SIAM J. OPTIM.
Vol. 31, No. 4, pp. 2558–2585

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 nonsmooth) 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 approximation, 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

DOI. 10.1137/20M1313222

1. Introduction.

The first goal of this paper is to present and study the complexity of an accelerated inexact proximal point smoothing (AIPP-S) scheme for finding approximate stationary points of the (potentially nonsmooth) min-max composite nonconvex optimization (CNO) problem

$$\min_{x\in X} \{\hat{p}(x) := p(x) + h(x)\}, \tag{1.1}$$

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

$$p(x) := \max_{y\in Y} \Phi(x,y) \quad \forall x \in X, \tag{1.2}$$

for some nonempty compact convex set Y and function Φ which, for some scalar $m > 0$ and open set $\Omega \supseteq X$, is such that (i) Φ is continuous on $\Omega \times Y$; (ii) the function $-\Phi(x,\cdot) : Y \mapsto \mathbb{R}$ 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 and 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) is a smooth CNO problem for which many algorithms have been developed for in the literature. In particular, accelerated inexact proximal point (AIPP) methods, i.e., methods which use an accelerated composite gradient variant to approximately solve a generated sequence of prox subproblems, have been developed for it (see, for example,

*Received by the editors January 16, 2020; accepted for publication (in revised form) July 9, 2021; published electronically October 25, 2021.
https://doi.org/10.1137/20M1313222
Funding: The authors were partially supported by ONR grant N00014-18-1-2077 and NSERC grant PGSD3-516700-2018.
†School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205 USA (wkong37@gatech.edu, monteiro@isye.gatech.edu).

[4,15]). 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 smooths the max term in (1.1) and solves the resulting CNO problem by an AIPP method.

Throughout our presentation, it is assumed that oracles for evaluating the quantities $\Phi(x,y)$, $\nabla_x \Phi(x,y)$, and $h(x)$ and for obtaining exact solutions of the problems

$$\min_{x\in X}\left\{\lambda h(x)+\frac{1}{2}\|x-x_0\|^2\right\},\quad \max_{y\in Y}\left\{\lambda\Phi(x_0,y)-\frac{1}{2}\|y-y_0\|^2\right\} \tag{1.3}$$

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 $\mathcal{O}(1)$ where each of them appears at least once. We refer to the computation of the solution of the first problem above as an 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 $(\rho_x,\rho_y)\in\mathbb{R}^2_{++}$, it is shown that an instance of this scheme obtains $(\bar u,\bar v,\bar x,\bar y)$ such that

$$\begin{pmatrix}\bar u\\ \bar v\end{pmatrix}\in\begin{pmatrix}\nabla_x\Phi(\bar x,\bar y)\\ 0\end{pmatrix}+\begin{pmatrix}\partial h(\bar x)\\ \partial\left[-\Phi(\bar x,\cdot)\right](\bar y)\end{pmatrix},\quad \|\bar u\|\le\rho_x,\quad \|\bar v\|\le\rho_y \tag{1.4}$$

in $\mathcal{O}(\rho_x^{-2}\rho_y^{-1/2})$ oracle calls, where $\partial\phi(z)$ is the subdifferential of a convex function ϕ 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 $\delta>0$, this instance computes $x\in X$ such that

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

in $\mathcal{O}(\delta^{-3})$ 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 for finding approximate stationary points of a linearly constrained version of (1.1), namely,

$$\min_{x\in X}\left\{p(x)+h(x):\mathcal{A}x=b\right\}, \tag{1.6}$$

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

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

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 [15]. For a given tolerance triple $(\rho_x,\rho_y,\eta)\in\mathbb{R}^3_{++}$, it is shown that the method computes a quintuple $(\bar u,\bar v,\bar x,\bar y,\bar r)$ satisfying

$$\begin{gathered}\begin{pmatrix}\bar u\\ \bar v\end{pmatrix}\in\begin{pmatrix}\nabla_x\Phi(\bar x,\bar y)+\mathcal{A}^*\bar r\\ 0\end{pmatrix}+\begin{pmatrix}\partial h(\bar x)\\ \partial\left[-\Phi(\bar x,\cdot)\right](\bar y)\end{pmatrix},\\ \|\bar u\|\le\rho_x,\quad \|\bar v\|\le\rho_y,\quad \|\mathcal{A}\bar x-b\|\le\eta\end{gathered} \tag{1.8}$$

in $\mathcal{O}(\rho_x^{-2}\rho_y^{-1/2}+\rho_x^{-2}\eta^{-1})$ oracle calls.

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 assumed neither 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, 11, 13, 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 $x \in X$.

Letting δ_C denote the indicator function of a closed convex set $C \subseteq \mathcal{X}$ (see subsection 1.1), $\overline{\text{Conv}}(\mathcal{X})$ denote the set of proper lower semicontinuous convex functions on $\mathcal{X}$, and $\rho := \min\{\rho_x, \rho_y\}$, Tables 1.1 and 1.2 compare the assumptions and iteration complexities obtained in this work with corresponding ones derived in the earlier papers [24, 26] and the subsequent works [17, 25, 30]. Note 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 actually equivalent to the ones in this paper up to multiplicative constants that are independent of the tolerances, i.e., ρ_x, ρ_y, δ.

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 [15] and its iteration complexity. The second one presents the AIPP-S scheme and 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 [15] and its iteration complexity. The second one presents the QP-AIPP-S scheme and its iteration complexity for finding points satisfying (1.8). Section 5 presents some computational results. Section 6 gives some

Table 1.1
Comparison of iteration complexities based on (1.4) *with* $\rho := \min\{\rho_x, \rho_y\}$.

Algorithm	Oracle complexity	Use cases			
		$D_h = \infty$	$h \equiv 0$	$h \equiv \delta_C$	$h \in \overline{\text{Conv}}(\mathcal{X})$
PGSF [24]	$\mathcal{O}(\rho^{-3})$	✗	✓	✓	✗
Minimax-PPA [17]	$\mathcal{O}(\rho^{-2.5}\log^2(\rho^{-1}))$	✗	✓	✓	✗
FNE search [25]	$\mathcal{O}(\rho_x^{-2}\rho_y^{-1/2}\log(\rho^{-1}))$	✓	✓	✓	✗
AIPP-S	$\mathcal{O}(\rho_x^{-2}\rho_y^{-1/2})$	✓	✓	✓	✓

Table 1.2
Comparison of iteration complexities based on (1.5).

Algorithm	Oracle complexity	Use cases			
		$D_h = \infty$	$h \equiv 0$	$h \equiv \delta_C$	$h \in \overline{\text{Conv}}(\mathcal{X})$
PG-SVRG [26]	$\mathcal{O}(\delta^{-6}\log\delta^{-1})$	✗	✓	✓	✓
Minimax-PPA [17]	$\mathcal{O}(\delta^{-3}\log^2(\delta^{-1}))$	✗	✓	✓	✗
Prox-DIAG [30]	$\mathcal{O}(\delta^{-3}\log^2(\delta^{-1}))$	✓	✓	✗	✗
AIPP-S	$\mathcal{O}(\delta^{-3})$	✓	✓	✓	✓

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. The set of real numbers is denoted by $\mathbb{R}$. The set of nonnegative real numbers and the set of positive real numbers are denoted by $\mathbb{R}_+$ and $\mathbb{R}_{++}$, respectively. The set of natural numbers is denoted by $\mathbb{N}$. For $t > 0$, define $\log_1^+(t) := \max\{1, \log(t)\}$. Let $\mathbb{R}^n$ denote a real-valued n-dimensional Euclidean space with standard norm $\|\cdot\|$. Given a linear operator $A : \mathbb{R}^n \mapsto \mathbb{R}^p$, the operator norm of A is denoted by $\|A\| := \sup\{\|Az\|/\|z\| : z \in \mathbb{R}^n, z \neq 0\}$.

The following are for a Euclidean space $\mathcal{Z}$ with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$. The effective domain of a function $\psi : \mathcal{Z} \mapsto (-\infty, \infty]$ is denoted as $\operatorname{dom}\psi := \{z \in \mathcal{Z} : \psi(z) < \infty\}$, and ψ is said to be proper if $\operatorname{dom}\psi \neq \emptyset$. The set of proper, lower semicontinuous, convex functions is denoted by $\overline{\operatorname{Conv}}(\mathcal{Z})$. Moreover, for convex $Z \subseteq \mathcal{Z}$, we denote $\overline{\operatorname{Conv}}(Z)$ to be set of functions in $\overline{\operatorname{Conv}}(\mathcal{Z})$ whose effective domain is equal to Z. For $\varepsilon \geq 0$, the *ε-subdifferential* of $\psi \in \overline{\operatorname{Conv}}(\mathcal{Z})$ at $z \in \operatorname{dom}\psi$ is denoted by

$$\partial_\varepsilon \psi(z) := \{w \in \mathbb{R}^n : \psi(z') \geq \psi(z) + \langle w, z' - z\rangle - \varepsilon \ \forall z' \in \mathcal{Z}\}, \tag{1.9}$$

and we denote $\partial\psi \equiv \partial_0\psi$. The *directional derivative* of ψ at $z \in \mathcal{Z}$ in the direction $d \in \mathcal{Z}$ is denoted by

$$\psi'(z; d) := \lim_{t \to 0} \frac{\psi(z + td) - \psi(z)}{t}. \tag{1.10}$$

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

For a given $Z \subseteq \mathcal{Z}$, the indicator function of Z, denoted by δ_Z, has value 0 if $z \in Z$ and value ∞ if $z \notin Z$. The closure, interior, and relative interior of Z are denoted by $\operatorname{cl} Z$, $\operatorname{int} Z$, and $\operatorname{ri} Z$, 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 and a closed convex set $C \subseteq \mathbb{R}^n$, the problem of interest is the minimization of $\max_{1 \leq i \leq k} f_i$ over the set C, i.e.,

$$\min_{x \in C} \max_{1 \leq i \leq k} f_i(x),$$

which is clearly an instance of (1.1) where $Y = \{y \in \mathbb{R}^k_+ : \sum_{i=1}^k y_i = 1\}$, $\Phi(x, y) = \sum_{i=1}^k y_i f_i(x)$, and $h(x) = \delta_C(x)$.

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 $\{\ell_\theta(\cdot|\sigma)\}_{\theta \in \Theta}$ be a family of nonconvex loss functions in which (i) $\ell_\theta(x|\sigma)$ is concave in θ 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 Θ, 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 $Y = \Theta$, $\Phi(x, y) = \ell_y(x|\sigma)$, and $h(x) = 0$.

1.2.3. Min-max games with an adversary. Let $\{\mathcal{U}_j(x_1,\ldots,x_k,y)\}_{j=1}^k$ be a set of utility functions in which (i) $\mathcal{U}_j$ is nonconvex and continuously differentiable in its first k arguments but concave in its last argument and (ii) $\nabla_{x_i}\mathcal{U}_j(x_1,\ldots,x_k,y)$ is Lipschitz continuous for every $1 \le i \le k$. 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.,

$$\min_{x_1,\ldots,x_k} \max_y \left\{ -\sum_{i=1}^k \mathcal{U}_j(x_1,\ldots,x_k,y) : x_i \in B_i, i = 0,\ldots,k \right\},$$

which is clearly an instance of (1.1) where $x = (x_1,\ldots,x_k)$, $Y = B_y$, $\Phi(x,y) = -\sum_{i=1}^k \mathcal{U}_j(x_1,\ldots,x_k,y)$, and $h(x) = \delta_{B_1\times\cdots\times B_k}(x)$.

2. Preliminaries.

We first present some preliminary material in two parts. The first one 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 an approximation of the max function p in (1.1) and of its properties.

2.1. Assumptions and notions of stationary points.

We present four notions of stationarity for (1.1). Two of these notions appear in the complexity results of of section 3, while the remaining two appear in related works. For the sake of comparison, the relationships between all four are discussed in this subsection.

Throughout our presentation, we let $\mathcal{X}$ and $\mathcal{Y}$ be Euclidean 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) $h \in \overline{\text{Conv}}(X)$ and $-\Phi(x,\cdot) \in \overline{\text{Conv}}(Y)$ 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, for every $x, x' \in X$ and $y, y' \in Y$, we have

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

$$\|\nabla_x \Phi(x,y) - \nabla_x \Phi(x',y')\| \le L_x\|x - x'\| + L_y\|y - y'\|; \tag{2.2}$$

- (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

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

for every $(x',x,y) \in X \times X \times Y$. Second, functions satisfying (2.1) are often referred to as weakly convex functions (see, for example, [5,6,7,8]). 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

$$\hat{\Phi}(x,y) := \Phi(x,y) + h(x) \quad \forall (x,y) \in X \times Y, \tag{2.4}$$

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

$$\hat{\Phi}(x^*,y) \le \hat{\Phi}(x^*,y^*) \le \hat{\Phi}(x,y^*) \tag{2.5}$$

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

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \nabla_x \Phi(x^*, y^*) \\ 0 \end{pmatrix} + \begin{pmatrix} \partial h(x^*) \\ \partial \left[-\Phi(x^*, \cdot)\right](y^*) \end{pmatrix}. \tag{2.6}$$

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

$$\inf_{\|d\| \le 1} \hat{p}'(x^*; d) \ge 0. \tag{2.7}$$

When $m = 0$, the above condition also becomes sufficient for x^* to be a global minimum of (1.1). Moreover, in view of Lemma 19 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 difficult. Hence, in this section and the next one, we only consider their approximate versions, which are (1.4) and (1.5). For ease of future reference, we say that

(i) a quadruple $(\bar{u}, \bar{v}, \bar{x}, \bar{y})$ is a (ρ_x, ρ_y)-*primal-dual stationary point* of (1.1) if (1.4) holds;

(ii) a point $\hat{x}$ is a δ-*directional stationary point* of (1.1) if the first inequality in (1.5) holds.

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 checking an infinite number of directional derivatives for a (potentially) nonsmooth function at points $\hat{x}$ near $\bar{x}$. In contrast, the definition of an approximate primal-dual stationary point is generally easier to verify because the quantities $\|\bar{u}\|$ and $\|\bar{v}\|$ 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 (ρ_x, ρ_y)-primal-dual stationary point, for small enough ρ_x and ρ_y, yields a point x satisfying (1.5). Its statement makes use of the diameter of Y defined as

$$D_y := \sup_{y, y' \in Y} \|y - y'\|. \tag{2.8}$$

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

$$\inf_{\|d\| \le 1} \hat{p}'(\hat{x}; d) \ge -\rho_x - 2\sqrt{2mD_y\rho_y}, \quad \|\bar{x} - \hat{x}\| \le \sqrt{\frac{2D_y\rho_y}{m}}.$$

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 (λ, ε)-prox stationary point of (1.1) if the function $\hat{p} + \|\cdot\|^2/(2\lambda)$ is strongly convex and

$$\frac{1}{\lambda}\|x - x_\lambda\| \le \varepsilon, \quad x_\lambda = \underset{u\in\mathcal{X}}{\operatorname{argmin}} \left\{ \hat{P}_\lambda(u) := \hat{p}(u) + \frac{1}{2\lambda}\|u - x\|^2 \right\}. \tag{2.9}$$

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*

$$0 < \delta \le \frac{\lambda^3\varepsilon}{\lambda^2 + 2(1-\lambda m)(1+\lambda)}, \tag{2.10}$$

then x is a (λ,ε)-prox stationary point;

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

Note that for a fixed $\lambda \in (0, 1/m)$ such that $\max\{\lambda^{-1}, (1-\lambda m)^{-1}\} = \mathcal{O}(1)$, the largest δ in part (a) is $\mathcal{O}(\varepsilon)$. Similarly, for part (b), if $\lambda^{-1} = \mathcal{O}(1)$, then the largest ε 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 (ρ_x, ρ_y)-first-order Nash equilibrium point of (1.1) if

$$\inf_{\|d_x\|\le 1} \mathcal{S}'_{\bar{y}}(\bar{x}; d_x) \ge -\rho_x, \quad \sup_{\|d_y\|\le 1} \mathcal{S}'_{\bar{x}}(\bar{y}; d_y) \le \rho_y, \tag{2.11}$$

where $\mathcal{S}_{\bar{y}} := \Phi(\cdot, \bar{y}) + h(\cdot)$ and $\mathcal{S}_{\bar{x}} := \Phi(\bar{x}, \cdot)$. 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 (ρ_x, ρ_y)-first-order Nash equilibrium point if and only if there exists $(\bar{u}, \bar{v}) \in \mathcal{X} \times \mathcal{Y}$ 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 [15] 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. We present an approximation of p in (1.1).

For every $\xi > 0$, consider the smoothed function p_ξ defined by

$$p_\xi(x) := \max_{y\in Y} \left\{ \Phi_\xi(x, y) := \Phi(x, y) - \frac{1}{2\xi}\|y - y_0\|^2 \right\} \quad \forall x \in X, \tag{2.12}$$

for some $y_0 \in Y$. The following proposition presents the properties of p_ξ.

Proposition 4. *Let $\xi > 0$ be given, and assume that the function Φ satisfies conditions* (A0)–(A3). *Let $p_\xi(\cdot)$ and $\Phi_\xi(\cdot,\cdot)$ be as defined in* (2.12), *and define*

$$\begin{gathered} Q_\xi := \xi L_y + \sqrt{\xi(L_x + m)}, \quad L_\xi := L_y Q_\xi + L_x \le \left(L_y\sqrt{\xi} + \sqrt{L_x} \right)^2, \\ y_\xi(x) := \underset{y'\in Y}{\operatorname{argmax}}\, \Phi_\xi(x, y') \end{gathered} \tag{2.13}$$

for every $x \in X$. Then, the following properties hold:

(a) *$y_\xi(\cdot)$ is Q_ξ-Lipschitz continuous on X;*
(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_ξ-Lipschitz continuous on X;*
(d) *for every $x, x' \in X$, we have*

$$p_\xi(x) - [p_\xi(x') + \langle \nabla p_\xi(x'), x - x' \rangle] \geq -\frac{m}{2}\|x - x'\|^2. \tag{2.14}$$

Proof. First, the inequality in (2.13) follows from (a), the bound $m \leq L_x$, and

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

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 Φ_ξ, and use the definition of ∇p_ξ in (b), then

$$\begin{aligned} &-\frac{m}{2}\|x - x'\|^2 \leq \Phi(x', y_\xi(x)) - [\Phi(x, y_\xi(x)) + \langle \nabla_x \Phi(x, y_\xi(x)), x' - x \rangle] \\ &= \Phi_\xi(x', y_\xi(x)) - [p_\xi(x) + \langle \nabla p_\xi(x), x' - x \rangle] \leq p_\xi(x') - [p_\xi(x) + \langle \nabla p_\xi(x), x' - x \rangle], \end{aligned}$$

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_ξ and ∇p_ξ depend on the value of ξ while the weak convexity constant m in (2.14) 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 $-\infty < p(x) - D_y^2/(2\xi) \leq p_\xi(x) \leq p(x)$ for every $x \in X$, 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 $y_0 \in Y$, we have

$$\Phi(x, y) + h(x) \geq \Phi(x, y) - \frac{1}{2\xi}\|y - y_0\|^2 + h(x) \geq \Phi(x, y) - \frac{D_y^2}{2\xi} + h(x)$$

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_ξ yields the remaining bounds. ∎

3. Unconstrained min-max optimization. We present the AIPP-S scheme for (1.1) in two parts. The first one reviews an AIPP method for solving CNO problems, while the second one presents the AIPP-S scheme and its complexity bounds. Throughout, $\mathcal{X}$ is a Euclidean space.

Before proceeding, we briefly outline the idea of the AIPP-S scheme. Essentially, it applies the AIPP method described in the next subsection to the CNO problem

$$\min_{x \in X} \left\{\hat{p}_\xi(x) := p_\xi(x) + h(x)\right\}, \tag{3.1}$$

where p_ξ is as in (2.12) and ξ 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 p_ξ used in both schemes is smooth, but the one here is nonconvex while the one in [23] is convex. Moreover, while [23] uses an accelerated composite gradient (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. We first describe the problem of interest. Consider the smooth CNO problem

$$\phi_* := \inf_{x \in \mathcal{X}} \left[\phi(x) := f(x) + h(x)\right], \tag{3.2}$$

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

(P1) $h \in \overline{\mathrm{Conv}}(\mathcal{X})$ and f is differentiable on $\mathrm{dom}\, h$;

(P2) for some $M \geq m > 0$ and every $x, x' \in \mathrm{dom}\, h$, the function f satisfies

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

$$\|\nabla f(x') - \nabla f(x)\| \leq M\|x' - x\|; \tag{3.4}$$

(P3) ϕ_* 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 $x^* \in \mathrm{dom}\, h$ to be a local minimum of (3.2) is that x^* is a stationary point of ϕ, 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 p_ξ in (2.12) satisfies assumption (P2) with $(M, f) = (L_\xi, p_\xi)$, where L_ξ is as in (2.13). Fourth, it is also easy to see that the function p_ξ in (2.12) satisfies assumption (P3) with $\phi_* = \inf_{x \in X} \hat{p}_\xi(x)$ in view of assumption (A4) and Lemma 5.

For the purpose of discussing future complexity results, we consider the following notion of an approximate stationary point of (3.2): given a tolerance $\bar{\rho} > 0$, a pair $(\bar{x}, \bar{u}) \in \mathrm{dom}\, h \times \mathcal{X}$ is said to be a $\bar{\rho}$-approximate stationary point of (3.2) if

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

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 $x_0 \in \mathrm{dom}\, h$, and a tolerance $\bar{\rho} > 0$;

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

(0) set $k = 1$ and define $\hat{\rho} := \bar{\rho}/4$, $\hat{\varepsilon} := \bar{\rho}^2/[32(M + \lambda^{-1})]$, and $M_\lambda := M + \lambda^{-1}$;

(1) call the ACG method in Appendix A with inputs $z_0 = x_{k-1}$, $(\mu, L) = (1/2, \lambda M + 1/2)$, $\psi_s = \lambda f + \|\cdot - x_{k-1}\|^2/4$, and $\psi_n = \lambda h + \|\cdot - x_{k-1}\|^2/4$ in order to obtain a triple $(x, u, \varepsilon) \in \mathcal{X} \times \mathcal{X} \times \mathbb{R}_+$ satisfying

$$u \in \partial_\varepsilon \left(\lambda\phi + \frac{1}{2}\|\cdot - x_{k-1}\|^2\right)(x), \quad \|u\|^2 + 2\varepsilon \leq \sigma\|x_{k-1} - x + u\|^2; \tag{3.6}$$

(2) if $\|x_{k-1} - x + u\| \leq \lambda\hat{\rho}/5$, 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 $\tilde{\varepsilon} \leq \hat{\varepsilon}\lambda$ and $(x, u, \varepsilon) = (\tilde{x}, \tilde{u}, \tilde{\varepsilon})$ satisfies (3.6);

(4) compute

$$\bar{x} := \operatorname*{argmin}_{x' \in \mathcal{X}} \left\{\langle \nabla f(x), x' - x\rangle + h(x') + \frac{M_\lambda}{2}\|x' - x\|^2\right\}, \tag{3.7}$$

$$\bar{u} := M_\lambda(x - \bar{x}) + \nabla f(\bar{x}) - \nabla f(x), \tag{3.8}$$

where M_λ 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 kth 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

$$\min_{x\in\mathcal{X}}\left\{\lambda\phi(x)+\frac{1}{2}\|x-x_{k-1}\|^2\right\} \tag{3.9}$$

according to (3.6). Second, Lemma 12 shows that every ACG iterate (z,u,ε) satisfies the inclusion in (3.6), and hence, only the inequality in (3.6) needs to be verified. Third, (3.4) implies that the gradient $\nabla\psi_s$ is $(\lambda M+1/2)$-Lipschitz continuous. Hence, Lemma 12 with $L=\lambda M+1/2$ implies that the triple (z,u,ε) obtained in step 1 requires $\mathcal{O}(\sqrt{[\lambda M+1]/\sigma})$ ACG iterations.

Note that the above method differs slightly from the one presented in [15] in that it adds step 4 in order to directly output a $\bar\rho$-approximate stationary point as in (3.5). The justification for the latter claim follows from [15, Lemma 12], [15, Theorem 13], and [15, Corollary 14], which also imply the following complexity result.

Proposition 6. *The AIPP method outputs a $\bar\rho$-approximate stationary point of* (3.2) *in*

$$\mathcal{O}\left(\sqrt{\lambda M+1}\left[\frac{R(\phi;\lambda)}{\sqrt{\sigma}(1-\sigma)^2\lambda^2\bar\rho^2}+\log_1^+(\lambda M)\right]\right) \tag{3.10}$$

ACG iterations, where

$$R(\phi;\lambda)=\inf_{x'}\left\{\frac{1}{2}\|x_0-x'\|^2+\lambda\left[\phi(x')-\phi_*\right]\right\}. \tag{3.11}$$

Note that scaling $R(\phi;\lambda)$ by $1/\lambda$ and then shifting by ϕ_* results in the λ-Moreau envelope[1] of ϕ. Moreover, $R(\phi;\lambda)$ admits the upper bound

$$R(\phi;\lambda)\le\min\left\{\frac{1}{2}d_0^2,\lambda\left[\phi(x_0)-\phi_*\right]\right\} \tag{3.12}$$

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

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 ξ and the tolerance ρ used in its step 2 are chosen. Two invocations of this method, with different choices of ξ and ρ, 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

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 $\rho>0$;

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

[1]See [28, Chapter 1.G] for an exact definition.

(0) set L_ξ as in (2.13), $\sigma = 1/2$, $\lambda = 1/(4m)$, and define p_ξ as in (2.12);

(1) apply the AIPP method with inputs (m, L_ξ), (p_ξ, h), λ, σ, x_0, and ρ to obtain a pair (x, u) satisfying

$$u \in \nabla p_\xi(x) + \partial h(x), \quad \|u\| \le \rho; \tag{3.13}$$

(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 [15, Theorem 13] and the third and fourth remarks following assumptions (P1)–(P3). Second, since the AIPP-S scheme is a one-pass method,[2] 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.

Our goal now is to show that a careful selection of the scalars ξ and ρ allows the AIPP-S method to output approximate stationary points as in (1.4) and (1.5). 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 (3.10) with $\phi = \hat{p}_\xi$.

LEMMA 7. *For every $\xi > 0$ and $\lambda \ge 0$, it holds that*

$$R(\hat{p}_\xi; \lambda) \le R(\hat{p}; \lambda) + \frac{\lambda D_y^2}{2\xi}, \tag{3.14}$$

where $R(\cdot, \cdot)$ and D_y are as in (3.11) and (2.8), respectively.

Proof. Using Lemma 5 and the definitions of $\hat{p}$ and $\hat{p}_\xi$, it holds that

$$\hat{p}_\xi(x) - \inf_{x'} \hat{p}_\xi(x') \le \hat{p}(x) - \inf_{x'} \hat{p}(x') + \frac{D_y^2}{2\xi} \quad \forall x \in X. \tag{3.15}$$

Multiplying the above expression by $(1 - \sigma)\lambda$ and adding the quantity $\|x_0 - x\|^2/2$ yields the inequality

$$\begin{aligned} &\frac{1}{2}\|x_0 - x\|^2 + (1 - \sigma)\lambda \left[\hat{p}_\xi(x) - \inf_{x'} \hat{p}_\xi(x')\right] \\ &\le \frac{1}{2}\|x_0 - x\|^2 + (1 - \sigma)\lambda \left[\hat{p}(x) - \inf_{\tilde{x}} \hat{p}(x')\right] + (1 - \sigma)\frac{\lambda D_y^2}{2\xi} \quad \forall x \in X, \end{aligned} \tag{3.16}$$

Taking the infimum of the above expression, and using the definition of $R(\cdot; \cdot)$ in (3.11) yields the desired conclusion. ∎

We now show how the AIPP-S scheme generates a (ρ_x, ρ_y)-primal-dual stationary point of (1.1), i.e., a quadruple $(\bar{u}, \bar{v}, \bar{x}, \bar{y})$ satisfying (1.4).

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 ξ and tolerance ρ satisfying $\xi \ge D_y/\rho_y$ and $\rho = \rho_x$. Moreover, define*

[2] As opposed to an iterative method.

$$(\bar{u}, \bar{v}) := \left(u, \frac{y_0 - y_\xi(x)}{\xi} \right), \quad (\bar{x}, \bar{y}) := (x, y_\xi(x)), \tag{3.17}$$

where y_ξ is as in (2.13)*. Then, the following statements hold:*

(a) *the AIPP-S scheme performs*

$$\mathcal{O}\left(\Omega_\xi \left[\frac{m^2 R(\hat{p}; 1/(4m))}{\rho_x^2} + \frac{m D_y^2}{\xi \rho_x^2} + \log_1^+(\Omega_\xi) \right] \right) \tag{3.18}$$

oracle calls, where $R(\cdot;\cdot)$ and D_y are as in (3.11) *and* (2.8)*, respectively, and*

$$\Omega_\xi := 1 + \frac{\sqrt{\xi} L_y + \sqrt{L_x}}{\sqrt{m}}; \tag{3.19}$$

(b) *the quadruple $(\bar{u}, \bar{v}, \bar{x}, \bar{y})$ is a (ρ_x, ρ_y)-primal-dual stationary point of* (1.1).

Proof. (a) Using the inequality in (2.13), it holds that

$$\sqrt{\frac{L_\xi}{4m} + 1} \le 1 + \sqrt{\frac{L_\xi}{4m}} \le 1 + \frac{\sqrt{\xi} L_y + \sqrt{L_x}}{2\sqrt{m}} = \Theta(\Omega_\xi). \tag{3.20}$$

Moreover, using Proposition 6 with $(\phi, M) = (\hat{p}_\xi, L_\xi)$, Lemma 7, and bound (3.20), it follows that the number of ACG iterations performed by the AIPP-S scheme is on the order given by (3.18). Since step 1 of the AIPP invokes once the ACG variant in Appendix A with a pair (ψ_s, ψ_n) of the form

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

for some $\tilde{z}$ and each iteration of this ACG variant performs $\mathcal{O}(1)$ gradient evaluations of ψ_s, $\mathcal{O}(1)$ function evaluations of ψ_s and ψ_n, and $\mathcal{O}(1)$ ψ_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 $\mathcal{O}(1)$ oracle calls. Statement (a) now follows from the above two conclusions.

(b) It follows from the definitions of p_ξ, tolerance ρ, and $(\bar{y}, \bar{u})$ in (2.12), the choice of ξ and ρ, and (3.17), respectively, Proposition 4(b), and the inclusion in (3.13) that $\|\bar{u}\| \le \rho_x$ 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 in (2.12) give

$$0 \in \partial\left[-\Phi(\bar{x}, \cdot)\right](\bar{y}) + \frac{\bar{y} - y_0}{\xi} = \partial\left[-\Phi(\bar{x}, \cdot)\right](\bar{y}) - \bar{v}. \tag{3.21}$$

Moreover, the definition of ξ implies that $\|\bar{v}\| = \|\bar{y} - y_0\|/\xi \le D_y/(D_y/\rho_y) = \rho_y$. Hence, combining (3.21) and the previous identity, we conclude that the bottom inclusion and the upper bound on $\|\bar{v}\|$ in (1.4) hold. ∎

We now make three remarks about Proposition 8. First, recall that $R(\hat{p}; 1/(4m))$ in the complexity (3.18) can be majorized by the rightmost quantity in (3.12) with

$(\phi, \lambda) = (\hat{p}, 1/(4m))$. Second, under the assumption that $\xi = D_y/\rho_y$, the complexity of AIPP-S scheme reduces to

$$\mathcal{O}\left(m^{3/2} \cdot R(\hat{p}; 1/(4m)) \cdot \left[\frac{L_x^{1/2}}{\rho_x^2} + \frac{L_y D_y^{1/2}}{\rho_x^2 \rho_y^{1/2}}\right]\right) \tag{3.22}$$

under the reasonable assumption that the $\mathcal{O}(\rho_x^{-2} + \rho_x^{-2}\rho_y^{-1/2})$ term in (3.18) dominates the other terms. Third, recall from the last remark following the previous proposition that when Y is a singleton, (1.1) is a special instance of (3.2) and the AIPP-S scheme is equivalent to the AIPP method of subsection 3.1. It similarly follows that the complexity in (3.22) reduces to $\mathcal{O}(\rho_x^{-2})$ and, hence, the $\mathcal{O}(\rho_x^{-2}\rho_y^{-1/2})$ term in (3.22) is attributed to the (possible) nonsmoothness in (1.1).

We next show how the AIPP-S scheme generates a point that is *near* a δ-directional stationary point of (1.1), i.e., a point $\hat{x}$ satisfying the first inequality in (1.5).

Proposition 9. *Let a tolerance pair $\delta > 0$ be given, and consider the AIPP-S scheme with input parameter ξ and tolerance ρ satisfying $\xi \geq D_y/\tau$ and $\rho = \delta/2$ for some $\tau \leq \min\left\{m\delta^2/2D_y, \delta^2/32mD_y\right\}$. Then, the following statements hold:*

(a) *the AIPP-S scheme performs*

$$\mathcal{O}\left(\Omega_\xi \left[\frac{R(\hat{p}; \lambda)}{\lambda^2\delta^2} + \frac{D_y^2}{\lambda\xi\delta^2} + \log_1^+(\Omega_\xi)\right]\right) \tag{3.23}$$

oracle calls where Ω_ξ, $R(\cdot;\cdot)$, and D_y are as in (3.19), (3.11), *and* (2.8)*;*

(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 τ it follows that the number of ACG iterations needed by AIPP-S is as in (3.23).

(b) Let (x, u) be the $\bar{\rho}$-approximate stationary point of (3.1) generated by the AIPP-S scheme (see step 2) under the given assumption on ξ and $\bar{\rho}$. Defining $(\bar{v}, \bar{y})$ as in (3.17), it follows from Proposition 8 with $(\rho_x, \rho_y) = (\delta/2, \tau)$ that $(u, \bar{v}, x, \bar{y})$ 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

$$\|\hat{x} - x\| \leq \sqrt{\frac{2D_y\tau}{m}}, \quad \inf_{\|d\|\leq 1} \hat{p}'(\hat{x}; d) \geq -\frac{\delta}{2} - 2\sqrt{2mD_y\tau}. \tag{3.24}$$

Combining the above bounds with our assumption on τ yields the desired conclusion in view of (1.5). ∎

We now give four remarks about the above result. First, recall that $R(\hat{p}; 1/(4m))$ in the complexity (3.23) is majorized by the rightmost quantity in (3.12) with $(\phi, \lambda) = (\hat{p}, 1/(4m))$. Second, Proposition 9(b) states that while x not a stationary point itself, it is near a δ-directional stationary point $\hat{x}$. Third, under the assumption that the bounds on ξ and τ in Proposition 9 hold at equality, the complexity of the AIPP-S scheme is

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

under the reasonable assumption that the $\mathcal{O}(\delta^{-2} + \delta^{-3})$ term in (3.23) dominates the other $\mathcal{O}(\delta^{-1})$ terms. Fourth, when Y is a singleton, it is easy to see that (1.1) is a special instance of (3.2), the AIPP-S scheme is equivalent to the AIPP method of subsection 3.1, and the complexity in (3.25) is $\mathcal{O}(\delta^{-2})$. In view of the last remark, the $\mathcal{O}(\delta^{-3})$ term in (3.25) is attributed to the (possible) nonsmoothness in (1.1).

4. Linearly constrained min-max optimization. We present the QP-AIPP-S scheme for (1.6) in two parts. The first one reviews a QP-AIPP method for linearly-constrained CNO problems, while the second presents the QP-AIPP-S scheme and its complexity bound. Throughout, $\mathcal{X}, \mathcal{Y}$, and $\mathcal{U}$ are Euclidean spaces.

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

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

(A6) there exists $\hat{c} \geq 0$ such that $\inf_{x\in X} \left\{\hat{p}(x) + \hat{c}\|\mathcal{A}x - b\|^2/2\right\} > -\infty$.

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

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

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \nabla_x \Phi(x^*, y^*) + A^* r^* \\ 0 \end{pmatrix} + \begin{pmatrix} \partial h(x^*) \\ \partial\left[-\Phi(x^*, \cdot)\right](y^*) \end{pmatrix}, \tag{4.1}$$

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

$$\min_{x\in X} \left\{p_\xi(x) + h(x) : \mathcal{A}x = b\right\}, \tag{4.2}$$

where p_ξ is as in (1.2) and ξ 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. We first describe the problem of interest. Consider the linearly constrained CNO problem

$$\hat{\phi}_* := \inf_{x\in\mathcal{X}} \{\phi(x) := f(x) + h(x) : \mathcal{A}x = b\}, \tag{4.3}$$

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

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

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

$$\hat{\phi}_c := \inf_{x\in\mathcal{X}} \left\{\phi_c(x) := \phi(x) + \frac{c}{2}\|\mathcal{A}x - b\|^2\right\} \quad \forall c \geq 0. \tag{4.4}$$

We now give some remarks about the above assumptions. First, similar to problem (3.2), it is well-known that a necessary condition for $x^* \in \operatorname{dom} h$ to be a local minimum of (4.3) is that x^* satisfies $0 \in \nabla f(x^*) + \partial h(x^*) + \mathcal{A}^* r^*$ for some $r^* \in \mathcal{U}$. Second, it is easy to see that $(p, h, \mathcal{A}, b)$ 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 assumption (Q2) that $\hat{\phi}_* \geq \hat{\phi}_{\hat{c}} > -\infty$. Fourth, if $\inf_{x\in\mathcal{X}} \phi(x) > -\infty$ (e.g., $\operatorname{dom} h$ 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 $(\bar{x}, \bar{u}, \bar{r}) \in \operatorname{dom} h \times \mathcal{X} \times \mathcal{U}$ is said to be a $(\bar{\rho}, \bar{\eta})$-approximate stationary point of (4.3) if

$$\bar{u} \in \nabla f(\bar{x}) + \partial h(\bar{x}) + \mathcal{A}^* \bar{r}, \quad \|\bar{u}\| \le \bar{\rho}, \quad \|\mathcal{A}\bar{x} - b\| \le \bar{\eta}. \tag{4.5}$$

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 $x_0 \in \operatorname{dom} h$, and a tolerance pair $(\bar{\rho}, \bar{\eta}) \in \mathbb{R}^2_{++}$;

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

(0) set $c = \hat{c} + M/\|\mathcal{A}\|^2$;

(1) define the quantities

$$M_c := M + c\|\mathcal{A}\|^2, \quad f_c := f + \frac{c}{2}\|\mathcal{A}(\cdot) - b\|^2, \quad \phi_c = f_c + h, \tag{4.6}$$

and apply the AIPP method with inputs (m, M_c), (f_c, h), λ, σ, x_0, and $\bar{\rho}$ to obtain a $\bar{\rho}$-approximate stationary point $(\bar{x}, \bar{u})$ of (3.2) with $f = f_c$;

(2) if $\|\mathcal{A}\bar{x} - b\| > \bar{\eta}$, then set $c = 2c$ and go to (1); otherwise, set $\bar{r} = c\,(\mathcal{A}\bar{x} - b)$ and output the triple $(\bar{x}, \bar{u}, \bar{r})$.

We now give two remarks about the above method. First, it terminates due to the results in [15, section 4]. Second, in view of Proposition 6 with $(\phi, M) = (\phi_c, M_c)$, the number of ACG iterations executed in step 1 at any iteration of the method is

$$\mathcal{O}\left(\sqrt{\lambda M_c + 1}\left[\frac{R(\phi_c; \lambda)}{\sqrt{\sigma}(1-\sigma)^2 \lambda^2 \bar{\rho}^2} + \log_1^+(\lambda M_c)\right]\right), \tag{4.7}$$

and the pair $(\bar{x}, \bar{u})$ 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

$$R_c(\phi; \lambda) := \inf_{x'}\left\{\frac{1}{2}\|x_0 - x'\|^2 + \lambda\left[\phi(x') - \hat{\phi}_c\right] : x' \in \mathcal{F}\right\} \tag{4.8}$$

for every $c \ge \hat{c}$, where ϕ_c is as defined in (4.4). The quantity in (4.8) plays an analogous role as (3.11) in (3.10), and, similar to the discussion following Proposition 6, it is a scaled and shifted λ-Moreau envelope of $\phi + \delta_{\mathcal{F}}$. Moreover, due to [15, Lemma 16], it also admits the upper bound

$$R_c(\phi; \lambda) \le R_{\hat{c}}(\phi; \lambda) \le \min\left\{\frac{1}{2}\hat{d}_0^2, \lambda\left[\hat{\phi}_* - \hat{\phi}_{\hat{c}}\right]\right\}, \tag{4.9}$$

where $\hat{\phi}_*$ is as defined in (4.3) and

$$\hat{d}_0 := \inf\left\{\|x_0 - x_*\| : x_* \text{ is an optimal solution of (4.3)}\right\}.$$

We now state the iteration complexity of the QP-AIPP method, whose proof follows from [15, Lemma 12] and [15, Theorem 18].

PROPOSITION 10. *Let a $\hat{c}$ as in* (Q2)*, scalar $\sigma \in (0,1)$, curvature pair $(m, M) \in \mathbb{R}^2_{++}$, and a tolerance pair $(\bar{\rho}, \bar{\eta}) \in \mathbb{R}^2_{+}$ be given. Moreover, define*

$$T_{\bar{\eta}} := \frac{2R_{\hat{c}}(\phi;\lambda)}{\bar{\eta}^2(1-\sigma)\lambda} + \hat{c}, \quad \Theta_{\bar{\eta}} := M + T_{\bar{\eta}}\|\mathcal{A}\|^2. \tag{4.10}$$

Then, the QP-AIPP method outputs a triple $(\bar{x}, \bar{u}, \bar{r})$ satisfying (4.5) *in*

$$\mathcal{O}\left(\sqrt{\lambda\Theta_{\bar{\eta}}+1}\left[\frac{R_{\hat{c}}(\phi;\lambda)}{\sqrt{\sigma}(1-\sigma)^2\lambda^2\bar{\rho}^2} + \log_1^+(\lambda\Theta_{\bar{\eta}})\right]\right) \tag{4.11}$$

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 $(m, L_x, L_y) \in \mathbb{R}^2_{++}$ as in (A3), a scalar $\hat{c}$ as in (A6), a scalar $\xi \geq D_y/\rho_y$, 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);

(0) set L_ξ as in (2.13), $\sigma = 1/2$, $\lambda = 1/(4m)$, and define p_ξ as in (2.12);

(1) apply the QP-AIPP method of subsection 4.1 with inputs (m, L_ξ), (p_ξ, h), λ, σ, $\hat{c}$, x_0, and (ρ_x, η) to obtain a triple $(\bar{u}, \bar{x}, \bar{r})$ satisfying

$$\bar{u} \in \nabla p_\xi(\bar{x}) + \partial h(\bar{x}) + A^*\bar{r}, \quad \|\bar{u}\| \leq \rho_x, \quad \|\mathcal{A}\bar{x} - b\| \leq \eta. \tag{4.12}$$

(2) define $(\bar{v}, \bar{y})$ as in (3.17) 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,[3] 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 that the output of the QP-AIPP-S scheme satisfies (1.8).

PROPOSITION 11. *Let a tolerance triple $(\rho_x, \rho_x, \eta) \in \mathbb{R}^3_{++}$ 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*

$$\mathcal{O}\left(\Omega_{\xi,\eta}\left[\frac{m^2 R_{\hat{c}}(\hat{p}; 1/(4m))}{\rho_x^2} + \frac{mD_y^2}{\xi\rho_x^2} + \log_1^+(\Omega_{\xi,\eta})\right]\right) \tag{4.13}$$

oracle calls, where

$$\Omega_{\xi,\eta} := \Omega_\xi + \left(R_{\hat{c}}(\hat{p}; 1/(4m)) + \frac{D_y^2}{m\xi}\right)^{1/2} \frac{\|\mathcal{A}\|}{\eta} \tag{4.14}$$

[3] As opposed to an iterative algorithm.

and Ω_ξ, $R(\cdot;\cdot)$, *and* D_y *are as in* (3.19), (3.11), *and* (2.8), *respectively*;
(b) *the quintuple* $(\bar{u},\bar{v},\bar{x},\bar{y},\bar{r})$ *satisfies* (1.8).

Proof. (a) Let Θ_η be as in (4.10) with $M = L_\xi$. Using the same arguments as in Lemma 7, it is easy to see that $R_{\hat{c}}(\hat{p}_\xi; 1/(4m)) \le R_{\hat{c}}(\hat{p}; 1/(4m)) + D_y^2/(8m\xi)$, and hence, using (3.20), we have

$$\sqrt{\frac{\Theta_\eta}{4m}+1} \le 1 + \sqrt{\frac{L_\xi}{4m}} + \sqrt{\frac{4R_{\hat{c}}(\hat{p}_\xi;1/(4m))\|\mathcal{A}\|^2}{\eta^2}}$$

$$\le 1 + \frac{\sqrt{\xi}L_y + \sqrt{L_x}}{2\sqrt{m}} + 2\left(R_{\hat{c}}(\hat{p};1/(4m)) + \frac{D_y^2}{8m\xi}\right)^{1/2}\frac{\|\mathcal{A}\|}{\eta} = \Theta(\Omega_{\xi,\eta}). \tag{4.15}$$

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

(b) The top inclusion and bounds involving $\|\bar{u}\|$ and $\|\mathcal{A}\bar{x} - b\|$ in (1.8) follow from Proposition 4(b), the definition of $\bar{y}$ in step 2 of the algorithm, and Proposition 10 with $f = p_\xi$. 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 $R_{\hat{c}}(p; 1/(4m))$ 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

$$\mathcal{O}\left(m^{3/2}\cdot R_{\hat{c}}(\hat{p};1/(4m))\cdot\left[\frac{L_x^{1/2}}{\rho_x^2} + \frac{L_y D_y^{1/2}}{\rho_y^{1/2}\rho_x^2} + \frac{m^{1/2}\|\mathcal{A}\|R_{\hat{c}}^{1/2}(p;1/(4m))}{\eta\rho_x^2}\right]\right) \tag{4.16}$$

under the reasonable assumption that the $\mathcal{O}(\rho_x^{-2} + \eta^{-1}\rho_x^{-2} + \rho_y^{-1/2}\rho_x^{-2})$ term in (4.13) dominates the other terms. Third, when Y is a singleton, it is easy to see that (1.6) is a special instance of the linearly constrained smooth CNO problem (4.3), the QP-AIPP-S of this subsection is equivalent to the QP-AIPP method of subsection 4.1, and the complexity in (4.16) is $\mathcal{O}(\eta^{-1}\rho_x^{-2})$. In view of the last remark, the $\mathcal{O}(\rho_x^{-2}\rho_y^{-1/2})$ term in (4.16) is attributed to the (possible) nonsmoothness in (1.6).

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

$$\min_{x\in X}\left\{p_\xi(x) + h(x) + \frac{c}{2}\|\mathcal{A}x - b\|^2\right\}, \tag{4.17}$$

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

$$\min_{x\in X}\max_{y\in Y, r\in\mathcal{U}}\left[\Psi(x,y,r) := \Phi(x,y) + h(x) + \langle r, \mathcal{A}x - b\rangle\right]. \tag{4.18}$$

Second, it is straightforward to verify that problem (4.17) is equivalent to

$$\min_{x\in X}\left\{\hat{p}_{c,\xi}(x) := p_{c,\xi}(x) + h(x)\right\}, \tag{4.19}$$

where the function $p_{c,\xi} : X \mapsto \mathbb{R}$ is given by $p_{c,\xi}(x) := \max_{y\in Y, r\in\mathcal{U}}\{\Psi(x,y,r) - \|r\|^2/(2c) - \|y - y_0\|^2/(2\xi)\}$ for every $x \in X$, and Ψ as in (4.18). As a consequence, problem (4.19) is similar to (3.1) in that a smooth approximate is used in place of the nonsmooth component of the underlying saddle function Ψ.

On the other hand, observe that we cannot directly apply the smoothing scheme developed in subsection 3.2 to (4.19) 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 (4.19) for increasing values of c. However, in view of the equivalence of (4.17) and (4.19), this is exactly the approach taken by the QP-AIPP-S scheme of this section.

5. Numerical experiments. We present numerical results that illustrate the computational efficiency of the our smoothing scheme in three parts. Each part presents computational results for a specific min-max optimization problem.

Each unconstrained problem considered in this section is of the form in (1.1) and is such that the computation of the function y_ξ 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 the inclusion of (1.4) and

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

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 relaxed AIPP smoothing method (abbreviated RA-S), (ii) the accelerated gradient smoothing (AG-S) scheme, and (iii) the projected gradient step framework (PGSF). Both the AG-S and RA-S schemes are modifications of the AIPP-S scheme which, instead of using the AIPP method in its step 1, use the accelerated gradient (AG) method of [10] and relaxed AIPP (R-AIPP) method of [16], respectively. The PGSF is a simplified variant of Algorithm 2 of [24, subsection 4.1] which explicitly evaluates the argmax function $\alpha^*(\cdot)$ in [24, section 4] instead of applying an ACG variant to estimate its evaluation.

Regarding the penalty solvers, the AG method is in [10, Algorithm 2] while the R-AIPP method is as in [14, section 5.3].

Note that, like the AIPP method, the R-AIPP similarly (i) invokes at each of its outer iterations an ACG method to inexactly solve the proximal subproblem (3.9) and (ii) outputs a $\bar{\rho}$-approximate stationary point of (3.2). However, the R-AIPP method is more efficient due to three practical improvements over the AIPP method, namely, (i) it allows the stepsize λ 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.

We next state some additional details about the numerical experiments. First, each algorithm is run with a time limit of 4000 seconds. Second, 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 $\hat{p}_\xi(\bar{x})$ that contains the smallest obtained value of the smoothed function in (3.1) across all of the tested algorithms. Third, the description of y_ξ and choice of the constants m, L_x, and L_y for each of the considered optimization problems can be found in [14, Appendix I]. Fourth, 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.[4]

5.1. Maximum of a finite number of nonconvex quadratic forms. 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

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

where, for every index $1 \le i \le k$, integer $p \in \mathbb{N}$, and $x \in \mathbb{R}^n$, we define $g_i(x) := \alpha_i\|C_i x - d_i\|^2/2 - \beta_i\|D_i B_i x\|^2/2$ and $\Delta^p := \left\{z \in \mathbb{R}^p_+ : \sum_{i=1}^p z_i = 1, z \ge 0\right\}$.

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 $\mathcal{U}[0, 1]$ (resp., $\mathcal{U}[1, 1000]$). 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, the inputs, for every $(x, y) \in \mathbb{R}^n \times \mathbb{R}^k$, are $\Phi(x, y) = \sum_{i=1}^k y_i g_i(x)$, $h(x) = \delta_{\Delta^n}(x)$, $\rho_x = 10^{-2}$,$\rho_y = 10^{-1}$, and $Y = \Delta^k$. Finally, each problem instance considered is based on a specific curvature pair (m, M) satisfying $m \le M$, for which each scalar pair $(\alpha_i, \beta_i) \in \mathbb{R}^2_{++}$ is selected so that $M = \lambda_{\max}(\nabla^2 g_i)$ and $-m = \lambda_{\min}(\nabla^2 g_i)$.

We now present the results in Table 5.1.

5.2. Truncated robust regression. Given a dimension pair $(n, k) \in \mathbb{N}^2$, a set of n data points $\{(a_j, b_j)\}_{i=1}^n \subseteq \mathbb{R}^k \times \{1, -1\}$, 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 Δ^n is as in subsection 5.1 with $p = n$, $\phi_\alpha(t) := \alpha \log(1 + t/\alpha)$, and $\ell_j(x) := \log\left(1 + e^{-b_j\langle a_j, x\rangle}\right)$ for every $(\alpha, t, x) \in \mathbb{R}_{++} \times \mathbb{R}_{++} \times \mathbb{R}^k$.

We now describe the experiment parameters for the instances considered. First, α is set to 10, and the data points $\{(a_i, b_i)\}$ are taken from different datasets in the LIBSVM library[5] from which each problem instance is based (see the "data name"

Table 5.1
Iteration counts and runtimes for QVM problems.

M	m	$\hat{p}_\xi(\bar{x})$	Iteration count			Runtime		
			RA-S	AG-S	PGSF	RA-S	AG	PGSF
10^0	10^0	2.85E-01	**23**	294	1591	**0.66**	5.72	22.60
10^1	10^0	2.88E+00	**86**	1371	14815	**1.37**	25.96	209.62
10^2	10^0	2.85E+01	**217**	6270	150493	**3.35**	118.32	2122.93
10^3	10^0	2.85E+02	**1417**	28989	-	**21.58**	546.25	4000.00*

[4]See https://github.com/wwkong/nc_opt/tree/master/examples/minmax.
[5]See https://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/datasets/binary.html.

Table 5.2
Iteration counts and runtimes for TRR problems.

Data name	$\hat{p}_\xi(\bar{x})$	Iteration count			Runtime		
		RA-S	AG-S	PGSF	RA-S	AG	PGSF
heart	6.70E-01	**425**	1747	6409	**6.37**	15.54	32.76
diabetes	6.70E-01	**852**	1642	3718	**8.61**	24.12	52.77
ionosphere	6.70E-01	**1197**	8328	54481	**8.26**	63.82	320.72
sonar	6.70E-01	**45350**	96209	-	**461.52**	580.37	4000.00*
breast-cancer	1.11E-03	**46097**	-	-	**476.59**	4000.00*	4000.00*

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, the inputs, for every $(x,y) \in \mathbb{R}^k \times \mathbb{R}^n$, are $\Phi(x,y) = \sum_{j=1}^n y_j(\phi_\alpha \circ \ell_j)(x)$, $h(x) = 0$, $\rho_x = 10^{-5}$, $\rho_y = 10^{-3}$, and $Y = \Delta^n$.

We now present the results in Table 5.2.

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 $\mathcal{O}(\rho^{-6})$ 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.

5.3. Power control in the presence of a jammer. Given a dimension pair $(N,K) \in \mathbb{N}^2$, a pair of parameters $(\sigma, R) \in \mathbb{R}^2_{++}$, a three-dimensional tensor $\mathcal{A} \in \mathbb{R}_+^{K\times K\times N}$, and a matrix $B \in \mathbb{R}_+^{K\times N}$, the problem of interest is the power control (PC) problem

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

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

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

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 $\mathcal{A}$ 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 $\mathcal{CN}(0,1)$ with $\mathcal{A}_{j,k,n} = |\mathcal{H}_{j,k,n}|^2$ and $B_{k,n} = |P_{k,n}|^2$ for every (j,k,n). Second, the initial starting point is $z_0 = 0$. Third, with respect to the termination criterion, the inputs are $\Phi(X,y) = \sum_{k=1}^K \sum_{n=1}^N f_{k,n}(X,y)$, $h(X) = \delta_{Q_R^{K\times N}}(X)$, $\rho_x = 10^{-1}$, $\rho_y = 10^{-1}$, and $Y = Q_{N/2}^{N\times 1}$ for every $(X,y) \in \mathbb{R}^{K\times N} \times \mathbb{R}^N$ and $(U,V) \in \mathbb{N}^2$, where $Q_T^{U\times V} := \{z \in \mathbb{R}^{p\times q} : 0 \le z \le T\}$ for every $T > 0$. Fourth, each problem instance considered is based on a specific dimension pair (N,K).

We now present the results in Table 5.3.

It is worth mentioning that [18] also presents a min-max algorithm for obtaining stationary points for the aforementioned problem. However, its notion of stationarity is significantly different than what is being considered in this paper, and, hence, we choose not to its algorithm in our benchmarks.

TABLE 5.3
Iteration counts and runtimes for PC problems.

N	K	$\hat{p}_\xi(\bar{x})$	Iteration count			Runtime		
			RA-S	AG-S	PGSF	RA-S	AG	PGSF
5	5	-3.64E+00	**37**	322832	-	**0.96**	2371.27	4000.00*
10	10	-2.82E+00	**54**	33399	-	**0.75**	293.60	4000.00*
25	25	-4.52E+00	**183**	-	-	**9.44**	4000.00*	4000.00*
50	50	-4.58E+00	**566**	-	-	**40.89**	4000.00*	4000.00*

6. 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 (3.13), or equivalently,[6]

$$\inf_{\|d\|\le 1} (\hat{p}_\xi)'(x;d) \ge -\rho. \tag{6.1}$$

Moreover, using Proposition 8 with $(\rho_x, \rho_y) = (\rho, D_y/\xi)$, it straightforward to see that that the number of oracle calls, in terms of (ξ, ρ), is $\mathcal{O}(\rho^{-2}\xi^{1/2})$, which reduces to $\mathcal{O}(\rho^{-2.5})$ if ξ 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 $\mathcal{O}(\rho^{-3})$ 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[7] of the related mathematical program with equilibrium constraints:

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

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 [15].

Part of the input of the ACG variant, which is described below, consists of a pair of functions (ψ_s, ψ_n) satisfying

(i) $\psi_n \in \overline{\text{Conv}}(\mathcal{Z})$ is μ-strongly convex for some $\mu \ge 0$;
(ii) ψ_s is a convex differentiable function on $\operatorname{dom}\psi_n$ 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 (ψ_n, ψ_s), and an initial point $z_0 \in \operatorname{dom}\psi_n$;

(0) set $y_0 = z_0$, $A_0 = 0$, $\Gamma_0 \equiv 0$, and $j = 0$;
(1) compute

$$A_{j+1} = A_j + \frac{\mu A_j + 1 + \sqrt{(\mu A_j + 1)^2 + 4L(\mu A_j + 1)A_j}}{2L},$$

[6]See Lemma 15 with $f = p_\xi$.
[7]See, for example, [19, Chapter 3].

$$\begin{aligned}
\tilde{z}_j &= \frac{A_j}{A_{j+1}} z_j + \frac{A_{j+1} - A_j}{A_{j+1}} y_j, \\
\Gamma_{j+1}(y) &= \frac{A_j}{A_{j+1}} \Gamma_j(y) + \frac{A_{j+1} - A_j}{A_{j+1}} \left[\psi_s(\tilde{z}_j) + \langle \nabla \psi_s(\tilde{z}_j), y - \tilde{z}_j \rangle\right] \quad \forall y, \\
y_{j+1} &= \operatorname*{argmin}_y \left\{ \Gamma_{j+1}(y) + \psi_n(y) + \frac{1}{2A_{j+1}} \|y - y_0\|^2 \right\}, \\
z_{j+1} &= \frac{A_j}{A_{j+1}} z_j + \frac{A_{j+1} - A_j}{A_{j+1}} y_{j+1};
\end{aligned}$$

(2) compute

$$\begin{aligned}
u_{j+1} &= \frac{y_0 - y_{j+1}}{A_{j+1}}, \\
\varepsilon_{j+1} &= \psi(z_{j+1}) - \Gamma_{j+1}(y_{j+1}) - \psi_n(y_{j+1}) - \langle u_{j+1}, z_{j+1} - y_{j+1} \rangle;
\end{aligned}$$

(3) increment $j = j + 1$ and go to (1).

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 ψ_n, ψ_s, $\nabla \psi_s$ at any point in $\operatorname{dom} \psi_n$ and (ii) the computation of the exact solution of subproblems of the form $\min_y \left\{ \psi_n(y) + \|y - a\|^2/(2\alpha) \right\}$ for any $a \in \mathcal{Z}$ and $\alpha > 0$. In particular, the latter is needed in the computation of y_{j+1}. Second, because Γ_{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

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

The following result, given in [15, Lemma 9], is used to establish the work needed to obtain (z, u, ε) 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, ε) satisfying*

$$u \in \partial_\varepsilon(\psi_s + \psi_n)(z) \quad \|u\|^2 + 2\varepsilon \le \sigma \|z_0 - z + u\|^2 \tag{A.1}$$

in at most $\lceil 2\sqrt{2L}(1 + \sqrt{\sigma})/\sqrt{\sigma} \rceil$ 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 [9, Theorem 10.2.1], presents a key property about max functions.

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

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

Then, for every $(x,d) \in X \times \mathcal{X}$, 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 ∇q. It is worth mentioning that it generalizes related results in [2, Theorem 5.26] (which covers the case where Ψ 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. *Assume that the triple (X, Y, Ψ) satisfies* (A0)–(A3) *in subsection* 2.1 *with $\Phi = \Psi$ and that, for some $\mu > 0$, the function $\Psi(x, \cdot)$ is μ-strongly concave on Y for every $x \in X$, and define*

$$Q_\mu := \frac{L_y}{\mu} + \sqrt{\frac{L_x + m}{\mu}}, \quad L_\mu := L_y Q_\mu + L_x, \quad y(x) := \operatorname*{argmax}_{y \in Y} \Psi(x,y) \tag{B.2}$$

for every $x \in X$. Then, the following properties hold:

(a) *$y(\cdot)$ is Q_μ-Lipschitz continuous on X;*

(b) *$\nabla q(\cdot)$ is L_μ-Lipschitz continuous on X where q is as in* (B.1).

Proof. (a) Let $x, \tilde{x} \in X$ be given, and denote $(y, \tilde{y}) = (y(x), y(\tilde{x}))$. Define $\alpha(u) := \Psi(u, y) - \Psi(u, \tilde{y})$ for every $u \in X$, and observe that the optimality conditions of y and $\tilde{y}$ imply that $\alpha(x) \geq \mu \|y - \tilde{y}\|^2/2$ and $-\alpha(\tilde{x}) \geq \mu\|y - \tilde{y}\|^2/2$. Using the previous inequalities, (2.1), (2.2), (2.3), and the Cauchy–Schwarz inequality, we conclude that

$$\begin{aligned}
\mu\|y - \tilde{y}\|^2 \leq \alpha(x) - \alpha(\tilde{x}) &\leq \langle \nabla_x \Psi(x,y) - \nabla_x \Psi(x, \tilde{y}), x - \tilde{x} \rangle + \frac{L_x + m}{2}\|x - \tilde{x}\|^2 \\
&\leq \|\nabla_x \Psi(x,y) - \nabla_x \Psi(x,\tilde{y})\| \cdot \|x - \tilde{x}\| + \frac{L_x + m}{2}\|x - \tilde{x}\|^2 \\
&\leq L_y \|y - \tilde{y}\| \cdot \|x - \tilde{x}\| + \frac{L_x + m}{2}\|x - \tilde{x}\|^2.
\end{aligned}$$

Considering the above as a quadratic inequality in $\|\tilde{y} - y\|$ yields the bound

$$\begin{aligned}
\|y - \tilde{y}\| &\leq \frac{1}{2\mu}\left[L_y\|x - \tilde{x}\| + \sqrt{L_y^2\|x - \tilde{x}\|^2 + 4\mu(L_x + m)\|x - \tilde{x}\|^2}\right] \\
&\leq \left[\frac{L_y}{\mu} + \sqrt{\frac{L_x + m}{\mu}}\right]\|x - \tilde{x}\| = Q_\mu \|x - \tilde{x}\|
\end{aligned}$$

which is the conclusion of (a).

(b) Let $x, \tilde{x} \in X$ be given, and denote $(y, \tilde{y}) = (y(x), y(\tilde{x}))$. Using part (a), Lemma 13, and (2.2) we have that

$$\begin{aligned}
\|\nabla q(x) - \nabla q(\tilde{x})\| &= \|\nabla_x \Psi(x,y) - \nabla_x \Psi(\tilde{x}, \tilde{y})\| \\
&\leq \|\nabla_x \Psi(x,y) - \nabla_x \Psi(x,\tilde{y})\| + \|\nabla_x \Psi(x, \tilde{y}) - \nabla_x \Psi(\tilde{x}, \tilde{y})\| \\
&\leq L_y\|y - \tilde{y}\| + L_x\|x - \tilde{x}\| \leq (L_y Q_\mu + L_x)\|x - \tilde{x}\| = L_\mu \|x - \tilde{x}\|,
\end{aligned}$$

which is the conclusion of (b). ∎

Appendix C. The main goal of this appendix is to prove Propositions 17 and 18, which are used in the proofs of Propositions 1, 2, and 3 given in Appendix D.

The following well-known result presents an important property about the directional derivative of a composite function $f + h$.

LEMMA 15. *Let $h : \mathcal{X} \mapsto (-\infty, \infty]$ be a proper convex function, and let f be a differentiable function on $\operatorname{dom} h$. Then, for any $x \in \operatorname{dom} h$, it holds that*

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

The proof of Lemma 15 can be found for example in [28, Exercise 8.8(c)]. An alternative and more direct proof is given in [14, Lemma F.1.2]. It is also worth mentioning that if we further assumed that $\operatorname{dom} h = \mathcal{X}$, then the above result would follow from [3, Lemma 5.1].

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

LEMMA 16. *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 such that $\psi(\cdot, y)$ is convex lower semicontinuous for every $y \in Y$ and $\psi(d, \cdot)$ is concave upper semicontinuous for every $d \in D$. Then,*

$$\inf_{d\in\mathcal{X}}\sup_{y\in\mathcal{Y}}\psi(d,y) = \sup_{y\in\mathcal{Y}}\inf_{d\in\mathcal{X}}\psi(d,y).$$

The next result establishes an identity similar to Lemma 15 but for the case where f is a max function.

PROPOSITION 17. *Assume the quadruple (Ψ, 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*

$$\inf_{\|d\|\le 1}(q+h)'(\bar{x};d) = -\inf_{u\in Q(\bar{x})}\|u\|, \tag{C.2}$$

where $Q(\bar{x}) := \partial h(\bar{x}) + \bigcup_{y\in Y(\bar{x})}$. Moreover, if $\partial h(\bar{x})$ is nonempty, then the infimum on the right-hand side of (C.2) *is achieved.*

Proof. Let $\bar{x} \in X$, and define

$$\psi(d,y) := (\Psi_y + h)'(\bar{x}; d) \quad \forall (d, x, y) \in \mathcal{X} \times \Omega \times Y. \tag{C.3}$$

We claim that ψ in (C.3) satisfies the assumptions on ψ in Lemma 16 with $Y = Y(\bar{x})$ and D given by

$$D := \{d \in \mathcal{Z} : \|d\| \le 1, d \in F_X(\bar{x})\},$$

where $F_X(\bar{x}) := \{t(x - \bar{x}) : x \in X, t \ge 0\}$ is the set of feasible directions at $\bar{x}$. Before showing this claim, we use it to show that (C.2) holds. First observe that (A1) and Lemma 13 imply that $q'(\bar{x}; d) = \sup_{y\in Y} \Psi'_y(\bar{x}; d)$ for every $d \in \mathcal{X}$. Using then Lemma 16 with $Y = Y(\bar{x})$, Lemma 15 with $(f, x) = (\Psi_{\bar{y}}, \bar{x})$ for every $\bar{y} \in Y(\bar{x})$, and the previous observation, we have that

$$\begin{aligned}
\inf_{\|d\|\le 1}(q+h)'(\bar{x};d) &= \inf_{d\in D}(q+h)'(\bar{x};d) = \inf_{d\in D}\sup_{y\in Y(\bar{x})}(\Psi_y+h)'(\bar{x};d)\\
&= \inf_{d\in D}\sup_{y\in Y(\bar{x})}\psi(d,y) = \sup_{y\in Y(\bar{x})}\inf_{d\in D}\psi(d,y) = \sup_{y\in Y(\bar{x})}\inf_{\|d\|\le 1}(\Psi_y+h)'(\bar{x};d)\\
&= \sup_{y\in Y(\bar{x})}\left[-\inf_{u\in\nabla_x\Phi(\bar{x},y)+\partial h(\bar{x})}\|u\|\right] = \left[-\inf_{u\in Q(\bar{x})}\|u\|\right].
\end{aligned} \tag{C.4}$$

Let us now assume that $\partial h(\bar{x})$ is nonempty and, hence, $Q(\bar{x})$ is nonempty as well. Note that continuity of the function $\nabla_x \Psi(\bar{x}, \cdot)$ from assumption (A1) and the compactness of $Y(\bar{x})$ imply that Q is closed. Moreover, since $\|u\| \geq 0$, it holds that any sequence $\{u_k\}_{k\geq 1}$ where $\lim_{k\to\infty} \|u_k\| = \inf_{u\in Q(\bar{x})} \|u\|$ is bounded. Combining the previous two remarks with the Bolzano–Weierstrass theorem, we conclude that $\inf_{u\in Q(\bar{x})} \|u\| = \min_{u\in Q(\bar{x})} \|u\|$, and hence (C.2) holds.

To complete the proof, we now justify the first claim on ψ. First, for any $y \in Y(\bar{x})$, it follows from [27, Theorem 23.1] with $f(\cdot) = \Psi_y(\cdot)$ and the definitions of q and $Y(\bar{x})$ that

$$\psi(d, \bar{y}) = \Psi'_{\bar{y}}(\bar{x}; d) = \inf_{t>0} \frac{\Psi_y(\bar{x} + td) - q(\bar{x})}{t} \quad \forall d \in \mathcal{X}. \tag{C.5}$$

Since assumption (A2) implies that $\Psi(\bar{x}, \cdot)$ is upper semicontinuous and concave on Y, it follows from (C.5), [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.5), the fact that $\bar{x} \in \operatorname{int} \Omega$, 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 [12, Theorem 4.2.1].

Proposition 18. *Let a proper closed function $\phi : \mathcal{X} \mapsto (-\infty, \infty]$, and assume that $\phi + \|\cdot\|^2/2\lambda$ is μ-strongly convex for some scalars $\mu, \lambda > 0$. If a quadruple $(x^-, x, u, \varepsilon) \in \mathcal{X} \times \operatorname{dom} \phi \times \mathcal{X} \times \mathbb{R}_+$ together with λ satisfy the inclusion $u \in \partial_\varepsilon \left(\phi + \|\cdot - x^-\|^2/[2\lambda]\right)(x)$, then the point $\hat{x} \in \operatorname{dom} \phi$ given by*

$$\hat{x} := \operatorname*{argmin}_{x'} \left\{ \phi_\lambda(x') := \phi(x') + \frac{1}{2\lambda}\|x' - x^-\|^2 - \langle u, x' \rangle \right\} \tag{C.6}$$

satisfies

$$\inf_{\|d\|\leq 1} \phi'(\hat{x}; d) \geq -\frac{1}{\lambda}\|x^- - x + \lambda u\| - \sqrt{\frac{2\varepsilon}{\lambda^2 \mu}}, \quad \|\hat{x} - x\| \leq \sqrt{\frac{2\varepsilon}{\mu}}. \tag{C.7}$$

Proof. We first observe that the assumed inclusion implies that $\phi_\lambda(x') \geq \phi_\lambda(x) - \varepsilon$ for every $x' \in X$. Using the previous inequality at $x' = \hat{x}$, the optimality of $\hat{x}$, and the μ-strong convexity of ϕ_λ, we have that $\mu\|\hat{x} - x\|^2/2 \leq \phi_\lambda(x) - \phi_\lambda(\hat{x}) \leq \varepsilon$ from which we conclude that $\|\hat{x} - x\| \leq \sqrt{2\varepsilon/\mu}$, i.e., the second inequality in (C.7).

To show the other inequality, let $n_\lambda := x^- - x + \lambda u$. Using the definition of ϕ_λ, the triangle inequality, and the previous bound on $\|\hat{x} - x\|$, we obtain

$$\begin{aligned} 0 &\leq \inf_{\|d\|\leq 1} \phi'_\lambda(\hat{x}; d) = \inf_{\|d\|\leq 1} \phi'(\hat{x}; d) - \frac{1}{\lambda}\langle d, n_\lambda \rangle \\ &\leq \inf_{\|d\|\leq 1} \phi'(\hat{x}; d) + \frac{\|n_\lambda\|}{\lambda} + \frac{\|x - \hat{x}\|}{\lambda} \leq \inf_{\|d\|\leq 1} \phi'(\hat{x}; d) + \frac{\|n_\lambda\|}{\lambda} + \sqrt{\frac{2\varepsilon}{\lambda^2 \mu}}, \end{aligned} \tag{C.8}$$

which clearly implies the first inequality in (C.7). ◻

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 19. *Suppose the quadruple* (Φ, h, X, Y) *satisfies assumptions* (A0)–(A3) *of subsection* 2.1*, and let* $(\bar{x}, \bar{u}, \bar{v}) \in X \times \mathcal{X} \times \mathcal{Y}$ *be given. Then, there exists* $\bar{y} \in Y$ *such that the quadruple* $(\bar{u}, \bar{v}, \bar{x}, \bar{y})$ *satisfies the inclusion in* (1.4) *if and only if*

$$\inf_{\|d\|\le 1}(p_{\bar{u},\bar{v}} + h)'(\bar{x}; d) \ge 0, \tag{D.1}$$

where $p_{\bar{u},\bar{v}} := \max_{y\in Y}[\Phi(x,y) + \langle \bar{v}, y\rangle - \langle \bar{u}, x\rangle]$ *for every* $x \in \Omega$.

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

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

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 Ψ satisfies the assumptions on Ψ in Proposition 17, and $\bar{x}$ satisfies (D.1) if and only if $\inf_{\|d\|\le 1}(q+h)'(\bar{x}; d) \ge 0$. The desired conclusion follows from Proposition 17, the previous observation, and the fact that $\bar{y} \in Y(\bar{x})$ if and only if $\bar{v} \in \partial[-\Phi(\bar{x}, \cdot)](\bar{y})$. ◻

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 (ρ_x, ρ_y)-primal-dual stationary point of (1.1). Moreover, let Ψ, q, and D_y be as in (D.2), (B.1), and (2.8), respectively, and define

$$\hat{q}(x) := q(x) + h(x) \quad \forall x \in X.$$

Using Lemma 19, we first observe that $\inf_{\|d\|\le 1} \hat{q}(\bar{x}; d) \ge 0$. Since $\hat{q}$ is convex from assumption (A3), it follows from the previous bound and Lemma 15 with $(f, h) = (0, \hat{q})$ that $\min_{u\in\partial\hat{q}(\bar{x})}\|u\| \le 0$ 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 Ψ, it follows that for every $x \in \mathcal{X}$,

$$\hat{p}(x) + D_y\rho_y - \langle \bar{u}, x\rangle + m\|x - \bar{x}\|^2 \ge \hat{q}(x) \ge \hat{q}(\bar{x}) \ge \hat{p}(\bar{x}) - D_y\rho_y - \langle \bar{u}, \bar{x}\rangle$$

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 18 with $(\phi, x, x^-, u) = (\hat{p}, \bar{x}, \bar{x}, \bar{u})$ and also $(\varepsilon, \lambda, \mu) = (D_y\rho_y, 1/(2m), m)$, we conclude that there exists $\hat{x}$ such that $\|\hat{x} - \bar{x}\| \le \sqrt{2D_y\rho_y/m}$ and

$$\inf_{\|d\|\le 1} \hat{p}'(\hat{x}; d) \ge -\|\bar{u}\| - 2\sqrt{2mD_y\rho_y} \ge -\rho_x - 2\sqrt{2mD_y\rho_y}.$$

◻

We next give the proof of Proposition 2.

Proof of Proposition 2. (a) We first claim that $\hat{P}_\lambda$ is α-strongly convex, where $\alpha = 1/\lambda - m$. To see this, note that $\Phi(\cdot, y) + m\|\cdot\|^2/2$ is convex for every $y \in Y$ from (A3). The claim now follows from (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, δ) 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 (λ, ε)-prox stationary point. Suppose now that $\hat{x} \ne x_\lambda$. Using the convexity of $\hat{P}_\lambda$, we first have that $\hat{P}'_\lambda(\hat{x}; d) = \inf_{t>0}[\hat{P}_\lambda(\hat{x} + td) - \hat{P}_\lambda(\hat{x})]/t$ for every $d \in \mathcal{X}$. Denoting $n_\lambda := (x_\lambda - \hat{x})/\|x_\lambda - \hat{x}\|$, using both inequalities in (1.5) and the previous identity, we then have that

$$\frac{\hat{P}_\lambda(x_\lambda) - \hat{P}_\lambda(\hat{x})}{\|x_\lambda - \hat{x}\|} \ge \hat{p}'(\hat{x}; n_\lambda) + \left\langle \frac{n_\lambda}{\lambda}, \hat{x} - x \right\rangle \ge -\delta - \frac{\|\hat{x} - x\|}{\lambda} \ge -\delta\left(\frac{1+\lambda}{\lambda}\right).$$

Using the optimality of x_λ, the α-strong convexity of $\hat{P}_\lambda$ (see our claim on $\hat{p}$ in the first paragraph), and the above bound, we conclude that

$$\frac{1}{2\alpha}\|\hat{x}-x_\lambda\|^2 \le \hat{P}_\lambda(\hat{x}) - \hat{P}_\lambda(x_\lambda) \le \delta\left(\frac{1+\lambda}{\lambda}\right)\|\hat{x}-x_\lambda\|.$$

Thus, $\|\hat{x}-x_\lambda\| \le 2\alpha\delta(1+\lambda)/\lambda$. Using the previous bound, the second inequality in (1.5), and (2.10) yields

$$\|x-x_\lambda\| \le \|x-\hat{x}\| + \|\hat{x}-x_\lambda\| \le \left(1+2\alpha\left[\frac{1+\lambda}{\lambda}\right]\right)\delta \le \lambda\varepsilon,$$

which implies (2.9) and, hence, that x is a (λ,ε)-prox stationary point.

(b) Suppose that the point x is a (λ,ε)-prox stationary point with $\varepsilon \le \delta \cdot \min\{1, 1/\lambda\}$. Then the optimality of x_λ, 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

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

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

Finally, we give the proof of Proposition 3.

Proof of Proposition 3. This follows by using Lemma 15 with $(f,h) = (\Phi(\cdot,\bar{y}), h)$ and $(f,h) = (0, -\Phi(\bar{x},\cdot))$. ∎

REFERENCES

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

[2] A. Beck, *First-order methods in optimization*, SIAM, 2017.

[3] J. V. Burke and J. J. Moré, *On the identification of active constraints*, SIAM J. Numer. Anal., 25 (1988), pp. 1197–1211.

[4] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, *Accelerated methods for nonconvex optimization*, SIAM J. Optim., 28 (2018), pp. 1751–1772.

[5] D. Davis and D. Drusvyatskiy, *Stochastic model-based minimization of weakly convex functions*, SIAM J. Optim., 29 (2019), pp. 207–239.

[6] D. Davis, D. Drusvyatskiy, K. J. MacPhee, and C. Paquette, *Subgradient methods for sharp weakly convex functions*, J. Optim. Theory Appl., 179 (2018), pp. 962–982.

[7] D. Drusvyatskiy and C. Paquette, *Efficiency of minimizing compositions of convex functions and smooth maps*, Math. Program., 178 (2019), pp. 503–558.

[8] J. C. Duchi and F. Ruan, *Stochastic methods for composite and weakly convex optimization problems*, SIAM J. Optim., 28 (2018), pp. 3229–3259.

[9] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer Science & Business Media, New York, 2007.

[10] S. Ghadimi and G. Lan, *Accelerated gradient methods for nonconvex nonlinear and stochastic programming*, Math. Program., 156 (2016), pp. 59–99.

[11] Y. He and R. D. C. Monteiro, *An accelerated HPE-type algorithm for a class of composite convex-concave saddle-point problems*, SIAM J. Optim., 26 (2016), pp. 29–56.

[12] J. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms* II, Springer, Berlin, 1993.

[13] O. Kolossoski and R. D. C. Monteiro, *An accelerated non-Euclidean hybrid proximal extragradient-type algorithm for convex-concave saddle-point problems*, Optim. Methods Softw., 32 (2017), pp. 1244–1272.

[14] W. Kong, *Accelerated Inexact First-Order Methods for Solving Nonconvex Composite Optimization Problems*, arXiv preprint, arXiv:2104.09685 [math.OC], 2021, https://arxiv.org/abs/2104.09685.

[15] W. Kong, J. G. Melo, and R. D. C. Monteiro, *Complexity of a quadratic penalty accelerated inexact proximal point method for solving linearly constrained nonconvex composite programs*, SIAM J. Optim., 29 (2019), pp. 2566–2593.

[16] W. Kong, J. G. Melo, and R. D. C. Monteiro, *An efficient adaptive accelerated inexact proximal point method for solving linearly constrained nonconvex composite problems*, Comput. Optim. Appl., 76 (2020), pp. 305–346.

[17] T. Lin, C. Jin, and M. Jordan, *Near-Optimal Algorithms for Minimax Optimization*, arXiv preprint, arXiv:2002.02417 [math.OC], 2020, https://arxiv.org/abs/2002.02417.

[18] S. Lu, I. Tsaknakis, and M. Hong, *Block alternating optimization for non-convex min-max problems: Algorithms and applications in signal processing and communications*, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 4754–4758.

[19] Z.-Q. Luo, J.-S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, UK, 1996.

[20] R. D. C. Monteiro and B. F. Svaiter, *Convergence Rate of Inexact Proximal Point Methods with Relative Error Criteria for Convex Optimization*, preprint, Optimization Online, 2010.

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

[22] A. Nemirovski and D. Yudin, *Cesari convergence of the gradient method of approximating saddle points of convex-concave functions*, Dokl. Akad. Nauk, 239 (1978), pp. 1056–1059.

[23] Y. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152.

[24] M. Nouiehed, M. Sanjabi, T. Huang, J. Lee, and M. Razaviyayn, *Solving a class of nonconvex min-max games using iterative first order methods*, in Advances in Neural Information Processing Systems, 2019, pp. 14905–14916.

[25] D. M. Ostrovskii, A. Lowy, and M. Razaviyayn, *Efficient Search of First-Order Nash Equilibria in Nonconvex-Concave Smooth Min-Max Problems*, arXiv preprint, arXiv:2002.07919 [math.OC], 2020, https://arxiv.org/abs/2002.07919.

[26] H. Rafique, M. Liu, Q. Lin, and T. Yang, *Weakly-convex-concave min-max optimization: Provable algorithms and applications in machine learning*, Optim. Methods Sofw., in press.

[27] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[28] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer Science & Business Media, New York, 2009.

[29] M. Sion, *On general minimax theorems.*, Pacific J. Math., 8 (1958), pp. 171–176.

[30] K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh, *Efficient algorithms for smooth minimax optimization*, in Advances in Neural Information Processing Systems, 2019, pp. 12680–12691.