Hybrid Variance-Reduced SGD Algorithms For Nonconvex-Concave Minimax Problems

Quoc Tran-Dinh* · Deyi Liu* · Lam M. Nguyen†

*Department of Statistics and Operations Research
The University of North Carolina at Chapel Hill
Chapel Hill, NC 27599
Email: quoctd@email.unc.edu, deyi@live.unc.edu

†IBM Research, Thomas J. Watson Research Center, NY10598
Email: lamnguyen.mltd@ibm.com

June 30, 2020

Abstract

We develop a novel variance-reduced algorithm to solve a stochastic nonconvex-concave minimax problem which has various applications in different fields. This problem has several computational challenges due to its nonsmoothness, nonconvexity, nonlinearity, and non-separability of the objective functions. Our approach relies on a novel combination of recent ideas, including smoothing and hybrid stochastic variance-reduced techniques. Our algorithm and its variants can achieve $\mathcal{O}(T^{-2/3})$-convergence rate in T, and the best-known oracle complexity under standard assumptions. They have several computational advantages compared to existing methods. They can also work with both single sample or mini-batch on derivative estimators, with constant or diminishing step-sizes. We demonstrate the benefits of our algorithms over existing methods through two numerical examples.

1 Introduction

We study the following stochastic nonconvex-concave saddle-point problem, which covers various practical problems in different fields, see, e.g., [4, 9, 11]:

$$\min_{x\in\mathbb{R}^p}\max_{y\in\mathbb{R}^n}\big\{\Psi(x,y) := \mathcal{R}(x) + \mathbb{E}_\xi\big[\langle Ky, \mathbf{F}(x,\xi)\rangle\big] - \psi(y)\big\}, \tag{1}$$

where $\mathbf{F}:\mathbb{R}^p\times\Omega\to\mathbb{R}^q$ is a stochastic vector function defined on a probability space $(\Omega,\mathbb{P})$, $K\in\mathbb{R}^{q\times n}$ is a given matrix, $\langle\cdot,\cdot\rangle$ is an inner product, and $\psi:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ and $\mathcal{R}:\mathbb{R}^p\to\mathbb{R}\cup\{+\infty\}$ are proper, closed, and convex functions [3]. Problem (1) is a special case of the nonconvex-concave minimax problem, where our bifunction $\mathcal{H}(x,y) = \mathbb{E}_\xi\big[\langle Ky, \mathbf{F}(x,\xi)\rangle\big]$ is linear in y.

Note that (1) can be reformulated into a general stochastic compositional non-convex problem:

$$\min_{x\in\mathbb{R}^p}\Big\{\Psi_0(x) := \phi_0(F(x)) + \mathcal{R}(x) \equiv \phi_0\left(\mathbb{E}_\xi\big[\mathbf{F}(x,\xi)\big]\right) + \mathcal{R}(x)\Big\}, \tag{2}$$

where ϕ_0 is a convex, but possibly nonsmooth function, defined as

$$\phi_0(u) := \max_{y\in\mathbb{R}^n}\Big\{\langle K^\top u, y\rangle - \psi(y)\Big\} \equiv \psi^*(K^\top u), \tag{3}$$

with ψ^* being the Fenchel conjugate of ψ [3]. Note that problem (2) is completely different from existing models such as [7, 8], where the expectation is inside the outer function ϕ_0, i.e., $\phi_0\left(\mathbb{E}_\xi\left[\mathbf{F}(x,\xi)\right]\right)$. We refer to this setting as a "non-separable" model. The template (3) also covers penalized formulations of a stochastic constrained optimization problem.

Challenges. Developing numerical methods for solving (1) or (2) faces several challenges. First, it is often nonconvex, i.e., F is not affine. Many recent papers consider special cases of (2) when Ψ_0 in (2) is convex by imposing restrictive conditions, which are unfortunately not realistic in applications. Second, the max-form ϕ_0 in (3) is often non-smooth if ψ is not strongly convex. This prevents the use of gradient-based methods. Third, since the expectation is inside ϕ_0, it is very challenging to form an unbiased estimate for [sub]gradients of Φ_0, making classical stochastic gradient-based methods inapplicable. Finally, prox-linear operator-based methods as in [7, 8, 28, 39] require large mini-batch evaluations of both function value $\mathbf{F}$ and its Jacobian $\mathbf{F}'$, see [28, 37, 39], instead of single sample or small mini-batch, making them less flexible and more expensive than gradient-based methods.

Related work. Problem (1) has recently attracted considerable attention due to key applications, e.g., in game theory, robust optimization, and generative adversarial nets (GANs) [4, 9, 11]. Various first-order methods have been developed to solve (1) during the past decades for both convex-concave models , e.g., [3, 13, 19, 30] and nonconvex-concave settings [16, 24]. Some recent works consider a nonnonvex-nonconcave formulation, e.g., [23, 34]. However, they still rely on additional assumptions to guarantee that the maximization problem in (3) can globally be solved. One well-known assumption is the Polyak-Łojasiewicz (PL) condition, which is rather strong and often used to guarantee linear convergence. A majority of these works focus on deterministic models, while some methods have been extended to stochastic settings, e.g., [16, 34]. Although (1) is a special case of a general model in [16, 34], it almost covers all examples in [16, 34]. Compared to these, our algorithm is rather simple with a single loop, and our oracle complexity is significantly improved over the ones in [16, 34].

Alternatively, the compositional reformulation (2) has been broadly studied in the literature under both deterministic and stochastic settings, see, e.g., [7, 8, 14, 21, 27, 31]. If $q = 1$ and $\phi_0(u) = u$, then (2) reduces to the standard stochastic optimization model studied e.g., in [10, 25]. In the deterministic setting, one common method to solve (2) is the prox-linear-type method, which is also known as a Gauss-Newton method [14, 21]. This method has been studied in several papers, including [7, 8, 14, 21, 27]. However, the prox-linear operator often does not have a closed form expression, and its evaluation may require solving a general nonsmooth strongly convex subproblem.

In the stochastic setting as (2), [31, 32] proposed stochastic compositional gradient methods to solve more general forms than (2), but they required a set of stronger assumptions than Assumptions 2.1-2.2 below, including the smoothness of ϕ_0. Recent related works include [15, 17, 33, 35, 36], which also rely on similar ideas. For instance, [16] proposed a double loop subgradient-based method with $\mathcal{O}\left(\varepsilon^{-6}\right)$ oracle complexity. Another subgradient-based method was recently proposed in [34] based on a two-side PL condition. Stochastic methods exploiting prox-linear operators have also been recently proposed in [28, 39], which are essentially extensions of existing deterministic methods to (2). Together with algorithms, convergence guarantees, stochastic oracle complexity bounds have also been estimated. For instance, [31] obtained $\mathcal{O}\left(\varepsilon^{-8}\right)$ oracle complexity for (2), while it was improved to $\mathcal{O}\left(\varepsilon^{-4.5}\right)$ in [32]. Recent works [37, 38] further improved the complexity to $\mathcal{O}\left(\varepsilon^{-3}\right)$. These methods require the smoothness of both ϕ_0 and F, use large batch sizes, and need a double loop scheme. In contrast, ***our method has single loop, can work with either single sample or mini-batch, and allows both constant or diminishing step-sizes***. For nonsmooth ϕ_0, under the same assumptions as [28, 39], our methods achieve $\mathcal{O}\left(\varepsilon^{-3}\right)$ Jacobian and $\mathcal{O}\left(\varepsilon^{-5}\right)$ function evaluation complexity as in those papers. However, our method is gradient-based, which only uses

proximal operator of ψ and $\mathcal{R}$ instead of a complex prox-linear operator as in [28, 39]. Moreover, it can work with both single sample and mini-batch for Jacobian $\mathbf{F}'$ compared to a large batch size as in [28, 39].

Our contribution. Our main contribution in this paper can be summarized as follows:

(a) We develop a new single-loop hybrid variance-reduced SGD algorithm to solve (1) under Assumptions 2.1 and 2.2 below. Under the strong convexity of ψ, our algorithm has $\mathcal{O}\left((bT)^{-2/3}\right)$ convergence rate to approximate a KKT (Karush-Kuhn-Tucker) point of (1), where b is the batch size and T is the iteration counter. We also estimate its $\mathcal{O}\left(\varepsilon^{-3}\right)$-oracle complexity to obtain an ε-KKT point, matching the best-known one as, e.g., in [37, 38]. Our complexity bound holds for a wide range of b as opposed to a specific choice in [37, 38].

(b) When ψ is non-strongly convex, we combine our approach with a smoothing technique to develop a gradient-based variant, that can achieve the best-known $\mathcal{O}\left(\varepsilon^{-3}\right)$ Jacobian and $\mathcal{O}\left(\varepsilon^{-5}\right)$ function evaluations of $\mathbf{F}$ for finding an ε-KKT point of (1). Moreover, our algorithm does not require prox-linear operators and large batches for Jacobian as in [28, 39].

(c) We also propose a simple restarting technique without sacrificing convergence guarantees to accelerate the practical performance of both cases (a) and (b).

Our methods exploit a recent biased hybrid estimators introduced in [29] as opposed to SARAH ones in [28, 37, 39]. This allows us to simplify our algorithm with a single loop and without large batches at each iteration compared to [37]. As indicated in [2], our $\mathcal{O}\left(\varepsilon^{-3}\right)$ oracle complexity is optimal under the considered assumptions. If ψ is non-strongly convex (i.e. ϕ_0 in (2) can be nonsmooth), then our algorithm is fundamentally different from the ones in [28, 39] as it does not use prox-linear operator. Note that evaluating a prox-linear operator requires to solve a general strongly convex but possible nonsmooth subproblem. In addition, they only work with large batch sizes of both $\mathbf{F}$ and $\mathbf{F}'$.

Content. The rest of this paper is organized as follows. Section 2 states our assumptions and recalls some mathematical tools. Section 3 develops a new algorithm and analyzes its convergence. Section 5 provides two numerical examples to compare our methods. All technical details and proofs are deferred to the appendices.

2 Basic assumptions, KKT points and smoothing technique

Notation. We work with finite-dimensional space $\mathbb{R}^p$ equipped with standard inner product $\langle\cdot,\cdot\rangle$ and Euclidean norm $\|\cdot\|$. For a function $\phi : \mathbb{R}^p \to \mathbb{R}\cup\{+\infty\}$, $\mathrm{dom}(\phi)$ denotes its domain. If ϕ is convex, then prox_ϕ denotes its proximal operator, $\partial\phi$ denotes its subdifferential, and $\nabla\phi$ is its [sub]gradient, see, e.g., [3]. ϕ is μ_ϕ-strongly convex with a strongly convex parameter $\mu_\phi > 0$ if $\phi(\cdot) - \frac{\mu_\phi}{2}\|\cdot\|^2$ remains convex. For a smooth vector function $F : \mathbb{R}^p \to \mathbb{R}^q$, F' denotes its Jacobian. We use $\mathrm{dist}\left(x, \mathcal{X}\right) := \inf_{y\in\mathcal{X}}\|x - y\|$ to denote the Euclidean distance from x to a convex set $\mathcal{X}$.

2.1 Model assumptions

Let $F(x) := \mathbb{E}_\xi\left[\mathbf{F}(x,\xi)\right]$ denote the expectation function of $\mathbf{F}$ and $\mathrm{dom}(\Psi_0)$ denote the domain of Ψ_0. Throughout this paper, we always assume that

$$\Psi_0^\star := \inf_{x\in\mathbb{R}^p}\left\{\Psi_0(x) := \phi_0(F(x)) + \mathcal{R}(x)\right\} > -\infty$$

in (2) and $\mathcal{R}$ is proper, closed, and convex without recalling them in the sequel. Our goal is to develop stochastic gradient-based algorithms to solve (1) relying on the following assumptions:

Assumption 2.1. The function $\mathbf{F}$ in problem (1) or (2) satisfies the following assumptions:

(a) ***Smoothness:*** $\mathbf{F}(\cdot,\cdot)$ is L_F-average smooth with $L_F \in (0,+\infty)$, i.e.:

$$\mathbb{E}_\xi\big[\,\big\|\mathbf{F}'(x,\xi)-\mathbf{F}'(y,\xi)\big\|^2\,\big] \le L_F^2\,\|x-y\|^2\,,\quad \forall x,y\in\mathrm{dom}(\Psi_0). \tag{4}$$

(b) ***Bounded variance:*** There exists two constants $\sigma_F,\sigma_J\in(0,+\infty)$ such that

$$\mathbb{E}_\xi\big[\,\|\mathbf{F}(x,\xi)-F(x)\|^2\,\big]\le\sigma_F^2 \quad\text{and}\quad \mathbb{E}_\xi\big[\,\big\|\mathbf{F}'(x,\xi)-F'(x)\big\|^2\,\big]\le\sigma_J^2,\quad \forall x\in\mathrm{dom}(\Psi_0).$$

(c) ***Lipschitz continuity:*** $F(\cdot)$ is M_F-average Lipschitz continuous with $M_F\in(0,+\infty)$, i.e.:

$$\mathbb{E}_\xi\big[\,\big\|\mathbf{F}'(x,\xi)\big\|^2\,\big]\le M_F^2,\quad \forall x\in\mathrm{dom}(\Psi_0). \tag{5}$$

Note that Assumptions 2.1 are standard in stochastic nonconvex optimization, see [28, 37, 38, 39]. If $\mathrm{dom}(\mathcal{R})$ is bounded, then $\mathrm{dom}(\Psi_0)$ is bounded, and this assumption automatically holds.

For ψ, we only require the following assumption, which is mild and holds for many applications.

Assumption 2.2. The function ψ in (1) is proper, closed, and convex. Moreover, $\mathrm{dom}(\psi)$ is bounded by $M_\psi\in(0,+\infty)$, i.e.: $\sup\{\|y\| : y\in\mathrm{dom}(\psi)\}\le M_\psi$.

An important special case of ψ is the indicator of convex and bounded sets. Hitherto, we do not require ϕ_0 and $\mathcal{R}$ in (2) to be smooth or strongly convex. They can be nonsmooth so that (2) can also cover constrained problems. Note that the boundedness of $\mathrm{dom}(\psi)$ is equivalent to the Lipschitz continuity of ϕ_0 (Lemma A.1). Simple examples of ϕ_0 include norms and gauge functions.

2.2 KKT points and approximate KKT points

Since (1) is nonconvex-concave, a pair $(x^\star,y^\star)$ is said to be a KKT point of (1) if

$$0\in F'(x^\star)^\top Ky^\star+\partial\mathcal{R}(x^\star)\qquad\text{and}\qquad 0\in K^\top F(x^\star)-\partial\psi(y^\star). \tag{6}$$

From (6), we have $y^\star\in\partial\psi^*(K^\top F(x^\star))$. Substituting this $y^\star$ into the first expression, we get

$$0\in F'(x^\star)^\top\partial\phi_0(F(x^\star))+\partial\mathcal{R}(x^\star). \tag{7}$$

Here, we have used $K^\top\partial\psi^*(K^\top u)=\partial\phi_0(u)$, where ϕ_0 is given by (3) This inclusion shows that $x^\star$ is a stationary point of (2). In the convex-concave case, under mild assumptions, a KKT point is also a saddle-point of (1). In particular, if (2) is convex, then $x^\star$ is also its global optimum of (2).

However, in practice, we can only find an approximation $(\tilde{x}_0^*,\tilde{y}_0^*)$ of a KKT point $(x^\star,y^\star)$ for (1).

Definition 2.1. Given any tolerance $\varepsilon>0$, $(\tilde{x}_0^*,\tilde{y}_0^*)$ is called an ε-KKT point of (1) if

$$\begin{aligned}&\mathbb{E}\big[\mathcal{E}(\tilde{x}_0^*,\tilde{y}_0^*)\big]\le\varepsilon,\\ \text{where}\quad &\mathcal{E}(x,y)\;:=\mathrm{dist}\big(0,F'(x)^\top Ky+\partial\mathcal{R}(x)\big)+\mathrm{dist}\big(0,K^\top F(x)-\partial\psi(y)\big)\,.\end{aligned} \tag{8}$$

Here, the expectation is taken overall the randomness from both model (1) and the algorithm. Clearly, if $\mathbb{E}\big[\mathcal{E}(\tilde{x}_0^*,\tilde{y}_0^*)\big]=0$, then $(\tilde{x}_0^*,\tilde{y}_0^*)$ is a KKT point of (1) as characterized by (6).

2.3 Smoothing techniques

Under Assumption 2.2, ϕ_0 defined by (3) can be nonsmooth. Hence, we can smooth ϕ_0 as follows:

$$\phi_\gamma(u) := \max_{y\in\mathbb{R}^n} \left\{\langle u, Ky\rangle - \psi(y) - \gamma b(y)\right\}, \tag{9}$$

where $b : \mathrm{dom}(\psi) \to \mathbb{R}_+$ is a continuously differentiable and 1-strongly convex function such that $\min_y b(y) = 0$, and $\gamma > 0$ is a smoothness parameter. For example, we can choose $b(y) := \frac{1}{2}\|y - \dot{y}\|^2$ for a fixed $\dot{y}$ or $b(y) := \log(n) + \sum_{j=1}^n y_j \log(y_j)$ defined on a standard simplex Δ_n [20].

Let $y^*_\gamma(u)$ be an optimal solution of the maximization problem in (9), which always exists and is unique. In particular, if $b(y) := \frac{1}{2}\|y - \dot{y}\|^2$, then $y^*_\gamma(u) := \mathrm{prox}_{\psi/\gamma}\left(\dot{y} - \gamma^{-1}K^\top u\right)$. Under Assumption 2.2, ϕ_γ possesses some useful properties as stated in Lemma A.1 (Appendix A.1).

Given ϕ_γ defined by (9), we consider the following functions:

$$\Phi_\gamma(x) := \phi_\gamma(F(x)) = \phi_\gamma\left(\mathbb{E}_\xi\left[\mathbf{F}(x,\xi)\right]\right) \quad \text{and} \quad \Psi_\gamma(x) := \Phi_\gamma(x) + \mathcal{R}(x). \tag{10}$$

In this case, under Assumptions 2.1 and 2.2, Φ_γ is continuously differentiable, and

$$\nabla\Phi_\gamma(x) = F'(x)^\top \nabla\phi_\gamma(F(x)) = F'(x)^\top K y^*_\gamma(F(x)). \tag{11}$$

Smoothness. Moreover, $\Phi_\gamma(\cdot)$ is L_{Φ_γ}-smooth with $L_{\Phi_\gamma} := M_{\phi_\gamma}L_F + M_F^2 L_{\phi_\gamma}$ (see [38]), i.e.:

$$\|\nabla\Phi_\gamma(x) - \nabla\Phi_\gamma(\hat{x})\| \le L_{\Phi_\gamma}\|x - \hat{x}\|, \quad \forall x, \hat{x} \in \mathrm{dom}(\Psi_0), \tag{12}$$

where $M_{\phi_\gamma} := M_\psi\|K\|$ and $L_{\phi_\gamma} := \frac{\|K\|^2}{\gamma+\mu_\psi}$ given in Lemma A.1.

Gradient mapping. Let us recall the following gradient mapping of $\Psi_\gamma(\cdot)$ given in (10):

$$\mathcal{G}_\eta(x) := \tfrac{1}{\eta}\left(x - \mathrm{prox}_{\eta\mathcal{R}}(x - \nabla\Phi_\gamma(x))\right), \quad \text{for any } \eta > 0. \tag{13}$$

This mapping will be used to characterize approximate KKT points of (1) in Definition 2.1.

3 The proposed algorithm and its convergence analysis

First, we introduce a stochastic estimator for $\nabla\Phi_\gamma$. Then, we develop our main algorithm and analyze its convergence and oracle complexity. Finally, we show how to construct an ϵ-KKT point of (1).

3.1 Stochastic estimators and the algorithm

Since F is the expectation of a stochastic function $\mathbf{F}$, we exploit the hybrid stochastic estimators for F and its Jacobian F' proposed in [29]. More precisely, given a sequence $\{x_t\}$ generated by a stochastic algorithm, these hybrid stochastic estimators $\tilde{F}_t$ and $\tilde{J}_t$ are defined as follows:

$$\begin{cases} \tilde{F}_t &:= \beta_{t-1}\tilde{F}_{t-1} + \frac{\beta_{t-1}}{b_1}\sum_{\xi_i\in\mathcal{B}_t^1}\left[\mathbf{F}(x_t,\xi_i) - \mathbf{F}(x_{t-1},\xi_i)\right] + \frac{(1-\beta_{t-1})}{b_2}\sum_{\zeta_i\in\mathcal{B}_t^2}\mathbf{F}(x_t,\zeta_i) \\ \tilde{J}_t &:= \hat{\beta}_{t-1}\tilde{J}_{t-1} + \frac{\hat{\beta}_{t-1}}{\hat{b}_1}\sum_{\hat{\xi}_i\in\hat{\mathcal{B}}_t^1}\left[\mathbf{F}'(x_t,\hat{\xi}_i) - \mathbf{F}'(x_{t-1},\hat{\xi}_i)\right] + \frac{(1-\hat{\beta}_{t-1})}{\hat{b}_2}\sum_{\hat{\zeta}_i\in\hat{\mathcal{B}}_t^2}\mathbf{F}'(x_t,\hat{\zeta}_i), \end{cases} \tag{14}$$

where $\beta_{t-1}, \hat{\beta}_{t-1} \in [0,1]$ are given weights, and the initial estimators $\tilde{F}_0$ and $\tilde{J}_0$ are defined as

$$\tilde{F}_0 := \tfrac{1}{b_0}\textstyle\sum_{\xi_i\in\mathcal{B}^0}\mathbf{F}(x_0,\xi_i) \quad \text{and} \quad \tilde{J}_0 := \tfrac{1}{\hat{b}_0}\textstyle\sum_{\hat{\xi}_i\in\hat{\mathcal{B}}^0}\mathbf{F}'(x_0,\hat{\xi}_i). \tag{15}$$

Here, $\mathcal{B}^0$, $\hat{\mathcal{B}}^0$, $\mathcal{B}_t^1$, $\hat{\mathcal{B}}_t^1$, $\mathcal{B}_t^2$, and $\hat{\mathcal{B}}_t^2$ are mini-batches of sizes b_0, $\hat{b}_0$, b_1, $\hat{b}_1$, b_2, and $\hat{b}_2$, respectively. We also require that $\mathcal{B}_t^1$ is independent of $\mathcal{B}_t^2$, and $\hat{\mathcal{B}}_t^1$ is independent of $\hat{\mathcal{B}}_t^2$, but not between the others.

For $\tilde{F}_t$ and $\tilde{J}_t$ defined by (14), we introduce a stochastic estimator for the gradient $\nabla\Phi_\gamma(x_t) = F'(x_t)^\top\nabla\phi_\gamma(F(x_t))$ of $\Phi_\gamma(\cdot)$ in (10) at x_t as follows:

$$v_t := \tilde{J}_t^\top\nabla\phi_\gamma(\tilde{F}_t) \equiv \tilde{J}_t^\top K y_\gamma^*(\tilde{F}_t). \tag{16}$$

To evaluate v_t, we need to solve a strongly convex problem (9) to find $y_\gamma^*(\tilde{F}_t)$, which often cheaper than prox-linear operators. Moreover, due to (15) and (16), evaluating v_0 does not require the full matrix $\tilde{J}_0$, but a matrix-vector product $\tilde{J}_0^\top K y_\gamma^*(\tilde{F}_0)$, which is often cheaper than evaluating $\tilde{J}_0$.

Using the new estimator v_t of $\nabla\Phi_\gamma(x_t)$ in (16), we propose Algorithm 1 to solve (1).

Algorithm 1 (Smoothing Hybrid Variance-Reduced SGD Algorithm for solving (1))

Inputs: An arbitrarily initial point $x_0 \in \mathrm{dom}(\Psi_0)$.
Input $\beta_0, \hat{\beta}_0 \in (0,1)$, $\gamma_0 \geq 0$, $\eta_0 > 0$, and $\theta_0 \in (0,1]$ (specified later).
Initialization: Generate $\tilde{F}_0$ and $\tilde{J}_0$ as in (15) with mini-batch sizes b_0 and $\hat{b}_0$, respectively.
Solve (9) to obtain $y_{\gamma_0}^*(\tilde{F}_0)$. Then, evaluate $v_0 := \tilde{J}_0^\top K y_{\gamma_0}^*(\tilde{F}_0)$.
Update $\hat{x}_1 := \mathrm{prox}_{\eta_0\mathcal{R}}(x_0 - \eta_0 v_0)$ and $x_1 := (1-\theta_0)x_0 + \theta_0\hat{x}_1$.
For $t := 1, \cdots, T$ **do**
  Construct $\tilde{F}_t$ and $\tilde{J}_t$ as in (14) and $v_t := \tilde{J}_t^\top K y_{\gamma_t}^*(\tilde{F}_t)$, where $y_{\gamma_t}^*(\tilde{F}_t)$ solves (9).
  Update $\hat{x}_{t+1} := \mathrm{prox}_{\eta_t\mathcal{R}}(x_t - \eta_t v_t)$ and $x_{t+1} := (1-\theta_t)x_t + \theta_t\hat{x}_{t+1}$.
  Update $\beta_{t+1}, \hat{\beta}_{t+1}, \theta_{t+1} \in (0,1)$, $\eta_{t+1} > 0$, and $\gamma_{t+1} \geq 0$ if necessary.
EndFor
Output: Choose $\bar{x}_T$ randomly from $\{x_0, x_1, \cdots, x_T\}$ with $\mathbf{Prob}\{\bar{x}_T = x_t\} = \frac{\theta_t/L_{\Phi_{\gamma_t}}}{\sum_{t=0}^T \theta_t/L_{\Phi_{\gamma_t}}}$.

Algorithm 1 is designed by adopting the idea in [29], where it can start from two initial mini-batches $\mathcal{B}^0$ and $\hat{\mathcal{B}}^0$ to generate a good approximation for the search direction v_0 before getting into the main loop. However, it has 3 major differences compared to [29]: dual step $y_{\gamma_t}^*(\tilde{F}_t)$, estimator v_t, and dynamic parameter updates. Note that Algorithm 1 is single loop, making it easy to implement in practice compared to SVRG [12] and SARAH [22], but it requires one additional sample in (14). Moreover, if we use a diminishing step-size (see Theorems 3.2 and 3.4 below), then the initial mini-batches $\mathcal{B}^0$ and $\hat{\mathcal{B}}^0$ are not required.

3.2 Convergence analysis of Algorithm 1

Let $\mathcal{F}_t$ be the σ-field generated by Algorithm 1 up to the t-th iteration, which is defined as follows:

$$\mathcal{F}_t := \sigma\big(x_0, \mathcal{B}^0, \hat{\mathcal{B}}^0, \mathcal{B}_1^1, \hat{\mathcal{B}}_1^1, \mathcal{B}_1^2, \hat{\mathcal{B}}_1^2, \cdots, \mathcal{B}_t^1, \hat{\mathcal{B}}_t^1, \mathcal{B}_t^2, \hat{\mathcal{B}}_t^2\big). \tag{17}$$

If ψ is strongly convex, then, without loss of generality, we can assume $\mu_\psi := 1$. Otherwise, we can rescale it. Moreover, for the sake of our presentation, for a given $c_0 > 0$, we introduce:

$$\begin{aligned} P &:= \tfrac{\sqrt{26}\|K\|}{3\sqrt{c_0}}\sqrt{M_F^4\|K\|^2 + c_0L_F^2M_\psi^2}, \qquad Q := \tfrac{26}{9c_0}\|K\|^2\big(M_F^4\|K\|^2\sigma_F^2 + c_0M_\psi^2\sigma_J^2\big), \\ L_{\Phi_0} &:= L_FM_\psi\|K\| + M_F^2\|K\|^2, \qquad \text{and} \quad L_{\Phi_\gamma} := L_FM_\psi\|K\| + \tfrac{M_F^2\|K\|^2}{\gamma}, \end{aligned} \tag{18}$$

where $\gamma > 0$, M_F, L_F, σ_F, and σ_J are given in Assumption 2.1 and M_ψ is in Assumption 2.2.

(a) The smooth case. Theorem 3.1, whose proof is in Appendix B.3, analyzes convergence rate and complexity of Algorithm 1 for the smooth case of ϕ_0 in (2) (i.e., ψ is strongly convex).

Theorem 3.1 (**Constant step-size**). *Suppose that Assumptions 2.1 and 2.2 hold, ψ is μ_ψ-strongly convex with $\mu_\psi := 1$, and P, Q, and L_{Φ_0} are defined in* (18). *Given a mini-batch $0 < b \leq \hat{b}_0(T+1)$, let $b_0 := c_0\hat{b}_0$, $\hat{b}_1 = \hat{b}_2 := b$, and $b_1 = b_2 := c_0 b$. Let $\{x_t\}_{t=0}^T$ be generated by Algorithm 1 using*

$$\gamma_t := 0, \quad \beta_t = \hat{\beta}_t := 1 - \tfrac{b^{1/2}}{[\hat{b}_0(T+1)]^{1/2}}, \quad \theta_t = \theta := \tfrac{L_{\Phi_0} b^{3/4}}{P[\hat{b}_0(T+1)]^{1/4}}, \quad \text{and} \quad \eta_t = \eta := \tfrac{2}{L_{\Phi_0}(3+\theta)}, \tag{19}$$

provided that $\frac{\hat{b}_0(T+1)}{b^3} > \frac{L_{\Phi_0}^4}{P^4}$. Let $b_0 := c_1^2[b(T+1)]^{1/3}$ for some $c_1 > 0$. Then, we have

$$\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_T)\|^2\big] \leq \frac{\Delta_0}{[b(T+1)]^{2/3}}, \quad \text{where} \quad \Delta_0 := 16P\sqrt{c_1}\big[\Psi_0(x_0) - \Psi_0^\star\big] + \frac{24Q}{c_1}. \tag{20}$$

For a given tolerance $\varepsilon > 0$, the total number of iterations T to obtain $\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_T)\|^2\big] \leq \varepsilon^2$ is at most $T := \big\lfloor \frac{\Delta_0^{3/2}}{b\varepsilon^3} \big\rfloor$. The total numbers of function evaluation $\mathbf{F}(x_t, \xi)$ and its Jacobian evaluations $\mathbf{F}'(x_t, \xi)$ are at most $\mathcal{T}_F := \big\lfloor \frac{c_0 c_1^2 \Delta_0^{1/2}}{\varepsilon} + \frac{3c_0\Delta_0^{3/2}}{\varepsilon^3} \big\rfloor$ and $\mathcal{T}_J := \big\lfloor \frac{c_1^2\Delta_0^{1/2}}{\varepsilon} + \frac{3\Delta_0^{3/2}}{\varepsilon^3} \big\rfloor$, respectively.

Theorem 3.2 states convergence of Algorithm 1 using diminishing step-size (see Appendix. B.4).

Theorem 3.2 (**Diminishing step-size**). *Suppose that Assumptions 2.1 and 2.2 hold, ψ is μ_ψ-strongly convex with $\mu_\psi := 1$ (i.e., ϕ_0 in* (2) *is smooth). Let $\{x_t\}_{t=0}^T$ be generated by Algorithm 1 using the mini-batch sizes as in Theorem 3.1, and increasing weight and diminishing step-sizes as*

$$\gamma_t := 0, \quad \beta_t = \hat{\beta}_t := 1 - \tfrac{1}{(t+2)^{2/3}}, \quad \theta_t := \tfrac{L_{\Phi_0}\sqrt{b}}{P(t+2)^{1/3}}, \quad \text{and} \quad \eta_t := \tfrac{2}{L_{\Phi_0}(3+\theta_t)}, \tag{21}$$

Then, for all $T \geq 0$, and $(\bar{x}_T, \bar{\eta}_T)$ chosen as $\mathbf{Prob}\big\{\mathcal{G}_{\bar{\eta}_T}(\bar{x}_T) = \mathcal{G}_{\eta_t}(x_t)\big\} = \frac{\theta_t}{\sum_{t=0}^T \theta_t}$, we have

$$\mathbb{E}\big[\|\mathcal{G}_{\bar{\eta}_T}(\bar{x}_T)\|^2\big] \leq \tfrac{32P[\Psi_0(x_0)-\Psi_0^\star]}{3\sqrt{b}\big[(T+3)^{2/3}-2^{2/3}\big]} + \tfrac{32Q}{3\big[(T+3)^{2/3}-2^{2/3}\big]}\Big[\tfrac{2^{1/3}}{\hat{b}_0} + \tfrac{2(1+\log(T+1))}{b}\Big]. \tag{22}$$

If we set $b = \hat{b}_0 = 1$, then our convergence rate is $\mathcal{O}\left(\frac{\log(T)}{T^{2/3}}\right)$ with a $\log(T)$ factor slower than (20). However, it does not require a large initial mini-batch $\hat{b}_0$ as in Theorem 3.1. In Theorems 3.1 and 3.2, we do not need to smooth ϕ_0. Hence, γ_t is absent in Algorithm 1, i.e., $\gamma_t = 0$ for $t \geq 0$.

(b) The non-smooth case. Now we consider the case $\mu_\psi = 0$, i.e., ϕ_0 in (2) is non-smooth. Theorem 3.3 states convergence of Algorithm 1 in this case, whose proof is in Appendix B.5.

Theorem 3.3 (**Constant step-size**). *Assume that Assumptions 2.1 and 2.2 hold, ψ in* (1) *is non-strongly convex (i.e., ϕ_0 is nonsmooth), and P, Q, and L_{Φ_γ} are defined in* (18). *Let b and $\hat{b}_0$ be two positive integers, $c_0 > 0$, and $\{x_t\}_{t=0}^T$ be generated by Algorithm 1 after T iterations using:*

$$\begin{cases} \hat{b}_1 = \hat{b}_2 := b, \quad b_1 = b_2 := \frac{c_0 b}{\gamma^2}, \quad \hat{b}_0 := c_1^2[b(T+1)]^{1/3}, \quad b_0 := \frac{c_0\hat{b}_0}{\gamma^2}, \quad \gamma_t := \gamma \in (0, 1], \\ \beta_t = \hat{\beta}_t = 1 - \frac{b^{1/2}}{[\hat{b}_0(T+1)]^{1/2}}, \quad \theta_t = \theta := \frac{L_{\Phi_\gamma} b^{3/4}}{P[\hat{b}_0(T+1)]^{1/4}}, \quad \text{and} \quad \eta_t = \eta := \frac{2}{L_{\Phi_\gamma}(3+\theta)}. \end{cases} \tag{23}$$

Then, with B_ψ defined in Lemma A.1, the following bound holds

$$\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_T)\|^2\big] \leq \frac{\hat{\Delta}_0}{[b(T+1)]^{2/3}}, \quad \text{where} \quad \hat{\Delta}_0 := 16\sqrt{c_1}P\big(\Psi_0(x_0) - \Psi_0^\star + B_\psi\big) + \frac{24Q}{c_1}. \tag{24}$$

The total number of iterations T to achieve $\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_T)\|^2\big] \le \varepsilon^2$ is at most $T := \big\lfloor \frac{\hat{\Delta}_0^{3/2}}{b\varepsilon^3} \big\rfloor$. The total numbers of function evaluations $\mathcal{T}_F$ and Jacobian evaluations $\mathcal{T}_J$ are respectively at most

$$\mathcal{T}_F := \tfrac{c_0\hat{\Delta}_0^{1/2}}{\gamma^2\varepsilon} + \tfrac{3c_0\hat{\Delta}_0^{3/2}}{\gamma^2\varepsilon^3} = \mathcal{O}\Big(\tfrac{\hat{\Delta}_0^{3/2}}{\gamma^2\varepsilon^3}\Big) \quad \text{and} \quad \mathcal{T}_J := \tfrac{\hat{\Delta}_0^{1/2}}{\varepsilon} + \tfrac{3\hat{\Delta}_0^{3/2}}{\varepsilon^3} = \mathcal{O}\left(\tfrac{\hat{\Delta}_0^{1.5}}{\varepsilon^3}\right).$$

If we choose $\gamma := c_2\varepsilon$ for some $c_2 > 0$, then $\mathcal{T}_F = \frac{c_0\hat{\Delta}_0^{1/2}}{c_2^2\varepsilon^3} + \frac{3c_0\hat{\Delta}_0^{3/2}}{c_2^2\varepsilon^5} = \mathcal{O}\big(\frac{\hat{\Delta}_0^{3/2}}{\varepsilon^5}\big)$.

Note that both convergence rate (24) and $\mathcal{T}_J$ in Theorem 3.3 are independent of γ. The choice of $\gamma := c_2\varepsilon$ is to achieve an ϵ-KKT point in the sense of Definition 2.1 by using Lemma 3.1 below.

Alternatively, we can also establish convergence and estimate the complexity of Algorithm 1 with diminishing step-size in Theorem 3.4, whose proof is in Appendix B.6.

Theorem 3.4 (**Diminishing step-size**). *Suppose that Assumptions 2.1 and 2.2 hold, ψ is non-strongly convex (i.e., ϕ_0 is possibly nonsmooth), and P, Q, $L_{\Phi_{\gamma_t}}$ are defined by (18). Given mini-batch sizes $b > 0$ and $\hat{b}_0 > 0$, let $b_0 := \frac{c_0\hat{b}_0}{\gamma_0^2}$, $b_1^t = b_2^t := \frac{c_0 b}{\gamma_t^2}$, and $\hat{b}_1 = \hat{b}_2 := b$ for some $c_0 > 0$. Let $\{x_t\}_{t=0}^T$ be generated by Algorithm 1 using increasing weight and diminishing step-sizes as*

$$\gamma_t := \tfrac{1}{(t+2)^{1/3}}, \quad \beta_t = \hat{\beta}_t := 1 - \tfrac{1}{(t+2)^{2/3}}, \quad \theta_t := \tfrac{L_{\Phi_{\gamma_t}} b^{1/2}}{P(t+2)^{1/3}}, \quad \text{and} \quad \eta_t := \tfrac{2}{L_{\Phi_{\gamma_t}}(3+\theta_t)}. \tag{25}$$

For $(\bar{x}_T, \bar{\eta}_T)$ chosen as $\mathbf{Prob}\big\{\mathcal{G}_{\bar{\eta}_T}(\bar{x}_T) = \mathcal{G}_{\eta_t}(x_t)\big\} = \big[\sum_{t=0}^T(\theta_t/L_{\Phi_{\gamma_t}})\big]^{-1}(\theta_t/L_{\Phi_{\gamma_t}})$, we have

$$\begin{aligned}\mathbb{E}\big[\|\mathcal{G}_{\bar{\eta}_T}(\bar{x}_T)\|^2\big] \;\le\; & \tfrac{32P}{3\sqrt{b}[(T+3)^{2/3}-2^{2/3}]}\big(\Psi_0(x_0) - \Psi_0^\star + \tfrac{B_\psi}{(T+2)^{1/3}}\big) \\ & + \tfrac{16Q}{3[(T+3)^{2/3}-2^{2/3}]}\Big(\tfrac{2^{1/3}}{\hat{b}_0} + \tfrac{2(1+\log(T+1))}{b}\Big) = \mathcal{O}\left(\tfrac{\log(T)}{T^{2/3}}\right).\end{aligned} \tag{26}$$

Note that since $\gamma_t := \frac{1}{(t+2)^{1/3}}$ (diminishing) and $b_1^t = b_2^t := \frac{c_0 b}{\gamma_t^2}$, we have $b_1^t = b_2^t = c_0 b(t+2)^{2/3}$, which shows that the mini-batch sizes of the function estimation $\tilde{F}_t$ are chosen in increasing manner (not fixed at a large size for all t), which can save computational cost for F. The batch sizes b and $\hat{b}_0$ in Theorems 3.3 and 3.4 must be chosen to guarantee $\beta_t, \theta_t \in (0, 1]$.

Remark 3.1. If we define an approximate gradient mapping $\widetilde{\mathcal{G}}_{\eta_t}$ for $\mathcal{G}_\eta$ in (13) as $\widetilde{\mathcal{G}}_{\eta_t}(x_t) := \frac{1}{\eta_t}\big(x_t - \mathrm{prox}_{\eta_t\mathcal{R}}(x_t - \eta_t v_t)\big)$. Clearly, if $\mathcal{R} = 0$, then $\widetilde{\mathcal{G}}_{\eta_t}(x_t) = v_t$, which reduces to an approximation of the gradient $\nabla\Phi_{\gamma_t}(x_t)$. Then, the update of Algorithm 1 becomes $x_{t+1} := x_t - \eta_t\theta_t\widetilde{\mathcal{G}}_{\eta_t}(x_t)$. Thus we can refer to $\hat{\theta}_t := \theta_t\eta_t$ as a combined step-size (also called a learning rate). Since $\eta_t := \frac{2}{L_{\Phi_{\gamma_t}}(3+\theta_t)}$ we have $\hat{\theta}_t = \frac{2\theta_t}{L_{\Phi_{\gamma_t}}(3+\theta_t)} \le \frac{2\theta_t}{3L_{\Phi_{\gamma_t}}}$, which is **diminishing** to zero in (21) or (25).

3.3 Constructing approximate KKT point for (1) from Algorithm 1

Existing works such as [37, 39] do not show how to construct an ϵ-KKT point of (1) or an ϵ-stationary point of (2) from $\bar{x}_T$ with $\mathbb{E}\big[\|\mathcal{G}_{\bar{\eta}_T}(\bar{x}_T)\|^2\big] \le \varepsilon^2$. Lemma 3.1, whose proof is in Appendix A.3, shows how to construct an ϵ-KKT point of (1) in the sense of Definition 2.1 with $\epsilon := \mathcal{O}(\varepsilon)$.

Lemma 3.1. *Let $\bar{x}_T$ be computed by Algorithm 1 up to an accuracy $\varepsilon > 0$ after T iterations. Assume that we can approximate $F'(\bar{x}_T)$, $F(\bar{x}_T)$, and $F(\tilde{x}^*_{\gamma_T})$, respectively such that*

$$\begin{aligned}&\mathbb{E}\big[\|\tilde{F}(\bar{x}_T) - F(\bar{x}_T)\|\big] \le (\mu_\psi + \gamma_T)\varepsilon, \quad \mathbb{E}\big[\|(\tilde{J}(\bar{x}_T) - F'(\bar{x}_T))^\top\nabla\phi_{\gamma_T}(\tilde{F}(\bar{x}_T))\|\big] \le \varepsilon, \\ &\text{and} \;\; \mathbb{E}\big[\|\tilde{F}(\tilde{x}^*_{\gamma_T}) - F(\tilde{x}^*_{\gamma_T})\|\big] \le \varepsilon.\end{aligned} \tag{27}$$

Let us denote $\widetilde{\nabla}\Phi_{\gamma_T}(\bar{x}_T) := \tilde{J}(\bar{x}_T)^\top \nabla\phi_\gamma(\tilde{F}(\bar{x}_T))$ *and compute* $(\tilde{x}^*_{\gamma_T}, \tilde{y}^*_{\gamma_T})$ *as*

$$\tilde{x}^*_{\gamma_T} := \mathrm{prox}_{\eta_T\mathcal{R}}(\bar{x}_T - \eta_T\widetilde{\nabla}\Phi_{\gamma_T}(\bar{x}_T)) \quad \textit{and} \quad \tilde{y}^*_{\gamma_T} := y^*_{\gamma_T}(\tilde{F}(\tilde{x}^*_{\gamma_T})) \textit{ by } (9). \tag{28}$$

Suppose that $\mathbb{E}\big[\|\mathcal{G}_{\eta_T}(\bar{x}_T)\|^2\big] \le \varepsilon^2$ *and* $0 \le \gamma_T \le c_2\varepsilon$ *for a constant* $c_2 \ge 0$. *Then*

$$\mathbb{E}\big[\mathcal{E}(\tilde{x}^*_{\gamma_T}, \tilde{y}^*_{\gamma_T})\big] \le \epsilon, \quad \textit{where} \quad \epsilon := \big[\tfrac{13}{3} + \tfrac{8}{3}M_F\|K\|^2 + c_2D_\psi\big]\varepsilon, \tag{29}$$

where D_ψ *is in Lemma A.1 and* $\mathcal{E}(\cdot)$ *is given by* (8). *In other words,* $(\tilde{x}^*_{\gamma_T}, \tilde{y}^*_{\gamma_T})$ *is an* ϵ*-KKT of* (1).

If we use stochastic estimators as in (15) to form $\tilde{F}(\bar{x}_T)$, $\tilde{J}(\bar{x}_T)$, and $\tilde{F}(\tilde{x}^*_{\gamma_T})$ with batch sizes b_T, $\hat{b}_T$, and $\tilde{b}_T$, respectively, then (27) holds if we choose $b_T := \big\lfloor \frac{\sigma_F^2}{(\mu_\psi+\gamma_T)^2\varepsilon^2}\big\rfloor$, $\hat{b}_T := \big\lfloor \frac{\sigma_J^2}{\varepsilon^2}\big\rfloor$, and $\tilde{b}_T := \big\lfloor \frac{\sigma_F^2}{\varepsilon^2}\big\rfloor$. We do not explicitly compute Jacobian $\tilde{J}(\bar{x}_T)$, but its matrix-vector product $\tilde{J}(\bar{x}_T)^\top\nabla\phi_{\gamma_T}(\tilde{F}(\bar{x}_T))$. This extra cost is dominated by $\mathcal{T}_J$ and $\mathcal{T}_F$ in Theorems 3.1, 3.2, 3.3, and 3.4. For $\bar{x}_T$ computed by Theorems 3.1 and 3.2, we can set $\gamma_T := 0$, or equivalently, $c_2 := 0$. For $\bar{x}_T$ computed by Theorem 3.3, since $\gamma_T := c_2\varepsilon$ and $\mu_\psi = 0$, we have $b_T = \big\lfloor \frac{\sigma_F^2}{c_2^2\varepsilon^4}\big\rfloor < \mathcal{T}_F = \mathcal{O}\big(\frac{\hat{\Delta}_0^{3/2}}{\varepsilon^5}\big)$.

4 Restarting variant of Algorithm 1 and its convergence

In this section., we propose a simple restarting variant, Algorithm 2, of Algorithm 1, prove its convergence, and estimate its oracle complexity bounds for both smooth ϕ_0 and non-smooth ϕ_0 in (2). For simplicity of our analysis, we only consider the constant step-size case, and omit the diminishing step-size analysis.

4.1 Restarting variant

Motivation. Since the constant step-size θ in (19) of Theorem 3.1 and (23) of Theorem 3.3 depends on the number of iterations T. Clearly, if T is large, then θ is small. To avoid using small step-size θ, we can restart Algorithm 1 by frequently resetting its initial point and parameters after T iterations. This variant is described in Algorithm 2. Algorithm 2 has two loops, where each iteration s of the outer loop is called the s-th stage. Unlike the outer loop in other variance-reduced methods relying on SVRG or SARAH estimators from the literature, which is mandatory to guarantee convergence, our outer loop is optional, since without it, Algorithm 2 reduces to Algorithm 1, and it still converges.

Algorithm 2 (Restarting Variant of Algorithm 1)

1: **Inputs:** An arbitrarily initial point $\tilde{x}^0 \in \mathrm{dom}(F)$, and a fixed number of iterations T.
2: **For** $s := 1, \cdots, S$ **do**
3: Run Algorithm 1 for T iterations starting from $x_0^{(s)} := \tilde{x}^{s-1}$.
4: Set $\tilde{x}^s := x_{T+1}^{(s)}$ as the last iterate of Algorithm 1.
5: **EndFor**
6: **Output:** Choose $\bar{x}_N$ randomly from $\{x_t^{(s)}\}_{t=0\to T}^{s=1\to S}$ such that $\mathbf{Prob}\left(\bar{x}_N = x_t^{(s)}\right) = \frac{\theta_t}{S\sum_{j=0}^T\theta_j}$.

4.2 The smooth case ϕ_0 with constant step-size

The smoothness of ϕ_0 is equivalent to the μ_ψ-strong convexity of ψ in (1). The following theorem, whose proof is in Appendix C, states convergence rate and estimates oracle complexity of Algorithm 2.

Theorem 4.1. *Suppose that Assumptions 2.1 and 2.2 hold, ψ is strongly convex (i.e., $\mu_\psi = 1 > 0$), and P, Q, and L_{Φ_0} are defined by (18). Let $\{x_t^{(s)}\}_{t=0\to T}^{s=1\to S}$ be generated by Algorithm 2 using $\gamma := 0$, $b_0 := c_0\hat{b}_0$, $b_1 = b_2 := c_0 b$, $\hat{b}_1 = \hat{b}_2 = b$ for some $c_0 > 0$ and given batch sizes $b > 0$ and $\hat{b}_0 > 0$, and the parameter configuration (19). Then, the following estimate holds*

$$\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_N)\|^2\big] \le \frac{16P\hat{b}_0^{1/4}}{S[b(T+1)]^{3/4}}\big[\Psi_0(\tilde{x}^0) - \Psi_0^\star\big] + \frac{24Q}{[\hat{b}_0 b(T+1)]^{1/2}}, \tag{30}$$

where $\bar{x}_N$ is uniformly randomly chosen from $\{x_t^{(s)}\}_{t=0\to T}^{s=1\to S}$.

Given $\varepsilon > 0$, if we choose $T := \big\lfloor \frac{48Q}{b\varepsilon^2}\big\rfloor$ and $\hat{b}_0 := \big\lfloor \frac{48Q}{\varepsilon^2}\big\rfloor$, then after at most $S := \big\lfloor \frac{8P}{\varepsilon\sqrt{3Q}}\big\rfloor$ outer iterations, we obtain $\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_N)\|^2\big] \le \varepsilon^2$. Consequently, the total number of function evaluations $\mathcal{T}_F$ and the total number of Jacobian evaluations $\mathcal{T}_J$ are at most $\mathcal{T}_F = \mathcal{T}_J := \big\lfloor \frac{400P\sqrt{3Q}}{\varepsilon^3}\big\rfloor$.

Theorem 4.1 holds for any mini-batch b such that $1 \le b \le \frac{48Q}{\varepsilon^2}$, which is different from, e.g., [37], where the complexity result holds under large batches. Moreover, the total oracle calls $\mathcal{T}_F$ and $\mathcal{T}_J$ are independent of b. In this case, the weight β and the step-size θ become

$$\beta := 1 - \frac{b\varepsilon^2}{48Q} \quad \text{and} \quad \theta := \frac{bL_{\Phi_0}}{4P\varepsilon\sqrt{3Q}}.$$

Clearly, if b is large, then our step-size θ is also large.

4.3 The non-smooth ϕ_0 with constant step-size

Finally, we prove the convergence of Algorithm 2 when ψ is non-strongly convex (i.e., ϕ_0 in (2) is possibly nonsmooth). The proof of the following theorem is in Appendix C.

Theorem 4.2. *Assume that Assumptions 2.1 and 2.2 hold, ψ in (1) is non-strongly convex (i.e., $\mu_\psi = 0$), and P, Q, and L_{Φ_γ} are defined by (18). Let $\{x_t^{(s)}\}_{t=0\to T}^{s=1\to S}$ be generated by Algorithm 2 after $N := S(T+1)$ iterations using:*

$$\left\{\begin{array}{ll} b_1 = b_2 := \frac{2c_0 b\hat{R}_0}{\varepsilon^2}, \quad \hat{b}_1 = \hat{b}_2 := b, \quad b_0 := \frac{4c_0\hat{R}_0^2}{\varepsilon^4}, \quad \hat{b}_0 := \frac{2\hat{R}_0}{\varepsilon^2}, \\ \gamma := \frac{\varepsilon}{\sqrt{2\hat{R}_0}}, \quad \textit{and} \quad \beta := 1 - \frac{b\varepsilon^2}{2\hat{R}_0}. \end{array}\right. \tag{31}$$

where $\varepsilon > 0$ is a given tolerance[1]*, and*

$$R_0 := 16\big[\Psi_0(\tilde{x}^0) - \Psi^\star + B_\psi\big] \quad \textit{and} \quad \hat{R}_0 := 24Q. \tag{32}$$

Then, if we choose $T := \big\lfloor \frac{2\hat{R}_0}{\varepsilon^2}\big\rfloor$, then after at most $S := \big\lfloor \frac{\sqrt{2}R_0}{b\varepsilon\sqrt{\hat{R}_0}}\big\rfloor$ outer iterations, we obtain $\bar{x}_T$ such that $\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_T)\|^2\big] \le \varepsilon^2$.

Consequently, the total number of function evaluations $\mathcal{T}_F$ and the total number of Jacobian evaluations $\mathcal{T}_J$ are respectively at most

$$\mathcal{T}_F := \frac{4\sqrt{2}c_0 R_0\hat{R}_0^{3/2}(3+b^{-1})}{\varepsilon^5} = \mathcal{O}\left(\frac{R_0\hat{R}_0^{3/2}}{\varepsilon^5}\right) \quad \textit{and} \quad \mathcal{T}_J := \frac{2\sqrt{2}R_0\hat{R}_0^{1/2}(3+b^{-1})}{\varepsilon^3} = \mathcal{O}\left(\frac{R_0\hat{R}_0^{1/2}}{\varepsilon^3}\right).$$

Remark 4.1. Note that we do not need to choose the batch sizes and parameters depending on R_0 as in (31), which is unknown since $\Psi_0^\star$ is unknown, but they are proportional to R_0. In this case, the complexity bounds in Theorem 4.2 will only be shifted by a constant factor.

As we can see from Theorem 4.2, the number of outer iterations S is divided by the batch size b. However, the terms $\frac{12\sqrt{2}c_0 R_0\hat{R}_0^{3/2}}{\varepsilon^5}$ and $\frac{6\sqrt{2}R_0\hat{R}_0^{1/2}}{\varepsilon^3}$ are independent of b and dominate the complexity bounds in both $\mathcal{T}_F$ and $\mathcal{T}_J$, respectively.

[1]The batch sizes and T in this paper must be integer, but for simplicity, we do not write their rounding form.

5 Numerical experiments

We use two examples to illustrate our algorithm and compare it with existing methods. Our code is implemented in Python 3.6.3, running on a Linux desktop (3.6GHz Intel Core i7 and 16Gb memory).

5.1 Risk-averse portfolio optimization

We consider a risk-averse portfolio optimization problem from [18], and recent used in [38]:

$$\max_{x\in\mathbb{R}^p}\Big\{\mathbb{E}_\xi\big[h_\xi(x)\big] - \rho\mathrm{Var}_\xi\big[h_\xi(x)\big] \equiv \mathbb{E}_\xi\big[h_\xi(x)\big] + \rho\mathbb{E}_\xi\big[h_\xi(x)\big]^2 - \rho\mathbb{E}_\xi\big[h^2_\xi(x)\big]\Big\}, \tag{33}$$

where $\rho > 0$ is a trade-off parameter and $h_\xi(x)$ is a reward for the portfolio vector x. Following [38], (33) can be reformulated into (2), where $\phi_0(u) = u_1 + \rho u_1^2 - \rho u_2$ is smooth, and $\mathbf{F}(x,\xi) = (h_\xi(x), h^2_\xi(x))^\top$. Suppose further that we only consider N periods of time. Then we can view $\xi \in \{1,\cdots,N\}$ as a discrete random variable and define $h_i(x) := \langle r_i, x\rangle$ as a linear reward function, where $r_i := (r_{i1},\cdots,r_{ip})^\top$ and r_{ij} represents the return per unit of j at time i. We also choose $\mathcal{R}(x) := \lambda\|x\|_1$ as a regularizer to promote sparsity as in [38].

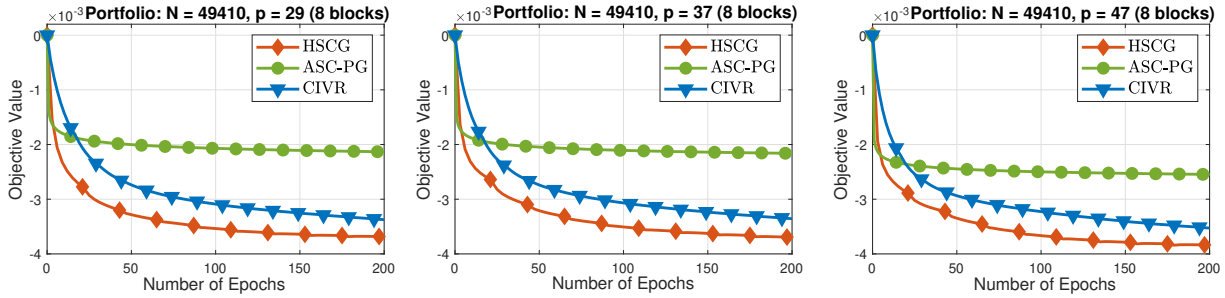

Figure 1: Comparison of three algorithms for solving (33) on 3 different datasets.

We implement our algorithm, abbreviated by `HSCG` (i.e., Hybrid Stochastic Compositional Gradient for short), and test it on three real-world portfolio datasets, which contain 29, 37, and 47 portfolios, respectively, from the Keneth R. French Data Library [1]. We set $\rho := 0.2$ and $\lambda := 0.01$ as in [38]. For comparison, we also implement 2 methods, called `CIVR` in [38] and `ASC-PG` in [32]. The step-size η of all algorithms are well tuned from a set of trials $\{1, 0.5, 0.1, 0.05, 0.01, 0.001, 0.0001\}$. The performance of 3 algorithms are shown in Figure 1 for three datasets using $b := \lfloor N/8\rfloor$ (8 blocks).

One can observe from Fig. 1 that both `HSCG` and `CIVR` highly outperform `ASC-PG` due to their variance-reduced property. `HSCG` is slightly better than `CIVR` since it has a flexible step-size θ_t. Note that, in theory, `CIVR` requires a large batch for both function values and Jacobian, which may affect its performance, while `HSCG` can work with a wide range of batches, including singe sample.

5.2 Stochastic minimax problem

We consider the following regularized stochastic minimax problem studied, e.g., in [26]:

$$\min_{x\in\mathbb{R}^p}\Big\{\max_{1\le i\le m}\big\{\mathbb{E}_\xi\big[\mathbf{F}_i(x,\xi)\big]\big\} + \tfrac{\lambda}{2}\|x\|^2\Big\}, \tag{34}$$

where $\mathbf{F}_i : \mathbb{R}^p \times \Omega \to \mathbb{R}_+$ can be taken as the loss function of the i-th model. If we define $\phi_0(u) := \max_{1\le i\le m}\{u_i\}$ and $\mathcal{R}(x) := \frac{\lambda}{2}\|x\|^2$, then (34) can be reformulated into (2). Since $u_i \ge 0$,

we have $\phi_0(u) := \max_{1\le i\le m}\{u_i\} = \|u\|_\infty = \max_{\|y\|_1\le 1}\{\langle u, y\rangle\}$, which is nonsmooth. Therefore, we can smooth ϕ_0 as $\phi_\gamma(u) := \max_{\|y\|_1\le 1}\{\langle u, y\rangle - (\gamma/2)\|y\|^2\}$ using $b(y) := \frac{1}{2}\|y\|^2$.

In this example, we employ (34) to solve a model selection problem in binary classification with nonconvex loss, see, e.g., [40]. Suppose that we have four ($m = 4$) different nonconvex losses: $\mathbf{F}_1(x,\xi) := 1 - \tanh(b\langle a, x\rangle)$, $\mathbf{F}_2(x,\xi) := \log(1+\exp(-b\langle a, x\rangle)) - \log(1+\exp(-b\langle a, x\rangle - 1))$, $\mathbf{F}_3(x,\xi) := (1 - 1/(\exp(-b\langle a, x\rangle) + 1))^2$, and $\mathbf{F}_4(x,\xi) := \log(1+\exp(-b\langle a, x\rangle))$ (see [40] for more details), where $\xi := (a, b)$ represents examples. We assume that we have N examples of ξ.

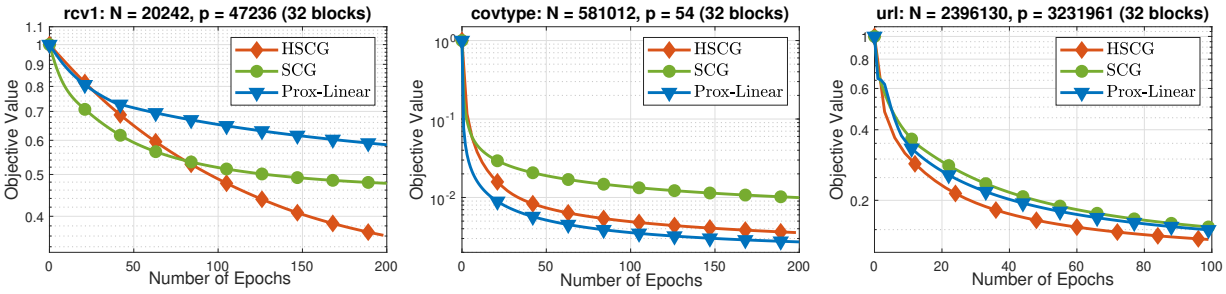

Figure 2: Comparison of three algorithms for solving (34) on 3 different datasets.

We implement three algorithms: `HSCG`, `SCG` in [31], and `Prox-Linear` in [39]. We test them on 3 datasets from LIBSVM [6]. We set $\lambda := 10^{-4}$ and update our γ_t parameter as $\gamma_t := \frac{1}{2(t+1)^{1/3}}$. The step-size η of all algorithms are well tuned from $\{1, 0.5, 0.1, 0.05, 0.01, 0.001, 0.0001\}$, and their performance is shown in Figure 2 for three datasets: **rcv1**, **covtype**, and **url** with 32 blocks.

One can observe from Figure 2 that `HSCG` outperforms `SCG` and `Prox-Linear` on **rcv1** and **url**. For **covtype**, since p is very small, allowing us to evaluate the prox-linear operator to a high accuracy, `Prox-Linear` slightly performs better than ours and much better than `SCG`. Note that solving the subproblem of `Prox-Linear` is expensive when p is large. Hence, if p is large, `Prox-Linear` becomes much slower than `HSCG` and `SCG` in terms of time.

6 Conclusions

We have proposed a new single loop hybrid variance-reduced SGD algorithm, Algorithm 1, to solve a class of nonconvex-concave saddle-point problems. The main idea is to combine both smoothing idea [20] and hybrid SGD approach in [29] to develop novel algorithms with less tuning effort. Our algorithm relies on standard assumptions, and can achieve the best-known oracle complexity, and in some cases, the optimal oracle complexity. It also has several computational advantages compared to existing methods such as avoiding expensive subproblems, working with both single sample and mini-batches, and using constant and diminishing step-sizes. We have also proposed a simple restarting variant, Algorithm 2, in Appendix 4 to improve practical performance in the constant step-size case without sacrificing complexity bounds. We believe that both algorithms and theoretical results are new, even in the smooth case, compared to [28, 37, 39]. Our future plan is to exploit this approach to solve some interesting applications, such as robust optimization and learning, and GANs.

Appendix

A Some technical results and proof of Lemma 3.1

In this appendix, we provide some useful properties of ϕ_0 in (3) and its smoothed approximation ϕ_γ defined by (9) in Section 2. Then we recall and prove some bounds of variance for $\tilde{F}_t$, $\tilde{J}_t$, and v_t.

Finally, we prove Lemma 3.1 in the main text.

A.1 Properties of the smoothed function ϕ_γ

Under Assumption 2.2, ϕ_0 in (3) and ϕ_γ defined by (9) have the following properties.

Lemma A.1. *Let ϕ_0 be defined by (3) and ϕ_γ be defined by (9). Then, the following statements hold:*

(a) $\mathrm{dom}(\psi)$ *is bounded by M_ψ iff ϕ_0 is M_{ϕ_0}-Lipschitz continuous with $M_{\phi_0} := M_\psi\|K\|$.*
(b) $\mathrm{dom}(\psi)$ *is bounded by M_ψ iff ϕ_γ is Lipschitz continuous with $M_{\phi_\gamma} := M_\psi\|K\|$.*
(c) *ϕ_γ is convex and L_{ϕ_γ}-smooth with $L_{\phi_\gamma} := \frac{\|K\|^2}{\gamma+\mu_\psi}$.*
(d) *It holds that $\phi_\gamma(u) \le \phi_0(u) \le \phi_\gamma(u) + \gamma B_\psi$ for all $u \in \mathbb{R}^q$, where $\gamma > 0$ and $B_\psi := \sup\{b(y) \mid y \in \mathrm{dom}(\psi)\}$. In addition, we have $D_\psi := \max_{v\in\mathrm{dom}(\psi)} \|\nabla b(v)\| < +\infty$.*
(e) *We have $\phi_\gamma(u) \le \phi_{\hat\gamma}(u) + (\hat\gamma - \gamma)b(y^*_\gamma(u)) \le \phi_{\hat\gamma}(u) + (\hat\gamma - \gamma)B_\psi$ for all $\hat\gamma \ge \gamma > 0$.*

Proof. The statement (a) can be found in [3, Corollary 17.19].

Since $\nabla\phi_\gamma(u) = Ky^*_\gamma(u)$ with $y^*_\gamma(u) \in \mathrm{dom}(\psi)$, we have $\|\nabla\phi_\gamma(u)\| \le \|K\|\|y^*_\gamma(u)\| \le M_\psi\|K\|$. Applying again [3, Corollary 17.19] we prove (b).

The statement (c) holds due to the well-known Baillon-Haddad theorem [3, Corollary 18.17].

The proof of the first part of (d) can be found in [20]. Under Assumption 2.2 and the continuous differentiability of b, we have $D_\psi := \max_{v\in\mathrm{dom}(\psi)} \|\nabla b(v)\| < +\infty$.

Finally, for any u and y, since $s(\gamma; u, y) := \langle u, Ky\rangle - \psi(y) - \gamma b(y)$ is linear in γ. Therefore, $\phi_\gamma(u) := \max_{y\in\mathbb{R}^n} s(\gamma; u, y)$ is convex in γ and $\frac{d}{d\gamma}\phi_\gamma(u) = -b(y^*_\gamma(u)) \le 0$. Consequently, we have $\phi_\gamma(u) + \frac{d}{d\gamma}\phi_\gamma(u)(\hat\gamma - \gamma) = \phi_\gamma(u) - (\hat\gamma - \gamma)b(y^*_\gamma(u)) \le \phi_{\hat\gamma}(u)$, which implies (e). □

One common example of ψ in Assumption 2.2 is $\psi(x) := \delta_{\mathcal{X}}(x)$, the indicator of a nonempty, closed, bounded, and convex set $\mathcal{X}$. For instance, $\mathcal{X} := \{y \in \mathbb{R}^n \mid \|y\|_* \le 1\}$ is a unit ball in the dual norm $\|\cdot\|_*$ of $\|\cdot\|$. Then, we have $\phi_0(u) := \|u\|$, which is clearly Lipschitz continuous. In particular, if $\mathcal{X} := \{y \in \mathbb{R}^n \mid \|y\|_\infty \le 1\}$, then $\phi_0(u) := \|u\|_1$.

A.2 Key bounds on the variance of estimators

Next, we provide some useful bounds for the estimators $\tilde F_t$ and $\tilde J_t$ defined in (14). The following lemma can be found in [29].

Lemma A.2. *Let $\tilde F_t$ and $\tilde J_t$ be defined by (14), and $\mathcal{F}_t$ be defined by (17). Then*

$$\begin{aligned}
\mathbb{E}_{(\mathcal{B}^1_t,\mathcal{B}^2_t)}\big[\|\tilde F_t - F(x_t)\|^2\big] &= \beta^2_{t-1}\|\tilde F_{t-1} - F(x_{t-1})\|^2 - \beta^2_{t-1}\|F(x_t) - F(x_{t-1})\|^2 \\
&\quad + (1-\beta_{t-1})^2\mathbb{E}_{\mathcal{B}^2_t}\big[\|\mathbf{F}(x_t,\zeta_t) - F(x_t)\|^2\big] \\
&\quad + \tfrac{\beta^2_{t-1}}{b_1}\mathbb{E}_\xi\big[\|\mathbf{F}(x_t,\xi) - \mathbf{F}(x_{t-1},\xi)\|^2\big], \\
\mathbb{E}_{(\hat{\mathcal{B}}^1_t,\hat{\mathcal{B}}^2_t)}\big[\|\tilde J_t - F'(x_t)\|^2\big] &= \hat\beta^2_{t-1}\|\tilde J_{t-1} - F'(x_{t-1})\|^2 - \hat\beta^2_{t-1}\|F'(x_t) - F'(x_{t-1})\|^2 \\
&\quad + (1-\hat\beta_{t-1})^2\mathbb{E}_{\hat{\mathcal{B}}^2_t}\big[\|\mathbf{F}'(x_t,\hat\zeta_t) - F'(x_t)\|^2\big] \\
&\quad + \tfrac{\hat\beta^2_{t-1}}{\hat b_1}\mathbb{E}_{\hat\xi}\big[\|\mathbf{F}'(x_t,\hat\xi) - \mathbf{F}'(x_{t-1},\hat\xi)\|^2\big].
\end{aligned} \tag{35}$$

Furthermore, we can bound the variance of the estimator v_t of $\nabla\Phi_{\gamma_t}(x_t)$ defined in (16) as follows.

Lemma A.3. *Let Φ_γ and v_t be defined by* (10) *and* (16)*, respectively. Then, under Assumptions 2.1 and 2.2, we have*

$$\mathbb{E}\big[\|v_t - \nabla\Phi_{\gamma_t}(x_t)\|^2\big] \le 2M_F^2 L_{\phi_{\gamma_t}}^2 \mathbb{E}\big[\|\tilde{F}_t - F(x_t)\|^2\big] + 2M_{\phi_{\gamma_t}}^2 \mathbb{E}\big[\|\tilde{J}_t - F'(x_t)\|^2\big]. \tag{36}$$

Proof. First, by the composition rule of derivatives, we can derive

$$\begin{aligned}
\|v_t - \nabla\Phi_{\gamma_t}(x_t)\|^2 &= \|\tilde{J}_t^\top \nabla\phi_{\gamma_t}(\tilde{F}_t) - F'(x_t)^\top \nabla\phi_{\gamma_t}(F(x_t))\|^2 \\
&= \big\|\tilde{J}_t^\top \nabla\phi_{\gamma_t}(\tilde{F}_t) - F'(x_t)^\top \nabla\phi_{\gamma_t}(\tilde{F}_t) + F'(x_t)^\top \nabla\phi_{\gamma_t}(\tilde{F}_t) \\
&\quad - F'(x_t)^\top \nabla\phi_{\gamma_t}(F(x_t))\big\|^2 \\
&\overset{(i)}{\le} 2\|(\tilde{J}_t - F'(x_t))^\top \nabla\phi_{\gamma_t}(\tilde{F}_t)\|^2 + 2\|F'(x_t)^\top\big(\nabla\phi_{\gamma_t}(\tilde{F}_t) - \nabla\phi_{\gamma_t}(F(x_t))\big)\|^2 \\
&\le 2\|\nabla\phi_{\gamma_t}(\tilde{F}_t)\|^2 \|\tilde{J}_t - F'(x_t)\|^2 + 2\|\nabla\phi_{\gamma_t}(\tilde{F}_t) - \nabla\phi_{\gamma_t}(F(x_t))\|^2 \|F'(x_t)\|^2 \\
&\overset{(ii)}{\le} 2M_{\phi_{\gamma_t}}^2 \|\tilde{J}_t - F'(x_t)\|^2 + 2L_{\phi_{\gamma_t}}^2 M_F^2 \|\tilde{F}_t - F(x_t)\|^2.
\end{aligned}$$

Here, we use $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ in *(i)* and the $M_{\phi_{\gamma_t}}$-Lipschitz continuity, $L_{\phi_{\gamma_t}}$-smoothness of ϕ_{γ_t}, and (5) in *(ii)*. Taking expectation over $\mathcal{F}_{t+1}$ on both sides the last inequality, we obtain

$$\mathbb{E}\big[\|v_t - \nabla\Phi_{\gamma_t}(x_t)\|^2\big] \le 2M_F^2 L_{\phi_{\gamma_t}}^2 \mathbb{E}\big[\|\tilde{F}_t - F(x_t)\|^2\big] + 2M_{\phi_{\gamma_t}}^2 \mathbb{E}\big[\|\tilde{J}_t - F'(x_t)\|^2\big],$$

which proves (36). □

A.3 The construction of approximate KKT points for (1)

Recall from (10) that $\Phi_\gamma(x) = \phi_\gamma(F(x))$ and $\nabla\Phi_\gamma(x) = F'(x)^\top \nabla\phi_\gamma(F(x))$, where ϕ_γ is defined by (9). We define a smoothed approximation problem of (2) as follows:

$$\min_{x\in\mathbb{R}^p} \Big\{\Psi_\gamma(x) := \Phi_\gamma(x) + \mathcal{R}(x) \equiv \phi_\gamma(F(x)) + \mathcal{R}(x)\Big\}. \tag{37}$$

Clearly, if $\gamma = 0$, then (37) reduces to (2). The optimality condition of (37) becomes

$$0 \in \nabla\Phi_\gamma(x_\gamma^\star) + \partial\mathcal{R}(x_\gamma^\star) \equiv F'(x_\gamma^\star)^\top \nabla\phi_\gamma(F(x_\gamma^\star)) + \partial\mathcal{R}(x_\gamma^\star). \tag{38}$$

Here, $x_\gamma^\star$ is called a stationary point of (37). Therefore, an ε-stationary point $\tilde{x}_\gamma^*$ is defined as

$$\mathbb{E}\big[\operatorname{dist}\big(0, \nabla\Phi_\gamma(\tilde{x}_\gamma^*) + \partial\mathcal{R}(\tilde{x}_\gamma^*)\big)\big] \le \varepsilon. \tag{39}$$

Again, the expectation $\mathbb{E}\big[\,\cdot\,\big]$ is taken over all the randomness generated by the model (37) and the algorithm for finding $\tilde{x}_\gamma^*$.

Alternatively, using the definition of ϕ_γ in (9), problem (37) can be written as

$$\min_{x\in\mathbb{R}^p}\max_{y\in\mathbb{R}^n} \Big\{\mathcal{R}(x) + \langle F(x), Ky\rangle - \psi(y) - \gamma b(y)\Big\}. \tag{40}$$

Its optimality condition becomes

$$0 \in \partial\mathcal{R}(x_\gamma^\star) + F'(x_\gamma^\star)Ky_\gamma^\star \quad \text{and} \quad 0 \in K^\top F(x_\gamma^\star) - \partial\psi(y_\gamma^\star) - \gamma\nabla b(y_\gamma^\star). \tag{41}$$

Using the definition of $\mathcal{E}$ in (8), we have

$$\mathcal{E}(x_\gamma^\star, y_\gamma^\star) := \operatorname{dist}\big(0, \partial\mathcal{R}(x_\gamma^\star) + F'(x_\gamma^\star)Ky_\gamma^\star\big) + \operatorname{dist}\Big(0, K^\top F(x_\gamma^\star) - \partial\psi(y_\gamma^\star)\Big) \le \gamma D_\psi. \tag{42}$$

Here, we use the fact that $\|\nabla b(y_\gamma^\star)\| \le D_\psi$ as stated in Lemma A.1.

Given $\bar{x} \in \mathrm{dom}(\Psi_0)$, let $\tilde{F}(\cdot)$ and $\tilde{J}(\cdot)$ be a stochastic approximation of $F(\cdot)$ and $F'(\cdot)$, respectively. We define $(\tilde{x}_\gamma^*, y_\gamma^*)$ as follows:

$$\begin{cases} \tilde{x}_\gamma^* &:= \mathrm{prox}_{\eta\mathcal{R}}\left(\bar{x} - \eta\widetilde{\nabla}\Phi_\gamma(\bar{x})\right), \quad \text{where} \quad \widetilde{\nabla}\Phi_\gamma(\bar{x}) := \tilde{J}(\bar{x})^\top\nabla\phi_\gamma(\tilde{F}(\bar{x})), \\ \tilde{y}_\gamma^* &:= y_\gamma^*(\tilde{F}(\tilde{x}_\gamma^*)) \equiv \arg\min_{y\in\mathbb{R}^n}\left\{\langle K^\top\tilde{F}(\tilde{x}_\gamma^*), y\rangle - \psi(y) - \gamma b(y)\right\}, \end{cases} \tag{43}$$

Note that $\tilde{x}_\gamma^*$ only depends on $\bar{x}$, while $\tilde{y}_\gamma^*$ depends on both $\bar{x}$ and $\tilde{x}_\gamma^*$. Hence, we first compute $\tilde{x}_\gamma^*$ and then compute $\tilde{y}_\gamma^*$.

The following lemma provides key estimates to prove Lemma 3.1 in the main text.

Lemma A.4. *Under Assumptions 2.1 and 2.2, for given $\bar{x}$ and $\eta > 0$, $\tilde{x}_\gamma^*$ defined by (43) satisfies*

$$\mathrm{dist}\big(0, \nabla\Phi_\gamma(\tilde{x}_\gamma^*) + \partial\mathcal{R}(\tilde{x}_\gamma^*)\big) \le \big(1 + \eta L_{\Phi_\gamma}\big)\|\mathcal{G}_\eta(\bar{x})\| + (2 + \eta L_{\Phi_\gamma})\|\nabla\Phi_\gamma(\bar{x}) - \widetilde{\nabla}\Phi_\gamma(\bar{x})\|. \tag{44}$$

Let $(\tilde{x}_\gamma^*, \tilde{y}_\gamma^*)$ be computed by (43), and $\mathcal{E}(x, y)$ be defined by (8). Then, we have

$$\begin{aligned} \mathcal{E}(\tilde{x}_\gamma^*, \tilde{y}_\gamma^*) \le\ & \big(1 + \eta L_{\Phi_\gamma}\big)\|\mathcal{G}_\eta(\bar{x})\| + \gamma D_\psi + \|K\|\|F(\tilde{x}_\gamma^*) - \tilde{F}(\tilde{x}_\gamma^*)\| \\ &+ \big(2 + \eta L_{\Phi_\gamma}\big)\big[\|(\tilde{J}(\bar{x}) - F'(\bar{x}))^\top\nabla\phi_\gamma(\tilde{F}(\bar{x})\| + L_{\phi_\gamma}M_F\|\tilde{F}(\bar{x}) - F(\bar{x})\|\big], \end{aligned} \tag{45}$$

where D_ψ is defined in Lemma A.1.

Proof. From (43), we have $\bar{x} - \eta\widetilde{\nabla}\Phi_\gamma(\bar{x}) \in \tilde{x}_\gamma^* + \eta\partial\mathcal{R}(\tilde{x}_\gamma^*)$, which is equivalent to

$$r_x^* := \frac{1}{\eta}(\bar{x} - \tilde{x}_\gamma^*) + \big(\nabla\Phi_\gamma(\tilde{x}_\gamma^*) - \widetilde{\nabla}\Phi_\gamma(\bar{x})\big) \in \nabla\Phi_\gamma(\tilde{x}_\gamma^*) + \partial\mathcal{R}(\tilde{x}_\gamma^*). \tag{46}$$

We can bound r_x^* in (46) as follows:

$$\begin{aligned} \|r_x^*\| &\le \tfrac{1}{\eta}\|\bar{x} - \tilde{x}_\gamma^*\| + \|\nabla\Phi_\gamma(\tilde{x}_\gamma^*) - \nabla\Phi_\gamma(\bar{x})\| + \|\nabla\Phi_\gamma(\bar{x}) - \widetilde{\nabla}\Phi_\gamma(\bar{x})\| \\ &\le \tfrac{1}{\eta}\big(1 + \eta L_{\Phi_\gamma}\big)\|\tilde{x}_\gamma^* - \bar{x}\| + \|\nabla\Phi_\gamma(\bar{x}) - \widetilde{\nabla}\Phi_\gamma(\bar{x})\|. \end{aligned} \tag{47}$$

Next, from (13), let us define $\bar{x}_\gamma^* := \bar{x} - \eta\mathcal{G}_\eta(\bar{x}) = \mathrm{prox}_{\eta\mathcal{R}}(\bar{x} - \eta\nabla\Phi_\gamma(\bar{x}))$. Then, we have

$$\begin{aligned} \|\tilde{x}_\gamma^* - \bar{x}\| &\le \|\tilde{x}_\gamma^* - \bar{x}_\gamma^*\| + \|\bar{x}_\gamma^* - \bar{x}\| \\ &= \|\mathrm{prox}_{\eta\mathcal{R}}(\bar{x} - \eta\widetilde{\nabla}\Phi_\gamma(\bar{x})) - \mathrm{prox}_{\eta\mathcal{R}}(\bar{x} - \eta\nabla\Phi_\gamma(\bar{x}))\| + \eta\|\mathcal{G}_\eta(\bar{x})\| \\ &\le \eta\|\widetilde{\nabla}\Phi_\gamma(\bar{x}) - \nabla\Phi_\gamma(\bar{x})\| + \eta\|\mathcal{G}_\eta(\bar{x})\|. \end{aligned} \tag{48}$$

Substituting this estimate into (47), we obtain

$$\|r_x^*\| \le \big(1 + \eta L_{\Phi_\gamma}\big)\|\mathcal{G}_\eta(\bar{x})\| + (2 + \eta L_{\Phi_\gamma})\|\nabla\Phi_\gamma(\bar{x}) - \widetilde{\nabla}\Phi_\gamma(\bar{x})\|.$$

Combining this inequality and (46), we obtain (44).

Now, since $\tilde{y}_\gamma^* = y_\gamma^*(\tilde{F}(\tilde{x}_\gamma^*))$, by the optimality condition of (9), we have

$$r_y^* := \gamma\nabla b(\tilde{y}_\gamma^*) + K^\top(F(\tilde{x}_\gamma^*) - \tilde{F}(\tilde{x}_\gamma^*)) \in K^\top F(\tilde{x}_\gamma^*) - \partial\psi(\tilde{y}_\gamma^*). \tag{49}$$

Utilizing Lemma A.1(d), we can bound r_y^* defined by (49) as

$$\|r_y^*\| \le \gamma\|\nabla b(\tilde{y}_\gamma^*)\| + \|K\|\|F(\tilde{x}_\gamma^*) - \tilde{F}(\tilde{x}_\gamma^*)\| \le \gamma D_\psi + \|K\|\|F(\tilde{x}_\gamma^*) - \tilde{F}(\tilde{x}_\gamma^*)\|.$$

Combining this estimate and (49), we get

$$\mathrm{dist}\Big(0, K^\top F(\tilde{x}^*_\gamma) - \partial\psi(\tilde{y}^*_\gamma)\Big) \le \|K\|\|F(\tilde{x}^*_\gamma) - \tilde{F}(\tilde{x}^*_\gamma)\| + \gamma D_\psi. \tag{50}$$

On the other hand, using the definition of $\widetilde{\nabla}\Phi_\gamma(\cdot)$ from (43), we can show that

$$\begin{aligned}
\|\widetilde{\nabla}\Phi_\gamma(\bar{x}) - \nabla\Phi_\gamma(\bar{x})\| &= \|\tilde{J}(\bar{x})^\top\nabla\phi_\gamma(\tilde{F}(\bar{x})) - F'(\bar{x})^\top\nabla\phi_\gamma(F(\bar{x}))\| \\
&\le \|(\tilde{J}(\bar{x}) - F'(\bar{x}))^\top\nabla\phi_\gamma(\tilde{F}(\bar{x}))\| + \|F'(\bar{x})^\top\big(\nabla\phi_\gamma(\tilde{F}(\bar{x})) - \nabla\phi_\gamma(F(\bar{x})\big)\| \\
&\le \|(\tilde{J}(\bar{x}) - F'(\bar{x}))^\top\nabla\phi_\gamma(\tilde{F}(\bar{x}))\| + \|\nabla\phi_\gamma(\tilde{F}(\bar{x})) - \nabla\phi_\gamma(F(\bar{x}))\|\|F'(\bar{x})\| \\
&\overset{(i)}{\le} \|(\tilde{J}(\bar{x}) - F'(\bar{x}))^\top\nabla\phi_\gamma(\tilde{F}(\bar{x}))\| + L_{\phi_\gamma}\|F'(\bar{x})\|\|\tilde{F}(\bar{x}) - F(\bar{x})\| \\
&\overset{(5)}{\le} \|(\tilde{J}(\bar{x}) - F'(\bar{x}))^\top\nabla\phi_\gamma(\tilde{F}(\bar{x}))\| + L_{\phi_\gamma}M_F\|\tilde{F}(\bar{x}) - F(\bar{x})\|.
\end{aligned}$$

Here, we have used the L_{ϕ_γ}-smoothness of ϕ_γ in *(i)*.

Finally, combining the last estimate, (44), and (50), and using the definition of $\mathcal{E}$ from (8), we have

$$\begin{aligned}
\mathcal{E}(\tilde{x}^*_\gamma, \tilde{y}^*_\gamma) &:= \mathrm{dist}\left(0, \nabla\Phi_\gamma(\tilde{x}^*_\gamma) + \partial\mathcal{R}(\tilde{x}^*_\gamma)\right) + \mathrm{dist}\left(0, K^\top F(\tilde{x}^*_\gamma) - \partial\psi(\tilde{y}^*_\gamma)\right) \\
&\le \left(1 + \eta L_{\Phi_\gamma}\right)\|\mathcal{G}_\eta(\bar{x})\| + (2 + \eta L_{\Phi_\gamma})\|\nabla\Phi_\gamma(\bar{x}) - \widetilde{\nabla}\Phi_\gamma(\bar{x})\| \\
&\quad + \|K\|\|F(\tilde{x}^*_\gamma) - \tilde{F}(\tilde{x}^*_\gamma)\| + \gamma D_\psi \\
&\le \left(1 + \eta L_{\Phi_\gamma}\right)\|\mathcal{G}_\eta(\bar{x})\| + \gamma D_\psi + \|K\|\|F(\tilde{x}^*_\gamma) - \tilde{F}(\tilde{x}^*_\gamma)\| \\
&\quad + \left(2 + \eta L_{\Phi_\gamma}\right)\left[\|(\tilde{J}(\bar{x}) - F'(\bar{x}))^\top\nabla\phi_\gamma(\tilde{F}(\bar{x}))\| + L_{\phi_\gamma}M_F\|\tilde{F}(\bar{x}) - F(\bar{x})\|\right],
\end{aligned}$$

which proves (45). □

The proof of Lemma 3.1. For notational simplicity, we drop the subscript T in this proof. Since $M_{\phi_\gamma} = M_\psi\|K\|$ and $L_{\phi_\gamma} = \frac{\|K\|^2}{\gamma+\mu_\psi}$, using the conditions in Lemma 3.1 and (27), we can derive from (45) after taking the full expectation that

$$\begin{aligned}
\mathbb{E}\big[\mathcal{E}(\tilde{x}^*_\gamma, \tilde{y}^*_\gamma)\big] &\le \left(1 + \eta L_{\Phi_\gamma}\right)\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x})\|\big] + \left(2 + \eta L_{\Phi_\gamma}\right)\mathbb{E}\big[\|(\tilde{J}(\bar{x}) - F'(\bar{x}))^\top\nabla\phi_\gamma(\tilde{F}(\bar{x}))\|\big] \\
&\quad + \|K\|\mathbb{E}\big[\|F(\tilde{x}^*_\gamma) - \tilde{F}(\tilde{x}^*_\gamma)\|\big] + \left(2 + \eta L_{\Phi_\gamma}\right)\tfrac{\|K\|^2M_F}{\mu_\psi+\gamma}\mathbb{E}\big[\|\tilde{F}(\bar{x}) - F(\bar{x})\|\big] + \gamma D_\psi.
\end{aligned}$$

Now, by the Jensen inequality $\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x})\|\big] \le \big(\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x})\|^2\big]\big)^{1/2} \le \varepsilon$. In addition, by (27), we also have $0 < \gamma \le c_2\epsilon$, $\mathbb{E}\big[\|(\tilde{J}(\bar{x}) - F'(\bar{x}))^\top\nabla\phi_\gamma(\tilde{F}(\bar{x}))\|\big] \le \varepsilon$, $\mathbb{E}\big[\|F(\tilde{x}^*_\gamma) - \tilde{F}(\tilde{x}^*_\gamma)\|\big] \le \varepsilon$, and $\frac{1}{\mu_\psi+\gamma}\mathbb{E}\big[\|\tilde{F}(\bar{x}) - F(\bar{x})\|\big] \le \varepsilon$. By the update rule of η in Theorems 3.1, 3.2, 3.3, and 3.4, we have $\eta L_{\Phi_\gamma} = \frac{2}{3+\theta} \le \frac{2}{3}$ since $\theta \in (0, 1]$. Substituting these expressions into the last inequality, we finally arrive at

$$\mathbb{E}\big[\mathcal{E}(\tilde{x}^*_\gamma, \tilde{y}^*_\gamma)\big] \le (1 + \tfrac{2}{3})\varepsilon + c_2D_\psi\varepsilon + \|K\|\varepsilon + (2 + \tfrac{2}{3})(1 + \|K\|^2M_F)\varepsilon,$$

which is exactly (29). □

B Convergence analysis of Algorithm 1 in Section 3

This appendix provides the full analysis of Algorithm 1, including convergence rates and oracle complexity for both strongly convex and non-strongly convex cases of ψ (or equivalently, the smoothness and the non-smoothness of ϕ_0, respectively).

B.1 Preparing technical results

Let us first recall and prove some technical results to prepare for our convergence analysis.

Lemma B.1. *Let $\{x_t\}$ be generated by Algorithm 1, $L_{\Phi_{\gamma_t}}$ be defined by (12), and B_ψ be given in Lemma A.1. Then, under Assumptions 2.1 and 2.2, for any $\eta_t > 0$ and $\theta_t \in [0,1]$, we have*

$$\begin{aligned}\mathbb{E}\big[\Psi_{\gamma_t}(x_{t+1})\big] \le\ & \mathbb{E}\big[\Psi_{\gamma_{t-1}}(x_t)\big] + \tfrac{\theta_t\left(1+L^2_{\Phi_{\gamma_t}}\eta_t^2\right)}{2L_{\Phi_{\gamma_t}}}\mathbb{E}\big[\|\nabla\Phi_{\gamma_t}(x_t) - v_t\|^2\big] + (\gamma_{t-1}-\gamma_t)B_\psi\\ & - \tfrac{L_{\Phi_{\gamma_t}}\eta_t^2\theta_t}{4}\mathbb{E}\big[\|\mathcal{G}_{\eta_t}(x_t)\|^2\big] - \tfrac{\theta_t}{2}\left(\tfrac{2}{\eta_t} - L_{\Phi_{\gamma_t}}\theta_t - 2L_{\Phi_{\gamma_t}}\right)\mathbb{E}\big[\|\hat{x}_{t+1}-x_t\|^2\big].\end{aligned} \tag{51}$$

Proof. Following the same line of proof of [29, Lemma 5], we can show that

$$\begin{aligned}\mathbb{E}\big[\Psi_{\gamma_t}(x_{t+1})\big] \ \le\ & \mathbb{E}\big[\Psi_{\gamma_t}(x_t)\big] + \tfrac{\theta_t\left(1+L^2_{\Phi_{\gamma_t}}\eta_t^2\right)}{2L_{\Phi_{\gamma_t}}}\mathbb{E}\big[\|\nabla\Phi_{\gamma_t}(x_t) - v_t\|^2\big]\\ & - \tfrac{L_{\Phi_{\gamma_t}}\eta_t^2\theta_t}{4}\mathbb{E}\big[\|\mathcal{G}_{\eta_t}(x_t)\|^2\big] - \tfrac{\theta_t}{2}\left(\tfrac{2}{\eta_t} - L_{\Phi_{\gamma_t}}\theta_t - 2L_{\Phi_{\gamma_t}}\right)\mathbb{E}\big[\|\hat{x}_{t+1}-x_t\|^2\big].\end{aligned}$$

Finally, since $\mathbb{E}\big[\Psi_{\gamma_t}(x_t)\big] \le \mathbb{E}\big[\Psi_{\gamma_{t-1}}(x_t)\big] + (\gamma_{t-1}-\gamma_t)B_\psi$ due to Lemma A.1(e), substituting this expression into the last inequality, we obtain (51). □

The Lyapunov function. To analyze Algorithm 1, we introduce the following Lyapunov function:

$$V_{\gamma_{t-1}}(x_t) := \mathbb{E}\big[\Psi_{\gamma_{t-1}}(x_t)\big] + \frac{\alpha_t}{2}\mathbb{E}\big[\|\tilde{F}_t - F(x_t)\|^2\big] + \frac{\hat{\alpha}_t}{2}\mathbb{E}\big[\|\tilde{J}_t - F'(x_t)\|^2\big], \tag{52}$$

where $\alpha_t > 0$ and $\hat{\alpha}_t > 0$ are given parameters, and the expectation is taken over $\mathcal{F}_{t+1}$. Lemma B.2 provides a key bound to estimate convergence rates and complexity bounds.

Lemma B.2. *Let $\{x_t\}$ be generated by Algorithm 1, and V_{γ_t} be the Lyapunov function defined by (52). Suppose further that the following conditions hold:*

$$\begin{cases} \frac{2}{\eta_t} \ge L_{\Phi_{\gamma_t}}\theta_t + 2L_{\Phi_{\gamma_t}} + \frac{M_F^2\beta_t^2\theta_t\alpha_{t+1}}{b_1} + \frac{L_F^2\hat{\beta}_t^2\theta_t\hat{\alpha}_{t+1}}{\hat{b}_1}\\ 2M_F^2L^2_{\phi_{\gamma_t}}\theta_t\Big(\frac{1+L^2_{\Phi_{\gamma_t}}\eta_t^2}{L_{\Phi_{\gamma_t}}}\Big) + \alpha_{t+1}\beta_t^2 \le \alpha_t \quad \textit{and} \quad 2M^2_{\phi_{\gamma_t}}\theta_t\Big(\frac{1+L^2_{\Phi_{\gamma_t}}\eta_t^2}{L_{\Phi_{\gamma_t}}}\Big) + \hat{\alpha}_{t+1}\hat{\beta}_t^2 \le \hat{\alpha}_t. \end{cases} \tag{53}$$

Then, for all $t \ge 0$, one has

$$\begin{aligned}V_{\gamma_t}(x_{t+1}) \ \le\ & V_{\gamma_{t-1}}(x_t) - \tfrac{L_{\Phi_{\gamma_t}}\eta_t^2\theta_t}{4}\mathbb{E}\big[\|\mathcal{G}_{\eta_t}(x_t)\|^2\big] + \tfrac{(1-\beta_t)^2\alpha_{t+1}\sigma_F^2}{b_2} + \tfrac{(1-\hat{\beta}_t)^2\hat{\alpha}_{t+1}\sigma_J^2}{\hat{b}_2}\\ & + (\gamma_{t-1}-\gamma_t)B_\psi.\end{aligned} \tag{54}$$

Proof. First of all, by combining (36) and (51), we obtain

$$\begin{aligned}\mathbb{E}\big[\Psi_{\gamma_t}(x_{t+1})\big] \le\ & \mathbb{E}\big[\Psi_{\gamma_{t-1}}(x_t)\big] - \tfrac{\theta_t}{2}\left(\tfrac{2}{\eta_t} - L_{\Phi_{\gamma_t}}\theta_t - 2L_{\Phi_{\gamma_t}}\right)\mathbb{E}\big[\|\hat{x}_{t+1}-x_t\|^2\big]\\ & - \tfrac{L_{\Phi_{\gamma_t}}\eta_t^2\theta_t}{4}\mathbb{E}\big[\|\mathcal{G}_{\eta_t}(x_t)\|^2\big] + (\gamma_{t-1}-\gamma_t)B_\psi\\ & + \theta_t\Big(\tfrac{1+L^2_{\Phi_{\gamma_t}}\eta_t^2}{L_{\Phi_{\gamma_t}}}\Big)\Big(M_F^2L^2_{\phi_{\gamma_t}}\mathbb{E}\big[\|\tilde{F}_t - F(x_t)\|^2\big] + M^2_{\phi_{\gamma_t}}\mathbb{E}\big[\|\tilde{J}_t - F'(x_t)\|^2\big]\Big).\end{aligned} \tag{55}$$

Due to the mini-batch estimators in (14), it is well-known that

$$\begin{aligned}\mathbb{E}_{\mathcal{B}_t^2}\big[\|\mathbf{F}(x_t,\zeta_t) - F(x_t)\|^2\big] &= \mathbb{E}\big[\big\|\tfrac{1}{b_2}\textstyle\sum_{\zeta_i\in\mathcal{B}_t^2}\mathbf{F}(x_t,\zeta_i) - F(x_t)\big\|^2\big] \le \tfrac{\sigma_F^2}{b_2}\\ \mathbb{E}_{\hat{\mathcal{B}}_t^2}\big[\|\mathbf{F}'(x_t,\hat{\zeta}_t) - F'(x_t)\|^2\big] &= \mathbb{E}\big[\big\|\tfrac{1}{\hat{b}_2}\textstyle\sum_{\hat{\zeta}_i\in\hat{\mathcal{B}}^2}\mathbf{F}'(x_t,\hat{\zeta}_i) - F'(x_t)\big\|^2\big] \le \tfrac{\sigma_J^2}{\hat{b}_2}.\end{aligned}$$

Substituting these bounds and $x_{t+1}-x_t=\theta_t(\hat{x}_{t+1}-x_t)$ into (35) and taking full expectation the resulting inequality over $\mathcal{F}_{t+1}$, we obtain

$$\begin{aligned}\mathbb{E}\big[\|\tilde{F}_{t+1}-F(x_{t+1})\|^2\big] &\leq \beta_t^2\mathbb{E}\big[\|\tilde{F}_t-F(x_t)\|^2\big]+\tfrac{\beta_t^2\theta_t^2M_F^2}{b_1}\mathbb{E}\big[\|\hat{x}_{t+1}-x_t\|^2\big]+\tfrac{(1-\beta_t)^2\sigma_F^2}{b_2}\\ \mathbb{E}\big[\|\tilde{J}_{t+1}-F'(x_{t+1})\|^2\big] &\leq \hat{\beta}_t^{\,2}\mathbb{E}\big[\|\tilde{J}_t-F'(x_t)\|^2\big]+\tfrac{\hat{\beta}_t^2\theta_t^2L_F^2}{\hat{b}_1}\mathbb{E}\big[\|\hat{x}_{t+1}-x_t\|^2\big]+\tfrac{(1-\hat{\beta}_t)^2\sigma_J^2}{\hat{b}_2}.\end{aligned}$$

Multiplying these inequalities by $\alpha_{t+1}>0$ and $\hat{\alpha}_{t+1}>0$, respectively, and adding the results to (55), we can further derive

$$\begin{aligned}V_{\gamma_t}(x_{t+1}) &\overset{(52)}{:=} \mathbb{E}\big[\Psi_{\gamma_t}(x_{t+1})\big]+\tfrac{\alpha_{t+1}}{2}\mathbb{E}\big[\|\tilde{F}_{t+1}-F(x_{t+1})\|^2\big]+\tfrac{\hat{\alpha}_{t+1}}{2}\mathbb{E}\big[\|\tilde{J}_{t+1}-F'(x_{t+1})\|^2\big]\\ &\leq \mathbb{E}\big[\Psi_{\gamma_{t-1}}(x_t)\big]+\left[M_F^2L_{\phi_{\gamma_t}}^2\theta_t\Big(\tfrac{1+L_{\Phi_{\gamma_t}}^2\eta_t^2}{L_{\Phi_{\gamma_t}}}\Big)+\tfrac{\alpha_{t+1}\beta_t^2}{2}\right]\mathbb{E}\big[\|\tilde{F}_t-F(x_t)\|^2\big]\\ &\quad+\left[M_{\phi_{\gamma_t}}^2\theta_t\Big(\tfrac{1+L_{\Phi_{\gamma_t}}^2\eta_t^2}{L_{\Phi_{\gamma_t}}}\Big)+\tfrac{\hat{\alpha}_{t+1}\hat{\beta}_t^2}{2}\right]\mathbb{E}\big[\|\tilde{J}_t-F'(x_t)\|^2\big]-\tfrac{L_{\Phi_{\gamma_t}}\eta_t^2\theta_t}{4}\mathbb{E}\big[\|\mathcal{G}_{\eta_t}(x_t)\|^2\big]\\ &\quad-\tfrac{\theta_t}{2}\left(\tfrac{2}{\eta_t}-L_{\Phi_{\gamma_t}}\theta_t-2L_{\Phi_{\gamma_t}}-\tfrac{M_F^2\beta_t^2\theta_t\alpha_{t+1}}{b_1}-\tfrac{L_F^2\hat{\beta}_t^2\theta_t\hat{\alpha}_{t+1}}{\hat{b}_1}\right)\mathbb{E}\big[\|\hat{x}_{t+1}-x_t\|^2\big]\\ &\quad+\tfrac{(1-\beta_t)^2\alpha_{t+1}\sigma_F^2}{b_2}+\tfrac{(1-\hat{\beta}_t)^2\hat{\alpha}_{t+1}\sigma_J^2}{\hat{b}_2}+(\gamma_{t-1}-\gamma_t)B_\psi.\end{aligned}$$

Let us choose $\alpha_t>0$ and $\hat{\alpha}_t>0$ and impose three conditions as in (53), i.e.:

$$\begin{cases}\frac{2}{\eta_t}\geq L_{\Phi_{\gamma_t}}\theta_t+2L_{\Phi_{\gamma_t}}+\frac{M_F^2\beta_t^2\theta_t\alpha_{t+1}}{b_1}+\frac{L_F^2\hat{\beta}_t^2\theta_t\hat{\alpha}_{t+1}}{\hat{b}_1},\\ 2M_F^2L_{\phi_{\gamma_t}}^2\theta_t\Big(\frac{1+L_{\Phi_{\gamma_t}}^2\eta_t^2}{L_{\Phi_{\gamma_t}}}\Big)+\alpha_{t+1}\beta_t^2\leq\alpha_t,\quad\text{and}\quad 2M_{\phi_{\gamma_t}}^2\theta_t\Big(\frac{1+L_{\Phi_{\gamma_t}}^2\eta_t^2}{L_{\Phi_{\gamma_t}}}\Big)+\hat{\alpha}_{t+1}\hat{\beta}_t^2\leq\hat{\alpha}_t.\end{cases}$$

Then, by using (52), the last inequality can be further upper bounded as

$$\begin{aligned}V_{\gamma_t}(x_{t+1}) &\leq V_{\gamma_{t-1}}(x_t)-\tfrac{L_{\Phi_{\gamma_t}}\eta_t^2\theta_t}{4}\mathbb{E}\big[\|\mathcal{G}_{\eta_t}(x_t)\|^2\big]+\tfrac{(1-\beta_t)^2\alpha_{t+1}\sigma_F^2}{b_2}\\ &\quad+\tfrac{(1-\hat{\beta}_t)^2\hat{\alpha}_{t+1}\sigma_J^2}{\hat{b}_2}+(\gamma_{t-1}-\gamma_t)B_\psi,\end{aligned}$$

which proves (54). □

B.2 A general key bound for Algorithm 1

Now, we are ready to prove one key result, Theorem B.1, for oracle complexity analysis of Algorithm 1. To simplify our expressions, let us introduce the following notations in advance:

$$\begin{cases}\omega_t &:= \frac{\theta_t}{L_{\Phi_{\gamma_t}}}\quad\text{and}\quad\Sigma_T:=\sum_{t=0}^T\omega_t,\\ \Theta_t &:= \frac{M_F^2L_{\phi_{\gamma_t}}^2\sqrt{26b_1\hat{b}_1}}{3\big(M_F^4L_{\phi_{\gamma_t}}^2\hat{b}_1+M_{\phi_{\gamma_t}}^2L_F^2b_1\big)^{1/2}},\\ \Pi_0 &:= \frac{\sqrt{26b_1\hat{b}_1}}{3\big(\hat{b}_1M_F^4L_{\phi_{\gamma_0}}^2+b_1L_F^2M_{\phi_{\gamma_0}}^2\big)^{1/2}}\left(\frac{M_F^2L_{\phi_{\gamma_0}}^2\sigma_F^2}{b_0}+\frac{M_{\phi_{\gamma_0}}^2\sigma_J^2}{\hat{b}_0}\right),\\ \Gamma_t &:= \frac{\sqrt{26b_1\hat{b}_1}}{3\big(\hat{b}_1M_F^4L_{\phi_{\gamma_t}}^2+b_1L_F^2M_{\phi_{\gamma_t}}^2\big)^{1/2}}\left(\frac{M_F^2L_{\phi_{\gamma_t}}^2\sigma_F^2}{b_2}+\frac{M_{\phi_{\gamma_t}}^2\sigma_J^2}{\hat{b}_2}\right).\end{cases}\tag{56}$$

Theorem B.1. *Suppose that Assumptions 2.1 and 2.2 hold, and* ω_t, Σ_T, Θ_t, Π_0, *and* Γ_t *are defined by* (56). *Let* $\{x_t\}_{t=0}^T$ *be generated by Algorithm 1 using the following step-sizes:*

$$\theta_t := \frac{3L_{\Phi_{\gamma_t}}\big[b_1\hat{b}_1(1-\beta_t)\big]^{1/2}}{\sqrt{26}(M_F^4L_{\phi_{\gamma_t}}^2\hat{b}_1 + M_{\phi_{\gamma_t}}^2L_F^2b_1)^{1/2}} \quad \textit{and} \quad \eta_t := \frac{2}{L_{\Phi_{\gamma_t}}(3+\theta_t)}, \tag{57}$$

where $\beta_t, \hat{\beta}_t \in (0,1]$ *are chosen such that* $\beta_t = \hat{\beta}_t$, $0 \le \gamma_{t+1} \le \gamma_t$, *and*

$$\frac{\beta_t^2(1-\beta_t)}{\Theta_t^2} \le \frac{1-\beta_{t+1}}{\Theta_{t+1}^2} \le \frac{1-\beta_t}{\Theta_t^2} \quad \textit{and} \quad \beta_t > \max\left\{0, 1 - \tfrac{26}{9L_{\Phi_{\gamma_t}}^2}\Big(\tfrac{M_F^4L_{\phi_{\gamma_t}}^2}{b_1} + \tfrac{L_F^2M_{\phi_{\gamma_t}}^2}{\hat{b}_1}\Big)\right\}. \tag{58}$$

Let $\bar{x}_T$ *be randomly chosen between* $\{x_0, \cdots, x_T\}$ *such that* $\mathbf{Prob}\,(\bar{x}_T = x_t) = \frac{\omega_t}{\Sigma_T}$, *and* $\bar{\eta}_T$ *be corresponding to* η_t *of* $\bar{x}_T$. *Then, the following estimate holds:*

$$\mathbb{E}\big[\|\mathcal{G}_{\bar{\eta}_T}(\bar{x}_T)\|^2\big] \le \frac{16}{\Sigma_T}\Big(\mathbb{E}\big[\Psi_0(x_0) - \Psi_0^\star\big] + \gamma_T B_\psi\Big) + \frac{8\Pi_0}{\Sigma_T\sqrt{1-\beta_0}} + \frac{16}{\Sigma_T}\sum_{t=0}^T\frac{\Gamma_{t+1}(1-\beta_t)^2}{\sqrt{1-\beta_{t+1}}}. \tag{59}$$

The proof of Theorem B.1. First, the conditions in (53) can be simplified as follows:

$$\begin{cases} L_{\Phi_{\gamma_t}}\theta_t + 2L_{\Phi_{\gamma_t}} + \big(\frac{M_F^2\beta_t^2\alpha_{t+1}}{b_1} + \frac{L_F^2\hat{\beta}_t^{\,2}\hat{\alpha}_{t+1}}{\hat{b}_1}\big)\theta_t &\le \frac{2}{\eta_t}, \\ 2M_F^2L_{\phi_{\gamma_t}}^2(1 + L_{\Phi_{\gamma_t}}^2\eta_t^2)\theta_t &\le L_{\Phi_{\gamma_t}}(\alpha_t - \beta_t^2\alpha_{t+1}), \\ 2M_{\phi_{\gamma_t}}^2(1 + L_{\Phi_{\gamma_t}}^2\eta_t^2)\theta_t &\le L_{\Phi_{\gamma_t}}(\hat{\alpha}_t - \hat{\beta}_t^2\hat{\alpha}_{t+1}). \end{cases} \tag{60}$$

Let us update $\eta_t := \frac{2}{(3+\theta_t)L_{\Phi_{\gamma_t}}}$ as (57). Since $\theta_t \in (0,1]$, we have

$$\frac{1}{2L_{\Phi_{\gamma_t}}} \le \eta_t < \frac{2}{3L_{\Phi_{\gamma_t}}} \quad \text{and} \quad 1 \le 1 + L_{\Phi_{\gamma_t}}^2\eta_t^2 < \frac{13}{9}.$$

Next, let us choose γ_t, β_t, $\hat{\beta}_t$, α_t, and $\hat{\alpha}_t$ such that

$$\hat{\beta}_t = \beta_t \in (0,1], \quad \hat{\alpha}_t = \frac{M_{\phi_{\gamma_t}}^2}{M_F^2L_{\phi_{\gamma_t}}^2}\alpha_t, \quad \frac{M_{\phi_{\gamma_{t+1}}}}{L_{\phi_{\gamma_{t+1}}}} \le \frac{M_{\phi_{\gamma_t}}}{L_{\phi_{\gamma_t}}}, \quad \text{and} \quad 0 < \alpha_t \le \alpha_{t+1} \le \frac{\alpha_t}{\beta_t}. \tag{61}$$

Then, we have

$$\begin{aligned} \alpha_t - \alpha_{t+1}\beta_t^2 &\ge \alpha_t(1-\beta_t) > 0, \\ \text{and} \quad \hat{\alpha}_t - \hat{\beta}_t^2\hat{\alpha}_{t+1} &= \tfrac{M_{\phi_{\gamma_t}}^2}{M_F^2L_{\phi_{\gamma_t}}^2}\alpha_t - \beta_t^2\tfrac{M_{\phi_{\gamma_{t+1}}}^2}{M_F^2L_{\phi_{\gamma_{t+1}}}^2}\alpha_{t+1} \ge \tfrac{M_{\phi_{\gamma_t}}^2}{M_F^2L_{\phi_{\gamma_t}}^2}(\alpha_t - \beta_t^2\alpha_{t+1}) \\ &\ge \tfrac{M_{\phi_{\gamma_t}}^2}{M_F^2L_{\phi_{\gamma_t}}^2}(1-\beta_t)\alpha_t = (1-\beta_t)\hat{\alpha}_t > 0. \end{aligned}$$

By using the last two inequalities, we can show that the conditions in (60) hold, if we have

$$\begin{aligned} &0 < \theta_t \le \tfrac{9L_{\Phi_{\gamma_t}}\alpha_t(1-\beta_t)}{26M_F^2L_{\phi_{\gamma_t}}^2}, \qquad 0 < \theta_t \le \tfrac{9L_{\Phi_{\gamma_t}}\hat{\alpha}_t(1-\beta_t)}{26M_{\phi_{\gamma_t}}^2}, \\ \text{and} \quad &0 < \theta_t \le L_{\Phi_{\gamma_t}}\Big(\tfrac{M_F^2\alpha_t}{b_1} + \tfrac{L_F^2\hat{\alpha}_t}{\hat{b}_1}\Big)^{-1}. \end{aligned} \tag{62}$$

Therefore, the three conditions in (62) hold if we choose

$$\frac{\alpha_t(1-\beta_t)}{M_F^2L_{\phi_{\gamma_t}}^2} = \frac{\hat{\alpha}_t(1-\beta_t)}{M_{\phi_{\gamma_t}}^2} \quad \text{and} \quad \left(\frac{M_F^2}{b_1} + \frac{L_F^2M_{\phi_{\gamma_t}}^2}{M_F^2L_{\phi_{\gamma_t}}^2\hat{b}_1}\right)\alpha_t = \frac{26M_F^2L_{\phi_{\gamma_t}}^2}{9\alpha_t(1-\beta_t)}.$$

These conditions show that we can choose

$$\alpha_t := \tfrac{\Theta_t}{\sqrt{1-\beta_t}} \quad \text{and} \quad \hat{\alpha}_t := \tfrac{M_{\phi_{\gamma_t}}^2\Theta_t}{M_F^2L_{\phi_{\gamma_t}}^2\sqrt{1-\beta_t}}, \quad \text{where} \quad \Theta_t := \tfrac{M_F^2L_{\phi_{\gamma_t}}^2\sqrt{26b_1\hat{b}_1}}{3\left(M_F^4L_{\phi_{\gamma_t}}^2\hat{b}_1+M_{\phi_{\gamma_t}}^2L_F^2b_1\right)^{1/2}}.$$

Clearly, this Θ_t is exactly given by (56). With this choice of α_t and $\hat{\alpha}_t$, we obtain

$$0 < \theta_t \le \bar{\theta}_t := \frac{9L_{\Phi_{\gamma_t}}\Theta_t\sqrt{(1-\beta_t)}}{26M_F^2L_{\phi_{\gamma_t}}^2} = \frac{3L_{\Phi_{\gamma_t}}\sqrt{b_1\hat{b}_1(1-\beta_t)}}{\sqrt{26}(M_F^4L_{\phi_{\gamma_t}}^2\hat{b}_1 + M_{\phi_{\gamma_t}}^2L_F^2b_1)^{1/2}}.$$

We then choose $\theta_t := \bar{\theta}_t$ at the upper bound as in (57).

Now, to guarantee that $0 < \bar{\theta}_t \le 1$, we impose the following condition as in (58), i.e.:

$$\beta_t > \max\left\{0, 1 - \tfrac{26}{9L_{\Phi_{\gamma_t}}^2}\left(\tfrac{M_F^4L_{\phi_{\gamma_t}}^2}{b_1} + \tfrac{L_F^2M_{\phi_{\gamma_t}}^2}{\hat{b}_1}\right)\right\}.$$

Due to the choice of α_t, the condition $\alpha_t \le \alpha_{t+1} \le \frac{\alpha_t}{\beta_t}$ in (61) is equivalent to

$$\frac{\beta_t^2(1-\beta_t)}{\Theta_t^2} \le \frac{1-\beta_{t+1}}{\Theta_{t+1}^2} \le \frac{1-\beta_t}{\Theta_t^2},$$

which is the first condition of (58). Moreover, since $M_{\phi_{\gamma_t}} = M_\psi\|K\|$ and $L_{\phi_{\gamma_t}} = \frac{\|K\|^2}{\mu_\psi+\gamma_t}$ due to Lemma A.1, the third condition of (61) reduces to $\gamma_{t+1} \le \gamma_t$, which is one of the conditions in Theorem B.1.

Next, under the choice of α_t and $\hat{\alpha}_t$, and $\eta_t \ge \frac{1}{2L_{\Phi_{\gamma_t}}}$, (54) implies

$$\begin{aligned}
&\tfrac{\theta_t}{16L_{\Phi_{\gamma_t}}}\mathbb{E}\big[\|\mathcal{G}_{\eta_t}(x_t)\|^2\big] \le V_{\gamma_{t-1}}(x_t) - V_{\gamma_t}(x_{t+1}) + (\gamma_{t-1}-\gamma_t)B_\psi \\
&+ \tfrac{\sqrt{26b_1\hat{b}_1}}{3\left(\hat{b}_1M_F^4L_{\phi_{\gamma_{t+1}}}^2+b_1L_F^2M_{\phi_{\gamma_{t+1}}}^2\right)^{1/2}}\left(\tfrac{M_F^2L_{\phi_{\gamma_{t+1}}}^2\sigma_F^2}{b_2} + \tfrac{M_{\phi_{\gamma_{t+1}}}^2\sigma_J^2}{\hat{b}_2}\right)\tfrac{(1-\beta_t)^2}{(1-\beta_{t+1})^{1/2}}.
\end{aligned} \tag{63}$$

Note that since $\Psi_{\gamma_0}(x_0) \le \Psi_0(x_0)$ due to Lemma A.1, and $\gamma_{-1} = \gamma_0$ by convention, we have

$$\begin{aligned}
V_{\gamma_0}(x_0) &= \mathbb{E}\big[\Psi_{\gamma_0}(x_0)\big] + \tfrac{\alpha_0}{2}\mathbb{E}\big[\|\tilde{F}_0 - F(x_0)\|^2\big] + \tfrac{\hat{\alpha}_0}{2}\mathbb{E}\big[\|\tilde{J}_0 - F'(x_0)\|^2\big] \\
&\le \mathbb{E}\big[\Psi_0(x_0)\big] + \tfrac{\sqrt{26b_1\hat{b}_1}}{6\left(\hat{b}_1M_F^4L_{\phi_{\gamma_0}}^2+b_1L_F^2M_{\phi_{\gamma_0}}^2\right)^{1/2}}\left(\tfrac{M_F^2L_{\phi_{\gamma_0}}^2\sigma_F^2}{b_0} + \tfrac{M_{\phi_{\gamma_0}}^2\sigma_J^2}{\hat{b}_0}\right)\tfrac{1}{(1-\beta_0)^{1/2}}.
\end{aligned} \tag{64}$$

Moreover, by Lemma A.1(d), we have

$$V_{\gamma_T}(x_{T+1}) \ge \mathbb{E}\big[\Psi_{\gamma_T}(x_{T+1})\big] \ge \mathbb{E}\big[\Psi_0(x_{T+1})\big] - \gamma_TB_\psi \ge \Psi_0^\star - \gamma_TB_\psi. \tag{65}$$

Let us define Γ_t and Π_0 as (56), i.e.:

$$\left\{\begin{array}{lcl}
\Gamma_t &:=& \tfrac{\sqrt{26b_1\hat{b}_1}}{3\left(\hat{b}_1M_F^4L_{\phi_{\gamma_t}}^2+b_1L_F^2M_{\phi_{\gamma_t}}^2\right)^{1/2}}\left(\tfrac{M_F^2L_{\phi_{\gamma_t}}^2\sigma_F^2}{b_2} + \tfrac{M_{\phi_{\gamma_t}}^2\sigma_J^2}{\hat{b}_2}\right), \\
\Pi_0 &:=& \tfrac{\sqrt{26b_1\hat{b}_1}}{3\left(\hat{b}_1M_F^4L_{\phi_{\gamma_0}}^2+b_1L_F^2M_{\phi_{\gamma_0}}^2\right)^{1/2}}\left(\tfrac{M_F^2L_{\phi_{\gamma_0}}^2\sigma_F^2}{b_0} + \tfrac{M_{\phi_{\gamma_0}}^2\sigma_J^2}{\hat{b}_0}\right).
\end{array}\right.$$

Then, summing up (63) from $t := 0$ to $t := T$, and using these expressions, (64), and (65), we get

$$\sum_{t=0}^{T}\frac{\theta_t}{16L_{\Phi_{\gamma_t}}}\mathbb{E}\big[\|\mathcal{G}_{\eta_t}(x_t)\|^2\big] \le \mathbb{E}\big[\Psi_0(x_0)-\Psi_0^\star\big] + \gamma_T B_\psi + \sum_{t=0}^{T}\frac{\Gamma_{t+1}(1-\beta_t)^2}{(1-\beta_{t+1})^{1/2}} + \frac{\Pi_0}{2(1-\beta_0)^{1/2}}.$$

Dividing this inequality by $\frac{\Sigma_T}{16}$, where $\Sigma_T := \sum_{t=0}^{T}\omega_t \equiv \sum_{t=0}^{T}\frac{\theta_t}{L_{\Phi_{\gamma_t}}}$, we obtain

$$\begin{aligned}\frac{1}{\Sigma_T}\sum_{t=0}^{T}\omega_t\mathbb{E}\big[\|\mathcal{G}_{\eta_t}(x_t)\|^2\big] \le{}& \frac{16}{\Sigma_T}\Big(\mathbb{E}\big[\Psi_0(x_0)-\Psi_0^\star\big]+\gamma_T B_\psi\Big) + \frac{8\Pi_0}{\Sigma_T(1-\beta_0)^{1/2}}\\ &+ \frac{16}{\Sigma_T}\sum_{t=0}^{T}\frac{\Gamma_{t+1}(1-\beta_t)^2}{(1-\beta_{t+1})^{1/2}}.\end{aligned}$$

Finally, due to the choice of $\bar{x}_T$ and $\bar{\eta}_T$, we have $\frac{1}{\Sigma_T}\sum_{t=0}^{T}\omega_t\mathbb{E}\big[\|\mathcal{G}_{\eta_t}(x_t)\|^2\big] = \mathbb{E}\big[\|\mathcal{G}_{\bar{\eta}_T}(\bar{x}_T)\|^2\big]$. This relation together with the above estimate prove (59). □

B.3 The proof of Theorem 3.1: The smooth case with constant step-size

Now, we prove our first main result in the main text.

The proof of Theorem 3.1 in the main text. First, since $\mu_\psi = 1 > 0$, we can set $\gamma_t := 0$ for all $t \ge 0$. That means, we do not need to smooth ϕ_0 in (2). Hence, from (56), $\Theta_t = \Theta_0 = \frac{M_F^2 L_{\phi_0}\sqrt{26b_1\hat{b}_1}}{3\big(M_F^4L_{\phi_0}^2\hat{b}_1+M_{\phi_0}^2L_F^2b_1\big)^{1/2}}$ and $\frac{\omega_t}{\Sigma_T} = \frac{\theta_t}{\sum_{t=0}^{T}\theta_t}$, where L_{Φ_0} is defined by (18).

Next, given a batch size $b > 0$, let us choose the mini-batch sizes $b_0 := c_0\hat{b}_0 > 0$, $\hat{b}_1 = \hat{b}_2 := b > 0$, and $b_1 = b_2 := c_0 b$ for some $c_0 > 0$. We also choose a constant step-size $\theta_t := \theta \in (0, 1]$ and a constant weight $\beta_t := \beta \in (0, 1]$ for all $t \ge 0$. We also recall P, Q, and L_{Φ_0} defined by (18).

With this configuration, the first condition of (58) and $0 \le \gamma_{t+1} \le \gamma_t$ are automatically satisfied, while the second one becomes

$$\beta > \max\left\{0, 1-\tfrac{26}{9c_0L_{\Phi_0}^2 b}\big(M_F^4\|K\|^4 + c_0\|K\|^2L_F^2M_\psi^2\big)\right\} = \max\left\{0, 1-\tfrac{P^2}{L_{\Phi_0}^2 b}\right\}. \tag{66}$$

Moreover, we also obtain from (56), (57), and (18) that

$$\left\{\begin{array}{lcll}\theta_t &=& \theta = \frac{3L_{\Phi_0}\sqrt{c_0b(1-\beta)}}{\sqrt{26}(M_F^4\|K\|^4+c_0\|K\|^2M_\psi^2L_F^2)^{1/2}} & \overset{(18)}{=} \frac{L_{\Phi_0}[b(1-\beta)]^{1/2}}{P},\\ \Gamma_t &=& \Gamma = \frac{\sqrt{26}\big(M_F^2\|K\|^4\sigma_F^2+c_0\|K\|^2M_\psi^2\sigma_J^2\big)}{3\sqrt{c_0b}\big(M_F^4\|K\|^4+c_0\|K\|^2L_F^2M_\psi^2\big)^{1/2}} & \overset{(18)}{=} \frac{Q}{P\sqrt{b}},\\ \Pi_0 &=& \frac{\sqrt{26b}\big(M_F^2\|K\|^4\sigma_F^2+c_0\|K\|^2M_\psi^2\sigma_J^2\big)}{3\sqrt{c_0}\hat{b}_0\big(M_F^4\|K\|^4+c_0\|K\|^2L_F^2M_\psi^2\big)^{1/2}} & \overset{(18)}{=} \frac{Q\sqrt{b}}{P\hat{b}_0},\\ \Sigma_T &=& \sum_{t=0}^{T}\frac{\theta}{L_{\Phi_0}} = \frac{\theta(T+1)}{L_{\Phi_0}} & = \frac{(T+1)[b(1-\beta)]^{1/2}}{P}.\end{array}\right.$$

Furthermore, with these expressions of Γ_t, Π_0, and Σ_T, (59) reduces to

$$\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_T)\|^2\big] \le \tfrac{16P}{(T+1)[b(1-\beta)]^{1/2}}\mathbb{E}\big[\Psi_0(x_0)-\Psi_0^\star\big] + \tfrac{8Q}{\hat{b}_0(T+1)(1-\beta)} + \tfrac{16Q(1-\beta)}{b}.$$

Trading-off the term $\frac{1}{\hat{b}_0(1-\beta)(T+1)} + \frac{2(1-\beta)}{b}$ over $\beta \in (0,1]$, we obtain $\beta := 1 - \frac{b^{1/2}}{[\hat{b}_0(T+1)]^{1/2}}$, which has shown in (19). In this case, $\theta_t = \theta = \frac{L_{\Phi_0}[b(1-\beta)]^{1/2}}{P} = \frac{L_{\Phi_0} b^{3/4}}{P[\hat{b}_0(T+1)]^{1/4}}$ as shown in (19).

Now, let us choose $\hat{b}_0 := c_1^2[b(T+1)]^{1/3}$ for some $c_1 > 0$. Then, the last inequality leads to

$$\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_T)\|^2\big] \le \tfrac{16P\sqrt{c_1}}{[b(T+1)]^{2/3}}\big[\Psi_0(x_0) - \Psi_0^\star\big] + \tfrac{24Q}{2c_1[b(T+1)]^{2/3}}.$$

Hence, if we define Δ_0 as in (20), i.e.:

$$\Delta_0 := 16P\sqrt{c_1}\big[\Psi_0(x_0) - \Psi_0^\star\big] + \frac{24Q}{c_1},$$

then we obtain from the last inequality that (20) holds, i.e.:

$$\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_T)\|^2\big] \le \frac{\Delta_0}{[b(T+1)]^{2/3}}.$$

Consequently, for a given tolerance $\varepsilon > 0$, to obtain $\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_T)\|^2\big] \le \varepsilon^2$, we need at most $T := \big\lfloor \frac{\Delta_0^{3/2}}{b\varepsilon^3} \big\rfloor$ iterations. In this case, the total number of function evaluations $\mathbf{F}(x_t, \xi)$ is at most

$$\mathcal{T}_F := b_0 + (T+1)(2b_1 + b_2) = c_0 c_1^2[b(T+1)]^{1/3} + 3c_0(T+1)b = \frac{c_0 c_1^2 \Delta_0^{1/2}}{\varepsilon} + \frac{3c_0\Delta_0^{3/2}}{\varepsilon^3}.$$

Alternatively, the total number of Jacobian evaluations $\mathbf{F}'(x_t, \xi)$ is at most

$$\mathcal{T}_J := \hat{b}_0 + (T+1)(2\hat{b}_1 + \hat{b}_2) = c_1^2[b(T+1)]^{1/3} + 3(T+1)b = \frac{c_1^2\Delta_0^{1/2}}{\varepsilon} + \frac{3\Delta_0^{3/2}}{\varepsilon^3}.$$

Finally, since $\beta := 1 - \frac{b^{1/2}}{[\hat{b}_0(T+1)]^{1/2}}$, the condition (66) leads to $\frac{b^{1/2}}{[\hat{b}_0(T+1)]^{1/2}} < \frac{P^2}{L_{\Phi_0}^2 b}$, which is equivalent to $\frac{\hat{b}_0(T+1)}{b^3} > \frac{L_{\Phi_0}^4}{P^4}$ as shown in Theorem 3.1. □

B.4 The proof of Theorem 3.2: The smooth case with diminishing step-size

The proof of Theorem 3.2 in the main text. Similar to the proof of Theorem 3.1, with $\mu_\psi = 1 > 0$, we set $\gamma_t = 0$. Hence, we obtain $\Theta_t = \Theta_0 = \frac{M_F^2 L_{\phi_0}\sqrt{26 b_1 \hat{b}_1}}{3\big(M_F^4 L_{\phi_0}^2 \hat{b}_1 + M_{\phi_0}^2 L_F^2 b_1\big)^{1/2}}$ and $\frac{\omega_t}{\Sigma_T} = \frac{\theta_t}{\sum_{t=0}^T \theta_t}$.

Next, given a mini-batch size $b > 0$, let us choose the mini-batch sizes $b_0 := c_0\hat{b}_0$, $\hat{b}_1 = \hat{b}_2 := b$, and $b_1 = b_2 := c_0 b > 0$ for some $c_0 > 0$. With these choices, the condition (58) becomes

$$\beta_t^2(1-\beta_t) \le 1 - \beta_{t+1} \le 1 - \beta_t \text{ and } \beta_t > \max\left\{0, 1 - \tfrac{26}{9c_0 L_{\Phi_0}^2 b}\big(c_0 M_F^4 L_{\phi_0}^2 + L_F^2 M_{\phi_0}^2\big)\right\}. \tag{67}$$

Moreover, from (56) and (57), we have

$$\left\{\begin{array}{lcll} \theta_t & = & \frac{3L_{\Phi_0}\sqrt{c_0 b(1-\beta_t)}}{\sqrt{26}(M_F^4\|K\|^4 + c_0\|K\|^2 M_\psi^2 L_F^2)^{1/2}} & \overset{(18)}{=} \frac{L_{\Phi_0}[b(1-\beta_t)]^{1/2}}{P}, \\ \Gamma_t & = & \Gamma = \frac{\sqrt{26}\big(M_F^2\|K\|^4\sigma_F^2 + c_0\|K\|^2 M_\psi^2\sigma_J^2\big)}{3\sqrt{c_0 b}\big(M_F^4\|K\|^4 + c_0\|K\|^2 L_F^2 M_\psi^2\big)^{1/2}} & \overset{(18)}{=} \frac{Q}{P\sqrt{b}}, \\ \Pi_0 & = & \frac{\sqrt{26b}\big(M_F^2\|K\|^4\sigma_F^2 + c_0\|K\|^2 M_\psi^2\sigma_J^2\big)}{3\sqrt{c_0}\hat{b}_0\big(M_F^4\|K\|^4 + c_0\|K\|^2 L_F^2 M_\psi^2\big)^{1/2}} & \overset{(18)}{=} \frac{Q\sqrt{b}}{P\hat{b}_0}, \\ \Sigma_T & = & \sum_{t=0}^T \omega_t = \sum_{t=0}^T \frac{\theta_t}{L_{\Phi_0}} & = \frac{\sqrt{b}}{P}\sum_{t=0}^T\sqrt{1-\beta_t}. \end{array}\right.$$

Furthermore, with these expressions of Γ_t, Π_0, and Σ_T, (59) reduces to

$$\begin{aligned}\frac{1}{\sum_{t=0}^T\theta_t}\sum_{t=0}^T\theta_t\mathbb{E}\big[\|\mathcal{G}_{\eta_t}(x_t)\|^2\big] \le& \frac{16P}{\sqrt{b}\sum_{t=0}^T\sqrt{1-\beta_t}}\big[\Psi_0(x_0)-\Psi_0^\star\big] + \frac{8Q}{\hat{b}_0\sqrt{1-\beta_0}\sum_{t=0}^T\sqrt{1-\beta_t}}\\ &+ \frac{16Q}{b\sum_{t=0}^T\sqrt{1-\beta_t}}\sum_{t=0}^T\frac{(1-\beta_t)^2}{(1-\beta_{t+1})^{1/2}}.\end{aligned} \tag{68}$$

Let us choose $\beta_t := 1-\frac{1}{(t+2)^{2/3}} \in (0,1)$ as in (21). Then, it is easy to check that $\beta_t^2(1-\beta_t) \le 1-\beta_{t+1} \le 1-\beta_t$ after a few elementary calculations.

Moreover, we have $\theta_t := \frac{L_{\Phi_0}\sqrt{b}}{P(t+2)^{1/3}}$ as (21). In addition, one can easily show that

$$\begin{cases} \sum_{t=0}^T\sqrt{1-\beta_t} = \sum_{t=0}^T\frac{1}{(t+2)^{1/3}} \ge \int_2^{T+3}\frac{ds}{s^{1/3}} = \frac{3}{2}[(T+3)^{2/3}-2^{2/3}],\\ \sum_{t=0}^T\frac{(1-\beta_t)^2}{\sqrt{1-\beta_{t+1}}} = \sum_{t=0}^T\frac{(t+3)^{1/3}}{(t+2)^{4/3}} \le \sum_{t=0}^T\frac{1}{(t+1)} \le 1+\log(T+1).\end{cases}$$

Here, we use the fact that $\int_t^{t+1} r(s)ds \le r(t) \le \int_{t-1}^t r(s)ds$ for a nonnegative and monotonically decreasing function r.

Substituting these estimates and $\sqrt{1-\beta_0} = \frac{1}{2^{1/3}}$ into (68), we eventually obtain

$$\begin{aligned}\frac{1}{\sum_{t=0}^T\theta_t}\sum_{t=0}^T\theta_t\mathbb{E}\big[\|\mathcal{G}_{\eta_t}(x_t)\|^2\big] \le& \frac{32P}{3\sqrt{b}\big[(T+3)^{2/3}-2^{2/3}\big]}\big[\Psi_0(x_0)-\Psi_0^\star\big]\\ &+ \frac{16Q}{3\big[(T+3)^{2/3}-2^{2/3}\big]}\left[\frac{2^{1/3}}{\hat{b}_0}+\frac{2(1+\log(T+1))}{b}\right].\end{aligned}$$

Combining this inequality and $\frac{1}{\sum_{t=0}^T\theta_t}\sum_{t=0}^T\theta_t\mathbb{E}\big[\|\mathcal{G}_{\eta_t}(x_t)\|^2\big] = \mathbb{E}\big[\|\mathcal{G}_{\bar{\eta}_T}(\bar{x}_T)\|^2\big]$, we have proved (22) for $T \ge 0$. □

B.5 The proof of Theorem 3.3: The non-smooth case with constant step-size

The proof of Theorem 3.3 in the main text. Since $\mu_\psi = 0$, let us fix the smoothness parameter $\gamma_t = \gamma > 0$ and the weights $\beta_t = \hat{\beta}_t = \beta \in (0,1]$ for all $t \ge 0$. By Lemma A.1, we have

$$M_{\phi_\gamma} = M_\psi\|K\|, \quad L_{\phi_\gamma} = \frac{\|K\|^2}{\gamma}, \quad \text{and} \quad L_{\Phi_\gamma} = L_F M_\psi\|K\| + \frac{M_F^2\|K\|^2}{\gamma}.$$

Given batch sizes $b > 0$ and $\hat{b}_0 > 0$, for some $c_0 > 0$, let us also choose the mini-batch sizes as

$$\hat{b}_1 = \hat{b}_2 := b, \quad b_1 = b_2 := \frac{c_0 b}{\gamma^2}, \quad \text{and} \quad b_0 := \frac{c_0\hat{b}_0}{\gamma^2}.$$

Recall that P, Q, and L_{Φ_γ} are defined by (18). In this case, the quantities in (56) become

$$\begin{cases} \Theta_t := \Theta = \frac{M_F^2 L_{\phi_\gamma}\sqrt{26b_1\hat{b}_1}}{3\left(M_F^4L_{\phi_\gamma}^2\hat{b}_1 + M_{\phi_\gamma}^2L_F^2b_1\right)^{1/2}} = \frac{\sqrt{26c_0b}M_F^2\|K\|^2}{3\gamma(M_F^4\|K\|^4+c_0\|K\|^2M_\psi^2L_F^2)^{1/2}} \overset{(18)}{=} \frac{M_F^2\|K\|^2b^{1/2}}{\gamma P},\\ \Gamma_t := \Gamma = \frac{\sqrt{26b_1\hat{b}_1}}{3\left(\hat{b}_1M_F^4L_{\phi_\gamma}^2+b_1L_F^2M_{\phi_\gamma}^2\right)^{1/2}}\left(\frac{M_F^2L_{\phi_\gamma}^2\sigma_F^2}{b_2}+\frac{M_{\phi_\gamma}^2\sigma_J^2}{\hat{b}_2}\right) \overset{(18)}{=} \frac{Q}{P\sqrt{b}},\\ \Pi_0 := \frac{\sqrt{26b_1\hat{b}_1}}{3\left(\hat{b}_1M_F^4L_{\phi_\gamma}^2+b_1L_F^2M_{\phi_\gamma}^2\right)^{1/2}}\left(\frac{M_F^2L_{\phi_\gamma}^2\sigma_F^2}{b_0}+\frac{M_{\phi_\gamma}^2\sigma_J^2}{\hat{b}_0}\right) \overset{(18)}{=} \frac{Q\sqrt{b}}{P\hat{b}_0}.\end{cases}$$

Furthermore, the step-sizes in (57) also become

$$\left\{\begin{array}{lcl}\theta_t & := & \theta = \frac{3L_{\Phi_\gamma}[b_1\hat{b}_1(1-\beta)]^{1/2}}{\sqrt{26}(M_F^4L_{\phi_\gamma}^2\hat{b}_1+M_{\phi_\gamma}^2L_F^2b_1)^{1/2}} \overset{(18)}{=} \frac{L_{\Phi_\gamma}[b(1-\beta)]^{1/2}}{P},\\ \eta_t & := & \eta = \frac{2}{L_{\Phi_\gamma}(3+\theta)}.\end{array}\right.$$

Therefore, we have $\omega_t := \frac{\theta}{L_{\Phi_\gamma}}$ and

$$\Sigma_T := \textstyle\sum_{t=0}^{T}\omega_t = \frac{\theta(T+1)}{L_{\Phi_\gamma}} = \frac{(T+1)[b(1-\beta)]^{1/2}}{P}.$$

Substituting these expressions into (59), we can further derive

$$\begin{array}{rcl}\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_T)\|^2\big] & \leq & \frac{16P}{(T+1)[b(1-\beta)]^{1/2}}\Big(\mathbb{E}\big[\Psi_0(x_0)-\Psi_0^\star\big]+\gamma B_\psi\Big)\\ & & + 8Q\left[\frac{1}{\hat{b}_0(1-\beta)(T+1)}+\frac{2(1-\beta)}{b}\right].\end{array} \tag{69}$$

From the last term of (69), we can choose β as $\beta = 1-\frac{b^{1/2}}{[\hat{b}_0(T+1)]^{1/2}}$. In this case, (69) reduces to

$$\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_T)\|^2\big] \leq \frac{16P\hat{b}_0^{1/4}}{[b(T+1)]^{3/4}}\Big(\mathbb{E}\big[\Psi_0(x_0)-\Psi_0^\star\big]+\gamma B_\psi\Big)+\frac{24Q}{[b\hat{b}_0(T+1)]^{1/2}}. \tag{70}$$

Clearly, from (70), to achieve the best convergence rate, we need to choose $\hat{b}_0 := c_1^2[b(T+1)]^{1/3}$. Then, since we choose $0<\gamma\leq 1$ and $\mathbb{E}\big[\Psi_0(x_0)\big]=\Psi_0(x_0)$, (70) can be overestimated as

$$\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_T)\|^2\big] \leq \frac{\hat{\Delta}_0}{[b(T+1)]^{2/3}},$$

which proves (24), where $\hat{\Delta}_0$ is defined by (24), i.e.:

$$\hat{\Delta}_0 := 16P\sqrt{c_1}\big(\Psi_0(x_0)-\Psi_0^\star+B_\psi\big)+\tfrac{24Q}{c_1}.$$

Now, for any tolerance $\varepsilon>0$, to obtain $\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_T)\|^2\big]\leq\varepsilon^2$, we require at most $T := \left\lfloor\frac{\hat{\Delta}_0^{3/2}}{b\varepsilon^3}\right\rfloor$ iterations. In this case, the total number of function evaluations $\mathcal{T}_F$ is at most

$$\mathcal{T}_F \;:=\; b_0+(T+1)(2b_1+b_2) = \tfrac{c_0}{\gamma^2}c_1^2[b(T+1)]^{1/3}+\tfrac{3c_0}{\gamma^2}[b(T+1)] = \tfrac{c_0c_1^2\hat{\Delta}_0^{1/2}}{\gamma^2\varepsilon}+\tfrac{3c_0\hat{\Delta}_0^{3/2}}{\gamma^2\varepsilon^3}.$$

Alternatively, the total number of Jacobian evaluations $\mathcal{T}_J$ is at most

$$\mathcal{T}_J \;:=\; \hat{b}_0+(T+1)(2\hat{b}_1+\hat{b}_2) = c_1[b(T+1)]^{1/3}+3b(T+1) = \tfrac{c_1^2\hat{\Delta}_0^{1/2}}{\varepsilon}+\tfrac{3\hat{\Delta}_0^{3/2}}{\varepsilon^3}.$$

If we choose $\gamma := c_2\varepsilon$ for some $c_2>0$, then

$$\mathcal{T}_F := \frac{c_0c_1^2\hat{\Delta}_0^{1/2}}{c_2^2\varepsilon^3}+\frac{3c_0\hat{\Delta}_0^{3/2}}{c_2^2\varepsilon^5} = \mathcal{O}\left(\frac{\hat{\Delta}_0^{3/2}}{\varepsilon^5}\right),$$

which proves the last statement. □

B.6 The proof of Theorem 3.4: The nonsmooth case with diminishing step-size

The proof of Theorem 3.4 in the main text. Using the fact that $\mu_\psi = 0$, from Lemma A.1, we have

$$M_{\phi_{\gamma_t}} = M_\psi\|K\|, \quad L_{\phi_{\gamma_t}} = \frac{\|K\|^2}{\gamma_t}, \quad \text{and} \quad L_{\Phi_{\gamma_t}} = L_F M_\psi\|K\| + \frac{M_F^2\|K\|^2}{\gamma_t},$$

where $\gamma_t > 0$, which will be appropriately updated. Moreover, let us choose $b_0 := \frac{c_0\hat{b}_0}{\gamma_0^2}$, $\hat{b}_1 = \hat{b}_2 := b$, and $b_1^t = b_2^t := \frac{c_0 b}{\gamma_t^2} > 0$, for some $b > 0$ and $c_0 > 0$. We also recall P, Q, and L_{Φ_γ} from (18).

With these expressions, the quantities defined by (56) and (57) become

$$\begin{cases} \theta_t &:= \frac{3L_{\Phi_{\gamma_t}}[b_1^t\hat{b}_1(1-\beta_t)]^{1/2}}{\sqrt{26}(M_F^4L_{\phi_{\gamma_t}}^2\hat{b}_1+M_{\phi_{\gamma_t}}^2L_F^2b_1^t)^{1/2}} \overset{(18)}{=} \frac{L_{\Phi_{\gamma_t}}[b(1-\beta_t)]^{1/2}}{P}, \\ \Theta_t &:= \frac{M_F^2L_{\phi_{\gamma_t}}\sqrt{26b_1^t\hat{b}_1}}{3\left(M_F^4L_{\phi_{\gamma_t}}^2\hat{b}_1+M_{\phi_{\gamma_t}}^2L_F^2b_1^t\right)^{1/2}} \overset{(18)}{=} \frac{M_F^2\|K\|^2b^{1/2}}{\gamma_t P}, \\ \Gamma_t &:= \frac{\sqrt{26b_1^t\hat{b}_1}}{3\left(\hat{b}_1M_F^4L_{\phi_{\gamma_t}}^2+b_1^tL_F^2M_{\phi_{\gamma_t}}^2\right)^{1/2}}\left(\frac{M_F^2L_{\phi_{\gamma_t}}^2\sigma_F^2}{b_2^t}+\frac{M_{\phi_{\gamma_t}}^2\sigma_J^2}{\hat{b}_2}\right) \overset{(18)}{=} \frac{Q}{P\sqrt{b}}, \\ \Pi_0 &:= \frac{\sqrt{26b_1^0\hat{b}_1}}{3\left(\hat{b}_1M_F^4L_{\phi_{\gamma_0}}^2+b_1^0L_F^2M_{\phi_{\gamma_0}}^2\right)^{1/2}}\left(\frac{M_F^2L_{\phi_{\gamma_0}}^2\sigma_F^2}{b_0}+\frac{M_{\phi_{\gamma_0}}^2\sigma_J^2}{\hat{b}_0}\right) \overset{(18)}{=} \frac{Q\sqrt{b}}{P\hat{b}_0}. \end{cases}$$

Let us choose $\beta_t := 1 - \frac{1}{(t+2)^{2/3}} \in (0,1)$ and $\gamma_t := \frac{1}{(t+2)^{1/3}}$ as in (25). Then, it is easy to check that

$$\frac{\beta_t^2(1-\beta_t)}{\Theta_t^2} \le \frac{1-\beta_{t+1}}{\Theta_{t+1}^2} \le \frac{1-\beta_t}{\Theta_t^2}.$$

In addition, as before, one can show that

$$\begin{cases} \sum_{t=0}^T\sqrt{1-\beta_t} = \sum_{t=0}^T\frac{1}{(t+2)^{1/3}} \ge \int_2^{T+3}\frac{ds}{s^{1/3}} = \frac{3}{2}[(T+3)^{2/3} - 2^{2/3}], \\ \sum_{t=0}^T\frac{(1-\beta_t)^2}{\sqrt{1-\beta_{t+1}}} = \sum_{t=0}^T\frac{(t+3)^{1/3}}{(t+2)^{4/3}} \le \sum_{t=0}^T\frac{1}{(t+1)} \le 1 + \log(T+1). \end{cases}$$

Using these estimates, we can easily prove

$$\begin{cases} \Sigma_T := \sum_{t=0}^T\omega_t &= \frac{\sqrt{b}}{P}\sum_{t=0}^T\sqrt{1-\beta_t} \ge \frac{3\sqrt{b}[(T+3)^{2/3}-2^{2/3}]}{2P}, \\ \sum_{t=0}^T\frac{\Gamma_{t+1}(1-\beta_t)^2}{\sqrt{1-\beta_{t+1}}} &\le \frac{Q[1+\log(T+1)]}{P\sqrt{b}} \end{cases}$$

Substituting these inequalities into (59) and using $\sqrt{1-\beta_0} = \frac{1}{2^{1/3}}$, we further upper bound

$$\begin{aligned} \mathbb{E}\left[\|\mathcal{G}_\eta(\bar{x}_T)\|^2\right] \le\ & \frac{32P}{3\sqrt{b}[(T+3)^{2/3}-2^{2/3}]}\left(\Psi_0(x_0) - \Psi_0^\star + \frac{B_\psi}{(T+2)^{1/3}}\right) \\ &+ \frac{16Q}{3[(T+3)^{2/3}-2^{2/3}]}\left(\frac{2^{1/3}}{\hat{b}_0} + \frac{2(1+\log(T+1))}{b}\right), \end{aligned}$$

which proves (26). □

C The proof of technical results in Section 4: Restarting variant

The proof of Theorem 4.1: Restarting variant. Since $\gamma := 0$, $\hat{b}_1 = \hat{b}_2 := b$ and $b_1 = b_2 := c_0 b$, from (63), using the superscript "$^{(s)}$" for the outer iteration s, and P and Q from (18), we have

$$\frac{\theta}{16L_{\Phi_0}}\mathbb{E}\big[\|\mathcal{G}_\eta(x_t^{(s)})\|^2\big] \le V_0(x_t^{(s)}) - V_0(x_{t+1}^{(s)}) + \frac{Q(1-\beta)^{3/2}}{Pb^{1/2}},$$

Summing up this inequality from $t := 0$ to $t := T$, and using the fact that $\tilde{x}^{s-1} := x_0^{(s)}$ and $\tilde{x}^s := x_{T+1}^{(s)}$, we get

$$\frac{\theta}{16L_{\Phi_0}}\sum_{t=0}^{T}\mathbb{E}\big[\|\mathcal{G}_\eta(x_t^{(s)})\|^2\big] \le V_0(\tilde{x}^{s-1}) - V_0(\tilde{x}^s) + \frac{Q(T+1)(1-\beta)^{3/2}}{Pb^{1/2}}.$$

Using the choice $b_0 := c_0\hat{b}_0$, similar to the proof of (64), we can show that

$$\begin{aligned} V_0(\tilde{x}^{s-1}) &= \mathbb{E}\big[\Psi_0(\tilde{x}^{s-1})\big] + \tfrac{\alpha}{2}\mathbb{E}\big[\|\tilde{F}_0^{(s)} - F(\tilde{x}^{s-1})\|^2\big] + \tfrac{\hat{\alpha}}{2}\mathbb{E}\big[\|\tilde{J}_0^{(s)} - F'(\tilde{x}^{s-1})\|^2\big] \\ &\le \mathbb{E}\big[\Psi_0(\tilde{x}^{s-1})\big] + \tfrac{Qb^{1/2}}{2P\hat{b}_0\sqrt{1-\beta}}. \end{aligned}$$

Using this estimate and and $V_0(\tilde{x}^s) \ge \Psi_0(\tilde{x}^s)$ into above inequality, we can further derive

$$\begin{aligned} \tfrac{1}{(T+1)}\textstyle\sum_{t=0}^{T}\mathbb{E}\big[\|\mathcal{G}_\eta(x_t^{(s)})\|^2\big] &\le \tfrac{16L_{\Phi_0}}{\theta(T+1)}\big[\Psi_0(\tilde{x}^{s-1}) - \Psi_0(\tilde{x}^s)\big] + \tfrac{16QL_{\Phi_0}(1-\beta)^{3/2}}{P\theta b^{1/2}} \\ &\quad + \tfrac{8QL_{\Phi_0}b^{1/2}}{P\theta(T+1)\hat{b}_0\sqrt{1-\beta}}. \end{aligned}$$

Due to the choice of b_1 and $\hat{b}_1$, it follows from (19) that $\beta := 1 - \frac{b^{1/2}}{[\hat{b}_0(T+1)]^{1/2}}$ and $\theta := \frac{L_{\Phi_0}b^{3/4}}{P[\hat{b}_0(T+1)]^{1/4}}$. Therefore, the last inequality becomes

$$\frac{1}{(T+1)}\sum_{t=0}^{T}\mathbb{E}\big[\|\mathcal{G}_\eta(x_t^{(s)})\|^2\big] \le \frac{16P\hat{b}_0^{1/4}}{[b(T+1)]^{3/4}}\big[\Psi_0(\tilde{x}^{s-1}) - \Psi_0(\tilde{x}^s)\big] + \frac{24Q}{[\hat{b}_0 b(T+1)]^{1/2}}.$$

Summing up this inequality from $s := 1$ to $s := S$ and multiplying the result by $\frac{1}{S}$, we get

$$\frac{1}{S(T+1)}\sum_{s=1}^{S}\sum_{t=0}^{T}\mathbb{E}\big[\|\mathcal{G}_\eta(x_t^{(s)})\|^2\big] \le \frac{16P\hat{b}_0^{1/4}}{S[b(T+1)]^{3/4}}\big[\Psi_0(\tilde{x}^0) - \Psi_0(\tilde{x}^S)\big] + \frac{24Q}{[\hat{b}_0 b(T+1)]^{1/2}}.$$

Substituting $\Psi_0(\tilde{x}^S) \ge \Psi_0^\star$ into the last inequality, and using the fact that $\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_N)\|^2\big] = \frac{1}{S(T+1)}\sum_{s=1}^{S}\sum_{t=0}^{T}\mathbb{E}\big[\|\mathcal{G}_\eta(x_t^{(s)})\|^2\big]$, we obtain

$$\begin{aligned} \mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_N)\|^2\big] &= \tfrac{1}{S(T+1)}\textstyle\sum_{s=1}^{S}\sum_{t=0}^{T}\mathbb{E}\big[\|\mathcal{G}_\eta(x_t^{(s)})\|^2\big] \\ &\le \tfrac{16P\hat{b}_0^{1/4}}{S[b(T+1)]^{3/4}}\big[\Psi_0(\tilde{x}^0) - \Psi_0^\star\big] + \tfrac{24Q}{[\hat{b}_0 b(T+1)]^{1/2}}, \end{aligned}$$

which is exactly (30).

Now, for a given tolerance $\varepsilon > 0$, to obtain $\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_K)\|^2\big] \le \varepsilon^2$, we need to impose

$$\frac{16P\hat{b}_0^{1/4}}{S[b(T+1)]^{3/4}} = \frac{\varepsilon^2}{2} \quad \text{and} \quad \frac{24Q}{[\hat{b}_0 b(T+1)]^{1/2}} = \frac{\varepsilon^2}{2}.$$

This condition leads to $N = S(T+1) = \frac{32P[\hat{b}_0(T+1)]^{1/4}}{b^{3/4}\varepsilon^2}$ and $\hat{b}_0 b(T+1) = \frac{48^2Q^2}{\varepsilon^4}$. Hence, the total number of iterations is $N := S(T+1) = \frac{32P[\hat{b}_0 b(T+1)]^{1/4}}{b\varepsilon^2} = \frac{128P\sqrt{3Q}}{b\varepsilon^3}$.

Clearly, to optimize the oracle complexity, we need to choose $T+1 := \frac{48Q}{b\varepsilon^2}$, then $\hat{b}_0 := \frac{48Q}{\varepsilon^2}$ and $S := \frac{8P}{\sqrt{3Q}\varepsilon}$. In this case, the total number of function evaluations is at most

$$\mathcal{T}_F := b_0 S + 3bS(T+1) = \frac{48Q}{\varepsilon^2}\cdot\frac{8P}{\sqrt{3Q}\varepsilon} + 3bN = \frac{16P\sqrt{3Q}}{\varepsilon^3} + \frac{384P\sqrt{3Q}}{\varepsilon^3} = \frac{400P\sqrt{3Q}}{\varepsilon^3}.$$

This is also the total number of Jacobian evaluations $\mathcal{T}_J$. □

The proof of Theorem 4.2. Let us first choose $\hat{b}_1 = \hat{b}_2 := b$, $b_1 = b_2 := \frac{c_0 b}{\gamma^2}$, and $b_0 := \frac{c_0\hat{b}_0}{\gamma^2}$. With the same line as the proof of (69), we can show that

$$\begin{aligned}\frac{1}{(T+1)}\sum_{t=0}^{T}\mathbb{E}\big[\|\mathcal{G}_\eta(x_t^{(s)})\|^2\big] &\le \frac{16P}{(T+1)[b(1-\beta)]^{1/2}}\big[\mathbb{E}\big[\Psi_0(x_0^{(s)})\big] - \mathbb{E}\big[\Psi_0(x_{T+1}^{(s)})\big] + \gamma B_\psi\big]\\ &\quad + 8Q\left[\frac{1}{\hat{b}_0(1-\beta)(T+1)} + \frac{2(1-\beta)}{b}\right].\end{aligned}$$

Here, we use the superscript "$^{(s)}$" to present the outer iteration s. Moreover, instead of Ψ_0^*, we keep $\Psi_0(x_{T+1}^{(s)})$ from (65). Now, using the fact that $\tilde{x}^{s-1} = x_0^{(s)}$ and $\tilde{x}^s = x_{T+1}^{(s)}$, we can further derive from the above inequality that

$$\begin{aligned}\frac{1}{(T+1)}\sum_{t=0}^{T}\mathbb{E}\big[\|\mathcal{G}_\eta(x_t^{(s)})\|^2\big] &\le \frac{16P}{(T+1)[b(1-\beta)]^{1/2}}\big[\mathbb{E}\big[\Psi_0(\tilde{x}^{s-1})\big] - \mathbb{E}\big[\Psi_0(\tilde{x}^s)\big] + \gamma B_\psi\big]\\ &\quad + 8Q\left[\frac{1}{\hat{b}_0(1-\beta)(T+1)} + \frac{2(1-\beta)}{b}\right].\end{aligned}$$

Summing up this inequality from $s := 1$ to $s := S$, and multiplying the result by $\frac{1}{S}$, and then using $0 < \gamma \le 1$, $\mathbb{E}\big[\Psi_0(\tilde{x}^0)\big] = \Psi_0(\tilde{x}^0)$, $\Psi_0(\tilde{x}^S) \ge \Psi_0^\star > -\infty$, and $\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_N)\|^2\big] = \frac{1}{S(T+1)}\sum_{s=1}^{S}\sum_{t=0}^{T}\mathbb{E}\big[\|\mathcal{G}_\eta(x_t^{(s)})\|^2\big]$, we arrive at

$$\begin{aligned}\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_N)\|^2\big] &= \frac{1}{(T+1)S}\sum_{s=1}^{S}\sum_{t=0}^{T}\mathbb{E}\big[\|\mathcal{G}_\eta(x_t^{(s)})\|^2\big]\\ &\le \frac{16P}{S(T+1)[b(1-\beta)]^{1/2}}\big[\Psi_0(\tilde{x}^0) - \Psi^\star + B_\psi\big]\\ &\quad + 8Q\left[\frac{1}{\hat{b}_0(1-\beta)(T+1)} + \frac{2(1-\beta)}{b}\right].\end{aligned}$$

Next, let us choose $\beta := 1 - \frac{b}{(T+1)}$ and $\hat{b}_0 := (T+1)$. Then, the above estimate becomes

$$\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_N)\|^2\big] \le \frac{16P}{bS(T+1)^{1/2}}\big[\Psi_0(\tilde{x}^0) - \Psi^\star + B_\psi\big] + \frac{24Q}{T+1}.$$

Let us define R_0 and $\hat{R}_0$ as in (32), i.e.:

$$R_0 := 16P\big[\Psi_0(\tilde{x}^0) - \Psi^\star + B_\psi\big] \quad\text{and}\quad \hat{R}_0 := 24Q.$$

In this case, for a given tolerance $\varepsilon > 0$, to achieve $\mathbb{E}\big[\|\mathcal{G}_\eta(\bar{x}_N)\|^2\big] \le \varepsilon^2$, we can impose

$$\frac{R_0}{bS(T+1)^{1/2}} = \frac{\varepsilon^2}{2} \quad\text{and}\quad \frac{\hat{R}_0}{(T+1)} = \frac{\varepsilon^2}{2}.$$

These conditions lead to $T+1 = \frac{2\hat{R}_0}{\varepsilon^2}$ and $S := \frac{2R_0}{b(T+1)^{1/2}\varepsilon^2} = \frac{\sqrt{2}R_0}{b\varepsilon\sqrt{\hat{R}_0}}$. Let us also choose $\gamma := \frac{\varepsilon}{\sqrt{2\hat{R}_0}}$. Then, we also obtain the parameters as in (31), i.e.:

$$\begin{cases} b_1 = b_2 := \frac{2c_0 b\hat{R}_0}{\varepsilon^2}, \quad \hat{b}_1 = \hat{b}_2 := b, \quad b_0 := \frac{4c_0\hat{R}_0^2}{\varepsilon^4}, \quad \hat{b}_0 := \frac{2\hat{R}_0}{\varepsilon^2}, \\ \gamma := \frac{\varepsilon}{\sqrt{2\hat{R}_0}}, \quad \text{and} \quad \beta := 1 - \frac{b\varepsilon^2}{2\hat{R}_0}. \end{cases}$$

The total number $\mathcal{T}_F$ of function evaluations $\mathbf{F}(x_t^{(s)}, \xi_t)$ is at most

$$\mathcal{T}_F := S[b_0 + (T+1)(2b_1 + b_2)] = \frac{\sqrt{2}R_0}{b\varepsilon\sqrt{\hat{R}_0}}\Big[\frac{4c_0\hat{R}_0^2}{\varepsilon^4} + \frac{2\hat{R}_0}{\varepsilon^2}\frac{6c_0 b\hat{R}_0}{\varepsilon^2}\Big] = \frac{4\sqrt{2}c_0R_0\hat{R}_0^{3/2}}{\varepsilon^5}\left(\frac{1}{b} + 3\right).$$

The total number $\mathcal{T}_J$ of Jacobian evaluations $\mathbf{F}'(x_t^{(s)}, \xi_t)$ is at most

$$\mathcal{T}_J := S[\hat{b}_0 + (T+1)(2\hat{b}_1 + \hat{b}_2)] = \frac{\sqrt{2}R_0}{b\varepsilon\sqrt{\hat{R}_0}}\Big[\frac{2\hat{R}_0}{\varepsilon^2} + \frac{6b\hat{R}_0}{\varepsilon^2}\Big] = \frac{2\sqrt{2}R_0\hat{R}_0^{1/2}}{b\varepsilon^3} + \frac{6\sqrt{2}R_0\hat{R}_0^{1/2}}{\varepsilon^3}.$$

These prove the last statement of Theorem 4.2. □

D Experiment setup and additional experiments

This appendix provides the details of configuration for our experiments in Section 5, and presents more numerical experiments to support our algorithms and theoretical results. As mentioned in the main text, all the algorithms used in this paper have been implemented in Python 3.6.3., running on a Linux desktop (3.6GHz Intel Core i7 and 16Gb memory).

Let us provide more details of our experiment configuration. We shorten the name of our algorithm, either Algorithm 1 or Algorithm 2, by Hybrid Stochastic Compositional Gradient, and abbreviate it by `HSCG` for both cases. We have implemented `CIVR` in [38] and `ASC-PG` in [32] to compare the smooth case of ϕ_0. For the nonsmooth case of ϕ_0, we have implemented two other algorithms, `SCG` in [31], and `Prox-Linear` in [28, 39]. While `SCG` only works for smooth ϕ_0, we have smoothed it as in our method, and used the estimator as well as the algorithm in [31], but update the smoothness parameter as in our method. We also omit comparison in terms of time since `Prox-Linear` becomes slower if p is large due to its expensive subproblem for evaluating the prox-linear operator. We only compare these algorithms in terms of epoch (i.e., the number of data passes).

Since both `CIVR` and `ASC-PG` are double loop, to be fair, we compare them with our restarting variant, Algorithm 2. To compare with `SCG` and `Prox-Linear`, we simply use Algorithm 1 since `SCG` has single loop. Since `Prox-Linear` requires to solve a nonsmooth convex subproblem, we have implemented a first-order primal-dual method in [5] to solve it. This algorithm has shown its efficiency in our test.

Note that the batch size b is determined as $b := \lfloor \frac{N}{n_b} \rfloor$, where N is the number of data points, and n_b is the number of blocks. In our experiments, we have varied the number of blocks n_b to observe the performance of these algorithms. Since we want to obtain the best performance, instead of using their theoretical step-sizes, we have carefully tuned the step-size η of three algorithms in a given set of candidates $\{1, 0.5, 0.1, 0.05, 0.01, 0.001, 0.0001\}$. For our algorithms, we have another step-size θ_t, which is also flexibly chosen from $\{0.1, 0.5, 1\}$. For the nonsmooth case, we update our smoothness parameter as $\gamma_t := \frac{1}{2(t+1)^{1/3}}$, which is proportional to the value in Theorems 3.2 and 3.4.

To further compare our algorithms with their competitors, we provide in the following subsections additional experiments for the two problems in the main text.

D.1 Risk-averse portfolio optimization: Additional experiments

Figure 1 in the main text has shown the performance of three algorithms on three different datasets using 8 blocks, i.e., $n_b = 8$. Unfortunately, since ASC-PG does not work well when the number of blocks is larger than 8, we skip showing it in our comparison. To obverse more performance of `HSCG` and `CIVR`, we have increased the number of blocks n_b from 8 to 32, 64, and 128. The convergence of the two algorithms is shown in Figure 3. As we can observe, `HSCG` remains slightly better than `CIVR` if $n_b = 32$ or 64. When $n_b = 128$, `CIVR` improves its performance and is slightly better than `HSCG`.

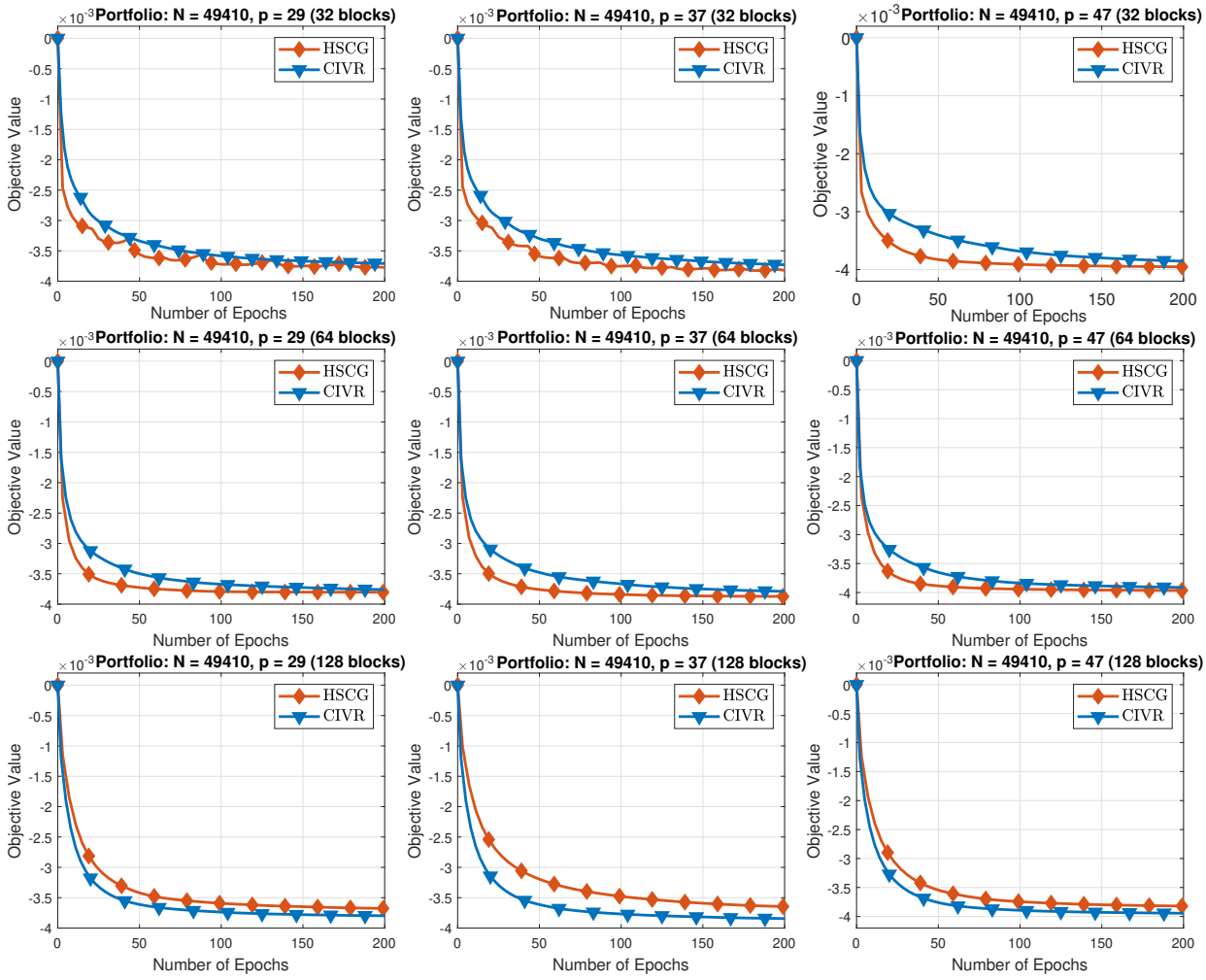

Figure 3: Comparison of two algorithms for solving (33) on larger blocks.

D.2 Stochastic minimax problem: Additional experiments

For the stochastic minimax problem (34), Figure 2 has shown the progress of the objective values of three algorithms on three different datasets. Figure 4 simultaneously shows both the objective values and the gradient mapping norms of this experiment.

Now, let us keep the same configuration as in Figure 2, but run one more case, where the number of blocks is increased to $n_b = 64$. The results are shown in Figure 5.

We again see that `HSCG` still highly outperforms the other two methods: `SCG` and `Prox-Linear` on **rcv1**. For **url**, `HSCG` is still slightly better than `Prox-Linear` as we have observed in Figure 2.

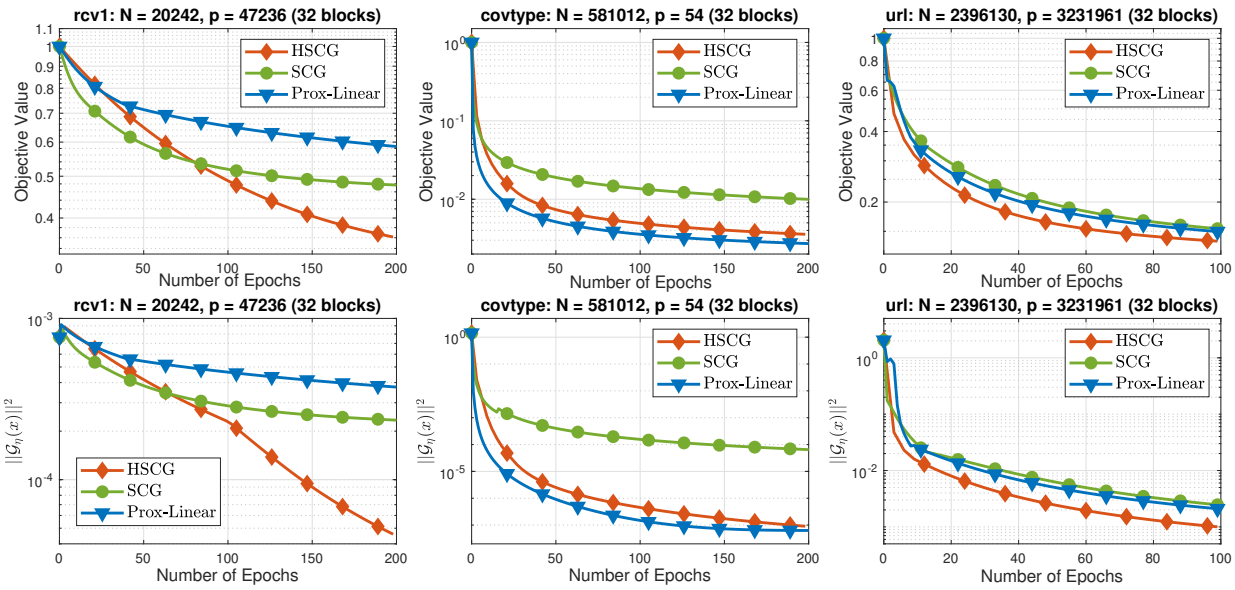

Figure 4: Comparison of three algorithms for solving (34) on 3 different datasets in Figure 2 with both objective values and gradient mapping norms.

However, for **covtype**, again, `Prox-Linear` shows a better performance than the other two competitors. Note that since $p = 54$ in this dataset, we can solve the subproblem in `Prox-Linear` up to a high accuracy without incurring too much computational cost. Therefore, the inexactness of evaluating the prox-linear operator does not really affect the performance in this example.

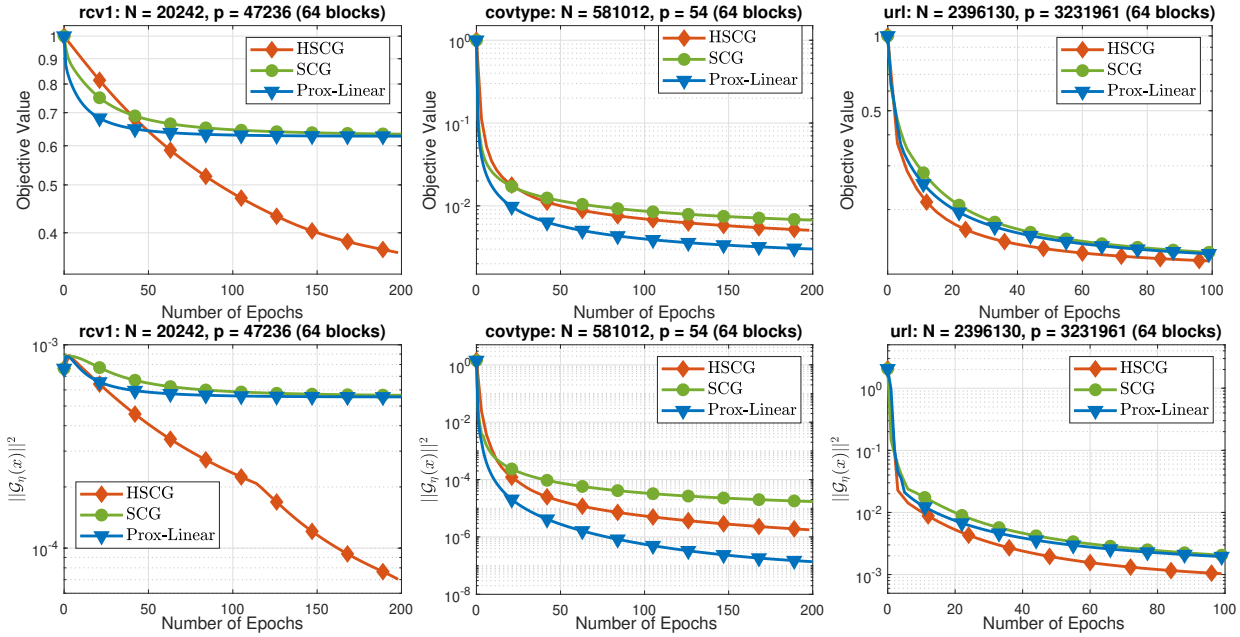

Figure 5: Comparison of three algorithms for solving (34) on 64 blocks.

Finally, we test three algorithms: `HSCG`, `SCG`, and `Prox-Linear` on other three datasets: **w8a**, **phishing**, and **mushrooms** from LIBSVM [6]. We use the same number of blocks $n_b = 32$, and the results are reported in Figure 6. Figure 6 shows that `HSCG` highly outperforms both `SCG` and

`Prox-Linear` on **w8a** and **phishing**. However, `Prox-Linear` becomes better than the other two on the **mushrooms** dataset.

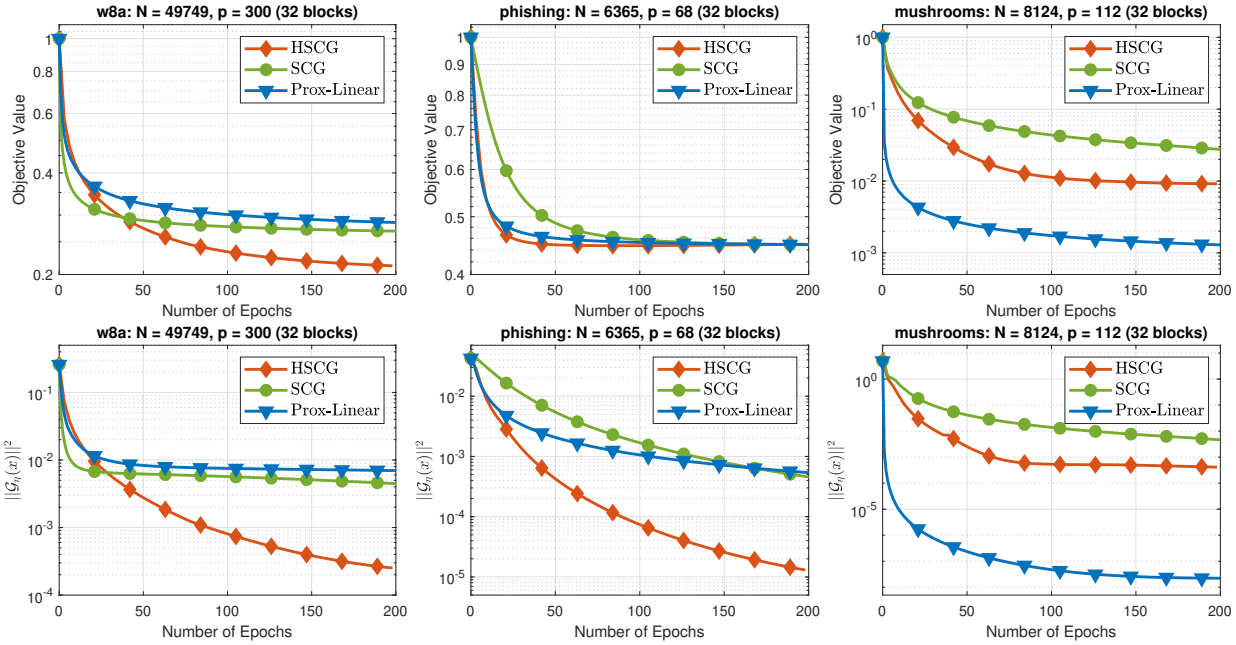

Figure 6: Comparison of three algorithms for solving (34) on three more different datasets.

References

[1] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. 2020.

[2] Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

[3] H. H. Bauschke and P. Combettes. *Convex analysis and monotone operators theory in Hilbert spaces*. Springer-Verlag, 2nd edition, 2017.

[4] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, 2009.

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

[6] C.-C. Chang and C.-J. Lin. LIBSVM: A library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[7] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Math. Program.*, 178(1-2):503–558, 2019.

[8] J. Duchi and F. Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM J. Optim.*, 28(4):3229–3259, 2018.

[9] F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*, volume 1-2. Springer-Verlag, 2003.

[10] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Program.*, 156(1-2):59–99, 2016.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[12] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 315–323, 2013.

[13] G. M. Korpelevic. An extragradient method for finding saddle-points and for other problems. *Èkonom. i Mat. Metody.*, 12(4):747–756, 1976.

[14] A. S. Lewis and S. J. Wright. A proximal method for composite minimization. *Math. Program.*, 158(1-2):501–546, 2016.

[15] X. Lian, M. Wang, and J. Liu. Finite-sum composition optimization via variance reduced gradient descent. In *Artificial Intelligence and Statistics*, pages 1159–1167, 2017.

[16] Q. Lin, M. Liu, H. Rafique, and T. Yang. Solving weakly-convex-weakly-concave saddle-point problems as weakly-monotone variational inequality. *arXiv preprint arXiv:1810.10207*, 2018.

[17] L. Liu, J. Liu, and D. Tao. Variance reduced methods for non-convex composition optimization. *arXiv preprint arXiv:1711.04416*, 2017.

[18] H. M. Markowitz. Portfolio Selection. 7:77–91.

[19] A. Nemirovskii. Prox-method with rate of convergence $\mathcal{O}(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. Op*, 15(1):229–251, 2004.

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

[21] Y. Nesterov. Modified Gauss-Newton scheme with worst case guarantees for global performance. *Optim. Method Softw.*, 22(3):469–483, 2007.

[22] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. *ICML*, 2017.

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

[24] D. M. Ostrovskii, A. Lowy, and M. Razaviyayn. Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems. *arXiv preprint arXiv:2002.07919*, 2020.

[25] H. N. Pham, M. L. Nguyen, T. D. Phan, and Q. Tran-Dinh. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *J. Mach. Learn. Res.*, (accepted), 2020.

[26] A. Shapiro and A. Kleywegt. Minimax analysis of stochastic problems. *Optim. Methods Softw.*, 17(3):523–542, 2002.

[27] Q. Tran-Dinh and M. Diehl. Proximal methods for minimizing the sum of a convex function and a composite function. Tech. report, KU Leuven, OPTEC and ESAT/SCD, Belgium, May 2011.

[28] Q. Tran-Dinh, N. H. Pham, and L. M. Nguyen. Stochastic Gauss-Newton algorithms for nonconvex compositional optimization. *Accepted for presentation at the International Conference on Machine Learning (ICML)*, 2020.

[29] Q. Tran-Dinh, N. H. Pham, D. T. Phan, and L. M. Nguyen. A hybrid stochastic optimization framework for stochastic composite nonconvex optimization. *Preprint: UNC-STOR 07.10.2019*, 2019.

[30] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *Submitted to SIAM J. Optim*, 2008.

[31] M. Wang, E. Fang, and L. Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Math. Program.*, 161(1-2):419–449, 2017.

[32] M. Wang, J. Liu, and E. X. Fang. Accelerating stochastic composition optimization. *The Journal of Machine Learning Research*, 18(1):3721–3743, 2017.

[33] Y. Xu and Y. Xu. Katyusha acceleration for convex finite-sum compositional optimization. *arXiv preprint arXiv:1910.11217*, 2019.

[34] J. Yang, N. Kiyavash, and N. He. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2002.09621*, 2020.

[35] S. Yang, M. Wang, and E. X. Fang. Multilevel stochastic gradient methods for nested composition optimization. *SIAM J. Optim.*, 29(1):616–659, 2019.

[36] Y. Yu and L. Huang. Fast stochastic variance reduced admm for stochastic composition optimization. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3364–3370. AAAI Press, 2017.

[37] J. Zhang and L. Xiao. Multi-level composite stochastic optimization via nested variance reduction. *arXiv preprint arXiv:1908.11468*, 2019.

[38] J. Zhang and L. Xiao. A stochastic composite gradient method with incremental variance reduction. *Advances in Neural Information Processing Systems*, 28:9078—-9088, 2019.

[39] J. Zhang and L. Xiao. Stochastic variance-reduced prox-linear algorithms for nonconvex composite optimization. *arXiv preprint arXiv:2004.04357*, 2020.

[40] L. Zhao, M. Mammadov, and J. Yearwood. From convex to nonconvex: a loss function analysis for binary classification. In *IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1281–1288. IEEE, 2010.