arXiv:2111.12743v1 [math.OC] 24 Nov 2021

ROBUST ACCELERATED PRIMAL-DUAL METHODS FOR COMPUTING SADDLE POINTS *

XUAN ZHANG†, NECDET SERHAT AYBAT†, AND MERT GÜRBÜZBALABAN ‡

Abstract. We consider strongly convex/strongly concave saddle point problems assuming we have access to unbiased stochastic estimates of the gradients. We propose a stochastic accelerated primal-dual (SAPD) algorithm and show that SAPD iterate sequence, generated using constant primal-dual step sizes, linearly converges to a neighborhood of the unique saddle point, where the size of the neighborhood is determined by the asymptotic variance of the iterates. Interpreting the asymptotic variance as a measure of robustness to gradient noise, we obtain explicit characterizations of robustness in terms of SAPD parameters and problem constants. Based on these characterizations, we develop computationally tractable techniques for optimizing the SAPD parameters, i.e., the primal and dual step sizes, and the momentum parameter, to achieve a desired trade-off between the convergence rate and robustness on the Pareto curve. This allows SAPD to enjoy fast convergence properties while being robust to noise as an accelerated method. We also show that SAPD admits convergence guarantees for the gap metric with a variance term optimal up to a logarithmic factor –which can be removed by employing a restarting strategy. Furthermore, to our knowledge, our work is the first one showing an iteration complexity result for the gap function on smooth SCSC problems without the bounded domain assumption. Finally, we illustrate the efficiency of our approach on distributionally robust logistic regression problems.

Key words. convex optimization, stochastic approximation, robust control theory, accelerated methods, matrix inequalities. saddle point problem

AMS subject classifications. 62L20, 90C15, 90C25, 90C30, 93C05, 93C10

1. Introduction.

We consider the following saddle point (SP) problem:

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

where $\mathcal{X}$ and $\mathcal{Y}$ are, n and m dimensional inner product spaces endowed with inner product norms $\|x\|_{\mathcal{X}}=\sqrt{\langle x,x\rangle_{\mathcal{X}}}$ and $\|y\|_{\mathcal{Y}}=\sqrt{\langle y,y\rangle_{\mathcal{Y}}}$, respectively; and $\Phi:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ is convex in x, concave in y with a Lipschitz gradient. We assume that $f:\mathcal{X}\to\mathbb{R}\cup\{\infty\}$ and $g:\mathcal{Y}\to\mathbb{R}\cup\{\infty\}$ are closed, convex functions with moduli $\mu_x,\mu_y\geq 0$, respectively. The SP problem in (1.1) has a wide range of applications; in fact, many problems in machine learning (ML) are convex optimization problems, which can be recast as (1.1) through Lagrangian duality. Prominent applications with SP formulations include empirical risk minimization (ERM) [42], supervised learning with non-separable losses, or regularizers [39, 32], distributionally robust ERM [25], and robust optimization [4].

In this paper, we consider first-order (gradient-based) methods to solve (1.1). That said, in various applications, we do not have access to exact values of the gradients $\nabla_x\Phi$ and $\nabla_y\Phi$; but, we have access to their unbiased stochastic estimates $\tilde{\nabla}_x\Phi$ and $\tilde{\nabla}_y\Phi$. This would typically be the case when the gradients are estimated from a subset of data points in the big-data regime (as in stochastic gradient, and stochastic approximation methods) or if noise is injected to the gradients on purpose to protect the privacy of the user data [38]. In a similar spirit, we assume that we have access

*This work was funded by ONR Grants N00014-21-1-2271 and N00014-21-1-2244.

†Department of Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, PA,USA. (xxz358@psu.edu, nsa10@psu.edu).

‡Department of Management Science and Information Systems, Rutgers University, Piscataway, NJ, USA.mg1366@rutgers.edu

to unbiased stochastic oracles $\tilde{\nabla}_x\Phi$ and $\tilde{\nabla}_y\Phi$ with a bounded variance, see Assumption 1.2 for the details. This setting is commonly considered in the literature and is relevant to a number of applications, e.g., training GANs, [44] and robust learning [36].

1.1. Related Work. When the coupling term $\Phi(x,y)$ is bilinear, i.e., $\Phi(x,y) = \langle Kx, y\rangle$ for some linear operator $K : \mathcal{X} \to \mathcal{Y}^*$, (1.1) is well-studied for both strongly convex/strongly concave (SCSC) problems ($\mu_x, \mu_y > 0$) as well as for merely convex/merely concave (MCMC) problems ($\mu_x = \mu_y = 0$). The convergence results cover both the stochastic case (when only stochastic estimates of the gradients are available) and the deterministic case (when the gradient information is exact). In our work, we do not assume bilinear Φ. When the coupling term Φ is non-bilinear, there exists some convergence results in the deterministic case; however, the stochastic setting remains relatively understudied.

Two standard metrics to measure the quality of a random $(\bar{x},\bar{y}) \in \mathcal{X}\times\mathcal{Y}$ returned by a stochastic algorithm are the *gap function* $\mathcal{G} : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}_+$ for the general case and *distance metric* $\mathcal{D} : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}_+$ when there is a unique SP (x^*, y^*), defined as

$$\mathcal{G}(\bar{x},\bar{y}) \triangleq \mathbb{E}\Big[\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \{\mathcal{L}(\bar{x},y) - \mathcal{L}(x,\bar{y})\}\Big], \quad \mathcal{D}(\bar{x},\bar{y}) \triangleq \mathbb{E}[\|\bar{x}-x^*\|^2 + \|\bar{y}-y^*\|^2], \tag{1.2}$$

where the expectation is taken with respect to the randomness encountered in the generation of the point $(\bar{x},\bar{y}) \in \mathcal{X}\times\mathcal{Y}$. In the following discussion, we summarize existing results closely related to our setting, and discuss our contributions.

1.1.1. The Deterministic Case. The bilinear structure has been thoroughly studied; some well-known algorithms include excessive gap technique [30, 29], primal-dual hybrid gradient (PDHG) [6, 7]. On the other extreme, when $\mathcal{L}$ owns a general form and the smoothness cannot be guaranteed, primal-dual subgradient algorithms have been proposed in several works, e.g., [26, 31, 17]. The iteration complexity of these subgradient-based methods can be significantly improved when $\mathcal{L}$ has further structure. Indeed, there are methods exploiting the structure when Φ is smooth, f and g have efficient prox maps, which include Mirror-Prox(MP) [27], Optimistic Gradient Descent Ascent (OGDA) and Extra-gradient (EG) [24] methods. Additionally, the effect of Lipschitz constants along different blocks of variables also has been explored recently; some new works account for the individual effects of L_{xx}, L_{yx} and L_{yy}, i.e., the Lipschitz constants of $\nabla_x\Phi(\cdot,y)$, $\nabla_y\Phi(x,\cdot)$ and $\nabla_y\Phi(\cdot,y)$, respectively, instead of using the worst-case parameters $L \triangleq \max\{L_{xx}, L_{xy}, L_{yx}, L_{yy}\}$, $\mu \triangleq \min\{\mu_x,\mu_y\}$. For bilinear SP problems, a lower complexity bound of $\Omega\left(\sqrt{1+\frac{L_{yx}^2}{\mu_x\mu_y}}\cdot\ln(1/\epsilon)\right)$ is shown in [41] for a class of first-order primal-dual algorithms employing proximal-gradient steps; on the other hand, the lower bound for gradient-based methods is $\Omega\left(\sqrt{\frac{L_{xx}}{\mu_x}+\frac{L_{yx}^2}{\mu_x\mu_y}+\frac{L_{yy}}{\mu_y}}\cdot\ln(1/\epsilon)\right)$ when $f(\cdot) = g(\cdot) = 0$ and Φ is SCSC [41]. In the rest, we focus on the SCSC setup.

Mokhtari *et al.* [24] show that both OGDA and EG have an iteration complexity of $\mathcal{O}\Big(\frac{L}{\mu}\ln(1/\varepsilon)\Big)$ for $\mathcal{D}$ metric defined in (1.2). Gidel *et al.* [12] also show the same rate for OGDA from a variational inequality (VI) perspective. In the analysis of these algorithms, primal and dual step sizes are set equal, which may lead to conservative steps whenever $L_{xx} \gg L_{yy}$, or vice versa. For instance, in the primal-dual formulation of empirical risk minimization problems in machine learning, choosing primal and dual step sizes to be different can lead to an improved convergence rate [42]. There are also some multi-loop algorithms. In particular, Lin *et al.* [22] proposed

an inexact proximal point algorithm, which consists of 3-nested loops. Indeed, each proximal step computations require calling AGD iteratively to solve strongly convex smooth (SCS) optimization subproblems with a high precision that can be impractical.[1] The computational complexity to compute $(\bar{x}, \bar{y})$ such that $\mathcal{G}(\bar{x}, \bar{y}) \leq \epsilon$ is $\mathcal{O}\left(\frac{L}{\sqrt{\mu_x \mu_y}} \cdot \ln^3(1/\epsilon)\right)$. Although they claim that their algorithm achieves the lower complexity bound in [41], this is not the case for problems with $L_{yx} \ll L$. The algorithm in [35] consists of 4-nested loops and has similar shortcomings in practice. The computational complexity to compute $(\bar{x}, \bar{y})$ such that $\mathcal{D}(\bar{x}, \bar{y}) \leq \epsilon$ is $\mathcal{O}(\sqrt{\frac{L_{xx}}{\mu_x} + \frac{LL_{xy}}{\mu_x \mu_y} + \frac{L_{yy}}{\mu_y}} \cdot \ln^3(1/\epsilon))$ – note that this is a weaker result in terms of the metric used, i.e., $\mathcal{G}(\bar{x}, \bar{y}) \leq \epsilon$ implies that $\mathcal{D}(\bar{x}, \bar{y}) = \mathcal{O}(\epsilon)$; but, the reverse is not true.

1.1.2. The Stochastic Case. While the deterministic SP problem has attracted huge attention, the study on the first-order stochastic methods for (1.1) is still relatively limited. For MCMC SP problems, proximal methods have been developed, e.g., Stochastic Mirror-Descent(SMD) [28], the Stochastic Mirror-Prox (SMP) [18] and its accelerated version (SAMP) [8]. In [43], Zhao considers MCMC and SCMC scenarios under *additive* unbiased noise with bounded variance setting and assuming $\mathcal{X}$ and $\mathcal{Y}$ are bounded. When $\mu_x > 0$, a multi-stage scheme achieving the best known complexity for the stochastic SP problems is proposed in [43]; however, this is a two-loop method and each outer iteration requires solving a non-trivial sub-problem with an increasing accuracy, which is a function of some problems parameters that may not be known in practice, e.g., Bregman diameters of $\mathcal{X}, \mathcal{Y}$ and noise variance. In addition, there are also some VI-based methods [10, 12] for the MCMC scenario.

Our focus in this paper will be on the SCSC case. Yan *et al.* [40] consider $\min_{x \in X} \max_{y \in Y} \Phi(x, y)$ for possibly non-smooth, SCSC Φ, and propose Epoch-GDA with an iteration complexity of $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ for the duality gap with high probability. When Φ is smooth, Hsieh *et al.* [16] show that a stochastic EG method can guarantee $\mathcal{D}(x_k, y_k) \leq \epsilon$ within $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ iterations. Fallah *et al.* [11] propose multi-stage variants (employing restarts) of Stochastic Gradient Descent Ascent (S-GDA) and Stochastic OGDA (S-OGDA) that can guarantee $\mathcal{D}(x_k, y_k) \leq \epsilon$ within $\mathcal{O}(\kappa^2 \ln(1/\epsilon) + \dfrac{\delta^2}{\mu^2 \epsilon})$ and $\mathcal{O}(\kappa \ln(1/\epsilon) + \dfrac{\delta^2}{\mu^2 \epsilon})$ iterations, respectively, where $\kappa = L/\mu$. Unlike our paper, in both [16, 11], Lipschitz constant of $\nabla \Phi$, i.e., L, is used to determine the step size, rather than exploiting the block Lipschitz structure.

1.2. Contributions. We propose the Stochastic Accelerated Primal-Dual (SAPD) algorithm which extends APD method proposed in [15] to the stochastic gradient setting. More precisely, we assume that the first-order oracles (SFO) $\tilde{\nabla} \Phi_x$ and $\tilde{\nabla} \Phi_y$ return noisy partial gradients that are unbiased and have finite variance δ_x^2 and δ_y^2, respectively. Let $\Omega_f \triangleq \sup_{x, x' \in \mathbf{dom}\, f} \|x - x'\|^2$ and $\Omega_g \triangleq \sup_{y, y' \in \mathbf{dom}\, g} \|y - y'\|^2$, for any $\epsilon > 0$, SAPD guarantees $\mathcal{G}(\bar{x}_K, \bar{y}_K) \leq \epsilon$ within

$$K \leq \mathcal{O}\Big(\Big(\frac{L_{xx}}{\mu_x} + \frac{L_{yx}}{\sqrt{\mu_x \mu_y}} + \frac{L_{yy}}{\mu_y} + \Big(\frac{\delta_x^2}{\mu_x} + \frac{\delta_y^2}{\mu_y}\Big)\frac{1}{\epsilon}\Big) \cdot \ln\Big(\frac{\mu_x \Omega_f + \mu_y \Omega_g}{\epsilon}\Big)\Big),$$

iterations, where $\{(\bar{x}_K, \bar{y}_K)\}$ is a particular weighted average sequence. The bound on the variance term $\tilde{\mathcal{O}}((\delta_x^2/\mu_x + \delta_y^2/\mu_y)/\epsilon)$ is optimal up to a log factor, which can be

[1] In each AGD call, an SCS function h is minimized with condition number $\kappa_x = L/\mu_x$ to compute $\bar{x} \approx \mathbf{argmin}_{x \in X}\, h(x)$ such that $\|\bar{x} - \Pi_X(\bar{x} - \nabla h(\bar{x})/L)\|^2 \leq \frac{\epsilon}{2(10\kappa_y)^{11}\kappa_x^{13}}$.

removed by employing a restarting strategy similar to [11]. Moreover, to the best of our knowledge, our work is the first one showing an iteration complexity result for the gap function $\mathcal{G}$ on smooth SCSC problems without the bounded domain assumption. Since we assume that the noise is persistent, linear convergence cannot be achieved – unlike the finite sum problems where variance reduction-based methods are applicable to obtain linear convergence [32]. However, for SCSC problems, SAPD with constant step size converges to a neighborhood of the saddle-point at a linear rate ρ, and the size of the neighborhood is determined by the asymptotic variance of the iterates.

The asymptotic variance of the iterates scales linearly with the gradient noise level. We interpret the ratio of asymptotic variance to the variance of gradient noise as a measure of *robustness*, which we denote with $\mathcal{J}$. We evaluate the overall algorithmic performance with two metrics: SAPD parameters should be tuned to achieve faster rate ρ with a smaller noise amplification $\mathcal{J}$. Our analysis leads to explicit characterizations of $\mathcal{J}$ for a particular problem class, and of an upper bound $\mathcal{R}$ on $\mathcal{J}$ for more general problems; both $\mathcal{J}$ and $\mathcal{R}$ are given as functions of SAPD parameters, strong convexity and Lipschitz constants of the problem in (1.1). Based on these characterizations, we develop computationally tractable techniques for optimizing the parameters of the SAPD algorithm to achieve a desired systematic trade-off between ρ and $\mathcal{J}$ without assuming the knowledge of neither the noise variances δ_x^2, δ_y^2 nor the domain diameters Ω_f and Ω_g. This allows SAPD to enjoy fast convergence with a robust performance in the presence of stochastic gradient noise. Achieving systematic trade-offs between the rate and robustness have been previously studied in [2] in the context of accelerated methods for smooth strongly convex minimization problems. To our knowledge, our work is the first one that can trade-off ρ with $\mathcal{J}$ in a systematic fashion in the context of primal-dual algorithms for solving the SP problem in (1.1).

The deterministic result can be derived from our stochastic results immediately by setting the noise variances $\delta_x^2 = \delta_y^2 = 0$. We note that our algorithm, when applied to (1.1) with a bilinear Φ, generates the same iterate sequence with [7] for a specific choice of step size parameters; therefore, SAPD can be viewed as a general form of the optimal method (CP) proposed by Chambolle and Pock [7] for bilinear SP problems. Thus, for SCSC bilinear problems, both CP and SAPD hit the lower complexity bound [41]. Moreover, when Φ is not assumed to be bilinear, SAPD guarantees $\mathcal{O}((\frac{L_{xx}}{\mu_x} + \frac{L_{yy}}{\mu_y} + \sqrt{\frac{L_{yx}^2}{\mu_x\mu_y}}) \cdot \ln(1/\epsilon))$, which is the best bound shown for (1.1) – we are not aware of any other single-loop methods achieving this complexity bound for (1.1). Furthermore, our framework exploiting block Lipschitz constants L_{xx}, L_{yx} and L_{yy}, provides larger step sizes compared to the traditional step size $\mathcal{O}(1/L)$.

1.3. Notation. Throughout the paper, $\mathbb{R}_{++}$ denotes the set of positive real numbers, and $\mathbb{R}_+ = \mathbb{R}_{++} \cup \{\mathbf{0}\}$. We adopted arithmetic using the extended reals with the convention that $\frac{1}{0} \triangleq \infty$, $\frac{0^2}{0} \triangleq 0$, $\frac{0}{0^2} \triangleq \infty$. We use $\|\cdot\|$ to denote the Euclidean norm. The proximal operator associated with a proper, closed convex $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ is given by $\mathbf{prox}_f(x) \triangleq \mathbf{argmin}_{v\in\mathcal{X}} f(v) + \frac{1}{2}\|x - v\|^2$, and $\mathbf{prox}_g(\cdot)$ is defined similarly. We let $\mathbb{S}^d$ denote the set of symmetric $d \times d$ real matrices.

1.4. Assumptions and Statement of SAPD Algorithm. In the following, we introduce the assumptions needed throughout this paper.

Assumption 1.1. *$f : \mathcal{X} \to \mathbb{R}\cup\{\infty\}$ and $g : \mathcal{Y} \to \mathbb{R}\cup\{\infty\}$ are proper, closed, convex functions with moduli $\mu_x, \mu_y \geq 0$. Moreover, $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is such that*

(i) for any $y \in \mathbf{dom}\, g \subset \mathcal{Y}$, $\Phi(\cdot, y)$ is convex and differentiable; and $\exists L_{xx} \geq 0$,

$\exists L_{xy} > 0$ *such that* $\forall x, \bar{x} \in \mathbf{dom}\, f \subset \mathcal{X}$ *and* $\forall y, \bar{y} \in \mathbf{dom}\, g \subset \mathcal{Y}$,

$$\|\nabla_x \Phi(x,y) - \nabla_x \Phi(\bar{x},\bar{y})\| \leq L_{xx}\|x-\bar{x}\| + L_{xy}\|y-\bar{y}\|, \tag{1.3}$$

(ii) for any $x \in \mathbf{dom}\, f \subset \mathcal{X}$, $\Phi(x,\cdot)$ *is concave and differentiable; and* $\exists\ L_{yx} > 0$ *and* $\exists L_{yy} \geq 0$ *such that* $\|\nabla_y \Phi(x,y) - \nabla_y \Phi(\bar{x},\bar{y})\| \leq L_{yx}\|x-\bar{x}\| + L_{yy}\|y-\bar{y}\|$ $\forall x, \bar{x} \in \mathbf{dom}\, f \subset \mathcal{X}$ *and* $\forall y, \bar{y} \in \mathbf{dom}\, g \subset \mathcal{Y}$.

Assumption 1.1 is standard in the analysis of first-order methods. It should be noticed that when $L_{yx} = L_{xy} = 0$, the coupling term $\Phi(x,y)$ can be written as the sum of two functions $\Phi_1(x)$ and $\Phi_2(y)$ in which case the min-max problem (1.1) becomes trivial. Therefore, our assumption $L_{yx} > 0$ is natural. In addition, for $y \in \mathbf{dom}\, g$, since $\Phi(\cdot, y)$ is convex, (1.3) implies that for all $x, \bar{x} \in \mathcal{X}$,

$$\Phi(x,y) - \Phi(\bar{x},y) - \langle \nabla_x \Phi(\bar{x},y), x-\bar{x}\rangle \leq \frac{L_{xx}}{2}\|x-\bar{x}\|^2. \tag{1.4}$$

In many ML applications, the full gradients are estimated through sampling from data as passing over the whole dataset to compute the full gradient may be computationally impractical. Within the context of SP problems, this setting arises in supervised learning tasks, e.g., [5, 32]. In the rest of the paper, we use $\tilde{\nabla}_x \Phi$ and $\tilde{\nabla}_y \Phi$ to denote such stochastic estimates of the true gradients $\nabla_x \Phi$ and $\nabla_y \Phi$. Given such stochastic oracles, we propose SAPD algorithm to tackle with (1.1), which is described in Algorithm 1.1. We note that when θ is zero, SAPD reduces to the well-known stochastic (proximal) gradient descent ascent (SGDA) method.

Algorithm 1.1 Stochastic Accelerated Primal-Dual (SAPD) Algorithm

Input: $\{\tau, \sigma, \theta\}$, $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$
$(x_{-1}, y_{-1}) \leftarrow (x_0, y_0)$
for $k \geq 0$ **do**
  $\tilde{q}_k \leftarrow \tilde{\nabla}_y \Phi(x_k, y_k; \omega_k^y) - \tilde{\nabla}_y \Phi(x_{k-1}, y_{k-1}; \omega_{k-1}^y)$
  $\tilde{s}_k \leftarrow \tilde{\nabla}_y \Phi(x_k, y_k; \omega_k^y) + \theta \tilde{q}_k$
  $y_{k+1} \leftarrow \mathbf{prox}_{\sigma g}(y_k + \sigma \tilde{s}_k)$
  $x_{k+1} \leftarrow \mathbf{prox}_{\tau f}(x_k - \tau \tilde{\nabla}_x \Phi(x_k, y_{k+1}; \omega_k^x))$
end for

We make the following assumption on the statistical nature of the gradient noise.

Assumption 1.2. *There exist* $\delta_x, \delta_y \geq 0$ *such that for all* $k \geq 0$, *given* (x_k, y_k, y_{k+1}), *the stochastic gradients* $\tilde{\nabla}_x \Phi(x_k, y_{k+1}; \omega_k^x)$, $\tilde{\nabla}_y \Phi(x_k, y_k; \omega_k^y)$ *and random sequences* $\{\omega_k^x\}_k$, $\{\omega_k^y\}_k$ *satisfy the following conditions: (i)* $\{\omega_k^x\}_k$ *and* $\{\omega_k^y\}_k$ *are independent sequences and also are independent from each other; (ii)* $\mathbb{E}[\tilde{\nabla}_x \Phi(x_k, y_{k+1}; \omega_k^x) | x_k, y_{k+1}] = \nabla_x \Phi(x_k, y_{k+1})$; *(iii)* $\mathbb{E}[\tilde{\nabla}_y \Phi(x_k, y_k; \omega_k^y) | x_k, y_k] = \nabla_y \Phi(x_k, y_k)$; *(iv)* $\mathbb{E}[\|\tilde{\nabla}_x \Phi(x_k, y_{k+1}; \omega_k^x) - \nabla_x \Phi(x_k, y_{k+1})\|^2 | x_k, y_{k+1}] \leq \delta_x^2$; *(v)* $\mathbb{E}[\|\tilde{\nabla}_y \Phi(x_k, y_k; \omega_k^y) - \nabla_y \Phi(x_k, y_k)\|^2 | x_k, y_k] \leq \delta_y^2$.

Assumption 1.2 applies to most unbiased situations. For example, when $\mathbf{dom}\, f \times \mathbf{dom}\, g$ is compact, for $\{\omega_k^x\} \subset \mathcal{X}^*$ and $\{\omega_k^y\} \subset \mathcal{Y}^*$ having zero-mean and finite-variance, the following additive noise model is a special case of Assumption 1.2: $\tilde{\nabla}_x \Phi(x_k, y_{k+1}) = \nabla_x \Phi(x_k, y_{k+1}) + \omega_k^x$, $\tilde{\nabla}_y \Phi(x_k, y_k) = \nabla_y \Phi(x_k, y_k) + \omega_k^y$. This type of noise arises in the context of privacy-preserving algorithms. Indeed, when $\nabla \Phi$ is associated with the user data, the user would inject additive noise to the gradients for protecting data privacy, see e.g., [23, Alg. 1], [20]. This unbiased noise structure

with a finite variance also arises in SP formulation of ERM problems if the gradients are estimated from mini-batches on bounded domains, e.g., [34, 42].

Finally, for the non-smooth case, i.e., $f(\cdot) \not\equiv 0$ or $g(\cdot) \not\equiv 0$, to derive bounds on the gap function $\mathcal{G}(\bar{x}_k, \bar{y}_k)$, defined in (1.2), we assume compact domains for f and g.

ASSUMPTION 1.3. *Suppose that* $\mathbf{dom}\, f \times \mathbf{dom}\, g \subset \mathcal{X} \times \mathcal{Y}$ *is compact; hence,* $\sup_{x_1,x_2 \in \mathbf{dom}\, f} \|x_1 - x_2\|^2 \leq \Omega_f$ *and* $\sup_{y_1,y_2 \in \mathbf{dom}\, g} \|y_1 - y_2\|^2 \leq \Omega_g$ *for some* $\Omega_f, \Omega_g > 0$.

It is worth pointing out that this compactness assumption is not required for deriving rate results in terms of $\mathcal{D}(x_k, y_k)$ for the SCSC scenario. Furthermore, in Theorem 2.6 we extend our rate results for the smooth case without assuming compactness – we are not aware of such result in the literature for the gap function. To our knowledge, existing results that can avoid the compactness assumption primarily work with the weaker[2] performance metric $\mathcal{D}$, see e.g. [11].

2. Performance Guarantees for SAPD. Under our noise model (Assumption 1.2), we next provide performance guarantees for the SAPD algorithm.

THEOREM 2.1. *Suppose Assumptions 1.1 and 1.2 hold, and* $\{x_k, y_k\}_{k\geq 0}$ *are generated by SAPD, stated in Algorithm 1.1, using* $\tau, \sigma > 0$ *and* $\theta \geq 0$ *that satisfy*

$$G \triangleq \begin{pmatrix} \frac{1}{\tau} + \mu_x - \frac{1}{\rho\tau} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\sigma} + \mu_y - \frac{1}{\rho\sigma} & (\frac{\theta}{\rho} - 1)L_{yx} & (\frac{\theta}{\rho} - 1)L_{yy} & 0 \\ 0 & (\frac{\theta}{\rho} - 1)L_{yx} & \frac{1}{\tau} - L_{xx} & 0 & -\frac{\theta}{\rho}L_{yx} \\ 0 & (\frac{\theta}{\rho} - 1)L_{yy} & 0 & \frac{1}{\sigma} - \alpha & -\frac{\theta}{\rho}L_{yy} \\ 0 & 0 & -\frac{\theta}{\rho}L_{yx} & -\frac{\theta}{\rho}L_{yy} & \frac{\alpha}{\rho} \end{pmatrix} \succeq 0 \tag{2.1}$$

for some $\alpha \in [0, \frac{1}{\sigma}]$ *and* $\rho \in (0, 1]$. *Then for any compact convex set* $X \times Y \subset \mathbf{dom}\, f \times \mathbf{dom}\, g$ *such that* $x_0 \in X$ *and* $y_0 \in Y$, *and for any* $\eta_x, \eta_y \geq 0$,

$$\mathbb{E}\Big[\sup_{(x,y)\in X\times Y} \Big\{\mathcal{L}(\bar{x}_N, y) - \mathcal{L}(x, \bar{y}_N) + \frac{\rho^{-N+1}}{K_N(\rho)}\Delta_N(x,y)\Big\}\Big] \leq \frac{1}{K_N(\rho)}\Omega_{\tau,\sigma,\theta} + \Xi_{\tau,\sigma,\theta}, \tag{2.2}$$

holds for $N \geq 1$, *where* $\Omega_X \triangleq \sup_{x_1,x_2\in X}\|x_1 - x_2\|^2$, $\Omega_Y \triangleq \sup_{y_1,y_2\in Y}\|y_1 - y_2\|^2$,

$$\Delta_N(x,y) \triangleq \frac{1}{2\rho\tau}\|x_N - x\|^2 + \frac{1}{2\rho\sigma}(1 - \alpha\sigma)\|y_N - y\|^2,$$

$$\Omega_{\tau,\sigma,\theta} \triangleq \frac{1}{2}(\frac{1}{\tau} + \eta_x)\Omega_X + \frac{1}{2}(\frac{1}{\sigma} + (1+2\theta)\eta_y)\Omega_Y,$$

$$\Xi_{\tau,\sigma,\theta} \triangleq \Big(\frac{\tau}{1+\tau\mu_x}\Xi^x_{\tau,\sigma,\theta} + \frac{1}{2\eta_x}\Big)\delta_x^2 + \Big(\frac{\sigma}{1+\sigma\mu_y}\Xi^y_{\tau,\sigma,\theta} + \frac{1+2\theta}{2\eta_y}\Big)\delta_y^2,$$

$$\Xi^x_{\tau,\sigma,\theta} \triangleq \left(1 + \frac{\sigma\theta(1+\theta)L_{yx}}{2(1+\sigma\mu_y)}\right),$$

$$\Xi^y_{\tau,\sigma,\theta} \triangleq \left(1 + 2\theta + \frac{\theta + \sigma\theta(1+\theta)L_{yy}}{1+\sigma\mu_y} + \frac{\tau\sigma\theta(1+\theta)L_{yx}L_{xy}}{(1+\tau\mu_x)(1+\sigma\mu_y)}\right)(1+2\theta) + \frac{\tau\theta(1+\theta)L_{yx}}{2(1+\tau\mu_x)},$$

and $\bar{z}_N = (\bar{x}_N, \bar{y}_N)$ *such that* $\bar{z}_N = \frac{1}{K_N(\rho)}\sum_{k=1}^N \rho^{-k+1} z_k$ *and* $K_N(\rho) \triangleq \sum_{k=0}^{N-1}\frac{1}{\rho^k}$; *thus,* $K_N(\rho) = \frac{1-\rho^N}{1-\rho}\frac{1}{\rho^{N-1}}$ *for* $\rho \in (0,1)$ *and* $K_N(1) = N$ *for* $\rho = 1$. *Furthermore, when* $\mu_x, \mu_y > 0$, *there exists* $\rho \in (0,1)$ *such that for* $N \geq 1$,

$$\mathbb{E}[\Delta_N(x^*, y^*)] \leq \rho^{N-1}\Delta_{\tau,\sigma,\theta} + \frac{1-\rho^N}{1-\rho}\Xi_{\tau,\sigma,\theta}, \tag{2.3}$$

[2]For SCSC problems, $\mathcal{G}(\bar{x}, \bar{y}) \leq \epsilon$ implies that $\mathcal{D}(\bar{x}, \bar{y}) = \mathcal{O}(\epsilon)$; but, the reverse is not true.

where $\Delta_{\tau,\sigma,\theta} \triangleq \frac{1}{2}(\frac{1}{\tau}+\eta_x)\|x_0-x^*\|^2+\frac{1}{2}(\frac{1}{\sigma}+(1+2\theta)\eta_y)\|y_0-y^*\|^2$.

Proof. See Appendix A.1. ∎

Remark 2.2. Since $\alpha \in [0, \frac{1}{\sigma}]$, we have $\Delta_N(x,y) \geq 0$ for all $(x,y) \in X \times Y$. Furthermore, for the unique saddle point (x^*, y^*), we also have $\mathcal{L}(\bar{x}_N, y^*) - \mathcal{L}(x^*, \bar{y}_N) \geq 0$ for $N \geq 1$. Therefore, if $(x^*, y^*) \in X \times Y$, then (2.2) implies that

$$\mathbb{E}\Big[\max\Big\{\sup_{(x,y)\in X\times Y}\{\mathcal{L}(\bar{x}_N, y) - \mathcal{L}(x, \bar{y}_N)\},\ \frac{\rho^{-N+1}}{K_N(\rho)}\Delta_N(x^*, y^*)\Big\}\Big] \leq \frac{1}{K_N(\rho)}\Omega_{\tau,\sigma,\theta} + \Xi_{\tau,\sigma,\theta}.$$

Remark 2.3. Consider the stochastic case, i.e., $\delta_x, \delta_y > 0$. The variance term $\Xi_{\tau,\sigma,\theta}$ depends on algorithm parameters $\{\tau, \sigma, \theta\}$ and two other free parameters $\eta_x, \eta_y > 0$. For this scenario, we set $\eta_x = \frac{1}{\tau} + \mu_x$ and $\eta_y = \frac{1}{\sigma} + \mu_y$. This choice of parameters implies that as the primal and dual step sizes approach 0, the variance term vanishes as expected, i.e., as $\tau \to 0$ and $\sigma \to 0$, we have $\Xi_{\tau,\sigma,\theta} \to 0$.

Remark 2.4. If we set $\theta = \rho$ in (2.1), we obtain a simpler matrix inequality:

$$\min\{\tau\mu_x,\ \sigma\mu_y\} \geq \frac{1-\theta}{\theta}, \qquad \begin{pmatrix} \frac{1}{\tau} - L_{xx} & 0 & -L_{yx} \\ 0 & \frac{1}{\sigma} - \alpha & -L_{yy} \\ -L_{yx} & -L_{yy} & \frac{\alpha}{\theta} \end{pmatrix} \succeq 0. \tag{2.4}$$

The SAPD complexity analysis is mainly based on the simpler system in (2.4).

Consider the SCSC setting. Under Assumption 1.3, $\mathbf{dom}\, f \times \mathbf{dom}\, g$ is compact; thus, setting $X = \mathbf{dom}\, f$ and $Y = \mathbf{dom}\, g$ together with $\Omega_X = \Omega_f$ and $\Omega_Y = \Omega_g$ within Theorem 2.1 implies that $\mathcal{G}(\bar{x}_N, \bar{y}_N)$ and $\mathcal{D}(x_N, y_N)$ are both bounded by a sum of two terms: *bias term* $\frac{1}{K_N(\rho)}\Omega_{\tau,\sigma,\theta}$ that goes to zero as $N \to \infty$ and a constant *variance term* $\Xi_{\tau,\sigma,\theta}$, which can be controlled by properly selecting τ, σ and θ.

When $\theta = 0$, SAPD reduces to SGDA, of which step size conditions can be obtained immediately from (2.1) by setting $\theta = \alpha = 0$. For details, see Appendix B.2.

2.1. Relaxing Compact Domain Assumption for Smooth Problems. For *smooth* problems, one can still bound $\mathcal{G}(\bar{x}_N, \bar{y}_N)$ without assuming a compact domain; indeed, the bound (2.2) provided in Theorem 2.1 would continue to hold with slightly modified bias and variance terms even when $X = \mathcal{X}$ and $Y = \mathcal{Y}$. First, we provide a technical result inspired by Wuertz's implicit function theorem [37]. This result establishes Lipschitz continuity of the optimizers of smooth SCSC coupling functions.

LEMMA 2.5. *Suppose* Φ^s *satisfies Assumption 1.1 over* $\mathcal{X} \times \mathcal{Y}$ *with constants* $\{L^s_{xx}, L^s_{xy}, L^s_{yx}, L^s_{yy}\}$ *and we also assume that* $\Phi^s(\cdot, y)$ *is strongly convex with modulus* $\mu_x > 0$ *for all* $y \in \mathcal{Y}$ *and* $\Phi^s(x, \cdot)$ *is strongly concave with modulus* $\mu_y > 0$ *for all* $x \in \mathcal{X}$. *For any given* $y \in \mathcal{Y}$, *let* $x^*(y) \triangleq \boldsymbol{\mathrm{argmin}}_{x\in\mathcal{X}}\, \Phi^s(x, y)$; *similarly, for any given* $x \in \mathcal{X}$, *let* $y^*(x) \triangleq \boldsymbol{\mathrm{argmax}}_{y\in\mathcal{Y}}\, \Phi^s(x, y)$. *Then* $x^*(\cdot)$ *and* $y^*(\cdot)$ *are Lipschitz maps on* $\mathcal{Y}$ *and* $\mathcal{X}$, *with constants* $L_{x^*} \triangleq \sqrt{2\kappa_{xy}^2 + 1}$ *and* $L_{y^*} \triangleq \sqrt{2\kappa_{yx}^2 + 1}$, *respectively, where* $\kappa_{xy} \triangleq L^s_{xy}/\mu_x$ *and* $\kappa_{yx} \triangleq L^s_{yx}/\mu_y$.

Based on this lemma, whose proof is provided in Appendix A.2, we can obtain the following performance bound without assuming compactness of the domain.

THEOREM 2.6. *Consider* (P) : $\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \Phi^s(x, y)$ *for* Φ^s *satisfying the premises of Lemma 2.5, and we assume that there is a stochastic oracle for* Φ^s *satisfying Assumption 1.2. Setting* $\Phi(x, y) \triangleq \Phi^s(x, y) - \frac{\mu_x}{2}\|x\|^2 + \frac{\mu_y}{2}\|y\|^2$, $f(x) = \frac{\mu_x}{2}\|x\|^2$

and $g(y) = \frac{\mu_y}{2}\|y\|^2$*, let* $\{x_k, y_k\}_{k\geq 0}$ *be generated by SAPD when applied to* (1.1) *using* $\tau, \sigma \geq 0$ *and* $\theta \geq 0$ *that satisfy* (2.1) *for some* $\alpha \in [0, \frac{1}{\sigma})$ *and* $\rho \in (0,1)$ *with* $L_{xx} = L^s_{xx} - \mu_x$*,* $L_{yy} = L^s_{yy} - \mu_y$*,* $L_{xy} = L^s_{xy}$ *and* $L_{yx} = L^s_{yx}$*. Then,*

$$\mathbb{E}\Big[\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \{\mathcal{L}(\bar{x}_N, y) - \mathcal{L}(x, \bar{y}_N)\}\Big] \leq \frac{(C_{\tau,\sigma,\theta} N\rho^N + 1)}{K_N(\rho)}\Delta_{\tau,\sigma,\theta} + (C_{\tau,\sigma,\theta}\rho^N + 1)\Xi_{\tau,\sigma,\theta},$$

for all $N \geq 1$*, where* $C_{\tau,\sigma,\theta} \triangleq \frac{2\sigma}{1-\alpha\sigma}(\frac{1}{\tau} + \eta_x)L^2_{x^*} + 2\tau(\frac{1}{\sigma} + \eta_y(1+2\theta))L^2_{y^*}$*, and both* $\Delta_{\tau,\sigma,\theta}$ *and* $K_N(\rho)$ *are defined in Theorem 2.1.*

Proof. Note that in the proof of Theorem 2.1, we only used compactness of $X \times Y$ to get the bound in (A.30). Here we show that for (P), we can remove this assumption by using the result in (2.3). Recall $\mathcal{L}$ defined in (1.1), we have $\mathcal{L} = \Phi^s$. Thus, given $(\bar{x}_N, \bar{y}_N) \in \mathcal{X}\times\mathcal{Y}$, $(x^*(\bar{y}_N), y^*(\bar{x}_N)) \triangleq \mathbf{argmax}_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \mathcal{L}(\bar{x}_N, y) - \mathcal{L}(x, \bar{y}_N)$ uniquely exists. Since $\Delta_N(x,y) \geq 0$ and $U_N(x,y) \leq \frac{1}{2\tau}\|x - x_0\|^2 + \frac{1}{2\sigma}\|y - y_0\|^2$ with probability (w.p.) 1 for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$, where U_N is as in (A.26). It follows from (A.26) that

$$K_N(\rho) \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \{\mathcal{L}(\bar{x}_N, y) - \mathcal{L}(x, \bar{y}_N)\} \leq \bar{U}(x^*(\bar{y}_N), y^*(\bar{x}_N)) + \sum_{k=0}^{N-1} \rho^{-k}(P_k + Q_k), \tag{2.5}$$

where $\bar{U}(x,y) \triangleq \frac{1}{2}(\frac{1}{\tau} + \eta_x)\|x - x_0\|^2 + \frac{1}{2}(\frac{1}{\sigma} + \eta_y(1+2\theta))\|y - y_0\|^2$. Using Lemma 2.5 and the relations $x^* = \mathbf{argmin}_{x\in\mathcal{X}} \Phi^s(x, y^*)$ and $y^* = \mathbf{argmax}_{y\in\mathcal{Y}} \Phi^s(x^*, y)$, we can bound $\bar{U}(x^*(\bar{y}_N), y^*(\bar{x}_N))$ as follows:

$$\begin{aligned}
&\bar{U}(x^*(\bar{y}_N), y^*(\bar{x}_N)) \\
&\leq (\frac{1}{\tau} + \eta_x)(\|x^*(\bar{y}_N) - x^*\|^2 + \|x_0 - x^*\|^2) + (\frac{1}{\sigma} + \eta_y(1+2\theta))(\|y^*(\bar{x}_N) - y^*\|^2 + \|y_0 - y^*\|^2) \\
&\leq \Big(\frac{1}{\tau} + \eta_x\Big)L^2_{x^*}\|\bar{y}_N - y^*\|^2 + \Big(\frac{1}{\sigma} + \eta_y(1+2\theta)\Big)L^2_{y^*}\|\bar{x}_N - x^*\|^2 + \Delta_{\tau,\sigma,\theta}.
\end{aligned}$$

When $\mu_x, \mu_y > 0$, one still has (2.3) even without the compactness assumption, which implies $\mathbb{E}[\frac{1}{2\tau}\|x_k - x^*\|^2 + \frac{1}{2\sigma}(1-\alpha\sigma)\|y_k - y^*\|^2] \leq \rho^k\Delta_{\tau,\sigma,\theta} + \frac{\rho}{1-\rho}\Xi_{\tau,\sigma,\theta}$ for all $k \geq 1$. Since $\|\cdot\|^2$ is convex and $\bar{z}_N = (\bar{x}_N, \bar{y}_N)$ such that $\bar{z}_N = \dfrac{1}{K_N(\rho)}\sum_{k=1}^N \rho^{-k+1} z_k$, we get

$$\mathbb{E}[\frac{1}{2\tau}\|\bar{x}_N - x^*\|^2 + \frac{1}{2\sigma}(1-\alpha\sigma)\|\bar{y}_N - y^*\|^2] \leq N\rho^N\Delta_{\tau,\sigma,\theta} + \frac{\rho}{1-\rho}\Xi_{\tau,\sigma,\theta}, \quad \forall N \geq 1.$$

Therefore, $\mathbb{E}[\bar{U}(x^*(\bar{y}_N), y^*(\bar{x}_N))] \leq (C_{\tau,\sigma,\theta}N\rho^N + 1)\Delta_{\tau,\sigma,\theta} + \frac{\rho}{1-\rho}C_{\tau,\sigma,\theta}\Xi_{\tau,\sigma,\theta}$ for all $N \geq 1$, which combined with (2.5) and (A.31) implies the desired result. ∎

2.2. Parameter Choices for SAPD. We employ the matrix inequality (MI) in (2.4) to describe the admissible set of algorithm parameters that guarantee convergence. Our aim is to enjoy as wide a range of parameters as possible to improve the robustness of SAPD, i.e., to control the noise amplification of the algorithm, in the presence of noisy gradients. Although, it seems difficult to find an explicit solution to the MI in Theorem 2.1, we can compute a particular solution to it by exploiting its structure. Next, in Lemma 2.7, we give an intermediate condition to help us construct the particular solution provided in Corollary 2.8 for the SCSC setting.

LEMMA 2.7. *Let* $\tau, \sigma > 0$*,* $\theta \in (0,1]$ *be a solution to the following system:*

$$\min\{\tau\mu_x, \sigma\mu_y\} \geq \frac{1-\theta}{\theta}, \quad \frac{1}{\tau} \geq L_{xx} + \pi_1 L_{yx}, \quad \frac{1}{\sigma} \geq \frac{\theta L_{yx}}{\pi_1} + (\frac{\theta}{\pi_2} + \pi_2)L_{yy}, \tag{2.6}$$

for some $\pi_1, \pi_2 > 0$*. Then* $\{\tau, \sigma, \theta, \alpha\}$ *is a solution to* (2.4) *for* $\alpha = \frac{\theta L_{yx}}{\pi_1} + \frac{\theta L_{yy}}{\pi_2}$*.*

Proof. Since the first inequalities in both (2.6) and (2.4) are the same, we only need to show the MI in (2.4) holds. Substituting $\alpha = \frac{\theta L_{yx}}{\pi_1} + \frac{\theta L_{yy}}{\pi_2}$ into (2.4), we get

$$\begin{pmatrix} \frac{1}{\tau} - L_{xx} & 0 & -L_{yx} \\ 0 & \frac{1}{\sigma} - \alpha & -L_{yy} \\ -L_{yx} & -L_{yy} & \frac{\alpha}{\theta} \end{pmatrix} = \begin{pmatrix} \frac{1}{\tau} - L_{xx} & 0 & -L_{yx} \\ 0 & \frac{1}{\sigma} - \frac{\theta L_{yx}}{\pi_1} - \frac{\theta L_{yy}}{\pi_2} & -L_{yy} \\ -L_{yx} & -L_{yy} & \frac{L_{yx}}{\pi_1} + \frac{L_{yy}}{\pi_2} \end{pmatrix} \triangleq \quad M_1 + M_2,$$

where $M_1 \triangleq \begin{pmatrix} \frac{1}{\tau} - L_{xx} & 0 & -L_{yx} \\ 0 & 0 & 0 \\ -L_{yx} & 0 & \frac{L_{yx}}{\pi_1} \end{pmatrix}$ and $M_2 \triangleq \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{\sigma} - \frac{\theta L_{yx}}{\pi_1} - \frac{\theta L_{yy}}{\pi_2} & -L_{yy} \\ 0 & -L_{yy} & \frac{L_{yy}}{\pi_2} \end{pmatrix}$. Therefore, since $\pi_1, \pi_2 > 0$, the Schur complement conditions in (2.6), i.e., the second and the third inequalities, imply $M_1 \succeq 0$ and $M_2 \succeq 0$, respectively. Thus, $M_1 + M_2 \succeq 0$.□

Lemma 2.7 shows that every solution to (2.6) can be converted to a solution to (2.4). Next, based on this lemma, we will give an explicit parameter choice for Algorithm 1.1.

COROLLARY 2.8. *Suppose $\mu_x, \mu_y > 0$. If $L_{yy} > 0$, for any given $\beta \in (0,1)$, let $\tau, \sigma > 0$ and $\theta \in (0,1)$ be chosen satisfying*

$$\tau = \frac{1-\theta}{\mu_x \theta}, \quad \sigma = \frac{1-\theta}{\mu_y \theta}, \quad \theta \geq \bar{\theta} \triangleq \max\{\bar{\theta}_1,\ \bar{\theta}_2\}, \tag{2.7}$$

where $\bar{\theta}_1,\ \bar{\theta}_2 \in (0,1)$, depending on the choice of β, are defined as

$$\bar{\theta}_1 \triangleq 1 - \frac{\beta(L_{xx}+\mu_x)\mu_y}{2L_{yx}^2}\Big(\sqrt{1 + \frac{4\mu_x L_{yx}^2}{\beta\mu_y(L_{xx}+\mu_x)^2}} - 1\Big), \quad \bar{\theta}_2 \triangleq 1 - \frac{(1-\beta)^2}{8}\frac{\mu_y^2}{L_{yy}^2}\Big(\sqrt{1 + \frac{16L_{yy}^2}{(1-\beta)^2\mu_y^2}} - 1\Big). \tag{2.8}$$

On the other hand, if $L_{yy} = 0$, let $\tau, \sigma > 0$ and $\theta \in (0,1)$ be chosen as in (2.7) for $\bar{\theta}_1$ in (2.8) with $\beta = 1$ and $\bar{\theta}_2 = 0$.[3] *Then $\alpha = \frac{1}{\sigma} - \sqrt{\theta} L_{yy} > 0$, and $\{\tau, \sigma, \theta, \alpha\}$ is a solution to (2.4). Moreover, when $L_{yy} > 0$, the minimum $\bar{\theta}$ is attained at unique $\beta^* \in (0,1)$ such that $\bar{\theta}_1 = \bar{\theta}_2$.*

Proof. Consider arbitrary $\tau, \sigma, \pi_1, \pi_2 > 0$ and $\theta \in (0,1)$. By a straightforward calculation, $\{\tau, \sigma, \theta, \pi_1, \pi_2\}$ is a solution to (2.6) if and only if

$$\tau \geq \frac{1-\theta}{\theta\mu_x}, \quad \sigma \geq \frac{1-\theta}{\theta\mu_y}, \quad \pi_1 \geq \frac{\sigma\theta L_{yx}}{1 - \sigma(\pi_2 + \frac{\theta}{\pi_2})L_{yy}}, \tag{2.9a}$$

$$\sigma(\pi_2 + \frac{\theta}{\pi_2})L_{yy} < 1, \quad \frac{1}{\tau} - L_{xx} \geq \pi_1 L_{yx}. \tag{2.9b}$$

In the remainder of the proof, we fix (π_1, π_2) as follows:

$$\pi_1 = \frac{\sigma\theta L_{yx}}{1 - \sigma\left(\pi_2 + \frac{\theta}{\pi_2}\right)L_{yy}}, \quad \pi_2 = \sqrt{\theta}. \tag{2.10}$$

Note the definition of $\bar{\theta}$ implies that $\bar{\theta} \in (0,1)$. Next, we show that $\theta \in [\bar{\theta}, 1)$ implies $\pi_1, \pi_2 > 0$; furthermore, we also show that $\tau, \sigma > 0$ defined as in (2.7) for $\theta \in [\bar{\theta}, 1)$ together with (π_1, π_2) as in (2.10) is a solution to (2.9).

[3] Our parameter selection when $L_{yy} = L_{xx} = 0$ recovers (τ, σ, θ) choice in [7, Eq.(49)].

First, setting τ, σ as in (2.7) and π_1, π_2 as in (2.10) imply that (2.9a) is trivially satisfied. Next, by substituting $\{\tau, \sigma, \pi_1, \pi_2\}$, chosen as in (2.7) and (2.10), into (2.9b), we conclude that $\{\tau, \sigma, \theta, \pi_1, \pi_2\}$ satisfies (2.9) for any $\theta \in (0,1)$ such that

$$\frac{2L_{yy}}{\mu_y} \cdot \frac{1-\theta}{\sqrt{\theta}} \leq 1-\beta, \tag{2.11}$$

$$\frac{\theta\mu_x}{1-\theta} - L_{xx} \geq (1-\theta)\frac{L_{yx}^2}{\mu_y} \cdot \Big(1 - \frac{2L_{yy}}{\mu_y} \cdot \frac{1-\theta}{\sqrt{\theta}}\Big)^{-1}, \tag{2.12}$$

for some $\beta \in (0,1]$. Clearly, a sufficient condition for (2.12) is

$$\frac{\theta\mu_x}{1-\theta} - L_{xx} \geq (1-\theta)\frac{L_{yx}^2}{\mu_y} \cdot \frac{1}{\beta}. \tag{2.13}$$

Note that (2.11) implies that $\pi_1 > 0$. We also have $\pi_2 = \sqrt{\theta} > 0$ trivially.

When $L_{yy} > 0$, given any $\beta \in (0,1)$, solving (2.11) and (2.13) for $\theta \in (0,1)$, we get the third condition in (2.7). Indeed, it can be checked that $\theta \in [\bar{\theta}_2, 1)$ satisfies (2.11) and $\theta \in [\bar{\theta}_1, 1)$ satisfies (2.13); thus, $\theta \in [\bar{\theta}, 1)$ satisfies (2.11) and (2.13) simultaneously. Moreover, when $L_{yy} = 0$, one does not need to solve (2.11) as the first inequality in (2.9b) holds trivially; thus, the only condition on θ comes from (2.12) which is equivalent to (2.13) with $\beta = 1$. The rest follows from Lemma 2.7 by setting $\alpha = \frac{\theta L_{yx}}{\pi_1} + \frac{\theta L_{yy}}{\pi_2}$. Indeed, the particular choice of (π_1, π_2) in (2.10) gives us $\alpha = \frac{1}{\sigma} - \sqrt{\theta}L_{yy}$. Finally, it can be verified that $\bar{\theta}_1 : [0,1] \to \mathbb{R}$ and $\bar{\theta}_2 : [0,1] \to \mathbb{R}$ are monotonically decreasing and monotonically increasing functions of β, respectively. Since $\bar{\theta}_1(0) = 1 > \bar{\theta}_2(0)$ and $\bar{\theta}_2(1) = 1 > \bar{\theta}_1(1)$, $\bar{\theta}$ obtains its minimum at the unique $\beta^* \in (0,1)$ such that $\bar{\theta}_1(\beta^*) = \bar{\theta}_2(\beta^*)$. ∎

To compute this particular solution, we set $\rho = \theta$; hence, θ is not only the momentum parameter, but also determines the linear rate for the bias term in (2.2).

2.3. Iteration Complexity Bound for SAPD. In this part, we study the iteration complexity bound for SAPD, to compute $(\bar{x}_N, \bar{y}_N)$ such that $\mathcal{G}(\bar{x}_N, \bar{y}_N) \leq \epsilon$ where $\epsilon > 0$ is a given tolerance and $\mathcal{G}(\cdot,\cdot)$ denotes the gap function defined in (1.2).

THEOREM 2.9. *Suppose $\mu_x, \mu_y > 0$, and Assumptions 1.1, 1.2 and 1.3 hold. For any $\epsilon > 0$ and $\beta \in (0,1]$, suppose the parameters $\{\tau, \sigma, \theta\}$ are chosen such that*

$$\tau = \frac{1-\theta}{\mu_x\theta}, \quad \sigma = \frac{1-\theta}{\mu_y\theta}, \quad \theta = \max\{\bar{\theta},\ \bar{\bar{\theta}}\} \tag{2.14}$$

where $\bar{\theta}$, defined in (2.7), depends on $\beta \in (0,1]$, and

$$\bar{\bar{\theta}} \triangleq \max\Big\{\bar{\bar{\theta}}_1,\ \bar{\bar{\theta}}_2\Big\}, \quad \bar{\bar{\theta}}_1 = \max\{0, 1 - \frac{2}{3\bar{\Xi}^x(\beta)}\frac{\mu_x}{\delta_x^2}\epsilon\}, \quad \bar{\bar{\theta}}_2 = \max\{0, 1 - \frac{2}{3\bar{\Xi}^y(\beta)}\frac{\mu_y}{\delta_y^2}\epsilon\}, \tag{2.15}$$

such that $\bar{\Xi}^x(\beta) \triangleq 3 + 2\Psi(\beta)$ and $\bar{\Xi}^y(\beta) \triangleq 33 + 6\beta\left(2\frac{L_{xy}}{L_{yx}} - 1\right) + 2\frac{\mu_y}{\mu_x}\Psi(\beta)$ with $\Psi(\beta) \triangleq \min\left\{\sqrt{\frac{\beta\mu_x}{\mu_y}}, \frac{1-\beta}{2}\frac{L_{yx}}{L_{yy}}\right\}$. Then, the iteration complexity of SAPD, as stated in Algorithm 1.1, to generate $(\bar{x}, \bar{y}) \in \mathcal{X} \times \mathcal{Y}$ such that $\mathcal{G}(\bar{x}, \bar{y}) \leq \epsilon$ is

$$\mathcal{O}\Big(\Big[\frac{L_{xx}}{\mu_x} + \frac{L_{yx}}{\sqrt{\mu_x\mu_y}} + \frac{L_{yy}}{\mu_y} + \Big(\Big(1 + \sqrt{\frac{\mu_x}{\mu_y}}\Big)\frac{\delta_x^2}{\mu_x} + \Big(1 + \frac{L_{xy}}{L_{yx}} + \sqrt{\frac{\mu_y}{\mu_x}}\Big)\frac{\delta_y^2}{\mu_y}\Big)\frac{1}{\epsilon}\Big] \cdot \ln\Big(\frac{\mu_x\Omega_f + \mu_y\Omega_g}{\epsilon} + 1\Big)\Big).$$

Proof. Given $\beta \in (0,1]$, let $\{\tau, \sigma, \theta, \alpha\}$ be a particular solution to (2.4) constructed according to Corollary 2.8. Therefore, using these particular parameter values together with $\rho = \theta$ within Theorem 2.1, we know that (2.2) holds. Since the result

is true for any $\eta_x, \eta_y > 0$, we set $\eta_x = \frac{1}{\tau} + \mu_x$ and $\eta_y = \frac{1}{\sigma} + \mu_y$. Thus, for any $\epsilon > 0$, the right side of (2.2) can be bounded by ϵ when

$$\frac{1}{K_N(\theta)}\Omega_{\tau,\sigma,\theta} \le \frac{\epsilon}{3}, \quad \frac{\tau}{1+\tau\mu_x}(1+2\Xi^x_{\tau,\sigma,\theta}) \le \frac{2}{3\delta_x^2}\epsilon, \quad \frac{\sigma}{1+\sigma\mu_y}(1+2\theta+2\Xi^y_{\tau,\sigma,\theta}) \le \frac{2}{3\delta_y^2}\epsilon. \tag{2.16}$$

Note (2.14) implies that $\frac{\tau}{1+\tau\mu_x} = \frac{1-\theta}{\mu_x}$ and $\frac{\sigma}{1+\sigma\mu_y} = \frac{1-\theta}{\mu_y}$; moreover, the other two terms that depend on θ satisfy $1 + \Xi^x_{\tau,\sigma,\theta} > 1$ and $1 + 2\theta + \Xi^y_{\tau,\sigma,\theta} > 1$. Therefore, to get a sufficient condition on θ for the last two inequalities in (2.16) to hold, we first upper bound $\Xi^x_{\tau,\sigma,\theta}$ and $\Xi^y_{\tau,\sigma,\theta}$, and then exploit $\frac{\tau}{1+\tau\mu_x} = \frac{1-\theta}{\mu_x}$ and $\frac{\sigma}{1+\sigma\mu_y} = \frac{1-\theta}{\mu_y}$ to get a lower bound on θ.

The parameter choice of τ and σ in (2.14) implies $\Xi^x_{\tau,\sigma,\theta} = 1 + \theta(1-\theta^2)\frac{L_{yx}}{2\mu_y}$, and

$$\Xi^y_{\tau,\sigma,\theta} = \Big((1+\theta)^2 + \theta(1-\theta^2)\frac{L_{yy}}{\mu_y} + \theta(1+\theta)(1-\theta)^2\frac{L_{yx}L_{xy}}{\mu_x\mu_y}\Big)(1+2\theta) + \theta(1-\theta^2)\frac{L_{yx}}{2\mu_x}.$$

Since $0 < \theta \le 1$, we have $1 - \theta^2 \le 2(1-\theta)$; thus, $1 + 2\Xi^x_{\tau,\sigma,\theta} \le 3 + 2(1-\theta)\frac{L_{yx}}{\mu_y}$ and

$$1 + 2\theta + 2\Xi^y_{\tau,\sigma,\theta} \le 3\Big(9 + 4(1-\theta)\frac{L_{yy}}{\mu_y} + 4(1-\theta)^2\frac{L_{yx}L_{xy}}{\mu_x\mu_y}\Big) + 2(1-\theta)\frac{L_{yx}}{\mu_x}. \tag{2.17}$$

On the other hand, since $\theta \ge \bar{\theta} = \max\{\bar{\theta}_1, \bar{\theta}_2\}$, the inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for all $a, b \ge 0$, and (2.8) together imply that $1 - \theta \le \min\{\frac{\sqrt{\beta\mu_x\mu_y}}{L_{yx}}, (1-\beta)\frac{\mu_y}{2L_{yy}}\}$. Thus,

$$1 + 2\Xi^x_{\tau,\sigma,\theta} \le 3 + 2(1-\theta)\frac{L_{yx}}{\mu_y} \le \bar{\Xi}^x(\beta), \tag{2.18}$$

and within (2.17) bounding $(1-\theta)\frac{L_{yx}}{\mu_x}$ similarly and using $(1-\theta)^2 \le \frac{\beta\mu_x\mu_y}{L_{yx}^2}$, we get

$$1 + 2\theta + 2\Xi^y_{\tau,\sigma,\theta} \le 3\Big(9 + 4(1-\theta)\frac{L_{yy}}{\mu_y} + \frac{4\beta L_{xy}}{L_{yx}}\Big) + \frac{2\mu_y}{\mu_x}\min\Big\{\sqrt{\frac{\beta\mu_x}{\mu_y}}, \frac{1-\beta}{2}\frac{L_{yx}}{L_{yy}}\Big\}.$$

Next, it follows from $1 - \theta \le (1-\beta)\frac{\mu_y}{2L_{yy}}$ that

$$1 + 2\theta + 2\Xi^y_{\tau,\sigma,\theta} \le \bar{\Xi}^y(\beta). \tag{2.19}$$

Our (τ, σ) choice in (2.14) together with (2.18) and (2.19) imply that requiring $\frac{1-\theta}{\mu_x}\frac{\delta_x^2}{2}\ \bar{\Xi}^x(\beta) \le \frac{\epsilon}{3}$ and $\frac{1-\theta}{\mu_y}\frac{\delta_y^2}{2}\ \bar{\Xi}^y(\beta) \le \frac{\epsilon}{3}$ is sufficient for the last two inequalities in (2.16) to hold. It follows from the definition of $\bar{\bar{\theta}}$ in (2.15) that the sufficient condition holds for $\theta \ge \bar{\bar{\theta}}$. Thus, combining this requirement with $\theta \ge \bar{\theta}$ in Corollary 2.8, it follows that (2.14) provides us with a particular parameter choice for SAPD such that the last two inequalities in (2.16) hold. Indeed, our choice in (2.14) satisfies (2.4) which is a simpler LMI obtained by setting $\theta = \rho$ in (2.1); therefore, $\rho = \theta \in (0, 1)$ provides us with an upper bound on the actual convergence rate –see (2.2).

Next, to derive an upper complexity bound for SAPD, we only need to analyze how N should grow depending on ϵ such that the first inequality in (2.16) holds. It follows from (2.14), Theorem 2.1 and Assumption 1.3 that

$$\frac{1}{K_N(\theta)}\Omega_{\tau,\sigma,\theta} = \frac{\theta^{N-1}}{1-\theta^N}\Big((1+\theta)\mu_x\Omega_f + (1+3\theta)\mu_y\Omega_g\Big)/2 \le \frac{\theta^{N-1}}{1-\theta^N}\left(\mu_x\Omega_f + 2\mu_y\Omega_g\right).$$

Therefore, the first inequality in (2.16) holds for $N \ge 1 + \ln(3\left(\mu_x\Omega_f + 2\mu_y\Omega_g\right)/\epsilon + 1)/\ln(\frac{1}{\theta})$ as it implies $\frac{\theta^{N-1}}{1-\theta^{N-1}}\left(\mu_x\Omega_f + 2\mu_y\Omega_g\right) \le \frac{\epsilon}{3}$. Thus, SAPD can generate a point

$(\bar{x},\bar{y}) \in \mathcal{X}\times\mathcal{Y}$ such that $\mathcal{G}(\bar{x},\bar{y}) \le \epsilon$ within $N_\epsilon \in \mathbb{Z}_+$ iterations where

$$N_\epsilon = \mathcal{O}\Big(1+\ln\Big(\frac{\mu_x\Omega_f+\mu_y\Omega_g}{\epsilon}+1\Big)/\ln(\frac{1}{\theta})\Big). \tag{2.20}$$

In the remaining part of the proof, we will bound the term $\ln(\frac{1}{\theta})^{-1}$ in terms of given $\epsilon > 0$. According to (2.14), $\theta = \max\{\overline{\theta},\ \overline{\overline{\theta}}\}$; hence, it follows from (2.7) and (2.15) that $\theta \in \{\overline{\theta}_1, \overline{\theta}_2, \overline{\overline{\theta}}_1, \overline{\overline{\theta}}_2\} \subset (0,1)$. Since $\ln(1/\theta)$ is convex for $\theta \in \mathbb{R}_{++}$, we immediately get $\frac{1}{\ln(\frac{1}{\theta})} \le \frac{1}{1-\theta}$ for $\theta \in (0,1)$. Therefore, we trivially get the bound

$$\frac{1}{\ln(\frac{1}{\theta})} \le \max\{(1-\overline{\theta}_1)^{-1}, (1-\overline{\theta}_2)^{-1}, (1-\overline{\overline{\theta}}_1)^{-1}, (1-\overline{\overline{\theta}}_2)^{-1}\}. \tag{2.21}$$

First, we equivalently rewrite $(1-\overline{\theta}_1)^{-1}$ and $(1-\overline{\theta}_2)^{-1}$ as follows:

$$(1-\overline{\theta}_1)^{-1} = \frac{1}{2}\left(\frac{L_{xx}}{\mu_x}+1\right)+\sqrt{\frac{1}{4}\left(\frac{L_{xx}}{\mu_x}+1\right)^2+\frac{L_{yx}^2}{\beta\mu_x\mu_y}}, \quad (1-\overline{\theta}_2)^{-1} = \frac{1}{2}+\sqrt{\frac{1}{4}+\frac{4L_{yy}^2}{(1-\beta)^2\mu_y^2}};$$

finally, $(1-\overline{\overline{\theta}}_1)^{-1} = \frac{3}{2}\overline{\Xi}^x(\beta)\frac{\delta_x^2}{\mu_x}\frac{1}{\epsilon}$ and $(1-\overline{\overline{\theta}}_2)^{-1} = \frac{3}{2}\overline{\Xi}^y(\beta)\frac{\delta_y^2}{\mu_y}\frac{1}{\epsilon}$. Thus, using four identities we derived above within (2.21) and combining it with (2.20), we achieve the desired upper complexity bound for SAPD. ∎

Remark 2.10. Given $\epsilon > 0$, for sufficiently small $\delta_x^2 > 0$, (2.15) implies that $\overline{\overline{\theta}}_1 = 0$; similarly, $\overline{\overline{\theta}}_2 = 0$ for sufficiently small $\delta_y^2 > 0$. Therefore, when there is no noise for $\nabla_x\Phi$, i.e., $\delta_x^2 = 0$, we get $\overline{\overline{\theta}}_1 = 0$; similarly, $\overline{\overline{\theta}}_2 = 0$ if $\delta_y^2 = 0$.

Our bound's variance term (the term that depends on the noise levels δ_x^2 and δ_y^2) in Theorem 2.9 is optimal with respect to its dependency to ε up to a log factor, which can further be eliminated through employing a restarting strategy in the lines of our previous work [1]. In the following proposition, we derive the iteration complexity bound for the stochastic merely convex case, i.e., when $\mu_x = \mu_y = 0$. This result can be viewed as an immediate extension of the deterministic complexity result in [15] to the stochastic gradient setting considered in this paper.

Remark 2.11. Suppose $\mu_x = \mu_y = 0$, and the parameters $\tau, \sigma > 0$ and $\theta \in (0,1]$ satisfy (2.4). The first condition (2.4) implies that $\theta = 1$.

PROPOSITION 2.12. *Suppose $\mu_x, \mu_y = 0$, and Assumptions 1.1, 1.2 and 1.3 hold. For any $\epsilon > 0$, suppose $\{\tau,\sigma,\theta\}$ are chosen such that*

$$\tau = \min\Big\{\frac{1}{L_{yx}+L_{xx}},\ \frac{2}{15}\cdot\frac{\epsilon}{\delta_x^2}\Big\}, \quad \sigma = \min\Big\{\frac{1}{L_{yx}+2L_{yy}},\ \frac{1}{L_{xy}},\ \frac{1}{72}\cdot\frac{\epsilon}{\delta_y^2}\Big\}, \quad \theta = 1. \tag{2.22}$$

Then for $\mathcal{G}(\cdot,\cdot)$ is defined in (1.2), $\mathcal{G}(\bar{x}_N,\bar{y}_N) \le \epsilon$ for all $N \ge N_\epsilon$ such that

$$N_\epsilon = \mathcal{O}\Big(\frac{(L_{yx}+L_{xx})\Omega_f+\max\{L_{yx}+L_{yy},\ L_{xy}\}\Omega_g}{\epsilon}+\frac{\delta_x^2\Omega_f+\delta_y^2\Omega_g}{\epsilon^2}\Big).$$

Proof. Since $\mu_x = \mu_y = 0$, the first condition in (2.4) trivially holds for (τ,σ,θ) as in (2.22). Furthermore, (2.22) implies that $\frac{1}{\tau} \ge L_{yx}+L_{xx}$ and $\frac{1}{\sigma} \ge L_{yx}+2L_{yy}$; therefore, (τ,σ,θ) in (2.22) with $\pi_1 = \pi_2 = 1$ satisfy the conditions in Lemma 2.7. Thus, (τ,σ,θ) together with $\alpha = L_{yx}+L_{yy}$ is a solution to (2.4). Note $\theta = 1$

implies $K_N(\theta) = N$. Therefore, if we substitute this solution into the dual gap bound in Theorem 2.1 together with $\eta_x = 1/\tau$ and $\eta_y = 1/\sigma$, we get

$$\begin{aligned}
&\mathbb{E}[\sup_{(x,y)\in X\times Y}\{\mathcal{L}(\bar{x}_N, y) - \mathcal{L}(x, \bar{y}_N) + \frac{1}{N}\Delta_N(x,y)\}] \le \frac{1}{N}\Omega_{\tau,\sigma,\theta} + \Xi_{\tau,\sigma,\theta} \\
&= \frac{1}{N}\Big(\frac{1}{\tau}\Omega_f + \frac{2}{\sigma}\Omega_g\Big) + \tau(1 + 2\Xi^x_{\tau,\sigma,\theta})\frac{\delta_x^2}{2} + \sigma(3 + 2\Xi^y_{\tau,\sigma,\theta})\frac{\delta_y^2}{2} \\
&= \frac{1}{N}\Big(\frac{1}{\tau}\Omega_f + \frac{2}{\sigma}\Omega_g\Big) + \tau(3 + 2\sigma L_{yx})\frac{\delta_x^2}{2} + \sigma(27 + 12\sigma L_{yy} + 12\tau\sigma L_{yx}L_{xy} + 2\tau L_{yx})\frac{\delta_y^2}{2} \\
&\le \frac{1}{N}\Big(\frac{1}{\tau}\Omega_f + \frac{2}{\sigma}\Omega_g\Big) + \frac{5\tau\delta_x^2}{2} + 24\sigma\delta_y^2,
\end{aligned} \tag{2.23}$$

where for the last inequality we use $\tau L_{yx} \le 1$, $\sigma\max\{L_{yx}, L_{xy}\} \le 1$ and $\sigma L_{yy} \le \frac{1}{2}$ due to (2.22). For any $\epsilon > 0$, to bound (2.23) from above by ϵ, we require

$$\frac{1}{N}\Big(\frac{1}{\tau}\Omega_f + \frac{2}{\sigma}\Omega_g\Big) \le \frac{\epsilon}{3}, \quad \frac{5\tau\delta_x^2}{2} \le \frac{\epsilon}{3}, \quad 24\sigma\delta_y^2 \le \frac{\epsilon}{3}. \tag{2.24}$$

Our parameter choice in (2.22) implies that the second and the third inequalities in (2.24) trivially hold. It suffices to choose N large enough depending on given ϵ so that (2.24) holds, i.e., $N \ge \frac{3}{\epsilon}(\frac{1}{\tau}\Omega_f + \frac{2}{\sigma}\Omega_g)$. From (2.22) we have $\frac{1}{\tau} \le L_{yx} + L_{xx} + \frac{15}{2}\frac{\delta_x^2}{\epsilon}$ and $\frac{1}{\sigma} \le \max\{L_{yx} + 2L_{yy},\ L_{xy}\} + 72\frac{\delta_y^2}{\epsilon}$; thus, $\mathcal{G}(\bar{x}_N, \bar{y}_N) \le \epsilon$ holds for all $N \ge N_\epsilon$. □

3. Robustness and Convergence Rate Trade-off. In this section, assuming $\mu_x, \mu_y > 0$, we study the trade-offs between robustness-to-gradient noise and the convergence rate depending on the choice of SAPD parameters. Given the saddle point $z^* \triangleq (x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ of the problem (1.1), we first define the *robustness* as follows:

$$\mathcal{J} \triangleq \limsup_{N\to\infty} \mathcal{J}_N, \quad \text{where} \quad \mathcal{J}_N \triangleq \mathbb{E}\Big[\frac{1}{\delta_x^2}\|x_N - x^*\|^2 + \frac{1}{\delta_y^2}\|y_N - y^*\|^2\Big]. \tag{3.1}$$

The quantity $\mathcal{J}_N$ reflects the variance of $z_N \triangleq (x_N, y_N)$ normalized by the level of gradient noise: δ_x^2 and δ_y^2. Thus, using $\mathcal{J}$ we measure how much SAPD amplifies the gradient noise asymptotically. Due to the persistent stochastic noise, $\{z_N\}$ does not typically converge to z^* but oscillate around it with a positive variance. The limit $\mathcal{J}$ provides a bound on the expected size of neighborhood $\{z_N\}$ accumulates in, i.e.,

$$\limsup_{N\to\infty} \mathbb{E}\|z_N - z^*\| \le \limsup_{N\to\infty} \sqrt{\mathbb{E}\left[\|x_N - x^*\|^2 + \|y_N - y^*\|^2\right]} \le \max\{\delta_x, \delta_y\}\sqrt{\mathcal{J}}$$

from Jensen's lemma. Therefore, smaller values of $\mathcal{J}$ will lead to better robustness to noise and will give better asymptotic performance. Below we derive an explicit characterization of $\mathcal{J}$ for a particular class of SCSC problems; and we will obtain an upper bound on $\mathcal{J}$ for more general SCSC problems in subsection 3.2.

3.1. Explicit Estimates for Robustness to Noise. We consider the special case of (1.1) when Φ is bilinear and f, g have simple quadratic forms, i.e.,

$$\Phi(x,y) = \langle Kx, y\rangle,\ f(x) = \frac{\mu_x}{2}\|x\|^2,\ g(y) = \frac{\mu_y}{2}\|y\|^2, \tag{3.2}$$

where $K \in \mathbb{R}^{d\times d}$ is a symmetric matrix, and the noise is additive, i.e.,

$$\tilde{\nabla}_x\Phi(x_k, y_{k+1}; \omega_k^x) = \nabla_x\Phi(x_k, y_{k+1}) + \omega_k^x, \quad \tilde{\nabla}_y\Phi(x_k, y_k; \omega_k^y) = \nabla_y\Phi(x_k, y_k) + \omega_k^y, \tag{3.3}$$

satisfying Assumption 1.2. We also assume that there exists $\delta > 0$ such that for each $k \ge 0$, w_k^x, w_k^y are Gaussian with an isotopic covariance, i.e.,

$$\mathbb{E}\left[w_k^x (w_k^x)^\top\right] = \frac{\delta^2}{d} I_d, \quad \mathbb{E}\left[w_k^y (w_k^y)^\top\right] = \frac{\delta^2}{d} I_d. \tag{3.4}$$

Clearly, the unique saddle point to (3.2) is the origin, i.e., $(x^*, y^*) = (\mathbf{0}, \mathbf{0})$. We will show that the robustness measure $\mathcal{J}$ defined in (3.1) is finite, and that it admits a closed form solution. We first note that according to Algorithm 1.1, for $k \geq 0$,

$$\begin{aligned} y_{k+1} &= \frac{1}{1+\sigma\mu_y}\left(y_k + \sigma(1+\theta)Kx_k - \sigma\theta Kx_{k-1} + \sigma(1+\theta)\omega_k^y - \sigma\theta\omega_{k-1}^y\right), \\ x_{k+1} &= \frac{1}{1+\tau\mu_x}(x_k - \tau K^\top y_{k+1} - \tau\omega_k^x). \end{aligned} \tag{3.5}$$

Next, for $k \geq 0$, we define $\tilde{z}_k \triangleq \begin{bmatrix} x_k; & y_k; & x_{k-1}; & y_{k-1} \end{bmatrix}$, which is a vertical concatenation of iterates from two consecutive iterations, and $w_k \triangleq \begin{bmatrix} w_k^x; & w_k^y; & w_{k-1}^x; & w_{k-1}^y \end{bmatrix}$, which is the vertical concatenation of the noise realization at step k and $k-1$. The column vector $\tilde{z}_k$ satisfies the following affine recursion:

$$\tilde{z}_{k+1} = A\tilde{z}_k + Bw_k, \quad \text{where} \tag{3.6}$$

$$A = \begin{bmatrix} \frac{1}{1+\tau\mu_x}I_d - \frac{\tau\sigma(1+\theta)}{(1+\sigma\mu_y)(1+\tau\mu_x)}K^\top K & -\frac{\tau}{(1+\sigma\mu_y)(1+\tau\mu_x)}K^\top & \frac{\tau\sigma\theta}{(1+\sigma\mu_y)(1+\tau\mu_x)}K^\top K & 0_d \\ \frac{\sigma(1+\theta)}{1+\sigma\mu_y}K & \frac{1}{1+\sigma\mu_y}I_d & -\frac{\sigma\theta}{1+\sigma\mu_y}K & 0_d \\ I_d & 0_d & 0_d & 0_d \\ 0_d & I_d & 0_d & 0_d \end{bmatrix},$$

$$B = \begin{bmatrix} -\frac{\tau}{1+\tau\mu_x}I_d & -\frac{\tau\sigma(1+\theta)}{(1+\sigma\mu_y)(1+\tau\mu_x)}K^\top & 0_d & \frac{\tau\sigma\theta}{(1+\sigma\mu_y)(1+\tau\mu_x)}K^\top \\ 0_d & \frac{\sigma(1+\theta)}{1+\sigma\mu_y}I_d & 0_d & -\frac{\sigma\theta}{1+\sigma\mu_y}I_d \\ 0_d & 0_d & 0_d & 0_d \\ 0_d & 0_d & 0_d & 0_d \end{bmatrix}.$$

From (3.6), it is easy to see that $\Sigma_k \triangleq \mathbb{E}[\tilde{z}_k\tilde{z}_k^\top]$ satisfies the recursion

$$\Sigma_{k+1} = A\Sigma_k A^\top + \frac{\delta^2}{d}BB^\top, \tag{3.7}$$

for $k \geq 0$, where we used the noise structure (3.4). Linear dynamical systems subject to Gaussian noise such as (3.6) has been well studied in the robust control literature. In fact, it is known that the limit $\Sigma_\infty \triangleq \lim_{k\to\infty}\Sigma_k$ exists if the spectral radius[4] of A, denoted by $\rho(A)$, is less than one, and it satisfies the Lyapunov equation:

$$\Sigma_\infty \triangleq A\Sigma_\infty A^\top + \frac{\delta^2}{d}BB^\top, \tag{3.8}$$

whose solution is given in the form of an infinite series (see e.g. [45]):

$$\Sigma_\infty = \frac{\delta^2}{d}\sum_{k=0}^{\infty} A^k B(A^k B)^\top. \tag{3.9}$$

It is also easy to see that $\mathcal{J} = \frac{1}{2\delta^2}\mathbf{Tr}(\Sigma_\infty) = \frac{1}{2d}\sum_{k=0}^{\infty}\|A^k B\|_F^2$. We also observe that $\mathcal{J}$ is *invariant* under orthogonal transformations, i.e., if we introduce $\tilde{A} \triangleq Z^\top AZ$ and $\tilde{B} \triangleq Z^\top BZ$ for an orthogonal matrix Z, then we have

$$\mathcal{J} = \frac{1}{2\delta^2}\mathbf{Tr}(\tilde{\Sigma}_\infty), \quad \text{where} \quad \tilde{\Sigma}_\infty \triangleq \frac{\delta^2}{d}\sum_{k=0}^{\infty}\tilde{A}^k\tilde{B}(A^k\tilde{B})^\top = Z^\top\Sigma_\infty Z \tag{3.10}$$

solves the transformed Lyapunov equation

$$\tilde{\Sigma}_\infty = \tilde{A}\tilde{\Sigma}_\infty\tilde{A}^\top + \frac{\delta^2}{d}\tilde{B}\tilde{B}^\top. \tag{3.11}$$

[4] The spectral radius of a square matrix A is defined as the largest absolute value of its eigenvalues.

In order to compute $\mathcal{J}$ explicitly, in the following we will choose a particular orthogonal matrix Z so that solving (3.11) explicitly will be simpler than solving (3.8).

First, we consider the eigenvalue decomposition $K = U\Lambda U^\top$ of K, where Λ is a diagonal matrix such that $\Lambda_{ii} = \lambda_i$, and $\{\lambda_i\}_{i=1}^d$ are the eigenvalues of K in increasing order, i.e., $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_d$. Then, we can write $A = VA_\Lambda V^\top$, where

$$A_\Lambda \triangleq \begin{bmatrix} \frac{1}{1+\tau\mu_x} I_d - \frac{\tau\sigma(1+\theta)}{(1+\sigma\mu_y)(1+\tau\mu_x)}\Lambda^2 & -\frac{\tau}{(1+\sigma\mu_y)(1+\tau\mu_x)}\Lambda & \frac{\tau\sigma\theta}{(1+\sigma\mu_y)(1+\tau\mu_x)}\Lambda^2 & 0_d \\ \frac{\sigma(1+\theta)}{1+\sigma\mu_y}\Lambda & \frac{1}{1+\sigma\mu_y} I_d & -\frac{\sigma\theta}{1+\sigma\mu_y}\Lambda & 0_d \\ I_d & 0_d & 0_d & 0_d \\ 0_d & I_d & 0_d & 0_d \end{bmatrix}$$

and $V \in \mathbb{S}^{4d}$ is a block-diagonal matrix with each block on the diagonal equal to U, i.e., $V = \mathrm{diag}(\{U\}_{i=1}^4)$. Furthermore, we can permute the entries of A_Λ so that it becomes a block diagonal matrix, i.e., there exists a permutation matrix P such that

$$PA_\Lambda P^\top = \mathrm{diag}(\{\tilde{A}_i\}_{i=1}^d) \triangleq \tilde{A}$$

where $\tilde{A}_i \in \mathbb{R}^{4\times 4}$ is defined by

$$\tilde{A}_i \triangleq \begin{bmatrix} \frac{1}{1+\tau\mu_x} - \frac{\tau\sigma(1+\theta)}{(1+\sigma\mu_y)(1+\tau\mu_x)}\lambda_i^2 & -\frac{\tau}{(1+\sigma\mu_y)(1+\tau\mu_x)}\lambda_i & \frac{\tau\sigma\theta}{(1+\sigma\mu_y)(1+\tau\mu_x)}\lambda_i^2 & 0 \\ \frac{\sigma(1+\theta)}{1+\sigma\mu_y}\lambda_i & \frac{1}{1+\sigma\mu_y} & -\frac{\sigma\theta}{1+\sigma\mu_y}\lambda_i & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Thus, for $Z = VP^\top$, we have $\tilde{A} = Z^\top AZ$, and $\tilde{B} = Z^\top BZ = \mathrm{diag}\{\tilde{B}_i\}_{i=1}^d$ such that

$$\tilde{B}_i \triangleq \begin{bmatrix} -\frac{\tau}{1+\tau\mu_x} & -\frac{\tau\sigma(1+\theta)}{(1+\sigma\mu_y)(1+\tau\mu_x)}\lambda_i & 0 & \frac{\tau\sigma\theta}{(1+\sigma\mu_y)(1+\tau\mu_x)}\lambda_i \\ 0 & \frac{\sigma(1+\theta)}{1+\sigma\mu_y} & 0 & -\frac{\sigma\theta}{1+\sigma\mu_y} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Note that both $\tilde{A}$ and $\tilde{B}$ have a block diagonal structure. Therefore, we see

$$\tilde{\Sigma}_\infty = \mathrm{diag}(\{\tilde{S}_i\}_{i=1}^d), \tag{3.12}$$

where for each $i \in \{1, \ldots, d\}$, S_i solves

$$\tilde{S}_i = \tilde{A}_i\tilde{S}_i\tilde{A}_i^\top + \frac{\delta^2}{d}\tilde{B}_i\tilde{B}_i^\top. \tag{3.13}$$

Equation (3.13) is a 4×4 system, which can be solved for $\tilde{S}_i$ explicitly. Using (3.10) and (3.12) will yield us an explicit formula for $\mathcal{J}$. We solve (3.13) with the "Matlab Symbolic Toolbox", this allows us to compute the robustness measure $\mathcal{J}$ explicitly. Next, for A in (3.6), we define $\rho_{\text{true}} \triangleq (\rho(A))^2$, which determines the exact (asymptotic) convergence rate of $\mathbb{E}[\|z_N - z^*\|^2]$; hence, $\mathbb{E}[\Delta_N(x^*, y^*)]$ in Theorem 2.1 also converges with this rate. Furthermore, it can also be shown that $\mathbb{E}[\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}}\{\mathcal{L}(x_k, y) - \mathcal{L}(x, y_k)\}]$ converges with the same rate (see Appendix C). This way of computing robustness and convergence rate is independent of our theoretical analysis of the SAPD algorithm. It reflects the exact behavior of the algorithm for a quadratic function in (3.2), which helps us understand some fundamental relations.

We numerically illustrate the fundamental rate-robustness trade-off in Figure 1 for (3.2) through plotting 3 curves: $\mathcal{J}_1$, $\mathcal{J}_2$ and $\mathcal{J}^*$. For $\mathcal{J}_1$, we uniformly grid the parameter space $(\tau, \sigma, \theta) \in [0, 0.5] \times [0, 0.5] \times [0, 2]$ using $500 \times 500 \times 200$ points;

then, for each grid point, we compute the corresponding $(\rho_{\text{true}}, \mathcal{J})$ values and plot it. For $\mathcal{J}_2$, we employ the step sizes suggested in [7, Algorithm 5] for the CP method[5] and plot $(\rho_{\text{true}}, \mathcal{J})$ (see Appendix C.1 for more details). For $\mathcal{J}^*$, defining $\mathcal{J}^*(\rho) \triangleq \min_{\tau,\sigma,\theta\geq 0}\{\mathcal{J} : \rho_{\text{true}} = \rho\}$, we plot $(\rho, \mathcal{J}^*(\rho))$ which illustrates the best robustness that can be achieved for a given rate. In the $\mathcal{J}_1$ plot, there are vertical lines as there exist many points in the grid sharing the same rate while they have very different robustness values. As seen in Figure 1(a), for great majority of parameter choice from the uniform grid, the corresponding robustness is very poor, i.e., very high $\mathcal{J}$ value. As a consequence, we infer that it is necessary to control the robustness through properly tuning the algorithm parameters. The $\mathcal{J}_2$ plot demonstrates that for the same rate CP parameter choice ensures relatively lower $\mathcal{J}$ values compared to the majority of points in the uniform grid; but, $\mathcal{J}_2$ is still far away from the efficient frontier $\mathcal{J}^*$. As indicated in $\mathcal{J}_2$ plot, the best convergence rate CP parameters can achieve is only around 0.83, while the best rate achieved among the uniform grid is 0.67.

We emphasize that while the parameter optimization problem to compute $\mathcal{J}^*(\rho)$ for a given convergence rate $\rho \in (0,1)$ can be done for (1.1) defined by (3.2), this is not a trivial task for a more general coupling function Φ; therefore, we provide an alternative formulation to achieve a similar trade-off result between an upper bound on $\mathcal{J}$ and a bound on the convergence rate in Proposition 3.1.

(a) (b)

FIG. 1. *The rate-robustness trade-off for* (3.2) *when* $\mu_x = \mu_y = 1$, $\|K\|_2 = 10$, $d = 30$ *and* $\delta_x = \delta_y = \delta = 10$. *The best achievable rate is* 0.67. *The point indicated with a red "$*$" in Figure* 1(a) *is the particular choice of CP parameters given in [7, Eq.(49)]. Figure* 1(b) *illustrates that employing the SAPD parameters obtained through minimizing* $\bar{\mathcal{R}}$, *an upper bound on* $\mathcal{J}$ *defined in subsection* 3.2, *one can closely track the efficient frontier* $\mathcal{J}^*$. *The best certifiable rate is* 0.9049.

Next, we analyze how primal-dual step sizes, τ and σ, affect the convergence rate and the robustness level. For any given $\epsilon \in (0,1)$, we define

$$\rho^*_\epsilon(\tau,\sigma) \triangleq \min_{\theta\geq 0}\{\rho_{\text{true}} : \rho_{\text{true}} \leq 1-\epsilon\}, \quad \mathcal{J}^*_\epsilon(\tau,\sigma) \triangleq \min_{\theta\geq 0}\{\mathcal{J} : \rho_{\text{true}} \leq 1-\epsilon\}.$$

We consider the same experiment described in the caption of Figure 1, setting x-axis as τ, and y-axis as σ, we plot $\rho^*_\epsilon(\tau,\sigma)$ in Figure 2(a) and $\mathcal{J}^*_\epsilon$ in Figure 2(b), for $\epsilon = 0.01$. We observe that except for the boundary points, simultaneously increasing τ and σ leads to a faster convergence rate at the expense of a decrease in robustness level –

[5] Although our method SAPD generalizes the CP method beyond bilinear problem, SAPD coincides with CP on this particular problem as it has a bilinear coupling function Φ.

as one approaches to the boundary, there is a significant increase in both convergence rate and $\mathcal{J}$ values. These results illustrate the fundamental trade-offs between the convergence rate versus robustness for SAPD. We obtained similar results for SGDA in Appendix B.3 and compared those with SAPD.

(a) (b)

FIG. 2. *The effect of the step sizes on rate and robustness of SAPD running on* (3.2) *when* $\mu_x = \mu_y = 1$, $\|K\|_2 = 10$, $d = 30$ *and* $\delta_x = \delta_y = \delta = 10$. *The best achievable rate is* 0.67.

3.2. An Upper Bound for the Robustness Measure $\mathcal{J}$. As we mentioned earlier, $\mathcal{J}$ is hard to compute in general. To alleviate this issue, we can alternatively minimize an upper bound on $\mathcal{J}$ to control the robustness level. We start with a proposition that provides an upper bound on the robustness measure $\mathcal{J}$.

PROPOSITION 3.1. *Suppose Assumptions 1.1, and 1.2 hold, and* $\{x_k, y_k\}_{k\geq 0}$ *are generated by SAPD stated in Algorithm 1.1, the parameters* $\{\tau, \sigma, \theta\}$ *satisfy the conditions in Theorem 2.1 for some* $\alpha \in [0, \frac{1}{\sigma})$ *and* $\rho \in (0,1)$. *Then*

$$\mathcal{J} = \limsup_{N\to\infty} \mathcal{J}_N \leq \frac{1}{(1-\rho)\delta^2} \cdot \frac{\Xi_{\tau,\sigma,\theta}}{C_{\tau,\sigma}}, \tag{3.14}$$

where $C_{\tau,\sigma} \triangleq \min\{\frac{1}{2\rho\tau}, \frac{1}{2\rho\sigma}(1-\alpha\sigma)\}$, $\delta \triangleq \min\{\delta_x, \delta_y\}$ *and* $\Xi_{\tau,\sigma,\theta}$ *is as in Theorem 2.1.*

Proof. From Theorem 2.1, we have that

$$\begin{aligned}
\frac{1-\rho}{1-\rho^N} C_{\tau,\sigma}\delta^2 \mathcal{J}_N \leq& \mathbb{E}\left[\frac{1-\rho}{1-\rho^N} C_{\tau,\sigma}(\|x_N - x^*\|^2 + \|y_N - y^*\|^2)\right] \\
\leq& \mathbb{E}\left[\mathcal{L}(\bar{x}_N, y^*) - \mathcal{L}(x^*, \bar{y}_N) + \frac{1-\rho}{1-\rho^N} C_{\tau,\sigma}(\|x_N - x^*\|^2 + \|y_N - y^*\|^2)\right] \\
\leq& \mathbb{E}\left[\mathcal{L}(\bar{x}_N, y^*) - \mathcal{L}(x^*, \bar{y}_N) + \frac{1-\rho}{1-\rho^N} \Delta_N(x^*, y^*)\right] \leq \frac{1}{K_N(\rho)} \Omega_{\tau,\sigma,\theta} + \Xi_{\tau,\sigma,\theta},
\end{aligned}$$

which implies (3.14) since $\frac{1}{K_N(\rho)}\Omega_{\tau,\sigma,\theta} \to 0$ as $N \to \infty$. ∎

This upper bound is theoretically correct only for parameters satisfying our step size conditions in (2.1), which are only sufficient for ensuring a linear rate; but, these conditions are not necessary. Next we investigate the trade-off between the convergence rate and the robustness upper bound provided in the above proposition.

LEMMA 3.2. *Given $\rho \in (0,1)$, let $(t_\rho, s_\rho, \theta_\rho, \alpha_\rho)$ be an element of $\mathcal{P}_\rho$, where*

$$\mathcal{P}_\rho \triangleq \{(t,s,\theta,\alpha): \ t,s,\theta,\alpha \geq 0, \ \alpha \leq s, \ G_\rho(t,s,\theta,\alpha) \succeq 0\} \text{ and}$$

$$G_\rho(t,s,\theta,\alpha) \triangleq \begin{pmatrix} (1-\frac{1}{\rho})t+\mu_x & 0 & 0 & 0 & 0 \\ 0 & (1-\frac{1}{\rho})s+\mu_y & (\frac{\theta}{\rho}-1)L_{yx} & (\frac{\theta}{\rho}-1)L_{yy} & 0 \\ 0 & (\frac{\theta}{\rho}-1)L_{yx} & t-L_{xx} & 0 & -\frac{\theta}{\rho}L_{yx} \\ 0 & (\frac{\theta}{\rho}-1)L_{yy} & 0 & s-\alpha & -\frac{\theta}{\rho}L_{yy} \\ 0 & 0 & -\frac{\theta}{\rho}L_{yx} & -\frac{\theta}{\rho}L_{yy} & \frac{\alpha}{\rho} \end{pmatrix}.$$

Then the bias term for SAPD employing $\tau = 1/t_\rho$, $\sigma = 1/s_\rho$ and $\theta = \theta_\rho$ converges to 0 with rate ρ. If $\mathcal{P}_\rho = \emptyset$, then (2.1) does not have a solution for the given ρ value.

Proof. This result immediately follows from Theorem 2.1. ∎

With the help of Lemma 3.2, one can do a binary search on $(0,1)$ interval to compute the best rate ρ^* we can justify using the matrix inequality in (2.1), i.e., $\rho^* \triangleq \min_{\rho \geq 0}\{\rho : \mathcal{P}_\rho \neq \emptyset\}$. For any $\rho \in (0,1)$, checking whether $\mathcal{P}_\rho$ is nonempty or not requires solving 4-dimensional SDP.

Next, we numerically illustrate that the explicit upper bound we derived in Proposition 3.1 provides a reasonable approximation to the actual robustness measure $\mathcal{J}$. For this purpose, we consider the same example from Section 3.1 where $\mathcal{J}$ can be explicitly computed, and compare $\mathcal{J}$ to its upper bound. In (3.2), we take $\mu_x = \mu_y = 1$, where we generate the symmetric matrix $K \in \mathbb{R}^{d\times d}$ randomly with $\|K\|_2 = 10$ and $d = 30$. We assume the noise model given in (3.3) and (3.4) with $\delta_x = \delta_y = 10$. Consequently, we have $L_{xx} = L_{yy} = 0$, $L_{yx} = 10$. Employing the particular parameter choice in Corollary 2.8, i.e., setting $\beta = 1$ and $\bar{\theta}_2 = 0$, we can certify that SAPD converges with rate $\rho \approx 0.9049$ using τ and σ as in (2.7) and $\theta = \rho$. We have found out that ρ^* obtained using the binary search for this example was also equal to 0.9049, i.e., our special solution in Corollary 2.8 leads to the optimal rate ρ^*.

For any $\rho \in [\rho^*, 1)$, we can optimize SAPD parameters minimizing the upper bound for robustness provided in (3.14) while ensuring that the bias term converges linearly with the given rate ρ, i.e.,

$$\mathcal{R}(\rho) \triangleq \min_{\tau,\sigma,\theta,\alpha \geq 0} \left\{ \frac{1}{(1-\rho)\delta^2} \cdot \Xi_{\tau,\sigma,\theta}/C_{\tau,\sigma} \ : \ (2.1) \text{ holds for } (\tau,\sigma,\theta,\rho,\alpha) \right\}. \tag{3.15}$$

Note that for solving this optimization problem, we do not need to know δ_x or δ_y. It is a non-convex problem; however, it has some structure. In the next lemma, we provide a simpler optimization problem exploiting this structure.

LEMMA 3.3. *Given $\rho \in (0,1)$ and $\tau > 0$, let $S_\rho(\tau) \triangleq \{(\sigma,\theta,\alpha) :$ (2.1) holds for $(\tau,\sigma,\theta,\rho,\alpha)\}$. Suppose $\cup_{\tau>0} S_\rho(\tau) \neq \emptyset$. Then, $\frac{1-\rho}{\mu_x \rho} = \min\{\tau : \ S_\rho(\tau) \neq \emptyset\}$. Moreover, for any $\tau_1 \geq \tau_2 \geq \frac{1-\rho}{\mu_x\rho}$, we have $S_\rho(\tau_1) \subset S_\rho(\tau_2)$. Finally, $\mathcal{R}(\rho)$ defined in (3.15) can also be computed as*

$$\mathcal{R}(\rho) = \min_{\sigma,\theta,\alpha \geq 0} \left\{ \frac{2\rho}{1-\rho} \cdot \max\left\{\tau, \frac{\sigma}{(1-\alpha\sigma)}\right\} \cdot \Xi_{\tau,\sigma,\theta}/\delta^2 \ : \ \tau = \frac{1-\rho}{\mu_x\rho}, \ (2.1) \text{ holds} \right\}. \tag{3.16}$$

Proof. Since $\cup_{\tau>0} S_\rho(\tau) \neq \emptyset$, there exists $(\tau,\sigma,\theta,\rho,\alpha)$ satisfying (2.1); thus, $\frac{1}{\tau}(1-\frac{1}{\rho}) + \mu_x \geq 0$, i.e., $\tau \geq \bar{\tau} \triangleq \frac{1-\rho}{\mu_x\rho}$. Say $\tau \geq \bar{\tau}$, then we have $\frac{1}{\bar{\tau}} - L_{xx} \geq \frac{1}{\tau} - L_{xx}$, which implies that $(\sigma,\theta,\alpha) \in S_\rho(\bar{\tau})$. Therefore, we can conclude that $\frac{1-\rho}{\mu_x\rho} = \min\{\tau : \ S_\rho(\tau) \neq \emptyset\}$ because (2.1) requires that $\tau \geq \bar{\tau}$.

Now suppose $(\sigma, \theta, \alpha) \in S_\rho(\tau_1)$ for some $\tau_1 > 0$. The same arguments above also show that for any $\tau_2 \in [\frac{1-\rho}{\mu_x \rho}, \tau_1]$, we have $(\sigma, \theta, \alpha) \in S_\rho(\tau_2)$; hence, $S_\rho(\tau_1) \subset S_\rho(\tau_2)$. Furthermore, the objective function in (3.15) is strictly increasing in τ; thus, we can conclude that the optimal value of (3.15) is equal to that of (3.16). ∎

For any given $\rho \in (0, 1]$, (2.1) implies that $\tau \geq \frac{1-\rho}{\mu_x \rho}$. Furthermore, another necessary condition for (2.1) is that $\begin{pmatrix} \mu_y & (\frac{\theta}{\rho}-1)L_{yx} \\ (\frac{\theta}{\rho}-1)L_{yx} & \frac{1}{\tau}-L_{xx} \end{pmatrix} \succeq 0$, which further implies that $(\frac{\theta}{\rho}-1)^2 L_{yx}^2 \leq \mu_y(\frac{1}{\tau}-L_{xx})$. Thus, either $\theta \in [0, \rho]$, or when $\theta \geq \rho$, the necessary condition above gives $\theta \leq \rho\left(1+\frac{\sqrt{\mu_y(\frac{1}{\tau}-L_{xx})}}{L_{yx}}\right) \leq \rho\left(1+\frac{\sqrt{\mu_y(\frac{\mu_x\rho}{1-\rho}-L_{xx})}}{L_{yx}}\right) \triangleq \bar{\Theta}_\rho$.

Next, we discuss how an upper bound on $\mathcal{J}$ can be computed efficiently by a bisection search over the rate parameter ρ and a grid search on θ. We start with a definition and provide a lemma for obtaining a bound on robustness.

Definition 3.4. *For $\rho \in (0, 1)$, let $\mathcal{C}_\rho \triangleq \{c \in (0, 1) : \mathcal{P}_\rho \cap L_c \neq \emptyset\}$, where $L_c \triangleq \{(t, s, \theta, \alpha) \in \mathbb{R}_+^4 : \alpha = cs\}$, and $\mathcal{P}_\rho$ is as in Lemma 3.2. Note $\mathcal{C}_\rho \neq \emptyset$ for all $\rho \in [\rho^*, 1)$.*

Remark 3.5. For any given $\rho \in [\rho^*, 1)$, $\mathcal{C}_\rho \neq \emptyset$ is a convex set;[6] hence, $\mathcal{C}_\rho \subset [0, 1]$ is an interval. Therefore, $\bar{c}_\rho \triangleq \sup \mathcal{C}_\rho$ and $\underline{c}_\rho \triangleq \inf \mathcal{C}_\rho$ can be computed using bisection. Each iteration of bisection involves checking the feasibility of $\{(t, s, \theta) \in \mathbb{R}_+^3 : G_\rho(t, s, \theta, cs) \succeq \mathbf{0}\}$ for a given $c \in (0, 1)$, which is a 3-dimensional SDP.

Lemma 3.6. *Given $\rho \in [\rho^*, 1)$ and $c \in \mathcal{C}_\rho$, define $S_{c,\rho} \triangleq \{(s, \theta) : \exists (t, s, \theta, \alpha) \in \mathcal{P}_\rho$ s.t. $t = \frac{\mu_x \rho}{1-\rho}$, $\alpha = cs\}$ and $\Theta_{c,\rho} \triangleq \{\theta : \exists (s, \theta) \in S_{c,\rho}\}$. Let $\underline{\theta}_{c,\rho} \triangleq \inf \Theta_{c,\rho}$ and $\bar{\theta}_{c,\rho} \triangleq \sup \Theta_{c,\rho}$. Given $K_\theta \in \mathbb{Z}_+$, let $\{\theta_k\}_{k=1}^{K_\theta} \subset [\underline{\theta}_{c,\rho}, \bar{\theta}_{c,\rho}] \subset [0, \Theta_\rho]$ be an arbitrary set of grid points such that $\theta_1 = \underline{\theta}_{c,\rho}$ and $\theta_{K_\theta} = \bar{\theta}_{c,\rho}$. Define $\bar{\mathcal{R}}_c(\rho) \triangleq \min_{k=1,\ldots K_\theta} \bar{\mathcal{R}}_c(\rho, \theta_k)$, where*

$$\bar{\mathcal{R}}_c(\rho, \theta) \triangleq \min_{\sigma \geq 0} \left\{ \max\left\{\frac{2}{\mu_x}, \frac{2\rho\sigma}{(1-c)(1-\rho)}\right\} \cdot \Xi_{\tau,\sigma,\theta}/\delta^2 : \tau = \frac{1-\rho}{\mu_x \rho}, \alpha = \frac{c}{\sigma}, \text{ (2.1) holds} \right\}. \tag{3.17}$$

Then, $\bar{\mathcal{R}}_c(\rho) \geq \mathcal{R}(\rho)$. Furthermore, for any fixed $\rho \in [\rho^*, 1)$, $c \in \mathcal{C}_\rho$ and $\theta \in [\underline{\theta}_{c,\rho}, \bar{\theta}_{c,\rho}]$, $\sigma_c(\rho, \theta) \triangleq 1/\max\{s : (s, \theta) \in S_{c,\rho}\}$ is the unique optimal solution to (3.17).

Proof. Given $\rho \in [\rho^*, 1)$ and $c \in \mathcal{C}_\rho$, since we fix θ and $\alpha = c/\sigma$ while deriving (3.17), we immediately get $\bar{\mathcal{R}}_c(\rho) \geq \mathcal{R}(\rho)$ due to Lemma 3.3. Lastly, after fixing $\rho \in [\rho^*, 1)$, $c \in \mathcal{C}_\rho$ and $\theta \in [\underline{\theta}_{c,\rho}, \bar{\theta}_{c,\rho}]$, the objective function in (3.17) is increasing in $\sigma > 0$, and $\sigma \mapsto 1/\sigma = s$ is a bijection between the feasible region of (3.17) and $\{s : (s, \theta) \in S_{c,\rho}\}$. Therefore, the unique optimal solution $\sigma_c(\rho, \theta)$ can be computed by solving a one-dimensional SDP, i.e., $\max\{s : (s, \theta) \in S_{c,\rho}\}$ for fixed ρ, c and θ. ∎

Given $K_c, K_\rho \in \mathbb{Z}_+$, let $P \triangleq \{\rho_k\}_{k=1}^{K_\rho} \subset [\rho^*, 1]$ and $C_\rho \triangleq \{c_k\}_{k=1}^{K_c} \subset \mathcal{C}_\rho$ be the grid points. Finally, for $\rho \in P$, we define $\bar{\mathcal{R}}(\rho) \triangleq \min_{c \in C_\rho} \bar{\mathcal{R}}_c(\rho)$, where $\bar{\mathcal{R}}_c(\rho)$ can be computed based on Lemma 3.6 for any $c \in C_\rho$. Therefore, for any $\rho \in P$, computing $\bar{\mathcal{R}}(\rho)$ using Lemma 3.6 will yield $(\tau_\rho, \sigma_\rho, \theta_\rho)$ achieving $\bar{\mathcal{R}}_c(\rho)$ for some $c \in C_\rho$ such that $\bar{\mathcal{R}}(\rho) = \bar{\mathcal{R}}_c(\rho)$. Thus, for the quadratic model assumed in subsection 3.1, we can compute the robustness measure, defined in (3.1), corresponding to $(\tau_\rho, \sigma_\rho, \theta_\rho)$,

[6] We skip the proof due to limited space; for details, see Lemma C.1.

FIG. 3. *Comparison of SAPD against SOGDA, SMP and SMD on a synthetic toy problem using the distance (left) and gap (right) metrics. The rates of SAPD are* $\rho_1 = 0.99$ *and* $\rho_2 = 0.995$.

which we call $\mathcal{J}(\rho)$. Recall that in Section 3.1, we defined $\mathcal{J}^*(\rho) \triangleq \min_{\tau,\sigma,\theta\geq 0}\{\mathcal{J} : \rho_{\text{true}} = \rho\}$. To numerically illustrate the rate vs robustness trade-off and also to demonstrate that we can control robustness through optimizing $\bar{\mathcal{R}}$, in Figure 1(b), we plot robustness measure $\mathcal{J}(\rho)$, corresponding to $(\tau_\rho, \sigma_\rho, \theta_\rho)$ computed by minimizing its upper bound $\bar{\mathcal{R}}(\rho)$, against the convergence rate values $\rho \in P$ in the x-axis, and compare $\mathcal{J}(\rho)$ with $\mathcal{J}^*(\rho)$ and $\bar{\mathcal{R}}(\rho)$, where we set $K_\rho = K_\theta = 100$ and $K_c = 50$. In Figure 1(b), we observe that $\mathcal{J}(\rho)$ computed for SAPD parameters optimizing $\bar{\mathcal{R}}(\rho)$ closely tracks $\mathcal{J}^*(\rho)$. Therefore, we infer that optimizing the upper bound is practical to optimize the robustness for the problem class used in these experiments.

4. Numerical experiments. In this section, we compared SAPD against S-OGDA [11], SMD [28] and SMP [18] for solving (1.1) with synthetic and real-data.

4.1. Regularized Bilinear SP Problem with Synthetic Data. We first tested SAPD, S-OGDA and SMP on the regularized bilinear SP problem defined in (3.2). In this experiment, we set $\mu_x = \mu_y = 1$, $\|K\|_2 = 10$, $d = 30$ and $\delta_x = \delta_y = \delta = 5$. Since SMD step size condition requires a bound on the stochastic gradients, SMD is implemented on (3.2) with additional $(x, y) \in X \times Y$ constraint where $X = \{x \in \mathcal{X} : \|x\| \leq \sqrt{d}\}$ and $Y = \{y \in \mathcal{Y} : \|y\| \leq \sqrt{d}\}$. Letting x-axis as the iteration counter, we plot the 50 sample paths for each algorithm in Figure 3. The step sizes for S-OGDA, SMD and SMP are selected as in [11], [28] and [18], respectively. Specifically, other than SAPD, both the primal and dual step sizes are set equal, and their value is a function of $L = \max\{\mu_x, \mu_y, L_{xy}, L_{yx}\}$; indeed, S-OGDA uses $\frac{1}{8L}$, SMP uses $\frac{1}{\sqrt{3}L}$, and SMD uses $\frac{2}{\sqrt{5GN}}$, where N denotes the total iteration budget for SMD, and $G > 0$ is a fixed constant such that $\mathbb{E}[2\|\tilde{\nabla}\mathcal{L}(x, y; \omega^x, \omega^y)\|^2] \leq G$ uniformly for all $(x, y) \in X \times Y$. The step sizes for SAPD are determined by minimizing $\bar{\mathcal{R}}(\rho)$ for $\rho \in \{\rho_1, \rho_2\}$, where $\rho_1 = 0.99$ and $\rho_2 = 0.995$. This process leads to $(\tau, \sigma, \theta) = (0.010, 0.012, 0.645)$ for SAPD(ρ_1), and to $(\tau, \sigma, \theta) = (0.005, 0.008, 0.174)$ for SAPD(ρ_2). In Figure 3, we observe that *SAPD* outperforms the other algorithms in both metrics (distance to the saddle point and the gap function). Since $\rho_1 < \rho_2$, we also observe that SAPD with $\rho = \rho_1$ leads to a faster decay of the bias term than that with $\rho = \rho_2$. However, due to rate and robustness trade-off, the choice of $\rho = \rho_2$ is more robust to noise, leading to a smaller asymptotic variance of $\{z_k\}$ as expected.

4.2. Distributionally Robust Optimization with Real Data. Next, we consider ℓ_2-regularized variant of the distributionally robust optimization problem

from [25], i.e., (DRO): $\min_{x\in\mathcal{S}} \max_{y\in\mathcal{P}_r} \frac{\mu_x}{2}\|x\|^2 + \sum_{i=1}^n y_i\phi_i(x)$ where $\phi_i : \mathbb{R}^d \to \mathbb{R}$ is a strongly convex smooth loss function corresponding to the i-th data point, $\mu_x > 0$ is a regularization parameter, $\mathcal{S} \triangleq \{x \in \mathbb{R}^d : \|x\|^2 \leq \mathcal{D}_x\}$ for some given model diameter $\mathcal{D}_x > 0$ and $\mathcal{P}_r \triangleq \{y \in \mathbb{R}^n_+ : \mathbf{1}^\top y = 1, \|y - \mathbf{1}/n\|^2 \leq \frac{r}{n^2}\}$ – here, $\mathbf{1}$ denotes the vector with all entries equal to one, and $\mathcal{P}_r$ is the uncertainty set around the uniform distribution $\mathbf{1}/n$ whose radius is determined by the parameter r. In the special case when $r = 0$, the problem recovers the ERM problems arising in supervised learning from labeled data which assigns uniform weights $y_i = 1/n$ to all data points. When $r > 0$, the problem is to minimize a worst-case objective to be robust against uncertainty in the underlying data distribution. (DRO) has several advantages to construct confidence intervals for the parameters of predictive models in supervised learning, see [25]. This SP problem is affine in the dual variable y; therefore, it is not strongly convex with respect to y. However, we can approximate it, in a similar spirit to Nesterov's smoothing technique in [30], with the following SCSC problem:

$$\min_{x\in\mathcal{S}} \max_{y\in\mathcal{P}_r} \mathcal{L}_{\mu_y}(x,y) \triangleq \frac{\mu_x}{2}\|x\|^2 + \Phi(x,y) - \frac{\mu_y}{2}\|y\|^2, \tag{4.1}$$

where $\Phi(x,p) = \sum_{i=1}^n y_i\phi_i(x)$ for some properly chosen smoothing parameter $\mu_y > 0$. In our tests, we consider the binary logistic loss with an l_2 regularizer, i.e., $\phi_i(x) = \log(1 + \exp(-b_i {a_i}^\top x))$ and set $r = 2\sqrt{n}$. We can then apply SAPD to the SCSC problem (4.1) which admits the Lipschitz constants $L_{xy} = L_{yx} = \|A\|_2$, $L_{xx} = \max_{i=1,\ldots,n}\{\frac{1}{4}\|a_i\|_2^2\}$, and $L_{yy} = 0$, where $A \in \mathbb{R}^{n\times d}$ is the data matrix with rows $\{a_i\}_{i=1}^n$ and columns $\{A_j\}_{j=1}^d$. Since $\mathcal{D}_y \triangleq \sup_{y\in\mathcal{P}_r}\|y\|^2 = 1$, for any given $\epsilon > 0$, one can compute $(\bar{x},\bar{y})$ such that $\mathcal{G}(\bar{x},\bar{y}) \leq \epsilon$ for (DRO) through solving (4.1) with $\mu_y = \frac{\epsilon}{2\mathcal{D}_y}$ and generating a point $(\bar{x},\bar{y}) \in \mathcal{X}\times\mathcal{Y}$ such that $\mathbb{E}[\sup_{(x,y)\in\mathcal{S}\times\mathcal{P}_r}\{\mathcal{L}_{\mu_y}(\bar{x},y) - \mathcal{L}_{\mu_y}(x,\bar{y})\}] \leq \frac{3\epsilon}{4}$, for which the iteration complexity of SAPD method is $\mathcal{O}\left(\left[\frac{L_{xx}}{\mu_x} + \frac{L_{yx}}{\sqrt{\mu_x}}\frac{1}{\sqrt{\epsilon}} + \frac{\delta_x^2}{\sqrt{\mu_x}}\frac{1}{\epsilon^{3/2}} + \frac{\delta_y^2}{\epsilon^2}\right]\cdot\ln\left(\frac{\mathcal{D}_x\mu_x}{\epsilon} + 1\right)\right)$. Note that for $\epsilon \geq \mu_x$, we have $\mathcal{O}(\frac{\delta_x^2}{\mu_x}\frac{1}{\epsilon} + \frac{\delta_y^2}{\epsilon^2})$ and this complexity is optimal.

We perform experiments on two data sets: 1) Dry Bean data set [19] with $d = 16$, $n = 9528$ with a test data set of 4083 points; 2) Arcene data set [14] with $d = 10,000$, $n = 97$ and test size of 96 points. In these experiments we set the regularization parameter μ_x through cross-validation. The source of noise in gradient computations is mini-batch sampling of data. For the Dry Bean data set, we set $\mu_x = 0.01$, $\mu_y = 10$ and use batch size 1, and normalize each feature column $A_j \in \mathbb{R}^n$ using $A_j \leftarrow \frac{A_j - \min(A_j)}{\max(A_j) - \min(A_j)}$, where both min and max are taken over the elements of A_j. For the Arcene data set, we set $\mu_x = 0.02$ and $\mu_y = 10$, and use batch size 10, and normalize the data matrix such that $A \leftarrow A/\min\{\sqrt{d},\sqrt{n}\}$.

As described in Section 3.2, given a desired rate $\rho \geq \rho^*$, we compute (τ,σ,θ) for SAPD that achieves $\bar{\mathcal{R}}(\rho)$. We plot SAPD statistics for two different rates to illustrate that our framework can trade-off rate and robustness in an effective manner. The other methods we tested set the primal and dual step sizes equal. Indeed, for S-OGDA [16, 11] the step size is $\frac{1}{8L}$, and for SMP [18], it is $\frac{1}{\sqrt{3}L}$, where $L = \max\{L_{xx} + \mu_x, \mu_y, L_{xy}, L_{yx}\}$. In Figure 4, we plotted the optimization error using the distance metric $\mathcal{D}$, training and the test errors. We reported the results for 50 sample paths. Our results show that for both test and training errors, reported in terms of distances to the solution, SAPD achieves a good performance for both rate and robustness.

(a) Dry Bean: $\rho_1 = 0.9986$, $\rho_2 = 0.9997$.

(b) Arcene: $\rho_1 = 0.989$, $\rho_2 = 0.992$.

FIG. 4. *Comparison of SAPD against S-OGDA [11] and SMP [18] on real-data sets.*

REFERENCES

[1] N. S. AYBAT, A. FALLAH, M. GURBUZBALABAN, AND A. OZDAGLAR, *A universally optimal multistage accelerated stochastic gradient method*, in Advances in Neural Information Processing Systems, 2019, pp. 8525–8536.

[2] N. S. AYBAT, A. FALLAH, M. GURBUZBALABAN, AND A. OZDAGLAR, *Robust accelerated gradient methods for smooth strongly convex functions*, SIAM Journal on Optimization, 30 (2020), pp. 717–751.

[3] A. BECK, *First-order methods in optimization*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.

[4] A. BEN-TAL, L. EL GHAOUI, AND A. NEMIROVSKI, *Robust optimization*, vol. 28, Princeton University Press, 2009.

[5] L. BOTTOU, F. E. CURTIS, AND J. NOCEDAL, *Optimization methods for large-scale machine learning*, Siam Review, 60 (2018), pp. 223–311.

[6] A. CHAMBOLLE AND T. POCK, *A first-order primal-dual algorithm for convex problems with applications to imaging*, Journal of Math. Imaging and Vision, 40 (2011), pp. 120–145.

[7] A. CHAMBOLLE AND T. POCK, *On the ergodic convergence rates of a first-order primal–dual algorithm*, Mathematical Programming, 159 (2016), pp. 253–287.

[8] Y. CHEN, G. LAN, AND Y. OUYANG, *Accelerated schemes for a class of variational inequalities*, Mathematical Programming, 165 (2017), pp. 113–149.

[9] L. CONDAT, *Fast projection onto the simplex and the l_1 ball*, Mathematical Programming, 158 (2016), pp. 575–585.

[10] S. CUI AND U. V. SHANBHAG, *On the analysis of reflected gradient and splitting methods for monotone stochastic variational inequality problems*, in 2016 IEEE 55th Conference on Decision and Control (CDC), IEEE, 2016, pp. 4510–4515.

[11] A. FALLAH, A. OZDAGLAR, AND S. PATTATHIL, *An optimal multistage stochastic gradient method for minimax problems*, in 2020 59th IEEE Conference on Decision and Control (CDC), IEEE, 2020, pp. 3573–3579.

[12] G. GIDEL, H. BERARD, G. VIGNOUD, P. VINCENT, AND S. LACOSTE-JULIEN, *A variational inequality perspective on generative adversarial networks*, arXiv:1802.10551, (2018).

[13] G. H. GOLUB AND C. F. VAN LOAN, *Matrix computations,. johns*, Hopkins Studies in Mathematical Sciences, 3rd edition edition, (1996).

[14] I. GUYON, S. GUNN, A. BEN-HUR, AND G. DROR, *Result analysis of the Nips 2003 feature selection challenge*, Advances in neural information processing systems, 17 (2004).

[15] E. Y. HAMEDANI AND N. S. AYBAT, *A primal-dual algorithm for general convex-concave saddle*

point problems, arXiv preprint arXiv:1803.01401, (2018).

[16] Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos, *On the convergence of single-call stochastic extra-gradient methods*, in Advances in Neural Information Processing Systems, 2019, pp. 6938–6948.

[17] A. Juditsky and A. Nemirovski, *First order methods for nonsmooth convex large-scale optimization, i: general purpose methods*, Optimization for Machine Learning, 30 (2011), pp. 121–148.

[18] A. Juditsky, A. Nemirovski, and C. Tauvel, *Solving variational inequalities with stochastic mirror-prox algorithm*, Stochastic Systems, 1 (2011), pp. 17–58.

[19] M. Koklu and I. A. Ozkan, *Multiclass classification of dry beans using computer vision and machine learning techniques*, Computers and Electronics in Agriculture, 174 (2020).

[20] N. Kuru, Ş. İ. Birbil, M. Gurbuzbalaban, and S. Yildirim, *Differentially private accelerated optimization algorithms*, arXiv preprint arXiv:2008.01989, (2020).

[21] T. Liang and J. Stokes, *Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks*, in The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, 2019, pp. 907–915.

[22] T. Lin, C. Jin, and M. I. Jordan, *Near-Optimal Algorithms for Minimax Optimization*, arXiv e-prints, (2020), arXiv:2002.02417, p. arXiv:2002.02417, https://arxiv.org/abs/2002.02417.

[23] Y. Liu, J. Peng, J. J. Q. Yu, and Y. Wu, *Ppgan: Privacy-preserving generative adversarial network*, in 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS), 2019, pp. 985–989.

[24] A. Mokhtari, A. Ozdaglar, and S. Pattathil, *A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach*, in International Conference on Artificial Intelligence and Statistics, 2020, pp. 1497–1507.

[25] H. Namkoong and J. C. Duchi, *Stochastic gradient methods for distributionally robust optimization with f-divergences*, in Advances in NeurIPS, 2016, pp. 2208–2216.

[26] A. Nedić and A. Ozdaglar, *Subgradient methods for saddle-point problems*, Journal of optimization theory and applications, 142 (2009), pp. 205–228.

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

[28] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stoc. programming*, SIAM Journal on Optimization, 19 (2009), pp. 1574–1609.

[29] Y. Nesterov, *Excessive gap technique in nonsmooth convex minimization*, SIAM Journal on Optimization, 16 (2005), pp. 235–249.

[30] Y. Nesterov, *Smooth minimization of non-smooth functions*, Mathematical Programming, 103 (2005), pp. 127–152.

[31] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, Mathematical Programming, 120 (2009), pp. 221–259.

[32] B. Palaniappan and F. Bach, *Stochastic variance reduction methods for saddle-point problems*, in Advances in Neural Information Processing Systems, 2016, pp. 1416–1424.

[33] G. Strang, *Linear algebra and its applications.: Thomson Brooks*, Cole, Belmont, CA, USA, (2005).

[34] C. Tan, T. Zhang, S. Ma, and J. Liu, *Stochastic primal-dual method for empirical risk minimization with $\mathcal{O}(1)$ per-iteration complexity*, in Advances in NeurIPS, vol. 31, 2018.

[35] Y. Wang and J. Li, *Improved Algorithms for Convex-Concave Minimax Optimization*, arXiv e-prints, (2020), arXiv:2006.06359, p. arXiv:2006.06359, https://arxiv.org/abs/2006.06359.

[36] J. Wen, C.-N. Yu, and R. Greiner, *Robust learning under uncertain test distributions: Relating covariate shift to model misspecification*, in Proceedings of the 31st International Conference on Machine Learning, E. P. Xing and T. Jebara, eds., vol. 32 of Proceedings of Machine Learning Research, Bejing, China, 22–24 Jun 2014, PMLR, pp. 631–639, https://proceedings.mlr.press/v32/wen14.html.

[37] M. Wuertz, *The implicit function theorem for Lipschitz functions and applications*, Ph.D. Thesis, University of Missouri-Columbia, 2008.

[38] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, *Differentially private generative adversarial network*, arXiv preprint arXiv:1802.06739, (2018).

[39] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans, *Maximum margin clustering*, in Advances in neural information processing systems, 2005, pp. 1537–1544.

[40] Y. Yan, Y. Xu, Q. Lin, W. Liu, and T. Yang, *Sharp analysis of epoch stochastic gradient descent ascent methods for min-max optimization*, arXiv:2002.05309, (2020).

[41] J. Zhang, M. Hong, and S. Zhang, *On lower iteration complexity bounds for the saddle point problems*, arXiv preprint arXiv:1912.07481, (2019).

[42] Y. Zhang and L. Xiao, *Stochastic primal-dual coordinate method for regularized empirical risk minimization*, The Journal of Machine Learning Research, 18 (2017), pp. 2939–2980.
[43] R. Zhao, *Optimal stochastic algorithms for convex-concave saddle-point problems*, arXiv preprint arXiv:1903.01687, (2019).
[44] J. Zhong, X. Liu, and C.-J. Hsieh, *Improving the speed and quality of GAN by adversarial training*, arXiv preprint arXiv:2008.03364, (2020).
[45] K. Zhou, J. C. Doyle, K. Glover, et al., *Robust and optimal control*, vol. 40, Prentice hall New Jersey, 1996.

Appendix A. Proofs. We will derive some key inequalities below for SAPD iterates $\{x_k, y_k\}_{k\geq 0}$ generated by Algorithm 1.1. For $k \geq 0$, let q_k and s_k be the noiseless versions of $\tilde{q}_k$ and $\tilde{s}_k$, i.e.,

$$q_k \triangleq \nabla_y \Phi(x_k, y_k) - \nabla_y \Phi(x_{k-1}, y_{k-1}), \qquad s_k \triangleq \nabla_y \Phi(x_k, y_k) + \theta q_k. \tag{A.1}$$

Recall $x_{-1} = x_0,\ y_{-1} = y_0$, thus $q_0 = \mathbf{0}$; and for $k \geq 0$, Assumption 1.1 implies that

$$\|q_{k+1}\| \leq L_{yx}\|x_{k+1} - x_k\| + L_{yy}\|y_{k+1} - y_k\|. \tag{A.2}$$

Lemma A.1. *Let $\{x_k, y_k\}_{k\geq 0}$ be SAPD iterates generated according to Algorithm 1.1. Then for all $x \in \mathbf{dom}\, f \subset \mathcal{X}$, $y \in \mathbf{dom}\, g \subset \mathcal{Y}$, and $k \geq 0$,*

$$\begin{aligned} &\mathcal{L}(x_{k+1}, y) - \mathcal{L}(x, y_{k+1}) \\ &\leq -\langle q_{k+1}, y_{k+1} - y\rangle + \theta\langle q_k, y_k - y\rangle + \Lambda_k(x, y) - \Sigma_{k+1}(x, y) + \Gamma_{k+1} + \varepsilon_k^x + \varepsilon_k^y, \end{aligned} \tag{A.3}$$

where $\varepsilon_k^x \triangleq \langle \tilde{\nabla}_x \Phi(x_k, y_{k+1}; \omega_k^x) - \nabla_x \Phi(x_k, y_{k+1}),\ x - x_{k+1}\rangle$ and $\varepsilon_k^y \triangleq \langle \tilde{s}_k - s_k, y_{k+1} - y\rangle$, q_k and s_k are defined as in (A.1), and

$$\begin{aligned} \Lambda_k(x, y) &\triangleq \frac{1}{2\tau}\|x - x_k\|^2 + \frac{1}{2\sigma}\|y - y_k\|^2, \\ \Sigma_{k+1}(x, y) &\triangleq (\frac{1}{2\tau} + \frac{\mu_x}{2})\|x - x_{k+1}\|^2 + (\frac{1}{2\sigma} + \frac{\mu_y}{2})\|y - y_{k+1}\|^2, \\ \Gamma_{k+1} &\triangleq (\frac{L_{xx}}{2} - \frac{1}{2\tau})\|x_{k+1} - x_k\|^2 - \frac{1}{2\sigma}\|y_{k+1} - y_k\|^2 \\ &\quad + \theta L_{yx}\|x_k - x_{k-1}\|\|y_{k+1} - y_k\| + \theta L_{yy}\|y_k - y_{k-1}\|\|y_{k+1} - y_k\|. \end{aligned}$$

Proof. Fix $x \in \mathbf{dom}\, f$ and $y \in \mathbf{dom}\, g$ arbitrary. Using Lemma 7.1 from [15] for the $y-$ and $x-$subproblems in Algorithm 1.1, we get two inequalities:

$$\begin{aligned} &f(x_{k+1}) + \langle \tilde{\nabla}_x \Phi(x_k, y_{k+1}; \omega_k^x), x_{k+1} - x\rangle \\ &\leq f(x) + \frac{1}{2\tau}\left[\|x - x_k\|^2 - \|x - x_{k+1}\|^2 - \|x_{k+1} - x_k\|^2\right] - \frac{\mu_x}{2}\|x - x_{k+1}\|^2, \end{aligned}$$

$$\begin{aligned} &g(y_{k+1}) - \langle \tilde{s}_k, y_{k+1} - y\rangle \\ &\leq g(y) + \frac{1}{2\sigma}\left[\|y - y_k\|^2 - \|y - y_{k+1}\|^2 - \|y_{k+1} - y_k\|^2\right] - \frac{\mu_y}{2}\|y - y_{k+1}\|^2. \end{aligned}$$

Using the following two equalities:

$$\begin{aligned} &\langle \tilde{\nabla}_x \Phi(x_k, y_{k+1}; \omega_k^x), x_{k+1} - x\rangle \\ &= \langle \tilde{\nabla}_x \Phi(x_k, y_{k+1}; \omega_k^x) - \nabla_x \Phi(x_k, y_{k+1}), x_{k+1} - x\rangle + \langle \nabla_x \Phi(x_k, y_{k+1}), x_{k+1} - x\rangle, \end{aligned}$$

and $\langle \tilde{s}_k, y_{k+1} - y\rangle = \langle s_k, y_{k+1} - y\rangle + \langle \tilde{s}_k - s_k, y_{k+1} - y\rangle$, we get

$$\begin{aligned} &f(x_{k+1}) + \langle \nabla_x \Phi(x_k, y_{k+1}),\ x_{k+1} - x\rangle \\ &\leq f(x) + \frac{1}{2\tau}(\|x - x_k\|^2 - \|x - x_{k+1}\|^2 - \|x_{k+1} - x_k\|^2) - \frac{\mu_x}{2}\|x - x_{k+1}\|^2 + \varepsilon_k^x, \end{aligned} \tag{A.4a}$$

$$
\begin{aligned}
&g(y_{k+1}) - \langle s_k, y_{k+1} - y\rangle \\
&\le g(y) + \frac{1}{2\sigma}(\|y - y_k\|^2 - \|y - y_{k+1}\|^2 - \|y_{k+1} - y_k\|^2) - \frac{\mu_y}{2}\|y - y_{k+1}\|^2 + \varepsilon_k^y.
\end{aligned} \tag{A.4b}
$$

Rearranging the terms in (A.4b), we get

$$
\begin{aligned}
&-g(y) + g(y_{k+1}) \\
&\le \langle s_k, y_{k+1} - y\rangle + \frac{1}{2\sigma}\left[\|y - y_k\|^2 - \|y - y_{k+1}\|^2 - \|y_{k+1} - y_k\|^2\right] - \frac{\mu_y}{2}\|y - y_{k+1}\|^2 + \varepsilon_k^y.
\end{aligned} \tag{A.5}
$$

Since $y_{k+1} \in \mathbf{dom}\, g$, the inner product in (A.4a) can be lower bounded using convexity of $\Phi(\cdot, y_{k+1})$ in Assumption 1.1 as follows:

$$
\begin{aligned}
\langle \nabla_x \Phi(x_k, y_{k+1}), x_{k+1} - x\rangle =& \langle \nabla_x \Phi(x_k, y_{k+1}), x_k - x\rangle + \langle \nabla_x \Phi(x_k, y_{k+1}), x_{k+1} - x_k\rangle \\
\ge& \Phi(x_k, y_{k+1}) - \Phi(x, y_{k+1}) + \langle \nabla_x \Phi(x_k, y_{k+1}),\ x_{k+1} - x_k\rangle.
\end{aligned}
$$

Using this inequality after adding $\Phi(x_{k+1}, y_{k+1})$ to both sides of (A.4a), we get

$$
\begin{aligned}
&\Phi(x_{k+1}, y_{k+1}) + f(x_{k+1}) \\
&\le \Phi(x, y_{k+1}) + f(x) \\
&\quad + \Phi(x_{k+1}, y_{k+1}) - \Phi(x_k, y_{k+1}) - \langle \nabla_x \Phi(x_k, y_{k+1}), x_{k+1} - x_k\rangle \\
&\quad + \frac{1}{2\tau}\left[\|x - x_k\|^2 - \|x - x_{k+1}\|^2 - \|x_{k+1} - x_k\|^2\right] - \frac{\mu_x}{2}\|x - x_{k+1}\|^2 + \varepsilon_k^x \\
&\le \Phi(x, y_{k+1}) + f(x) + \frac{L_{xx}}{2}\|x_{k+1} - x_k\|^2 \\
&\quad + \frac{1}{2\tau}\left[\|x - x_k\|^2 - \|x - x_{k+1}\|^2 - \|x_{k+1} - x_k\|^2\right] - \frac{\mu_x}{2}\|x - x_{k+1}\|^2 + \varepsilon_k^x,
\end{aligned} \tag{A.6}
$$

where the last step uses (1.4). Rearranging the terms gives us

$$
\begin{aligned}
&f(x_{k+1}) - f(x) - \Phi(x, y_{k+1}) \\
&\le -\Phi(x_{k+1}, y_{k+1}) + \frac{L_{xx}}{2}\|x_{k+1} - x_k\|^2 \\
&\quad + \frac{1}{2\tau}\left[\|x - x_k\|^2 - \|x - x_{k+1}\|^2 - \|x_{k+1} - x_k\|^2\right] - \frac{\mu_x}{2}\|x - x_{k+1}\|^2 + \varepsilon_k^x.
\end{aligned} \tag{A.7}
$$

Then, for $k \ge 0$, by summing (A.5) and (A.7), we obtain

$$
\begin{aligned}
&\mathcal{L}(x_{k+1}, y) - \mathcal{L}(x, y_{k+1}) \\
&= f(x_{k+1}) + \Phi(x_{k+1}, y) - g(y) - f(x) - \Phi(x, y_{k+1}) + g(y_{k+1}) \\
&\le \Phi(x_{k+1}, y) - \Phi(x_{k+1}, y_{k+1}) + \langle s_k, y_{k+1} - y\rangle + \frac{L_{xx}}{2}\|x_{k+1} - x_k\|^2 \\
&\quad + \frac{1}{2\sigma}\left[\|y - y_k\|^2 - \|y - y_{k+1}\|^2 - \|y_{k+1} - y_k\|^2\right] - \frac{\mu_y}{2}\|y - y_{k+1}\|^2 + \varepsilon_k^y \\
&\quad + \frac{1}{2\tau}\left[\|x - x_k\|^2 - \|x - x_{k+1}\|^2 - \|x_{k+1} - x_k\|^2\right] - \frac{\mu_x}{2}\|x - x_{k+1}\|^2 + \varepsilon_k^x.
\end{aligned} \tag{A.8}
$$

From Assumption 1.1, the concavity of $\Phi(x, \cdot)$ for fixed $x \in \mathbf{dom}\, f \subset \mathcal{X}$ implies

$$
\begin{aligned}
&\Phi(x_{k+1}, y) - \Phi(x_{k+1}, y_{k+1}) + \langle s_k, y_{k+1} - y\rangle \\
&\le \langle \nabla_y \Phi(x_{k+1}, y_{k+1}), y - y_{k+1}\rangle + \langle \nabla_y \Phi(x_k, y_k) + \theta q_k, y_{k+1} - y\rangle \\
&= \langle \nabla_y \Phi(x_{k+1}, y_{k+1}) - \nabla_y \Phi(x_k, y_k), y - y_{k+1}\rangle + \theta\langle q_k, y_{k+1} - y\rangle \\
&= -\langle q_{k+1}, y_{k+1} - y\rangle + \theta\langle q_k, y_k - y\rangle + \theta\langle q_k, y_{k+1} - y_k\rangle.
\end{aligned}
$$

Thus, using the above inequality within (A.8), we get

$$\begin{aligned}
&\mathcal{L}(x_{k+1}, y) - \mathcal{L}(x, y_{k+1}) \\
&\leq -\langle q_{k+1}, y_{k+1} - y\rangle + \theta\langle q_k, y_k - y\rangle + \theta\langle q_k, y_{k+1} - y_k\rangle + \frac{L_{xx}}{2}\|x_{k+1} - x_k\|^2 \\
&\quad + \frac{1}{2\sigma}\left[\|y - y_k\|^2 - \|y - y_{k+1}\|^2 - \|y_{k+1} - y_k\|^2\right] - \frac{\mu_y}{2}\|y - y_{k+1}\|^2 \\
&\quad + \frac{1}{2\tau}\left[\|x - x_k\|^2 - \|x - x_{k+1}\|^2 - \|x_{k+1} - x_k\|^2\right] - \frac{\mu_x}{2}\|x - x_{k+1}\|^2 + \varepsilon_k^x + \varepsilon_y^x.
\end{aligned}$$

Finally, (A.3) follows from using Cauchy-Schwarz for $\langle q_k, y_{k+1} - y_k\rangle$ and (A.2). ∎

LEMMA A.2. *Let f be a proper closed and strongly convex function with convexity modulus $\mu > 0$. Then for any $x, x' \in \mathbf{dom}\, f$, and positive constant $c > 0$,*

$$\|\mathbf{prox}_{cf}(x) - \mathbf{prox}_{cf}(x')\| \leq \frac{1}{1 + c\mu}\|x - x'\|.$$

Because Lemma A.2 is a simple extension of [3, Theorem 6.42] to the strongly convex scenario, we do not provide its proof. Next, based on the above inequality, we prove an intermediate result that assists to bound the variance of the SAPD iterate sequence.

LEMMA A.3. *Let $\{x_k, y_k\}_{k\geq 0}$ be SAPD iterates generated according to Algorithm 1.1. For $k \geq 0$, let q_k and s_k be defined as in (A.1), and let*

$$\begin{aligned}
\hat{x}_{k+1} &\triangleq \mathbf{prox}_{\tau f}\left(x_k - \tau\nabla_x\Phi(x_k, y_{k+1})\right), \\
\hat{y}_{k+1} &\triangleq \mathbf{prox}_{\sigma g}\left(y_k + \sigma s_k\right), \\
\hat{\hat{x}}_{k+1} &\triangleq \mathbf{prox}_{\tau f}\left(x_k - \tau\nabla_x\Phi(x_k, \hat{y}_{k+1})\right), \\
\hat{\hat{y}}_{k+1} &\triangleq \mathbf{prox}_{\sigma g}\left(\hat{y}_k + \sigma(1+\theta)\nabla_y\Phi(\hat{\hat{x}}_k, \hat{y}_k) - \sigma\theta\nabla_y\Phi(x_{k-1}, y_{k-1})\right),
\end{aligned} \tag{A.9}$$

then the following inequalities hold for $k \geq 0$:

$$\|x_{k+1} - \hat{x}_{k+1}\| \leq \frac{\tau}{1 + \tau\mu_x}\|\Delta_k^{\Phi,x}\|, \tag{A.10a}$$

$$\|y_{k+1} - \hat{y}_{k+1}\| \leq \frac{\sigma}{1 + \sigma\mu_y}\left((1+\theta)\|\Delta_k^{\Phi,y}\| + \theta\|\Delta_{k-1}^{\Phi,y}\|\right), \tag{A.10b}$$

$$\begin{aligned}
\|y_{k+1} - \hat{\hat{y}}_{k+1}\| \leq \frac{\sigma}{1 + \sigma\mu_y}\Bigg(&(1+\theta)\|\Delta_k^{\Phi,y}\| + \theta\|\Delta_{k-1}^{\Phi,y}\| + \frac{\tau(1+\theta)L_{yx}}{1 + \tau\mu_x}\|\Delta_{k-1}^{\Phi,x}\| \\
&+ \left(\frac{1 + \sigma(1+\theta)L_{yy}}{1 + \sigma\mu_y} + \frac{\tau\sigma(1+\theta)L_{yx}L_{xy}}{(1+\tau\mu_x)(1+\sigma\mu_y)}\right)\left((1+\theta)\|\Delta_{k-1}^{\Phi,y}\| + \theta\|\Delta_{k-2}^{\Phi,y}\|\right)\Bigg),
\end{aligned} \tag{A.10c, A.10d}$$

where

$$\Delta_k^{\Phi,x} \triangleq \tilde{\nabla}_x\Phi(x_k, y_{k+1}; \omega_k^x) - \nabla_x\Phi(x_k, y_{k+1}), \tag{A.11a}$$

$$\Delta_k^{\Phi,y} \triangleq \tilde{\nabla}_y\Phi(x_k, y_k; \omega_k^y) - \nabla_y\Phi(x_k, y_k). \tag{A.11b}$$

Proof. First, (A.10a) is a direct result of Lemma A.2. For (A.10b), we have that

$$\|y_{k+1} - \hat{y}_{k+1}\| \leq \frac{\sigma}{1 + \sigma\mu_y}\|\tilde{s}_k - s_k\| \leq \frac{\sigma}{1 + \sigma\mu_y}\left((1+\theta)\|\Delta_k^{\Phi,y}\| + \theta\|\Delta_{k-1}^{\Phi,y}\|\right),$$

which follows from Lemma A.2 and the triangle inequality. To show (A.10d), we bound $\|y_{k+1} - \hat{y}_{k+1}\|$ and $\|\hat{y}_{k+1} - \hat{\hat{y}}_{k+1}\|$ separately. It follows from Lemma A.2 that

$$\|x_{k+1} - \hat{\hat{x}}_{k+1}\| \leq \frac{\tau}{1 + \tau\mu_x}\|\tilde{\nabla}_x\Phi(x_k, y_{k+1}; \omega_k^x) - \nabla_x\Phi(x_k, \hat{y}_{k+1})\|.$$

After adding and subtracting $\nabla_x \Phi(x_k, y_{k+1})$, Assumption 1.1 implies that

$$\|x_{k+1} - \hat{\hat{x}}_{k+1}\| \leq \frac{\tau}{1+\tau\mu_x}\left(\|\Delta_k^{\Phi,x}\| + L_{xy}\|y_{k+1} - \hat{y}_{k+1}\|\right). \tag{A.12}$$

We will use this relation to bound $\|\hat{y}_{k+1} - \hat{\hat{y}}_{k+1}\|$. Indeed, using Lemma A.2, we have

$$\begin{aligned}
\|\hat{y}_{k+1} - \hat{\hat{y}}_{k+1}\| \leq & \frac{1}{1+\sigma\mu_y}\|y_k - \hat{y}_k + \sigma(1+\theta)\left(\nabla_y \Phi(x_k, y_k) - \nabla_y \Phi(\hat{\hat{x}}_k, \hat{y}_k)\right)\| \\
\leq & \frac{1}{1+\sigma\mu_y}\left((1+\sigma(1+\theta)L_{yy})\|y_k - \hat{y}_k\| + \sigma(1+\theta)L_{yx}\|x_k - \hat{\hat{x}}_k\|\right) \\
\leq & \frac{1}{1+\sigma\mu_y}\left(\left(1+\sigma(1+\theta)L_{yy} + \frac{\tau\sigma(1+\theta)L_{yx}L_{xy}}{1+\tau\mu_x}\right)\|y_k - \hat{y}_k\|\right. \\
& \left. + \frac{\tau\sigma(1+\theta)L_{yx}}{1+\tau\mu_x}\|\Delta_{k-1}^{\Phi,x}\|\right) \\
\leq & \frac{\sigma}{1+\sigma\mu_y}\left(\left(\frac{1+\sigma(1+\theta)L_{yy}}{1+\sigma\mu_y} + \frac{\tau\sigma(1+\theta)L_{yx}L_{xy}}{(1+\tau\mu_x)(1+\sigma\mu_y)}\right)\cdot\right. \\
& \left.\left((1+\theta)\|\Delta_{k-1}^{\Phi,y}\| + \theta\|\Delta_{k-2}^{\Phi,y}\|\right) + \frac{\tau(1+\theta)L_{yx}}{1+\tau\mu_x}\|\Delta_{k-1}^{\Phi,x}\|\right),
\end{aligned} \tag{A.13}$$

where the second, the third and the fourth inequalities follow from Assumption 1.1, (A.12) and (A.10b), respectively. The triangle inequality $\|y_{k+1} - \hat{\hat{y}}_{k+1}\| \leq \|y_{k+1} - \hat{y}_{k+1}\| + \|\hat{y}_{k+1} - \hat{\hat{y}}_{k+1}\|$, (A.10b) and (A.13) give us the desired bound. ∎

Next, we provide some inequalities to bound the variance later in our analysis.

LEMMA A.4. *Let $\{x_k, y_k\}_{k\geq 0}$ be SAPD iterates generated according to Algorithm 1.1. The following inequality holds for all $k \geq 0$:*

$$\mathbb{E}\left[\langle \Delta_k^{\Phi,x}, \hat{x}_{k+1} - x_{k+1}\rangle\right] \leq \frac{\tau}{1+\tau\mu_x}\delta_x^2, \qquad \mathbb{E}\left[\langle \Delta_k^{\Phi,y}, y_{k+1} - \hat{y}_{k+1}\rangle\right] \leq \frac{\sigma(1+2\theta)}{1+\sigma\mu_y}\delta_y^2,$$

$$\begin{aligned}
\mathbb{E}\left[\langle \Delta_{k-1}^{\Phi,y}, \hat{\hat{y}}_{k+1} - y_{k+1}\rangle\right] \leq & \frac{\sigma}{1+\sigma\mu_y}\left[\left(\left(1 + \frac{1+\sigma(1+\theta)L_{yy}}{1+\sigma\mu_y} + \frac{\tau\sigma(1+\theta)L_{yx}L_{xy}}{(1+\tau\mu_x)(1+\sigma\mu_y)}\right)\cdot\right.\right. \\
& \left.\left.(1+2\theta) + \frac{\tau(1+\theta)L_{yx}}{2(1+\tau\mu_x)}\right)\delta_y^2 + \frac{\tau(1+\theta)L_{yx}}{2(1+\tau\mu_x)}\delta_x^2\right].
\end{aligned}$$

Proof. With the convention that $y_{-2} = y_{-1} = y_0$, and $x_{-2} = x_{-1} = x_0$, Lemma A.3 and Cauchy-Schwarz inequality imply for all $k \geq 0$ that

$$\begin{aligned}
& \langle \Delta_k^{\Phi,x}, x_{k+1} - \hat{x}_{k+1}\rangle \leq \frac{\tau}{1+\tau\mu_x}\|\Delta_k^{\Phi,x}\|^2, \\
& \langle \Delta_k^{\Phi,y}, y_{k+1} - \hat{y}_{k+1}\rangle \leq \frac{\sigma}{1+\sigma\mu_y}\left((1+\theta)\|\Delta_k^{\Phi,y}\|^2 + \theta\|\Delta_{k-1}^{\Phi,y}\|\|\Delta_k^{\Phi,y}\|\right), \\
& \langle \Delta_{k-1}^{\Phi,y}, y_{k+1} - \hat{\hat{y}}_{k+1}\rangle \leq \frac{\sigma}{1+\sigma\mu_y}\left((1+\theta)\|\Delta_k^{\Phi,y}\|\|\Delta_{k-1}^{\Phi,y}\| + \theta\|\Delta_{k-1}^{\Phi,y}\|^2\right. \\
& + \frac{\tau(1+\theta)L_{yx}}{1+\tau\mu_x}\|\Delta_{k-1}^{\Phi,x}\|\|\Delta_{k-1}^{\Phi,y}\| + \left(\frac{1+\sigma(1+\theta)L_{yy}}{1+\sigma\mu_y} + \frac{\tau\sigma(1+\theta)L_{yx}L_{xy}}{(1+\tau\mu_x)(1+\sigma\mu_y)}\right) \\
& \left.\cdot\left((1+\theta)\|\Delta_{k-1}^{\Phi,y}\|^2 + \theta\|\Delta_{k-2}^{\Phi,y}\|\|\Delta_{k-1}^{\Phi,y}\|\right)\right).
\end{aligned}$$

Next, using Assumption 1.2 and $\|a\|\|b\| \leq \frac{1}{2}\|b\|^2 + \frac{1}{2}\|b\|^2$, which holds for $a, b \in \mathbb{R}^n$, and taking the expectation leads to the desired result. ∎

Before we move on to prove our main result in Theorem 2.1, we give two technical lemmas that help us simplify the SAPD parameter selection rule to the LMI in (2.1).

LEMMA A.5. *Let $G \in \mathbb{R}^{5\times 5}$ be the matrix on the left-hand-side of (2.1), and*

$$G' \triangleq \begin{pmatrix} \frac{1}{\tau}+\mu_x-\frac{1}{\rho\tau} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\sigma}+\mu_y-\frac{1}{\rho\sigma} & -|1-\frac{\theta}{\rho}|\, L_{yx} & -|1-\frac{\theta}{\rho}|\, L_{yy} & 0 \\ 0 & -|1-\frac{\theta}{\rho}|\, L_{yx} & \frac{1}{\tau}-L_{xx} & 0 & -\frac{\theta}{\rho}L_{yx} \\ 0 & -|1-\frac{\theta}{\rho}|\, L_{yy} & 0 & \frac{1}{\sigma}-\alpha & -\frac{\theta}{\rho}L_{yy} \\ 0 & 0 & -\frac{\theta}{\rho}L_{yx} & -\frac{\theta}{\rho}L_{yy} & \frac{\alpha}{\rho} \end{pmatrix},$$

then $G \succeq 0$ is equivalent to $G' \succeq 0$.

Proof. $\forall\ \mathbf{y} = (y_1, y_2, y_3, y_4, y_5)^\top \in \mathbb{R}^5$, letting $\tilde{\mathbf{y}} = (y_1, -y_2, y_3, y_4, y_5)^\top$, we have

$$\mathbf{y}^\top G'\mathbf{y} = \begin{cases} \mathbf{y}^\top G\mathbf{y} & \text{if } \theta \le \rho, \\ \tilde{\mathbf{y}}^\top G\tilde{\mathbf{y}} & \text{else;} \end{cases} \qquad \mathbf{y}^\top G\mathbf{y} = \begin{cases} \mathbf{y}^\top G'\mathbf{y} & \text{if } \theta \le \rho, \\ \tilde{\mathbf{y}}^\top G'\tilde{\mathbf{y}} & \text{else.} \end{cases}$$

Thus, $G_1 \succeq 0$ is equivalent to $G_2 \succeq 0$. ∎

LEMMA A.6. *Suppose the parameters $\tau, \sigma, \theta > 0$ satisfy (2.1) for some $\alpha \in [0, \frac{1}{\sigma})$ and $\rho \in (0, 1]$, then it follows that*

$$G'' \triangleq \begin{pmatrix} \frac{1}{\sigma}(1-\frac{1}{\rho})+\mu_y+\frac{\alpha}{\rho} & (-|1-\frac{\theta}{\rho}|-\frac{\theta}{\rho})L_{yx} & (-|1-\frac{\theta}{\rho}|-\frac{\theta}{\rho})L_{yy} \\ (-|1-\frac{\theta}{\rho}|-\frac{\theta}{\rho})L_{yx} & \frac{1}{\tau}-L_{xx} & 0 \\ (-|1-\frac{\theta}{\rho}|-\frac{\theta}{\rho})L_{yy} & 0 & \frac{1}{\sigma}-\alpha \end{pmatrix} \succeq 0. \tag{A.14}$$

Proof. Since (2.1) holds, $\forall (x_1, x_2, x_3) \in \mathbb{R}^3$, by Lemma A.5 we have that

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^\top G'' \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ x_1 \\ x_2 \\ x_3 \\ x_1 \end{pmatrix}^\top G' \begin{pmatrix} 0 \\ x_1 \\ x_2 \\ x_3 \\ x_1 \end{pmatrix} \ge 0, \tag{A.15}$$

where G' is defined in Lemma A.5. ∎

Finally, with the following observation, we will be ready to proceed to the proof of Theorem 2.1. Let $\{\mathcal{F}_k^x\}$ and $\{\mathcal{F}_k^y\}$ be the filtrations such that $\mathcal{F}_k^x \triangleq \mathcal{F}(\{x_i\}_{i=0}^k, \{y_i\}_{i=0}^{k+1})$ and $\mathcal{F}_k^y \triangleq \mathcal{F}(\{x_i\}_{i=0}^k, \{y_i\}_{i=0}^k)$ denote the σ-algebras generated by the random variables in their arguments. A consequence of Assumption 1.2 is that for $\mathcal{F}_k^x$-measurable random variable v, i.e., $v \in \mathcal{F}_k^x$, we have that $\mathbb{E}\left[\langle \tilde{\nabla}\Phi_x(x_k, y_{k+1}; \omega_k^x) - \nabla\Phi_x(x_k, y_{k+1}), v\rangle\right] = 0$; similarly, for $v \in \mathcal{F}_k^y$, it holds that $\mathbb{E}\left[\langle \tilde{\nabla}\Phi_y(x_k, y_k; \omega_k^y) - \nabla\Phi_y(x_k, y_k), v\rangle\right] = 0$.

A.1. Proof of Theorem 2.1.

Proof. Fix arbitrary $(x, y) \in X \times Y \subset \mathbf{dom}\, f \times \mathbf{dom}\, g$. Since $(x_{k+1}, y_{k+1}) \in \mathbf{dom}\, f \times \mathbf{dom}\, g$, using the concavity of $\mathcal{L}(x_{k+1}, \cdot)$ and the convexity of $\mathcal{L}(\cdot, y_{k+1})$, Jensen's lemma immediately implies that

$$K_N(\rho)\left(\mathcal{L}(\bar{x}_N, y) - \mathcal{L}(x, \bar{y}_N)\right) \le \sum_{k=0}^{N-1} \rho^{-k}\left(\mathcal{L}(x_{k+1}, y) - \mathcal{L}(x, y_{k+1})\right), \ \forall \rho \in (0, 1]. \tag{A.16}$$

Thus, if we multiply ρ^{-k} for both sides of (A.3) and sum the resulting inequality from $k=0$ to $N-1$, then using (A.16) we get

$$\begin{aligned} &K_N(\rho)\left(\mathcal{L}(\bar{x}_N,y)-\mathcal{L}(x,\bar{y}_N)\right)\\ &\le \sum_{k=0}^{N-1}\rho^{-k}\Big(\underbrace{-\langle q_{k+1},y_{k+1}-y\rangle+\theta\langle q_k,y_k-y\rangle}_{\textbf{part 1}}+\Lambda_k(x,y)-\Sigma_{k+1}(x,y)+\Gamma_{k+1}\\ &\quad\underbrace{-\langle\tilde{\nabla}_x\Phi(x_k,y_{k+1};\omega_k^x)-\nabla_x\Phi(x_k,y_{k+1}),x_{k+1}-x\rangle}_{\textbf{part 2}}+\underbrace{\langle\tilde{s}_k-s_k,y_{k+1}-y\rangle}_{\textbf{part 3}}\Big). \end{aligned} \tag{A.17}$$

Using Cauchy–Schwarz inequality and (A.2) leads to

$$|\langle q_{k+1},y_{k+1}-y\rangle|\le S_{k+1}\triangleq L_{yx}\|x_{k+1}-x_k\|\|y_{k+1}-y\|+L_{yy}\|y_{k+1}-y_k\|\|y_{k+1}-y\| \tag{A.18}$$

for $k\ge -1$. Recall $x_{-1}=x_0,\ y_{-1}=y_0$, thus $q_0=\mathbf{0}$; therefore, for **part 1**,

$$\begin{aligned} &\sum_{k=0}^{N-1}\rho^{-k}\Big(-\langle q_{k+1},y_{k+1}-y\rangle+\theta\langle q_k,y_k-y\rangle\Big)\\ &=\sum_{k=0}^{N-2}\rho^{-k}(\frac{\theta}{\rho}-1)\langle q_{k+1},y_{k+1}-y\rangle-\rho^{-N+1}\langle q_N,y_N-y\rangle\\ &\le\sum_{k=0}^{N-2}\rho^{-k}|1-\frac{\theta}{\rho}|\ S_{k+1}+\rho^{-N+1}S_N\\ &=\sum_{k=0}^{N-1}\rho^{-k}|1-\frac{\theta}{\rho}|\ S_{k+1}+\rho^{-N+1}\frac{\theta}{\rho}S_N, \end{aligned} \tag{A.19}$$

where the first inequality follows from (A.18). Next we consider **part 2**, let $\Delta_k^{\Phi,x}$ be defined as in (A.11a). For some arbitrary $\eta_x>0$, define $\{\tilde{x}_k\}$ sequence as follows:

$$\tilde{x}_0\triangleq x_0,\quad \tilde{x}_{k+1}\triangleq\mathbf{argmin}_{x'\in X}-\langle\Delta_k^{\Phi,x},x'\rangle+\frac{\eta_x}{2}\|x'-\tilde{x}_k\|^2,\quad\forall\ k\ge 0. \tag{A.20}$$

Then by [28, Lemma 2.1], for all $k\ge 0$ and $x\in X$, we have that

$$\langle\Delta_k^{\Phi,x},x-\tilde{x}_k\rangle\le\frac{\eta_x}{2}\|x-\tilde{x}_k\|^2-\frac{\eta_x}{2}\|x-\tilde{x}_{k+1}\|^2+\frac{1}{2\eta_x}\|\Delta_k^{\Phi,x}\|^2.$$

Thus, using $\tilde{x}_0=x_0$ we get

$$\begin{aligned} &\sum_{k=0}^{N-1}\rho^{-k}\langle\Delta_k^{\Phi,x},x-\tilde{x}_k\rangle\le\sum_{k=0}^{N-1}\rho^{-k}\Big(\frac{\eta_x}{2}\|x-\tilde{x}_k\|^2-\frac{\eta_x}{2}\|x-\tilde{x}_{k+1}\|^2+\frac{1}{2\eta_x}\|\Delta_k^{\Phi,x}\|^2\Big)\\ &=\frac{\eta_x}{2}(\|x-x_0\|^2-\rho^{-N+1}\|x-\tilde{x}_N\|^2)+\sum_{k=0}^{N-1}\frac{\rho^{-k}}{2\eta_x}\|\Delta_k^{\Phi,x}\|^2-\sum_{k=1}^{N-1}\frac{\eta_x}{2}\rho^{-k}(1-\rho)\|x-\tilde{x}_k\|^2\\ &\le\frac{\eta_x}{2}\|x-x_0\|^2+\frac{1}{2\eta_x}\sum_{k=0}^{N-1}\rho^{-k}\|\Delta_k^{\Phi,x}\|^2, \end{aligned} \tag{A.21}$$

where for the last inequality we use $\rho\le 1$; hence, **part 2** becomes

$$\begin{aligned} &\sum_{k=0}^{N-1}-\rho^{-k}\langle\Delta_k^{\Phi,x},x_{k+1}-x\rangle\\ &=\sum_{k=0}^{N-1}\rho^{-k}\Big(\langle\Delta_k^{\Phi,x},\hat{x}_{k+1}-x_{k+1}\rangle-\langle\Delta_k^{\Phi,x},\hat{x}_{k+1}-\tilde{x}_k\rangle+\langle\Delta_k^{\Phi,x},x-\tilde{x}_k\rangle\Big)\\ &\le\sum_{k=0}^{N-1}\rho^{-k}\Big(\langle\Delta_k^{\Phi,x},\hat{x}_{k+1}-x_{k+1}\rangle-\langle\Delta_k^{\Phi,x},\hat{x}_{k+1}-\tilde{x}_k\rangle+\frac{1}{2\eta_x}\|\Delta_k^{\Phi,x}\|^2\Big)+\frac{\eta_x}{2}\|x-x_0\|^2, \end{aligned} \tag{A.22}$$

where $\hat{x}_{k+1}$ is defined in (A.9) and the inequality follows from (A.21). When $\delta_x = 0$, clearly $\Delta_k^{\Phi,x} = \mathbf{0}$; thus, **part 2** is equal to 0 and we can set $\eta_x = 0$ for which (A.22) trivially becomes $0 \le 0$. Next, we consider **part 3**, let $\Delta_k^{\Phi,y}$ be defined as in (A.11b).

For some arbitrary $\eta_y > 0$, we construct two auxiliary sequences as follows: let $\tilde{y}_0^+ = \tilde{y}_0^- = y_0$, and we define

$$\tilde{y}_{k+1}^+ \triangleq \underset{y' \in Y}{\mathbf{argmin}} \langle \Delta_k^{\Phi,y}, y' \rangle + \frac{\eta_y}{2}\|y' - \tilde{y}_k^+\|^2, \quad \tilde{y}_{k+1}^- \triangleq \underset{y' \in Y}{\mathbf{argmin}} -\langle \Delta_k^{\Phi,y}, y' \rangle + \frac{\eta_y}{2}\|y' - \tilde{y}_k^-\|^2$$

for $k \ge 0$. Thus, it follows from [28, Lemma 2.1] that for $y \in Y$,

$$\begin{aligned} \langle \Delta_k^{\Phi,y}, \tilde{y}_k^+ - y \rangle &\le \frac{\eta_y}{2}\|y - \tilde{y}_k^+\|^2 - \frac{\eta_y}{2}\|y - \tilde{y}_{k+1}^+\|^2 + \frac{1}{2\eta_y}\|\Delta_k^{\Phi,y}\|^2, \\ \langle \Delta_k^{\Phi,y}, y - \tilde{y}_k^- \rangle &\le \frac{\eta_y}{2}\|y - \tilde{y}_k^-\|^2 - \frac{\eta_y}{2}\|y - \tilde{y}_{k+1}^-\|^2 + \frac{1}{2\eta_y}\|\Delta_k^{\Phi,y}\|^2. \end{aligned} \tag{A.23}$$

Therefore, similar to (A.21), using (A.23) we get[7]

$$\begin{aligned} &\sum_{k=0}^{N-1} \rho^{-k}\Big[(1+\theta)\langle \Delta_k^{\Phi,y}, \tilde{y}_k^+ - y \rangle + \theta \langle \Delta_{k-1}^{\Phi,y}, y - \tilde{y}_{k-1}^- \rangle\Big] \\ &\qquad \le \frac{\eta_y}{2}(1+2\theta)\|y - y_0\|^2 + \frac{1}{2\eta_y}\sum_{k=0}^{N-1} \rho^{-k}\Big((1+\theta)\|\Delta_k^{\Phi,y}\|^2 + \theta\|\Delta_{k-1}^{\Phi,y}\|^2\Big). \end{aligned} \tag{A.24}$$

Next, using (A.24), we can bound **part 3** as follows:

$$\begin{aligned} &\sum_{k=0}^{N-1} \rho^{-k}\langle \tilde{s}_k - s_k, y_{k+1} - y \rangle \\ &= \sum_{k=0}^{N-1} \rho^{-k}\Big[(1+\theta)\langle \Delta_k^{\Phi,y}, y_{k+1} - y \rangle - \theta\langle \Delta_{k-1}^{\Phi,y}, y_{k+1} - y \rangle\Big] \\ &= \sum_{k=0}^{N-1} \rho^{-k}\Big[(1+\theta)\langle \Delta_k^{\Phi,y}, y_{k+1} - \hat{y}_{k+1} + \hat{y}_{k+1} - \tilde{y}_k^+ + \tilde{y}_k^+ - y \rangle \\ &\quad - \theta\langle \Delta_{k-1}^{\Phi,y}, y_{k+1} - \hat{\hat{y}}_{k+1} + \hat{\hat{y}}_{k+1} - \tilde{y}_{k-1}^- + \tilde{y}_{k-1}^- - y \rangle\Big] \\ &\le \sum_{k=0}^{N-1} \rho^{-k}\Big[(1+\theta)\langle \Delta_k^{\Phi,y}, y_{k+1} - \hat{y}_{k+1} + \hat{y}_{k+1} - \tilde{y}_k^+ \rangle \\ &\quad - \theta\langle \Delta_{k-1}^{\Phi,y}, y_{k+1} - \hat{\hat{y}}_{k+1} + \hat{\hat{y}}_{k+1} - \tilde{y}_{k-1}^- \rangle\Big] \\ &\quad + \frac{1}{2\eta_y}\sum_{k=0}^{N-1} \rho^{-k}\Big((1+\theta)\|\Delta_k^{\Phi,y}\|^2 + \theta\|\Delta_{k-1}^{\Phi,y}\|^2\Big) + \frac{\eta_y}{2}(1+2\theta)\|y - y_0\|^2, \end{aligned} \tag{A.25}$$

where $\hat{y}_{k+1}$ and $\hat{\hat{y}}_{k+1}$ are defined in (A.9). Adding $\rho^{-N+1}\Delta_N(x,y)$ to both sides of (A.17), then using (A.19), (A.22) and (A.25), for any fixed $(x,y) \in X \times Y$, we get

$$\begin{aligned} &K_N(\rho)\left(\mathcal{L}(\bar{x}_N, y) - \mathcal{L}(x, \bar{y}_N)\right) + \rho^{-N+1}\Delta_N(x,y) \\ &\qquad \le U_N(x,y) + \frac{\eta_x}{2}\|x - x_0\|^2 + \frac{\eta_y}{2}(1+2\theta)\|y - y_0\|^2 + \sum_{k=0}^{N-1} \rho^{-k}(P_k + Q_k), \end{aligned} \tag{A.26}$$

[7] Similar to **part 2**, when $\delta_y = 0$, we can set $\eta_y = 0$.

where $U_N(x,y)$, P_k and Q_k for $k=0,\dots,N-1$ are defined as

(A.27)
$$\begin{aligned}
U_N(x,y) &\triangleq \sum_{k=0}^{N-1}\rho^{-k}\left(\Gamma_{k+1}+\Lambda_k(x,y)-\Sigma_{k+1}(x,y)+|1-\frac{\theta}{\rho}|\ S_{k+1}\right)-\rho^{-N+1}(-\Delta_N(x,y)-\frac{\theta}{\rho}S_N),\\
P_k &\triangleq -\langle\Delta_k^{\Phi,x},\hat{x}_{k+1}-\tilde{x}_k\rangle+(1+\theta)\langle\Delta_k^{\Phi,y},\hat{y}_{k+1}-\tilde{y}_k^{+}\rangle-\theta\langle\Delta_{k-1}^{\Phi,y},\hat{\hat{y}}_{k+1}-\tilde{y}_{k-1}^{-}\rangle,\\
Q_k &\triangleq \langle\Delta_k^{\Phi,x},\hat{x}_{k+1}-x_{k+1}\rangle+(1+\theta)\langle\Delta_k^{\Phi,y},y_{k+1}-\hat{y}_{k+1}\rangle-\theta\langle\Delta_{k-1}^{\Phi,y},y_{k+1}-\hat{\hat{y}}_{k+1}\rangle\\
&\quad+\frac{1}{2\eta_x}\|\Delta_k^{\Phi,x}\|^2+\frac{1+\theta}{2\eta_y}\|\Delta_k^{\Phi,y}\|^2+\frac{\theta}{2\eta_y}\|\Delta_{k-1}^{\Phi,y}\|^2.
\end{aligned}$$

For fixed $(x,y)\in X\times Y$, we first analyze $U_N(x,y)$. For any $\alpha>0$, after adding and subtracting $\frac{\alpha}{2}\|y_{k+1}-y_k\|^2$, and rearranging the terms, we get

(A.28)
$$\begin{aligned}
U_N(x,y) =&\frac{1}{2}\sum_{k=0}^{N-1}\rho^{-k}\Big(\xi_k^\top A\xi_k-\xi_{k+1}^\top B\xi_{k+1}\Big)-\rho^{-N+1}(-\Delta_N(x,y)-\frac{\theta}{\rho}S_N)\\
=&\frac{1}{2}\xi_0^\top A\xi_0-\frac{1}{2}\sum_{k=1}^{N-1}\rho^{-k+1}[\xi_k^\top(B-\frac{1}{\rho}A)\xi_k]-\rho^{-N+1}(\frac{1}{2}\xi_N^\top B\xi_N-\Delta_N(x,y)-\frac{\theta}{\rho}S_N),
\end{aligned}$$

where $A,B\in\mathbb{R}^{5\times5}$ and $\xi_k\in\mathbb{R}^5$ are defined for $k\geq0$ as follows:

$$A\triangleq\begin{pmatrix}\frac{1}{\tau}&0&0&0&0\\0&\frac{1}{\sigma}&0&0&0\\0&0&0&0&\theta L_{yx}\\0&0&0&0&\theta L_{yy}\\0&0&\theta L_{yx}&\theta L_{yy}&-\alpha\end{pmatrix},\quad \xi_k\triangleq\begin{pmatrix}\|x_k-x\|\\\|y_k-y\|\\\|x_k-x_{k-1}\|\\\|y_k-y_{k-1}\|\\\|y_{k+1}-y_k\|\end{pmatrix},$$

$$B\triangleq\begin{pmatrix}\frac{1}{\tau}+\mu_x&0&0&0&0\\0&\frac{1}{\sigma}+\mu_y&-|1-\frac{\theta}{\rho}|\ L_{yx}&-|1-\frac{\theta}{\rho}|\ L_{yy}&0\\0&-|1-\frac{\theta}{\rho}|\ L_{yx}&\frac{1}{\tau}-L_{xx}&0&0\\0&-|1-\frac{\theta}{\rho}|\ L_{yy}&0&\frac{1}{\sigma}-\alpha&0\\0&0&0&0&0\end{pmatrix},$$

such that $x_{-1}=x_0$ and $y_{-1}=y_0$. Below, in Lemma A.5 we show that (2.1) is equivalent to $B-\frac{1}{\rho}A\succeq0$; therefore, it follows from (A.28) that

$$U_N(x,y)\leq\frac{1}{2}\xi_0^\top A\xi_0-\rho^{-N+1}(\frac{1}{2}\xi_N^\top B\xi_N-\Delta_N(x,y)-\frac{\theta}{\rho}S_N).$$

holds w.p. 1 for any given $(x,y)\in X\times Y$. Furthermore, for $(x,y)\in X\times Y$, we have

$$\begin{aligned}
&\frac{1}{2}\xi_N^\top B\xi_N-\frac{\theta}{\rho}S_N\\
&=\frac{1}{2\rho\tau}\|x_N-x\|^2+\frac{1}{2}\left(\frac{1}{\rho\sigma}-\frac{\alpha}{\rho}\right)\|y_N-y\|^2+\frac{1}{2}\xi_N^\top\begin{pmatrix}\frac{1}{\tau}(1-\frac{1}{\rho})+\mu_x&\mathbf{0}_{1\times3}&0\\\mathbf{0}_{3\times1}&G''&\mathbf{0}_{3\times1}\\0&\mathbf{0}_{1\times3}&0\end{pmatrix}\xi_N\\
&\geq\frac{1}{2\rho\tau}\|x_N-x\|^2+\frac{1}{2\rho\sigma}(1-\alpha\sigma)\|y_N-y\|^2=\Delta_N(x,y),
\end{aligned}\tag{A.29}$$

where the inequality follows from (2.1) and Lemma A.6, and G'' is defined in (A.14). Finally, $\frac{1}{2}\xi_0^\top A\xi_0=\frac{1}{2\tau}\|x-x_0\|^2+\frac{1}{2\sigma}\|y-y_0\|^2$. Now, we are ready to show (2.2).

Since $(x,y)\in X\times Y$ is fixed arbitrarily, all the results we have derived so far holds for any $(x,y)\in X\times Y$; thus, $U_N(x,y)\leq\frac{1}{2\tau}\|x-x_0\|^2+\frac{1}{2\sigma}\|y-y_0\|^2$ for any $(x,y)\in X\times Y$ w.p. 1. Therefore, $X\subset\mathcal{X}$ and $Y\subset\mathcal{Y}$ both being compact implies

$$\mathbb{E}[\sup_{(x,y)\in X\times Y}\{U_N(x,y)+\frac{\eta_x}{2}\|x-x_0\|^2+\frac{\eta_y}{2}(1+2\theta)\|y-y_0\|^2\}]\leq\Omega_{\tau,\sigma,\theta}.\tag{A.30}$$

Note from Assumption 1.2, for $k \geq -1$, we have

$$\mathbb{E}\left[\langle \Delta_k^{\Phi,x}, \hat{x}_{k+1} - \tilde{x}_k\rangle\right] = 0, \quad \mathbb{E}\left[\langle \Delta_k^{\Phi,y}, \hat{y}_{k+1} - \tilde{y}_k^{\pm}\rangle\right] = 0, \quad \mathbb{E}\left[\langle \Delta_{k-1}^{\Phi,y}, \hat{y}_{k+1} - \tilde{y}_{k-1}^{-}\rangle\right] = 0.$$

Thus, it follows that $\mathbb{E}[P_k] = 0$. Also, from Assumption 1.2, for $k \geq -1$, we have

$$\mathbb{E}\left[\|\Delta_k^{\Phi,x}\|^2\right] \leq \delta_x^2, \quad \mathbb{E}\left[\|\Delta_k^{\Phi,y}\|^2\right] \leq \delta_y^2.$$

Next, we uniformly upper bound $\mathbb{E}\left[Q_k\right]$ for $k \geq 0$ using Lemma A.4, i.e.,

$$\mathbb{E}[\sum_{k=0}^{N-1} \rho^{-k} Q_k] \leq \Big[\Big(\frac{\tau}{1+\tau\mu_x}\Xi^x_{\tau,\sigma,\theta} + \frac{1}{2\eta_x}\Big)\delta_x^2 + \Big(\frac{\sigma}{1+\sigma\mu_y}\Xi^y_{\tau,\sigma,\theta} + \frac{1+2\theta}{2\eta_y}\Big)\delta_y^2\Big] \sum_{k=0}^{N-1} \rho^{-k}.$$

Therefore, combining this result with $\mathbb{E}[P_k] = 0$ for any $k \in \{0, \ldots, N-1\}$, we get

$$\mathbb{E}[\sum_{k=0}^{N-1} \rho^{-k}(P_k + Q_k)] \leq K_N(\rho)\ \Xi_{\tau,\sigma,\theta}. \tag{A.31}$$

Therefore, the desired result in (2.2) follows from (A.26), (A.30) and (A.31).

Recall that until (A.30), we have not used compactness of X and Y; indeed, without using compactness we were able to show that $U(x,y) \leq \frac{1}{2\tau}\|x - x_0\|^2 + \frac{1}{2\sigma}\|y - y_0\|^2$ for any $(x,y) \in X \times Y$ w.p. 1. This result holds for arbitrary closed convex sets $X \subset \mathbf{dom}\, f$ and $Y \subset \mathbf{dom}\, g$ such that $x_0 \in X$ and $y_0 \in Y$. Thus, picking X and Y such that $x^* \in X$ and $y^* \in Y$, we trivially have $U_N(x^*, y^*) \leq \frac{1}{2\tau}\|x^* - x_0\|^2 + \frac{1}{2\sigma}\|y^* - y_0\|^2$ w.p. 1. Next, we will analyze (A.26) for $x = x^*$ and $y = y^*$ in the expected sense. It follows from (A.26) that

$$\mathbb{E}[K_N(\rho)\left(\mathcal{L}(\bar{x}_N, y^*) - \mathcal{L}(x^*, \bar{y}_N)\right) + \rho^{-N+1}\Delta_N(x^*, y^*)] \leq K_N(\rho)\ \Xi_{\tau,\sigma,\theta} + \Delta_{\tau,\sigma,\theta}.$$

Since $\mathcal{L}(\bar{x}_N, y^*) - \mathcal{L}(x^*, \bar{y}_N) \geq 0$, (2.3) immediately follows from above inequality. ∎

A.2. Proof of Lemma 2.5.

Proof. Given some $\epsilon > 0$ (which will be specified later), define $h : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y} \times \mathcal{X}$ such that $h_\epsilon(x,y) = (y, \epsilon\nabla_x\Phi^s(x,y))$. For any $y \in \mathcal{Y}$, $x^*(y)$ exists due to strong convexity and it satisfies $\nabla_x\Phi^s(x^*(y), y) = \mathbf{0}$ due to first-order optimality conditions. In addition, we have $\|\nabla_x\Phi^s(x,y) - \nabla_x\Phi^s(\bar{x},y)\| \geq \mu_x\|x - \bar{x}\|$ for all $x, \bar{x} \in \mathcal{X}$. Consequently, given $x, \bar{x} \in \mathcal{Y}$ and $y, \bar{y} \in \mathcal{Y}$, we have

$$\frac{\mu_x^2}{2}\|x - \bar{x}\|^2 \leq \|\nabla_x\Phi^s(x,y) - \nabla_x\Phi^s(\bar{x},\bar{y})\|^2 + \|\nabla_x\Phi^s(\bar{x},\bar{y}) - \nabla_x\Phi^s(\bar{x},y)\|^2. \tag{A.32}$$

Due to the Lipschitz continuity of $\nabla_x\Phi^s$, this implies $\frac{\mu_x^2}{2}\|x - \bar{x}\|^2 - (L^s_{xy})^2\|y - \bar{y}\|^2 \leq \|\nabla_x\Phi^s(x,y) - \nabla_x\Phi^s(\bar{x},\bar{y})\|^2$. Then, it follows that for any $\epsilon \in (0, \frac{1}{L^s_{xy}})$, we get

$$\frac{\epsilon^2\mu_x^2}{2}\|x - \bar{x}\|^2 + (1 - (\epsilon L^s_{xy})^2)\|y - \bar{y}\|^2 \leq \|h(x,y) - h(\bar{x},\bar{y})\|^2, \tag{A.33}$$

where we used the definition of h. This leads to $\|h(x,y) - h(\bar{x},\bar{y})\|/C_\epsilon \geq \|(x,y) - (\bar{x},\bar{y})\|$, where $C_\epsilon \triangleq \sqrt{\min\{\frac{\epsilon^2\mu_x^2}{2},\ 1 - (\epsilon L^s_{xy})^2\}} > 0$. Then, given any $y, \bar{y} \in \mathcal{Y}$, plugging in $x = x^*(y), \bar{x} = x^*(\bar{y})$; we can conclude that,

$$\|x^*(y) - x^*(\bar{y})\| \leq \frac{1}{C_\epsilon}\|h(x^*(y), y) - h(x^*(\bar{y}), \bar{y})\| = \frac{1}{C_\epsilon}\|y - \bar{y}\|.$$

where in the last equality we used the fact that $\nabla_x\Phi(x^*(y), y) = \mathbf{0}$ for all $y \in \mathcal{Y}$. Setting $\epsilon = \sqrt{2/\big(\mu_x^2 + 2(L^s_{xy})^2\big)}$ leads to the desired result. The result for Lipschitz continuity of $y^*(\cdot)$ follows from exactly the same argument. ∎

Appendix B. Extensions and Special Cases. In this section, we discuss the deterministic case, i.e., $\delta_x = \delta_y = 0$, and we also go over a special case of SAPD when $\theta = 0$, i.e., SGDA.

B.1. A Deterministic Primal-Dual Method (APD). When $\delta_x = \delta_y = 0$, i.e., $\tilde{\nabla}_x \Phi = \nabla_x \Phi$ and $\tilde{\nabla}_y \Phi = \nabla_y \Phi$, we call this deterministic variant of SAPD as APD. APD, when applied to (1.1) with a bilinear Φ, generates the same iterate sequence with [7] for a specific choice of step sizes; therefore, APD can be viewed as a general form of the method proposed by Chambolle and Pock [7] for bilinear SP problems. For bilinear problems as in [7], APD hits the lower complexity bound when $\mathcal{L}$ is strongly convex in x and strongly concave in y. Moreover, when Φ is not assumed to be bilinear, APD has the best iteration complexity bound shown for single-loop primal-dual first-order algorithm applied to (1.1) – we are not aware of any other single-loop method achieving this complexity bound for (1.1). The convergence guarantees for the deterministic scenario follows directly from Theorem 2.1 by setting $\delta_x = \delta_y = 0$.

COROLLARY B.1. *Suppose Assumption 1.1 hold, $\delta_x = \delta_y = 0$, and $\{x_k, y_k\}_{k\geq 0}$ be the iterates generated by APD, which is the deterministic version of Algorithm 1.1. The parameter $\tau, \sigma > 0$ and $\theta \geq 0$ satisfy (2.1) for some $\alpha \in [0, \frac{1}{\sigma}]$ and $\rho \in (0, 1]$. Then, for any (x, y) and $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$,*

$$\mathcal{L}(\bar{x}_N, y) - \mathcal{L}(x, \bar{y}_N) + \frac{\rho^{-N+1}}{K_N(\rho)} \Delta_N(x, y) \leq \frac{1}{K_N(\rho)} \Delta_0(x, y),$$

for all $N \geq 1$, where $\Delta_0(x, y) \triangleq \frac{1}{2\tau}\|x - x_0\|^2 + \frac{1}{2\sigma}\|y - y_0\|^2$, $\Delta_N(x, y)$ and $K_N(\rho)$ are defined in Theorem 2.1.

Corollary B.1 immediately follows from the proof for Theorem 2.1.

Remark B.2. This result extends the Accelerated Primal-Dual (APD) method proposed in [15] for MCMC and SCMC SP problems to cover the SCSC scenario as well. Indeed, the result for the MCMC case in [15] can be recovered from Corollary B.1 immediately by setting $\theta = \rho = 1$ and $\mu_x = \mu_y = 0$. Furthermore, the step sizes suggested in [15, Remark 2.3] satisfy (2.1) for a particular choice of $\alpha > 0$. Finally, since $K_N(1) = N$, APD achieves the sublinear rate of $\mathcal{O}(1/N)$ for the MCMC scenario.

In the rest, we consider SP problem in (1.1) under SCSC scenario. Let (x^*, y^*) denote the unique saddle point of (1.1). Next, we account for the individual effects of L_{xx}, L_{yx}, L_{yy} as well as μ_x, μ_y on the iteration complexity of APD. When Φ is bilinear, APD requires $\mathcal{O}\left(\sqrt{1 + \frac{L_{yx}^2}{\mu_x \mu_y}} \cdot \ln(1/\epsilon)\right)$ iteration to compute $(\bar{x}, \bar{y})$ such that $\mathcal{G}(\bar{x}, \bar{y}) \leq \epsilon$; this complexity is shown to be optimal in [41]. Moreover, for the general case, i.e., Φ may not be bilinear, the iteration complexity of APD is $\mathcal{O}\left(\left(\frac{L_{xx}}{\mu_x} + \frac{L_{yx}}{\sqrt{\mu_x \mu_y}} + \frac{L_{yy}}{\mu_y}\right) \cdot \ln(1/\epsilon)\right)$.

PROPOSITION B.3. *Suppose $\mu_x, \mu_y > 0$, and Assumption 1.1 hold. Let $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ denote the unique SP of (1.1). For any $\epsilon > 0$, and for any given $\beta \in (0, 1)$, suppose the APD parameters $\{\tau, \sigma, \theta\}$ are chosen such that*

$$\tau = \frac{1 - \theta}{\mu_x \theta}, \quad \sigma = \frac{1 - \theta}{\mu_y \theta}, \quad \theta = \bar{\theta} \tag{B.1}$$

where $\bar{\theta}$ is defined in (2.7). Then, the iteration complexity of APD to generate a point

$(\bar{x},\bar{y}) \in \mathcal{X}\times\mathcal{Y}$ such that $\mathcal{D}(\bar{x},\bar{y}) \le \epsilon$ is

$$\mathcal{O}\left(\left(1+\frac{L_{xx}}{\mu_x}+\frac{L_{yx}}{\sqrt{\mu_x\mu_y}}+\frac{L_{yy}}{\mu_y}\right)\cdot \ln(1/\epsilon)\right). \tag{B.2}$$

Moreover, when $\Phi(x,y)$ is a bilinear function, the iteration complexity of APD reduces to $\mathcal{O}\left(\left(1+\frac{L_{yx}}{\sqrt{\mu_x\mu_y}}\right)\cdot\ln(1/\epsilon)\right)$. Furthermore, assuming $\mathbf{dom}\, f\times \mathbf{dom}\, g$ is compact, APD can compute $(\bar{x},\bar{y})\in\mathcal{X}\times\mathcal{Y}$ such that $\mathcal{G}(\bar{x},\bar{y})\le\epsilon$ with the same iteration complexity stated above for both bilinear and general cases of Φ.

Proof. Using the particular parameters given in Corollary 2.8 within Corollary B.1, and following the similar arguments as in the proof of Theorem 2.9, we immediately get the result. When $\Phi(x,y)$ is bilinear, we only need to set $L_{xx}=L_{yy}=0$ in the general result to get the complexity for the bilinear case. ∎

According to [41], the complexity of APD for the bilinear case is optimal in terms of μ_x,μ_y,L_{yx} and ϵ. Furthermore, the complexity in (B.2) for the general case is the best we know for a single-loop first-order primal-dual algorithm.

B.2. Stochastic Gradient Descent Ascent Method (SGDA). The SGDA algorithm can be analyzed as a special case of SAPD with $\theta=0$. Our analysis leads to a wider range of admissible step sizes and establishes the iteration complexity bound for SGDA and shows its dependence on L_{xx},L_{yx},L_{yy} and μ_x,μ_y explicitly.

COROLLARY B.4. *Suppose Assumptions 1.1, 1.2 hold, and $\{x_k,y_k\}_{k\ge0}$ are generated by SAPD, stated in Algorithm 1.1, using parameters $\theta=0$ and $\tau,\sigma>0$ satisfying*

$$\begin{pmatrix} \frac{1}{\sigma}+\mu_y-\frac{1}{\rho\sigma} & -L_{yx} & -L_{yy}\\ -L_{yx} & \frac{1}{\tau}-L_{xx} & 0\\ -L_{yy} & 0 & \frac{1}{\sigma}\end{pmatrix}\succeq 0, \qquad \tau\mu_x\ge\frac{1-\rho}{\rho}, \tag{B.3}$$

for some $\rho\in(0,1)$. Then, for any compact set $X\times Y\subset\mathbf{dom}\, f\times\mathbf{dom}\, g$ such that $x_0\in X$ and $y_0\in Y$, and for any $\eta_x,\eta_y\ge0$, the following bound holds for $N\ge1$:

$$\mathbb{E}\left[\sup_{(x,y)\in X\times Y}\left\{\mathcal{L}\left(\bar{x}_N,y\right)-\mathcal{L}\left(x,\bar{y}_N\right)+\frac{\rho^{-N+1}}{K_N(\rho)}\Delta'_N(x,y)\right\}\right]\le\frac{1}{K_N(\rho)}\Omega'_{\tau,\sigma}+\Xi'_{\tau,\sigma},$$

where $\Delta'_N(x,y)\triangleq\frac{1}{2\rho\tau}\|x_N-x\|^2+\frac{1}{2\rho\sigma}\|y_N-y\|^2$, $\Omega'_{\tau,\sigma}\triangleq\frac{1}{2}(\frac{1}{\tau}+\eta_x)\Omega_X+\frac{1}{2}(\frac{1}{\sigma}+\eta_y)\Omega_Y$, and $\Xi'_{\tau,\sigma}\triangleq\left(\frac{\tau}{1+\tau\mu_x}+\frac{1}{\eta_x}\right)\delta_x^2+\left(\frac{\sigma}{1+\sigma\mu_y}+\frac{1}{\eta_y}\right)\delta_y^2$.

Proof. Setting θ and α to 0 in (2.1) immediately leads to the above result. ∎

Remark B.5. When $\mu_x=\mu_y=0$, unlike SAPD, SGDA does not have an admissible (τ,σ) pair with convergence guarantees. Indeed, from (B.3), $\mu_x=0$ implies that $\rho=1$ so that the second inequality is satisfied; furthermore, $\rho=1$ and $\mu_y=0$ imply that first diogonal element in the matrix inequality (MI) becomes 0; thus, there is no (τ,σ) such that the MI holds. It is worth emphasizing that (B.3) not having a solution when $\mu_x=\mu_y=0$ is not because our analysis is not tight enough; indeed, there are examples for which SGDA iterate sequence does not converge to a saddle point when $\mu_x=\mu_y=0$.

B.2.1. Parameter Choices for SGDA. We provide a particular solution to the matrix inequality (B.3) following a similar technique we used for deriving a particular parameter choice for SAPD. Next, in Lemma B.6, we provide an auxiliary system, simpler than (B.3), to construct the particular solution given in Corollary B.7.

LEMMA B.6. *Let $\tau, \sigma > 0$, $\rho \in (0,1)$, and $\pi_1, \pi_2 > 0$ satisfy*

$$\frac{1}{\tau} - L_{xx} - \pi_1 L_{yx} \geq 0, \tag{B.4a}$$

$$\frac{1}{\sigma} - \pi_2 L_{yy} \geq 0, \tag{B.4b}$$

$$\frac{1}{\sigma}(1 - \frac{1}{\rho}) + \mu_y \geq \frac{L_{yx}}{\pi_1} + \frac{L_{yy}}{\pi_2}, \tag{B.4c}$$

$$\tau \mu_x \geq \frac{1-\rho}{\rho}. \tag{B.4d}$$

Then $\{\tau, \sigma, \rho\}$ is a solution to (B.3).

Proof. We only need to verify that the matrix inequality in (B.3) holds. Permuting the rows and columns in (B.4c), it follows that

$$\begin{pmatrix} \frac{1}{\sigma}(1-\frac{1}{\rho}) + \mu_y & -L_{yx} & -L_{yy} \\ -L_{yx} & \frac{1}{\tau} - L_{xx} & 0 \\ -L_{yy} & 0 & \frac{1}{\sigma} \end{pmatrix} \succeq \begin{pmatrix} \frac{L_{yx}}{\pi_1} + \frac{L_{yy}}{\pi_2} & -L_{yx} & -L_{yy} \\ -L_{yx} & \frac{1}{\tau} - L_{xx} & 0 \\ -L_{yy} & 0 & \frac{1}{\sigma} \end{pmatrix} \triangleq \tilde{M}.$$

Note $\tilde{M} = \tilde{M}_1 + \tilde{M}_2$ for

$$\tilde{M}_1 \triangleq \begin{pmatrix} \frac{L_{yx}}{\pi_1} & -L_{yx} & 0 \\ -L_{yx} & \frac{1}{\tau} - L_{xx} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \tilde{M}_2 \triangleq \begin{pmatrix} \frac{L_{yy}}{\pi_2} & 0 & -L_{yy} \\ 0 & 0 & 0 \\ -L_{yy} & 0 & \frac{1}{\sigma} \end{pmatrix}.$$

The condition in (B.4a) and (B.4b) imply that

$$\tilde{M}_1 \succeq \begin{pmatrix} \frac{L_{yx}}{\pi_1} & -L_{yx} & 0 \\ -L_{yx} & \pi_1 L_{yx} & 0 \\ 0 & 0 & 0 \end{pmatrix} \succeq 0, \qquad \tilde{M}_2 \succeq \begin{pmatrix} \frac{L_{yy}}{\pi_2} & 0 & -L_{yy} \\ 0 & 0 & 0 \\ -L_{yy} & 0 & \pi_2 L_{yy} \end{pmatrix} \succeq 0,$$

respectively. Thus $\tilde{M}_1 + \tilde{M}_2 \succeq 0$, which completes the proof. ∎

Lemma B.6 helps us describe a subset of solutions to the matrix inequality system in (B.3) using the solutions of an inequality system in (B.4) that is easier to deal with. Next, based on based on Lemma B.6, we will construct a family of admissible parameters for SGDA, i.e., SAPD with $\theta = 0$, such that the iterate sequence will exhibit the desired convergence behavior.

COROLLARY B.7. *Suppose $\mu_x, \mu_y > 0$. For any $\beta_1, \beta_2 \in (0,1)$ such that $\beta_1 + \beta_2 < 1$, $\{\tau, \sigma, \rho\}$ chosen satisfying*

$$\tau = \frac{1-\rho}{\rho \mu_x}, \quad \sigma = \frac{1}{1-\beta_1-\beta_2} \cdot \frac{1-\rho}{\rho \mu_y}, \quad \rho \geq \bar{\rho} \triangleq \Big(1 + \frac{1}{L(\beta_1, \beta_2)}\Big)^{-1} \tag{B.5}$$

is a solution to (B.3), where $L(\beta_1, \beta_2) \triangleq \max\Big\{\frac{L_{xx}}{\mu_x} + \frac{1}{\beta_1} \cdot \frac{L_{yx}^2}{\mu_x \mu_y}, \frac{1}{\beta_2(1-\beta_1-\beta_2)} \cdot \frac{L_{yy}^2}{\mu_y^2}\Big\}$.

Proof. The proof is based on the result in Lemma B.6. Let $\beta_1, \beta_2, \beta_3 \in (0,1)$ such that $\beta_1 + \beta_2 < 1$. Given any $\rho \in (0,1)$, let $\tau = \frac{1-\rho}{\rho\mu_x}$, $\sigma = \frac{1}{\beta_3} \cdot \frac{1-\rho}{\rho\mu_y}$ and let $\pi_1 = \frac{L_{yx}}{\beta_1 \mu_y}$, $\pi_2 = \frac{L_{yy}}{\beta_2 \mu_y}$. If we substitute $(\tau, \sigma, \rho, \pi_1, \pi_2)$ into (B.4a)-(B.4c), we get

$$\mu_x \frac{\rho}{1-\rho} - L_{xx} - \frac{L_{yx}^2}{\beta_1 \mu_y} \geq 0, \quad \beta_3 \mu_y \frac{\rho}{1-\rho} - \frac{L_{yy}^2}{\beta_2 \mu_y} \geq 0, \tag{B.6a}$$

$$\beta_3\mu_y\frac{\rho}{1-\rho}(1-\frac{1}{\rho})+\mu_y\geq(\beta_1+\beta_2)\mu_y. \tag{B.6b}$$

Next, we solve this inequality system in terms of $\rho\in(0,1)$. Note (B.6a) holds for

$$\rho\geq\max\left\{\left(1+\frac{1}{\frac{L_{xx}}{\mu_x}+\frac{L_{yx}^2}{\beta_1\mu_x\mu_y}}\right)^{-1},\ \left(1+\frac{\beta_2\beta_3}{L_{yy}^2/\mu_y^2}\right)^{-1}\right\},$$

and (B.6b) holds whenever $\beta_1+\beta_2+\beta_3\leq 1$. To minimize the lower bound on ρ, the optimal choice for β_3 is $\beta_3=1-\beta_1-\beta_2>0$. Thus, $\{\tau,\sigma,\rho,\pi_1,\pi_2\}$ satisfying (B.5) is a solution to (B.4), which implies that $\{\tau,\sigma,\rho\}$ is a solution to (B.3). ∎

To determine the best certifiable convergence rate, i.e., the smallest ρ, one can optimize β_1 and β_2. Finally, using the above parameter choice, we establish the iteration complexity bound for SGDA in the next subsection.

B.2.2. Iteration Complexity Bound for SGDA. In this part, we study the iteration complexity bound for SGDA to generate a point $(\bar{x},\bar{y})\in\mathcal{X}\times\mathcal{Y}$ such that $\mathcal{G}(\bar{x},\bar{y})\leq\epsilon$. The proof technique is very similar to that for SAPD.

PROPOSITION B.8. *Suppose $\mu_x,\mu_y>0$, and Assumption 1.1 to 1.3 hold. For any $\epsilon>0$, and for any given $\beta_1,\beta_2\in(0,1)$ satisfying $\beta_1+\beta_2<1$, suppose the parameters $\{\tau,\sigma\}$ are chosen such that*

$$\tau=\frac{1-\rho}{\mu_x\rho},\quad \sigma=\frac{1}{1-\beta_1-\beta_2}\cdot\frac{1-\rho}{\mu_y\rho},\quad \rho=\max\{\overline{\rho},\overline{\overline{\rho}}\}, \tag{B.7}$$

where $\overline{\rho}$ is defined in (B.5) and $\overline{\overline{\rho}}\triangleq\max\{\overline{\overline{\rho}}_1,\overline{\overline{\rho}}_2\}$ such that

$$\overline{\overline{\rho}}_1=\max\left\{0,\ 1-\frac{\mu_x}{6\delta_x^2}\ \epsilon\right\},\quad \overline{\overline{\rho}}_2=\max\left\{0,\ \frac{6\delta_y^2-\epsilon\mu_y}{6\delta_y^2-\epsilon\mu_y(\beta_1+\beta_2)}\right\}$$

with the convention that $\overline{\overline{\rho}}_1=0$ if $\delta_x^2=0$ and $\overline{\overline{\rho}}_2=0$ if $\delta_y^2=0$. Then the iteration complexity of SGDA method, i.e., SAPD with $\theta=0$, as stated in Algorithm 1.1, to generate a point $(\bar{x},\bar{y})\in\mathcal{X}\times\mathcal{Y}$ such that $\mathcal{G}(\bar{x},\bar{y})\leq\epsilon$ is

$$\mathcal{O}\left(\left(\frac{L_{xx}}{\mu_x}+\frac{L_{yx}^2}{\mu_x\mu_y}+\frac{L_{yy}^2}{\mu_y^2}+\left(\frac{\delta_x^2}{\mu_x}+\frac{\delta_y^2}{\mu_y}\right)\frac{1}{\epsilon}\right)\cdot\ln\left(\frac{\mu_x\Omega_f+\mu_y\Omega_g}{\epsilon}+1\right)\right). \tag{B.8}$$

Proof. Given $\beta_1,\beta_2\in(0,1)$ such that $\beta_1+\beta_2<1$, let $\{\tau,\sigma,\rho\}$ be a particular solution to (B.3) chosen according to Corollary B.7, and we set $\eta_x=\frac{1}{\tau}+\mu_x$, $\eta_y=\frac{1}{\sigma}+\mu_y$. Using these particular parameter values and setting $X=\mathbf{dom}\,f$ and $Y=\mathbf{dom}\,g$ within Corollary B.4, it follows from Assumption 1.3 that

$$\begin{aligned}&\mathbb{E}[\sup_{(x,y)\in X\times Y}\{\mathcal{L}(\bar{x}_N,y)-\mathcal{L}(x,\bar{y}_N)+\frac{1-\rho}{1-\rho^N}\Delta_N'(x,y)\}]\\ \leq&\frac{1}{K_N(\rho)}\Big((\frac{1}{\tau}+\frac{\mu_x}{2})\Omega_f+(\frac{1}{\sigma}+\frac{\mu_y}{2})\Omega_g\Big)+\frac{2\tau}{1+\tau\mu_x}\delta_x^2+\frac{2\sigma}{1+\sigma\mu_y}\delta_y^2.\end{aligned} \tag{B.9}$$

For any $\epsilon>0$, the right side of (B.9) can be bounded by $\epsilon>0$ when

$$\frac{1}{K_N(\rho)}\Big((\frac{1}{\tau}+\frac{\mu_x}{2})\Omega_f+(\frac{1}{\sigma}+\frac{\mu_y}{2})\Omega_g\Big)\leq\frac{\epsilon}{3},\quad \frac{2\tau}{1+\tau\mu_x}\delta_x^2\leq\frac{\epsilon}{3},\quad \frac{2\sigma}{1+\sigma\mu_y}\delta_y^2\leq\frac{\epsilon}{3}. \tag{B.10}$$

Substituting τ and σ values given in (B.5) into the second and the third conditions in (B.10), we observe that these conditions will hold when $\rho\geq\overline{\overline{\rho}}$. Thus, combining this requirement with $\rho\geq\overline{\rho}$ in Corollary B.7, it follows that (B.7) provides a particular parameter choice for SGDA. By mimicking the steps in Theorem 2.9, one can easily show that the number of iterations for (B.10) to be satisfied is bounded by (B.8). ∎

Since we do not use a symmetric update rule, the effect of Lipschitz constants in the complexity bound are different, i.e., compare $\frac{L_{xx}}{\mu_x}$ with $\frac{L_{yy}^2}{\mu_y^2}$. Furthermore, we also observe that adopting a momentum term as in SAPD, i.e., $\theta > 0$, the $\mathcal{O}(1)$ constant improves from $\frac{L_{xx}}{\mu_x} + \frac{L_{yx}^2}{\mu_x\mu_y} + \frac{L_{yy}^2}{\mu_y^2}$ for SGDA to $\frac{L_{xx}}{\mu_x} + \frac{L_{yx}}{\sqrt{\mu_x\mu_y}} + \frac{L_{yy}}{\mu_y}$ for SAPD.

B.3. Robustness Analysis of SGDA. Since SGDA is a special case of SAPD with momentum parameter θ set to zero, using similar arguments as in subsection 3.1, we demonstrate the rate-robustness trade-off for SGDA and compare the Pareto curves of both methods. In the rest we illustrate the rate-robustness trade-off for (3.2) when $\mu_x = \mu_y = 1$, $\|K\|_2 = 10$, $d = 30$ and $\delta_x = \delta_y = \delta = 10$ as in subsection 3.1.

We numerically illustrate the fundamental rate-robustness trade-off in Figure 5(c) for (3.2) through plotting 4 curves: $\mathcal{J}_1$, $\mathcal{J}_2$, $\mathcal{J}^*$ and $\mathcal{J}^*_{SGDA}$. For $\mathcal{J}_1$, we uniformly grid the parameter space $(\tau, \sigma, \theta) \in [0, 0.2] \times [0, 0.2] \times [0, 2]$ using $200 \times 200 \times 200$ points, then for each grid point, we compute the corresponding $(\rho_{\text{true}}, \mathcal{J})$ value and plot it – recall that for SP problems of the form (3.2), we discussed how to compute $\mathcal{J}$ explicitly using symbolic math in subsection 3.1 and $\rho_{\text{true}} \triangleq \rho(A)^2$ where A is as in (3.6). For $\mathcal{J}_2$, we employ the step sizes suggested in [7, Algorithm 5] for the CP method and plot $(\rho_{\text{true}}, \mathcal{J})$ (see Appendix C.1 for more details). For $\mathcal{J}^*$, recall that $\mathcal{J}^*(\rho) = \min_{\tau,\sigma,\theta\geq 0}\{\mathcal{J} : \rho_{\text{true}} = \rho\}$, and we plot $(\rho, \mathcal{J}^*(\rho))$ which illustrates the best robustness that can be achieved for a given rate. We further define the robustness of SGDA as $\mathcal{J}^*_{SGDA}(\rho) \triangleq \min_{\tau,\sigma,\theta\geq 0}\{\mathcal{J} : \rho_{\text{true}} = \rho,\ \theta = 0\}$.

Note that the patterns of $\mathcal{J}^*$, $\mathcal{J}_1$ and $\mathcal{J}_2$ are similar to those we observed in Figure 1(a). When ρ is close to 1, $\mathcal{J}^*_{SGDA}$ and $\mathcal{J}^*(\rho)$ exhibit a similar behavior, and both curves lie strictly below the $\mathcal{J}_2$ curve corresponding to CP parameters. However, as ρ gets smaller, $\mathcal{J}^*_{SGDA}$ increases monotonically and gets larger than $\mathcal{J}_2$ in which case the robustness gets worse, i.e., corresponds to a very high $\mathcal{J}$ value. On the other hand, we observe that SAPD is able to provide the same convergence rate with a better robustness level (with a smaller $\mathcal{J}$ value) compared to SGDA which does not have a momentum term. As a consequence, we infer that using a momentum term leads to improvements in both the convergence rate and the robustness.

In Figure 5(a) and Figure 5(b), we further analyze how the choice of primal-dual step sizes, τ and σ, affect the convergence rate and the robustness level where we set x-axis as τ, and y-axis as σ, and we plot $\rho_{\text{true}}(\tau, \sigma)$ and $\mathcal{J}(\tau, \sigma)$, respectively. Note that given τ and σ, the corresponding ρ_{true} and $\mathcal{J}$ values are fixed because $\theta = 0$ for SGDA and there is no degree of freedom left. We observe from these two figures that simultaneously increasing τ and σ leads to a faster convergence at the expense of a decrease in robustness level. These results illustrate the fundamental trade-offs between the convergence rate versus robustness.

B.4. Further Discussions for SGDA. In this section, considering the SCSC setting, we compare the convergence rate we established for SGDA with those rate results in [11] and [24] established for the problem:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \hat{\Phi}(x, y) \tag{B.11}$$

under the following assumptions on the coupling function $\hat{\Phi}$.

Assumption B.9. *The function $\hat{\Phi}(x, y)$ is strongly convex and strongly concave in x and y with moduli $\hat{\mu}_x$ and $\hat{\mu}_y$, respectively. Furthermore, there exist positive*

FIG. 5. *The rate-robustness trade-off for* (3.2) *when* $\mu_x = \mu_y = 1$, $\|K\|_2 = 10$, $d = 30$ *and* $\delta_x = \delta_y = \delta = 10$. *The best rate is* 0.67 *for SAPD while the best rate for SGDA is* 0.7. *The point indicated with a red "$*$" in Figure* 5(c) *is the particular choice of CP parameters given in [7, Eq.(49)].*

constants $\{\hat{L}_{xx}, \hat{L}_{yy}, \hat{L}_{xy}, \hat{L}_{yx}\}$ *such that for all* $(x,y),(x',y') \in \mathcal{X}\times\mathcal{Y}$, *we have*

$$\|\nabla_x \hat{\Phi}(x,y) - \nabla_x \hat{\Phi}(x',y')\| \leq \hat{L}_{xx}\|x-x'\| + \hat{L}_{xy}\|y-y'\|,$$
$$\|\nabla_y \hat{\Phi}(x,y) - \nabla_y \hat{\Phi}(x',y')\| \leq \hat{L}_{yx}\|x-x'\| + \hat{L}_{xy}\|y-y'\|.$$

Under Assumption B.9, the analysis in [11] uses constant primal and dual step sizes equal to each other, $\hat{\alpha}_{\text{GDA}} > 0$; and it is shown that for all $N \geq 1$, the iterates $z_N = (x_N, y_N) \in \mathcal{X}\times\mathcal{Y}$ of the SGDA algorithm satisfy

$$\mathbb{E}\left[\|z_N - z^*\|^2\right] \leq \hat{C}_{\text{GDA}}\hat{\rho}^N_{\text{GDA}}\mathbb{E}\left[\|z_0 - z^*\|^2\right] + \hat{\varepsilon}_{\text{GDA}}, \tag{B.12}$$

where $(x^*, y^*) \in \mathcal{X}\times\mathcal{Y}$ is the unique saddle point for the SCSC SP problem in (B.11), $\hat{C}_{\text{GDA}}$ is the $\mathcal{O}(1)$ constant, $\hat{\rho}_{\text{GDA}} \in (0,1)$ denotes the convergence rate and $\hat{\varepsilon}_{\text{GDA}}$ is an error term due to the presence of noise in gradients – the constants $\hat{C}_{\text{GDA}}$, $\hat{\rho}_{\text{GDA}}$ and $\hat{\varepsilon}_{\text{GDA}}$ depend on the constant step size $\hat{\alpha}_{\text{GDA}}$ as well as on the parameters $\hat{L}_{xy}$, $\hat{L}_{yx}$, $\hat{L}_{xx}$, $\hat{L}_{yy}$ and $\hat{\mu}_x, \hat{\mu}_y$. More specifically, consider the worst-case constants

$$\hat{L} \triangleq \max\{\hat{L}_{xx}, \hat{L}_{yy}, \hat{L}_{xy}, \hat{L}_{yx}\}, \quad \hat{\mu} \triangleq \min\{\hat{\mu}_x, \hat{\mu}_y\}, \quad \hat{\kappa} \triangleq \frac{\hat{L}}{\hat{\mu}} \geq 1. \tag{B.13}$$

It is shown in [11] that SGDA with a constant step size $\hat{\alpha}_{GDA} \in (0, \frac{\mu}{4L^2}]$ satisfies (B.12) with $\hat{C}_{\text{GDA}} = 1$, $\hat{\rho}_{\text{GDA}} = \hat{\rho}_1 \triangleq 1 - \hat{\alpha}_{\text{GDA}}\mu$ and $\hat{\varepsilon}_{\text{GDA}} = 2\hat{\alpha}_{\text{GDA}}\sigma^2/\hat{\mu}$. In the

deterministic case (when $\sigma^2 = 0$), assuming $\hat{\Phi}$ is twice differentiable, it is shown in [21] that for GDA with a constant step size, (B.12) holds with $\hat{\varepsilon}_{\text{GDA}} = 0$ and $\hat{\rho}_{\text{GDA}} = \hat{\rho}_2 \triangleq 1 - \frac{\lambda_{\min}}{\lambda_{\max}}$ when the primal and dual step sizes are equal and set to $\frac{\sqrt{\lambda_{\min}}}{\lambda_{\max}}$, where $\lambda_{\min} \triangleq \inf\{\lambda_{\min}(H_1(x,y)) : (x,y) \in \mathcal{X} \times \mathcal{Y}\}$ with $H_1(x,y) \triangleq \operatorname{diag}(\nabla_{xx}\hat{\Phi}^2, \nabla_{yy}\hat{\Phi}^2)(x,y)$, and similarly $\lambda_{\max} \triangleq \sup\{\lambda_{\max}(H_2(x,y)) : (x,y) \in \mathcal{X} \times \mathcal{Y}\}$ with

$$H_2(x,y) \triangleq \begin{bmatrix} \nabla_{xx}\hat{\Phi}^2 + \nabla_{xy}\hat{\Phi}\nabla_{xy}\hat{\Phi}^\top & -\nabla_{xx}\hat{\Phi}\nabla_{xy}\hat{\Phi} - \nabla_{xy}\hat{\Phi}\nabla_{yy}\hat{\Phi} \\ -\nabla_{xy}\hat{\Phi}^\top\nabla_{xx}\hat{\Phi} - \nabla_{yy}\hat{\Phi}\nabla_{xy}\hat{\Phi}^\top & \nabla_{yy}\hat{\Phi}^2 + \nabla_{xy}\hat{\Phi}^\top\nabla_{xy}\hat{\Phi} \end{bmatrix}(x,y).$$

The constants $\lambda_{\min}, \lambda_{\max}$ are clearly related to the problem constants such as Lipschitz constants and moduli of strong convexity. However, computing $\lambda_{\min}$ and $\lambda_{\max}$ in practice exactly is not trivial, and one could argue that this task is in general not easier than solving the original problem (B.11). One exception would be $\hat{\Phi}(x,y)$ is quadratic function, in which case the matrices $H_1(x,y)$ and $H_2(x,y)$ are constant matrices.

We will next compare our rate guarantees for SGDA solving (B.11) with the existing ones in [11, 24]. Recall that our results on SGDA convergence guarantees require Assumption 1.1. On the other hand, (B.11) can be rewritten equivalently as

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \hat{\Phi}(x,y) = f(x) + \Phi(x,y) - g(y) \tag{B.14}$$

where $f(x) = \frac{\hat{\mu}_x}{2}\|x\|^2$, $g(y) = \frac{\hat{\mu}_y}{2}\|y\|^2$, and $\Phi(x,y) = \hat{\Phi}(x,y) - \frac{\hat{\mu}_x}{2}\|x\|^2 + \frac{\hat{\mu}_y}{2}\|y\|^2$. Therefore, under Assumption B.9, Assumption 1.1 hold for Φ with constants

$$L_{xx} = \hat{L}_{xx} - \hat{\mu}_x,\ L_{yy} = \hat{L}_{yy} - \hat{\mu}_y,\ L_{yx} = \hat{L}_{yx},\ L_{xy} = \hat{L}_{xy},\ \mu_x = \hat{\mu}_x,\ \mu_y = \hat{\mu}_y. \tag{B.15}$$

Note that if we multiply the inequality in Corollary B.4 by $\frac{K_N(\rho)}{\rho^{-N+1}}$ and substitute $(x,y) = (x^*, y^*)$, it follows that

$$\mathbb{E}\left[\Delta'_N(x^*, y^*)\right] \le \rho^{N-1}\Omega'_{\tau,\sigma} + \frac{1-\rho^N}{1-\rho} \cdot \Xi'_{\tau,\sigma},$$

where we used the fact that $\mathcal{L}(\bar{x}_N, y^*) - \mathcal{L}(x^*, \bar{y}_N) \ge 0$. Furthermore, defining $m_{\tau,\rho} \triangleq \max\{\tau, \sigma\}$ and multiplying the above inequality by $2\rho m_{\tau,\sigma}$, we obtain

$$\mathbb{E}\left[\|z^N - z^*\|^2\right] \le 2m_{\tau,\sigma}\Omega'_{\tau,\sigma}\rho^N + \frac{2m_{\tau,\sigma}\rho}{1-\rho} \cdot \Xi'_{\tau,\sigma},$$

for $\rho \in (0,1)$ where $z_N = (x_N, y_N)$ denotes the SGDA iterate. Therefore, our analysis implies that (B.12) holds with $\hat{C}_{\text{GDA}} = 2m_{\tau,\sigma}\Omega'_{\tau,\sigma}$, $\hat{\rho}_{\text{GDA}} = \rho$ and $\hat{\varepsilon}_{\text{GDA}} = \frac{2m_{\tau,\sigma}\rho}{1-\rho} \cdot \Xi'_{\tau,\sigma}$.

In the rest of this section, we will numerically illustrate that our analysis recovers the existing results or leads to an improvement upon those results in certain regimes for the parameters $\hat{L}_{xx}, \hat{L}_{yy}, \hat{L}_{xy}$ and $\hat{L}_{yx}$. For this purpose, we consider a quadratic function over $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{Y} = \mathbb{R}^2$: for any $x = (x_1, x_2)$ and $y = (y_1, y_2)$, define

$$\hat{\Phi}(x,y) = \frac{1}{2}x_1^2 + \frac{\gamma_x}{2}x_2^2 + \langle Kx, y\rangle - \frac{1}{2}y_1^2 - \frac{\gamma_y}{2}y_2^2, \tag{B.16}$$

where $\gamma_x, \gamma_y \ge 1$ and $K \in \mathbb{R}^{2\times 2}$ such that $\|K\|_2 = \gamma_{yx}$ with $\gamma_{yx} > 0$. Then, the Lipschitz constants and strong convexity moduli in Assumption B.9 are $\hat{L}_{xx} = \gamma_x$, $\hat{L}_{yy} = \gamma_y$, $\hat{L}_{xy} = \hat{L}_{yx} = \gamma_{yx}$, and $\hat{\mu}_x = \hat{\mu}_y = 1$. Thus, according to (B.15), we get

$$L_{xx} = \gamma_x - 1, \quad L_{yy} = \gamma_y - 1, \quad L_{yx} = \gamma_{yx}, \quad \mu_x = \mu_y = 1. \tag{B.17}$$

Next, we investigate how $\hat{\rho}_{\text{GDA}}$ derived in [11, 21] and in this paper, i.e., $\hat{\rho}_1$, $\hat{\rho}_2$ and $\hat{\rho}_{\text{our}}$, respectively, change as the condition number of $\hat{\Phi}$ in (B.16), i.e., $\hat{\kappa} \ge 1$, increases.

(a) $\hat{L}_{xx} > \hat{L}_{yx} > \hat{L}_{yy}$. (b) $\hat{L}_{xx} > \hat{L}_{yy} > \hat{L}_{yx}$. (c) $\hat{L}_{yx} > \hat{L}_{xx} > \hat{L}_{yy}$.

FIG. 6. *The convergence comparison of SGDA*

Given the problem parameters, using bisection on ρ within (B.3), we can compute the best SGDA rate we can certify through Corollary B.4, i.e., $\hat{\rho}_{\text{our}}$. More precisely, we can substitute the constants in (B.17) into (B.3) and find the smallest $\hat{\rho}_{\text{our}} \in (0,1)$ satisfying the matrix inequality (B.3) using bisection, where each bisection step requires solving a 2-dimensional SDP obtained via the change of variables $t = 1/\tau$ and $s = 1/\sigma$. Now, we compare our rate $\hat{\rho}_{\text{our}}$ with $\hat{\rho}_1$ from [11] and $\hat{\rho}_2$ from [21]. Note that although $\hat{\rho}_1$ and $\hat{\rho}_{\text{our}}$ are for the stochastic GDA algorithm, here they represent the rate for the bias term, which is denoted by $\hat{\rho}_{\text{GDA}}$ in (B.12). Therefore, $\hat{\rho}_{our}$, $\hat{\rho}_1$ and $\hat{\rho}_2$ are directly comparable. Furthermore, as mentioned above, due to the quadratic structure in (B.16), $\hat{\rho}_2$ is easy to compute.

Since the effects of $\hat{L}_{xx}$ and $\hat{L}_{yy}$ on the rate are symmetric, we consider three cases for any given $\hat{\kappa} \in [4, 100]$: *Case i.* $(\hat{L}_{xx}, \hat{L}_{yy}, \hat{L}_{yx}) = (\hat{\kappa}, \frac{\hat{\kappa}}{2}, \frac{\hat{\kappa}}{4})$, *Case ii.* $(\hat{L}_{xx}, \hat{L}_{yy}, \hat{L}_{yx}) = (\hat{\kappa}, \frac{\hat{\kappa}}{4}, \frac{\hat{\kappa}}{2})$, and *Case iii.* $(\hat{L}_{xx}, \hat{L}_{yy}, \hat{L}_{yx}) = (\frac{\hat{\kappa}}{2}, \frac{\hat{\kappa}}{4}, \hat{\kappa})$. In Figure 6(a) and Figure 6(b), $\hat{\rho}_{\text{our}}$ is always strictly less than both $\hat{\rho}_1$ and $\hat{\rho}_2$. In Figure 6(c), although $\hat{\rho}_{our}$ and $\hat{\rho}_2$ exhibit a similar behavior as in Figure 6(a), $\hat{\rho}_{\text{our}}$ is not strictly better than $\hat{\rho}_2$ for this scenario. We conclude that, our analysis can improve upon $\hat{\rho}_1$ and $\hat{\rho}_2$ for certain scenarios. For more general SP problems beyond the bilinear setting, computing the step size leading to $\hat{\rho}_2$ can be expensive, especially in high dimensions –as it requires estimating the smallest and largest eigenvalues of the matrices $H_1(x,y)$ and $H_2(x,y)$ over the domain; however, SGDA paramaters leading to $\hat{\rho}_{\text{our}}$ can be computed efficiently with our line search techniques even in high dimensions.

Appendix C. Supporting Results for the Robustness Analysis. In this section, we provide some details about our robustness analysis.

C.1. CP parameters. Consider (1.1) with Φ, f and g defined as in (3.2). Using the notations in our paper, the step size condition in [7, Algorithm 5] can be summarized as

$$1 + \mu_x \tau = 1 + \mu_y \sigma = \frac{1}{\theta}, \qquad \frac{1}{\tau} \geq \theta L_{yx}^2 \sigma. \tag{C.1}$$

In fact, the above condition is a quadratic inequality of θ, which is $\frac{L_{yx}^2}{\mu_y \mu_y}(1-\theta)^2 - \theta \leq 0$; thus, $\theta \in [1 + \frac{\mu_x \mu_y}{2L_{yx}^2} - \sqrt{(1 + \frac{\mu_x \mu_y}{2L_{yx}^2})^2 - 1}, 1]$. In Figure 1(a), we compute and plot the $(\rho_{\text{true}}, \mathcal{J})$ for all possible (τ, σ, θ) satisfying (C.1). Moreover, condition (C.1) holds with equality at the point indicated with " $*$ " in red color.

C.2. Convergence of the Gap Function Bias Term for (3.2). Consider (1.1) for Φ, f and g as defined in (3.2). We will show $\mathcal{G}(x_k, y_k)$ and $\mathcal{D}(x_k, y_k)$ converge

with the same rate, where $\mathcal{G}$ and $\mathcal{D}$ are defined in (1.2). First, we can compute $\mathcal{G}(x_k, y_k)$ explicitly, i.e.,

$$\mathcal{G}(x_k, y_k) = \mathbb{E}\left[\frac{\mu_x}{2}\|x_k\|^2 + \frac{1}{2\mu_y}\|Kx_k\|^2 + \frac{\mu_y}{2}\|y_k\|^2 + \frac{1}{2\mu_x}\|K^\top y_k\|^2\right]. \tag{C.2}$$

Recall the augmented vector $\tilde{z}_k = [x_k; y_k; x_{k-1}; y_{k-1}]$ obtained by vertical concatenation for all $k \geq 0$ such that $x_{-1} = x_0$ and $y_{-1} = y_0$, and $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$ is a given initial point. Let P_x and P_y be matrices with appropriate dimensions such that $x_k = P_x\tilde{z}_k$ and $y_k = P_y\tilde{z}_k$. Note (3.6) implies $\tilde{z}_k = A^k\tilde{z}_0 + \sum_{i=1}^k A^{i-1}B\omega_{k-i}$; thus, $\|x_k\|^2 = \|P_x\tilde{z}_k\|^2 = \|P_xA^k\tilde{z}_0 + \sum_{i=1}^k P_xA^{i-1}B\omega_{k-i}\|^2$. The noise model assumed in (3.3) and (3.4) implies that

$$\mathbb{E}\left[\|x_k\|^2\right] = \|P_xA^k\tilde{z}_0\|^2 + \delta^2\sum_{i=1}^k \mathbf{Tr}((P_xA^{i-1}B)^\top P_xA^{i-1}B) = \|P_xA^k\tilde{z}_0\|^2 + \delta^2\sum_{i=1}^k\|P_xA^{i-1}B\|_F^2.$$

We can also write the other terms in (C.2) using the same argument as above:

$$\mathbb{E}\left[\|Kx_k\|^2\right] = \|KP_xA^k\tilde{z}_0\|^2 + \delta^2\sum_{i=1}^k\|KP_xA^{i-1}B\|_F^2,$$

$$\mathbb{E}\left[\|y_k\|^2\right] = \|P_yA^k\tilde{z}_0\|^2 + \delta^2\sum_{i=1}^k\|P_yA^{i-1}B\|_F^2,$$

$$\mathbb{E}\left[\|K^\top y_k\|^2\right] = \|K^\top P_yA^k\tilde{z}_0\|^2 + \delta^2\sum_{i=1}^k\|K^\top P_yA^{i-1}B\|_F^2.$$

Therefore,

$$\begin{aligned}
&\mathcal{G}(x_k, y_k)\\
=&\frac{\mu_x}{2}\|P_xA^k\tilde{z}_0\|^2 + \frac{\mu_y}{2}\|P_yA^k\tilde{z}_0\|^2 + \frac{1}{2\mu_y}\|KP_xA^k\tilde{z}_0\|^2 + \frac{1}{2\mu_x}\|K^\top P_yA^k\tilde{z}_0\|^2\\
&+\delta^2\sum_{i=1}^k\left(\frac{\mu_x}{2}\|P_xA^{i-1}B\|_F^2 + \frac{\mu_y}{2}\|P_yA^{i-1}B\|_F^2 + \frac{1}{2\mu_y}\|KP_xA^{i-1}B\|_F^2 + \frac{1}{2\mu_x}\|K^\top P_yA^{i-1}B\|_F^2\right).
\end{aligned}$$

The matrix A is non-symmetric in general. By considering the Jordan decomposition of the $4d \times 4d$ matrix A, it is known that there exists a positive constant c_1 and a non-negative integer $0 \leq m_1 < 4d$ such that $\|A^k\| \leq c_1k^{m_1}\rho(A)^k$, for all $k \geq 1$ (see e.g. [13, 33]). Therefore, the bias term of $\mathcal{G}(x_k, y_k)$ is bounded by $c_2k^{2m_1}\rho(A)^{2k}$ for some positive constant c_2. Thus, we conclude that the bias diminishes exponentially with rate $\rho(A)^2$.

C.3. $\mathcal{C}_\rho$ is a connected set. Given ρ, we next show that the set $\mathcal{C}_\rho$ is connected. This result allows us to use Lemma 3.6 for optimizing the robustness.

LEMMA C.1. *For any $\rho \in [\rho^*, 1)$, $\mathcal{C}_\rho \subset (0, 1)$ is a non-empty convex set where $\mathcal{C}_\rho$ is defined by (3.4). Hence, it is connected.*

Proof. Since $\mathcal{C}_\rho \neq \emptyset$, let $c_1, c_2 \in \mathcal{C}_\rho$. Without loss of generality, suppose $c_2 \geq c_1$. We aim to show that for any $\beta \in [0, 1]$, we have $c^* \triangleq \beta c_1 + (1-\beta)c_2 \in \mathcal{C}_\rho$. Since $c_1, c_2 \in \mathcal{C}_\rho$, there exist $t_i, s_i, \theta_i > 0$ such that $G_\rho(t_i, s_i, \theta_i, c_is_i) \succeq \mathbf{0}$ for $i = 1, 2$. For a given $\lambda \in [0, 1]$, let $(t^*, s^*, \theta^*) \triangleq \lambda(t_1, s_1, \theta_1) + (1-\lambda)(t_2, s_2, \theta_2)$. It suffices to construct $\lambda \in [0, 1]$ such that $G_\rho(t^*, s^*, \theta^*, c^*s^*) \succeq \mathbf{0}$. This will show that $c^* \in \mathcal{C}_\rho$.

It follows from the definition of G_ρ that

$$
\begin{aligned}
&G_\rho(t^*, s^*, \theta^*, c^* s^*) = \\
&\lambda \begin{pmatrix} (1-\frac{1}{\rho})t_1 + \mu_x & 0 & 0 & 0 & 0 \\ 0 & (1-\frac{1}{\rho})s_1 + \mu_y & (\frac{\theta_1}{\rho}-1)L_{yx} & (\frac{\theta_1}{\rho}-1)L_{yy} & 0 \\ 0 & (\frac{\theta_1}{\rho}-1)L_{yx} & t_1 - L_{xx} & 0 & -\frac{\theta_1}{\rho}L_{yx} \\ 0 & (\frac{\theta_1}{\rho}-1)L_{yy} & 0 & (1-c_1)s_1 & -\frac{\theta_1}{\rho}L_{yy} \\ 0 & 0 & -\frac{\theta_1}{\rho}L_{yx} & -\frac{\theta_1}{\rho}L_{yy} & \frac{c_1 s_1}{\rho} \end{pmatrix} \\
&+ \lambda \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & (c_1 - c^*)s_1 & 0 \\ 0 & 0 & 0 & 0 & \frac{(c^*-c_1)s_1}{\rho} \end{pmatrix} \\
&+ (1-\lambda) \begin{pmatrix} (1-\frac{1}{\rho})t_2 + \mu_x & 0 & 0 & 0 & 0 \\ 0 & (1-\frac{1}{\rho})s_2 + \mu_y & (\frac{\theta_2}{\rho}-1)L_{yx} & (\frac{\theta_2}{\rho}-1)L_{yy} & 0 \\ 0 & (\frac{\theta_2}{\rho}-1)L_{yx} & t_2 - L_{xx} & 0 & -\frac{\theta_2}{\rho}L_{yx} \\ 0 & (\frac{\theta_2}{\rho}-1)L_{yy} & 0 & (1-c_2)s_2 & -\frac{\theta_2}{\rho}L_{yy} \\ 0 & 0 & -\frac{\theta_2}{\rho}L_{yx} & -\frac{\theta_2}{\rho}L_{yy} & \frac{c_2 s_2}{\rho} \end{pmatrix} \\
&+ (1-\lambda) \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & (c_2 - c^*)s_2 & 0 \\ 0 & 0 & 0 & 0 & \frac{(c^*-c_2)s_2}{\rho} \end{pmatrix} \\
&= \lambda G_\rho(t_1, s_1, \theta_1, c_1 s_1) + (1-\lambda) G_\rho(t_2, s_2, \theta_2, c_2 s_2) \\
&+ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda(c_1 - c^*)s_1 + (1-\lambda)(c_2 - c^*)s_2 & 0 \\ 0 & 0 & 0 & 0 & \frac{\lambda(c^*-c_1)s_1 + (1-\lambda)(c^*-c_2)s_2}{\rho} \end{pmatrix}
\end{aligned}
$$

Since $c_i \in \mathcal{C}_\rho$ implies $G_\rho(t_i, s_i, \theta_i, c_i s_i) \succeq \mathbf{0}$ for $i = 1, 2$, we have $G_\rho(t^*, s^*, \theta^*, c^* s^*) \succeq 0$ if

$$
\begin{aligned}
&\lambda(c_1 - \beta c_1 - (1-\beta)c_2)s_1 + (1-\lambda)(c_2 - \beta c_1 - (1-\beta)c_2)s_2 \geq 0 \\
&\lambda(\beta c_1 + (1-\beta)c_2 - c_1)s_1 + (1-\lambda)(\beta c_1 + (1-\beta)c_2 - c_2)s_2 \geq 0.
\end{aligned}
$$

Therefore, to show the desired result, it is sufficient to find $\lambda \in [0, 1]$ such that

$$
\begin{aligned}
&\lambda(1-\beta)(c_1 - c_2)s_1 + (1-\lambda)\beta(c_2 - c_1)s_2 \geq 0 \\
&\lambda(1-\beta)(c_2 - c_1)s_1 + (1-\lambda)\beta(c_1 - c_2)s_2 \geq 0.
\end{aligned}
$$

This system is equivalent to $\lambda(1-\beta)s_1 - (1-\lambda)\beta s_2 = 0$, which yields $\lambda = \frac{\beta s_2}{(1-\beta)s_1 + \beta s_2}$ and this completes the proof. ∎

Appendix D. Euclidean projection onto the Intersection of the Simplex and the f-divergence Ball. In this section, we show an efficient method to solve the proximal problems $\min_{y \in \mathcal{P}_r} \frac{\mu_y}{2}\|y\|^2 + \dfrac{1}{2\sigma}\|y - (y_k + \sigma \tilde{s}_k)\|^2$ arising when SAPD is applied to (4.1). In the rest, we consider a generic form of this problem. Indeed, given some $\bar{p} \in \mathbb{R}^n$ and $R > 0$, we aim to solve

$$
p^* \triangleq \underset{p \in \mathcal{P}}{\mathbf{argmin}} \|p - \bar{p}\|^2, \quad \text{where} \quad \mathcal{P} \triangleq \{p \in \mathbb{R}^n_+ : \ \mathbf{1}^\top p = 1, \ \|p - \mathbf{1}/n\|^2 \leq R^2\}. \tag{D.1}
$$

Next, we construct an equivalent problem to (D.1), mainly because computing a dual optimal solution for the new formulation would be easier. Let $\mathcal{S} \triangleq \{p \in \mathbb{R}^n_+ : \ \mathbf{1}^\top p =$

1}. For $p \in \mathcal{P} \subset \mathcal{S}$, we have $\|p - \mathbf{1}/n\|^2 = \|p\|^2 - 1/n$ since $\mathbf{1}^\top p = 1$. Therefore, (D.1) is equivalent to

$$p^* = \underset{p\in\mathcal{S}}{\mathbf{argmin}}\{\frac{1}{2}\|p-\bar{p}\|^2:\ \|p\|^2 \leq R^2 + \frac{1}{n} \triangleq \bar{R}^2\} \tag{D.2}$$

In the literature, many efficient methods are provided to compute the Euclidean projection of a given point onto a unit simplex, e.g., see [9]. Therefore, we assume that $\|p^s\| > \bar{R}$, where $p^s \triangleq \mathbf{argmin}\{\|p-\bar{p}\|^2:\ p \in \mathcal{S}\}$; otherwise, i.e., $\|p^s\| \leq \bar{R}$, we trivially have $p^* = p^s$; thus, p^* can be efficiently computed with one of these simplex projection methods from the literature. Since we assume that $\|p^s\| > \bar{R}$, p^* must satisfy $\|p^*\| = \bar{R}$. The Lagrangian function for the problem in (D.2) can be written as

$$\mathcal{L}(p,\lambda) \triangleq \mathbb{1}_{\mathcal{S}}(p) + \frac{1}{2}\|p-\bar{p}\|^2 + \frac{\lambda}{2}\Big(\|p\|^2 - \bar{R}^2\Big), \tag{D.3}$$

where $\mathbb{1}_{\mathcal{S}}(\cdot)$ denotes the indicator function of $\mathcal{S}$.[8]

$$p^*(\lambda) \triangleq \underset{p\in\mathbb{R}^n}{\mathbf{argmin}}\, \mathcal{L}(p,\lambda) = \underset{p\in\mathcal{S}}{\mathbf{argmin}}\|p - \frac{\bar{p}}{1+\lambda}\|^2. \tag{D.4}$$

The aim is to compute $\lambda^* \geq 0$ such that $\|p^*(\lambda^*)\| = \bar{R}$, considering that $(p^*(\lambda^*), \lambda^*)$ is a KKT point; thus, $p^* = p^*(\lambda^*)$. It is essential to observe three critical points: *i)* $p^*(0) = p^s$, which implies $\|p^*(0)\| > \bar{R}$; *ii)* $\|p^*(\lambda_1)\| \leq \|p^*(\lambda_2)\|$ for all $\lambda_2 \geq \lambda_1 \geq 0$; *iii)* $p^*(\lambda) \to \mathbf{1}/n$ as $\lambda \nearrow \infty$, which also implies that $\|p^*(\lambda)\| < \bar{R}$ for sufficiently large $\lambda > 0$ since $\mathbf{1}/n \in \mathcal{P}$. These observations show that we can start from $\lambda = 0$ and keep gradually increasing it until the first time $\|p^*(\lambda)\| = \bar{R}$.

For numerical stability, i.e., for avoiding $\lambda \to \infty$, instead of (D.4), we will consider an equivalent problem: for $\gamma \in (0,1]$,

$$p^*_\gamma \triangleq \underset{p\in\mathcal{S}}{\mathbf{argmin}}\frac{1}{2}\|p-\gamma\bar{p}\|^2. \tag{D.5}$$

Our aim is to compute $\bar{\gamma} \in (0,1)$ such that $\|p^*_{\bar{\gamma}}\| = \bar{R}$. Let $u \in \mathbb{R}^n$ consist of elements of $\bar{p}$ in the descending order, i.e., $u_1 \geq u_2 \geq ... \geq u_n$. Define $K_\gamma \triangleq \max_{k\in[n]}\{k : (\sum_{i=1}^k \gamma u_i - 1)/k < \gamma u_k\}$ and $q_\gamma \triangleq (\sum_{i=1}^{K_\gamma} \gamma u_i - 1)/K_\gamma$, where $[n] \triangleq \{1,\ldots,n\}$. Note that $K_\gamma \geq 1$ is well-defined since $\gamma u_1 - 1 < \gamma u_1$. From [9, Algorithm 1], we know that

$$p^*_\gamma = \begin{cases} \gamma u_i - q_\gamma & i = 1,2,...,K_\gamma, \\ 0 & \text{otherwise.} \end{cases} \tag{D.6}$$

It follows from (D.6) that the equation $\|p^*_{\bar{\gamma}}\| = \bar{R}$ has a unique positive solution,

$$\bar{\gamma} = \sqrt{\frac{\bar{R}^2 - 1/K_{\bar{\gamma}}}{\sum_{i=1}^{K_{\bar{\gamma}}} u_i^2 + (\sum_{i=1}^{K_{\bar{\gamma}}} u_i)^2/K_{\bar{\gamma}}}}. \tag{D.7}$$

Since $K_{\bar{\gamma}}$ depends on $\bar{\gamma} \in (0,1)$, we cannot solve (D.7) for $\bar{\gamma}$ immediately.

At this point, it is essential to observe that for any $\gamma \in (0,1)$, the definitions of K_γ and q_γ imply that $K_\gamma = k \in [n]$ if and only if

$$\gamma u_{k+1} \leq q_\gamma < \gamma u_k, \tag{D.8}$$

where we define $u_{n+1} \triangleq 0$. Since $K_{\bar{\gamma}} \in [n]$, we can set $K_{\bar{\gamma}} = k$ for $k = 1,2,...,n$ and check whether $K_{\bar{\gamma}}$ satisfies the condition in (D.8) for $\bar{\gamma}$ computed by (D.7). Then substituting such $K_{\bar{\gamma}}$ into (D.6) yields the solution p^*.

[8] The indicator function $\mathbb{1}_{\mathcal{S}}(\cdot)$ is defined as $\mathbb{1}_{\mathcal{S}}(x) = 0$ if $x \in \mathcal{S}$, $\mathbb{1}_{\mathcal{S}}(x) = +\infty$ otherwise.