities
\[
S_{k,n}^{-}(X, y) := \sigma^2 + B_{k,n} y_n + \sum_{j=1, j \neq k}^K A_{j,k,n} X_{j,n}, \quad S_{k,n}(X, y) := A_{k,k,n} X_{k,n} + S_{k,n}^{-},
\]
as well as
\[
T_{j,n}(X, y) := \left[S_{j,n}^{-}(X, y) + S_{j,n}(X, y)\right] / \left[S_{j,n}(X, y) S_{j,n}^{-}(X, y)\right]^2,
\]
for every $(X, y) \in \mathbb{R}^{K \times N} \times \mathbb{R}^N$. Observe now that
\[
(E.2) \quad \frac{\partial \Phi}{\partial y_n}(X, y) = \frac{B_{k,n}}{S_{k,n}(X, y) S_{k,n}^{-}(X, y)} \quad \forall n \in \{1, \ldots, N\}.
\]
The form in (E.2) implies that \( \nabla_y \Phi(X, y) \) is a separable function in \( y \) where each component is a monotonically decreasing function in its argument. Hence, since \( Y = Q_{N/2}^{N \times 1} \), the computation of \( y_k \) reduces to an \( N \)-dimensional bisection search on the functions

\[
F_n(y; \xi) = \left[ \sum_{k=1}^{K} \frac{B_{k,n}}{S_{k,n}(X, y)S'_{k,n}(X, y)} \right] - \frac{y_n}{\xi} \quad \forall n \in \{1, ..., N\}.
\]

For the validity of the constants \( m, L_x, \) and \( L_y \), we first observe that, for every \( 1 \leq k \leq K \) and \( 1 \leq n \leq N \), and also \( (X, y) \in \mathbb{R}^{K \times N} \times \mathbb{R}^N \), we have

\[
\frac{\partial \Phi}{\partial X_{k,n}}(X, y) = -\frac{A_{k,n}}{S_{k,n}(X, y)} + \sum_{j=1, j \neq k}^{K} \frac{A_{k,j,n}}{S_{j,n}(X, y)S'_{j,n}(X, y)} \quad \forall (X, y) \in \mathbb{R}^{K \times N} \times \mathbb{R}^N.
\]

Using the Mean Value Theorem with respect to \( X_{k,n} \) on \( \partial \Phi / \partial X_{k,n} \), we have that for every \( X, X' \in \mathbb{R}^{K \times N} \),

\[
\left| \frac{\partial}{\partial X_{k,n}} f(X, y) - \frac{\partial}{\partial X_{k,n}} f(X', y) \right| \leq \sup_{(X, y) \in \mathbb{R}^{K \times N} \times \mathbb{R}^N} \left| \frac{\partial^2}{\partial X_{k,n}^2} f(X, y) \right| |X_{k,n} - X'_{k,n}|
\]

\[
= \sup_{(X, y) \in \mathbb{R}^{K \times N} \times \mathbb{R}^N} \left| \frac{A_{k,n}^2}{S_{k,n}(X, y)S'_{k,n}(X, y)} - \sum_{j=1, j \neq k}^{K} \frac{A_{k,j,n}^2}{S_{j,n}(X, y)S'_{j,n}(X, y)} \right| |X_{k,n} - X'_{k,n}|
\]

\[
\leq \frac{2 \sum_{j=1}^{K} A_{k,j,n}^2}{\min\{\sigma^4, \sigma^6\}} |X_{k,n} - X'_{k,n}| \leq L_x |X_{k,n} - X'_{k,n}|,
\]

and hence we conclude that the choice of \( L_x \) in (5.8) is valid. On the other hand, using the Mean Value Theorem with respect to \( y_n \) on \( \partial \Phi / \partial X_{k,n} \), we have that for every \( y, y' \in \mathbb{R}^{K \times N} \),

\[
\left| \frac{\partial}{\partial X_{k,n}} f(X, y) - \frac{\partial}{\partial X_{k,n}} f(X', y) \right| \leq \sup_{(X, y) \in \mathbb{R}^{K \times N} \times \mathbb{R}^N} \left| \frac{\partial^2}{\partial y_n X_{k,n}} f(X, y) \right| |y_n - y'_{n}|
\]

\[
= \sup_{(X, y) \in \mathbb{R}^{K \times N} \times \mathbb{R}^N} \left| \frac{B_{k,n} A_{k,n}}{S_{k,n}(X, y)} - \sum_{j=1, j \neq k}^{K} \frac{B_{k,n} A_{k,j,n}}{S_{j,n}(X, y)} \right| |y_n - y'_{n}|
\]

\[
\leq \frac{2 \sum_{j=1}^{K} B_{k,n} A_{k,j,n}}{\min\{\sigma^4, \sigma^6\}} |y_n - y'_{n}| \leq L_y |y_n - y'_{n}|,
\]

and hence we conclude that the choice of \( L_y \) in (5.8) is valid.

REFERENCES


36

This manuscript is for review purposes only.