Zeroth-order Nonconvex Stochastic Optimization: Handling Constraints, High-Dimensionality and Saddle-Points*

Krishnakumar Balasubramanian[†1] and Saeed Ghadimi[‡2]

[1]Department of Statistics, University of California, Davis
[2]Department of Operations Research and Financial Engineering, Princeton University

Abstract

In this paper, we propose and analyze zeroth-order stochastic approximation algorithms for nonconvex and convex optimization, with a focus on addressing constrained optimization, high-dimensional setting and saddle-point avoiding. To handle constrained optimization, we first propose generalizations of the conditional gradient algorithm achieving rates similar to the standard stochastic gradient algorithm using only zeroth-order information. To facilitate zeroth-order optimization in high-dimensions, we explore the advantages of structural sparsity assumptions. Specifically, (i) we highlight 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 and (ii) propose a truncated stochastic gradient algorithm with zeroth-order information, whose rate of convergence depends only poly-logarithmically on the dimensionality. We next focus on avoiding saddle-points in non-convex setting. Towards that, we interpret the Gaussian smoothing technique for estimating gradient based on zeroth-order information as an instantiation of first-order Stein's identity. Based on this, we provide a novel linear-(in dimension) time estimator of the Hessian matrix of a function using only zeroth-order information, which is based on second-order Stein's identity. We then provide an algorithm for avoiding saddle-points, which is based on a zeroth-order cubic regularization Newton's method and discuss its convergence rates.

---

*Both authors contributed equally and are listed in alphabetical order
†kbala@ucdavis.edu
‡sghadimi@princeton.edu

1 Introduction

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

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

where $\mathcal{X}$ is a closed convex subset of $\mathbb{R}^d$. The case of nonconvex objective function f is ubiquitous in modern statistical machine learning problems and developing provable algorithms for such problems has been a topic of intense research in the recent years [NW06, Ber16, JK17], along with the more standard convex case [BTN01, BV04, Nes13, BS15, Bec17]. 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), Hessian-queries (second-order oracle), and similar 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 [RK16], selecting the tuning parameters of deep neural networks [SLA12] and design of black-box attacks to deep networks [CZS+17]. It is worth noting that recently such zeroth-order optimization techniques have also been applied in the field of reinforcement learning [SHC+17, CRS+18, MGR18]. Furthermore, methods using similar oracles have been studied in the literature under the name of derivative-free optimization [Spa05, CSV09], bayesian optimization [Moc12] and optimization with bandit feedback [BCB12].

Algorithms available for solving problem (1.1) depend crucially on the constraint set $\mathcal{X}$, along with the structure imposed on the objective function, f. Despite decades of work in zeroth-order optimization literature, there still exists several challenges, primarily motivated by contemporary statistical machine learning problems. A majority of the existing zeroth-order algorithms are predominantly analyzed in the low-dimensional unconstrained setting. Furthermore, when f is nonconvex, apart from the first-order stationarity result for gradient descent (GD) algorithm in [GL13], other meaningful theoretical results are lacking in the zeroth-order optimization literature. In this work, we provide theoretically sound algorithms to address the following three main drawbacks of existing zeroth-order optimization methods.

The first issue we address is that of **constrainted zeroth-order stochastic optimization**. For the problem in (1.1), depending on the geometry of the constraint set $\mathcal{X}$, the cost of computing the projection to the set might be prohibitive. In the first-order oracle setting, this lead to the re-emergence of Conditional Gradient (CG) algorithms recently [HK12, Jag13]. 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 convergence rates compared to zeroth-order GD algorithms for the unconstrained case. To address this question, we propose and analyze in Section 2 a classical version of CG algorithm with zeroth-order information and provide convergence results. We then propose a modification in Section 2.2 that has improved rates, when f is convex. Notably, we demonstrate that with zeroth-order information, the complexity of CG algorithms also depend linearly on the dimensionality, similar to the GD algorithms, thereby facilitating constrained zeroth-order optimization.

Next, we consider the **impact of dimensionality in zeroth-order optimization**. Considering the unconstrained case of $\mathcal{X} = \mathbb{R}^d$, recall that when first-order information is available, the rate of convergence of the standard Gradient Descent (GD) algorithm is dimension-independent [Nes13]. Whereas when only the zeroth-order information is available, any algorithm (with estimated gradients) has (at least) linear dependence on d [GL13, JNR12, DJWW15]. This illustrates yet another difference between the availability of first and zeroth-order oracle information. 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 [WDBS18], 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 [WDBS18]), 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 [WDBS18].

Finally, we address the issue of **avoiding saddle-points in zeroth-order stochastic optimization**. When the function f is non-convex, designing algorithms that avoid saddle-points and converge to local minimizers is challenging, as exemplified by worst-case computational hardness results [MK87, CGT18]. Hence, it is necessary to impose further structure on the problem to obtain meaningful results. A particularly interesting structure on f is the so-called strict saddle property, which necessitates that all local minima are global minima. This structure has regained popularity as several useful stochastic optimization problems in statistical machine learning are shown to posses this property; for example, phase retrieval [SQW18], tensor decomposition [GHJY15], matrix completion and sensing [BNS16, GLM16] and training deep neural networks [KK19]. See also the survey article [SQW15]. Motivated by this, algorithms that avoid saddle-points and converge to second-order stationary points have re-gained popularity as well. Indeed, methods based on exact or in-exact second-order oracle naturally converge to second-order stationary points [NP06, CGT11a, CGT11b, XRKM17, TSJ$^+$17, CDHS18, AZ18]. Furthermore, first-order methods escape saddle points by leveraging an additional noise term in each iteration; for example [GHJY15, JGN$^+$17, RZS$^+$18] and the references therein. But to the best of our knowledge, there is no algorithm for efficiently avoiding saddle-points under zeroth-order oracle information. In this work, we propose a zeroth-order cubic regularized Newton method, that converges efficiently to second-order stationary points with just noisy function evaluations. In order to do so, we interpret the Gaussian smoothing for zeroth-order gradient estimation [NS17], as an instantiation of Stein's identity [Ste72, Ste81]. Based on this interpretation, we develop provable techniques for estimating the Hessian of a function at a point with just function queries, leveraging higher-order Stein's identity. Notably, our Hessian estimator is based only on inner-product evaluations thereby having a linear-in-dimension time runtime. We also provide a comprehensive complexity analysis of the proposed algorithm in terms of achieving second-order stationary points.

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 5)	Nonconvex, s-sparse	$\mathcal{O}\left((s\log d)^2/\epsilon^4\right)$	Theorem 3.1
Truncated ZSGD (Alg 6)	Convex, s-sparse	$\mathcal{O}\left(s(\log d/\epsilon)^2\right)$	Theorem 3.2
ZSGD	Convex	$\mathcal{O}(d/\epsilon^2)$	[JNR12, DJWW15, GL13]
	Nonconvex	$\mathcal{O}(d/\epsilon^4)$	[GL13]
ZCRN (Alg 7)	Nonconvex	$\mathcal{O}\left(\frac{d}{\epsilon^{3.5}}\right) + \tilde{\mathcal{O}}\left(\frac{d^4}{\epsilon^{2.5}}\right)$	Theorem 4.1

Table 1: A list of complexity bounds for stochastic zeroth-order methods to find an ϵ-optimal or ϵ-stationaly or ϵ-local optimal (see Definition 1.1) point of problem (1.1). Here, $\tilde{\mathcal{O}}$ hides log factors in d.

Our contributions: To summarize the above discussion, in this paper we make the following contributions to the literature on zeroth-order stochastic optimization.

1. 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. 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.

2. Next, we consider a zeroth-order stochastic gradient algorithm in the high-dimensional non-convex setting and illustrate an implicit regularization phenomenon –the algorithm converges to first-order stationary points with rates that depend only poly-logarithmically on dimensionality. We also propose a truncated zeroth-order stochastic gradient algorithm in the convex setting which also depends only poly-logarithmically on the dimensionally but has improved dependence on the error-tolerance.

3. Finally, we propose a zeroth-order Stochastic cubic regularized Newton method that avoids saddle points and converges to second-order stationary points efficiently. Our algorithm is based on a novel technique for estimating the Hessian of a function from function queries based on Stein's identities.

Our contributions extend the applicability of zeroth-order stochastic optimization to the constrained, high-dimensional and non-convex settings 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 $\mathbf{E}[F(x,\xi)] = f(x), \mathbf{E}[\nabla F(x,\xi) = \nabla f(x), \mathbf{E}[\|\nabla F(x,\xi) - \nabla f(x)\|_*^2] \le \sigma^2$, 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 [NS17] 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

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

We leverage the following properties of f_ν due to Nesterov [NS17] in our proofs later, which we replicate below for completeness.

Theorem 1.1 ([NS17]) *For a Gaussian random vector $u \sim N(0, I_d)$ we have that*

$$\mathbf{E}[\|u\|^k] \leq (d + k)^{k/2} \tag{1.5}$$

for any $k \geq 2$. Moreover, 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 $x \in \mathbb{R}^d$,*

$$|f_\nu(x) - f(x)| \leq \frac{\nu^2}{2} L d, \tag{1.6}$$

$$\|\nabla f_\nu(x) - \nabla f(x)\| \leq \frac{\nu}{2} L(d + 3)^{\frac{3}{2}}. \tag{1.7}$$

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

$$\frac{1}{\nu^2}\mathbf{E}_u[\{f(x+\nu u) - f(x)\}^2\|u\|^2] \leq \frac{\nu^2}{2}L^2(d+6)^3 + 2(d+4)\|\nabla f(x)\|^2. \qquad (1.8)$$

We next introduce the Stein's identity, popular in the statistics and probability theory literature.

Theorem 1.2 ([Ste72, Ste81]) *Let $u \sim N(0, I_d)$, be a standard Gaussian random vector and let $g : \mathbb{R}^d \to \mathbb{R}$, be an almost-differentiable function* [1] *with $\mathbf{E}\left[\|\nabla g(u)\|\right] < \infty$, we have*

$$\mathbf{E}\left[u\ g(u)\right] = \mathbf{E}\left[\nabla g(u)\right]. \qquad (1.9)$$

Furthermore, when the function g has a twice continuously differentiable Hessian, $\nabla^2 g(\cdot)$, we have the following (where the Expectation is assumed to exist):

$$\mathbf{E}[(uu^\top - I_d)\ g(u)] = \mathbf{E}[\nabla^2 g(u)]. \qquad (1.10)$$

Based on the above theorem, the Gaussian smoothing approach of estimating gradients from function queries proposed by [NS17], is indeed based on Stein's identity. Indeed, if we let $g(u) = f(x + \nu u)$ in Equation 1.9, it is easy to see that the identity in Equation 1.3 holds by simply evaluating the Gaussian Stein's identity in Equation 1.9. We elaborate more on this connection and extensions in Section 4.1. We conclude the section, by defining 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:*

- *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.*

- *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 $u \in \mathcal{X}$.*

- *If f is nonconvex, $\bar{x}$ is called an ϵ-local optima of the unconstrained variant of problem (1.1) if*

$$\max\left\{ \sqrt{\mathbf{E}[\|\nabla f(\bar{x})\|_*]}, \frac{-1}{\sqrt{\lambda_{\max}(\nabla^2 f)}}\ \mathbf{E}[\lambda_{\min}\left(\nabla^2 f(\bar{x})\right)] \right\} \leq \sqrt{\epsilon}$$

where for a symmetric matrix A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denotes the minimum and maximum eigenvalue.

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 [GL13, LZ16], 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 [BCB12, Sha13] which may have a different behavior than our performance measures.

---

[1]For a definition of almost-differentiable function, we refer the reader to Definition 1 in [Ste81]

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.
---

2 Handling Constraints: 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.

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 $B := L D_\mathcal{X} + \|\nabla f(x_*)\|$. 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., [GL13, RSPS16]). 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., [LZ16, HL16, RSPS16, MHK18a, MHK18b, Gha18]). 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

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

has been widely use in the literature to show rate of convergence of the CG methods when f is convex (see e.g., [FW56, DR70, Hea82]). 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*

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

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

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

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

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

8

Then we have

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

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

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

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

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

In order to prove Theorem 2.1, we need the following result that provides upper bounds for the variance of our gradient estimator.

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

$$\mathbf{E}[\|\bar{G}_\nu^k - \nabla f_\nu(z_{k-1})\|^2] \leq \frac{2(d+5)(B^2 + \sigma^2)}{m_k} + \frac{\nu^2}{2m_k}L^2(d+3)^3, \qquad (2.14)$$

$$\mathbf{E}[\|\bar{G}_\nu^k - \nabla f(z_{k-1})\|^2] \leq \frac{4(d+5)(B^2 + \sigma^2)}{m_k} + \frac{3\nu^2}{2}L^2(d+3)^3. \qquad (2.15)$$

Proof. [Proof of Lemma 2.1] First note that using (1.8) for function F instead of f, under Assumptions 1 and 2, we obtain

$$\mathbf{E}[\|G_\nu^{k,j}\|^2] \leq \frac{\nu^2 L^2}{2}(d+6)^3 + 2\left[\|\nabla f(z_{k-1})\|^2 + \sigma^2\right](d+4)$$

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 (2.14). We can then obtain (2.15) by noting (1.7) and the fact that

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

$\blacksquare$

Proof. [Proof of Theorem 2.1] Denoting $\Delta_k = \bar{G}_\nu^k - \nabla f(z_{k-1})$, noting (1.2), (2.3), and (2.5), we have

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

$$= f(z_{k-1}) + \alpha_k \langle \nabla f(z_{k-1}), x_k - z_{k-1} \rangle + \frac{L\alpha_k^2}{2}\|x_k - z_{k-1}\|^2$$

$$\leq f(z_{k-1}) + \alpha_k \langle \nabla f(z_{k-1}), \hat{x}_k - z_{k-1} \rangle + \frac{L\alpha_k^2}{2}\left[\|x_k - z_{k-1}\|^2 + \|x_k - \hat{x}_k\|^2\right] + \frac{\|\Delta_k\|^2}{2L}$$

$$\leq f(z_{k-1}) - \alpha_k g_\mathcal{X}^k + LD_\mathcal{X}^2 \alpha_k^2 + \frac{\|\Delta_k\|^2}{2L}, \qquad (2.16)$$

9

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

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

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 2.1, 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 (2.16), 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

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

which together with the fact that

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

due to (2.12), imply that

$$\mathbf{E}[f(z_N)] - f^* + \mathbf{E}[g_{\mathcal{X}}^R] \leq \frac{\Gamma_N}{1 - \Gamma_N} \left[f(z_0) - f^* + LD_{\mathcal{X}}^2 \sum_{k=1}^{N} \frac{\alpha_k^2}{\Gamma_k} + \frac{2(d+5)(B^2+\sigma^2)}{L} \sum_{k=1}^{N} \frac{1}{\Gamma_k m_k} \right.$$
$$\left. + \frac{\nu^2}{2} L(d+3)^3 \sum_{k=1}^{N} \frac{1}{\Gamma_k} \right] \tag{2.17}$$

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). ∎

10

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**

$$y_t = \operatorname*{argmin}_{u \in \mathcal{X}} \{h_\gamma(u) := \langle g + \gamma(\bar{y}_{t-1} - x), u - \bar{y}_{t-1} \rangle\} \tag{2.18}$$

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 $\bar{y}_t$.

Remark 1 *Observe that the complexity bounds in (2.9), in terms of ϵ, match the ones obtained in [Gha18, RSPS16, MHK18b] 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 [HL16, Gha18]. 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 [DJWW15]. 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 [LZ16], 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

$$P_{\mathcal{X}}(x, g, \gamma) = \operatorname*{argmin}_{u \in \mathcal{X}} \left\{ \langle g, u \rangle + \frac{\gamma}{2}\|u - x\|^2 \right\}, \tag{2.19}$$

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 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

$$w_k = (1 - \alpha_k)z_{k-1} + \alpha_k x_{k-1} \tag{2.20}$$

    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

$$x_k = ICG(x_{k-1}, \bar{G}_\nu^k, \gamma_k, \mu_k), \tag{2.21}$$

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

    3. Set

$$z_k = (1 - \alpha_k)z_{k-1} + \alpha_k x_k \tag{2.22}$$

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*

$$\alpha_k = \frac{2}{k+1}, \quad \gamma_k = \frac{4L}{k}, \quad \mu_k = \frac{LD_X^0}{kN}, \quad \nu = \frac{1}{\sqrt{2N}} \max\left\{ \frac{1}{d+3}, \sqrt{\frac{D_X^0}{d(N+1)}} \right\}$$

$$m_k = \frac{k(k+1)}{D_X^0} \max\{(d+5)B_{L\sigma}N, d+3\}, \quad \forall k \geq 1, \tag{2.23}$$

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*

$$\mathbf{E}[f(z_N) - f(x_*)] \leq \frac{12LD_X^0}{N(N+1)}. \tag{2.24}$$

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

$$\mathcal{O}\left(\frac{d}{\epsilon^2}\right), \quad \mathcal{O}\left(\frac{1}{\epsilon}\right). \tag{2.25}$$

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

$$
\begin{aligned}
f_\nu(z_k) &\leq f_\nu(w_k) + \langle \nabla f_\nu(w_k), z_k - w_k \rangle + \frac{L}{2}\|z_k - w_k\|^2 \\
&\leq (1 - \alpha_k)f_\nu(z_{k-1}) + \alpha_k\left[f_\nu(w_k) + \langle \nabla f_\nu(w_k), x_k - w_k \rangle\right] \\
&\quad + \frac{L\alpha_k^2}{2}\|x_k - x_{k-1}\|^2,
\end{aligned}
\tag{2.26}
$$

where the second inequality follows from convexity of f_ν, (2.20), and (2.22). Also note that by (2.18) and (2.21), we have

$$-\mu_k \leq \langle \bar{G}_\nu^k + \gamma_k(x_k - x_{k-1}), u - x_k \rangle \qquad \forall u \in \mathcal{X}. \tag{2.27}$$

12

Letting $u = x_*$ in the above inequality and multiplying it by α_k, summing it up with (2.26), and denoting $\bar{\Delta}_k = \bar{G}_\nu^k - \nabla f_\nu(w_k)$, we obtain

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

which together with the facts that

$$\|x_{k-1} - x_*\|^2 = \|x_k - x_{k-1}\|^2 + \|x_k - x_*\|_2^2 + 2\langle x_{k-1} - x_k, x_k - x_*\rangle,$$

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

imply

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

$$+ \quad \frac{\alpha_k\gamma_k}{2}\left[\|x_{k-1} - x_*\|^2 - \|x_k - x_*\|^2\right] + \frac{\|\bar{\Delta}_k\|^2}{2L}. \quad (2.28)$$

Defining

$$\hat{\Gamma}_k = \prod_{i=2}^{k}(1 - \alpha_i), \quad \hat{\Gamma}_1 = 1, \quad (2.29)$$

subtracting $f_\nu(x_*)$ from both sides of the above inequality, diving them by $\hat{\Gamma}_k$, taking expectation, summing them up, noting (1.6) assuming that $\alpha_1 = 1$, $\gamma_k \geq 2L\alpha_k$, and $\gamma_k\alpha_k/\hat{\Gamma}_k$ is constant for any $k \geq 1$, we obtain

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

Now noticing that

$$\hat{\Gamma}_k = \frac{2}{k(k+1)}, \qquad \frac{\alpha_k\gamma_k}{\hat{\Gamma}_k} = 4L, \qquad \frac{\alpha_k\mu_k}{\hat{\Gamma}_k} = \frac{LD_0^2}{N},$$

$$\frac{1}{m_k\hat{\Gamma}_k} \leq \frac{2D_0^2}{\max\{(d+5)B_{L\sigma}N, d+3\}}$$

due to (2.23) and (2.29), we obtain (2.24).

Furthermore, note that the function h_γ defined in Algorithm 2 is indeed negative the FW-gap of the CG method applied to problem (2.19). 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 (2.19) is Lipschitz continuous with constant γ, we have

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

which together with the choice of μ_k and γ_k in (2.23), 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 $\sum_{k=1}^{N} T_k \leq 48LD_\mathcal{X}^2/\epsilon^2$ due to (2.25).

∎

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 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 in Section 2.3.* ∎

2.3 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 4 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 $\{0, \dots, N-1\}$
for $k = 1, \dots, N$ **do**
    Generate $u_k = [u_{k,1}, \dots, u_{k,m_k}]$, 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 (2.21).
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 (2.19), 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

$$GP_{\mathcal{X}}(x, g, \gamma) = \gamma(x - P_{\mathcal{X}}(x, g, \gamma)), \qquad (2.30)$$

where $P_{\mathcal{X}}$ is the solution to (2.19). This quantity which has been widely used in the literature as a convergence criteria for solving nonconvex problems (see, e.g., [NY83, Nes04]), plays an analogues role of the gradient in constrained problems. Next result provides some properties for this criteria.

Lemma 2.2 *Let $P_{\mathcal{X}}(\cdot)$ be defined in (2.19), $\gamma > 0$, 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 $P_{\mathcal{X}}^\mu$ be the inexact solution of (2.19) such that

$$\langle g + \gamma(P_{\mathcal{X}}^\mu(x, g, \gamma) - x), u - P_{\mathcal{X}}^\mu(x, g, \gamma) \rangle \geq -\mu \qquad \forall u \in \mathcal{X} \qquad (2.31)$$

14

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

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

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 (2.19) 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 (2.19), we have

$$\langle g + \gamma(P_{\mathcal{X}}(x, g, \gamma) - x), u - P_{\mathcal{X}}(x, g, \gamma) \rangle \geq 0 \qquad \forall \tilde{u} \in \mathcal{X}. \tag{2.32}$$

Letting $\tilde{u} = P_{\mathcal{X}}^{\mu}(x, g, \gamma)$ in the above inequality and $u = P_{\mathcal{X}}(x, g, \gamma)$ and $g = \nabla f(x)$ in (2.31) and summing them up, we clear get the result in part b). Third, letting $\tilde{u} = x$ in (2.32), 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, (2.32) also implies that

$$g_{\mathcal{X}}(x) + \frac{1}{\gamma}\|GP_{\mathcal{X}}(x, \nabla f(x), \gamma)\|^2 \leq \langle \nabla f(x) + \gamma(x - u), x - P_{\mathcal{X}}(x, \nabla f(x), \gamma) \rangle$$

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

where the last inequality follows from Assumption 3.

∎

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

Theorem 2.3 *Let $\{x_k\}$ be generated by Algorithm 4, the function f be nonconvex, and*

$$\nu = \sqrt{\frac{1}{2N(d+3)^3}}, \quad \gamma_k = 2L, \quad \mu_k = \frac{1}{4N}, \quad m_k = 6(d+5)N, \quad \forall k \geq 1. \tag{2.33}$$

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

$$\mathbf{E}[\|GP_{\mathcal{X}}(x_R, \nabla f(x_R), \gamma_R)\|^2] \leq \frac{8L}{N}\left(f(x_0) - f^* + L + B^2 + \sigma^2\right). \tag{2.34}$$

where R is uniformly distributed over $\{0, \ldots, N-1\}$ and $g_{\mathcal{X}}$ is defined in (2.30). 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

$$\mathcal{O}\left(\frac{d}{\epsilon^2}\right), \quad \mathcal{O}\left(\frac{1}{\epsilon^2}\right). \tag{2.35}$$

15

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

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

$$\leq f(x_{k-1}) - \gamma_k \left(1 - \frac{L}{\gamma_k} \right) \|x_k - x_{k-1}\|^2 + \frac{\|\Delta_k\|^2}{2L} + \mu_k.$$

Taking expectation from the above inequalities, summing them up, re-arranging the terms, and in the view of Lemma 2.1, we have

$$\sum_{k=1}^{N} \gamma_k \left(1 - \frac{L}{\gamma_k} \right) \mathbf{E}[\|x_k - x_{k-1}\|^2]$$

$$\leq f(x_0) - f^* + \sum_{k=1}^{N} \mu_k + \frac{\nu^2 L (d+3)^3 N}{2} + \frac{2(d+5)(B^2 + \sigma^2)}{L} \sum_{k=1}^{N} \frac{1}{m_k},$$

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

$$\frac{1}{\gamma_k^2} \|GP_{\mathcal{X}}(x_{k-1}, \nabla f(x_{k-1}), \gamma_k)\|^2$$

$$= \|x_{k-1} - P_{\mathcal{X}}(x_{k-1}, \nabla f(x_{k-1}), \gamma_k)\|^2$$

$$\leq 2\|x_k - x_{k-1}\|^2 + \frac{4\mu_k}{\gamma_k} + \frac{4\nu^2 L^2 (d+3)^3}{\gamma_k^2} + \frac{16(d+5)(B^2 + \sigma^2)}{\gamma_k^2 m_k},$$

imply that

$$\sum_{k=1}^{N} \left(\frac{\gamma_k - L}{2\gamma_k^2} \right) \mathbf{E}[\|GP_{\mathcal{X}}(x_{k-1}, \nabla f(x_{k-1}), \gamma_k)\|^2] \leq f(x_0) - f^* + \sum_{k=1}^{N} \left(\frac{3\gamma_k - 2L}{\gamma_k} \right) \mu_k$$

$$+ \frac{\nu^2 L (d+3)^3}{2} \sum_{k=1}^{N} \left(1 + \frac{4L(\gamma_k - L)}{\gamma_k^2} \right) + \frac{2(d+5)(B^2 + \sigma^2)}{L} \sum_{k=1}^{N} \frac{1}{m_k} \left(1 + \frac{8L(\gamma_k - L)}{\gamma_k^2} \right).$$

Hence, noting (2.33), we obtain

$$\mathbf{E}[\|GP_{\mathcal{X}}(x_R, \nabla f(x_R), \gamma_R)\|^2]$$

$$\leq \frac{8L[f(x_0) - f^*]}{N} + 16L^2 \mu_1 + 8\nu^2 L^2 (d+3)^3 + \frac{48(d+5)(B^2 + \sigma^2)}{m_1},$$

which implies (2.34). Rest of the proof is similar to that of Theorem 2.2 and hence we skip the details. ∎

Remark 3 *We point out that while the complexity bounds in (2.35) 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 2.2.* ∎

Algorithm 5 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.

---

3 Handling High-Dimensionality: 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, [WDBS18] considered the convex case and proposed algorithms for high-dimensional zeroth-order stochastic optimization. Motivated by [WDBS18], 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 $\nabla f_\nu(x) = \mathbf{E}_u[\nabla f(x + \nu u)]$. 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 [GL13] (provided in Algorithm 5 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 5 under the gradient sparsity assumption is stated below.

Theorem 3.1 *Let $\{x_k\}_{k \geq 0}$ be generated by Algorithm 5 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 3.1), 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

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

Before proving the theorem, we first present two technical results which play key roles in our convergence analysis.

Lemma 3.1 *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

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

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,

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

$\blacksquare$

Lemma 3.2 *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*

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

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 3.1. Second, noting this lemma again, Assumption 4, and part a), we have

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

Furthermore, by (1.4), Holder inequality, Lemma 3.1, and under Assumption 4 we have

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

∎

Proof. [Proof of Theorem 3.1] Noting (1.4), Lemma 3.2.a), and with the notion of $G_{\nu,k} \equiv G_\nu(x_k, \xi_k, u_k)$, we have

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

which after taking expectation imply that

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

where the last inequality follow from Holder inequality and Lemma 3.2.b). Summing both sides of the above inequality over the iterations and rearranging terms, we get

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

19

Algorithm 6 Truncated Zeroth-Order Stochastic Gradient Method

Given a positive integer $\hat{s}$, replace updating step of Algorithm 5 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.

---

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

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

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

Remark 4 *Note that the above theorem establishes rate of convergence of Algorithm 5 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 [GL13] for general nonconvex smooth problems.* ∎

Remark 5 *Remarkably, Algorithm 5 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 5 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 6. Note that this algorithm could be considered as a truncated variant of Algorithm 5 and a zeroth-order stochastic variant of the truncated gradient descent algorithm [JTK14]. In the next result, we present convergence analysis of this algorithm.

Theorem 3.2 *Let $\{x_k\}_{k\geq 1}$ be generated by Algorithm 5, 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

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

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

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

where the inequality follows from the facts that $|J^k| \leq 2\hat{s} + s^*$ and $\|G_{\nu,k}^{J^k}\| \leq \|G_{\nu,k}\|$. Taking expectation from both sides of the above inequality, summing them up, noting Lemma 3.2, convexity of f_ν (due to convexity of f), we have

$$\mathbf{E}\left[\|x_N - x^*\|_2^2\right] \leq \|x_0 - x^*\|_2^2 + (2\hat{s} + s^*)\sum_{k=0}^{N-1}\gamma_k^2\mathbf{E}\left[\|G_{\nu,k}\|_\infty^2\right] - 2\sum_{k=0}^{N-1}\gamma_k\langle x_k - x^*, \nabla f_\nu(x_k)\rangle$$

$$\leq \|x_0 - x^*\|_2^2 + 4C(2\hat{s} + s^*)(\log d)^2\sum_{k=0}^{N-1}\gamma_k^2\left[L^2\nu^2(\log d) + 4\|\nabla f(x_k)\|_1^2 + 4\sigma^2\right]$$

$$- 2\sum_{k=0}^{N-1}\gamma_k\left[f_\nu(x_k) - f_\nu(x_*)\right]$$

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

$$- 2\sum_{k=0}^{N-1}\gamma_k[1 - 16LCs(2\hat{s} + s^*)(\log d)^2\gamma_k][f(x_k) - f(x_*)],$$

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

21

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. $\blacksquare$

Remark 6 *While for convex case, similar to the nonconvex case, the complexity of Algorithm 6 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.* $\blacksquare$

Remark 7 *As discussed in detail in [WDBS18], 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 [WDBS18], we do not make any assumption on the sparsity or smoothness of the second-order derivative of the objective function f for our results.* $\blacksquare$

Remark 8 *As mentioned before, [WDBS18] 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, [WDBS18] requires computationally expensive debiased Lasso estimators whereas our method requires only simple thresholding operations (for convex case) to handle sparsity.* $\blacksquare$

4 Handling Saddle-Points: Zeroth-Order Cubic Regularization Method

In this section, we study zeroth-order stochastic cubic regularized Newton method for unconstrained version of Problem 1.1. Throughout this section, we equip our space with the self- dual Euclidean norm, i.e., $\|\cdot\| = \|\|_2$. Furthermore, for a matrix A, we denote by $\|A\|_F$, its Frobenious norm and by $\|A\|$, its operator norm. We also make the following smoothness assumption on the Hessian of the objective function f, which is a generalization of the assumption in Equation 1.2.

Assumption 6 *The function f is twice differentiable and has Lipschitz continuous Hessian i.e., there exists $L_H > 0$ such that*

$$\|\nabla^2 f(y) - \nabla^2 f(y)\| \leq L_H \|y - x\| \quad \forall x, y \in \mathbb{R}^d.$$

It can be easily seen that the above assumption is equivalent to

$$\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x)\| \leq \frac{L_H}{2}\|y - x\|^2, \tag{4.1}$$

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle - \frac{1}{2}\langle y - x, \nabla^2 f(x)(y - x)\rangle| \leq \frac{L_H}{6}\|y - x\|^3. \tag{4.2}$$

Note that such an assumption in standard in the analysis of second-order optimization techniques [NP06]. We next describe a general technique for estimating the Hessian of a function based on Stein's identity in Section 4.1 and use it to provide a zeroth-order cubic regularization method and its analysis in Section 4.2.

4.1 Estimating Hessian with Zeroth-Order Information

Charles Stein, in his seminal paper [Ste72], proposed a method for characterizing Gaussian random variables. Specifically, a random vector, $u \sim N(0, I_d)$, is standard Gaussian *if and only if,* $\mathbf{E}\left[u\ g(u)\right] = \mathbf{E}\left[\nabla g(u)\right]$, for all absolutely continuous function g. Note that Stein's identity, naturally relate function queries (left hand side of Equation 1.9) to gradients (right hand side of Equation 1.9) and thus is naturally suited for zeroth-order optimization. Indeed the Gaussian smoothing technique proposed by [NS17], is based on the Stein's identity. Indeed, if we let $g(u) = f(x + \nu u)$ in Equation 1.9, it is easy to see that the identity in Equation 1.3 holds by simply evaluating the Gaussian Stein's identity in Equation 1.9. Recall that the results in Sections 2 and 3 are essentially based on approximately estimating the gradient information based on the Gaussian smoothing technique [NS17]. In this section, we develop techniques for approximately estimating the Hessian using zeroth-order oracle, based on second-order Stein's identities. It is worth noting that [Erd16] also use Stein's identities to estimate the Hessian but they only work in the restricted framework of generalized linear models with Gaussian data. Our use of Stein's identity to estimate Hessians, is completely different and we provide Hessian estimators for a general class of non-covnex, smooth functions, even for deterministic functions.

The second-order Gaussian Stein's identity, that we provide here informally for convenience, states thats $\mathbf{E}[(uu^\top - I_d)\ g(u)] = \mathbf{E}[\nabla^2 g(u)]$, for all functions g with well-defined Hessians. Similar to first-order Stein's identity, this naturally relates function queries to Hessians. In order to leverage this, similar to the previous case, we let $g(u) = f(x + \nu u)$ and note that we have

$$\mathbf{E}\left[\frac{(uu^\top - I_d)f(x + \nu u)}{\nu^2}\right] = \mathbf{E}[\nabla^2 f(x + \nu u)] = \nabla^2 f_\nu(x) = H_{f_\nu}. \tag{4.3}$$

This provides a way of approximately estimating the Hessian of the function f_ν by approximating the expectation on the left hand side using Gaussian samples. Hence, we can leverage this estimate of Hessian of the smoothed function to get an approximate estimate of Hessian of f. Similar to the gradient-free setting, we now have the following estimates of the Hessian.

$$H_\nu(x, \xi, u) = \frac{1}{\nu^2}\left(uu^\top - I_d\right)F(x + \nu u, \xi),$$

$$H_\nu(x, \xi, u) = \frac{1}{\nu^2}\left(uu^\top - I_d\right)\left[F(x + \nu u, \xi) - F(x, \xi)\right],$$

$$H_\nu(x, \xi, u) = \frac{1}{2\nu^2}\left(uu^\top - I_d\right)\left[F(x + \nu u, \xi) - F(x, \xi) + F(x - \nu u, \xi) - F(x, \xi)\right], \tag{4.4}$$

$$\bar{H}_\nu = \frac{1}{b}\sum_{i=1}^{b} H_\nu(x, \xi, u_i). \tag{4.5}$$

Note that above quantities are all unbiased estimators of H_{f_ν}, with the last one, also being the variance reduced version. Note however that the above two estimators, unlike the gradient case, are not robust w.r.t the smoothing parameter ν in the sense that their variances blow up when ν converges to 0. Hence, for the rest of this section we only focus on the Hessian estimator defined in (4.4). The above Hessian-estimator has several advantages that we elaborate now. Recall that for second-order optimization algorithms, for example, cubic regularization, it is important to be able to compute the Hessian matrix operating on a vector $v \in \mathbb{R}^d$, efficiently. For the proposed estimator above, such a Hessian-vector product boils down to just inner-product based operations

and could be done in time linear in the dimensionality. To the best of our knowledge, no such estimator for computing the hessian of a function exists. We now present the following results that characterize the estimation and approximation capability of the $H_\nu(x, \xi, u)$.

Lemma 4.1 (Variance Bound) *Let the Hessian estimator be defined in (4.4) and Assumption 6 hold for $F(x, \xi)$. Then, we have*

$$\mathbf{E}[\|H_\nu(x, \xi, u)\|_F^4] \leq (d + 16)^8 \left(\frac{L_H^4 (d + 16)^2 \nu^4}{9} + L^4 \right). \tag{4.6}$$

As a consequence, we have

$$\mathbf{E}[\|H_\nu(x, \xi, u)\|^2] \leq (d + 16)^4 \left(\frac{L_H^2 (d + 16) \nu^2}{3} + L^2 \right). \tag{4.7}$$

Proof. Noting (4.4) and Holder's inequality, we have

$$\mathbf{E}[\|H_\nu(x, \xi, u)\|_F^4] = \mathbf{E}\left[\left\| \frac{1}{2\nu^2} \left(uu^\top - I_d \right) [F(x + \nu u, \xi) + F(x - \nu u, \xi) - 2F(x, \xi)] \right\|_F^4 \right]$$

$$\leq \frac{\left(\mathbf{E}\left[|F(x + \nu u, \xi) + F(x - \nu u, \xi) - 2F(x, \xi)|^8 \right] \cdot \mathbf{E}\left[\|uu^\top - I_d\|_F^8 \right] \right)^{\frac{1}{2}}}{16\nu^8},$$

which together with Assumptions 2, assumption (4.2) for $F(x, \xi)$, and the fact that

$$\mathbf{E}\left[\|uu^\top - I_d\|_F^8 \right] = \mathbf{E}\left[\left(\|u\|^4 - 2\|u\|^2 + d \right)^4 \right] \leq \mathbf{E}\left[\left(\|u\|^4 + d \right)^4 \right] \leq 2(d + 16)^8,$$

due to (1.5), imply

$$\mathbf{E}[\|H_\nu(x, \xi, u)\|_F^4] \leq \frac{(d + 16)^4}{8\sqrt{2}} \left(\mathbf{E}\left[\left| \frac{L_H \nu \|u\|^3}{3} + \langle u, \nabla^2 F(x, \xi) u \rangle \right|^8 \right] \right)^{\frac{1}{2}}$$

$$\leq (d + 16)^4 \left(\frac{L_H^4 \nu^4 \sqrt{\mathbf{E}[\|u\|^{24}]}}{81} + L^4 \sqrt{\mathbf{E}[\|u\|^{16}]} \right)$$

$$\leq (d + 16)^8 \left(\frac{L_H^4 \nu^4 (d + 16)^2}{9} + L^4 \right).$$

Moreover, by Holder's inequality, we have

$$\mathbf{E}[\|H_\nu(x, \xi, u)\|^2] \leq \mathbf{E}[\|H_\nu(x, \xi, u)\|_F^2] \leq \left(\mathbf{E}[\|H_\nu(x, \xi, u)\|_F^4] \right)^{\frac{1}{2}},$$

which together with (4.6), imply (4.7). ∎

Lemma 4.2 (Approximation Error) *Under Assumption 6, denoting the Hessian of f by H_f for simplicity, we have*

$$\|H_{f_\nu} - H_f\| \leq \frac{L_H \nu (d + 6)^{\frac{5}{2}}}{4}. \tag{4.8}$$

Proof. Taking $y = x + \nu u$ in Equation 4.2, note that we have

$$|f(x+\nu u) - f(x) - \nu\langle\nabla f(x), u\rangle - \frac{\nu^2}{2}\langle y - x, \nabla^2 f(x)(y-x)\rangle| \le \frac{L_H \nu^3}{6}\|u\|^3. \quad (4.9)$$

Furthermore, note that

$$H_{f_\nu} - H_f = \frac{\mathbf{E}[(uu^\top - I_d)(f(x+\nu u) - f(x) + f(x-\nu u) - f(x))]}{2\nu^2} - H_f$$

$$= \mathbf{E}\left[\left(\frac{f(x+\nu u) - f(x) - \frac{\nu^2}{2}\langle u, H_f u\rangle + f(x-\nu u) - f(x) - \frac{\nu^2}{2}\langle u, H_f u\rangle}{2\nu^2}\right)\left(uu^\top - I_d\right)\right],$$

which together with (4.9) and (1.5), imply that

$$\|H_{f_\nu} - H_f\| \le \frac{L_H \nu}{6}\mathbf{E}\left[\|u\|^3\|uu^\top - I_d\|\right] \le \frac{L_H \nu (d+6)^{\frac{3}{2}}}{6}\left[\frac{\sqrt{5}}{2}(d+4)+1\right].$$

∎

Remark 9 *Note that (4.8) is obtained only under Assumption 6. However one could obtain an improved bound on the approximation error, by making the more restrictive assumption of interchangeability of differentiation and expectation as follows: $\|H_{f_\nu} - H_f\| = \|\mathbf{E}[\nabla^2 f(x+\nu u)] - \nabla^2 f(x)\| \le \mathbf{E}\|\nabla^2 f(x+\nu u) - \nabla^2 f(x)\| \le L_H \nu \mathbf{E}\|u\| \le L_H \nu \sqrt{d}$. While this provides an improved dependency on d, we remark that this improvement does not translate to the improvement in the number of zeroth-order oracle calls, at least for the cubic regularized method as discussed in Section 4.2.* ∎

Remark 10 *Recall from Section 1.1 that one could obtain high-probability results via the approach proposed in [GL13, LZ16] under sub-Gaussian tail assumption on the function F. To allow for functions F, that have heavy-tails, one could also leverage the spectral truncation argument to construct a robust Hessian estimators; see for example, [Min18]. Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a non-decreasing function such that*

$$-\log(1 - x + x^2/2) \le \phi(x) \le \log(1 + x + x^2/2), \quad \forall x \in \mathbb{R}.$$

Recall the definition of a spectral function below.

Definition 4.1 *Let $A \in \mathbb{R}^{d \times d}$ be a real symmetric matrix with eigenvalue decomposition $A = U\Lambda U^\top$ where U is the matrix of eigenvectors of A and Λ is a diagonal matrix of eigenvalues $(\lambda_1, \ldots, \lambda_d)$. A real-valued function $\phi(\cdot)$ is a spectral function if it acts on the matrix as follows: $\phi(A) = U\phi(\Lambda)U^\top$ where*

$$\phi(\Lambda) = \phi\left(\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix}\right) = \begin{bmatrix} \phi(\lambda_1) & & \\ & \ddots & \\ & & \phi(\lambda_d) \end{bmatrix}$$

Then, we define the robust Hessian estimator as

$$\tilde{H}(x, \xi, u) = \frac{1}{\kappa} \cdot \phi\big[\kappa \cdot H_\nu(x, \xi, u)\big],$$

Algorithm 7 Zeroth-order Stochastic Cubic Regularized Newton Method

---

Input: $x_0 \in \mathbb{R}^d$, smoothing parameter $\nu > 0$, non-negative sequence α_k, positive integer sequences m_k and b_k, iteration limit $N \geq 1$ and probability distribution $P_R(\cdot)$ over $\{1, \ldots, N\}$.

for $k = 1, \ldots, N$ **do**

1. Generate $u_k^{G(H)} = [u_{k,1}^{G(H)}, \ldots, u_{k,m_k(b_k)}^{G(H)}]$, where $u_{k,j}^{G(H)} \sim N(0, I_d)$, call the stochastic oracle to compute m_k stochastic gradients $G_\nu^{k,j}$ and b_k stochastic Hessians $H_\nu^{k,j}$ according to (1.4) and (4.4), respectively, and take their averages:

$$\bar{G}_\nu^k = \frac{1}{m_k} \sum_{j=1}^{m_k} \frac{F(x_{k-1} + \nu u_{k,j}, \xi_{k,j}) - F(x_{k-1}, \xi_{k,j}^G)}{\nu} \, u_{k,j}^G, \tag{4.10}$$

$$\bar{H}_\nu^k = \frac{1}{b_k} \sum_{i=1}^{b_k} \frac{[F(x_{k-1} + \nu u_{k,i}^H, \xi_{k,i}^H) + F(x_{k-1} - \nu u_{k,i}^H, \xi_{k,i}^H) - 2F(x_{k-1}, \xi_{k,i}^H)]}{2\nu^2} \left(u_{k,i}^H (u_{k,i}^H)^\top - I_d \right). \tag{4.11}$$

2. Compute
$$x_k = \underset{x \in \mathbb{R}^d}{\mathrm{argmin}} \left\{ \tilde{f}^k(x) \equiv \tilde{f}(x, x_{k-1}, \bar{H}_\nu^k, \bar{G}_\nu^k, \alpha_k) \right\}, \tag{4.12}$$

where
$$\tilde{f}(x, y, H, g, \alpha) = \langle g, x - y \rangle + \frac{1}{2} \langle H(x - y), x - y \rangle + \frac{\alpha}{6} \|x - y\|^3. \tag{4.13}$$

end for

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

---

where $\kappa > 0$ is a tuning parameter. This provides us with a robust Hessian estimator that allows for the function F to have heavy tails. Furthermore, the more standard median-of-means estimator [NY83] provides a robust gradient estimator as well. A thorough treatment of the estimation error of the robust Hessian and gradient follows from an analysis similar to that of [Min18] and [NY83] respectively, although we do not outline the details in the current paper. We also remark that while the spectral truncation argument makes the estimator robust, the computational advantage of the vanilla estimator in Equation 4.4 is lost. ∎

4.2   Zeroth-Order Stochastic Cubic Regularized Newton Method

Our goal in this subsection is to provide a second-order algorithmic framework using the estimated gradient and Hessian based on Stein's identities. In particular, we present a zeroth-order stochastic cubic regularized Newton method in Algorithm 7. Note that the output of this algorithm, similar to the other algorithms presented in this paper for nonconvex problems, is a random index from the generated trajectory. In order to analyze its complexity, we first state a result due to [NP06] that provides optimality conditions of the cubic regularized subproblem in step 2 of Algorithm 7.

Lemma 4.3 ([NP06]) *Let* $\bar{x} = \operatorname*{argmin}_{x \in \mathbb{R}^d} \tilde{f}(x, y, H, g, \alpha)$. *Then, we have*

$$g + H(\bar{x} - y) + \frac{\alpha}{2}\|\bar{x} - y\|(\bar{x} - y) = 0,$$

$$H + \frac{\alpha}{2}\|\bar{x} - y\|I_d \succeq 0.$$

Our next result is the analogous result of Lemma 2.1 for the averaged Hessian matrices.

Lemma 4.4 *Let* $\bar{H}_\nu^k$ *be computed by (4.11),* $b_k \geq 4(1 + 2\log 2d)$. *Then under Assumptions 1 and 2, we have*

$$\mathbf{E}[\|\bar{H}_\nu^k - \nabla^2 f(x_{k-1})\|^2] \leq \frac{128(1 + 2\log 2d)(d + 16)^4 L^2}{3b_k} + 3L_H^2(d + 16)^5 \nu^2, \tag{4.14}$$

$$\mathbf{E}[\|\bar{H}_\nu^k - \nabla^2 f(x_{k-1})\|^3] \leq \frac{160\sqrt{1 + 2\log 2d}(d + 16)^6 L^3}{b_k^{\frac{3}{2}}} + 21L_H^3(d + 16)^{\frac{15}{2}} \nu^3. \tag{4.15}$$

Proof. First, note that by Theorem 1 in [Tro16], we have

$$\mathbf{E}[\|\bar{H}_\nu^k - \nabla^2 f_\nu(x_{k-1})\|^2] \leq \frac{2C(d)}{b_k^2}\left(\left\|\sum_{i=1}^{b_k} \mathbf{E}[\Delta_{k,i}^2]\right\| + C(d)\mathbf{E}\left[\max_i \|\Delta_{k,i}\|^2\right]\right),$$

where $\Delta_{k,i} = H_\nu(x_{k-1}, \xi_{k,i}^H, u_{k,i}^H) - \nabla^2 f_\nu(x_{k-1})$ and $C(d) = 4(1 + 2\log 2d)$. Now, noting (4.7), we have

$$\mathbf{E}[\|\Delta_{k,i}\|^2] \leq \frac{2(d + 16)^4(L_H^2(d + 16)\nu^2 + 4L^2)}{3},$$

which together with the above inequality and the fact that

$$\left\|\sum_{i=1}^{b_k} \mathbf{E}[\Delta_{k,i}^2]\right\| \leq \sum_{i=1}^{b_k} \left\|\mathbf{E}[\Delta_{k,i}^2]\right\| \leq \sum_{i=1}^{b_k} \mathbf{E}[\|\Delta_{k,i}\|^2],$$

imply

$$\mathbf{E}[\|\bar{H}_\nu^k - \nabla^2 f_\nu(x_{k-1})\|^2] \leq \frac{16(1 + 2\log 2d)(d + 16)^4}{3b_k}\left[4L^2 + L_H^2(d + 16)\nu^2\right]. \tag{4.16}$$

Combining this inequality with (4.8), we obtain (4.14). Moreover, by Holder's inequality we have

$$\mathbf{E}\left[\|\bar{H}_\nu^k - \nabla^2 f_\nu(x_{k-1})\|^3\right] \leq \mathbf{E}\left[\|\bar{H}_\nu^k - \nabla^2 f_\nu(x_{k-1})\| \cdot \|\bar{H}_\nu^k - \nabla^2 f_\nu(x_{k-1})\|_F^2\right]$$

$$\leq \left(\mathbf{E}\left[\|\bar{H}_\nu^k - \nabla^2 f_\nu(x_{k-1})\|^2\right] \cdot \mathbf{E}\left[\|\bar{H}_\nu^k - \nabla^2 f_\nu(x_{k-1})\|_F^4\right]\right)^{\frac{1}{2}}.$$

Now, by vector-valued Rosenthal's inequality (see, for example, Theorem 5.2 in [Pin94]) and (4.6), we obtain

$$\mathbf{E}\left[\|\bar{H}_\nu^k - \nabla^2 f_\nu(x_{k-1})\|_F^4\right] \leq \frac{3\mathbf{E}[\|\Delta_{k,i}\|_F^4]}{b_k^2} \leq \frac{3(d + 16)^8\left(L_H^4(d + 16)^2\nu^4 + 9L^4\right)}{b_k^2},$$

which together with the above inequality and (4.16) imply (4.15). ∎

We now proceed to provide the complexity results for Algorithm 7. We first require two intermediate results.

Lemma 4.5 *Let $\{x_k\}$ be computed by Algorithm 7. Then under Assumptions 1 and 2, we have*

$$\sqrt{\mathbf{E}[\|x_k - x_{k-1}\|^2]}$$
$$\geq \max\left\{\sqrt{\frac{\left(\mathbf{E}[\|\nabla f(x_k)\|] - \delta_k^g - \delta_k^H\right)}{L_H + \alpha_k}}, \frac{-2}{\alpha_k + 2L_H}\left[\mathbf{E}[\lambda_{\min}\left(\nabla^2 f(x_k)\right)] + \sqrt{2(\alpha_k + L_H)\delta_k^H}\right]\right\},$$
$$(4.17)$$

where $\delta_k^g, \delta_k^H > 0$ are chosen such that

$$\mathbf{E}[\|\nabla f(x_{k-1}) - \bar{G}_\nu^k\|^2] \leq (\delta_k^g)^2, \qquad \mathbf{E}[\|\nabla^2 f(x_{k-1}) - \bar{H}_\nu^k\|^3] \leq \left(2(L_H + \alpha_k)\delta_k^H\right)^{\frac{3}{2}}. \qquad (4.18)$$

Proof. By the equality condition in Lemma 4.3 and (4.1), we have

$$\|\nabla f(x_k)\| \leq \|\nabla f(x_k) - \nabla f(x_{k-1}) - \nabla^2 f(x_{k-1})(x_k - x_{k-1})\| + \|\nabla f(x_{k-1}) - \bar{G}_\nu^k\|$$
$$+ \|\nabla^2 f(x_{k-1}) - \bar{H}_\nu^k\| \cdot \|x_k - x_{k-1}\| + \frac{\alpha_k}{2}\|x_k - x_{k-1}\|^2$$
$$\leq \frac{(L_H + \alpha_k)}{2}\|x_k - x_{k-1}\|^2 + \|\nabla f(x_{k-1}) - \bar{G}_\nu^k\| + \|\nabla^2 f(x_{k-1}) - \bar{H}_\nu^k\| \cdot \|x_k - x_{k-1}\|$$
$$\leq (L_H + \alpha_k)\|x_k - x_{k-1}\|^2 + \|\nabla f(x_{k-1}) - \bar{G}_\nu^k\| + \frac{\|\nabla^2 f(x_{k-1}) - \bar{H}_\nu^k\|^2}{2(L_H + \alpha_k)}.$$

Taking expectation from both sides of the above inequality and noting that δ_k^g, δ_k^H given in (4.18) are well-defined by properly choosing m_k and b_k in Lemmas 2.1 and 4.4, we obtain

$$\frac{\left(\mathbf{E}[\|\nabla f(x_k)\|] - \delta_k^g - \delta_k^H\right)}{L_H + \alpha_k} \leq \mathbf{E}[\|x_k - x_{k-1}\|^2]. \qquad (4.19)$$

Also, by smoothness assumption of the Hessian and the inequality relation in Lemma 4.3

$$\nabla^2 f(x_k) \succeq \nabla^2 f(x_{k-1}) - L_H\|x_k - x_{k-1}\|I_d \succeq \nabla^2 f(x_{k-1}) - \bar{H}_\nu^k + \bar{H}_\nu^k - L_H\|x_k - x_{k-1}\|I_d$$
$$\succeq \nabla^2 f(x_{k-1}) - \bar{H}_\nu^k - \frac{(\alpha_k + 2L_H)\|x_k - x_{k-1}\|}{2}I_d,$$

which implies that

$$\frac{(\alpha_k + 2L_H)\|x_k - x_{k-1}\|}{2} \geq \lambda_{\min}\left(f(x_{k-1}) - \bar{H}_\nu^k\right) - \lambda_{\min}\left(\nabla^2 f(x_k)\right).$$

Taking expectation from both sides of the above inequality and noting definition of δ_k^H in (4.18), we obtain

$$\sqrt{\mathbf{E}[\|x_k - x_{k-1}\|^2]} \geq \mathbf{E}[\|x_k - x_{k-1}\|] \geq \frac{-2}{\alpha_k + 2L_H}\left[\sqrt{2(\alpha_k + L_H)\delta_k^H} + \mathbf{E}[\lambda_{\min}\left(\nabla^2 f(x_k)\right)]\right].$$

Combining the above inequality with (4.19), we obtain (4.17). ∎

Lemma 4.6 *Let $\{x_k\}$ be computed by Algorithm 7 for a given iteration limit $N \geq 1$. Then under Assumptions 1 and 2, we have*

$$\mathbf{E}[\|x_R - x_{R-1}\|^3] \leq \frac{36}{\sum_{k=1}^{N} \alpha_k} \left[f(x_0) - f^* + \sum_{k=1}^{N} \frac{4(\delta_k^g)^{\frac{3}{2}}}{\sqrt{3}\alpha_k} + \sum_{k=1}^{N} \left(\frac{18\sqrt[4]{2}}{\alpha_k} \right)^2 \left((L_H + \alpha_k)\delta_k^H \right)^{\frac{3}{2}} \right],$$
(4.20)

where R is an integer random variable whose probability distribution $P_R(\cdot)$ is supported on $\{1, \ldots, N\}$ and given by

$$P_R(R = k) = \frac{\alpha_k}{\sum_{k=1}^{N} \alpha_k} \qquad k = 1, \ldots, N,$$
(4.21)

and $\delta_k^g, \delta_k^H > 0$ are defined in (4.18).

Proof. First, note that by (4.2), (4.12), and the fact that $\alpha_k \geq L_H$, we have

$$f(x_k) \leq f(x_{k-1}) + \tilde{f}^k(x_k) + \|\nabla f(x_{k-1}) - \bar{G}_\nu^k\| \cdot \|x_k - x_{k-1}\| + \frac{1}{2}\|\nabla^2 f(x_{k-1}) - \bar{H}_\nu^k\| \cdot \|x_k - x_{k-1}\|^2.$$

Moreover, by Lemma 4.3, we have

$$\tilde{f}^k(x_k) = -\frac{1}{2}\langle \bar{H}_\nu^k(x_k - x_{k-1}), (x_k - x_{k-1})\rangle - \frac{\alpha_k}{3}\|x_k - x_{k-1}\|^3 \leq -\frac{\alpha_k}{12}\|x_k - x_{k-1}\|^3.$$

Combining the above two relations, we obtain

$$\frac{\alpha_k}{12}\|x_k - x_{k-1}\|^3 \leq f(x_{k-1}) - f(x_k) + \|\nabla f(x_{k-1}) - \bar{G}_\nu^k\| \cdot \|x_k - x_{k-1}\|$$

$$+ \frac{1}{2}\|\nabla^2 f(x_{k-1}) - \bar{H}_\nu^k\| \cdot \|x_k - x_{k-1}\|^2$$

$$\leq f(x_{k-1}) - f(x_k) + \frac{4}{\sqrt{3}\alpha_k}\|\nabla f(x_{k-1}) - \bar{G}_\nu^k\|^{\frac{3}{2}} + \left(\frac{9\sqrt{2}}{\alpha_k} \right)^2 \|\nabla^2 f(x_{k-1}) - \bar{H}_\nu^k\|^3$$

$$+ \frac{\alpha_k}{18}\|x_k - x_{k-1}\|^3,$$

where the last inequality follows from the Young's inequality. Taking expectation from both sides, re-arranging the terms, and noting (4.18), we obtain

$$\frac{\alpha_k}{36}\mathbf{E}[\|x_k - x_{k-1}\|^3] \leq f(x_0) - f^* + \frac{4(\delta_k^g)^{\frac{3}{2}}}{\sqrt{3}\alpha_k} + \left(\frac{18\sqrt[4]{2}}{\alpha_k} \right)^2 \left((L_H + \alpha_k)\delta_k^H \right)^{\frac{3}{2}}.$$

Summing up the above inequalities, dividing both sides by $\sum_{k=1}^{N} \alpha_k$, and noting (4.21), we obtain (4.20). ∎

Theorem 4.1 *Let $\{x_k\}$ be computed by Algorithm 7 for a given iteration limit $N \geq 1$. Moreover, assume that the parameters are set to*

$$\alpha_k = L_H, \qquad \nu \leq \frac{1}{2}\min\left\{ \sqrt{\frac{L_H\epsilon}{36(d+16)^5}}, \frac{\epsilon}{L(d+3)^{\frac{3}{2}}} \right\}, \qquad N = \frac{12\sqrt{L_H}(f(x_0) - f^*)}{\epsilon^{\frac{3}{2}}},$$

$$b_k = \frac{2L^2}{L_H}\left(4(d+16)^2 \right)^4 \left(\frac{\sqrt[3]{1 + 2\log 2d}}{\epsilon} \right), \qquad m_k = \frac{26(d+5)(B^2 + \sigma^2)}{\epsilon^2}.$$
(4.22)

Then under Assumptions 1 and 2, we have

$$5\sqrt{\epsilon} \geq \max\left\{ \sqrt{\mathbf{E}[\|\nabla f(x_R)\|]}, \frac{-5}{8\sqrt{L_H}}\mathbf{E}[\lambda_{\min}\left(\nabla^2 f(x_R)\right)]\right\}, \qquad (4.23)$$

where R is uniformly distributed over $\{1, \ldots, N\}$. As a consequence, to obtain an ϵ second-order stationary point of the problem, the total number of samples required to compute the gradient and Hessian are, respectively, bounded by

$$\mathcal{O}\left(\frac{d}{\epsilon^{\frac{7}{2}}}\right), \qquad \tilde{\mathcal{O}}\left(\frac{d^4}{\epsilon^{\frac{5}{2}}}\right).$$

Proof. First, note that by (4.22), Lemmas 2.1, and 4.4, we can ensure that (4.18) is satisfied by $\delta_k^g = 2\epsilon/5$ and $\delta_k^H = \epsilon/138$. Moreover, by Lemma 4.6, we have

$$\mathbf{E}[\|x_R - x_{R-1}\|^3] \leq \frac{1}{L_H^{\frac{3}{2}}}\left[\frac{12\sqrt{L_H}(f(x_0) - f^*)}{N} + 7\epsilon^{\frac{3}{2}}\right].$$

Hence, by choosing N according (4.22), and noting Lemma 4.6, we obtain (4.23). Therefore, x_R is an 4ϵ second-order stationary point of the problem. Finally, note that the total number of required samples to obtain such a solution is bounded by

$$\sum_{k=1}^{N} m_k = \mathcal{O}\left(\frac{d}{\epsilon^{\frac{7}{2}}}\right), \qquad \sum_{k=1}^{N} b_k = \tilde{\mathcal{O}}\left(\frac{d^4}{\epsilon^{\frac{5}{2}}}\right)$$

∎

Remark 11 *Note that [TSJ+17] provide a high-probability complexity result for stochastic Newton method with inexact gradient and Hessian information of the order $\tilde{\mathcal{O}}\left(\epsilon^{-3.5}\right)$. This dependence on ϵ is better compared to algorithms that only use stochastic first-order information to avoid saddle points. They mainly focus on sub-sampled Newton method common in the finite-sum setting and require their stochastic Hessians to be almost-surely bounded. However, this assumption does not imply the zeroth-order Hessian estimators in Equation 4.4 are bounded almost-surely, which complicates the analysis.* ∎

Remark 12 *Note that by Theorem 4.1, the total number of calls to the zeroth-order oracle is of the order $\tilde{\mathcal{O}}\left(\frac{d}{\epsilon^{3.5}}\right)$ when $\epsilon \approx d^{-3}$. This shows the advantage of using the (estimated) second-order information for converging to high-accuracy second-order stationary points. The linear dependence on d is the price to pay for having access to only zeroth-order information, similar to the previous sections. Furthermore, depending on the quality of the solution required, a wide variety of intermediate complexity results are possible, thereby providing practical flexibility.* ∎

5 Discussion

In this work, we propose and analyze zeroth-order stochastic approximation algorithms for convex and nonconvex problems motivated by modern machine learning challenges. Specifically, we

provide zeroth-order algorithms to deal with constraints, dimensionality and saddle-points in non-convex stochastic optimization problems. While our focus was on general stochastic optimization problems, one could naturally obtain better rates in the case of finite-sum optimization problems with various variance reduction techniques. Several concrete extensions are possible for future work. 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. Obtaining regret bounds for the non-convex problems considered in this work is more challenging. Furthermore, lower bounds can be explored for the cases considered in this paper when f is nonconvex. Finally, obtaining second-order stationarity results in the constrained setting is more challenging. We plan to extend our results for these setting in the future.

References

[AZ18]   Zeyuan Allen-Zhu. Natasha 2: Faster non-convex optimization than sgd. In *Advances in Neural Information Processing Systems*, pages 2680–2691, 2018.

[BCB12]   Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and non-stochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[Bec17]   Amir Beck. *First-Order Methods in Optimization*, volume 25. SIAM, 2017.

[Ber16]   Dimitri P Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 2016.

[BNS16]   Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

[BS15]   Dimitri P Bertsekas and Athena Scientific. *Convex optimization algorithms*. Athena Scientific Belmont, 2015.

[BTN01]   Ahron Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, volume 2. Siam, 2001.

[BV04]   Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[CDHS18]   Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

[CGT11a]   Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. Adaptive cubic regularisation methods for unconstrained optimization. part i: motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295, 2011.

[CGT11b]   Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. Adaptive cubic regularisation methods for unconstrained optimization. part ii: worst-case function-and derivative-evaluation complexity. *Mathematical programming*, 130(2):295–319, 2011.

[CGT18]   Coralia Cartis, Nick IM Gould, and Philippe L Toint. Second-order optimality and beyond: Characterization and evaluation complexity in convexly constrained nonlinear optimization. *Foundations of Computational Mathematics*, 18(5):1073–1107, 2018.

[CRS+18]   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.

[CSV09]   Andrew Conn, Katya Scheinberg, and Luis Vicente. *Introduction to derivative-free optimization*, volume 8. Siam, 2009.

[CZS+17]   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.

[DJWW15]   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.

[DR70]   V. Demyanov and A. Rubinov. *Approximate methods in optimization problems*. American Elsevier Publishing Co, 1970.

[Erd16]   Murat A Erdogdu. Newton-stein method: an optimization method for glms via stein's lemma. *The Journal of Machine Learning Research*, 17(1):7565–7616, 2016.

[FW56]   Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3:95–110, 1956.

[Gha18]   Saeed Ghadimi. Conditional gradient type methods for composite nonlinear and stochastic optimization. *Mathematical Programming*, 2018.

[GHJY15]   Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle pointsonline stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[GL13]   S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[GLM16]   Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[Hea82]   Donald Hearn. The gap function of a convex program. *Operations Research Letters*, 2, 1982.

[HK12]   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.

[HL16]   Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In *International Conference on Machine Learning*, pages 1263–1271, 2016.

[Jag13]   Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

[JGN⁺17]   Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732, 2017.

[JK17]   Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017.

[JNR12]   Kevin Jamieson, Robert Nowak, and Ben Recht. Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2012.

[JTK14]   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.

[KK19]   Kenji Kawaguchi and Leslie Pack Kaelbling. Elimination of all bad local minima in deep learning. *arXiv preprint arXiv:1901.00279*, 2019.

[LZ16]   Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *SIAM Journal on Optimization*, 26(2):1379–1409, 2016.

[MGR18]   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.

[MHK18a]   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.

[MHK18b]   Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *arXiv preprint arXiv:1804.09554*, 2018.

[Min18]   Stanislav Minsker. Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics*, 46(6A):2871–2903, 2018.

[MK87]   Katta G Murty and Santosh N Kabadi. Some np-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 39(2):117–129, 1987.

[Moc12]   Jonas Mockus. *Bayesian approach to global optimization: theory and applications*, volume 37. Springer Science & Business Media, 2012.

[Nes04]   Y. E. Nesterov. *Introductory Lectures on Convex Optimization: a basic course*. Kluwer Academic Publishers, Massachusetts, 2004.

[Nes13]   Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[NP06]   Yurii Nesterov and Boris Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[NS17]   Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[NW06]   Jorge Nocedal and Stephen J Wright. *Nonlinear Equations*. Springer, 2006.

[NY83]   A. S. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley, XV, 1983.

[Pin94]   Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

[RK16]   Reuven Rubinstein and Dirk Kroese. *Simulation and the Monte Carlo method*, volume 10. John Wiley & Sons, 2016.

[RSPS16]   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.

[RZS+18]   Sashank Reddi, Manzil Zaheer, Suvrit Sra, Barnabas Poczos, Francis Bach, Ruslan Salakhutdinov, and Alex Smola. A generic approach for escaping saddle points. In *International Conference on Artificial Intelligence and Statistics*, pages 1233–1242, 2018.

[Sha13]   Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, pages 3–24, 2013.

[SHC+17]   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.

[SLA12]   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.

[Spa05]   James Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons, 2005.

[SQW15]   Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

[SQW18]   Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.

[Ste72]   Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California, 1972.

[Ste81]   Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.

[Tro16]   Joel A Tropp. The expected norm of a sum of independent random matrices: An elementary approach. In *High Dimensional Probability VII*, pages 173–202. Springer, 2016.

[TSJ+17]   Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan. Stochastic cubic regularization for fast nonconvex optimization. *arXiv preprint arXiv:1711.02838*, 2017.

[WDBS18]   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.

[XRKM17]   Peng Xu, Farbod Roosta-Khorasani, and Michael W Mahoney. Newton-type methods for non-convex optimization under inexact hessian information. *arXiv preprint arXiv:1708.07164*, 2017.