Optimal Algorithms for Stochastic Three-Composite Convex-Concave Saddle Point Problems

Renbo Zhao
Operations Research Center, Massachusetts Institute of Technology, renboz@mit.edu

We develop stochastic first-order primal-dual algorithms to solve a class of stochastic three-composite convex-concave saddle point problems, which subsumes many previously studied problems as special cases. To obtain an $\epsilon$-expected duality gap, when the saddle function is non-strongly convex in the primal variable, we design an algorithm based on the primal-dual hybrid gradient framework, that achieves the state-of-the-art oracle complexity. By using this algorithm as subroutine, when the saddle function is strongly convex in the primal variable and the gradient noises follow the sub-Gaussian distribution, we develop a novel stochastic restart scheme, whose oracle complexity is strictly better than any of the existing ones, even in the deterministic case. Moreover, for each problem parameter of interest, whenever the lower complexity bound exists in the literature, the complexity obtained by our algorithm is optimal in the non-strongly convex regime and nearly optimal (up to log-factor) in the strongly convex regime.

1. Introduction.
Let $\mathcal{X}$ and $\mathcal{Y}$ be two finite-dimensional real Banach spaces with dual spaces $\mathcal{X}^*$ and $\mathcal{Y}^*$ respectively. Consider the following saddle-point problem (SPP)
\[
\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} [S(x, y) \triangleq f(x) + g(x) + \Phi(x, y) - J(y)],
\]
where $\mathcal{X} \subseteq \mathcal{X}$ and $\mathcal{Y} \subseteq \mathcal{Y}$ are nonempty, closed and convex sets, and the functions $f : \mathcal{X} \to \mathbb{R} \triangleq (-\infty, +\infty)\bigcup\{\infty\}$, $g : \mathcal{X} \to \mathbb{R}$ and $J : \mathcal{Y} \to \mathbb{R}$ are convex, closed and proper (CCP). In addition, the function $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is convex-concave, i.e., $\Phi(\cdot, y)$ is convex for any $y \in \mathcal{Y}$ and $\Phi(x, \cdot)$ is concave for any $x \in \mathcal{X}$. We assume that $f$, $g$, $J$ and $\Phi$ satisfy the following regularity conditions:

- $f$ is differentiable$^1$ on $\mathcal{X}' \supseteq \mathcal{X}'$, where $\mathcal{X}'$ is an open set in $\mathcal{X}$, and its gradient $\nabla f : \mathcal{X} \to \mathcal{X}^*$ is $L$-Lipschitz on $\mathcal{X}'$ (where $L \geq 0$), i.e.,
\[
\|\nabla f(x) - \nabla f(x')\|_{\mathcal{X}^*} \leq L\|x - x'\|_{\mathcal{X}}, \quad \forall x, x' \in \mathcal{X}',
\]
where $\|\cdot\|_{\mathcal{X}^*}$ and $\|\cdot\|_{\mathcal{X}}$ denote the norms on $\mathcal{X}^*$ and $\mathcal{X}$ respectively.

- $f$ is $\mu$-strongly convex (s.c.) on $\mathcal{X}'$ (where $\mu \geq 0$), i.e., for any $x, x' \in \mathcal{X}'$,\[
f(x) \geq f(x') + \langle \nabla f(x'), x - x' \rangle + \frac{\mu}{2}\|x - x'\|^2.
\]

In this work, we will consider both cases where $\mu = 0$ and $\mu > 0$.

$^1$Throughout this work, differentiability is in the sense of Fréchet.
• $g$ and $J$ admit tractable Bregman proximal projections on $\mathcal{X}$ and $\mathcal{Y}$ respectively (see Section 2 for details).

• $\Phi$ is differentiable on $\mathcal{X}' \times \mathcal{Y}'$, where $\mathcal{Y}' \supseteq \mathcal{Y}$ is an open set in $\mathcal{Y}$. For any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, denote the gradient of $\Phi(\cdot, y)$ and $\Phi(x, \cdot)$ by $x \mapsto \nabla_x \Phi(x, y)$ and $y \mapsto \nabla_y \Phi(x, y)$ respectively. For all $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$, we assume that there exist constants $L_{xx}, L_{xy}, L_{yx}, L_{yy} \geq 0$ such that

\[
\begin{align*}
\|\nabla_x \Phi(x, y) - \nabla_x \Phi(x', y)\|_{\mathcal{X}'} &\leq L_{xx} \|x - x'\|_X, \\
\|\nabla_x \Phi(x, y) - \nabla_x \Phi(x', y')\|_{\mathcal{X}'} &\leq L_{xy} \|y - y'\|_Y, \\
\|\nabla_y \Phi(x, y) - \nabla_y \Phi(x', y')\|_{\mathcal{Y}'} &\leq L_{yx} \|x - x'\|_X, \\
\|\nabla_y \Phi(x, y) - \nabla_y \Phi(x', y')\|_{\mathcal{Y}'} &\leq L_{yy} \|y - y'\|_Y.
\end{align*}
\]

Note that $L_{xy} = L_{yx}$ and the gradient operator $(x, y) \mapsto [\nabla_x \Phi(x, y), -\nabla_y \Phi(x, y)]$ is $M$-Lipschitz on $\mathcal{X} \times \mathcal{Y}$, where $M \triangleq L_{xx} + 2L_{yx} + L_{yy}$.

Based on the assumptions above, we aim to design optimal (or nearly optimal) first-order algorithms that find a saddle point $(x^\dagger, y^\dagger) \in \mathcal{X} \times \mathcal{Y}$ of Problem (1.1), i.e., $(x^\dagger, y^\dagger)$ satisfies that

\[
S(x^\dagger, y^\dagger) \leq S(x^\dagger, y^\dagger) \leq S(x, y^\dagger), \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}.
\]

For well-posedness, we assume that such a saddle point exists (see Assumption 3.1 for conditions that guarantee the existence).

1.1. Stochastic first-order oracles.

Since we aim to solve (1.1) via first-order information, we need to properly set up the oracle model. For generality, we do not assume that the exact gradients of $f$, $\Phi(\cdot, y)$ and $\Phi(x, \cdot)$ can be obtained. Rather, we only assume that we have access to the unbiased estimators of $\nabla f$, $\nabla \Phi(\cdot, y)$ and $\nabla \Phi(x, \cdot)$ (a.k.a., stochastic gradients), which we denote by $\hat{\nabla} f$, $\hat{\nabla} \Phi(\cdot, y)$ and $\hat{\nabla} \Phi(x, \cdot)$ respectively. In addition, we assume that the gradient noise on $\nabla f$, i.e., $\hat{\nabla} f - \nabla f$, has bounded second moment and denote this bound as $\sigma_{x,f}^2$. Similarly, we also assume that gradient noises on $\nabla \Phi(\cdot, y)$ and $\nabla \Phi(x, \cdot)$ have bounded second-moments and denote the bounds as $\sigma_{x,\Phi}^2$ and $\sigma_{y,\Phi}^2$ respectively. In some situations, we will further assume that the gradient noises have sub-Gaussian distributions. For a formal description of the oracle model and a precise statement of the aforementioned assumptions, readers are referred to Section 3 and Assumption 3.2 respectively.

Indeed, the oracles described above are standard in the literature on stochastic approximation, which dates back to Robbins and Monro [34] and since then, has become a popular approach to

---

2 This holds true if the norm on $\mathcal{X} \times \mathcal{Y}$ is defined as $\|(x, y)\|_{\mathcal{X} \times \mathcal{Y}} \triangleq (a\|x\|_X^p + b\|y\|_Y^p)^{1/p}$, for any $a, b \geq 0$, and for any $p \geq 1$ or $p \to +\infty$. 
solve stochastic programming (SP) problems. In the standard SP formulation, the smooth functions $f$ and $\Phi$ are typically represented as expectations (see e.g., Nemirovski et al. [23]), i.e.,

$$f(x) \triangleq \mathbb{E}_{\xi \sim P}[\tilde{f}(x, \xi)] \quad \text{and} \quad \Phi(x, y) \triangleq \mathbb{E}_{\zeta \sim Q}[\tilde{\Phi}(x, y, \zeta)], \quad \forall x \in \mathbb{X}, y \in \mathbb{Y}, \quad (1.6)$$

where $\xi$ and $\zeta$ denote the random variables with distributions $P$ (supported on $\Xi$) and $Q$ (supported on $\mathcal{Z}$) respectively, and the functions $\tilde{f} : \mathbb{X} \times \Xi \to \mathbb{R}$ and $\tilde{\Phi} : \mathbb{X} \times \mathbb{Y} \times \mathcal{Z} \to \mathbb{R}$ are chosen such that $f$ and $\Phi$ satisfy the convexity and smoothness assumptions above. In particular, if we take $P = n^{-1} \sum_{i=1}^{n} \delta_{\xi_{i}}$ and $Q = m^{-1} \sum_{i=1}^{m} \delta_{\zeta_{i}}$, where $\{\xi_{i}\}_{i=1}^{n}$ and $\{\zeta_{i}\}_{i=1}^{m}$ are deterministic points in $\Xi$ and $\mathcal{Z}$ respectively and $\delta_{\xi_{i}}$ denotes the Dirac measure at $\xi_{i}$ (and same for $\delta_{\zeta_{i}}$), then $f$ and $\Phi$ in (1.6) assume the finite-sum forms, i.e.,

$$f(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} \tilde{f}(x, \xi_{i}) \quad \text{and} \quad \Phi(x, y) \triangleq \frac{1}{m} \sum_{i=1}^{m} \tilde{\Phi}(x, y, \zeta_{i}), \quad \forall x \in \mathbb{X}, y \in \mathbb{Y}. \quad (1.7)$$

In this case, we can construct the stochastic (first-order) oracle for $f$ by first sampling an index set $B$ from $[n] \triangleq \{1, \ldots, n\}$ uniformly randomly, and then output the gradient of $f_{B} \triangleq |B|^{-1} \sum_{i \in B} \tilde{f}(x, \xi_{i})$. The stochastic oracle for $\Phi$ can also be constructed in the same way.

In the sequel, we will refer to the class of problems in (1.1) as SPP($L$, $L_{xx}$, $L_{yx}$, $L_{yy}$, $\sigma$, $\mu$), where $\sigma \triangleq \sigma_{x,f} + \sigma_{x,\Phi} + \sigma_{y,\Phi}$ represents the collected stochasticity in the (stochastic) gradients of $f$, $\Phi(\cdot, y)$ and $\Phi(x, \cdot)$. If $\sigma = 0$, then (1.1) degenerates to a deterministic optimization problem.

**Oracle complexity.** For the algorithms in this work and almost all the works in the literature, the number of calls to each of the oracle described above (which returns $\nabla f$, $\nabla \Phi(\cdot, y)$ or $\nabla \Phi(x, \cdot)$ or their stochastic versions) is the same. Therefore, in our complexity analysis and comparison of complexities with other algorithms, we do not distinguish among the these oracles. Instead, the word “oracle complexity” refers to the complexity of each of them.

**1.2. Applications.**

Due to its generality, problem (1.1) has a wide range of applications across many fields, including statistics, machine learning, operations research and game theory. When $\Phi$ is bilinear, i.e., there exists a (bounded) linear operator $A : \mathbb{X} \to \mathbb{Y}^{*}$ such that $\Phi(x, y) = \langle Ax, y \rangle$ (where $\langle \cdot, \cdot \rangle : \mathbb{Y}^{*} \times \mathbb{Y} \to \mathbb{R}$ denotes the duality pairing between $\mathbb{Y}^{*}$ and $\mathbb{Y}$), the applications of (1.1) can be found in numerous previous works, e.g., Juditsky and Nemirovski [15, 16], Chambolle and Pock [5] and Zhao et al. [37]. Beyond bilinear $\Phi$, there are also rich applications. For details, we refer readers to Balamurugan and Bach [1] and Hien et al. [14].

**1.3. Related work.**

Depending on whether $\Phi$ is bilinear, the previous works on solving convex-concave SPPs mainly fall into two categories. In the following, we will briefly review the works for bilinear SPPs (i.e., the problem where $\Phi$ is bilinear), and then focus on the works for non-bilinear SPPs.
1.3.1. Bilinear SPPs. This class of problems is indeed a special case of (1.1), i.e., when $L_{xx} = L_{yy} = 0$. In recent years, both deterministic (i.e., $\sigma = 0$) and stochastic (i.e., $\sigma > 0$) versions of this problem have been thoroughly studied, for both $\mu = 0$ and $\mu > 0$. For the deterministic problems, some well-known algorithms include Nesterov smoothing (a.k.a., excessive gap technique, Nesterov [26, 25]), primal-dual hybrid gradient (PDHG, Chambolle and Pock [4, 5]), hybrid proximal extragradient-type algorithm (HPE-type, He and Monteiro [13]) and primal-dual operator splitting (e.g., Condat [8], Vû [35] and Davis [9]). In addition, to tackle the stochastic problems, stochastic versions of these algorithms have also been developed, e.g., Chen et al. [6], Zhao and Cevher [36] and Zhao et al. [37].

1.3.2. Non-bilinear SPPs. We first focus on the case where $\mu = 0$ and $\sigma = 0$ (i.e., no primal strong convexity and (1.1) is deterministic). When $\Phi$ is possibly nonsmooth, algorithms based on primal-dual subgradient have been developed in several works, including Nedić and Ozdaglar [21], Nesterov [27] and Juditsky and Nemirovski [15]. However, these methods typically incur high oracle complexity, i.e., $O(\epsilon^{-2})$ (where $\epsilon$ denotes the desired accuracy for the duality gap). As a result, they are not competitive when $\Phi$ is smooth (cf. (1.4a) to (1.4d)). The smoothness of $\Phi$ has been exploited in many algorithms to achieve better complexity results. These methods include Mirror-Prox (Nemirovski [22]), HPE-type algorithm (Kolossoski and Monteiro [18]) and PDHG-type algorithm (Hamedani and Aybat [12]). In particular, the last two algorithms are the extensions of their counterparts for solving bilinear SPPs. When $\sigma > 0$, stochastic extensions of Mirror-Prox have been developed in the literature. Some representative works include the stochastic Mirror-Prox (SMP) method (Juditsky et al. [17]) and the stochastic accelerated Mirror-Prox (SAMP) method (Chen et al. [7]).

Unlike the case where $\mu = 0$, there exist very few works that have considered the case where $\mu > 0$. When $L_{yy} = 0$ and $\sigma = 0$ (i.e., the function $\Phi(x, \cdot)$ is linear and (1.1) is deterministic), Juditsky and Nemirovski [16] and Hamedani and Aybat [12] have proposed algorithms, which are based on Mirror-Prox and PDHG respectively, that achieve better complexity than their counterparts that are designed for $\mu = 0$. However, despite their success, two questions remain:

(i) Can we improve the oracle complexities of these two algorithms under the conditions above?
(ii) Can we develop an algorithm that works for all the cases where $\mu > 0$, $L_{yy} > 0$ and $\sigma > 0$?

In this work, we will provide affirmative answers both questions above, by developing an algorithm that is not only sufficiently general to deal all with the cases listed in (ii), but also significantly improves the oracle complexities of the two algorithms introduced above.
1.4. Main Contribution.

Our main contribution is summarized below.

First, when \( \mu = 0 \) (i.e., \( f \) is non-strongly-convex), we develop a stochastic algorithm (i.e., Algorithm 1) by extending PDHG, which was originally developed for bilinear SPPs (cf. Section 1.3.1), to handle the non-bilinear case. In addition, we innovatively incorporate the stochastic acceleration technique (see e.g., Lan [19]) into our algorithm, which indeed enables us to obtain the optimal oracle complexity for the smooth function \( f \). By a judicious and intricate choice of the algorithm parameters, we are able to obtain an \( \epsilon \)-expected duality gap (defined in Section 3.4) with the state-of-the-art oracle complexity

\[
O \left( \frac{\sqrt{L}}{\epsilon} + \frac{L_{xx} + L_{yx} + L_{yy}}{\epsilon} + \frac{(\sigma_{x,f} + \sigma_{x,\Phi})^2 + \sigma_{y,\Phi}^2}{\epsilon^2} \right),
\]

(1.8)

when the second moments of all the gradient noises are bounded (cf. Section 1.1). Previously, the complexity in (1.8) has been achieved by the SAMP algorithm introduced in Chen et al. [7]. However, since this algorithm is based on Mirror-Prox, it significantly differs from our algorithm (which is based on PDHG). Consequently, our work affirms the power of the PDHG framework on the stochastic non-bilinear SPPs.

Regarding the optimality of the complexity in (1.8), we notice that the complexities for \( L \) and \( L_{yx} \) match the lower bounds derived in Ouyang and Xu [30], and the complexities for \( \sigma_{x,f} + \sigma_{x,\Phi} \) and \( \sigma_{y,\Phi} \) match the lower bounds derived in Nemirovskii and Yudin [24]. Therefore, all of these complexities are optimal. In addition, the complexities of \( L_{xx} \) and \( L_{yy} \) are also best-known, although no lower bounds have been derived in the literature.\(^3\)

If in addition, all the gradient noises follow sub-Gaussian distributions, we can obtain an \( \epsilon \)-duality gap with probability (w.p.) at least \( 1 - \zeta \), again, with the state-of-the-art oracle complexity

\[
O \left( \sqrt{\frac{L}{\epsilon}} + \frac{L_{xx} + L_{yx} + L_{yy}}{\epsilon} + \frac{(\sigma_{x,f} + \sigma_{x,\Phi})^2 + \sigma_{y,\Phi}^2}{\epsilon^2} \log \left( \frac{1}{\zeta} \right) \right).
\]

(1.9)

In particular, the \( \log(1/\zeta) \) factor in (1.9) indicates that our algorithm achieves the large-deviation-type convergence results (see e.g., Nemirovski et al. [23]).

Second, when \( \mu > 0 \) (i.e., \( f \) is strongly-convex), we design a novel (multi-stage) stochastic restart scheme (i.e., Algorithm 2S) by using a modified version of Algorithm 1 (developed for the case where \( \mu = 0 \)) as the subroutine. Since SPPs have different structures from convex optimization problems (COPs), our restart scheme is different from that for COPs (e.g., Ghadimi and Lan [11]).

\(^3\)Note that in this case, by taking \( \Phi(x,y) = \phi^P(x,x) + \langle A x, y \rangle - \phi^D(y) \), where the convex functions \( \phi^P \) and \( \phi^D \) are \( L_{xx} \)- and \( L_{yy} \)-smooth respectively, we can obtain trivial lower complexity bounds \( O(\sqrt{L_{xx}/\epsilon}) \) and \( O(\sqrt{L_{yy}/\epsilon}) \). However, this essentially returns to the bilinear case. Therefore, these lower bounds may not be tight for the non-bilinear case.
Specifically, we use a distance-based quantity as the restart criterion, instead of the objective error. In addition, rather than expectation, we analyze the stochasticity via the error probability, which is obtained using techniques in analyzing the finite-state Markov chains. (For detailed explanations, we refer readers to Section 4.4.1.) Since our restart scheme can also interface with other subroutines (e.g., those based on Mirror-Prox), we believe that our restart scheme is of independent interest for solving stochastic SPPs with primal strong convexity.

When the gradient noises have bounded second moments and follow sub-Gaussian distributions, our scheme achieves an $\epsilon$-duality gap w.p. at least $1 - \zeta$, with oracle complexity

$$O\left(\left(\sqrt{\frac{L}{\mu}} + \frac{L_{xx}}{\mu}\right) \log \left(\frac{1}{\epsilon}\right) + \frac{L_{yx}}{\sqrt{\mu} \epsilon} + \frac{L_{yy}}{\epsilon} + \left(\frac{(\sigma_{x,f} + \sigma_{x,\phi})^2}{\mu \epsilon} + \frac{\sigma^2_{y,\phi}}{\epsilon^2}\right) \log \left(\frac{\log(1/\epsilon)}{\zeta}\right)\right). \tag{1.10}$$

Note that even in the deterministic case (i.e., $\sigma = \sigma_{x,f} + \sigma_{x,\phi} + \sigma_{y,\phi} = 0$), this complexity is strictly better than any one in the previous works (cf. Table 2). Based on the complexity in (1.9), under rather mild assumptions on the nonsmooth functions $g$ and $J$ (cf. Assumption 4.1), our scheme can be shown to obtain an $\epsilon$-expected duality gap with oracle complexity

$$O\left(\left(\sqrt{\frac{L}{\mu}} + \frac{L_{xx}}{\mu}\right) \log \left(\frac{1}{\epsilon}\right) + \frac{L_{yx}}{\sqrt{\mu} \epsilon} + \frac{L_{yy}}{\epsilon} + \left(\frac{(\sigma_{x,f} + \sigma_{x,\phi})^2}{\mu \epsilon} + \frac{\sigma^2_{y,\phi}}{\epsilon^2}\right) \log \left(\frac{1}{\epsilon}\right)\right). \tag{1.11}$$

Compared to the complexity in (1.8), we observe that the complexities of $L$, $L_{xx}$, $L_{yx}$ and $\sigma_{x,f} + \sigma_{x,\phi}$ have been greatly improved, due to primal strong convexity. (For detailed discussions, we refer readers to Section 4.4.3.) In (1.11), the complexities of $L_{xx}$ and $L_{yx}$ match the lower bounds derived in Nemirovskii and Yudin [24] and Ouyang and Xu [30] respectively. Additionally, the complexities of $\sigma_{x,f} + \sigma_{x,\phi}$ and $\sigma_{y,\phi}$ nearly match (up to $\log(1/\epsilon)$ factor) the lower bounds derived in Raginsky and Rakhlin [31]. Similar to the case where $\mu = 0$, the complexities of $L_{xx}$ and $L_{yy}$ are the state-of-the-art, although their lower bounds are not available in the literature.
1.5. Notations.

Denote the set of natural numbers by $\mathbb{N} \triangleq \{1, 2, \ldots\}$ and define $\mathbb{Z}_+ \triangleq \mathbb{N} \cup \{0\}$. For any finite-dimensional real normed space $\mathbb{U}$, we denote its dual space by $\mathbb{U}^*$.

We denote the norms on $\mathbb{U}$ and $\mathbb{U}^*$ by $\|\cdot\|$ and $\|\cdot\|_*$ respectively. In addition, denote the duality pairing between $\mathbb{U}^*$ and $\mathbb{U}$ by $\langle \cdot, \cdot \rangle : \mathbb{U}^* \times \mathbb{U} \to \mathbb{R}$. For any CCP function $h : \mathbb{U} \to \mathbb{R}$, define its domain as $\text{dom } h \triangleq \{ u \in \mathbb{U} : h(u) < +\infty \}$.

In addition, for any nonempty set $\mathcal{V}$ in $\mathbb{U}$, denote its interior by $\text{int } \mathcal{V}$ and boundary by $\text{bd } \mathcal{V}$.

2. Preliminaries.

We first introduce the distance generating function and Bregman proximal projection, followed by the primal and dual functions associated with $S(\cdot, \cdot)$ in (1.1).

2.1. Distance generating function and Bregman proximal projection.

Let $\mathbb{U}$ and $h$ be given in Section 1.5. We say that $h$ is essentially smooth if $h$ is continuously differentiable on int dom $h \neq \emptyset$ and for any $u \in \text{bd dom } h$ and any sequence $\{u^k\}_{k \in \mathbb{Z}_+} \subseteq \text{int dom } h$ such that $u^k \to u$, $\|\nabla h(u^k)\|_* \to +\infty$. Let $\mathcal{U}$ be any nonempty, closed and convex set in $\mathbb{U}$. We call $h_{\mathcal{U}}$ a distance generating function (DGF) on $\mathcal{U}$ if it is essentially smooth and continuous on $\mathcal{U}$ and

$$D_{h_{\mathcal{U}}}(u, u') \triangleq h_{\mathcal{U}}(u) - h_{\mathcal{U}}(u') - \langle \nabla h_{\mathcal{U}}(u'), u - u' \rangle \geq (1/2)\|u - u'\|^2, \forall u \in \mathcal{U}, \forall u' \in \mathcal{U}^c,$$  

(2.1)

where $\mathcal{U}^c \triangleq \mathcal{U} \cap \text{int dom } h_{\mathcal{U}}$ and $D_{h_{\mathcal{U}}} : \mathcal{U} \times \mathcal{U}^c \to \mathbb{R}$ is called the Bregman distance associated with $h_{\mathcal{U}}$. Note that (2.1) in particular implies the 1-strong convexity of $h_{\mathcal{U}}$ on $\mathcal{U}^c$. Based on $D_{h_{\mathcal{U}}} (\cdot, \cdot)$, we define the Bregman diameter of $\mathcal{U}$ under $h_{\mathcal{U}}$ as

$$\Omega_{h_{\mathcal{U}}} \triangleq \sup_{u \in \mathcal{U}, u' \in \mathcal{U}^c} D_{h_{\mathcal{U}}}(u, u').$$  

(2.2)

In addition, for any $u' \in \mathcal{U}^c$ and CCP function $\varphi : \mathbb{U} \to \mathbb{R}$, define the Bregman proximal projection (BPP) of $u'$ on $\mathcal{U}$ under $\varphi$ (associated with DGF $h_{\mathcal{U}}$, $u^* \in \mathbb{U}^*$ and $\lambda > 0$) as

$$u' \mapsto u^+ \triangleq \arg \min_{u \in \mathcal{U}} \left[ P_\lambda(u) \triangleq \varphi(u) + \langle u^*, u \rangle + \lambda^{-1} D_{h_{\mathcal{U}}}(u, u') \right].$$  

(2.3)

Note that if $\inf_{u \in \mathcal{U}} \varphi(u) > -\infty$ and $\mathcal{U} \cap \text{dom } P_\lambda \neq \emptyset$, then the minimization problem in (2.3) always has a unique solution in $\mathcal{U}^c \cap \text{dom } \varphi$; see Lemma A.1 for details. We say that the function $\varphi$ has a tractable BPP on $\mathcal{U}$ if there exists a DGF $h_{\mathcal{U}}$ on $\mathcal{U}$ such that the minimization problem in (2.3) has a unique closed-form solution in $\mathcal{U}^c \cap \text{dom } \varphi$, for any $u^* \in \mathbb{U}^*$ and $\lambda > 0$. 

Table 2: Comparison of oracle complexities of all the algorithms to obtain $\varepsilon$-expected duality gap when $\mu > 0$

| Algorithm          | Problem Class  | Oracle Complexity                               |
|--------------------|----------------|-----------------------------------------------|
| PDHG-type [12]     | SPP($L$, $L_{xx}$, $L_{yx}$, 0, 0, $\mu$) | $O\left(\frac{L + L_{xx} + L_{yx}}{\sqrt{\mu}}\right)$ |
| Mirror-Prox-B [16] | SPP($L$, $L_{xx}$, $L_{yx}$, 0, 0, $\mu$) | $O\left(\frac{L + L_{xx}}{\mu} \log \left(\frac{1}{\varepsilon}\right) + \frac{L_{yx}}{\sqrt{\mu}}\right)$ |
| Algorithm 2S       | SPP($L$, $L_{xx}$, $L_{yx}$, $L_{yy}$, $\sigma$, $\mu$) | (1.11) |
Algorithm 1 Optimal Stochastic Primal-Dual Algorithm for Convex $f$

**Input**: Interpolation sequence $\{\beta_t\}_{t \in \mathbb{N}}$, dual stepsizes $\{\alpha_t\}_{t \in \mathbb{N}}$, primal stepsizes $\{\tau_t\}_{t \in \mathbb{N}}$, relaxation sequence $\{\theta_t\}_{t \in \mathbb{N}}$, DGFs $h_Y : Y \rightarrow \mathbb{R}$ and $h_X : X \rightarrow \mathbb{R}$

**Initialize**: $x^1 \in X^\circ$, $y^1 \in Y^\circ$, $\bar{x}^1 = x^1$, $\bar{y}^1 = y^1$, $s^1 = \nabla_y \Phi(x^1, y^1, \zeta^1)$, $t = 1$

**Repeat** (until some convergence criterion is met)

\[
y^{t+1} := \arg\min_{y \in Y} J(y) - \langle s^t, y - y^t \rangle + \alpha_t^{-1} D_{h_Y}(y, y^t)
\]

\[
\bar{x}^{t+1} := (1 - \beta_t)\bar{x}^t + \beta_t x^t
\]

\[
x^{t+1} := \arg\min_{x \in X} g(x) + \langle \nabla_x \Phi(x^t, y^{t+1}, \zeta^t) \rangle + \nabla f(\bar{x}^{t+1}, \xi^t), x - x^t \rangle + \tau_t^{-1} D_{h_X}(x, x^t)
\]

\[
s^{t+1} := (1 + \theta_{t+1}) \nabla_y \Phi(x^{t+1}, y^{t+1}, \zeta^{t+1}) - \theta_{t+1} \nabla_y \Phi(x^t, y^t, \zeta^t)
\]

\[
\bar{y}^{t+1} := (1 - \beta_t)\bar{y}^t + \beta_t y^{t+1}
\]

\[
t := t + 1
\]

**Output**: $(\bar{x}^t, \bar{y}^t)$

2.2. Primal function, dual function and duality gap.

For the SPP in (1.1), we define the associated primal and dual problem as

\[
(P) : \min_{x \in \mathcal{X}} \left[ \tilde{S}(x) \triangleq \sup_{y \in \mathcal{Y}} S(x, y) \right], \quad (D) : \max_{y \in \mathcal{Y}} \left[ \check{S}(x) \triangleq \inf_{x \in \mathcal{X}} S(x, y) \right].
\]  

(2.4)

From the definition in (1.5), we can easily prove the following: Given that a saddle point $(x^t, y^t)$ exists in (1.1), both $(P)$ and $(D)$ have nonempty solution sets $\mathcal{P}^*$ and $\mathcal{D}^*$ respectively. Furthermore, $x^t \in \mathcal{P}^*$ and $y^t \in \mathcal{D}^*$ and $\tilde{S}(x^t) = S(x^t, y^t) = \check{S}(y^t)$. Based on the functions $\tilde{S}$ and $\check{S}$, we define the duality gap

\[
G(x, y) \triangleq \tilde{S}(x) - \check{S}(y) = \sup_{x' \in \mathcal{X}, y' \in \mathcal{Y}} S(x, y') - S(x', y).
\]  

(2.5)

3. Convex $f$: algorithm and convergence analysis.

We first consider the case where $\mu = 0$. We begin with introducing our algorithm, followed by the assumptions needed to analyze it, and finally its convergence results and detailed analysis.

3.1. Algorithm.

The pseudo-code of our algorithm is shown in Algorithm 1. For input, we require two CCP functions $h_Y : Y \rightarrow \mathbb{R}$ and $h_X : X \rightarrow \mathbb{R}$ which are DGFs on $Y$ and $X$ respectively, i.e., they are essentially
smooth on their respective domains and

\[ D_{h_Y}(y, y') \geq (1/2)\|y - y'\|^2, \forall y \in \mathcal{Y}, \forall y' \in \mathcal{Y}^o, \]  

\[ D_{h_X}(x, x') \geq (1/2)\|x - x'\|^2, \forall x \in \mathcal{X}, \forall x' \in \mathcal{X}^o, \]  

where \( \mathcal{Y}^o \triangleq \mathcal{Y} \cap \text{int dom } h_Y \) and \( \mathcal{X}^o \triangleq \mathcal{X} \cap \text{int dom } h_X \). In addition, \( h_Y \) and \( h_X \) are chosen such that the minimization problems in (2.6) and (2.8) have (unique) closed-form solutions in \( \mathcal{Y}^o \cap \text{dom } J \) and \( \mathcal{X}^o \cap \text{dom } g \) respectively (cf. Section 2). In addition, according to (2.2), we define the Bregman diameters of \( \mathcal{X} \) and \( \mathcal{Y} \) under \( h_X \) and \( h_Y \) as \( \Omega_{h_X} \) and \( \Omega_{h_Y} \) respectively, i.e.,

\[ \Omega_{h_X} \triangleq \sup_{x, x' \in \mathcal{X}^o} D_{h_X}(x, x'), \quad \Omega_{h_Y} \triangleq \sup_{y, y' \in \mathcal{Y}^o} D_{h_Y}(y, y'). \]  

We next introduce the stochastic first-order oracles. For each \( t \in \mathbb{N} \), it returns \( \hat{\nabla}_y \Phi(x^t, y^t, \zeta^t) \), \( \hat{\nabla}_x \Phi(x^t, y^{t+1}, \zeta^t) \) and \( \hat{\nabla} f(\bar{x}^{t+1}, \xi^t) \), which are the unbiased estimators for the gradients \( \nabla_y \Phi(x^t, y^t) \), \( \nabla_x \Phi(x^t, y^{t+1}) \) and \( \nabla f(\bar{x}^t) \) respectively, conditioned on the past information. Here \( \zeta^t, \xi^t \) are the underlying random variables that generate the stochasticity. For analysis purposes, let us define a filtration \( \{\mathcal{F}_t\}_{t \in \mathbb{Z}_+} \) based on the stochastic process \( \{(\zeta^t, \xi^t)\}_{t \in \mathbb{N}} \). Specifically, we first define the nested sequence of sets of random variables \( \{\Xi_t\}_{t \in \mathbb{Z}_+} \) such that \( \Xi_0 \triangleq \{0\} \), for any \( t \in \mathbb{N} \), \( \Xi_t \triangleq \{(\zeta^t, \xi^t)\} \). Then, for any \( t \in \mathbb{Z}_+ \), define \( \mathcal{F}_t \) to be the \( \sigma \)-algebra generated by \( \Xi_t \), i.e., the minimal \( \sigma \)-algebra with respect to (w.r.t.) which \( \Xi_t \) is measurable. In addition, for any \( t \in \mathbb{N} \), we define the stochastic gradient “noises”

\[ \delta^t_{y, \Phi} \triangleq \hat{\nabla}_y \Phi(x^t, y^t, \zeta^t) - \nabla_y \Phi(x^t, y^t), \]  

\[ \delta^t_{x, \Phi} \triangleq \hat{\nabla}_x \Phi(x^t, y^{t+1}, \zeta^t) - \nabla_x \Phi(x^t, y^{t+1}), \]  

\[ \delta^t_{x, f} \triangleq \hat{\nabla} f(\bar{x}^{t+1}, \xi^t) - \nabla f(\bar{x}^{t+1}). \]

To measure the progress of Algorithm 1, we adopt the duality gap (defined in (2.5)) and analyze the convergence rate of the sequence \( \{G(x^t, y^t)\}_{t \in \mathbb{N}} \) (in expectation or with high probability). Using the definition in (1.5), we easily see that if \( E[G(x^t, y^t)] \to 0 \) and \( E[(x^t, y^t)] \to (x^\dagger, y^\dagger) \), then \( (x^\dagger, y^\dagger) \) must be a saddle point of Problem (1.1).

3.2. Assumptions.

Before presenting our convergence results, we first place assumptions on the constraint sets \( \mathcal{X} \) and \( \mathcal{Y} \), as well as the stochastic gradient noises \( \delta^t_{y, \Phi}, \delta^t_{x, \Phi} \) and \( \delta^t_{x, f} \).

**Assumption 3.1.**

(A) The Bregman diameters \( \Omega_{h_X} \) and \( \Omega_{h_Y} \) in (3.3) are bounded.

(B) The set \( \mathcal{X} \) and the Bregman diameter \( \Omega_{h_X} \) are bounded.
Assumption 3.2. For any \( x \in \mathcal{X} \) and \( y \in \mathcal{Y} \) and any \( t \in \mathbb{N} \), there exist positive constants \( \sigma_{y, \Phi} \), \( \sigma_{x, \Phi} \) and \( \sigma_{x,f} \) such that

(A) \( \mathbb{E}_{t-1}[\delta_{y, \Phi}^t] = 0 \), \( \mathbb{E}_{t-1}[\delta_{x, \Phi}^t] = 0 \), \( \mathbb{E}_{t-1}[\delta_{x,f}^t] = 0 \) a.s.,

(B) \( \mathbb{E}_{t-1}[\|\delta_{y, \Phi}^t\|^2] \leq \sigma_{y, \Phi}^2 \), \( \mathbb{E}_{t-1}[\|\delta_{x, \Phi}^t\|^2] \leq \sigma_{x, \Phi}^2 \), \( \mathbb{E}_{t-1}[\|\delta_{x,f}^t\|^2] \leq \sigma_{x,f}^2 \) a.s.,

(C) \( \mathbb{E}_{t-1}[\exp(\|\delta_{y, \Phi}^t\|^2 / \sigma_{y, \Phi}^2)] \leq \exp(1) \), \( \mathbb{E}_{t-1}[\exp(\|\delta_{x, \Phi}^t\|^2 / \sigma_{x, \Phi}^2)] \leq \exp(1) \),

\[ \mathbb{E}_{t-1}[\exp(\|\delta_{x,f}^t\|^2 / \sigma_{x,f}^2)] \leq \exp(1) \) a.s.,

where the conditional expectation \( \mathbb{E}_t[\cdot] \triangleq \mathbb{E}[\cdot | \mathcal{F}_t] \), for any \( t \in \mathbb{Z}_+ \).

3.3. Remarks about the assumptions.

We make several remarks about the assumptions above. First, Assumption 3.1(A) implies Assumption 3.1(B) (cf. (3.2)). These two assumptions will be used in proving different convergence results. Note that in many scenarios, Assumption 3.1(A) is equivalent to the boundedness of \( \mathcal{X} \) and \( \mathcal{Y} \). For example, if both \( \mathcal{X} \) and \( \mathcal{Y} \) are finite dimensional real Hilbert spaces (with inner products \( \langle \cdot, \cdot \rangle_{\mathcal{X}} \) and \( \langle \cdot, \cdot \rangle_{\mathcal{Y}} \) and their induced norms \( \|\cdot\|_{\mathcal{X}} \) and \( \|\cdot\|_{\mathcal{Y}} \) respectively) and we take \( h_{\mathcal{X}}(x) = (1/2) \|x\|_{\mathcal{X}}^2 \) and \( h_{\mathcal{Y}}(y) = (1/2) \|y\|_{\mathcal{Y}}^2 \), then

\[ D_{h_{\mathcal{X}}}(x, x') = (1/2) \|x - x'\|_{\mathcal{X}}^2, \quad D_{h_{\mathcal{Y}}}(y, y') = (1/2) \|y - y'\|_{\mathcal{Y}}^2. \]  

(3.7)

The boundedness of \( \mathcal{X} \) and \( \mathcal{Y} \), together with other structural assumptions stated in Section 1, ensures that at least one saddle point of Problem (1.1) exists. In addition, the compactness of \( \mathcal{X} \) and \( \mathcal{Y} \) ensures that the sequence \( \{\mathbb{E}[G(\mathbf{x}, \mathbf{y})]\}_{t \in \mathbb{N}} \) has at least one limit point in \( \mathcal{X} \times \mathcal{Y} \). Hence if \( \mathbb{E}[G(\mathbf{x}, \mathbf{y})] \rightarrow 0 \) (which will be shown in Theorem 3.1), then any limit point of \( \{\mathbb{E}[G(\mathbf{x}, \mathbf{y})]\}_{t \in \mathbb{N}} \) is a saddle point of Problem (1.1).

Moreover, since the definition of the duality gap (cf. (2.5)) involves taking supremum over \( \mathcal{X} \) and \( \mathcal{Y} \), Assumption 3.1 is also needed in our analysis. Note that by using a perturbation-based variant of the duality gap as the convergence criterion, some previous works (e.g., Monteiro and Svaiter [20] and Chen et al. [6]) manage to get rid of this assumption. However, discussions on this approach are out the scope of this work.

Second, in Assumption 3.2, part (A) states that the stochastic noise process \( \{(\delta_{y, \Phi}^t, \delta_{x, \Phi}^t, \delta_{x,f}^t)\}_{t \in \mathbb{Z}_+} \) forms a (vector-valued) martingale difference sequence (MDS) w.r.t. the filtration \( \{\mathcal{F}_t\}_{t \in \mathbb{Z}_+} \).

Part (B) states that the (conditional) second-moment of each of \( \{\delta_{y, \Phi}^t\}_{t \in \mathbb{N}} \), \( \{\delta_{x, \Phi}^t\}_{t \in \mathbb{N}} \) and \( \{\delta_{x,f}^t\}_{t \in \mathbb{N}} \) is uniformly bounded. This assumption is sufficient for proving convergence of \( \{G(\mathbf{x}, \mathbf{y})\}_{t \in \mathbb{N}} \) in expectation, but not enough for showing its convergence with high probability. To achieve this, we need to assume that the (conditional) distributions of these stochastic noises are “light-tailed”. Specifically, in part (C), we assume that \( \delta_{y, \Phi}^t \), \( \delta_{x, \Phi}^t \) and \( \delta_{x,f}^t \) are (conditional) sub-Gaussian random vectors with variance proxies \( \sigma_{y, \Phi}^2 \), \( \sigma_{x, \Phi}^2 \) and \( \sigma_{x,f}^2 \) respectively (Rigollet and Hutter [33]). As we will see in Section 3.5, such an assumption allows us to invoke concentration inequalities (e.g., Asuma-Hoeffding) to obtain large-deviation-type convergence results on the sequence \( \{G(\mathbf{x}, \mathbf{y})\}_{t \in \mathbb{N}} \).
3.4. Convergence results.

We analyze the convergence of the duality gap in two aspects, i.e., in expectation and with high probability.

**Theorem 3.1.** Let Assumptions 3.1 and 3.2(A) hold. In Algorithm 1, for any $t \in \mathbb{N}$, choose

$$
\theta_t = \frac{t - 1}{t}, \quad \beta_t = \frac{2}{t + 1}, \quad \alpha_t = \frac{1}{16 (L_{yx} + L_{yy} + \rho \sigma_{y, \phi} \sqrt{t})},
$$

$$
\tau_t = \frac{t}{2(2L + (L_{xx} + L_{yx})t + \rho' \sigma_{x, \phi} + \sigma_{x, f})^{3/2}},
$$

where $\rho, \rho' > 0$ are constants independent of $(L, L_{xx}, L_{yx}, L_{yy}, \sigma_{x, f}, \sigma_{x, \phi}, \sigma_{y, \phi}, t)$.

(A) If Assumption 3.2(B) also holds, then for any $T \geq 3$, we have

$$
\mathbb{E}[G(\bar{x}^T, \bar{y}^T)] \leq B_\epsilon(T) = \frac{16L}{T(T - 1)} \Omega_{h_X} + \frac{8(L_{xx} + L_{yx})}{T} \Omega_{h_X} + \frac{128(L_{yx} + L_{yy})}{T} \Omega_{h_Y} + \frac{8\sigma_{y, \phi}}{\sqrt{T}} \left( \frac{1}{\rho} + 16\rho \Omega_{h_Y} \right) + \frac{8(\sigma_{x, f} + \sigma_{x, \phi})}{\sqrt{T}} \left( \frac{1}{\rho'} + \rho' \Omega_{h_X} \right).
$$

(B) Let $\varsigma \in (0, 1/6]$. If Assumption 3.2(C) also holds, then w.p. at least $1 - 6\varsigma$,

$$
G(\bar{x}^T, \bar{y}^T) \leq B_{\varsigma}(T) + \frac{8\sigma_{y, \phi}}{\sqrt{T}} \left( \frac{\log(1/\varsigma)}{\rho} + \sqrt{\log(1/\varsigma) \Omega_{h_Y}} \right) + \frac{8(\sigma_{x, f} + \sigma_{x, \phi})}{\sqrt{T}} \left( \frac{\log(1/\varsigma)}{\rho'} + \sqrt{\log(1/\varsigma) \Omega_{h_X}} \right).
$$

Recall that the output of Algorithm 1 is denoted by $(\bar{x}^T, \bar{y}^T)$. Theorem 3.1 indicates that to obtain an $\epsilon$-expected duality gap (i.e., $\mathbb{E}[G(\bar{x}^T, \bar{y}^T)] \leq \epsilon$), the oracle complexity of Algorithm 1 is

$$
O\left( \sqrt{\frac{L}{\epsilon}} + \frac{L_{xx} + L_{yx} + L_{yy}}{\epsilon} + \frac{(\sigma_{x, f} + \sigma_{x, \phi})^2 + \sigma_{y, \phi}^2}{\epsilon^2} \right),
$$

and to obtain $\epsilon$-duality gap w.p. at least $1 - \varsigma$ (i.e., $\Pr\{G(\bar{x}^T, \bar{y}^T) \leq \epsilon\} \geq 1 - \varsigma$), the oracle complexity of Algorithm 1 is

$$
O\left( \sqrt{\frac{L}{\epsilon}} + \frac{L_{xx} + L_{yx} + L_{yy}}{\epsilon} + \frac{(\sigma_{x, f} + \sigma_{x, \phi})^2 + \sigma_{y, \phi}^2}{\epsilon^2} \log \left( \frac{1}{\varsigma} \right) \right).
$$

**Remark 3.1.** Note that the parameter choices in Theorem (3.1) do not involve the Bregman diameter $\Omega_{h_X}$ and $\Omega_{h_Y}$. However, if they are known (or can be estimated), we can choose $\rho = 1/(4\sqrt{\Omega_{h_Y}})$ and $\rho' = 1/\sqrt{\Omega_{h_X}}$ to “optimize” the bound in (3.10) (and 3.11).

3.5. Analysis.

We first present the Bregman proximal inequality associated with the Bregman proximal projection in (2.3) and its corollary. The proof of this inequality can be found in many previous works, e.g., Ghadimi and Lan [10, Lemma 2].
Lemma 3.1. In (2.3), for any \( u \in \mathcal{U} \), we have
\[
\varphi(u^+) - \varphi(u) \leq \langle u^*, u - u^+ \rangle + \lambda^{-1}(D_{h_1}(u, u') - D_{h_1}(u, u^+)) - (2\lambda)^{-1}\|u^+ - u^\|^2. \tag{3.14}
\]

Corollary 3.1. In (2.3), for any \( u \in \mathcal{U} \), we have
\[
\varphi(u^+) - \varphi(u) \leq \langle u^*, u - u^+ \rangle + \lambda^{-1}(D_{h_1}(u, u') - D_{h_1}(u, u^+)) + (\lambda/2)\|u^*\|^2. \tag{3.15}
\]

Proof. First, by Young’s inequality, for any \( u \in \mathcal{U} \) and \( u^* \in \mathcal{U}^* \), we have
\[
|\langle u^*, u \rangle| \leq \|u^*\|_* \|u\| \leq (\eta/2)\|u^*\|^2_2 + (2\eta)^{-1}\|u\|^2, \quad \forall \eta > 0.
\]
Therefore, \( \langle u^*, u^* - u^+ \rangle \leq (2\lambda)^{-1}\|u^+ - u^\|^2 + (\lambda/2)\|u^*\|^2_2 \). Add up this inequality with (3.14), we then complete the proof. \( \square \)

Proposition 3.1. Let Assumptions 3.1 and 3.2(A) hold. If \( \theta_0 = 0, \beta_0 = 2, \alpha_0 = \tau_0 = 1^4 \) and there exists a nonnegative sequence \( \{\gamma_t\}_{t \in \mathbb{Z}^+} \) that satisfies \( \gamma_0 = 0 \) and
\[
0 \leq \theta_t \leq 1, \quad \theta_{t-1} \leq \theta_t, \quad \alpha_t \theta_t \leq \alpha_{t-1}, \quad \gamma_t \theta_t = \gamma_{t-1}, \tag{3.17}
\]
\[
\gamma_{t-1} \beta_{t-1}^{-1} = \gamma_t (\beta_t^{-1} - 1), \quad \gamma_{t-1}/\tau_{t-1} \leq \gamma_t/\tau_t, \quad \alpha_t \leq (2L_{yy})^{-1}, \tag{3.18}
\]
\[
L_{\beta_t} + L_{xx} - (2\tau_t^{-1}) + 4\alpha_t L_{yxx}^2 \leq 0, \quad (1 + \theta_t)L_{yy} - (8\alpha_t)^{-1} \leq 0, \tag{3.19}
\]
for any \( t \in \mathbb{N} \), then for any \( T \geq 3 \), we have the following:

(A) If Assumption 3.2(B) also holds, then
\[
\mathbb{E}[G(\overline{x}^T, \overline{y}^T)] \leq \frac{1}{\beta_T^{-1} - 1} \left\{ \frac{2\theta_T}{\tau_{t-1}} \Omega_{h_1} + \frac{4\theta_T}{\alpha_{t-1}} \Omega_{h_2} \right\}_{\leq B_1(T)} + \frac{4(\sigma_{x,y}^2 + \sigma_{x,f}^2)}{\gamma_T} \sum_{t=1}^{T-1} \gamma_t \tau_t + \frac{22\sigma_{y}^2}{\gamma_T} \sum_{t=1}^{T-1} \gamma_t \alpha_t \right\}_{\leq B_2(T)}. \tag{3.20}
\]

(B) Let \( \varsigma \in (0, 1/5) \). If Assumption 3.2(C) also holds, then w.p. at least \( 1 - 5\varsigma \),
\[
G(\overline{x}^T, \overline{y}^T) \leq \frac{1}{\beta_T^{-1} - 1} \{ B_1(T) + (1 + \log(1/\varsigma))B_2(T) \}
\]
\[
+ \frac{2 \sqrt{\log(1/\varsigma)} \gamma_T (\beta_{t-1}^{-1} - 1)}{\gamma_T} \left( \sum_{t=1}^{T-1} \gamma_t \right)^{1/2} \left( \sigma_{y,f} \Omega_{h_2} + (\sigma_{x,y} + \sigma_{x,f}) \Omega_{h_1} \right). \tag{3.21}
\]

Proof. For convenience, for any \( t \in \mathbb{Z}^+ \) and \( (x, y) \in \mathcal{X} \times \mathcal{Y} \), let us define
\[
\bar{G}(\overline{x}, \overline{y}; x, y) \triangleq S(\overline{x}, y) - S(x, \overline{y})
\]
\[
= (f(\overline{x}) - f(x)) + (g(\overline{x}) - g(x)) + (\Phi(\overline{x}, y) - \Phi(x, \overline{y})) + (J(\overline{y}) - J(y)). \tag{3.22}
\]

\(^4\text{Note that these parameters are not used in Algorithm 1 (which starts with } t = 1\text{), and hence not affect the complexity results — they just make the analysis more convenient.}\)
The most crucial step in our proof is to establish the recursion between \( \widetilde{G}(\overline{x}^{t+1}, \overline{y}^{t+1}; x, y) \) and \( \widetilde{G}(\overline{x}, \overline{y}; x, y) \), which requires us to establish the recursion for each of the four terms in (3.22). By the convexity of \( g \) and \( J \) and convexity-concavity of \( \Phi \), we easily see that

\[
\begin{align*}
g(\overline{x}^{t+1}) - g(x) &\leq (1 - \beta_t)(g(\overline{x}^t) - g(x)) + \beta_t(g(x^{t+1}) - g(x)), \\
J(\overline{y}^{t+1}) - J(y) &\leq (1 - \beta_t)(J(\overline{y}^t) - J(y)) + \beta_t(J(y^{t+1}) - J(y)), \\
\Phi(\overline{x}^{t+1}, y) - \Phi(x, \overline{y}^{t+1}) &\leq [(1 - \beta_t)\Phi(\overline{x}^t, y) + \beta_t\Phi(x^{t+1}, y)] - [(1 - \beta_t)\Phi(x, \overline{y}^t) + \beta_t\Phi(x, y^{t+1})] \\
&\leq (1 - \beta_t)(\Phi(\overline{x}^t, y) - \Phi(x, \overline{y}^t)) + \beta_t(\Phi(x^{t+1}, y) - \beta_t\Phi(x, y^{t+1})).
\end{align*}
\]  

To connect \( f(\overline{x}^{t+1}) \) and \( f(\overline{x}^t) \), we have

\[
\begin{align*}
f(\overline{x}^{t+1}) - f(x) &\leq f(\overline{x}^{t+1}) + \langle \nabla f(\overline{x}^{t+1}), \overline{x}^{t+1} - x^t \rangle + (L/2)\|\overline{x}^{t+1} - x^t\|^2 - f(x) \\
&\leq (1 - \beta_t)(f(\overline{x}^t) + \langle \nabla f(\overline{x}^t), \overline{x}^t - x^t \rangle - f(x)) \\
&+ \beta_t(f(x^{t+1}) + \langle \nabla f(x^{t+1}), x^{t+1} - x^t \rangle - f(x)) + (L/2)\|x^{t+1} - x^t\|^2 \\
&\leq (1 - \beta_t)(f(\overline{x}^t) - f(x)) \\
&+ \beta_t(f(x^{t+1}) + \langle \nabla f(x^{t+1}), x^{t+1} - x^t \rangle - f(x)) + (L\beta_t^2/2)\|x^{t+1} - x^t\|^2,
\end{align*}
\]  

where in (a) we use the descent lemma (e.g., Bertsekas [3]), resulted from the \( L \)-smoothness of \( f \); in (b) we use the step (2.10) in Algorithm 1; in (c) we use the convexity of \( f \) and that \( x^{t+1} - \overline{x}^{t+1} = \beta_t(x^{t+1} - x^t) \) (resulted from the steps (2.7) and (2.10)); in (d) we again use the convexity of \( f \).

Combining (3.23), (3.24), (3.25) and (3.29), we have the following recursion

\[
\begin{align*}
\widetilde{G}(\overline{x}^{t+1}, \overline{y}^{t+1}; x, y) &\leq (1 - \beta_t)\widetilde{G}(\overline{x}^t, \overline{y}^t; x, y) + \beta_t(g(x^{t+1}) - g(x)) + \beta_t(J(y^{t+1}) - J(y)) \\
&+ \beta_t(\Phi(x^{t+1}, y) - \Phi(x, y^{t+1})) + \beta_t(\nabla f(\overline{x}^{t+1}), x^{t+1} - x) + (L\beta_t^2/2)\|x^{t+1} - x^t\|^2.
\end{align*}
\]  

We now can apply the Bregman proximal inequality in (3.14) to the steps (2.7) and (2.10), and obtain bounds on \( g(x^{t+1}) - g(x) \) and \( J(y^{t+1}) - J(y) \), i.e.,

\[
\begin{align*}
J(y^{t+1}) - J(y) &\leq -\langle s^t, y - y^{t+1} \rangle + \alpha_t^{-1}(D_{h,y}(y, y^t) - D_{h,y}(y, y^{t+1})) - (2\alpha_t)^{-1}\|y^{t+1} - y^t\|^2 \\
&= -\langle (1 + \theta_t)\nabla_y \Phi(x^t, y^t, \zeta^t_y) - \theta_t\nabla_y \Phi(x^{t-1}, y^{t-1}, \zeta^t_{y-1}), y - y^{t+1} \rangle \\
&+ \alpha_t^{-1}(D_{h,y}(y, y^t) - D_{h,y}(y, y^{t+1})) - (2\alpha_t)^{-1}\|y^{t+1} - y^t\|^2 \\
&= -\langle (1 + \theta_t)\delta_{y,\Phi} - \theta_t\delta_{y,\Phi}, y - y^{t+1} \rangle - \langle \nabla_y \Phi(x^t, y^t), y - y^{t+1} \rangle \\
&- \theta_t(\nabla_y \Phi(x^t, y^t) - \nabla_y \Phi(x^{t-1}, y^{t-1}), y - y^{t+1})
\end{align*}
\]  

(3.31)
\[ g(x^{t+1}) - g(x) \leq \langle \nabla_x \Phi(x^t, y^{t+1}, \zeta^t) + \nabla f(x^{t+1}, \xi^t), x - x^{t+1} \rangle \]

\[ + \tau_t^{-1}(D_{h_x}(x, x^t) - D_{h_x}(x, x^{t+1})) - (2\tau_t)^{-1}\|x^{t+1} - x^t\|^2 \]

\[ \leq \langle \delta_t^g, \delta_t, x - x^{t+1} \rangle + \langle \nabla_x \Phi(x^t, y^{t+1}) + \nabla f(x^{t+1}), x - x^{t+1} \rangle \]

\[ + \tau_t^{-1}(D_{h_x}(x, x^t) - D_{h_x}(x, x^{t+1})) - (2\tau_t)^{-1}\|x^{t+1} - x^t\|^2, \tag{3.34} \]

Note that when \( t = 1 \), (3.31) holds for any \( \hat{\nabla}_y \Phi(x^0, y^0, \zeta^0_y) \in \mathbb{Y}^* \) since \( \theta_1 = 0 \).

To bound \( \Phi(x^{t+1}, y) - \Phi(x, y^{t+1}) \), we have

\[ \Phi(x^{t+1}, y) - \Phi(x, y^{t+1}) = [\Phi(x^{t+1}, y) - \Phi(x^{t+1}, y^{t+1})] \]

\[ + [\Phi(x^t, y^{t+1}) - \Phi(x, y^{t+1})] + [\Phi(x^{t+1}, y^{t+1}) - \Phi(x^{t+1}, y^{t+1})] \tag{3.35} \]

\[ \leq \langle \nabla_y \Phi(x^{t+1}, y^{t+1}), y - y^{t+1} \rangle + \langle \nabla_x \Phi(x^t, y^{t+1}), x^t - x \rangle \]

\[ + \langle \nabla_x \Phi(x^t, y^{t+1}) \rangle - \langle \nabla_x \Phi(x^t, y^{t+1}) \rangle \]

\[ + \langle L_{xx}/2 \rangle \|x^{t+1} - x^t\|^2 \tag{3.36} \]

\[ \leq \langle \nabla_y \Phi(x^{t+1}, y^{t+1}), y - y^{t+1} \rangle + \langle \nabla_x \Phi(x^t, y^{t+1}), x^t - x \rangle \]

\[ + \langle L_{xx}/2 \rangle \|x^{t+1} - x^t\|^2, \tag{3.37} \]

where in (3.36), we use the concavity of \( \Phi(x^{t+1}, \cdot) \), the convexity of \( \Phi(\cdot, y^{t+1}) \) and the \( L_{xx} \)

smoothness of \( \Phi(\cdot, y^{t+1}) \), respectively.

If we multiply both sides of (3.30) by \( \gamma_t \beta_t^{-1} \), and then substitute (3.32), (3.34) and (3.37) into the resulting inequality, we have

\[ \gamma_t \beta_t^{-1} \tilde{G}(x^{t+1}, y^{t+1}, x, y) \leq \gamma_t \beta_t^{-1} \tilde{G}(x, y^t, x, y) + \langle \gamma_t / 2 \rangle (L \beta_t + L_{xx} - 1/\tau_t) \|x^{t+1} - x^t\|^2 \]

\[ + \gamma_t \tau_t^{-1}(D_{h_x}(x, x^t) - D_{h_x}(x, x^{t+1})) + \gamma_t \beta_t^{-1}(D_{h_y}(y, y^t) - D_{h_y}(y, y^{t+1})) - \gamma_t (2\alpha_t)^{-1} \|y^{t+1} - y^t\|^2 \]

\[ + \gamma_t \langle \nabla_y \Phi(x^{t+1}, y^{t+1}) - \nabla_y \Phi(x^t, y^t), y - y^{t+1} \rangle - \gamma_t \langle \nabla_y \Phi(x^t, y^t) - \nabla_y \Phi(x^{t-1}, y^{t-1}), y - y^t \rangle \]

\[ + \gamma_t \langle \nabla_x \Phi(x^t, y^t) - \nabla_x \Phi(x^{t-1}, y^{t-1}), y - y^{t-1} \rangle \]

\[ + \gamma_t (1 + \theta_t) \langle \delta_t^g, \delta_t^g, y^{t+1} - y^t \rangle \]

\[ + \gamma_t \langle \delta_t^g, \delta_t^g, y^{t+1} - y^t \rangle \tag{3.38} \]

Before we proceed, recall that \( \gamma_t \theta_t = \gamma_{t-1} \) and \( \gamma_t \beta_t^{-1} = \gamma_{t+1} (\beta_{t+1}^{-1} - 1) \), for any \( t \in \mathbb{N} \). This enables us to observe certain recursion patterns (e.g., on \( \gamma_t (\beta_t^{-1} - 1) \tilde{G}(x^t, y^t, x, y) \)) in (3.38). We now bound the terms (I), (II) and (III) in (3.38).
To bound (I), we make use of Young’s inequality (cf. (3.16)), i.e.,
\[
(I) = -\gamma_t \langle \nabla_y \Phi(x^t, y^t) - \nabla_y \Phi(x^t, y^{t-1}), y^t - y^{t+1} \rangle - \gamma_t \langle \nabla_y \Phi(x^t, y^{t-1}) - \nabla_y \Phi(x^{t-1}, y^{t-1}), y^t - y^{t+1} \rangle \\
\leq \gamma_t \langle (2L_{yy})^{-1} \| \nabla_y \Phi(x^t, y^t) - \nabla_y \Phi(x^t, y^{t-1}) \|^2 + (L_{yy}/2)\| y^{t+1} - y^t \|^2 \rangle \\
+ \gamma_t \| 2\alpha_t \| \nabla_y \Phi(x^t, y^{t-1}) - \nabla_y \Phi(x^{t-1}, y^{t-1}) \|^2 + (8\alpha_t \gamma_t)^{-1}\| y^{t+1} - y^t \|^2 \rangle \\
\leq (\gamma_t/2)(\| y^t - y^{t-1} \|^2 + (L_{yy}/2)\| y^{t+1} - y^t \|^2) \\
+ \gamma_t \| 2\alpha_t \| \nabla_y \Phi(x^t, y^{t-1}) - \nabla_y \Phi(x^{t-1}, y^{t-1}) \|^2 + (8\alpha_t \gamma_t)^{-1}\| y^{t+1} - y^t \|^2 \rangle \\
\leq (\gamma_t/2)(\| y^t - y^{t-1} \|^2 + (\gamma_t/2L_{yy} + (4\alpha_t)^{-1})\| y^{t+1} - y^t \|^2) \\
+ (\gamma_t - L_{yy}/2)\| y^t - y^{t-1} \|^2 \leq 2\gamma_t - \alpha_t L_{yy} \| x^t - x^{t-1} \|^2,
\]
where in (a) we use the Lipschitz continuity of $\nabla_y \Phi(x^t, \cdot)$ and $\nabla_y \Phi(\cdot, y^{t-1})$ respectively and in (b) we use the conditions that $\gamma_t \theta_t = \gamma_{t-1}$ and $\alpha_t \theta_t \leq \alpha_{t-1}$ for any $t \in \mathbb{N}$.

To bound (II), we need to use a technique introduced in Nemirovski et al. [23]. Namely, we introduce an auxiliary (stochastic) sequence $\{\hat{x}^t\}_{t \in \mathbb{N}}$ such that $\hat{x}^1 = x^1$ and
\[
\hat{x}^{t+1} \triangleq \arg\min_{x \in \mathcal{X}} \langle \delta_{x, \Phi}^t + \delta_{x, f}^t, x \rangle + \tau_t^{-1}D_{h_X}(x, \hat{x}^t).
\]
Based on $\{\hat{x}^t\}_{t \in \mathbb{N}}$, we can decompose (II) into three parts, i.e.,
\[
(II) = \underbrace{\gamma_t \langle \delta_{x, \Phi}^t + \delta_{x, f}^t, \hat{x}^t - x^t \rangle}_{(II.A)} + \underbrace{\gamma_t \langle \delta_{x, \Phi}^t + \delta_{x, f}^t, \hat{x}^t - x^t \rangle}_{(II.B)} + \underbrace{\gamma_t \langle \delta_{x, \Phi}^t + \delta_{x, f}^t, x^t - x^{t+1} \rangle}_{(II.C)}.
\]
To see the benefit of doing this, note that in (II.B), both $\hat{x}^t, x^t \in \mathcal{F}_{t-1}$, i.e., $\hat{x}^t$ and $x^t$ are measurable w.r.t. $\mathcal{F}_{t-1}$. Therefore by Assumption 3.2(A), $\{\langle \delta_{x, \Phi}^t + \delta_{x, f}^t, \hat{x}^t - x^t \rangle\}_{t \in \mathbb{N}}$ is a MDS adapted to $\{\mathcal{F}_t\}_{t \in \mathbb{N}}$. Moreover, (II.A) and (II.C) can also be bounded using Corollary 3.1 and Young’s inequality respectively, i.e.,
\[
(II.A) \leq \gamma_t \tau_t^{-1}(D_{h_X}(x, \hat{x}^t) - D_{h_X}(x, \hat{x}^{t+1})) + (\gamma_t \tau_t/2)\| \delta_{x, \Phi}^t + \delta_{x, f}^t \|^2,
\]
\[
(II.C) \leq \gamma_t \tau_t \| \delta_{x, \Phi}^t + \delta_{x, f}^t \|^2 + \gamma_t (4\tau_t)^{-1}\| x^t - x^{t+1} \|^2.
\]
In summary, we have
\[
(II) \leq \gamma_t \tau_t^{-1}(D_{h_X}(x, \hat{x}^t) - D_{h_X}(x, \hat{x}^{t+1})) + 2\gamma_t \tau_t \| \delta_{x, \Phi}^t + \delta_{x, f}^t \|^2 \]
\[
+ \gamma_t (4\tau_t)^{-1}\| x^t - x^{t+1} \|^2 + \gamma_t (\delta_{x, \Phi}^t + \delta_{x, f}^t, \hat{x}^t - x^t).
\]
We can bound (III) and (IV) in a similar fashion. Indeed, define $\bar{y}^0 \triangleq y^0$ and for any $t \in \mathbb{N}$,
\[
\bar{y}^{t+1} \triangleq \arg\min_{y \in \mathcal{Y}} \langle \delta_{y, \Psi}^t, y \rangle + \alpha_t^{-1}D_{h_Y}(y, \bar{y}^t).
\]
We then have

\[(I) = \gamma_t (1 + \theta_t) (\delta_{y, \Phi, t} (y^{t+1} - y^t) + \gamma_t (1 + \theta_t) (\delta_{y, \Phi, t} (y^t - \hat{y}^t) + \gamma_t (1 + \theta_t) (\delta_{y, \Phi, t} (\hat{y}^t - y))

\leq (1 + \theta_t) ((1 + \theta_t) \gamma_t \alpha_t \| \delta_{y, \Phi, t} \|^2 + (1 + \theta_t)^{-1} \gamma_t (4 \alpha_t)^{-1} \| y^{t+1} - y^t \|^2)

+ (1 + \theta_t) (\gamma_t \alpha_t^{-1} (D_{h y} (y, \hat{y}^t) - D_{h y} (y, \hat{y}^t)) + (\gamma_t \alpha_t / 2) \| \delta_{y, \Phi, t} \|^2) + \gamma_t (1 + \theta_t) (\delta_{y, \Phi, t} (y^t - \hat{y}^t))

\leq 5 \gamma_t \alpha_t \| \delta_{y, \Phi, t} \|^2 + (4 \alpha_t)^{-1} \| y^{t+1} - y^t \|^2

+ (1 + \theta_t) \gamma_t \alpha_t^{-1} (D_{h y} (y, \hat{y}^t) - D_{h y} (y, \hat{y}^t)) + \gamma_t (1 + \theta_t) (\delta_{y, \Phi, t} (y^t - \hat{y}^t)), \quad (3.46)\]

where in the last inequality we use the fact that \( \theta_t \in [0, 1] \). In addition,

\[(IV) = \gamma_{t-1} (\delta_{y, \Phi, t-1} (y - \hat{y}^{t-1}) + \gamma_{t-1} (\delta_{y, \Phi, t-1} (\hat{y}^{t-1} - y^{t-1}) + \gamma_{t-1} (\delta_{y, \Phi, t-1} (y^{t-1} - y^{t+1})

\leq \gamma_{t-1} (\alpha_{t-1}^{-1} (D_{h y} (y, \hat{y}^{t-1}) - D_{h y} (y, \hat{y}^t)) + \alpha_{t-1}^{-1} (\delta_{y, \Phi, t-1} (y^{t-1} - y^{t-1})

+ \gamma_{t-1} (16 \alpha_{t-1} \| \delta_{y, \Phi, t-1} \|^2 + (64 \alpha_{t-1})^{-1} \| y^{t-1} - y^{t+1} \|^2) + \gamma_{t-1} (\delta_{y, \Phi, t-1} (\hat{y}^{t-1} - y^{t-1}))

\leq \gamma_{t-1} \alpha_{t-1}^{-1} (D_{h y} (y, \hat{y}^{t-1}) - D_{h y} (y, \hat{y}^t)) + 17 \gamma_{t-1} \alpha_{t-1}^{-1} \| \delta_{y, \Phi, t-1} \|^2 + \gamma_{t-1} (32 \alpha_{t-1})^{-1} \| y^{t-1} - y^t \|^2

+ \gamma_{t} (32 \alpha_{t})^{-1} \| y^{t+1} - y^t \|^2 + \gamma_{t-1} (\delta_{y, \Phi, t-1} (\hat{y}^{t-1} - y^{t-1})), \quad (3.47)\]

where in the last inequality we use \( \|a + b\|^2 \leq 2(\|a\|^2 \|b\|^2) \), for any \( a, b \in \mathbb{Y} \) and \( \gamma_{t-1} / \alpha_{t-1} \leq \gamma_t / \alpha_t \) (since \( \alpha_t \theta_t \leq \alpha_{t-1} \) and \( \gamma_t \theta_t = \gamma_{t-1} \)).

Now we can substitute (3.39), (3.44), (3.46) and (3.47) into the recursion (3.38) to obtain

\[
\gamma_{t+1} (\beta_{t+1}^{-1} - 1) \tilde{G}(\bar{x}^{t+1}, \bar{y}^{t+1}; x, y) \leq \gamma_t (\beta_t^{-1} - 1) \tilde{G}(\bar{x}^t, \bar{y}^t; x, y)

+ (\gamma_t / 2) (L \beta_t + L_{xx} - 1 / (2 \tau_t)) \| x^{t+1} - x^t \|^2 + 2 \gamma_{t-1} \alpha_{t-1} L_{yx}^2 \| x^t - x^{t-1} \|^2

+ (\gamma_t / 2) (\theta_t L_{yy} - 3 (16 \alpha_t)^{-1}) \| y^{t+1} - y^t \|^2 + (\gamma_t / 2) (L_{yy} + 1 / (16 \alpha_t)) \| y^{t+1} - y^t \|^2

+ \gamma_t \alpha_t^{-1} \{ (D_{h x} (x, x^t) + D_{h x} (x, \hat{x}^t)) - (D_{h x} (x, x^{t+1}) + D_{h x} (x, \hat{x}^{t+1})) \} \}

+ \gamma_t \alpha_t^{-1} (D_{h y} (y, \hat{y}^t) - D_{h y} (y, y^{t+1})) + (1 + \theta_t) \gamma_t \alpha_t^{-1} (D_{h y} (y, \hat{y}^t) - D_{h y} (y, y^{t+1}))

+ \gamma_{t-1} \alpha_{t-1}^{-1} (D_{h y} (y, \hat{y}^{t-1}) - D_{h y} (y, \hat{y}^t))

+ 2 \gamma_t \alpha_t \| \delta_{x, \Phi} + \delta_{x, \Phi}^t \|^2 + 5 \gamma_t \alpha_t \| \delta_{y, \Phi} \|^2 + 17 \gamma_{t-1} \alpha_{t-1} \| \delta_{y, \Phi} \|^2

+ \gamma_t (\nabla_y \Phi(x^{t+1}, y^{t+1}) - \nabla_y \Phi(x^t, y^t) - y + y^{t+1} - y^{t+1} - y^{t+1} - y^t - y^t) \}

+ \gamma_t (\delta_{x, \Phi}^t + \delta_{x, \Phi}^t, \hat{x}^t - x^t) + \gamma_t (1 + \theta_t) (\delta_{y, \Phi}^t, y^t - \hat{y}^t) + \gamma_{t-1} (\delta_{y, \Phi}^{t-1}, \hat{y}^{t-1} - y^{t-1}). \quad (3.48)\]
We then sum up the inequality (3.48) over \( t = 1, \ldots, T - 1 \) to obtain

\[
\gamma_T \left( \beta_{T-1}^{-1} - 1 \right) \tilde{G}(\mathbf{x}^T, \mathbf{y}^T; x, y) \\
\leq \sum_{t=1}^{T-1} \frac{\gamma_t}{2} \left( L_{\beta_t} + L_{xx} - \frac{1}{2\tau_t} + \frac{4\alpha_t L_{yx}^2}{1} \right) \|x^{t+1} - x^t\|^2 - 2\gamma_T \alpha_{T-1} L_{yx}^2 \|x^T - x^{T-1}\|^2 \\
+ \sum_{t=1}^{T-1} \frac{\gamma_t}{2} \left( 1 + \theta_t \right) L_{yy} - \frac{1}{8\alpha_t} \|y^{t+1} - y^t\|^2 - \frac{\gamma_{T-1} \alpha_T - 1}{2} \left( L_{yy} + \frac{1}{16\alpha_T} \right) \|y^T - y^{T-1}\|^2 \\
+ \sum_{t=1}^{T-1} \left( \frac{\gamma_t}{\tau_t} - \frac{\gamma_{t-1}}{\tau_{t-1}} \right) \left( D_{hx}(x, x^t) + D_{hx}(x, \tilde{x}^t) \right) + \sum_{t=1}^{T-1} \left( \frac{\gamma_t}{\alpha_t} - \frac{\gamma_{t-1}}{\alpha_{t-1}} \right) \left( D_{hy}(y, y^t) - \frac{\gamma_{T-1}}{\alpha_T} D_{hy}(y, y^T) \right) \\
+ \sum_{t=1}^{T-1} \left( (2 + \theta\gamma_t) \frac{\gamma_t}{\alpha_t} - (2 + \theta\gamma_{t-1}) \frac{\gamma_{t-1}}{\alpha_{t-1}} \right) \left( D_{hy}(y, y^t) + 2 \sum_{i=1}^{T-1} \gamma_{t_1} \| \delta_{x,\Phi}^i + \delta_{x,\tilde{x}}^i \|^2 + 22 \sum_{i=1}^{T-1} \gamma_{t_1} \| \delta_{y,\Phi}^i \|^2 \right) \\
+ \gamma_T (\nabla_y \Phi(x^T, y^T) - \nabla_y \Phi(x^{T-1}, y^{T-1}), y - y^T) \frac{1}{\alpha_T} \sum_{t=1}^{T-2} \gamma_t \left( \delta_{x,\Phi}^t + \delta_{x,\tilde{x}}^t \right) + \sum_{t=1}^{T-1} \gamma_t \left( \delta_{y,\Phi}^t, y^t - \tilde{y}^t \right) + \gamma_T - (1 + \theta_T) \left( \delta_{y,\Phi}^{T-1}, y^{T-1} - \tilde{y}^{T-1} \right),
\]

where we have used the facts that \( \gamma_{t-1}/\alpha_{t-1} \leq \gamma_t/\alpha_t \).

In addition, we can bound (V) in a similar fashion to bounding (I) (cf. (3.39)), i.e.,

\[
(V) = \gamma_T - \left\{ \langle \nabla_y \Phi(x^T, y^T) - \nabla_y \Phi(x^T, y^{T-1}), y - y^T \rangle + \langle \nabla_y \Phi(x^T, y^{T-1}) - \nabla_y \Phi(x^{T-1}, y^{T-1}), y - y^T \rangle \right\} \\
\leq \gamma_T \left\{ \alpha_T^{-1} L_{yy}^2 \|y^T - y^{T-1}\|^2 + (4\alpha_T^{-1})^{-1} \|y^T - y^{T-1}\|^2 \\
+ \alpha_T^{-1} L_{yx}^2 \|x^T - x^{T-1}\|^2 + (4\alpha_T^{-1})^{-1} \|y^T - y^{T-1}\|^2 \right\} \\
\leq \left( \gamma_T L_{yy}^2 / 2 \right) \|y^T - y^{T-1}\|^2 + \gamma_T - \alpha_T^{-1} L_{yx}^2 \|x^T - x^{T-1}\|^2 + \gamma_T - \alpha_T^{-1} D_{hy}(y, y^T),
\]

where in the last inequality we have use \( \alpha_{T-1} \leq (2L_{yy})^{-1} \) and \( (1/2) \|y^T - y^{T-1}\|^2 \leq D_{hy}(y, y^T) \).

We now substitute (3.50) into (3.49) and simplify the resulting inequality by noting that

1) \( \beta_{T-1}^{-1} - 1 \leq \frac{1}{2}\tau_1 - (4\alpha_1 L_{yx}^2) \leq 0, \) \( (1 + \theta_1) L_{yy} - (8\alpha_1)^{-1} \leq 0. \)
2) \( \gamma_{t-1}/\tau_{t-1} \leq \gamma_t/\tau_t, D_{hx}(x, x^t), D_{hx}(x, \tilde{x}^t) \leq \Omega_{hx}, \gamma_{t-1}/\alpha_{t-1} \leq \gamma_t/\alpha_t, \)

\( (2 + \theta_T) \gamma_{t-1} \gamma_t \leq (2 + \theta_T) \gamma_t \alpha_t \) (since \( \gamma_{t-1} \leq \gamma_t \), \( D_{hy}(y, y^T), D_{hy}(\tilde{y}, \tilde{y}^T) \leq \Omega_{hy} \).

As a result, we have

\[
\gamma_T \left( \beta_{T-1}^{-1} - 1 \right) G(\mathbf{x}^T, \mathbf{y}^T; x, y) \leq \frac{2\gamma_T \Omega_{hx}}{\tau_{T-1}} + \frac{4\gamma_T \Omega_{hy}}{\alpha_T \tau_{T-1}} + 2 \sum_{t=1}^{T-1} \gamma_t \tau_t \| \delta_{x,\Phi}^t + \delta_{x,\tilde{x}}^t \|^2 + 22 \sum_{t=1}^{T-1} \gamma_t \| \delta_{y,\Phi}^t \|^2 \\
+ \sum_{t=1}^{T-1} \gamma_t \left( \delta_{x,\Phi}^t + \delta_{x,\tilde{x}}^t - \tilde{x}^t \right) + \sum_{t=1}^{T-2} \gamma_t \left( \delta_{y,\Phi}^t, y^t - \tilde{y}^t \right) + \gamma_T (1 + \theta_T) \left( \delta_{y,\Phi}^{T-1}, y^{T-1} - \tilde{y}^{T-1} \right).
\]

Furthermore, note that (3.51) holds for any \((x, y) \in \mathcal{X} \times \mathcal{Y}\), so we can replace the LHS of (3.51) by \( \gamma_T (\beta_{T-1}^{-1} - 1) G(\mathbf{x}^T, \mathbf{y}^T; x, y) \). As a result, (3.51) becomes an almost sure bound on \( G(\mathbf{x}^T, \mathbf{y}^T) \).
The proof of parts (A) and (B) differ by the way we treat the inequality (3.51). For part (A), we simply take expectation on both sides; whereas for part (B), we need to apply concentration inequalities and the Chernoff bound. The details are shown below.

**Proof of Part (A).** For any \( t \in \mathbb{N} \), since \( x^t, \hat{x}^t, y^t, \hat{y}^t \in \mathcal{F}_{t-1} \), both \( \{ \langle \delta^t_{x,\Phi} + \delta^t_{x,f}, \hat{x}^t - x^t \rangle \}_t \in \mathbb{N} \) and \( \{ \langle \delta^t_{y,\Phi}, y^t - \hat{y}^t \rangle \}_t \in \mathbb{N} \) are MDS adapted to \( \{ \mathcal{F}_t \}_t \in \mathbb{Z}_+ \). Therefore,

\[
\mathbb{E}[\langle \delta^t_{x,\Phi} + \delta^t_{x,f}, \hat{x}^t - x^t \rangle] = \mathbb{E}[\langle \delta^t_{y,\Phi}, y^t - \hat{y}^t \rangle] = 0.
\] (3.52)

In addition, by Assumption 3.2(B), we have

\[
\mathbb{E}[\| \delta^t_{y,\Phi} \|^2] = \mathbb{E}[\mathbb{E}_{t-1}[\| \delta^t_{y,\Phi} \|^2]] \leq \sigma^2_{y,\Phi},
\] (3.53)

\[
\mathbb{E}[\| \delta^t_{x,f} + \delta^t_{x,\Phi} \|^2] = \mathbb{E}[\mathbb{E}_{t-1}[\| \delta^t_{x,f} + \delta^t_{x,\Phi} \|^2]] \leq 2(\sigma^2_{x,f} + \sigma^2_{x,\Phi}),
\] (3.54)

We then take expectation on both sides of (3.51) and substitute (3.52), (3.53) and (3.54) to the resulting inequality to obtain (3.20).

**Proof of Part (B).** We first present the Asuma-Hoeffding lemma for sub-Gaussian MDS.

**Lemma 3.2 (Nemirovski et al. [23]).** Let \( \{ \epsilon_t \}_t \in \mathbb{N} \) be a real-valued MDS adapted to a filtration \( \{ \mathcal{F}_t \}_t \in \mathbb{Z}_+ \), such that for any \( t \in \mathbb{N} \), \( \mathbb{E}_{t-1}[\epsilon_t] = 0 \) and there exists a constant \( d_t > 0 \) such that \( \mathbb{E}_{t-1}[\epsilon_t^2/d_t^2] \leq \exp(1) \). Then for any \( p > 0 \) and \( T \in \mathbb{N} \),

\[
\Pr \left\{ \sum_{t=1}^{T-1} \epsilon_t > p \left( \sum_{t=1}^{T-1} d_t^2 \right)^{1/2} \right\} \leq \exp(-p^2/4).
\] (3.55)

For convenience, define \( C_T \triangleq \sum_{t=1}^{T-1} \gamma_t \alpha_t \) and \( C'_T \triangleq \sum_{t=1}^{T-1} \gamma_t \tau_t \). Then

\[
\Pr \left\{ \sum_{t=1}^{T-1} \gamma_t \alpha_t \| \delta^t_{y,\Phi} \|_x^2 > (1+p)C_T \sigma^2_{y,\Phi} \right\} = \Pr \left\{ \exp \left( \frac{1}{C_T} \sum_{t=1}^{T-1} \gamma_t \tau_t \frac{\| \delta^t_{y,\Phi} \|_x^2}{\sigma^2_{y,\Phi}} \right) > \exp(1+p) \right\}
\]

\[
\overset{(a)}{\leq} \exp(-1-p) \mathbb{E} \left[ \exp \left( \frac{1}{C_T} \sum_{t=1}^{T-1} \gamma_t \tau_t \frac{\| \delta^t_{y,\Phi} \|_x^2}{\sigma^2_{y,\Phi}} \right) \right]
\]

\[
\overset{(b)}{\leq} \exp(-1-p) \frac{1}{C_T} \sum_{t=1}^{T-1} \gamma_t \tau_t \mathbb{E} \left[ \exp \left( \frac{\| \delta^t_{y,\Phi} \|_x^2}{\sigma^2_{y,\Phi}} \right) \right]
\]

\[
\overset{(c)}{\leq} \exp(-p),
\] (3.56)

where in (a) we use Markov’s inequality, in (b) we use the convexity of \( \exp(\cdot) \) and in (c) we use Assumption 3.2(C). Similarly, we can also show that for any \( \bar{p} > 0 \),

\[
\Pr \left\{ \sum_{t=1}^{T-1} \gamma_t \alpha_t \| \hat{\delta}^t_{x,f} \|_x^2 > (1+\bar{p})C'_T \sigma^2_{x,f} \right\} \leq \exp(-\bar{p}),
\] (3.57)

\[
\Pr \left\{ \sum_{t=1}^{T-1} \gamma_t \alpha_t \| \hat{\delta}^t_{x,\Phi} \|_x^2 > (1+\bar{p})C'_T \sigma^2_{x,\Phi} \right\} \leq \exp(-\bar{p}).
\] (3.58)
Next, since \( \{\gamma_t, \delta_{x,f}^t, x^t - \tilde{x}^t\}\) and \( \{\gamma_t, \delta_{y,\Phi}^t, y^t - \tilde{y}^t\}\) are MDS, we can apply Lemma 3.2 to the last three terms in (3.51). Specifically, let us define
\[
d_{t,y}^2 \triangleq 4\gamma_{t-1}^2 \sigma_{y,\Phi}^2 \Omega_{h,y} + d_{t,x,\Phi}^2 \triangleq 2\gamma_t^2 \sigma_{x,\Phi}^2 \Omega_{h,x}, \quad d_{t,\tilde{x},f}^2 \triangleq 2\gamma_t^2 \sigma_{x,f}^2 \Omega_{h,x}.
\]
Then by Assumption 3.2(C), for any \( t = 1, \ldots, T - 2 \), we have
\[
\mathbb{E}_{t-1} \left[ \exp \left( |\gamma_{t-1}(\delta_{y,\Phi}^t, \tilde{y}^t - y^t)^2/d_{t,y}^2| \right) \right] \leq \mathbb{E}_{t-1} \left[ \exp \left( 2\gamma_{t-1}^2 \|\delta_{y,\Phi}^t\|^2\|\tilde{y}^t - y^t\|^2/d_{t,y}^2 \right) \right] \leq \exp(1),
\]
and
\[
\mathbb{E}_{t-1} \left[ \exp \left( |(\gamma_{T-2} + \gamma_{T-1})(\delta_{y,\Phi}^{T-1}, \tilde{y}^{T-1} - y^{T-1})|^2/(4(\gamma_{T-2}^2 + \gamma_{T-1}^2)\sigma_{y,\Phi}^2 \Omega_{h,y}) \right) \right] \leq \exp(1).
\]
Similarly, for any \( t = 1, \ldots, T - 1 \), we have
\[
\mathbb{E}_{t-1} \left[ \exp \left( |\gamma_t(\delta_{x,f}^t, \tilde{x}^t - x^t)^2/d_{t,x,f}^2| \right) \right] \leq \exp(1),
\]
and
\[
\mathbb{E}_{t-1} \left[ \exp \left( |\gamma_t(\delta_{x,f}^t, \tilde{x}^t - x^t)^2/d_{t,x,f}^2| \right) \right] \leq \exp(1).
\]
Thus by Lemma 3.2, for any \( q, \tilde{q} > 0 \), we have
\[
\Pr \left\{ \sum_{t=1}^{T-2} \gamma_{t-1} \delta_{y,\Phi}^t, \tilde{y}^t - y^t + (\gamma_{T-2} + \gamma_{T-1})(\delta_{y,\Phi}^{T-1}, \tilde{y}^{T-1} - y^{T-1}) > 2q\sigma_{y,\Phi} \sqrt{\Omega_{h,y} \left( \sum_{t=1}^{T-1} \gamma_t^2 \right)^{1/2}} \right\} \leq \exp(-q^2/4),
\]
and
\[
\Pr \left\{ \sum_{t=1}^{T-1} \gamma_t \delta_{x,f}^t, \tilde{x}^t - x^t > 2\tilde{q}(\sigma_{x,\Phi} + \sigma_{x,f}) \sqrt{\Omega_{h,x} \left( \sum_{t=1}^{T-1} \gamma_t^2 \right)^{1/2}} \right\} \leq 2\exp(-\tilde{q}^2/4).
\]
We then combine (3.56), (3.57), (3.58), (3.64) and (3.65), and take \( p = \bar{p} = \log(1/\varsigma) \) and \( q = \tilde{q} = 2\sqrt{\log(1/\varsigma)} \) to obtain (3.21).

\textbf{Proof of Theorem 3.1.} We first verify the choices of the input sequences \( \{\beta_t\}_{t \in \mathbb{N}}, \{\alpha_t\}_{t \in \mathbb{N}}, \{\tau_t\}_{t \in \mathbb{N}} \) and \( \{\theta_t\}_{t \in \mathbb{N}} \) in Theorem 3.1 indeed satisfy the conditions required in Proposition 3.1. Indeed, based on these choices, we can choose \( \gamma_t = t \), for any \( t \in \mathbb{N} \). We only show the steps to verify the conditions in (3.19). First, since \( \tau_{t-1} \geq (4L + 2(L_{yx} + L_{xx})t)/t \) and \( \alpha_t \leq 1/(16L_{yx}) \), we have
\[
L\beta_t + L_{xx} - (1/2)\tau_{t-1}^2 + 4\alpha_t L_{yx}^2 \leq 2L/(t + 1) + L_{xx} - (2L + (L_{yx} + L_{xx})t)/t + L_{yx}/4 \leq 0.
\]
Also, since \( \alpha_t^{-1} \geq 16L_{yy} \) and \( \theta_t \leq 1 \), we have
\[
(1 + \theta_t)L_{yy} - (1/8)\alpha_t^{-1} \leq (1 + \theta_t)L_{yy} - 2L_{yy} \leq 0.
\]
Next, we bound the summation terms appearing in (3.20) and (3.21). Specifically, by noting that 
\[ \alpha_t \leq (16\rho_s \sqrt{t})^{-1} \text{ and } \tau_t \leq (2\rho' (\sigma_x f + \sigma_z \Phi) \sqrt{t})^{-1}, \]
we have
\begin{align*}
\sum_{t=1}^{T-1} \gamma_t \alpha_t &\leq \frac{1}{16\rho_s \sqrt{t}} \sum_{t=1}^{T-1} \sqrt{t} \leq \frac{1}{16\rho_s \sqrt{t}} \int_{t=0}^{T} \sqrt{t} dt \leq \frac{T^{3/2}}{16\rho_s \sqrt{t}}, \\
\sum_{t=1}^{T-1} \gamma_t \tau_t &\leq \frac{1}{2\rho' (\sigma_x f + \sigma_z \Phi)} \sum_{t=1}^{T-1} \sqrt{t} \leq \frac{T^{3/2}}{2\rho' (\sigma_x f + \sigma_z \Phi)}, \\
\sum_{t=1}^{T-1} \gamma_t^2 &\leq \sum_{t=1}^{T-1} t^2 \leq \int_{t=0}^{T} t^2 dt = \frac{1}{3}T^3.
\end{align*}
(3.67) (3.68) (3.69)
We then substitute (3.8), (3.9), (3.67), (3.68) and (3.69) into (3.20) and (3.21) to obtain (3.10) and (3.11).

4. **Strongly Convex \( f \): restart scheme and complexity analysis.**

We next consider the case where \( \mu > 0 \). We aim to develop restart schemes based on Algorithm 1 that can significantly improve the oracle complexities in (3.12) and (3.13).

4.1. **Algorithm 1 with scaled geometry.**

We first introduce a variant of Algorithm 1 (for non-s.c. \( f \)) that can be used as the subroutine in our restart scheme. To do so, let us define different norms on \( \mathcal{X} \) and \( \mathcal{Y} \), i.e., \[ \|\cdot\|_{\mathcal{X},R} \triangleq (\sqrt{\eta_x}/R) \|\cdot\|_{\mathcal{X}} \]
and \[ \|\cdot\|_{\mathcal{Y}} \triangleq \sqrt{\eta_y} \|\cdot\|_{\mathcal{Y}}, \]
where \( \eta_x, \eta_y > 0 \). Consequently, the dual norms \[ \|\cdot\|'_{\mathcal{X},R} = (R/\sqrt{\eta_x}) \|\cdot\|_{\mathcal{X}}, \]
and \[ \|\cdot\|'_{\mathcal{Y}} = (1/\sqrt{\eta_y}) \|\cdot\|_{\mathcal{Y}}. \]
Under such a geometry, if we define \( L' \triangleq LR^2/\eta_x, L'_{xx} \triangleq L_{xx}R^2/\eta_x, L'_y \triangleq L_yR/\sqrt{\eta_x\eta_y} \) and \( L'_{yy} = L_{yy}/\eta_y \), then
\begin{align*}
\|\nabla f(x) - \nabla f(x')\|'_{\mathcal{X},R} &\leq L' \|x - x'\|'_{\mathcal{X},R}, \\
\|\nabla_x \Phi(x, y) - \nabla_x \Phi(x', y)\|'_{\mathcal{X},R} &\leq L'_{xx} \|x - x'\|'_{\mathcal{X},R}, \\
\|\nabla_x \Phi(x, y) - \nabla_x \Phi(x', y)\|'_{\mathcal{X},R} &\leq L'_{xx} \|x - x'\|'_{\mathcal{X},R}, \\
\|\nabla_y \Phi(x, y) - \nabla_y \Phi(x', y)\|'_{\mathcal{Y}} &\leq L'_{yy} \|y - y'\|'_{\mathcal{Y}}.
\end{align*}
(4.1) (4.2) (4.3) (4.4) (4.5)
(Again, we abbreviate \( \|\cdot\|'_{\mathcal{X},R} \) and \( \|\cdot\|'_{\mathcal{X},R} \) as \( \|\cdot\|'_{\mathcal{R}} \) and \( \|\cdot\|'_{\mathcal{X},R} \), and \( \|\cdot\|'_{\mathcal{Y}} \) and \( \|\cdot\|'_{\mathcal{Y}} \) as \( \|\cdot\|' \) and \( \|\cdot\|' \), whenever no confusion is caused.) Fix any \( x_0 \in \mathcal{X}^o \) and define \( \mathcal{X}(x_0, R) \triangleq R\mathcal{X} + x_0 \). We then define two new DGFs on \( \mathcal{X} \) and \( \mathcal{Y} \) respectively, i.e.,
\[ \bar{h}_{\mathcal{X}(x_0, R)}(x) \triangleq \eta_x h_x((x - x_0)/R), \quad \bar{h}_{\mathcal{Y}}(y) \triangleq \eta_y h_y(y), \quad \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}. \]
(4.6)
We can easily see that \( \bar{h}_{\mathcal{X}(x_0, R)} \) and \( \bar{h}_{\mathcal{Y}} \) are 1-s.c. on \( \mathcal{X}(x_0, R)^o \) and \( \mathcal{Y}^o \) w.r.t. \( \|\cdot\|'_{\mathcal{X}} \) and \( \|\cdot\|'_{\mathcal{Y}} \) respectively. In addition, we have \( D_{\bar{h}_{\mathcal{Y}}}(y, y') = \eta_y D_{h_y}(y, y') \) and
\[ D_{\bar{h}_{\mathcal{X}(x, R)}}(x, x') = \eta_x \left\{ h_x \left( \frac{x - x_0}{R} \right) - h_x \left( \frac{x' - x_0}{R} \right) - \langle \nabla h_x \left( \frac{x' - x_0}{R} \right), \frac{x - x'}{R} \rangle \right\}. \]
Algorithm 1S: Optimal Stochastic Primal-Dual Algorithm for Convex \( f \) with Scaled Geometry

**Input:** Starting primal variable \( x^0 \in X^0 \), positive radius \( R \geq 2\|x^0 - x^1\| \), primal constraint set \( X' \) (which satisfies \( x^0 \in X' \subseteq X \) ), number of iterations \( T \geq 3 \), interpolation sequence \( \{ \beta_t \}_{t \in \mathbb{N}} \), dual stepsizes \( \{ \alpha_t \}_{t \in \mathbb{N}} \), primal stepsizes \( \{ \tau_t \}_{t \in \mathbb{N}} \), relaxation sequence \( \{ \theta_t \}_{t \in \mathbb{N}} \), DGFs \( \tilde{h}_X : Y \rightarrow \mathbb{R} \) and \( \tilde{h}_{X(x_c,R)} : X \rightarrow \mathbb{R} \).

**Initialize:** \( x^1 = x^0 \), \( y^1 \in Y^0 \), \( \pi^1 = x^1 \), \( \pi^1 = y^1 \), \( s^1 = \nabla_y \Phi(x^1, y^1, \zeta^1_y) \)

**For** \( t = 1, \ldots, T - 1 \)

\[
\begin{align*}
y^{t+1} &= \arg \min_{y \in Y} J(y) - \langle s^t, y - y^t \rangle + \alpha_t^{-1} D_{\tilde{h}_Y}(y, y^t) \quad (4.8) \\
x^{t+1} &= \arg \min_{x \in X'} \langle \nabla_x \Phi(x^t, y^{t+1}, \zeta^t_x) + \tilde{\nabla} f(\hat{x}^{t+1}, \xi^t), x - x^t \rangle + \tau_t^{-1} D_{\tilde{h}_{X(x^t,R)}}(x, x^t) \quad (4.9)
\end{align*}
\]

**Output:** \( (\pi^T, y^T) \)

Define \( B_{\| \cdot \|}^{x_c,R} \triangleq \{ x \in X : \| x - x_c \| \leq R \} \). If \( B_{\| \cdot \|}^{0,1} \subseteq \text{dom} h_X \), we then have

\[
\sup_{x \in X'} \| D_{\tilde{h}_{X(x, R)}}(x, x_c) \| \leq \eta_x \Omega'_{h_X}, \\
\text{where } \Omega'_{h_X} \triangleq \sup_{z \in B_{\| \cdot \|}^{0,1}} h_X(z) - h_X(0) - \langle \nabla h_X(0), z \rangle < +\infty. \quad (4.7)
\]

Note that the condition \( B_{\| \cdot \|}^{0,1} \subseteq \text{dom} h_X \) is satisfied when \( h_X = (1/2) \| \cdot \|^2_X \) (where \( X \) is a Hilbert space), in which case \( \text{dom} h_X = X \). For some other examples, we refer readers to Nesterov [26, Section 4] and Nemirovski [22, Section 5].

Our modified algorithm is shown in Algorithm 1S. We make two remarks about it. First, in the input, \( x^0 \) denotes primal part of any saddle point \( (x^0, y^0) \) of (1.1) (cf. (1.5)). In addition, it is the unique minimizer of the primal function \( \bar{S} \) on \( X \) (cf. (2.4)). Second, note that if step (2.8) has a closed-form solution, then in many cases, so does step (4.9). We illustrate three cases here. The first one is when \( X \) is a Hilbert space, so we can take \( X' = X \) and \( h_X = (1/2) \| \cdot \|_X^2 \) (cf. Section 3.3). The second one is when \( g = 0 \) and \( X' = X = \mathbb{R} \). Finally, if \( X \) is a Hilbert space and \( g = 0 \), we can take \( h_X = (1/2) \| \cdot \|_X^2 \) and \( X' \) to be any set that admits a closed-form orthogonal projection.

### 4.2. Deterministic restart scheme.

For ease of exposition, we first develop our restart scheme when we can obtain the exact gradients of \( f \), \( \Phi(\cdot, y) \) and \( \Phi(x, \cdot) \), i.e., when \( \sigma_{x,f} = \sigma_{x,\Phi} = \sigma_{y,\Phi} = 0 \). (The restart scheme for the stochastic case, where \( \sigma_{x,f}, \sigma_{x,\Phi} > 0 \), will be developed in Section 4.3.) We start with analyzing the convergence properties of Algorithm 1S.
Theorem 4.1. Assume that $\sigma_{x,t} = \sigma_{y,\Phi} = \sigma_{y,0} = 0$ and $B_{\|\|}^{0,1} \subseteq \text{dom} \ h_X$. In addition, let Assumption 3.1(B) hold. In Algorithm 1S, fix any $T \geq \max \left\{ 3.64 \sqrt{(L/\mu)} \Omega_{h,x}, 1024(L_{xx}/\mu) \Omega_{h,x}, 4096L_{yx}(\mu R)^{-1} \sqrt{\Omega_{h,x} \Omega_{h,y}}, 8192L_{yy}(\mu R^2)^{-1} \Omega_{h,y} \right\}$. (4.10)

If we choose $\eta_x = 16\Omega_{h,y}$, $\eta_y = \Omega_{h,x}'$, $\{\beta_t\}_{t \in \mathbb{N}}$ and $\{\theta_t\}_{t \in \mathbb{N}}$ as in (3.8), and $\alpha_t = \alpha$ and $\tau_t = \tau$ for any $t \in [T]$, where

$$\alpha = 1 / (16(L_{yx}^T + L_{yy}^T)),$$  

$$\tau = 1 / (4L' + 2(L_{xx}^T + L_{yx}^T)T),$$  

then

$$G(\bar{\mathbf{f}}^T, \bar{\mathbf{y}}^T) \leq B^\text{det}_R(T) \triangleq \frac{16L^2R^2}{(T - 1)T - 1} \Omega_{h,x}' + \frac{8L_{xx}^2R^2}{T - 1} \Omega_{h,x}$$

$$+ \frac{8L_{yx}R}{T - 1} \left( \sqrt{\frac{\eta_x}{\eta_y}} \Omega_{h,x}' + 16 \sqrt{\frac{\eta_y}{\eta_x}} \Omega_{h,y} \right) + \frac{128L_{yy}}{T} \Omega_{h,y} \leq \mu R^2 / 16,$$  

and $\|\mathbf{x}^T - \mathbf{x}\| \leq \sqrt{(2/\mu)B^\text{det}_R(T)} \leq R/(2\sqrt{2})$. (4.12)

Proof. From the choices of $\{\beta_t\}_{t \in \mathbb{N}}$, $\{\alpha_t\}_{t \in \mathbb{N}}$, $\{\tau_t\}_{t \in \mathbb{N}}$ and $\{\theta_t\}_{t \in \mathbb{N}}$, we can easily verify that the conditions (3.17) to (3.19) in Proposition 3.1 continue to hold with $\gamma_t = t$, for any $t \in \mathbb{N}$ and $L$, $L_{xx}$, $L_{yx}$, and $L_{yy}$ being replaced by $L'$, $L_{xx}'$, $L_{yx}'$ and $L_{yy}'$. In particular, $\gamma_t / \tau_t = \tau^{-1}$, for any $t \in \mathbb{N}$ and $\gamma_0 / \tau_0 = 0$. Therefore, by substituting the parameter choices in Theorem 4.1, (3.51) now becomes

$$\frac{T(T - 1)}{2} \bar{G}(\bar{\mathbf{x}}^T, \bar{\mathbf{y}}^T; x, y) \geq \frac{2}{T} D_{h,x}(x^1, y^1)(x, x^1) + \frac{4(T - 1) \eta_y}{\alpha} \Omega_{h,y}, \ \forall T \geq 3.$$  

(4.13)

Next, recall from (2.4) that $\tilde{S}(x) = f(x) + g(x) + \max_{y \in \mathcal{Y}} \Phi(x, y) - J(y)$ and define

$$\tilde{S}_{x^1,R}(y) \triangleq \min_{x \in \mathcal{X} \cap B_{\|\|}^{1, R}} S(x, y), \quad x_{x^1,R}(y) \triangleq \arg \min_{x \in \mathcal{X} \cap B_{\|\|}^{1, R}} S(x, y).$$  

(4.14)

Note that since $f$ is $\mu$-s.c. (w.r.t. $\|\|\|)$ on $\mathcal{X}$, the same holds for $\tilde{S}$. We then take supremum over $x \in \mathcal{X} \cap B_{\|\|}^{1, R}$ and $y \in \mathcal{Y}$ on both sides of (4.13) and define

$$\hat{G}_R(\bar{\mathbf{x}}^T, \bar{\mathbf{y}}^T) \triangleq \sup_{x \in \mathcal{X} \cap B_{\|\|}^{1, R}, y \in \mathcal{Y}} \hat{G}(\bar{\mathbf{x}}^T, \bar{\mathbf{y}}^T; x, y) = \tilde{S}(\bar{\mathbf{x}}^T) - \tilde{S}_{x^1,R}(\bar{\mathbf{y}}^T).$$  

(4.15)

By the choices of $\alpha$ and $\tau$ in (4.11), we have that for any $T \geq 3$,

$$\hat{G}_R(\bar{\mathbf{x}}^T, \bar{\mathbf{y}}^T) \leq \frac{16L^2R^2}{T(T - 1)T - 1} \Omega_{h,x}' + \frac{8L_{xx}^2R^2}{T - 1} \Omega_{h,x}$$

$$+ \frac{8L_{yx}R}{T - 1} \left( \sqrt{\frac{\eta_x}{\eta_y}} \Omega_{h,x}' + 16 \sqrt{\frac{\eta_y}{\eta_x}} \Omega_{h,y} \right) + \frac{128L_{yy}}{T} \Omega_{h,y}.$$  

(4.16)

On the other hand, by the $\mu$-strong convexity of $\tilde{S}$, $\mathbf{x}^1$ is the unique minimizer of $\tilde{S}$ on $\mathcal{X}$, and

$$\|\mathbf{x}^T - \mathbf{x}^1\|^2 \leq \frac{2}{\mu} (\tilde{S}(\mathbf{x}^T) - \tilde{S}(\mathbf{x}^1)) \leq \frac{2}{\mu} (\tilde{S}(\mathbf{x}^T) - \tilde{S}_{x^1,R}(\mathbf{y}^T)) = \frac{2}{\mu} \hat{G}_R(\mathbf{x}^T, \mathbf{y}^T),$$  

(4.17)
where (a) follows from $\tilde{S}(x^1) = S(x^1, y^1) \geq \min_{x \in \mathcal{X} \cap B^{z_1}_R} S(x, y) = \hat{S}_{z_1, R}(\tilde{y}^T)$ as $x^1 \in \mathcal{X} \cap B^{z_1}_R$.

Next, we aim to show that $\hat{G}_R(\bar{x}^T, \tilde{y}^T) = G(\bar{x}^T, \tilde{y}^T)$. To start, suppose that

$$
\|x^*_{z_1, R}(\tilde{y}^T) - x^1\| \leq R/(2\sqrt{2}).
$$

(4.18)

By the input condition $\|x^1 - x^\| \leq R/2$, we have $\|x^*_{z_1, R}(\tilde{y}^T) - x^1\| < R$. In other words, $x^*_{z_1, R}(\tilde{y}^T) \in \mathcal{X} \cap \text{int } B^{z_1}_R$. Note that by its definition, there exists $d \in \partial_\mathcal{X} S(x^*_{z_1, R}(\tilde{y}^T), \tilde{y}^T)$

$$
\langle d, x - x^*_{z_1, R}(\tilde{y}^T) \rangle \geq 0, \quad \forall x \in \mathcal{X} \cap B^{z_1}_R.
$$

(4.19)

Since $x^*_{z_1, R}(\tilde{y}^T) \in \text{int } B^{z_1}_R$, for any $x \in \mathcal{X} \setminus B^{z_1}_R$, there exists $\lambda \in (0, 1)$ such that $\bar{x} \triangleq \lambda x + (1 - \lambda)x^*_{z_1, R}(\tilde{y}^T) \in B^{z_1}_R$. Moreover, $\bar{x} \in \mathcal{X}$ since $\mathcal{X}$ is convex. Thus $\bar{x} \in \mathcal{X} \cap B^{z_1}_R$ and

$$
\langle d, \bar{x} - x^*_{z_1, R}(\tilde{y}^T) \rangle \geq 0.
$$

(4.20)

On the other hand, we have $\bar{x} - x^*_{z_1, R}(\tilde{y}^T) = \lambda(x - x^*_{z_1, R}(\tilde{y}^T))$. Consequently,

$$
\langle d, x - x^*_{z_1, R}(\tilde{y}^T) \rangle \geq 0, \quad \forall x \in \mathcal{X} \setminus B^{z_1}_R.
$$

(4.21)

Combining (4.19) and (4.21), we conclude that $x^*_{z_1, R}(\tilde{y}^T) = \arg \min_{x \in \mathcal{X}} S(x, \tilde{y}^T)$ and hence $\hat{G}_R(\bar{x}^T, \tilde{y}^T) = G(\bar{x}^T, \tilde{y}^T)$.

Therefore, it remains to show (4.18). First, since $S(\cdot, \tilde{y}^T)$ is $\mu$-s.c. on $\mathcal{X}$, by the definition of $x^*_{z_1, R}(\tilde{y}^T)$ and the fact that $x^1 \in \mathcal{X} \cap B^{z_1}_R$, we have

$$
S(x^1, \tilde{y}^T) - \hat{S}_{z_1, R}(\tilde{y}^T) = S(x^1, \tilde{y}^T) - S(x^*_{z_1, R}(\tilde{y}^T), \tilde{y}^T) \geq \frac{\mu}{2} \|x^*_{z_1, R}(\tilde{y}^T) - x^1\|^2.
$$

(4.22)

On the other hand, $S(x^1, \tilde{y}^T) \leq \max_{y \in \mathcal{Y}} S(x^1, y) = \tilde{S}(x^1) \leq \bar{S}(\bar{x}^T)$, since $x^1$ minimizes $\bar{S}$ on $\mathcal{X}$. Thus

$$
\hat{G}_R(\bar{x}^T, \tilde{y}^T) = \bar{S}(\bar{x}^T) - \hat{S}_{z_1, R}(\tilde{y}^T) \geq \frac{\mu}{2} \|x^*_{z_1, R}(\tilde{y}^T) - x^1\|^2.
$$

(4.23)

However, note that from (4.16) and the choice of $T$ in (4.10), we have $\hat{G}_R(\bar{x}^T, \tilde{y}^T) \leq \mu R^2/16$. We hence complete the proof. \(\blacksquare\)

From Theorem 4.1, we observe that $\|x^1 - x^1\| \leq R/2$ and $\|\bar{x}^T - x^1\| \leq R/(2\sqrt{2})$. This suggests that if Algorithm 1S is used as the subroutine in a restart scheme, then at each stage, the radius $R$ will be reduced by a factor of $\sqrt{2}$, and accordingly, the duality gap ($\leq \mu R^2/16$) will be halved. This observation naturally leads us to the restart scheme in Algorithm 2, which comprises $K$ stages. At each stage $k$, using the output primal variable $\bar{x}_{k-1}^{T_{k-1}}$ from the last stage as the input, we run Algorithm 1S for a sufficiently large number of iterations (i.e., $T_k$ iterations), so as to ensure the output primal variable $\bar{x}_k^{T_k}$ in the current stage satisfies $\|\bar{x}_k^{T_k} - x^1\| \leq R_k/\sqrt{2}$. 
Algorithm 2 Deterministic restart scheme

**Input:** Diameter estimate $U_X \geq D_X \triangleq \sup_{x,x' \in X} \|x-x'\|$, starting primal variable $x_0 \in X^o$, desired accuracy $\epsilon > 0$, $K = \lceil \max \{0, \log_2 (\mu U_X^2 / (4\epsilon)) \} \rceil + 1$

**Initialize:** $R_1 = 2U_X$, $x_1 = x_0$

For $k = 1, \ldots, K$

1. $T_k := \left\lceil \max \left\{ 3, 64 \sqrt{(L/\mu)\Omega_{h_X}}, 1024(L_{xx}/\mu)\Omega'_{h_X}, 4096L_{yx}(\mu R)_{-1}^{-1} \sqrt{\Omega'_{h_X} \Omega_{h_Y}}, 8192L_{yy}(\mu R)_{-1}^{-1} \Omega_{h_Y} \right\} \right\rceil$. \hspace{1cm} (4.24)

2. Run Algorithm 1S for $T_k$ iterations with starting primal variable $x_k$, radius $R_k$, constraint set $X_k \equiv X$ and other input parameters set as in Theorem 4.1. Denote the output as $(\tilde{x}^k, \tilde{y}^k)$.

3. $R_{k+1} := R_k / \sqrt{2}$, $x_{k+1} := \tilde{x}^k_{\Theta}$

**Output:** $(x_{K+1}, y_{K+1})$

**Theorem 4.2.** In Algorithm 2, for any $x_0 \in X^o$ and any accuracy $\epsilon \in (0, \mu U_X^2 / 4]$, we have $G(x_{K+1}, y_{K+1}) \leq \epsilon$ and the total number of oracle calls, i.e.,

$$C_{\epsilon}^{\text{det}} = \sum_{k=1}^{K} T_k \leq \left( 3 + 64 \sqrt{(L/\mu)\Omega_{h_X}} + 1024(L_{xx}/\mu)\Omega'_{h_X} \right) \left( \lceil \log_2 (\mu U_X^2 / (4\epsilon)) \rceil + 1 \right) + 8192(L_{yx}/\sqrt{\mu'}) \Omega'_{h_X} \Omega_{h_Y} + 2048(L_{yy}/\epsilon) \Omega_{h_Y}. \hspace{1cm} (4.25)$$

**Proof.** Note that $R_k = 2^{(3-k)/2} U_X$, for any $k \in [K]$. Therefore,

$$G(x_{K+1}, y_{K+1}) = G(\tilde{x}^K, \tilde{y}^K) \overset{(a)}{=} G(\tilde{x}^K, \tilde{y}^K) \leq \mu R_{K}^2 / 16 = \mu U_X^2 2^{-(K+1)} \overset{(b)}{=} \epsilon, \hspace{1cm} (4.26)$$

where (a) follows from (4.12) and (b) follows from $K \geq \log_2 (\mu U_X^2 / (4\epsilon)) + 1$. In addition, we can also substitute the value of $R_k$ into (4.24) to obtain (4.25). \hspace{1cm} \square

**Remark 4.1.** By restricting $\epsilon \in (0, \mu U_X^2 / 4]$, we see that $\max \{0, \log_2 (\mu U_X^2 / (4\epsilon)) \}$ simply becomes $\log_2 (\mu U_X^2 / (4\epsilon))$. By doing so, we have indeed simplified the bound in (4.25) (as compared to the bound derived for $\epsilon > 0$). On the other hand, note that if $\epsilon > \mu U_X^2 / 4$, then $K = 1$. As a result, Algorithm 2 degenerates to Algorithm 1S. For the same reason, we will also focus on analyzing the regime $\epsilon \in (0, \mu U_X^2 / 4]$ in the stochastic restart scheme (see Theorem 4.4).

Theorem 4.2 indicates that to obtain an $\epsilon$-duality gap, the oracle complexity is

$$O \left( \sqrt{\frac{L}{\mu} \log \left( \frac{1}{\epsilon} \right)} + \frac{L_{xx}}{\mu} \log \left( \frac{1}{\epsilon} \right) + \frac{L_{yx}}{\sqrt{\mu} \epsilon} + \frac{L_{yy}}{\epsilon} \right). \hspace{1cm} (4.27)$$

### 4.3. Stochastic restart scheme.

We next consider the general case where we only have access to the stochastic gradients of $f$, $\Phi(\cdot, y)$ and $\Phi(x, \cdot)$. 

4.3.1. **Intuition.** At the first attempt, one may try to combine the techniques used in proving Theorem 3.1(A) and Theorem 4.1, so as to analyze the convergence of $E[G(x^T, y^T)]$ in Algorithm 1S. However, a close inspection shows that such a combination does not work. Indeed, with this combination, although one can ensure $E[\hat{G}_R(x^T, y^T)] \leq \mu R^2/16$ and hence $E[\|x^*_1 - x^1\|] \leq R/(2\sqrt{2})$ (cf. (4.23)) by choosing $T$ properly, these conditions cannot guarantee that $E[\hat{G}_R(x^T, y^T)] = E[G(x^T, y^T)]$. This is because it is unclear how to ensure that $x^*_1 \in X \cap \text{int} B_{\| \cdot \|}^{x^1, R}$ holds in the “expectation” sense, in contrast to the deterministic setting, where this condition holds a.s. With that said, when Assumption 3.2(C) holds (i.e., the gradient noises are “light-tailed”), it is possible to show that $\hat{G}_R(x^T, y^T) \leq \mu R^2/16$ (and hence all the rest steps in the proof of Theorem 4.1, including $G(x^T, y^T) \leq \mu R^2/16$) holds with high probability (cf. Theorem 3.1(B)). Moreover, if we “concatenate” this result in our restart scheme, we will still end up with a high-probability bound on the duality gap, as long as we keep the number of stages “reasonably” small.

4.3.2. **Algorithmic details.** To start with, note that under the geometry in Section 4.3, Assumption 3.2(C) now becomes

\[
E_{t-1}\left[\exp\left(\frac{(\|\delta_{y,\phi}\|^2 (\sigma'_{y,\phi})^2)}{2}\right)\right] \leq \exp(1) \ a.s., \tag{4.28}
\]

\[
E_{t-1}\left[\exp\left(\frac{(\|\delta_{x,\phi}\|^2 (\sigma'_{x,\phi})^2)}{2}\right)\right] \leq \exp(1) \ a.s., \tag{4.29}
\]

\[
E_{t-1}\left[\exp\left(\frac{(\|\delta_{x,f}\|^2 (\sigma'_{x,f})^2)}{2}\right)\right] \leq \exp(1) \ a.s., \tag{4.30}
\]

where $\sigma'_{y,\phi} \triangleq \sigma_{y,\phi}/(\sqrt{\delta_y})$, $\sigma'_{x,\phi} \triangleq (R/(\sqrt{\delta_x})\sigma_{x,\phi}$ and $\sigma'_{x,f} \triangleq (R/(\sqrt{\delta_x})\sigma_{x,f}$. Based on these definitions, we state the convergence results of Algorithm 1S in the stochastic setting.

**Theorem 4.3.** Assume that $B_{\| \cdot \|}^{0,1} \subseteq \text{dom } h_{\chi}$. Also, let Assumptions 3.1(B), 3.2(A) and (C) hold. In Algorithm 1S, choose $X'$ such that $D_{X'} \supseteq \{x, \alpha \in X' | \|x - x'\| \leq R \}$. Fix any $\varsigma \in (0, 1/6]$ and

\[
T \geq \left\lceil \max \left\{3, 64\sqrt{(L/\mu)} \Omega_{h_X}', 2048(L_{xx}/\mu) \Omega_{h_X}', 4096L_{yx}(\mu R)^{-1} \Omega_{h_X}' \Omega_{h_Y}, 128^2L_{yy}(\mu R^2)^{-1} \Omega_{h_Y}, \right. \right.
\]

\[
512^2(\sigma_{x,f} + \sigma_{x,\phi})^2(\mu R)^{-2}(4\sqrt{(1 + \log(1/\varsigma)) \Omega_{h_X}' \Omega_{h_Y}} + 2\sqrt{\log(1/\varsigma)})^2, \right. \right.
\]

\[
512^2\sigma_{y,\phi}^2(\mu R^2)^{-2}(8\sqrt{(1 + \log(1/\varsigma)) \Omega_{h_Y} + 2\sqrt{\log(1/\varsigma)} \Omega_{h_Y}^2})^2 \right\} \right\}. \tag{4.31}
\]

If we choose $\eta_x = 16\Omega_{h_Y}$, $\eta_y = \Omega_{h_X}'$, $\{\beta_t\}_{t \in N}$ and $\{\theta_t\}_{t \in N}$ as in (3.8), and $\alpha_t = \alpha$ and $\tau_t = \tau$ for any $t \in [T]$, where

\[
\alpha = 1/(16(L_{yy} + \rho' \sigma_{y,\phi} \sqrt{T})), \quad \rho = (1/4)\sqrt{(1 + \log(1/\varsigma))/(2\Omega_{h_X}' \Omega_{h_Y})},
\]

\[
\tau = 1/(4L' + 2(L_{xx}' + L_{yx}'))T + \rho(\sigma_{x,\phi} + \sigma_{x,f})T^{3/2}), \quad \rho' = (1/8)\sqrt{(1 + \log(1/\varsigma))/(\Omega_{h_X}' \Omega_{h_Y})}, \tag{4.32}
\]
then w.p. at least $1 - 6\varsigma$,

$$G(\mathbf{x}^T, \mathbf{y}^T) \leq B_{R}^{\text{det}}(T) + B_{R}^{\text{var}}(T) \leq \mu R^2/16, \quad (4.33)$$

where $B_{R}^{\text{det}}(T)$ is defined in (4.12) and

$$B_{R}^{\text{var}}(T) \triangleq \frac{4(\sigma_x + \sigma_y)R}{\sqrt{T}} \{4\sqrt{(1 + \log(1/\varsigma))\Omega_{h_X}} + 2\sqrt{\log(1/\varsigma)}\}$$

$$\quad + \frac{4\sigma_y R}{\sqrt{T}} \{8\sqrt{2(1 + \log(1/\varsigma))\Omega_{h_Y}} + 2\sqrt{\log(1/\varsigma)\Omega_{h_Y}}\}. \quad (4.34)$$

Furthermore, $\|\mathbf{x}^T - x^\dagger\| \leq \sqrt{(2/\mu)(B_{R}^{\text{det}}(T) + B_{R}^{\text{var}}(T))} \leq R/(2\sqrt{2})$ w.p. at least $1 - 6\varsigma$.

**Proof.** Following the proof of Theorem 3.1(B), by the assumption that $D_{X'} \leq R$ and the choices of $\{\beta_t\}_{t \in \mathbb{N}}, \{\alpha_t\}_{t \in \mathbb{N}}, \{\tau_t\}_{t \in \mathbb{N}}$ and $\{\theta_t\}_{t \in \mathbb{N}}$ in Theorem 4.3, we can easily show that

$$\hat{G}_R(\mathbf{x}^T, \mathbf{y}^T) \leq B_{R}^{\text{det}}(T) + B_{R}^{\text{var}}(T), \quad (4.35)$$

w.p. at least $1 - 6\varsigma$. Denote the event in (4.35) as $A_T$. From now on, let us condition on $A_T$. By the choice of $T$ in (4.31), we have w.p. 1 that $\hat{G}_R(\mathbf{x}^T, \mathbf{y}^T) \leq \mu R^2/16$. Consequently, from (4.17) and (4.23), we see that $\|\mathbf{x}^T - x^\dagger\| \leq R/(2\sqrt{2})$ and $\|x^*_{1,R}(\mathbf{y}^T) - x^\dagger\| \leq R/(2\sqrt{2})$ respectively w.p. 1. As a result, we can conclude that $x^*_{1,R}(\mathbf{y}^T) \in \mathcal{X} \cap \text{int} B_{\|\cdot\|}^{x^\dagger}$ and hence $\hat{G}_R(\mathbf{x}^T, \mathbf{y}^T) = G(\mathbf{x}^T, \mathbf{y}^T)$. We then complete the proof. □

Based on Theorem 4.3, we present our stochastic restart scheme in Algorithm 2S. Compared to the deterministic restart scheme (i.e., Algorithm 2), a notable difference is that at each stage $k$, the constraint set $\mathcal{X}'$ becomes $\mathcal{X} \cap B_{\|\cdot\|}^{x^\dagger,R_k/\sqrt{2}}$ (as opposed to $\mathcal{X}$ as in Algorithm 2). As will be shown in Theorem 4.4, this step enables us to obtain the nearly optimal $O(\log \log (1/\epsilon) / \epsilon)$ complexity on the primal noise term $\sigma_x + \sigma_y$. To discuss the computational issue of the step (4.9) in Algorithm 1S, suppose $\mathcal{X}$ is a Hilbert space and take $h_X = (1/2)\|\cdot\|^2_X$. If $g \equiv 0$ and $\mathcal{X} = \mathcal{X}$, then (4.9) becomes an orthogonal projection onto $B_{\|\cdot\|}^{x^\dagger,R_k/\sqrt{2}}$, which clearly admits closed-form solution. If not, we can consider the Lagrangian form of this problem, i.e.,

$$\min_{x \in \mathcal{X}} \max_{\lambda \geq 0} g(x) + \langle \pi^t, x - x^t \rangle + \frac{\eta_x}{2\tau_k R_k^2} \|x - x^t\|^2 + \lambda(\|x - x_k\|^2 - R_k^2/4), \quad (4.36)$$

where $\pi^t \triangleq \nabla_x \Phi(x^t, y^{t+1}, \zeta^t) + \nabla f(\mathbf{x}^t, \xi^t)$. We observe that for any fixed $\lambda \geq 0$, the minimization problem in (4.36) has the same form as that in (2.3), hence has closed-form solution. Since $\lambda$ is a scalar, we can solve the maximization problem in (4.36) efficiently (e.g., via bisection).

For the subsequent analysis, in Algorithm 2S, define the event

$$E_1 = \Omega, \quad E_k \triangleq \{ \hat{G}_{R_{k-1}}(x_k, y_k) \leq \mu R_{k-1}^2/16 \}, \quad \forall k = 2, \ldots, K + 1, \quad (4.37)$$
Algorithm 2S Stochastic restart scheme

**Input:** Diameter estimate $U_X \geq D_X$, starting primal variable $x_0 \in X^o$, desired accuracy $\epsilon > 0$, error probability $\nu \in (0, 1]$, $K = \lceil \max \{0, \log_2 (3\mu U_X^2/(4\epsilon))\} \rceil + 1$, $\varsigma = \nu/(6K)$

**Initialize:** $R_1 = 2U_X$, $x_1 = x_0$, $y_0 \in Y^o$

For $k = 1, \ldots, K$

1. $T_k := \max \left\{3, 64\sqrt{(L/\mu)\Omega_{h_X}^y}, 2048(L_{xx}/\mu)\Omega_{h_X}^y, 4096L_{xy}(\mu R_K)^{-1}\Omega_{h_X}^y, 128^2L_{yy}(\mu R_K)^{-1}\Omega_{h_y}^y, 512^2(\sigma_{x,f} + \sigma_{x,\Phi})^2(\mu R_K)^{-2}(4\sqrt{(1 + \log(1/\varsigma))\Omega_{h_X}^y} + 2\sqrt{\log(1/\varsigma)\Omega_{h_y}^y})^2, 512^2\sigma_{y,\Phi}^2(\mu R_K)^{-2}(8\sqrt{2(1 + \log(1/\varsigma))\Omega_{h_y}^y} + 2\sqrt{\log(1/\varsigma)\Omega_{h_y}^y})^2 \right\}$. \hspace{1cm} (4.39)

2. Run Algorithm 1S for $T_k$ iterations with starting primal variable $x_k$, radius $R_k$, constraint set $\mathcal{X}_k = \{x \in \mathcal{X} : \|x - x_k\| \leq R_k/2\}$ and other input parameters set as in Theorem 4.3. Denote the output as $(\pi_k^1, \pi_k^2)$.

3. $R_{k+1} := R_k / \sqrt{2}$, $x_{k+1} := \pi_k^1$.

**Output:** $(x_{K+1}, y_{K+1})$

where $\Omega$ denotes the underlying sample space for the stochastic process $\{(x_k, y_k)\}_{k=1}^{K+1}$. For any $k \geq 2$, by conditioning on $\mathcal{E}_k$, from the proof of Theorem 4.3, we see that both $\|x_k - x^\dagger\| \leq R_k/2$ (since $R_k = R_{k-1}/\sqrt{2}$) w.p. 1 and

$$G(x_k, y_k) = \tilde{G}_{R_{k-1}}(x_k, y_k) \leq \mu R_{k-1}^2/16 \hspace{0.5cm} \text{w.p. 1.} \hspace{1cm} (4.38)$$

Therefore, given $\mathcal{E}_k$, $\mathcal{X}_k$ satisfies all the requirements stated in Algorithm 1S and Theorem 4.3, i.e., $x^\dagger \in \mathcal{X}_k \subseteq \mathcal{X}$ and $D_{\mathcal{X}_k} \leq R_k$. (Note that when $k = 1$, we have $\mathcal{X}_1 = \mathcal{X}$, thus these requirements are clearly satisfied.) As a result, Theorem 4.3 can be applied to the $k$-th stage. Based on this observation, we derive the oracle complexity of Algorithm 2S below.

**Theorem 4.4.** Assume that $\mathcal{B}^0 \cap \|\| \subseteq \text{dom}\ h_X$. Also, let Assumptions 3.1(B), 3.2(A) and (C) hold. In Algorithm 2S, for any $x_0 \in X^o$, accuracy $\epsilon \in (0, \mu U_X^2/4]$ and error probability $\nu \in (0, 1]$, we have $G(x_{K+1}, y_{K+1}) \leq \epsilon$ w.p. at least $1 - \nu$. The total number of oracle calls to achieve this, i.e.,

$$C^\text{est}_\epsilon \leq \left(3 + 64\sqrt{(L/\mu)\Omega_{h_X}^y} + 2048(L_{xx}/\mu)\Omega_{h_X}^y \right) \left(\lceil \log_2 (\mu U_X^2/(4\epsilon)) \rceil + 1 \right) + 256^2L_{yy}/\sqrt{\epsilon \mu \Omega_{h_X}^y} \Omega_{h_Y}^y + 64^2L_{xy}/\epsilon \Omega_{h_Y}^y + 1024^2 \left\{ (\sigma_{x,f} + \sigma_{x,\Phi})^2 / (\epsilon \mu) \right\} \left\{ 4\Omega_{h_X}^y + 1 \right\} \log \left(6 \lceil \log_2 (\mu U_X^2/(4\epsilon)^{-1}) \rceil + 2 \right) / \nu + 4\Omega_{h_X}^y \right) + 1024^2 \left( \sigma_{y,\Phi}^2 / \epsilon^2 \right) \left\{ 1 + \log \left(6 \lceil \log_2 (\mu U_X^2/(4\epsilon)^{-1}) \rceil + 2 \right) / \nu \right\} \Omega_{h_Y}^y. \hspace{1cm} (4.40)$$
Proof. For any \( k = 1, \ldots, K + 1 \), let \( \mathbb{I}_{E_k} \) denote the indicator function of the event \( E_k \) (cf. (4.37)). It is clear that \( \{\mathbb{I}_{E_k}\}_{k=1}^{K+1} \) forms a (finite-horizon) Markov chain, and therefore
\[
\Pr\{\bigcap_{k=1}^{K+1} E_k\} = \Pr\{\mathbb{I}_{E_2} = 1, \forall k = 2, \ldots, K + 1\} \\
= \mathbb{P}\{\mathbb{I}_{E_2} = 1\} \prod_{k=3}^{K+1} \Pr\{\mathbb{I}_{E_k} = 1 | \mathbb{I}_{E_{k-1}} = 1\} \\
\overset{(a)}{\geq} (1 - 6\varsigma)^K \overset{(b)}{\geq} 1 - 6K\varsigma \overset{(c)}{=} 1 - \nu, \quad (4.41)
\]
where (a) follows from Theorem 4.3, (b) follows from Bernoulli’s inequality and (c) follows from the choice of \( \varsigma \) in Algorithm 2S. By the choice of \( \{R_k\}_{k=1}^K \) in Algorithm 2S, we have
\[
R_k = 2^{(3-k)/2}U_X, \forall k \in [K]. \quad (4.42)
\]
Therefore, from (4.26), we know that \( \mu R_k^2/16 \leq \epsilon \). As a result,
\[
\Pr\{G(x_{K+1}, y_{K+1}) \leq \epsilon\} \geq \Pr\{G(x_{K+1}, y_{K+1}) \leq \mu R_k^2/16\} \\
\geq \Pr\{G(x_{K+1}, y_{K+1}) \leq \mu R_k^2/16 | E_{K+1}\} \Pr\{E_{K+1}\} \\
\overset{(a)}{=} \Pr\{E_{K+1}\} \geq \Pr\{\bigcap_{k=1}^{K+1} E_k\} \geq 1 - \nu, \quad (4.43)
\]
where (a) follows from (4.38). From the choice of \( \varsigma \) and \( K \), we have
\[
\log(1/\varsigma) = \log(6K/\nu) \leq \log \left(6 \left[ \log_2 (\mu U_X^2 (4\epsilon)^{-1}) + 2 \right] / \nu \right). \quad (4.44)
\]
Now, we substitute both (4.42) and (4.44) into the choice of \( T_k \) in (4.39), and then obtain the oracle complexity in (4.40). \( \square \)

From Theorem 4.4, we see that in order to obtain an \( \epsilon \)-duality gap w.p. at least \( 1 - \nu \), the oracle complexity is
\[
O\left(\left(\sqrt{\frac{L}{\mu}} + \frac{L_{xx}}{\mu}\right) \log \left(\frac{1}{\epsilon}\right) + \frac{L_{yx}}{\sqrt{\mu\epsilon}} + \frac{L_{yy}}{\epsilon} + \left(\frac{\sigma_{x,f} + \sigma_{x,\Phi}}{\mu\epsilon} + \frac{\sigma_{y,\Phi}^2}{\epsilon^2}\right) \log \left(\frac{\log(1/\epsilon)}{\nu}\right)\right). \quad (4.45)
\]

4.3.3. Complexity of convergence in expectation. Based on the result in Theorem 4.4, we aim to analyze the oracle complexity to obtain an \( \epsilon \)-expected duality gap. To do so, we need some additional assumptions on the nonsmooth functions \( g \) and \( J \).

Assumption 4.1. The nonsmooth functions \( g \) and \( J \) have closed domains, and are continuous on \( X \) and \( Y \) respectively.

Remark 4.2. Note that the closed-domain assumption is satisfied by most of the nonsmooth functions, e.g., any indicator function of a closed convex set and any absolutely homogeneous function (e.g., the norm function). Regarding the continuity assumption, since \( g \) (resp. \( J \)) is convex, it suffices for \( X \) (resp. \( Y \)) to reside in \( \text{int dom} \ g \) (resp. \( \text{int dom} \ J \)).
To see the implication of Assumption 4.1, let us write down the explicit form of the primal and dual functions $\bar{S}$ and $\underline{S}$ in (2.4):

$$\bar{S}(x) = f(x) + g(x) + \left\{ \max_{y \in Y \cap \text{dom} J} \Phi(x, y) - J(y) \right\}, \quad (4.46)$$

$$\underline{S}(y) = \left\{ \min_{x \in X \cap \text{dom} g} f(x) + g(x) + \Phi(x, y) \right\} - J(y). \quad (4.47)$$

In (4.46), by the closedness of $\text{dom} J$ and the compactness of $Y$ (cf. Assumption 3.1(B)), we see that $Y \cap \text{dom} J$ is compact, on which $J$ is continuous. Therefore, we can invoke Berge’s maximum theorem (see Ok [29, Section E.3]) to conclude that the function in \{\cdot\} is continuous on $X$. Combined with the continuity of $g$ on $X$, we have $\bar{S}$ is continuous on $X$. Similarly, we can show that $\underline{S}$ is continuous on $Y$. As a result, there exists a positive constant $\Gamma < +\infty$ such that

$$\sup_{(x,y) \in X \times Y} G(x, y) = \sup_{y' \in Y} \bar{S}(x, y') - \inf_{x' \in X} \underline{S}(x', y) \leq \Gamma. \quad (4.48)$$

Based on this observation, we can easily arrive at the following result.

**Theorem 4.5.** Assume that $B^{0,1}_{\| \cdot \|} \subseteq \text{dom} h_K$. In addition, let Assumptions 3.1(B), 3.2(A), (C), and 4.1 hold. In Algorithm 2S, for any $x_0 \in X^o$ and accuracy $\epsilon \in (0, \mu U^2_K/2]$, if we choose $K = \lceil \log_2 (\mu U^2_K/(2\epsilon)) \rceil + 1$ and $\nu = \min \{\epsilon/(2\Gamma), 1\}$, then $\mathbb{E}[G(x_{K+1}, y_{K+1})] \leq \epsilon$. Moreover, the oracle complexity to achieve this is

$$O\left( \left( \sqrt{\frac{L}{\mu}} + \frac{L_{xy}}{\mu} \right) \log \left( \frac{1}{\epsilon} \right) + \frac{L_{xy}}{\sqrt{\mu \epsilon}} + \frac{L_{yy}}{\epsilon} + \left( \frac{(\sigma_x + \sigma_y) \mu}{\epsilon^2} \right) \log \left( \frac{1}{\epsilon} \right) \right). \quad (4.49)$$

**Proof.** Define the event $\mathcal{G}_{K,\epsilon} \triangleq \{G(x_{K+1}, y_{K+1}) \leq \epsilon/2\}$ and denote its complement as $\bar{\mathcal{G}}_{K,\epsilon}^c$. From Theorem 4.4, we know that by the choice of $K$, $\Pr\{\mathcal{G}_{K,\epsilon}\} \geq 1 - \nu$, for any $\nu \in (0, 1]$. Therefore,

$$\mathbb{E}[G(x_{K+1}, y_{K+1})] = \mathbb{E}[G(x_{K+1}, y_{K+1})I_{\mathcal{G}_{K,\epsilon}}] + \mathbb{E}[G(x_{K+1}, y_{K+1})I_{\bar{\mathcal{G}}_{K,\epsilon}^c}]$$

$$\leq (\epsilon/2) \Pr\{\mathcal{G}_{K,\epsilon}\} + \Gamma \Pr\{\bar{\mathcal{G}}_{K,\epsilon}^c\} \leq \epsilon/2 + \Gamma \epsilon/(2\Gamma) \leq \epsilon, \quad (4.50)$$

where (a) follows from (4.48) and (b) follows from the choice of $\nu$ (which implies that $\nu \leq \epsilon/(2\Gamma)$).

To derive the oracle complexity in (4.49), we simply substitute $\epsilon = \epsilon/2$ and $\nu = \epsilon/(2\Gamma)$ into (4.45), and note that $O(\log(\log(1/\epsilon)/\epsilon)) = O(\log(1/\epsilon))$. □

4.4. **Discussions.**

We conclude this section with discussions on several technical issues.
4.4.1. Restart techniques. Note that our restart scheme, designed for stochastic SPPs, greatly differs from the usual approaches for convex minimization problems (see e.g., Nesterov [28], Ghadimi and Lan [11] and in particular, Renegar and Grimmer [32] for a comprehensive review). For illustration, consider \( \min_{x \in \mathcal{X}} f(x) \), where \( f \) is \( L \)-smooth and \( \mu \)-s.c. on \( \mathcal{X} \) (cf. Section 1). Denote its minimizer by \( x^* \). Almost all the existing approaches use the (expected) objective error as the reduction criterion. Specifically, they require the subroutine to satisfy that for any \( \epsilon, \delta > 0 \), there exists \( T(\epsilon, \delta) \in \mathbb{N} \) such that

\[
\mathbb{E}[\|x^1 - x^*\|^2] \leq \delta \quad \implies \quad \mathbb{E}[f(\bar{x}^{T(\epsilon, \delta)}) - f(x^*)] < \epsilon, \tag{4.51}
\]

where \( x^1 \) denotes the starting point and \( \bar{x}^{T(\epsilon, \delta)} \) denotes the (weighted) average of \( \{x^t\}_{t=1}^{T(\epsilon, \delta)} \). By the strong convexity of \( f \), we can bound

\[
\mathbb{E}[\|x^1 - x^*\|^2] \leq \frac{2}{\mu} \mathbb{E}[f(x^1) - f(x^*)] \tag{4.52}
\]

and thus establish a recursion (indeed, contraction by properly choosing \( T(\epsilon, \delta) \)) between \( \mathbb{E}[f(\bar{x}^{T(\epsilon, \delta)}) - f(x^*)] \) and \( \mathbb{E}[f(x^1) - f(x^*)] \). However, somewhat interestingly, for (stochastic) SPPs, we cannot use the seemingly analogous (expected) duality gap as the reduction criterion. The reasons are twofold. First, by the definition of the duality gap in (2.5), we do not have the analog of (4.51), i.e., for any \( \epsilon, \delta > 0 \), there exists \( T^*(\epsilon, \delta) \in \mathbb{N} \) such that

\[
\max \left\{ \mathbb{E}[\|x^1 - x^\dagger\|^2], \mathbb{E}[\|y^1 - y^\dagger\|^2] \right\} \leq \delta \quad \implies \quad \mathbb{E}[G(\bar{x}^{T^*(\epsilon, \delta)}, \bar{y}^{T^*(\epsilon, \delta)})] < \epsilon, \tag{4.53}
\]

where recall that \( (x^\dagger, y^\dagger) \) denotes any saddle point of (1.1). Second, even if (4.53) holds, since we do not assume strong convexity on the dual function \( \mathcal{S} \), there is no analog of (4.52) on the dual variable \( y^1 \). Therefore, we still cannot establish a recursion between \( \mathbb{E}[G(\bar{x}^{T^*(\epsilon, \delta)}, \bar{y}^{T^*(\epsilon, \delta)})] \) and \( \mathbb{E}[G(x^1, y^1)] \). Therefore, more sophisticated restart schemes need to be designed.

Inspired by Juditsky and Nemirovski [16], in the deterministic setting, we use the quantity \( R \) as the reduction criterion (cf. Section 4.2). The quantity \( R \) not only serves an upper bound on the distance between the primal iterate \( \bar{x}^T \) to \( x^1 \), but also on the duality gap \( G(\bar{x}^T, \bar{y}^T) \) (cf. Theorem 4.1). Thus by sufficiently reducing \( R \) (possibly over multiple stages), we can drive the duality gap below any desired accuracy \( \epsilon > 0 \). To achieve this, we need to design the subroutine Algorithm 1S, whose convergence bounds depend on \( R \), based on Algorithm 1. As a side note, since Algorithm 1S is based on PDHG, it is very different from the one in Juditsky and Nemirovski [16], which is based on Mirror-Prox.

In the stochastic setting, the situation becomes even more challenging. As discussed above (see also Section 4.3.1), the restart scheme based on expectation (of the duality gap or the iterates) fails
to work. Therefore, we innovatively design the restart scheme as a Markov chain and accordingly, analyze the oracle complexity to drive the duality gap below \( \epsilon \) with high probability. In addition, if \( G \) is bounded on \( \mathcal{X} \times \mathcal{Y} \), we can also derive the oracle complexity under which the expected duality gap falls below \( \epsilon \).

### 4.4.2. “Light-tailed” noises.

In our stochastic restart scheme, we have assumed that the gradient noises \( \{\delta^t_{y,\Phi}\}_{t \in \mathbb{N}}, \{\delta^t_{x,\Phi}\}_{t \in \mathbb{N}} \) not only have uniformly bounded (conditional) second moments, but also follow sub-Gaussian distributions. The reason is that sub-Gaussianity leads to concentration, so that we have large deviation results at each stage (cf. Theorem 4.3). Moreover, we still have such type of results over multiple stages (cf. Theorem 4.4). However, if the noises are “heavy-tailed”, following the same approach, the oracle complexity in (4.45) will have a poor dependence on \( \nu \) (e.g., \( 1/\nu \)). This in turn causes the complexity in (4.49) to have a worse dependence on \( \epsilon \), since in Theorem 4.5 we need to choose \( \nu = \Theta(\epsilon) \). It is interesting to develop new methodology that can effectively deal with “heavy-tailed” noises, and we leave this to future work.

### 4.4.3. Complexity results.

It is instructive to compare the oracle complexity in (4.45) with that in (3.13). We observe that when \( \mu > 0 \), the oracle complexity for \( L_{xx}, L_{yx} \) and \( \sigma_{x,f} + \sigma_{x,\Phi} \) has been significantly improved. (Indeed, the complexity is optimal for \( L_{xx} \) and \( L_{yx} \) and nearly optimal for \( \sigma_{x,f} + \sigma_{x,\Phi} \); see Section 1.4 for details.) In addition, we notice that the complexity for \( L_{yy} \) and \( \sigma_{y,\Phi} \) remains the same (modulo the log log(1/\( \epsilon \)) factor). This is rather intuitive: since we have no strong convexity on the dual side, the complexity for the terms involving only the dual variables cannot be improved in general. Similar observations and reasonings also apply when we compare the complexity in (4.49) with that in (3.12).

In terms of optimality, as introduced in Section 1.4, the complexities of \( L_{xx}, L_{yx}, \sigma_{x,f} + \sigma_{x,\Phi} \) and \( \sigma_{y,\Phi} \) in 4.49 are either optimal or nearly optimal. However, it is unknown whether the complexities of \( L_{xx} \) and \( L_{yy} \) are optimal (the same also applies to the complexity result in (3.12)). Therefore, as future work, we aim to derive lower bounds for the complexities of \( L_{xx} \) and \( L_{yy} \), and at the same time, design accelerated algorithms (if any) that can improve the existing complexity results (for \( L_{xx} \) and \( L_{yy} \)).

Additionally, we observe that there exist two artifacts in our complexity results. The first one is the additional log log(1/\( \epsilon \)) factor associated with the noise terms (i.e., \( \sigma_{x,f} + \sigma_{x,\Phi} \) and \( \sigma_{y,\Phi} \)) in (4.45) (resp. the log(1/\( \epsilon \)) factor in (4.49)). The second one is Assumption 4.1, which together with Assumption 3.1(B), ensures the boundedness of \( G \) on \( \mathcal{X} \times \mathcal{Y} \). Although these artifacts are rather moderate, we believe they can be removed by a more intelligent and careful design of our stochastic restart scheme. We defer this to future work.
Appendix A: Unique output of the Bregman proximal projection

**Lemma A.1.** For the minimization problem in (2.3), if $\varphi^* \triangleq \inf_{u \in \mathcal{U}} \varphi(u) > -\infty$ and $U \cap \text{dom } h \cap \text{dom } \varphi \neq \emptyset$, then a unique solution exists in $U^\circ \cap \text{dom } \varphi$.

**Proof.** We first prove the existence of solutions. By condition (2.1) and $\varphi^* > -\infty$, we see that $P_\lambda$ is coercive on $U$, i.e., $\lim_{\|u\| \to +\infty} P_\lambda(u) = +\infty$. In addition, $P_\lambda$ is CCP since both $\varphi$ and $h$ are CCP. Choose any point $u \in U \cap \text{dom } P_\lambda \neq \emptyset$ (note that $\text{dom } P_\lambda = \text{dom } \varphi \cap \text{dom } h$) and any $\alpha \geq P_\lambda(u)$. The closedness and coercivity of $P_\lambda$ together imply that the sub-level set $S_\alpha(P_\lambda) \triangleq \{u \in U : P_\lambda(u) \leq \alpha\}$ is compact. Since $U$ is closed, $U \cap S_\alpha(P_\lambda)$ is compact and nonempty. This, together with the closedness of $P_\lambda$, implies that the solution set of Problem (2.3), denoted by $U_{\text{opt}}$, is nonempty and contained in $U \cap \text{dom } h \cap \text{dom } \varphi$. Since $h$ is essentially smooth, then for any $u \in \text{dom } h \setminus \text{int dom } h$, $\partial h(u) = \emptyset$ hence $\partial P_\lambda(u) = \emptyset$. By Bauschke et al. [2, Theorem 4.1], $U_{\text{opt}} \subseteq \text{int dom } h \cap U \cap \text{dom } \varphi = U^\circ \cap \text{dom } \varphi$. Since $h$ is strongly convex on $U^\circ$, so is $P_\lambda$ and hence $U_{\text{opt}}$ must be a singleton. □

**References**

[1] Balamurugan P, Bach F (2016) Stochastic variance reduction methods for saddle-point problems. *Proc. NIPS*.

[2] Bauschke HH, Borwein JM, Combettes PL (2001) Essential smoothness, essential strict convexity, and legendre functions in Banach spaces. *Commun. Contemp. Math.* 3(4):615–647.

[3] Bertsekas DP (1999) *Nonlinear Programming* (Athena Scitific).

[4] Chambolle A, Pock T (2011) A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.* 40(1):120–145.

[5] Chambolle A, Pock T (2016) On the ergodic convergence rates of a first-order primal–dual algorithm. *Math. Program.* 159(1):253–287.

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

[7] Chen Y, Lan G, Ouyang Y (2017) Accelerated schemes for a class of variational inequalities. *Math. Program.* 165(1):113–149.

[8] Condat L (2013) A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *J. Optim. Theory Appl.* 158(2):460–479.

[9] Davis D (2015) Convergence rate analysis of primal-dual splitting schemes. *SIAM J. Optim.* 25(3):1912–1943.

[10] Ghadimi S, Lan G (2012) Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM J. Optim.* 22(4):1469–1492.

[11] Ghadimi S, Lan G (2013) Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization II: Shrinking procedures and optimal algorithms. *SIAM J. Optim.* 23(4):2061–2089.
[12] Hamedani EY, Aybat NS (2018) A primal-dual algorithm for general convex-concave saddle point problems. arXiv:1803.01401.

[13] He Y, Monteiro RDC (2016) An accelerated HPE-type algorithm for a class of composite convex-concave saddle-point problems. *SIAM J. Optim.* 26(1):29–56.

[14] Hien LTK, Zhao R, Haskell WB (2017) An inexact primal-dual smoothing framework for large-scale non-bilinear saddle point problems. arXiv:1711.03669.

[15] Juditsky A, Nemirovski A (2012) First-order methods for nonsmooth convex large-scale optimization, I: General purpose methods. *Optimization for Machine Learning*, 121–148 (MIT Press).

[16] Juditsky A, Nemirovski A (2012) First-order methods for nonsmooth convex large-scale optimization, II: Utilizing problems structure. *Optimization for Machine Learning*, 149–184 (MIT Press).

[17] Juditsky A, Nemirovski A, Tauvel C (2011) Solving variational inequalities with stochastic mirror-prox algorithm. *Stoch. Syst.* 1(1):17–58.

[18] Kolossoski O, Monteiro R (2017) An accelerated non-euclidean hybrid proximal extragradient-type algorithm for convexconcave saddle-point problems. *Optim. Methods Softw.* 32(6):1244–1272.

[19] Lan G (2012) An optimal method for stochastic composite optimization. *Math. Program.* 133(1-2):365–397.

[20] Monteiro R, Svaiter B (2011) Complexity of variants of tseng’s modified f-b splitting and korpelevich’s methods for hemivariational inequalities with applications to saddle-point and convex optimization problems. *SIAM J. Optim.* 21(4):1688–1720.

[21] Nedić A, Ozdaglar A (2009) Subgradient methods for saddle-point problems. *J. Optim. Theory Appl.* 142(1):205–228.

[22] Nemirovski A (2005) Prox-method with rate of convergence $O(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. Optim.* 15(1):229–251.

[23] Nemirovski A, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* 19(4):1574–1609.

[24] Nemirovskii A, Yudin D (1979) Efficient methods for large-scale convex problems. *Ekonomika i Matematicheskie Metody (in Russian)* 15:135–152.

[25] Nesterov Y (2005) Excessive gap technique in nonsmooth convex minimization. *SIAM J. Optim.* 16(1):235–249.

[26] Nesterov Y (2005) Smooth minimization of non-smooth functions. *Math. Program.* 103(1):127–152.

[27] Nesterov Y (2009) Primal-dual subgradient methods for convex problems. *Math. Program.* 120(1):221–259.
[28] Nesterov Y (2013) Gradient methods for minimizing composite functions. *Math. Program.* 140(1):125–161.

[29] Ok EA (2007) *Real Analysis with Economic Applications* (Princeton University Press).

[30] Ouyang Y, Xu Y (2018) Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. arXiv:1808.02901.

[31] Raginsky M, Rakhlin A (2011) Information-based complexity, feedback and dynamics in convex programming. *IEEE Trans. Inf. Theory* 57(10):7036–7056.

[32] Renegar J, Grimmer B (2018) A simple nearly-optimal restart scheme for speeding-up first order methods. arXiv:1803.00151.

[33] Rigollet P, Hutter JC (2018) Lecture notes in high dimensional statistics.

[34] Robbins H, Monro S (1951) A stochastic approximation method. *Ann. Math. Statist.* 22(3):400–407.

[35] Vù BC (2013) A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Adv. Comput. Math.* 38(3):667–681.

[36] Zhao R, Cevher V (2018) Stochastic three-composite convex minimization with a linear operator. *Proc. AISTATS* (Lanzarote, Spain).

[37] Zhao R, Haskell WB, Tan VYF (2019) An optimal algorithm for stochastic three-composite optimization. *Proc. AISTATS* (Okinawa, Japan).