Zeroth-order (Non)-Convex Stochastic Optimization via Conditional Gradient and Gradient Updates

Krishnakumar Balasubramanian
Department of Statistics
University of California, Davis
kbala@ucdavis.edu

Saeed Ghadimi *
Department of Operations Research and Financial Engineering
Princeton University
sghadimi@princeton.edu

Abstract

In this paper, we propose and analyze zeroth-order stochastic approximation algorithms for nonconvex and convex optimization. Specifically, we propose generalizations of the conditional gradient algorithm achieving rates similar to the standard stochastic gradient algorithm using only zeroth-order information. Furthermore, under a structural sparsity assumption, we first illustrate an implicit regularization phenomenon where the standard stochastic gradient algorithm with zeroth-order information adapts to the sparsity of the problem at hand by just varying the step-size. Next, we propose a truncated stochastic gradient algorithm with zeroth-order information, whose rate depends only poly-logarithmically on the dimensionality.

1 Introduction

In this work, we propose and analyze algorithms for solving the following stochastic optimization problem

$$\min_{x\in\mathcal{X}} \left\{ f(x) = \mathbf{E}_\xi[F(x,\xi)] = \int F(x,\xi)\, dP(\xi) \right\}, \tag{1.1}$$

where $\mathcal{X}$ is a closed convex subset of $\mathbb{R}^d$. The case of nonconvex objective function f is ubiquitous in modern deep learning problems and developing provable algorithms for such problems has been a topic of intense research in the recent years [16, 11], along with the more standard convex case [1]. Several methods are available for solving such stochastic optimization problems under access to different oracle information, for example, function queries (zeroth-order oracle), gradient queries (first-order oracle), and higher-order oracles. In this work, we assume that we only have access to noisy evaluation of f through a stochastic zeroth-order oracle described in detail in Assumption 1. This oracle setting is motivated by several applications where only noisy function queries of problem (1.1) is available and obtaining higher-order information might not be possible. Such a situation occurs frequently for example, in simulation based modeling [29], selecting the tuning parameters of deep neural networks [32] and design of black-box attacks to deep networks [3]. It is worth noting that recently such zeroth-order optimization techniques have also been applied in the field of reinforcement learning [30, 4, 20]. Furthermore, methods using similar oracles have been studied in the literature under the name of derivative-free optimization [33, 5], bayesian optimization [21] and optimization with bandit feedback [2].

*Both authors contributed equally and are listed in alphabetical order.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

Algorithm	Structure	Function Queries	References
ZSCG (Alg 1)	Nonconvex	$\mathcal{O}(d/\epsilon^4)$	Theorem 2.1
	Convex	$\mathcal{O}(d/\epsilon^3)$	
Modified ZSCG (Alg 3)	Convex	$\mathcal{O}(d/\epsilon^2)$	Theorem 2.2
ZSGD (Alg 4)	Nonconvex, s-sparse	$\mathcal{O}\left((s\log d)^2/\epsilon^4\right)$	Theorem 3.1
Truncated ZSGD (Alg 5)	Convex, s-sparse	$\mathcal{O}\left(s(\log d/\epsilon)^2\right)$	Theorem 3.2
ZSGD	Convex	$\mathcal{O}(d/\epsilon^2)$	[18, 7, 9]
	Nonconvex	$\mathcal{O}(d/\epsilon^4)$	[9]

Table 1: A list of complexity bounds for stochastic zeroth-order methods to find an ϵ-stationary or ϵ-optimal (see Definition 1.1) point of problem (1.1).

Algorithms available for solving problem (1.1) also depend crucially on the constraint set $\mathcal{X}$. First, consider the case of $\mathcal{X} = \mathbb{R}^d$. When first-order information is available, the rate of convergence of the standard Gradient Descent (GD) algorithm is dimension-independent [26]. Whereas when only the zeroth-order information is available, any algorithm (with estimated gradients) has (at least) linear dependence on d [9, 18, 7]. This illustrates the main difference between the availability of different oracle information. Next, note that depending on the geometry of the constraint set $\mathcal{X}$, the cost of computing the projection to the set might be prohibitive. This lead to the re-emergence of Conditional Gradient (CG) algorithms recently [12, 15]. But the performance of the CG algorithm under the zeroth-order oracle is unexplored in the literature to the best of our knowledge, both under convex and nonconvex settings. Hence it is natural to ask if CG algorithms, with access to zeroth-order oracle has similar (or better) convergence rates compared to GD algorithms with zeroth-order information. We propose and analyze in Section 2 a classical version of CG algorithm with zeroth-order information and present convergence results. We then propose a modification in Section 2.2 that has improved rates, when f is convex.

Notably, with zeroth-order information, the complexity of CG algorithms also depend linearly on the dimensionality, similar to the GD algorithms. We refer to this situation as the low-dimensional setting in the rest of the paper. This motivates us to examine assumptions under which one can achieve weaker dependence on the dimensionality while optimizing with zeroth-order information. In a recent work [34], the authors used a *functional sparsity* assumption, under which the function $f : \mathbb{R}^d \to \mathbb{R}$ to be optimized depends only on s of the d components, and proposed a LASSO based algorithm that has poly-logarithmic dependence on the dimensionality when f is convex. We refer to this situation as the high-dimensional setting. In this work, we perform a refined analysis under a similar sparsity assumption for both convex and nonconvex objective functions. When the performance is measured by the size of the gradient, we show in Section 3 that zeroth-order GD algorithm (without using thresholding or LASSO approach of [34]), has poly-logarithmic dependence on the dimensionality thereby demonstrating an *implicit regularization* phenomenon in this setting. Note that this is applicable for both convex and nonconvex objectives. When the performance is measured by function values (as in the case of convex objective), we show that a simple thresholded zeroth-order GD algorithm achieves a poly-logarithmic dependence on dimensionality. This algorithm is notably less expensive than the algorithm proposed by [34].

Our contributions: To summarize the above discussion, in this paper we make the following contributions to the literature on zeroth-order stochastic optimization: (i) We first analyze a classical version of CG algorithm in the nonconvex (and convex) setting, under access to zeroth-order information and provide results on the convergence rates in the low-dimensional setting; (ii) We then propose and analyze a modified CG algorithm in the convex setting with zeroth-order information and show that it attains improved rates in the low-dimensional setting; (iii) Finally, we consider a zeroth-order stochastic gradient algorithm in the high-dimensional nonconvex (and convex) setting and illustrate an implicit regularization phenomenon. We also show that this algorithm achieves rates that depend only poly-logarithmically on dimensionality. Our contributions extend the applicability of zeroth-order stochastic optimization to the constrained and high-dimensional setting and also provide theoretical insights in the form of rates of convergence. A summary of the results is provided in Table 1.

1.1 Preliminaries

We now list the main assumptions we make in this work. Additional assumptions will be introduced in the appropriate sections as needed. We start with the assumption on the zeroth-order oracle.

Assumption 1 *Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$. For any $x \in \mathbb{R}^d$, the zeroth-order oracle outputs an estimator $F(x,\xi)$ of $f(x)$ such that $\mathbf{E}[F(x,\xi)] = f(x), \mathbf{E}[\nabla F(x,\xi)] = \nabla f(x), \mathbf{E}[\|\nabla F(x,\xi) - \nabla f(x)\|_*^2] \leq \sigma^2$, where $\|\cdot\|_*$ denotes the dual norm.*

It should be noted that in the above assumption, we do not observe $\nabla F(x,\xi)$ and we just assume that it is an unbiased estimator of gradient of f and its variance is bounded. Furthermore, we make the following smoothing assumption about the noisy estimation of f.

Assumption 2 *Function F has Lipschitz continuous gradient with constant L, almost surely for any ξ, i.e., $\|\nabla F(y,\xi) - \nabla F(x,\xi)\|_* \leq L\|y-x\|$, which consequently implies that $|F(y,\xi) - F(x,\xi) - \langle \nabla F(x,\xi), y-x\rangle| \leq \frac{L}{2}\|y-x\|^2$.*

It is easy to see that the above two assumptions imply that f also has Lipschitz continuous gradient with constant L since

$$\|\nabla f(y) - \nabla f(x)\|_* \leq \mathbf{E}\left[\|\nabla F(y,\xi) - \nabla F(x,\xi)\|_*\right] \leq L\|y-x\| \tag{1.2}$$

due the Jensen's inequality for the dual norm. We now collect some facts about a gradient estimator based on the above zeroth-order information. Let $u \sim N(0, I_d)$ be a standard Gaussian random vector. For some $\nu \in (0,\infty)$ consider the smoothed function $f_\nu(x) = \mathbf{E}_u\left[f(x+\nu u)\right]$. Nesterov [27] has shown that $\nabla f_\nu(x) =$

$$\mathbf{E}_u\left[\frac{f(x+\nu u)}{\nu}\, u\right] = \mathbf{E}_u\left[\frac{f(x+\nu u) - f(x)}{\nu}\, u\right] = \frac{1}{(2\pi)^{d/2}} \int \frac{f(x+\nu u) - f(x)}{\nu}\, u\, e^{-\frac{\|u\|_2^2}{2}}\, du. \tag{1.3}$$

This relation implies that we can estimate gradient of f_ν by only using evaluations of f. In particular, one can define stochastic gradient of $f_\nu(x)$ as

$$G_\nu(x,\xi,u) = \frac{F(x+\nu u,\xi) - F(x,\xi)}{\nu}\, u, \tag{1.4}$$

which is an unbiased estimator of $\nabla f_\nu(x)$ under Assumption 1 since

$$\mathbf{E}_{u,\xi}[G_\nu(x,\xi,u)] = \mathbf{E}_u[\tfrac{f(x+\nu u) - f(x)}{\nu}\, u] = \nabla f_\nu(x).$$

We leverage some properties of f_ν due to Nesterov [27] in our proofs later, that we replicate in the supplementary material (Section A) for convenience. Finally, we define the following criterion which are used to analyze the complexity of our proposed algorithms.

Definition 1.1 *Assume that a solution $\bar{x} \in \mathcal{X}$ as output of an algorithm and a target accuracy $\epsilon > 0$ are given. Then: (i) If f is nonconvex, $\bar{x}$ is called an ϵ-stationary point of the unconstrained variant of problem (1.1) if $\mathbf{E}[\|\nabla f(\bar{x})\|_*] \leq \epsilon$. For the constrained case, $\bar{x}$ should satisfies $\mathbf{E}[\langle \nabla f(\bar{x}), \bar{x} - u\rangle] \leq \epsilon$ for all $u \in \mathcal{X}$; (ii) If f is convex, $\bar{x}$ is called an ϵ-optimal point of problem (1.1) if $\mathbf{E}[f(\bar{x})] - f(x_*) \leq \epsilon$, where x_* denotes an optimal solution of the problem.*

It should be pointed out that while the above performance measures are presented in expectation form, one can also use their high probability counterparts. Since, convergence results in this case can be obtained by making sub-Gaussian tail assumptions on the output of the zeroth-order oracle and using the standard two-stage process presented in [9, 19], we do not elaborate more on this approach. Furthermore, note that the aforementioned measures for evaluating the algorithms are from the derivative-free optimization point of view. In the literature on optimization with bandit feedback, the preferred performance measure is the so-called regret of the algorithm [2, 31] which may have a different behavior than our performance measures.

2 Zeroth-order Stochastic Conditional Gradient Type Method

In this section, we study zeroth-order stochastic conditional gradient (ZSCG) algorithms in the low-dimensional setting for solving constrained stochastic optimization problems. In particular, we incorporate a variant of the gradient estimate defined in (1.4) into the framework of the classical CG method and provide its convergence analysis in Subsection 2.1. We also present improved rates for a variant of this method in Subsection 2.2 when f is convex. Throughout this section, we assume that $\mathbb{R}^d$ is equipped with the self-dual Euclidean norm i.e., $\|\cdot\| = \|\cdot\|_2$. We also make the following natural boundedness assumption.

Algorithm 1 Zeroth-order Stochastic Conditional Gradient Method

Input: $z_0 \in \mathcal{X}$, smoothing parameter $\nu > 0$, non-negative sequence α_k, positive integer sequence m_k, iteration limit $N \geq 1$ and probability distribution $P_R(\cdot)$ over $\{1, \ldots, N\}$.
for $k = 1, \ldots, N$ **do**

1. Generate $u_k = [u_{k,1}, \ldots, u_{k,m_k}]$, where $u_{k,j} \sim N(0, I_d)$, call the stochastic oracle to compute m_k stochastic gradient $G_\nu^{k,j}$ according to (1.4) and take their average:

$$\bar{G}_\nu^k \equiv \bar{G}_\nu(z_{k-1}, \xi_k, u_k) = \frac{1}{m_k} \sum_{j=1}^{m_k} \frac{F(z_{k-1} + \nu u_{k,j}, \xi_{k,j}) - F(z_{k-1}, \xi_{k,j})}{\nu} u_{k,j}. \tag{2.1}$$

2. Compute

$$x_k = \operatorname*{argmin}_{u \in \mathcal{X}} \langle \bar{G}_\nu^k, u \rangle, \tag{2.2}$$

$$z_k = (1 - \alpha_k) z_{k-1} + \alpha_k x_k. \tag{2.3}$$

end for
Output: Generate R according to $P_R(\cdot)$ and output z_R.

Assumption 3 *The feasible set $\mathcal{X}$ is bounded such that $\max_{x,y \in \mathcal{X}} \|y - x\| \leq D_{\mathcal{X}}$ for some $D_{\mathcal{X}} > 0$. Moreover, for all $x \in \mathcal{X}$, there exists a constant $B > 0$ such that $\|\nabla f(x)\| \leq B$.*

We should point out that under Assumptions 1 and 2, the second statement in Assumption 3 follows immediately by the first one and choosing $B := LD_{\mathcal{X}} + \|\nabla f(x_*)\|$. However, we just use B in our analysis for simplicity.

2.1 Zeroth-order Stochastic Conditional Gradient Method

The vanilla ZSCG method is formally presented in Algorithm 1 and a few remarks about it follows. First, note that this algorithm differs from the classical CG method in estimating the gradient using zeroth-order information and in outputting a random solution from the generated trajectory. This randomization scheme is the current practice in the literature to provide convergence results for nonconvex stochastic optimization (see e.g., [9, 28]). Second, $\bar{G}_\nu^k$ is the averaged variant of the gradient estimator presented in Subsection 1.1 and is still an unbiased estimator of $\nabla f_\nu(z_{k-1})$. Moreover, it can be easily seen that it has a reduced variance with respect to the individual estimators i.e.,

$$\mathbf{E}[\|\bar{G}_\nu^k - \nabla f_\nu(z_{k-1})\|^2] \leq \frac{1}{m_k} \mathbf{E}[\|G_\nu^{k,j} - \nabla f_\nu(z_{k-1})\|^2]. \tag{2.4}$$

We emphasize that the use of the above variance reduction technique in stochastic CG methods is standard and has been previously proposed and leveraged in several works (see e.g., [19, 13, 28, 22, 23, 10]). Indeed, when exact gradient is not available, an error term appears in the convergence analysis which should converge to 0 at a certain rate as the algorithm moves forward. Hence, the choice of m_k plays a key role in the convergence analysis of Algorithm 1. $\bar{G}_\nu^k$ can be also viewed as a biased estimator for $\nabla f(z_{k-1})$. Finally, since f is possibly nonconvex, we need a different criteria than the optimality gap to provide convergence analysis of Algorithm 1. The well-known Frank-Wolfe Gap given by

$$g_{\mathcal{X}}^k \equiv g_{\mathcal{X}}(z_{k-1}) := \langle \nabla f(z_{k-1}), z_{k-1} - \hat{x}_k \rangle, \ \text{ where } \ \hat{x}_k = \operatorname*{argmin}_{u \in \mathcal{X}} \langle \nabla f(z_{k-1}), u \rangle, \tag{2.5}$$

has been widely use in the literature to show rate of convergence of the CG methods when f is convex (see e.g., [8, 6, 14]). In this case, it is easy to see that

$$f(z_{k-1}) - f^* \leq g_{\mathcal{X}}(z_{k-1}). \tag{2.6}$$

When f is nonconvex, this criteria is still useful since $\langle \nabla f(z_{k-1}), z_{k-1} - u \rangle \leq g_{\mathcal{X}}(z_{k-1}), \ \forall u \in \mathcal{X}$, which implies that one can obtain an approximate stationary point of problem (1.1) by minimizing $g_{\mathcal{X}}^k$, in the view of Definition 1.1. Note that in our setting, this quantity is not exactly computable and it is only used to provide convergence analysis of Algorithm 1 as shown in the next result.

Theorem 2.1 *Let $\{z_k\}_{k \geq 0}$ be generated by Algorithm 1 and Assumptions 1, 2, and 3 hold.*

1. *Let f be nonconvex, bounded from below by f^*, and let the parameters of the algorithm be set as*

$$\nu = \sqrt{\frac{2B_{L\sigma}}{N(d+3)^3}}, \;\; \alpha_k = \frac{1}{\sqrt{N}}, \;\; m_k = 2B_{L\sigma}(d+5)N, \;\; \forall k \geq 1 \tag{2.7}$$

for some constant $B_{L\sigma} \geq \max\{\sqrt{B^2+\sigma^2}/L, 1\}$ and a given iteration bound $N \geq 1$. Then we have

$$\mathbf{E}[g_{\mathcal{X}}^R] \leq \frac{f(z_0) - f^* + LD_{\mathcal{X}}^2 + 2\sqrt{B^2+\sigma^2}}{\sqrt{N}}, \tag{2.8}$$

where R is uniformly distributed over $\{1, \ldots, N\}$ and g_k is defined in (2.5). Hence, the total number of calls to the zeroth-order stochastic oracle and linear subproblems required to be solved to find an ϵ-stationary point of problem (1.1) are, respectively, bounded by

$$\mathcal{O}\left(\frac{d}{\epsilon^4}\right), \;\; \mathcal{O}\left(\frac{1}{\epsilon^2}\right). \tag{2.9}$$

2. *Let f be convex and let the parameters be set to*

$$\nu = \sqrt{\frac{2B_{L\sigma}}{N^2(d+3)^3}}, \;\; \alpha_k = \frac{6}{k+5}, \;\; m_k = 2B_{L\sigma}(d+5)N^2, \;\; \forall k \geq 1. \tag{2.10}$$

Then we have

$$\mathbf{E}[f(z_N)] - f^* + \mathbf{E}[g_{\mathcal{X}}^R] \leq \frac{120[f(z_0) - f(x_*)]}{(N+3)^3} + \frac{36LD_{\mathcal{X}}^2}{N+5} + \frac{\sqrt{B^2+\sigma^2}}{N} \tag{2.11}$$

where R is random variable from $\{1, \ldots, N\}$ whose probability distribution is given by

$$P_R(R=k) = \frac{\alpha_k \Gamma_N}{2\Gamma_N(1-\Gamma_N)}, \qquad \Gamma_k = \prod_{i=1}^{k}\left(1 - \frac{\alpha_i}{2}\right), \;\; \Gamma_0 = 1. \tag{2.12}$$

Hence, the total number of calls to the zeroth-order stochastic oracle and linear subproblems required to be solved to find and ϵ-optimal solution of problem (1.1) are, respectively, bounded by

$$\mathcal{O}\left(\frac{d}{\epsilon^3}\right), \;\; \mathcal{O}\left(\frac{1}{\epsilon}\right). \tag{2.13}$$

Remark 1 *Observe that the complexity bounds in (2.9), in terms of ϵ, match the ones obtained in [10, 28, 23] for stochastic CG method with first-order oracle applied to nonconvex problems. For convex problems, similar observation can be made for terms in (2.13) which match the ones in [13, 10]. Note that the linear dependence of our complexity bounds on d is unimprovable due to the lower bounds for zeorth-order algorithms applied to convex optimization problems [7]. We conjecture that this is also the case for nonconvex problems.* ∎

2.2 Improved Rates for Convex Problems

Our goal in this subsection is to improve the complexity bounds of the ZCSG method when f is convex. Recall that the ZSCG method presented in Section 2.1 involves two main steps: the gradient evaluation step and the linear optimization step. Motivated by [19], we now propose a modified algorithm that allows one to skip the gradient evaluation from time to time. Notice that, as our gradients are estimated by calling the zeroth-order oracle, this directly reduces the number of calls to the zeroth-order oracle. We first state a subroutine in Algorithm 2 used in our modified algorithm. Note that Algorithm 2 is indeed the zeroth-order conditional gradient method for inexactly solving the following quadratic program

$$P_{\mathcal{X}}(x, g, \gamma) = \operatorname*{argmin}_{u \in \mathcal{X}} \left\{ \langle g, u \rangle + \frac{\gamma}{2}\|u - x\|^2 \right\}, \tag{2.15}$$

which is the standard subproblem of stochastic first-order methods applied to a minimization problem when g is an unbiased stochastic gradient of the objective function at x. We now present Algorithm 3 which applies the CG method to inexactly solve subproblems of the stochastic accelerated gradient

Algorithm 2 Inexact Conditional Gradient (ICG) method

Input: (x, g, γ, μ).
Set $\bar{y}_0 = x$, $t = 1$, and $\kappa = 0$..
while $\kappa = 0$ **do**

$$y_t = \underset{u \in \mathcal{X}}{\operatorname{argmin}} \{h_\gamma(u) := \langle g + \gamma(\bar{y}_{t-1} - x), u - \bar{y}_{t-1} \rangle\} \tag{2.14}$$

If $h_\gamma(y_t) \geq -\mu$, set $\kappa = 1$.
Else $\bar{y}_t = \frac{t-1}{t+1}\bar{y}_{t-1} + \frac{2}{t+1}y_t$ and $t = t + 1$.
end while
Output $\bar{y}_t$.

method. This way of using CG methods can significantly improve the total number of calls to the stochastic oracle. Our next result provides convergence analysis of this algorithm.

Algorithm 3 Zeroth-order Stochastic Accelerated Gradient Method with Inexact Updates

Input:$z_0 = x_0 \in \mathcal{X}$, smoothing parameter $\nu > 0$, sequences α_k, m_k, γ_k, μ_k, and iteration limit $N \geq 1$.
for $k = 1, \ldots, N$ **do**
1. Set

$$w_k = (1 - \alpha_k)z_{k-1} + \alpha_k x_{k-1} \tag{2.16}$$

2. Generate $u_k = [u_{k,1}, \ldots, u_{k,m_k}]$, where $u_{k,j} \sim N(0, I_d)$, call the stochastic oracle m_k times to compute $\bar{G}_\nu^k \equiv \bar{G}_\nu(w_k, \xi_k, u_k)$ as given by (2.1), and set

$$x_k = ICG(x_{k-1}, \bar{G}_\nu^k, \gamma_k, \mu_k), \tag{2.17}$$

where $ICG(\cdot)$ is the output of Algorithm 2 with input $(x_{k-1}, \bar{G}_\nu^k, \gamma_k)$.
3. Set

$$z_k = (1 - \alpha_k)z_{k-1} + \alpha_k x_k \tag{2.18}$$

end for
Output: z_N

Theorem 2.2 *Let $\{z_k\}_{k\geq 1}$ be generated by Algorithm 3, the function f be convex, and*

$$\begin{aligned} \alpha_k &= \frac{2}{k+1}, \ \gamma_k = \frac{4L}{k}, \ \mu_k = \frac{LD_X^0}{kN}, \ \nu = \frac{1}{\sqrt{2N}} \max\left\{\frac{1}{d+3}, \sqrt{\frac{D_X^0}{d(N+1)}}\right\} \\ m_k &= \frac{k(k+1)}{D_X^0} \max\left\{(d+5)B_{L\sigma}N, d+3\right\}, \ \forall k \geq 1, \end{aligned} \tag{2.19}$$

and for some constants $D_X^0 \geq \|x_0 - x_*\|^2$ and $B_{L\sigma} \geq \max\{\sqrt{B^2 + \sigma^2}/L, 1\}$. Then under Assumptions 1, 2, and 3, we have

$$\mathbf{E}[f(z_N) - f(x_*)] \leq \frac{12LD_X^0}{N(N+1)}. \tag{2.20}$$

Hence, the total number of calls to the stochastic oracle and linear subproblems solved to find and ϵ-stationary point of problem (1.1) are, respectively, bounded by

$$\mathcal{O}\left(\frac{d}{\epsilon^2}\right), \ \mathcal{O}\left(\frac{1}{\epsilon}\right). \tag{2.21}$$

Remark 2 *Observe that while the number of linear subproblems required to find an ϵ-optimal solution of problem (1.1) is the same for both Algorithms 1 and 3, the number of calls to the stochastic zeroth-order oracle in Algorithm 3 is significantly smaller than that of Algorithm 1. It is also natural to ask if such an improvement is achievable when f is nonconvex. This situation is more subtle and the answer depends on the performance measure used to measure the rate of convergence. Indeed, we can obtain improved complexity bounds for a different performance measure than the Frank-Wolfe*

Algorithm 4 Zeroth-Order Stochastic Gradient Method

Input: $x_0 \in \mathbb{R}^d$, smoothing parameter $\nu > 0$, iteration limit $N \geq 1$, a probability distribution P_R supported on $\{0, \ldots, N-1\}$.
for k =1, ..., N **do**
Generate $u_k \sim N(0, I_d)$, call the stochastic oracle, and compute $G_\nu(x_{k-1}, \xi_k, u_k)$ as defined in (1.4) and set $x_k = x_{k-1} - \gamma_k G_\nu(x_{k-1}, \xi_k; u_k)$.
end for
Output: Generate R according to $P_R(\cdot)$ and output x_R.

gap with a modified algorithm. However, the complexity bounds are of the same order as (2.9) in terms of the Frank-Wolfe gap for the modified algorithm. For the sake of completeness, we add this algorithm and its convergence analysis in the supplementary material in Section D. ∎

3 Zeroth-order Stochastic Gradient Methods

In this section, we study unconstrained variant of problem 1.1 i.e, $\mathcal{X} = \mathbb{R}^d$, under certain sparsity assumptions on the objective function f to facilitate zeroth-order optimization in high-dimensions. Recently, [34] considered the convex case and proposed algorithms for high-dimensional zeroth-order stochastic optimization. Motivated by [34], we make the following assumption.

Assumption 4 *For any $x \in \mathbb{R}^d$, we have $\|\nabla f(x)\|_0 \leq s$, i.e., the gradient is s-sparse, where $s \ll d$.*

Note that the above assumption implies $\|\nabla f(x)\|_2 \leq \sqrt{s}\|\nabla f(x)\|_\infty$ and $\|\nabla f(x)\|_1 \leq s\|\nabla f(x)\|_\infty$, for all $x \in \mathbb{R}^d$. Furthermore, this assumption also implies that $\|\nabla f_\nu(x)\|_0 \leq s$ for all $x \in \mathbb{R}^d$ since $\nabla f_\nu(x) = \mathbf{E}_u\left[\nabla f(x + \nu u)\right]$. To exploit the above sparsity assumption, we assume that the primal space $\mathbb{R}^d$ is equipped with the l_∞ norm throughout this section. More specifically, we assume that Assumptions 1 and 2 hold with the choice of $\|\cdot\| = \|\cdot\|_\infty$ and its dual norm $\|\cdot\|_* = \|\cdot\|_1$. We now present zeroth-order stochastic gradient methods for solving problem (1.1) when f is nonconvex and convex, in Subsections 3.1 and 3.2 respectively.

3.1 Zeroth-order Stochastic Gradient Method for Nonconvex Problems

In this subsection, we consider the zeroth-order stochastic gradient method presented in [9] (provided in Algorithm 4 for convenience) and provide a refined convergence analysis for it under the sparsity assumption 1, when f is nonconvex. Our main convergence result for Algorithm 4 under the gradient sparsity assumption is stated below.

Theorem 3.1 *Let $\{x_k\}_{k\geq 0}$ be generated by Algorithm 4 and stepsizes are chosen such that $\forall k \geq 1$,*

$$\gamma_k = \frac{1}{2L\hat{C}\log d}\min\left\{\frac{1}{12\hat{s}\log d}, \sqrt{\frac{D_0 L\hat{C}}{2N\sigma^2}}\right\}, \quad \nu \leq \frac{1}{\sqrt{L\hat{C}\log d}}\min\left\{\sqrt{\frac{2\sigma^2}{L}}, \sqrt{\frac{D_0}{N}}\right\} \tag{3.1}$$

for some $\hat{s} \geq s$, $\hat{C} \geq C$ (the universal constant defined in Lemma C.1), and $D_0 \geq f(x_0) - f^*$. Assume that f is nonconvex. Then under Assumptions 1, 2, and 4, we have

$$\mathbf{E}_\zeta\left[\|\nabla f(x_R)\|_1^2\right] \leq \frac{150L\hat{C}D_0\hat{s}s(\log d)^2}{N} + \frac{54\sigma\sqrt{2L\hat{C}D_0}\, s\log d}{\sqrt{N}}, \tag{3.2}$$

where $\zeta = \{\xi, u, R\}$ and R is uniformly distributed over $\{0, \ldots, N-1\}$. Hence, the total number of calls to the stochastic oracle (number of iterations) required to find an ϵ-stationary point of problem (1.1), in the view of Definition 1.1, is bounded by

$$\mathcal{O}\left(\frac{(\hat{s}\log d)^2}{\epsilon^4}\right). \tag{3.3}$$

Remark 3 *Note that the above theorem establishes rate of convergence of Algorithm 4 which only poly-logarithmically depends on the problem dimension d, by just selecting the step-size appropriately, under additional assumption that the gradient is sparse. This significantly improves the linear dimensionality dependence of the rate of convergence of this algorithm as presented in [9] for general nonconvex smooth problems.* ∎

Algorithm 5 Truncated Zeroth-Order Stochastic Gradient Method

Given a positive integer $\hat{s}$, replace updating step of Algorithm 4 with

$$x_k = P_{\hat{s}}\left(x_{k-1} - \gamma_k G_\nu(x_{k-1}, \xi_k; u_k)\right), \tag{3.4}$$

where $P_{\hat{s}}(x)$ keeps the top $\hat{s}$ largest absolute value of components of x and make the others 0.

Remark 4 *Remarkably, Algorithm 4 does not require any special operation to adapt to the sparsity assumption. This demonstrates an* implicit regularization *phenomenon exhibited by the zeroth-order stochastic gradient method in the high-dimensional setting when the performance is measured by the size of the gradient in the dual norm. We emphasize that the choice of the performance measure is motivated by the fact that we allow f to be nonconvex. Trivially, the result also applies to the case when f is convex, for the same performance measure.* ∎

3.2 Zeroth-order Stochastic Gradient Method for Convex Problems

We now consider the case when the function f is convex. In this setting, a more natural performance measure is the convergence of optimality gap in terms of the function values. For this situation, we propose and analyze a truncate variant of Algorithm 4 that demonstrates similar poly-logarithmic dependence on the dimensionality. To proceed, in addition to Assumption 4, we also make the following sparsity assumption on the optimal solution of problem (1.1).

Assumption 5 *Problem (1.1) has a sparse optimal solution x_* such that $\|x_*\|_0 \leq s^*$, where $s^* \approx s$.*

Our algorithm for the convex setting is presented in Algorithm 5. Note that this algorithm could be considered as a truncated variant of Algorithm 4 and a zeroth-order stochastic variant of the truncated gradient descent algorithm [17]. In the next result, we present convergence analysis of this algorithm.

Theorem 3.2 *Let $\{x_k\}_{k\geq 1}$ be generated by Algorithm 4, f is convex, Assumptions 1, 2, 4, and 5 hold. Also assume the stepsizes are chosen such that, $\forall k \geq 1$,*

$$\gamma_k = \frac{1}{4\hat{C}\hat{s}\log d}\min\left\{\frac{1}{12L\hat{s}\log d}, \sqrt{\frac{D_X^0\hat{C}\hat{s}}{3N\sigma^2}}\right\}, \quad \nu \leq \sqrt{\log d}\min\left\{\frac{\sigma}{\log d}, \sqrt{\frac{\hat{s}^2 D_X^0}{N}}\right\} \tag{3.5}$$

for some $\hat{C} \geq C$, $\hat{s} \geq \max\{s, s^*\}$, and $D_X^0 \geq \|x_0 - x_*\|^2$.

$$\mathbf{E}\left[f(\bar{x}_N) - f^*\right] \leq \frac{52L\hat{C}D_X^0\hat{s}^2(\log d)^2}{N} + \frac{69\sigma\sqrt{3\hat{C}D_X^0\hat{s}}\,\log d}{\sqrt{N}}, \tag{3.6}$$

where $\bar{x}_N = \frac{\sum_{k=0}^{N-1} x_k}{N}$. Hence, the total number of calls to the stochastic oracle (number of iterations) required to find an ϵ-optimal point of problem (1.1) is bounded by

$$\mathcal{O}\left(\hat{s}\left(\frac{\log d}{\epsilon}\right)^2\right). \tag{3.7}$$

Remark 5 *While for convex case, similar to the nonconvex case, the complexity of Algorithm 5 depends poly-logarithmically on d, it only linearly depends on the choice of $\hat{s}$, facilitating zeroth-order stochastic optimization in high-dimensions under sparsity assumptions.* ∎

Remark 6 *As discussed in detail in [34], both Assumption 4 and 5 are implied when we assume the function f depends on only s of the d coordinates. But, both Assumption 4 and 5 are comparatively weaker than that assumption. Furthermore, unlike [34], we do not make any assumption on the sparsity or smoothness of the second-order derivative of the objective function f for our results.* ∎

Remark 7 *As mentioned before, [34] considers only the convex case. Furthermore, their gradient estimator with zeroth-order oracle requires poly$(s, s^*, \log d)$ function queries in each iteration whereas our estimator is based on only one function query per iteration. Moreover, [34] requires computationally expensive debiased Lasso estimators whereas our method requires only simple thresholding operations (for convex case) to handle sparsity.* ∎

4 Future Work

Two concrete extensions are possible for future work. First, for our results, we focus on performance measures common in the optimization setting. It is interesting to extend our results to the bandit setting, where the performance is measured via regret of the algorithm. Next, the performance of conditional gradient algorithm in the high-dimensional constrained optimization setting is not well-explored; the interaction between the geometry of the constraint set, sparsity structure and zeroth-order information is extremely interesting to explore. Finally, lower bounds can be explored for the cases considered in this paper when f is nonconvex.

References

[1] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[2] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[3] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26. ACM, 2017.

[4] Krzysztof Choromanski, Mark Rowland, Vikas Sindhwani, Richard Turner, and Adrian Weller. Structured evolution with compact architectures for scalable policy optimization. In *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2018.

[5] Andrew Conn, Katya Scheinberg, and Luis Vicente. *Introduction to derivative-free optimization*, volume 8. Siam, 2009.

[6] V. Demyanov and A. Rubinov. *Approximate methods in optimization problems*. American Elsevier Publishing Co, 1970.

[7] John Duchi, Michael Jordan, Martin Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[8] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3:95–110, 1956.

[9] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[10] Saeed Ghadimi. Conditional gradient type methods for composite nonlinear and stochastic optimization. *Mathematical Programming*, 2018.

[11] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[12] Elad Hazan and Satyen Kale. Projection-free online learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1843–1850. Omnipress, 2012.

[13] Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In *International Conference on Machine Learning*, pages 1263–1271, 2016.

[14] Donald Hearn. The gap function of a convex program. *Operations Research Letters*, 2, 1982.

[15] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

[16] Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017.

[17] Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems*, pages 685–693, 2014.

[18] Kevin Jamieson, Robert Nowak, and Ben Recht. Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2012.

[19] Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *SIAM Journal on Optimization*, 26(2):1379–1409, 2016.

[20] Horia Mania, Aurelia Guy, and Benjamin Recht. Simple random search provides a competitive approach to reinforcement learning. In *Advances in Neural Information Processing Systems*, 2018.

[21] Jonas Mockus. *Bayesian approach to global optimization: theory and applications*, volume 37. Springer Science & Business Media, 2012.

[22] Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Conditional gradient method for stochastic submodular maximization: Closing the gap. In *International Conference on Artificial Intelligence and Statistics*, pages 1886–1895, 2018.

[23] Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *arXiv preprint arXiv:1804.09554*, 2018.

[24] A. S. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley, XV, 1983.

[25] Y. E. Nesterov. *Introductory Lectures on Convex Optimization: a basic course*. Kluwer Academic Publishers, Massachusetts, 2004.

[26] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[27] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[28] Sashank Reddi, Suvrit Sra, Barnabás Póczos, and Alexander Smola. Stochastic Frank-Wolfe Methods for Nonconvex Optimization. *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1244–1251, 2016.

[29] Reuven Rubinstein and Dirk Kroese. *Simulation and the Monte Carlo method*, volume 10. John Wiley & Sons, 2016.

[30] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

[31] Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, pages 3–24, 2013.

[32] Jasper Snoek, Hugo Larochelle, and Ryan Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[33] James Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons, 2005.

[34] Yining Wang, Simon Du, Sivaraman Balakrishnan, and Aarti Singh. Stochastic zeroth-order optimization in high dimensions. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, 2018.

A Relevant Results from [27]

In this section, for completeness, we replicate relevant results from [27], that are required for our proofs.

Theorem A.1 *The following statements hold for any function f whose gradient is Lipschitz continuous with constant L.*

a) The gradient of f_ν is Lipschitz continuous with constant L_ν such that $L_\nu \leq L$.

b) For any $x \in \mathbb{R}^d$,

$$|f_\nu(x) - f(x)| \leq \frac{\nu^2}{2} Ld, \tag{A.8}$$

$$\|\nabla f_\nu(x) - \nabla f(x)\| \leq \frac{\nu}{2} L(d+3)^{\frac{3}{2}}. \tag{A.9}$$

c) For any $x \in \mathbb{R}^n$,

$$\frac{1}{\nu^2}\mathbf{E}_u[\{f(x+\nu u) - f(x)\}^2 \|u\|^2] \leq \frac{\nu^2}{2} L^2 (d+6)^3 + 2(d+4)\|\nabla f(x)\|^2. \tag{A.10}$$

B Proofs for Section 2

We present all the proofs for Section 2 below. Recall that, we assumed that $\|\cdot\| = \|\cdot\|_2$ in Section 2. In order to prove Theorem 2.1, we need the following result that provides upper bounds for the variance of our gradient estimator.

Lemma B.1 *Let $\bar{G}_\nu^k$ be computed by (2.1). Then under Assumptions 1, 2 and 3, we have*

$$\mathbf{E}[\|\bar{G}_\nu^k - \nabla f_\nu(z_{k-1})\|^2] \leq \frac{2(d+5)(B^2+\sigma^2)}{m_k} + \frac{\nu^2}{2m_k} L^2 (d+3)^3, \tag{B.11}$$

$$\mathbf{E}[\|\bar{G}_\nu^k - \nabla f(z_{k-1})\|^2] \leq \frac{4(d+5)(B^2+\sigma^2)}{m_k} + \frac{3\nu^2}{2} L^2 (d+3)^3. \tag{B.12}$$

Proof. First note that using (A.10) for function F instead of f, under Assumptions 1 and 2, we obtain

$$\begin{aligned}\mathbf{E}[\|G_\nu^{k,j}\|^2] &\leq \tfrac{\nu^2 L^2}{2}\mathbf{E}\left[\|u\|^6\right] + 2\mathbf{E}_\xi[\|\nabla F(z_{k-1},\xi_k)\|^2]\mathbf{E}_u\left[\|u\|^4\right] \\ &\leq \tfrac{\nu^2 L^2}{2}(d+6)^3 + 2\left[\|\nabla f(z_{k-1})\|^2 + \sigma^2\right](d+4),\end{aligned}$$

where the second inequality follows from the fact that $\mathbf{E}[\|u\|^k] \leq (d+k)^{k/2}$ for any $k \geq 2$ due to Nesterov [27]. Also noting (1.4), (2.4), and the fact that $\|\nabla f_\nu\| \leq B$ under Assumption 3, we have

$$\mathbf{E}[\|\bar{G}_\nu^k - \nabla f_\nu(z_{k-1})\|^2] \leq \frac{1}{m_k}\left(\mathbf{E}[\|G_\nu^{k,j}\|^2] + B^2\right),$$

which together with the above relation clearly imply (B.11). We can then obtain (B.12) by noting (A.9) and the fact that

$$\mathbf{E}[\|\bar{G}_\nu^k - \nabla f(z_{k-1})\|^2] \leq 2\mathbf{E}[\|\bar{G}_\nu^k - \nabla f_\nu(z_{k-1})\|^2] + 2\mathbf{E}[\|\nabla f_\nu(z_{k-1}) - \nabla f(z_{k-1})\|^2].$$

■

B.1 Proof of Theorem 2.1

Proof. Denoting $\Delta_k = \bar{G}_\nu^k - \nabla f(z_{k-1})$, noting (1.2), (2.3), and (2.5), we have

$$\begin{aligned}f(z_k) &\leq f(z_{k-1}) + \langle \nabla f(z_{k-1}), z_k - z_{k-1}\rangle + \frac{L}{2}\|z_k - z_{k-1}\|^2 \\ &= f(z_{k-1}) + \alpha_k\langle \nabla f(z_{k-1}), x_k - z_{k-1}\rangle + \frac{L\alpha_k^2}{2}\|x_k - z_{k-1}\|^2 \\ &\leq f(z_{k-1}) + \alpha_k\langle \nabla f(z_{k-1}), \hat{x}_k - z_{k-1}\rangle + \frac{L\alpha_k^2}{2}\left[\|x_k - z_{k-1}\|^2 + \|x_k - \hat{x}_k\|^2\right] + \frac{\|\Delta_k\|^2}{2L} \\ &\leq f(z_{k-1}) - \alpha_k g_{\mathcal{X}}^k + LD_{\mathcal{X}}^2\alpha_k^2 + \frac{\|\Delta_k\|^2}{2L},\end{aligned} \tag{B.13}$$

where the last inequality follows from boundedness of the feasible set, (2.5), and the fact that

$$\langle \nabla f(z_{k-1}) + \Delta_k, x_k - u \rangle \leq 0 \ \ \forall u \in \mathcal{X}$$

due to the optimality condition of (2.2). Taking expectation from both sides of the above inequality, summing them up, rearranging the terms, and noting Lemma B.1, we obtain

$$\sum_{k=1}^{N} \alpha_k \mathbf{E}[g_{\mathcal{X}}^k] \leq f(z_0) - f^* + LD_{\mathcal{X}}^2 \sum_{k=1}^{N} \alpha_k^2 + \frac{\nu^2}{2} LN(d+3)^3 + \frac{2(d+5)(B^2+\sigma^2)}{L} \sum_{k=1}^{N} \frac{1}{m_k}.$$

Hence, choosing $\alpha_k = \alpha_1$ and $m_k = m_1$ for all $k \geq 1$, and noting that R is a uniform random variable, we have

$$\mathbf{E}[g_{\mathcal{X}}^R] = \frac{\sum_{k=1}^N \mathbf{E}[g_{\mathcal{X}}^k]}{N} = \frac{\sum_{k=1}^N \alpha_k \mathbf{E}[g_{\mathcal{X}}^k]}{\sum_{k=1}^N \alpha_k} \leq \frac{f(z_0) - f^*}{N\alpha_1} + LD_{\mathcal{X}}^2 \alpha_1 + \frac{\nu^2}{2\alpha_1} L(d+3)^3 + \frac{2(d+5)(B^2+\sigma^2)}{L\alpha_1 m_1},$$

which together with (2.7) imply (2.8). Hence, (2.9) follows by noting that the total number of calls to the stochastic oracle is bounded by $\sum_{k=1}^N m_k$.

Now assume that f is convex. Hence, by (2.6) and (B.13), we have

$$f(z_k) - f(x_*) \leq (1 - \tfrac{\alpha_k}{2})(f(z_{k-1} - f(x_*)) - \frac{\alpha_k g_{\mathcal{X}}^k}{2} + LD_{\mathcal{X}}^2 \alpha_k^2 + \frac{\|\Delta_k\|^2}{2L}$$

Taking expectation from both sides of the above inequality, dividing them by T_k, and summing them up, and noting (2.12), we obtain

$$\frac{\mathbf{E}[f(z_N)] - f^*}{\Gamma_N} + \sum_{k=1}^{N} \frac{\alpha_k \mathbf{E}[g_{\mathcal{X}}^k]}{2\Gamma_k} \leq f(z_0) - f^* + LD_{\mathcal{X}}^2 \sum_{k=1}^{N} \frac{\alpha_k^2}{\Gamma_k} + \frac{1}{2L} \sum_{k=1}^{N} \frac{\mathbf{E}\left[\|\Delta_k\|^2\right]}{\Gamma_k},$$

which together with the fact that

$$\sum_{k=1}^{N} \frac{\alpha_k}{2\Gamma_k} = \frac{1 - \Gamma_N}{\Gamma_N}, \ \ 1 - \Gamma_1 \leq 1 - \Gamma_N \leq 1$$

due to (2.12), imply that

$$\mathbf{E}[f(z_N)] - f^* + \mathbf{E}[g_{\mathcal{X}}^R] \leq \frac{\Gamma_N}{1 - \Gamma_N} \left[f(z_0) - f^* + LD_{\mathcal{X}}^2 \sum_{k=1}^{N} \frac{\alpha_k^2}{\Gamma_k} + \frac{2(d+5)(B^2+\sigma^2)}{L} \sum_{k=1}^{N} \frac{1}{\Gamma_k m_k} + \frac{\nu^2}{2} L(d+3)^3 \sum_{k=1}^{N} \frac{1}{\Gamma_k} \right] \tag{B.14}$$

Now noting (2.10) and (2.12), we have

$$\Gamma_k = \frac{60}{(k+3)(k+4)(k+5)}, \qquad \sum_{k=1}^{N} \frac{\alpha_k^2}{\Gamma_k} \leq \sum_{k=1}^{N} \frac{3(k+3)}{5} = \frac{3N(N+7)}{10},$$

$$\Gamma_N \sum_{k=1}^{N} \frac{1}{\Gamma_k m_k} \leq \frac{1}{4(d+5)B_{L\sigma}N}, \qquad \Gamma_N \sum_{k=1}^{N} \frac{1}{\Gamma_k} \leq N.$$

Combining the above relations, we get (2.11) and (2.13). ∎

B.2 Proof of Theorem 2.2

Proof. First, note that by (1.2), we have

$$\begin{aligned} f_\nu(z_k) &\leq f_\nu(w_k) + \langle \nabla f_\nu(w_k), z_k - w_k \rangle + \frac{L}{2}\|z_k - w_k\|^2 \\ &\leq (1-\alpha_k) f_\nu(z_{k-1}) + \alpha_k \left[f_\nu(w_k) + \langle \nabla f_\nu(w_k), x_k - w_k \rangle \right] \\ &+ \frac{L\alpha_k^2}{2}\|x_k - x_{k-1}\|^2, \end{aligned} \tag{B.15}$$

where the second inequality follows from convexity of f_ν, (2.16), and (2.18). Also note that by (2.14) and (2.17), we have

$$-\mu_k \le \langle \bar{G}_\nu^k + \gamma_k(x_k - x_{k-1}), u - x_k\rangle \qquad \forall u \in \mathcal{X}. \tag{B.16}$$

Letting $u = x_*$ in the above inequality and multiplying it by α_k, summing it up with (B.15), and denoting $\bar{\Delta}_k = \bar{G}_\nu^k - \nabla f_\nu(w_k)$, we obtain

$$f_\nu(z_k) \le (1-\alpha_k) f_\nu(z_{k-1}) + \alpha_k f_\nu(x_*) + \alpha_k \left[\mu_k + \langle \bar{\Delta}_k + \gamma_k(x_k - x_{k-1}), x_* - x_k\rangle\right] + \frac{L\alpha_k^2}{2}\|x_k - x_{k-1}\|^2,$$

which together with the facts that

$$\|x_{k-1} - x_*\|^2 = \|x_k - x_{k-1}\|^2 + \|x_k - x_*\|_2^2 + 2\langle x_{k-1} - x_k, x_k - x_*\rangle,$$
$$\alpha_k\langle \bar{\Delta}_k, x_* - x_k\rangle \le \alpha_k\langle \bar{\Delta}_k, x_* - x_{k-1}\rangle + \frac{\|\bar{\Delta}_k\|^2}{2L} + \frac{L\alpha_k^2}{2}\|x_k - x_{k-1}\|^2,$$

imply

$$\begin{aligned} f_\nu(z_k) \le (1-\alpha_k) f_\nu(z_{k-1}) + \alpha_k f_\nu(x_*) \quad &+ \quad \alpha_k\left[\mu_k + \frac{2L\alpha_k - \gamma_k}{2}\|x_k - x_{k-1}\|^2 + \langle \bar{\Delta}_k, x_* - x_{k-1}\rangle\right] \\ &+ \quad \frac{\alpha_k\gamma_k}{2}\left[\|x_{k-1} - x_*\|^2 - \|x_k - x_*\|^2\right] + \frac{\|\bar{\Delta}_k\|^2}{2L}. \end{aligned} \tag{B.17}$$

Defining

$$\hat{\Gamma}_k = \prod_{i=2}^{k}(1-\alpha_i), \;\; \hat{\Gamma}_1 = 1, \tag{B.18}$$

subtracting $f_\nu(x_*)$ from both sides of the above inequality, diving them by $\hat{\Gamma}_k$, taking expectation, summing them up, noting (A.8) assuming that $\alpha_1 = 1$, $\gamma_k \ge 2L\alpha_k$, and $\gamma_k\alpha_k/\hat{\Gamma}_k$ is constant for any $k \ge 1$, we obtain

$$\frac{\mathbf{E}\left[f_(z_N)\right] - f(x_*) - \nu^2 Ld}{\hat{\Gamma}_N} \le \frac{\gamma_1}{2}\|x_0 - x_*\|^2 + \sum_{k=1}^{N}\frac{\alpha_k\mu_k}{\hat{\Gamma}_k} + \left[\frac{(d+5)(B^2+\sigma^2)}{L} + \frac{\nu^2 L(d+3)^3}{2}\right]\sum_{k=1}^{N}\frac{1}{m_k\hat{\Gamma}_k}.$$

Now noticing that

$$\hat{\Gamma}_k = \frac{2}{k(k+1)}, \qquad \frac{\alpha_k\gamma_k}{\hat{\Gamma}_k} = 4L, \qquad \frac{\alpha_k\mu_k}{\hat{\Gamma}_k} = \frac{LD_0^2}{N},$$
$$\frac{1}{m_k\hat{\Gamma}_k} \le \frac{2D_0^2}{\max\{(d+5)B_{L\sigma}N, d+3\}}$$

due to (2.19) and (B.18), we obtain (2.20).

Furthermore, note that the function h_γ defined in Algorithm 2 is indeed negative the FW-gap of the CG method applied to problem (2.15). From classical analysis of the CG method and similar to our result in Theorem 2.1, one can show that the FW-gap is bounded by $LD_{\mathcal{X}}^2/T$ if the CG method runs for T iteration. Since the gradient of the objective function in (2.15) is Lipschitz continuous with constant γ, we have

$$-h_{\gamma_k}(\bar{y}_{T_k}) \le \frac{\gamma_k D_{\mathcal{X}}^2}{T_k},$$

which together with the choice of μ_k and γ_k in (2.19), imply that at iteration k of Algorithm 1, we need to run Algorithm 2 for at most $T_k = 4D_{\mathcal{X}}^2 N/D_0^2$ iterations. Therefore, the total number of iterations of Algorithm 2 to find an ϵ-stationary point of problem (1.1) is bounded by $\sum_{k=1}^{N} T_k \le 48LD_{\mathcal{X}}^2/\epsilon^2$ due to (2.21).

■

C Proofs for Section 3

We now present the proofs for section 3. Recall that, we assumed that $\|\cdot\| = \|\cdot\|_\infty$ in Section 3 We first present two technical results which play key roles in our convergence analysis.

Lemma C.1 *Let $u \sim N(0, I_d)$ be a d-dimensional standard Gaussian vector. Then for all integer $k \geq 1$ and for some universal constant C, we have $\mathbf{E}\left[\|u\|_\infty^k\right] \leq C(2\log d)^{k/2}$.*

Proof. Let $Z = \|u\|_\infty$ and denote by $p(x)$ the standard normal pdf. Note that we have

$$\begin{aligned}\mathbf{E}Z^k &= \int_0^\infty kx^{k-1}P(Z > x)\,dx \\ &\leq \int_0^{x_d} kx^{k-1}dx + \int_{x_d}^\infty x^{k-2}p(x)dx\end{aligned}$$

where we define $x_d = \sqrt{2\ln d}$. Now we have

$$\int_0^{x_d} kx^{k-1}dx = x_d^k = (2\log d)^{k/2}$$

and by l'Hospital's rule, for large d we have

$$\int_{x_d}^\infty x^{k-2}p(x)dx \approx x_d^{k-3}p(x_d) \ll (\log d)^{(k-3)/2} = o\left(\frac{(\log d)^{k/2}}{d}\right)$$

Hence we have for some universal constant C,

$$\mathbf{E}\left[\|u\|_\infty^k\right] \leq C(2\log d)^{k/2}.$$

■

Lemma C.2 *The following statements hold for function f and its smooth approximation f_ν.*

a) Under Assumptions 1 and 2, gradient of f is Lipschitz continuous with constant L and

$$|f_\nu(x) - f(x)| \leq \nu^2 CL\log d.$$

b) If Assumption 4 also holds, we have

$$\begin{aligned}&\|\nabla f_\nu(x) - \nabla f(x)\|_2 \leq C\nu L\sqrt{2s}(\log d)^{3/2} \\ &\mathbf{E}\left[\|G_\nu(x,\xi;u)\|_\infty^2\right] \leq 4C(\log d)^2\left[L^2\nu^2(\log d) + 4\|\nabla f(x)\|_1^2 + 4\sigma^2\right].\end{aligned}$$

Proof. First note that

$$\begin{aligned}|f_\nu(x) - f(x)| &= |\mathbf{E}\left[f(x+\nu u) - f(x) - \nu\langle\nabla f(x), u\rangle\right]| \\ &\leq \mathbf{E}\left|f(x+\nu u) - f(x) - \nu\langle\nabla f(x), u\rangle\right| \\ &\leq \frac{\nu^2 L}{2}\mathbf{E}\left[\|u\|_\infty^2\right] \leq C\nu^2 L\log d,\end{aligned}$$

where the last inequality follows from Lemma C.1. Second, noting this lemma again, Assumption 4, and part a), we have

$$\begin{aligned}\|\nabla f_\nu(x) - \nabla f(x)\|_2 &\leq \sqrt{s^*}\|\nabla f_\nu(x) - \nabla f(x)\|_\infty \\ &\leq \frac{\sqrt{s}}{\nu(2\pi)^{d/2}}\int \left|f(x+\nu u) - f(x) - \nu\langle\nabla f(x), u\rangle\right| \|u\|_\infty e^{-\frac{\|u\|_2^2}{2}}\,du \\ &\leq \frac{\nu L\sqrt{s}}{2(2\pi)^{d/2}}\int \|u\|_\infty^3 e^{-\frac{\|u\|_2^2}{2}}\,du \leq C\nu L\sqrt{2s}(\log d)^{3/2}.\end{aligned}$$

Furthermore, by (1.4), Holder inequality, Lemma C.1, and under Assumption 4 we have

$$\begin{aligned}
&\mathbf{E}\left[\|G_\nu(x,\xi;u)\|_\infty^2\right] \\
&= \frac{2}{\nu^2}\mathbf{E}\left[|F(x+\nu u,\xi) - F(x,\xi) - \nu\langle\nabla F(x,\xi),u\rangle|^2\|u\|_\infty^2\right] + 2\mathbf{E}\left[\langle\nabla F(x,\xi),u\rangle^2\|u\|_\infty^2\right] \\
&\le \tfrac{\nu^2L^2}{2}\mathbf{E}\left[\|u\|_\infty^6\right] + 2\mathbf{E}_\xi[\|\nabla F(x,\xi)\|_1^2]\mathbf{E}_u\left[\|u\|_\infty^4\right] \\
&\le 4CL^2\nu^2(\log d)^3 + 8C(\log d)^2\mathbf{E}_\xi[\|\nabla F(x,\xi)\|_1^2] \\
&\le 4C(\log d)^2\left[L^2\nu^2(\log d) + 4\|\nabla f(x)\|_1^2 + 4\sigma^2\right].
\end{aligned}$$

■

C.1 Proof of Theorem 3.1

Proof. Noting (1.4), Lemma C.2.a), and with the notion of $G_{\nu,k} \equiv G_\nu(x_k,\xi_k,u_k)$, we have

$$\begin{aligned}
f(x_{k+1}) &\le f(x_k) + \langle\nabla f(x_k), x_{k+1}-x_k\rangle + \frac{L}{2}\|x_{k+1}-x_k\|_\infty^2 \\
&\le f(x_k) - \gamma_k\langle\nabla f(x_k), G_{\nu,k}\rangle + \frac{L\gamma_k^2}{2}\|G_{\nu,k}\|_\infty^2,
\end{aligned}$$

which after taking expectation imply that

$$\begin{aligned}
\mathbf{E}[f(x_{k+1})] &\le f(x_k) - \gamma_k\|\nabla f(x_k)\|_2^2 + \gamma_k\langle\nabla f(x_k), \nabla f(x_k) - \nabla f_\nu(x_k)\rangle + \frac{L\gamma_k^2}{2}\mathbf{E}[\|G_{\nu,k}\|_\infty^2] \\
&\le f(x_k) - \frac{\gamma_k}{2}\|\nabla f(x_k)\|_2^2 + \frac{\gamma_k}{2}\|\nabla f(x_k) - \nabla f_\nu(x_k)\|_2^2 + \frac{L\gamma_k^2}{2}\mathbf{E}[\|G_{\nu,k}\|_\infty^2] \\
&\le f(x_k) - \frac{\gamma_k}{2s}\left(1 - 16LCs(\log d)^2\gamma_k\right)\|\nabla f(x_k)\|_1^2 + (\nu LC)^2 s(\log d)^3\gamma_k \\
&\quad + 2LC(\log d)^2\left[L^2\nu^2(\log d) + 4\sigma^2\right]\gamma_k^2,
\end{aligned}$$

where the last inequality follow from Holder inequality and Lemma C.2.b). Summing both sides of the above inequality over the iterations and rearranging terms, we get

$$\mathbf{E}[\|\nabla f(x_R)\|_1^2] \le \frac{6s\left[f(x_0) - f^* + (\nu LC)^2 s(\log d)^3\sum_{k=1}^N\gamma_k + 2CL(\log d)^2\left(L^2\nu^2(\log d) + 4\sigma^2\right)\sum_{k=0}^{N-1}\gamma_k^2\right]}{\sum_{k=0}^{N-1}\gamma_k},$$

where R is uniformly distributed over $\{0,\ldots,N-1\}$ since

$$\mathbf{E}[\|\nabla f(x_R)\|_1^2] = \frac{1}{N}\sum_{k=0}^{N-1}\|\nabla f(x_k)\|_1^2 = \frac{\sum_{k=0}^{N-1}\gamma_k\left(1 - 16LCs(\log d)^2\gamma_k\right)\|\nabla f(x_k)\|_1^2}{\sum_{k=0}^{N-1}\gamma_k\left(1 - 16LCs(\log d)^2\gamma_k\right)},$$

due to the constant choice of γ_k in (3.1). Therefore, we have

$$\mathbf{E}[\|\nabla f(x_R)\|_1^2] \le 6s\left[\frac{f(x_0) - f^*}{N\gamma_1} + (\nu LC)^2 s(\log d)^3 + 2CL(\log d)^2\left(L^2\nu^2(\log d) + 4\sigma^2\right)\gamma_1\right],$$

which together with the choice of smoothing parameter in (3.1) imply (3.2). ■

C.2 Proof of Theorem 3.2

Proof. Denoting the index set of nonzero elements of x_k and x_* by $Z^k \subseteq \mathbb{R}^{\hat{s}}$ and $Z^* \subseteq \mathbb{R}^{s^*}$, respectively, and $J^k = Z^k \cup Z^{k+1} \cup Z^*$, we have

$$\begin{aligned}
&\|x_{k+1} - x_*\|_2^2 \\
&= \|x_{k+1}^{J^k} - x_*^{J^k}\|_2^2 = \|x_k^{J^k} - x_*^{J^k} - \gamma_k G_{\nu,k}^{J^k}\|_2^2 = \|x_k^{J^k} - x_*^{J^k}\|_2^2 + \gamma_k^2\|G_{\nu,k}^{J^k}\|_2^2 - 2\gamma_k\langle x_k^{J^k} - x_*^{J^k}, \gamma_k G_{\nu,k}^{J^k}\rangle \\
&\le \|x_k - x^*\|_2^2 + (2\hat{s} + s^*)\gamma_k^2\|G_{\nu,k}\|_\infty^2 - 2\gamma_k\langle x_k - x_*, G_{\nu,k}\rangle,
\end{aligned}$$

where the inequality follows from the facts that $|J^k| \leq 2\hat{s} + s^*$ and $\|G^{J^k}_{\nu,k}\| \leq \|G_{\nu,k}\|$. Taking expectation from both sides of the above inequality, summing them up, noting Lemma C.2, convexity of f_ν (due to convexity of f), we have

$$\begin{aligned}
\mathbf{E}\left[\|x_N - x^*\|_2^2\right] &\leq \|x_0 - x^*\|_2^2 + (2\hat{s} + s^*) \sum_{k=0}^{N-1} \gamma_k^2 \mathbf{E}\left[\|G_{\nu,k}\|_\infty^2\right] - 2\sum_{k=0}^{N-1} \gamma_k \langle x_k - x^*, \nabla f_\nu(x_k)\rangle \\
&\leq \|x_0 - x^*\|_2^2 + 4C(2\hat{s} + s^*)(\log d)^2 \sum_{k=0}^{N-1} \gamma_k^2 \left[L^2\nu^2(\log d) + 4\|\nabla f(x_k)\|_1^2 + 4\sigma^2\right] \\
&\quad - 2\sum_{k=0}^{N-1} \gamma_k \left[f_\nu(x_k) - f_\nu(x_*)\right] \\
&\leq \|x_0 - x^*\|_2^2 + 4C(2\hat{s} + s^*)(\log d)^2 \sum_{k=0}^{N-1} \gamma_k^2 \left[L^2\nu^2(\log d) + 4\sigma^2\right] + 4\nu^2 CL \log d \sum_{k=0}^{N-1} \gamma_k \\
&\quad - 2\sum_{k=0}^{N-1} \gamma_k [1 - 16LCs(2\hat{s} + s^*)(\log d)^2 \gamma_k][f(x_k) - f(x_*)],
\end{aligned}$$

where the last inequality follows from the fact that $f(x_k) - f(x_*) \geq 1/(2Ls)\|\nabla f(x_k)\|_2^2$ due to the convexity of f and sparsity of its gradient. Rearranging the terms in the above inequality and noting that $\bar{x}_N = \frac{\sum_{k=0}^{N-1} x_k}{N}$, we obtain

$$f(\bar{x}_N) - f(x_*) \leq \frac{\|x_0 - x^*\|_2^2 + 4C(2\hat{s} + s^*)(\log d)^2 \sum_{k=0}^{N-1} \gamma_k^2 \left[L^2\nu^2(\log d) + 4\sigma^2\right] + 4\nu^2 CL \log d \sum_{k=0}^{N-1} \gamma_k}{2\sum_{k=0}^{N-1} \gamma_k [1 - 16LCs(2\hat{s} + s^*)(\log d)^2 \gamma_k]}$$

since

$$\bar{x}_N = \frac{\sum_{k=0}^{N-1} x_k}{N} = \frac{\gamma_k [1 - 16LCs(2\hat{s} + s^*)(\log d)^2 \gamma_k] x_k}{\sum_{k=0}^{N-1} \gamma_k [1 - 16LCs(2\hat{s} + s^*)(\log d)^2 \gamma_k]}$$

due to the constant choice of γ_k in (3.5). Hence, (3.6) follows by using the choice of parameters in (3.5) into the above relation. ∎

D Zeroth-order Stochastic Gradient Method with Inexact Updates-Nonconvex case

In this section, we present a zeroth-order stochastic gradient method which applies the CG method to solve the subproblems. This algorithm shares the main idea of Algorithm 3, but for nonconvex problems. We show while this algorithm enjoys better complexity bound than Algorithm 3, it possess the same one when the same performance measure is employed.

Algorithm 6 Zeroth-order Stochastic Gradient Method with Inexact Updates

Input: $x_0 \in \mathcal{X}$, smoothing parameter $\nu > 0$, positive integer sequence m_k, and sequences γ_k and μ_k and a probability distribution $P_R(\cdot)$ over $\{0, \ldots, N-1\}$
for $k = 1, \ldots, N$ **do**
Generate $u_k = [u_{k,1}, \ldots, u_{k,m_k}]$, where $u_{k,j} \sim N(0, I_d)$, call the stochastic oracle m_k times, compute $\bar{G}_\nu^k \equiv \bar{G}_\nu(x_{k-1}, \xi_k, u_k)$ as given by (2.1), and set x_k to (2.17).
end for
Output: Generate R according to $P_R(\cdot)$ and output x_R.

Since we are now using the CG method for inexactly solving (2.15), we can provide an alternative termination criterion than the FW-gap given in (2.5) to provide our convergence analysis. In particular, we use the gradient mapping defined as

$$GP_{\mathcal{X}}(x, g, \gamma) = \gamma(x - P_{\mathcal{X}}(x, g, \gamma)), \tag{D.19}$$

where $P_{\mathcal{X}}$ is the solution to (2.15). This quantity which has been widely used in the literature as a convergence criteria for solving nonconvex problems (see, e.g., [24, 25]), plays an analogues role of the gradient in constrained problems. Next result provides some properties for this criteria.

Lemma D.1 *Let $P_{\mathcal{X}}(\cdot)$ be defined in (2.15), $\gamma > 0$, and $x \in \mathcal{X}$ are given.*

a) for and $\hat{g} \in \mathbb{R}^d$, we have

$$\|P_{\mathcal{X}}(x, g, \gamma) - P_{\mathcal{X}}(x, \hat{g}, \gamma)\| \leq \frac{\|g - \hat{g}\|}{\gamma}.$$

b) Let $P^{\mu}_{\mathcal{X}}$ be the inexact solution of (2.15) such that

$$\langle g + \gamma(P^{\mu}_{\mathcal{X}}(x, g, \gamma) - x), u - P^{\mu}_{\mathcal{X}}(x, g, \gamma)\rangle \geq -\mu \qquad \forall u \in \mathcal{X} \tag{D.20}$$

for some $\mu \geq 0$. Then, we have

$$\|P_{\mathcal{X}}(x, g, \gamma) - P^{\mu}_{\mathcal{X}}(x, g, \gamma)\|^2 \leq \frac{\mu}{\gamma}.$$

c) Let $g_{\mathcal{X}}(\cdot)$ be the Frank-Wolfe gap defined in (2.5). Then we have

$$\|GP_{\mathcal{X}}(x, \nabla f(x), \gamma)\|^2 \leq g_{\mathcal{X}}(x).$$

Moreover, under Assumption 3, we have

$$g_{\mathcal{X}}(x) \leq (B/\gamma + D_{\mathcal{X}})\|GP_{\mathcal{X}}(x, \nabla f(x), \gamma)\|.$$

Proof. First note that (2.15) implies

$$\|P_{\mathcal{X}}(x, g, \gamma) - P_{\mathcal{X}}(x, \hat{g}, \gamma)\| = \|\Pi_{\mathcal{X}}(x - g/\gamma) - \Pi_{\mathcal{X}}(x - \hat{g}/\gamma)\| \leq \frac{\|g - \hat{g}\|}{\gamma},$$

where the last inequality follows from Lipschitz continuity of the Euclidian projection over the feasible set $\Pi_{\mathcal{X}}$. Second, by optimality condition of (2.15), we have

$$\langle g + \gamma(P_{\mathcal{X}}(x, g, \gamma) - x), u - P_{\mathcal{X}}(x, g, \gamma)\rangle \geq 0 \qquad \forall \tilde{u} \in \mathcal{X}. \tag{D.21}$$

Letting $\tilde{u} = P^{\mu}_{\mathcal{X}}(x, g, \gamma)$ in the above inequality and $u = P_{\mathcal{X}}(x, g, \gamma)$ and $g = \nabla f(x)$ in (D.20) and summing them up, we clear get the result in part b). Third, letting $\tilde{u} = x$ in (D.21), we have

$$\|GP_{\mathcal{X}}(x, \nabla f(x), \gamma)\|^2 \leq \gamma\langle \nabla f(x), x - P_{\mathcal{X}}(x, \nabla f(x), \gamma)\rangle \leq \gamma g_{\mathcal{X}}(x),$$

where the last inequality follows from (2.5). Furthermore, (D.21) also implies that

$$\begin{aligned} g_{\mathcal{X}}(x) + \frac{1}{\gamma}\|GP_{\mathcal{X}}(x, \nabla f(x), \gamma)\|^2 &\leq \langle \nabla f(x) + \gamma(x - u), x - P_{\mathcal{X}}(x, \nabla f(x), \gamma)\rangle \\ &\leq (B/\gamma + D_{\mathcal{X}})\|GP_{\mathcal{X}}(x, \nabla f(x), \gamma)\|, \end{aligned}$$

where the last inequality follows from Assumption 3.

■

Now we are ready to state the main result for the nonconvex case.

Theorem D.1 *Let $\{x_k\}$ be generated by Algorithm 6, the function f be nonconvex, and*

$$\nu = \sqrt{\frac{1}{2N(d+3)^3}}, \;\; \gamma_k = 2L, \;\; \mu_k = \frac{1}{4N}, \;\; m_k = 6(d+5)N, \;\; \forall k \geq 1. \tag{D.22}$$

Then under Assumptions 1, 2, and 3, we have

$$\mathbf{E}[\|GP_{\mathcal{X}}(x_R, \nabla f(x_R), \gamma_R)\|^2] \leq \frac{8L}{N}\left(f(x_0) - f^* + L + B^2 + \sigma^2\right). \tag{D.23}$$

where R is uniformly distributed over $\{0, \ldots, N-1\}$ and $g_{\mathcal{X}}$ is defined in (D.19). Hence, the total number of calls to the stochastic oracle and linear subproblems solved to find and ϵ-stationary point of problem (1.1) are, respectively, bounded by

$$\mathcal{O}\left(\frac{d}{\epsilon^2}\right), \;\; \mathcal{O}\left(\frac{1}{\epsilon^2}\right). \tag{D.24}$$

Proof. First note that by (1.2), we have

$$f(x_k) \leq f(x_{k-1}) + \langle \nabla f(x_{k-1}), x_k - x_{k-1} \rangle + \frac{L}{2}\|x_k - x_{k-1}\|^2.$$

Letting $u = x_{k-1}$ in (B.16), summing it up with the above inequality, and denoting $\Delta_k = \bar{G}_\nu^k - \nabla f(x_{k-1})$, we obtain

$$\begin{aligned} f(x_k) &\leq f(x_{k-1}) - \gamma_k \left(1 - \frac{L}{2\gamma_k}\right) \|x_k - x_{k-1}\|^2 + \langle \Delta_k, x_{k-1} - x_k \rangle + \mu_k \\ &\leq f(x_{k-1}) - \gamma_k \left(1 - \frac{L}{\gamma_k}\right) \|x_k - x_{k-1}\|^2 + \frac{\|\Delta_k\|^2}{2L} + \mu_k. \end{aligned}$$

Taking expectation from the above inequalities, summing them up, re-arranging the terms, and in the view of Lemma B.1, we have

$$\begin{aligned} &\sum_{k=1}^N \gamma_k \left(1 - \frac{L}{\gamma_k}\right) \mathbf{E}[\|x_k - x_{k-1}\|^2] \\ &\leq f(x_0) - f^* + \sum_{k=1}^N \mu_k + \frac{\nu^2 L (d+3)^3 N}{2} + \frac{2(d+5)(B^2+\sigma^2)}{L} \sum_{k=1}^N \frac{1}{m_k}, \end{aligned}$$

which together with the facts that $x_k = P_{\mathcal{X}}^{\mu_k}(x_{k-1}, \bar{G}_\nu^k, \gamma_k)$ and

$$\begin{aligned} &\frac{1}{\gamma_k^2}\|GP_{\mathcal{X}}(x_{k-1}, \nabla f(x_{k-1}), \gamma_k)\|^2 \\ &= \|x_{k-1} - P_{\mathcal{X}}(x_{k-1}, \nabla f(x_{k-1}), \gamma_k)\|^2 \\ &\leq 2\|x_k - x_{k-1}\|^2 + \frac{4\mu_k}{\gamma_k} + \frac{4\nu^2 L^2 (d+3)^3}{\gamma_k^2} + \frac{16(d+5)(B^2+\sigma^2)}{\gamma_k^2 m_k}, \end{aligned}$$

imply that

$$\begin{aligned} \sum_{k=1}^N \left(\frac{\gamma_k - L}{2\gamma_k^2}\right) \mathbf{E}[\|GP_{\mathcal{X}}(x_{k-1}, \nabla f(x_{k-1}), \gamma_k)\|^2] &\leq f(x_0) - f^* + \sum_{k=1}^N \left(\frac{3\gamma_k - 2L}{\gamma_k}\right) \mu_k \\ + \frac{\nu^2 L (d+3)^3}{2} \sum_{k=1}^N \left(1 + \frac{4L(\gamma_k - L)}{\gamma_k^2}\right) + \frac{2(d+5)(B^2+\sigma^2)}{L} \sum_{k=1}^N \frac{1}{m_k} \left(1 + \frac{8L(\gamma_k - L)}{\gamma_k^2}\right). \end{aligned}$$

Hence, noting (D.22), we obtain

$$\begin{aligned} &\mathbf{E}[\|GP_{\mathcal{X}}(x_R, \nabla f(x_R), \gamma_R)\|^2] \\ &\leq \frac{8L[f(x_0) - f^*]}{N} + 16L^2\mu_1 + 8\nu^2 L^2 (d+3)^3 + \frac{48(d+5)(B^2+\sigma^2)}{m_1}, \end{aligned}$$

which implies (D.23). Rest of the proof is similar to that of Theorem 2.2 and hence we skip the details. ∎

Remark 8 *We point out that while the complexity bounds in (D.24) are better than those in (2.9) in terms of dependence on the target accuracy ϵ, they have been obtained for a different performance measure. Indeed, if only the Frank-Wolfe gap is considered then it is easy to see that both bounds are of the same order of magnitude due to part c of Lemma D.1.* ∎