AN INEXACT PRIMAL-DUAL SMOOTHING FRAMEWORK FOR LARGE-SCALE NON-BILINEAR SADDLE POINT PROBLEMS*

LE THI KHANH HIEN†, RENBO ZHAO‡, AND WILLIAM B. HASKELL§

Abstract. We develop an inexact primal-dual first-order smoothing framework to solve a class of non-bilinear saddle point problems with primal strong convexity. Compared with existing methods, our framework yields a significant improvement over the primal oracle complexity, while it has competitive dual oracle complexity. In addition, we consider the situation where the primal-dual coupling term has a large number of component functions. To efficiently handle this situation, we develop a randomized version of our smoothing framework, which allows the primal and dual subproblems in each iteration to be solved by randomized algorithms inexactly in expectation. The convergence of this framework is analyzed both in expectation and with high probability. In terms of the primal and dual oracle complexities, this framework significantly improves over its deterministic counterpart. As an important application, we adapt both frameworks for solving convex optimization problems with many functional constraints. To obtain an ε-optimal and ε-feasible solution, both frameworks achieve the best-known oracle complexities (in terms of their dependence on ε).

Key words. Non-bilinear saddle point problems, Inexact primal-dual smoothing, Convex optimization with functional constraints, Stochastic optimization, Large-scale optimization

1. Introduction.

Let $(\mathbb{E}_1, \|\cdot\|_{\mathbb{E}_1})$ and $(\mathbb{E}_2, \|\cdot\|_{\mathbb{E}_2})$ be finite-dimensional real normed spaces, with dual spaces $(\mathbb{E}_1^*, \|\cdot\|_{\mathbb{E}_1^*})$ and $(\mathbb{E}_2^*, \|\cdot\|_{\mathbb{E}_2^*})$, respectively. We consider the following convex-concave saddle point problem (SPP)

$$\min_{x\in\mathcal{X}} \max_{\lambda\in\Lambda} \big\{S(x,\lambda) \triangleq f(x) + g(x) + \Phi(x,\lambda) - h(\lambda)\big\}, \tag{1.1}$$

where $\mathcal{X} \subseteq \mathbb{E}_1$ and $\Lambda \subseteq \mathbb{E}_2$ are nonempty, convex and closed sets, $\mathcal{X}$ is bounded and the functions $f, g : \mathbb{E}_1 \to \overline{\mathbb{R}} \triangleq \mathbb{R} \cup \{+\infty\}$ and $h : \mathbb{E}_2 \to \overline{\mathbb{R}}$ are convex, closed and proper (CCP) functions. Define $\mathsf{dom}\, g \triangleq \{x \in \mathbb{E}_1 : g(x) < +\infty\}$ and $\mathsf{dom}\, h$ similarly. We assume that $\mathcal{X} \subseteq \mathsf{dom}\, g$ and $\Lambda \subseteq \mathsf{dom}\, h$. (Otherwise, we can take $\mathcal{X} \cap \mathsf{dom}\, g$ and $\Lambda \cap \mathsf{dom}\, h$ to be the new constraint sets that satisfy the above assumptions on $\mathcal{X}$ and Λ.) We also assume that both g and h admit tractable Bregman proximal projections (BPP) on $\mathcal{X}$ and Λ, respectively. (See Section 2 for its precise definition.) In addition, we assume that f is differentiable on an open set $\mathcal{X}' \supseteq \mathcal{X}$, and μ-strongly convex (s.c.) and L-smooth on $\mathcal{X}$ (where $L \geq \mu > 0$), i.e.,

$$\frac{\mu}{2}\|x - x'\|_{\mathbb{E}_1}^2 \leq f(x) - f(x') - \langle \nabla f(x'), x - x'\rangle \leq \frac{L}{2}\|x - x'\|_{\mathbb{E}_1}^2, \quad \forall\, x, x' \in \mathcal{X}, \tag{1.2}$$

where $\langle\cdot,\cdot\rangle$ denotes the duality pairing between $\mathbb{E}_1^*$ (resp. $\mathbb{E}_2^*$) and $\mathbb{E}_1$ (resp. $\mathbb{E}_2$).

We next state our assumptions on the function $\Phi : \mathbb{E}_1 \times \mathbb{E}_2 \to [-\infty, +\infty]$. First, it is convex-concave, i.e., for any $(x,\lambda) \in \mathbb{E}_1 \times \mathbb{E}_2$, $\Phi(\cdot,\lambda)$ is convex and differentiable on $\mathbb{E}_1$ and $\Phi(x,\cdot)$ is concave and differentiable on $\mathbb{E}_2$. In addition, Φ satisfies the $(L_{xx}, L_{\lambda x}, L_{\lambda\lambda})$-smoothness condition (where $L_{xx}, L_{\lambda x}, L_{\lambda\lambda} \geq 0$) on $\mathcal{X} \times \Lambda$, i.e., for

---

* L. T. K. Hien and R. Zhao contribute equally to this work.

Funding: This work was funded by A*STAR Project Number 1421200078 "ABCD: Analyzing Big Corrupted Data".

†Department of Mathematics and Operations Research, University of Mons, Belgium (thikhanhhien.le@umons.ac.be).

‡Operations Research Center, Massachusetts Institute of Technology, USA (renboz@mit.edu).

§Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore (isehwb@nus.edu.sg).

any $x, x' \in \mathcal{X}$ and $\lambda, \lambda' \in \Lambda$,

$$\|\nabla_x \Phi(x,\lambda) - \nabla_x \Phi(x',\lambda)\|_{\mathbb{E}_1^*} \leq L_{xx} \|x - x'\|_{\mathbb{E}_1}, \tag{1.3a}$$

$$\|\nabla_x \Phi(x,\lambda) - \nabla_x \Phi(x,\lambda')\|_{\mathbb{E}_1^*} \leq L_{\lambda x} \|\lambda - \lambda'\|_{\mathbb{E}_2}, \tag{1.3b}$$

$$\|\nabla_\lambda \Phi(x,\lambda) - \nabla_\lambda \Phi(x',\lambda)\|_{\mathbb{E}_2^*} \leq L_{\lambda x} \|x - x'\|_{\mathbb{E}_1}, \tag{1.3c}$$

$$\|\nabla_\lambda \Phi(x,\lambda) - \nabla_\lambda \Phi(x,\lambda')\|_{\mathbb{E}_2^*} \leq L_{\lambda\lambda} \|\lambda - \lambda'\|_{\mathbb{E}_2}, \tag{1.3d}$$

where $x \mapsto \nabla_x \Phi(x,\lambda)$ and $\lambda \mapsto \nabla_\lambda \Phi(x,\lambda)$ denote the gradients of $\Phi(\cdot,\lambda)$ and $\Phi(x,\cdot)$, respectively. For later use, let us define the (primal) condition number

$$\kappa_{\mathcal{X}} \triangleq (L + L_{xx})/\mu. \tag{1.4}$$

In this work, we assume that the saddle function $\Phi(\cdot,\cdot)$ in (1.1) has the following finite-sum structure, i.e.,

$$\Phi(x,\lambda) \triangleq (1/n)\textstyle\sum_{i=1}^n \Phi_i(x,\lambda), \tag{1.5}$$

where for each $i \in [n] \triangleq \{1,\ldots,n\}$ and any $(x,y) \in \mathbb{E}_1 \times \mathbb{E}_2$, $\Phi_i : \mathbb{E}_1 \times \mathbb{E}_2 \to [-\infty,+\infty]$ is convex-concave and satisfies the $(L_{xx}^i, L_{\lambda x}^i, L_{\lambda\lambda}^i)$-smoothness condition on $\mathcal{X} \times \Lambda$. As a result, the smoothness parameters of Φ can be bounded as $L_{xx} \leq (1/n)\sum_{i=1}^n L_{xx}^i$, $L_{\lambda x} \leq (1/n)\sum_{i=1}^n L_{\lambda x}^i$ and $L_{\lambda\lambda} \leq (1/n)\sum_{i=1}^n L_{\lambda\lambda}^i$. In addition, we are particularly interested in the setting where the number of component functions (i.e., n) is *large*.

For well-posedness, we assume that for the SPP in (1.1), at least one saddle point (x^*,λ^*) exists, i.e., there exists $(x^*,\lambda^*) \in \mathcal{X} \times \Lambda$ such that

$$S(x^*,\lambda) \leq S(x^*,\lambda^*) \leq S(x,\lambda^*), \quad \forall\, (x,\lambda) \in \mathcal{X} \times \Lambda. \tag{1.6}$$

1.1. Primal and Dual First-order Oracles. Since we are interested in developing primal-dual first-order methods to solve the SPP in (1.1), where the function $\Phi(\cdot,\cdot)$ has the finite-sum structure as in (1.5), we set up the primal and dual first-order oracles as follows: Upon receiving $(x,\lambda,i) \in \mathcal{X} \times \Lambda \times [n]$ (where $[n] \triangleq \{1,\ldots,n\}$), the primal oracle $\mathscr{O}^{\mathrm{P}}$ returns $\nabla_x \Phi_i(x,\lambda)$ and the dual oracle $\mathscr{O}^{\mathrm{D}}$ returns $\nabla_\lambda \Phi_i(x,\lambda)$. In addition, $\mathscr{O}^{\mathrm{P}}$ returns $\nabla f(x)$ upon receiving $(x,0)$. Accordingly, we define the *primal oracle complexity* and *dual oracle complexity* to be the number of oracle calls to $\mathscr{O}^{\mathrm{P}}$ and $\mathscr{O}^{\mathrm{D}}$, respectively.

1.2. Applications. The SPP in (1.1), where $\Phi(\cdot,\cdot)$ has the finite-sum structure as in (1.5), has applications in numerous fields, including game theory, image processing, machine learning and statistics. In Section 5, we will focus on one of the most important ones, i.e., convex optimization problems with functional constraints (and their stochastic extension). Apart from this, we also illustrate another important application below. For more applications, we refer readers to [12, 9, 3, 11].

Maximum Margin Clustering [30]. Let $\mathcal{D}$ denote a set of m objects, which consists of two unknown disjoint subsets, denoted by $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively (i.e., $\mathcal{D}_1 \cap \mathcal{D}_2 = \emptyset$ and $\mathcal{D}_1 \cup \mathcal{D}_2 = \mathcal{D}$). Given n noisy samples $\{\mathcal{S}_i\}_{i=1}^n$ of $\mathcal{D}$ (where each $\mathcal{S}_i$ consists of m data points), we wish to find a label kernel matrix $M \in \mathbb{S}_+^m$ that can assign each object in $\mathcal{D}$ to $\mathcal{D}_1$ or $\mathcal{D}_2$ (where $\mathbb{S}_+^m$ denotes the positive semi-definite cone of dimension m). To do so, we first compute a kernel matrix $K_i \in \mathbb{S}_+^m$ from each $\mathcal{S}_i$ ($i \in [n]$). Define $e \triangleq (1,1,\ldots,1) \in \mathbb{R}^m$, we then solve the following SPP:

$$\min_{M \in \mathcal{M}} \max_{\lambda_i \in \Lambda,\, \forall\, i \in [n]} (1/n)\textstyle\sum_{i=1}^n - \langle \lambda_i \lambda_i^T, K_i \circ M + \alpha I \rangle + 2\lambda_i^T e,$$

TABLE 1
Comparison of primal and dual oracle complexities with existing methods.

Algorithms	Primal Oracle Comp.	Dual Oracle Comp.
PDHG-type [11]	$O(n/\varepsilon)$	$O(n/\varepsilon)$
Mirror-Prox [20]	$O(n/\varepsilon)$	$O(n/\varepsilon)$
Det. IPDS (Algo. 1)	$\widetilde{O}(n\sqrt{\kappa_{\mathcal{X}}/\varepsilon})$	$\widetilde{O}(n/\varepsilon)$
Rand. IPDS (Algo. 2)[1]	$\widetilde{O}((n+\sqrt{n\kappa_{\mathcal{X}}})/\sqrt{\varepsilon})$	$\widetilde{O}(n/\sqrt{\varepsilon}+\sqrt{n}/\varepsilon)$

[1] Both primal and dual oracle complexities of Rand. IPDS correspond to obtaining *expected* duality gap (cf. Theorem 4.2).

where $\Lambda \triangleq \{\lambda \in \mathbb{R}^m : 0 \le \lambda_i \le C < +\infty, \forall\, i \in [m]\}$, $\mathcal{M} \triangleq \{M \in \mathbb{S}^m_+ : \mathrm{diag}(M) = e, |e^T M| \le \ell < +\infty\}$ and $\alpha > 0$. In addition, $\circ$ denotes the Hadamard product and $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product. We can easily see that (1.2) is a convex-concave SPP with finite-sum structure. Moreover, both constraint sets $\mathcal{M}$ and Λ^n are nonempty, closed and convex. In addition, to achieve high clustering accuracy, the required number of samples n may potentially be large.

1.3. Related Works. We focus on reviewing the works on solving (convex-concave) *non-bilinear non-smooth* SPPs, where the primal-dual coupling term $\Phi(\cdot,\cdot)$ is not bilinear. (Note that $\Phi(\cdot,\cdot)$ is bilinear if $\Phi(x,y) = \langle \mathsf{A}x, y\rangle$, where $\mathsf{A} : \mathbb{E}_1 \to \mathbb{E}_2^*$ is a (bounded) linear operator.) In the past few years, bilinear SPPs have been extensively studied, with many efficient algorithms proposed, e.g., Nesterov smoothing [22] and primal-dual hybrid gradient (PDHG) [7]. For details, see [8, 32].

For non-bilinear non-smooth SPPs without composite structure, i.e., the saddle function $S(\cdot,\cdot)$ cannot be decomposed into a smooth term $\Phi(\cdot,\cdot)$ and non-smooth functions g and h with tractable BPPs, the existing algorithms are mainly based on primal-dual subgradient (e.g., [19, 23]). To reach an ε-duality gap (defined in (2.13)), these works achieve the optimal oracle complexity $O(1/\varepsilon^2)$.

However, if the non-smooth SPPs do possess composite structure, e.g., the one in (1.1), then the algorithms for smooth SPPs can be extended to this case and their oracle complexities can be much better than $O(1/\varepsilon^2)$. For example, the Mirror-Prox method [20] was extended in [12] to solve (1.1) with an oracle complexity of $O(1/\varepsilon)$. As another example, Nesterov smoothing was extended in [14] to solve a special case of (1.1), where $L_{\lambda\lambda} = 0$ (i.e., $\Phi(x,\cdot)$ is linear). Recently, the PDHG method, originally developed for the bilinear SPPs, was extended in [11] to solve (1.1). In addition, the stochastic (and accelerated) versions of the aforementioned methods have also been developed in [13, 9, 32] to tackle the situation where only stochastic first-order oracles are available. (See Section 4.5 for details.) However, most of these methods focus on the case where $\mu = 0$, so the favorable condition $\mu > 0$ may not be exploited (to further improve the oracle complexity). The only two works that have utilized this condition are [12, 11], which are based on the Mirror-Prox and PDHG methods, respectively. However, both works only consider the special case where $L_{\lambda\lambda} = 0$. Therefore, in this work, we aim to answer the following question:

Can we develop an algorithmic framework that works for both $\mu > 0$ and $L_{\lambda\lambda} \ge 0$, yet with improved oracle complexity over the existing methods?

1.4. Main Contributions. We make the following three main contributions.

First, we develop a novel (deterministic) inexact primal-dual smoothing (IPDS)

framework (i.e., Algorithm 1) for solving the non-bilinear SPP in (1.1) with primal strong convexity (i.e., $\mu > 0$). To the best of our knowledge, this is the first inexact smoothing framework developed for such a problem. To reach an ε-duality gap (defined in (2.13)), the primal and dual oracle complexities are $\widetilde{O}(n\sqrt{\kappa_{\mathcal{X}}/\varepsilon})$ and $\widetilde{O}(n/\varepsilon)$, respectively (where $\widetilde{O}(\cdot)$ hides the $\log n$ and $\log(1/\varepsilon)$ factors). Compared with existing works (cf. Table 1), the primal oracle complexity of our framework is significantly better, while the dual oracle complexity is competitive. In addition, in contrast to the methods in [12, 11], which can *only* improve the primal oracle complexity $O(n/\varepsilon)$ (to e.g., $O(n\sqrt{\kappa_{\mathcal{X}}/\varepsilon})$) when $L_{\lambda\lambda} = 0$, our framework applies to any value $L_{\lambda\lambda} \geq 0$.

Second, we develop a randomized version of our IPDS framework (i.e., Algorithm 2), by allowing each sub-problem to be solved inexactly *in expectation*. This framework is particularly useful in the regime where n is large in (1.5). Indeed, to reach an ε-expected duality gap, the primal and dual oracle complexities are $\widetilde{O}((n + \sqrt{n\kappa_{\mathcal{X}}})/\sqrt{\varepsilon})$ and $\widetilde{O}(n/\sqrt{\varepsilon} + \sqrt{n}/\varepsilon)$, respectively, which significantly improve over those of Algorithm 1. In addition, we show that Algorithm 2 also converges *with high probability*. We believe that the techniques used in our stochastic analysis are of independent interest, as they can be applied to other inexact frameworks (e.g., the inexact augmented Lagrangian method (ALM) [29, 25, 31]).

Finally, we apply both of our aforementioned frameworks (i.e., Algorithms 1 and 2) to convex optimization problems with (potentially many) functional constraints (more precisely, to their associated Lagrangian problems). To do so, we manage to overcome two challenges: the unboundedness of the constraint set Λ and the dependence of the smoothness parameter L_{xx} on the dual variable λ. (For details, see Section 5.3.) To obtain an ε-optimal and ε-feasible solution (cf. (5.6)), both Algorithms 1 and 2 achieve the state-of-the-art (primal) oracle complexities $\widetilde{O}(1/\sqrt{\varepsilon})$. (Note that the dual oracle complexity is $O(1)$ in this case; see Section 5.1.) Compared to other first-order methods with similar oracle complexities but *specifically* designed for constrained convex optimization problems, our frameworks enjoy much wider applicability.

2. Preliminaries.

We first define the primal and dual functions and the duality gap associated with the saddle function $S(\cdot,\cdot)$, their smoothed versions, and the inexact solutions of the optimization problems appearing in these definitions. We then prove some smoothness properties of the formerly defined quantities.

2.1. Definitions.

First, we define the primal function $\psi^{\mathrm{P}} : \mathbb{E}_1 \to \overline{\mathbb{R}}$ and the dual function $\psi^{\mathrm{D}} : \mathbb{E}_2 \to \overline{\mathbb{R}}$ associated with $S(\cdot,\cdot)$ as

$$\psi^{\mathrm{P}}(x) \triangleq \sup_{\lambda\in\Lambda} S(x,\lambda) = f(x) + g(x) + \widehat{\psi}^{\mathrm{P}}(x), \qquad \forall\, x \in \mathbb{E}_1, \tag{2.1}$$

$$\psi^{\mathrm{D}}(\lambda) \triangleq \inf_{x\in\mathcal{X}} S(x,\lambda) = \widehat{\psi}^{\mathrm{D}}(\lambda) - h(\lambda), \qquad \forall\, \lambda \in \mathbb{E}_2, \tag{2.2}$$

where

$$\widehat{\psi}^{\mathrm{P}}(x) \triangleq \sup_{\lambda\in\Lambda}\{\widehat{S}^{\mathrm{D}}(x,\lambda) \triangleq \Phi(x,\lambda) - h(\lambda)\}, \qquad \forall\, x \in \mathbb{E}_1, \tag{2.3}$$

$$\widehat{\psi}^{\mathrm{D}}(\lambda) \triangleq \inf_{x\in\mathcal{X}}\{\widehat{S}^{\mathrm{P}}(x,\lambda) \triangleq f(x) + g(x) + \Phi(x,\lambda)\}, \qquad \forall\, \lambda \in \mathbb{E}_2. \tag{2.4}$$

Let $\omega : \mathbb{E}_2 \to \overline{\mathbb{R}}$ be a CCP function that is 1-s.c. and continuous on Λ and essentially smooth, i.e., ω is continuously differentiable on $\mathsf{int}\,\mathsf{dom}\,\omega \neq \emptyset$, and $\|\nabla w(\lambda_k)\|_* \to +\infty$ if $\lambda_k \to \lambda \in \mathsf{bd}\,\mathsf{dom}\,\omega$ [4]. (Note that for any set $\mathcal{K}$, $\mathsf{int}\,\mathcal{K}$ and $\mathsf{bd}\,\mathcal{K}$ denote the interior and boundary of $\mathcal{K}$, respectively.) In addition, for any $\alpha > 0$ and $v \in \mathbb{E}_2^*$, the following

problem (which is defined by the triple (ω, h, Λ))

$$\min_{\lambda\in\Lambda} h(\lambda) + \langle v, \lambda\rangle + \alpha^{-1}\omega(\lambda) \tag{2.5}$$

has a (unique) *easily computable* solution in $\Lambda^o \triangleq \Lambda \cap \mathsf{int}\,\mathsf{dom}\,\omega$. (Note that (2.5) is equivalent to BPP; see [2] for details and examples.) We call ω a *distance generating function* (DGF) w.r.t. (h, Λ). Additionally, we say that h has a tractable BPP on Λ if and only if such an ω exists. Since we also assume that g has a tractable BPP on $\mathcal{X}$ (cf. Section 1), there exists a DGF $\overline{\omega} : \mathbb{E}_1 \to \overline{\mathbb{R}}$ w.r.t. $(g, \mathcal{X})$. Similar to Λ^o, we define $\mathcal{X}^o \triangleq \mathcal{X} \cap \mathsf{int}\,\mathsf{dom}\,\overline{\omega}$. The assumption that (2.5) can be solved easily is typical in the first-order methods for composite optimization on normed spaces (see e.g., [28]). Since we will employ such methods to solve the sub-problems in our framework (cf. Sections 3.2 and 4.1), we also make this assumption throughout the whole work.

Based on ω, we define the dual-regularized saddle function

$$S_\rho(x, \lambda) \triangleq S(x, \lambda) - \rho\omega(\lambda), \tag{2.6}$$

where $\rho > 0$ is the smoothing parameter. Accordingly, the primal function $\psi_\rho^{\mathrm{P}} : \mathbb{E}_1 \to \overline{\mathbb{R}}$ associated with $S_\rho(\cdot,\cdot)$ is

$$\psi_\rho^{\mathrm{P}}(x) \triangleq \sup_{\lambda\in\Lambda} S_\rho(x, \lambda) = f(x) + g(x) + \widehat{\psi}_\rho^{\mathrm{P}}(x), \quad \forall\, x \in \mathbb{E}_1, \tag{2.7}$$

where

$$\widehat{\psi}_\rho^{\mathrm{P}}(x) \triangleq \sup_{\lambda\in\Lambda}\{\widehat{S}_\rho^{\mathrm{D}}(x, \lambda) \triangleq \Phi(x, \lambda) - h(\lambda) - \rho\omega(\lambda)\}, \quad \forall\, x \in \mathbb{E}_1. \tag{2.8}$$

Next, we introduce the optimal solutions of the optimization problems in (2.4) and (2.8). Since f is μ-s.c. on $\mathcal{X}$, the minimization problem in (2.2) has the unique solution

$$x^*(\lambda) \triangleq \arg\min_{x\in\mathcal{X}} \widehat{S}^{\mathrm{P}}(x, \lambda), \quad \forall\, \lambda \in \mathbb{E}_2. \tag{2.9}$$

In addition, for any $\lambda \in \mathbb{E}_2$, we call $\widetilde{x}_\gamma(\lambda) \in \mathcal{X}$ an γ-inexact solution (where $\gamma \geq 0$) if

$$\widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda) - \widehat{\psi}^{\mathrm{D}}(\lambda) = \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda) - \widehat{S}^{\mathrm{P}}(x^*(\lambda), \lambda) \leq \gamma. \tag{2.10}$$

Similar to (2.4), since $\widehat{S}_\rho^{\mathrm{D}}(x, \cdot)$ is ρ-strongly concave on Λ, the maximization problem in (2.8) has the unique solution

$$\lambda_\rho^*(x) \triangleq \arg\max_{\lambda\in\Lambda} \widehat{S}_\rho^{\mathrm{D}}(x, \lambda), \quad \forall\, x \in \mathbb{E}_1. \tag{2.11}$$

Additionally, for any $x \in \mathbb{E}_1$, we call $\widetilde{\lambda}_{\rho,\eta}(x) \in \Lambda$ an η-inexact solution (where $\eta \geq 0$) if

$$\widehat{\psi}_\rho^{\mathrm{P}}(\lambda) - \widehat{S}_\rho^{\mathrm{D}}(x, \widetilde{\lambda}_{\rho,\eta}(x)) = \widehat{S}_\rho^{\mathrm{D}}(x, \lambda_\rho^*(x)) - \widehat{S}_\rho^{\mathrm{D}}(x, \widetilde{\lambda}_{\rho,\eta}(x)) \leq \eta. \tag{2.12}$$

Finally, we define the duality gap $\Delta : \mathbb{E}_1 \times \mathbb{E}_2 \to \overline{\mathbb{R}}$ associated with $S(\cdot,\cdot)$ as

$$\Delta(x, \lambda) \triangleq \psi^{\mathrm{P}}(x) - \psi^{\mathrm{D}}(\lambda), \quad \forall\, (x, \lambda) \in \mathbb{E}_1 \times \mathbb{E}_2. \tag{2.13}$$

Clearly, $(x^*, \lambda^*) \in \mathcal{X} \times \Lambda$ is a saddle point of (1.1) if and only if $\Delta(x^*, \lambda^*) = 0$. As a consequence, for any $\varepsilon > 0$, we call $(\bar{x}, \bar{\lambda}) \in \mathcal{X} \times \Lambda$ an ε-saddle point of (1.1) if $\Delta(\bar{x}, \bar{\lambda}) \leq \varepsilon$. Additionally, we define the (ρ-)smoothed duality gap $\Delta_\rho : \mathbb{E}_1 \times \mathbb{E}_2 \to \overline{\mathbb{R}}$ as

$$\Delta_\rho(x, \lambda) \triangleq \psi_\rho^{\mathrm{P}}(x) - \psi^{\mathrm{D}}(\lambda), \quad \forall\, (x, \lambda) \in \mathbb{E}_1 \times \mathbb{E}_2. \tag{2.14}$$

2.2. Smoothness Properties. We first show that the optimal solution $x^*(\cdot)$ in (2.9) is Lipschitz on $\mathbb{E}_2$ and the function $\widehat{\psi}^{\mathrm{D}}$ in (2.4) is smooth on $\mathbb{E}_2$.

PROPOSITION 2.1. *The function $\widehat{\psi}^{\mathrm{D}}$ is differentiable on $\mathbb{E}_2$ and for any $\lambda \in \mathbb{E}_2$, $\nabla\widehat{\psi}^{\mathrm{D}}(\lambda) = \nabla_\lambda \widehat{S}^{\mathrm{P}}(x^*(\lambda),\lambda) = \nabla_\lambda \Phi(x^*(\lambda),\lambda)$. In addition, $x^*(\cdot)$ is $(2L_{\lambda x}/\mu)$-Lipschitz on $\mathbb{E}_2$ and $\nabla\widehat{\psi}^{\mathrm{D}}$ is L_{D}-Lipschitz on $\mathbb{E}_2$, where*

$$L_{\mathrm{D}} \triangleq L_{\lambda\lambda} + 2L_{\lambda x}^2/\mu. \tag{2.15}$$

Proof. Since for any $(x,\lambda) \in \mathbb{E}_1 \times \mathbb{E}_2$, $\mathcal{X}$ is compact, $\widehat{S}^{\mathrm{P}}(\cdot,\lambda)$ is μ-s.c. on $\mathcal{X}$, and $\widehat{S}^{\mathrm{P}}(x,\cdot)$ is differentiable on $\mathbb{E}_2$, we can invoke Danskin's Theorem [5, Proposition B.25] to conclude that ψ^{D} is differentiable on $\mathbb{E}_2$ and $\nabla\psi^{\mathrm{D}}(\lambda) = \nabla_\lambda \widehat{S}^{\mathrm{P}}(x^*(\lambda),\lambda)$. To show Lipschitz continuity of $x^*(\cdot)$, note that for any $\lambda_1,\lambda_2 \in \mathbb{E}_2$, since $\widehat{S}^{\mathrm{P}}(\cdot,\lambda_1)$ is μ-s.c.,

$$\|x^*(\lambda_2) - x^*(\lambda_1)\|^2 \le \frac{2}{\mu}\big(\widehat{S}^{\mathrm{P}}(x^*(\lambda_2),\lambda_1) - \widehat{S}^{\mathrm{P}}(x^*(\lambda_1),\lambda_1)\big). \tag{2.16}$$

On the other hand,

$$\begin{aligned}
&\widehat{S}^{\mathrm{P}}(x^*(\lambda_2),\lambda_1) - \widehat{S}^{\mathrm{P}}(x^*(\lambda_1),\lambda_1) \\
&\overset{(a)}{\le} \big(\widehat{S}^{\mathrm{P}}(x^*(\lambda_2),\lambda_1) - \widehat{S}^{\mathrm{P}}(x^*(\lambda_1),\lambda_1)\big) - \big(\widehat{S}^{\mathrm{P}}(x^*(\lambda_2),\lambda_2) - \widehat{S}^{\mathrm{P}}(x^*(\lambda_1),\lambda_2)\big) \\
&\overset{(b)}{=} \big(\Phi(x^*(\lambda_2),\lambda_1) - \Phi(x^*(\lambda_1),\lambda_1)\big) - \big(\Phi(x^*(\lambda_2),\lambda_2) - \Phi(x^*(\lambda_1),\lambda_2)\big) \\
&= \int_0^1 \langle \nabla_\lambda \Phi(x^*(\lambda_2),\lambda_2 + t(\lambda_1-\lambda_2)) \\
&\qquad\qquad -\nabla_\lambda \Phi(x^*(\lambda_1),\lambda_2 + t(\lambda_1-\lambda_2)), \lambda_1-\lambda_2\rangle \,\mathrm{d}t \\
&\overset{(c)}{\le} \|\nabla_\lambda \Phi(x^*(\lambda_2),\lambda_2 + t(\lambda_1-\lambda_2)) - \nabla_\lambda \Phi(x^*(\lambda_1),\lambda_2 + t(\lambda_1-\lambda_2))\|_* \|\lambda_1-\lambda_2\| \\
&\overset{(d)}{\le} L_{\lambda x}\|x^*(\lambda_2) - x^*(\lambda_1)\|\|\lambda_1-\lambda_2\|,
\end{aligned} \tag{2.17}$$

where in (a) we use the fact that $\widehat{S}^{\mathrm{P}}(x^*(\lambda_2),\lambda_2) - \widehat{S}^{\mathrm{P}}(x^*(\lambda_1),\lambda_2) \le 0$ since $x^*(\lambda_2)$ is the minimizer of $\widehat{S}^{\mathrm{P}}(\cdot,\lambda_2)$ on $\mathcal{X}$, in (b) we use the definition of $\widehat{S}^{\mathrm{P}}(\cdot,\cdot)$ in (2.9), in (c) we use the definition of the dual norm $\|\cdot\|_*$ and in (d) we use the Lipschitz continuity of $\nabla_\lambda\Phi(\cdot,\lambda)$ in (1.3c). Therefore, by combining (2.16) and (2.17), we have

$$\|x^*(\lambda_1) - x^*(\lambda_2)\| \le \frac{2L_{\lambda x}}{\mu}\|\lambda_1-\lambda_2\|. \tag{2.18}$$

As a result,

$$\begin{aligned}
\big\|\nabla\psi^{\mathrm{D}}(\lambda_1) - \nabla\psi^{\mathrm{D}}(\lambda_2)\big\|_* &= \|\nabla_\lambda\Phi(x^*(\lambda_1),\lambda_1) - \nabla_\lambda\Phi(x^*(\lambda_2),\lambda_2)\|_* \\
&\le \|\nabla_\lambda\Phi(x^*(\lambda_1),\lambda_1) - \nabla_\lambda\Phi(x^*(\lambda_2),\lambda_1)\|_* \\
&\qquad + \|\nabla_\lambda\Phi(x^*(\lambda_2),\lambda_1) - \nabla_\lambda\Phi(x^*(\lambda_2),\lambda_2)\|_* \\
&\le L_{\lambda x}\|x^*(\lambda_1) - x^*(\lambda_2)\| + L_{\lambda\lambda}\|\lambda_1-\lambda_2\| \\
&\le \big(2L_{\lambda x}^2/\mu + L_{\lambda\lambda}\big)\,\|\lambda_1-\lambda_2\|,
\end{aligned}$$

where in the last inequality we use (2.18). ∎

By a symmetric argument, we can also conclude that $\widehat{\psi}_\rho^{\mathrm{P}}$ is differentiable on $\mathbb{E}_1$ and $\nabla\widehat{\psi}_\rho^{\mathrm{P}}$ is $(L_{xx} + 2L_{\lambda x}^2/\rho)$-Lipschitz on $\mathbb{E}_1$. For this reason, we will call $\widehat{\psi}_\rho^{\mathrm{P}}$ the (ρ-)*smoothed primal function* and consequently, Δ_ρ the (ρ-)*smoothed duality gap*.

Based on Proposition 2.1, we can establish the following results involving $\widetilde{x}_\gamma(\lambda)$, i.e., the γ-inexact solution of (2.2).

PROPOSITION 2.2. *For any $\gamma \geq 0$ and $\lambda, \lambda' \in \mathbb{E}_2$, we have*

$$\|\nabla_\lambda \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda) - \nabla \widehat{\psi}^{\mathrm{D}}(\lambda)\|_* \leq L_{\lambda x}\sqrt{2\gamma/\mu}, \tag{2.19}$$

$$0 \leq \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda) - \widehat{\psi}^{\mathrm{D}}(\lambda') + \left\langle \nabla_\lambda \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda), \lambda' - \lambda \right\rangle \leq L_{\mathrm{D}}\|\lambda - \lambda'\|^2 + 2\gamma. \tag{2.20}$$

Proof. First, for any $\lambda \in \mathbb{E}_2$, we note that $\nabla_\lambda \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda) = \nabla_\lambda \Phi(\widetilde{x}_\gamma(\lambda), \lambda)$ and $\nabla \widehat{\psi}^{\mathrm{D}}(\lambda) = \nabla_\lambda \Phi(x^*(\lambda), \lambda)$ (cf. Proposition 2.1). Therefore,

$$\begin{aligned} \|\nabla_\lambda \Phi(\widetilde{x}_\gamma(\lambda), \lambda) - \nabla \psi^{\mathrm{D}}(\lambda)\|_*^2 &= \|\nabla_\lambda \Phi(\widetilde{x}_\gamma(\lambda), \lambda) - \nabla_\lambda \Phi(x^*(\lambda), \lambda)\|_*^2 \\ &\overset{(a)}{\leq} L_{\lambda x}^2 \|\widetilde{x}_\gamma(\lambda) - x^*(\lambda)\|^2 \\ &\overset{(b)}{\leq} (2L_{\lambda x}^2/\mu)\big(\widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda) - \widehat{S}^{\mathrm{P}}(x^*(\lambda), \lambda)\big) \\ &\overset{(c)}{\leq} 2L_{\lambda x}^2 \gamma/\mu, \end{aligned} \tag{2.21}$$

where in (a) we use (1.3c), in (b) we use the μ-strong convexity of $\widehat{S}^{\mathrm{P}}(\cdot, \lambda)$ on $\mathcal{X}$ and in (c) we use the definition of $\widetilde{x}_\gamma(\lambda)$ in (2.10). This proves (2.19).

We next prove (2.20). First, for any $\lambda, \lambda' \in \mathbb{E}_2$,

$$\widehat{\psi}^{\mathrm{D}}(\lambda') \overset{(a)}{\leq} \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda') \overset{(b)}{\leq} \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda) + \langle \nabla_\lambda \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda), \lambda' - \lambda \rangle, \tag{2.22}$$

where (a) follows from (2.4) and (b) follows from the concavity of $\widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \cdot)$ on $\mathbb{E}_2$. This proves the left-hand side (LHS) of (2.20). To show the right-hand side (RHS), we note that $\widehat{\psi}^{\mathrm{D}}$ is concave and L_{D}-smooth on $\mathbb{E}_2$ (cf. Proposition 2.1). Thus we can invoke the descent lemma [5], such that for all $\lambda, \lambda' \in \mathbb{E}_2$,

$$\begin{aligned} \widehat{\psi}^{\mathrm{D}}(\lambda') &\geq \widehat{\psi}^{\mathrm{D}}(\lambda) + \langle \nabla \widehat{\psi}^{\mathrm{D}}(\lambda), \lambda' - \lambda \rangle - (L_{\mathrm{D}}/2)\|\lambda - \lambda'\|^2 \\ &\geq \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda) - \gamma + \langle \nabla_\lambda \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda), \lambda' - \lambda \rangle \\ &\quad + \langle \nabla \psi^{\mathrm{D}}(\lambda) - \nabla_\lambda \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda), \lambda' - \lambda \rangle - (L_{\mathrm{D}}/2)\|\lambda - \lambda'\|^2 \\ &\geq \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda) - \gamma + \langle \nabla_\lambda \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda), \lambda' - \lambda \rangle \\ &\quad - \|\nabla \psi^{\mathrm{D}}(\lambda) - \nabla_\lambda \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda)\|_* \|\lambda - \lambda'\| - (L_{\mathrm{D}}/2)\|\lambda - \lambda'\|^2 \\ &\overset{(a)}{\geq} \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda) - \gamma + \langle \nabla_\lambda \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda), \lambda' - \lambda \rangle \\ &\qquad - L_{\lambda x}\sqrt{2\gamma/\mu}\|\lambda - \lambda'\| - (L_{\mathrm{D}}/2)\|\lambda - \lambda'\|^2, \\ &\overset{(b)}{\geq} \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda) - 2\gamma + \langle \nabla_\lambda \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda), \lambda' - \lambda \rangle - L_{\mathrm{D}}\|\lambda - \lambda'\|^2, \end{aligned} \tag{2.23}$$

where (a) follows from (2.19) and (b) follows from the AM-GM inequality, i.e.,

$$L_{\lambda x}\sqrt{2\gamma/\mu}\|\lambda - \lambda'\| \leq (L_{\lambda x}^2/\mu)\|\lambda - \lambda'\|^2 + \gamma \leq (L_{\mathrm{D}}/2)\|\lambda - \lambda'\|^2 + \gamma. \tag{2.24}$$

We then rearrange (2.23) to obtain the RHS of (2.20). ∎

3. Deterministic IPDS Framework. We develop the IPDS framework based on the idea of *smoothed duality gap reduction*. First, we make an important assumption, followed by a few remarks.

ASSUMPTION 3.1. *The set Λ is bounded.*

Algorithm 1 Deterministic inexact primal-dual smoothing framework

Input: Initial smoothing parameter $\rho_0 > 0$, nonnegative error sequences $\{\eta_k\}_{k\in\mathbb{Z}_+}$ and $\{\gamma_k\}_{k\in\mathbb{Z}_+}$, interpolation sequence $\{\tau_k\}_{k\in\mathbb{Z}_+} \subseteq (0,1)$ and deterministic first-order algorithms N_1 and N_2.

Initialize: $x^0 \in \mathcal{X}$, $\lambda^0 \in \Lambda$ and $k = 0$.

Repeat (until some convergence criterion is met)

1. Use N_1 to find $\widetilde{\lambda}_{\rho_k,\eta_k}(x^k) \in \Lambda$ such that
$$\widehat{\psi}^{\mathrm{P}}_{\rho_k}(x^k) - \widehat{S}^{\mathrm{D}}_{\rho_k}(x^k, \widetilde{\lambda}_{\rho_k,\eta_k}(x^k)) \le \eta_k. \tag{3.3}$$
2. Set $\hat{\lambda}^k = \tau_k\lambda^k + (1-\tau_k)\widetilde{\lambda}_{\rho_k,\eta_k}(x^k)$.
3. Use N_2 to find $\widetilde{x}_{\gamma_k}(\hat{\lambda}^k) \in X$ such that
$$\widehat{S}^{\mathrm{P}}(\widetilde{x}_{\gamma_k}(\hat{\lambda}^k), \hat{\lambda}^k) - \widehat{\psi}^{\mathrm{D}}(\hat{\lambda}^k) \le \gamma_k. \tag{3.4}$$
4. Set $x^{k+1} = \tau_k x^k + (1-\tau_k)\widetilde{x}_{\gamma_k}(\hat{\lambda}^k)$.
5. Set $\rho_{k+1} = \tau_k\rho_k$.
6. Use N_1 to find $\widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1}) \in \Lambda$ such that
$$\widehat{\psi}^{\mathrm{P}}_{\rho_{k+1}}(x^{k+1}) - \widehat{S}^{\mathrm{D}}_{\rho_{k+1}}(x^{k+1}, \widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1})) \le \eta_k. \tag{3.5}$$
7. Set $\lambda^{k+1} = \tau_k\lambda^k + (1-\tau_k)\widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1})$.
8. Set $k = k+1$.

Output: $(x^{\mathrm{out}}, \lambda^{\mathrm{out}}) \triangleq (x^k, \lambda^k)$.

To see the implication of this assumption, for any $(x,\lambda) \in \mathcal{X}\times\Lambda$, we may bound

$$\begin{aligned} |\Delta(x,\lambda) - \Delta_\rho(x,\lambda)| &= |\sup_{\lambda\in\Lambda} S(x,\lambda) - \sup_{\lambda\in\Lambda} S_\rho(x,\lambda)| \\ &\le \sup_{\lambda\in\Lambda} |S(x,\lambda) - S_\rho(x,\lambda)| \\ &= \rho \sup_{\lambda\in\Lambda} |\omega(\lambda)| \\ &< +\infty, \end{aligned} \tag{3.1}$$

where the last inequality follows from Assumption 3.1, the closedness of Λ and the continuity of ω on Λ. For convenience, define

$$B_{\omega,\Lambda} \triangleq \sup_{\lambda\in\Lambda} |\omega(\lambda)|. \tag{3.2}$$

Since ω is also 1-s.c. on Λ, we conclude that $B_{\omega,\Lambda} < +\infty$ if and only if Λ is bounded.

Remark 3.2. We provide a few remarks about Assumption 3.1 (which is equivalent to $B_{\omega,\Lambda} < +\infty$). First, it is important in connecting the smoothed duality gap Δ_ρ to the duality gap Δ. Indeed, in our analysis, we will first analyze the convergence rate of the smoothed duality gap, and then show that the same rate holds for the (original) duality gap if $B_{\omega,\Lambda} < +\infty$. Note that we *do not* need this assumption to derive any convergence results regarding the smoothed duality gap. Finally, we note that Assumption 3.1 also appears in many other algorithms for solving SPPs, e.g., Mirror-Prox [20], HPE-type [14] and PDHG [32], although for different reasons. Indeed, it is typical in the works where the duality gap is used as the convergence criterion, and is not specific to our work.

3.1. Framework Description. The framework is presented in Algorithm 1. We now describe the main ideas behind this framework. From (3.1), we observe that $\Delta(x,\lambda) \le \Delta_\rho(x,\lambda)+\rho B_{\omega,\Lambda}$. Therefore, if there exists a primal-dual pair $(x,\lambda)\in\mathcal{X}\times\Lambda$ such that the smoothed duality gap $\Delta_\rho(x,\lambda)$ is small, then with a small smoothing parameter ρ, the duality gap $\Delta(x,\lambda)$ will also be small. This leads us to develop a framework that "sufficiently" reduces both the smoothed duality gap and smoothing parameter in each iteration. Indeed, in step 5 of Algorithm 1, the smoothing parameter ρ_k is reduced by a factor of $(1-\tau_k)\in(0,1)$. The same factor is also used to interpolate the primal and dual iterates (cf. steps 2, 4 and 7). As a key difference with the smoothing frameworks for bilinear SPPs (e.g., [22, 21]), our framework does not require the sub-problems (2.4) and (2.8) to be solved exactly. Instead, we only need the inexact solutions satisfying certain accuracy criteria (involving parameters γ_k and η_k; cf. (3.3), (3.4) and (3.5)). In principle, such solutions can be computed via any first-order method. (For the implementation details, we refer readers to Section 3.2.) From the description above, we see that the success of our framework hinges upon the proper choices of τ_k, γ_k and η_k, which ensure the reduction of the smoothed duality gap $\Delta_{\rho_k}(x^k,\lambda^k)$ in each iteration, and simultaneously decrease ρ_k (cf. Section 3.3).

3.2. Solving sub-problems. In Algorithm 1, it is important for us to solve the sub-problems in steps 1, 3 and 6 inexactly in an efficient manner. Since both of the optimization problems in (2.4) and (2.8) have composite forms, it is natural for us to employ the optimal first-order algorithm in [24] to solve them. More specifically, we choose N_1 and N_2 to be the accelerated proximal gradient method (denoted as APG) in [24, Equation (4.9)].

We now briefly review the convergence rate of this method. Let $\mathcal{U}$ be a nonempty, convex and compact set (in $\mathbb{E}_1$ or $\mathbb{E}_2$). In addition, let ϕ_1 and ϕ_2 be CCP functions such that ϕ_1 is L'-smooth on $\mathcal{U}$, and ϕ_2 is μ'-s.c. on $\mathcal{U}$ (where $\mu'>0$) and admits an easily computable Bregman projection on $\mathcal{U}$ (with DGF π; cf. Section 2.1). Consider

$$\min_{u\in\mathcal{U}}\big\{\Psi(u)\triangleq\phi_1(u)+\phi_2(u)\big\}, \tag{3.6}$$

whose (unique) solution is denoted by $u^*\in\mathcal{U}$. Define $\kappa'\triangleq L'/\mu'$. From [24, Theorem 6], if we use APG to solve (3.6), then for any starting point $u^0\in\mathcal{U}^o\triangleq\mathcal{U}\cap\mathsf{int\,dom}\,\pi$,

$$\begin{aligned}\Psi(u^N)-\Psi(u^*) &\le (L'/4)(1+1/\sqrt{2\kappa})^{-2(N-1)}\|u^0-u^*\|^2\\ &\le (L'/4)(1+1/\sqrt{2\kappa})^{-2(N-1)}D_{\mathcal{U}}^2,\quad \forall\, N\in\mathbb{N},\end{aligned} \tag{3.7}$$

where u^N denotes the N-th iterate of APG (where $N\in\mathbb{N}$) and

$$D_{\mathcal{U}}\triangleq\max_{u,u'\in\mathcal{U}}\|u-u'\|<+\infty \tag{3.8}$$

denotes the diameter of the constraint set $\mathcal{U}$. In other words, to obtain an ε-optimality gap, the number of iterations of APG that we need is

$$\left\lceil\sqrt{\frac{\kappa'}{2}}\log\left(\frac{L'D_{\mathcal{U}}^2}{4\varepsilon}\right)\right\rceil+1=O\left(\sqrt{\kappa'}\log\left(\frac{L'}{\varepsilon}\right)\right). \tag{3.9}$$

Note that (3.9) also holds for when ϕ_1, instead of ϕ_2, is μ-s.c. on $\mathcal{U}$; see [24, Section 5.1].

Therefore, based on (3.9), to find an η-inexact solution of (2.8) (cf. (2.12)), the number of dual oracle calls (cf. Section 1.1) made by N_1 is

$$C_{\mathsf{N}_1}\triangleq n\left\{\left\lceil\sqrt{\frac{L_{\lambda\lambda}}{2\rho}}\log\left(\frac{L_{\lambda\lambda}D_\Lambda^2}{4\eta}\right)\right\rceil+1\right\}=O\left(n\sqrt{\frac{L_{\lambda\lambda}}{\rho}}\log\left(\frac{L_{\lambda\lambda}}{\eta}\right)\right), \tag{3.10}$$

and to find a γ-inexact solution of (2.4) (cf. (2.10)), the number of primal oracle calls made by N_2 is

$$C_{\mathsf{N}_2} \triangleq (n+1)\left\{\left\lceil\sqrt{\frac{L+L_{xx}}{2\mu}}\log\left(\frac{(L+L_{xx})D_{\mathcal{X}}^2}{4\gamma}\right)\right\rceil+1\right\} = O\left(n\sqrt{\kappa_{\mathcal{X}}}\log\left((L+L_{xx})/\gamma\right)\right). \tag{3.11}$$

Note that in (3.10) and (3.11), D_Λ and $D_{\mathcal{X}}$ denote the diameters of the sets Λ and $\mathcal{X}$, respectively (defined similarly to $D_{\mathcal{U}}$ as in (3.8)), and are both *finite* (cf. Assumption 3.1). Therefore, we do *not* need to know the optimal solution or the optimal objective value of the problem in (2.8) (resp. (2.4)) in order to find $\widetilde{\lambda}_{\rho,\eta}(x)$ (resp. $\widetilde{x}_\gamma(\lambda)$).

3.3. Convergence Analysis. As the first step of our analysis, we prove that in each iteration k, if the smoothing parameter ρ_k is chosen to be sufficiently large, then the smoothed duality gap is reduced.

LEMMA 3.3. *In Algorithm 1, for any $k \in \mathbb{Z}_+ \triangleq \mathbb{N}\cup\{0\}$, if $\rho_{k+1} \geq 4(1-\tau_k)^2 L_{\mathrm{D}}$, then*

$$\Delta_{\rho_{k+1}}(x^{k+1},\lambda^{k+1}) \leq \tau_k\Delta_{\rho_k}(x^k,\lambda^k) + 2\gamma_k + 2\eta_k. \tag{3.12}$$

Proof. Fix any $k\in\mathbb{Z}_+$. Since $\widehat{S}^{\mathrm{D}}_{\rho_k}(x^k,\cdot)$ is ρ_k-strongly concave on Λ,

$$\begin{aligned}\widehat{\psi}^{\mathrm{P}}_{\rho_k}(x^k) - \widehat{S}^{\mathrm{D}}_{\rho_k}(x^k,\lambda) &= \widehat{S}^{\mathrm{D}}_{\rho_k}(x^k,\lambda^*_{\rho_k}(x^k)) - \widehat{S}^{\mathrm{D}}_{\rho_k}(x^k,\lambda)\\ &\geq \frac{\rho_k}{2}\|\lambda-\lambda^*_{\rho_k}(x^k)\|^2, \quad \forall\,\lambda\in\Lambda.\end{aligned} \tag{3.13}$$

As a result, for all $\lambda\in\Lambda$, we have

$$\begin{aligned}&\Delta_{\rho_k}(x^k,\lambda^k)\\ &= \psi^{\mathrm{P}}_{\rho_k}(x^k) - \psi^{\mathrm{D}}(\lambda^k)\\ &= f(x^k)+g(x^k)+\widehat{\psi}^{\mathrm{P}}_{\rho_k}(x^k) - \psi^{\mathrm{D}}(\lambda^k)\\ &\overset{\text{(a)}}{\geq} f(x^k)+g(x^k)+\widehat{S}^{\mathrm{D}}_{\rho_k}(x^k,\lambda)+\frac{\rho_k}{2}\|\lambda-\lambda^*_{\rho_k}(x^k)\|^2 - \psi^{\mathrm{D}}(\lambda^k)\\ &= S(x^k,\lambda)-\rho_k\omega(\lambda)+\frac{\rho_k}{2}\|\lambda-\lambda^*_{\rho_k}(x^k)\|^2 - (\widehat{\psi}^{\mathrm{D}}(\lambda^k)-h(\lambda^k))\\ &= \widehat{S}^{\mathrm{P}}(x^k,\lambda)-h(\lambda)-\rho_k\omega(\lambda)+\frac{\rho_k}{2}\|\lambda-\lambda^*_{\rho_k}(x^k)\|^2 - \widehat{\psi}^{\mathrm{D}}(\lambda^k)+h(\lambda^k),\end{aligned} \tag{3.14}$$

where in (a) we use (3.13). Define $z_k(\lambda)\triangleq\tau_k\lambda^k+(1-\tau_k)\lambda$. We then multiply both sides of (3.14) by $\tau_k>0$, and obtain

$$\begin{aligned}&\tau_k\Delta_{\rho_k}(x^k,\lambda^k)\\ &\overset{\text{(a)}}{\geq}\tau_k\widehat{S}^{\mathrm{P}}(x^k,\lambda)-\tau_k h(\lambda)-\rho_{k+1}\omega(\lambda)\\ &\quad+\frac{\rho_{k+1}}{2}\|\lambda-\lambda^*_{\rho_k}(x^k)\|^2-\tau_k\widehat{\psi}^{\mathrm{D}}(\lambda^k)+\tau_k h(\lambda^k)\\ &=\tau_k\widehat{S}^{\mathrm{P}}(x^k,\lambda)+(1-\tau_k)\widehat{S}^{\mathrm{P}}(\widetilde{x}_{\gamma_k}(\hat{\lambda}^k),\lambda)-\tau_k h(\lambda)-\rho_{k+1}\omega(\lambda)\\ &\quad+\frac{\rho_{k+1}}{2}\|\lambda-\lambda^*_{\rho_k}(x^k)\|^2-\tau_k\widehat{\psi}^{\mathrm{D}}(\lambda^k)-(1-\tau_k)\widehat{S}^{\mathrm{P}}(\widetilde{x}_{\gamma_k}(\hat{\lambda}^k),\lambda)+\tau_k h(\lambda^k)\\ &\overset{\text{(b)}}{\geq}\widehat{S}^{\mathrm{P}}(x^{k+1},\lambda)-\tau_k h(\lambda)-\rho_{k+1}\omega(\lambda)+\frac{\rho_{k+1}}{2}\|\lambda-\lambda^*_{\rho_k}(x^k)\|^2\end{aligned} \tag{3.15}$$

$$
\begin{aligned}
&- \tau_k\big(\widehat{S}^{\mathrm{P}}(\widetilde{x}_{\gamma_k}(\hat{\lambda}^k),\hat{\lambda}^k) + \langle \nabla_\lambda \widehat{S}^{\mathrm{P}}(\widetilde{x}_{\gamma_k}(\hat{\lambda}^k),\hat{\lambda}^k),\, \lambda^k - \hat{\lambda}^k\rangle\big)\\
&- (1-\tau_k)\big(\widehat{S}^{\mathrm{P}}(\widetilde{x}_{\gamma_k}(\hat{\lambda}^k),\hat{\lambda}^k) + \langle \nabla_\lambda \widehat{S}^{\mathrm{P}}(\widetilde{x}_{\gamma_k}(\hat{\lambda}^k),\hat{\lambda}^k),\, \lambda - \hat{\lambda}^k\rangle\big) + \tau_k h(\lambda^k)\\
\overset{(c)}{=}& S_{\rho_{k+1}}(x^{k+1},\lambda) + (1-\tau_k)h(\lambda) - \widehat{S}^{\mathrm{P}}(\widetilde{x}_{\gamma_k}(\hat{\lambda}^k),\hat{\lambda}^k)\\
&+ \frac{\rho_{k+1}}{2}\|\lambda - \lambda^*_{\rho_k}(x^k)\|^2 - \langle \nabla_\lambda \widehat{S}^{\mathrm{P}}(\widetilde{x}_{\gamma_k}(\hat{\lambda}^k),\hat{\lambda}^k),\, z_k(\lambda) - \hat{\lambda}^k\rangle + \tau_k h(\lambda^k)\\
\overset{(d)}{\geq}& S_{\rho_{k+1}}(x^{k+1},\lambda) + \frac{\rho_{k+1}}{2}\|\lambda - \lambda^*_{\rho_k}(x^k)\|^2\\
&- \big(\widehat{\psi}^{\mathrm{D}}(z_k(\lambda)) + L_{\mathrm{D}}\|\hat{\lambda}^k - z_k(\lambda)\|^2 + 2\gamma_k\big) + h(z_k(\lambda)),
\end{aligned}
$$

where in (a) we use $\tau_k\rho_k = \rho_{k+1}$, in (b) we use the convexity of $\widehat{S}^{\mathrm{P}}(\cdot,\lambda)$, the LHS of (2.20) and the concavity of $\widehat{S}^{\mathrm{P}}(\widetilde{x}_{\gamma_k}(\hat{\lambda}^k),\cdot)$, in (c) we use the definition of $\widehat{S}^{\mathrm{P}}$ and $S_{\rho_{k+1}}$ (in (2.4) and (2.6), respectively) and in (d) we use the RHS of (2.20) and the convexity of h. Note that if we take $\lambda = \widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1})$, then $z_k(\lambda) = \lambda^{k+1}$ by step 7. In addition, from steps 2 and 7, we have

$$
\hat{\lambda}^k - \lambda^{k+1} = (1-\tau_k)\big(\widetilde{\lambda}_{\rho_k,\eta_k}(x^k) - \widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1})\big), \tag{3.16}
$$

and from (3.3) and (3.13), we have

$$
\frac{\rho_k}{2}\|\widetilde{\lambda}_{\rho_k,\eta_k}(x^k) - \lambda^*_{\rho_k}(x^k)\|^2 \leq \widehat{\psi}^{\mathrm{P}}_{\rho_k}(x^k) - \widehat{S}^{\mathrm{D}}_{\rho_k}(x^k,\widetilde{\lambda}_{\rho_k,\eta_k}(x^k)) \leq \eta_k. \tag{3.17}
$$

This observation leads us to bound $\|\widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1}) - \lambda^*_{\rho_k}(x^k)\|^2$ as

$$
\begin{aligned}
&\|\widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1}) - \lambda^*_{\rho_k}(x^k)\|^2\\
\overset{(a)}{\geq}& \frac{1}{2}\|\widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1}) - \widetilde{\lambda}_{\rho_k,\eta_k}(x^k)\|^2 - \|\widetilde{\lambda}_{\rho_k,\eta_k}(x^k) - \lambda^*_{\rho_k}(x^k)\|^2\\
\overset{(b)}{\geq}& \frac{1}{2}\|\widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1}) - \widetilde{\lambda}_{\rho_k,\eta_k}(x^k)\|^2 - \frac{2\eta_k}{\rho_k}\\
\overset{(c)}{=}& \frac{\|\lambda^{k+1} - \hat{\lambda}^k\|^2}{2(1-\tau_k)^2} - \frac{2\eta_k}{\rho_k},
\end{aligned} \tag{3.18}
$$

where in (a) we use $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$, in (b) we use (3.17) and in (c) we use (3.16). We then substitute $\lambda = \widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1})$ and (3.18) into (3.15), and obtain

$$
\begin{aligned}
\tau_k\Delta_{\rho_k}(x^k,\lambda^k) \geq& S_{\rho_{k+1}}(x^{k+1},\widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1})) + \frac{\rho_{k+1}}{2}\left(\frac{\|\lambda^{k+1} - \hat{\lambda}^k\|^2}{2(1-\tau_k)^2} - \frac{2\eta_k}{\rho_k}\right)\\
&- \big(\psi^{\mathrm{D}}(\lambda^{k+1}) + L_{\mathrm{D}}\|\hat{\lambda}^k - \lambda^{k+1}\|^2 + 2\gamma_k\big)\\
\overset{(a)}{\geq}& \psi^{\mathrm{P}}_{\rho_{k+1}}(x^{k+1}) - (1+\tau_k)\eta_k + \left(\frac{\rho_{k+1}}{4(1-\tau_k)^2} - L_{\mathrm{D}}\right)\|\lambda^{k+1} - \hat{\lambda}^k\|^2\\
&- \psi^{\mathrm{D}}(\lambda^{k+1}) - 2\gamma_k\\
\overset{(b)}{\geq}& \Delta_{\rho_{k+1}}(x^{k+1},\lambda^{k+1}) - 2\eta_k - 2\gamma_k,
\end{aligned}
$$

where in (a) we use (3.5) and in (b) we use $\tau_k \in (0,1)$ and $\rho_{k+1} \geq 4(1-\tau_k)^2 L_{\mathrm{D}}$. We hence complete the proof. ∎

In Lemma 3.3, we notice that if $\Delta_{\rho_k}(x^k,\lambda^k) > 2(\gamma_k+\eta_k)/(1-\tau_k)$, then the smoothed duality gap will be reduced, i.e., $\Delta_{\rho_{k+1}}(x^{k+1},\lambda^{k+1}) < \Delta_{\rho_k}(x^k,\lambda^k)$. Indeed, from our choices of τ_k, γ_k and η_k in Theorem 3.5 (see below), the reduction holds as

long as $\Delta_{\rho_k}(x^k, \lambda^k) > \varepsilon/2$. This corroborates our description in Section 3.1.

Before proving our main convergence results, let us state a result about linear recursion, whose proof simply follows from induction.

LEMMA 3.4. *Let $\{\alpha_k\}_{k\in\mathbb{Z}_+}$, $\{\beta_k\}_{k\in\mathbb{Z}_+}$ and $\{a_k\}_{k\in\mathbb{Z}_+}$ be real sequences. If for all $k \in \mathbb{Z}_+$, $\alpha_k \geq 0$ and*

$$a_{k+1} \leq \alpha_k a_k + \beta_k, \tag{3.19}$$

then for all $K \in \mathbb{N}$,

$$a_K \leq \left(\prod_{k=0}^{K-1} \alpha_k\right) a_0 + \sum_{k=1}^{K} \left(\prod_{j=k}^{K-1} \alpha_j\right) \beta_{k-1}, \tag{3.20}$$

where we define the empty product $\prod_{j=K}^{K-1} \alpha_j \triangleq 1$.

Based on Lemmas 3.3 and 3.4, our main results follow immediately.

THEOREM 3.5. *In Algorithm 1, if we choose $\rho_0 = 8L_{\mathrm{D}}$ and for any $k \in \mathbb{Z}_+$,*

$$\tau_k = \frac{k+1}{k+3}, \quad \gamma_k = \frac{\varepsilon}{4(k+3)} \quad \text{and} \quad \eta_k = \frac{\varepsilon}{4(k+3)}, \tag{3.21}$$

then for any starting point $(x^0, \lambda^0) \in \mathcal{X} \times \Lambda$ and $K \in \mathbb{N}$,

$$\Delta_{\rho_K}(x^K, \lambda^K) \leq B'_{\Delta}(K, \varepsilon) \triangleq \frac{2\Delta_{\rho_0}(x^0, \lambda^0)}{(K+1)(K+2)} + \frac{\varepsilon}{2}. \tag{3.22}$$

Furthermore, if Assumption 3.1 holds, then

$$\Delta(x^K, \lambda^K) \leq B_{\Delta}(K, \varepsilon) \triangleq \frac{32L_{\mathrm{D}}B_{\omega,\Lambda} + 2\Delta(x^0, \lambda^0)}{(K+1)(K+2)} + \frac{\varepsilon}{2}. \tag{3.23}$$

Proof. Based on the choice of ρ_0 and $\{\tau_k\}_{k\in\mathbb{Z}_+}$, for any $K \in \mathbb{N}$, we have

$$\prod_{k=0}^{K-1} \tau_k = \frac{2}{(K+1)(K+2)} \quad \Longrightarrow \quad \rho_K = \rho_0 \prod_{k=0}^{K-1} \tau_k = \frac{16L_{\mathrm{D}}}{(K+1)(K+2)}. \tag{3.24}$$

Therefore, we can easily verify that the condition $\rho_K \geq 4(1-\tau_{K-1})^2 L_{\mathrm{D}}$ in Lemma 3.3 is satisfied. Consequently, by the recursion in (3.12) and Lemma 3.4, we have

$$\begin{aligned} \Delta_{\rho_K}(x^K, \lambda^K) \leq & \Delta_{\rho_0}(x^0, \lambda^0) \prod_{k=0}^{K-1} \tau_k + \sum_{k=1}^{K} 2(\gamma_{k-1} + \eta_{k-1}) \prod_{j=k}^{K-1} \tau_j \\ = & \frac{2\Delta_{\rho_0}(x^0, \lambda^0)}{(K+1)(K+2)} + \sum_{k=1}^{K} \frac{\varepsilon}{k+2} \cdot \frac{(k+1)(k+2)}{(K+1)(K+2)} \\ = & \frac{2\Delta_{\rho_0}(x^0, \lambda^0)}{(K+1)(K+2)} + \frac{\varepsilon}{2} \frac{K(K+3)}{(K+1)(K+2)}. \end{aligned} \tag{3.25}$$

We then obtain (3.22) by noticing that $K(K+3) \leq (K+1)(K+2)$. Based on (3.22), to derive (3.23), we simply use (3.1) and (3.24). ∎

Remark 3.6. From (3.23), since L_{D} only depends on $L_{\lambda x}$ and $L_{\lambda\lambda}$ (cf. (2.15)), we note that the convergence of the duality gap in Algorithm 1 only requires the Lipschitz continuity of $\nabla_\lambda \Phi(\cdot, \lambda)$ and $\nabla_\lambda \Phi(x, \cdot)$ (cf. (1.3c) and (1.3d)), but not the Lipschitz continuity of ∇f, $\nabla_x \Phi(\cdot, \lambda)$ and $\nabla_x \Phi(x, \cdot)$ (cf. (1.2), (1.3a) and (1.3b)). However, the latter smoothness conditions are needed in order to use the optimal first-order algorithm, as introduced in Section 3.2, to solve the sub-problems in (3.4). By doing so, in Algorithm 1, we can achieve the overall primal oracle complexity $\widetilde{O}(1/\sqrt{\varepsilon})$. For details, we refer readers to Section 3.4.

Note that Theorem 3.5 indicates that in Algorithm 1, to achieve an ε-duality gap, the number of iterations we need is

$$K_{\mathsf{det}} \triangleq \left\lceil \frac{2\sqrt{16L_{\mathrm{D}}B_{\omega,\Lambda}+\Delta(x^0,\lambda^0)}}{\sqrt{\varepsilon}} \right\rceil + 1 = O\left(\sqrt{\frac{L_{\mathrm{D}}}{\varepsilon}}\right). \tag{3.26}$$

Furthermore, by the definition of L_{D} in (2.15), we have $K_{\mathsf{det}} = O(\sqrt{L_{\lambda\lambda}/\varepsilon}+L_{\lambda x}/\sqrt{\mu\varepsilon})$.

3.4. Oracle Complexity. Based on the results in Sections 3.2 and 3.3 (specifically, (3.10), (3.11) and (3.26)), we may analyze the primal and dual oracle complexities needed in Algorithm 1 to achieve an ε-duality gap (i.e., $\Delta(x^{\mathrm{out}},\lambda^{\mathrm{out}}) \le \varepsilon$).

THEOREM 3.7. *Let Assumption 3.1 hold. In Algorithm 1, for any starting point $(x^0,\lambda^0) \in \mathcal{X}\times\Lambda$, let $C^{\mathrm{P}}_{\mathsf{det}}$ and $C^{\mathrm{D}}_{\mathsf{det}}$ denote the primal and dual oracle complexities (cf. Section 1.1) to achieve an ε-duality gap, respectively. Then we have*

$$C^{\mathrm{P}}_{\mathsf{det}} = O\left(n\sqrt{\kappa_{\mathcal{X}}L_{\mathrm{D}}/\varepsilon}\log\left((L+L_{xx})L_{\mathrm{D}}/\varepsilon\right)\right), \tag{3.27}$$

$$C^{\mathrm{D}}_{\mathsf{det}} = O\left(n\left(\sqrt{L_{\lambda\lambda}L_{\mathrm{D}}}/\varepsilon\right)\log\left(L_{\lambda\lambda}L_{\mathrm{D}}/\varepsilon\right)\right). \tag{3.28}$$

Proof. Since $\gamma_k = \varepsilon/(4(k+3)) = O(\varepsilon/k)$ (cf. (3.21)), based on (3.11), we have

$$\begin{aligned} C^{\mathrm{P}}_{\mathsf{det}} &= \sum\nolimits_{k=1}^{K_{\mathsf{det}}} O\left(n\sqrt{\kappa_{\mathcal{X}}}\log\left((L+L_{xx})k/\varepsilon\right)\right) \\ &= O\left(n\sqrt{\kappa_{\mathcal{X}}}\Big(K_{\mathsf{det}}\log\left(\kappa_{\mathcal{X}}\right)+\log\left(K_{\mathsf{det}}!\right)\Big)\right) \\ &\overset{\text{(a)}}{=} O\left(n\sqrt{\kappa_{\mathcal{X}}}\sqrt{L_{\mathrm{D}}/\varepsilon}\Big(\log\left((L+L_{xx})/\varepsilon\right)+\log\left(L_{\mathrm{D}}/\varepsilon\right)\Big)\right) \\ &= O\left(n\sqrt{\kappa_{\mathcal{X}}L_{\mathrm{D}}/\varepsilon}\log\left((L+L_{xx})L_{\mathrm{D}}/\varepsilon\right)\right), \end{aligned} \tag{3.29}$$

where in (a) we use the fact that $\log(K!)=\Theta(K\log K)$ for any $K\in\mathbb{N}$ and (3.26).

Similarly, we can analyze the dual oracle complexity for solving (3.3). Since $\rho_k = O(L_{\mathrm{D}}/k^2)$ (cf. (3.24)) and $\eta_k = O(\varepsilon/k)$ (cf. (3.21)), based on (3.10), we have

$$\begin{aligned} C^{\mathrm{D}}_{\mathsf{det},1} &= \sum\nolimits_{k=1}^{K_{\mathsf{det}}} O\Big(n\sqrt{L_{\lambda\lambda}k^2/L_{\mathrm{D}}}\log(L_{\lambda\lambda}k/\varepsilon)\Big) \\ &= O\Big(n\sqrt{L_{\lambda\lambda}/L_{\mathrm{D}}}\Big(\log(L_{\lambda\lambda}/\varepsilon)\sum\nolimits_{k=1}^{K_{\mathsf{det}}}k+\sum\nolimits_{k=1}^{K_{\mathsf{det}}}k\log k\Big)\Big) \\ &\overset{\text{(a)}}{=} O\Big(n\sqrt{L_{\lambda\lambda}/L_{\mathrm{D}}}(L_{\mathrm{D}}/\varepsilon)\Big(\log(L_{\lambda\lambda}/\varepsilon)+\log(L_{\mathrm{D}}/\varepsilon)\Big)\Big) \\ &= O\Big(n\sqrt{L_{\lambda\lambda}L_{\mathrm{D}}}/\varepsilon\log(L_{\lambda\lambda}L_{\mathrm{D}}/\varepsilon)\Big), \end{aligned} \tag{3.30}$$

where in (a) we use $\sum_{k=1}^{K}k=\Theta(K^2)$, $\sum_{k=1}^{K}k\log k=\Theta(K^2\log K)$ and (3.26). We can also repeat this analysis to conclude that the dual oracle complexity for solving (3.5), i.e., $C^{\mathrm{D}}_{\mathsf{det},2}$ has the same order as $C^{\mathrm{D}}_{\mathsf{det},1}$. Since $C^{\mathrm{D}}_{\mathsf{det}} = C^{\mathrm{D}}_{\mathsf{det},1}+C^{\mathrm{D}}_{\mathsf{det},2}$, we complete the proof. ∎

4. Randomized IPDS Framework. When the saddle function $\Phi(\cdot,\cdot)$ has a large number of components (i.e., n is large), we propose to find the inexact solutions in steps 1, 3 and 6 of Algorithm 1 using randomized first-order methods. This is because randomized first-order methods, especially those incorporating the variance-reduction techniques (e.g., [26, 15]), enjoy superior oracle complexities compared to their deterministic counterparts, for solving finite-sum convex composite problems. Based on this idea, we develop our randomized IPDS framework, which is shown in

Algorithm 2 Randomized primal-dual smoothed gap reduction framework

Input: Initial smoothing parameter $\rho_0 > 0$, nonnegative error sequences $\{\eta_k\}_{k\in\mathbb{Z}_+}$ and $\{\gamma_k\}_{k\in\mathbb{Z}_+}$, interpolation sequence $\{\tau_k\}_{k\in\mathbb{Z}_+} \subseteq (0,1)$ and randomized first-order algorithms M_1 and M_2.

Initialize: $x^0 \in \mathcal{X}$, $\lambda^0 \in \Lambda$ and $k = 0$.

Repeat (until some convergence criterion is met)

1. Use M_1 to find $\widetilde{\lambda}_{\rho_k,\eta_k}(x^k) \in \Lambda$ such that

$$\mathbb{E}\big[\psi^{\mathrm{P}}_{\rho_k}(x^k) - S_{\rho_k}(x^k, \widetilde{\lambda}_{\rho_k,\eta_k}(x^k)) \,\big|\, \mathcal{F}_{k,0}\big] \le \eta_k \quad \text{a.s.} \tag{4.1}$$

2. Set $\hat{\lambda}^k = \tau_k\lambda^k + (1-\tau_k)\widetilde{\lambda}_{\rho_k,\eta_k}(x^k)$.
3. Use M_2 to find $\widetilde{x}_{\gamma_k}(\hat{\lambda}^k) \in X$ such that

$$\mathbb{E}\big[S(\widetilde{x}_{\gamma_k}(\hat{\lambda}^k), \hat{\lambda}^k) - \psi^{\mathrm{D}}(\hat{\lambda}^k) \,\big|\, \mathcal{F}_{k,1}\big] \le \gamma_k \quad \text{a.s.} \tag{4.2}$$

4. Set $x^{k+1} = \tau_k x^k + (1-\tau_k)\widetilde{x}_{\gamma_k}(\hat{\lambda}^k)$.
5. Set $\rho_{k+1} = \tau_k\rho_k$.
6. Use M_1 to find $\widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1}) \in \Lambda$ such that

$$\mathbb{E}\big[\psi^{\mathrm{P}}_{\rho_{k+1}}(x^{k+1}) - S_{\rho_{k+1}}(x^{k+1}, \widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1})) \,\big|\, \mathcal{F}_{k,2}\big] \le \eta_k \quad \text{a.s.} \tag{4.3}$$

7. Set $\lambda^{k+1} = \tau_k\lambda^k + (1-\tau_k)\widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1})$.
8. Set $k = k+1$.

Output: (x^k, λ^k).

Algorithm 2.

Note that at each iteration k, in steps 1, 3 and 6 of Algorithm 2, the inexact solutions that we aim to find are functions of some *stochastic* iterates, i.e., x^k, $\hat{\lambda}^k$ and x^{k+1}. Therefore, to analyze such inexact solutions, we need to properly condition on the past information. To this end, let us denote the probability space for all the stochastic processes in Algorithm 2 by $(\Omega, \mathcal{B}, \Pr)$ (where $\mathcal{B}$ denotes the Borel σ-field on Ω) and define a filtration $\bigcup_{k\in\mathbb{Z}_+}\{\mathcal{F}_{k,i}\}_{i=0}^2$, where $\mathcal{F}_{0,0} \triangleq \{\emptyset, \Omega\}$ and for any $k \in \mathbb{Z}_+$,

$$\mathcal{F}_{k,1} \triangleq \sigma\big\{\mathcal{F}_{k,0} \cup \sigma\big\{\widetilde{\lambda}_{\rho_k,\eta_k}(x^k)\big\}\big\}, \tag{4.4}$$

$$\mathcal{F}_{k,2} \triangleq \sigma\big\{\mathcal{F}_{k,1} \cup \sigma\big\{\widetilde{x}_{\gamma_k}(\hat{\lambda}^k)\big\}\big\}, \tag{4.5}$$

$$\mathcal{F}_{k+1,0} \triangleq \sigma\big\{\mathcal{F}_{k,2} \cup \sigma\big\{\widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1})\big\}\big\}. \tag{4.6}$$

Here we overload the notation $\sigma\{\cdot\}$ to represent the σ-field generated by either a family of (Borel-measurable) sets or a random variable. From this definition, we clearly have

$$\mathcal{F}_{k,0} \subseteq \mathcal{F}_{k,1} \subseteq \mathcal{F}_{k,2} \subseteq \mathcal{F}_{k+1,0}, \quad \forall\, k \in \mathbb{Z}_+. \tag{4.7}$$

In addition, we have $x^0, \lambda^0 \in \mathcal{F}_{0,0}$ and for any $k \in \mathbb{Z}_+$,

$$\widetilde{\lambda}_{\rho_k,\eta_k}(x^k), \hat{\lambda}^k \in \mathcal{F}_{k,1}, \;\; \widetilde{x}_{\gamma_k}(\hat{\lambda}^k), x^{k+1} \in \mathcal{F}_{k,2}, \;\; \widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1}), \lambda^{k+1} \in \mathcal{F}_{k+1,0}, \tag{4.8}$$

where for any function ξ and σ-field $\mathcal{F}$, $\xi \in \mathcal{F}$ denotes that ξ is measurable w.r.t. $\mathcal{F}$.

4.1. Solving sub-problems. Similar to the deterministic case (cf. Section 3.2), we choose both M_1 and M_2 to be the optimal first-order randomized method in [15, Algorithm 3]. We denote this method as RPD since it is based on the randomized

primal-dual gradient. Consider the optimization problem in (3.6), where the smooth function ϕ_1 has a finite-sum structure, i.e.,

$$\phi_1(u) = (1/m)\textstyle\sum_{i=1}^m \varphi_i(u) \tag{4.9}$$

and each φ_i is convex and L_i'-smooth on $\mathcal{U}$ (so that $L' = (1/m)\sum_{i=1}^n L_i'$). From the convergence results in [15, Corollary 1], for any starting point $u^0 \in \mathcal{U}^o$, to have $\mathbb{E}[\Psi(\widetilde{u}^N) - \Psi(u^*)] \le \varepsilon$ (where $\widetilde{u}^N$ denotes the N-th iterate of RPD), it suffices to let

$$\begin{aligned} N &= 2(m + \sqrt{8m\kappa'}) \log\Big(2(L'/\sqrt{\mu'} + \sqrt{\mu'})^2 (m + \sqrt{8m\kappa'}) D_\varpi(u^*, u^0)/\varepsilon\Big) \\ &\le 2(m + \sqrt{8m\kappa'}) \log\Big(2(L'/\sqrt{\mu'} + \sqrt{\mu'})^2 (m + \sqrt{8m\kappa'}) R_{\varpi,\mathcal{U}}(u^0)/\varepsilon\Big) \\ &= O\big((m + \sqrt{m\kappa'}) \log(L'\kappa'(m + \sqrt{m\kappa'}) R_{\varpi,\mathcal{U}}(u^0)/\varepsilon)\big), \end{aligned} \tag{4.10}$$

where ϖ denotes the DGF w.r.t. $(\phi_2, \mathcal{U})$, $D_\varpi(u, u') \triangleq \varpi(u) - \varpi(u') - \langle \nabla\varpi(u'), u - u'\rangle$ (for any $u \in \mathcal{U}$ and $u' \in \mathcal{U}^o$) denotes the Bregman distance induced by ϖ on $\mathcal{U}$, and

$$R_{\varpi,\mathcal{U}}(u^0) \triangleq \sup_{u\in\mathcal{U}} D_\varpi(u, u^0) < +\infty, \tag{4.11}$$

since ϖ is continuous on the compact set $\mathcal{U}$. Therefore, similar to the (deterministic) APG algorithm, we do not need to know u^* or $\Psi(u^*)$ to use RPD for solving (3.6).

In the context of Algorithm 2, it follows that we should fix $\bar{x} \in \mathcal{X}^o$ and use it as the starting point to solve (4.2) in each iteration k. Similarly, we can solve (4.1) and (4.3) in each iteration k by using a fixed $\overline{\lambda} \in \Lambda^o$ as the starting point. By doing so, both $R_{\overline{\omega},\mathcal{X}}(\bar{x})$ and $R_{\omega,\Lambda}(\overline{\lambda})$ are finite constants that are independent of k.

Based on (4.10), for *any* $x \in \mathcal{X}$, if we denote C_{M_1} as the number of dual oracle calls of M_1 to find an η-inexact solution of (2.8) in expectation, i.e., $\widetilde{\lambda}_{\rho,\eta}(x)$ such that $\mathbb{E}[\widehat{\psi}_\rho^{\mathrm{P}}(\lambda) - \widehat{S}_\rho^{\mathrm{D}}(x, \widetilde{\lambda}_{\rho,\eta}(x))] \le \eta$, then

$$C_{\mathsf{M}_1} = O\left(\left(n + \sqrt{nL_{\lambda\lambda}/\rho}\right) \log\left(L_{\lambda\lambda}\big(n + \sqrt{nL_{\lambda\lambda}/\rho}\big)/(\rho\eta)\right)\right), \tag{4.12}$$

where $\overline{\lambda}$ is any point in Λ^o. Similarly, for *any* $\lambda \in \Lambda$, if we denote C_{M_2} as the number of primal oracle calls of M_2 to find a γ-inexact solution of (2.4), i.e., $\widetilde{x}_\gamma(\lambda)$ such that $\mathbb{E}[\widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\lambda), \lambda) - \widehat{\psi}^{\mathrm{D}}(\lambda)] \le \gamma$, then

$$C_{\mathsf{M}_2} = O\big((n + \sqrt{n\kappa_{\mathcal{X}}}) \log\big((L + L_{xx})(n + \sqrt{n\kappa_{\mathcal{X}}})/(\mu\gamma)\big)\big). \tag{4.13}$$

4.2. Convergence Analysis. We analyze the convergence rate of Algorithm 2 in expectation. For convergence results w.h.p., we refer readers to Section 4.4.

LEMMA 4.1. *In Algorithm 2, for any* $k \in \mathbb{Z}_+$, *if* $\rho_{k+1} \ge 4(1 - \tau_k)^2 L_{\mathrm{D}}$, *then*

$$\mathbb{E}[\Delta_{\rho_{k+1}}(x^{k+1}, \lambda^{k+1}) \,|\, \mathcal{F}_{k,0}] \le \tau_k \Delta_{\rho_k}(x^k, \lambda^k) + 2\gamma_k + 2\eta_k \quad a.s. \tag{4.14}$$

Proof. To prove this lemma, one only needs to properly incorporate the inexact criteria in (4.1), (4.2) and (4.3) (which involve conditional expectations) into the proof of Lemma 3.3. The key steps are: i) taking conditional expectation over the steps in the proof of Lemma 3.3 by using the measurability results in (4.8) and ii) applying the tower property of conditional expectation by using the nested relation in (4.7).

Specifically, at the k-th iteration, we first modify the proof of Proposition 2.1 and show that

$$\begin{aligned} &\mathbb{E}[\widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\hat{\lambda}^k), \hat{\lambda}^k) - \widehat{\psi}^{\mathrm{D}}(\lambda^{k+1}) + \big\langle \nabla_\lambda \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\hat{\lambda}^k), \hat{\lambda}^k), \lambda^{k+1} - \hat{\lambda}^k\big\rangle \,|\, \mathcal{F}_{k,1}] \\ &\qquad \le L_{\mathrm{D}} \mathbb{E}[\|\hat{\lambda}^k - \lambda^{k+1}\|^2 \,|\, \mathcal{F}_{k,1}] + 2\gamma_k. \end{aligned} \tag{4.15}$$

(For notational brevity, we omit 'a.s.' here and for all the inequalities below.) Furthermore, since $\mathcal{F}_{k,0} \subseteq \mathcal{F}_{k,1}$, if we take $\mathbb{E}[\cdot|\mathcal{F}_{k,0}]$ over (4.15), then we have

$$\mathbb{E}[\widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\hat{\lambda}^k), \hat{\lambda}^k) - \widehat{\psi}^{\mathrm{D}}(\lambda^{k+1}) + \langle \nabla_\lambda \widehat{S}^{\mathrm{P}}(\widetilde{x}_\gamma(\hat{\lambda}^k), \hat{\lambda}^k), \lambda^{k+1} - \hat{\lambda}^k \rangle \,|\, \mathcal{F}_{k,0}] \leq L_{\mathrm{D}} \mathbb{E}[\|\hat{\lambda}^k - \lambda^{k+1}\|^2 \,|\, \mathcal{F}_{k,0}] + 2\gamma_k. \tag{4.16}$$

In addition, from (3.18), we have

$$\mathbb{E}[\|\widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1}) - \lambda^*_{\rho_k}(x^k)\|^2 \,|\, \mathcal{F}_{k,0}] \geq \frac{\mathbb{E}[\|\lambda^{k+1} - \hat{\lambda}^k\|^2 \,|\, \mathcal{F}_{k,0}]}{2(1-\tau_k)^2} - \frac{2\eta_k}{\rho_k}. \tag{4.17}$$

Now, we can take $\mathbb{E}[\cdot|\mathcal{F}_{k,0}]$ over Equation (c) in (3.15), and use (4.16), (4.17) and the fact that $x^k, \lambda^k \in \mathcal{F}_{k,0}$ to obtain

$$\begin{aligned}\tau_k \Delta_{\rho_k}(x^k, \lambda^k) \geq & \mathbb{E}[S_{\rho_{k+1}}(x^{k+1}, \widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1})) + \frac{\rho_{k+1}\|\lambda^{k+1} - \hat{\lambda}^k\|^2}{4(1-\tau_k)^2} - \tau_k\eta_k \\ & - \left(\psi^{\mathrm{D}}(\lambda^{k+1}) + L_{\mathrm{D}}\|\hat{\lambda}^k - \lambda^{k+1}\|^2 + 2\gamma_k\right) | \, \mathcal{F}_{k,0}].\end{aligned} \tag{4.18}$$

Again, since $\mathcal{F}_{k,0} \subseteq \mathcal{F}_{k,2}$, if we take $\mathbb{E}[\cdot|\mathcal{F}_{k,0}]$ over (4.3), then we have

$$\mathbb{E}\left[\psi^{\mathrm{P}}_{\rho_{k+1}}(x^{k+1}) - S_{\rho_{k+1}}(x^{k+1}, \widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1})) \,\middle|\, \mathcal{F}_{k,0}\right] \leq \eta_k. \tag{4.19}$$

We then substitute (4.19) into (4.18), and use the condition $\rho_{k+1} \geq 4(1-\tau_k)^2 L_{\mathrm{D}}$ to conclude that

$$\tau_k \Delta_{\rho_k}(x^k, \lambda^k) \geq \mathbb{E}[\Delta_{\rho_{k+1}}(x^{k+1}, \lambda^{k+1}) \,|\, \mathcal{F}_{k,0}] - 2\eta_k - 2\gamma_k. \tag{4.20}$$

∎

Based on Lemma 4.1, we can derive the convergence rate of Algorithm 2 in expectation. The proof directly follows that of Theorem 3.5 and the tower property of conditional expectation, hence it is omitted.

THEOREM 4.2. *In Algorithm 2, if we choose the input parameters ρ_0, $\{\tau_k\}_{k\in\mathbb{Z}_+}$, $\{\gamma_k\}_{k\in\mathbb{Z}_+}$ and $\{\eta_k\}_{k\in\mathbb{Z}_+}$ in the same way as in Theorem 3.5, then for any starting point $(x^0, \lambda^0) \in \mathcal{X} \times \Lambda$ and $K \in \mathbb{N}$, $\mathbb{E}[\Delta_{\rho_K}(x^K, \lambda^K)] \leq B'_\Delta(K, \varepsilon)$ (defined in (3.22)). Moreover, if Assumption 3.1 holds, then $\mathbb{E}[\Delta(x^K, \lambda^K)] \leq B_\Delta(K, \varepsilon)$ (defined in (3.23)).*

Denote K_{stoc} as the number of iterations needed to achieve an ε-expected duality gap in Algorithm 2. Based on Theorem 4.2, we have that $K_{\mathsf{stoc}} = K_{\mathsf{det}} = O(\sqrt{L_{\mathrm{D}}/\varepsilon})$.

4.3. Oracle Complexity. We analyze the primal and dual oracle complexities of Algorithm 2 to achieve an ε-expected duality gap, i.e., $\mathbb{E}[\Delta(x^{\mathrm{out}}, \lambda^{\mathrm{out}})] \leq \varepsilon$.

THEOREM 4.3. *Let Assumption 3.1 hold. In Algorithm 2, for any starting point $(x^0, \lambda^0) \in \mathcal{X} \times \Lambda$, denote $C^{\mathrm{P}}_{\mathsf{stoc}}$ and $C^{\mathrm{D}}_{\mathsf{stoc}}$ as the primal and dual oracle complexities to achieve an ε-expected duality gap, respectively. Then we have*

$$C^{\mathrm{P}}_{\mathsf{stoc}} = O\left((n + \sqrt{n\kappa_{\mathcal{X}}})\sqrt{\frac{L_{\mathrm{D}}}{\varepsilon}} \log\left(\frac{\kappa_{\mathcal{X}} L_{\mathrm{D}}(n + \sqrt{n\kappa_{\mathcal{X}}})}{\varepsilon}\right)\right), \tag{4.21}$$

$$C^{\mathrm{D}}_{\mathsf{stoc}} = O\left(\left(n\sqrt{\frac{L_{\mathrm{D}}}{\varepsilon}} + \frac{\sqrt{nL_{\lambda\lambda}L_{\mathrm{D}}}}{\varepsilon}\right) \log\left(\frac{L_{\lambda\lambda}(n + \sqrt{nL_{\lambda\lambda}/L_{\mathrm{D}}})}{\varepsilon}\right)\right). \tag{4.22}$$

Proof. The proof follows the same line of argument as that of Theorem 3.7, hence we only outline the important steps. Based on the choice of γ_k in (3.21) and the complexity of M_2 in (4.13), we have

$$C^{\mathrm{P}}_{\mathsf{stoc}} = \sum\nolimits_{k=1}^{K_{\mathsf{det}}} O\big((n + \sqrt{n\kappa_{\mathcal{X}}}) \log\left((L + L_{xx})(n + \sqrt{n\kappa_{\mathcal{X}}})k/(\mu\varepsilon)\right)\big)$$

$$
\begin{aligned}
&= O\left((n+\sqrt{n\kappa_{\mathcal{X}}})\Big(K_{\mathsf{stoc}}\log\big((L+L_{xx})(n+\sqrt{n\kappa_{\mathcal{X}}})/(\mu\varepsilon)\big)+\log\big(K_{\mathsf{stoc}}!\big)\Big)\right)\\
&= O\left((n+\sqrt{n\kappa_{\mathcal{X}}})\sqrt{L_{\mathrm{D}}/\varepsilon}\Big(\log\big((L+L_{xx})(n+\sqrt{n\kappa_{\mathcal{X}}})/(\mu\varepsilon)\big)+\log\big(L_{\mathrm{D}}/\varepsilon\big)\Big)\right)\\
&= O\big((n+\sqrt{n\kappa_{\mathcal{X}}})\sqrt{L_{\mathrm{D}}/\varepsilon}\log\big((L+L_{xx})L_{\mathrm{D}}(n+\sqrt{n\kappa_{\mathcal{X}}})/(\mu\varepsilon)\big)\big).
\end{aligned}
$$

Using the same reasoning as in the proof of Theorem 3.7, the dual oracle complexities for solving both (4.1) and (4.3) have the same order, so it suffices to only analyze the complexity for solving (4.1). Specifically, based on (4.12), we have

$$
\begin{aligned}
C^{\mathrm{D}}_{\mathsf{stoc},1} &= \textstyle\sum_{k=1}^{K_{\mathsf{stoc}}} O\Big(\big(n+\sqrt{nL_{\lambda\lambda}k^2/L_{\mathrm{D}}}\big)\log\big(L_{\lambda\lambda}(n+\sqrt{nL_{\lambda\lambda}k^2/L_{\mathrm{D}}})k/(L_{\mathrm{D}}\varepsilon)\big)\Big)\\
&\overset{\text{(a)}}{=} \textstyle\sum_{k=1}^{K_{\mathsf{stoc}}} O\Big(\big(n+k\sqrt{nL_{\lambda\lambda}/L_{\mathrm{D}}}\big)\big(\log\big(L_{\lambda\lambda}(n+\sqrt{nL_{\lambda\lambda}/L_{\mathrm{D}}})/(L_{\mathrm{D}}\varepsilon)\big)+\log k\big)\Big)\\
&= O\Big(\big(K_{\mathsf{stoc}}n+\sqrt{nL_{\lambda\lambda}/L_{\mathrm{D}}}\textstyle\sum_{k=1}^{K_{\mathsf{stoc}}}k\big)\log\big(L_{\lambda\lambda}(n+\sqrt{nL_{\lambda\lambda}/L_{\mathrm{D}}})/(L_{\mathrm{D}}\varepsilon)\big)\\
&\qquad\qquad + \big(n\textstyle\sum_{k=1}^{K_{\mathsf{stoc}}}\log k+\sqrt{L_{\lambda\lambda}/L_{\mathrm{D}}}\sum_{k=1}^{K_{\mathsf{stoc}}}k\log k\big)\Big)\\
&= O\Big(\big(n\sqrt{L_{\mathrm{D}}/\varepsilon}+\sqrt{nL_{\lambda\lambda}L_{\mathrm{D}}}/\varepsilon\big)\log\big(L_{\lambda\lambda}(n+\sqrt{nL_{\lambda\lambda}/L_{\mathrm{D}}})/(L_{\mathrm{D}}\varepsilon)\big)\\
&\qquad\qquad + \sqrt{L_{\mathrm{D}}/\varepsilon}\log(L_{\mathrm{D}}/\varepsilon)(n+\sqrt{nL_{\lambda\lambda}/\varepsilon})\Big)\\
&= O\Big(\big(n\sqrt{L_{\mathrm{D}}/\varepsilon}+\sqrt{nL_{\lambda\lambda}L_{\mathrm{D}}}/\varepsilon\big)\log\big(L_{\lambda\lambda}(n+\sqrt{nL_{\lambda\lambda}/L_{\mathrm{D}}})/\varepsilon\big)\Big),
\end{aligned}
$$

where (a) holds since $n \le kn$. We obtain (4.22) by noting that $C^{\mathrm{D}}_{\mathsf{stoc}} = \Theta\big(C^{\mathrm{D}}_{\mathsf{stoc},1}\big)$. ∎

If we compare the results in Theorem 4.3 with those in Theorem 3.7, in terms of the dependence of the primal oracle complexity on n, $\kappa_{\mathcal{X}}$ and ε, the randomized framework (i.e., Algorithm 2) indeed yields improvement upon the deterministic one (i.e., Algorithm 1), from $\widetilde{O}(n\sqrt{\kappa_{\mathcal{X}}/\varepsilon})$ to $\widetilde{O}((n+\sqrt{n\kappa_{\mathcal{X}}})/\sqrt{\varepsilon})$ (where recall that $\widetilde{O}(\cdot)$ omits the the log-factors in n, $\kappa_{\mathcal{X}}$ and ε). Similarly, the dual oracle complexity has also been improved from $\widetilde{O}(n/\varepsilon)$ to $\widetilde{O}(n/\sqrt{\varepsilon}+\sqrt{n}/\varepsilon)$.

4.4. Convergence with High Probability. Apart from convergence in expectation, given an error probability $\delta \in (0,1)$, we can modify the inexact criteria in Algorithm 2 (i.e., (4.1), (4.2) and (4.3)) to obtain an ε-duality gap w.p. at least $1-\delta$, i.e., $\Pr\{\Delta(x^{\mathrm{out}},\lambda^{\mathrm{out}}) \le \varepsilon\} \ge 1-\delta$.

THEOREM 4.4. *Let Assumption 3.1 hold and $\varepsilon > 0$ and $\delta \in (0,1)$ be given. In Algorithm 2, choose the input parameters ρ_0, $\{\tau_k\}_{k\in\mathbb{Z}_+}$, $\{\gamma_k\}_{k\in\mathbb{Z}_+}$ and $\{\eta_k\}_{k\in\mathbb{Z}_+}$ in the same way as in Theorem 3.5, fix the total number of iterations $K \in \mathbb{N}$ and modify the inexact criteria (4.1), (4.2) and (4.3) to*

$$
\mathbb{E}\big[\psi^{\mathrm{P}}_{\rho_k}(x^k) - S_{\rho_k}(x^k, \widetilde{\lambda}_{\rho_k,\eta_k}(x^k)) \,\big|\, \mathcal{F}_{k,0}\big] \le \eta_k\delta/(3K) \quad a.s., \tag{4.23}
$$

$$
\mathbb{E}\big[S(\widetilde{x}_{\gamma_k}(\hat{\lambda}^k),\hat{\lambda}^k) - \psi^{\mathrm{D}}(\hat{\lambda}^k) \,\big|\, \mathcal{F}_{k,1}\big] \le \gamma_k\delta/(3K) \quad a.s., \tag{4.24}
$$

$$
\mathbb{E}\big[\psi^{\mathrm{P}}_{\rho_{k+1}}(x^{k+1}) - S_{\rho_{k+1}}(x^{k+1}, \widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1})) \,\big|\, \mathcal{F}_{k,2}\big] \le \eta_k\delta/(3K) \quad a.s., \tag{4.25}
$$

respectively. If we set $K = K'_{\mathsf{det}} \triangleq 2\left\lceil\sqrt{\max\{\Delta_{\rho_0}(x^0,\lambda^0),0\}/\varepsilon}\right\rceil + 1$, then

$$
\Pr\{\Delta_{\rho_K}(x^K,\lambda^K) \le \varepsilon\} \ge 1-\delta. \tag{4.26}
$$

Furthermore, if we set $K = K_{\mathsf{det}}$ as in (3.26), then

$$
\Pr\{\Delta(x^K,\lambda^K) \le \varepsilon\} \ge 1-\delta. \tag{4.27}
$$

Proof. First, let us define the events $\mathcal{A}_{0,0} \triangleq \Omega$, and for any $k \in \mathbb{Z}_+$,

$$\begin{aligned}
\mathcal{A}_{k,1} &\triangleq \{\psi^{\mathrm{P}}_{\rho_k}(x^k) - S_{\rho_k}(x^k, \widetilde{\lambda}_{\rho_k,\eta_k}(x^k)) \le \eta_k\},\\
\mathcal{A}_{k,2} &\triangleq \{S(\widetilde{x}_{\gamma_k}(\hat{\lambda}^k), \hat{\lambda}^k) - \psi^{\mathrm{D}}(\hat{\lambda}^k) \le \gamma_k\},\\
\mathcal{A}_{k+1,0} &\triangleq \{\psi^{\mathrm{P}}_{\rho_{k+1}}(x^{k+1}) - S_{\rho_{k+1}}(x^{k+1}, \widetilde{\lambda}_{\rho_{k+1},\eta_k}(x^{k+1})) \le \eta_k\}.
\end{aligned}$$

Also, for any measurable event $\mathcal{A}$, denote its complement as $\mathcal{A}^{\mathrm{c}}$ and its indicator function as $\mathbb{I}_{\mathcal{A}}$, i.e., $\mathbb{I}_{\mathcal{A}}(z) = 1$ if $z \in \mathcal{A}$ and 0 otherwise.

Fix any $k \in \{0, \ldots, K-1\}$. From Markov's inequality and (4.23), we have

$$\Pr\big\{\mathcal{A}^{\mathrm{c}}_{k,1} \,\big|\, \mathcal{F}_{k,0}\big\} \le \mathbb{E}\big[\psi^{\mathrm{P}}_{\rho_k}(x^k) - S_{\rho_k}(x^k, \widetilde{\lambda}_{\rho_k,\eta_k}(x^k)) \,\big|\, \mathcal{F}_{k,0}\big]/\eta_k \le \delta/(3K) \ \text{ a.s.} \tag{4.28}$$

Since $\bigcup_{i=0}^{k-1}\{\mathcal{A}_{i,1}, \mathcal{A}_{i,2}, \mathcal{A}_{i+1,0}\} \subseteq \mathcal{F}_{k,0}$, we have that

$$\mathcal{C}_{k,0} \in \mathcal{F}_{k,0}, \quad \text{where} \quad \mathcal{C}_{k,0} \triangleq \textstyle\bigcap_{i=0}^{k-1} \big(\mathcal{A}_{i,1} \cap \mathcal{A}_{i,2} \cap \mathcal{A}_{i+1,0}\big). \tag{4.29}$$

(When $k = 0$, define $\mathcal{C}_{0,0} \triangleq \mathcal{A}_{0,0}$.) In addition, note that $\Pr\{\mathcal{C}_{k,0}\} > 0$, since

$$\Pr\big\{\mathcal{C}^{\mathrm{c}}_{k,0}\big\} = \Pr\big\{\textstyle\bigcup_{i=0}^{k-1}\big(\mathcal{A}^{\mathrm{c}}_{k-1,1} \cup \mathcal{A}^{\mathrm{c}}_{k-1,2} \cup \mathcal{A}^{\mathrm{c}}_{k,0}\big)\big\} \le (3k)\delta/(3K) \le \delta < 1. \tag{4.30}$$

We then take conditional expectation $\mathbb{E}[\cdot \,|\, \mathcal{C}_{k,0}]$ in (4.28) to obtain

$$\mathbb{E}\big[\Pr\big\{\mathcal{A}_{k,1} \,\big|\, \mathcal{F}_{k,0}\big\} \,|\, \mathcal{C}_{k,0}\big] \ge 1 - \delta/(3K). \tag{4.31}$$

On the other hand,

$$\begin{aligned}
\mathbb{E}\big[\Pr\big\{\mathcal{A}_{k,1} \,\big|\, \mathcal{F}_{k,0}\big\} \,|\, \mathcal{C}_{k,0}\big] &= \mathbb{E}\big[\Pr\big\{\mathcal{A}_{k,1} \,\big|\, \mathcal{F}_{k,0}\big\}\mathbb{I}_{\mathcal{C}_{k,0}}\big]/\mathbb{P}(\mathcal{C}_{k,0})\\
&\overset{\text{(a)}}{=} \mathbb{E}\big[\mathbb{I}_{\mathcal{A}_{k,1}}\mathbb{I}_{\mathcal{C}_{k,0}}\big]/\mathbb{P}(\mathcal{C}_{k,0}) = \Pr\big\{\mathcal{A}_{k,1} \,|\, \mathcal{C}_{k,0}\big\},
\end{aligned}$$

where (a) follows since $\mathcal{C}_{k,0} \in \mathcal{F}_{k,0}$. Therefore, we have

$$\Pr\{\mathcal{A}_{k,1} \,|\, \mathcal{C}_{k,0}\} \ge 1 - \delta/(3K). \tag{4.32}$$

Similarly, if we define $\mathcal{C}_{k,1} \triangleq \mathcal{C}_{k,0} \cap \mathcal{A}_{k,1}$ and $\mathcal{C}_{k,2} \triangleq \mathcal{C}_{k,1} \cap \mathcal{A}_{k,2}$, then we also have

$$\Pr\{\mathcal{A}_{k,2} \,|\, \mathcal{C}_{k,1}\} \ge 1 - \delta/(3K), \quad \Pr\{\mathcal{A}_{k+1,0} \,|\, \mathcal{C}_{k,2}\} \ge 1 - \delta/(3K). \tag{4.33}$$

From Theorem 3.5, we know that if $K = K'_{\mathsf{det}}$ and the event $\bigcap_{k=0}^{K-1}\big(\mathcal{A}_{k,1} \cap \mathcal{A}_{k,2} \cap \mathcal{A}_{k+1,0}\big)$ occurs, then $\Delta_{\rho_K}(x^K, \lambda^K) \le \varepsilon$. Therefore,

$$\begin{aligned}
\Pr\big\{\Delta_{\rho_K}(x^K, \lambda^K) \le \varepsilon\big\} &\ge \Pr\big\{\textstyle\bigcap_{k=0}^{K-1}\big(\mathcal{A}_{k,1} \cap \mathcal{A}_{k,2} \cap \mathcal{A}_{k+1,0}\big)\big\}\\
&= \textstyle\prod_{k=0}^{K-1} \Pr\big\{\mathcal{A}_{k+1,0} \,|\, \mathcal{C}_{k,2}\big\} \Pr\big\{\mathcal{A}_{k,2} \,|\, \mathcal{C}_{k,1}\big\} \Pr\{\mathcal{A}_{k,1} \,|\, \mathcal{C}_{k,0}\}\\
&\overset{\text{(a)}}{\ge} \big(1 - \delta/(3K)\big)^{3K} \overset{\text{(b)}}{\ge} 1 - \delta,
\end{aligned}$$

where (a) follows from (4.32) and (4.33) and (b) follows from Bernoulli's inequality. By the same reasoning, we can also show that if $K = K_{\mathsf{det}}$, then (4.27) holds. ∎

Based on the inexact criteria in Theorem 4.4, we can also derive the primal and dual oracle complexities of obtaining an ε-duality gap w.p. at least $1 - \delta$. The derivation is essentially the same as that of Theorem 4.3, hence it is omitted.

THEOREM 4.5. *Let Assumption 3.1 hold and $\varepsilon > 0$ and $\delta \in (0, 1)$ be given. In Algorithm 2, modify the inexact criteria (4.1), (4.2) and (4.3) in the same way as in Theorem 4.4. For any starting point $(x^0, \lambda^0) \in \mathcal{X} \times \Lambda$, denote $C^{\mathrm{P}}_{\mathsf{hp}}$ and $C^{\mathrm{D}}_{\mathsf{hp}}$ as the primal and dual oracle complexities to achieve an ε-duality gap w.p. at least $1 - \delta$.*

Then we have

$$C^{\mathrm{P}}_{\mathsf{hp}} = O\left((n+\sqrt{n\kappa_{\mathcal{X}}})\sqrt{L_{\mathrm{D}}/\varepsilon}\log\left(\kappa_{\mathcal{X}}L_{\mathrm{D}}(n+\sqrt{n\kappa_{\mathcal{X}}})/(\varepsilon\delta)\right)\right), \tag{4.34}$$

$$C^{\mathrm{D}}_{\mathsf{hp}} = O\Big(\big(n\sqrt{L_{\mathrm{D}}/\varepsilon}+\sqrt{nL_{\lambda\lambda}L_{\mathrm{D}}}/\varepsilon\big)\log\Big(L_{\lambda\lambda}L_{\mathrm{D}}(n+\sqrt{nL_{\lambda\lambda}/L_{\mathrm{D}}})/(\varepsilon\delta)\Big)\Big). \tag{4.35}$$

4.5. Interface with Stochastic Approximation (SA). We can also use our randomized framework (i.e., Algorithm 2) to solve the stochastic version of (1.1), where we only have access to stochastic gradients of f, $\Phi(\cdot,\lambda)$ and $\Phi(x,\cdot)$, i.e., unbiased estimators of ∇f, $\nabla_x\Phi(\cdot,\lambda)$ and $\nabla_\lambda\Phi(x,\cdot)$ with bounded second moments (and potentially other distributional assumptions; for details, we refer readers to [32]). Indeed, in each iteration k, the sub-problems in steps 1, 3 and 6 can be solved by the optimal stochastic first-order algorithms for stochastic strongly-convex composite problems, e.g., those in [10]. Specifically, in [10], the optimality gaps of the proposed algorithms were analyzed both in expectation (cf. (4.1)) and w.h.p. (cf. (4.28)). These two convergence results can interface with our convergence analyses in Sections 4.2 and 4.4, respectively. Thus, our convergence results for Algorithm 2, both in expectation (cf. Theorem 4.2) and w.h.p. (cf. Theorem 4.4), still hold in this case.

5. Convex Optimization with Functional Constraints. In this section, we apply our IPDS frameworks (i.e., Algorithms 1 and 2) to the Lagrangian (saddle point) problems associated with the constrained convex problems.

5.1. Problem Setup. We consider

$$\min_{x\in\mathcal{X}} f(x)+r(x) \qquad \text{s.t.} \qquad g_i(x)\le 0,\ \forall\, i\in[n], \tag{5.1}$$

where $\mathcal{X}\neq\emptyset$ is convex and compact, f is μ-s.c. and L-smooth on $\mathcal{X}$, r is CCP and admits a tractable BPP on $\mathcal{X}$ (with DGF $\overline{\omega}$; cf. Section 2.1), and for each $i\in[n]$, g_i is convex and α_i-smooth on $\mathcal{X}$ (where $\alpha_i\ge 0$). We assume that there exists $\bar{x}\in\mathcal{X}^o$ (recall that $\mathcal{X}^o=\mathcal{X}\cap\mathsf{int}\,\mathsf{dom}\,\overline{\omega}$) such that $g_i(\bar{x})<0$, for any $i\in[n]$, so Slater's condition holds for (5.1). Under these conditions, (5.1) has the unique primal optimal solution $x^*\in\mathcal{X}$, a dual optimal solution $\lambda^*\in\mathbb{R}^n_+$ (where $\mathbb{R}_+\triangleq[0,+\infty)$) and zero duality gap. Moreover, any such (x^*,λ^*) is a saddle point of the Lagrangian problem associated with (5.1):

$$\min_{x\in X}\max_{\lambda\in\mathbb{R}^n_+}\big\{S(x,\lambda)=f(x)+r(x)+(1/n)\textstyle\sum_{i=1}^n n\lambda_i g_i(x)\big\}, \tag{5.2}$$

where λ_i denotes the i-th entry of λ. In addition, any saddle point (x^*,λ^*) of (5.2) is a primal-dual optimal solution pair for (5.1) with zero duality gap [6, Section 5.4]. This establishes the *equivalence* of solving (5.1) and (5.2).

Indeed, we observe that (5.2) has the same form as the SPP in (1.1). Specifically, if we set $g=r$, $h\equiv 0$, $\Lambda=\mathbb{R}^n_+$ and $\Phi_i(x,\lambda)=n\lambda_i g_i(x)$ in (1.1), then we recover (5.2). As a result, $L^i_{xx}(\lambda)=n\lambda_i\alpha_i$, $L^i_{\lambda\lambda}=0$ and

$$L^i_{\lambda x}=nM_i, \quad \text{where} \quad M_i\triangleq\alpha_i D_{\mathcal{X}}+\inf_{x\in\mathcal{X}}\|\nabla g_i(x)\|_*. \tag{5.3}$$

In (5.3), we recall that $D_{\mathcal{X}}<+\infty$ denotes the diameter of the set $\mathcal{X}$. (To obtain (5.3), we note that $M_i\le\sup_{x\in\mathcal{X}}\|\nabla g_i(x)\|_*$. Then, by the α_i-smoothness of g_i, we have $\|\nabla g_i(x)\|_*\le\alpha_i\|x-x'\|+\|\nabla g_i(x')\|_*\le\alpha_i D_{\mathcal{X}}+\|\nabla g_i(x')\|_*$, for any $x,x'\in\mathcal{X}$.) Thus,

$$L_{xx}(\lambda)=\textstyle\sum_{i=1}^n\lambda_i\alpha_i, \quad L_{\lambda x}=M\triangleq\sum_{i=1}^n M_i, \quad L_{\mathrm{D}}=2M^2/\mu. \tag{5.4}$$

In addition, since $\Phi(x,\cdot)$ is linear, we can choose $\mathbb{E}_2=(\mathbb{R}^n,\|\cdot\|_2)$, where $\|\cdot\|_2$ denotes the Euclidean norm. Subsequently, the problem in (2.8) now has closed-form

solution, i.e.,

$$\big([g_i(x)]_+/\rho\big)_{i=1}^n = \arg\max_{\lambda\in\mathbb{R}^n_+} \textstyle\sum_{i=1}^n \lambda_i g_i(x) - (\rho/2)\|\lambda\|_2^2, \tag{5.5}$$

where $[\cdot]_+ \triangleq \max\{0,\cdot\}$ and we choose the DGF w.r.t. $(0,\mathbb{R}^n_+)$ as $\omega(\cdot) = (1/2)\|\cdot\|_2^2$.

However, note that two of the assumptions that we make about (1.1) fail to hold for (5.2). First, in Assumption 3.1, we assume that Λ is bounded in (1.1), but it is unbounded in (5.2). Second, we assume that L_{xx} is a constant (w.r.t. λ) in (1.1), but it depends (linearly) on λ in (5.2). That said, both of these challenges can be overcome. Before we present the details below, we first provide some intuition. For the first challenge, recall from Remark 3.2 that the boundedness of Λ is needed for two purposes, i.e., solving the problem in (2.8) inexactly and bounding the duality gap Δ via its smoothed counterpart Δ_ρ. In the case of (5.2), from (5.5), we see that the problem in (2.8) can be solved *exactly*. In addition, in Section 5.2, we will use a convergence criterion different from the duality gap Δ. For these reasons, Λ need not be bounded for (5.2). For the second challenge, we propose to properly bound the growth of $L_{xx}(\lambda)$ in each iteration of our frameworks via bounding $\|\lambda\|_\infty$, i.e., the ℓ_∞-norm of λ.

5.2. Convergence Analysis. For the constrained problem in (5.1), instead of the duality gap, it is more common to use the (primal) optimality gap and constraint violation together as the convergence criterion (e.g., [31]). Specifically, for any $\varepsilon > 0$, $\bar{x} \in \mathcal{X}$ is an ε-optimal and ε-feasible solution of (5.1) if

$$f(\bar{x}) - f(x^*) \le \varepsilon, \quad \text{and} \quad [g_i(\bar{x})]_+ \le \varepsilon,\ \forall\, i \in [n]. \tag{5.6}$$

Note that this is a primal convergence criterion. However, if we apply Algorithm 1 or 2 to (5.2), the established convergence results (in Theorems 3.5, 4.2 and 4.4) are all in terms of the smoothed duality gap Δ_ρ. Thus, we need to relate Δ_ρ to the criteria in (5.6). Indeed, in the following lemma, we will show that if there exists $\overline{\lambda} \in \mathbb{R}^n_+$ such that both $\Delta_\rho(\bar{x},\overline{\lambda})$ and ρ are sufficiently small, then $\bar{x} \in \mathcal{X}$ satisfies (5.6).

LEMMA 5.1. *Let $(x^*,\lambda^*) \in \mathcal{X}\times\mathbb{R}^n_+$ be a saddle point of (5.2), so in particular x^* is the optimal solution of (5.1). For any $\rho > 0$ and $\epsilon \ge 0$, if there exist $\bar{x}\in\mathcal{X}$ and $\overline{\lambda}\in\mathbb{R}^n_+$ that satisfy $\Delta_\rho(\bar{x},\overline{\lambda}) \le \epsilon$, then*

$$f(\bar{x}) - f(x^*) \le \epsilon, \quad [g_i(\bar{x})]_+ \le V_i(\epsilon,\rho) \triangleq (\lambda_i^* + \|\lambda^*\|_2)\rho + \sqrt{2\epsilon\rho},\ \forall\, i \in [n]. \tag{5.7}$$

Proof. Since $\lambda^* \in \mathbb{R}^n_+$ is an optimal solution of the dual problem $\max_{\lambda\in\mathbb{R}^n_+} \psi^{\rm D}(\lambda)$, we have that $\psi^{\rm D}(\overline{\lambda}) \le \psi^{\rm D}(\lambda^*)$. This implies $\Delta_\rho(\bar{x},\lambda^*) \le \Delta_\rho(\bar{x},\overline{\lambda}) \le \epsilon$. From the definition of Δ_ρ in (2.14), we have

$$\epsilon \ge \Delta_\rho(\bar{x},\lambda^*) \ge S(\bar{x},\lambda) - (\rho/2)\|\lambda\|_2^2 - S(x,\lambda^*), \quad \forall\, x\in\mathcal{X},\ \forall\,\lambda\in\mathbb{R}^m_+. \tag{5.8}$$

We then choose $x{=}x^*$ and $\lambda{=}0$ in (5.8) to obtain

$$\epsilon \ge S(\bar{x},0) - S(x^*,\lambda^*) = f(\bar{x}) - (f(x^*) + \textstyle\sum_{i=1}^n \lambda_i^* g_i(x^*)) \ge f(\bar{x}) - f(x^*), \tag{5.9}$$

where the last step follows from $\lambda_i^* \ge 0$ and $g_i(x^*) \le 0$, for any $i \in [n]$.

Now fix any $\theta > 0$ and $i \in [n]$. Let $e_i \in \mathbb{R}^n$ denote the i-th standard basis vector, i.e., $(e_i)_i = 1$ and $(e_i)_j = 0$ for any $j \in [n]\setminus\{i\}$. In (5.8), if we choose $x = x^*$ and $\lambda = \lambda^* + \theta_i e_i$, where $\theta_i = \theta$ if $g_i(\bar{x}) > 0$ and 0 otherwise, then

$$\begin{aligned}\epsilon &\ge S(\bar{x},\lambda^*) + \theta_i g_i(\bar{x}) - (\rho/2)\|\lambda^* + \theta_i e_i\|_2^2 - S(x^*,\lambda^*)\\ &\ge \theta_i g_i(\bar{x}) - (\rho/2)\|\lambda^* + \theta_i e_i\|_2^2 \ge \theta[g_i(\bar{x})]_+ - (\rho/2)\|\lambda^* + \theta e_i\|_2^2,\end{aligned}$$

where in the last step we use $\lambda^* \geq 0$ and $\theta \geq \theta_i \geq 0$. After rearranging, we have

$$[g_i(\bar{x})]_+ \leq \rho\lambda_i^* + \rho\theta/2 + (\rho\|\lambda^*\|_2^2 + 2\epsilon)/(2\theta) \tag{5.10}$$
$$\overset{\text{(a)}}{\leq} \rho\lambda_i^* + \sqrt{\rho(\rho\|\lambda^*\|_2^2 + 2\epsilon)} \overset{\text{(b)}}{\leq} (\lambda_i^* + \|\lambda^*\|_2)\rho + \sqrt{2\epsilon\rho},$$

where in (a) we take the infimum over $\theta > 0$ and in (b) we use $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, for any $a, b \geq 0$. ∎

Using similar arguments, we can derive a stochastic version of Lemma 5.1.

LEMMA 5.2. *Let $(x^*, \lambda^*) \in \mathcal{X} \times \mathbb{R}^n_+$ be a saddle point of (5.2) and $(\bar{x}, \bar{\lambda})$ be a (stochastic) primal-dual pair such that $(\bar{x}, \bar{\lambda}) \in \mathcal{X} \times \mathbb{R}^n_+$ a.s. For any $\rho > 0$ and $\epsilon \geq 0$, if $(\bar{x}, \bar{\lambda})$ satisfies $\mathbb{E}[\Delta_\rho(\bar{x}, \bar{\lambda})] \leq \epsilon$, then*

$$\mathbb{E}[f(\bar{x})] - f(x^*) \leq \epsilon, \quad \mathbb{E}[[g_i(\bar{x})]_+] \leq V_i(\epsilon, \rho), \ \forall\, i \in [n],$$

where $V_i(\epsilon, \rho)$ is defined in (5.7). For any $\delta \in (0,1)$, if we have $\Pr\{\Delta_\rho(\bar{x}, \bar{\lambda}) \leq \epsilon\} \geq 1 - \delta$ (rather than $\mathbb{E}[\Delta_\rho(\bar{x}, \bar{\lambda})] \leq \epsilon$), then

$$\Pr\{f(\bar{x}) - f(x^*) \leq \epsilon\} \geq 1 - \delta, \quad \Pr\left\{[g_i(\bar{x})]_+ \leq V_i(\epsilon, \rho)\right\} \geq 1 - \delta, \forall\, i \in [n].$$

Based on Lemma 5.1 and the convergence results of Algorithm 1 in terms of the smoothed duality gap (cf. Theorem 3.5), we can easily derive the following results.

THEOREM 5.3. *Let $(x^*, \lambda^*) \in \mathcal{X} \times \mathbb{R}^n_+$ be a saddle point of (5.2) and $\varepsilon > 0$ be given. If we apply Algorithm 1 to solving (5.2), with the input parameters ρ_0, $\{\tau_k\}_{k\in\mathbb{Z}_+}$, $\{\gamma_k\}_{k\in\mathbb{Z}_+}$ chosen in the same way as in Theorem 3.5 and $\eta_k = 0$ for any $k \in \mathbb{Z}_+$, then for any starting point $(x^0, \lambda^0) \in \mathcal{X} \times \Lambda$ and $K \in \mathbb{N}$, we have*

$$f(x^K) - f(x^*) \leq W_f(K, \varepsilon) \triangleq \frac{2[\Delta_{\rho_0}(x^0, \lambda^0)]_+}{(K+1)(K+2)} + \frac{\varepsilon}{2}, \tag{5.11}$$

$$[g_i(x^K)]_+ \leq W_{g_i}(K, \varepsilon) \tag{5.12}$$
$$\triangleq \frac{16\,(\lambda_i^* + \|\lambda^*\|_2)\, L_{\mathrm{D}} + 8\sqrt{L_{\mathrm{D}}[\Delta_{\rho_0}(x^0, \lambda^0)]_+}}{(K+1)(K+2)} + \frac{4\sqrt{L_{\mathrm{D}}\varepsilon}}{K+1}, \ \forall\, i \in [m].$$

Proof. In Lemma 5.1, let us take $\bar{x} = x^K$, $\bar{\lambda} = \lambda^K$, $\rho = \rho_K$ and $\epsilon = [B'_\Delta(K, \varepsilon)]_+$ (where $B'_\Delta(K, \varepsilon)$ is defined in (3.22)). Using that $[a+b]_+ \leq [a]_+ + [b]_+$ (for any $a, b \in \mathbb{R}$), we have $[B'_\Delta(K, \varepsilon)]_+ \leq W_f(K, \varepsilon)$. Thus we obtain (5.11). Using the analytic expression of ρ_K in (3.24), and $\epsilon \leq W_f(K, \varepsilon)$, we also obtain (5.12). ∎

Similarly, based on Lemma 5.2 and the convergence results of Algorithm 2 in Theorems 4.2 and 4.4, we can show the following results using the same reasoning that leads to Theorem 5.3.

THEOREM 5.4. *Let $(x^*, \lambda^*) \in \mathcal{X} \times \mathbb{R}^n_+$ be a saddle point of (5.2) and $\varepsilon > 0$ be given. Let us apply Algorithm 2 to solving (5.2), with the input parameters ρ_0, $\{\tau_k\}_{k\in\mathbb{Z}_+}$, $\{\gamma_k\}_{k\in\mathbb{Z}_+}$ and $\{\eta_k\}_{k\in\mathbb{Z}_+}$ chosen in the same way as in Theorem 5.3, and the starting point chosen to be any $(x^0, \lambda^0) \in \mathcal{X} \times \Lambda$. Then for any $K \in \mathbb{N}$,*

$$\mathbb{E}[f(x^K)] - f(x^*) \leq W_f(K, \varepsilon), \tag{5.13}$$

$$\mathbb{E}[[g_i(x^K)]_+] \leq W_{g_i}(K, \varepsilon), \quad \forall\, i \in [m]. \tag{5.14}$$

In addition, for any $\delta \in (0,1)$, if we choose $K = K'_{\mathsf{det}}$ and modify the inexact crite-

ria (4.1), (4.2) *and* (4.3) *in the same way as in Theorem 4.4, then*

$$\Pr\{f(x^K) - f(x^*) \le W_f(K,\varepsilon)\} \ge 1-\delta, \tag{5.15}$$

$$\Pr\{[g_i(x^K)]_+ \le W_{g_i}(K,\varepsilon)\} \ge 1-\delta, \quad \forall\, i \in [m]. \tag{5.16}$$

From Theorems 5.3 and 5.4, we see that for Algorithm 1 to find an ε-optimal and ε-feasible solution of (5.1), or for Algorithm 2 to find such a solution in expectation (i.e., a solution that satisfies both (5.13) and (5.14)), the number of iterations needed is the same, which is denoted by K_{cons}. Furthermore, we have

$$K_{\mathsf{cons}} = O(\sqrt{L_{\rm D}/\varepsilon}) = O(M/\sqrt{\varepsilon\mu}). \tag{5.17}$$

5.3. Oracle Complexity. From Section 5.1 (specifically, (5.5)), we notice that in Algorithms 1 and 2, the dual sub-problems can be solved exactly. Therefore, we focus on analyzing their primal oracle complexities, where the sub-routines N_2 and M_2 remain the same as the ones in Sections 3.2 and 4.1, respectively. Note that based on our oracle model in Section 1.1, in the case of (5.2), the primal oracle $\mathscr{O}^{\rm P}$ returns $\nabla f(x)$ with input $(x,0)$ and $\lambda_i \nabla g_i(x)$ with input (x,λ,i).

Compared to the complexity analyses in Sections 3.4, 4.3 and 4.4, the challenge here is that $L_{xx}(\lambda)$ now depends on λ. This implies that in Algorithm 1 or 2, as $\hat{\lambda}^k$ changes over iterations, $L_{xx}(\hat{\lambda}^k)$ also changes. Although this does not affect the iteration complexity of Algorithm 1 or 2 (since $L_{\rm D}$ depends only on $L_{\lambda x}$), it does affect the oracle complexity of solving the primal sub-problem at each iteration k. To overcome this challenge, we propose to bound $\|\hat{\lambda}^k\|_\infty$ for each $k \in \mathbb{N}$ (in either the deterministic or stochastic sense).

LEMMA 5.5. *In Algorithm 1, if we choose the input parameters* ρ_0, $\{\tau_k\}_{k\in\mathbb{Z}_+}$, $\{\gamma_k\}_{k\in\mathbb{Z}_+}$ *and* $\{\eta_k\}_{k\in\mathbb{Z}_+}$ *in the same way as in Theorem 5.3, then for any* $k \in \mathbb{N}$,

$$\|\hat{\lambda}^k\|_\infty = O(1 + k\sqrt{\varepsilon\mu}/M). \tag{5.18}$$

Proof. From (5.5), we have that $\widetilde{\lambda}_{\rho_k,0}(x^k) = \big([g_i(x^k)]_+/\rho_k\big)_{i=1}^n$. By the bound on $[g_i(x^k)]_+$ in (5.12) and the expression of ρ_k in (3.24), we have

$$\|\widetilde{\lambda}_{\rho_k,0}(x^k)\|_\infty = O(1 + k\sqrt{\varepsilon\mu}/M). \tag{5.19}$$

By step 7 and the choice of $\{\tau_k\}_{k\in\mathbb{Z}_+}$ in Theorem 5.3, we have

$$\|\lambda^k\|_\infty \le \frac{k}{k+2}\|\lambda^{k-1}\|_\infty + \frac{2}{k+2}\|\widetilde{\lambda}_{\rho_k,0}(x^k)\|_\infty. \tag{5.20}$$

Based on (5.20), we use Lemma 3.4 to conclude that

$$\begin{aligned}\|\lambda^K\|_\infty &\le \frac{2}{(K+1)(K+2)}\left(\|\lambda^0\|_\infty + \sum\nolimits_{k=0}^K (k+1)\|\widetilde{\lambda}_{\rho_k,0}(x^k)\|_\infty\right) \\ &\overset{\rm (a)}{=} O\left(1 + K\sqrt{\varepsilon\mu}/M\right),\end{aligned} \tag{5.21}$$

where in (a) we use (5.19). Finally, by step 2, we have

$$\|\hat{\lambda}^k\|_\infty \le \tau_k\|\lambda^k\|_\infty + (1-\tau_k)\|\widetilde{\lambda}_{\rho_k,0}(x^k)\|_\infty. \tag{5.22}$$

We then substitute (5.19) and (5.21) into (5.22), and obtain (5.18). ∎

Based on Lemma 5.5, we can bound $L_{xx}(\hat{\lambda}^k)$ via

$$L_{xx}(\hat{\lambda}^k) \le \alpha\|\hat{\lambda}^k\|_\infty = O(\alpha + k\alpha\sqrt{\varepsilon\mu}/M), \quad \text{where} \quad \alpha \triangleq \textstyle\sum_{i=1}^n \alpha_i. \tag{5.23}$$

Based on this bound, the oracle complexity of N_2 for (approximately) solving the sub-problem in (2.4) (cf. (3.11)), and the iteration complexity of Algorithm 1, i.e., K_{cons} in (5.17), we can derive the following result.

THEOREM 5.6. *In Algorithm 1, for any starting point $(x^0, \lambda^0) \in \mathcal{X} \times \mathbb{R}_+$, denote $\overline{C}_{\mathsf{det}}$ as the oracle complexity to obtain an ε-optimal and ε-feasible solution. Then*

$$\overline{C}_{\mathsf{det}} = O\left(\frac{nM}{\sqrt{\mu\varepsilon}}\sqrt{(L+\alpha)/\mu}\log\left(\frac{L+\alpha}{\varepsilon}\right)\right). \tag{5.24}$$

Proof. Similar to the analysis in Section 3.4, we have

$$\begin{aligned}
\overline{C}_{\mathsf{det}} &\overset{\text{(a)}}{=} O\Big(n\sum\nolimits_{k=1}^{K_{\mathsf{cons}}}\sqrt{(L+\alpha)/\mu + k\alpha\sqrt{\varepsilon}/(M\sqrt{\mu})}\log\big(k\big((L+\alpha)/\varepsilon + k\alpha\sqrt{\mu}/(M\sqrt{\varepsilon})\big)\big)\Big)\\
&\overset{\text{(b)}}{=} O\Big(n\sum\nolimits_{k=1}^{K_{\mathsf{cons}}}\big(\sqrt{(L+\alpha)/\mu} + \sqrt{k\alpha/M}(\varepsilon/\mu)^{1/4}\big)\big(\log k + \log\big((L+\alpha)\mu/(M\varepsilon)\big)\big)\Big)\\
&\overset{\text{(c)}}{=} O\Big(n\big(\sqrt{(L+\alpha)/\mu}K_{\mathsf{cons}}\big(\log K_{\mathsf{cons}} + \log\big((L+\alpha)\mu/(M\varepsilon)\big)\big)\\
&\qquad\qquad + \sqrt{\alpha/M}(\varepsilon/\mu)^{1/4}K_{\mathsf{cons}}^{3/2}\big(\log K_{\mathsf{cons}} + \log\big((L+\alpha)\mu/(M\varepsilon)\big)\big)\big)\Big)\\
&= O\Big(n\big(M\sqrt{L+\alpha}/(\mu\sqrt{\varepsilon}) + M\sqrt{\alpha}/(\mu\sqrt{\varepsilon})\big)\log\big((L+\alpha)/\varepsilon\big)\Big),
\end{aligned}$$

where in (a) we use $\gamma_k = \Theta(\varepsilon/k)$, in (b) we use $\alpha/\sqrt{\varepsilon} = O((L+\alpha)/\varepsilon)$ and in (c) we use $\sum_{k=1}^{K} k^\nu \log k = \Theta(K^{\nu+1}\log K)$, for any $\nu \geq 0$. Finally, by noting that $\alpha \leq L+\alpha$, we obtain (5.24). ∎

Based on the oracle complexity of M_2 for (approximately) solving the sub-problem in (2.4) (cf. (4.13)), by using the same arguments as in Theorem 5.6, we can also derive the following oracle complexity for Algorithm 2.

THEOREM 5.7. *In Algorithm 2, for any starting point $(x^0, \lambda^0) \in \mathcal{X} \times \mathbb{R}_+$, denote $\overline{C}_{\mathsf{stoc}}$ as the oracle complexity to obtain an ε-optimal and ε-feasible solution in expectation. Then we have*

$$\overline{C}_{\mathsf{stoc}} = O\left(\frac{\sqrt{n}M}{\sqrt{\mu\varepsilon}}\left(\sqrt{n} + \sqrt{(L+\alpha)/\mu}\right)\log\left(\frac{nM(L+\alpha)}{\mu\varepsilon}\right)\right). \tag{5.25}$$

Let us compare the complexity results in Theorems 5.6 and 5.7. If we interpret the factor $\kappa_{\mathsf{cons}} \triangleq (L+\alpha)/\mu$ as the "condition number" of the constrained problem in (5.1), and recall that $K_{\mathsf{cons}} = O(M/\sqrt{\mu\varepsilon})$ (cf. (5.17)), then the oracle complexity of Algorithm 1, i.e., $\widetilde{O}(nK_{\mathsf{cons}}\sqrt{\kappa_{\mathsf{cons}}})$, has been reduced to $\widetilde{O}\big(\sqrt{n}K_{\mathsf{cons}}(\sqrt{n}+\sqrt{\kappa_{\mathsf{cons}}})\big)$ in Algorithm 2. This is indeed consistent with our observation in Section 4.3, which concerns the oracle complexities of Algorithms 1 and 2 for the SPP in (1.1).

5.4. Related Works. Although numerous first-order methods have been proposed for solving the constrained problem in (5.1) when f is non-strongly convex (i.e., $\mu = 0$), it appears that there exist few methods that tackle the case where f is strongly convex (i.e., $\mu > 0$). Among these, the best-known oracle complexity is $\widetilde{O}(\varepsilon^{-1/2})$, which has been achieved by three methods, which include the inexact ALM [31], the inexact dual gradient method [18] and the level-set method [17].[1] From Theorems 5.6 and 5.7, we observe that this complexity is also achieved by Algorithms 1 and 2.

[1] Note that the method in [17] requires to start with a strictly feasible solution of (5.1). However, obtaining such a solution "may be as costly as computing an optimal solution" [1]. In contrast, this is not required in any of the other methods discussed in this section (including ours).

Although all of these methods achieve the same complexity, there are two important features that distinguish our methods from the rest. First, our randomized framework (i.e., Algorithm 2) can effectively handle the case where the number of constraints n is extremely large (cf. Theorem 5.7). Second, both of our frameworks (i.e., Algorithms 1 and 2) are developed for solving the *general* SPP in (1.1), not only the Lagrangian problem in (5.2). Therefore, they have much wider applicability compared to the other methods.

5.5. Stochastic Convex Optimization with Expectation Constraints. We consider a problem related to (5.1), where $f, g_1, \ldots, g_n$ are given in terms of expectation (see e.g., [16]). Specifically, let $\xi, \zeta_1, \ldots, \zeta_n$ be random vectors with sample spaces denoted by $\Xi, \mathcal{Z}_1, \ldots, \mathcal{Z}_n$, respectively, and $F : \mathbb{E}_1 \times \Xi \to \mathbb{R}$ and $\{G_i : \mathbb{E}_1 \times \mathcal{Z}_i \to \mathbb{R}\}_{i=1}^n$ be chosen such that

$$f(x) \triangleq \mathbb{E}_\xi[F(x,\xi)], \qquad g_i(x) \triangleq \mathbb{E}_{\zeta_i}[G_i(x,\zeta_i)], \quad \forall\, i \in [n], \tag{5.26}$$

and $f, g_1, \ldots, g_n$ satisfy the structural assumptions stated in Section 5.1. Since the expectations in (5.26) cannot be evaluated with high accuracy in general, from the stochastic programming literature (e.g., [27]), there are two standard approaches to tackle this difficulty, i.e., SA and sample average approximation (SAA). The SA approach involves employing certain mechanisms to generate unbiased estimators of ∇f or ∇g_i at any $x \in \mathcal{X}$, whereas the SAA approach involves generating i.i.d. samples of $\xi, \zeta_1, \ldots, \zeta_n$ (denoted as $\{\xi^j\}_{j=1}^{m_0}, \{\zeta_1^j\}_{j=1}^{m_1}, \ldots, \{\zeta_n^j\}_{j=1}^{m_n}$, respectively) to approximate the expectations in (5.26), i.e.,

$$\begin{aligned} &\min_{x\in\mathcal{X}} \{\widehat{f}(x) \triangleq (1/m_0)\textstyle\sum_{j=1}^{m_0} F(x,\xi_j)\}, \\ &\text{s.t.} \quad \{\widehat{g}_i(x) \triangleq (1/m_i)\textstyle\sum_{j=1}^{m_i} G_i(x,\zeta_i^j)\} \le 0,\ \forall\, i \in [n]. \end{aligned} \tag{5.27}$$

Note that to achieve high approximation accuracy, the sample sizes $\{m_i\}_{i=0}^n$ are typically very large. We illustrate that Algorithm 2 can solve (5.26), no matter whether SA or SAA is used. As in Section 1, we still apply Algorithm 2 to the Lagrangian problem associated with (5.26). The ability of Algorithm 2 to interface with SA can be readily seen from Section 4.5. For the case of SAA, we see that (5.27) has a Lagrangian problem that fits into the finite-sum structure in (1.5), which again can be efficiently solved by Algorithm 2 (cf. Theorem 5.7).

Acknowledgments. The authors would like to express sincere gratitude to Robert M. Freund and Yangyang Xu for helpful discussions and constructive feedback during the preparation of this manuscript.

REFERENCES

[1] Aleksandr Y. Aravkin, James V. Burke, Dmitriy Drusvyatskiy, Michael P. Friedlander, and Scott Roy. Level-set methods for convex optimization. arXiv:1602.01506, 2016.
[2] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM J. Optim.*, 16(3):697–725, 2006.
[3] P. Balamurugan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. In *Proc. NIPS*, 2016.
[4] Heinz H. Bauschke, Jonathan M. Borwein, and Patrick L. Combettes. Essential smoothness, essential strict convexity, and legendre functions in Banach spaces. *Commun. Contemp. Math.*, 3(4):615–647, 2001.
[5] Dimitri P. Bertsekas. *Nonlinear Programming.* Athena Scitific, 1999.
[6] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[7] Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Math. Program.*, 159(1):253–287, 2016.
[8] Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM J. Optim.*, 24(4):1779–1814, 2014.
[9] Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. Accelerated schemes for a class of variational inequalities. *Math. Program.*, 165(1):113–149, 2017.
[10] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: Shrinking procedures and optimal algorithms. *SIAM J. Optim.*, 23(4):2061–2089, 2013.
[11] Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm for general convex-concave saddle point problems. arXiv:1803.01401, 2018.
[12] Anatoli Juditsky and Arkadi Nemirovski. *First-Order Methods for Nonsmooth Convex Large-Scale Optimization, II: Utilizing Problems Structure*, pages 149–184. MIT Press, 2012.
[13] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stoch. Syst.*, 1(1):17–58, 2011.
[14] O. Kolossoski and R.D.C. Monteiro. An accelerated non-euclidean hybrid proximal extragradient-type algorithm for convexconcave saddle-point problems. *Optim. Methods Softw.*, 32(6):1244–1272, 2017.
[15] Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Math. Program.*, 171(1):167–215, 2018.
[16] Guanghui Lan and Zhiqiang Zhou. Algorithms for stochastic optimization with expectation constraints. arXiv:1604.03887, 2016.
[17] Q. Lin, S. Nadarajah, and N. Soheili. A level-set method for convex optimization with a feasible solution path. *SIAM J. Optim.*, 28(4):3290–3311, 2018.
[18] I. Necoara and V. Nedelcu. Rate analysis of inexact dual first-order methods application to dual decomposition. *IEEE Trans. Autom. Control*, 59(5):1232–1243, 2014.
[19] A. Nedić and A. Ozdaglar. Subgradient methods for saddle-point problems. *J. Optim. Theory Appl.*, 142(1):205–228, 2009.
[20] Arkadi Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. Optim.*, 15(1):229–251, 2005.
[21] Yu. Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM J. Optim.*, 16(1):235–249, 2005.
[22] Yu. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.
[23] Yu. Nesterov. Primal-dual subgradient methods for convex problems. *Math. Program.*, 120(1):221–259, 2009.
[24] Yu. Nesterov. Gradient methods for minimizing composite functions. *Math. Program.*, 140(1):125–161, 2013.
[25] Andrei Patrascu, Ion Necoara, and Quoc Tran-Dinh. Adaptive inexact fast augmented lagrangian methods for constrained convex optimization. *Optim. Lett.*, 11(3):609–626, 2017.
[26] Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Math. Program.*, 155(1):105–145, 2016.
[27] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming Modeling and Theory*. SIAM and MPS, 2009.
[28] Marc Teboulle. A simplified view of first order methods for optimization. *Math. Program.*, 170(1):67–96, 2018.
[29] Q. Tran-Dinh, I. Necoara, and M. Diehl. Fast inexact decomposition algorithms for large-scale separable convex optimization. *Optim.*, 65(2):325–356, 2016.
[30] Linli Xu, James Neufeld, Bryce Larson, and Dale Schuurmans. Maximum margin clustering. In *Proc. NIPS*, pages 1537–1544, 2005.
[31] Yangyang Xu. Iteration complexity of inexact augmented lagrangian methods for constrained convex programming. arXiv:1711.05812, 2017.
[32] Renbo Zhao. Optimal algorithms for stochastic three-composite convex-concave saddle point problems. arXiv:1903.01687, 2019.