Zeroth-Order Methods for Constrained Nonconvex Nonsmooth Stochastic Optimization

Zhuanghua Liu [1 2] **Cheng Chen** [3] **Luo Luo** [4 5] **Bryan Kian Hsiang Low** [1]

Abstract

This paper studies the problem of solving nonconvex nonsmooth optimization over a closed convex set. Most previous works tackle such problems by transforming the constrained problem into an unconstrained problem. However, they only provide asymptotic convergence analysis for their methods. In this work, we provide the non-asymptotic convergence analysis for solving constrained nonconvex nonsmooth optimization. We first generalize classical gradient mapping and the Frank–Wolfe gap in the nonsmooth setting. Then we introduce novel notions of approximate stationarity concerning such generalized quantities. We also propose several stochastic zeroth-order algorithms for the problem, along with their non-asymptotic convergence guarantees of obtaining the proposed approximate stationarity. Finally, we conduct numerical experiments that demonstrate the effectiveness of our algorithms.

1. Introduction

This paper considers the following constrained stochastic optimization problem

$$\min_{x\in\Omega} F(x) \coloneqq \mathbb{E}_{\xi\sim\mathcal{P}}\left[f(x;\xi)\right] \tag{1}$$

where the stochastic component $f(x;\xi)$, indexed by some random variable ξ , is probably nonconvex and nonsmooth, and the feasible set $\Omega \subseteq \mathbb{R}^d$ is convex and compact. Such problems are very common in many real-world machine learning applications including adversarial attack (Carlini & Wagner, 2017; Madry et al., 2017), regularized support vector machine (Smola & Schölkopf, 1998; Zhang, 2010) and training Generative Adversarial Networks (GANs) (Gulrajani et al., 2017; Miyato et al., 2018).

Existing research (Curtis & Overton, 2012; Tang et al., 2014; Hare et al., 2016; Curtis et al., 2017; Hoseini Monjezi & Nobakhtian, 2021) for constrained nonconvex nonsmooth optimization mainly focuses on transforming the constrained problem into an unconstrained problem that can be solved with techniques developed in the unconstrained setting. Though their methods enjoy asymptotic convergence, little is known about the non-asymptotic convergence rates of these algorithms. One of the difficulties is the choice of convergence criteria that measure the progress of the algorithm. The gradient mapping and Frank–Wolfe gap, which are widely used as the convergence criteria in the constrained smooth problem, are unfortunately not suitable for the nonsmooth setting because their definitions require the gradient to be well-defined at every point in the feasible set. One possible solution is to generalize such quantities with the Clarke subdifferential, which considers the set of generalized gradients at the current iterate. Nevertheless, we show that the approximate stationarity concerning such quantities does not permit a finite-time analysis for any algorithm. Inspired by the definition of (δ,ϵ)-Goldstein stationary points for unconstrained nonconvex nonsmooth problems (Zhang et al., 2020; Lin et al., 2022; Chen et al., 2023; Cutkosky et al., 2023; Kornowski & Shamir, 2023), we propose the generalized gradient mapping and δ-Frank–Wolfe gap as the extensions of the gradient mapping and the Frank–Wolfe gap by leveraging the Goldstein δ-subdifferential (Goldstein, 1977), which considers the convex combination of all generalized gradients in the neighborhood of current iterate. Furthermore, we define the (γ,δ,ϵ)-generalized Goldstein stationary point and the (δ,ϵ)-Goldstein Frank–Wolfe stationary point as the approximate stationarity for our problem.

Armed with the refined approximate stationarity, we propose zeroth-order projection-based and projection-free stochastic optimization algorithms for solving the problem (1) in finite time. Specifically, we rigorously show that the zeroth-order

[1]Department of Computer Science, National University of Singapore, Singapore, Singapore [2]CNRS@CREATE LTD, 1 Create Way, #08-01 CREATE Tower, Singapore 138602 [3]Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai, China [4]School of Data Science, Fudan University, Shanghai, China [5]Shanghai Key Laboratory for Contemporary Applied Mathematics. Correspondence to: Luo Luo <luoluo@fudan.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Table 1. We present the non-asymptotic convergence rates of proposed algorithms for constrained nonconvex nonsmooth problems. FQO stands for the function query oracle calls. GGSP stands for generalized Goldstein stationary point, and GFWSP stands for Goldstein Frank–Wolfe stationary point.

METHODS	CRITERION	FQO	REFERENCE
MB-ZOSPGD	$(\gamma, \delta, \epsilon)$-GGSP	$\mathcal{O}(d^{\frac{3}{2}}\delta^{-1}\epsilon^{-4})$	COROLLARY 5.2
VR-ZOSPGD	$(\gamma, \delta, \epsilon)$-GGSP	$\mathcal{O}(d^{\frac{3}{2}}\delta^{-1}\epsilon^{-3})$	COROLLARY 5.4
MB-ZOSFW	(δ, ϵ)-GFWSP	$\mathcal{O}(d^{\frac{3}{2}}\delta^{-1}\epsilon^{-4})$	COROLLARY 5.7
VR-ZOSFW	(δ, ϵ)-GFWSP	$\mathcal{O}(d^{\frac{3}{2}}\delta^{-1}\epsilon^{-3})$	COROLLARY 5.9

stochastic projected gradient descent algorithm with a minibatch gradient estimator obtains the $(\gamma, \delta, \epsilon)$-generalized Goldstein stationary point through $\mathcal{O}(d^{\frac{3}{2}}\delta^{-1}\epsilon^{-4})$ function query oracle calls. To tackle the case where the feasible set is so complicated that projection onto it is rather expensive or even intractable, we also propose a zeroth-order stochastic Frank–Wolfe algorithm with a minibatch gradient estimator for problem (1) which attains the (δ, ϵ)-Goldstein Frank–Wolfe stationary point through $\mathcal{O}(d^{\frac{3}{2}}\delta^{-1}\epsilon^{-4})$ function query oracle calls. Furthermore, the convergence rate of both algorithms can be improved through the use of variance-reduction techniques. The complexity results of the algorithms are summarized in Table 1. Finally, we perform numerical experiments to validate the effectiveness of the proposed approaches.

Paper Organization In Section 2, we present a literature review on the convergence analysis for minimizing unconstrained and constrained nonconvex nonsmooth problems. In Section 3, we formalize the notations and assumptions of our problem and introduce the background for nonsmooth analysis. In Section 4, we propose the approximate stationarity for our problem and provide some fundamental properties of these notions. In Section 5, we present the zeroth-order stochastic projection-based and projection-free algorithms for solving our problem in finite time. In Section 6, we conduct numerical experiments to demonstrate the effectiveness of the proposed algorithms. We conclude this work in Section 7.

2. Related Works

In this section, we review prior works on nonconvex nonsmooth optimization.

2.1. Non-Asymptotic Convergence Analysis of Nonconvex Nonsmooth Functions

The development of non-asymptotic convergence analysis of nonsmooth optimization only emerged recently. Some recent works (Davis & Grimmer, 2019; Davis & Drusvyatskiy, 2019) showed that a (δ, ϵ)-near approximate stationary (NAS) point is obtainable through $\mathcal{O}(\rho^4\delta^{-4} + \epsilon^{-4})$ oracles calls for ρ-weakly convex functions. However, a large class of nonconvex nonsmooth functions does not belong to the class of ρ-weakly convex functions, including deep neural networks with RELU activations. Even worse, Tian & So (2021) showed that it is impossible to find NAS points for ρ-weakly convex functions when ρ is unbounded using dimension-free algorithms in finite time. Zhang et al. (2020) gave the first dimension independent non-asymptotic complexity analysis to compute the (δ, ϵ)-Goldstein stationary point. Their methods were improved by introducing perturbations to remove the unrealistic subgradient oracle (Davis et al., 2022; Tian et al., 2022). Furthermore, Tian & So (2022); Jordan et al. (2023) showed that randomization is necessary to obtain the dimension-free complexity for nonconvex nonsmooth optimization. Cutkosky et al. (2023) proposed the optimal algorithm via the reduction from nonconvex nonsmooth optimization to online learning.

The first non-asymptotic convergence analysis of zeroth-order methods for nonconvex nonsmooth functions was introduced by Nesterov & Spokoiny (2017). Lin et al. (2022) proposed the gradient-free approaches for finding a (δ, ϵ)-Goldstein stationary point of the problem and Chen et al. (2023) improved their results by leveraging the variance-reduction technique. Kornowski & Shamir (2023) applied the reduction technique introduced by Cutkosky et al. (2023) to the gradient-free setting for achieving a shaper bound.

2.2. Convergence Analysis of Constrained Nonconvex Nonsmooth Functions

We divide the existing literature for constrained nonconvex nonsmooth optimization into two categories.

Asymptotic Analysis Most previous literature analyzes the asymptotic convergence properties of various optimization algorithms including bundle methods and gradient sampling methods. A nonsmooth problem over a closed convex set is usually reformulated as an unconstrained nonsmooth

problem through the penalty or filter method (Hare et al., 2016; Dao et al., 2016), which can be solved using the bundle method. Besides this approach, Curtis & Overton (2012) formulated the inequality-constrained nonconvex nonsmooth problem as a sequential quadratic programming (SQP) problem, which was solved by the gradient sampling (GS) method. The convergence result showed that accumulation points were stationary points of the reformulated function but could possibly be infeasible. Their result was improved by Tang et al. (2014) that generated a sequence of feasible iterates using the SQP-GS methodology. A more efficient BFGS-SQP was proposed by Curtis et al. (2017) which showed faster convergence behavior without requiring the existence of the Hessian. Xu et al. (2015) considered a smoothing augmented lagrangian method for solving the inequality-constrained problem.

Non-Asymptotic Analysis The first non-asymptotic convergence analysis of the constrained nonconvex nonsmooth optimization was proposed by Davis & Grimmer (2019) for ρ-weakly convex functions. Vladarean et al. (2023) proposed a Frank-Wolfe algorithm for constrained nonconvex nonsmooth stochastic compositional optimization problems. They considered the composite functions where the outer function is convex but possibly non-differentiable and the inner function is smooth. Very recently, Grimmer & Jia (2023) proposed the non-asymptotic convergence analysis of minimizing the inequality-constrained problem using the subgradient method. We point out several differences between the approximate stationarity proposed in their work and ours below. Our problem can be reformulated in their setting by the introduction of some Lipschitz functions $g_1, \ldots, g_m$. However, the proposed (δ, ϵ, η)-Goldstein KKT (GKKT) stationary point by Grimmer & Jia (2023) requires fulfillment of a certain constraint qualification while the (δ, ϵ, η)-Goldstein Fritz-John (GKJ) stationary point does not. In addition, the corresponding GKKT stationary point and GFJ stationary point depend on how to describe the constrained set Ω by the specific choice of Lipschitz functions $g_1, \ldots, g_m$ which may not be unique, while the definitions of our GGSP and GFWSP mainly depend on the constrained set Ω.

3. Preliminaries

In this section, we first present the notations and assumptions used in the paper, then introduce the background of nonsmooth analysis, and finally review the randomized smoothing technique that is widely used in zeroth-order optimization.

3.1. Problem Assumptions

In this paper, we assume the problem (1) satisfies the following two assumptions.

Assumption 3.1. The feasible set $\Omega \in \mathbb{R}^d$ of the problem (1) is convex and compact with diameter bounded by B.

Assumption 3.2. The stochastic component $f(\cdot, \xi)$ of the problem (1) is $L(\xi)$-Lipschitz for a given ξ, i.e., it holds that

$$|f(x, \xi) - f(y, \xi)| \leq L(\xi) \|x - y\|,$$

for any $x, y \in \mathbb{R}^d$, where $L(\xi)$ has bounded second-order moment such that $\mathbb{E}_\xi[L(\xi)^2] \leq G^2$ for some $G > 0$.

Remark 3.3. Assumption 3.2 implies the objective function $F(\cdot)$ is G-Lipschitz by Jensen's inequality.

In this paper, we study zeroth-order stochastic optimization algorithms that can access one or both of the following two oracles:

- Function Query Oracle (FQO): Given a point $x \in \Omega$ and ξ sampled from the distribution $\mathcal{P}$, FQO returns the value of $f(x, \xi)$.
- Linear Maximization Oracle (LMO): Given a vector $v \in \mathbb{R}^d$, LMO returns a solution of the linear optimization problem: $\arg\max_{u \in \Omega} \langle u, v \rangle$.

3.2. Background for Nonsmooth Analysis

We can define generalized directional derivatives and generalized gradients for general nondifferentiable functions as follows.

Definition 3.4 (Clarke (1990))**.** Given a point $x \in \mathbb{R}^d$ and a direction $v \in \mathbb{R}^d$, the generalized directional derivative of a nondifferentiable function f is defined as $Df(x; v) := \limsup_{y \to x, t \downarrow 0} \frac{f(y+tv)-f(y)}{t}$. Then the Clarke subdifferential of f is defined as the set $\partial f(x) := \{g \in \mathbb{R}^d \colon g^\top v \leq Df(x; v), \forall v \in \mathbb{R}^d\}$. Each element $g \in \partial f(x)$ is called a generalized gradient of f.

Given the definition of the Clarke subdifferential and generalized gradients, the convergence criterion of solving the unconstrained nonconvex nonsmooth problem $\min_{x \in \mathbb{R}^d} f(x)$ can be characterized as finding the ϵ-Clarke stationary point x of the function f, specifically,

$$\min\{\|g\| : g \in \partial f(x)\} \leq \epsilon.$$

Unfortunately, Zhang et al. (2020) showed that no algorithm can find an ϵ-Clarke stationary point of the unconstrained problem in finite time. Thus they considered a refined notion of approximate stationarity concerning the Goldstein δ-subdifferential.

Definition 3.5 (Goldstein (1977))**.** Denote $\mathbb{B}(x, \delta) = \{y : \|y - x\| \leq \delta\}$. Given a Lipschitz function $f \colon \mathbb{R}^d \to \mathbb{R}$, a point $x \in \mathbb{R}^d$ and $\delta \geq 0$, the Goldstein δ-subdifferential of f at x is defined as

$$\partial_\delta f(x) := \text{conv}\left(\cup_{y \in \mathbb{B}(x, \delta)} \partial f(y)\right),$$

which is the convex combination of the generalized gradients at points in the δ-neighbourhood of x.

We have the following relationship between Goldstein δ-subdifferential and Clarke subdifferential.

Lemma 3.6 (Makela & Neittaanmaki (1992)). *The Goldstein δ-subdifferential is equivalent to the Clarke subdifferential when $\delta = 0$, i.e., $\partial_0 f(x) = \partial f(x)$.*

Accordingly, a refined approximate stationarity for the unconstrained nonconvex nonsmooth problem is defined below.

Definition 3.7 (Zhang et al. (2020)). Given a Lipschitz function $f: \mathbb{R}^d \to \mathbb{R}$, a point $x \in \mathbb{R}^d$ and $\delta \geq 0$, a point x is called a (δ, ϵ)-Goldstein stationary point of $f(\cdot)$ if

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

Zhang et al. (2020) and its follow-up works (Tian et al., 2022; Davis et al., 2022; Cutkosky et al., 2023) proposed a series of algorithms with non-asymptotic convergence analysis of finding a (δ, ϵ)-Goldstein stationary point for $\delta > 0$.

3.3. Randomized Smoothing

The randomized smoothing technique is widely used in nonsmooth analysis (Duchi et al., 2012) and zeroth-order optimization (Nesterov & Spokoiny, 2017). Formally, given a L-Lipschitz function f and a distribution $\mathcal{Q}$, we define the smoothing function as $f_\delta(x) = \mathbb{E}_{u\sim\mathcal{Q}}[f(x + \delta u)]$, which enjoys the following properties.

Lemma 3.8 (Lin et al. (2022)). *Let $f_\delta(x) = \mathbb{E}_{u\sim\mathcal{Q}}[f(x + \delta u)]$ where $\mathcal{Q}$ is a uniform distribution on a unit ball in ℓ_2-norm. Suppose the function f is L-Lipschitz, then we have (a) $|f_\delta(x) - f(x)| \leq \delta L$; (b) f_δ is differentiable everywhere and L-Lipschitz with $(cL\sqrt{d}/\delta)$-Lipschitz gradient where $c > 0$ is a constant; (c) $\nabla f_\delta(x) \in \partial_\delta f(x)$ for all $x \in \mathbb{R}^d$.*

An unbiased estimation of $\nabla f_\delta(x)$ can be obtained by making two function query oracle calls on points randomly sampled from the unit sphere, which induces the zeroth-order gradient estimator (Agarwal et al., 2010).

Definition 3.9. Given a stochastic function component $f(\cdot;\xi): \mathbb{R}^d \to \mathbb{R}$, we denote its zeroth-order stochastic gradient estimator at $x \in \mathbb{R}^d$ by

$$\hat{g}(x; w, \xi) = \frac{d}{2\delta}(f(x + \delta w; \xi) - f(x - \delta w; \xi))w,$$

where w is sampled from a uniform distribution on a unit sphere in $\mathbb{R}^d$.

4. The Approximate Stationarity for Constrained Nonsmooth Optimization

In this section, we first formally define our notions of approximate stationarity for constrained nonconvex nonsmooth optimization and then present several properties of our definitions. These notions can help us achieve the non-asymptotic convergence rate of stochastic optimization algorithms to be introduced in the later sections.

4.1. Definitions of Approximate Stationarity

We first introduce the notion of generalized gradient mapping as follows.

Definition 4.1. Given some $\delta \geq 0$, $\gamma > 0$ and a convex compact set $\Omega \subseteq \mathbb{R}^d$, the generalized gradient mapping of a Lipschitz function $f: \mathbb{R}^d \to \mathbb{R}$ at a point $x \in \Omega$ associated with some $g \in \partial_\delta f(x)$ is defined as:

$$\mathcal{G}(x, g, \gamma) \coloneqq \frac{1}{\gamma}(x - \psi(x, g, \gamma)),$$

where

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

If the function $f(\cdot)$ is differentiable, the generalized gradient mapping with $\delta = 0$ is equivalent to the vanilla gradient mapping (Nesterov et al., 2018, Section 2.2.4)

$$\frac{1}{\gamma}(x - \psi(x, \nabla f(x), \gamma)),$$

which is a popular criterion for measuring the convergence rate of stochastic projection-based algorithms in the smooth setting (Nemirovskij & Yudin, 1983; Ghadimi et al., 2016; Nesterov et al., 2018; Wang et al., 2019). This property can be proved by the fact that $\partial f(x) = \{\nabla f(x)\}$ when f is a differentiable function (Makela & Neittaanmaki, 1992, Theorem 3.1.7) and Lemma 3.6.

We define the following approximate stationary point w.r.t. the generalized gradient mapping.

Definition 4.2. We say the point $x \in \Omega$ is a $(\gamma, \delta, \epsilon)$-generalized Goldstein stationary point (GGSP) of the problem (1) if it satisfies

$$\min_{g\in\partial_\delta f(x)} \|\mathcal{G}(x, g, \gamma)\| \leq \epsilon.$$

By setting $\delta = 0$, we denote the point $x \in \Omega$ is a (γ, ϵ)-generalized Clarke stationary point (GCSP) if it satisfies

$$\min_{g\in\partial f(x)} \|\mathcal{G}(x, g, \gamma)\| \leq \epsilon,$$

where $\partial f(\cdot) = \partial_0 f(\cdot)$ is the Clarke subdifferential.

Although the (γ, ϵ)-GCSP is a more natural generalization of the approximate stationarity w.r.t. the gradient mapping, we will show that it does not admit non-asymptotic convergence for our problem. Instead, we use the $(\gamma, \delta, \epsilon)$-GGSP as the approximate stationarity to analyze the finite-time convergence rate of the stochastic projection-based algorithms for any $\delta > 0$. We remark that when $\Omega = \mathbb{R}^d$, the $(\gamma, \delta, \epsilon)$-GGSP is reduced to the (δ, ϵ)-Goldstein stationary point for the unconstrained nonconvex nonsmooth optimization. To analyze the projection-free stochastic optimization algorithms, we rely on the following notion of the δ-Frank–Wolfe gap.

Definition 4.3. Given some $\delta \geq 0$ and a convex compact set $\Omega \subseteq \mathbb{R}^d$, the δ-Frank–Wolfe gap of a Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ at $x \in \Omega$ is defined as:

$$\min_{g \in \partial_\delta f(x)} \max_{u \in \Omega} \langle u - x, -g \rangle, \tag{2}$$

where $\partial_\delta f(\cdot)$ is the Goldstein δ-subdifferential.

For a differentiable function $f(\cdot)$, the δ-Frank–Wolfe gap when $\delta = 0$ is equivalent to the vanilla Frank–Wolfe gap

$$\max_{u \in \Omega} \langle u - x, -\nabla f(x) \rangle,$$

which is a common criterion for measuring the convergence rate of stochastic projection-free algorithms in the smooth setting (Lacoste-Julien, 2016; Reddi et al., 2016; Yurtsever et al., 2019; Gao & Huang, 2020). It is natural to define the approximate stationary point w.r.t. the δ-Frank–Wolfe gap as follows.

Definition 4.4. We say the point $x \in \Omega$ is a (δ, ϵ)-Goldstein Frank–Wolfe stationary point (GFWSP) of the problem (1) if it satisfies

$$\min_{g \in \partial_\delta f(x)} \max_{u \in \Omega} \langle u - x, -g \rangle \leq \epsilon.$$

where $\partial_\delta f(\cdot)$ is the Goldstein δ-subdifferential. By setting $\delta = 0$, we denote the point $x \in \Omega$ is an ϵ-Clarke Frank–Wolfe stationary point (CFWSP) if it satisfies

$$\min_{g \in \partial f(x)} \max_{u \in \Omega} \langle u - x, -g \rangle \leq \epsilon.$$

where $\partial f(\cdot) = \partial_0 f(\cdot)$ is the Clarke subdifferential.

Similar to the (γ, ϵ)-GCSP, we will show that ϵ-CFWSP does not admit non-asymptotic convergence for solving our problem. Alternatively, we will use the (δ, ϵ)-GFWSP as the approximate stationarity to obtain the finite-time convergence rate of stochastic projection-free algorithms for any $\delta > 0$.

4.2. Properties of Proposed Approximate Stationary Points

Apparently, $(\gamma, \delta, \epsilon)$-GGSP appears to be a weaker notion since if x is a (γ, ϵ)-GCSP, then it is also a $(\gamma, \delta, \epsilon)$-GGSP for any $\delta \geq 0$, but not vice versa. We show that the converse implication indeed still holds, assuming that $f(\cdot)$ is a differentiable function.

Proposition 4.5. *Given a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ and a convex compact set Ω, the following statements hold:*

- *Suppose a point $x \in \Omega$ is a (γ, ϵ)-generalized Clarke stationary point, then it is a $(\gamma, \delta, \epsilon)$-generalized Goldstein stationary point for any $\delta \geq 0$.*
- *Suppose $f(\cdot)$ has L-Lipschitz gradient and the point $x \in \Omega$ is a $(\gamma, \epsilon/(2L), \epsilon/2)$-generalized Goldstein stationary point, then the corresponding vanilla gradient mapping at x satisfies $\|\mathcal{G}(x, \nabla f(x), \gamma)\| \leq \epsilon$.*

We can infer the equivalence between the (γ, ϵ)-GCSP and the $(\gamma, \delta, \epsilon)$-GGSP when $f(\cdot)$ is a differentiable function. Similar to Proposition 4.5, we can show the connection between ϵ-CFWSP and (δ, ϵ)-GFWSP as follows.

Proposition 4.6. *Given a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ and a convex compact set Ω, we have the following statements:*

- *Suppose a point $x \in \Omega$ is an ϵ-Clarke Frank–Wolfe stationary point, then it is a (δ, ϵ)-Goldstein Frank–Wolfe stationary point for any $\delta \geq 0$.*
- *Suppose $f(\cdot)$ has L-Lipschitz gradient and the point $x \in \Omega$ is a $\big(\epsilon/(3BL), 2\epsilon/3\big)$-Goldstein Frank–Wolfe stationary point, then the corresponding vanilla Frank–Wolfe Gap at x satisfies $\max_{u \in \Omega} \langle u - x, -\nabla f(x) \rangle \leq \epsilon$.*

Consequently, the ϵ-CFWSP is equivalent to the (δ, ϵ)-GFWSP assuming $f(\cdot)$ is a differentiable function. However, neither (γ, ϵ)-GCSP nor ϵ-CFWSP permits a finite-time analysis for any deterministic or randomized algorithm interacting with a local oracle[1] in the nonsmooth setting.

Theorem 4.7. *For any algorithm $\mathcal{A}$ interacting with a local oracle, and any $T \in \mathbb{N}$, $d \geq 2$, there is a function $f(\cdot)$ on $\mathbb{R}^d$ such that*

1. *$f(\cdot)$ is $\frac{15}{4}$-Lipschitz, $f(0) - \inf_x f(x) \leq 2$,*

[1] We consider oracles that given a function f and a point x, return some quantity $\mathbb{O}_f(x)$ which conveys local information about the function near that point (Kornowski & Shamir, 2022). A typical example is the first-order oracle $(f(x), \partial f(x))$.

2. *With probability at least* $1 - 2T\exp(-d/36)$ *over the algorithm's randomness, the iterates* $x_1, \ldots, x_T$ *produced by the algorithm do not belong to the set of* (γ, ϵ)*-GCSP or* ϵ*-CFWSP for* $\epsilon < 1/(4\sqrt{2})$.

On the contrary, we will show that both $(\gamma, \delta, \epsilon)$-GGSP and (δ, ϵ)-GFWSP can help us achieve finite-time convergence in the next section.

5. Zeroth-Order Stochastic Methods

In this section, we study stochastic zeroth-order methods for the nonconvex nonsmooth optimization. We first introduce two zeroth-order gradient estimators used in our algorithms. Then we propose several stochastic zeroth-order algorithms which non-asymptotically converge to the approximate stationary points defined in Section 4.

5.1. Zeroth-Order Gradient Estimators

We borrow the idea of the classical two-point gradient estimator and propose two stochastic zeroth-order gradient estimators. The first estimator approximates the gradient by the mean of estimated gradients of a small batch of samples:

$$v_t = \frac{1}{b}\sum_{i=1}^{b} g_{i,t},$$

where $g_{i,t} = \hat{g}(x_t; w_{i,t}, \xi_{i,t})$ is one single stochastic gradient estimator at iteration t. The complete procedure of the minibatch stochastic gradient estimator (MB-SGrad) is shown in Algorithm 1. For the second gradient estimator, we leverage the idea of variance reduction (VR) to approximate $\nabla F_\delta(x_t)$ by a recursive gradient estimator v_t with the following update

$$v_t = \frac{1}{b_2}\sum_{i=1}^{b_2}(g_{i,t} - g_{i,t-1}) + v_{t-1},$$

where $g_{i,t-1}$ and $g_{i,t}$ are the stochastic gradient estimators at two consecutive iterations. The complete procedure of the variance-reduced stochastic gradient estimator (VR-SGrad) is presented in Algorithm 2.

Both algorithms employ the randomized smoothing technique to approximate the gradient of the smoothing function F_δ. The obtained approximations are verified to belong to the Goldstein δ-subdifferential of $F(\cdot)$ by Lemma 3.8.

Algorithm 1 $v_t = \text{MB-SGrad}(x_t)$

Input: Parameter b_t.
Sample $\xi_{1,t}, \ldots, \xi_{b_t,t} \sim \mathcal{P}$ independently.
Sample $w_{1,t}, \ldots, w_{b_t,t}$ independently and uniformly from a unit sphere in $\mathbb{R}^d$.
Let $g_{i,t} = \hat{g}(x_t; w_{i,t}, \xi_{i,t})$ for each $i \in [b_t]$.
return $v_t = \frac{1}{b_t}\sum_{i=1}^{b_t} g_{i,t}$.

Algorithm 2 $v_t = \text{VR-SGrad}(x_t, x_{t-1}, t)$

Input: Parameters b_1, b_2, q.
if $\mod(t, q) = 0$ **then**
  Sample $\xi_{1,t}, \ldots, \xi_{b_1,t} \sim \mathcal{P}$ independently.
  Sample $w_{1,t}, \ldots, w_{b_1,t}$ independently and uniformly from a unit sphere in $\mathbb{R}^d$.
  Let $g_{i,t} = \hat{g}(x_t; w_{i,t}, \xi_{i,t})$ for each $i \in [b_1]$.
  return $v_t = \frac{1}{b_1}\sum_{i=1}^{b_1} g_{i,t}$.
else
  Sample $\xi_{1,t}, \ldots, \xi_{b_2,t} \sim \mathcal{P}$ independently.
  Sample $w_{1,t}, \ldots, w_{b_2,t}$ independently and uniformly from a unit sphere in $\mathbb{R}^d$.
  Let $g_{i,t} = \hat{g}(x_t; w_{i,t}, \xi_{i,t})$ for each $i \in [b_2]$.
  Let $g_{i,t-1} = \hat{g}(x_{t-1}; w_{i,t}, \xi_{i,t})$ for each $i \in [b_2]$.
  return $v_t = \frac{1}{b_2}\sum_{i=1}^{b_2}(g_{i,t} - g_{i,t-1}) + v_{t-1}$.
end if

Algorithm 3 ZOSPGD Method

Input: Initial points $x_0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, smoothing parameter δ and iteration number $T \geq 1$.
for $t = 0, 1, \ldots, T-1$ **do**
  Option I: Set $v_t = \text{MB-SGrad}(x_t)$.
  Option II: Set $v_t = \text{VR-SGrad}(x_t, x_{t-1}, t)$.
  Set $x_{t+1} = \arg\min_{u\in\Omega}\{\langle v_t, u\rangle + \frac{1}{2\gamma}\|u - x_t\|^2\}$.
end for
return x_R where R is uniformly sampled from the set $\{1, 2, \ldots, T\}$.

5.2. Zeroth-Order Stochastic Projection-based Algorithms

We present the details of the zeroth-order stochastic projected gradient descent (ZOSPGD) algorithms for our problem in Algorithm 3. The algorithm leverages the classic projected gradient descent algorithm except for the gradient computation step. Since the gradient oracle is unavailable in our setting, our algorithm estimates the gradient by the two gradient estimators introduced in Section 5.1.

The following theorem shows the convergence rate of solving the problem (1) by Algorithm 3 with Option I (which we call the MB-ZOSPGD algorithm).

Theorem 5.1. *Running the MB-ZOSPGD algorithm (Algorithm 3 with Option I) with* $\gamma = \frac{\delta}{cG\sqrt{d}}$ *and the subroutine MB-SGrad (Algorithm 1) with* $b_t = b$ *where* b *is some constant, then the output* x_R *holds that*

$$\mathbb{E}[\|\mathcal{G}(x_R, \nabla_\delta F(x_R), \gamma)\|] = \mathcal{O}\left(\frac{G\sqrt{B}d^{\frac{1}{4}}}{\sqrt{T\delta}} + \frac{\sqrt{d}G}{\sqrt{b}}\right).$$

Theorem 5.1 and Lemma 3.8(c) imply the following oracle complexity of Algorithm 3 with Option I.

Corollary 5.2. *The MB-ZOSPGD algorithm (Algorithm 3 with Option I) requires at most $\mathcal{O}(G^4Bd^{\frac{3}{2}}\delta^{-1}\epsilon^{-4})$ FQO calls to obtain a (γ,δ,ϵ)-GGSP.*

The VR technique has been shown to improve the convergence rate of the stochastic projected gradient descent algorithm on the constrained nonconvex smooth problem in the existing literature. It is interesting to see whether this technique can improve the convergence rate of our algorithm. Accordingly, we provide the following convergence analysis of solving the problem (1) by Algorithm 3 with Option II (which we call the VR-ZOSPGD algorithm).

Theorem 5.3. *Running the VR-ZOSPGD algorithm (Algorithm 3 with Option II) with $\gamma = \frac{\delta}{2dG}$ and the subroutine VR-SGrad (Algorithm 2) with $b_2 = q$, then the output x_R holds that*

$$\mathbb{E}\left[\|\mathcal{G}(x_R,\nabla_\delta F(x_R),\gamma)\|\right] = \mathcal{O}\left(\frac{\sqrt{dB}G}{\sqrt{\delta T}} + \frac{\sqrt{d}G}{\sqrt{b_1}}\right).$$

Theorem 5.3 and Lemma 3.8(c) imply the following oracle complexity of Algorithm 3 with Option II.

Corollary 5.4. *The VR-ZOSPGD algorithm (Algorithm 3 with Option II) requires at most $\mathcal{O}(G^3Bd^{\frac{3}{2}}\delta^{-1}\epsilon^{-3})$ FQO calls to obtain a (γ,δ,ϵ)-GGSP.*

We remark on the connection between our result and existing work on unconstrained nonconvex nonsmooth optimization.

Remark 5.5. When the feasible set $\Omega = \mathbb{R}^d$, Lin et al. (2022) proved that $\mathcal{O}(d^{\frac{3}{2}}\delta^{-1}\epsilon^{-4})$ FQO calls suffice to obtain a (δ,ϵ)-Goldstein stationary point, which matches the complexity of our result in Corollary 5.2. Chen et al. (2023) improved their complexity to $\mathcal{O}(d^{\frac{3}{2}}\delta^{-1}\epsilon^{-3})$ with the VR technique, which matches our result in Corollary 5.4.

5.3. Zeroth-Order Stochastic Projection-Free Algorithms

In this subsection, we focus on the cases where projection onto the feasible sets could be rather expensive. For example, projection onto the nuclear norm constraints is an essential step of the matrix completion problem, and this step requires computing the full singular value decomposition, which takes $\mathcal{O}(d^3)$ time. The Frank–Wolfe method (Frank & Wolfe, 1956; Jaggi, 2013) has recently become popular in constrained smooth optimization because it can avoid such an expensive projection step by utilizing an efficient linear maximization oracle (LMO). We present a zeroth-order stochastic Frank–Wolfe (ZOSFW) algorithm for our problem in Algorithm 4. The algorithm is built upon the classical Frank–Wolfe algorithm with the two gradient estimators introduced in Section 5.1.

Algorithm 4 ZOSFW Method

1: **Input**: Initial point $x_0 \in \mathbb{R}^d$, sequence of stepsizes $\{\gamma_t\colon \gamma_t > 0\}_{t=0}^{T-1}$, smoothing parameter δ and iteration number $T \geq 1$.
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: **Option I**: Set $v_t = \text{MB-SGrad}(x_t)$.
4: **Option II**: Set $v_t = \text{VR-SGrad}(x_t, x_{t-1}, t)$.
5: Set $u_t = \arg\max_{u\in\Omega}\langle u, -v_t\rangle$.
6: Set $x_{t+1} = x_t + \gamma_t(u_t - x_t)$.
7: **end for**
8: **return** x_R where R is uniformly sampled from the set $\{1, 2, \ldots, T\}$.

The following theorem shows the convergence rate of solving the problem (1) by Algorithm 4 with Option I (which we call the MB-ZOSFW algorithm).

Theorem 5.6. *Running the MB-ZOSFW algorithm (Algorithm 4 with Option I) with $\gamma_t = \delta^{\frac{1}{2}}T^{-\frac{1}{2}}B^{-\frac{1}{2}}d^{-\frac{1}{4}}$ and the subroutine MB-SGrad (Algorithm 1) with $b_t = b$ where b is some constant, then the output x_R holds that*

$$\mathbb{E}\left[\max_{u\in\Omega}\langle -\nabla F_\delta(x_R), u - x_R\rangle\right] = \mathcal{O}\left(\frac{GB^{\frac{3}{2}}d^{\frac{1}{4}}}{\sqrt{T\delta}} + \frac{GB\sqrt{d}}{\sqrt{b}}\right).$$

Theorem 5.6 and Lemma 3.8(c) imply the following oracle complexity of Algorithm 4 with Option I.

Corollary 5.7. *The MB-ZOSFW algorithm (Algorithm 4 with Option I) requires at most $\mathcal{O}(B^5G^4d^{\frac{3}{2}}\delta^{-1}\epsilon^{-4})$ FQO calls and $\mathcal{O}(B^3G^2d^{\frac{1}{2}}\delta^{-1}\epsilon^{-2})$ LMO calls to obtain a (δ,ϵ)-GFWSP.*

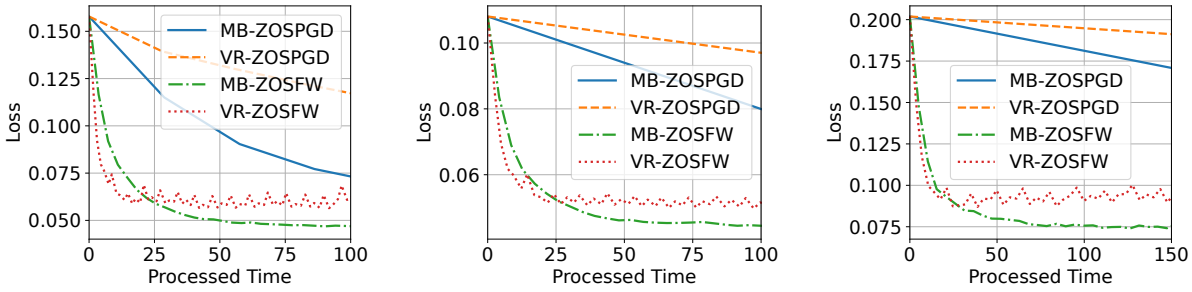
Then, we provide the convergence analysis of solving the problem (1) by Algorithm 4 with Option II (which we call the VR-ZOSFW algorithm), which show that VR technique can be used to improve the convergence rate of the ZOSFW algorithm.

Theorem 5.8. *Running the VR-ZOSFW algorithm (Algorithm 4 with Option II) with $\gamma_t = \sqrt{\delta d^{-1}T^{-1}B^{-1}}$ and the subroutine VR-SGrad (Algorithm 2) with $b_2 = q$, then the output x_R holds that*

$$\mathbb{E}\left[\max_{u\in\Omega}\langle -\nabla F_\delta(x_R), u - x_R\rangle\right] = \mathcal{O}\left(\frac{GB^{\frac{3}{2}}\sqrt{d}}{\sqrt{\delta T}} + \frac{BG\sqrt{d}}{\sqrt{b_1}}\right).$$

Theorem 5.8 and Lemma 3.8(c) imply the following oracle complexity of Algorithm 4 with Option II.

Corollary 5.9. *The VR-ZOSFW algorithm (Algorithm 4 with Option II) requires at most $\mathcal{O}(G^3B^4d^{\frac{3}{2}}\delta^{-1}\epsilon^{-3})$ FQO calls and $\mathcal{O}(G^2B^3d\delta^{-1}\epsilon^{-2})$ LMO calls to obtain a (δ,ϵ)-GFWSP.*



(a) $d = 3000$ (b) $d = 4000$ (c) $d = 5000$

Figure 1. Loss vs. Processed time on the synthetic dataset.

One notable difference in the analysis between our algorithm and the zeroth-order methods for the constrained smooth optimization (Balasubramanian & Ghadimi, 2018; Gao & Huang, 2020) is that we take δ as any positive constant to get a (δ, ϵ)-GFWSP for the problem (1) while algorithms for constrained smooth optimization use δ as a hyperparameter that is set arbitrarily small to obtain the ϵ-stationary point.

6. Experiments

In this section, we conduct numerical experiments to validate the effectiveness of proposed approaches. We note that Chen et al. (2023) has shown the improved convergence result of the VR-ZOSPGD method compared with the MB-ZOSPGD on simple constraints. However, their theoretical results only consider the convergence analysis under the unconstrained setting. Beyond their empirical findings, we aim to demonstrate the improved time efficiency of stochastic projection-free methods compared with the stochastic projection-based methods on more complex constraints. In particular, we evaluate the proposed algorithms on the application of a robust low-rank matrix recovery problem. Formally, we consider the following objective function:

$$\begin{aligned} \min_{X \in \mathbb{R}^{m \times n}} \quad & \sum_{i,j \in \Delta} 1 - \exp(-|X_{i,j} - Y_{i,j}|/\sigma), \\ \text{s.t.} \quad & \|X\|_* \leq B, \end{aligned} \tag{3}$$

where σ is a tunable parameter, $X_{i,j}$ is the i,j-th element of the matrix X, and Δ is the set of observed indices in target matrix $Y \in \mathbb{R}^{m \times n}$. This loss is less sensitive to the discrepancy $X_{i,j} - Y_{i,j}$ compared with the common least square loss, and hence more robust to adversarial outliers (Qu et al., 2018; Shen et al., 2019). We conduct experiments on both synthetic and real-world datasets. For all the experiments, we set the parameter $\sigma = 1$.

For the synthetic dataset, we follow a similar setup of Shen et al. (2019). We first generate an underlying matrix Y of size $d \times d$ with rank $\gamma = 20$ where d is chosen from the set $\{3000, 4000, 5000\}$. We also set singular values of Y as $2^{\lceil \gamma \rceil}/2^{\gamma+1} \times B$ and hence $\|Y\|_* \leq B$. For $d = 3000$ or 4000, we choose the parameter $B = 2000$ while for $d = 5000$, we choose $B = 4000$. We then inject noise into Y by uniformly sampling 5% of the entries in Y and adding random noise uniformly sampled from $[-3, 3]$ to each selected entry. After that, we uniformly sample 10% of the entries in the noise-injected matrix Y as the observations. In terms of hyperparameter setting for the algorithms, we choose the minibatch size $b = 100,000$ for both MB-ZOSPGD and MB-ZOSFW methods. We set $b_1 = 100,000$, $b_2 = 10,000$ and $q = b_1/b_2$ for VR-ZOSPGD and VR-ZOSFW methods. The number of iterations T is set to be 300 for all algorithms. The step size is tuned from the set $\{0.1, 0.03, \ldots, 3\times10^{-7}, 1\times10^{-7}\}$ for each algorithm. The experimental result is demonstrated in Figure 1. We find that MB-ZOSFW and VR-ZOSFW have faster convergence than the ZOSPGD methods because they are projection-free. Interestingly, VR-ZOSPGD converges slower than the MB-ZOSPGD although VR-ZOSPGD has a better theoretical convergence rate. We conjecture that it is due to the expensive projection operation that undermines the efficiency of cheap gradient calculation.

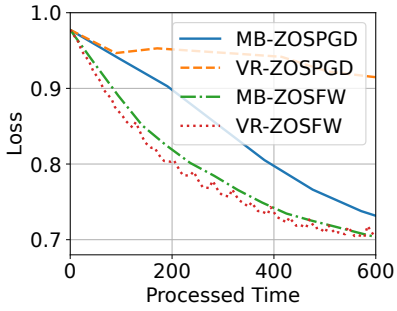


Figure 2. Loss vs. Processed time on the real-world dataset.

For the real-world dataset, we validate our methods on the "MovieLens 1M"[2] dataset. The dataset is a sparse movie

[2] https://grouplens.org/datasets/movielens/1m/

rating matrix Y with 6040 users and 3952 movies. Each rating of the matrix Y is an integer ranging from 1 to 5. We set the parameter $B = 7000$. In terms of the hyperparameter setting for the algorithms, we choose $b = 1,000,000$ for both MB-ZOSPGD and MB-ZOSFW methods. We set $b_1 = 1,000,000$, $b_2 = 100,000$ and $q = b_1/b_2$ for VR-ZOSPGD and VR-ZOSFW methods. For other hyperparameters including the stepsize and the number of iterations, we use the same parameter setting in the synthetic dataset experiment. We present the experimental results on this dataset in Figure 2. We also find that ZOSFW methods converge much faster than ZOSPGD methods. In addition, the VR-ZOSPGD method converges even slower than the MB-ZOSPGD method due to the high projection cost.

7. Conclusion

In this work, we introduce the novel notions of $(\gamma, \delta, \epsilon)$-generalized Goldstein stationary points and (δ, ϵ)-Goldstein Frank–Wolfe stationary points for solving the constrained nonconvex nonsmooth problem. We also propose zeroth-order stochastic projected gradient descent algorithms and stochastic Frank–Wolfe algorithms with non-asymptotic convergence guarantees for obtaining the proposed approximate stationary points. We provide numerical experiments on the robust low-rank matrix recovery problem to show the convergence behavior of the proposed algorithms empirically.

In future work, it is interesting to study the lower bound of the zeroth-order stochastic optimization algorithms for solving unconstrained or constrained nonconvex nonsmooth problems. It is also interesting to investigate whether the $\mathcal{O}(d^{\frac{3}{2}}\delta^{-1}\epsilon^{-3})$ complexity of zeroth-order stochastic algorithms for our problem can be further improved.

Acknowledgement

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2023-08-043T-J). This research is part of the programme DesCartes and is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. Luo Luo is supported by National Natural Science Foundation of China (No. 62206058), Shanghai Sailing Program (22YF1402900), and Shanghai Basic Research Program (23JC1401000). Cheng Chen is supported by National Natural Science Foundation of China (No. 62306116) and the Dean's fund of Shanghai Key Laboratory of Trustworthy Computing.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agarwal, A., Dekel, O., and Xiao, L. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proc. COLT*, pp. 28–40, 2010.

Balasubramanian, K. and Ghadimi, S. Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. In *Proc. NeurIPS*, pp. 3459–3468, 2018.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Chen, L., Xu, J., and Luo, L. Faster gradient-free algorithms for nonsmooth nonconvex stochastic optimization. *arXiv preprint arXiv:2301.06428*, 2023.

Clarke, F. H. *Optimization and nonsmooth analysis*. SIAM, 1990.

Curtis, F. E. and Overton, M. L. A sequential quadratic programming algorithm for nonconvex, nonsmooth constrained optimization. *SIAM Journal on Optimization*, 22 (2):474–500, 2012.

Curtis, F. E., Mitchell, T., and Overton, M. L. A bfgs-sqp method for nonsmooth, nonconvex, constrained optimization and its evaluation using relative minimization profiles. *Optimization Methods and Software*, 32(1):148–181, 2017.

Cutkosky, A., Mehta, H., and Orabona, F. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. *arXiv preprint arXiv:2302.03775*, 2023.

Dao, M. N., Gwinner, J., Noll, D., and Ovcharova, N. Nonconvex bundle method with application to a delamination problem. *Computational Optimization and Applications*, 65(1):173–203, 2016.

Davis, D. and Drusvyatskiy, D. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Davis, D. and Grimmer, B. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

Davis, D., Drusvyatskiy, D., Lee, Y. T., Padmanabhan, S., and Ye, G. A gradient sampling method with complexity guarantees for Lipschitz functions in high and low dimensions. In *Proc. NeurIPS*, 2022.

Duchi, J. C., Bartlett, P. L., and Wainwright, M. J. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

Frank, M. and Wolfe, P. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2): 95–110, 1956.

Gao, H. and Huang, H. Can stochastic zeroth-order Frank-Wolfe method converge faster for non-convex problems? In *Proc. ICML*, pp. 3377–3386, 2020.

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155 (1-2):267–305, 2016.

Goldstein, A. Optimization of Lipschitz continuous functions. *Mathematical Programming*, 13:14–22, 1977.

Grimmer, B. and Jia, Z. Goldstein stationarity in lipschitz constrained optimization. *arXiv preprint arXiv:2310.03690*, 2023.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Hare, W., Sagastizábal, C., and Solodov, M. A proximal bundle method for nonsmooth nonconvex functions with inexact information. *Computational Optimization and Applications*, 63:1–28, 2016.

Hoseini Monjezi, N. and Nobakhtian, S. A filter proximal bundle method for nonsmooth nonconvex constrained optimization. *Journal of Global Optimization*, 79:1–37, 2021.

Jaggi, M. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proc. ICML*, pp. 427–435, 2013.

Jordan, M., Kornowski, G., Lin, T., Shamir, O., and Zampetakis, M. Deterministic nonsmooth nonconvex optimization. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 4570–4597. PMLR, 2023.

Juditsky, A. and Nemirovski, A. S. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

Kornowski, G. and Shamir, O. Oracle complexity in nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 23(314):1–44, 2022.

Kornowski, G. and Shamir, O. An algorithm with optimal dimension-dependence for zero-order nonsmooth nonconvex stochastic optimization. *arXiv preprint arXiv:2307.04504*, 2023.

Lacoste-Julien, S. Convergence rate of Frank-Wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

Lin, T., Zheng, Z., and Jordan, M. I. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. In *Proc. NeurIPS*, 2022.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Makela, M. M. and Neittaanmaki, P. *Nonsmooth optimization: analysis and algorithms with applications to optimal control*. World Scientific, 1992.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Nemirovskij, A. S. and Yudin, D. B. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

Nesterov, Y. and Spokoiny, V. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Nesterov, Y. et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Qu, C., Li, Y., and Xu, H. Non-convex conditional gradient sliding. In *Proc. ICML*, pp. 4208–4217, 2018.

Reddi, S. J., Sra, S., Póczos, B., and Smola, A. Stochastic frank-wolfe methods for nonconvex optimization. In *2016 54th annual Allerton conference on communication, control, and computing (Allerton)*, pp. 1244–1251. IEEE, 2016.

Shen, Z., Fang, C., Zhao, P., Huang, J., and Qian, H. Complexities in projection-free stochastic non-convex minimization. In *Proc. AISTATS*, pp. 2868–2876, 2019.

Smola, A. J. and Schölkopf, B. *Learning with kernels*, volume 4. Citeseer, 1998.

Tang, C.-m., Liu, S., Jian, J.-b., and Li, J.-l. A feasible sqp-gs algorithm for nonconvex, nonsmooth constrained optimization. *Numerical Algorithms*, 65(1):1–22, 2014.

Tian, L. and So, A. M.-C. On the hardness of computing near-approximate stationary points of Clarke regular nonsmooth nonconvex problems and certain DC programs. In *ICML Workshop on Beyond First-Order Methods in ML Systems*, 2021.

Tian, L. and So, A. M.-C. No dimension-free deterministic algorithm computes approximate stationarities of lipschitzians. *arXiv preprint arXiv:2210.06907*, 2022.

Tian, L., Zhou, K., and So, A. M.-C. On the finite-time complexity and practical computation of approximate stationarity concepts of Lipschitz functions. In *Proc. ICML*, pp. 21360–21379, 2022.

Vladarean, M.-L., Doikov, N., Jaggi, M., and Flammarion, N. Linearization algorithms for fully composite optimization. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 3669–3695. PMLR, 2023.

Wang, Z., Ji, K., Zhou, Y., Liang, Y., and Tarokh, V. SpiderBoost and momentum: Faster variance reduction algorithms. In *Proc. NeurIPS*, pp. 2406–2416, 2019.

Xu, M., Ye, J. J., and Zhang, L. Smoothing sqp methods for solving degenerate nonsmooth constrained optimization problems with applications to bilevel programs. *SIAM Journal on Optimization*, 25(3):1388–1410, 2015.

Yurtsever, A., Sra, S., and Cevher, V. Conditional gradient methods via stochastic path-integrated differential estimator. In *Proc. ICML*, pp. 7282–7291, 2019.

Zhang, J., Lin, H., Jegelka, S., Jadbabaie, A., and Sra, S. Complexity of finding stationary points of nonsmooth nonconvex functions. In *Proc. ICML*, pp. 11173–11182, 2020.

Zhang, T. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11(3), 2010.

The appendix is organized as below. Section A introduces several key lemmas essential for the convergence analysis of proposed zeroth-order stochastic optimization methods. Section B presents the proof of the properties introduced in Section 4. Section C proves the convergence rate of MB-ZOSPGD and VR-ZOSPGD methods proposed in Section 5.2. Section D proves the convergence rate of MB-ZOSFW and VR-ZOSFW methods proposed in Section 5.3. For all the proposed algorithms, we also provide the improved large-deviation estimation results of the algorithms with the two-phase postprocessing technique.

A. Supporting Lemmas

In this section, we first review some key lemmas which are essential for the analysis of proposed methods. The following result shows some basic properties of the zeroth-order gradient estimator.

Lemma A.1 (Lin et al. (2022)). *Suppose that f is G-Lipschitz and let $\{g_t\}_{t=0}^{T-1}$ be defined as*

$$g_t = \frac{d}{2\delta}(f(x_t + \delta w_{i,t}) - f(x_t - \delta w_{i,t}))w_{i,t}.$$

where $w_{i,t}$ is uniformly sampled from a unit sphere in $\mathbb{R}^d$. Then, we have $\mathbb{E}[g_t \mid x_t] = \nabla f_\delta(x_t)$ and $\mathbb{E}[\|g_t\|^2 \mid x_t] \leq 16\sqrt{2\pi}dG^2$.

We find the following result useful for the proof of the large-deviation estimation bound of the proposed method.

Proposition A.2 (Juditsky & Nemirovski (2008)). *Suppose that Ω is a Polish space with a Borel probability measure $\mathbb{P}$ and let $\{\emptyset, \Omega\} = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots$ be a sequence of filtration. For an integer $N \geq 1$, we define a martingale difference sequence of Borel functions $\{\zeta_k\}_{k=1}^N \subseteq \mathbb{R}^d$ such that ζ_k is $\mathcal{F}_k$-measurable and $\mathbb{E}[\zeta_k \mid \mathcal{F}_{k-1}] = 0$. Then, if $\mathbb{E}\big[\|\zeta_k\|^2\big] \leq \sigma_k^2$ for all $k \geq 1$, we have $\mathbb{E}\Big[\Big\|\sum_{k=1}^N \zeta_k\Big\|^2\Big] \leq \sum_{k=1}^N \sigma_k^2$ and the following statement holds true*

$$\text{Prob}\left(\left\|\sum_{k=1}^N \zeta_k\right\|^2 \geq \lambda \sum_{k=1}^N \sigma_k^2\right) \leq \frac{1}{\lambda}, \quad \text{for all } \lambda \geq 0.$$

Here we show that the variance of the VR-GRAD estimator (Algorithm 2) can be bounded with the following lemmas.

Lemma A.3. *Assume $g_{i,t}$ is the i-th function call evaluated at t-th iteration for Algorithm 2, then it follows that:*

$$\mathbb{E}[\|g_{i,t} - g_{i,t-1}\|^2] \leq \frac{d^2G^2}{\delta^2}\|x_t - x_{t-1}\|^2.$$

Proof. By the definition of $g_{i,t}$, we have

$$\begin{aligned}
&\|g_{i,t} - g_{i,t-1}\|^2 \\
=&\frac{d^2}{4\delta^2}|F(x_t + \delta w_{i,t}, \xi_{i,t}) - F(x_t - \delta w_{i,t}, \xi_{i,t}) - (F(x_{t-1} + \delta w_{i,t}, \xi_{i,t}) - F(x_{t-1} - \delta w_{i,t}, \xi_{i,t}))|^2 \|w_{i,t}\|^2 \\
=&\frac{d^2}{4\delta^2}|F(x_t + \delta w_{i,t}, \xi_{i,t}) - F(x_{t-1} + \delta w_{i,t}, \xi_{i,t}) - (F(x_t - \delta w_{i,t}, \xi_{i,t}) - F(x_{t-1} - \delta w_{i,t}, \xi_{i,t}))|^2 \\
\leq&\frac{d^2}{2\delta^2}(|F(x_t + \delta w_{i,t}, \xi_{i,t}) - F(x_{t-1} + \delta w_{i,t}, \xi_{i,t})|^2 + |F(x_t - \delta w_{i,t}, \xi_{i,t}) - F(x_{t-1} - \delta w_{i,t}, \xi_{i,t})|^2) \\
\leq&\frac{d^2L(\xi_{i,t})^2}{\delta^2}\|x_t - x_{t-1}\|^2.
\end{aligned}$$

The first inequality is due to $|a+b|^2 \leq 2|a|^2 + 2|b|^2$. The second inequality follows from the assumption that $F(\cdot, \xi_{i,t})$ is $L(\xi_{i,t})$-Lipschitz and the diameter of the feasible set is bounded by B. Taking expectations on both sides of the equation and using the assumption that $\mathbb{E}[L(\xi)^2] \leq G^2$, we have

$$\mathbb{E}[\|g_{i,t} - g_{i,t-1}\|^2] \leq \frac{d^2G^2}{\delta^2}\|x_t - x_{t-1}\|^2.$$

□

Lemma A.4. *Define $n_t = \lfloor t/q \rfloor$ for any t in Algorithm 2, then we have*

$$\mathbb{E}\left[\left\|v_{n_t q} - \nabla F_\delta(x_{n_t q})\right\|^2\right] \leq \frac{16\sqrt{2\pi} d G^2}{b_1}. \tag{4}$$

Proof. Let $n_t = \lfloor t/q \rfloor$ such that $n_t q \leq t \leq (n_t+1)q - 1$

$$\begin{aligned}
&\mathbb{E}\left[\left\|v_{n_t q} - \nabla F_\delta(x_{n_t q})\right\|^2\right] \\
=&\mathbb{E}\left[\left\|\frac{1}{b_1}\sum_{i\in[b_1]} g_{i,n_t q} - \nabla F_\delta(x_{n_t q})\right\|^2\right] \\
\leq&\frac{1}{b_1}\mathbb{E}\left[\left\|g_{1,n_t q} - \nabla F_\delta(x_{n_t q})\right\|^2\right] \\
\leq&\frac{1}{b_1}\mathbb{E}\left[\left\|g_{1,n_t q}\right\|^2\right] \\
\leq&\frac{16\sqrt{2\pi} d G^2}{b_1}.
\end{aligned}$$

The first inequality follows because $g_{i,n_t q}$ are i.i.d. random variables, and a sequence of i.i.d. random variables $\{\zeta_i\}_{i=1}^b$ satisfies that $\mathbb{E}\Big[\left\|\frac{1}{b}\sum_{i=1}^b \zeta_i - \mathbb{E}[\zeta_i]\right\|^2\Big] \leq \frac{1}{b}\mathbb{E}\Big[\left\|\zeta_1 - \mathbb{E}[\zeta_1]\right\|^2\Big]$. The second inequality follows from $\mathbb{E}\big[\left\|\xi - \mathbb{E}[\xi]\right\|^2\big] \leq \mathbb{E}\big[\left\|\xi\right\|^2\big]$ for any random variable ξ. The last inequality is due to Lemma A.1. □

Lemma A.5. *In Algorithm 2, we can bound the variance of the gradient estimator v_t for any t as follows:*

$$\mathbb{E}\left[\left\|v_t - \nabla F_\delta(x_t)\right\|^2\right] \leq \frac{d^2 G^2}{\delta^2 b_2}\sum_{j=n_t q+1}^{t}\left\|x_j - x_{j-1}\right\|^2 + \frac{16\sqrt{2\pi} d G^2}{b_1}.$$

Proof. Let $n_t q \leq t \leq (n_t+1)q - 1$ where $n_t \geq 0$, we have:

$$v_t - \nabla F_\delta(x_t) = v_{n_t q} - \nabla F_\delta(x_{n_t q}) + \sum_{i=n_t q+1}^{t}(v_i - v_{i-1} - (\nabla F_\delta(x_i) - \nabla F_\delta(x_{i-1}))).$$

In addition,

$$v_t = \frac{1}{b_2}\sum_{i\in[b_2]}(g_{i,t} - g_{i,t-1}) + v_{t-1}.$$

Taking expectations on both sides, we have:

$$\mathbb{E}\left[v_t - v_{t-1} - (\nabla F_\delta(x_t) - \nabla F_\delta(x_{t-1}))\right] = 0.$$

As a result, $v_t - \nabla F_\delta(x_t)$ is a martingale. Therefore, we have:

$$\mathbb{E}\left[\left\|v_t - \nabla F_\delta(x_t)\right\|^2\right] = \mathbb{E}\left[\left\|v_{n_t q} - \nabla F_\delta(x_{n_t q})\right\|^2\right] + \sum_{j=n_t q+1}^{t}\mathbb{E}\left[\left\|v_j - v_{j-1} - (\nabla F_\delta(x_j) - \nabla F_\delta(x_{j-1}))\right\|^2\right].$$

We can expand the second term on the right-hand side with:

$$\mathbb{E}\left[\left\|v_j - v_{j-1} - (\nabla F_\delta(x_j) - \nabla F_\delta(x_{j-1}))\right\|^2\right]$$

$$
\begin{aligned}
=&\mathbb{E}\left[\left\|\frac{1}{b_2}\sum_{i\in[b_2]}(g_{i,j}-g_{i,j-1})-(\nabla F_\delta(x_j)-\nabla F_\delta(x_{j-1}))\right\|^2\right]\\
\leq&\frac{1}{b_2}\mathbb{E}\left[\|(g_{1,j}-g_{1,j-1})-(\nabla F_\delta(x_j)-\nabla F_\delta(x_{j-1}))\|^2\right]\\
\leq&\frac{1}{b_2}\mathbb{E}\left[\|g_{1,j}-g_{1,j-1}\|^2\right]\\
\leq&\frac{d^2G^2}{\delta^2 b_2}\|x_j-x_{j-1}\|^2.
\end{aligned}
$$

The first inequality follows because $g_{i,j}$ are i.i.d. random variables, and a sequence of i.i.d. random variables $\{\zeta_i\}_{i=1}^b$ satisfies that $\mathbb{E}\Big[\Big\|\frac{1}{b}\sum_{i=1}^b\zeta_i-\mathbb{E}[\zeta_i]\Big\|^2\Big]\leq\frac{1}{b}\mathbb{E}\Big[\|\zeta_1-\mathbb{E}[\zeta_1]\|^2\Big]$. The second inequality follows from $\mathbb{E}\big[\|\xi-\mathbb{E}[\xi]\|^2\big]\leq\mathbb{E}\big[\|\xi\|^2\big]$ for any random variable ξ. The last inequality is due to Lemma A.3. Combining the above two inequalities with Lemma A.4, it follows that:

$$
\mathbb{E}[\|v_t-\nabla F_\delta(x_t)\|^2]\leq\frac{d^2G^2}{\delta^2 b_2}\sum_{j=n_tq+1}^{t}\|x_j-x_{j-1}\|^2+\frac{16\sqrt{2\pi}dG^2}{b_1}.
$$

□

B. Properties of the Refined Approximate Stationarity

In this section, we present the proof of propositions and theorems proposed in Section 4.

B.1. Proof of Proposition 4.5

The proof of proposition 4.5 (i) is trivial. For (ii), according to the definition of $(\gamma,\frac{\epsilon}{2L},\epsilon/2)$-GGSP, we have $g\in\mathrm{conv}\{\nabla f(y)\colon\ \|y-x\|\leq\frac{\epsilon}{2L}\}$ such that $\|\mathcal{G}(x,g,\gamma)\|\leq\epsilon/2$. The conv operation means that there exists k coefficients $\alpha_1,\ldots,\alpha_k$ satisfying $\sum_{i=1}^k\alpha_i=1$ and k points $x_1,\ldots,x_k\in\{y\colon\ \|y-x\|\leq\frac{\epsilon}{2L}\}$ satisfying:

$$
g=\sum_{i=1}^k\alpha_i\nabla f(x_i).
$$

Assume $y_1=\psi(x,g,\gamma)$ and $y_2=\psi(x,\nabla f(x),\gamma)$, then by the definition of $\psi(\cdot)$ it follows that:

$$
\langle g+\frac{y_1-x}{\gamma},y_1-y_2\rangle\leq 0,
$$

and

$$
\langle\nabla f(x)+\frac{y_2-x}{\gamma},y_2-y_1\rangle\leq 0.
$$

Add the above two inequalities together, and we have:

$$
\langle g-\nabla f(x)+\frac{y_1-y_2}{\gamma},y_1-y_2\rangle\leq 0.
$$

Rearrange the terms, we have:

$$
\begin{aligned}
\|y_1-y_2\|^2\leq&\gamma\langle\nabla f(x)-g,y_1-y_2\rangle\\
\leq&\gamma\|\nabla f(x)-g\|\,\|y_1-y_2\|\\
\leq&\sum_{i=1}^k\gamma\alpha_i\|\nabla f(x)-\nabla f(x_i)\|\,\|y_1-y_2\|
\end{aligned}
$$

$$\begin{aligned}
&\leq \sum_{i=1}^{k} \gamma L \alpha_i \left\| x - x_i \right\| \left\| y_1 - y_2 \right\| \\
&\leq \frac{\gamma \epsilon \left\| y_1 - y_2 \right\|}{2}.
\end{aligned}$$

The second inequality is due to the Cauchy–Schwartz inequality, the third inequality is due to the Jensen's inequality, and the fourth inequality follows the definition that f is L-smooth. Therefore, we have

$$\begin{aligned}
&\left\| \mathcal{G}(x, \nabla f(x), \gamma) \right\| \\
=&\left\| \frac{x - y_2 + y_1 - y_1}{\gamma} \right\| \\
\leq&\left\| \frac{x - y_1}{\gamma} \right\| + \left\| \frac{y_2 - y_1}{\gamma} \right\| \\
\leq&\frac{\epsilon}{2} + \frac{\epsilon}{2} \\
=&\epsilon.
\end{aligned}$$

The first equality is due to the definition of y_1 and y_2. The first inequality is due to $\|a - b\| \leq \|a\| + \|b\|$.

B.2. Proof of Proposition 4.6

The proof of Proposition 4.6 (i) is trivial. For (ii), according to the definition of $(\epsilon/(3BL), 2\epsilon/3)$-GFWSP, we have $g \in \operatorname{conv}\{\nabla f(y) \colon \|y - x\| \leq \frac{\epsilon}{3BL}\}$ such that $\max_{u \in \Omega} \langle u - x, -g \rangle \leq 2\epsilon/3$. The conv operation means that there exists k coefficients $\alpha_1, \ldots, \alpha_k$ satisfying $\sum_{i=1}^{k} \alpha_i = 1$ and k points $x_1, \ldots, x_k \in \{y \colon \|y - x\| \leq \frac{\epsilon}{3BL}\}$ satisfying:

$$g = \sum_{i=1}^{k} \alpha_i \nabla f(x_i).$$

Then it follows that

$$\begin{aligned}
&\max_{u \in \Omega} \langle u - x, -\nabla f(x) \rangle \\
=&\max_{u \in \Omega} (\langle u - x, -\nabla f(x) + g \rangle + \langle u - x, -g \rangle) \\
\leq&\max_{u \in \Omega} \left\| u - x \right\| \left\| g - \nabla f(x) \right\| + \max_{u' \in \Omega} \langle u' - x, -g \rangle \\
\leq&B \left\| \sum_{i=1}^{k} \alpha_i \nabla f(x_i) - \nabla f(x) \right\| + \frac{2\epsilon}{3} \\
\leq&\sum_{i=1}^{k} \alpha_i BL \left\| x_i - x \right\| + \frac{2\epsilon}{3} \\
\leq&\frac{\epsilon}{3} + \frac{2\epsilon}{3} \\
=&\epsilon.
\end{aligned}$$

The first inequality is due to the Cauchy–Schwartz inequality, the second inequality follows from domain Ω has diameter B, and the third inequality is the result of applying Jensen's Inequality.

B.3. Proof of Theorem 4.7

We first introduce some notations. Function $h(\cdot)$ is a hard and non-negative function sampled from a uniform distribution of hard functions h_σ. In particular, the hard functions h_σ is defined in the following recursive manner,

$$h_0^1(x) = \begin{cases} 1-x & x \in (-\infty, 0] \\ 1-2x & x \in (0, \frac{3}{8}] \\ \frac{6}{5}x - \frac{1}{5} & x \in (\frac{3}{8}, 1] \\ x & x \in (1, \infty) \end{cases}, \quad h_1^1(x) = \begin{cases} 1-x & x \in (-\infty, 0] \\ -\frac{6}{5}x + 1 & x \in (0, \frac{5}{8}] \\ 2x-1 & x \in (\frac{5}{8}, 1] \\ x & x \in (1, \infty) \end{cases}$$

For any $\hat{\sigma} := (\sigma_2, \ldots, \sigma_N) \in \{0,1\}^{N-1}$, we define

$$h_{0,\hat{\sigma}}^N(x) = \begin{cases} 1-x & x \in (-\infty, 0] \\ 1-2x & x \in (0, \frac{1}{4}] \\ \frac{1}{4}h_{\hat{\sigma}}^{(N-1)}(4x-1) + \frac{1}{4} & x \in (\frac{1}{4}, \frac{1}{2}] \\ x & x \in (\frac{1}{2}, 1] \\ x & x \in (1, \infty) \end{cases},$$

$$h_{1,\hat{\sigma}}^N(x) = \begin{cases} 1-x & x \in (-\infty, 0] \\ 1-x & x \in (0, \frac{1}{2}] \\ \frac{1}{4}h_{\hat{\sigma}}^{(N-1)}(4x-2) + \frac{1}{4} & x \in (\frac{1}{2}, \frac{3}{4}] \\ 2x-1 & x \in (\frac{3}{4}, 1] \\ x & x \in (1, \infty) \end{cases}$$

Let $\tilde{x} \in (0,1)$ denote the global minima of $h(\cdot)$. Then we define

$$\bar{f}(x) := h(x_d) + \frac{1}{4}\sqrt{\sum_{i=1}^{d-1} x_i^2}$$

$$f_w(x) := h(x_d + \tilde{x}) + \frac{1}{4}\sqrt{\sum_{i=1}^{d-1} x_i^2} - [\langle \bar{w}, x + w\rangle - \frac{1}{2}\|x + w\|]_+$$

$$F_w(x_1, \ldots, x_d) := \max\{-1, f_w(x - x^*)\}$$

where $x^* = (0, \ldots, 0, \tilde{x})$ is the global minima of $\bar{f}(x)$. For a vector x, we use

$$\bar{x} = \frac{x}{\|x\|}$$

to denote the normalized vector of x.

Now we consider solving $\min_{x\in\Omega} F_w(x)$ over the feasible set $\Omega = [-100, 100]^d$.

We first show that The function $F_w(\cdot)$ satisfies following properties:

Lemma B.1 ((Lemma 14 and Lemma 15 of (Kornowski & Shamir, 2022))). *$F_w(\cdot)$ satisfies the following properties:*

1. *$F_w(\cdot)$ is $\frac{15}{4}$-Lipschitz, $F_w(0) - \inf_x F_w(x) \le 2$ and $\inf\{\|x\| \mid \partial F_w(x) = \{0\}\} \le 13$.*
2. *f_w has no ϵ-stationary points for any $\epsilon < \frac{1}{4\sqrt{2}}$.*
3. *Any ϵ-stationary point x of $F_w(\cdot)$ for $\epsilon < \frac{1}{4\sqrt{2}}$ satisfies $F_w(x) = -1$.*
4. *For all vector x which satisfies $x \neq x^*$ and $\langle \bar{w}, \overline{x - x^*}\rangle \le \frac{1}{2} - \frac{3\|w\|}{2\|x-x^*\|}$, we have $F_w(x) = \bar{f}(x) := h(x_d) + \frac{1}{4}\sqrt{\sum_{i=1}^{d-1} x_i^2}$.*

5. *There exists a choice of w, such that if we run any algorithm $\mathcal{A}$ with a local oracle on $F_w(\cdot)$, then with probability at least $1-2T\exp(-d/36)$, the algorithm's iterates $x_1^{F_w},\dots,x_T^{F_w}$ satisfy $\min_{t\in[T]} F_w(\mathbf{x}_t^{F_w})>0$.*

Based on Lemma 1, we can obtain the following results:

Lemma B.2. *If a vector x satisfies*

$$x\neq x^* \ \text{ and } \ \langle \bar{w}, \overline{x-x^*}\rangle \leq \frac{1}{2}-\frac{3\|w\|}{2\|x-x^*\|} \tag{5}$$

then the norm of the generalized gradient mapping with $\gamma\leq 0.1$ and 0-Frank–Wolfe gap of F_w at point x is not less than $1/(4\sqrt{2})$.

Proof. **step 1:** We first consider the subgradient of point x satisfying condition (5).

According to (4) of Lemma B.1 and the fact that $h(\cdot)$ is non-negative, we have $F_w(x)=\overline{f}(x)\geq 0$ for all x satisfying the condition (5). In addition, by (3) of Lemma B.1, all ϵ-stationary points x with $\epsilon<\frac{1}{4\sqrt{2}}$ satisfy $F_w(x)=-1$. Therefore, every point x satisfying condition (5) meets $\inf\{\|g\|\colon g\in\partial F_w(x)\}\geq\frac{1}{4\sqrt{2}}$.

step 2: Then we consider the norm of the generalized gradient mapping.

Let $g\in\partial F_w(x)$ and g_i be the i-th coordinate of g. By the definition of $\overline{f}(x)$, g_i has the same sign as the x_i and $|g_i|\leq 1$ for $i\in[d-1]$. In addition, by (1) of Lemma B.1 and the definition of $h(\cdot)$, g_d satisfies

$$\begin{cases} g_d=-1 & x_d<0\\ g_d=1 & x_d>1\\ |g_d|\leq\frac{15}{4} & x_d\in[0,1]\end{cases} \tag{6}$$

If x is in the feasible set, then $x-\gamma g$ is also in the feasible set since $\gamma\leq 0.1$. Then we can get $\mathcal{G}(x,g,\gamma)=g$, which indicates that $\|\mathcal{G}(x,g,\gamma)\|=\|g\|$.

If x is not in the feasible set, then there exists $i\in[d]$ such that $|x_i|>100$. Since x_i and g_i have the same sign, $\gamma\leq 0.1$, and $|g_i|\leq 1$, we have $|x_i-\gamma g_i|\leq|x_i|$. If $|x_i-\gamma g_i|\leq 100$, then $\mathrm{proj}_\Omega(x_i-\gamma g_i)=x_i-\gamma g_i$. Otherwise, $\mathrm{proj}_\Omega(x_i-\gamma g_i)=sign(x_i)\cdot 100$. Consequently, $|x_i-\mathrm{proj}_\Omega(x_i-\gamma g_i)|\geq\gamma|g_i|$. By the definition of generalized gradient mapping, we have $\|\mathcal{G}(x,g,\gamma)\|\geq\|g\|$.

To sum up, we can get $\|\mathcal{G}(x,g,\gamma)\|\geq\|g\|\geq\frac{1}{4\sqrt{2}}$ where we use the result of step 1.

step 3: Finally we consider the 0-Frank–Wolfe gap.

If $x_d\neq\tilde{x}$, we have $|g_d|\geq 1$ by Proposition 11 of (Kornowski & Shamir, 2022). Then we can get $\max_{u\in\Omega}\langle u-x,-g\rangle\geq\max_{u_d\in[-100,100]}-g_d(u_d-x_d)\geq\frac{1}{4\sqrt{2}}$, where the first inequality is because we can set $u_i=x_i$ for $i\in[d-1]$ and the second inequality is due to property (6).

If $x_d=\tilde{x}$, there exists $x_1,\dots,x_{d-1}$ cannot be all zero since $x\neq x^*$. Then we can obtain the norm of the gradient of $\frac{1}{4}\sqrt{\sum_{i=1}^{d-1}x_i^2}$ is $1/4$, which means $\sqrt{\sum_{i=1}^{d-1}g_i^2}=1/4$ and $g_i\in[0,1/4]$. Thus we have $\max_{u\in\Omega}\langle u-x,-g\rangle\geq\sum_{i=1}^{d-1}4g_i^2=\frac{1}{4}\geq\frac{1}{4\sqrt{2}}$, where we choose $u_d=x_d$ and $u_i=x_i-4g_i$ (Notice that we can easily get $(x_i-4g_i)\in[-100,100]$ for $i\in[d-1]$ since $g_i\in[0,1/4]$).

According to the Eq.(12) of (Kornowski & Shamir, 2022), we have

$$Pr_{\mathcal{A}}\left[\left(\min_{t\in[T]}\left\|x_t^{\bar{f}}-x^*\right\|\geq\rho\right)\wedge\left(\max_{t\in[T]}\langle\bar{w},\overline{x_t^{\bar{f}}-x^*}\rangle<\frac{1}{3}\right)\right]>1-2T\exp(-d/36),$$

where $\rho>0$ is a constant, $\|w\|=\rho/99$ and $x_t^{\bar{f}}$ is the t-th iteration of algorithm $\mathcal{A}$ minimizing $\bar{f}$. Thus by Lemma B.2, we have $\mathbf{x}_t^{\bar{f}}$ is not a (γ,ϵ)-GCSP or ϵ-CFWSP for $\epsilon<\frac{1}{4\sqrt{2}}$ with probability $1-2T\exp(-d/36)$. □

Algorithm 5 Two-Phase Zeroth-Order Stochastic Projected Gradient Descent Method

Input: Initial point $x_0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, problem dimension $d \geq 1$, smoothing parameter δ, iteration number $T \geq 1$, number of rounds $S \geq 1$ and sample size $\hat{B}$.
for $s = 0, 1, \ldots, S-1$ **do**
  Call Algorithm 3 with x_0, γ, d, δ and T and let $\bar{x}^s$ be an output.
end for
for $s = 0, 1, \ldots, S-1$ **do**
  for $k = 0, 1, \ldots, \hat{B}-1$ **do**
    Simulate $\xi_k \sim \mathcal{P}$
    Sample $w_k \in \mathbb{R}^d$ uniformly from a unit sphere in $\mathbb{R}^d$.
    $g_k^s = \frac{d}{2\delta}(f(\bar{x}^s + \delta w_k, \xi_k) - f(\bar{x}^s - \delta w_k, \xi_k))w_k$
  end for
  $g^s = \frac{1}{\hat{B}} \sum_{k=0}^{\hat{B}-1} g_k^s$
  $\bar{g}^s = \frac{1}{\gamma}(\bar{x}^s - \psi(\bar{x}^s, g^s, \gamma))$
end for
Let $\hat{S} \coloneqq \{0, 1, \ldots, S-1\}$, and choose $s^* = \arg\min_{s \in \hat{S}} \|\bar{g}^s\|$
return $\bar{x}^{s^*}$.

C. Convergence Analysis of ZOSPGD Methods

In this section, we prove the convergence rate of MB-ZOSPGD and VR-ZOSPGD methods introduced in Section 5.2. First, we review some fundamental lemmas demonstrating the properties of the gradient mapping operator.

Lemma C.1 (Ghadimi et al. (2016)). *For arbitrary $g_1, g_2 \in \mathbb{R}^d$, we have*

$$\|\mathcal{G}(x, g_1, \gamma) - \mathcal{G}(x, g_2, \gamma)\| \leq \|g_1 - g_2\|. \tag{7}$$

Lemma C.2 (Ghadimi et al. (2016)). *For arbitrary $g \in \mathbb{R}^d$, we can show that*

$$\langle g, \mathcal{G}(x, g, \gamma)\rangle \geq \|\mathcal{G}(x, g, \gamma)\|^2. \tag{8}$$

Proof. Denote $x^+ \coloneqq \psi(x, g, \gamma)$. By the optimality of x^+ on the convex set Ω, for $\forall u \in \Omega$ we have,

$$\langle g + \frac{1}{\gamma}(x^+ - x), u - x^+\rangle \geq 0.$$

Let $u = x$, it follows

$$\langle g, x - x^+\rangle \geq \frac{1}{\gamma}\left\|x - x^+\right\|^2.$$

By dividing both sides by γ, we get the desired result. □

C.1. Convergence Analysis of the MB-ZOSPGD method

We can now show the convergence rate of the MB-ZOSPGD method with the following lemma.

Lemma C.3. *Running the MB-ZOSPGD method (Algorithm 3 with Option I), then the output x_R holds that*

$$\mathbb{E}[\|\mathcal{G}(x_R, v_R, \gamma)\|^2] \leq \mathbb{E}\left[\frac{F_\delta(x_0) - F_\delta(x_T)}{T(\gamma - \frac{c\gamma^2 G\sqrt{d}}{2\delta})} + \frac{\gamma}{T(\gamma - \frac{c\gamma^2 G\sqrt{d}}{2\delta})}\sum_{t=0}^{T-1}\|v_t - \nabla F_\delta(x_t)\|^2\right].$$

Proof. Since F_δ is $\frac{cG\sqrt{d}}{\delta}$-smooth, we have

$$F_\delta(x_{t+1}) \leq F_\delta(x_t) + \langle \nabla F_\delta(x_t), x_{t+1} - x_t\rangle + \frac{cG\sqrt{d}}{2\delta}\|x_{t+1} - x_t\|^2$$

$$\begin{aligned}
\leq & F_\delta(x_t) - \gamma\langle \nabla F_\delta(x_t), \mathcal{G}(x_t, v_t, \gamma)\rangle + \frac{c\gamma^2 G\sqrt{d}}{2\delta}\left\|\mathcal{G}(x_t, v_t, \gamma)\right\|^2 \\
= & F_\delta(x_t) - \gamma\langle v_t, \mathcal{G}(x_t, v_t, \gamma)\rangle + \gamma\langle v_t - \nabla F_\delta(x_t), \mathcal{G}(x_t, v_t, \gamma)\rangle + \frac{c\gamma^2 G\sqrt{d}}{2\delta}\left\|\mathcal{G}(x_t, v_t, \gamma)\right\|^2.
\end{aligned}$$

The second inequality follows from the definition of the $\mathcal{G}(x_t, v_t, \gamma)$. Using Lemma C.2, we have

$$\begin{aligned}
F_\delta(x_{t+1}) \leq & F_\delta(x_t) - \left(\gamma - \frac{c\gamma^2 G\sqrt{d}}{2\delta}\right)\left\|\mathcal{G}(x_t, v_t, \gamma)\right\|^2 + \gamma\langle v_t - \nabla F_\delta(x_t), \mathcal{G}(x_t, v_t, \gamma)\rangle \\
\leq & F_\delta(x_t) - \left(\gamma - \frac{c\gamma^2 G\sqrt{d}}{2\delta}\right)\left\|\mathcal{G}(x_t, v_t, \gamma)\right\|^2 + \gamma\langle v_t - \nabla F_\delta(x_t), \mathcal{G}(x_t, \nabla F_\delta(x_t), \gamma)\rangle \\
& + \gamma\left\|v_t - \nabla F_\delta(x_t)\right\|\left\|\mathcal{G}(x_t, v_t, \gamma) - \mathcal{G}(x_t, \nabla F_\delta(x_t), \gamma)\right\| \\
\leq & F_\delta(x_t) - \left(\gamma - \frac{c\gamma^2 G\sqrt{d}}{2\delta}\right)\left\|\mathcal{G}(x_t, v_t, \gamma)\right\|^2 + \gamma\langle v_t - \nabla F_\delta(x_t), \mathcal{G}(x_t, \nabla F_\delta(x_t), \gamma)\rangle \\
& + \gamma\left\|v_t - \nabla F_\delta(x_t)\right\|^2.
\end{aligned}$$

The second inequality follows from the Cauchy–Schwarz inequality. The last inequality is due to Lemma C.1. Take expectations on both sides and rearrange the terms, and note that $\mathbb{E}[v_t] = \nabla F_\delta(x_t)$, then we have

$$\left(\gamma - \frac{c\gamma^2 G\sqrt{d}}{2\delta}\right)\mathbb{E}\left[\left\|\mathcal{G}(x_t, v_t, \gamma)\right\|^2\right] \leq \mathbb{E}\left[F_\delta(x_t) - F_\delta(x_{t+1}) + \gamma\left\|v_t - \nabla F_\delta(x_t)\right\|^2\right].$$

Sum up both sides of the inequality from $t = 0$ to $T - 1$, and divide both sides by T

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(\gamma - \frac{c\gamma^2 G\sqrt{d}}{2\delta}\right)\mathbb{E}\left[\left\|\mathcal{G}(x_t, v_t, \gamma)\right\|^2\right] \leq \mathbb{E}\left[\frac{F_\delta(x_0) - F_\delta(x_T)}{T} + \frac{\gamma}{T}\sum_{t=0}^{T-1}\left\|v_t - \nabla F_\delta(x_t)\right\|^2\right].$$

□

Now we can prove Theorem 5.1 with the above result.

C.1.1. Proof of Theorem 5.1

Substituting $\gamma = \frac{\delta}{cG\sqrt{d}}$ into the above lemma, we have the following result:

$$\begin{aligned}
\mathbb{E}[\left\|\mathcal{G}(x_R, v_R, \gamma)\right\|^2] \leq & \mathbb{E}\left[\frac{2cG\sqrt{d}(F_\delta(x_0) - F_\delta(x_T))}{T\delta} + \frac{2}{T}\sum_{t=0}^{T-1}\left\|v_t - \nabla F_\delta(x_t)\right\|^2\right] \\
\leq & \mathbb{E}\left[\frac{2cG\sqrt{d}(F_\delta(x_0) - F_\delta(x_T))}{T\delta} + \frac{2}{T}\sum_{t=0}^{T-1}\left\|v_t\right\|^2\right] \\
\leq & \frac{2cG^2\sqrt{d}B}{T\delta} + \frac{32\sqrt{2\pi}dG^2}{b}.
\end{aligned}$$

The second inequality follows from $\mathbb{E}\left[\left\|\xi - \mathbb{E}[\xi]\right\|^2\right] \leq \mathbb{E}\left[\left\|\xi\right\|^2\right]$ for any random variable ξ. The last inequality follows from the G-Lipschitzness of the function F, B is the upper bound of the diameter of the feasible set Ω and Lemma A.1.

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathcal{G}(x_R, \nabla F_\delta(x_R), \gamma)\right\|^2\right] \leq & \mathbb{E}\left[2\left\|\mathcal{G}(x_R, \nabla F_\delta(x_R), \gamma) - \mathcal{G}(x_R, v_R, \gamma)\right\|^2 + 2\left\|\mathcal{G}(x_R, v_R, \gamma)\right\|^2\right] \\
\leq & 2\mathbb{E}\left[\left\|v_R - \nabla F_\delta(x_R)\right\|^2\right] + 2\mathbb{E}\left[\left\|\mathcal{G}(x_R, v_R, \gamma)\right\|^2\right] \\
\leq & \frac{4cG^2\sqrt{d}B}{T\delta} + \frac{96\sqrt{2\pi}dG^2}{b}.
\end{aligned}$$

The first inequality is due to the fact $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$. The second inequality follows from Lemma C.1. By Lemma 3.8, we have $\nabla F_\delta(x_R) \in \partial_\delta F(x_R)$. This together with the above inequality implies that

$$\mathbb{E}[\min\{\|\mathcal{G}(x_R, g, \gamma)\| : g \in \partial_\delta F(x_R)\}] \le \mathbb{E}[\|\mathcal{G}(x_R, \nabla F_\delta(x_R), \gamma)\|] \le \frac{2G\sqrt{cB}d^{\frac{1}{4}}}{\sqrt{T\delta}} + \frac{24\sqrt{d}G}{\sqrt{b}}.$$

C.1.2. Proof of Corollary 5.2

To ensure $\mathbb{E}[\min\{\|\mathcal{G}(x_R, g, \gamma)\| : g \in \partial_\delta F(x_R)\}] \le \epsilon$, we choose $b = \frac{dG^2}{\epsilon^2}$, the total number of the function value oracle calls is bounded by

$$bT = \mathcal{O}\left(\frac{G^2 B d^{\frac{1}{2}}}{\delta\epsilon^2} \cdot \frac{dG^2}{\epsilon^2}\right) = \mathcal{O}\left(\frac{G^4 B d^{\frac{3}{2}}}{\delta\epsilon^4}\right).$$

C.1.3. Large-deviation Estimation of the Two-Phase MB-ZOSPGD Method

While Theorem 5.1 and 5.3 establish the expected convergence rate over many runs of Algorithm 3, we are also interested in the large-deviation properties for a single run. To show such a bound, we combine Algorithm 3 with a post-optimization procedure (Ghadimi et al., 2016), leading to a two-phase zeroth-order stochastic projected gradient descent method (2-ZOSPGD) that is shown in Algorithm 5. Formally, we provide the large-deviation estimation of the two-phase MB-ZOSPGD method as follows.

Theorem C.4. *Let $\delta > 0$ and $0 < \epsilon, \Lambda < 1$, then there exists some $T, S, \hat{B} > 0$ such that the output $\bar{x}^{s^*}$ of Algorithm 5 with MB-ZOSPGD satisfies that* $\text{Prob}\left(\min\{\left\|\mathcal{G}(\bar{x}^{s^*}, g, \gamma)\right\| \ge \epsilon \colon g \in \partial_\delta F(\bar{x}^{s^*})\}\right) \le \Lambda$ *and the total number of calls of the FQO is bounded by*

$$\mathcal{O}\left(\frac{G^4 B d^{\frac{3}{2}}}{\delta\epsilon^4}\log\left(\frac{1}{\Lambda}\right) + \frac{dG^2}{\epsilon^2\Lambda}\log\left(\frac{1}{\Lambda}\right)^2\right).$$

Proof. By the definition of s^* and the Cauchy–Schwarz inequality, we have

$$\begin{aligned}\left\|\mathcal{G}(\bar{x}^{s^*}, g^{s^*}, \gamma)\right\|^2 =& \min_{s\in\{0,1,\ldots,S-1\}} \|\mathcal{G}(\bar{x}^s, g^s, \gamma)\|^2 \le \min_{s\in\{0,1,\ldots,S-1\}} (2\|\mathcal{G}(\bar{x}^s, \nabla F_\delta(\bar{x}^s), \gamma)\|^2 + 2\|g^s - \nabla F_\delta(\bar{x}^s)\|^2) \\ \le& 2\min_{s\in\{0,1,\ldots,S-1\}} \|\mathcal{G}(\bar{x}^s, \nabla F_\delta(\bar{x}^s), \gamma)\|^2 + 2\max_{s\in\{0,1,\ldots,S-1\}} \|g^s - \nabla F_\delta(\bar{x}^s)\|^2.\end{aligned}$$

The first inequality is due to Lemma C.1 and the fact that $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$. It implies that

$$\begin{aligned}\left\|\mathcal{G}(\bar{x}^{s^*}, \nabla F_\delta(\bar{x}^{s^*}), \gamma)\right\|^2 \le& 2\left\|\mathcal{G}(\bar{x}^{s^*}, g^{s^*}, \gamma)\right\| + 2\left\|g^{s^*} - \nabla F_\delta(\bar{x}^{s^*})\right\|^2 \\ \le& 4\min_{s\in\{0,1,\ldots,S-1\}} \|\mathcal{G}(\bar{x}^s, \nabla F_\delta(\bar{x}^s), \gamma)\|^2 + 4\max_{s\in\{0,1,\ldots,S-1\}} \|g^s - \nabla F_\delta(\bar{x}^s)\|^2 \\ &+ 2\left\|g^{s^*} - \nabla F_\delta(\bar{x}^{s^*})\right\|^2.\end{aligned}$$

The first inequality follows from Cauchy–Schwarz inequality and Lemma C.1. The next step is to provide the probabilistic bounds on all the terms on the right-hand side of the above inequality. Theorem 5.1 implies that

$$\mathbb{E}\left[\|\mathcal{G}(\bar{x}^s, \nabla F_\delta(\bar{x}^s), \gamma)\|^2\right] \le \frac{4cG^2\sqrt{d}B}{T\delta} + \frac{96\sqrt{2\pi}dG^2}{b}.$$

Using Markov's inequality, we have

$$\text{Prob}\left(\|\mathcal{G}(\bar{x}^s, \nabla F_\delta(\bar{x}^s), \gamma)\|^2 \ge \frac{8cG^2\sqrt{d}B}{T\delta} + \frac{192\sqrt{2\pi}dG^2}{b}\right) \le \frac{1}{2}.$$

Thus, we have

$$\text{Prob}\left(\min_{s\in\{0,1,\ldots,S-1\}} \|\mathcal{G}(\bar{x}^s, \nabla F_\delta(\bar{x}^s), \gamma)\|^2 \ge \frac{8cG^2\sqrt{d}B}{T\delta} + \frac{192\sqrt{2\pi}dG^2}{b}\right) \le \frac{1}{2^S}.$$

Furthermore, for each $s \in \{0, 1, \ldots, S-1\}$, we have

$$g^s - \nabla F_\delta(\bar{x}^s) = \frac{1}{\hat{B}} \sum_{k=0}^{\hat{B}-1} (g_k^s - \nabla F_\delta(\bar{x}^s)).$$

By Lemma A.1, we have $\mathbb{E}\left[g_k^s \mid \bar{x}^s\right] = \nabla F_\delta(\bar{x}^s)$ and $\mathbb{E}\left[\|g_k^s\|^2 \mid \bar{x}^s\right] \leq 16\sqrt{2\pi} dG^2$. By Markov's inequality and Proposition A.2, it implies that

$$\text{Prob}\left(\|g^s - \nabla F_\delta(\bar{x}^s)\|^2 \geq \frac{\lambda(16\sqrt{2\pi} dG^2)}{\hat{B}}\right) = \text{Prob}\left(\left\|\sum_{k=0}^{\hat{B}-1} (g_k^s - \nabla F_\delta(\bar{x}^s))\right\|^2 \geq \lambda \hat{B}(16\sqrt{2\pi} dG^2)\right) \leq \frac{1}{\lambda}.$$

Thus, we conclude that

$$\text{Prob}\left(\max_{s\in\{0,1,\ldots,S-1\}} \|g^s - \nabla F_\delta(\bar{x}^s)\|^2 \geq \frac{\lambda(16\sqrt{2\pi} dG^2)}{\hat{B}}\right) \leq \frac{S}{\lambda}.$$

By a similar argument, one has

$$\text{Prob}\left(\left\|g^{s^*} - \nabla F_\delta(\bar{x}^{s^*})\right\|^2 \geq \frac{\lambda(16\sqrt{2\pi} dG^2)}{\hat{B}}\right) \leq \frac{1}{\lambda}.$$

Combining the above inequalities yields that

$$\text{Prob}\left(\left\|\mathcal{G}(x^{s^*}, \nabla F_\delta(x^{s^*}), \gamma)\right\|^2 \geq \frac{32cG^2\sqrt{d}B}{T\delta} + \frac{768\sqrt{2\pi} dG^2}{b} + \frac{\lambda 96\sqrt{2\pi} dG^2}{\hat{B}}\right) \leq \frac{S+1}{\lambda} + \frac{1}{2^S}.$$

If we set $\lambda = \frac{2(S+1)}{\Lambda}$, $S = \lceil \log(2/\Lambda) \rceil$ and the parameters $(T, b, \hat{B})$ as follows

$$T = \mathcal{O}\left(\frac{G^2 B d^{\frac{1}{2}}}{\delta\epsilon^2}\right), \quad b = \mathcal{O}\left(\frac{dG^2}{\epsilon^2}\right), \quad \hat{B} = \mathcal{O}\left(\frac{dG^2}{\epsilon^2\Lambda} \log\left(\frac{1}{\Lambda}\right)\right).$$

To satisfy $\text{Prob}(\min\{\left\|\mathcal{G}(\bar{x}^{s^*}, g, \gamma)\right\| : g \in \partial_\delta F(\bar{x}^{s^*})\} \geq \epsilon) \leq \Lambda$, the total number of function oracle calls is bounded by

$$S(Tb + \hat{B}) = \mathcal{O}\left(\frac{G^4 B d^{\frac{3}{2}}}{\delta\epsilon^4} \log\left(\frac{1}{\Lambda}\right) + \frac{dG^2}{\epsilon^2\Lambda} \log\left(\frac{1}{\Lambda}\right)^2\right).$$

□

C.2. Convergence Analysis of VR-ZOSPGD

We can prove the convergence rate of the VR-ZOSPGD method as follows.

Lemma C.5. *Running the VR-ZOSPGD method (Algorithm 3 with Option II), then the output x_R holds that*

$$\mathbb{E}\left[\|\mathcal{G}(x_R, v_R, \gamma)\|^2\right] \leq \left(\frac{\gamma}{2} - \frac{c\gamma^2 G\sqrt{d}}{2\delta} - \frac{\gamma^3 d^2 G^2 q}{2\delta^2 b_2}\right)^{-1} \left(\frac{\mathbb{E}[F_\delta(x_0)] - \mathbb{E}[F_\delta(x_T)]}{T} + \frac{8\sqrt{2\pi} d\gamma G^2}{b_1}\right).$$

Proof. By the Lipschitz continuity of ∇F_δ, we have

$$F_\delta(x_{t+1}) \leq F_\delta(x_t) + \langle \nabla F_\delta(x_t), x_{t+1} - x_t \rangle + \frac{cG\sqrt{d}}{2\delta} \|x_{t+1} - x_t\|^2$$

$$
\begin{aligned}
\leq & F_\delta(x_t) - \gamma\langle \nabla F_\delta(x_t), \mathcal{G}(x_t, v_t, \gamma)\rangle + \frac{c\gamma^2 G\sqrt{d}}{2\delta}\left\|\mathcal{G}(x_t, v_t, \gamma)\right\|^2 \\
= & F_\delta(x_t) - \gamma\langle \nabla F_\delta(x_t) - v_t, \mathcal{G}(x_t, v_t, \gamma)\rangle - \gamma\langle v_t, \mathcal{G}(x_t, v_t, \gamma)\rangle + \frac{c\gamma^2 G\sqrt{d}}{2\delta}\left\|\mathcal{G}(x_t, v_t, \gamma)\right\|^2 \\
\leq & F_\delta(x_t) + \frac{\gamma}{2}\left\|\nabla F_\delta(x_t) - v_t\right\|^2 - \gamma\langle v_t, \mathcal{G}(x_t, v_t, \gamma)\rangle + \left(\frac{c\gamma^2 G\sqrt{d}}{2\delta} + \frac{\gamma}{2}\right)\left\|\mathcal{G}(x_t, v_t, \gamma)\right\|^2 \\
\leq & F_\delta(x_t) + \frac{\gamma}{2}\left\|\nabla F_\delta(x_t) - v_t\right\|^2 + \left(\frac{c\gamma^2 G\sqrt{d}}{2\delta} - \frac{\gamma}{2}\right)\left\|\mathcal{G}(x_t, v_t, \gamma)\right\|^2.
\end{aligned}
$$

The second inequality follows from $x_{t+1} = x_t - \gamma\mathcal{G}(x_t, v_t, \gamma)$. The third inequality is due to Young's inequality. The last inequality follows from Lemma C.2.

Denote $n_t = \lfloor t/q \rfloor$. Taking expectations on both sides, we obtain

$$
\begin{aligned}
\mathbb{E}[F_\delta(x_{t+1})] \leq & \mathbb{E}[F_\delta(x_t)] + \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla F_\delta(x_t) - v_t\right\|^2\right] - \left(\frac{\gamma}{2} - \frac{c\gamma^2 G\sqrt{d}}{2\delta}\right)\mathbb{E}\left[\left\|\mathcal{G}(x_t, v_t, \gamma)\right\|^2\right] \\
\leq & \mathbb{E}[F_\delta(x_t)] + \frac{\gamma d^2 G^2}{2\delta^2 b_2}\sum_{i=n_t q}^{t-1}\mathbb{E}\left[\left\|x_{i+1} - x_i\right\|^2\right] + \frac{8\sqrt{2\pi}d\gamma G^2}{b_1} - \left(\frac{\gamma}{2} - \frac{c\gamma^2 G\sqrt{d}}{2\delta}\right)\mathbb{E}\left[\left\|\mathcal{G}(x_t, v_t, \gamma)\right\|^2\right] \\
\leq & \mathbb{E}[F_\delta(x_t)] + \frac{\gamma^3 d^2 G^2}{2\delta^2 b_2}\sum_{i=n_t q}^{t-1}\mathbb{E}\left[\left\|\mathcal{G}(x_i, v_i, \gamma)\right\|^2\right] + \frac{8\sqrt{2\pi}d\gamma G^2}{b_1} - \left(\frac{\gamma}{2} - \frac{c\gamma^2 G\sqrt{d}}{2\delta}\right)\mathbb{E}\left[\left\|\mathcal{G}(x_t, v_t, \gamma)\right\|^2\right].
\end{aligned}
$$

The second inequality is due to Lemma A.5. The last inequality follows from $x_{i+1} = x_i - \gamma\mathcal{G}(x_i, v_i, \gamma)$. Telescoping the above inequality over t from $n_t q$ to t where $t \leq (n_t + 1)q - 1$, we have

$$
\begin{aligned}
& \mathbb{E}\left[F_\delta(x_{t+1})\right] - \mathbb{E}\left[F_\delta(x_{n_t q})\right] \\
\leq & \frac{\gamma^3 d^2 G^2}{2\delta^2 b_2}\sum_{j=n_t q}^{t}\sum_{i=n_t q}^{j-1}\mathbb{E}\left[\left\|\mathcal{G}(x_i, v_i, \gamma)\right\|^2\right] + \sum_{j=n_t q}^{t}\frac{8\sqrt{2\pi}d\gamma G^2}{b_1} \\
& - \left(\frac{\gamma}{2} - \frac{c\gamma^2 G\sqrt{d}}{2\delta}\right)\sum_{j=n_t q}^{t}\mathbb{E}\left[\left\|\mathcal{G}(x_j, v_j, \gamma)\right\|^2\right] \\
\leq & \frac{\gamma^3 d^2 G^2}{2\delta^2 b_2}\sum_{j=n_t q}^{t}\sum_{i=n_t q}^{t}\mathbb{E}\left[\left\|\mathcal{G}(x_i, v_i, \gamma)\right\|^2\right] + \sum_{j=n_t q}^{t}\frac{8\sqrt{2\pi}d\gamma G^2}{b_1} \\
& - \left(\frac{\gamma}{2} - \frac{c\gamma^2 G\sqrt{d}}{2\delta}\right)\sum_{j=n_t q}^{t}\mathbb{E}\left[\left\|\mathcal{G}(x_j, v_j, \gamma)\right\|^2\right] \\
\leq & \frac{\gamma^3 d^2 G^2 q}{2\delta^2 b_2}\sum_{j=n_t q}^{t}\mathbb{E}\left[\left\|\mathcal{G}(x_j, v_j, \gamma)\right\|^2\right] + \sum_{j=n_t q}^{t}\frac{8\sqrt{2\pi}d\gamma G^2}{b_1} \\
& - \left(\frac{\gamma}{2} - \frac{c\gamma^2 G\sqrt{d}}{2\delta}\right)\sum_{j=n_t q}^{t}\mathbb{E}\left[\left\|\mathcal{G}(x_j, v_j, \gamma)\right\|^2\right] \\
= & -\left(\frac{\gamma}{2} - \frac{c\gamma^2 G\sqrt{d}}{2\delta} - \frac{\gamma^3 d^2 G^2 q}{2\delta^2 b_2}\right)\sum_{j=n_t q}^{t}\mathbb{E}\left[\left\|\mathcal{G}(x_j, v_j, \gamma)\right\|^2\right] \\
& + \sum_{j=n_t q}^{t}\frac{8\sqrt{2\pi}d\gamma G^2}{b_1}.
\end{aligned}
$$

The third inequality uses the fact that $n_t q \leq t \leq (n_t + 1)q$. Thus, we have $t - n_t q + 1 \leq q$. Now if we sum up the above inequality over all epochs, we obtain

$$
\mathbb{E}[F_\delta(x_T)] - \mathbb{E}[F_\delta(x_0]
$$

$$
\begin{aligned}
=&(\mathbb{E}[F_\delta(x_q)] - \mathbb{E}[F_\delta(x_0)]) + (\mathbb{E}[F_\delta(x_{2q})] - \mathbb{E}[F_\delta(x_q)]) + \cdots + (\mathbb{E}[F_\delta(x_T)] - \mathbb{E}[F_\delta(x_{n_T q})]) \\
\leq& -\left(\frac{\gamma}{2} - \frac{c\gamma^2 G\sqrt{d}}{2\delta} - \frac{\gamma^3 d^2 G^2 q}{2\delta^2 b_2}\right) \sum_{t=0}^{T-1} \mathbb{E}\left[\|\mathcal{G}(x_t, v_t, \gamma)\|^2\right] + \sum_{t=0}^{T-1} \frac{8\sqrt{2\pi} d\gamma G^2}{b_1} \\
=& -\left(\frac{\gamma}{2} - \frac{c\gamma^2 G\sqrt{d}}{2\delta} - \frac{\gamma^3 d^2 G^2 q}{2\delta^2 b_2}\right) \sum_{t=0}^{T-1} \mathbb{E}\left[\|\mathcal{G}(x_t, v_t, \gamma)\|^2\right] + \frac{8\sqrt{2\pi} d\gamma G^2 T}{b_1}.
\end{aligned}
$$

Rearrange the terms, and divide both sides by $(\frac{\gamma}{2} - \frac{c\gamma^2 G\sqrt{d}}{2\delta} - \frac{\gamma^3 d^2 G^2 q}{2\delta^2 b_2})T$, we can obtain

$$
\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\mathcal{G}(x_t, v_t, \gamma)\|^2\right] \leq \left(\frac{\gamma}{2} - \frac{c\gamma^2 G\sqrt{d}}{2\delta} - \frac{\gamma^3 d^2 G^2 q}{2\delta^2 b_2}\right)^{-1} \left(\frac{\mathbb{E}[F_\delta(x_0)] - \mathbb{E}[F_\delta(x_T)]}{T} + \frac{8\sqrt{2\pi} d\gamma G^2}{b_1}\right).
$$

Since R is chosen uniformly from $0, 1, \ldots, T-1$, it completes the proof. □

C.2.1. Proof of Theorem 5.3

To bound $\mathbb{E}\left[\|\mathcal{G}(x_R, \nabla F_\delta(x_R), \gamma)\|^2\right]$, one has

$$
\begin{aligned}
\mathbb{E}\left[\|\mathcal{G}(x_R, \nabla F_\delta(x_R), \gamma)\|^2\right] \leq& 2\mathbb{E}\left[\|\mathcal{G}(x_R, v_R, \gamma)\|^2\right] + 2\mathbb{E}\left[\|\mathcal{G}(x_R, v_R, \gamma) - \mathcal{G}(x_R, \nabla F_\delta(x_R), \gamma)\|^2\right] \\
\leq& 2\mathbb{E}\left[\|\mathcal{G}(x_R, v_R, \gamma)\|^2\right] + 2\mathbb{E}\left[\|v_R - \nabla F_\delta(x_R)\|^2\right].
\end{aligned}
$$

The first inequality is due to $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for $\forall a, b \in \mathbb{R}^d$. We use Lemma C.1 for the second inequality. To bound the second term on the right-hand side, one has

$$
\begin{aligned}
\mathbb{E}\left[\|v_R - \nabla F_\delta(x_R)\|^2\right] \leq& \frac{d^2 G^2}{\delta^2 b_2} \mathbb{E}\left[\sum_{j=n_R q+1}^{R} \|x_{j+1} - x_j\|^2\right] + \frac{16\sqrt{2\pi} d G^2}{b_1} \\
=& \frac{d^2 G^2 \gamma^2}{\delta^2 b_2} \mathbb{E}\left[\sum_{j=n_R q+1}^{R} \|\mathcal{G}(x_j, v_j, \gamma)\|^2\right] + \frac{16\sqrt{2\pi} d G^2}{b_1} \\
\leq& \frac{d^2 G^2 \gamma^2 q}{\delta^2 b_2 T} \mathbb{E}\left[\sum_{j=0}^{T-1} \|\mathcal{G}(x_j, v_j, \gamma)\|^2\right] + \frac{16\sqrt{2\pi} d G^2}{b_1}.
\end{aligned}
$$

The first inequality is due to Lemma A.5. The first equality follows from $x_{j+1} = x_j - \gamma\mathcal{G}(x_j, v_j, \gamma)$. The second inequality holds because each term $j \in [T]$ is chosen with a probability less than q/T. Therefore, one has

$$
\mathbb{E}\left[\|\mathcal{G}(x_R, \nabla F_\delta(x_R), \gamma)\|^2\right] \leq 2\left(\frac{d^2 G^2 \gamma^2 q}{\delta^2 b_2 T} + \frac{1}{T}\right) \mathbb{E}\left[\sum_{j=0}^{T-1} \|\mathcal{G}(x_j, v_j, \gamma)\|^2\right] + \frac{32\sqrt{2\pi} d G^2}{b_1}.
$$

If we choose $b_2 = q$, and we can show that

$$
\frac{\gamma}{2} - \frac{c\gamma^2 G\sqrt{d}}{2\delta} - \frac{\gamma^3 d^2 G^2 q}{2\delta^2 b_2} = \frac{\gamma}{2}\left(1 - \frac{c\gamma G\sqrt{d}}{\delta} - \frac{\gamma^2 d^2 G^2}{\delta^2}\right).
$$

If we further choose $\gamma = \frac{\delta}{2dG}$ and assume $d \geq 4c^2$, then we have

$$
\frac{\gamma}{2} - \frac{c\gamma^2 G\sqrt{d}}{2\delta} - \frac{\gamma^3 d^2 G^2 q}{2\delta^2 b_2} = \frac{\delta}{4dG}\left(\frac{3}{4} - \frac{c}{2\sqrt{d}}\right) \geq \frac{\delta}{8dG}.
$$

Now we can bound $\mathbb{E}\left[\|\mathcal{G}(x_R, \nabla F_\delta(x_R), \gamma)\|^2\right]$ with

$$\begin{aligned}\mathbb{E}\left[\|\mathcal{G}(x_R, \nabla F_\delta(x_R), \gamma)\|^2\right] \leq & \frac{5}{2T}\mathbb{E}\left[\sum_{j=0}^{T-1}\|\mathcal{G}(x_j, v_j, \gamma)\|^2\right] + \frac{32\sqrt{2\pi}dG^2}{b_1} \\ \leq & \frac{20dG}{\delta}\left(\frac{\mathbb{E}[F_\delta(x_0)] - \mathbb{E}[F_\delta(x_T)]}{T} + \frac{4\sqrt{2\pi}G\delta}{b_1}\right) + \frac{32\sqrt{2\pi}dG^2}{b_1} \\ \leq & \frac{20dG}{\delta}\left(\frac{GB}{T} + \frac{4\sqrt{2\pi}G\delta}{b_1}\right) + \frac{32\sqrt{2\pi}dG^2}{b_1} \\ \leq & \frac{20dG^2B}{\delta T} + \frac{192\sqrt{2\pi}dG^2}{b_1}.\end{aligned}$$

The second inequality follows from Lemma C.5. The third inequality is due to the assumption that the function $F_\delta(\cdot)$ is G-Lipschitz and the diameter of the feasible set Ω is bounded by B. It further implies that

$$\mathbb{E}\left[\min\{\|\mathcal{G}(x_R, g, \gamma)\| : g \in \partial_\delta F(x_R)\}\right] \leq \mathbb{E}\left[\|\mathcal{G}(x_R, \nabla F_\delta(x_R), \gamma)\|\right] \leq \frac{5\sqrt{dB}G}{\sqrt{\delta T}} + \frac{16\sqrt{d}G}{\sqrt{b_1}}.$$

C.2.2. Proof of Corollary 5.4

To obtain an ϵ-approximate solution, we choose $T = \mathcal{O}(\frac{dBG^2}{\delta\epsilon^2})$, $b_1 = \mathcal{O}(\frac{dG^2}{\epsilon^2})$ and $b_2 = q = \mathcal{O}(\frac{\sqrt{d}G}{\epsilon})$. The total number of the function value oracle calls is bounded by

$$\mathcal{O}(b_1 T/q + b_2 T) = \mathcal{O}\left(\frac{d^{\frac{3}{2}}G^3B}{\delta\epsilon^3}\right).$$

C.2.3. Large-deviation Estimation of The Two-Phase VR-ZOSPGD Method

Similar to the result of Theorem C.4 for the two-phase MB-ZOSPGD method, we present the large-deviation estimation for the two-phase VR-ZOSPGD method as follows.

Theorem C.6. *Let $\delta > 0$ and $0 < \epsilon, \Lambda < 1$, then there exists some $T, S, \hat{B} > 0$ such that the output $\bar{x}^{s^*}$ of Algorithm 5 with the VR-ZOSPGD satisfies that* $\mathrm{Prob}\left(\min\{\left\|\mathcal{G}(\bar{x}^{s^*}, g, \gamma)\right\| \geq \epsilon : g \in \partial_\delta F(\bar{x}^{s^*})\}\right) \leq \Lambda$ *and the total number of calls of the FQO is bounded by*

$$\mathcal{O}\left(\frac{d^{\frac{3}{2}}BG^3}{\delta\epsilon^3}\log\left(\frac{1}{\Lambda}\right) + \frac{dG^2}{\epsilon^2\Lambda}\log\left(\frac{1}{\Lambda}\right)^2\right).$$

Proof. By the definition of s^* and the Cauchy–Schwarz inequality, we have

$$\begin{aligned}\left\|\mathcal{G}(\bar{x}^{s^*}, g^{s^*}, \gamma)\right\|^2 = & \min_{s\in\{0,1,\ldots,S-1\}}\|\mathcal{G}(\bar{x}^s, g^s, \gamma)\|^2 \leq \min_{s\in\{0,1,\ldots,S-1\}}\left(2\|\mathcal{G}(\bar{x}^s, \nabla F_\delta(\bar{x}^s), \gamma)\|^2 + 2\|g^s - \nabla F_\delta(\bar{x}^s)\|^2\right) \\ \leq & 2\min_{s\in\{0,1,\ldots,S-1\}}\|\mathcal{G}(\bar{x}^s, \nabla F_\delta(\bar{x}^s), \gamma)\|^2 + 2\max_{s\in\{0,1,\ldots,S-1\}}\|g^s - \nabla F_\delta(\bar{x}^s)\|^2.\end{aligned}$$

The first inequality is due to Lemma C.1 and the fact that $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. It implies that

$$\begin{aligned}\left\|\mathcal{G}(\bar{x}^{s^*}, \nabla F_\delta(\bar{x}^{s^*}), \gamma)\right\|^2 \leq & 2\left\|\mathcal{G}(\bar{x}^{s^*}, g^{s^*}, \gamma)\right\| + 2\left\|g^{s^*} - \nabla F_\delta(\bar{x}^{s^*})\right\|^2 \\ \leq & 4\min_{s\in\{0,1,\ldots,S-1\}}\|\mathcal{G}(\bar{x}^s, \nabla F_\delta(\bar{x}^s), \gamma)\|^2 + 4\max_{s\in\{0,1,\ldots,S-1\}}\|g^s - \nabla F_\delta(\bar{x}^s)\|^2 \\ & + 2\left\|g^{s^*} - \nabla F_\delta(\bar{x}^{s^*})\right\|^2.\end{aligned}$$

The first inequality follows from Cauchy–Schwarz inequality and Lemma C.1. The next step is to provide the probabilistic bounds on all the terms on the right-hand side of the above inequality. Theorem 5.3 implies that

$$\mathbb{E}\left[\left\|\mathcal{G}(\bar{x}^s, \nabla F_\delta(\bar{x}^s), \gamma)\right\|^2\right] \leq \frac{20dG^2B}{\delta T} + \frac{192\sqrt{2\pi}dG^2}{b_1}.$$

Using Markov's inequality, we have

$$\text{Prob}\left(\left\|\mathcal{G}(\bar{x}^s, \nabla F_\delta(\bar{x}^s), \gamma)\right\|^2 \geq \frac{40dG^2B}{\delta T} + \frac{384\sqrt{2\pi}dG^2}{b_1}\right) \leq \frac{1}{2}.$$

Thus, we have

$$\text{Prob}\left(\min_{s\in\{0,1,\ldots,S-1\}}\left\|\mathcal{G}(\bar{x}^s, \nabla F_\delta(\bar{x}^s), \gamma)\right\|^2 \geq \frac{40dG^2B}{\delta T} + \frac{384\sqrt{2\pi}dG^2}{b_1}\right) \leq \frac{1}{2^S}.$$

Furthermore, for each $s \in \{0, 1, \ldots, S-1\}$, we have

$$g^s - \nabla F_\delta(\bar{x}^s) = \frac{1}{\hat{B}}\sum_{k=0}^{\hat{B}-1}(g_k^s - \nabla F_\delta(\bar{x}^s)).$$

By Lemma A.1, we have $\mathbb{E}\left[g_k^s \mid \bar{x}^s\right] = \nabla F_\delta(\bar{x}^s)$ and $\mathbb{E}\left[\left\|g_k^s\right\|^2 \mid \bar{x}^s\right] \leq 16\sqrt{2\pi}dG^2$. By Markov's inequality and Proposition A.2, it implies that

$$\text{Prob}\left(\left\|g^s - \nabla F_\delta(\bar{x}^s)\right\|^2 \geq \frac{\lambda(16\sqrt{2\pi}dG^2)}{\hat{B}}\right) = \text{Prob}\left(\left\|\sum_{k=0}^{\hat{B}-1}(g_k^s - \nabla F_\delta(\bar{x}^s))\right\|^2 \geq \lambda\hat{B}(16\sqrt{2\pi}dG^2)\right) \leq \frac{1}{\lambda}.$$

Thus, we conclude that

$$\text{Prob}\left(\max_{s\in\{0,1,\ldots,S-1\}}\left\|g^s - \nabla F_\delta(\bar{x}^s)\right\|^2 \geq \frac{\lambda(16\sqrt{2\pi}dG^2)}{\hat{B}}\right) \leq \frac{S}{\lambda}.$$

By a similar argument, one has

$$\text{Prob}\left(\left\|g^{s^*} - \nabla F_\delta(\bar{x}^{s^*})\right\|^2 \geq \frac{\lambda(16\sqrt{2\pi}dG^2)}{\hat{B}}\right) \leq \frac{1}{\lambda}.$$

Combining the above inequalities yields that

$$\text{Prob}\left(\left\|\mathcal{G}(x^{s^*}, \nabla F_\delta(x^{s^*}), \gamma)\right\|^2 \geq \frac{160dG^2B}{\delta T} + \frac{1536\sqrt{2\pi}dG^2}{b_1} + \frac{\lambda 96\sqrt{2\pi}dG^2}{\hat{B}}\right) \leq \frac{S+1}{\lambda} + \frac{1}{2^S}.$$

If we set $\lambda = \frac{2(S+1)}{\Lambda}$, $S = \lceil\log(2/\Lambda)\rceil$ and the parameters $(T, b_1, b_2, q, \hat{B})$ as follows

$$T = \mathcal{O}\left(\frac{dBG^2}{\delta\epsilon^2}\right), \quad b_1 = \mathcal{O}\left(\frac{dG^2}{\epsilon^2}\right), \quad b_2 = q = \mathcal{O}\left(\frac{\sqrt{d}G}{\epsilon}\right), \quad \hat{B} = \mathcal{O}\left(\frac{dG^2}{\epsilon^2\Lambda}\log\left(\frac{1}{\Lambda}\right)\right).$$

To satisfy $\text{Prob}\left(\min\{\left\|\mathcal{G}(\bar{x}^{s^*}, g, \gamma)\right\| \geq \epsilon\colon g \in \partial_\delta F(\bar{x}^{s^*})\}\right) \leq \Lambda$, the total number of function oracle calls is bounded by

$$S(T(b_1/q + b_2) + \hat{B}) = \mathcal{O}\left(\frac{d^{\frac{3}{2}}BG^3}{\delta\epsilon^3}\log\left(\frac{1}{\Lambda}\right) + \frac{dG^2}{\epsilon^2\Lambda}\log\left(\frac{1}{\Lambda}\right)^2\right).$$

□

Algorithm 6 Two-Phase Zeroth-Order Stochastic Frank–Wolfe Method

Input: Initial point $x_0 \in \mathbb{R}^d$, sequence of stepsizes $\{\gamma_t : \gamma_t > 0\}_{t=0}^{T-1}$, problem dimension $d \geq 1$, smoothing parameter δ, iteration number $T \geq 1$, number of rounds $S \geq 1$ and sample size $\hat{B}$.
for $s = 0, 1, \ldots, S-1$ **do**
  Call Algorithm 4 with x_0, $\{\gamma_t\}_{t=0}^{T-1}$, d, δ and T and let $\bar{x}^s$ be an output.
end for
for $s = 0, 1, \ldots, S-1$ **do**
  for $k = 0, 1, \ldots, \hat{B}-1$ **do**
    Simulate $\xi_k \sim \mathcal{P}$
    Sample $w_k \in \mathbb{R}^d$ uniformly from a unit sphere in $\mathbb{R}^d$.
    $g_k^s = \frac{d}{2\delta}(f(\bar{x}^s + \delta w_k, \xi_k) - f(\bar{x}^s - \delta w_k, \xi_k))w_k$
  end for
  $g^s = \frac{1}{\hat{B}} \sum_{k=0}^{\hat{B}-1} g_k^s$
  $u^s = \arg\max_{u \in \Omega} \langle u, -g^s \rangle$
end for
Choose $s^* = \arg\min_{s \in \{0,1,\ldots,S-1\}} \langle -g^s, u^s - \bar{x}^s \rangle$
return $\bar{x}^{s^*}$.

D. Convergence Analysis of ZOSFW Methods

In this section, we present the analysis of the convergence rate of both MB-ZOSFW and VR-ZOSFW methods. In addition, we show the large-deviation estimation results of the proposed projection-free stochastic optimization methods with the two-phase post-processing technique.

D.1. Convergence Analysis of the MB-ZOSFW method

In this subsection, we show the convergence rate of the MB-ZOSFW method as follows.

D.1.1. Proof of Theorem 5.6

We define $\tilde{u}_t = \arg\max_{u \in \Omega} \langle u, -\nabla F_\delta(x_t) \rangle$, it follows that:

$$\begin{aligned}
&\langle \nabla F_\delta(x_t), u_t - x_t \rangle \\
=&\langle \nabla F_\delta(x_t), \tilde{u}_t - x_t \rangle + \langle \nabla F_\delta(x_t), u_t - \tilde{u}_t \rangle \\
=&\langle \nabla F_\delta(x_t), \tilde{u}_t - x_t \rangle + \langle v_t, u_t - \tilde{u}_t \rangle + \langle \nabla F_\delta(x_t) - v_t, u_t - \tilde{u}_t \rangle \\
\leq&\langle \nabla F_\delta(x_t), \tilde{u}_t - x_t \rangle + \|\nabla F_\delta(x_t) - v_t\| \, \|u_t - \tilde{u}_t\| .
\end{aligned}$$

The first inequality is due to $\langle v_t, u_t \rangle \leq \langle v_t, \tilde{u}_t \rangle$ by the optimality of u_t.

$$\begin{aligned}
&F_\delta(x_{t+1}) \\
\leq&F_\delta(x_t) + \gamma_t \langle \nabla F_\delta(x_t), u_t - x_t \rangle + \frac{cG\sqrt{d}\gamma_t^2}{2\delta} \|u_t - x_t\|^2 \\
\leq&F_\delta(x_t) + \gamma_t \langle \nabla F_\delta(x_t), \tilde{u}_t - x_t \rangle + \frac{cG\sqrt{d}\gamma_t^2}{2\delta} \|u_t - x_t\|^2 + \gamma_t \|\nabla F_\delta(x_t) - v_t\| \, \|u_t - \tilde{u}_t\| \\
\leq&F_\delta(x_t) + \gamma_t \langle \nabla F_\delta(x_t), \tilde{u}_t - x_t \rangle + \frac{cG\sqrt{d}\gamma_t^2 B^2}{2\delta} + \gamma_t B \|\nabla F_\delta(x_t) - v_t\| .
\end{aligned}$$

The first inequality follows from Lemma 3.8 such that F_δ is a $cG\sqrt{d}/\delta$-smooth objective function. The last inequality is due to the assumption that the diameter of the domain is bounded by B. Taking expectations on both sides of the inequality, we have

$$\begin{aligned}
&\mathbb{E}[F_\delta(x_{t+1})] \\
\leq&\mathbb{E}[F_\delta(x_t)] + \gamma_t \mathbb{E}[\langle \nabla F_\delta(x_t), \tilde{u}_t - x_t \rangle] + \frac{cG\sqrt{d}\gamma_t^2 B^2}{2\delta} + \gamma_t B \mathbb{E}[\|\nabla F_\delta(x_t) - v_t\|]
\end{aligned}$$

$$\leq \mathbb{E}[F_\delta(x_t)] + \gamma_t \mathbb{E}[\langle \nabla F_\delta(x_t), \tilde{u}_t - x_t\rangle] + \frac{cG\sqrt{d}\gamma_t^2 B^2}{2\delta} + \frac{8\gamma_t BG\sqrt{d}}{\sqrt{b_t}}.$$

The last inequality follows from Lemma A.1 and the fact $\mathbb{E}[\|\zeta\|]^2 \leq \mathbb{E}[\|\zeta\|^2]$ for any random variable ζ. Fix $\gamma_t = \gamma$, and telescope the above results through $t = 0$ to $T-1$, we have

$$\gamma \sum_{t=0}^{T-1} \mathbb{E}[\langle -\nabla F_\delta(x_t), \tilde{u}_t - x_t\rangle] \leq \mathbb{E}[F_\delta(x_0)] - \mathbb{E}[F_\delta(x_T)] + \frac{cG\sqrt{d}\gamma^2 B^2 T}{2\delta} + \sum_{t=0}^{T-1} \frac{8\gamma BG\sqrt{d}}{\sqrt{b_t}}.$$

Divide both sides by γT, we can obtain:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\langle -\nabla F_\delta(x_t), \tilde{u}_t - x_t\rangle] \\
\leq &\frac{\mathbb{E}[F_\delta(x_0)] - \mathbb{E}[F_\delta(x_T)]}{\gamma T} + \frac{cG\sqrt{d}\gamma B^2}{2\delta} + 8BG\sqrt{d}\frac{\sum_{t=0}^{T-1} 1/\sqrt{b_t}}{T} \\
\leq &\frac{GB}{\gamma T} + \frac{cG\sqrt{d}\gamma B^2}{2\delta} + 8BG\sqrt{d}\frac{\sum_{t=0}^{T-1} 1/\sqrt{b_t}}{T}.
\end{aligned}$$

The last inequality follows from the assumption that the function $F_\delta(\cdot)$ is G-Lipschitz and the diameter of the domain Ω is bound by B. Set $b_t = b$ and $\gamma = \sqrt{\delta}/(\sqrt{TB}d^{\frac{1}{4}})$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\langle -\nabla F_\delta(x_t), \tilde{u}_t - x_t\rangle] \leq \frac{2\sqrt{c}GB^{\frac{3}{2}}d^{\frac{1}{4}}}{\delta^{\frac{1}{2}}T^{\frac{1}{2}}} + \frac{8BG\sqrt{d}}{\sqrt{b}}.$$

D.1.2. Proof of Corollary 5.7

If we choose $T = \mathcal{O}(B^3G^2d^{\frac{1}{2}}\delta^{-1}\epsilon^{-2})$ and $b = \mathcal{O}(B^2G^2d\epsilon^{-2})$ for Theorem 5.6, then the total function oracle call is $Tb = \mathcal{O}(B^5G^4d^{\frac{3}{2}}\delta^{-1}\epsilon^{-4})$.

D.1.3. Large-deviation Estimation of the Two-Phase MB-ZOSFW method

Similar to the large-deviation estimation of the two-phase MB-ZOSPGD method, we combine Algorithm 4 with a post-optimization procedure, leading to a two-phase zeroth-order stochastic Frank–Wolfe method (2-ZOSFW) that is shown in Algorithm 6. Formally, we provide the large-deviation estimation of the two-phase MB-ZOSFW method as follows.

Theorem D.1. *Let $\delta > 0$ and $0 < \epsilon, \Lambda < 1$, then there exists some $T, S, \hat{B} > 0$ such that the output $\bar{x}^{s^*}$ of Algorithm 6 with the MB-ZOSFW satisfies that* $\mathrm{Prob}(\min\{\max_{u\in\Omega}\langle -g, u - \bar{x}^{s^*}\rangle \geq \epsilon \colon g \in \partial_\delta F(\bar{x}^{s^*})\}) \leq \Lambda$ *and the total number of calls of the FQO is bounded by*

$$\mathcal{O}\left(\frac{B^5G^4d^{\frac{3}{2}}}{\delta\epsilon^4}\log\left(\frac{1}{\Lambda}\right) + \frac{dG^2B^2}{\epsilon^2\Lambda}\log\left(\frac{1}{\Lambda}\right)^2\right).$$

Proof. We define $\tilde{u}^s = \arg\max_{u\in\Omega}\langle u, -\nabla F_\delta(\bar{x}^s)\rangle$, it follows that

$$\begin{aligned}
\langle -\nabla F_\delta(\bar{x}^{s^*}), \tilde{u}^{s^*} - \bar{x}^{s^*}\rangle =&\langle -g^{s^*}, \tilde{u}^{s^*} - \bar{x}^{s^*}\rangle + \langle g^{s^*} - \nabla F_\delta(\bar{x}^{s^*}), \tilde{u}^{s^*} - \bar{x}^{s^*}\rangle \\
\leq&\langle -g^{s^*}, u^{s^*} - \bar{x}^{s^*}\rangle + \langle -g^{s^*}, \tilde{u}^{s^*} - u^{s^*}\rangle + \left\|g^{s^*} - \nabla F_\delta(\bar{x}^{s^*})\right\| \left\|\tilde{u}^{s^*} - \bar{x}^{s^*}\right\| \\
\leq&\langle -g^{s^*}, u^{s^*} - \bar{x}^{s^*}\rangle + \left\|g^{s^*} - \nabla F_\delta(\bar{x}^{s^*})\right\| \left\|\tilde{u}^{s^*} - \bar{x}^{s^*}\right\|.
\end{aligned}$$

The first inequality follows from Cauchy–Schwarz inequality. The last inequality is due to $u^{s^*} = \arg\max_{u\in\Omega}\langle -g^{s^*}, u\rangle$. The first term on the right-hand side of the inequality can be bound by

$$\langle -g^{s^*}, u^{s^*} - \bar{x}^{s^*}\rangle = \min_{s\in\{0,1,\ldots,S-1\}}\langle -g^s, u^s - \bar{x}^s\rangle$$

$$
\begin{aligned}
&= \min_{s\in\{0,1,\ldots,S-1\}} \left(\langle -\nabla F_\delta(x^s), u^s - \bar{x}^s\rangle + \langle \nabla F_\delta(x^s) - g^s, u^s - \bar{x}^s\rangle\right)\\
&\le \min_{s\in\{0,1,\ldots,S-1\}} \left(\langle -\nabla F_\delta(x^s), \tilde{u}^s - \bar{x}^s\rangle + \langle \nabla F_\delta(x^s) - g^s, u^s - \bar{x}^s\rangle\right)\\
&\le \min_{s\in\{0,1,\ldots,S-1\}} \langle -\nabla F_\delta(x^s), \tilde{u}^s - \bar{x}^s\rangle + \max_{s\in\{0,1,\ldots,S-1\}} \langle \nabla F_\delta(x^s) - g^s, u^s - \bar{x}^s\rangle\\
&\le \min_{s\in\{0,1,\ldots,S-1\}} \langle -\nabla F_\delta(x^s), \tilde{u}^s - \bar{x}^s\rangle + \max_{s\in\{0,1,\ldots,S-1\}} \|\nabla F_\delta(x^s) - g^s\| \, \|u^s - \bar{x}^s\| .
\end{aligned}
$$

The first inequality follows from $\tilde{u}^s = \arg\max_{u\in\Omega}\langle u, -\nabla F_\delta(\bar{x}^s)\rangle$, and the last inequality is due to Cauchy–Schwarz inequality. The proof in Theorem 5.6 implies that

$$
\mathbb{E}[\langle -\nabla F_\delta(x^s), \tilde{u}^s - x^s\rangle] \le \frac{GB^{\frac{3}{2}}d^{\frac{1}{4}}}{\delta^{\frac{1}{2}}T^{\frac{1}{2}}} + \frac{8BG\sqrt{d}}{\sqrt{b}}.
$$

Using Markov's inequality, we have

$$
\text{Prob}\left(\langle -\nabla F_\delta(x^s), \tilde{u}^s - x^s\rangle \ge \frac{2GB^{\frac{3}{2}}d^{\frac{1}{4}}}{\delta^{\frac{1}{2}}T^{\frac{1}{2}}} + \frac{16BG\sqrt{d}}{\sqrt{b}}\right) \le \frac{1}{2}.
$$

Therefore, we can deduce that

$$
\text{Prob}\left(\min_{s\in\{0,1,\ldots,S-1\}}\langle -\nabla F_\delta(x^s), \tilde{u}^s - \bar{x}^s\rangle \ge \frac{2GB^{\frac{3}{2}}d^{\frac{1}{4}}}{\delta^{\frac{1}{2}}T^{\frac{1}{2}}} + \frac{16BG\sqrt{d}}{\sqrt{b}}\right) \le \frac{1}{2^S}.
$$

By Lemma A.1, we have $\mathbb{E}[g_k^s \mid \bar{x}^s] = \nabla F_\delta(\bar{x}^s)$ and $\mathbb{E}\left[\|g_k^s\|^2 \mid \bar{x}^s\right] \le 16\sqrt{2\pi}dG^2$. By Markov's inequality and Proposition A.2, it yields that

$$
\begin{aligned}
&\text{Prob}\left(\|g^s - \nabla F_\delta(\bar{x}^s)\| \, \|u^s - \bar{x}^s\| \ge \frac{\sqrt{\lambda}(8\sqrt{d}GB)}{\sqrt{\hat{B}}}\right)\\
=&\text{Prob}\left(\|g^s - \nabla F_\delta(\bar{x}^s)\|^2 \, \|u^s - \bar{x}^s\|^2 \ge \frac{\lambda(16\sqrt{2\pi}dG^2B^2)}{\hat{B}}\right) \le \frac{1}{\lambda}.
\end{aligned}
$$

Therefore, we can conclude that

$$
\text{Prob}\left(\max_{s\in[S-1]}\|g^s - \nabla F_\delta(\bar{x}^s)\| \, \|u^s - \bar{x}^s\| \ge \frac{\sqrt{\lambda}(8\sqrt{d}GB)}{\sqrt{\hat{B}}}\right) \le \frac{S}{\lambda}.
$$

Using a similar argument, one has

$$
\text{Prob}\left(\left\|g^{s^*} - \nabla F_\delta(\bar{x}^{s^*})\right\| \left\|\tilde{u}^{s^*} - \bar{x}^{s^*}\right\| \ge \frac{\sqrt{\lambda}(8\sqrt{d}GB)}{\sqrt{\hat{B}}}\right) \le \frac{1}{\lambda}.
$$

Combining the above inequalities, for all $\lambda > 0$ we have

$$
\text{Prob}\left(\langle -\nabla F_\delta(\bar{x}^{s^*}), \tilde{u}^{s^*} - \bar{x}^{s^*}\rangle \ge \frac{2GB^{\frac{3}{2}}d^{\frac{1}{4}}}{\delta^{\frac{1}{2}}T^{\frac{1}{2}}} + \frac{16BG\sqrt{d}}{\sqrt{b}} + \frac{\sqrt{\lambda}(16\sqrt{d}GB)}{\sqrt{\hat{B}}}\right) \le \frac{S+1}{\lambda} + \frac{1}{2^S}.
$$

We set $\lambda = \frac{2(S+1)}{\Lambda}$, $S = \lceil \log\frac{2}{\Lambda}\rceil$, and the parameters $(T, b, \hat{B})$ as follows

$$
T = \mathcal{O}\left(\frac{B^3G^2d^{\frac{1}{2}}}{\delta\epsilon^2}\right), \quad b = \mathcal{O}\left(\frac{B^2G^2d}{\epsilon^2}\right), \quad \hat{B} = \mathcal{O}\left(\frac{dG^2B^2(S+1)}{\epsilon^2\Lambda}\right).
$$

Therefore, we have

$$
\text{Prob}\left(\langle -\nabla F_\delta(\bar{x}^{s^*}), \tilde{u}^{s^*} - \bar{x}^{s^*}\rangle \ge \epsilon\right) \le \Lambda.
$$

The total number of function oracle calls is therefore bounded by

$$S(Tb+\hat{B}) = \mathcal{O}\left(\frac{B^5G^4d^{\frac{3}{2}}}{\delta\epsilon^4}\log\left(\frac{1}{\Lambda}\right) + \frac{dG^2B^2}{\epsilon^2\Lambda}\log\left(\frac{1}{\Lambda}\right)^2\right).$$

□

D.2. Convergence Analysis of the VR-ZOSFW Method

In this subsection, we provide the convergence analysis of the VR-ZOSFW method.

D.2.1. Proof of Theorem 5.8

We define $\tilde{u}_t = \arg\max_{u\in\Omega}\langle u, -\nabla F_\delta(x_t)\rangle$, it follows that:

$$\begin{aligned}
&\langle \nabla F_\delta(x_t), u_t - x_t\rangle \\
=&\langle \nabla F_\delta(x_t), \tilde{u}_t - x_t\rangle + \langle \nabla F_\delta(x_t), u_t - \tilde{u}_t\rangle \\
=&\langle \nabla F_\delta(x_t), \tilde{u}_t - x_t\rangle + \langle v_t, u_t - \tilde{u}_t\rangle + \langle \nabla F_\delta(x_t) - v_t, u_t - \tilde{u}_t\rangle \\
\leq&\langle \nabla F_\delta(x_t), \tilde{u}_t - x_t\rangle + \|\nabla F_\delta(x_t) - v_t\|\,\|u_t - \tilde{u}_t\|.
\end{aligned}$$

The first inequality is due to $\langle v_t, u_t\rangle \leq \langle v_t, \tilde{u}_t\rangle$ by the optimality of u_t and Cauchy–Schwarz inequality.

$$\begin{aligned}
&F_\delta(x_{t+1}) \\
\leq&F_\delta(x_t) + \gamma_t\langle \nabla F_\delta(x_t), u_t - x_t\rangle + \frac{cG\sqrt{d}\gamma_t^2}{2\delta}\|u_t - x_t\|^2 \\
\leq&F_\delta(x_t) + \gamma_t\langle \nabla F_\delta(x_t), \tilde{u}_t - x_t\rangle + \frac{cG\sqrt{d}\gamma_t^2}{2\delta}B^2 + \gamma_t\|\nabla F_\delta(x_t) - v_t\|\,\|u_t - \tilde{u}_t\|.
\end{aligned}$$

The first inequality is due to the $\frac{cG\sqrt{d}}{\delta}$-smoothness of the function $F_\delta(\cdot)$. The last inequality follows from the Cauchy–Schwarz inequality and the assumption that the diameter of the domain Ω is bounded by B. Taking expectations on both sides and fix $\gamma_t = \gamma$, we have:

$$\begin{aligned}
&\mathbb{E}[F_\delta(x_{t+1})] \\
\leq&\mathbb{E}[F_\delta(x_t)] + \gamma\mathbb{E}[\langle \nabla F_\delta(x_t), \tilde{u}_t - x_t\rangle] + \frac{cG\sqrt{d}\gamma^2B^2}{2\delta} + \gamma\mathbb{E}[\|\nabla F_\delta(x_t) - v_t\|\,\|u_t - \tilde{u}_t\|] \\
\leq&\mathbb{E}[F_\delta(x_t)] + \gamma\mathbb{E}\left[\langle \nabla F_\delta(x_t), \tilde{u}_t - x_t\rangle\right] + \frac{cG\sqrt{d}\gamma^2B^2}{2\delta} + \gamma B\mathbb{E}\left[\sqrt{\frac{d^2G^2}{\delta^2 b_2}\sum_{j=n_tq+1}^{t}\|x_j - x_{j-1}\|^2 + \frac{16\sqrt{2\pi}dG^2}{b_1}}\right] \\
\leq&\mathbb{E}[F_\delta(x_t)] + \gamma\mathbb{E}[\langle \nabla F_\delta(x_t), \tilde{u}_t - x_t\rangle] + \frac{cG\sqrt{d}\gamma^2B^2}{2\delta} + \gamma B\left(\frac{dG\gamma B\sqrt{q}}{\delta\sqrt{b_2}} + \frac{20\sqrt{d}G}{\sqrt{b_1}}\right) \\
=&\mathbb{E}[F_\delta(x_t)] + \gamma\mathbb{E}[\langle \nabla F_\delta(x_t), \tilde{u}_t - x_t\rangle] + \frac{cG\sqrt{d}\gamma^2B^2}{2\delta} + \frac{dG\gamma^2B^2\sqrt{q}}{\delta\sqrt{b_2}} + \frac{20\gamma\sqrt{d}BG}{\sqrt{b_1}}.
\end{aligned}$$

The second inequality is due to Lemma A.5. The last inequality follows from the fact $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ if $a, b \geq 0$. Rearrange the terms and telescope t from 0 to $T-1$ gives:

$$\begin{aligned}
&\sum_{t=0}^{T-1}\gamma\mathbb{E}[\langle -\nabla F_\delta(x_t), \tilde{u}_t - x_t\rangle] \\
\leq&\mathbb{E}[F_\delta(x_0)] - \mathbb{E}[F_\delta(x_T)] + \frac{cG\sqrt{d}\gamma^2B^2T}{2\delta} + \frac{dG\gamma^2B^2\sqrt{q}T}{\delta\sqrt{b_2}} + \frac{20\gamma\sqrt{d}BGT}{\sqrt{b_1}}.
\end{aligned}$$

Dividing both sides by γT, then

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\langle -\nabla F_\delta(x_t), \tilde{u}_t - x_t\rangle]$$

$$\leq \frac{\mathbb{E}[F_\delta(x_0)] - \mathbb{E}[F_\delta(x_T)]}{\gamma T} + \frac{cG\sqrt{d}\gamma B^2}{2\delta} + \frac{dG\gamma B^2\sqrt{q}}{\delta\sqrt{b_2}} + \frac{20\sqrt{d}BG}{\sqrt{b_1}}.$$

If we choose $b_2 = q$, then

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\langle -\nabla F_\delta(x_t), \tilde{u}_t - x_t\rangle] \\
\leq &\frac{\mathbb{E}[F_\delta(x_0)] - \mathbb{E}[F_\delta(x_T)]}{\gamma T} + \frac{cG\sqrt{d}\gamma B^2}{2\delta} + \frac{dG\gamma B^2}{\delta} + \frac{20\sqrt{d}BG}{\sqrt{b_1}} \\
\leq &\frac{GB}{\gamma T} + \frac{cG\sqrt{d}\gamma B^2}{2\delta} + \frac{dG\gamma B^2}{\delta} + \frac{20\sqrt{d}BG}{\sqrt{b_1}}.
\end{aligned}$$

The last inequality follows from the assumption that the function $F_\delta(\cdot)$ is G-Lipschitz and the diameter of the domain Ω is bound by B. We can further set $\gamma = \sqrt{\delta/(dTB)}$, then we can obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\langle -\nabla F_\delta(x_t), \tilde{u}_t - x_t\rangle] \leq \frac{3GB^{\frac{3}{2}}\sqrt{d}}{\sqrt{\delta T}} + \frac{20\sqrt{d}BG}{\sqrt{b_1}}.$$

D.2.2. Proof of Corollary 5.9

To obtain an ϵ-approximate solution, we have to set $T = \mathcal{O}(G^2B^3d\delta^{-1}\epsilon^{-2})$ and $b_1 = \mathcal{O}(dB^2G^2\epsilon^{-2})$. In addition, we set $b_2 = q = \sqrt{b_1}$, then the total function calls is

$$Tb_1/q + Tb_2 = \mathcal{O}(G^3B^4d^{\frac{3}{2}}\delta^{-1}\epsilon^{-3}).$$

D.2.3. Large-deviation Estimation of the Two-Phase VR-ZOSFW Method

Similar to the result of Theorem D.1 for the two-phase MB-ZOSFW method, we present the large-deviation estimation for the two-phase VR-ZOSFW method as follows.

Theorem D.2. *Let $\delta > 0$ and $0 < \epsilon, \Lambda < 1$, then there exists some $T, S, \hat{B} > 0$ such that the output $\bar{x}^{s^*}$ of Algorithm 6 with the VR-ZOSFW satisfies that* $\mathrm{Prob}(\min\{\max_{u\in\Omega}\langle -g, u - \bar{x}^{s^*}\rangle \geq \epsilon \colon g \in \partial_\delta F(\bar{x}^{s^*})\}) \leq \Lambda$ *and the total number of calls of the FQO is bounded by*

$$\mathcal{O}\left(\frac{B^4G^3d^{\frac{3}{2}}}{\delta\epsilon^3}\log\left(\frac{1}{\Lambda}\right) + \frac{dG^2B^2}{\epsilon^2\Lambda}\log\left(\frac{1}{\Lambda}\right)^2\right).$$

Proof. We define $\tilde{u}^s = \arg\max_{u\in\Omega}\langle u, -\nabla F_\delta(\bar{x}^s)\rangle$, it follows that

$$\begin{aligned}
\langle -\nabla F_\delta(\bar{x}^{s^*}), \tilde{u}^{s^*} - \bar{x}^{s^*}\rangle =&\langle -g^{s^*}, \tilde{u}^{s^*} - \bar{x}^{s^*}\rangle + \langle g^{s^*} - \nabla F_\delta(\bar{x}^{s^*}), \tilde{u}^{s^*} - \bar{x}^{s^*}\rangle \\
\leq&\langle -g^{s^*}, u^{s^*} - \bar{x}^{s^*}\rangle + \langle -g^{s^*}, \tilde{u}^{s^*} - u^{s^*}\rangle + \left\|g^{s^*} - \nabla F_\delta(\bar{x}^{s^*})\right\|\left\|\tilde{u}^{s^*} - \bar{x}^{s^*}\right\| \\
\leq&\langle -g^{s^*}, u^{s^*} - \bar{x}^{s^*}\rangle + \left\|g^{s^*} - \nabla F_\delta(\bar{x}^{s^*})\right\|\left\|\tilde{u}^{s^*} - \bar{x}^{s^*}\right\|.
\end{aligned}$$

The last inequality is due to $u^{s^*} = \arg\max_{u\in\Omega}\langle -g^{s^*}, u\rangle$. The first term on the right-hand side of the inequality can be bound by

$$\begin{aligned}
\langle -g^{s^*}, u^{s^*} - \bar{x}^{s^*}\rangle =& \min_{s\in\{0,1,\ldots,S-1\}}\langle -g^s, u^s - \bar{x}^s\rangle \\
=& \min_{s\in\{0,1,\ldots,S-1\}}(\langle -\nabla F_\delta(x^s), u^s - \bar{x}^s\rangle + \langle \nabla F_\delta(x^s) - g^s, u^s - \bar{x}^s\rangle) \\
\leq& \min_{s\in\{0,1,\ldots,S-1\}}(\langle -\nabla F_\delta(x^s), \tilde{u}^s - \bar{x}^s\rangle + \langle \nabla F_\delta(x^s) - g^s, u^s - \bar{x}^s\rangle) \\
\leq& \min_{s\in\{0,1,\ldots,S-1\}}\langle -\nabla F_\delta(x^s), \tilde{u}^s - \bar{x}^s\rangle + \max_{s\in\{0,1,\ldots,S-1\}}\langle \nabla F_\delta(x^s) - g^s, u^s - \bar{x}^s\rangle
\end{aligned}$$

$$\leq \min_{s\in\{0,1,\ldots,S-1\}} \langle -\nabla F_\delta(x^s), \tilde{u}^s - \bar{x}^s\rangle + \max_{s\in\{0,1,\ldots,S-1\}} \|\nabla F_\delta(x^s) - g^s\| \, \|u^s - \bar{x}^s\|.$$

The first inequality follows from $\tilde{u}^s = \arg\max_{u\in\Omega}\langle u, -\nabla F_\delta(\bar{x}^s)\rangle$, and the last inequality is due to Cauchy–Schwarz inequality. The proof in Theorem 5.8 implies that

$$\mathbb{E}[\langle -\nabla F_\delta(x^s), \tilde{u}^s - x^s\rangle] \leq \frac{3GB^{\frac{3}{2}}\sqrt{d}}{\sqrt{\delta T}} + \frac{20\sqrt{d}BG}{\sqrt{b_1}}.$$

Using Markov's inequality, we have

$$\text{Prob}\left(\langle -\nabla F_\delta(x^s), \tilde{u}^s - x^s\rangle \geq \frac{6GB^{\frac{3}{2}}\sqrt{d}}{\sqrt{\delta T}} + \frac{40\sqrt{d}BG}{\sqrt{b_1}}\right) \leq \frac{1}{2}.$$

Therefore, we can deduce that

$$\text{Prob}\left(\min_{s\in\{0,1,\ldots,S-1\}}\langle -\nabla F_\delta(x^s), \tilde{u}^s - \bar{x}^s\rangle \geq \frac{6GB^{\frac{3}{2}}\sqrt{d}}{\sqrt{\delta T}} + \frac{40\sqrt{d}BG}{\sqrt{b_1}}\right) \leq \frac{1}{2^S}.$$

By Lemma A.1, we have $\mathbb{E}[g_k^s \mid \bar{x}^s] = \nabla F_\delta(\bar{x}^s)$ and $\mathbb{E}[\|g_k^s\|^2 \mid \bar{x}^s] \leq 16\sqrt{2\pi}dG^2$. By Markov's inequality and Proposition A.2, it yields that

$$\begin{aligned}&\text{Prob}\left(\|g^s - \nabla F_\delta(\bar{x}^s)\| \, \|u^s - \bar{x}^s\| \geq \frac{\sqrt{\lambda}(8\sqrt{d}GB)}{\sqrt{\hat{B}}}\right)\\ =&\text{Prob}\left(\|g^s - \nabla F_\delta(\bar{x}^s)\|^2 \, \|u^s - \bar{x}^s\|^2 \geq \frac{\lambda(16\sqrt{2\pi}dG^2B^2)}{\hat{B}}\right) \leq \frac{1}{\lambda}.\end{aligned}$$

Therefore, we can conclude that

$$\text{Prob}\left(\max_{s\in\{0,1,\ldots,S-1\}} \|g^s - \nabla F_\delta(\bar{x}^s)\| \, \|u^s - \bar{x}^s\| \geq \frac{\sqrt{\lambda}(8\sqrt{d}GB)}{\sqrt{\hat{B}}}\right) \leq \frac{S}{\lambda}.$$

Using a similar argument, one has

$$\text{Prob}\left(\left\|g^{s^*} - \nabla F_\delta(\bar{x}^{s^*})\right\| \left\|\tilde{u}^{s^*} - \bar{x}^{s^*}\right\| \geq \frac{\sqrt{\lambda}(8\sqrt{d}GB)}{\sqrt{\hat{B}}}\right) \leq \frac{1}{\lambda}.$$

Combining the above inequalities, for all $\lambda > 0$ we have

$$\text{Prob}\left(\langle -\nabla F_\delta(\bar{x}^{s^*}), \tilde{u}^{s^*} - \bar{x}^{s^*}\rangle \geq \frac{6GB^{\frac{3}{2}}\sqrt{d}}{\sqrt{\delta T}} + \frac{40\sqrt{d}BG}{\sqrt{b_1}} + \frac{\sqrt{\lambda}(16\sqrt{d}GB)}{\sqrt{\hat{B}}}\right) \leq \frac{S+1}{\lambda} + \frac{1}{2^S},$$

We set $\lambda = \frac{2(S+1)}{\Lambda}$, $S = \lceil \log\frac{2}{\Lambda}\rceil$, and the parameters $(T, b_1, b_2, q, \hat{B})$ as follows

$$T = \mathcal{O}\left(\frac{B^3G^2d}{\delta\epsilon^2}\right), \quad b_1 = \mathcal{O}\left(\frac{B^2G^2d}{\epsilon^2}\right), \quad b_2 = q = \sqrt{b_1}, \quad \hat{B} = \mathcal{O}\left(\frac{dG^2B^2(S+1)}{\epsilon^2\Lambda}\right).$$

Therefore, we have

$$\text{Prob}\left(\langle -\nabla F_\delta(\bar{x}^{s^*}), \tilde{u}^{s^*} - \bar{x}^{s^*}\rangle \geq \epsilon\right) \leq \Lambda.$$

The total number of function oracle calls is therefore bounded by

$$S(T(b_1/q + b_2) + \hat{B}) = \mathcal{O}\left(\frac{B^4G^3d^{\frac{3}{2}}}{\delta\epsilon^3}\log\left(\frac{1}{\Lambda}\right) + \frac{dG^2B^2}{\epsilon^2\Lambda}\log\left(\frac{1}{\Lambda}\right)^2\right).$$

□