

ZEROTH-ORDER METHODS FOR STOCHASTIC NONCONVEX NONSMOOTH COMPOSITE OPTIMIZATION

Anonymous authors

Paper under double-blind review

ABSTRACT

This work aims to solve a stochastic nonconvex nonsmooth composite optimization problem. Previous works on composite optimization problem requires the major part to satisfy Lipschitz smoothness or some relaxed smoothness conditions, which excludes some machine learning examples such as regularized ReLU network and sparse support matrix machine. In this work, we focus on stochastic nonconvex composite optimization problem without any smoothness assumptions. In particular, we propose two new notions of approximate stationary points for such optimization problem and obtain finite-time convergence results of two zeroth-order algorithms to these two approximate stationary points respectively. Finally, we demonstrate that these algorithms are effective using numerical experiments.

1 INTRODUCTION

This work focuses on the following stochastic nonconvex nonsmooth composite optimization problem.

$$\min_{x \in \mathbb{R}^d} \phi(x) := F(x) + h(x), \text{ where } F(x) = \mathbb{E}_{\xi \sim \mathcal{P}}[f_{\xi}(x)], \quad (1)$$

where the individual function $f_{\xi}(x)$ is nonconvex and nonsmooth associated with a stochastic sample ξ from the distribution $\mathcal{P}$, and h is a convex regularizer. This problem covers many machine learning examples such as regularized ReLU network (Mazumdar & Rawat, 2019; Wang et al., 2021b) and sparse support matrix machine (Zheng et al., 2018; Gu et al., 2021; Li et al., 2022).

Existing approaches to the stochastic nonconvex composite optimization problem (1) require the major part F to satisfy either Lipschitz smooth conditions (Nitanda, 2014; Li & Lin, 2015; Ghadimi et al., 2016; Ghadimi & Lan, 2016; Li et al., 2017; Pham et al., 2020), or some relaxed notions of smoothness such as relative smoothness (Bauschke et al., 2017; Lu et al., 2018; Latafat et al., 2022), smooth adaptivity (Wang & Han, 2023; Ding et al., 2025), anisotropic smoothness (Laude & Patrinos, 2025), weak convexity (Davis & Drusvyatskiy, 2019; Davis & Grimmer, 2019) and Holder continuous gradient (Guo et al., 2022), which cannot cover the applications with discontinuous gradient, such as regularized ReLU network (Mazumdar & Rawat, 2019; Wang et al., 2021b) and sparse support matrix machine (Zheng et al., 2018; Gu et al., 2021; Li et al., 2022).

To solve such a stochastic nonconvex nonsmooth composite optimization problem, the first challenging step is to propose proper and feasible convergence criteria. The existing notions of proximal gradient mapping (Ghadimi et al., 2016; Reddi et al., 2016; Li & Li, 2018) and Frank-Wolfe gap (Jiang & Zhang, 2014; Guo et al., 2022) requiring F to be differentiable everywhere are not suitable for nonsmooth composite optimization. Even after extending the gradient to the Clarke subdifferential, we will prove that convergence under the corresponding generalized stationary notions is intractable (see Theorem 1). Fortunately, Zhang et al. (2020) proposes the notion of (δ, ϵ) -Goldstein stationary point which has been achieved by various nonconvex nonsmooth optimization algorithms (Zhang et al., 2020; Lin et al., 2022; Chen et al., 2023; Cutkosky et al., 2023; Kornowski & Shamir, 2024), and the Goldstein stationary notion is extended to nonconvex nonsmooth constrained optimization (Liu et al., 2024). Inspired by these stationary notions, we propose $(\gamma, \delta, \epsilon)$ -proximal Goldstein stationary point (PGSP) and (δ, ϵ) -conditional gradient Goldstein stationary point (CGGSP) as the approximate notions of stationarity for our nonconvex nonsmooth composite optimization problem (1), by using the Goldstein δ -subdifferential (Goldstein, 1977) as a convex combination of the gradients in the neighborhood around the point of interest.

Table 1: Function evaluation complexity results of zeroth-order proximal gradient descent (0-PGD) and zeroth-order generalized conditional gradient algorithms (0-GCG).

Main algorithm	Gradient estimation	Criterion	Complexity	Reference
0-PGD (Algorithm 1)	Minibatch	$(\gamma, \delta, \epsilon)$ -PGSP	$\mathcal{O}(d^{3/2}\delta^{-1}\epsilon^{-4})$	Theorem 2
0-PGD (Algorithm 1)	Variance reduction	$(\gamma, \delta, \epsilon)$ -PGSP	$\mathcal{O}(d^{3/2}\delta^{-1}\epsilon^{-3})$	Theorem 3
0-GCG (Algorithm 2)	Minibatch	(δ, ϵ) -CGGSP	$\mathcal{O}(d^{3/2}\delta^{-1}\epsilon^{-4})$	Theorem 4
0-GCG (Algorithm 2)	Variance reduction	(δ, ϵ) -CGGSP	$\mathcal{O}(d^{3/2}\delta^{-1}\epsilon^{-3})$	Theorem 5

Using our proposed stationary notions above, we prove that the zeroth-order proximal gradient descent algorithm (0-PGD, see Algorithm 1) converges to our proposed $(\gamma, \delta, \epsilon)$ -proximal Goldstein stationary point (PGSP), obtain the convergence rate and function evaluation complexity result using minibatch zeroth-order gradient estimation, and then improve these results using variance-reduced gradient estimation. Furthermore, we study a zeroth-order generalized conditional gradient algorithm (0-GCG, Algorithm 2) which avoids the possibly expensive proximal operator used by 0-PGD, and obtain similar convergence rate and complexity result of 0-GCG to achieve our proposed (δ, ϵ) -CGGSP. We summarize these convergence results of both algorithms in Table 1.

1.1 PAPER ORGANIZATION

Section 2 introduces the basic backgrounds including problem formulation, fundamentals for nonsmooth analysis and zeroth-order gradient estimation. Section 3 proposes our generalized stationary notions for composite optimization. Section 4 presents our zeroth-order proximal gradient descent (0-PGD) algorithm and its finite-time convergence results. Section 5 presents our zeroth-order generalized conditional gradient (0-GCG) algorithm and its finite-time convergence results. Section 6 shows the experimental results. Section 7 concludes this work.

2 PRELIMINARIES

In this section, we will introduce the problem formulation (Section 2.1), review fundamentals of nonsmooth analysis (Section 2.2), and introduce zeroth-order gradient estimation (Section 2.3).

2.1 PROBLEM FORMULATION

Throughout this work, we make the following two standard assumptions on the stochastic nonconvex nonsmooth composite optimization problem (1).

Assumption 1. For any stochastic sample ξ , $f_\xi(x) : \mathbb{R}^d \rightarrow \mathbb{R}$ is an L_ξ -Lipschitz continuous for some $L_\xi > 0$ (i.e., $|f_\xi(y) - f_\xi(x)| \leq L_\xi \|y - x\|$ for any $x, y \in \mathbb{R}^d$) and $\mathbb{E}_\xi(L_\xi^2) \leq G^2$ for some $G > 0$.

Assumption 2. $h : \mathbb{R}^d \rightarrow \mathbb{R}$ is a proper closed convex function with at least one feasible point $x^{(h)} \in \mathbb{R}^d$ such that $h(x^{(h)}) < +\infty$.

Assumption 3. There exists $R > 0$ such that $h(x) > h(x^{(h)}) + G\|x - x^{(h)}\|$ for the feasible point $x^{(h)}$ defined in Assumption 2 and any $x \in \mathbb{R}^d$ satisfying $\|x - x^{(h)}\| > R$.

Assumption 1 has also been used by (Davis & Drusvyatskiy, 2019; Davis & Grimmer, 2019; Liu et al., 2024). It implies that $F(x) = \mathbb{E}_\xi[f_\xi(x)]$ is G -Lipschitz continuous¹. Such Lipschitz continuous but possibly nonsmooth functions have been widely used in optimization and machine learning, including any neural networks with ReLU activation (Krizhevsky et al., 2017; Mazumdar & Rawat, 2019; Ghosh et al., 2024; Shen et al., 2024), ramp loss (Gu et al., 2021; Wang & Shao, 2024), capped ℓ_1 penalty (Xu et al., 2014; Zhang, 2008; Kumar et al., 2021), etc. Many commonly used convex regularizers h satisfy Assumptions 2 and 3, including ℓ_p regularizer with $p > 1$ (McCulloch et al., 2024; Lu et al., 2024), ℓ_1 regularizer $\lambda\|x\|_1$ with $\lambda > G$ to induce the sparsity of the parameter vector x (Mazumdar & Rawat, 2019; Ali et al., 2024), super-coercive regularizer satisfying $\lim_{\|x\| \rightarrow +\infty} [h(x)/\|x\|] = +\infty$

¹Assumption 1 implies that F is G -Lipschitz continuous because for any $x, x' \in \mathbb{R}^d$,
 $|F(x') - F(x)| \leq \mathbb{E}_\xi[f_\xi(x') - f_\xi(x)] \leq \mathbb{E}_\xi[L_\xi\|x' - x\|] \leq \|x' - x\| \sqrt{\mathbb{E}_\xi[L_\xi^2]}.$

(Bredies et al., 2005; Yu et al., 2017; Bredies et al., 2009), and the following constraint regularizer which enforces the constraint $x \in \Omega$ where $\Omega \subset \mathbb{R}^d$ is a convex and compact set (Jaggi, 2013; Rakotomamonjy et al., 2015; Nesterov, 2018; Liu et al., 2024; Assunção et al., 2025).

$$h_\Omega(x) \stackrel{\text{def}}{=} \begin{cases} 0; & x \in \Omega \\ +\infty; & x \notin \Omega \end{cases}, \quad (2)$$

Proposition 1. *Under Assumptions 1-3, the original objective function (1) has a non-empty solution set $\arg \min_{x \in \mathbb{R}^d} \phi(x)$, which is a subset of $\mathcal{B}_d(x^{(h)}, R) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d : \|x - x^{(h)}\| \leq R\}$.*

Remark: Assumption 3 requires $h(x)$ to outgrow $F(x)$ as $\|x - x^{(h)}\| \rightarrow +\infty$, such that the objective $\phi(x) = F(x) + g(x)$ has minimizers and they are not too far from the feasible point $x^{(h)}$.

2.2 FUNDAMENTALS OF NONSMOOTH ANALYSIS

In this subsection, we will introduce some basic concepts for the unconstrained nonconvex nonsmooth optimization problem $\min_{x \in \mathbb{R}^d} F(x)$, a special case of the composite problem (1) with $h = 0$.

For a nondifferentiable function $F : \mathbb{R}^d \rightarrow \mathbb{R}$, we can define generalized directional derivatives and generalized gradients as follows.

Definition 1. *The **generalized directional derivative** of a function F at point $x \in \mathbb{R}^d$ and direction $v \in \mathbb{R}^d$ is defined as $DF(x; v) \stackrel{\text{def}}{=} \limsup_{y \rightarrow x, t \downarrow 0} \frac{F(x+tv) - F(x)}{t}$. The **Clark subdifferential** of F is defined as the set $\partial F(x) \stackrel{\text{def}}{=} \{g \in \mathbb{R}^d : \langle g, v \rangle \leq DF(x; v), \forall v \in \mathbb{R}^d\}$.*

For the unconstrained nonconvex nonsmooth optimization problem $\min_{x \in \mathbb{R}^d} F(x)$, one may aim to find an ϵ -Clarke stationary point defined as $x \in \mathbb{R}^d$ satisfying $\min\{\|g\| : g \in \partial F(x)\} \leq \epsilon$. However, Zhang et al. (2020) proves that such an ϵ -Clarke stationary point cannot be obtained in finite time for general Lipschitz continuous function F . Hence, they focus on more tractable and relaxed concepts of subdifferential and stationary solution, as defined below.

Definition 2 (Goldstein (1977)). *The **Goldstein δ -subdifferential** of a function F at $x \in \mathbb{R}^d$ with radius $\delta \geq 0$ is defined as*

$$\partial_\delta F(x) \stackrel{\text{def}}{=} \text{conv}[\cup_{y \in \mathcal{B}_d(x, \delta)} \partial F(y)],$$

where $\text{conv}(A)$ denotes the set of every convex combination of the elements in A .

Definition 3 (Zhang et al. (2020)). *For any $\delta \geq 0$ and $\epsilon > 0$, a (δ, ϵ) -**Goldstein stationary point** of F is defined as any $x \in \mathbb{R}^d$ satisfying*

$$\min\{\|g\| : g \in \partial_\delta F(x)\} \leq \epsilon. \quad (3)$$

Note that $\partial_0 F(x) = \partial F(x)$ (Makela & Neittaanmaki, 1992). Hence, as $\delta = 0$, Goldstein δ -subdifferential and (δ, ϵ) -Goldstein stationary point respectively reduce to Clark subdifferential and ϵ -Clarke stationary point. Such a (δ, ϵ) -Goldstein stationary point can be achieved at finite time by various algorithms (Zhang et al., 2020; Tian et al., 2022; Davis et al., 2022; Cutkosky et al., 2023).

2.3 ZERO-ORDER GRADIENT ESTIMATION

Zeroth-order gradient estimation with random smoothing technique has been widely used when direct computation of gradient is costly or impossible. To estimate the gradient of a function F , we can approximate F by its smoothing function $F_\delta(x) = \mathbb{E}_{u \sim Q}[F(x + \delta u)]$ with a small radius $\delta > 0$, with a certain distribution Q . We focus on the case where Q is uniform distribution on the unit sphere $\mathcal{S}_d(1) \stackrel{\text{def}}{=} \{u \in \mathbb{R}^d : \|u\| = 1\}$ (Duchi et al., 2015; Lin et al., 2020), since the corresponding smoothing function $F_\delta(x) \stackrel{\text{def}}{=} \mathbb{E}_{u \sim \text{Uniform}(\mathcal{S}_d(1))} F(x + \delta u)$ has the following amenable properties.

Lemma 1 (Proposition 2.3 of (Lin et al., 2022)). *For any G -Lipschitz continuous function F , its smoothing function $F_\delta(x) \stackrel{\text{def}}{=} \mathbb{E}_{u \sim \text{Uniform}(\mathcal{S}_d(1))} F(x + \delta u)$ satisfies: (1) $\sup_{x \in \mathbb{R}^d} |F_\delta(x) - F(x)| \leq \delta G$; (2) F_δ is G -Lipschitz continuous and differentiable everywhere with $cG\sqrt{d}/\delta$ -Lipschitz continuous gradient for an absolute constant $c > 0$; (3) $\nabla F_\delta(x) \in \partial_\delta F(x)$ for any $x \in \mathbb{R}^d$.*

$\nabla F_\delta(x)$ admits the following unbiased two-point estimator, which is widely used in zeroth-order gradient estimation (Duchi et al., 2015; Lin et al., 2022; Ma & Huang, 2025).

$$\hat{g}_\delta(x; u, \xi) = \frac{d}{2\delta} [f_\xi(x + \delta u) - f_\xi(x - \delta u)]u, \quad (4)$$

where $\xi \sim \mathcal{P}$ and $u \sim \text{Uniform}(\mathcal{S}_d(1))$.

3 GENERALIZED GOLDSTEIN STATIONARY POINTS FOR COMPOSITE OPTIMIZATION

In this section, we propose two new notions of stationary points for the stochastic nonconvex nonsmooth composite optimization problem (1), proximal Goldstein stationary point (PGSP) and conditional gradient Goldstein stationary point (CGGSP), the targets of zeroth-order algorithms in Sections 4 and 5. Then we show some properties of these new stationary points.

3.1 PROXIMAL GOLDSTEIN STATIONARY POINT (PGSP)

Definition 4. For any stepsize $\gamma > 0$ and convex regularizer $h : \mathbb{R}^d \rightarrow \mathbb{R}$, we define the **proximal operator** of γh at point $x \in \mathbb{R}^d$ as follows (Parikh et al., 2014; Nitanda, 2014; Mardani et al., 2018; Yang & Yu, 2020).

$$\text{prox}_{\gamma h}(x) = \arg \min_{y \in \mathbb{R}^n} \left[h(y) + \frac{1}{2\gamma} \|y - x\|^2 \right]. \quad (5)$$

The proximal operator (5) returns a unique solution since $h(y) + \frac{1}{2\gamma} \|y - x\|^2$ is a strongly convex function of y . Proximal operator is essential in the popular proximal gradient descent algorithms for composite optimization (Parikh et al., 2014; Nitanda, 2014; Mardani et al., 2018; Yang & Yu, 2020). We will use the proximal operator to propose a generalized notion of stationary point as follows.

Definition 5. For any stepsize $\gamma > 0$ and convex regularizer $h : \mathbb{R}^d \rightarrow \mathbb{R}$, we define the **proximal gradient mapping** at point $x \in \mathbb{R}^d$ and gradient $g \in \mathbb{R}^d$ as follows (Ghadimi et al., 2016; Reddi et al., 2016; Li & Li, 2018).

$$\mathcal{G}_{\gamma h}(x, g) = \frac{1}{\gamma} [x - \text{prox}_{\gamma h}(x - \gamma g)]. \quad (6)$$

Furthermore, for any $\epsilon \geq 0$, we define $x \in \mathbb{R}^d$ as a $(\gamma, \delta, \epsilon)$ -**proximal Goldstein stationary point (PGSP)** if $\min_{g \in \partial_\delta F(x)} \|\mathcal{G}_{\gamma h}(x, g)\| \leq \epsilon$. Specifically, we call a $(\gamma, 0, \epsilon)$ -PGSP as (γ, ϵ) -**PGSP**, defined by $x \in \mathbb{R}^d$ such that $\min_{g \in \partial F(x)} \|\mathcal{G}_{\gamma h}(x, g)\| \leq \epsilon$.

Our proposed notions of PGSP for nonsmooth composite optimization problem generalize existing stationary notions for the following special cases.

- For constrained optimization problem $\min_{x \in \Omega} F(x)$, a special case of the nonconvex nonsmooth composite optimization problem (1) with $h = h_\Omega$ defined by Eq. (2), $(\gamma, \delta, \epsilon)$ -PGSP reduces to the $(\gamma, \delta, \epsilon)$ -generalized Goldstein stationary point (Liu et al., 2024), where the proximal operator is reduced to the projection onto Ω . Furthermore, when $h = 0$, $(\gamma, \delta, \epsilon)$ -PGSP reduces to (δ, ϵ) -Goldstein stationary point (see Definition 3) (Zhang et al., 2020).

- When F is differentiable, (γ, ϵ) -PGSP has simplified definition that $\|\mathcal{G}_{\gamma h}[x, \nabla F(x)]\| \leq \epsilon^2$, which can be achieved by proximal gradient descent algorithms within finite-time (Ghadimi et al., 2016; Reddi et al., 2016; Li & Li, 2018). Furthermore, when $h = 0$, (γ, ϵ) -PGSP reduces to ϵ -stationary point defined as $x \in \mathbb{R}^d$ satisfying $\|\mathcal{G}_{\gamma h}[x, \nabla F(x)]\| \leq \epsilon$ which can be also achieved in finite time by many first-order algorithms. In contrast, (γ, ϵ) -PGSP is intractable for our setting with Lipschitz continuous and nondifferentiable F , as will be shown later in Theorem 1. Therefore, we aim at $(\gamma, \delta, \epsilon)$ -PGSP, a relaxed notion of stationarity, and will propose a zeroth-order proximal gradient descent algorithm (Algorithm 1) that achieves this point in finite time.

Our proposed notions of PGSP satisfy the following properties.

²For differentiable function F , we have $\partial F(x) = \{\nabla F(x)\}$ (Makela & Neittaanmaki, 1992).

Proposition 2. Suppose the function $F : \mathbb{R}^d \rightarrow \mathbb{R}$ is differentiable and $h : \mathbb{R}^d \rightarrow \mathbb{R}$ is a convex function. Then, $(\gamma, \delta, \epsilon)$ -PGSP has the following properties.

1. A (γ, ϵ) -PGSP is also a $(\gamma, \delta, \epsilon)$ -PGSP.
2. If ∇F is L -Lipschitz continuous, then a $(\gamma, \epsilon/(2L), \epsilon/2)$ -PGSP is also a (γ, ϵ) -PGSP, i.e., $\|\mathcal{G}_{\gamma h}[x, \nabla F(x)]\| \leq \epsilon$.
3. If $x \in \mathbb{R}^d$ satisfies $\|\mathcal{G}_{\gamma h}[x, \nabla F_\delta(x)]\| \leq \epsilon$, then x is a $(\gamma, \delta, \epsilon)$ -PGSP.

Remark: In Proposition 2, items 1 and 2 imply that for L -Lipschitz smooth functions F , our notions of $(\gamma, \delta, \epsilon)$ -PGSP and (γ, ϵ') -PGSP are equivalent (for possibly different $\epsilon, \epsilon' \geq 0$). Item 3 implies that we can obtain a $(\gamma, \delta, \epsilon)$ -PGSP by solving $\min_{x \in \mathbb{R}^d} F_\delta(x)$, which is important for designing the zeroth-order proximal gradient descent algorithm (Algorithm 1).

3.2 CONDITIONAL GRADIENT GOLDSTEIN STATIONARY POINT (CGGSP)

Note that the PGSP defined in the previous subsection relies on the stepsize γ . In this subsection, we will define conditional gradient Goldstein stationary point (CGGSP), a notion of stationary point that does not rely on the stepsize γ and can be computationally cheaper.

Definition 6. For any convex regularizer $h : \mathbb{R}^d \rightarrow \mathbb{R}$, we define the **linear minimization oracle (LMO)** of h at the gradient $g \in \mathbb{R}^d$ as follows.

$$\mathcal{L}_h(g) \stackrel{\text{def}}{=} \arg \min_{y \in \mathbb{R}^d} [h(y) + \langle y, g \rangle]. \quad (7)$$

The LMO defined above always exist and is bounded as shown below.

Proposition 3. Under Assumptions 2-3, for any $\|g\| \leq G$, the LMO (7) yields a non-empty set $\mathcal{L}_h(g) \subset \mathcal{B}_d(x^{(h)}, R) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d : \|x - x^{(h)}\| \leq R\}$.

Linear minimization oracle (LMO) has been adopted to develop generalized conditional gradient methods for composite optimization (Jiang & Zhang, 2014; Ghadimi, 2019). In the constrained optimization $\min_{x \in \Omega} F(x)$ as a special case, LMO reduces to $\arg \min_{y \in \Omega} \langle y, g \rangle$ used by the Frank-Wolfe algorithm (Frank et al., 1956; Lan & Zhou, 2016). Compared with the proximal operator (5), LMO can be computationally cheaper Juditsky & Nemirovski (2016). We will also use LMO to propose a computationally cheaper notion of stationary point as follows.

Definition 7. For any convex regularizer $h : \mathbb{R}^d \rightarrow \mathbb{R}$, we define the **δ -regularized Frank-Wolfe gap** of h at point $x \in \mathbb{R}^d$ and gradient $g \in \mathbb{R}^d$ as follows.

$$\mathcal{W}_h(x, g) \stackrel{\text{def}}{=} \max_{y \in \mathbb{R}^d} [h(x) - h(y) + \langle y - x, -g \rangle] \stackrel{\text{Eq. (7)}}{=} h(x) - h(y') + \langle y' - x, -g \rangle, \quad (8)$$

for any $y' \in \mathcal{L}_h(g)$. Furthermore, for any $\epsilon \geq 0$, we define $x \in \mathbb{R}^d$ as a **(δ, ϵ) -conditional gradient Goldstein stationary point (CGGSP)** if $\min_{g \in \partial_\delta F(x)} \mathcal{W}_h(x, g) \leq \epsilon$. Specifically, a $(0, \epsilon)$ -CGGSP is also called an ϵ -CGGSP, defined by $x \in \mathbb{R}^d$ such that $\min_{g \in \partial F(x)} \mathcal{W}_h(x, g) \leq \epsilon$.

Our proposed notions of CGGSP for nonsmooth composite optimization problem generalizes existing stationary notions for the following special cases.

- For constrained optimization problem $\min_{x \in \Omega} F(x)$, a special case of the nonconvex nonsmooth composite optimization problem (1) with $h = h_\Omega$ defined by Eq. (2), (δ, ϵ) -CGGSP reduces to the (δ, ϵ) -Clarke Frank-Wolfe stationary point (Liu et al., 2024).
- When F is differentiable, ϵ -CGGSP has simplified definition that $\|\mathcal{W}_h[x, \nabla F(x)]\| \leq \epsilon^3$, which can be achieved by conditional gradient descent algorithms within finite-time (Jiang & Zhang, 2014; Guo et al., 2022). In contrast, (γ, ϵ) -CGGSP is intractable for our setting with Lipschitz continuous and nondifferentiable F , as will be shown later in Theorem 1. Therefore, we aim at (δ, ϵ) -CGGSP, a relaxed notion of stationarity, and will propose a zeroth-order generalized conditional gradient algorithm (Algorithm 2) that achieves this point in finite time.

Our proposed notions of CGGSP satisfy the following properties.

³For differentiable function F , we have $\partial F(x) = \{\nabla F(x)\}$ (Makela & Neittaanmaki, 1992).

Proposition 4. Suppose the function $F : \mathbb{R}^d \rightarrow \mathbb{R}$ is differentiable and $h : \mathbb{R}^d \rightarrow \mathbb{R}$ satisfies Assumption 2. Then (δ, ϵ) -CGGSP has the following properties.

1. An ϵ -CGGSP is also a (δ, ϵ) -CGGSP.
2. Suppose ∇F is L -Lipschitz continuous. Then a $(\epsilon/(2RL), \epsilon/2)$ -CGGSP is also an ϵ -CGGSP, i.e., $\mathcal{W}_h[x, \nabla F(x)] \leq \epsilon$.
3. If $x \in \mathbb{R}^d$ satisfies $\mathcal{W}_h[x, \nabla F_\delta(x)] \leq \epsilon$, then x is a (δ, ϵ) -CGGSP.

Remark: Items 1-3 of Proposition 4 for CGGSP are analogous to items 1-3 of Proposition 2 for PGSP. Specifically, items 1 and 2 imply that for L -Lipschitz smooth functions F , our notions of (δ, ϵ) -CGGSP and ϵ' -CGGSP are equivalent (for possibly different $\epsilon, \epsilon' \geq 0$). Item 3 implies that we can obtain a (δ, ϵ) -CGGSP by solving $\min_{x \in \mathbb{R}^d} F_\delta(x)$, which is important for later designing the zeroth-order generalized conditional gradient algorithm (Algorithm 2).

Finally, Theorem 1 below shows that for composite problem (1) with general nonsmooth Lipschitz continuous function F , (γ, ϵ) -PGSP and ϵ -CGGSP are intractable. In contrast, $(\gamma, \delta, \epsilon)$ -PGSP and (δ, ϵ) -CGGSP can be achieved in finite time, by two zeroth-order algorithms presented in the two consequent sections respectively.

Theorem 1. Consider any $T \in \mathbb{N}$, $d \geq 2$ and any randomized algorithm $\mathcal{A}$ with access to a local oracle of the objective function (1)⁴ Then there exist functions F and h satisfying Assumptions 1-2 such that $\phi(0) - \inf_{x \in \mathbb{R}^d} \phi(x) \leq 2$ but with probability at least $1 - 2T \exp(-d/36)$, none of $\{x_t\}_{t=1}^T$ generalized by $\mathcal{A}$ belongs to the set of (γ, ϵ) -PGSP or ϵ -CGGSP for $\epsilon \in (0, \frac{1}{4\sqrt{2}})$ and $\gamma \in (0, 0.1]$.

4 ZERO-ORDER PROXIMAL GRADIENT DESCENT (0-PGD) ALGORITHM

In this section we study a zeroth-order proximal gradient descent (0-PGD) algorithm, as shown in Algorithm 1. The main algorithm framework is proximal gradient descent update (2) on the composite optimization problem $\min_{x \in \mathbb{R}^d} [F_\delta(x) + h(x)]$ that approximates the original problem (1), where the zeroth-order stochastic gradient estimator $g_t \approx \nabla F_\delta(x_t)$ is obtained using either minibatch estimation (option G1) or variance-reduced estimation (option G2).

We first present the convergence results of Algorithm 1 with minibatch estimation as follows.

Theorem 2 (Convergence of 0-PGD Algorithm with Minibatch Gradient Estimation). *Implement Algorithm 1 with Option G1, stepsize $\gamma = \frac{\delta}{cG\sqrt{d}}$ and constant batchsize $B_t \equiv B$. Then under Assumptions 1-2, the output $x_{\tilde{T}}$ has the following convergence rate.*

$$\mathbb{E}[\|\mathcal{G}_{\gamma h}(x_{\tilde{T}}, \nabla F_\delta(x_{\tilde{T}}))\|] \leq \frac{\sqrt{2cGd^{1/4}}}{\sqrt{T\delta}} \sqrt{\mathbb{E}[\phi(x_0)] - \phi_{\min} + 2\delta G} + \frac{16G\sqrt{d}}{\sqrt{B}} \quad (12)$$

where $\phi_{\min} \stackrel{\text{def}}{=} \min_{x \in \mathbb{R}^d} [F(x) + h(x)]$. Furthermore, we can obtain a $(\gamma, \delta, \epsilon)$ -PGSP by using hyperparameters $T = \mathcal{O}(Gd^{1/2}\delta^{-1}\epsilon^{-2})$, $B = \mathcal{O}(G^2d\epsilon^{-2})$ (see their full expressions in Eqs. (35) and (36) in Appendix I), which requires at most $2TB = \mathcal{O}(G^3d^{3/2}\delta^{-1}\epsilon^{-4})$ function evaluations and $T = \mathcal{O}(Gd^{1/2}\delta^{-1}\epsilon^{-2})$ proximal updates (2).

Comparison with Constrained Optimization: The stochastic constrained optimization problem $\min_{x \in \Omega} \{F(x) \stackrel{\text{def}}{=} \mathbb{E}_\xi[F_\xi(x)]\}$ on a convex and compact set Ω is a special case of the composition optimization problem (1) by using $h = h_\Omega$ defined in Eq. (2). Liu et al. (2024) studies this constrained optimization with also nonconvex, nonsmooth and G -Lipschitz continuous F , proposes a stochastic projected gradient descent algorithm as a special case of our Algorithm 1, and obtains the function evaluation complexity result $\mathcal{O}(G^4Rd^{3/2}\delta^{-1}\epsilon^{-4})$ to achieve a $(\gamma, \delta, \epsilon)$ -generalized Goldstein stationary point as a special case of our $(\gamma, \delta, \epsilon)$ -PGSP (see Corollary 5.2 of (Liu et al., 2024)). This complexity result requires Ω to be bounded with radius R and is higher than our $\mathcal{O}(G^3d^{3/2}\delta^{-1}\epsilon^{-4})$ that does not require R . Our improvement is obtained by replacing their bound $F_\delta(x_0) - F_\delta(x_T) \leq G\|x_0 - x_T\| \leq 2GR$ with the tighter bound $F_\delta(x_0) + h(x_0) - F_\delta(x_T) - h(x_T) \leq \phi(x_0) - \phi_{\min} + 2\delta G$.

⁴A local oracle means a quantity $\mathcal{O}_{F,h}(x)$ (e.g. $F(x), \nabla F(x) + \partial h(x)$) that reveals local information about the function values of F and h around a certain point $x \in \mathbb{R}^d$.

Algorithm 1 Zeroth-order proximal gradient descent (0-PGD) algorithm

```

1: Inputs: Number of iterations  $T$ , stepsize  $\gamma > 0$ , batchsizes  $B_t$ , radius  $\delta > 0$ , period  $q$  for
   variance reduction.
2: Initialize:  $x_0 \in \mathbb{R}^d$ .
3: for iterations  $t = 0, 1, \dots, T - 1$  do
4:   Obtain i.i.d. samples  $\{u_{i,t}\}_{i=1}^{B_t} \sim \text{Uniform}(\mathcal{S}_d(1))$  and  $\{\xi_{i,t}\}_{i=1}^{B_t} \sim \mathcal{P}$ .
5:   Obtain stochastic gradient estimation  $g_t \approx \nabla F_\delta(x_t)$  by either option below.
6:
7:   Option G1: Minibatch Estimation.

$$g_t = \frac{1}{B_t} \sum_{i=1}^{B_t} \hat{g}_\delta(x_t, u_{i,t}, \xi_{i,t}), \quad (9)$$

   where  $\hat{g}_\delta$  is defined by Eq. (4).
8:
9:   Option G2: Variance-reduced Estimation.
10:  if  $t \bmod q = 0$  then
11:    Obtain  $g_t$  by Eq. (9).
12:  else
13:    Obtain  $g_t$  by the following variance-reduced estimation.

$$g_t = g_{t-1} + \frac{1}{B_t} \sum_{i=1}^{B_t} [\hat{g}_\delta(x_t, u_{i,t}, \xi_{i,t}) - \hat{g}_\delta(x_{t-1}, u_{i,t}, \xi_{i,t})], \quad (10)$$

   where  $\hat{g}_\delta$  is defined by Eq. (4).
14:  end if
15:
16:  Update  $x_t$  by proximal gradient descent as follows.

$$x_{t+1} = \text{prox}_{\gamma h}(x_t - \gamma g_t) \stackrel{\text{def}}{=} \arg \min_{y \in \mathbb{R}^n} \left[ h(y) + \frac{1}{2\gamma} \|y - x_t + \gamma g_t\|^2 \right], \quad (11)$$

   where the proximal operator  $\text{prox}_{\gamma h}$  is defined by Eq. (5).
17: end for
18: Output:  $x_{\tilde{T}}$  where  $\tilde{T}$  is uniformly obtained from  $\{0, 1, \dots, T - 1\}$  at random.

```

Then using variance reduced gradient estimation, we obtain the following improved convergence rate and complexity results of Algorithms 1 as follows.

Theorem 3 (Convergence of 0-PGD Algorithm with Variance Reduction). *Implement Algorithm 1 with Option G2, stepsize $\gamma = \frac{\delta}{2G(d+c\sqrt{d})}$, batchsize $B_t = B_0$ for any $t \bmod q = 0$ and $B_t = B_1 = q$ for other t . Then under Assumptions 1-2, the output $x_{\tilde{T}}$ has the following convergence rate.*

$$\mathbb{E}[\|\mathcal{G}_{\gamma h}(x_{\tilde{T}}, \nabla F_\delta(x_{\tilde{T}}))\|] \leq \frac{\sqrt{2cG}d^{1/4}}{\sqrt{T}\delta} \sqrt{\mathbb{E}[\phi(x_0)] - \phi_{\min} + 2\delta G} + \frac{16G\sqrt{d}}{\sqrt{B}} \quad (13)$$

Furthermore, we can obtain a $(\gamma, \delta, \epsilon)$ -PGSP by using hyperparameters $B_0 = 1764dG^2\epsilon^{-2}$, $B_1 = q = 42\sqrt{d}G\epsilon^{-1}$, $T = \mathcal{O}(Gd\delta^{-1}\epsilon^{-2})$ (see their full expressions in Eq. (39) in Appendix J), which requires at most $2B_0\lfloor T/q \rfloor + 4B_1(T - \lfloor T/q \rfloor) = \mathcal{O}(G^2d^{3/2}\delta^{-1}\epsilon^{-3})$ function evaluations and $T = \mathcal{O}(Gd\delta^{-1}\epsilon^{-2})$ proximal updates (2).

Comparison with Existing Results: For stochastic nonconvex nonsmooth constrained optimization $\min_{x \in \Omega} \{F(x) \stackrel{\text{def}}{=} \mathbb{E}_\xi[F_\xi(x)]\}$ as a special case, Liu et al. (2024) obtains function evaluation complexity $\mathcal{O}(G^3Rd^{3/2}\delta^{-1}\epsilon^{-3})$, higher than our $\mathcal{O}(G^2d^{3/2}\delta^{-1}\epsilon^{-3})$ (see their Corollary 5.4). For unconstrained optimization $\min_{x \in \mathbb{R}^d} F(x)$, a smaller special case, Chen et al. (2023) uses variance reduction to achieve a (δ, ϵ) -Goldstein stationary point using variance reduction with also $\mathcal{O}(G^2d^{3/2}\delta^{-1}\epsilon^{-3})$ function evaluations that matches our complexity result (see their Theorem 1).

5 ZERO-ORDER GENERALIZED CONDITIONAL GRADIENT (0-GCG) ALGORITHM

In this section, we consider the case where the proximal operator (5) is costly. For example, when $h(x)$ is a nuclear norm of regularizer, the proximal operator requires full singular value decomposition (Wang et al., 2021a). The popular generalized conditional gradient method (Bredies et al., 2005; Jiang & Zhang, 2014; Rakotomamonjy et al., 2015; Bach, 2015; Nesterov, 2018; Ghadimi, 2019; Guo et al., 2022; Ito et al., 2023) uses a cheaper linear minimization oracle (LMO, defined by Eq. (7)) to replace the proximal operator. We propose a zeroth-order generalized conditional gradient method, using also two options of the zeroth-order gradient estimations, minibatch estimation (option G1) and variance-reduced estimation (option G2), as shown in Algorithm 2.

Algorithm 2 Zeroth-order generalized conditional gradient algorithm (0-GCG)

- 1: **Inputs:** Number of iterations T , stepsize $\gamma > 0$, batchsizes B_t , radius $\delta > 0$, period q for variance reduction.
- 2: **Initialize:** $x_0 \in \mathbb{R}^d$.
- 3: **for** iterations $t = 0, 1, \dots, T - 1$ **do**
- 4: Obtain i.i.d. samples $\{u_{i,t}\}_{i=1}^{B_t} \sim \text{Uniform}(\mathcal{S}_d(1))$ and $\{\xi_{i,t}\}_{i=1}^{B_t} \sim \mathcal{P}$.
- 5: Obtain stochastic gradient estimation $g_t \approx \nabla F_\delta(x_t)$ by option G1 or G2 in Algorithm 1.
- 6: Update x_t using LMO as follows.

$$y_t \in \mathcal{L}_h(g_t) \stackrel{\text{def}}{=} \arg \min_{y \in \mathbb{R}^d} [h(y) + \langle g_t, y \rangle], \quad (14)$$

$$x_{t+1} = x_t + \gamma(y_t - x_t). \quad (15)$$

- 7: **end for**
- 8: **Output:** $x_{\tilde{T}}$ where $\tilde{T}$ is uniformly obtained from $\{0, 1, \dots, T - 1\}$ at random.

We present the convergence rate and complexity results of Algorithm 2 in the following two theorems, for the two gradient estimation options respectively.

Theorem 4 (Convergence of 0-GCG Algorithm with Minibatch Gradient Estimation). *Implement Algorithm 2 with Option G1, stepsize $\gamma = \frac{1}{R} \sqrt{\frac{2\delta}{TcG\sqrt{d}}} \mathbb{E}[\phi(x_0) - \phi_{\min} + 2\delta G]$, constant batchsize $B_t \equiv B$ and initial point x_0 satisfying $\|x_0 - x^{(h)}\| \leq R$. Then under Assumptions 1-3, the output $x_{\tilde{T}}$ has the following convergence rate.*

$$\mathbb{E}[\mathcal{W}_h(x_{\tilde{T}}, \nabla F_\delta(x_{\tilde{T}}))] \leq R \sqrt{\frac{8cG\sqrt{d}}{T\delta} \mathbb{E}[\phi(x_0) - \phi_{\min} + 2\delta G]} + \frac{21RG\sqrt{d}}{\sqrt{B}}. \quad (16)$$

Furthermore, we can obtain a (δ, ϵ) -CGGSP by using hyperparameters $T = \mathcal{O}(GR^2 d^{1/2} \delta^{-1} \epsilon^{-2})$ (see its full expression in Eq. (43) in Appendix K), $B = 1764G^2 d R^2 \epsilon^{-2}$, which requires at most $2TB = \mathcal{O}(G^3 R^4 d^{3/2} \delta^{-1} \epsilon^{-4})$ function evaluations and $T = \mathcal{O}(GR^2 d^{1/2} \delta^{-1} \epsilon^{-2})$ LMO updates (14).

Theorem 5 (Convergence of 0-GCG Algorithm with Variance Reduction). *Implement Algorithm 2 with Option G2, stepsize $\gamma = \frac{1}{R} \sqrt{\frac{\delta \mathbb{E}[\phi(x_0) - \phi_{\min} + 2\delta G]}{TG(4d + 2c\sqrt{d})}}$, batchsize $B_t = B_0$ for any $t \bmod q = 0$ and $B_t = B_1 = q$ for other t . The initial point x_0 satisfies $\|x_0 - x^{(h)}\| \leq R$. Then under Assumptions 1-3, the output $x_{\tilde{T}}$ has the following convergence rate.*

$$\mathbb{E}[\|\mathcal{W}_h(x_{\tilde{T}}, \nabla F_\delta(x_{\tilde{T}}))\|] \leq 2R \sqrt{\frac{G(4d + 2c\sqrt{d})}{T\delta} \mathbb{E}[\phi(x_0) - \phi_{\min} + 2\delta G]} + \frac{13RG\sqrt{d}}{\sqrt{B_0}} \quad (17)$$

Furthermore, we can obtain a (δ, ϵ) -CGGSP by using hyperparameters $B_0 = 676dR^2 G^2 \epsilon^{-2}$, $B_1 = q = 26RG\epsilon^{-1}\sqrt{d}$, $T = \mathcal{O}(GR^2 d \delta^{-1} \epsilon^{-2})$ (see its full expression in Eq. (45) in Appendix L), which requires at most $2B_0 \lfloor T/q \rfloor + 4B_1(T - \lfloor T/q \rfloor) = \mathcal{O}(G^2 R^3 d^{3/2} \delta^{-1} \epsilon^{-3})$ function evaluations and $T = \mathcal{O}(GR^2 d \delta^{-1} \epsilon^{-2})$ LMO updates (14).

Comparison with Constrained Optimization: With minibatch gradient estimation, our function evaluation complexity $\mathcal{O}(G^3 R^4 d^{3/2} \delta^{-1} \epsilon^{-4})$ in Theorem 4 is more efficient than the complexity $\mathcal{O}(G^4 R^5 d^{3/2} \delta^{-1} \epsilon^{-4})$ to achieve (δ, ϵ) -Goldstein Frank–Wolfe stationary point of the stochastic nonconvex nonsmooth constrained optimization $\min_{x \in \Omega} \{F(x) \stackrel{\text{def}}{=} \mathbb{E}_\xi[F_\xi(x)]\}$, a special case of our (δ, ϵ) -CGGSP of the composite optimization problem (1) (Corollary 5.7 of (Liu et al., 2024)). Using variance reduction, our complexity improves to $\mathcal{O}(G^2 R^3 d^{3/2} \delta^{-1} \epsilon^{-3})$, which is also lower than $\mathcal{O}(G^3 R^4 d^{3/2} \delta^{-1} \epsilon^{-3})$ for $\min_{x \in \Omega} \{F(x) \stackrel{\text{def}}{=} \mathbb{E}_\xi[F_\xi(x)]\}$ (Corollary 5.9 of (Liu et al., 2024)).

6 EXPERIMENTS

We apply our zeroth-order algorithms to train a two-layer ReLU network $r_\xi(x) = W_2\sigma(W_1\xi + b_1) + b_2$. Here, $\xi \in \mathbb{R}^{d_\xi}$ is an input sample. The network parameters include the weight matrices ($W_1 \in \mathbb{R}^{d_1 \times d_\xi}$ and $W_2 \in \mathbb{R}^{d_2 \times d_1}$) and bias vectors ($b_1 \in \mathbb{R}^{d_1}$ and $b_2 \in \mathbb{R}^{d_2}$). $x \in \mathbb{R}^d$ denotes the total parameter which is concatenated by b_1 , b_2 and flattened W_1 , W_2 , so the total dimensionality is $d = d_1d_\xi + d_1d_2 + d_1 + d_2$. $\sigma : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_1}$ is the widely used ReLU activation function which maps each entry u to $\max(u, 0)$.

We select $d_\xi = 5$, $d_1 = 4$ and $d_2 = 2$ which imply $d = 34$, and generate the underlying sparse parameters $x^* \in \mathbb{R}^d$ by randomly selecting half of the entries to be 0 and generating the other half from standard Gaussian. Then we construct the binary classification dataset $\{(\xi_i, y_i)\}_{i=1}^N$ with sample size $N = 1000$, where the inputs $\xi_i \in \mathbb{R}^5$ are i.i.d. standard Gaussian, and the label $y_i = 0$ if the first entry of $f_{\xi_i}(x^*) \in \mathbb{R}^2$ is larger, otherwise $y_i = 1$. Then we train the regularized ReLU network via the following composite optimization problem.

$$\min_{x \in \mathbb{R}^d} \phi(x) = \frac{1}{N} \sum_{i=1}^N \ell[r_{\xi_i}(x), y_i] + \lambda_1 \|x\|_1 + \frac{\lambda_2}{2} \|x\|_2^2, \quad (18)$$

This can be seen as an instance of the problem (1), where the main part $F(x) = \frac{1}{N} \sum_{i=1}^N \ell[r_{\xi_i}(x), y_i]$ denotes the average cross-entropy loss between the prediction $r_{\xi_i}(x)$ and the true label y_i , and is nonsmooth due to the ReLU activation σ . In the convex regularizer $h(x) = \lambda_1 \|x\|_1 + \frac{\lambda_2}{2} \|x\|_2^2$, we select $\lambda_1 = \lambda_2 = 0.01$, $\|x\|_1$ induces sparse parameters and $\|x\|_2$ controls the parameter magnitude.

We implement our Algorithms 1 and 2, and for each algorithm we test both gradient estimation options, G1 (minibatch) and G2 (variance reduction), all with radius $\delta = 0.001$. For both algorithms with option G1 we select batchsize 500 and run 100 iterations. For both algorithms with option G2 we run 523 iterations, start each epoch of 10 iterations with batchsize 500, and use batchsize 50 for the rest iterations. We use fine-tuned stepsizes 0.005 for 0-PGD with G1, 0.001 for 0-PGD with G2, 5×10^{-5} for 0-GCG with G1, and 10^{-5} for 0-GCG with G2. The experiment is conducted on Python 3.9 using Apple M1 Pro with 8 cores and 16 GB memory, which costs about half a minute.

At each iteration t , we evaluate the training objective function $\phi(x_t)$ (Eq. (18)) as well as the classification accuracies on both the 1000 training samples and the 1000 heldout test samples generated in the same way as that of the training samples. In Figure 1, we plot these metrics VS the function evaluation complexity (the total number of function evaluations up to each iteration), which shows that all the algorithms converge well with over 90% accuracy on both training and test samples. In particular, compared with minibatch gradient estimation (option G1), after improving gradient estimation with variance reduction (option G2), both algorithms 0-PGD and 0-GCG converge faster.

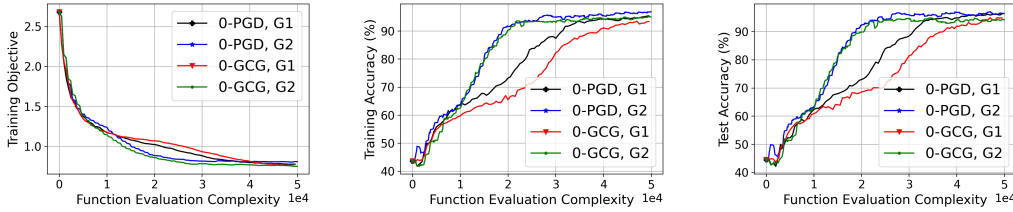


Figure 1: Experimental results on regularized ReLU network.

7 CONCLUSION

In this work, we have proposed two new notions of stationary points for stochastic nonconvex nonsmooth composite optimization, the $(\gamma, \delta, \epsilon)$ -proximal Goldstein stationary point (PGSP) and the (δ, ϵ) -conditional gradient Goldstein stationary point (CGGSP). We have also proved that the zeroth-order proximal gradient descent algorithm (0-PGD) and the zeroth-order generalized conditional gradient algorithm (0-GCG) converge to a $(\gamma, \delta, \epsilon)$ -PGSP and a (δ, ϵ) -CGGSP respectively, and obtained the convergence rates and complexity results. The experimental results on regularized ReLU network show that these algorithms converge well.

REFERENCES

- Liaqat Ali, Ashir Javeed, Adeeb Noor, Hafiz Tayyab Rauf, Seifedine Kadry, and Amir H Gandomi. Parkinson’s disease detection based on features refinement through l1 regularized svm and deep neural network. *Scientific Reports*, 14(1):1333, 2024.
- PB Assunção, OP Ferreira, and LF Prudente. A generalized conditional gradient method for multiobjective composite optimization problems. *Optimization*, 74(2):473–503, 2025.
- Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical programming*, 137(1):91–129, 2013.
- Francis Bach. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1):115–129, 2015.
- Heinz H Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.
- Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.
- Gábor Braun, Alejandro Carderera, Cyrille W Combettes, Hamed Hassani, Amin Karbasi, Aryan Mokhtari, and Sebastian Pokutta. Conditional gradient methods. *ArXiv:2211.14103*, 2022.
- Kristian Bredies, Dirk Lorenz, and Peter Maass. *Equivalence of a generalized conditional gradient method and the method of surrogate functionals*. Citeseer, 2005.
- Kristian Bredies, Dirk A Lorenz, and Peter Maass. A generalized conditional gradient method and its connection to an iterative shrinkage method. *Computational Optimization and applications*, 42: 173–193, 2009.
- Lesi Chen, Jing Xu, and Luo Luo. Faster gradient-free algorithms for nonsmooth nonconvex stochastic optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 5219–5233, 2023.
- Wang Chen, Liping Tang, and Xinmin Yang. Generalized conditional gradient methods for multiobjective composite optimization problems with h $\{\setminus\}$ o} lder condition. *ArXiv:2410.18465*, 2024.
- Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 6643–6670, 2023.
- Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.
- Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.
- Damek Davis, Dmitriy Drusvyatskiy, Yin Tat Lee, Swati Padmanabhan, and Guanghao Ye. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. In *Proceedings of the International Conference on Neural Information Processing Systems (Neurips)*, 2022.
- Kuangyu Ding, Jingyang Li, and Kim-Chuan Toh. Nonconvex stochastic bregman proximal gradient method with application to deep learning. *Journal of Machine Learning Research*, 26(39):1–44, 2025.
- John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

- Marguerite Frank et al. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.
- Masao Fukushima and Hisashi Mine. A generalized proximal point algorithm for certain non-convex minimization problems. *International Journal of Systems Science*, 12(8):989–1000, 1981.
- Saeed Ghadimi. Conditional gradient type methods for composite nonlinear and stochastic optimization. *Mathematical Programming*, 173:431–464, 2019.
- Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, 2016.
- Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1):267–305, 2016.
- Avrajit Ghosh, Xitong Zhang, Kenneth K Sun, Qing Qu, Saiprasad Ravishankar, and Rongrong Wang. Optimal eye surgeon: Finding image priors through sparse generators at initialization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.
- Allen A Goldstein. Optimization of lipschitz continuous functions. *Mathematical Programming*, 13: 14–22, 1977.
- Mingen Gu, Jinde Zheng, Haiyang Pan, and Jinyu Tong. Ramp sparse support matrix machine and its application in roller bearing fault diagnosis. *Applied Soft Computing*, 113:107928, 2021.
- Jiahong Guo, Huiling Liu, and Xiantao Xiao. A unified analysis of stochastic gradient-free frank–wolfe methods. *International Transactions in Operational Research*, 29(1):63–86, 2022.
- Zaid Harchaoui, Anatoli Juditsky, and Arkadi Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1):75–112, 2015.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Masaru Ito, Zhaosong Lu, and Chuan He. A parameter-free conditional gradient method for composite minimization under hölder condition. *Journal of Machine Learning Research*, 24(166):1–34, 2023.
- Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 427–435, 2013.
- Bo Jiang and Shuzhong Zhang. Iteration bounds for finding the ϵ -stationary points for structured nonconvex optimization. *ArXiv:1410.4066*, 2014.
- Michael Jordan, Guy Kornowski, Tianyi Lin, Ohad Shamir, and Manolis Zampetakis. Deterministic nonsmooth nonconvex optimization. In *Proceedings of the Conference on Learning Theory (COLT)*, pp. 4570–4597, 2023.
- Anatoli Juditsky and Arkadi Nemirovski. Solving variational inequalities with monotone operators on domains given by linear minimization oracles. *Mathematical Programming*, 156(1):221–256, 2016.
- Guy Kornowski and Ohad Shamir. An algorithm with optimal dimension-dependence for zero-order nonsmooth nonconvex stochastic optimization. *Journal of Machine Learning Research*, 25(122): 1–14, 2024.
- A Krizhevsky. Learning multiple layers of features from tiny images. *Master’s thesis, University of Toronto*, 2009.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
- Aakash Kumar, Ali Muhammad Shaikh, Yun Li, Hazrat Bilal, and Baoqun Yin. Pruning filters with l1-norm and capped l1-norm for cnn compression. *Applied Intelligence*, 51:1152–1160, 2021.

- Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *SIAM Journal on Optimization*, 26(2):1379–1409, 2016.
- Puya Latafat, Andreas Themelis, Masoud Ahookhosh, and Panagiotis Patrinos. Bregman finito/miso for nonconvex regularized finite sum minimization without lipschitz gradient continuity. *SIAM Journal on Optimization*, 32(3):2230–2262, 2022.
- Emanuel Laude and Panagiotis Patrinos. Anisotropic proximal gradient. *Mathematical Programming*, pp. 1–45, 2025.
- Huan Li and Zhouchen Lin. Accelerated proximal gradient methods for nonconvex programming. In *Proceedings of the International Conference on Neural Information Processing Systems (Neurips)*, pp. 379–387, 2015.
- Qunwei Li, Yi Zhou, Yingbin Liang, and Pramod K Varshney. Convergence analysis of proximal gradient with momentum for nonconvex optimization. In *International Conference on Machine Learning*, pp. 2111–2119, 2017.
- Yuxiang Li, Dongqing Wang, and Feng Liu. The auto-correlation function aided sparse support matrix machine for eeg-based fatigue detection. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 70(2):836–840, 2022.
- Zhize Li and Jian Li. A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. In *Proceedings of the International Conference on Neural Information Processing Systems (Neurips)*, pp. 5569–5579, 2018.
- Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 6083–6093, 2020.
- Tianyi Lin, Zeyu Zheng, and Michael I Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 26160–26175, 2022.
- Zhuanghua Liu, Cheng Chen, Luo Luo, and Bryan Kian Hsiang Low. Zeroth-order methods for constrained nonconvex nonsmooth stochastic optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.
- Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.
- Junqi Lu, Xijiong Xie, and Yujie Xiong. Multi-view hypergraph regularized lp norm least squares twin support vector machines for semi-supervised learning. *Pattern Recognition*, 156:110753, 2024.
- Shaocong Ma and Heng Huang. Revisiting zeroth-order optimization: Minimum-variance two-point estimators and directionally aligned perturbations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.
- Marko M Makela and Pekka Neittaanmaki. *Nonsmooth optimization: analysis and algorithms with applications to optimal control*. World Scientific, 1992.
- Morteza Mardani, Qingyun Sun, Shreyas Vasawanala, Vardan Papyan, Hatef Monajemi, John Pauly, and David Donoho. Neural proximal gradient descent for compressive imaging. In *Proceedings of the International Conference on Neural Information Processing Systems (Neurips)*, pp. 9596–9606, 2018.
- Arya Mazumdar and Ankit Singh Rawat. Learning and recovery in the relu model. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 108–115. IEEE, 2019.
- Jeremy A McCulloch, Skyler R St. Pierre, Kevin Linka, and Ellen Kuhl. On sparse regression, lp-regularization, and automated model discovery. *International Journal for Numerical Methods in Engineering*, 125(14):e7481, 2024.

- Yu Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. *Mathematical Programming*, 171(1):311–330, 2018.
- Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.
- Atsushi Nitanda. Stochastic proximal gradient descent with acceleration techniques. In *Proceedings of the International Conference on Neural Information Processing Systems (Neurips)*, pp. 1574–1582, 2014.
- Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014.
- Nhan H Pham, Lam M Nguyen, Dzung T Phan, and Quoc Tran-Dinh. Proxsarah: An efficient algorithmic framework for stochastic composite nonconvex optimization. *Journal of Machine Learning Research*, 21(110):1–48, 2020.
- Alain Rakotomamonjy, Rémi Flamary, and Nicolas Courty. Generalized conditional gradient: analysis of convergence and applications. *ArXiv:1510.06567*, 2015.
- Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alexander J Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Proceedings of the International Conference on Neural Information Processing Systems (Neurips)*, pp. 1153–1161, 2016.
- R Tyrrell Rockafellar. *Convex analysis*. Number 28. Princeton university press, 1970.
- Maohao Shen, Subhro Das, Kristjan Greenewald, Prasanna Sattigeri, Gregory Wornell, and Soumya Ghosh. Thermometer: towards universal calibration for large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 44687–44711, 2024.
- Lai Tian and Anthony Man-Cho So. No dimension-free deterministic algorithm computes approximate stationarities of lipschitzians. *Mathematical Programming*, 208(1):51–74, 2024.
- Lai Tian, Kaiwen Zhou, and Anthony Man-Cho So. On the finite-time complexity and practical computation of approximate stationarity concepts of lipschitz functions. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 21360–21379, 2022.
- Huajun Wang and Yuanhai Shao. Fast generalized ramp loss support vector machine for pattern classification. *Pattern Recognition*, 146:109987, 2024.
- Qingsong Wang and Deren Han. A bregman stochastic method for nonconvex nonsmooth problem beyond global lipschitz gradient continuity. *Optimization Methods and Software*, 38(5):914–946, 2023.
- Yaqing Wang, Quanming Yao, and James Kwok. A scalable, adaptive and sound nonconvex regularizer for low-rank matrix learning. In *Proceedings of the Web Conference*, pp. 1798–1808, 2021a.
- Yifei Wang, Jonathan Lacotte, and Mert Pilanci. The hidden convex optimization landscape of regularized two-layer relu networks: an exact characterization of optimal solutions. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021b.
- Zhixiang Xu, Gao Huang, Kilian Q Weinberger, and Alice X Zheng. Gradient boosted feature selection. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 522–531, 2014.
- Yingzhen Yang and Jiahui Yu. Fast proximal gradient descent for a class of non-convex and non-smooth sparse learning problems. In *Uncertainty in Artificial Intelligence (UAI)*, pp. 1253–1262, 2020.
- Yaoliang Yu, Xinhua Zhang, and Dale Schuurmans. Generalized conditional gradient for sparse estimation. *Journal of Machine Learning Research*, 18(144):1–46, 2017.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 11173–11182, 2020.

Tong Zhang. Multi-stage convex relaxation for learning with sparse regularization. In *Proceedings of the International Conference on Neural Information Processing Systems (Neurips)*, pp. 1929–1936, 2008.

Qingqing Zheng, Fengyuan Zhu, Jing Qin, Badong Chen, and Pheng-Ann Heng. Sparse support matrix machine. *Pattern Recognition*, 76:715–726, 2018.

Appendix

Table of Contents

A Related Works	15
B Experiments on Regularized Resnet	16
C Supporting Lemmas	18
D Proof of Proposition 1	19
E Proof of Proposition 2	19
F Proof of Proposition 3	20
G Proof of Proposition 4	20
H Proof of Theorem 1	21
I Proof of Theorem 2	22
J Proof of Theorem 3	23
K Proof of Theorem 4	25
L Proof of Theorem 5	26
M Use of Large Language Models (LLMs)	27

A RELATED WORKS

Finite-time Convergence Results on Nonconvex Nonsmooth Optimization: Only recently are finite-time convergence results obtained on nonconvex nonsmooth optimization, a special case of our nonconvex nonsmooth composite optimization (1). Davis & Drusvyatskiy (2019); Davis & Grimmer (2019) obtain finite-time convergence for stochastic optimization of a ρ -weakly convex function. Zhang et al. (2020) obtains the first dimension-free convergence result to achieve a (δ, ϵ) -Goldstein stationary point, which involves impractical subgradient computation. Such subgradient computation is removed by (Davis et al., 2022; Tian et al., 2022) using perturbations. Jordan et al. (2023); Tian & So (2024) prove that deterministic algorithms cannot obtain dimension-free convergence for nonconvex nonsmooth optimization. Cutkosky et al. (2023) obtains the optimal complexity result using online learning.

Nesterov & Spokoiny (2017) obtains the first finite-time convergence result of zeroth-order methods for stochastic nonconvex nonsmooth optimization. Lin et al. (2022) designs zeroth-order algorithms with provable finite-time convergence to (δ, ϵ) -Goldstein stationary point. Their oracle complexity is improved by (Chen et al., 2023) using variance reduction, and further improved to the optimal complexity $\mathcal{O}(d\delta^{-1}\epsilon^{-3})$ by (Kornowski & Shamir, 2024) using the online learning technique in (Cutkosky et al., 2023).

Proximal Gradient Methods: Various proximal gradient methods are very popular for various composite optimization problem (1). For example, Fukushima & Mine (1981) derives asymptotic convergence of proximal gradient method for smooth composite optimization problem ⁵ under

⁵Here, smooth composite optimization means the major part F of the composite optimization problem (1) is Lipschitz smooth.

both convex and nonconvex settings. Attouch et al. (2013) analyzes the convergence of multiple variants of inexact proximal algorithms⁶ on smooth nonconvex composite optimization satisfying Kurdyka-Łojasiewicz geometry. Ghadimi et al. (2016) analyzes the convergence of proximal gradient method for non-stochastic smooth composite optimization, and that of the two algorithm variants with minibatch and zeroth-order gradient estimations for stochastic smooth nonconvex and convex composite optimization problems. Beck & Teboulle (2009) applies Nesterov’s acceleration to proximal gradient method⁷ and uses backtracking technique to estimate Lipschitz constant, for smooth convex composite optimization. Nitanda (2014) combines Nesterov’s acceleration and SVRG variance reduction in proximal gradient methods and thus improved convergence rate for stochastic smooth convex composite optimization. (Li & Lin, 2015; Ghadimi & Lan, 2016; Li et al., 2017) study accelerated proximal gradient (APG) algorithms for nonconvex smooth composite optimization. Stochastic proximal gradient methods have been accelerated by variance reduction techniques, such as SVRG (Reddi et al., 2016; Li et al., 2017), SAGA (Reddi et al., 2016), SARAH (Pham et al., 2020), and adaptive APG algorithm with Spider variance reduction. Proximal gradient methods have also been extended from Euclidean distance to Bregman distance. For example, Bregman distance based proximal gradient method has provable convergence results for Bregman distance based relatively smooth composite optimization under both convex (Bauschke et al., 2017) and nonconvex settings (Latafat et al., 2022; Wang & Han, 2023). Ding et al. (2025) obtains the optimal sample complexity results of both Bregman proximal gradient method and its momentum variant for smooth adaptable composite optimization. Laude & Patrinos (2025) analyzes an anisotropic proximal gradient method for anisotropic smooth composite optimization.

Conditional Gradient Methods: Frank et al. (1956) proposes conditional gradient method (also known as Frank-Wolfe algorithm) for quadratic programming. Lan & Zhou (2016) extends conditional gradient method to convex optimization by skipping gradient evaluations, and achieved optimal computation complexity results. Bredies et al. (2005) proposes a generalized conditional gradient method which extends to composite optimization, the focus of this work, and obtains asymptotic convergence result for nonconvex setting. Since then, generalized conditional gradient methods have been applied to various composite optimization problems. For example, Jiang & Zhang (2014) studies nonconvex and nonsmooth composite optimization with block-structure. Bach (2015); Nesterov (2018) focus on general convex composite optimization problems. Harchaoui et al. (2015) studies norm-regularized convex optimization. Rakotomamonjy et al. (2015) obtains the non-asymptotic convergence rate of generalized conditional gradient method for convex composite optimization. Bach (2015) shows that the non-projected subgradient method for the primal convex composite optimization problem is equivalent to the conditional gradient applied to the dual optimization problem. Yu et al. (2017) improves generalized conditional gradient method for sparse optimization problems with convex gauge regularizers. Ghadimi (2019) focuses on smooth and weakly smooth nonconvex composite optimization problems. Ito et al. (2023) studies weakly convex composite optimization under Holder condition. Guo et al. (2022) provides a unified convergence analysis for zeroth-order conditional gradient methods on both stochastic constrained and composite optimization problems, under both convex and nonconvex settings. Recently, Chen et al. (2024); Assunção et al. (2025) extends conditional gradient methods to multiobjective composite optimization. See Braun et al. (2022) for a survey of conditional gradient methods.

B EXPERIMENTS ON REGULARIZED RESNET

We train a regularized Resnet-20 (He et al., 2016)⁸ with cross-entropy loss for classification task on the Cifar 10 image data (Krizhevsky, 2009), using our 0-PGD algorithm (Algorithm 1) and 0-GCG algorithm (Algorithm 2). In particular, we use the objective function (18) where ξ_i denotes an image-label pair in the Cifar 10 training set, and we select $\lambda_1 = \lambda_2 = 0.01$.

We implement our Algorithm 1 (0-PGD) and Algorithm 2 (0-GCG), and for each algorithm we test both gradient estimation options, G1 (minibatch) and G2 (variance reduction), all with radius $\delta = 0.001$. For both algorithms with option G1 we select batchsize 5000 and run 500 iterations.

⁶Proximal gradient method is called forward-backward splitting in (Attouch et al., 2013).

⁷Proximal gradient method is called iterative shrinkage-thresholding algorithms (ISTA) in (Beck & Teboulle, 2009)

⁸The Resnet-20 code comes from https://github.com/sarwaridas/ResNet20_PyTorch/blob/main/resnet_cifar10_TRIAL.ipynb

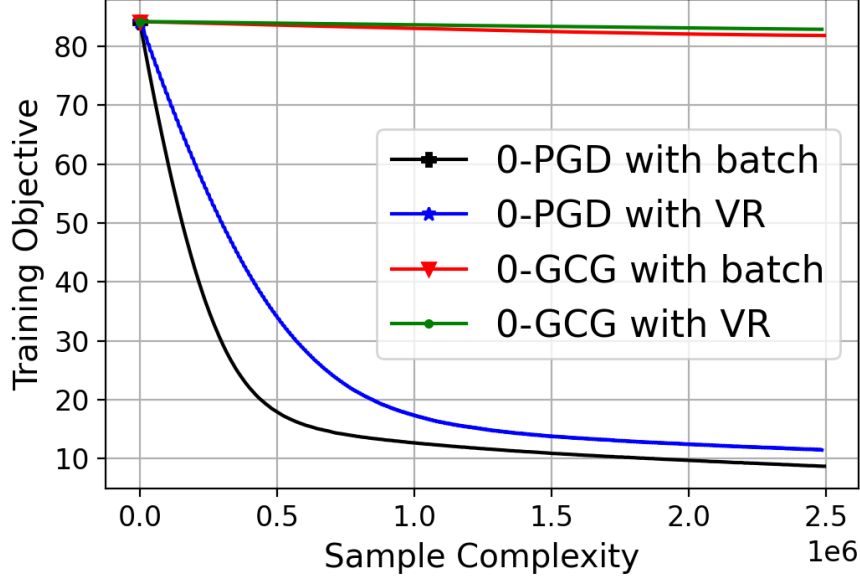


Figure 2: Experimental results on regularized Resnet-20.

For both algorithms with option G2 we run 2615 iterations, start each epoch of 10 iterations with batchsize 5000, and use batchsize 500 for the rest iterations. We use fine-tuned stepsizes 0.005 for 0-PGD with G1, 0.001 for 0-PGD with G2, 5×10^{-5} for 0-GCG with G1, and 10^{-5} for 0-GCG with G2. The experiment is conducted on Python 3.11.5 in Red Hat Enterprise Linux 8.10 (Ootpa), using 1 RTX A6000 GPU (48GB memory) and 4 CPU cores (20GB memory).

At each iteration t , we evaluate the objective function $\phi(x_t)$ (Eq. (18)) on the Cifar-10 training data. In Figure 2, we plot $\phi(x_t)$ VS the function evaluation complexity (the total number of function evaluations up to each iteration t), which shows that Algorithm 1 (0-PGD) converges well while Algorithm 2 (0-GCG) descends on the objective very slowly. When tuning hyperparameters for 0-GCG, we found that 0-GCG ascends and diverges even with slightly larger stepsizes, and descends slightly faster with fine-tuned stepsizes when using larger batchsizes (e.g. 50000 for option G1 and start of each epoch of option G2, and 5000 for the rest iterations for option G2, which is time consuming). This phenomenon can be largely explained by comparing the theoretical batchsizes required by these algorithms, as shown in Table 2 below, which indicates that the batchsizes required by 0-GCG depend on the regularizer-dependent radius $R > 0$ defined by Assumption 3 while 0-PGD does not depend on R . Therefore, when R is very large, 0-PGD can be much more efficient than 0-GCG.

Table 2: Batchsizes required by zeroth-order proximal gradient descent (0-PGD) and zeroth-order generalized conditional gradient algorithms (0-GCG).

	0-PGD (Algorithm 1)	0-GCG (Algorithm 2)
Batchsize B for option G1	$\mathcal{O}(G^2\epsilon^{-2}d)$ (Theorem 2)	$\mathcal{O}(R^2G^2\epsilon^{-2}d)$ (Theorem 4)
Large batchsize B_0 for option G2	$\mathcal{O}(G^2\epsilon^{-2}d)$ (Theorem 3)	$\mathcal{O}(R^2G^2\epsilon^{-2}d)$ (Theorem 5)
Small batchsize B_1 for option G2	$\mathcal{O}(G\epsilon^{-1}\sqrt{d})$ (Theorem 3)	$\mathcal{O}(RG\epsilon^{-1}\sqrt{d})$ (Theorem 5)

C SUPPORTING LEMMAS

Lemma 2 (Proposition 1 of (Ghadimi et al., 2016)). *For any $x, g, g' \in \mathbb{R}^d$, $\gamma > 0$ and proper convex function h , we have*

$$\|\mathcal{G}_{\gamma h}(x, g') - \mathcal{G}_{\gamma h}(x, g)\| \leq \|g' - g\| \quad (19)$$

Lemma 3 (Lemma 1 of (Ghadimi et al., 2016)). *For any $x, g \in \mathbb{R}^d$, $\gamma > 0$ and proper convex function h , we have*

$$\langle g, \mathcal{G}_{\gamma h}(x, g) \rangle \geq \|\mathcal{G}_{\gamma h}(x, g)\|^2 + \frac{1}{\gamma} [h[\text{prox}_{\gamma h}(x - \gamma g)] - h(x)]. \quad (20)$$

Lemma 4. *For i.i.d. random variables $\{X_i\}_{i=1}^N$, we have*

$$\mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N [X_i - \mathbb{E}(X_i)] \right\|^2 \right] = \frac{1}{N} \mathbb{E} [\|X_1 - \mathbb{E}X_1\|^2] \leq \frac{1}{N} \mathbb{E} [\|X_1\|^2] \quad (21)$$

Proof.

$$\begin{aligned} & \mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N [X_i - \mathbb{E}(X_i)] \right\|^2 \right] \\ &= \mathbb{E} \left[\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \langle X_i - \mathbb{E}(X_i), X_j - \mathbb{E}(X_j) \rangle \right] \\ &= \frac{1}{N^2} \sum_{i=1}^N \mathbb{E} [\|X_i - \mathbb{E}(X_i)\|^2] + \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1, j \neq i}^N \mathbb{E} [\langle X_i - \mathbb{E}(X_i), X_j - \mathbb{E}(X_j) \rangle] \\ &\stackrel{(a)}{=} \frac{1}{N^2} \sum_{i=1}^N \mathbb{E} [\|X_1 - \mathbb{E}(X_1)\|^2] + \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1, j \neq i}^N [\langle \mathbb{E}[X_i - \mathbb{E}(X_i)], \mathbb{E}[X_j - \mathbb{E}(X_j)] \rangle] \\ &\stackrel{(b)}{=} \frac{1}{N} \mathbb{E} [\|X_1 - \mathbb{E}(X_1)\|^2] \\ &= \frac{1}{N} \mathbb{E} [\langle X_1 - \mathbb{E}(X_1), X_1 - \mathbb{E}(X_1) \rangle] \\ &= \frac{1}{N} \mathbb{E} [\|X_1\|^2 - \langle \mathbb{E}(X_1), X_1 \rangle - \langle X_1, \mathbb{E}(X_1) \rangle + \|\mathbb{E}(X_1)\|^2] \\ &= \frac{1}{N} \mathbb{E} [\|X_1\|^2 - \|\mathbb{E}(X_1)\|^2] \\ &\stackrel{(c)}{\leq} \frac{1}{N} \mathbb{E} [\|X_1\|^2] \end{aligned} \quad (22)$$

where (a) uses the fact that $\{X_i\}_{i=1}^N$ are i.i.d. samples, (b) proves the "=" of Eq. (21), and (c) proves the "<=" of Eq. (21). $\square$

Lemma 5 (Lemma E.1 of (Lin et al., 2022)). *Suppose Assumption 1 holds and $\xi \sim \mathcal{P}$ and $u \in \text{Uniform}(\mathcal{S}_d(1))$ (uniformly distribution in $\mathcal{S}_d(1) := \{y \in \mathbb{R}^d : \|y\| = 1\}$). Then for any $x \in \mathbb{R}^d$, the stochastic gradient estimator (4) satisfies $\mathbb{E}[\hat{g}_\delta(x; u, \xi)|x] = \nabla F_\delta(x)$ and $\mathbb{E}[\|\hat{g}_\delta(x; u, \xi)\|^2|x] \leq 16\sqrt{2\pi}dG^2$.*

Lemma 6. *Suppose Assumption 1 holds and we have i.i.d. samples $\{u_{i,t}\}_{i=1}^{B_t} \sim \text{Uniform}(\mathcal{S}_d(1))$ and $\{\xi_{i,t}\}_{i=1}^{B_t} \sim \mathcal{P}$. Then the stochastic gradient estimator (9) satisfies $\mathbb{E}[g_t|x_t] = \nabla F_\delta(x_t)$ and $\mathbb{E}[\|g_t - \nabla F_\delta(x_t)\|^2|x_t] \leq \frac{16\sqrt{2\pi}dG^2}{B_t}$.*

Proof.

$$\mathbb{E}[g_t|x_t] = \mathbb{E} \left[\frac{1}{B_t} \sum_{i=1}^{B_t} \hat{g}_\delta(x_t, u_{i,t}, \xi_{i,t}) | x_t \right] = \frac{1}{B_t} \sum_{i=1}^{B_t} \mathbb{E}[\hat{g}_\delta(x_t, u_{i,t}, \xi_{i,t}) | x_t] \stackrel{(a)}{=} \nabla F_\delta(x_t) \quad (23)$$

where (a) uses Lemma 5.

$$\begin{aligned}\mathbb{E}[\|g_t - \nabla F_\delta(x_t)\|^2 | x_t] &= \mathbb{E}\left[\left\|\frac{1}{B_t} \sum_{i=1}^{B_t} [\hat{g}_\delta(x_t, u_{i,t}, \xi_{i,t}) - \nabla F_\delta(x_t)]\right\|^2 \middle| x_t\right] \\ &\stackrel{(a)}{\leq} \frac{1}{B_t} \mathbb{E}[\|\hat{g}_\delta(x_t, u_{1,t}, \xi_{1,t})\|^2] \\ &\stackrel{(b)}{\leq} \frac{16\sqrt{2\pi}dG^2}{B_t}\end{aligned}\tag{24}$$

where (a) uses $\mathbb{E}[\hat{g}_\delta(x_t, u_{i,t}, \xi_{i,t}) | x_t] = \nabla F_\delta(x_t)$ (based on Lemma 5) and Lemma 4, and (b) uses Lemma 5. $\square$

Lemma 7 (Lemma A.5 of (Liu et al., 2024)). *Implement Algorithm 1 or 2 with Option G2. Select batchsize $B_t = B_0$ for any $t \bmod q = 0$ and $B_t = B_1$ for other t . Then under Assumption 1, the stochastic gradient estimation $g_t \approx \nabla F_\delta(x_t)$ has the following error bound.*

$$\mathbb{E}[\|g_t - \nabla F_\delta(x_t)\|^2] \leq \frac{d^2 G^2}{\delta^2 B_1} \sum_{j=n_t q+1}^t \|x_j - x_{j-1}\|^2 + \frac{16\sqrt{2\pi}dG^2}{B_0}\tag{25}$$

where $n_t = \lfloor t/q \rfloor$. Specifically, when $t \bmod q = 0$ (i.e., $t = n_t q$), the upper bound above reduces to $\frac{16\sqrt{2\pi}dG^2}{B_0}$.

Lemma 8. *Implement Algorithm 2 with either Option G1 or G2, and the initialization x_0 satisfies $\|x_0 - x^{(h)}\| \leq R$. Then all the points x_t, y_t generated from Algorithm 2 satisfies $\|x_t - x^{(h)}\| \leq R$ and $\|y_t - x^{(h)}\| \leq R$.*

Proof. Based on Proposition 3, $y_t \in \mathcal{L}_h(g_t)$ satisfies $\|y_t - x^{(h)}\| \leq R$.

We will prove $\|x_t - x^{(h)}\| \leq R$ by induction. Suppose $\|x_k - x^{(h)}\| \leq R$ for a certain natural number k . Then the update rule (15) implies that

$$\|x_{k+1} - x^{(h)}\| = \|(1 - \gamma)x_k + \gamma y_k\| \leq (1 - \gamma)\|x_k\| + \gamma\|y_k\| \leq R.$$

Since $\|x_0 - x^{(h)}\| \leq R$, we have proved that $\|x_t - x^{(h)}\| \leq R$ for all t . $\square$

D PROOF OF PROPOSITION 1

Since h is a proper and closed convex function based on Assumption 2, the sub-level set $A = \{x \in \mathbb{R}^d : \phi(x) \leq \phi(x^{(h)})\}$ is a closed set in which h is continuous, based on Corollary 10.1.1 of (Rockafellar, 1970).

For any $x \in \mathbb{R}^d$ satisfying $\|x - x^{(h)}\| > R$, we have

$$\phi(x) - \phi(x^{(h)}) \stackrel{(a)}{=} F(x) - F(x^{(h)}) + h(x) - h(x^{(h)}) \stackrel{(b)}{>} -G\|x - x^{(h)}\| + G\|x - x^{(h)}\| = 0,$$

where (a) uses the objective function (1), (b) uses Assumption 3 and the G -Lipschitz continuity of F (based on Assumption 1). Therefore, $A \subset \mathcal{B}_d(x^{(h)}, R) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d : \|x - x^{(h)}\| \leq R\}$, so A is a compact set. Note that $\phi = F + h$ is continuous in A , so $\arg \max_{x \in A} \phi(x)$ is non-empty. Based on the definition of A , $\arg \max_{x \in \mathbb{R}^d} \phi(x) = \arg \max_{x \in A} \phi(x) \subset A \subset \mathcal{B}_d(x^{(h)}, R)$.

E PROOF OF PROPOSITION 2

Proof of Item 1: Item 1 is directly implied by Definition 5 and $\partial F(x) \subset \partial_\delta F(x)$.

Proof of Item 2: Since $x \in \mathbb{R}^d$ is a $(\gamma, \epsilon/(2L), \epsilon/2)$ -PGSP, there exists $g \in \partial_\delta F(x) = \text{conv}\{\nabla F(y) : \|y - x\| \leq \frac{\epsilon}{2L}\}$ such that $\|\mathcal{G}_{\gamma h}(x, g)\| \leq \frac{\epsilon}{2}$. Therefore, g can be written as the following convex combination of gradients

$$g = \sum_{k=1}^n \alpha_k \nabla F(x_k),$$

where $\alpha_k \geq 0$, $\sum_{k=1}^n \alpha_k = 1$ and $x_k \in \mathbb{R}^d$ satisfies $\|x_k - x\| \leq \frac{\epsilon}{2L}$. Then, we prove that x is a (γ, ϵ) -PGSP as follows.

$$\begin{aligned}
& \|\mathcal{G}_{\gamma h}[x, \nabla F(x)]\| \\
&= \frac{1}{\gamma} \|x - \text{prox}_{\gamma h}[x - \gamma \nabla F(x)]\| \\
&= \frac{1}{\gamma} \|x - \text{prox}_{\gamma h}(x - \gamma g)\| + \frac{1}{\gamma} \|\text{prox}_{\gamma h}(x - \gamma g) - \text{prox}_{\gamma h}[x - \gamma \nabla F(x)]\| \\
&\stackrel{(a)}{\leq} \|\mathcal{G}_{\gamma h}(x, g)\| + \frac{1}{\gamma} \|(x - \gamma g) - [x - \gamma \nabla F(x)]\| \\
&\leq \frac{\epsilon}{2} + \left\| \sum_{k=1}^n \alpha_k [\nabla F(x_k) - \nabla F(x)] \right\| \\
&\leq \frac{\epsilon}{2} + \sum_{k=1}^n \alpha_k \|\nabla F(x_k) - \nabla F(x)\| \\
&\stackrel{(b)}{\leq} \frac{\epsilon}{2} + L \sum_{k=1}^n \alpha_k \|x_k - x\| \\
&\leq \frac{\epsilon}{2} + L \sum_{k=1}^n \alpha_k \left(\frac{\epsilon}{2L} \right) = \epsilon,
\end{aligned}$$

where (a) uses Lemma 2, and (b) uses the L -Lipschitz continuity of ∇F .

Proof of Item 3: Item 3 is directly implied by Definition 5 and item 3 of Lemma 1.

F PROOF OF PROPOSITION 3

We will prove that Proposition 3 is a special case of Proposition 1. Specifically, in the original objective function (1), let $f_{\xi}(x) \equiv \langle x, g \rangle$ that does not depend on the stochastic sample ξ , which is L_{ξ} -Lipschitz where $L_{\xi} \equiv G$ and thus satisfies Assumption 1. Therefore, by applying Proposition 3, we conclude that $\arg \min_{y \in \mathbb{R}^d} [h(y) + \langle y, g \rangle] = \arg \min_{y \in \mathbb{R}^d} \phi(x)$ is a non-empty subset of $\mathcal{B}_d(x^{(h)}, R)$.

G PROOF OF PROPOSITION 4

Proof of Item 1: Item 1 is directly implied by Definition 7 and $\partial F(x) \subset \partial_{\delta} F(x)$.

Proof of Item 2: If $x \in \mathbb{R}^d$ is an $(\epsilon/(2RL), \epsilon/2)$ -CGGSP, then based on Definition 7, there exists $g \in \partial_{\delta} F(x)$ such that

$$\max_{y \in \mathbb{R}^d} [h(x) - h(y) + \langle y - x, -g \rangle] \leq \frac{\epsilon}{2}. \quad (26)$$

As F is differentiable, g can be written as the following convex combination of gradients

$$g = \sum_{k=1}^n \alpha_k \nabla F(x_k),$$

where $\alpha_k \geq 0$, $\sum_{k=1}^n \alpha_k = 1$ and $x_k \in \mathbb{R}^d$ satisfies $\|x_k - x\| \leq \frac{\epsilon}{2RL}$. Then we can prove Item 2 as follows.

$$\begin{aligned}
\mathcal{W}_h[x, \nabla F(x)] &\stackrel{(a)}{=} h(x) - h(y) + \langle y - x, -\nabla F(x) \rangle \\
&= h(x) - h(y) + \langle y - x, -g \rangle + \langle y - x, g - \nabla F(x) \rangle \\
&\leq h(x) - h(y) + \langle y - x, -g \rangle + \|y - x\| \cdot \|g - \nabla F(x)\| \\
&\stackrel{(b)}{\leq} \frac{\epsilon}{2} + R \left\| \sum_{k=1}^n \alpha_k [\nabla F(x_k) - \nabla F(x)] \right\|
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{\epsilon}{2} + R \sum_{k=1}^n \alpha_k \|\nabla F(x_k) - \nabla F(x)\| \\
&\stackrel{(c)}{\leq} \frac{\epsilon}{2} + R \sum_{k=1}^n \alpha_k L \|x_k - x\| \\
&\stackrel{(d)}{\leq} \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,
\end{aligned}$$

where (a) holds for any $y \in \mathcal{L}_h[x, \nabla F(x)]$ based on Eqs. (8) and (7), (b) uses Eq. (26), Proposition 3 and $\sum_{k=1}^n \alpha_k = 1$, (c) uses the L -Lipschitz continuity of ∇F , and (d) uses $\|x_k - x\| \leq \frac{\epsilon}{2RL}$ and $\sum_{k=1}^n \alpha_k = 1$.

Proof of Item 3: Item 3 is directly implied by Definition 7 and item 3 of Lemma 1.

H PROOF OF THEOREM 1

The proof of Theorem 4.7 in (Liu et al., 2024) has designed a function F^9 that satisfies Assumption 1 that F is Lipschitz continuous, and satisfies $F(0) - \inf_{x \in \Omega} F(x) \leq 2$ where $\Omega = [-100, 100]^d$ is a convex and compact set. The conclusion of their Theorem 4.7 states each point x_t generated from the algorithm $\mathcal{A}$ satisfies the following inequalities for $\gamma \in (0, 0.1]$.

$$\min_{g \in \partial F(x_t)} \left[\frac{1}{\gamma} \|x_t - \psi(x_t, g, \gamma)\| \right] \geq \frac{1}{4\sqrt{2}}, \quad (27)$$

$$\min_{g \in \partial F(x_t)} \max_{u \in \Omega} \langle u - x_t, -g \rangle \geq \frac{1}{4\sqrt{2}}, \quad (28)$$

where

$$\psi(x, g, \gamma) = \arg \min_{y \in \Omega} \left(\langle g, y \rangle + \frac{1}{2\gamma} \|y - x\|^2 \right). \quad (29)$$

Select $h = h_\Omega$ defined by Eq. (2), which yields the following equations.

$$\begin{aligned}
\text{prox}_{\gamma h}(x_t - \gamma g) &\stackrel{(a)}{=} \arg \min_{y \in \mathbb{R}^n} \left[h_\Omega(y) + \frac{1}{2\gamma} \|y - x_t + \gamma g\|^2 \right] \\
&\stackrel{(b)}{=} \arg \min_{y \in \Omega} \left[\frac{1}{2\gamma} \|y - x_t\|^2 + \langle g, y - x_t \rangle \right] \stackrel{(c)}{=} \psi(x_t, g, \gamma),
\end{aligned} \quad (30)$$

$$\begin{aligned}
\min_{g \in \partial F(x_t)} \mathcal{W}_h(x_t, g) &\stackrel{(d)}{=} \min_{g \in \partial F(x_t)} \max_{y \in \mathbb{R}^d} [h_\Omega(x_t) - h_\Omega(y) + \langle y - x_t, -g \rangle] \\
&\stackrel{(b)}{\geq} \min_{g \in \partial F(x_t)} \max_{y \in \Omega} \langle y - x_t, -g \rangle \stackrel{(e)}{\geq} \frac{1}{4\sqrt{2}},
\end{aligned} \quad (31)$$

$$\phi(0) - \inf_{x \in \mathbb{R}^d} \phi(x) = F(0) + h_\Omega(0) - \inf_{x \in \mathbb{R}^d} [F(x) + h_\Omega(x)] \stackrel{(b)}{=} F(0) - \inf_{x \in \Omega} F(x) \leq 2, \quad (32)$$

where (a)-(e) use Eqs. (5), (2), (29), (8) and (28) respectively. Therefore,

$$\begin{aligned}
\min_{g \in \partial F(x_t)} \|\mathcal{G}_{\gamma h}(x_t, g)\| &\stackrel{(a)}{=} \min_{g \in \partial F(x_t)} \left[\frac{1}{\gamma} \|x_t - \text{prox}_{\gamma h}(x_t - \gamma g)\| \right] \\
&\stackrel{(b)}{=} \min_{g \in \partial F(x_t)} \left[\frac{1}{\gamma} \|x_t - \psi(x_t, g, \gamma)\| \right] \stackrel{(c)}{\geq} \frac{1}{4\sqrt{2}},
\end{aligned} \quad (33)$$

where (a) uses Eq. (6), (b) uses Eq. (30) and (c) uses Eq. (27).

Eq. (33) implies that x_t is not a (γ, ϵ) -PGSP for $\epsilon \leq \frac{1}{4\sqrt{2}}$. Eq. (31) implies that x_t is not an ϵ -CGGSP for $\epsilon \leq \frac{1}{4\sqrt{2}}$. These implications along with Eq. (32) conclude the proof.

⁹This function is denoted as F_w in (Liu et al., 2024).

I PROOF OF THEOREM 2

Since ∇F_δ is $\frac{cG\sqrt{d}}{\delta}$ -Lipschitz continuous based on item 2 of Lemma 1, we obtain that

$$\begin{aligned}
F_\delta(x_{t+1}) &\leq F_\delta(x_t) + \langle \nabla F_\delta(x_t), x_{t+1} - x_t \rangle + \frac{cG\sqrt{d}}{2\delta} \|x_{t+1} - x_t\|^2 \\
&\stackrel{(a)}{=} F_\delta(x_t) - \gamma \langle \nabla F_\delta(x_t), \mathcal{G}_{\gamma h}(x_t, g_t) \rangle + \frac{cG\gamma^2\sqrt{d}}{2\delta} \|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2 \\
&= F_\delta(x_t) - \gamma \langle g_t, \mathcal{G}_{\gamma h}(x_t, g_t) \rangle + \gamma \langle g_t - \nabla F_\delta(x_t), \mathcal{G}_{\gamma h}(x_t, g_t) \rangle + \frac{cG\gamma^2\sqrt{d}}{2\delta} \|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2 \\
&\stackrel{(b)}{\leq} F_\delta(x_t) - \gamma \left[\|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2 + \frac{1}{\gamma} [h[\text{prox}_{\gamma h}(x_t - \gamma g_t)] - h(x_t)] \right] \\
&\quad + \gamma \langle g_t - \nabla F_\delta(x_t), \mathcal{G}_{\gamma h}(x_t, g_t) \rangle + \frac{\gamma}{2} \|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2 \\
&\stackrel{(c)}{=} F_\delta(x_t) - \frac{\gamma}{2} \|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2 + h(x_t) - h(x_{t+1}) + \gamma \langle g_t - \nabla F_\delta(x_t), \mathcal{G}_{\gamma h}[x_t, \nabla F_\delta(x_t)] \rangle \\
&\quad + \gamma \langle g_t - \nabla F_\delta(x_t), \mathcal{G}_{\gamma h}(x_t, g_t) - \mathcal{G}_{\gamma h}[x_t, \nabla F_\delta(x_t)] \rangle \\
&\leq F_\delta(x_t) - \frac{\gamma}{2} \|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2 + h(x_t) - h(x_{t+1}) + \gamma \langle g_t - \nabla F_\delta(x_t), \mathcal{G}_{\gamma h}[x_t, \nabla F_\delta(x_t)] \rangle \\
&\quad + \gamma \|g_t - \nabla F_\delta(x_t)\| \cdot \|\mathcal{G}_{\gamma h}(x_t, g_t) - \mathcal{G}_{\gamma h}[x_t, \nabla F_\delta(x_t)]\| \\
&\stackrel{(d)}{\leq} F_\delta(x_t) - \frac{\gamma}{2} \|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2 + h(x_t) - h(x_{t+1}) + \gamma \langle g_t - \nabla F_\delta(x_t), \mathcal{G}_{\gamma h}[x_t, \nabla F_\delta(x_t)] \rangle \\
&\quad + \gamma \|g_t - \nabla F_\delta(x_t)\|^2,
\end{aligned}$$

where (a) and (c) use the update rule (11) that $x_{t+1} = \text{prox}_{\gamma h}(x_t - \gamma g_t) = x_t - \gamma \mathcal{G}_{\gamma h}(x_t, g_t)$, (b) uses Lemma 3 and the stepsize $\gamma = \frac{\delta}{cG\sqrt{d}}$, and (d) uses Lemma 2. Rearranging the inequality above, and taking expectation, we obtain that

$$\begin{aligned}
\frac{\gamma}{2} \mathbb{E}[\|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2] &\leq \mathbb{E}[F_\delta(x_t) + h(x_t) - F_\delta(x_{t+1}) - h(x_{t+1})] \\
&\quad + \gamma \mathbb{E}[\langle g_t - \nabla F_\delta(x_t), \mathcal{G}_{\gamma h}[x_t, \nabla F_\delta(x_t)] \rangle] + \gamma \mathbb{E}[\|g_t - \nabla F_\delta(x_t)\|^2] \\
&\leq \mathbb{E}[F_\delta(x_t) + h(x_t) - F_\delta(x_{t+1}) - h(x_{t+1})] + \frac{16\gamma\sqrt{2\pi}dG^2}{B_t},
\end{aligned}$$

where the second $\leq$ uses Lemma 6. Rearranging the inequality above and summing over $t = 0, 1, \dots, T-1$, we have

$$\begin{aligned}
\mathbb{E}[\|\mathcal{G}_{\gamma h}(x_{\tilde{T}}, g_{\tilde{T}})\|^2] &= \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2] \\
&\leq \frac{1}{T} \sum_{t=0}^{T-1} \left[\frac{2}{\gamma} \mathbb{E}[F_\delta(x_t) + h(x_t) - F_\delta(x_{t+1}) - h(x_{t+1})] + \frac{32\sqrt{2\pi}dG^2}{B_t} \right] \\
&\stackrel{(a)}{\leq} \frac{2}{T\gamma} \mathbb{E}[F_\delta(x_0) + h(x_0) - F_\delta(x_T) - h(x_T)] + \frac{32\sqrt{2\pi}dG^2}{B} \\
&\stackrel{(b)}{\leq} \frac{2cG\sqrt{d}}{T\delta} \mathbb{E}[F(x_0) + h(x_0) - F(x_T) - h(x_T) + 2\delta G] + \frac{32\sqrt{2\pi}dG^2}{B} \\
&\stackrel{(c)}{\leq} \frac{2cG\sqrt{d}}{T\delta} \mathbb{E}[\phi(x_0) - \phi_{\min} + 2\delta G] + \frac{32\sqrt{2\pi}dG^2}{B}, \tag{34}
\end{aligned}$$

where (a) uses constant batchsize $B_t \equiv B$ and stepsize $\gamma = \frac{\delta}{cG\sqrt{d}}$, (b) uses item 1 of Lemma 1, (c) uses $\phi \stackrel{\text{def}}{=} F + g$ and $\phi_{\min} \stackrel{\text{def}}{=} \min_{x \in \mathbb{R}^d} \phi(x)$. Therefore, we can obtain the convergence rate (12) as follows.

$$\mathbb{E}[\|\mathcal{G}_{\gamma h}(x_{\tilde{T}}, \nabla F_\delta(x_{\tilde{T}}))\|] \leq \mathbb{E}[\|\mathcal{G}_{\gamma h}(x_{\tilde{T}}, g_{\tilde{T}})\|] + \mathbb{E}[\|\mathcal{G}_{\gamma h}(x_{\tilde{T}}, \nabla F_\delta(x_{\tilde{T}})) - \mathcal{G}_{\gamma h}(x_{\tilde{T}}, g_{\tilde{T}})\|]$$

$$\begin{aligned}
&\stackrel{(a)}{\leq} \sqrt{\mathbb{E}[\|\mathcal{G}_{\gamma h}(x_{\tilde{T}}, g_{\tilde{T}})\|^2]} + \sqrt{\mathbb{E}[\|\nabla F_{\delta}(x_{\tilde{T}}) - g_{\tilde{T}}\|^2]} \\
&\stackrel{(b)}{\leq} \sqrt{\frac{2cG\sqrt{d}}{T\delta} \mathbb{E}[\phi(x_0) - \phi_{\min} + 2\delta G] + \frac{32\sqrt{2\pi}dG^2}{B}} + \sqrt{\frac{16\sqrt{2\pi}dG^2}{B}} \\
&\stackrel{(c)}{\leq} \frac{\sqrt{2cGd^{1/4}}}{\sqrt{T\delta}} \sqrt{\mathbb{E}[\phi(x_0)] - \phi_{\min} + 2\delta G} + \frac{16G\sqrt{d}}{\sqrt{B}},
\end{aligned}$$

where (a) uses Lemma 2, (b) uses Eq. (34) and Lemma 6, (c) uses $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$ and $\sqrt{32\sqrt{2\pi}} + \sqrt{16\sqrt{2\pi}} < 16$

Furthermore, we can select the following hyperparameters.

$$T = \frac{8cG\sqrt{d}}{\delta\epsilon^2} [\mathbb{E}[\phi(x_0)] - \phi_{\min} + 2\delta G] = \mathcal{O}(Gd^{1/2}\delta^{-1}\epsilon^{-2}), \quad (35)$$

$$B = \frac{1024dG^2}{\epsilon^2} = \mathcal{O}(G^2d\epsilon^{-2}). \quad (36)$$

Then substituting the hyperparameters above into the convergence rate (12), we obtain the following bound, which based on item 3 of Proposition 2 implies that there exists at least one $(\gamma, \delta, \epsilon)$ -PGSP in $\{x_t\}_{t=0}^{T-1}$.

$$\min_{0 \leq t \leq T-1} \mathbb{E}[\|\mathcal{G}_{\gamma h}(x_t, \nabla F_{\delta}(x_t))\|] \leq \mathbb{E}[\|\mathcal{G}_{\gamma h}(x_{\tilde{T}}, \nabla F_{\delta}(x_{\tilde{T}}))\|] \leq \epsilon.$$

J PROOF OF THEOREM 3

Since ∇F_{δ} is $\frac{cG\sqrt{d}}{\delta}$ -Lipschitz continuous based on item 2 of Lemma 1, we obtain that

$$\begin{aligned}
F_{\delta}(x_{t+1}) &\leq F_{\delta}(x_t) + \langle \nabla F_{\delta}(x_t), x_{t+1} - x_t \rangle + \frac{cG\sqrt{d}}{2\delta} \|x_{t+1} - x_t\|^2 \\
&\stackrel{(a)}{=} F_{\delta}(x_t) - \gamma \langle \nabla F_{\delta}(x_t), \mathcal{G}_{\gamma h}(x_t, g_t) \rangle + \frac{cG\gamma^2\sqrt{d}}{2\delta} \|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2 \\
&= F_{\delta}(x_t) - \gamma \langle g_t, \mathcal{G}_{\gamma h}(x_t, g_t) \rangle + \gamma \langle g_t - \nabla F_{\delta}(x_t), \mathcal{G}_{\gamma h}(x_t, g_t) \rangle + \frac{cG\gamma^2\sqrt{d}}{2\delta} \|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2 \\
&\stackrel{(b)}{\leq} F_{\delta}(x_t) - \gamma [\|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2 + \frac{1}{\gamma} [h[\text{prox}_{\gamma h}(x_t - \gamma g_t)] - h(x_t)]] \\
&\quad + \frac{\gamma}{2} \|g_t - \nabla F_{\delta}(x_t)\|^2 + \left(\frac{cG\gamma^2\sqrt{d}}{2\delta} + \frac{\gamma}{2} \right) \|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2 \\
&\stackrel{(c)}{=} F_{\delta}(x_t) + h(x_t) - h(x_{t+1}) + \frac{\gamma}{2} \|g_t - \nabla F_{\delta}(x_t)\|^2 + \left(\frac{cG\gamma^2\sqrt{d}}{2\delta} - \frac{\gamma}{2} \right) \|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2,
\end{aligned}$$

where (a) and (c) use the update rule (11) that $x_{t+1} = \text{prox}_{\gamma h}(x_t - \gamma g_t) = x_t - \gamma \mathcal{G}_{\gamma h}(x_t, g_t)$, and (b) uses Lemma 3 and the inequality that $\langle u, v \rangle \leq (\|u\|^2 + \|v\|^2)/2$ for $u = g_t - \nabla F_{\delta}(x_t)$, $v = \mathcal{G}_{\gamma h}(x_t, g_t)$. Rearranging the inequality above, and taking expectation, we obtain that

$$\begin{aligned}
&\left(\frac{\gamma}{2} - \frac{cG\gamma^2\sqrt{d}}{2\delta} \right) \mathbb{E}[\|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2] \\
&\leq \mathbb{E}[F_{\delta}(x_t) + h(x_t) - F_{\delta}(x_{t+1}) - h(x_{t+1})] + \frac{\gamma}{2} \mathbb{E}[\|g_t - \nabla F_{\delta}(x_t)\|^2] \\
&\stackrel{(a)}{\leq} \mathbb{E}[F_{\delta}(x_t) + h(x_t) - F_{\delta}(x_{t+1}) - h(x_{t+1})] + \frac{\gamma d^2 G^2}{2\delta^2 B_1} \sum_{j=n_t q+1}^t \mathbb{E}[\|x_j - x_{j-1}\|^2] + \frac{8\gamma\sqrt{2\pi}dG^2}{B_0},
\end{aligned}$$

where (a) uses Lemma 7. Summing the inequality above over $t = 0, 1, \dots, T-1$, we have

$$\left(\frac{\gamma}{2} - \frac{cG\gamma^2\sqrt{d}}{2\delta} \right) \sum_{t=0}^{T-1} \mathbb{E}[\|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2]$$

$$\begin{aligned}
&\leq \mathbb{E}[F_\delta(x_0) + h(x_0) - F_\delta(x_T) - h(x_T)] + \frac{\gamma d^2 G^2}{2\delta^2 B_1} \sum_{t=0}^{T-1} \sum_{j=n_t q+1}^t \|x_j - x_{j-1}\|^2 + \frac{8T\gamma\sqrt{2\pi}dG^2}{B_0} \\
&\stackrel{(a)}{\leq} \mathbb{E}[F(x_0) + h(x_0) - F(x_T) - h(x_T)] + 2\delta G \\
&\quad + \frac{\gamma^3 d^2 G^2}{2\delta^2 B_1} \sum_{t=0}^{T-1} \sum_{j=n_t q+1}^{(n_t+1)q} \|\mathcal{G}_{\gamma h}(x_j, g_j)\|^2 + \frac{8T\gamma\sqrt{2\pi}dG^2}{B_0} \\
&\stackrel{(b)}{\leq} \mathbb{E}[\phi(x_0) - \phi_{\min}] + 2\delta G + \frac{q\gamma^3 d^2 G^2}{2\delta^2 B_1} \sum_{t=0}^{T-1} \mathbb{E}[\|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2] + \frac{8T\gamma\sqrt{2\pi}dG^2}{B_0}, \tag{37}
\end{aligned}$$

where (a) uses $t < (n_t + 1)q$ ($n_t = \lfloor t/q \rfloor$), item 1 of Lemma 1 and the update rule (11) that $x_{j+1} = \text{prox}_{\gamma h}(x_j - \gamma g_j) = x_j - \gamma \mathcal{G}_{\gamma h}(x_j, g_j)$, and (b) uses $\phi \stackrel{\text{def}}{=} F + g$ and $\phi_{\min} \stackrel{\text{def}}{=} \min_{x \in \mathbb{R}^d} \phi(x)$. Rearranging the inequality above, we obtain that

$$\begin{aligned}
&\mathbb{E}[\|\mathcal{G}_{\gamma h}(x_{\tilde{T}}, g_{\tilde{T}})\|^2] \\
&= \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2] \\
&\leq \frac{1}{T} \left(\frac{\gamma}{2} - \frac{cG\gamma^2\sqrt{d}}{2\delta} - \frac{q\gamma^3 d^2 G^2}{2\delta^2 B_1} \right)^{-1} \left[\mathbb{E}[\phi(x_0)] - \phi_{\min} + 2\delta G + \frac{8T\gamma\sqrt{2\pi}dG^2}{B_0} \right] \\
&\leq \frac{16G(d + c\sqrt{d})}{T\delta} \left[\mathbb{E}[\phi(x_0)] - \phi_{\min} + 2\delta G \right] + \frac{64\sqrt{2\pi}dG^2}{B_0} \tag{38}
\end{aligned}$$

where (a) uses the following inequality with stepsize $\gamma = \frac{\delta}{2G(d+c\sqrt{d})}$ and batchsize $B_1 = q$, and (b) uses stepsize $\gamma = \frac{\delta}{2G(d+c\sqrt{d})}$.

$$\begin{aligned}
\frac{\gamma}{2} - \frac{cG\gamma^2\sqrt{d}}{2\delta} - \frac{q\gamma^3 d^2 G^2}{2\delta^2 B_1} &= \frac{\gamma}{2} \left(1 - \frac{cG\gamma\sqrt{d}}{\delta} - \frac{\gamma^2 d^2 G^2}{\delta^2} \right) \\
&\geq \frac{\delta}{4G(d + c\sqrt{d})} \left(1 - \frac{c\sqrt{d}}{2(d + c\sqrt{d})} - \frac{d^2}{4(d + c\sqrt{d})^2} \right) \\
&\geq \frac{\delta}{4G(d + c\sqrt{d})} \left(1 - \frac{1}{2} - \frac{1}{4} \right) = \frac{\delta}{16G(d + c\sqrt{d})}.
\end{aligned}$$

Then we have the following bound.

$$\begin{aligned}
&\mathbb{E}[\|\mathcal{G}_{\gamma h}(x_{\tilde{T}}, \nabla F_\delta(x_{\tilde{T}}))\|^2] \\
&\leq 2\mathbb{E}[\|\mathcal{G}_{\gamma h}(x_{\tilde{T}}, g_{\tilde{T}})\|^2] + 2\mathbb{E}[\|\mathcal{G}_{\gamma h}(x_{\tilde{T}}, \nabla F_\delta(x_{\tilde{T}})) - \mathcal{G}_{\gamma h}(x_{\tilde{T}}, g_{\tilde{T}})\|^2] \\
&\stackrel{(a)}{\leq} 2\mathbb{E}[\|\mathcal{G}_{\gamma h}(x_{\tilde{T}}, g_{\tilde{T}})\|^2] + 2\mathbb{E}[\|\nabla F_\delta(x_{\tilde{T}}) - g_{\tilde{T}}\|^2] \\
&= \frac{2}{T} \sum_{t=0}^{T-1} [\mathbb{E}[\|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2] + \mathbb{E}[\|\nabla F_\delta(x_t) - g_t\|^2]] \\
&\stackrel{(b)}{\leq} \frac{2}{T} \sum_{t=0}^{T-1} \left[\mathbb{E}[\|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2] + \frac{d^2 G^2}{\delta^2 B_1} \sum_{j=n_t q+1}^t \|x_j - x_{j-1}\|^2 + \frac{16\sqrt{2\pi}dG^2}{B_0} \right] \\
&\stackrel{(c)}{\leq} \frac{2}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2] + \frac{2d^2 G^2 \gamma^2}{T\delta^2 q} \sum_{t=0}^{T-1} \sum_{j=n_t q+1}^{(n_t+1)q} \mathbb{E}[\|\mathcal{G}_{\gamma h}(x_j, g_j)\|^2] \\
&\leq \left(\frac{2}{T} + \frac{2d^2 G^2 \gamma^2}{T\delta^2} \right) \sum_{t=0}^{T-1} \mathbb{E}[\|\mathcal{G}_{\gamma h}(x_t, g_t)\|^2] \\
&\stackrel{(d)}{\leq} \left(2 + \frac{d^2}{2(d + c\sqrt{d})^2} \right) \left\{ \frac{16G(d + c\sqrt{d})}{T\delta} \left[\mathbb{E}[\phi(x_0)] - \phi_{\min} + 2\delta G \right] + \frac{64\sqrt{2\pi}dG^2}{B_0} \right\}
\end{aligned}$$

$$\leq \frac{40G(d + c\sqrt{d})}{T\delta} \left[\mathbb{E}[\phi(x_0)] - \phi_{\min} + 2\delta G \right] + \frac{160\sqrt{2\pi}dG^2}{B_0},$$

where (a) uses Lemma 2, (b) uses Lemma 7, (c) uses $t < (n_t + 1)q$, $B_1 = q$ and the update rule that $x_j = x_{j-1} + \gamma \mathcal{G}_{\gamma h}(x_j, g_j)$, and (d) uses Eq. (38) and $\gamma = \frac{\delta}{2G(d+c\sqrt{d})}$. Therefore, by taking square root of the inequality above, we obtain the convergence rate (13) as follows.

$$\begin{aligned} \mathbb{E}[\|\mathcal{G}_{\gamma h}(x_{\tilde{T}}, \nabla F_{\delta}(x_{\tilde{T}}))\|] &\leq \sqrt{\mathbb{E}[\|\mathcal{G}_{\gamma h}(x_{\tilde{T}}, \nabla F_{\delta}(x_{\tilde{T}}))\|^2]} \\ &\leq \sqrt{\frac{40G(d + c\sqrt{d})}{T\delta} \left[\mathbb{E}[\phi(x_0)] - \phi_{\min} + 2\delta G \right] + \frac{160\sqrt{2\pi}dG^2}{B_0}} \\ &\leq \frac{\sqrt{40G}(\sqrt{d} + \sqrt{cd^{1/4}})}{\sqrt{T\delta}} \sqrt{\mathbb{E}[\phi(x_0)] - \phi_{\min} + 2\delta G} + \frac{21G\sqrt{d}}{\sqrt{B_0}}, \end{aligned}$$

where the final $\leq$ uses $\sqrt{160\sqrt{2\pi}} < 21$ and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$.

Furthermore, we can select $B_0 = 1764dG^2\epsilon^{-2}$, $B_1 = q = 42\sqrt{d}G\epsilon^{-1}$ and the following T

$$T = 320\delta^{-1}\epsilon^{-2}G(d + c\sqrt{d}) \left[\mathbb{E}[\phi(x_0)] - \phi_{\min} + 2\delta G \right] = \mathcal{O}(Gd\delta^{-1}\epsilon^{-2}). \quad (39)$$

Then substituting the hyperparameters above into the convergence rate (13), we obtain the following bound, which based on item 3 of Proposition 2 implies that there exists at least one $(\gamma, \delta, \epsilon)$ -PGSP in $\{x_t\}_{t=0}^{T-1}$.

$$\min_{0 \leq t \leq T-1} \mathbb{E}[\|\mathcal{G}_{\gamma h}(x_t, \nabla F_{\delta}(x_t))\|] \leq \mathbb{E}[\|\mathcal{G}_{\gamma h}(x_{\tilde{T}}, \nabla F_{\delta}(x_{\tilde{T}}))\|] \leq \epsilon.$$

K PROOF OF THEOREM 4

Denote $\tilde{y}_t \in \mathcal{L}_h[\nabla F_{\delta}(x_t)] \stackrel{\text{def}}{=} \arg \min_{y \in \mathbb{R}^d} [h(y) + \langle y, \nabla F_{\delta}(x_t) \rangle]$. Then the δ -regularized Frank-Wolfe gap (8) can be expressed as follows.

$$\begin{aligned} \mathcal{W}_h[x_t, \nabla F_{\delta}(x_t)] &\stackrel{\text{def}}{=} \max_{y \in \mathbb{R}^d} [h(x_t) - h(y) + \langle y - x_t, -\nabla F_{\delta}(x_t) \rangle] \\ &= h(x_t) - h(\tilde{y}_t) - \langle \tilde{y}_t - x_t, \nabla F_{\delta}(x_t) \rangle. \end{aligned} \quad (40)$$

Then we have

$$\begin{aligned} \langle \nabla F_{\delta}(x_t), y_t - x_t \rangle &= \langle \nabla F_{\delta}(x_t), \tilde{y}_t - x_t \rangle + \langle \nabla F_{\delta}(x_t), y_t - \tilde{y}_t \rangle \\ &= \langle \nabla F_{\delta}(x_t), \tilde{y}_t - x_t \rangle + \langle g_t, y_t - \tilde{y}_t \rangle + \langle \nabla F_{\delta}(x_t) - g_t, y_t - \tilde{y}_t \rangle \\ &\stackrel{(a)}{\leq} \langle \nabla F_{\delta}(x_t), \tilde{y}_t - x_t \rangle + h(\tilde{y}_t) - h(y_t) + \|\nabla F_{\delta}(x_t) - g_t\| \|y_t - \tilde{y}_t\| \\ &\stackrel{(b)}{\leq} -\mathcal{W}_h[x_t, \nabla F_{\delta}(x_t)] + h(x_t) - h(y_t) + 2R\|\nabla F_{\delta}(x_t) - g_t\| \end{aligned} \quad (41)$$

where (a) uses $h(y_t) + \langle y_t, g_t \rangle \leq h(\tilde{y}_t) + \langle \tilde{y}_t, g_t \rangle$ based on the update rule (14), and (b) uses Eq. (40) as well as $\|y_t - x^{(h)}\| \leq R$ and $\|\tilde{y}_t - x^{(h)}\| \leq R$ (based on Proposition 3 and item 2 of Lemma 1).

Then since ∇F_{δ} is $\frac{cG\sqrt{d}}{\delta}$ -Lipschitz continuous based on item 2 of Lemma 1, we obtain that

$$\begin{aligned} &F_{\delta}(x_{t+1}) \\ &\leq F_{\delta}(x_t) + \langle \nabla F_{\delta}(x_t), x_{t+1} - x_t \rangle + \frac{cG\sqrt{d}}{2\delta} \|x_{t+1} - x_t\|^2 \\ &\stackrel{(a)}{=} F_{\delta}(x_t) + \gamma \langle \nabla F_{\delta}(x_t), y_t - x_t \rangle + \frac{cG\sqrt{d}\gamma^2}{2\delta} \|y_t - x_t\|^2 \\ &\stackrel{(b)}{\leq} F_{\delta}(x_t) - \gamma \mathcal{W}_h[x_t, \nabla F_{\delta}(x_t)] + \gamma h(x_t) - \gamma h(y_t) + 2R\gamma \|\nabla F_{\delta}(x_t) - g_t\| + \frac{cG\sqrt{d}(2R)^2\gamma^2}{2\delta} \\ &\stackrel{(c)}{\leq} F_{\delta}(x_t) - \gamma \mathcal{W}_h[x_t, \nabla F_{\delta}(x_t)] + h(x_t) - h(x_{t+1}) + 2R\gamma \|\nabla F_{\delta}(x_t) - g_t\| + \frac{2cG\sqrt{d}R^2\gamma^2}{\delta}, \end{aligned}$$

where (a) uses the update rule (15), (b) uses Eq. (41) and Lemma 8, and (c) uses $h(x_{t+1}) \leq (1 - \gamma)h(x_t) + \gamma h(y_t)$ which holds for convex function h and x_{t+1} obtained from the update rule (15). Rearranging the inequality above and averaging it over $t = 0, 1, \dots, T - 1$, we obtain the convergence rate (16) as follows.

$$\begin{aligned}
& \mathbb{E}[\mathcal{W}_h[x_{\tilde{T}}, \nabla F_\delta(x_{\tilde{T}})]] \\
&= \frac{1}{T} \sum_{t=0}^{T-1} \mathcal{W}_h[x_t, \nabla F_\delta(x_t)] \\
&\leq \frac{1}{T\gamma} \mathbb{E}[F_\delta(x_0) + h(x_0) - F_\delta(x_T) - h(x_T)] + \frac{2R}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla F_\delta(x_t) - g_t\|] + \frac{2cG\sqrt{d}R^2\gamma}{\delta} \\
&\stackrel{(a)}{\leq} \frac{1}{T\gamma} \mathbb{E}[F(x_0) + h(x_0) - F(x_T) - h(x_T) + 2\delta G] + \frac{2R}{T} \sum_{t=0}^{T-1} \sqrt{\mathbb{E}[\|\nabla F_\delta(x_t) - g_t\|^2]} \\
&\quad + \frac{2cG\sqrt{d}R^2\gamma}{\delta} \\
&\stackrel{(b)}{\leq} \frac{1}{T\gamma} \mathbb{E}[\phi(x_0) - \phi_{\min} + 2\delta G] + \frac{2R}{T} \sum_{t=0}^{T-1} \sqrt{\frac{16\sqrt{2\pi}dG^2}{B_t}} + \frac{cG\sqrt{d}R^2\gamma}{2\delta} \\
&\stackrel{(c)}{\leq} R\sqrt{\frac{8cG\sqrt{d}}{T\delta} \mathbb{E}[\phi(x_0) - \phi_{\min} + 2\delta G]} + \frac{21RG\sqrt{d}}{\sqrt{B}},
\end{aligned} \tag{42}$$

where (a) uses item 1 of Proposition 1, (b) uses $\phi \stackrel{\text{def}}{=} F + g$, $\phi_{\min} \stackrel{\text{def}}{=} \min_{x \in \mathbb{R}^d} \phi(x)$ and Lemma 6, (c) uses stepsize $\gamma = \frac{1}{R} \sqrt{\frac{2\delta}{TcG\sqrt{d}}} \mathbb{E}[\phi(x_0) - \phi_{\min} + 2\delta G]$ and constant batchsize $B_t \equiv B$.

Furthermore, we can select the following hyperparameters.

$$T = \frac{32cGR^2\sqrt{d}}{\delta\epsilon^2} [\mathbb{E}[\phi(x_0)] - \phi_{\min} + 2\delta G] = \mathcal{O}(GR^2d^{1/2}\delta^{-1}\epsilon^{-2}), \tag{43}$$

$$B = 1764dR^2G^2\epsilon^{-2} = \mathcal{O}(dR^2G^2\epsilon^{-2}). \tag{44}$$

Then substituting the hyperparameters above into the convergence rate (16), we obtain the following bound, which based on item 3 of Proposition 4 implies that there exists at least one (δ, ϵ) -CGGSP in $\{x_t\}_{t=0}^{T-1}$.

$$\min_{0 \leq t \leq T-1} \mathbb{E}[\mathcal{W}_h(x_t, \nabla F_\delta(x_t))] \leq \mathbb{E}[\mathcal{W}_h(x_{\tilde{T}}, \nabla F_\delta(x_{\tilde{T}}))] \leq \epsilon.$$

L PROOF OF THEOREM 5

We can prove the convergence rate (17) as follows.

$$\begin{aligned}
& \mathbb{E}[\mathcal{W}_h[x_{\tilde{T}}, \nabla F_\delta(x_{\tilde{T}})]] \\
&\stackrel{(a)}{\leq} \frac{1}{T\gamma} \mathbb{E}[F(x_0) + h(x_0) - F(x_T) - h(x_T) + 2\delta G] + \frac{2R}{T} \sum_{t=0}^{T-1} \sqrt{\mathbb{E}[\|\nabla F_\delta(x_t) - g_t\|^2]} \\
&\quad + \frac{2cG\sqrt{d}R^2\gamma}{\delta} \\
&\stackrel{(b)}{\leq} \frac{1}{T\gamma} \mathbb{E}[\phi(x_0) - \phi_{\min} + 2\delta G] + \frac{2R}{T} \sum_{t=0}^{T-1} \sqrt{\frac{d^2G^2}{\delta^2B_1} \sum_{j=n_tq+1}^t \|x_j - x_{j-1}\|^2 + \frac{16\sqrt{2\pi}dG^2}{B_0}} \\
&\quad + \frac{2cG\sqrt{d}R^2\gamma}{\delta} \\
&\stackrel{(c)}{\leq} \frac{1}{T\gamma} \mathbb{E}[\phi(x_0) - \phi_{\min} + 2\delta G] + \frac{2R}{T} \sum_{t=0}^{T-1} \sqrt{\frac{d^2G^2}{\delta^2} (2R\gamma)^2 + \frac{16\sqrt{2\pi}dG^2}{B_0}} + \frac{2cG\sqrt{d}R^2\gamma}{\delta}
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{1}{T\gamma} \mathbb{E}[\phi(x_0) - \phi_{\min} + 2\delta G] + \frac{GR^2\gamma}{\delta}(4d + 2c\sqrt{d}) + \frac{13RG\sqrt{d}}{\sqrt{B_0}} \\
&\stackrel{(d)}{=} 2R\sqrt{\frac{G(4d + 2c\sqrt{d})}{T\delta}} \mathbb{E}[\phi(x_0) - \phi_{\min} + 2\delta G] + \frac{13RG\sqrt{d}}{\sqrt{B_0}},
\end{aligned}$$

where (a) uses Eq. (42) (we can see it still holds by following its proof in Appendix K), (b) uses $\phi \stackrel{\text{def}}{=} F + g$, $\phi_{\min} \stackrel{\text{def}}{=} \min_{x \in \mathbb{R}^d} \phi(x)$ and Lemma 7, (c) uses $t < (n_t + 1)q$, $B_1 = q$, $\|x_j - x_{j-1}\| = \gamma\|y_{j-1} - x_{j-1}\| \leq 2R\gamma$ (based on Eq. (15) and Lemma 8), and (d) uses stepsize $\gamma = \frac{1}{R} \sqrt{\frac{\delta \mathbb{E}[\phi(x_0) - \phi_{\min} + 2\delta G]}{TG(4d + 2c\sqrt{d})}}$.

Furthermore, we can select the following hyperparameters.

$$T = \frac{16GR^2(4d + 2c\sqrt{d})}{\delta\epsilon^2} [\mathbb{E}[\phi(x_0)] - \phi_{\min} + 2\delta G] = \mathcal{O}(GR^2d\delta^{-1}\epsilon^{-2}), \quad (45)$$

$$B_0 = 676dR^2G^2\epsilon^{-2}, \quad (46)$$

$$B_1 = q = 26RG\epsilon^{-1}\sqrt{d}. \quad (47)$$

Then substituting the hyperparameters above into the convergence rate (16), we obtain the following bound, which based on item 3 of Proposition 4 implies that there exists at least one (δ, ϵ) -CGGSP in $\{x_t\}_{t=0}^{T-1}$.

$$\min_{0 \leq t \leq T-1} \mathbb{E}[\mathcal{W}_h(x_t, \nabla F_\delta(x_t))] \leq \mathbb{E}[\mathcal{W}_h(x_{\tilde{T}}, \nabla F_\delta(x_{\tilde{T}}))] \leq \epsilon.$$

M USE OF LARGE LANGUAGE MODELS (LLMs)

We used LLMs to generate some functions in the experimental code, and then checked and edited the code to ensure that it exactly implements the algorithms.