Zeroth-Order Negative Curvature Finding: Escaping Saddle Points without Gradients

Hualin Zhang[1]   **Huan Xiong**[2,3]   **Bin Gu**[1,3]

[1] Nanjing University of Information Science & Technology
[2] Harbin Institute of Technology
[3] Mohamed bin Zayed University of Artificial Intelligence
{zhanghualin98,huan.xiong.math,jsgubin}@gmail.com

Abstract

We consider escaping saddle points of nonconvex problems where only the function evaluations can be accessed. Although a variety of works have been proposed, the majority of them require either second or first-order information, and only a few of them have exploited zeroth-order methods, particularly the technique of negative curvature finding with zeroth-order methods which has been proven to be the most efficient method for escaping saddle points. To fill this gap, in this paper, we propose two zeroth-order negative curvature finding frameworks that can replace Hessian-vector product computations without increasing the iteration complexity. We apply the proposed frameworks to ZO-GD, ZO-SGD, ZO-SCSG, ZO-SPIDER and prove that these ZO algorithms can converge to (ϵ, δ)-approximate second-order stationary points with less query complexity compared with prior zeroth-order works for finding local minima.

1 Introduction

Nonconvex optimization has received wide attention in recent years due to its popularity in modern machine learning (ML) and deep learning (DL) tasks. Specifically, in this paper, we study the following unconstrained optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x), \tag{1}$$

where both $f_i(\cdot)$ and $f(\cdot)$ can be nonconvex. In general, finding the global optima of nonconvex functions is NP-hard. Fortunately, finding local optima is an alternative because it has been shown in theory and practice that local optima have comparable performance capabilities to global optima in many machine learning problems [18, 19, 30, 21, 20, 23, 31]. Gradient-based methods have been shown to be able to find an ϵ-approximate first-order stationary point ($\|\nabla f(x)\| \leq \epsilon$) efficiently, both in the deterministic setting (*e.g.*, gradient descent [37]; accelerated gradient descent [8, 33]) and stochastic setting (*e.g.*, stochastic gradient descent [37, 43]; SCSG [32]; SPIDER [16]). However, in nonconvex settings, first-order stationary points can be local minima, global minima, or even saddle points. Converging to saddle points will lead to highly suboptimal solutions [24, 45] and destroy the model's performance. Thus, escaping saddle points has recently become an important research topic in nonconvex optimization.

Several classical results have shown that, for ρ-Hessian Lipschitz functions (see Definition 1), using the second-order information like computing the Hessian [39] or Hessian-vector products [1, 9, 2], one can find an ϵ-approximate second-order stationary point (SOSP, $\|\nabla f(x)\| \leq \epsilon$ and $\nabla^2 f(x) \succeq -\sqrt{\rho\epsilon}\mathbf{I}$). However, when the dimension of x is large, even once access to the Hessian

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

is computationally infeasible. A recent line of work shows that, by adding uniform random perturbations, first-order (FO) methods can efficiently escape saddle points and converge to SOSP. In the deterministic setting, [26] proposed the perturbed gradient descent (PGD) algorithm with gradient query complexity $\tilde{\mathcal{O}}(\log^4 d/\epsilon^2)$ by adding uniform random perturbation into the standard gradient descent algorithm. This complexity is later improved to $\tilde{\mathcal{O}}(\log^6 d/\epsilon^{1.75})$ by the perturbed accelerated gradient descent [28] which replaces the gradient descent step in PGD by Nesterov's accelerated gradient descent.

Table 1: A summary of the results of finding (ϵ, δ)-approximate SOSPs (see Definition 2) by the zeroth-order algorithms. (CoordGE, GaussGE, and RandGE are abbreviations of "coordinate-wise gradient estimator", "Gaussian random gradient estimator" and "uniform random gradient estimator", respectively. RP, RS, and CR are abbreviations of "random perturbation", "random search" and "cubic regularization", respectively.)

Algorithm	Setting	ZO Oracle	Main Techniques	Function Queries
ZPSGD [27]	Deterministic	GaussGE + Noise	RP	$\tilde{\mathcal{O}}\left(\frac{d^2}{\epsilon^5}\right)$ †
PAGD [47]	Deterministic	CoordGE	RP	$\mathcal{O}\left(\frac{d\log^4 d}{\epsilon^2}\right)$ †
RSPI [35]	Deterministic	CoordGE	RS + NCF	$\mathcal{O}(\frac{d\log d}{\epsilon^{8/3}})$ ‡
Theorem. 4	Deterministic	CoordGE	NCF	$\mathcal{O}\left(\frac{d}{\epsilon^2} + \frac{d\log d}{\delta^{3.5}}\right)$
ZO-SCRN [5]	Stochastic	GaussGE	CR	$\tilde{\mathcal{O}}\left(\frac{d}{\epsilon^{3.5}} + \frac{d^4}{\epsilon^{2.5}}\right)$ †
Theorem. 3	Stochastic	CoordGE	NCF	$\tilde{\mathcal{O}}\left(\frac{d}{\epsilon^4} + \frac{d}{\epsilon^2\delta^3} + \frac{d}{\delta^5}\right)$
Theorem. 5	Stochastic	CoordGE + (RandGE)	NCF	$\tilde{\mathcal{O}}\left(\frac{d}{\epsilon^{10/3}} + \frac{d}{\epsilon^2\delta^3} + \frac{d}{\delta^5}\right)$
Theorem. 6	Stochastic	CoordGE	NCF	$\tilde{\mathcal{O}}\left(\frac{d}{\epsilon^3} + \frac{d}{\epsilon^2\delta^2} + \frac{d}{\delta^5}\right)$

† guarantees $(\epsilon, \mathcal{O}(\sqrt{\epsilon}))$-approximate SOSP, and ‡ guarantees $(\epsilon, \epsilon^{2/3})$-approximate SOSP.

Another line of work for escaping saddle points is to utilize the negative curvature finding (NCF), which can be combined with ϵ-approximate first-order stationary point (FOSP) finding algorithms to find an (ϵ, δ)-approximate SOSP. The main task of NCF is to calculate the approximate smallest eigenvector of the Hessian for a given point. Classical methods for solving NCF like the power method and Oja's method need the computation of Hessian-vector products. Based on the fact the Hessian-vector product can be approximated by the finite difference between two gradients, [49, 4] proposed the FO NCF frameworks Neon+ and Neon2, respectively. In general, adding perturbations in the negative curvature direction can escape saddle points more efficiently than adding random perturbations by a factor of $\tilde{\mathcal{O}}(\text{poly}(\log d))$ in theory. Specifically, in the deterministic setting, CDHS [9] combined with Neon2 can find an (ϵ, δ)-approximate SOSP in gradient query complexity $\tilde{\mathcal{O}}(\log d/\epsilon^{1.75})$. Recently, the same result was achieved by a simple single-loop algorithm [51], which combined the techniques of perturbed accelerated gradient descent and accelerated negative curvature finding. In the online stochastic setting, the best gradient query complexity result $\tilde{\mathcal{O}}(1/\epsilon^3)$ is achieved by SPIDER-SFO$^+$ [16], which combined the near-optimal ϵ-approximate FOSP finding algorithm SPIDER and the NCF framework Neon2 to find an (ϵ, δ)-approximate SOSP.

However, the gradient information is not always accessible. Many machine learning and deep learning applications often encounter situations where the calculation of explicit gradients is expensive or even infeasible, such as black-box adversarial attack on deep neural networks [42, 36, 13, 6, 46] and policy search in reinforcement learning [44, 14, 29]. Thus, zeroth-order (ZO) optimization, which uses function values to estimate the explicit gradients as an important gradient-based black-box method, is one of the best options for solving this type of ML/DL problem. A considerable body of work has shown that ZO algorithms based on gradient estimation have comparable convergence rates to their gradient-based counterparts. Although many gradient estimation-based ZO algorithms have been proposed in recent years, most of them focus on the performance of converging to FOSPs [40, 22, 25, 16], and only a few of them on SOSPs [27, 47, 35, 5].

As mentioned above, although there have been several works of finding local minima via ZO methods, they utilized the techniques of random perturbations [27, 47], random search [35], and cubic regularization [5], as shown in Table 1, which are not the most efficient ones of escaping saddle points as discussed before. Specifically, in the deterministic setting, [27] proposed the ZO perturbed

2

stochastic gradient (ZPSGD) method, which uses a batch of Gaussian smoothing based stochastic ZO gradient estimators and adds a random perturbation in each iteration. As a result, ZPSGD can find an ϵ-approximate SOSP using $\tilde{\mathcal{O}}\left(d^2/\epsilon^5\right)$ function queries. [47] proposed the perturbed approximate gradient descent (PAGD) method which iteratively conducts the gradient descent steps by utilizing the forward difference version of the coordinate-wise gradient estimators until it reaches a point with a small gradient. Then, PAGD adds a uniform perturbation and continues the gradient descent steps. The total function queries of PAGD to find an ϵ-approximate SOSP is $\tilde{\mathcal{O}}\left(d\log^4 d/\epsilon^2\right)$. Recently, [35] proposed the random search power iteration (RSPI) method, which alternately performs random search steps and power iteration steps. The power iteration step contains an inexact power iteration subroutine using only the ZO oracle to conduct the NCF, and the core idea is to use a finite difference approach to approximate the Hessian-vector product. In the stochastic setting, [5] proposed a zeroth-order stochastic cubic regularization newton (ZO-SCRN) method with function query complexity $\tilde{\mathcal{O}}\left(d/\epsilon^{7/2}\right)$ using Gaussian sampling-based gradient estimator and Hessian estimator. Unfortunately, each iteration of ZO-SCRN needs to solve a cubic minimization subproblem, which does not have a closed-form solution. Typically, inexact solvers for solving the cubic minimization subproblem need additional computations of the Hessian-vector product [1] or the gradient [7].

Thus, it is then natural to explore faster ZO negative curvature finding based algorithms to make escaping saddle points more efficient. To the best of our knowledge, negative curvature finding algorithms with access only to ZO oracle is still a vacancy in the stochastic setting. Inspired by the fact that the gradient can be approximated by the finite difference of function queries with high accuracy, a question is: *Can we turn FO NCF methods (especially the state-of-the-art Neon2) into ZO methods without increasing the iteration complexity and turn ZO algorithms of finding FOSPs into the ones of finding SOSPs?*

Contributions. We summarize our main contributions as follows:

- We give an affirmative answer to the above question. We propose two ZO negative curvature finding frameworks, which use only function queries and can detect whether there is a negative curvature direction at a given point x on a smooth, Hessian-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ in offline deterministic and online stochastic settings, respectively.

- We apply the proposed frameworks to four ZO algorithms and prove that these ZO algorithms can converge to (ϵ, δ)-approximate SOSPs, which are ZO-GD, ZO-SGD, ZO-SCSG, and ZO-SPIDER.

- In the deterministic setting, compared with the classical setting where $\delta = \mathcal{O}(\sqrt{\epsilon})$ [26, 28, 27, 47], or the special case $\delta = \epsilon^{2/3}$ [35], our Theorem 4 is always not worse than other algorithms in Table 1. In the online stochastic setting, all of our algorithms don't need to solve the cubic subproblem as in ZO-SCRN and our Theorem 6 improves the best function query complexity by a factor of $\tilde{\mathcal{O}}(1/\sqrt{\epsilon})$.

2 Preliminaries

Throughout this paper, we use $\|\cdot\|$ to denote the Euclidean norm of a vector and the spectral norm of a matrix. We use $\tilde{\mathcal{O}}(\cdot)$ to hide the poly-logarithmic terms. For a given set $\mathcal{S}$ drawn from $[n] := \{1, 2, \ldots, n\}$, define $f_{\mathcal{S}}(\cdot) := \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} f_i(\cdot)$.

Definition 1. *For a twice differentiable nonconvex function $f \colon \mathbb{R}^d \to \mathbb{R}$,*

- *f is ℓ-Lipschitz smooth if $\forall x, y \in \mathbb{R}^d, \|\nabla f(x) - \nabla f(y)\| \le \ell\|x - y\|$.*

- *f is ρ-Hessian Lipschitz if $\forall x, y \in \mathbb{R}^d, \|\nabla^2 f(x) - \nabla^2 f(y)\| \le \rho\|x - y\|$.*

Definition 2. *For a twice differentiable nonconvex function $f \colon \mathbb{R}^d \to \mathbb{R}$, we say*

- *$x \in \mathbb{R}^d$ is an ϵ-approximate first-order stationary point if $\|\nabla f(x)\| \le \epsilon$.*

- *$x \in \mathbb{R}^d$ is an (ϵ, δ)-approximate second-order stationary point if $\|\nabla f(x)\| \le \epsilon, \nabla^2 f(x) \succeq -\delta\mathbf{I}$.*

We need the following assumptions which are standard in the literature of finding SOSPs [4, 16, 51].

Assumption 1. *We assume that $f(\cdot)$ in (1) satisfies:*

3

- $\Delta_f := f(x_0) - f(x^*) < \infty$ *where* $x^* := \operatorname{argmin}_x f(x)$.

- *Each component function $f_i(x)$ is ℓ-Lipschitz smooth and ρ-Hessian Lipschitz.*

- *(For online case only) The variance of the stochastic gradient is bounded: $\forall x \in \mathbb{R}^d$, $\mathbb{E}\|\nabla f_i(x) - \nabla f(x)\|^2 \leq \sigma^2$.*

We'll also need the following more stringent assumption to get high-probability convergence results of ZO-SPIDER.

Assumption 2. *We assume that Assumption 1 holds, and in addition, the gradient of each component function $f_i(x)$ satisfies $\forall i, x \in \mathbb{R}^d$, $\|\nabla f_i(x) - \nabla f(x)\|^2 \leq \sigma^2$.*

2.1 ZO Gradient Estimators

Given a smooth, Hessian Lipschitz function f, a central difference version of the deterministic coordinate-wise gradient estimator is defined by

$$\hat{\nabla}_{coord} f(x) = \sum_{i=1}^{d} \frac{f(x + \mu e_i) - f(x - \mu e_i)}{2\mu} e_i, \qquad \text{(CoordGradEst)}$$

where e_i denotes a standard basis vector with 1 at its i-th coordinate and 0 otherwise; μ is the smoothing parameter, which is a sufficient small positive constant. A central difference version of the random gradient estimator is defined by

$$\hat{\nabla}_{rand} f(x) = d \frac{f(x + \mu u) - f(x - \mu u)}{2\mu} u, \qquad \text{(RandGradEst)}$$

where $u \in \mathbb{R}^d$ is a random direction drawn from a uniform distribution over the unit sphere; μ is the smoothing parameter, which is a sufficient small positive constant.

Remark 1. *Deterministic vs. Random: CoordGradEst needs d times more function queries than RandGradEst. However, as will be discussed in section 4, it has a lower approximation error and thus can reduce the iteration complexity. **Central Difference vs. Forward Difference** (please refer to Appendix A.1): Under the assumption of Hessian Lipschitz, a smaller approximation error bound can be obtained by the central difference version of both CoordGradEst and RandGradEst.*

2.2 ZO Hessian-Vector Product Estimator

By the definition of derivative: $\nabla^2 f(x) \cdot v = \lim_{\mu \to 0} \frac{\nabla f(x + \mu v) - \nabla f(x)}{\mu}$, we have $\nabla^2 f(x) \cdot v$ can be approximated by the difference of two gradients $\nabla f(x + v) - \nabla f(x)$ for some v with small magnitude. On the other hand, $\nabla f(x + v), \nabla f(x)$ can be approximated by $\hat{\nabla}_{coord} f(x + v), \hat{\nabla}_{coord} f(x)$ with high accuracy, respectively. Then the coordinate-wise Hessian-vector product estimator is defined by:

$$\mathcal{H}_f(x)v \triangleq \sum_{i=1}^{d} \frac{f(x + v + \mu e_i) - f(x + v - \mu e_i) + f(x - \mu e_i) - f(x + \mu e_i)}{2\mu} e_i. \qquad (2)$$

Note that we do not need to know the explicit representation of $\mathcal{H}_f(x)$. It is merely used as a notation for a virtual matrix and can be viewed as the Hessian $\nabla^2 f(x_0)$ with minor perturbations. As stated in the following lemma, the approximation error is efficiently upper bounded.

Lemma 1. *Assume that f is ρ-Hessian Lipschitz, then for any smoothing parameter μ and $x \in \mathbb{R}^d$, we have*

$$\|\mathcal{H}_f(x)v - \nabla^2 f(x)v\| \leq \rho \left(\|v\|^2/2 + \sqrt{d}\mu^2/3 \right). \qquad (3)$$

The ZO Hessian-vector product estimator was previously studied in [50, 35], but we provide a tighter bound than that in Lemma 6 in [35]. This is because we utilize properties of the central difference version of the coordinate-wise gradient estimator under the Hessian Lipschitz assumption. It is then directly concluded that, if $f(\cdot)$ is quadratic, we have $\rho = 0$ and $\|\mathcal{H}_f(x)v - \nabla^2 f(x)v\| = 0$.

4

3 Zeroth-Order Negative Curvature Finding

In this section, we introduce how to find the negative curvature direction near the saddle point using zeroth-order methods. Recently, based on the fact that the Hessian-vector product $\nabla^2 f(x) \cdot v$ can be approximated by $\nabla f(x + v) - \nabla f(x)$ with approximation error up to $\mathcal{O}(\|v\|^2)$, [4] proposed a FO framework named Neon2 that can replace the Hessian-vector product computations in NCF subroutine with gradient computations and thus can turn a FO algorithm for finding FOSPs into a FO algorithm for finding SOSPs. Enlightened by Neon2, we propose two zeroth-order NCF frameworks (*i.e.*, *ZO-NCF-Online* and *ZO-NCF-Deterministic*) using only function queries to solve nonconvex problems in the online stochastic setting and offline deterministic setting, respectively.

3.1 Stochastic Setting

In this subsection, we focus on solving the NCF problem with zeroth-order methods under the online stochastic setting and propose *ZO-NCF-Online*. Before introducing *ZO-NCF-Online*, we first introduce *ZO-NCF-Online-Weak* with weak confidence of $2/3$ for solving the NCF problem.

We summarize *ZO-NCF-Online-Weak* in Algorithm 1. Specifically, *ZO-NCF-Online-Weak* consists of at most $T = \mathcal{O}(\frac{\log^2 d}{\delta^2})$ iterations and works as follows: Given a detection point x_0, add a random perturbation with small magnitude σ as the starting point. At the t-th iteration where $t = 1, \ldots, T$, set $\mu_t = \|x_t - x_0\|$ to be the smoothing parameter μ in (2). Then we keep updating $x_{t+1} = x_t - \eta \mathcal{H}_{f_i}(x_0)(x_t - x_0)$ where $\mathcal{H}_{f_i}(x_0)(x_t - x_0)$ is the ZO Hessian-vector product estimator and stops whenever $\|x_{t+1} - x_0\| \geq r$ or the maximum iteration number T is reached. Thus as long as Algorithm 1 does not terminate, we have that the approximation error $\|\mathcal{H}_{f_i}(x_0)(x_t - x_0) - \nabla^2 f_i(x_0)(x_t - x_0)\|$ can be bounded by $\mathcal{O}(\sqrt{d}r^2)$ according to Lemma 1. Note that, although the error bound is poorer by a factor of $\mathcal{O}(\sqrt{d})$ as compared to $Neon_{weak}^{online}$ in [4] which used the difference of two gradients to approximate the Hessian-vector product and achieve an approximation error up to $\mathcal{O}(r^2)$, with our choice of r in Algorithm 1, the error term is still efficiently upper bounded.

Algorithm 1 ZO-NCF-Online-Weak (f, x_0, δ)

1: $\eta \leftarrow \frac{\delta}{C_0^2 \ell^2 \log(100d)}, T \leftarrow \frac{C_0^2 \log(100d)}{\eta \delta}, \sigma \leftarrow \frac{\eta^2 \delta^3}{(100d)^{3C_0}\rho}, r \leftarrow (100d)^{C_0}\sigma$
2: $\xi \leftarrow \sigma \frac{\xi'}{\|\xi'\|}$, with $\xi' \sim \mathcal{N}(0, \mathbf{I})$
3: $x_1 \leftarrow x_0 + \xi$
4: **for** $t = 1, \ldots, T$ **do**
5: $\quad \mu_t \leftarrow \|x_t - x_0\|$
6: $\quad x_{t+1} = x_t - \eta \mathcal{H}_{f_i}(x_0)(x_t - x_0)$ with $\mu = \mu_t$ and $i \in [n]$
7: $\quad$ **if** $\|x_{t+1} - x_0\| \geq r$ **then return** $v = \frac{x_s - x_0}{\|x_s - x_0\|}$ for a uniformly random $s \in [t]$
Return $v = \perp$

Other than the additional error term caused by ZO approximation, the motivation of *ZO-NCF-Online-Weak* is almost the same as $Neon_{weak}^{online}$. That is, under reasonable control of the approximation error of the Hessian-vector product, using the update rule of Oja's method [41] to approximately calculate the eigenvector corresponding to the minimum eigenvalue of $\nabla^2 f(x_0) = \frac{1}{n} \sum_{i=1}^n \nabla^2 f_i(x_0)$. Under similar analysis, we conclude that as long as the minimum eigenvalue of $\nabla^2 f(x_0)$ satisfies $\lambda_{min}(\nabla^2 f(x_0)) \leq -\delta$, *ZO-NCF-Online-Weak* will stop before T and find a negative curvature direction that aligns well with the eigenvector corresponding to the minimum eigenvalue of $\nabla f^2(x_0)$. Then we have the following lemma:

Lemma 2 (ZO-NCF-Online-Weak). *The output v of Algorithm 1 satisfies: If $\lambda_{min}(\nabla^2 f(x_0)) \leq -\delta$, then with probability at least $2/3$, $v \neq \perp$ and $v^\mathsf{T} \nabla^2 f(x_0)v \leq -\frac{3}{4}\delta$.*

We summarize *ZO-NCF-Online* in Algorithm 2. Specifically, *ZO-NCF-Online* repeatedly calls *ZO-NCF-Online-Weak* for $\Theta(\log(1/p))$ times to boost the confidence of solving the NCF problem from $2/3$ to $1 - p$. We have the following results:

Lemma 3. *In the same setting as in Algorithm 2, define $z = \frac{1}{m} \sum_{j=1}^m v^\mathsf{T}(\mathcal{H}_{f_{i_j}}(x_0))v$. Then, if $\|v\| \leq \frac{\delta}{16d\rho}$ and $m = \Theta(\frac{\ell^2}{\delta^2})$, with probability at least $1 - p$, we have $\left| \frac{z}{\|v\|^2} - \frac{v^\mathsf{T} \nabla^2 f(x)v}{\|v\|^2} \right| \leq \frac{\delta}{4}$.*

5

Algorithm 2 ZO-NCF-Online

Input: $f(\cdot) = \frac{1}{n}\sum_{i=1}^{n} f_i(\cdot)$, x_0, $\delta > 0$, $p \in (0, 1]$.

1: **for** $j = 1, 2, \cdots, \Theta(\log(1/p))$ **do**

2:      $v_j \leftarrow$ ZO-NCF-Online-Weak (f, x_0, δ)

3:      **if** $v_j \neq \perp$ **then**

4:          $m \leftarrow \Theta(\frac{\ell^2 \log(1/p)}{\delta^2})$, $v' \leftarrow \Theta(\frac{\delta}{d\rho})v_j$

5:          Draw $i_1, \ldots, i_m$ uniformly randomly from $[n]$

6:          $z_j = \frac{1}{m}\sum_{k=1}^{m} \frac{(v')^T \mathcal{H}_{f_{i_k}}(x_0)v'}{\|v'\|^2}$

7:          **if** $z_j \leq -\frac{3\delta}{4}$ **then return** $v = v_j$

Return $v = \perp$

Theorem 1. *Let $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$ where each f_i is ℓ-smooth and ρ-Hessian Lipschitz. For every point $x_0 \in \mathbb{R}^d$, every $\delta \in (0, \ell]$, the output of Algorithm 2 v satisfies that, with probability at least $1 - p$: If $v = \perp$, then $\nabla^2 f(x_0) \succeq -\delta \mathbf{I}$; If $v \neq \perp$, then $\|v\| = 1$ and $v^\mathsf{T}\nabla^2 f(x_0)v \leq -\frac{\delta}{2}$. The total function query complexity is*

$$\mathcal{O}\left(\frac{d\log^2(d/p)\ell^2}{\delta^2}\right).$$

3.2 Deterministic Setting

In this subsection, we focus on solving the NCF problem with zeroth-order methods under the offline deterministic setting and propose *ZO-NCF-Deterministic*. We summarize *ZO-NCF-Deterministic* in Algorithm 3. Since we want to compute the eigenvector corresponding to the most negative eigenvalue of $\nabla^2 f(x_0)$ approximately, one can convert it into an approximated top eigenvector computation problem of $\mathbf{M} := -\frac{1}{\ell}\nabla^2 f(x_0) + (1 - \frac{3\delta}{4\ell})\mathbf{I}$. This is because all eigenvalues of $\nabla^2 f(x_0)$ in $[-\frac{3\delta}{4}, \ell]$ will be mapped to eigenvalues of $\mathbf{M}$ in $[-1, 1]$, and all eigenvalues of $\nabla^2 f(x_0)$ smaller than $-\delta$ will be mapped to eigenvalues of $\mathbf{M}$ greater than $1 + \frac{\delta}{4\ell}$.

Algorithm 3 ZO-NCF-Deterministic

Input: Function $f(\cdot)$, point x_0, negative curvature $\delta > 0$, confidence $p \in (0, 1]$.

1: $T \leftarrow \frac{C_1^2 \log \frac{d}{p}\sqrt{\ell}}{\sqrt{\delta}}$, $\sigma \triangleq (d/p)^{-2C_1} \frac{\delta}{T^4 \rho}$, $r \triangleq (d/p)^{C_1}\sigma$

2: $\xi \leftarrow \sigma \frac{\xi'}{\|\xi'\|}$, with $\xi' \sim \mathcal{N}(0, \mathbf{I})$

3: $x_1 \leftarrow x_0 + \xi$, $y_0 \leftarrow 0$, $y_1 \leftarrow \xi$

4: **for** $t = 1, \ldots, T$ **do**

5:      $\mu_t = \|y_t\|$

6:      $y_{t+1} = 2\mathcal{M}(y_t) - y_{t-1}$ where $\mathcal{M}(y) = (-\frac{1}{\ell}\mathcal{H}_f(x_0) + (1 - \frac{3\delta}{4\ell}))y$

7:      $x_{t+1} = x_0 + y_{t+1} - \mathcal{M}(y_t)$

8:      **if** $\|x_{t+1} - x_0\| \geq r$ **then return** $v = \frac{x_{t+1} - x_0}{\|x_{t+1} - x_0\|}$

Return $v = \perp$.

Similar to *ZO-NCF-Online-Weak*, *ZO-NCF-Deterministic* starts by adding a random perturbation ξ to the detection point x_0. To find the negative curvature direction v of $\nabla^2 f(x_0)$ such that $v^\mathsf{T}\nabla^2 f(x_0)v \leq -\frac{\delta}{2}$, the classical power method which updates through $x_{T+1} = x_0 + \mathbf{M}^T\xi$ [35] will take $T \geq \tilde{\Omega}(\frac{\ell}{\delta})$ number of iterations since eigenvalues of $\mathbf{M}$ greater than $1 + \frac{\delta}{4\ell}$ grows in a speed $(1 + \delta/\ell)^T$. To reduce the iteration complexity T, we can replace the matrix polynomial $\mathbf{M}^T$ with the matrix Chebyshev polynomial $\mathcal{T}_T(\mathbf{M})$ and virtually update $x_{T+1} = x_0 + \mathcal{T}_T(\mathbf{M})\xi$.

Definition 3. *Chebyshev polynomial $\{\mathcal{T}_n(x)\}_{n \geq 0}$ of the first kind is*

$$\mathcal{T}_0(x) = 1, \quad \mathcal{T}_1(x) = x, \quad \mathcal{T}_{n+1}(x) = 2x \cdot \mathcal{T}_n(x) - \mathcal{T}_{n-1}(x),$$

then it satisfies $\mathcal{T}_t(x) = \begin{cases} \cos(n\arccos(x)), & x \in [-1, 1] \\ \frac{1}{2}[(x - \sqrt{x^2 - 1})^n + (x + \sqrt{x^2 - 1})^n], & x > 1 \end{cases}$.

6

In the matrix case, we have the so-called matrix Chebyshev polynomial $\mathcal{T}_t(\mathbf{M})$ [3], which satisfies: $\mathcal{T}_{t+1}(\mathbf{M})\xi = 2\mathbf{M}\mathcal{T}_t(\mathbf{M})\xi - \mathcal{T}_{t-1}(\mathbf{M})\xi$. Thus, eigenvalues of $\mathbf{M}$ greater than $1 + \frac{\delta}{4\ell}$ will grow to $\left(1 + \delta/4\ell + \sqrt{(\delta/4\ell)^2 + \delta/2\ell}\right)^T \approx \left(1 + \sqrt{\delta/\ell}\right)^T$, so we only need to choose $T \geq \sqrt{\ell/\delta}$.

On the other hand, since we only have access to the zeroth-order information, we need to stably compute the matrix Chebyshev polynomial. In algorithm 3, we set $\mu_t = \|y_t\|$ and use $\mathcal{M}(y_t) = (-\frac{1}{\ell}\mathcal{H}_f(x_0) + (1 - \frac{3\delta}{4\ell}))y_t$ to approximate $\mathbf{M}y_t$ with approximation error up to $\frac{2\rho\sqrt{d}rt}{\ell}\|y_t\|$. With proper choice of r, it allows us to use the inexact backward recurrence [3] to ensure a stable computation of matrix Chebyshev polynomial:

$$y_0 = 0, \quad y_1 = \xi, \quad y_{t+1} = 2\mathcal{M}(y_t) - y_{t-1}.$$

Then the output $x_{T+1} = x_0 + y_{T+1} - \mathcal{M}(y_T)$ is close to $x_0 + \mathcal{T}_T(\mathbf{M})\xi$ with a small approximation error. Finally, we have the following theorem:

Theorem 2. *Let $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ where each f_i is ℓ-smooth and ρ-Hessian Lipschitz. For every point $x_0 \in \mathbb{R}^d$, every $\delta \in (0, \ell]$, the output of Algorithm 3 v satisfies that, with probability at least $1 - p$: If $v = \perp$, then $\nabla^2 f(x_0) \succeq -\delta\mathbf{I}$; If $v \neq \perp$, then $\|v\| = 1$ and $v^\mathsf{T}\nabla^2 f(x_0)v \leq -\frac{\delta}{2}$. The function query complexity is*

$$\mathcal{O}(\frac{d\log\frac{d}{p}\sqrt{\ell}}{\sqrt{\delta}}).$$

4 Applications of Zeroth-Order Negative Curvature Finding

In this section, we focus on applying the zeroth-order negative curvature frameworks to the following ZO algorithms: ZO-GD, ZO-SGD, ZO-SCSG, and ZO-SPIDER. The following result shows that one can verify if a point x is an ϵ-approximate FOSP using CoordGradEst.

Proposition 1. *In the online setting, using CoordGradEst with a batch size of $\mathcal{O}\left(\left(\frac{128\sigma^2}{\epsilon^2} + 1\right)\log\frac{1}{p}\right)$ and smoothing parameter $\mu \leq \sqrt{\frac{3\epsilon}{4\rho\sqrt{d}}}$, we can verify with probability at least $1 - p$, either $\|\nabla f(x)\| \geq \epsilon/2$ or $\|\nabla f(x)\| \leq \epsilon$. In the deterministic setting, using once computation of CoordGradEst with smoothing parameter $\mu \leq \sqrt{\frac{3\epsilon}{2\rho\sqrt{d}}}$, we can verify with probability 1, either $\|\nabla f(x)\| \geq \epsilon/2$ or $\|\nabla f(x)\| \leq \epsilon$.*

4.1 Applying Zeroth-Order Negative Curvature Finding to ZO-GD and ZO-SGD

We apply ZO-NCF-Online to ZO-SGD to turn it into a local minima finding algorithm, and propose ZO-SGD-NCF in Algorithm 4. At each iteration, we use a batch size of $\mathcal{O}\left(\frac{\sigma^2}{\epsilon^2}\log\left(\frac{2K}{p}\right)\right)$ Coord-GradEst to verify if x_t is an ϵ-approximate stationary point. If not, ZO-SGD-NCF either estimates the gradient $\nabla f_S(x_t) = \frac{1}{|S|}\sum_{i \in S}\nabla f_i(x_t)$ by CoordGradEst (**Option I**) or RandGradEst (**Option II**) with both mini-batch size $\mathcal{O}(\frac{\sigma^2}{\epsilon^2})$; If so, we call the ZO-NCF-Online subroutine. Then, If we find an approximate negative curvature direction v around x_t, then we update x_{t+1} by moving from x_t in the direction v with step-size δ/ρ. We have the following theorem:

Theorem 3. *Under Assumption 1, we set $\mu_1 = \sqrt{\frac{3\epsilon}{2\rho\sqrt{d}}}$ and other parameters as follows,*

Option I: $|S| = \max\{\frac{32\sigma^2}{\epsilon^2}, 1\}, K = \mathcal{O}(\frac{\rho^2\Delta_f}{\delta^3} + \frac{\ell\Delta_f}{\epsilon^2}), \eta = \frac{1}{4\ell}, \mu_2 = \sqrt{\frac{3\epsilon}{4\rho\sqrt{d}}};$

Option II: $|S| = \max\{\frac{8\sigma^2}{\epsilon^2}, 1\}, K = \mathcal{O}(\frac{\rho^2\Delta_f}{\delta^3} + \frac{d\ell\Delta_f}{\epsilon^2}), \eta = \frac{1}{32d\ell}, \mu_2 = \min\left\{\sqrt{\frac{3\epsilon}{4\rho d}}, \frac{\epsilon}{32\sqrt{d\ell}}\right\},$

where μ_1 and μ_2 are only used in Line 3 and Line 5 (or Line 6) of Algorithm 4, respectively. With probability at least $1 - p$, Algorithm 4 outputs an (ϵ, δ)-approximate local minimum in function query complexity

Option I: $\tilde{\mathcal{O}}(\frac{d\sigma^2\ell\Delta_f}{\epsilon^4} + \frac{d\sigma^2\rho^2\Delta_f}{\epsilon^2\delta^3} + \frac{d\ell^2\rho^2\Delta_f}{\delta^5})$; ***Option II:*** $\tilde{\mathcal{O}}(\frac{d^2\sigma^2\ell\Delta_f}{\epsilon^4} + \frac{d\sigma^2\rho^2\Delta_f}{\epsilon^2\delta^3} + \frac{d\ell^2\rho^2\Delta_f}{\delta^5})$.

7

Algorithm 4 ZO-SGD-NCF

Input: Function f, starting point x_0, confidence $p \in (0,1)$, $\epsilon > 0$ and $\delta > 0$.

 1: **for** $t = 0, \dots, K-1$ **do**

 2:      uniformly randomly choose a set $\mathcal{B}$ with batch size $\mathcal{O}(\frac{\sigma^2}{\epsilon^2} \log(2K/p))$

 3:      **if** $\|\hat{\nabla}_{coord} f_{\mathcal{B}}(x_t)\| \geq \frac{3\epsilon}{4}$ **then**

 4:          uniformly randomly choose $S \subseteq [n]$

 5:          **Option I :** $x_{t+1} \leftarrow x_t - \eta \hat{\nabla}_{coord} f_S(x_t)$

 6:          **Option II :** $x_{t+1} \leftarrow x_t - \eta \hat{\nabla}_{rand} f_S(x_t)$

 7:      **else**

 8:          $v \leftarrow$ ZO-NCF-Online $(f, x_t, \delta, \frac{p}{2K})$

 9:          **if** $v = \perp$ **then return** x_t

10:          **else** $x_{t+1} = x_t \pm \frac{\delta}{\rho} v$

Algorithm 5 ZO-GD-NCF

Input: Function f, starting point x_0, confidence $p \in (0,1)$, $\epsilon > 0$ and $\delta > 0$.

 1: **for** $t = 0, \dots, K-1$ **do**

 2:      **if** $\|\hat{\nabla}_{coord} f(x_t)\| \geq \frac{3\epsilon}{4}$ **then**

 3:          **Option I :** $x_{t+1} \leftarrow x_t - \eta \hat{\nabla}_{coord} f(x_t)$

 4:          **Option II :** $x_{t+1} \leftarrow x_t - \eta \hat{\nabla}_{rand} f(x_t)$

 5:      **else**

 6:          $v \leftarrow$ ZO-NCF-Deterministic $(f, x_t, \delta, \frac{p}{K})$

 7:          **if** $v = \perp$ **then return** x_t

 8:          **else** $x_{t+1} = x_t \pm \frac{\delta}{\rho} v$

Remark 2. *Note that the dominant term of the function query complexity in **Option I** is $\tilde{\mathcal{O}}(\frac{d}{\epsilon^4})$, while in **Option II** is $\tilde{\mathcal{O}}(\frac{d^2}{\epsilon^4})$. This is because CoordGradEst has a lower approximation error and thus can reduce the iteration complexity by a factor of d. Then the function query complexity of **Option II** is dominated by evaluating the magnitude of the gradient (Line 3 in Algorithm 4).*

In the Deterministic setting, we apply ZO-NCF-Deterministic to ZO-GD to turn it into a local minima finding algorithm and propose ZO-GD-NCF in Algorithm 5. The update rule of ZO-GD-NCF is similar to that in ZO-SGD-NCF, the only difference is that we don't need to use mini-batch sampling of the stochastic gradient. Similarly, we have the following theorem:

Theorem 4. *Under Assumption 1, we set $\mu_1 = \sqrt{\frac{3\epsilon}{2\rho\sqrt{d}}}$ and other parameters as follows,*

$$\textbf{\textit{Option I:}} \ K = \mathcal{O}(\frac{\rho^2 \Delta_f}{\delta^3} + \frac{\ell \Delta_f}{\epsilon^2}), \eta = \frac{1}{4\ell}, \mu_2 = \sqrt{\frac{3\epsilon}{4\rho\sqrt{d}}};$$

$$\textbf{\textit{Option II:}} \ K = \mathcal{O}(\frac{\rho^2 \Delta_f}{\delta^3} + \frac{d\ell \Delta_f}{\epsilon^2}), \eta = \frac{1}{8d\ell}, \mu_2 = \min\left\{\sqrt{\frac{3\epsilon}{4\rho d}}, \frac{\epsilon}{16\sqrt{d}\ell}\right\},$$

where μ_1 and μ_2 are only used in Line 2 and Line 3 (or Line 4) of Algorithm 5, respectively. With probability at least $1 - p$, Algorithm 5 outputs an (ϵ, δ)-approximate local minimum in function query complexity

$$\textbf{\textit{Option I:}} \quad \tilde{\mathcal{O}}(\frac{d\ell\Delta_f}{\epsilon^2} + d\frac{\sqrt{\ell}}{\sqrt{\delta}}\frac{\rho^2 \Delta_f}{\delta^3}); \quad \textbf{\textit{Option II:}} \quad \tilde{\mathcal{O}}(\frac{d^2\ell\Delta_f}{\epsilon^2} + d\frac{\sqrt{\ell}}{\sqrt{\delta}}\frac{\rho^2 \Delta_f}{\delta^3}).$$

4.2 Applying Zeroth-Order Negative Curvature Finding to ZO-SCSG and ZO-SPIDER

In the stochastic setting, we can also apply the zeroth-order negative curvature finding to the variance reduction-based algorithms: SCSG [32] and SPIDER [16]. Due to space limitation, We defer the detailed discussions of these applications to Appendix E and F.

To apply ZO-NCF-Online to SCSG, we first propose a zeroth-order variant of the SCSG [32] method in Algorithm 6. At the beginning of the j-th epoch, we estimate the gradient $\nabla f_{\mathcal{I}_j}(\tilde{x}_{j-1})$ by

8

CoordGradEst over a batch sampling set $\mathcal{I}_j$ with size B. In the inner loop iterations, the stochastic gradient estimator v_{k-1}^j is either constructed by CoordGradEst or by RandGradEst over a mini-batch sampling set $\mathcal{I}_{k-1}^j$ with size b. Then we apply ZO-NCF-Online to ZO-SCSG and propose the ZO-SCSG-NCF method (see Algorithm 7).

Theorem 5 (informal, full version deferred to Appendix E). *With probability at least $\frac{2}{3}$, for both **Option I** and **Option II**, Algorithm 7 outputs an (ϵ, δ)-approximate local minimum in function query complexity*

$$\tilde{\mathcal{O}}(d(\frac{\ell\Delta_f}{\epsilon^{\frac{4}{3}}\sigma^{\frac{2}{3}}} + \frac{\rho^2\Delta_f}{\delta^3})(\frac{\sigma^2}{\epsilon^2} + \frac{\ell^2}{\delta^2}) + d\frac{\ell\Delta_f}{\epsilon^2}\frac{\ell^2}{\delta^2}).$$

We apply ZO-NCF-Online to ZO-SPIDER to turn it into a local minima finding algorithm and propose ZO-SPIDER-NCF in Algorithm 8. As a by-product, we also propose a zeroth-order variant of the SPIDER method in Appendix G that can converge to an ϵ-approximate FOSP with high probability rather than expectation. Using the same technique as in SPIDER-SFO$^+$ [16], that is, instead of moving in a large single step with size δ/ρ along the approximate negative curvature direction as in ZO-SGD-NCF and ZO-SCSG-NCF, we can split it into $\delta/(\rho\eta)$ equal length mini-steps with size η. As a result, we can maintain the SPIDER estimates and improve the so-called non-improvable coupling term $\frac{1}{\delta^3\epsilon^2}$ by a fact of δ.

Theorem 6 (informal, full version deferred to Appendix F). *With probability at least $\frac{3}{4}$, Algorithm 8 outputs an (ϵ, δ)-approximate local minimum in function query complexity*

$$\tilde{\mathcal{O}}\left(d\left(\frac{\sigma\ell\Delta_f}{\epsilon^3} + \frac{\sigma\ell\rho\Delta_f}{\epsilon^2\delta^2} + \frac{\ell^2\rho\Delta_f}{\delta^3\epsilon} + \frac{\ell^2\rho^2\Delta_f}{\delta^5} + \frac{\sigma^2}{\epsilon^2} + \frac{\sigma\delta\ell}{\rho\epsilon^2} + \frac{\ell^2}{\delta^2}\right)\right).$$

Remark 3. *We can boost the confidence the of Theorem 5 and 6 to $1 - p$ by running $\log(1/p)$ copies of Algorithm 7 and 8.*

5 Numerical Experiments

Octopus Function. We first consider the octopus function proposed by Du et al. [15]. The octopus function has 2^d local optimum: $x^* = (\pm 4\tau, \ldots, \pm 4\tau)^\mathsf{T}$ and $2^d - 1$ saddle points:

$$(0, \ldots, 0)^\mathsf{T}, (\pm 4\tau, 0, \ldots, 0)^\mathsf{T}, \ldots, (\pm 4\tau, \ldots, \pm 4\tau, 0)^\mathsf{T}.$$

We compare ZO-GD-NCF, ZPSGD, PAGD, and RSPI on the octopus function with growing dimensions. The parameters corresponding to the octopus function are set with $\tau = e, L = e, \gamma = 1$. All algorithms are initialized at point $(0, \ldots, 0)^\mathsf{T}$, which is a strict saddle point and the one farthest from the optimal points among the $2^d - 1$ saddle points.

We set $\epsilon = 1e - 4, \delta = \sqrt{\rho\epsilon}$ for all experiments and report the function value v.s. the number of function queries in Figure 1. For RSPI, we follow the hyperparameter update strategy as described in ([35], Appendix, Section F): We keep σ_2 constant and update $\sigma_1 = \rho_{\sigma_1}\sigma_1$ every T_{σ_1} iterations. We conduct a grid search for T_{σ_1} and ρ_{σ_1}.

| (a) d=10 | (b) d=30 | (c) d=50 | (d) d=100 |
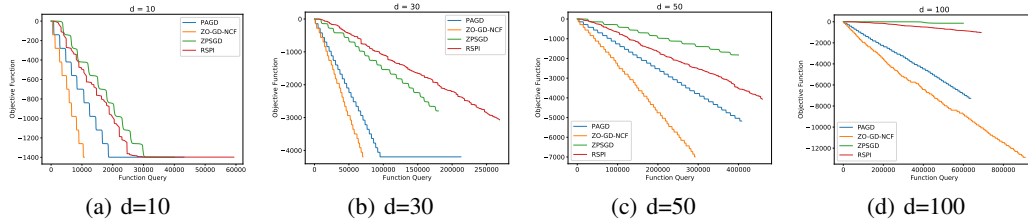
Figure 1: Performance of ZO-GD-NCF, ZPSGD, PAGD, and RSPI on the octopus function with growing dimensions.

The results in Figure 1 illustrate that all algorithms are able to escape saddle points. With the increase of the dimension of the octopus function, more function queries are needed for each algorithm to converge to the local minimum. Note that in all experiments, RSPI performs worse than PAGD and

9

ZO-GD-NCF. This is because RSPI is not a gradient based algorithm. Although it can efficiently escape from the saddle point using the negative curvature finding, it converges very slowly when the current point is far from the saddle point due to the random search.

We defer more experimental results to Appendix G.

6 Conclusion

In this paper, we analyse two types of ZO negative curvature finding frameworks, which can be used to find the negative curvature directions near a saddle point in the deterministic setting and stochastic setting, respectively. We apply the two frameworks to four ZO algorithms and analyse the complexities for converging to (ϵ, δ)-approximate SOSPs. Finally, we conduct several numerical experiments to verify the effectiveness of the proposed method in escaping saddle points.

As a future work, it would be interesting to study the (zeroth-order) unified negative curvature finding frameworks with generic analysis that can be applied to any FOSPs finding algorithms.

Acknowledgments and Disclosure of Funding

The authors thank four anonymous reviewers for their constructive comments and suggestions. Bin Gu was partially supported by the National Natural Science Foundation of China under Grant 62076138.

References

[1] Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1195–1199, 2017.

[2] Zeyuan Allen-Zhu. Natasha 2: Faster non-convex optimization than sgd. *Advances in Neural Information Processing Systems*, 31, 2018.

[3] Zeyuan Allen-Zhu and Yuanzhi Li. Faster principal component regression and stable matrix chebyshev approximation. In *International Conference on Machine Learning*, pages 107–115. PMLR, 2017.

[4] Zeyuan Allen-Zhu and Yuanzhi Li. Neon2: Finding local minima via first-order oracles. *Advances in Neural Information Processing Systems*, 31, 2018.

[5] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, 22(1):35–76, 2022.

[6] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 154–169, 2018.

[7] Yair Carmon and John C Duchi. Gradient descent efficiently finds the cubic-regularized non-convex newton step. *arXiv preprint arXiv:1612.00547*, 2016.

[8] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. "convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *International Conference on Machine Learning*, pages 654–663. PMLR, 2017.

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

[10] 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.

10

[11] 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.

[12] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[13] 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, 2017.

[14] Krzysztof Choromanski, Mark Rowland, Vikas Sindhwani, Richard Turner, and Adrian Weller. Structured evolution with compact architectures for scalable policy optimization. In *International Conference on Machine Learning*, pages 970–978. PMLR, 2018.

[15] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. *Advances in Neural Information Processing Systems*, 30, 2017.

[16] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.

[17] Xiang Gao, Bo Jiang, and Shuzhong Zhang. On the information-adaptive variants of the admm: an iteration complexity perspective. *Journal of Scientific Computing*, 76(1):327–363, 2018.

[18] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points-online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pages 797–842. PMLR, 2015.

[19] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *Advances in Neural Information Processing Systems*, 29, 2016.

[20] Rong Ge, Jason D. Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. In *International Conference on Learning Representations*, 2018.

[21] Rong Ge and Tengyu Ma. On the optimization landscape of tensor decompositions. *Advances in Neural Information Processing Systems*, 30, 2017.

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

[23] Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

[24] Prateek Jain, Chi Jin, Sham Kakade, and Praneeth Netrapalli. Global convergence of non-convex gradient descent for computing matrix squareroot. In *Artificial Intelligence and Statistics*, pages 479–488. PMLR, 2017.

[25] Kaiyi Ji, Zhe Wang, Yi Zhou, and Yingbin Liang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International Conference on Machine Learning*, pages 3100–3109. PMLR, 2019.

[26] 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. PMLR, 2017.

[27] Chi Jin, Lydia T Liu, Rong Ge, and Michael I Jordan. On the local minima of the empirical risk. *Advances in Neural Information Processing Systems*, 31, 2018.

[28] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference On Learning Theory*, pages 1042–1085. PMLR, 2018.

11

[29] Gangshan Jing, He Bai, Jemin George, Aranya Chakrabortty, and Piyush K Sharma. Asynchronous distributed reinforcement learning for lqr control via zeroth-order block coordinate descent. *arXiv preprint arXiv:2107.12416*, 2021.

[30] Kenji Kawaguchi. Deep learning without poor local minima. *Advances in Neural Information Processing Systems*, 29, 2016.

[31] Kenji Kawaguchi, Jiaoyang Huang, and Leslie Pack Kaelbling. Every local minimum value is the global minimum value of induced model in nonconvex machine learning. *Neural Computation*, 31(12):2293–2323, 2019.

[32] Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via scsg methods. *Advances in Neural Information Processing Systems*, 30, 2017.

[33] Huan Li and Zhouchen Lin. Restarted nonconvex accelerated gradient descent: No more polylogarithmic factor in the $\mathcal{O}(\epsilon^{-7/4})$ complexity. *arXiv preprint arXiv:2201.11411*, 2022.

[34] Mingrui Liu, Zhe Li, Xiaoyu Wang, Jinfeng Yi, and Tianbao Yang. Adaptive negative curvature descent with applications in non-convex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

[35] Aurelien Lucchi, Antonio Orvieto, and Adamos Solomou. On the second-order convergence properties of random search methods. *Advances in Neural Information Processing Systems*, 34, 2021.

[36] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[37] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[38] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

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

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

[41] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, 1982.

[42] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 506–519, 2017.

[43] Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning*, pages 314–323. PMLR, 2016.

[44] 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.

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

[46] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:742–749, 07 2019.

[47] Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, and Georgios Piliouras. Efficiently avoiding saddle points with zero order methods: No gradients required. *Advances in Neural Information Processing Systems*, 32, 2019.

12

[48] Wikipedia contributors. Azuma's inequality — Wikipedia, the free encyclopedia, 2021. [Online; accessed 22-March-2022].

[49] Yi Xu, Rong Jin, and Tianbao Yang. Neon+: Accelerated gradient methods for extracting negative curvature for non-convex optimization. *arXiv preprint arXiv:1712.01033*, 2017.

[50] Haishan Ye, Zhichao Huang, Cong Fang, Chris Junchi Li, and Tong Zhang. Hessian-aware zeroth-order optimization for black-box adversarial attack. *arXiv preprint arXiv:1812.11377*, 2018.

[51] Chenyi Zhang and Tongyang Li. Escape saddle points by a simple gradient-descent based algorithm. *Advances in Neural Information Processing Systems*, 34, 2021.

Checklist

1. For all authors...

    (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

    (b) Did you describe the limitations of your work? [Yes] In the conclusion, we mention a way that could improve the current work.

    (c) Did you discuss any potential negative societal impacts of your work? [No]

    (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

    (a) Did you state the full set of assumptions of all theoretical results? [Yes]

    (b) Did you include complete proofs of all theoretical results? [Yes] Please refer to the Appendix for complete proofs of the theoretical results.

3. If you ran experiments...

    (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]

    (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

    (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]

    (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

    (a) If your work uses existing assets, did you cite the creators? [Yes]

    (b) Did you mention the license of the assets? [Yes]

    (c) Did you include any new assets either in the supplemental material or as a URL? [No]

    (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

    (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

    (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

    (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

    (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

13

Appendix

A   Auxiliary Lemmas

Lemma 4 ([38], Lemma 1.2.3 & 1.2.4). *If f is ℓ-Lipschitz smooth, then for all $x, y \in \mathbb{R}^d$,*

$$|f(y) - f(x) - \nabla f(x)^T(y - x)| \leq \frac{\ell}{2}\|y - x\|^2.$$

If f is ρ-Hessian Lipschitz, then for all $x, y \in \mathbb{R}^d$,

$$\left\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x)\right\| \leq \frac{\rho}{2}\|y - x\|^2$$

and

$$\left|f(y) - f(x) - \nabla f(x)^T(y - x) - \frac{1}{2}(y - x)^T\nabla^2 f(x)(y - x)\right| \leq \frac{\rho}{6}\|y - x\|^3$$

Lemma 5 ([4], Fact 2.2). *If $v_1, \ldots, v_n \in \mathbb{R}^d$ satisfy $\sum_{i=1}^n v_i = 0$, and S is a non-empty, uniform random subset of $[n]$. Then*

$$\mathbb{E}\left[\left\|\frac{1}{|S|}\sum_{i \in S} v_i\right\|^2\right] \leq \frac{\mathbb{I}[|S| < n]}{|S|}\frac{1}{n}\sum_{i \in [n]}\|v_i\|^2 \tag{4}$$

Lemma 6 ([4], Claim C.2). *If v is a unit vector and $v^\mathsf{T}\nabla^2 f(y)v \leq -\frac{\delta}{2}$, suppose we choose $y' = y \pm \frac{\delta}{\rho}v$ where the sign is random, then $f(y) - \mathbb{E}[f(y')] \geq \frac{\delta^3}{12\rho^2}$.*

A.1   Lemmas of ZO Gradient Estimators

Lemma 7. *For any given smoothing parameter μ and any $x \in \mathbb{R}^d$, if f is ℓ-Lipschitz smooth, then for both CoordGradEst and the forward difference version of the coordinate-wise gradient: $\hat{\nabla}_{coord}f(x) = \sum_{i=1}^d \frac{f(x + \mu e_i) - f(x)}{\mu}e_i$, we have*

$$\|\hat{\nabla}_{coord}f(x) - \nabla f(x)\|^2 \leq \ell^2 d\mu^2. \tag{5}$$

If we further assume that f has ρ-Lipschitz Hessian, we have

$$\|\hat{\nabla}_{coord}f(x) - \nabla f(x)\|^2 \leq \frac{1}{36}\rho^2 d\mu^4 \tag{6}$$

Proof. For the ℓ-Lipschitz gradient case, the proof directly follows from Lemma 3 in [25]. For the ρ-Hessian Lipschitz case, we have

$$\left\|\nabla f(x) - \hat{\nabla}_{coord}f(x)\right\| = \left\|\sum_{i=1}^d \frac{f(x + \mu e_i) - f(x - \mu e_i)}{2\mu}e_i - \nabla f(x)\right\|$$

$$= \frac{1}{2\mu}\left\|\sum_{i=1}^d (f(x + \mu e_i) - f(x - \mu e_i) - 2\mu\nabla_i f(x))e_i\right\|$$

For all $i \in [d]$, we have

$$f(x + \mu e_i) - f(x - \mu e_i) - 2\mu\nabla_i f(x)$$

$$= \left[f(x + \mu e_i) - f(x) - \mu\nabla_i f(x) - \frac{\mu^2}{2}\nabla_{ii}^2 f(x)\right] - \left[f(x - \mu e_i) - f(x) + \mu\nabla_i f(x) - \frac{\mu^2}{2}\nabla_{ii}^2 f(x)\right]$$

$$\leq \left|f(x + \mu e_i) - f(x) - \mu\nabla_i f(x) - \frac{\mu^2}{2}\nabla_{ii}^2 f(x)\right| + \left|f(x - \mu e_i) - f(x) + \mu\nabla_i f(x) - \frac{\mu^2}{2}\nabla_{ii}^2 f(x)\right|$$

$$\overset{\text{①}}{\leq} 2 \cdot \frac{\rho}{6}\mu^3 = \frac{\rho}{3}\mu^3$$

14

where ① is due to Lemma 4.

$$\left\| \nabla f(x) - \hat{\nabla}_{coord} f(x) \right\|$$

$$= \frac{1}{2\mu} \left\| \sum_{i=1}^{d} (f(x + \mu e_i) - f(x - \mu e_i) - 2\mu \nabla_i f(x)) e_i \right\|$$

$$= \frac{1}{2\mu} \sqrt{\sum_{i=1}^{d} (f(x + \mu e_i) - f(x - \mu e_i) - 2\mu \nabla_i f(x))^2}$$

$$\leq \frac{1}{2\mu} \sqrt{d \left(\frac{\rho \mu^3}{3} \right)^2} = \frac{\sqrt{d} \rho \mu^2}{6} \tag{7}$$

$\square$

Lemma 8. *Let $f_\mu(x) = \mathbb{E}_{u \sim U_B} f(x + \mu u)$ be a smooth approximation of $f(x)$, where U_B is the uniform distribution over the d-dimension unit Euclidean ball B. Given the gradient estimator $\hat{\nabla}_{rand} f(x) = d \frac{f(x + \mu u) - f(x - \mu u)}{2\mu} u$, we have*

If we assume that f is ℓ-Lipshitz smooth, then it has similar properties as its forward version :
$\hat{\nabla}_{rand} f(x) = d \frac{f(x + \mu u) - f(x)}{\mu} u$.

  (1) $|f_\mu(x) - f(x)| \leq \frac{\ell \mu^2}{2}$.

  (2) $\mathbb{E} \hat{\nabla}_{rand} f_S(x) = \mathbb{E} \frac{1}{|S|} \sum_{i \in S} f_i(x) = \nabla f_\mu(x)$, *where $S \in [n]$.*

  (3) $\|\nabla f_\mu(x) - \nabla f(x)\| \leq \frac{\ell d \mu}{2}$ *for any $x \in \mathbb{R}^d$.*

  (4) $\mathbb{E}[\|\hat{\nabla}_{rand} f(x)\|^2] \leq 2d \|\nabla f(x)\|^2 + \frac{\ell^2 d^2 \mu^2}{2}$ *for any $x \in \mathbb{R}^d$.*

  (5) $\mathbb{E}\|\hat{\nabla}_{rand} f(x) - \hat{\nabla}_{rand} f(y)\|^2 \leq 3d\ell^2 \|x - y\|^2 + \frac{3\ell^2 d^2 \mu^2}{2}$ *for any $x, y \in \mathbb{R}^d$*

If we further assume that f is ρ-Hessian Lipschitz, we have

  (1) $\left| f_\mu(x) - f(x) - \frac{\mu^2}{2} \frac{\sum_{i=1}^{d} \lambda_i}{d} \right| \leq \frac{\rho \mu^3}{6}$, *where $\lambda_i, i = 1, \ldots, d$ are eigenvalues of $\nabla^2 f(x)$.*

  (2) $\|\nabla f_\mu(x) - \nabla f(x)\| \leq \frac{\rho d \mu^2}{6}$ *for any $x \in \mathbb{R}^d$.*

  (3) $\mathbb{E}\|\hat{\nabla}_{rand} f(x)\|^2 \leq d \|\nabla f(x)\|^2 + \frac{\rho^2 d^2 \mu^4}{36}$ *for any $x \in \mathbb{R}^d$.*

  (4) $\mathbb{E}\|\hat{\nabla}_{rand} f(x) - \hat{\nabla}_{rand} f(y)\|^2 \leq 2d\ell^2 \|x - y\|^2 + \frac{\rho^2 d^2 \mu^4}{18}$ *for any $x, y \in \mathbb{R}^d$*

Remark 4. *The inequality of item (1) shows that increasing the level of smoothness of function f beyond ℓ-smooth cannot improve the approximation ability of f by f_μ. Consider a special case that $f^*(x) = \frac{\ell \|x\|^2}{2}$, then we have f^* is ℓ-Lipschitz smooth and 0-Hessian Lipschitz, and according to item (1)*

$$|f_\mu^*(x) - f^*(x)| = \left| \frac{\mu^2}{2} \frac{\sum_{i=1}^{d} \lambda_i}{d} \right| = \frac{\ell \mu^2}{2}.$$

This means that $\frac{\ell \mu^2}{2}$ is a tight bound of $|f_\mu(x) - f(x)|$ when f is assumed to be ℓ-smooth and ρ-Hessian Lipschitz.

Proof. Let $\alpha(d)$ be the volume of the unit ball in $\mathbb{R}^d$, and $\beta(d)$ be the surface area of the unit sphere in $\mathbb{R}^d$. Denote by B and S_p the unit ball and unit sphere, respectively.

When f is ℓ-Lipschitz smooth, the proof directly follow from Lemma4.1 in [17] and [25].

15

When f is ρ-Hessian Lipschitz, We first prove item (1).

$$\frac{1}{\alpha(d)} \int_B \left[f(x + \mu u) - f(x) - \langle \nabla f(x), \mu u \rangle - \frac{\mu^2}{2} \langle \nabla^2 f(x) u, u \rangle \right] du$$

$$= f_\mu(x) - f(x) - \frac{\mu^2}{2} \mathbb{E}_{u \sim U_B} \langle \nabla^2 f(x) u, u \rangle$$

$$= f_\mu(x) - f(x) - \frac{\mu^2}{2} \frac{\text{tr}\left(\nabla^2 f(x)\right)}{d} = f_\mu(x) - f(x) - \frac{\mu^2}{2} \frac{\sum_{i=1}^d \lambda_i}{d}$$

where $\lambda_i, i = 1, \ldots, d$ are eigenvalues of $\nabla^2 f(x)$. Therefore

$$\left| f_\mu(x) - f(x) - \frac{\mu^2}{2} \frac{\sum_{i=1}^d \lambda_i}{d} \right|$$

$$\leq \frac{1}{\alpha(d)} \int_B \left| f(x + \mu u) - f(x) - \langle \nabla f(x), \mu u \rangle - \frac{\mu^2}{2} \langle \nabla^2 f(x) u, u \rangle \right| du$$

$$\leq \frac{1}{\alpha(d)} \int_B \frac{\rho \mu^3}{6} \|u\|^3 du \overset{\text{①}}{\leq} \frac{\rho \mu^3}{6} \frac{d}{d+3} \leq \frac{\rho \mu^3}{6}$$

where ① is due to $\frac{1}{\alpha(d)} \int_B \|u\|^p du = \frac{d}{d+p}$ as proved in Lemma 7.3(a) in [17].

Then we prove item (2). Denote

$$a_u(x, \mu) = f(x + \mu u) - f(x) - \langle \nabla f(x), \mu u \rangle - \frac{\mu^2}{2} u^\mathsf{T} \nabla^2 f(x) u,$$

we have $|a_u(x, \mu)| = |a_u(x, -\mu)| \leq \frac{\mu^3}{6} \rho \|u\|^3$ according to Lemma 4. Then we have

$$\|\nabla f_\mu(x) - \nabla f(x)\|$$

$$= \left\| \frac{1}{\beta(d)} \int_{S_p} \frac{d}{\mu} f(x + \mu u) u \, du - \nabla f(x) \right\|$$

$$= \left\| \frac{1}{\beta(d)} \int_{S_p} \frac{d}{2\mu} \left(f(x + \mu u) - f(x - \mu u) \right) u \, du - \nabla f(x) \right\|$$

$$= \left\| \frac{1}{\beta(d)} \int_{S_p} \frac{d}{2\mu} \left(f(x + \mu u) - f(x - \mu u) \right) u \, du - 2 \frac{1}{\beta(d)} \int_{S_p} \frac{d}{2\mu} \langle \nabla f(x), \mu u \rangle u \, du \right\|$$

$$= \frac{d}{2\beta(d)\mu} \left\| \int_{S_p} \left(f(x + \mu u) - f(x - \mu u) - 2 \langle \nabla f(x), \mu u \rangle \right) u \, du \right\|$$

$$\leq \frac{d}{2\beta(d)\mu} \int_{S_p} \left(|a_u(x, \mu)| + |a_u(-\mu)| \right) \|u\| du \leq \frac{\rho d \mu^2}{6\beta(d)} \int_{S_p} \|u\|^4 du = \frac{\rho d \mu^2}{6}$$

For item (3). we have

$$\mathbb{E}\|\hat{\nabla}_{rand} f(x)\|^2$$

$$= \frac{1}{\beta(d)} \int_{S_p} \frac{d^2}{4\mu^2} |f(x + \mu u) - f(x - \mu u)|^2 \|u\|^2 du$$

$$= \frac{d^2}{\beta(d)4\mu^2} \int_{S_p} |a_u(x, \mu) - a_u(x, -\mu) + 2 \langle \nabla f(x), \mu u \rangle|^2 du$$

$$\leq \frac{d^2}{\beta(d)4\mu^2} \left[\int_{S_p} 2 \left(|a_u(x, \mu)|^2 + |a_u(x, -\mu)|^2 \right) du + 4\mu^2 \int_{S_p} \nabla^\mathsf{T} f(x) u u^\mathsf{T} \nabla f(x) du \right]$$

$$\overset{\text{①}}{=} \frac{d^2}{\beta(d)4\mu^2} \left[\int_{S_p} 2 \left(|a_u(x, \mu)|^2 + |a_u(x, -\mu)|^2 \right) du + 4\mu^2 \frac{\beta(d)}{d} \|\nabla f(x)\|^2 \right]$$

16

$$\leq \frac{d^2}{\beta(d)4\mu^2}\left[\int_{S_p}4\left(\frac{\mu^3}{6}\rho\|u\|^3\right)^2 du + 4\mu^2\frac{\beta(d)}{d}\|\nabla f(x)\|^2\right]$$

$$=d\|\nabla f(x)\|^2 + \frac{\rho^2 d^2\mu^4}{36}$$

where ① is due to $\mathbb{E}(uu^\mathsf{T}) = \frac{1}{\beta(d)}\int_{S_p}uu^\mathsf{T} = \frac{1}{d}\mathbf{I}$ as proved in Lemma 6.3 in [17].

For item (4), we have

$$\mathbb{E}\|\hat{\nabla}_{rand}f(x) - \hat{\nabla}_{rand}f(y)\|^2$$

$$=\mathbb{E}\left\|d\frac{f(x+\mu u)-f(x-\mu u)}{2\mu}u - d\frac{f(y+\mu u)-f(y-\mu u)}{2\mu}u\right\|^2$$

$$=\frac{d^2}{4\mu^2}\mathbb{E}\|[f(x+\mu u)-f(x-\mu u)]u - [f(y+\mu u)-f(y-\mu u)]u\|^2$$

$$=\frac{d^2}{4\mu^2}\mathbb{E}\|(a_u(x,\mu)-a_u(x,-\mu)-(a_u(y,\mu)+a_u(y,-\mu))u + 2\langle\nabla f(x)-\nabla f(y),\mu u\rangle u\|^2$$

$$\leq\frac{d^2}{4\mu^2}\mathbb{E}\left[2\left(|a_u(x,\mu)|^2+|a_u(x,-\mu)|^2+|a_u(y,\mu)|^2+|a_u(y,-\mu)|^2\right)\|u\|^2\right.$$

$$\left.+8\|\langle\nabla f(x)-\nabla f(y),\mu u\rangle u\|^2\right]$$

$$\leq\frac{d^2}{4\mu^2}\mathbb{E}\left[8\frac{\mu^6}{36}\rho^2\|u\|^6\cdot\|u\|^2 + 8\|\langle\nabla f(x)-\nabla f(y),\mu u\rangle u\|^2\right]$$

$$=2d^2\mathbb{E}\|\langle\nabla f(x)-\nabla f(y),u\rangle u\|^2 + \frac{\rho^2 d^2\mu^4}{18}\mathbb{E}\|u\|^8 = 2d^2\mathbb{E}\langle\nabla f(x)-\nabla f(y),u\rangle^2 + \frac{\rho^2 d^2\mu^4}{18}$$

$$=2d^2(\nabla f(x)-\nabla f(y))^\mathsf{T}\mathbb{E}(uu^\mathsf{T})(\nabla f(x)-\nabla f(y)) + \frac{\rho^2 d^2\mu^4}{18}$$

$$\overset{①}{=}2d\|\nabla f(x)-\nabla f(y)\|^2 + \frac{\rho^2 d^2\mu^4}{18}$$

$$\leq 2d\ell^2\|x-y\|^2 + \frac{\rho^2 d^2\mu^4}{18}$$

where ① is due to $\mathbb{E}(uu^\mathsf{T}) = \frac{1}{\beta(d)}\int_{S_p}uu^\mathsf{T} = \frac{1}{d}\mathbf{I}$ as proved in Lemma 6.3 in [17].

$\square$

A.2 ZO Hessian-Vector Product Estimator

Proof of Lemma 1. If f is ρ-Hessian Lipschitz,

$$\left\|\nabla f(x+v)-\nabla f(x)-\nabla^2 f(x)v\right\|\leq\frac{\rho}{2}\|v\|^2 \tag{8}$$

this inequality uses Lemma 4. Then we have

$$\left\|\nabla^2 f(x_0)v - \mathcal{H}_f(x)v\right\|$$

$$=\left\|\nabla^2 f(x_0)v - (\nabla f(x+v)-\nabla f(x)) + (\nabla f(x+v)-\nabla f(x)) - \mathcal{H}_f(x)v\right\|$$

$$\overset{①}{\leq}\left\|\nabla^2 f(x_0)v - (\nabla f(x+v)-\nabla f(x))\right\| + \|(\nabla f(x+v)-\nabla f(x))-\mathcal{H}_f(x)v\|$$

$$=\left\|\nabla^2 f(x_0)v - (\nabla f(x+v)-\nabla f(x))\right\| + \left\|(\nabla f(x+v)-\nabla f(x))-(\hat{\nabla}_{coord}f(x+v)-\hat{\nabla}_{coord}f(x))\right\|$$

$$\overset{②}{\leq}\left\|\nabla^2 f(x_0)v - (\nabla f(x+v)-\nabla f(x))\right\| + \left\|\nabla f(x+v)-\hat{\nabla}_{coord}f(x+v)\right\| + \left\|\nabla f(x)-\hat{\nabla}_{coord}f(x)\right\|$$

$$\overset{③}{\leq}\frac{\rho}{2}\|v\|^2 + \frac{\sqrt{d}\rho\mu^2}{3}$$

where ① and ② are due to the triangle inequality; ③ is due to Eq. (8) and Lemma 7. $\square$

17

B Proof of Proposition 1

Proof. **Online stochastic setting.** Let $S := \{S_1, \ldots, S_m\}$ be $m = \mathcal{O}(\log \frac{1}{p})$ random uniform subsets of $[n]$, each of cardinality $B = \max\{\frac{128\sigma^2}{\epsilon^2}, 1\}$. Denote by $v_j = \frac{1}{B}\sum_{i \in S_j} \nabla f_i(x)$ and $\hat{v}_j = \frac{1}{B}\sum_{i \in S_j} \hat{\nabla}_{coord} f_i(x)$, according to Lemma 5 we have

$$\mathbb{E}_{S_j}\left[\|v_j - \nabla f(x)\|^2\right] \leq \frac{1}{B} \cdot \frac{1}{n}\sum_{i \in [n]} \|v_j - \nabla f(x)\|^2 \leq \frac{\sigma^2}{B} = \frac{\epsilon^2}{128}$$

Then, according to the Chebyshev's inequality: $P(|x - \mathbb{E}(x)| > u) \leq \frac{var(x)}{u^2}$, with probability at least $1/2$ over the randomness of S_j we have

$$\left|\|\hat{v}_j\| - \|\nabla f(x)\|\right| \leq \left|\|\hat{v}_j\| - \|v_j\|\right| + \left|\|v_j\| - \nabla f(x)\|\right|$$

$$\leq \|\hat{v}_j - v_j\| + \|v_j - \nabla f(x)\| \leq \frac{\rho\sqrt{d}\mu^2}{6} + \frac{\epsilon}{8} \leq \frac{\epsilon}{4}$$

where the third inequality comes from Lemma 7 and the Chebyshev's inequality by setting $u = \frac{\epsilon}{8}$; the last inequality is because we choose the smoothing parameter μ such that $\mu \leq \sqrt{\frac{3\epsilon}{4\rho\sqrt{d}}}$. We denote the non-decreasing order

$$\pi \circ S := \{S_{\pi(1)}, S_{\pi(2)}, \ldots, S_{\pi(m)}\}$$
$$\text{s.t.} \quad \left|\|\hat{v}_{\pi(1)}\| - \|\nabla f(x)\|\right| \geq \cdots \geq \left|\|\hat{v}_{\pi(m)}\| - \|\nabla f(x)\|\right|$$

Then we define the event

$$\mathcal{H}_j = \left(\left|\|\hat{v}_{\pi(j)}\| - \|\nabla f(x)\|\right| \geq \frac{\epsilon}{4}\right)$$

We have $\Pr(\mathcal{H}_j) \leq \frac{1}{2}$ for all $j \in [m]$. Using the fact that

$$\mathcal{H}_{\pi(\lfloor \frac{m}{2} \rfloor)} \subseteq \left(\left|\|\hat{v}_{\pi(j)}\| - \|\nabla f(x)\|\right| \geq \frac{\epsilon}{4}, \forall j \leq \left\lfloor \frac{m}{2} \right\rfloor\right) = \bigcap_{j=1}^{\lfloor \frac{m}{2} \rfloor} \mathcal{H}_{\pi(j)}$$

we have

$$\Pr\left(\mathcal{H}_{\pi(\lfloor \frac{m}{2} \rfloor)}\right) \leq \Pr\left(\bigcap_{j=1}^{\lfloor \frac{m}{2} \rfloor} \mathcal{H}_{\pi(j)}\right) = \prod_{j=1}^{\lfloor \frac{m}{2} \rfloor} \Pr\left(\mathcal{H}_{\pi(j)}\right) \leq \left(\frac{1}{2}\right)^{\lfloor \frac{m}{2} \rfloor}$$

If we choose $m = \mathcal{O}(\log(1/p))$, we have with probability at least $1 - p$, it satisfies that at least $\lfloor \frac{m}{2} \rfloor + 1$ of the vectors v_j satisfy $\left|\|\hat{v}_j\| - \|\nabla f(x)\|\right| \leq \frac{\epsilon}{4}$. Then we select $v^* = v_j$ where $j \in [m]$ is index that gives the median value of $\|\hat{v}_j\|$, then it satisfies $\left|\|\hat{v}_j\| - \|\nabla f(x)\|\right| \leq \frac{\epsilon}{4}$. Finally, we can check if $\|\hat{v}_j\| \leq \frac{3\epsilon}{4}$, then $\|\nabla f(x)\| \leq \epsilon$, and if not, then $\|\nabla f(x)\| \geq \frac{\epsilon}{2}$.

Deterministic. The case in the offline deterministic setting is much simpler than the online setting. According to Lemma 7 we have

$$\|\hat{\nabla}_{coord} f(x) - \nabla f(x)\| \leq \frac{\rho\sqrt{d}\mu^2}{6}$$

If we choose $\mu \leq \sqrt{\frac{3\epsilon}{2\rho\sqrt{d}}}$ we get $\|\hat{\nabla}_{coord} f(x) - \nabla f(x)\| \leq \frac{\epsilon}{4}$. Thus, we can check if $\|\hat{\nabla}_{coord} f(x)\|$, then $\|\nabla f(x)\| \leq \epsilon$, and if not, then $\|\nabla f(x)\| \geq \frac{\epsilon}{2}$. $\qquad\square$

C Proof of Zeroth-Order Negative Curvature Search

C.1 Proof of Online setting

Proof of Lemma 2. We first recall the parameter settings in Algorithm 1:

$$\eta \leftarrow \frac{\delta}{C_0^2 \ell^2 \log(100d)}, T \leftarrow \frac{C_0^2 \log(100d)}{\eta\delta}, \sigma \leftarrow (100d)^{-3C_0}\frac{\eta^2\delta^3}{\rho}, r \leftarrow (100d)^{C_0}\sigma \qquad \text{(T1)}$$

18

Denote by $i_t \in [n]$ the random index i chosen at the t-th iteration in Algorhtm 1. Let ξ_t be the error vector such that

$$\xi_t := \nabla^2 f_{i_t}(x_0)(x_t - x_0) - \mathcal{H}_{f_{i_t}}(x_0)(x_t - x_0). \tag{T2}$$

Then the error vector can be bounded by

$$\|\xi_t\| \le \rho\left(\frac{\|x_t - x_0\|^2}{2} + \frac{\sqrt{d}\mu_t^2}{3}\right) \le \rho\sqrt{d}\|x_t - x_0\|^2.$$

where the first inequality is due to By Lemma 1; the second inequality is due to $\mu_t = \|x_t - x_0\|$ and $d \ge 1$. According to the definition of ξ_t in (T2), we have

$$x_{t+1} = x_t - \eta\mathcal{H}_{f_{i_t}}(x_0)(x_t - x_0) = x_t - \eta\nabla^2 f_{i_t}(x_0)(x_t - x_0) + \eta\xi_t. \tag{T3}$$

Then we define the following notations,

$$z_t = x_t - x_0, \quad \mathbf{A}_t = \mathbf{B}_t + \mathbf{R}_t \quad \text{where} \quad \mathbf{B}_t = \nabla^2 f_{i_t}(x_0), \quad \mathbf{R}_t = -\frac{\xi_t z_t^\mathsf{T}}{\|z_t\|^2},$$

From (T3), we have

$$z_{t+1} = z_t - \eta\mathbf{B}_t z_t + \eta\xi_t = (\mathbf{I} - \eta\mathbf{A}_t)z_t.$$

As long as Algorithm 1 does not terminate, we have

$$\|\mathbf{R}_t\| \le \rho\sqrt{d}\|z_t\| \overset{①}{\le} \rho\sqrt{d}r, \quad \|\mathbf{B}_t\| \le \ell, \quad \|\mathbf{A}_t\| \le \|\mathbf{B}_t\| + \|\mathbf{R}_t\| \le \|\mathbf{B}_t\| + \rho\sqrt{d}r \overset{②}{\le} 2\ell$$

where ① holds since we always have $\|z_t\| = \|x_t - x_0\| \le r$ as long as Algorithm 1 does not terminate; ② holds since the parameter setting of r as in Algorithm 1 such that $\rho\sqrt{d}r = \frac{\sqrt{d}\delta^5}{C_0^4(100d)^{2C_0}\log^2(100d)\ell^4} \le \frac{\delta^5}{\ell^4} \le \ell$.

Define

$$\begin{aligned}
\Phi_t &= z_{t+1}z_{t+1}^\mathsf{T} \overset{①}{=} (\mathbf{I} - \eta\mathbf{A}_t)\cdots(\mathbf{I} - \eta\mathbf{A}_1)\xi\xi^\mathsf{T}(\mathbf{I} - \eta\mathbf{A}_1)^\mathsf{T}\cdots(\mathbf{I} - \eta\mathbf{A}_t)^\mathsf{T} \\
&= (\mathbf{I} - \eta\mathbf{A}_t)\Phi_{t-1}(\mathbf{I} - \eta\mathbf{A}_t)^\mathsf{T} \\
&= (\mathbf{I} - \eta\mathbf{A}_t)\Phi_{t-1}(\mathbf{I} - \eta\mathbf{A}_t) \\
w_t &= \frac{z_t}{\|z_t\|} = \frac{z_t}{\sqrt{\mathrm{tr}(\Phi_{t-1})}}
\end{aligned}$$

where ① is because $z_1 = x_1 - x_0 = \xi$. As long as Algorithm 1 does not terminate, we have

$$\begin{aligned}
\mathrm{tr}(\Phi_t) &= \|z_{t+1}\|^2 = \|(\mathbf{I} - \eta\mathbf{A}_t)z_t\|^2 \\
&= \|z_t\|^2 - 2\eta z_t^\mathsf{T}\mathbf{A}_t z_t + \eta^2 z_t^\mathsf{T}\mathbf{A}_t^2 z_t \\
&= \|z_t\|^2\left(1 - 2\eta w_t^\mathsf{T}\mathbf{A}_t w_t + \eta^2 w_t^\mathsf{T}\mathbf{A}_t^2 w_t\right) \\
&= \mathrm{tr}(\Phi_{t-1})\left(1 - 2\eta w_t^\mathsf{T}\mathbf{A}_t w_t + \eta^2 w_t^\mathsf{T}\mathbf{A}_t^2 w_t\right) \\
&\le \mathrm{tr}(\Phi_{t-1})\left(1 - 2\eta w_t^\mathsf{T}\mathbf{A}_t w_t + 4\eta^2\ell^2\right) \\
&\le \mathrm{tr}(\Phi_{t-1})\left(1 - 2\eta w_t^\mathsf{T}\mathbf{B}_t w_t + 2\eta\|\mathbf{R}_t\| + 4\eta^2\ell^2\right) \\
&\le \mathrm{tr}(\Phi_{t-1})\left(1 - 2\eta w_t^\mathsf{T}\mathbf{B}_t w_t + 2\eta\rho\sqrt{d}r + 4\eta^2\ell^2\right) \\
&\le \mathrm{tr}(\Phi_{t-1})\left(1 - 2\eta w_t^\mathsf{T}\mathbf{B}_t w_t + 8\eta^2\ell^2\right)
\end{aligned}$$

where the last inequality is because $\frac{\rho\sqrt{d}r}{\eta} = \frac{\sqrt{d}\delta^4}{(100d)^{2C_0}C_0^2\log(100d)\ell^2} \le \frac{\delta^4}{\ell^2} \le \ell^2$. On the other hand,

$$\begin{aligned}
\mathrm{tr}(\Phi_t) &= \mathrm{tr}(\Phi_{t-1})\left(1 - 2\eta w_t^\mathsf{T}\mathbf{A}_t w_t + \eta^2 w_t^\mathsf{T}\mathbf{A}_t^2 w_t\right) \\
&\ge \mathrm{tr}(\Phi_{t-1})\left(1 - 2\eta w_t^\mathsf{T}\mathbf{A}_t w_t\right) \\
&\ge \mathrm{tr}(\Phi_{t-1})\left(1 - 2\eta w_t^\mathsf{T}\mathbf{B}_t w_t - 2\eta\|\mathbf{R}_t\|\right) \\
&\ge \mathrm{tr}(\Phi_{t-1})\left(1 - 2\eta w_t^\mathsf{T}\mathbf{B}_t w_t - 4\eta\rho dr\right) \\
&\ge \mathrm{tr}(\Phi_{t-1})\left(1 - 2\eta w_t^\mathsf{T}\mathbf{B}_t w_t - 8\eta^2\ell^2\right)
\end{aligned}$$

19

Then take the logarithm on both sides of the inequality, we get

$$\log\left(1 - 2\eta w_t^\mathsf{T} \mathbf{B}_t w_t - 8\eta^2 \ell^2\right) \leq \log\left(\operatorname{tr}(\Phi_t)\right) - \log\left(\operatorname{tr}(\Phi_{t-1})\right) \leq \log\left(1 - 2\eta w_t^\mathsf{T} \mathbf{B}_t w_t + 8\eta^2 \ell^2\right).$$

Define
$$\lambda = -\lambda_{min}(\nabla^2 f(x_0)) = -\lambda_{min}(\mathbb{E}[\mathbf{B}_t]), \quad \mathbf{A} := \nabla^2 f(x_0) = \mathbb{E}[\mathbf{B}_t].$$

We know $w_t^\mathsf{T} \mathbf{B}_t w_t \in [-\ell, \ell]$ and $\mathbb{E}[w_t^\mathsf{T} \mathbf{B}_t w_t] = w_t^\mathsf{T} \mathbf{A} w_t \geq -\lambda$. Then we have

$$\log\left(\operatorname{tr}(\Phi_t)\right) - \log\left(\operatorname{tr}(\Phi_{t-1})\right) \in [\log\left(1 - 2\eta w_t^\mathsf{T} \mathbf{B}_t w_t - 8\eta^2 \ell^2\right), \log\left(1 - 2\eta w_t^\mathsf{T} \mathbf{B}_t w_t + 8\eta^2 \ell^2\right)]$$

$$\overset{①}{\subset} [-2(2\eta\ell + 8\eta^2 \ell^2), 2\eta\ell + 8\eta^2 \ell^2]$$

$$\overset{②}{\subset} [-6\eta\ell, 3\eta\ell]$$

$$\mathbb{E}\left[\log\left(1 - 2\eta w_t^\mathsf{T} \mathbf{B}_t w_t + 8\eta^2 \ell^2\right)\right] \overset{③}{\leq} \log\left(\mathbb{E}\left[\log\left(1 - 2\eta w_t^\mathsf{T} \mathbf{B}_t w_t + 8\eta^2 \ell^2\right)\right]\right) \leq 2\eta\lambda + 8\eta^2 \ell^2$$

above ① uses the fact that $|\log(1-x)| \leq 2|x|, \forall x \in [-1/2, 1/2]$ and $\log(1+x) < x, \forall x > 0$; ② is because $\eta = \frac{\delta}{C_0^2 \ell^2 \log(100d)} \leq \frac{1}{8\ell}$; ③ is due to the concavity of log and Jensen's inequality. Applying Lemma 9 by setting $\rho = 3\eta\ell, \beta = 2\eta\lambda + 8\eta^2 \ell^2$, we have

$$\Pr\left[\log(\operatorname{tr}(\Phi_t)) - \log(\operatorname{tr}(\Phi_0)) \geq (2\eta\lambda + 8\eta^2 \ell^2)t + 6\eta\ell\sqrt{t\log(\frac{t}{p})}\right] \leq p/t.$$

then we define

$$T_0 = \min\left\{\frac{1}{2\eta\lambda} \cdot \frac{\log(\frac{r^2}{\sigma^2})}{3}, \frac{1}{8\eta^2 \ell^2} \cdot \frac{\log(\frac{r^2}{\sigma^2})}{3}, T_0'\right\}$$

where T_0' is the largest positive integer such that $6\eta\ell\sqrt{T_0'\log(\frac{T_0'}{p})} \leq \frac{\log(\frac{r^2}{\sigma^2})}{3}$. Thus, with the choice of T_0 we have

$$(2\eta\lambda + 8\eta^2 \ell^2)T_0 + 6\eta\ell\sqrt{T_0\log(\frac{T_0}{p})} \leq \frac{\log(\frac{r^2}{\sigma^2})}{3} + \frac{\log(\frac{r^2}{\sigma^2})}{3} + \frac{\log(\frac{r^2}{\sigma^2})}{3} = \log(\frac{r^2}{\sigma^2})$$

Therefore, for $t \leq T_0$, we have

$$\Pr\left[\log(\operatorname{tr}(\Phi_t)) - \log(\operatorname{tr}(\Phi_0)) \geq \log(\frac{r^2}{\sigma^2})\right] \leq \frac{p}{T_0}$$

Taking a union bound over $t \in [T_0]$, we have

$$\Pr\left[\bigcup_{t=1}^{T_0} \log(\operatorname{tr}(\Phi_t)) - \log(\operatorname{tr}(\Phi_0)) \geq \log(\frac{r^2}{\sigma^2})\right] \leq \sum_{t=1}^{T_0} \Pr\left[\log(\operatorname{tr}(\Phi_t)) - \log(\operatorname{tr}(\Phi_0)) \geq \log(\frac{r^2}{\sigma^2})\right] \leq p$$

which is equivalent to

$$\Pr\left[\bigcap_{t=1}^{T_0} \log(\operatorname{tr}(\Phi_t)) - \log(\operatorname{tr}(\Phi_0)) \geq \log(\frac{r^2}{\sigma^2})\right] \geq 1 - p$$

By definition, $\operatorname{tr}(\Phi_t) = \|x_t - x_0\|^2$ and $\operatorname{tr}(\Phi_0) = \|\xi\|^2 = \sigma^2$. Then we know with probability at least $1 - p$, for every $t \in [T_0], \|x_{t+1} - x_t\| < r$. Thus Algorithm 1 will not terminate before iteration T_0.

> Then we prove that when $\lambda > \delta$, Algorithm 1 outputs a vector v, with probability at least $\frac{2}{3}$, $v^\mathsf{T} \mathbf{A} v \leq -\frac{3}{4}\delta$.

We first note when $\lambda \geq \delta$, using our choice of η and T_0, we have $T_0 \leq \frac{\log\left(\frac{r^2}{\sigma^2}\right)}{6\eta\lambda}$. Denote by $v_{t+1} \overset{\text{def}}{=} (\mathbf{I} - \eta \mathbf{B}_t) \cdots (\mathbf{I} - \eta \mathbf{B}_1)\xi$ and $u_t \overset{\text{def}}{=} z_t - v_t$ with $u_1 = 0$, we have

$$u_{t+1} = z_{t+1} - v_{t+1} = \prod_{s=1}^{t} (\mathbf{I} - \eta \mathbf{A}_s)\xi - \prod_{s=1}^{t} (\mathbf{I} - \eta \mathbf{B}_s)\xi$$

20

$$= \left(\mathbf{I} - \eta\mathbf{B}_t\right)\left(z_t - v_t\right) - \eta\mathbf{R}_t z_t$$
$$= \left(\mathbf{I} - \eta\mathbf{B}_t\right) u_t - \eta\mathbf{R}_t z_t$$

then, before Algorithm 1 stops, we have:

$$\|u_{t+1} - \left(\mathbf{I} - \eta\mathbf{B}_t\right) u_t\| = \eta \|\mathbf{R}_t z_t\| \overset{①}{\leq} 2d\eta\rho r^2 \tag{T4}$$

above, ① is due to $\|\mathbf{R}_t z_t\| \leq \|\mathbf{R}_t\| \cdot \|z_t\| \leq 2\rho d\|z_t\|^2$. Using Young's inequality $\|a + b\|_2^2 \leq (1+\beta)\|a\|_2^2 + \left(1 + \frac{1}{\beta}\right)\|b\|_2^2$ for every $\beta > 0$, we have:

$$\|u_{t+1}\|^2 \leq (1 + \eta^2\ell^2)\|\left(\mathbf{I} - \eta\mathbf{B}_t\right) u_t\|^2 + (1 + \frac{1}{\eta^2\ell^2})\|u_{t+1} - \left(\mathbf{I} - \eta\mathbf{B}_t\right) u_t\|^2$$

$$\overset{①}{\leq} (1 + \eta^2\ell^2)\|\left(\mathbf{I} - \eta\mathbf{B}_t\right) u_t\|^2 + (1 + \frac{1}{\eta^2\ell^2})4\eta^2 d^2 \rho^2 r^4$$

$$= (1 + \eta^2\ell^2)\|\left(\mathbf{I} - \eta\mathbf{B}_t\right) u_t\|^2 + \frac{1 + \ell^2\eta^2}{\ell^2}4d^2\rho^2 r^4$$

$$\overset{②}{\leq} \left(1 + \eta^2\ell^2\right)\|(\mathbf{I} - \eta\mathbf{B}_t)u_t\|^2 + 8d^2\frac{\rho^2}{\ell^2}r^4$$

$$= \left(1 + \eta^2\ell^2\right)\|u_t\|^2 \left(1 - 2\eta\frac{u_t^\top\mathbf{B}_t u_t}{\|u_t\|^2} + \eta^2\frac{u_t^\top\mathbf{B}_t^2 u_t}{\|u_t\|^2}\right) + 8d^2\left(\frac{\rho}{\ell}\right)^2 r^4$$

$$\overset{②}{\leq} \|u_t\|^2 \left(1 - 2\eta\frac{u_t^\top\mathbf{B}_t u_t}{\|u_t\|^2} + 10\eta^2\ell^2\right) + 8d^2\left(\frac{\rho}{\ell}\right)^2 r^4.$$

above, ① holds due to (T4); ② is because $\eta\ell \leq 1$; ③ is because $\|\mathbf{B}_t\| \leq \ell$. Apply Lemma 10 by setting $\alpha = 2\eta\ell + 10\eta^2\ell^2 \leq 2.5\eta\ell$ and $\beta = 2\eta\lambda$, we have

$$\Pr\left[\|u_t\| \geq 16d\frac{\rho}{\ell}r^2 t e^{\eta\lambda t + 8\eta\ell\sqrt{t\log\frac{t}{p}}}\right] \leq p \tag{9}$$

Apply Lemma 11 we conclude that for each $t \in [T]$, w.p. at least $99/100$:

- Norm growth: $\|v_t\| \geq \frac{1}{C}\left(e^{(\eta\lambda - 32\eta^2\ell^2)t}\sigma/\sqrt{d}\right)$.

- Negative curvature: $\frac{-v_{t+1}^\top\mathbf{A}v_{t+1}}{\|v_{t+1}\|^2} \leq -(1 - C\eta\ell)\lambda + C\left(\frac{\log d}{\eta t} + \sqrt{\frac{\ell^2}{t}} + \lambda\eta^2\ell^2 t\right)$.

Define

$$T_1 = \frac{\log\frac{2C\sqrt{d}r}{\sigma}}{\eta\lambda - 32\eta^2\ell^2} = \frac{C_0(\log d/p) + \log(2C\sqrt{d})}{\eta\lambda - 32\eta^2\ell^2} \leq \frac{2C_0 \cdot \log d/p}{\eta\lambda - 32\eta^2\ell^2} \leq \frac{4C_0 \cdot \log d/p}{\eta\lambda} \leq \frac{4C_0 \cdot \log d/p}{\eta\delta} < T \tag{10}$$

where the second equality is because $r = (d/p)^{C_0}\sigma$. When $t = T_1$, by "norm growth" property, we know that w.p. at least $99/100$:

$$\|v_{T_1}\| \geq \frac{1}{C}\left(e^{(\eta\lambda - 32\eta^2\ell^2)T_1}\sigma/\sqrt{d}\right) \overset{(10)}{=} 2r \tag{11}$$

Combing with (9), we have w.p. at least $98/100$,

$$\frac{\|u_{T_1}\|}{\|v_{T_1}\|} \leq \frac{16d\rho r^2 T_1 e^{\eta\lambda T_1 + 8\eta\ell\sqrt{T_1\log\frac{T_1}{p}}}}{\ell \cdot 2r}$$

$$\leq 16C\left(\frac{d^{\frac{3}{2}}\rho r^2 T_1 e^{\eta\lambda T_1 + 8\eta\ell\sqrt{T_1\log\frac{T_1}{p}}}}{\ell e^{(\eta\lambda - 32\eta^2\ell^2)T_1}\sigma}\right)$$

$$\leq 16C\left(\frac{d^{\frac{3}{2}}\rho r^2 T_1}{\ell\sigma} e^{8\eta\ell\sqrt{T_1\log\frac{T_1}{p}} + 32\eta^2\ell^2 T_1}\right)$$

21

$$\overset{①}{\leq} 16C\left(\frac{d^{\frac{3}{2}}\rho r^2 T_1}{\ell\sigma}e^{16\sqrt{\log\frac{T_1}{p}}}\right)$$

$$\overset{②}{\leq} 16C\left(\frac{r^2}{\sigma}\cdot\frac{d^{\frac{3}{2}}\rho T_1^2}{\ell p}\right)\overset{③}{\leq}\frac{\delta}{100\ell}\leq\frac{1}{100} \tag{12}$$

where inequality ① is because $\eta^2\ell^2 T_1\leq\eta^2\ell^2 T=1$; Inequality ② is because $e^{16\sqrt{\log(x)}}<x$ for sufficiently large x; Inequality ③ is because

$$\frac{r^2}{\sigma}=(\frac{d}{p})^{2C_0}\sigma=(\frac{d}{p})^{-C_0}\frac{\eta^2\delta^3}{\rho}$$

and thus, for sufficiently large C_0 and $p=1/100$, we have

$$(\frac{d}{p})^{-C_0}\leq\frac{p}{d^{3/2}}\cdot\frac{1}{C_0(4C_0\log(\frac{d}{p}))^2}$$

This implies

$$\frac{r^2}{\sigma}\leq\frac{\eta^2\delta^3 p}{C_0 d^{3/2}\rho(4C_0\log(\frac{d}{p}))^2}=\frac{\delta p}{C_0 d^{3/2}\rho(\frac{4C_0\log(\frac{d}{p})}{\eta\delta})^2}\overset{(10)}{\leq}\frac{\delta p}{C_0 d^{3/2}\rho T_1^2}.$$

Thus for sufficiently large $C_0\geq 1600C$ ③ holds.

Putting together with $\|v_{T_1}\|\geq 2r$, we have w.p. at least $97/100$, $\|z_{T_1}\|=\|u_{T_1}+v_{T_1}\|\geq r$. This means that Algorithm 1 will terminate within $T_1\leq T$ iterations.

Since w.p. $\geq 99/100$, Algorithm 1 will not terminate before $T_0\geq\frac{\log(r^2/\sigma^2)}{6\eta\lambda}$. Thus w.p. at least $96/100$, Algorithm 1 will terminate at $t\in[T_0,T_1]$.

Using the "negative curvature" property, we have w.p. at least $\geq 99/100$,

$$\frac{v_t^\top\mathbf{A}v_t}{\|v_t\|^2}\leq-(1-C\eta\ell)\lambda+C\left(\frac{\log d}{\eta T_0}+\sqrt{\frac{\ell^2}{T_0}}+\lambda\eta^2\ell^2 T_1\right).$$

Since $T_0\geq\frac{\log(\frac{r}{\sigma})^2}{6\eta\lambda}=\frac{C_0\log(d/p)}{3\eta\lambda}\geq\frac{C_0\log d}{3\eta\lambda}$, thus,

$$\frac{\log d}{\eta T_0}\leq\frac{3\lambda}{C_0}$$

By the choice of η we have $\eta\ell^2\leq\frac{\delta}{C_0^2}\leq\frac{\lambda}{C_0^2}$ and

$$\frac{\ell^2}{T_0}\leq\frac{3\eta\lambda\ell^2}{C_0}\leq\frac{\lambda^2}{C_0},$$

$$\lambda\eta^2\ell^2 T_1\leq\lambda\eta^2\ell^2\frac{4C_0\log(d/p)}{\eta\lambda}\leq\lambda\frac{4C_0\log d}{C_0^2\log d}\leq\frac{4\lambda}{C_0}$$

Then we have w.p. at least $95/100$, Algorithm 1 terminates at $t\in[T_0,T_1]$ and

$$\frac{v_t^\top\mathbf{A}v_t}{\|v_t\|^2}\leq-\frac{15}{16}\lambda\leq-\frac{15}{16}\delta.$$

Since $\|u_t+v_t\|=\|z_t\|\geq r$, we have w.p. at least $99/100$,

$$\frac{\|u_t\|}{\|u_t\|+\|v_t\|}\leq\frac{\|u_t\|}{\|u_t+v_t\|}\leq\frac{16d\rho r^2 T_1 e^{\eta\lambda T_1+8\eta\ell\sqrt{T_1\log\frac{T_1}{p}}}}{\ell r}\overset{(12)}{\leq}\frac{\delta}{50\ell}.$$

This implies $\frac{\|u_t\|}{\|v_t\|}\leq\frac{\delta}{49\ell}$. In sum, we have w.p. at least $94/100$:

$$\frac{z_t^\top\mathbf{A}z_t}{\|z_t\|^2}=\frac{\|v_t\|^2}{\|z_t\|^2}\cdot\frac{z_t^\top\mathbf{A}z_t}{\|v_t\|^2}=\frac{\|v_t\|^2}{\|z_t\|^2}\cdot\frac{(v_t+u_t)^\top\mathbf{A}(v_t+u_t)}{\|v_t\|^2}$$

22

$$\leq \frac{\|v_t\|^2}{\|z_t\|^2} \cdot \frac{v_t^\top \mathbf{A} v_t + 4\ell \|u_t\| \|v_t\|}{\|v_t\|^2}$$

$$\leq \frac{\|v_t\|^2}{\|z_t\|^2} \cdot \left(\frac{v_t^\top \mathbf{A} v_t}{\|v_t\|^2} + \frac{4\ell \|u_t\|}{\|v_t\|} \right) \leq \frac{\|v_t\|^2}{\|z_t\|^2} \cdot \left(-\frac{15}{16}\delta + \frac{4}{49}\delta \right) \leq -\frac{17\delta}{20} \frac{\|v_t\|^2}{\|z_t\|^2}$$

$$\leq -\frac{17\delta}{20} \left(1 - \frac{\|u_t\|^2}{\|z_t\|^2} \right) \leq -\frac{17\delta}{20} \cdot \frac{49}{50} < -\frac{3}{4}\delta$$

$\square$

Proof of Lemma 3. By Lemma 1 with $\mu = \|v\|$ we have

$$\left\| v^\top (\mathcal{H}_{f_i}(x) - \nabla^2 f_i(x))v \right\| \leq \rho\sqrt{d}\|v\|^3.$$

Define

$$z_j = v^\top \mathcal{H}_{f_{i_j}} v,$$

then $z_1, \dots, z_m$ are i.i.d. random variables with

$$|z_j| \leq \left\| v^\top (\mathcal{H}_{f_{i_j}}(x) - \nabla^2 f_i(x))v \right\| + \left\| v^\top \nabla^2 f_{i_j}(x)v \right\|$$

$$\leq \rho\sqrt{d}\|v\|^3 + \ell\|v\|^2.$$

By Chernoff inequality, we have

$$\Pr\left[|z - \mathbb{E}[z]| \geq 2(\rho\sqrt{d}\|v\|^3 + \ell\|v\|^2)\sqrt{\frac{1}{m}\log\frac{1}{p}} \right] \leq p.$$

Since $|\mathbb{E}[z] - v^\top \nabla^2 f(x)v| \leq \rho\sqrt{d}\|v\|^3$, we have

$$\Pr\left[\left| \frac{z}{\|v\|^2} - \frac{v^\top \nabla^2 f(x)v}{\|v\|^2} \right| \leq 2(\rho\sqrt{d}\|v\| + \ell)\sqrt{\frac{1}{m}\log\frac{1}{p}} + \rho\sqrt{d}\|v\| \right] \geq 1 - p.$$

$\square$

Lemma 9. *Consider the random variables $\{x_t\}_{t=0}^T$ with respect to random events $\{\mathcal{F}\}_{t=0}^T$ and $\log(1-a_0), \log(1-a_1), \dots, \log(1-a_T) \in [-2\rho, \rho]$ where each x_t and a_t only depend on $\mathcal{F}_1, \dots, \mathcal{F}_t$.*

$$\log x_t = \log x_{t-1} + \log(1 - a_t) \quad and \quad \mathbb{E}\left[\log(1 - a_t) \mid \mathcal{F}_1, \dots, \mathcal{F}_{t-1}\right] \leq \beta.$$

Then we have for every $p \in (0, 1)$,

$$\Pr\left[\log x_t - \log x_0 \geq \beta T + 2\rho\sqrt{T\log\frac{T}{p}} \right] \leq p/T$$

Proof. Applying a general form of Azuma-Hoeffding inequality [48], we have

$$\Pr\left[\log x_t - \log x_0 \geq \epsilon_1\right] \leq \exp\left\{-\frac{2\epsilon_1^2}{((\beta + 2\rho)^2)T}\right\}$$

Let $\exp\left\{-\frac{2\epsilon_1^2}{(\mu + 2\rho)^2 T}\right\} = p/T$, we get

$$\epsilon_1 = (\beta + 2\rho)\sqrt{\frac{1}{2}T\log\frac{T}{p}}$$

$$\leq \beta T + 2\rho\sqrt{T\log\frac{T}{p}} := \epsilon$$

So we have

$$\Pr[\log x_t - \log x_0 \geq \epsilon] \leq \Pr[\log x_t - \log x_0 \geq \epsilon_1] \leq p/T$$

$\square$

23

Lemma 10 ([4]). *Consider random events $\{\mathcal{F}_t\}_{t\geq 1}$ and random variables $x_1, \ldots, x_T \geq 0$ and $a_1, \ldots, a_T \in [-\alpha, \alpha]$ for $\alpha \in [0, 1/2]$ where each x_t and a_t only depend on $\mathcal{F}_1, \ldots, \mathcal{F}_t$. Letting $x_0 = 0$ and suppose there exist constant $b \geq 0$ and $\beta > 0$ such that for every $t \geq 1$:*

$$x_t \leq x_{t-1}(1 - a_t) + b \quad \text{and} \quad \mathbb{E}[a_t \mid \mathcal{F}_1, \ldots, \mathcal{F}_{t-1}] \geq -\beta.$$

Then, we have for every $p \in (0, 1)$: $\Pr\left[x_T \geq T \cdot b \cdot e^{\left(\beta T + 2\alpha\sqrt{T \log \frac{T}{p}}\right)}\right] \leq p$.

Lemma 11 ([4]). *There exists an absolute constant $C > 0$ such that the following holds: Suppose $\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_t$ are i.i.d. random matrices with $\|\mathbf{B}_i\| \leq \ell$ and $\mathbb{E}[\mathbf{B}_i] = -\mathbf{A}$. Suppose also $\lambda_{\max}(\mathbf{A}) = \lambda \geq 0$. Let*

$$\forall i = 0, 1, \ldots, t: \quad v_{i+1} \stackrel{def}{=} (\mathbf{I} - \eta\mathbf{B}_i) \cdots (\mathbf{I} - \eta\mathbf{B}_1)\xi$$

where ξ is a random Gaussian vector with norm σ, and $\eta \in \left(0, \sqrt{\frac{1}{1350000t\ell^2}}\right]$ is the learning rate. Then, with probability at least $99/100$:

1. Norm growth: $\|v_{t+1}\|_2 \geq \frac{1}{C}\left(e^{\left(\eta\lambda - 32\eta^2\ell^2\right)t}\sigma/\sqrt{d}\right)$.

2. Negative curvature: $-\frac{v_{t+1}^\top \mathbf{A} v_{t+1}}{\|v_{t+1}\|_2^2} \leq -(1 - C\eta\ell)\lambda + C\left(\frac{\log d}{\eta t} + \sqrt{\frac{\ell^2}{t}} + \lambda\eta^2\ell^2 t\right)$.

C.2 Proof of Deterministic setting

Proof of Theorem 2. For notation simplicity, we denote

$$\mathbf{A} = \nabla^2 f(x_0), \quad \mathbf{M} = -\frac{1}{\ell}\nabla^2 f(x_0) + (1 - \frac{3\delta}{4\ell})\mathbf{I}, \quad \lambda = -\lambda_{min}(\mathbf{A}).$$

Then, we know that all the eigenvalues of $\mathbf{M}$ lie in $[-1, 1 + \frac{\lambda - 3\delta/4}{\ell}]$. Define

$$\mathcal{M}(y) = \left(-\frac{1}{\ell}\mathcal{H}_f(x_0) + (1 - \frac{3\delta}{4\ell})\right)y$$

and use it to approximate $\mathbf{M}y$. Recall that

$$y_0 = 0, \quad y_1 = \xi, \quad y_t = 2\mathcal{M}(y_{t-1}) - y_{t-2}.$$

If we set $x_{t+1} = x_0 + y_{t+1} - \mathcal{M}(y_t)$, then it satisfies $x_{t+1} - x_0 \approx \mathcal{T}_t(\mathbf{M})\xi$ according to the Definition 5.

Denote by

$$x_{t+1}^* \triangleq x_0 + \mathcal{T}_t(\mathbf{M})\xi$$

the exact solution. We have

$$y_t = 2\mathcal{M}(y_{t-1}) - y_{t-2} = 2(x_0 - x_t + y_t) - y_{t-2} \implies y_t - y_{t-2} = 2(x_t - x_0).$$

Since $\|x_t - x_0\| \leq r$ for each t before Algorithm 3 terminates, we have

$$
\begin{aligned}
\|y_t\| &\leq \|y_{t-2}\| + \|y_t - y_{t-2}\| \leq \|y_{t-2}\| + 2r \\
&\leq \|y_{t-4}\| + 4r \\
&\leq \cdots \\
&\leq 2tr
\end{aligned}
$$

From Lemma 4, we have

$$\|\mathcal{M}(y_t) - \mathbf{M}y_t\| = \left\| -\frac{\mathcal{H}_f(x_0)}{\ell}y_t + \frac{\nabla^2 f(x_0)}{\ell}y_t \right\| \leq \frac{1}{\ell}\rho\left(\frac{\|y_t\|^2}{2} + \frac{\sqrt{d}\mu_t^2}{3}\right) \leq \frac{\rho\sqrt{d}}{\ell}\|y_t\|^2 \leq \frac{2\rho\sqrt{d}rt}{\ell}\|y_t\|.$$

Recall from Definition 4 that

$$\mathcal{T}_t(x) \in \begin{cases} [-1, 1] & \text{if } x \in [-1, 1] \\ \left[\frac{1}{2}\left(x - \sqrt{x^2 - 1}\right)^t, \left(x + \sqrt{x^2 - 1}\right)^t\right] & \text{if } x > 1 \end{cases}$$

24

On the other hand, we have for every $x > 1, a = x + \sqrt{x^2 - 1}, b = x - \sqrt{x^2 - 1}$, it satisfies

$$\mathcal{U}_t(x) = \frac{1}{a-b}(a^{t+1} - b^{t+1}) = \sum_{i=0}^{t} a^i b^{t-i} \leq (t+1)a^t$$

Then we apply Lemma 12 with eigenvalues of $\mathbf{M}$ in $[a,b] = \left[0, 1 + \frac{\lambda - 3\delta/4}{\ell}\right]$ and

$$\gamma \triangleq \max\left\{1 + \frac{\lambda - 3\delta/4}{\ell} + \sqrt{(1 + \frac{\lambda - 3\delta/4}{\ell})^2 - 1}, 1\right\}, \quad C_c = \gamma^t\|\xi\| = \gamma^t\sigma, \quad C_T = 2,$$

$$C_U = t + 2, \quad \epsilon = \frac{2\rho\sqrt{d}rt}{\ell}.$$

Then according to Lemma 12, we have

$$\|x_{t+1}^* - x_{t+1}\| \leq \frac{40\sqrt{d}\rho rt^4\gamma^t\sigma}{\ell}$$

Then we prove that if $\lambda_{min}(\nabla^2 f(x_0)) \leq -\delta$, then with probability at least $1 - p$, it satisfies that if $v \neq \bot, \|v\| = 1$, and $v^{\mathsf{T}}\nabla^2 f(x_0)v \leq -\frac{1}{2}\delta$. In other words, we can assume that $\lambda \geq \delta$.

$\lambda \geq \delta$ implies $\gamma > 1$, so we can let

$$T_1 \triangleq \frac{\log\frac{4dr}{p\sigma}}{\log\gamma} \leq T.$$

By Definition 4 we know that $\|\mathcal{T}_{T_1}(\mathbf{M})\| \geq \frac{1}{2}\gamma^{T_1} = \frac{2dr}{p\sigma}$. Thus, with probability at least $1 - p$, $\|x_{T_1+1}^* - x_0\| = \|\mathcal{T}_{T_1}(\mathbf{M})\xi\| \geq 2r$. Morever, at iteration T_1, we have

$$\|x_{T_1+1}^* - x_{T_1+1}\| \leq \frac{80d\rho rT_1^4\gamma^{T_1}\sigma}{\ell} \leq \frac{80d\rho rT_1^4\sigma}{\ell} \cdot \frac{4dr}{p\sigma} \leq \frac{512d^2\rho T_1^4}{\ell} \cdot \frac{r^2}{p} \overset{①}{\leq} \frac{\delta}{100\ell}r \leq \frac{1}{16}r,$$

where ① is because $r \leq \frac{\delta p}{51200d^2\rho T_1^4}$. This means $\|x_{T_1+1} - x_0\| \geq r$ so the algorithm must terminate before iteration $T_1 \leq T$.

On the other hand, since $\|\mathcal{T}_t(\mathbf{M})\| \leq \gamma^t$, we know that the algorithm will not terminate until $t \geq T_0$:

$$T_0 \triangleq \frac{\log\frac{r}{2\sigma}}{\log\gamma}$$

At the time of $t \geq T_0$ of termination, define $\gamma' = 1 + \frac{\lambda - 3\delta/4}{\ell}$, by Definition 4 we have

- $\mathcal{T}_t(\gamma') \geq \frac{1}{2}\gamma^t \geq \frac{1}{2}\gamma^{T_0} \geq \frac{r}{4\sigma} = (\frac{d}{p})^{\Theta(1)}$.

- $\forall x \in [-1, 1], \mathcal{T}_t(x) \in [-1, 1]$.

Since all the eigenvalues of $\mathbf{A}$ that are $\geq -3\delta/4$ are mapped to the eigenvalues of $\mathbf{M}$ that are in $[-1, 1]$, and the smallest eigenvalue of $\mathbf{A}$ is mapped to the largest eigenvalue γ' of $\mathbf{M}$. So we have, with probability at least $1 - p$, letting $v_t \triangleq x_{t+1}^* - x_0 = \mathcal{T}_t(\mathbf{M})\xi$, then it satisfies

$$\frac{v_t^{\mathsf{T}}\mathbf{A}v_t}{\|v_t\|^2} \leq -\frac{3}{4}\delta \leq -\frac{5}{8}\delta$$

Therefore, denoting by $z_t \triangleq x_{t+1} - x_0$, we have

$$\frac{\|z_t - v_t\|}{\|z_t - v_t\| + \|v_t\|} \leq \frac{\|z_t - v_t\|}{\|z_t\|} \leq \frac{\|x_{t+1}^* - x_{t+1}\|}{r} \leq \frac{\delta}{100\ell} \leq \frac{1}{16},$$

Finally, we have

$$\frac{z_t^{\mathsf{T}}\mathbf{A}z_t}{\|z_t\|^2} = \frac{\|v_t\|^2}{\|z_t\|^2} \cdot \frac{z_t^{\mathsf{T}}\mathbf{A}z_t}{\|v_t\|^2} \leq \frac{\|v_t\|^2}{\|z_t\|^2} \cdot \frac{v_t^{\mathsf{T}}\mathbf{A}v_t + 4\ell\|z_t - v_t\|\|v_t\|}{\|v_t\|^2}$$

25

$$\leq \frac{\|v_t\|^2}{\|z_t\|^2} \cdot \left(\frac{v_t^\top \mathbf{A} v_t}{\|v_t\|^2} + \frac{4\ell \|z_t - v_t\|}{\|v_t\|} \right)$$

$$\leq \frac{\|v_t\|^2}{\|z_t\|^2} \left(-\frac{5}{8}\delta + \frac{1}{25}\delta \right) \leq (1 - \frac{\|z_t - v_t\|}{\|z_t\|}) \left(-\frac{5}{8}\delta + \frac{1}{25}\delta \right)$$

$$\leq \frac{15}{16} \left(-\frac{5}{8}\delta + \frac{1}{25}\delta \right) \leq -\frac{1}{2}\delta.$$

$\square$

Definition 4. *Let $\mathcal{T}_t(x)$ be the t-th Chebyshev polynomial of the first kind and $\mathcal{U}_t(x)$ be the t-th Chebyshev polynomial of the second kind, defined as:*

$$\mathcal{T}_0(x) := 1, \quad \mathcal{T}_1(x) := x, \quad \mathcal{T}_{n+1}(x) := 2x \cdot \mathcal{T}_n(x) - \mathcal{T}_{n-1}(x)$$
$$\mathcal{U}_0(x) := 1, \quad \mathcal{U}_1(x) := 2x, \quad \mathcal{U}_{n+1}(x) := 2x \cdot \mathcal{U}_n(x) - \mathcal{U}_{n-1}(x)$$

then $\mathcal{U}_n(x)$ satisfies: $\frac{d}{dx}\mathcal{T}_n(x) = n\mathcal{U}_{n-1}(x)$ and:

$$\mathcal{T}_n(x) = \begin{cases} \cos(n\arccos(x)) \in [-1,1] & \text{if } x \in [-1,1] \\ \frac{1}{2}\left[\left(x - \sqrt{x^2-1}\right)^n + \left(x + \sqrt{x^2-1}\right)^n \right] & \text{if } x > 1 \end{cases}$$

$$\mathcal{U}_n(x) = \begin{cases} \in [-t,t] & \text{if } x \in [-1,1] \\ \frac{1}{2\sqrt{x^2-1}}\left[\left(x + \sqrt{x^2-1}\right)^{n+1} - \left(x - \sqrt{x^2-1}\right)^{n+1} \right] & \text{if } x > 1 \end{cases}$$

Definition 5 (Inexact backward recurrence, [4])**.** *Suppose we want to compute*

$$\vec{s}_N \triangleq \sum_{k=0}^{N} \mathcal{T}_k(\mathbf{M})\vec{c}_k, \quad \text{where} \quad \mathbf{M} \in \mathbb{R}^{d \times d} \quad \text{is symmetric and each} \quad \vec{c}_k \in \mathbb{R}^d.$$

Let $\mathcal{M}$ be an approximate algorithm that satisfies $\|\mathcal{M}(u) - \mathbf{M}u\| \leq \epsilon\|u\|$ for every $u \in \mathbb{R}^d$. Then, define inexact backward recurrence to be

$$\hat{b}_{N+1} = 0, \hat{b}_N = \vec{c}_N, \quad \text{and} \quad \forall r \in \{N-1, \dots, 0\} : \hat{b}_r \triangleq 2\mathcal{M}(\hat{b}_{r+1}) - \hat{b}_{r+2} + \vec{c}_r \in \mathbb{R}^d,$$

and define the output as $\hat{s}_N \triangleq \hat{b}_0 - \mathcal{M}(\hat{b}_1)$. If $\epsilon = 0$, then $\hat{s}_N = \vec{s}_N$.

Lemma 12 (Stable computation of Chebyshev Polynomials, [4])**.** *For every $N \in \mathbb{N}^*$, suppose the eigenvalues of $\mathbf{M}$ are in $[a, b]$ and suppose there are parameters $C_U \geq 1, C_T \geq 1, \gamma \geq 1, C_c \geq 0$ satisfying*

$$\forall k \in \{0, 1, \dots, N\} : \left\{ \gamma^k\|\vec{c}_k\| \leq C_c \quad \text{and} \quad \forall x \in [a,b] : |\mathcal{T}_k(x)| \leq C_T\gamma^k, |\mathcal{U}_k(x)| \leq C_U\gamma^k \right\}.$$

Then, if $\epsilon \leq \frac{1}{4NC_U}$, we have

$$\|\hat{s}_N - \vec{s}_N\| \leq \epsilon \cdot 2(1 + 2NC_T)NC_UC_c.$$

D Proof of Results of ZO-GD and ZO-SGD

D.1 Proof of Theorem 3 (Option I)

If we update $x_{t+1} = x_t - \frac{\eta}{|S|}\sum_{i \in S}\hat{\nabla}_{coord}f_i(x_t)$, then according to the smoothness of $f(\cdot)$ we have

$$f(x_t) - \mathbb{E}_S[f(x_{t+1})] \geq \mathbb{E}_S\left[\langle \nabla f(x_t), x_t - x_{t+1}\rangle - \frac{\ell}{2}\|x_t - x_{t+1}\|^2 \right]$$

$$= \eta \left\langle \nabla f(x_t), \hat{\nabla}_{coord}f(x)\right\rangle - \frac{\eta^2\ell}{2}\mathbb{E}_S\left[\left\| \frac{1}{|S|}\sum_{i \in S}\hat{\nabla}_{coord}f_i(x_t)\right\|^2 \right]$$

$$\overset{\text{①}}{\geq} \frac{\eta}{2}(\|\nabla f(x_t)\|^2 - \|\nabla f(x_t) - \hat{\nabla}_{coord}f(x_t)\|^2) - \frac{\eta^2\ell}{2}\mathbb{E}_S\left[\left\| \frac{1}{|S|}\sum_{i \in S}\hat{\nabla}_{coord}f_i(x_t)\right\|^2 \right]$$

26

$$\overset{②}{\geq} \frac{\eta}{2}(\|\nabla f(x_t)\|^2 - \frac{\rho^2 d\mu^4}{36}) - \eta^2 \ell \left(\|\nabla f(x)\|^2 + \mathbb{E}_S \left[\left\| \nabla f(x) - \frac{1}{|S|} \sum_{i \in S} \hat{\nabla}_{coord} f_i(x_t) \right\|^2 \right] \right)$$

$$= \frac{\eta}{2}((1 - 2\eta\ell)\|\nabla f(x_t)\|^2 - \frac{\rho^2 d\mu^4}{36}) - \eta^2 \ell \mathbb{E}_S \left[\left\| \nabla f(x) - \frac{1}{|S|} \sum_{i \in S} \hat{\nabla}_{coord} f_i(x_t) \right\|^2 \right]$$

$$\overset{③}{\geq} \frac{\eta}{2}((1 - 2\eta\ell)\|\nabla f(x_t)\|^2 - \frac{\rho^2 d\mu^4}{36}) - 2\eta^2 \ell \mathbb{E}_S \left[\left\| \nabla f(x) - \frac{1}{|S|} \sum_{i \in S} \nabla f_i(x_t) \right\|^2 \right.$$

$$\left. + \left\| \frac{1}{|S|} \sum_{i \in S} \left(\nabla f_i(x_t) - \hat{\nabla}_{coord} f_i(x_t) \right) \right\|^2 \right]$$

$$\overset{④}{\geq} \frac{\eta}{2}((1 - 2\eta\ell)\|\nabla f(x_t)\|^2 - \frac{\rho^2 d\mu^4}{36}) - 2\eta^2 \ell(\frac{\sigma^2}{B} + \frac{\rho^2 d\mu^4}{36})$$

$$= \frac{\eta - 2\eta^2 \ell}{2} \|\nabla f(x_t)\|^2 - (\frac{\eta}{2} + 2\eta^2 \ell)\frac{\rho^2 d\mu^4}{36} - \frac{2\eta^2 \ell \sigma^2}{B},$$

where inequality ① holds since $-2\langle a, b \rangle \leq \|a - b\|^2 - \|a\|^2$; ② and ③ holds since $\|a + b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$ and Lemma 7; inequality ④ holds since Lemma 5 and Lemma 7. With the choice of $\eta = \frac{1}{4\ell}, \mu \leq \sqrt{\frac{3\epsilon}{4\rho\sqrt{d}}}$ and $B = \max\{\frac{32\sigma^2}{\epsilon^2}, 1\}$ we have

$$f(x_t) - \mathbb{E}_S[f(x_{t+1})] \geq \frac{1}{16\ell}(\|\nabla f(x)\|^2 - \frac{\epsilon^2}{8})$$

Thus as long as Line 5 of Algorithm 4 is reached, we have $f(x_t) - \mathbb{E}f(x_{t+1}) \geq \Omega(\frac{\epsilon^2}{\ell})$. On the other hand, whenever line 10 is reached, we have $v^\mathsf{T} \nabla f(x_t) v \leq -\frac{\delta}{2}$. By Lemma 6, we have $f(x_t) - \mathbb{E}f(x_{t+1}) \geq \Omega(\frac{\delta^3}{\rho^2})$.

Then we choose $K = \mathcal{O}\left(\frac{\rho^2 \Delta_f}{\delta^3} + \frac{\ell \Delta_f}{\epsilon^2} \right)$, then the algorithm must terminate. As for the total query complexity, we note that each iteration of Algorithm 4 needs $\tilde{\mathcal{O}}(B) = \tilde{\mathcal{O}}(\frac{\sigma^2}{\epsilon^2} + 1)$ stochastic gradient estimators in Line 3 and Line 5, totaling $\tilde{\mathcal{O}}(d(\frac{\sigma^2}{\epsilon^2} + 1)K)$ function queries, as well as $\tilde{\mathcal{O}}(\frac{\ell^2}{\delta^2})$ stochastic gradient estimators computations with no more than $\mathcal{O}(\frac{\rho^2 \Delta_f}{\delta^3})$ times. Therefore, the total function query complexity is

$$\tilde{\mathcal{O}}\left(d(\frac{\sigma^2}{\epsilon^2} + 1)K + d\frac{\ell^2}{\delta^2}\frac{\rho^2 \Delta_f}{\delta^3} \right) = \tilde{\mathcal{O}}\left(d(\frac{\sigma^2}{\epsilon^2} + 1)(\frac{\rho^2 \Delta_f}{\delta^3} + \frac{\ell \Delta_f}{\epsilon^2}) + d\frac{\ell^2 \rho^2 \Delta_f}{\delta^5} \right)$$

D.2 Proof of Theorem 3 (Option II)

Lemma 13. *For any $x \in \mathbb{R}^d$, we have*

$$\mathbb{E}\|\frac{1}{|S|} \sum_{i \in S} \hat{\nabla}_{rand} f_i(x_t) - \hat{\nabla}_{rand} f(x_t)\|^2 \leq \frac{2d}{|S|}\|\nabla f(x_t)\|^2 + \frac{2d\sigma^2}{|S|} + \frac{\rho^2 d^2 \mu^4}{36|S|}$$

Proof. Let $z_i = \hat{\nabla}_{rand} f_i(x_t) - \hat{\nabla}_{rand} f(x_t)$ and $I_i = I(i \in S)$, where $I(\cdot)$ is the indicator function. Then we have $\mathbb{E}_i(I_i^2)\frac{|S|}{n}$ and $\mathbb{E}_i(I_i I_j) = \frac{\binom{|S|}{2}}{\binom{n}{2}} = \frac{|S|(|S|-1)}{n(n-1)}, i \neq j$. Then we have

$$\mathbb{E}\|\frac{1}{|S|} \sum_{i \in S} \hat{\nabla}_{rand} f_i(x_t) - \hat{\nabla}_{rand} f(x_t)\|^2 = \mathbb{E}\|\frac{1}{|S|} \sum_{i \in S} z_i\|^2 = \frac{1}{|S|^2}\mathbb{E}\|\sum_{i=1}^n z_i I_i\|^2$$

$$= \frac{1}{|S|^2}\left(\sum_{i=1}^n \mathbb{E}I_i^2\|z_i\|^2 + \sum_{i \neq j}\mathbb{E}I_i I_j \langle z_i, z_j \rangle \right) = \frac{1}{|S|^2}\mathbb{E}_u\left(\frac{|S|}{n}\sum_{i=1}^n\|z_i\|^2 + \frac{|S|(|S|-1)}{n(n-1)}\sum_{i \neq j}\langle z_i, z_j \rangle \right)$$

27

$$= \frac{1}{|S|^2} \mathbb{E}_u \left(\left(\frac{|S|}{n} - \frac{|S|(|S|-1)}{n(n-1)} \right) \sum_{i=1}^{n} \|z_i\|^2 + \frac{|S|(|S|-1)}{n(n-1)} \| \sum_{i=1}^{n} z_i \|^2 \right)$$

$$\overset{①}{=} \mathbb{E}_u \frac{n-|S|}{n(n-1)|S|} \sum_{i=1}^{n} \|z_i\|^2 \leq \frac{1}{|S|} \mathbb{E}_u \frac{1}{n} \sum_{i=1}^{n} \|z_i\|^2 = \frac{1}{|S|} \mathbb{E}_u \mathbb{E}_i \|\hat{\nabla}_{rand} f_i(x_t) - \hat{\nabla}_{rand} f(x_t)\|^2$$

$$\overset{②}{\leq} \frac{1}{|S|} \mathbb{E}_u \mathbb{E}_i \|\hat{\nabla}_{rand} f_i(x_t)\|^2 \overset{\text{Lemma 8}}{\leq} \frac{1}{|S|} \mathbb{E}_i \left(d\|\nabla f_i(x_t)\|^2 + \frac{\rho^2 d^2 \mu^4}{36} \right)$$

$$\leq \frac{d}{|S|} \mathbb{E}_i \left(2\|\nabla f(x_t)\|^2 + 2\|\nabla f(x_t) - \nabla f_i(x_t)\|^2 \right) + \frac{\rho^2 d^2 \mu^4}{36|S|}$$

$$\leq \frac{2d}{|S|} \|\nabla f(x_t)\|^2 + \frac{2d\sigma^2}{|S|} + \frac{\rho^2 d^2 \mu^4}{36|S|}$$

<div align="right">□</div>

If we update $x_{t+1} = x_t - \frac{\eta}{|S|} \sum_{i \in S} \hat{\nabla}_{rand} f_i(x_t)$, then according to the smoothness of $f_\mu(\cdot)$, we have

$$f_\mu(x_t) - \mathbb{E}[f_\mu(x_{t+1})]$$

$$\geq \mathbb{E} \left[\langle \nabla f_\mu(x_t), x_t - x_{t+1} \rangle - \frac{\ell}{2} \|x_t - x_{t+1}\|^2 \right]$$

$$= \mathbb{E} \left\langle \nabla f_\mu(x_t), \frac{\eta}{|S|} \sum_{i \in S} \hat{\nabla}_{rand} f_i(x_t) \right\rangle - \frac{\eta^2 \ell}{2} \mathbb{E} \left\| \frac{1}{|S|} \sum_{i \in S} \hat{\nabla}_{rand} f_i(x_t) \right\|^2$$

$$\overset{①}{\geq} \eta \|\nabla f_\mu(x_t)\|^2 - \frac{3\eta^2 \ell}{2} \mathbb{E} \left(\left\| \frac{1}{|S|} \sum_{i \in S} \hat{\nabla}_{rand} f_i(x_t) - \hat{\nabla}_{rand} f(x_t) \right\|^2 + \left\| \hat{\nabla}_{rand} f(x_t) - \nabla f_\mu(x_t) \right\|^2 \right.$$

$$\left. + \|\nabla f_\mu(x_t)\|^2 \right)$$

$$\overset{②}{\geq} \eta \left(1 - \frac{3\eta\ell}{2} \right) \|\nabla f_\mu(x_t)\|^2 - \frac{3\eta^2 \ell}{2} \left(\frac{2d}{|S|} \|\nabla f(x_t)\|^2 + \frac{2d\sigma^2}{|S|} + \frac{\rho^2 d^2 \mu^4}{36|S|} + d\|\nabla f(x_t)\|^2 + \frac{\rho^2 d^2 \mu^4}{36} \right)$$

$$\geq \eta \left(1 - \frac{3\eta\ell}{2} \right) \|\nabla f_\mu(x_t)\|^2 - \eta^2 \ell \left(\frac{9d}{2} \|\nabla f(x_t)\|^2 + \frac{3d\sigma^2}{|S|} + \frac{\rho^2 d^2 \mu^4}{24} \right)$$

$$\geq \eta \left(1 - \frac{3\eta\ell}{2} \right) \left(\frac{1}{2} \|\nabla f(x_t)\|^2 - \|\nabla f(x_t) - \nabla f_\mu(x_t)\|^2 \right) - \eta^2 \ell \left(\frac{9d}{2} \|\nabla f(x_t)\|^2 + \frac{3d\sigma^2}{|S|} + \frac{\rho^2 d^2 \mu^4}{24} \right)$$

$$\geq \eta \left(1 - \frac{3\eta\ell}{2} \right) \left(\frac{1}{2} \|\nabla f(x_t)\|^2 - \frac{\rho^2 d^2 \mu^4}{36} \right) - \eta^2 \ell \left(\frac{9d}{2} \|\nabla f(x_t)\|^2 + \frac{3d\sigma^2}{|S|} + \frac{\rho^2 d^2 \mu^4}{24} \right)$$

$$\geq \eta (\frac{1}{2} - 8d\eta\ell) \|\nabla f(x_t)\|^2 - 3d\eta^2 \ell \frac{\sigma^2}{|S|} - \eta \frac{\rho^2 d^2 \mu^4}{36}$$

where ① is due to $\|a+b+c\|^2 \leq 3(\|a\|^2 + \|b\|^2 + \|c\|^2)$; ② is due to Lemma 8 and Lemma 13.

Since $|f(x) - f_\mu(x)| \leq \frac{\ell\mu^2}{2}$, we have

$$f(x_t) - \mathbb{E}[f(x_{t+1})]$$

$$\geq f_\mu(x_t) - \mathbb{E}[f_\mu(x_{t+1})] - \ell\mu^2$$

$$\geq \eta (\frac{1}{2} - 8d\eta\ell) \|\nabla f(x_t)\|^2 - 3d\eta^2 \ell \frac{\sigma^2}{|S|} - \eta \frac{\rho^2 d^2 \mu^4}{36} - \ell\mu^2$$

With the choice of $\eta = \frac{1}{32d\ell}, \mu = \min \left\{ \sqrt{\frac{3\epsilon}{4\rho d}}, \frac{\epsilon}{32\sqrt{d}\ell} \right\}, B = \max\{\frac{8\sigma^2}{\epsilon^2}, 1\}$, we have

$$f(x_t) - \mathbb{E}[f(x_{t+1})]$$

$$\geq \frac{1}{128d\ell} \|\nabla f(x_t)\|^2 - \frac{1}{256d\ell} \frac{\sigma^2}{B} - \frac{1}{128d\ell} \frac{\rho^2 d^2 \mu^4}{9} - \ell\mu^2$$

<div align="center">28</div>

$$= \frac{1}{128d\ell} \left(\|\nabla f(x_t)\|^2 - \frac{\sigma^2}{2B} - \frac{\rho^2 d^2 \mu^4}{9} - 128d\ell^2 \mu^2 \right)$$

$$\geq \frac{1}{128d\ell} \left(\|\nabla f(x_t)\|^2 - \frac{\epsilon^2}{8} \right)$$

Thus as long as Line 6 of Algorithm 4 is reached, we have $f(x_t) - \mathbb{E}f(x_{t+1}) \geq \Omega(\frac{\epsilon^2}{d\ell})$. On the other hand, whenever line 10 is reached, we have $v^\mathsf{T} \nabla f(x_t) v \leq -\frac{\delta}{2}$. By Lemma 6, we have $f(x_t) - \mathbb{E}f(x_{t+1}) \geq \Omega(\frac{\delta^3}{\rho^2})$.

Then we choose $K = \mathcal{O}\left(\frac{\rho^2 \Delta_f}{\delta^3} + \frac{d\ell \Delta_f}{\epsilon^2} \right)$, then the algorithm must terminate. As for the total query complexity, we note that each iteration of Algorithm 4 needs $\tilde{\mathcal{O}}(B) = \tilde{\mathcal{O}}(\frac{\sigma^2}{\epsilon^2} + 1)$ stochastic gradient estimators in Line 6 and $\tilde{\mathcal{O}}(\frac{\sigma^2}{\epsilon^2} + 1)$ deterministic coordinate-wise gradient estimators Line 3, totaling $\tilde{\mathcal{O}}(d(\frac{\sigma^2}{\epsilon^2} + 1)K)$ function queries, as well as $\tilde{\mathcal{O}}(\frac{\ell^2}{\delta^2})$ stochastic gradient estimators computations with no more than $\mathcal{O}(\frac{\rho^2 \Delta_f}{\delta^3})$ times. Therefore, the total function query complexity is

$$\tilde{\mathcal{O}}\left((d\frac{\sigma^2}{\epsilon^2} + d)K + d\frac{\ell^2}{\delta^2}\frac{\rho^2 \Delta_f}{\delta^3} \right) = \tilde{\mathcal{O}}\left(d(\frac{\sigma^2}{\epsilon^2} + 1)(\frac{\rho^2 \Delta_f}{\delta^3} + \frac{d\ell \Delta_f}{\epsilon^2}) + d\frac{\ell^2 \rho^2 \Delta_f}{\delta^5} \right)$$

E Applying Zeroth-Order Negative Curvature Finding to ZO-SCSG

In this section, we first propose a zeroth-order variant of the SCSG [32] method in Algorithm 6. At the beginning of the j-th epoch, we estimate the gradient $\nabla f_{\mathcal{I}_j}(\tilde{x}_{j-1})$ by CoordGradEst over a batch sampling set $\mathcal{I}_j$ with size B. In the inner loop iterations, the stochastic gradient estimator v_{k-1}^j is either constructed by CoordGradEst or by RandGradEst over a mini-batch sampling set $\mathcal{I}_{k-1}^j$ with size b. After running ZO-SCSG for one epoch, we have the following lemma:

Algorithm 6 ZO-SCSG

Input: Number of stages T, initial point $\tilde{x}_0$, batch size B, mini-bath size b, learning rate $\eta > 0$.
1: **for** $j = 1, \dots, T$ **do**
2:      Uniformly randomly sample a batch $\mathcal{I}_j \subset [n]$ with $|\mathcal{I}_j| = B$
3:      $v_j = \hat{\nabla}_{coord} f_{\mathcal{I}_j}(\tilde{x}_{j-1})$
4:      $x_0^j = \tilde{x}_{j-1}$
5:      **Option I:** $N_j \sim \text{Geom}(\frac{B}{B+b})$ **Option II:** $N_j \sim \text{Geom}(\frac{B}{B+b/d})$
6:      **for** $k = 1, \dots, N_j$ **do**
7:          Randomly pick $\mathcal{I}_{k-1}^j \subset [n]$ with size b
8:          **Option I:** $v_{k-1}^j = \hat{\nabla}_{coord} f_{\mathcal{I}_{k-1}^j}(x_{k-1}^j) - \hat{\nabla}_{coord} f_{\mathcal{I}_{k-1}^j}(x_0^j) + v_j$
9:          **Option II:** $v_{k-1}^j = \hat{\nabla}_{rand} f_{\mathcal{I}_{k-1}^j}(x_{k-1}^j) - \hat{\nabla}_{rand} f_{\mathcal{I}_{k-1}^j}(x_0^j) + v_j$
10:          $x_k^j = x_{k-1}^j - \eta v_{k-1}^j$
11:      $\tilde{x}_j = x_{N_j}^j$

Lemma 14 (One epoch analysis). *Under Assumption 1, **Option I:** Let $\eta\ell = \gamma \left(\frac{b}{B} \right)^{\frac{2}{3}}$. Suppose $\gamma \leq \frac{1}{4}$, $B \geq 8b$ and $b \geq 1$, then after running ZO-SCSG for one epoch, we have*

$$\left(\frac{B}{b} \right)^{\frac{1}{3}} \mathbb{E}\|\nabla f(\tilde{x}_j)\|^2 \leq \frac{4\ell}{\gamma} (f(\tilde{x}_{j-1}) - f(\tilde{x}_j)) + \frac{30\sigma^2}{b^{\frac{1}{3}}B^{\frac{2}{3}}} + c\left(\frac{B}{b} \right)^{\frac{1}{3}} \ell^2 d\mu^2, \qquad (13)$$

*where c is a sufficiently large constant. **Option II:** Let $\eta\ell = \gamma \left(\frac{b/d}{B} \right)^{\frac{2}{3}}$. Suppose $\gamma \leq \frac{1}{8}$, $B \geq 8b/d$ and $b \geq d$, then after running ZO-SCSG for one epoch, we have*

$$\left(\frac{B}{b/d} \right)^{\frac{1}{3}} \mathbb{E}\|\nabla f(\tilde{x}_j)\|^2 \leq \frac{8\ell}{\gamma} \mathbb{E}(f(\tilde{x}_{j-1}) - f(\tilde{x}_j)) + \frac{72\sigma^2}{(b/d)^{\frac{1}{3}}B^{\frac{2}{3}}} + c\left(\frac{B}{b/d} \right)^{\frac{1}{3}} \ell^2 d^2\mu^2, \qquad (14)$$

29

Algorithm 7 ZO-SCSG-NCF

Input: Function f, starting point x_0, batch size B, mini-batch size b, K, $\epsilon > 0$ and $\delta > 0$.

1: **if** $b > B$ **then return** ZO-SGD$(f, x_0, \frac{2}{3}, \epsilon, \delta)$

2: **for** $t = 0, \ldots, K - 1$ **do**

3:     uniformly randomly choose a set $\mathcal{B}$ with batch size $\mathcal{O}(\frac{\sigma^2}{\epsilon^2} \log K)$

4:     **if** $\|\hat{\nabla}_{coord} f_{\mathcal{B}}(x_t)\| \geq \frac{3\epsilon}{4}$ **then**

5:         $x_{t+1} \leftarrow$ apply ZO-SCSG on x_t for one epoch with batch size B and mini-batch size b

6:     **else**

7:         $v \leftarrow$ ZO-NCF-Online $(f, x_t, \delta, \frac{1}{20K})$

8:         **if** $v = \perp$ **then return** x_t

9:         **else** $x_{t+1} = x_t \pm \frac{\delta}{\rho} v$

---

where c is a sufficiently large constant.

Remark 5. *The epoch size N_j obeys the Geometric distribution, i.e., $N_j \sim Geom(\frac{B}{B+b})$ in **Option I** and $N_j \sim Geom(\frac{B}{B+b/d})$ in **Option II**. Since in expectation we have $\mathbb{E}(N_j)_{N_j \sim Geom(\theta)} = \frac{\theta}{1-\theta}$ [32], then for both **Option I** and **Option II**, the function query complexity in each epoch is $\mathcal{O}(d \cdot B)$.*

Theorem 7. *Under Assumption 1, if we set $\mu_1 = \sqrt{\frac{3\epsilon}{4\rho\sqrt{d}}}$ and other parameters as follows,*

$$\textbf{Option I:} B = \max\{\frac{480\sigma^2}{\epsilon^2}, 1\}, b = \max\{1, \Theta(\frac{(\epsilon^2 + \sigma^2)\epsilon^4\rho^6}{\delta^9\ell^3})\}, K = \Theta(\frac{\ell b^{\frac{1}{3}}\Delta_f}{\epsilon^2 B^{\frac{1}{3}}}), \mu_2 = \frac{\epsilon}{4\sqrt{cd}\ell};$$

$$\textbf{Option II:} B = \max\{1, \frac{1152\sigma^2}{\epsilon^2}\}, b = \max\{1, d\Theta(\frac{(\epsilon^2 + \sigma^2)\epsilon^4\rho^6}{\delta^9\ell^3})\}, K = \Theta(\frac{\ell(b/d)^{\frac{1}{3}}\Delta_f}{\epsilon^2 B^{\frac{1}{3}}}), \mu_2 = \frac{\epsilon}{4\sqrt{cd}\ell},$$

*where μ_1 and μ_2 are only use in Line 4 and Line 5 of Algorithm 7, respectively. With probability at least $\frac{2}{3}$, for both **Option I** and **Option II**, Algorithm 7 outputs an (ϵ, δ)-approximate local minimum in function query complexity*

$$\tilde{\mathcal{O}}(d(\frac{\ell\Delta_f}{\epsilon^{\frac{4}{3}}\sigma^{\frac{2}{3}}} + \frac{\rho^2\Delta_f}{\delta^3})(\frac{\sigma^2}{\epsilon^2} + \frac{\ell^2}{\delta^2}) + d\frac{\ell\Delta_f}{\epsilon^2}\frac{\ell^2}{\delta^2}).$$

Remark 6. *The problem described in Remark 2 doesn't exist in ZO-SCSG-NCF as we only evaluate the magnitude of the gradient after each epoch (i.e., Line 4 in Algorithm 7), and the function query complexity is almost the same in the inner loop for both **Option I** and **Option II**. We can boost the confidence in Theorem 5 from $2/3$ to $1 - p$ by running $\log 1/p$ copies of ZO-SCSG-NCF.*

E.1 One Epoch Analysis of ZO-SCSG (Option I)

Lemma 15 ([32]). *Let $N \sim Geom(\gamma)$ for $\gamma > 0$. Then for any sequence $D_0, D_1, \ldots$ with $\mathbb{E}|D_N| < \infty$*

$$\mathbb{E}(D_N - D_{N+1}) = \left(\frac{1}{\gamma} - 1\right)(D_0 - \mathbb{E}D_N).$$

Proof. Then proof directly follows from Lemma A.2 in [32]. $\qquad\square$

Lemma 16. *Suppose $\eta\ell < 1$, then under Assumption 1,*

$$\eta(1 - \ell\eta)B\mathbb{E}\|\nabla f(\tilde{x}_j)\|^2 + \eta B\mathbb{E}\left\langle \hat{e}_{N_j}^j, \nabla f(\tilde{x}_j)\right\rangle$$

$$\leq b\left(f(\tilde{x}_{j-1}) - \mathbb{E}f(\tilde{x}_j)\right) + \frac{\ell^3\eta^2 B}{b}\mathbb{E}\|\tilde{x}_j - \tilde{x}_{j-1}\|^2 + \frac{4\ell^3\eta^2 d\mu^2 B}{b} + \ell\eta^2 B\mathbb{E}\|\hat{e}_{N_j}^j\|^2$$

Proof. By Lemma 4, we have

$$f(x_{k+1}^j) \leq f(x_k^j) - \left\langle x_{k+1}^j - x_k^j, \nabla f(x_k^j)\right\rangle + \frac{\ell}{2}\|x_{k+1}^j - x_k^j\|^2 \leq f(x_k^j) - \eta\left\langle v_k^j, \nabla f(x_k^j)\right\rangle + \frac{\ell\eta^2}{2}\|v_k^j\|^2$$

30

Define the following notation,

$$\hat{e}_j = v_j - \hat{\nabla}_{coord} f(x_0^j)$$
$$\hat{e}_k^j = \hat{\nabla}_{coord} f(x_k^j) - \nabla f(x_k^j) + \hat{e}_j$$

Then we have

$$\mathbb{E}_{\mathcal{I}_k^j} v_k^j = \mathbb{E}_{\mathcal{I}_k^j} \left(\hat{\nabla}_{coord} f_{\mathcal{I}_k^j}(x_k^j) - \hat{\nabla}_{coord} f_{\mathcal{I}_k^j}(x_0^j) + v_j \right) = \nabla f(x_k^j) + \hat{e}_k^j$$

Taking expectation over the above inequality we have

$$\mathbb{E}_{\mathcal{I}_k^j} f(x_{k+1}^j)$$
$$\leq f(x_k^j) - \eta \left\langle \mathbb{E}_{\mathcal{I}_k^j} v_k^j, \nabla f(x_k^j) \right\rangle + \frac{\ell \eta^2}{2} \mathbb{E}_{\mathcal{I}_k^j} \|v_k^j\|^2$$
$$= f(x_k^j) - \eta \left\langle \nabla f(x_k^j) + \hat{e}_k^j, \nabla f(x_k^j) \right\rangle + \frac{\ell \eta^2}{2} \mathbb{E}_{\mathcal{I}_k^j} \|v_k^j\|^2$$
$$= f(x_k^j) - \eta \|\nabla f(x_k^j)\|^2 - \eta \left\langle \hat{e}_k^j, \nabla f(x_k^j) \right\rangle + \frac{\ell \eta^2}{2} \mathbb{E}_{\mathcal{I}_k^j} \|v_k^j\|^2$$

Then we bound the term $\mathbb{E}_{\mathcal{I}_k^j} \|v_k^j\|^2$ by using the fact that $\mathbb{E}\|a\|^2 = \mathbb{E}\|a - \mathbb{E}a\|^2 + \|\mathbb{E}a\|^2$.

$$\mathbb{E}_{\mathcal{I}_k^j} \|v_k^j\|^2 = \mathbb{E}_{\mathcal{I}_k^j} \|v_k^j - \mathbb{E}_{\mathcal{I}_k^j} v_k^j\|^2 + \|\mathbb{E}_{\mathcal{I}_k^j} v_k^j\|^2$$
$$= \mathbb{E}_{\mathcal{I}_k^j} \|\hat{\nabla}_{coord} f_{\mathcal{I}_k^j}(x_k^j) - \hat{\nabla}_{coord} f_{\mathcal{I}_k^j}(x_0^j) - \left(\hat{\nabla}_{coord} f(x_k^j) - \hat{\nabla}_{coord} f(x_0^j) \right) \|^2 + \|\nabla f(x_k^j) + \hat{e}_k^j\|^2$$
$$\leq \mathbb{E}_{\mathcal{I}_k^j} \|\hat{\nabla}_{coord} f_{\mathcal{I}_k^j}(x_k^j) - \hat{\nabla}_{coord} f_{\mathcal{I}_k^j}(x_0^j) - \left(\hat{\nabla}_{coord} f(x_k^j) - \hat{\nabla}_{coord} f(x_0^j) \right) \|^2 + 2\|\nabla f(x_k^j)\|^2 + 2\|\hat{e}_k^j\|^2$$

By Lemma 5,

$$\mathbb{E}_{\mathcal{I}_k^j} \|\hat{\nabla}_{coord} f_{\mathcal{I}_k^j}(x_k^j) - \hat{\nabla}_{coord} f_{\mathcal{I}_k^j}(x_0^j) - \left(\hat{\nabla}_{coord} f(x_k^j) - \hat{\nabla}_{coord} f(x_0^j) \right) \|^2$$
$$\leq \frac{1}{b} \cdot \frac{1}{n} \sum_{i=1}^{n} \|\hat{\nabla}_{coord} f_i(x_k^j) - \hat{\nabla}_{coord} f_i(x_0^j) - \left(\hat{\nabla}_{coord} f(x_k^j) - \hat{\nabla}_{coord} f(x_0^j) \right) \|^2$$
$$= \frac{1}{b} \left(\frac{1}{n} \sum_{i=1}^{n} \|\hat{\nabla}_{coord} f_i(x_k^j) - \hat{\nabla}_{coord} f_i(x_0^j)\|^2 - \|\hat{\nabla}_{coord} f(x_k^j) - \hat{\nabla}_{coord} f(x_0^j)\|^2 \right)$$
$$\leq \frac{1}{b} \cdot \frac{1}{n} \sum_{i=1}^{n} \|\hat{\nabla}_{coord} f_i(x_k^j) - \hat{\nabla}_{coord} f_i(x_0^j)\|^2$$
$$\leq \frac{1}{b} \cdot \frac{1}{n} \sum_{i=1}^{n} \left(2\|\nabla f_i(x_k^j) - \nabla f_i(x_0^j)\|^2 + 2\|\hat{\nabla}_{coord} f_i(x_k^j) - \hat{\nabla}_{coord} f_i(x_0^j) - \left(\nabla f_i(x_k^j) - \nabla f_i(x_0^j) \right) \|^2 \right)$$
$$\leq \frac{1}{b} \cdot \frac{1}{n} \sum_{i=1}^{n} \left(2\|\nabla f_i(x_k^j) - \nabla f_i(x_0^j)\|^2 + 4\|\hat{\nabla}_{coord} f_i(x_k^j) - \nabla f_i(x_k^j)\|^2 + 4\|\hat{\nabla}_{coord} f_i(x_0^j) - \nabla f_i(x_0^j)\|^2 \right)$$
$$\leq \frac{1}{b} \cdot \frac{1}{n} \sum_{i=1}^{n} \left(2\|\nabla f_i(x_k^j) - \nabla f_i(x_0^j)\|^2 + 8\ell^2 d\mu^2 \right)$$
$$\leq \frac{1}{b} \cdot \frac{1}{n} \sum_{i=1}^{n} \left(2\ell^2 \|x_k^j - x_0^j\|^2 + 8\ell^2 d\mu^2 \right)$$
$$= \frac{2\ell^2}{b} \|x_k^j - x_0^j\|^2 + \frac{8\ell^2 d\mu^2}{b}$$

Therefore,

$$\mathbb{E}_{\mathcal{I}_k^j} \|v_k^j\|^2 \leq \frac{2\ell^2}{b} \|x_k^j - x_0^j\|^2 + \frac{8}{b} \ell^2 d\mu^2 + 2\|\nabla f(x_k^j)\|^2 + 2\|\hat{e}_k^j\|^2$$

So we have

$$\mathbb{E}_{\mathcal{I}_k^j} f(x_{k+1}^j)$$

31

$$\leq f(x_k^j) - \eta\|\nabla f(x_k^j)\|^2 - \eta\left\langle \hat{e}_k^j, \nabla f(x_k^j)\right\rangle + \frac{\ell\eta^2}{2}\left(\frac{2\ell^2}{b}\|x_k^j - x_0^j\|^2 + \frac{8}{b}\ell^2 d\mu^2 + 2\|\nabla f(x_k^j)\|^2 + 2\|\hat{e}_k^j\|^2\right)$$

$$= f(x_k^j) - \eta(1 - \ell\eta)\|\nabla f(x_k^j)\|^2 - \eta\left\langle \hat{e}_k^j, \nabla f(x_k^j)\right\rangle + \frac{\ell^3\eta^2}{b}\|x_k^j - x_0^j\|^2 + \frac{4\ell^3\eta^2 d\mu^2}{b} + \ell\eta^2\|\hat{e}_k^j\|^2$$

Let $\mathbb{E}_j$ denotes the expectation over $\mathcal{I}_0^k, \mathcal{I}_1^k, \ldots$, given N_j. Since $\mathcal{I}_{k+1}^j, \mathcal{I}_{k+2}^j, \ldots$ are independent of x_k^j, the above inequality implies that

$$\eta(1 - \ell\eta)\mathbb{E}_j\|\nabla f(x_k^j)\|^2 + \eta\mathbb{E}_j\left\langle \hat{e}_k^j, \nabla f(x_k^j)\right\rangle$$

$$\leq \mathbb{E}_j f(x_k^j) - \mathbb{E}_j f(x_{k+1}^j) + \frac{\ell^3\eta^2}{b}\mathbb{E}_j\|x_k^j - x_0^j\|^2 + \frac{4\ell^3\eta^2 d\mu^2}{b} + \ell\eta^2\mathbb{E}_j\|\hat{e}_k^j\|^2$$

Let $k = N_j$, by taking expectation to N_j and using Fubini's theorem, we have

$$\eta(1 - \ell\eta)\mathbb{E}_{N_j}\mathbb{E}_j\|\nabla f(x_{N_j}^j)\|^2 + \eta\mathbb{E}_{N_j}\mathbb{E}_j\left\langle \hat{e}_{N_j}^j, \nabla f(x_{N_j}^j)\right\rangle$$

$$\leq \mathbb{E}_{N_j}\left(\mathbb{E}_j f(x_{N_j}^j) - \mathbb{E}_j f(x_{N_j+1}^j)\right) + \frac{\ell^3\eta^2}{b}\mathbb{E}_{N_j}\mathbb{E}_j\|x_{N_j}^j - x_0^j\|^2 + \frac{4\ell^3\eta^2 d\mu^2}{b} + \ell\eta^2\mathbb{E}_{N_j}\mathbb{E}_j\|\hat{e}_{N_j}^j\|^2$$

$$= \frac{b}{B}\left(f(x_0^j) - \mathbb{E}_{N_j}\mathbb{E}_j f(x_{N_j}^j)\right) + \frac{\ell^3\eta^2}{b}\mathbb{E}_j\mathbb{E}_{N_j}\|x_{N_j}^j - x_0^j\|^2 + \frac{4\ell^3\eta^2 d\mu^2}{b} + \ell\eta^2\mathbb{E}_{N_j}\mathbb{E}_j\|\hat{e}_{N_j}^j\|^2$$

Substituting $X_{N_j}^j, x_0^j$ by $\tilde{x}_j, \tilde{x}_{j-1}$ and take a further expectation to the past randomness, we get

$$\eta(1 - \ell\eta)\mathbb{E}\|\nabla f(\tilde{x}_j)\|^2 + \eta\mathbb{E}\left\langle \hat{e}_{N_j}^j, \nabla f(\tilde{x}_j)\right\rangle$$

$$\leq \frac{b}{B}\left(f(\tilde{x}_{j-1}) - \mathbb{E}f(\tilde{x}_j)\right) + \frac{\ell^3\eta^2}{b}\mathbb{E}\|\tilde{x}_j - \tilde{x}_{j-1}\|^2 + \frac{4\ell^3\eta^2 d\mu^2}{b} + \ell\eta^2\mathbb{E}\|\hat{e}_{N_j}^j\|^2$$

Multiplying both sides by B, we have

$$\eta(1 - \ell\eta)B\mathbb{E}\|\nabla f(\tilde{x}_j)\|^2 + \eta B\mathbb{E}\left\langle \hat{e}_{N_j}^j, \nabla f(\tilde{x}_j)\right\rangle$$

$$\leq b\left(f(\tilde{x}_{j-1}) - \mathbb{E}f(\tilde{x}_j)\right) + \frac{\ell^3\eta^2 B}{b}\mathbb{E}\|\tilde{x}_j - \tilde{x}_{j-1}\|^2 + \frac{4\ell^3\eta^2 d\mu^2 B}{b} + \ell\eta^2 B\mathbb{E}\|\hat{e}_{N_j}^j\|^2$$

$$\square$$

Lemma 17. *Suppose* $2\eta^2\ell^2 B < b^2$, *under Assumption 1*

$$\left(\frac{b}{B} - \frac{2\eta^2\ell^2}{b}\right)\mathbb{E}\|\tilde{x}_j - \tilde{x}_{j-1}\|^2 + 2\eta\mathbb{E}\left\langle \hat{e}_{N_j}^j, \tilde{x}_j - \tilde{x}_{j-1}\right\rangle$$

$$\leq -2\eta\mathbb{E}\left\langle \nabla f(\tilde{x}_j), \tilde{x}_j - \tilde{x}_{j-1}\right\rangle + \frac{8\eta^2\ell^2 d\mu^2}{b} + 2\eta^2\mathbb{E}\|\nabla f(\tilde{x}_j)\|^2 + 2\eta^2\mathbb{E}\|\hat{e}_{N_j}^j\|^2$$

Proof. Since $x_{k+1}^j = x_k^j - \eta v_k^j$, we have

$$\mathbb{E}_{\mathcal{I}_k^j}\|x_{k+1}^j - x_0^j\|^2$$

$$= \|x_k^j - x_0^j\|^2 - 2\eta\left\langle \mathbb{E}_{\mathcal{I}_k^j}v_k^j, x_k^j - x_0^j\right\rangle + \eta^2\mathbb{E}_{\mathcal{I}_k^j}\|v_k^j\|^2$$

$$= \|x_k^j - x_0^j\|^2 - 2\eta\left\langle \nabla f(x_k^j), x_k^j - x_0^j\right\rangle - 2\eta\left\langle \hat{e}_k^j, x_k^j - x_0^j\right\rangle + \eta^2\mathbb{E}_{\mathcal{I}_k^j}\|v_k^j\|^2$$

$$\leq \|x_k^j - x_0^j\|^2 - 2\eta\left\langle \nabla f(x_k^j), x_k^j - x_0^j\right\rangle - 2\eta\left\langle \hat{e}_k^j, x_k^j - x_0^j\right\rangle$$

$$+ \eta^2\left(\frac{2\ell^2}{b}\|x_k^j - x_0^j\|^2 + \frac{8}{b}\ell^2 d\mu^2 + 2\|\nabla f(x_k^j)\|^2 + 2\|\hat{e}_k^j\|^2\right)$$

$$= \left(1 + \frac{2\eta^2\ell^2}{b}\right)\|x_k^j - x_0^j\|^2 - 2\eta\left\langle \nabla f(x_k^j), x_k^j - x_0^j\right\rangle - 2\eta\left\langle \hat{e}_k^j, x_k^j - x_0^j\right\rangle$$

$$+ \frac{8\eta^2\ell^2 d\mu^2}{b} + 2\eta^2\|\nabla f(x_k^j)\|^2 + 2\eta^2\|\hat{e}_k^j\|^2$$

32

Using the notation $\mathbb{E}_j$ we have

$$2\eta\mathbb{E}_j\left\langle \nabla f(x_k^j), x_k^j - x_0^j \right\rangle + 2\eta\mathbb{E}_j\left\langle \hat{e}_k^j, x_k^j - x_0^j \right\rangle$$

$$\leq \left(1 + \frac{2\eta^2\ell^2}{b}\right)\mathbb{E}_j\|x_k^j - x_0^j\|^2 - \mathbb{E}_j\|x_{k+1}^j - x_0^j\|^2 + \frac{8\eta^2\ell^2 d\mu^2}{b} + 2\eta^2\|\nabla f(x_k^j)\|^2 + 2\eta^2\mathbb{E}_j\|\hat{e}_k^j\|^2$$

Let $k = N_j$, by taking expectation with respect to N_j and using Fubini's theorem, we have

$$2\eta\mathbb{E}_{N_j}\mathbb{E}_j\left\langle \nabla f(x_{N_j}^j), x_{N_j}^j - x_0^j \right\rangle + 2\eta\mathbb{E}_{N_j}\mathbb{E}_j\left\langle \hat{e}_{N_j}^j, x_{N_j}^j - x_0^j \right\rangle$$

$$\leq \left(1 + \frac{2\eta^2\ell^2}{b}\right)\mathbb{E}_{N_j}\mathbb{E}_j\|x_{N_j}^j - x_0^j\|^2 - \mathbb{E}_{N_j}\mathbb{E}_j\|x_{N_j+1}^j - x_0^j\|^2 + \frac{8\eta^2\ell^2 d\mu^2}{b}$$

$$+ 2\eta^2\mathbb{E}_{N_j}\|\nabla f(x_{N_j}^j)\|^2 + 2\eta^2\mathbb{E}_{N_j}\mathbb{E}_j\|\hat{e}_{N_j}^j\|^2$$

$$= \left(-\frac{b}{B} + \frac{2\eta^2\ell^2}{b}\right)\mathbb{E}_{N_j}\mathbb{E}_j\|x_{N_j}^j - x_0^j\|^2 + \frac{8\eta^2\ell^2 d\mu^2}{b} + 2\eta^2\mathbb{E}_{N_j}\|\nabla f(x_{N_j}^j)\|^2 + 2\eta^2\mathbb{E}_{N_j}\mathbb{E}_j\|\hat{e}_{N_j}^j\|^2$$

Substituting $x_{N_j}^j, x_0^j$ by $\tilde{x}_j, \tilde{x}_{j-1}$ and take a further expectation to the past randomness, we get

$$2\eta\mathbb{E}\left\langle \nabla f(\tilde{x}_j), \tilde{x}_j - \tilde{x}_{j-1} \right\rangle + 2\eta\mathbb{E}\left\langle \hat{e}_{N_j}^j, \tilde{x}_j - \tilde{x}_{j-1} \right\rangle$$

$$\leq \left(-\frac{b}{B} + \frac{2\eta^2\ell^2}{b}\right)\mathbb{E}\|\tilde{x}_j - \tilde{x}_{j-1}\|^2 + \frac{8\eta^2\ell^2 d\mu^2}{b} + 2\eta^2\mathbb{E}\|\nabla f(\tilde{x}_j)\|^2 + 2\eta^2\mathbb{E}\|\hat{e}_{N_j}^j\|^2$$

Swapping the order we get

$$\left(\frac{b}{B} - \frac{2\eta^2\ell^2}{b}\right)\mathbb{E}\|\tilde{x}_j - \tilde{x}_{j-1}\|^2 + 2\eta\mathbb{E}\left\langle \hat{e}_{N_j}^j, \tilde{x}_j - \tilde{x}_{j-1} \right\rangle$$

$$\leq -2\eta\mathbb{E}\left\langle \nabla f(\tilde{x}_j), \tilde{x}_j - \tilde{x}_{j-1} \right\rangle + \frac{8\eta^2\ell^2 d\mu^2}{b} + 2\eta^2\mathbb{E}\|\nabla f(\tilde{x}_j)\|^2 + 2\eta^2\mathbb{E}\|\hat{e}_{N_j}^j\|^2$$

$$\square$$

Lemma 18.

$$\frac{b}{B}\mathbb{E}\left\langle \hat{e}_k^j, \tilde{x}_j - \tilde{x}_{j-1} \right\rangle = -\eta\mathbb{E}\left\langle \hat{e}_{N_j}^j, \nabla f(\tilde{x}_j) \right\rangle - \eta\mathbb{E}\|\hat{e}_{N_j}^j\|^2$$

Proof. Let $M_k^j = \left\langle \hat{e}_k^j, x_k^j - x_0^j \right\rangle$. Then we have

$$\mathbb{E}_{N_j}\left\langle \hat{e}_k^j, \tilde{x}_j - \tilde{x}_{j-1} \right\rangle = \mathbb{E}_{N_j}M_{N_j}^j$$

Since N_j is independent of $x_0^j, \hat{e}_k^j$, we have

$$\mathbb{E}\left\langle \hat{e}_k^j, \tilde{x}_j - \tilde{x}_{j-1} \right\rangle = \mathbb{E}M_{N_j}^j$$

Also we have $M_0^j = 0$. On the other hand,

$$\mathbb{E}_{\mathcal{I}_k^j}\left(M_{k+1}^j - M_k^j\right) = \mathbb{E}_{\mathcal{I}_k^j}\left\langle \hat{e}_k^j, x_{k+1}^j - x_k^j \right\rangle = -\eta\left\langle \hat{e}_k^j, \mathbb{E}_{\mathcal{I}_k^j}v_k^j \right\rangle$$

$$= -\eta\left\langle \hat{e}_k^j, \nabla f(x_k^j) \right\rangle - \eta\|\hat{e}_k^j\|^2$$

Using the same notation $\mathbb{E}_j$ as in the proof in Lemma 16 and Lemma 17, we have

$$\mathbb{E}_j\left(M_{k+1}^j - M_k^j\right) = -\eta\left\langle \hat{e}_k^j, \mathbb{E}_j\nabla f(x_k^j) \right\rangle - \eta\mathbb{E}_j\|\hat{e}_k^j\|^2$$

Let $k = N_j$, by taking the expectation with respect to N_j and using Fubini's theorem, we have , we have

$$\frac{b}{B}\mathbb{E}_{N_j}\mathbb{E}_j M_{N_j}^j = -\eta\left\langle \hat{e}_{N_j}^j, \mathbb{E}_{N_j}\mathbb{E}_j\nabla f(x_{N_j}^j) \right\rangle - \eta\mathbb{E}_{N_j}\mathbb{E}_j\|\hat{e}_{N_j}^j\|^2$$

33

Substituting $x_{N_j}^j, x_0^j$ by $\tilde{x}_j, \tilde{x}_{j-1}$ and take a further expectation to the past randomness, we get

$$\frac{b}{B}\mathbb{E}\left\langle \hat{e}_k^j, \tilde{x}_j - \tilde{x}_{j-1}\right\rangle = -\eta\mathbb{E}\left\langle \hat{e}_{N_j}^j, \nabla f(\tilde{x}_j)\right\rangle - \eta\mathbb{E}\|\hat{e}_{N_j}^j\|^2$$

$\square$

Lemma 19 ([25]). *Define* $\hat{e}_j = v_j - \hat{\nabla}_{coord}f(x_0^j) = \hat{\nabla}_{coord}f_{\mathcal{I}_j}(\tilde{x}_{j-1}) - \hat{\nabla}_{coord}f(\tilde{x}_{j-1})$, *we have*

$$\|\hat{e}_j\|^2 \leq \frac{3(2\ell^2 d\mu^2 + \sigma^2)}{B}$$

Proof. The proof directly follows from Lemma 4 in [25]. $\square$

Proof of Lemma 14 (Option I). Multiplying Lemma 16 by 2, Lemma 17 by $\frac{b}{\eta}$ and summing them up, we have

$$2\eta B\left(1 - \eta\ell - \frac{b}{B}\right)\mathbb{E}\|\nabla f(\tilde{x}_j)\|^2 + \frac{b^3 - 2\eta^2\ell^2 bB - 2\ell^3\eta^3 B^2}{\eta bB}\mathbb{E}\|\tilde{x}_j - \tilde{x}_{j-1}\|^2$$
$$+ 2\eta B\mathbb{E}\left\langle \hat{e}_{N_j}^j, \nabla f(\tilde{x}_j)\right\rangle + 2b\left\langle \hat{e}_{N_j}^j, \tilde{x}_j - \tilde{x}_{j-1}\right\rangle$$
$$\leq -2b\mathbb{E}\left\langle \nabla f(\tilde{x}_j), \tilde{x}_j - \tilde{x}_{j-1}\right\rangle + 2b\left(f(\tilde{x}_{j-1}) - f(\tilde{x}_j)\right) + (\ell\frac{B}{b} + \frac{1}{\eta})8\eta^2\ell^2 d\mu^2 + (2\eta^2\ell B + 2\eta b)\mathbb{E}\|\hat{e}_{N_j}^j\|^2$$

By Lemma 18,

$$2\eta B\mathbb{E}\left\langle \hat{e}_{N_j}^j, \nabla f(\tilde{x}_j)\right\rangle + 2b\left\langle \hat{e}_{N_j}^j, \tilde{x}_j - \tilde{x}_{j-1}\right\rangle = -2\eta B\mathbb{E}\|\hat{e}_{N_j}^j\|^2$$

So the above inequality can simplified as

$$2\eta B\left(1 - \eta\ell - \frac{b}{B}\right)\mathbb{E}\|\nabla f(\tilde{x}_j)\|^2 + \frac{b^3 - 2\eta^2\ell^2 bB - 2\ell^3\eta^3 B^2}{\eta bB}\mathbb{E}\|\tilde{x}_j - \tilde{x}_{j-1}\|^2$$
$$\leq -2b\mathbb{E}\left\langle \nabla f(\tilde{x}_j), \tilde{x}_j - \tilde{x}_{j-1}\right\rangle + 2b\left(f(\tilde{x}_{j-1}) - f(\tilde{x}_j)\right) + (\ell\frac{B}{b} + \frac{1}{\eta})8\eta^2\ell^2 d\mu^2$$
$$+ (2\eta^2\ell B + 2\eta b + 2\eta B)\mathbb{E}\|\hat{e}_{N_j}^j\|^2$$

Using the fact that $2\left\langle a, b\right\rangle \leq \beta\|a\|^2 + \frac{1}{\beta}\|b\|^2$ for any $\beta > 0$, we have

$$-2b\mathbb{E}\left\langle \nabla f(\tilde{x}_j), \tilde{x}_j - \tilde{x}_{j-1}\right\rangle$$
$$\leq \frac{\eta bB}{b^3 - 2\eta^2\ell^2 bB - 2\ell^3\eta^3 B^2}b^2\mathbb{E}\|\nabla f(\tilde{x}_j)\|^2 + \frac{b^3 - 2\eta^2\ell^2 bB - 2\ell^3\eta^3 B^2}{\eta bB}\mathbb{E}\|\tilde{x}_j - \tilde{x}_{j-1}\|^2$$

Then we conclude that

$$\eta B\left(2 - 2\eta\ell - 2\frac{b}{B} - \frac{b^3}{b^3 - 2\eta^2\ell^2 bB - 2\ell^3\eta^3 B^2}\right)\mathbb{E}\|\nabla f(\tilde{x}_j)\|^2$$
$$\leq 2b\left(f(\tilde{x}_{j-1}) - f(\tilde{x}_j)\right) + (\ell\frac{B}{b} + \frac{1}{\eta})8\eta^2\ell^2 d\mu^2 + (2\eta^2\ell B + 2\eta b + 2\eta B)\mathbb{E}\|\hat{e}_{N_j}^j\|^2$$

Multiplying both sides by $\frac{\ell}{b}$, we have

$$\eta\ell\frac{B}{b}\left(2 - 2\eta\ell - 2\frac{b}{B} - \frac{b^3}{b^3 - 2\eta^2\ell^2 bB - 2\ell^3\eta^3 B^2}\right)\mathbb{E}\|\nabla f(\tilde{x}_j)\|^2$$
$$\leq 2\ell\left(f(\tilde{x}_{j-1}) - f(\tilde{x}_j)\right) + \left(\frac{\ell^2 B}{b^2} + \frac{\ell}{b\eta}\right)8\eta^2\ell^2 d\mu^2 + (2\eta^2\ell^2\frac{B}{b} + 2\eta\ell + 2\eta\ell\frac{B}{b})\mathbb{E}\|\hat{e}_{N_j}^j\|^2$$

Let $\eta\ell = \gamma\left(\frac{b}{B}\right)^{\frac{2}{3}}$ and $b \geq 1, \frac{B}{b} \geq 8 \geq \frac{8}{b}$, we have

$$b^3 - 2\eta^2\ell^2 bB - 2\ell^3\eta^3 B^2 = b^3\left(1 - 2\gamma^2 b(\frac{b}{B})^{\frac{1}{3}} - 2\gamma^3 b^{-1}\right)$$

34

$$\geq b^3(1 - \gamma^2 - 2\gamma^3)$$

Then the above inequality can be simplified as

$$\gamma \left(\frac{B}{b}\right)^{\frac{1}{3}} \left(2 - 2\gamma(\frac{b}{B})^{\frac{2}{3}} - 2\frac{b}{B} - \frac{1}{1-\gamma^2-2\gamma^3}\right) \mathbb{E}\|\nabla f(\tilde{x}_j)\|^2$$

$$\leq 2\ell \left(f(\tilde{x}_{j-1}) - f(\tilde{x}_j)\right) + 2\gamma \left(1 + \gamma \left(\frac{b}{B}\right)^{\frac{2}{3}} + \frac{b}{B}\right) \left(\frac{B}{b}\right)^{\frac{1}{3}} \mathbb{E}\|\hat{e}^j_{N_j}\|^2 + \gamma \left(\gamma(\frac{b}{B})^{\frac{1}{3}} + (\frac{b}{B})^{\frac{2}{3}}\right) \frac{8\ell^2 d\mu^2}{b}$$

Since $B \geq 8b, \gamma \leq \frac{1}{4}$, we have

$$2 - 2\gamma(\frac{b}{B})^{\frac{2}{3}} - 2\frac{b}{B} - \frac{1}{1-\gamma^2-2\gamma^3} \geq 2 - \frac{\gamma}{2} - \frac{1}{4} - \frac{1}{1-\gamma^2-2\gamma^3} \geq 0.5$$

$$1 + \gamma \left(\frac{b}{B}\right)^{\frac{2}{3}} + \frac{b}{B} \leq 1 + \frac{\gamma}{4} + \frac{1}{8} \leq \frac{19}{16}$$

$$\gamma(\frac{b}{B})^{\frac{1}{3}} + (\frac{b}{B})^{\frac{2}{3}} \leq \frac{1}{2}\gamma + \frac{1}{4} \leq \frac{3}{8}$$

Thus we have

$$\left(\frac{B}{b}\right)^{\frac{1}{3}} \mathbb{E}\|\nabla f(\tilde{x}_j)\|^2 \leq \frac{4\ell}{\gamma} \left(f(\tilde{x}_{j-1}) - f(\tilde{x}_j)\right) + 5 \left(\frac{B}{b}\right)^{\frac{1}{3}} \mathbb{E}\|\hat{e}^j_{N_j}\|^2 + \frac{6\ell^2 d\mu^2}{b}$$

Using Lemma 19, we have

$$\mathbb{E}\|\hat{e}^j_{N_j}\|^2 = \mathbb{E}\|\nabla f(x^j_{N_j}) - \hat{\nabla}_{coord} f(x^j_{N_j}) + \hat{e}_j\|^2$$

$$\leq 2\|\nabla f(x^j_{N_j}) - \hat{\nabla}_{coord} f(x^j_{N_j})\|^2 + 2\mathbb{E}\|\hat{e}_j\|^2$$

$$\leq 2\ell^2 d\mu^2 + 2 \left(\frac{3(2\ell^2 d\mu^2 + \sigma^2)}{B}\right)$$

Thus we obtain

$$\left(\frac{B}{b}\right)^{\frac{1}{3}} \mathbb{E}\|\nabla f(\tilde{x}_j)\|^2 \leq \frac{4\ell}{\gamma} \left(f(\tilde{x}_{j-1}) - f(\tilde{x}_j)\right) + \frac{30\sigma^2}{b^{\frac{1}{3}} B^{\frac{2}{3}}} + c \left(\frac{B}{b}\right)^{\frac{1}{3}} \ell^2 d\mu^2$$

where c is a sufficient large constant. Telescope the sum in $j = 1, \dots, T$, and using the definition of $\tilde{x}^*_T$, we finally get

$$\mathbb{E}\|\tilde{x}^*_T\|^2 = \frac{1}{T} \sum_{j=1}^{T} \mathbb{E}\|\nabla f(\tilde{x}_j)\|^2 \leq \frac{\frac{4\ell}{\gamma}}{T \left(\frac{B}{b/d}\right)^{\frac{1}{3}}} \cdot \mathbb{E}\left(f(\tilde{x}_0) - f(\tilde{x}_T)\right) + \frac{30\sigma^2}{B} + c\ell^2 d\mu^2$$

$\square$

E.2 Proof of Second-Order Stationary Point (Option I)

Proof of Theorem 5. Let N_1 and N_2 be the number of times we reach Line 7 and 9 of Algorithm 7. From Lemma 14 of ZO-SCSG we know that for one epoch with size $B = \max\{1, \frac{480\sigma^2}{\epsilon^2}\}$, mini-batch size $b \geq 1$ and the smoothing parameter $\mu = \frac{\epsilon}{4\sqrt{cd\ell}}$, we have

$$\mathbb{E}\|x_{t+1}\|^2 \leq \frac{4\ell}{\gamma} \left(\frac{b}{B}\right)^{\frac{1}{3}} \mathbb{E}\left(f(x_t) - f(x_{t+1})\right) + \frac{\epsilon^2}{8}$$

Then, if $\|\nabla f(x_{t+1})\| \geq \frac{\epsilon}{2}$, we have $x_{t+1} = x_{t+1}$; if $v = \perp$, we set $x_{t+1} = x_{t+1}$ for if $v \neq \perp$, we have $f(x_{t+1}) - \mathbb{E}f(x_{t+1}) \geq \frac{\delta^3}{12\rho^2}$ (here the expectation is taken on the randomness of sign of v). Thus we have

$$\frac{\gamma B^{\frac{1}{3}}}{4\ell b^{\frac{1}{3}}} \mathbb{E}\left[\sum_{t=0}^{K-1} \left(\|\nabla f(x_{t+1})\|^2 - \frac{\epsilon^2}{8}\right)\right] + \frac{\delta^3}{12\rho^2} \mathbb{E}[N_2] \leq \Delta_f$$

35

On one hand, since we have chosen K such that $K \geq \Omega\left(\frac{\ell b^{\frac{1}{3}} \Delta_f}{\epsilon^2 B^{\frac{1}{3}}}\right) = \Omega\left(\frac{\ell b^{\frac{1}{3}} \Delta_f}{\epsilon^2 (1 + \frac{\sigma^2}{\epsilon^2})^{\frac{1}{3}}}\right)$, then by Markov's inequality, with probability at least $\frac{5}{6}$, it satisfies $\sum_{t=0}^{K-1} \|\nabla f(x_{t+1})\|^2 \leq \frac{\epsilon}{4} K$. As a sequence, at least half of the indices $t = 0, \ldots, K-1$ will satisfy $\|\nabla f(x_{t+1})\| \leq \frac{\epsilon}{2}$, which means that $N_1 \geq \frac{K}{2}$.

On the other hand, we have $\frac{\delta^3}{12\rho^2} \mathbb{E}[N_2] \leq \Delta_f + \frac{K\gamma B^{\frac{1}{3}} \epsilon^2}{32 \ell b^{\frac{1}{3}}}$. Since $K \geq \Omega\left(\frac{\ell b^{\frac{1}{3}} \Delta_f}{\epsilon^2 B^{\frac{1}{3}}}\right) = \Omega\left(\frac{\ell b^{\frac{1}{3}} \Delta_f}{\epsilon^2 (1 + \frac{\sigma^2}{\epsilon^2})^{\frac{1}{3}}}\right)$, we have $\mathbb{E}[N_2] \leq \frac{K\gamma B^{\frac{1}{3}} \epsilon^2 \rho^2}{\ell \delta^3 b^{\frac{1}{3}}}$. As long as $B \leq \mathcal{O}\left(\frac{\ell^3 \delta^9 b}{\epsilon^6 \rho^6}\right)$, or equivalently $b \geq \Omega\left(\frac{B \epsilon^6 \rho^6}{\delta^9 \ell^3}\right)$, we have $\mathbb{E}[N_2] \leq \frac{K}{12}$. Therefore, with provability at least $\frac{5}{6}$, it satisfies $N_2 \leq \frac{k}{2}$.

Since $N_1 \geq N_2$, this means with probability at least $\frac{2}{3}$ the algorithm must terminate and output some x_{t+1} in an iteration.

Finally, the per-iteration complexity of Algorithm 7 is dominated by $\tilde{\mathcal{O}}(B)$ stochastic gradient estimators per iteration for both ZO-SCSG and estimating $\|\nabla f(x_{t+1})\|$, as well as $\tilde{\mathcal{O}}(\frac{\ell^2}{\delta^2})$ invoking Algorithm 2. The total function query complexity is

$$
\begin{aligned}
\tilde{\mathcal{O}}\left(dK\left(B + \frac{\ell^2}{\delta^2}\right)\right) &= \tilde{\mathcal{O}}\left(d\left(\frac{\ell b^{\frac{1}{3}} \Delta_f}{\epsilon^2 B^{\frac{1}{3}}}\right)\left(B + \frac{\ell^2}{\delta^2}\right)\right) \\
&= \tilde{\mathcal{O}}\left(d\left(\frac{\ell \max\{1, \frac{B^{\frac{1}{3}} \epsilon^2 \rho^2}{\delta^3 \ell}\} \Delta_f}{\epsilon^2 B^{\frac{1}{3}}}\right)\left(B + \frac{\ell^2}{\delta^2}\right)\right) \\
&= \tilde{\mathcal{O}}\left(d\left(\frac{\ell(1 + \frac{B^{\frac{1}{3}} \epsilon^2 \rho^2}{\delta^3 \ell}) \Delta_f}{\epsilon^2 B^{\frac{1}{3}}}\right)\left(B + \frac{\ell^2}{\delta^2}\right)\right) \\
&= \tilde{\mathcal{O}}\left(d\left(\frac{\ell \Delta_f}{\epsilon^2 B^{\frac{1}{3}}} + \frac{\rho^2 \Delta_f}{\delta^3}\right)\left(B + \frac{\ell^2}{\delta^2}\right)\right) \\
&= \tilde{\mathcal{O}}\left(d\left(\frac{\ell \Delta_f}{\epsilon^2 \max\{1, \frac{\sigma^2}{\epsilon^2}\}^{\frac{1}{3}}} + \frac{\rho^2 \Delta_f}{\delta^3}\right)\left(\max\{1, \frac{\sigma^2}{\epsilon^2}\} + \frac{\ell^2}{\delta^2}\right)\right) \\
&= \tilde{\mathcal{O}}\left(d\left(\frac{\ell \Delta_f}{\epsilon^{\frac{4}{3}} \sigma^{\frac{2}{3}}} + \frac{\rho^2 \Delta_f}{\delta^3}\right)\left(\frac{\sigma^2}{\epsilon^2} + \frac{\ell^2}{\delta^2}\right) + d\frac{\ell \Delta_f}{\epsilon^2} \frac{\ell^2}{\delta^2}\right)
\end{aligned}
$$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \square$

E.3 One Epoch Analysis of ZO-SCSG (Option II)

From Algorithm 6, we know that all randomness in epoch j, iteration k come from three part: 1) random selection of $\mathcal{I}_{k-1}^j$ in Line 7; 2) random direction of u in estimating the gradient in Line 9; 3) random generation of N_j.

Lemma 20. *Under Assumption 1,*

$$
\mathbb{E}_u \mathbb{E}_{\mathcal{I}_k^j} \|v_k^j\|^2 \leq \frac{3d\ell^2}{b} \|x_k^j - x_0^j\|^2 + \frac{3\ell^2 d^2 \mu^2}{2b} + 2\|\nabla f_\mu(x_k^j)\|^2 + 2\|\hat{e}_j\|^2
$$

Proof. Define the following notation,

$$
\hat{e}_j = v_j - \nabla f_\mu(x_0^j)
$$

Then we have

$$
\mathbb{E}_u \mathbb{E}_{\mathcal{I}_k^j} v_k^j = \mathbb{E}_u \mathbb{E}_{\mathcal{I}_k^j} \left(\hat{\nabla}_{rand} f_{\mathcal{I}_k^j}(x_k^j) - \hat{\nabla}_{rand} f_{\mathcal{I}_k^j}(x_0^j) + v_j\right) = \nabla f_\mu(x_k^j) + \hat{e}_j
$$

Using the fact that $\mathbb{E}\|a\|^2 = \mathbb{E}\|a - \mathbb{E}a\|^2 + \|\mathbb{E}a\|^2$, we have

$$
\mathbb{E}_u \mathbb{E}_{\mathcal{I}_k^j} \|v_k^j\|^2 = \mathbb{E}_u \mathbb{E}_{\mathcal{I}_k^j} \|v_k^j - \mathbb{E}_u \mathbb{E}_{\mathcal{I}_k^j} v_k^j\|^2 + \|\mathbb{E}_u \mathbb{E}_{\mathcal{I}_k^j} v_k^j\|^2
$$

36

$$=\mathbb{E}_u\mathbb{E}_{\mathcal{I}_k^j}\|\hat{\nabla}_{rand}f_{\mathcal{I}_k^j}(x_k^j) - \hat{\nabla}_{rand}f_{\mathcal{I}_k^j}(x_0^j) - \left(\nabla f_\mu(x_k^j) - \nabla f_\mu(x_0^j)\right)\|^2 + \|\nabla f_\mu(x_k^j) + \hat{e}_j\|^2$$

$$\leq\mathbb{E}_u\mathbb{E}_{\mathcal{I}_k^j}\|\hat{\nabla}_{rand}f_{\mathcal{I}_k^j}(x_k^j) - \hat{\nabla}_{rand}f_{\mathcal{I}_k^j}(x_0^j) - \left(\nabla f_\mu(x_k^j) - \nabla f_\mu(x_0^j)\right)\|^2 + 2\|\nabla f_\mu(x_k^j)\|^2 + 2\|\hat{e}_j\|^2.$$

By Lemma 5, we have

$$\mathbb{E}_u\mathbb{E}_{\mathcal{I}_k^j}\|\hat{\nabla}_{rand}f_{\mathcal{I}_k^j}(x_k^j) - \hat{\nabla}_{rand}f_{\mathcal{I}_k^j}(x_0^j) - \left(\nabla f_\mu(x_k^j) - \nabla f_\mu(x_0^j)\right)\|^2$$

$$=\frac{1}{b^2}\mathbb{E}_u\mathbb{E}_{\mathcal{I}_k^j}\|\sum_{i\in\mathcal{I}_k^j}\left(\hat{\nabla}_{rand}f_i(x_k^j) - \hat{\nabla}_{rand}f_i(x_0^j) - \left(\nabla f_\mu(x_k^j) - \nabla f_\mu(x_0^j)\right)\right)\|^2$$

$$=\frac{1}{b^2}\mathbb{E}_u\mathbb{E}_{\mathcal{I}_k^j}\sum_{i\in\mathcal{I}_k^j}\|\hat{\nabla}_{rand}f_i(x_k^j) - \hat{\nabla}_{rand}f_i(x_0^j) - \left(\nabla f_\mu(x_k^j) - \nabla f_\mu(x_0^j)\right)\|^2$$

$$+\frac{1}{b^2}\mathbb{E}_u\mathbb{E}_{\mathcal{I}_k^j}\sum_{i\neq j}\left\langle \hat{\nabla}_{rand}f_i(x_k^j) - \hat{\nabla}_{rand}f_i(x_0^j) - \left(\nabla f_\mu(x_k^j) - \nabla f_\mu(x_0^j)\right),\right.$$

$$\left.\hat{\nabla}_{rand}f_j(x_k^j) - \hat{\nabla}_{rand}f_j(x_0^j) - \left(\nabla f_\mu(x_k^j) - \nabla f_\mu(x_0^j)\right)\right\rangle$$

$$=\frac{1}{b^2}\mathbb{E}_u\mathbb{E}_{\mathcal{I}_k^j}\sum_{i\in\mathcal{I}_k^j}\|\hat{\nabla}_{rand}f_i(x_k^j) - \hat{\nabla}_{rand}f_i(x_0^j) - \left(\nabla f_\mu(x_k^j) - \nabla f_\mu(x_0^j)\right)\|^2$$

$$=\frac{1}{b}\mathbb{E}_u\frac{1}{n}\sum_{i\in[n]}\|\hat{\nabla}_{rand}f_i(x_k^j) - \hat{\nabla}_{rand}f_i(x_0^j) - \left(\nabla f_\mu(x_k^j) - \nabla f_\mu(x_0^j)\right)\|^2$$

$$=\frac{1}{b}\cdot\mathbb{E}_u\left(\frac{1}{n}\sum_{i\in[n]}\|\hat{\nabla}_{rand}f_i(x_k^j) - \hat{\nabla}_{rand}f_i(x_0^j)\|^2\right) - \frac{1}{b}\|\nabla f_\mu(x_k^j) - \nabla f_\mu(x_0^j)\|^2$$

$$\leq\frac{1}{b}\cdot\mathbb{E}_u\left(\frac{1}{n}\sum_{i\in[n]}\|\hat{\nabla}_{rand}f_i(x_k^j) - \hat{\nabla}_{rand}f_i(x_0^j)\|^2\right)$$

$$\overset{\text{Lemma 8}}{\leq}\frac{3d\ell^2}{b}\|x_k^j - x_0^j\|^2 + \frac{3\ell^2 d^2\mu^2}{2b}$$

Therefore

$$\mathbb{E}_u\mathbb{E}_{\mathcal{I}_k^j}\|v_k^j\|^2 \leq \frac{3d\ell^2}{b}\|x_k^j - x_0^j\|^2 + \frac{3\ell^2 d^2\mu^2}{2b} + 2\|\nabla f_\mu(x_k^j)\|^2 + 2\|\hat{e}_j\|^2$$

$\square$

Lemma 21. *Suppose $\eta\ell < 1$, then under Assumption 1,*

$$\eta(1-\ell\eta)B\mathbb{E}\|\nabla f_\mu(\tilde{x}_j)\|^2 + \eta B\mathbb{E}\left\langle\hat{e}_j, \nabla f_\mu(\tilde{x}_j)\right\rangle$$

$$\leq b/d\mathbb{E}\left(f_\mu(\tilde{x}_{j-1}) - f_\mu(\tilde{x}_j)\right) + \frac{3d\ell^3\eta^2 B}{2b}\mathbb{E}\|\tilde{x}_j - \tilde{x}_{j-1}\|^2 + \frac{3\ell^3\eta^2 d^2\mu^2 B}{4b} + \ell\eta^2 B\mathbb{E}\|\hat{e}_j\|^2$$

Proof. By Lemma 4, we have

$$f_\mu(x_{k+1}^j) \leq f_\mu(x_k^j) - \left\langle x_{k+1}^j - x_k^j, \nabla f(x_k^j)\right\rangle + \frac{\ell}{2}\|x_{k+1}^j - x_k^j\|^2 \leq f_\mu(x_k^j) - \eta\left\langle v_k^j, \nabla f_\mu(x_k^j)\right\rangle + \frac{\ell\eta^2}{2}\|v_k^j\|^2$$

Taking expectation over the above inequality we have

$$f_\mu(x_{k+1}^j) \leq f_\mu(x_k^j) - \eta\left\langle\nabla f_\mu(x_k^j) + \hat{e}_j, \nabla f_\mu(x_k^j)\right\rangle + \frac{\ell\eta^2}{2}\mathbb{E}_u\mathbb{E}_{\mathcal{I}_k^j}\|v_k^j\|^2$$

$$=f_\mu(x_k^j) - \eta\|\nabla f_\mu(x_k^j)\|^2 - \eta\left\langle\hat{e}_j, \nabla f_\mu(x_k^j)\right\rangle + \frac{\ell\eta^2}{2}\mathbb{E}_u\mathbb{E}_{\mathcal{I}_k^j}\|v_k^j\|^2$$

$$\leq f_\mu(x_k^j) - \eta(1-\ell\eta)\|\nabla f_\mu(x_k^j)\|^2 - \eta\left\langle\hat{e}_j, \nabla f_\mu(x_k^j)\right\rangle$$

37

$$+ \frac{3d\ell^3\eta^2}{2b}\|x_k^j - x_0^j\|^2 + \frac{3\ell^3\eta^2 d^2\mu^2}{4b} + \ell\eta^2\|\hat{e}_j\|^2$$

Let $\mathbb{E}_j$ denote the expectation over $\mathcal{I}_0^k, \mathcal{I}_1^k, \ldots$, given N_j. Since $\mathcal{I}_{k+1}^j, \mathcal{I}_{k+2}^j, \ldots$ are independent of x_k^j, the above inequality implies that

$$\eta(1 - \ell\eta)\mathbb{E}_j\|\nabla f_\mu(x_k^j)\|^2 + \eta\mathbb{E}_j\left\langle \hat{e}_j, \nabla f_\mu(x_k^j)\right\rangle$$

$$\leq \mathbb{E}_j f_\mu(x_k^j) - \mathbb{E}_j f_\mu(x_{k+1}^j) + \frac{3d\ell^3\eta^2}{2b}\mathbb{E}_j\|x_k^j - x_0^j\|^2 + \frac{3\ell^3\eta^2 d^2\mu^2}{4b} + \ell\eta^2\|\hat{e}_j\|^2$$

Let $k = N_j$, by taking expectation to N_j and using Fubini's theorem, we have

$$\eta(1 - \ell\eta)\mathbb{E}_{N_j}\mathbb{E}_j\|\nabla f_\mu(x_{N_j}^j)\|^2 + \eta\mathbb{E}_{N_j}\mathbb{E}_j\left\langle \hat{e}_j, \nabla f_\mu(x_{N_j}^j)\right\rangle$$

$$\leq \mathbb{E}_{N_j}\left(\mathbb{E}_j f_\mu(x_{N_j}^j) - \mathbb{E}_j f_\mu(x_{N_j+1}^j)\right) + \frac{3d\ell^3\eta^2}{2b}\mathbb{E}_{N_j}\mathbb{E}_j\|x_{N_j}^j - x_0^j\|^2 + \frac{3\ell^3\eta^2 d^2\mu^2}{4b} + \ell\eta^2\|\hat{e}_j\|^2$$

$$= \frac{b/d}{B}\mathbb{E}_{N_j}\left(f_\mu(x_0^j) - f_\mu(x_{N_j}^j)\right) + \frac{3d\ell^3\eta^2}{2b}\mathbb{E}_j\mathbb{E}_{N_j}\|x_{N_j}^j - x_0^j\|^2 + \frac{3\ell^3\eta^2 d^2\mu^2}{4b} + \ell\eta^2\|\hat{e}_j\|^2$$

then we have

$$\eta(1 - \ell\eta)B\mathbb{E}\|\nabla f_\mu(\tilde{x}_j)\|^2 + \eta B\mathbb{E}\left\langle \hat{e}_j, \nabla f_\mu(\tilde{x}_j)\right\rangle$$

$$\leq b/d\,\mathbb{E}\left(f_\mu(\tilde{x}_{j-1}) - f_\mu(\tilde{x}_j)\right) + \frac{3d\ell^3\eta^2 B}{2b}\mathbb{E}\|\tilde{x}_j - \tilde{x}_{j-1}\|^2 + \frac{3\ell^3\eta^2 d^2\mu^2 B}{4b} + \ell\eta^2 B\mathbb{E}\|\hat{e}_j\|^2$$

$\square$

Lemma 22. *Suppose* $3d^2\ell^2\eta^2 B < b^2$, *then under Assumption 1,*

$$\left(\frac{b/d}{B} - \frac{3d\ell^2\eta^2}{b}\right)\mathbb{E}\|\tilde{x}_j - \tilde{x}_{j-1}\|^2 + 2\eta\mathbb{E}\left\langle \hat{e}_j, \tilde{x}_j - \tilde{x}_{j-1}\right\rangle$$

$$\leq -2\eta\mathbb{E}\left\langle \nabla f_\mu(\tilde{x}_j), \tilde{x}_j - \tilde{x}_{j-1}\right\rangle + \frac{3\ell^2\eta^2 d^2\mu^2}{2b} + 2\eta^2\mathbb{E}\|\nabla f_\mu(\tilde{x}_j)\|^2 + 2\eta^2\mathbb{E}\|\hat{e}_j\|^2$$

Proof. Since $x_{k+1}^j = x_k^j - \eta v_k^j$, we have

$$\mathbb{E}_u\mathbb{E}_{\mathcal{I}_k^j}\|x_{k+1}^j - x_0^j\|^2$$

$$= \|x_k^j - x_0^j\|^2 - 2\eta\left\langle \mathbb{E}_u\mathbb{E}_{\mathcal{I}_k^j}v_k^j, x_k^j - x_0^j\right\rangle + \eta^2\mathbb{E}_u\mathbb{E}_{\mathcal{I}_k^j}\|v_k^j\|^2$$

$$= \|x_k^j - x_0^j\|^2 - 2\eta\left\langle \nabla f_\mu(x_k^j), x_k^j - x_0^j\right\rangle - 2\eta\left\langle \hat{e}_j, x_k^j - x_0^j\right\rangle + \eta^2\mathbb{E}_u\mathbb{E}_{\mathcal{I}_k^j}\|v_k^j\|^2$$

$$\leq \left(1 + \frac{3d\ell^2\eta^2}{b}\right)\|x_k^j - x_0^j\|^2 - 2\eta\left\langle \nabla f_\mu(x_k^j), x_k^j - x_0^j\right\rangle - 2\eta\left\langle \hat{e}_j, x_k^j - x_0^j\right\rangle + \frac{3\ell^2\eta^2 d^2\mu^2}{2b}$$

$$+ 2\eta^2\|\nabla f_\mu(x_k^j)\|^2 + 2\eta^2\|\hat{e}_j\|^2$$

Using the notation $\mathbb{E}_j$ we have

$$2\eta\mathbb{E}_j\left\langle \nabla f_\mu(x_k^j), x_k^j - x_0^j\right\rangle + 2\eta\mathbb{E}_j\left\langle \hat{e}_j, x_k^j - x_0^j\right\rangle$$

$$\leq \left(1 + \frac{3d\ell^2\eta^2}{b}\right)\mathbb{E}_j\|x_k^j - x_0^j\|^2 - \mathbb{E}_j\|x_{k+1}^j - x_0^j\|^2 + \frac{3\ell^2\eta^2 d^2\mu^2}{2b} + 2\eta^2\|\nabla f_\mu(x_k^j)\|^2 + 2\eta^2\|\hat{e}_j\|^2$$

Let $k = N_j$, by taking expectation to N_j and using Fubini's theorem, we have

$$2\eta\mathbb{E}_{N_j}\mathbb{E}_j\left\langle \nabla f_\mu(x_{N_j}^j), x_{N_j}^j - x_0^j\right\rangle + 2\eta\mathbb{E}_{N_j}\mathbb{E}_j\left\langle \hat{e}_j, x_{N_j}^j - x_0^j\right\rangle$$

$$\leq \left(1 + \frac{3d\ell^2\eta^2}{b}\right)\mathbb{E}_{N_j}\mathbb{E}_j\|x_{N_j}^j - x_0^j\|^2 - \mathbb{E}_{N_j}\mathbb{E}_j\|x_{N_j+1}^j - x_0^j\|^2 + \frac{3\ell^2\eta^2 d^2\mu^2}{2b}$$

$$+ 2\eta^2\mathbb{E}_{N_j}\|\nabla f_\mu(x_{N_j}^j)\|^2 + 2\eta^2\|\hat{e}_j\|^2$$

38

$$= \left(-\frac{b/d}{B} + \frac{3d\ell^2\eta^2}{b} \right) \mathbb{E}_{N_j}\mathbb{E}_j \|x_{N_j}^j - x_0^j\|^2 + \frac{3\ell^2\eta^2 d^2\mu^2}{2b} + 2\eta^2\mathbb{E}_{N_j}\|\nabla f_\mu(x_{N_j}^j)\|^2 + 2\eta^2\|\hat{e}_j\|^2$$

Substituting $x_{N_j}^j, x_0^j$ by $\tilde{x}_j, \tilde{x}_{j-1}$ and take a further expectation to the past randomness, we get

$$\left(\frac{b/d}{B} - \frac{3d\ell^2\eta^2}{b} \right) \mathbb{E}\|\tilde{x}_j - \tilde{x}_{j-1}\|^2 + 2\eta\mathbb{E}\langle \hat{e}_j, \tilde{x}_j - \tilde{x}_{j-1} \rangle$$
$$\leq -2\eta\mathbb{E}\langle \nabla f_\mu(\tilde{x}_j), \tilde{x}_j - \tilde{x}_{j-1} \rangle + \frac{3\ell^2\eta^2 d^2\mu^2}{2b} + 2\eta^2\mathbb{E}\|\nabla f_\mu(\tilde{x}_j)\|^2 + 2\eta^2\mathbb{E}\|\hat{e}_j\|^2$$

$\square$

Lemma 23.
$$\frac{b/d}{B}\mathbb{E}\langle \hat{e}_j, \tilde{x}_j - \tilde{x}_{j-1} \rangle = -\eta\mathbb{E}\langle \hat{e}_j, \nabla f_\mu(\tilde{x}_j) \rangle - \eta\mathbb{E}\|\hat{e}_j\|^2$$

Proof. Let $M_k^j = \left\langle \hat{e}_j, x_k^j - x_0^j \right\rangle$. Then we have

$$\mathbb{E}_{N_j}\langle \hat{e}_j, \tilde{x}_j - \tilde{x}_{j-1} \rangle = \mathbb{E}_{N_j} M_{N_j}^j$$

Since N_j is independent of $x_0^j, \hat{e}_j$, we have

$$\mathbb{E}\langle \hat{e}_j, \tilde{x}_j - \tilde{x}_{j-1} \rangle = \mathbb{E} M_{N_j}^j$$

Also we have $M_0^j = 0$. On the other hand,

$$\mathbb{E}_u\mathbb{E}_{\mathcal{I}_k^j}\left(M_{k+1}^j - M_k^j \right) = \mathbb{E}_u\mathbb{E}_{\mathcal{I}_k^j}\left\langle \hat{e}_j, x_{k+1}^j - x_k^j \right\rangle = -\eta\left\langle \hat{e}_j, \mathbb{E}_u\mathbb{E}_{\mathcal{I}_k^j} v_k^j \right\rangle$$
$$= -\eta\left\langle \hat{e}_j, \nabla f_\mu(x_k^j) \right\rangle - \eta\|\hat{e}_j\|^2$$

Using the notation E_j, we have

$$\mathbb{E}_j\left(M_{k+1}^j - M_k^j \right) = -\eta\left\langle \hat{e}_j, \mathbb{E}_j \nabla f_\mu(x_k^j) \right\rangle - \eta\|\hat{e}_j\|^2$$

Let $k = N_j$, by taking the expectation with respect to N_j and using Fubini's theorem, we have

$$\frac{b/d}{B}\mathbb{E}_{N_j} M_{N_j}^j = -\eta\left\langle \hat{e}_j, \mathbb{E}_{N_j}\mathbb{E}_j \nabla f_\mu(x_{N_j}^j) \right\rangle - \eta\|\hat{e}_j\|^2$$

Substituting $x_{N_j}^j, x_0^j$ by $\tilde{x}_j, \tilde{x}_{j-1}$ and take a further expectation to the past randomness, we get

$$\frac{b/d}{B}\mathbb{E}\langle \hat{e}_j, \tilde{x}_j - \tilde{x}_{j-1} \rangle = -\eta\mathbb{E}\langle \hat{e}_j, \nabla f_\mu(\tilde{x}_j) \rangle - \eta\mathbb{E}\|\hat{e}_j\|^2$$

$\square$

Proof of Lemma 14 (Option II). Multiplying Lemma 21 by 2, Lemma 22 by $\frac{b/d}{\eta}$ and summing them up, we have

$$2\eta B(1 - \eta\ell - \frac{b/d}{B})\mathbb{E}\|\nabla f_\mu(\tilde{x}_j)\|^2 + \frac{b^3/d^2 - 3\ell^2\eta^2 bB - 3d\ell^3\eta^3 B^2}{\eta bB}\mathbb{E}\|\tilde{x}_j - \tilde{x}_{j-1}\|^2$$
$$+ 2\eta B\mathbb{E}\langle \hat{e}_j, \nabla f_\mu(\tilde{x}_j) \rangle + 2b/d\langle \hat{e}_j, \tilde{x}_j - \tilde{x}_{j-1} \rangle$$
$$\leq -2b/d\mathbb{E}\langle \nabla f_\mu(\tilde{x}_j), \tilde{x}_j - \tilde{x}_{j-1} \rangle + 2b/d\mathbb{E}\left(f_\mu(\tilde{x}_{j-1}) - f_\mu(\tilde{x}_j) \right) + \frac{3\ell^2\eta^2 d^2\mu^2}{2b}\left(B\ell + \frac{b/d}{\eta} \right)$$
$$+ (2\ell\eta^2 B + 2\eta b/d)\mathbb{E}\|\hat{e}_j\|^2$$

By Lemma 23,

$$2\eta B\mathbb{E}\langle \hat{e}_j, \nabla f_\mu(\tilde{x}_j) \rangle + 2b/d\mathbb{E}\langle \hat{e}_j, \tilde{x}_j - \tilde{x}_{j-1} \rangle = -2\eta B\mathbb{E}\|\hat{e}_j\|^2$$

39

So the above inequality can be simplified as

$$2\eta B(1 - \eta\ell - \frac{b/d}{B})\mathbb{E}\|\nabla f_\mu(\tilde{x}_j)\|^2 + \frac{b^3/d^2 - 3\ell^2\eta^2 bB - 3d\ell^3\eta^3 B^2}{\eta bB}\mathbb{E}\|\tilde{x}_j - \tilde{x}_{j-1}\|^2$$

$$\leq -2b/d\mathbb{E}\langle\nabla f_\mu(\tilde{x}_j), \tilde{x}_j - \tilde{x}_{j-1}\rangle + 2b/d\mathbb{E}(f_\mu(\tilde{x}_{j-1}) - f_\mu(\tilde{x}_j)) + \frac{3\ell^2\eta^2 d^2\mu^2}{2b}\left(B\ell + \frac{b/d}{\eta}\right)$$

$$+ (2\ell\eta^2 B + 2\eta b/d + 2\eta B)\mathbb{E}\|\hat{e}_j\|^2$$

Using the fact that $2\langle a, b\rangle \leq \beta\|a\|^2 + \frac{1}{\beta}\|b\|^2$ for any $\beta > 0$, we have

$$-2b/d\mathbb{E}\langle\nabla f_\mu(\tilde{x}_j), \tilde{x}_j - \tilde{x}_{j-1}\rangle$$

$$\leq \frac{\eta bB}{b^3/d^2 - 3\ell^2\eta^2 bB - 3d\ell^3\eta^3 B^2}(\frac{b}{d})^2\mathbb{E}\|\nabla f_\mu(\tilde{x}_j)\|^2 + \frac{b^3/d^2 - 3\ell^2\eta^2 bB - 3d\ell^3\eta^3 B^2}{\eta bB}\mathbb{E}\|\tilde{x}_j - \tilde{x}_{j-1}\|^2$$

Then we conclude that

$$\frac{\eta\ell B}{b/d}\left(2 - 2\eta\ell - 2\frac{b/d}{B} - \frac{b^3/d^2}{b^3/d^2 - 3\ell^2\eta^2 bB - 3d\ell^3\eta^3 B^2}\right)\mathbb{E}\|\nabla f_\mu(\tilde{x}_j)\|^2$$

$$\leq 2\ell\mathbb{E}(f_\mu(\tilde{x}_{j-1}) - f_\mu(\tilde{x}_j)) + \frac{3\ell^2\eta^2 d^2\mu^2}{2b}\left(\frac{B\ell^2}{b/d} + \frac{\ell}{\eta}\right) + \frac{2\eta\ell B}{b/d}\left(1 + \eta\ell + \frac{b/d}{B}\right)\mathbb{E}\|\hat{e}_j\|^2$$

Let $\eta\ell = \gamma\left(\frac{b/d}{B}\right)^{\frac{2}{3}}$ and $b/d \geq 1, \frac{B}{b/d} \geq 8 \geq \frac{8}{b/d}$,

$$b^3/d^2 - 3\ell^2\eta^2 bB - 3d\ell^3\eta^3 B^2 = b^3/d^2\left(1 - 3\gamma^2(b/d)^{-1}(\frac{b/d}{B})^{\frac{1}{3}} - 3\gamma^3(b/d)^{-1}\right)$$

$$\geq b^3/d^2\left(1 - \frac{3}{2}\gamma^2 - 3\gamma^3\right)$$

Then the above inequality can be simplified as

$$\gamma\left(\frac{B}{b/d}\right)^{\frac{1}{3}}\left(2 - 2\gamma\left(\frac{b/d}{B}\right)^{\frac{2}{3}} - 2\frac{b/d}{B} - \frac{1}{1 - \frac{3}{2}\gamma^2 - 3\gamma^3}\right)\mathbb{E}\|\nabla f_\mu(\tilde{x}_j)\|^2$$

$$\leq 2\ell\mathbb{E}(f_\mu(\tilde{x}_{j-1}) - f_\mu(\tilde{x}_j)) + \frac{3\ell^2 d^2\mu^2}{2b}\gamma\left((\frac{b/d}{B})^{\frac{1}{3}}\gamma + (\frac{b/d}{B})^{\frac{2}{3}}\right)$$

$$+ 2\gamma\left(\frac{B}{b/d}\right)^{\frac{1}{3}}(1 + \gamma(\frac{b/d}{B})^{\frac{2}{3}} + \frac{b/d}{B})\mathbb{E}\|\hat{e}_j\|^2$$

Since $\frac{B}{b/d} \geq 8, \gamma \leq \frac{1}{8}$, we have

$$2 - 2\gamma\left(\frac{b/d}{B}\right)^{\frac{2}{3}} - 2\frac{b/d}{B} - \frac{1}{1 - \frac{3}{2}\gamma^2 - 3\gamma^3} \geq 2 - \frac{\gamma}{2} - \frac{1}{4} - \frac{1}{1 - \frac{3}{2}\gamma^2 - 3\gamma^3} \geq 0.65$$

$$(\frac{b/d}{B})^{\frac{1}{3}}\gamma + (\frac{b/d}{B})^{\frac{2}{3}} \leq \frac{\gamma}{2} + \frac{1}{4} \leq \frac{5}{16}$$

$$1 + \gamma(\frac{b/d}{B})^{\frac{2}{3}} + \frac{b/d}{B} \leq 1 + \frac{\gamma}{4} + \frac{1}{8} \leq \frac{37}{32}$$

Thus we have

$$\left(\frac{B}{b/d}\right)^{\frac{1}{3}}\mathbb{E}\|\nabla f_\mu(\tilde{x}_j)\|^2 \leq \frac{4\ell}{\gamma}\mathbb{E}(f_\mu(\tilde{x}_{j-1}) - f_\mu(\tilde{x}_j)) + \frac{\ell^2 d^2\mu^2}{b} + 4\left(\frac{B}{b/d}\right)^{\frac{1}{3}}\mathbb{E}\|\hat{e}_j\|^2$$

Using Lemma 19, we have

$$\mathbb{E}\|\hat{e}_j\|^2$$
$$= \mathbb{E}\|v_j - \nabla f_\mu(x_0^j)\|^2 = \mathbb{E}\|\hat{\nabla}_{coord}f_{\mathcal{I}_j}(x_0^j)\|^2$$
$$\leq 3\mathbb{E}\|\hat{\nabla}_{coord}f_{\mathcal{I}_j}(x_0^j) - \hat{\nabla}_{coord}f(x_0^j)\|^2 + 3\mathbb{E}\|\hat{\nabla}_{coord}f(x_0^j) - \nabla f(x)\|^2 + 3\mathbb{E}\|\nabla f(x_0^j) - \nabla f_\mu(x_0^j)\|^2$$

40

$$\leq \frac{9(2\ell^2 d\mu^2 + \sigma^2)}{B} + 3\ell^2 d\mu^2 + \frac{3\ell^2 d^2 \mu^2}{4}$$

Using Lemma 8, we have

$$\mathbb{E}\left(f_\mu(\tilde{x}_{j-1}) - f_\mu(\tilde{x}_j)\right) \leq \mathbb{E}\left(f(\tilde{x}_{j-1}) - f(\tilde{x}_j)\right) + \ell\mu^2$$

$$\mathbb{E}\|\nabla f_\mu(\tilde{x}_j)\|^2 \geq \frac{1}{2}\mathbb{E}\|\nabla f(\tilde{x}_j)\|^2 - \frac{1}{2}\mathbb{E}\|\nabla f(\tilde{x}_j) - \nabla f_\mu(\tilde{x}_j)\|^2 \geq \frac{1}{2}\mathbb{E}\|\nabla f(\tilde{x}_j)\|^2 - \frac{\ell^2 d^2 \mu^2}{8}$$

Thus we obtain

$$\left(\frac{B}{b/d}\right)^{\frac{1}{3}}\left(\frac{1}{2}\mathbb{E}\|\nabla f(\tilde{x}_j)\|^2 - \frac{\ell^2 d^2 \mu^2}{8}\right)$$

$$\leq \frac{4\ell}{\gamma}\left(\mathbb{E}\left(f(\tilde{x}_{j-1}) - f(\tilde{x}_j)\right) + \ell\mu^2\right) + \frac{\ell^2 d^2 \mu^2}{b} + 4\left(\frac{B}{b/d}\right)^{\frac{1}{3}}\left(\frac{9(2\ell^2 d\mu^2 + \sigma^2)}{B} + 3\ell^2 d\mu^2 + \frac{3\ell^2 d^2 \mu^2}{4}\right)$$

$$\leq \frac{4\ell}{\gamma}\left(\mathbb{E}\left(f(\tilde{x}_{j-1}) - f(\tilde{x}_j)\right) + \ell\mu^2\right) + \frac{\ell^2 d^2 \mu^2}{b} + 4\left(\frac{B}{b/d}\right)^{\frac{1}{3}}\left(\frac{9(2\ell^2 d\mu^2 + \sigma^2)}{B} + 4\ell^2 d^2 \mu^2\right)$$

Finally we get

$$\left(\frac{B}{b/d}\right)^{\frac{1}{3}}\mathbb{E}\|\nabla f(\tilde{x}_j)\|^2 \leq \frac{8\ell}{\gamma}\mathbb{E}\left(f(\tilde{x}_{j-1}) - f(\tilde{x}_j)\right) + \frac{72\sigma^2}{(b/d)^{\frac{1}{3}}B^{\frac{2}{3}}} + c\left(\frac{B}{b/d}\right)^{\frac{1}{3}}\ell^2 d^2 \mu^2$$

where c is a sufficient large constant. Telescope the sum in $j = 1, \ldots, T$, and using the definition of $\tilde{x}_T^*$, we finally get

$$\mathbb{E}\|\tilde{x}_T^*\|^2 \leq \frac{\frac{8\ell}{\gamma}}{T\left(\frac{B}{b/d}\right)^{\frac{1}{3}}} \cdot \mathbb{E}\left(f(\tilde{x}_0) - f(\tilde{x}_T)\right) + \frac{72\sigma^2}{B} + c\ell^2 d^2 \mu^2$$

$\square$

E.4 Proof of Second-Order Stationary Point (Option II)

Proof of Theorem 5. Let N_1 and N_2 be the number of times we reach Line 7 and 9 of Algorithm 7. From Lemma 14 of ZO-SCSG we know that for one epoch with size $B = \max\{1, \frac{1152\sigma^2}{\epsilon^2}\}$, mini-batch size $b \geq 1$ and the smoothing parameter $\mu = \frac{\epsilon}{4\sqrt{c}\ell d}$, we have

$$\mathbb{E}\|x_{t+1}\|^2 \leq \frac{8\ell}{\gamma}\left(\frac{b/d}{B}\right)^{\frac{1}{3}}\mathbb{E}\left(f(x_t) - f(x_{t+1})\right) + \frac{\epsilon^2}{8}$$

Then, if $\|\nabla f(x_{t+1})\| \geq \frac{\epsilon}{2}$, we have $x_{t+1} = x_{t+1}$; if $v = \perp$, we set $x_{t+1} = x_{t+1}$ for if $v \neq \perp$, we have $f(x_{t+1}) - \mathbb{E}f(x_{t+1}) \geq \frac{\delta^3}{12\rho^2}$ (here the expectation is taken on the randomness of sign of v). Thus we have

$$\frac{\gamma B^{\frac{1}{3}}}{8\ell(b/d)^{\frac{1}{3}}}\mathbb{E}\left[\sum_{t=0}^{K-1}\left(\|\nabla f(x_{t+1})\|^2 - \frac{\epsilon^2}{8}\right)\right] + \frac{\delta^3}{12\rho^2}\mathbb{E}[N_2] \leq \Delta_f$$

On one hand, since we have chosen K such that $K \geq \Omega\left(\frac{\ell(b/d)^{\frac{1}{3}}\Delta_f}{\epsilon^2 B^{\frac{1}{3}}}\right) = \Omega\left(\frac{\ell(b/d)^{\frac{1}{3}}\Delta_f}{\epsilon^2(1+\frac{\sigma^2}{\epsilon^2})^{\frac{1}{3}}}\right)$, then by Markov's inequality, with probability at least $\frac{5}{6}$, it satisfies $\sum_{t=0}^{K-1}\|\nabla f(x_{t+1})\|^2 \leq \frac{\epsilon}{4}K$. As a sequence, at least half of the indices $t = 0, \ldots, K-1$ will satisfy $\|\nabla f(x_{t+1})\| \leq \frac{\epsilon}{2}$, which means that $N_1 \geq \frac{K}{2}$.

On the other hand, we have $\frac{\delta^3}{12\rho^2}\mathbb{E}[N_2] \leq \Delta_f + \frac{K\gamma B^{\frac{1}{3}}\epsilon^2}{64\ell(b/d)^{\frac{1}{3}}}$. Since $K \geq \Omega\left(\frac{\ell(b/d)^{\frac{1}{3}}\Delta_f}{\epsilon^2 B^{\frac{1}{3}}}\right) = \Omega\left(\frac{\ell(b/d)^{\frac{1}{3}}\Delta_f}{\epsilon^2(1+\frac{\sigma^2}{\epsilon^2})^{\frac{1}{3}}}\right)$, we have $\mathbb{E}[N_2] \leq \frac{K\gamma B^{\frac{1}{3}}\epsilon^2\rho^2}{\ell\delta^3(b/d)^{\frac{1}{3}}}$. As long as $B \leq \mathcal{O}\left(\frac{\ell^3\delta^9 b/d}{\epsilon^6\rho^6}\right)$, or equivalently $b/d \geq \Omega\left(\frac{B\epsilon^6\rho^6}{\delta^9\ell^3}\right)$, we have $\mathbb{E}[N_2] \leq \frac{K}{12}$. Therefore, with provability at least $\frac{5}{6}$, it satisfies $N_2 \leq \frac{k}{2}$.

41

Since $N_1 \geq N_2$, this means with probability at least $\frac{2}{3}$ the algorithm must terminate and output some x_{t+1} in an iteration.

Finally, the per-iteration complexity of Algorithm 7 is dominated by $\tilde{\mathcal{O}}(d \cdot B + \frac{B}{b/d} \cdot b) = \tilde{\mathcal{O}}(dB)$ function queries for ZO-SCSG and $\tilde{\mathcal{O}}(dB)$ function queries for estimating $\|\nabla f(x_{t+1})\|$, as well as $\tilde{\mathcal{O}}(\frac{\ell^2}{\delta^2})$ invoking Algorithm 2. Thus the total function query complexity is

$$
\begin{aligned}
\tilde{\mathcal{O}}\left(K\left(dB + d\frac{\ell^2}{\delta^2} \right) \right) &= \tilde{\mathcal{O}}\left(\left(\frac{\ell(b/d)^{\frac{1}{3}}\Delta_f}{\epsilon^2 B^{\frac{1}{3}}} \right) \left(dB + d\frac{\ell^2}{\delta^2} \right) \right) \\
&= \tilde{\mathcal{O}}\left(d\left(\frac{\ell \max\{1, \frac{B^{\frac{1}{3}}\epsilon^2\rho^2}{\delta^3\ell}\}\Delta_f}{\epsilon^2 B^{\frac{1}{3}}} \right) \left(B + \frac{\ell^2}{\delta^2} \right) \right) \\
&= \tilde{\mathcal{O}}\left(d\left(\frac{\ell(1 + \frac{B^{\frac{1}{3}}\epsilon^2\rho^2}{\delta^3\ell})\Delta_f}{\epsilon^2 B^{\frac{1}{3}}} \right) \left(B + \frac{\ell^2}{\delta^2} \right) \right) \\
&= \tilde{\mathcal{O}}\left(d\left(\frac{\ell\Delta_f}{\epsilon^2 B^{\frac{1}{3}}} + \frac{\rho^2\Delta_f}{\delta^3} \right) \left(B + \frac{\ell^2}{\delta^2} \right) \right) \\
&= \tilde{\mathcal{O}}\left(d\left(\frac{\ell\Delta_f}{\epsilon^2 \max\{1, \frac{\sigma^2}{\epsilon^2}\}^{\frac{1}{3}}} + \frac{\rho^2\Delta_f}{\delta^3} \right) \left(\max\{1, \frac{\sigma^2}{\epsilon^2}\} + \frac{\ell^2}{\delta^2} \right) \right) \\
&= \tilde{\mathcal{O}}\left(d\left(\frac{\ell\Delta_f}{\epsilon^{\frac{4}{3}}\sigma^{\frac{2}{3}}} + \frac{\rho^2\Delta_f}{\delta^3} \right) \left(\frac{\sigma^2}{\epsilon^2} + \frac{\ell^2}{\delta^2} \right) + d\frac{\ell\Delta_f}{\epsilon^2}\frac{\ell^2}{\delta^2} \right)
\end{aligned}
$$

$\square$

F Applying Zeroth-Order Negative Curvature Finding to ZO-SPIDER

In this section, we apply ZO-NCF-Online to ZO-SPIDER to turn it into a local minima finding algorithm and propose ZO-SPIDER-NCF in Algorithm 8. As a by-product, we also propose a zeroth-order variant of the SPIDER method in Appendix G that can converge to an ϵ-approximate FOSP with high probability rather than expectation. Using the same technique as in SPIDER-SFO$^+$ [16], that is, instead of moving in a large single step with size δ/ρ along the approximate negative curvature direction as in ZO-SGD-NCF and ZO-SCSG-NCF, we can split it into $\delta/(\rho\eta)$ equal length mini-steps with size η. As a result, we can maintain the SPIDER estimates and improve the so-called non-improvable coupling term $\frac{1}{\delta^3\epsilon^2}$ by a fact of δ.

Algorithm 8 ZO-SPIDER-NCF

Input: Function f, starting point x_0, $\epsilon > 0$ and $\delta > 0$.
1: **for** $j = 0$ to J **do**
2:      $w_1 \leftarrow$ ZO-NCF-Online $(f, x_k, \delta, \frac{1}{16J})$
3:      Randomly flip a sign, and set $w_2 = \pm\eta w_1$
4:      **for** k to $k + \mathcal{K}$ **do**
5:          **if** k mod q = 0 **then**
6:              Sample $\mathcal{S}_1$ from $[n]$ without replacement, $v_k = \hat{\nabla}_{coord} f_{\mathcal{S}_1}(x_k)$
7:          **else**
8:              Sample $\mathcal{S}_2$ from $[n]$ with replacement,
9:              $v_k = \hat{\nabla}_{coord} f_{\mathcal{S}_2}(x_k) - \hat{\nabla}_{coord} f_{\mathcal{S}_2}(x_{k-1}) + v_{k-1}$
10:          **if** $w_1 \neq \perp$ **then**
11:              $x_{k+1} = x_k - w_2$
12:          **else**
13:              **if** $\|v_k\| \leq 2\tilde{\epsilon}$ **then**
14:                  **return** x_k
15:              $x_{k+1} = x_k - \eta(v_k/\|v_k\|)$

42

Theorem 8. *Under Assumption 2, if we set $\tilde{\epsilon} = 10\epsilon \log(128(K_0 + 1))$, $|\mathcal{S}_1| = \frac{16\sigma^2}{\epsilon^2}$, $|\mathcal{S}_2| = \frac{16\sigma}{\epsilon n_0}$, $\eta = \frac{\epsilon}{\ell n_0}$, $q = \frac{\sigma n_0}{\epsilon}$, $\mu = \left(\frac{\epsilon \cdot \tilde{\epsilon}}{8q^2\rho^2 d}\right)^{\frac{1}{4}} = \tilde{O}(\frac{\epsilon}{d^{1/4}})$, $\mathscr{K} = \frac{\delta \ell n_0}{\rho \epsilon}$, where $n_0 \in [1, \frac{2\sigma}{\epsilon}]$, then with probability at least $\frac{3}{4}$, Algorithm 8 outputs x_k with $j \leq J = 8\left(\left\lfloor \max\left(\frac{12\rho^2\Delta_f}{\delta^3}, \frac{4\rho\Delta_f}{\delta\epsilon}\right)\right\rfloor + 1\right)$, $k \leq K_0 = J\mathscr{K}$ satisfying $\|\nabla f(x_k)\| \leq 3\tilde{\epsilon}$, $\lambda_{min}(\nabla^2 f(x_k)) \geq -2\delta$ with $\tilde{\epsilon} = 10\epsilon \log(128(K_0 + 1)) = \tilde{O}(\epsilon)$. The total function query complexity is bounded by*

$$\tilde{O}\left(d\left(\frac{\sigma\ell\Delta_f}{\epsilon^3} + \frac{\sigma\ell\rho\Delta_f}{\epsilon^2\delta^2} + \frac{\ell^2\rho\Delta_f}{\delta^3\epsilon} + \frac{\ell^2\rho^2\Delta_f}{\delta^5} + \frac{\sigma^2}{\epsilon^2} + \frac{\sigma\delta\ell}{\rho\epsilon^2} + \frac{\ell^2}{\delta^2}\right)\right).$$

Remark 7. *We can boost the confidence of Theorem 6 from $3/4$ to $1 - p$ by running $\log 1/p$ copies of ZO-SPIDER-NCF*

Algorithm 9 ZO-SPIDER-Coord (For convergence rates in high probability)

Input: function f, starting point x_0, $\epsilon > 0$.
1: **for** $k = 0$ to K **do**
2:     **if** k mod q = 0 **then**
3:         Sample $\mathcal{S}_1$ from $[n]$, $v_k = \hat{\nabla}_{coord} f_{\mathcal{S}_1}(x^k)$
4:     **else**
5:         Sample $\mathcal{S}_2$ from $[n]$ with replacement,
6:         $v_k = \hat{\nabla}_{coord} f_{\mathcal{S}_2}(x^k) - \hat{\nabla}_{coord} f_{\mathcal{S}_2}(x^{k-1}) + v_{k-1}$
7:     **if** $\|v_k\| \leq 2\tilde{\epsilon}$ **then**
8:         **return** x_k
9:     $x_{k+1} = x_k - \eta(v_k/\|v_k\|)$
Output: x_K ▷ this line is not reached with high probability

Theorem 9. *Under the settings of Algorithm 9, if we set*

$$\tilde{\epsilon} = 10\epsilon \log(4(K_0 + 1)/p), \quad |\mathcal{S}_1| = \frac{16\sigma^2}{\epsilon^2}, \quad |\mathcal{S}_2| = \frac{16\sigma}{\epsilon n_0}, \quad \eta = \frac{\epsilon}{\ell n_0}, \quad q = \frac{\sigma n_0}{\epsilon},$$

$$\mu = \left(\frac{\epsilon \cdot \tilde{\epsilon}}{8q^2\rho^2 d}\right)^{\frac{1}{4}} = \tilde{O}(\frac{\epsilon}{d^{1/4}})$$

where $n_0 \in [1, \frac{2\sigma}{\epsilon}]$. Then under Assumption 2, with probability at least $1 - p$, Algorithm 9 terminates before $K_0 = \left\lfloor \frac{4\ell n_0}{\epsilon^2}\right\rfloor + 2$ iterations and outputs $x_\mathcal{K}$ satisfying

$$\|v_\mathcal{K}\| \leq 2\tilde{\epsilon}, \quad \|\nabla f(x_\mathcal{K})\| \leq 3\tilde{\epsilon}.$$

The function query complexity is $O\left(d\left(\frac{128\ell\sigma\Delta_f}{\epsilon^3} + \frac{16\sigma^2}{\epsilon^2} + \frac{32\sigma}{\epsilon n_0}\right)\right)$.

Lemma 24 (Proposition 2 in [16]). *Let $\epsilon_{1:K}$ be a vector-valued martingale difference sequence with respect to $\mathcal{F}_k$, i.e., for each $k = 1, \ldots, K$, $\mathbb{E}[\epsilon_k|\mathcal{F}_{k-1}] = 0$ and $\|\epsilon_k\|^2 \leq B_k^2$. We have*

$$Pr\left(\left\|\sum_{k=1}^K \epsilon_k \geq \lambda\right\|\right) \leq \exp\left(-\frac{\lambda^2}{4\sum_{k=1}^K B_k^2}\right),$$

where λ is an arbitrary real positive value.

F.1 Proof of high probability results for First-Order Stationary Point of Theorem 9

Define $\mathcal{K}$ to be the time when Algorithm 9 stops. We have $\mathcal{K} = 0$ if $\|v_0\| \leq 2\epsilon$, and $\mathcal{K} = \inf\{k \geq 0 : \|v_k\| \leq 2\epsilon\} + 1$ if $\|v_0\| \geq 2\epsilon$.

Lemma 25. *Define the event,*

$$\mathcal{H}_{K_0} = \left(\|v_k - \nabla f(x_k)\|^2 \leq \epsilon \cdot \tilde{\epsilon}, \forall k \leq \min\{\mathcal{K}, K_0\}\right),$$

then $\mathcal{H}_{K_0}$ will happen with probability at least $1 - p$.

43

Proof. When $k \geq \mathcal{K}$, the algorithm has already stopped. Define virtual update when $x_{k+1} = x_k$, and v_k is generated by Line 3 and Line 6 in Algorithm 9.

Define the event $\tilde{\mathcal{H}}_k = (\|v_k - \nabla f(x_k)\|^2 \geq \epsilon \cdot \tilde{\epsilon}), k \in [K_0]$. Then if we can prove that for any $k \in [K_0]$, the probability of $\tilde{\mathcal{H}}_k$ occurring is no more than $p/(K_0 + 1)$, *i.e.*, $\Pr(\tilde{\mathcal{H}}_k) \leq \frac{p}{K_0+1}$. Then we have

$$\Pr(\mathcal{H}_{K_0}) \geq 1 - \Pr(\bigcup_{k=0}^{K_0} \tilde{\mathcal{H}}_k) \geq 1 - \sum_{k=0}^{K_0} \Pr(\tilde{\mathcal{H}}_k) \geq 1 - p$$

Now, we prove that $\Pr(\tilde{\mathcal{H}}_k) \leq \frac{p}{K_0+1}, \forall k \in [K_0]$.

First, we have

$$\|v_k - \nabla f(x_k)\|^2 \leq 2\|v_k - \hat{\nabla}_{coord} f(x_k)\|^2 + 2\|\hat{\nabla}_{coord} f(x_k) - \nabla f(x_k)\|^2$$

$$\leq 2\|v_k - \hat{\nabla}_{coord} f(x_k)\|^2 + \frac{\rho^2 d\mu^4}{18}$$

Denote by ξ_k the randomness in maintaining SPIDER v_k at iteration k, and $\mathcal{F}_k = \sigma\{\xi_0, \ldots, \xi_k\}$, where $\sigma(\cdot)$, where $\sigma\{\cdot\}$ denotes the sigma field. We know that x_k and v_k are measurable on $\mathcal{F}_{k-1}$.

❶ Then given $\mathcal{F}_{k-1}$, if $k = \lfloor k/q \rfloor q$, we define

$$\hat{\epsilon}_{k,i} = \frac{1}{|\mathcal{S}_1|} \left(\hat{\nabla}_{coord} f_{\mathcal{S}_1(i)}(x_k) - \hat{\nabla}_{coord} f(x_k)\right), \quad \epsilon_{k,i} = \frac{1}{|\mathcal{S}_1|} \left(\nabla f_{\mathcal{S}_1(i)}(x_k) - \nabla f(x_k)\right)$$

where $\mathcal{S}_1(i)$ denotes the i-th random component function seleted at iteration k and $1 \leq i \leq |\mathcal{S}_1|$. We have

$$\mathbb{E}[\hat{\epsilon}_{k,i}|\mathcal{F}_{k-1}] = 0, \quad \mathbb{E}[\epsilon_{k,i}|\mathcal{F}_{k-1}] = 0, \quad \|\epsilon_{k,i}\| \overset{\text{Assumption 2}}{\leq} \frac{\sigma}{|\mathcal{S}_1|}$$

and

$$\|v_k - \hat{\nabla}_{coord} f(x_k)\|^2$$

$$= \|\sum_{i=1}^{|\mathcal{S}_1|} \hat{\epsilon}_{k,i}\|^2$$

$$\leq 2\|\sum_{i=1}^{|\mathcal{S}_1|} \epsilon_{k,i}\|^2 + 2\|\sum_{i=1}^{|\mathcal{S}_1|} (\epsilon_{k,i} - \hat{\epsilon}_{k,i})\|^2$$

$$= 2\|\sum_{i=1}^{|\mathcal{S}_1|} \epsilon_{k,i}\|^2 + 2\|\hat{\nabla}_{coord} f_{\mathcal{S}_1}(x_k) - \hat{\nabla}_{coord} f(x_k) + \nabla f_{\mathcal{S}_1}(x_k) - \nabla f(x_k)\|^2$$

$$\leq 2\|\sum_{i=1}^{|\mathcal{S}_1|} \epsilon_{k,i}\|^2 + 2\left(2\|\hat{\nabla}_{coord} f_{\mathcal{S}_1}(x_k) - \nabla f_{\mathcal{S}_1(x_k)}\|^2 + 2\|\hat{\nabla}_{coord} f(x_k) - \nabla f(x_k)\|^2\right)$$

$$\leq 2\|\sum_{i=1}^{|\mathcal{S}_1|} \epsilon_{k,i}\|^2 + \frac{2}{9}\rho^2 d\mu^4$$

Then we have

$$\|v_k - \nabla f(x_k)\|^2 \leq 2\|v_k - \hat{\nabla}_{coord} f(x_k)\|^2 + \frac{\rho^2 d\mu^4}{18}$$

$$\leq 4\|\sum_{i=1}^{|\mathcal{S}_1|} \epsilon_{k,i}\|^2 + \frac{1}{2}\rho^2 d\mu^4$$

Then we have

$$\Pr\left(\|v_k - \nabla f(x_k)\|^2 \geq \epsilon \cdot \tilde{\epsilon}|\mathcal{F}_{k-1}\right)$$

44

$$\leq \Pr\left(4\|\sum_{i=1}^{|\mathcal{S}_1|}\epsilon_{k,i}\|^2 + \frac{1}{2}\rho^2 d\mu^4 \geq \epsilon\cdot\tilde{\epsilon}|\mathcal{F}_{k-1}\right) = \Pr\left(\|\sum_{i=1}^{|\mathcal{S}_1|}\epsilon_{k,i}\|^2 \geq \frac{\epsilon\cdot\tilde{\epsilon} - \frac{1}{2}\rho^2 d\mu^4}{4}|\mathcal{F}_{k-1}\right)$$

$$\leq 4\exp\left(-\frac{\frac{\epsilon\cdot\tilde{\epsilon}-\frac{1}{2}\rho^2 d\mu^4}{4}}{4|\mathcal{S}_1|\frac{\sigma^2}{|\mathcal{S}_1|^2}}\right) = 4\exp\left(-\frac{\epsilon\cdot\tilde{\epsilon} - \frac{1}{2}\rho^2 d\mu^4}{16\frac{\sigma^2}{|\mathcal{S}_1|}}\right)$$

❷ When $k \neq \lfloor k/q \rfloor q$, set $k_0 = \lfloor k/q \rfloor q$ and define

$$\epsilon_{j,i} = \frac{1}{|\mathcal{S}_2|}\left(\nabla f_{\mathcal{S}_2(i)}(x_j) - \nabla f_{\mathcal{S}_2(i)}(x_{j-1}) - \nabla f(x_j) + \nabla f(x_{j-1})\right)$$

$$\hat{\epsilon}_{j,i} = \frac{1}{|\mathcal{S}_2|}\left(\hat{\nabla}_{coord}f_{\mathcal{S}_2(i)}(x_j) - \hat{\nabla}_{coord}f_{\mathcal{S}_2(i)}(x_{j-1}) - \hat{\nabla}_{coord}f(x_j) + \hat{\nabla}_{coord}f(x_{j-1})\right)$$

where $\mathcal{S}_2(i)$ denotes the i-th random component function selected at iteration k and $1 \leq i \leq |\mathcal{S}_2|, k_0 \leq j \leq k$. We have

$$\mathbb{E}[\epsilon_{j,i}|\mathcal{F}_{j-1}] = 0, \quad \mathbb{E}[\hat{\epsilon}_{j,i}|\mathcal{F}_{j-1}] = 0$$

From the update rule if $k < \mathcal{K}$, we have $\|x_{k+1} - x_k\| = \|\eta v_k/\|v_k\|\| = \eta = \frac{\epsilon}{\ell n_0}$, if $k > \mathcal{K}$, we have $\|x_{k+1} - x_k\| = 0 \leq \frac{\epsilon}{\ell n_0}$. We have

$$\|\epsilon_{j,i}\|$$
$$\leq \frac{1}{|\mathcal{S}_2|}\left(\|\nabla f_{\mathcal{S}_2(i)}(x_j) - \nabla f_{\mathcal{S}_2(i)}(x_{j-1})\| + \|\nabla f(x_j) - \nabla f(x_j)\|\right)$$
$$\leq \frac{2\ell}{|\mathcal{S}_2|}\|x_j - x_{j-1}\| \leq \frac{2\epsilon}{|\mathcal{S}_2|n_0}$$

for all $k_0 < k \leq k$ and $1 \leq i \leq |\mathcal{S}_2|$, and

$$\|v_k - \hat{\nabla}_{coord}f(x_k)\|^2$$
$$= \left\|\hat{\nabla}_{coord}f_{\mathcal{S}_2}(x_k) - \hat{\nabla}_{coord}f_{\mathcal{S}_2}(x_{k-1}) - \hat{\nabla}_{coord}f(x_k) - \hat{\nabla}_{coord}f(x_{k-1}) + (v_k - \hat{\nabla}_{coord}f(x_{k-1}))\right\|^2$$
$$= \left\|\sum_{i=1}^{|\mathcal{S}_2|}\hat{\epsilon}_{k,i} + v_{k-1} - \hat{\nabla}_{coord}f(x_{k-1})\right\|^2$$
$$= \left\|\sum_{j=k_0+1}^{k}\sum_{i=1}^{|\mathcal{S}_2|}\hat{\epsilon}_{j,i} + v_{k_0} - \hat{\nabla}_{coord}f(x_{k_0})\right\|^2 = \left\|\sum_{j=k_0+1}^{k}\sum_{i=1}^{|\mathcal{S}_2|}\hat{\epsilon}_{j,i} + \sum_{i=1}^{|\mathcal{S}_1|}\hat{\epsilon}_{k_0,i}\right\|^2$$
$$\leq 2\left\|\sum_{j=k_0+1}^{k}\sum_{i=1}^{|\mathcal{S}_2|}\epsilon_{j,i} + \sum_{i=1}^{|\mathcal{S}_1|}\epsilon_{k_0,i}\right\|^2 + 2\left\|\sum_{j=k_0+1}^{k}\sum_{i=1}^{|\mathcal{S}_2|}(\epsilon_{j,i} - \hat{\epsilon}_{j,i}) + \sum_{i=1}^{|\mathcal{S}_1|}(\epsilon_{k_0,i} - \hat{\epsilon}_{k_0,i})\right\|^2$$
$$\leq 2\left\|\sum_{j=k_0+1}^{k}\sum_{i=1}^{|\mathcal{S}_2|}\epsilon_{j,i} + \sum_{i=1}^{|\mathcal{S}_1|}\epsilon_{k_0,i}\right\|^2 + 2\left(2\left\|\sum_{j=k_0+1}^{k}\sum_{i=1}^{|\mathcal{S}_2|}(\epsilon_{j,i} - \hat{\epsilon}_{j,i})\right\|^2 + 2\left\|\sum_{i=1}^{|\mathcal{S}_1|}(\epsilon_{k_0,i} - \hat{\epsilon}_{k_0,i})\right\|^2\right)$$
$$\leq 2\left\|\sum_{j=k_0+1}^{k}\sum_{i=1}^{|\mathcal{S}_2|}\epsilon_{j,i} + \sum_{i=1}^{|\mathcal{S}_1|}\epsilon_{k_0,i}\right\|^2 + 2\left(2\left\|\sum_{j=k_0+1}^{k}\sum_{i=1}^{|\mathcal{S}_2|}(\epsilon_{j,i} - \hat{\epsilon}_{j,i})\right\|^2 + \frac{2}{9}\rho^2 d\mu^4\right)$$

The second term can be bounded by

$$\left\|\sum_{j=k_0+1}^{k}\sum_{i=1}^{|\mathcal{S}_2|}(\epsilon_{j,i} - \hat{\epsilon}_{j,i})\right\|^2$$

45

$$\leq (k - k_0) \sum_{j=k_0+1}^{k} \left\| \sum_{i=1}^{\mathcal{S}_2} (\epsilon_{j,i} - \hat{\epsilon}_{j,i}) \right\|^2$$

$$= (k - k_0) \sum_{j=k_0+1}^{k} 4 \left(\left\| \nabla f_{\mathcal{S}_2}(x_j) - \hat{\nabla}_{coord} f_{\mathcal{S}_2}(x_j) \right\|^2 + \left\| \nabla f_{\mathcal{S}_2}(x_{j-1}) - \hat{\nabla}_{coord} f_{\mathcal{S}_2}(x_{j-1}) \right\|^2 \right.$$

$$\left. + \left\| \nabla f(x_j) - \hat{\nabla}_{coord} f(x_j) \right\|^2 + \left\| \nabla f(x_{j-1}) - \hat{\nabla}_{coord} f(x_{j-1}) \right\|^2 \right)$$

$$\leq (k - k_0) \sum_{j=k_0+1}^{k} 4 \cdot 4 \frac{\rho^2 d \mu^4}{36} = (k - k_0)^2 \frac{4}{9} \rho^2 d \mu^4$$

Thus we have

$$\|v_k - \nabla f(x_k)\|^2$$

$$\leq 2 \|v_k - \hat{\nabla}_{coord} f(x_k)\|^2 + \frac{\rho^2 d \mu^4}{18}$$

$$\leq 4 \left\| \sum_{j=k_0+1}^{k} \sum_{i=1}^{|\mathcal{S}_2|} \epsilon_{j,i} + \sum_{i=1}^{|\mathcal{S}_1|} \epsilon_{k_0,i} \right\|^2 + 4 \left(2(k - k_0)^2 \frac{4}{9} \rho^2 d \mu^4 + \frac{2}{9} \rho^2 d \mu^4 \right) + \frac{\rho^2 d \mu^4}{18}$$

$$\leq 4 \left\| \sum_{j=k_0+1}^{k} \sum_{i=1}^{|\mathcal{S}_2|} \epsilon_{j,i} + \sum_{i=1}^{|\mathcal{S}_1|} \epsilon_{k_0,i} \right\|^2 + 4(k - k_0)^2 \rho^2 d \mu^4$$

Using Lemma 24, we have

$$\Pr \left(\|v_k - \nabla f(x_k)\|^2 \geq \epsilon \cdot \tilde{\epsilon} \,|\, \mathcal{F}_{k_0-1} \right)$$

$$\leq \Pr \left(4 \left\| \sum_{j=k_0+1}^{k} \sum_{i=1}^{|\mathcal{S}_2|} \epsilon_{j,i} + \sum_{i=1}^{|\mathcal{S}_1|} \epsilon_{k_0,i} \right\|^2 + 4(k - k_0)^2 \rho^2 d \mu^4 \geq \epsilon \cdot \tilde{\epsilon} \,|\, \mathcal{F}_{k_0-1} \right)$$

$$= \Pr \left(\left\| \sum_{j=k_0+1}^{k} \sum_{i=1}^{|\mathcal{S}_2|} \epsilon_{j,i} + \sum_{i=1}^{|\mathcal{S}_1|} \epsilon_{k_0,i} \right\|^2 \geq \frac{\epsilon \cdot \tilde{\epsilon}}{4} - (k - k_0)^2 \rho^2 d \mu^4 \,|\, \mathcal{F}_{k_0-1} \right)$$

$$\leq 4 \exp \left(- \frac{\frac{\epsilon \cdot \tilde{\epsilon}}{4} - (k - k_0)^2 \rho^2 d \mu^4}{4 |\mathcal{S}_1| \frac{\sigma^2}{|\mathcal{S}_1|^2} + 4 |\mathcal{S}_2| (k - k_0) \frac{4 \epsilon^2}{|\mathcal{S}_2|^2 n_0^2}} \right)$$

$$\leq 4 \exp \left(- \frac{\frac{\epsilon \cdot \tilde{\epsilon}}{4} - q^2 \rho^2 d \mu^4}{4 \frac{\sigma^2}{|\mathcal{S}_1|} + 4q \frac{4 \epsilon^2}{|\mathcal{S}_2| n_0^2}} \right) = 4 \exp \left(- \frac{\epsilon \cdot \tilde{\epsilon} - 4 q^2 \rho^2 d \mu^4}{16 \frac{\sigma^2}{|\mathcal{S}_1|} + 16 q \frac{4 \epsilon^2}{|\mathcal{S}_2| n_0^2}} \right)$$

If we set

$$\tilde{\epsilon} = 10 \epsilon \log \frac{4(K_0 + 1)}{p}, |\mathcal{S}_1| = \frac{16 \sigma^2}{\epsilon^2}, |\mathcal{S}_2| = \frac{16 \sigma}{\epsilon n_0}, q = \frac{\sigma n_0}{\epsilon}, \mu = \left(\frac{\epsilon \cdot \tilde{\epsilon}}{8 q^2 \rho^2 d} \right)^{\frac{1}{4}} = \tilde{O}(\frac{\epsilon}{d^{1/4}})$$

We will get

- When $k \neq \lfloor k/q \rfloor q$,

$$\Pr \left(\|v_k - \nabla f(x_k)\|^2 \geq \epsilon \cdot \tilde{\epsilon} \,|\, \mathcal{F}_{k_0-1} \right) \leq 4 \exp \left(- \frac{\epsilon \cdot \tilde{\epsilon} - 4 q^2 \rho^2 d \mu^4}{16 \frac{\sigma^2}{|\mathcal{S}_1|} + 16 q \frac{4 \epsilon^2}{|\mathcal{S}_2| n_0^2}} \right) \leq \frac{p}{K_0 + 1}$$

- When $k = \lfloor k/q \rfloor q$,

$$\Pr \left(\|v_k - \nabla f(x_k)\|^2 \geq \epsilon \cdot \tilde{\epsilon} \,|\, \mathcal{F}_{k_0-1} \right) \leq 4 \exp \left(- \frac{\epsilon \cdot \tilde{\epsilon} - \frac{1}{2} \rho^2 d \mu^4}{16 \frac{\sigma^2}{|\mathcal{S}_1|}} \right)$$

46

$$\leq 4\exp\left(-\frac{\epsilon\cdot\tilde{\epsilon}-4q^2\rho^2d\mu^4}{16\frac{\sigma^2}{|\mathcal{S}_1|}+16q\frac{4\epsilon^2}{|\mathcal{S}_2|n_0^2}}\right)\leq\frac{p}{K_0+1}$$

This completes the whole proof.

$\square$

Lemma 26. *Under assumption 2, we have that on $\mathcal{H}_{K_0}\cap(\mathcal{K}>K_0)$, for all $0\leq k\leq K_0$,*

$$f(x_{k+1})-f(x_k)\leq-\frac{\epsilon\cdot\tilde{\epsilon}}{4\ell n_0}.$$

and

$$f(x_{K_0+1}-f(x_0))\leq-\frac{\epsilon\cdot\tilde{\epsilon}}{4\ell n_0}K_0$$

Proof. Let $\eta_k=\eta/\|v_k\|$ and since f has ℓ-Lipschitz continuous gradient, we have

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

$$=f(x_k)-\eta_k\langle\nabla f(x_k),v_k\rangle+\frac{\ell\eta_k^2}{2}\|v_k\|^2$$

$$=f(x_k)-\eta_k\langle\nabla f(x_k)-v_k,v_k\rangle-\eta_k\|v_k\|^2+\frac{\ell\eta_k^2}{2}\|v_k\|^2$$

$$\leq f(x_k)+\frac{\eta_k}{2}\|v_k-\nabla f(x_k)\|^2-\eta_k\left(\frac{1}{2}-\frac{\eta_k\ell}{2}\right)\|v_k\|^2 \qquad (15)$$

where the last inequality uses the the Cauchy-Schwarz inequality. Because we are on the event $\mathcal{H}_{K_0}\cap(\mathcal{K}>K_0)$, so $\mathcal{K}-1\geq K_0$, then for all $0\leq k\leq K_0$, we have $\|v_k\|\geq2\tilde{\epsilon}$, thus

$$\eta_k=\frac{\epsilon}{\ell n_0}\frac{1}{\|v_k\|}\leq\frac{\epsilon}{\ell n_0}\frac{1}{2\tilde{\epsilon}}\leq\frac{1}{2\ell n_0}\leq\frac{1}{2\ell}$$

we have

$$\eta_k\left(\frac{1}{2}-\frac{\eta_k\ell}{2}\right)\|v_k\|^2\geq\eta_k\frac{1}{4}\|v_k\|^2=\frac{1}{4}\frac{\epsilon}{\ell n_0}\frac{1}{\|v_k\|}\|v_k\|^2\geq\frac{\epsilon\cdot\tilde{\epsilon}}{2\ell n_0}$$

and when $\mathcal{H}_{K_0}$ happens, we also have

$$\frac{\eta_k}{2}\|v_k-\nabla f(x_k)\|^2\leq\frac{\eta_k}{2}\epsilon\cdot\tilde{\epsilon}\overset{\eta_k\leq\frac{1}{2\ell n_0}}{\leq}\frac{\epsilon\cdot\tilde{\epsilon}}{4\ell n_0}$$

Hence

$$f(x_{k+1})\leq f(x_k)+\frac{\epsilon\cdot\tilde{\epsilon}}{4\ell n_0}-\frac{\epsilon\cdot\tilde{\epsilon}}{2\ell n_0}=f(x_k)-\frac{\epsilon\cdot\tilde{\epsilon}}{4\ell n_0}$$

By telescoping the above the inequality, we have

$$f(x_{K_0+1})-f(x_0)\leq-\frac{\epsilon\cdot\tilde{\epsilon}}{4\ell n_0}K_0.$$

$\square$

Proof of Theorem 9 . If $\mathcal{K}\leq K_0\cap\mathcal{H}_{K_0}$, we have $\|v_\mathcal{K}\|\leq2\tilde{\epsilon}$. Because $\|v_\mathcal{K}-\nabla f(x_\mathcal{K})\|\leq\sqrt{\epsilon\cdot\tilde{\epsilon}}\leq\tilde{\epsilon}$ if $\mathcal{H}_{K_0}$ occurs, so $\|\nabla f(x_\mathcal{K})\|\leq2\tilde{\epsilon}$.

If $\mathcal{K}>K_0\cap\mathcal{H}_{K_0}$, we have

$$-\Delta_f\leq f^*-f(x_0)\leq f(x_{K_0+1})-f(x_0)\leq-\frac{\epsilon\cdot\tilde{\epsilon}}{4\ell n_0}K_0\leq-\frac{\epsilon\cdot\tilde{\epsilon}}{4\ell n_0}\left(\frac{4\ell\Delta_f n_0}{\epsilon^2}+1\right)$$

$$\leq-\frac{\epsilon\cdot\tilde{\epsilon}}{4\ell n_0}\left(\frac{4\ell\Delta_f n_0}{\epsilon\cdot\tilde{\epsilon}}+1\right)=-\Delta_f-\frac{\epsilon\cdot\tilde{\epsilon}}{4\ell n_0}$$

47

which is contradict with the fact that $-\Delta_f > -\Delta_f - \frac{\epsilon \cdot \tilde{\epsilon}}{4\ell n_0}$. This means that when $\mathcal{H}_{K_0}$ happens, then the algorithm must terminate before K_0 iterations.

Therefore, the total function query complexity can be bounded by:

$$
\begin{aligned}
d\left(\lceil K_0/q \rceil |\mathcal{S}_1| + K_0|\mathcal{S}_2|\right) &\leq d\left((K_0/q + 1)|\mathcal{S}_1| + K_0|\mathcal{S}_2|\right) \\
&\overset{①}{\leq} d\left(|\mathcal{S}_1| + 2K_0|\mathcal{S}_2|\right) \\
&\leq d\left(\frac{16\sigma^2}{\epsilon^2} + 2\left(\frac{4\ell\Delta_f n_0}{\epsilon^2} + 1\right)\frac{16\sigma}{\epsilon n_0}\right) \\
&= d\left(\frac{128\ell\sigma\Delta_f}{\epsilon^3} + \frac{16\sigma^2}{\epsilon^2} + \frac{32\sigma}{\epsilon n_0}\right)
\end{aligned}
$$

where ① is because $|\mathcal{S}_1| = q|\mathcal{S}_2|$.

$\square$

F.2 Proof of high probability results for Second-Order Stationary Point of Theorem 6

From Algorithm 8, we know that all randomness in iteration k come from three parts: 1) maintaining SPIDER v_k in Line 5-10; 2) to conduct the zeroth-order negative-curvature search in Line 2; choosing a random sign of w_2 in Line 3. Denote by $\xi_k^1; \xi_k^2; \xi_k^3$ the randomness of from the three parts, respectively. Let $\mathcal{F}_k$ be the filtration involving the full information of $x_{0:k}, v_{0:k}$, i.e., $\mathcal{F}_k = \sigma\{\xi_{0:k}^1, \xi_{0:k}^2, \xi_{0:k-1}^3\}$. So the randomness in iteration k given $\mathcal{F}_k$ only comes from ξ_k^3.

Denote the random index

$$
\mathcal{I}_k = \begin{cases} 1 & \text{if} \quad w_1 = \perp \\ 2 & \text{if} \quad w_1 \neq \perp \end{cases}
$$

then we know that $\mathcal{I}_k$ is measurable on $\mathcal{F}_{\lfloor k/\mathscr{K}\rfloor\mathscr{K}}$ and also on $\mathcal{F}_k$. When the event $(\mathcal{I}_k = 1 \bigcap \|v_k\| \leq 2\tilde{\epsilon})$ happens, then the algorithm will be stopped. In this case, we define a virtual update $x_{k+1} = x_k$ in Line 13 and Line 18.

Let $\mathcal{H}_k^1$ denotes the event that algorithm has not stopped before k, i.e.,

$$
\mathcal{H}_k^1 = \bigcap_{i=1}^{k}\left(\left(\|v_i\| \geq 2\tilde{\epsilon}\bigcap\mathcal{I}_i = 1\right)\bigcup\mathcal{I}_i = 2\right)
$$

Let $\mathcal{H}_{\mathscr{K}j}^2$ denotes the event that the Zeroth-Order Curvature Finding in iteration $\mathscr{K}j$ runs successfully.

Let $\mathcal{H}_k^3$ denotes the event that

$$
\mathcal{H}_k^3 = \left(\bigcap_{i=0}^{k}(\|v_i - \nabla f(x_i)\| \leq \epsilon \cdot \tilde{\epsilon})\right) \bigcap \left(\bigcap_{j=0}^{\lfloor k/\mathscr{K}\rfloor}\mathcal{H}_{\mathscr{K}j}^2\right)
$$

Then we have $\mathcal{H}_3^k \in \mathcal{F}_k$, and $\mathcal{H}_1^3 \supseteq \mathcal{H}_2^3 \supseteq \cdots \supseteq \mathcal{H}_k^3$.

Lemma 27. *With the setting of Theorem 6, and under the Assumption 2, we have*

$$
Pr\left(\mathcal{H}_{K_0}^3\right) \geq \frac{15}{16}
$$

Proof. Denote the event $\tilde{H}_k = (\|v_k - \nabla f(x_k)\| \leq \epsilon \cdot \tilde{\epsilon}), 0 \leq k \leq K_0$. If we can prove that $\Pr(\tilde{H}_k) \geq 1 - \frac{1}{32(K_0+1)}$, we have $\Pr\left(\bigcap_{i=0}^{K_0}\tilde{H}_i\right) \geq 1 - \frac{1}{32}$. On the other hand, from Theorem 1, we know that each time $\Pr\left(\mathcal{H}_{\mathscr{K}j}^2\right) \geq 1 - \frac{1}{32J}$, so $\Pr\left(\bigcap_{j=0}^{\lfloor K_0/\mathscr{K}\rfloor}\mathcal{H}_{\mathscr{K}j}^2\right) \geq 1 - \frac{1}{32}$. So we have $\Pr(\mathcal{H}_{K_0}^3) \geq \frac{15}{16}$.

Now, we prove $\Pr(\tilde{H}_k) \geq 1 - \frac{1}{32(K_0+1)}$.

Consider the filtration of full information of $x_{0:k}$, $\mathcal{F}_k^2 = \sigma\{\xi_{0:k-1}^1, \xi_{0:k}^2, \xi_{0:k-1}^3\}$. We know x_k is measurable on $\mathcal{F}_k^2$. Given $\mathcal{F}_k^2$, we have

48

- when $k = \lfloor k/q \rfloor q$,

$$\mathbb{E}_i \left[\hat{\nabla}_{coord} f_i(x_k) - \hat{\nabla}_{coord} f(x_k) \mid \mathcal{F}_k^2 \right] = 0, \quad \mathbb{E}_i \left[\nabla f_i(x_k) - \nabla f(x_k) \mid \mathcal{F}_k^2 \right] = 0$$

- when $k \neq \lfloor k/q \rfloor q$,

$$\mathbb{E}_i \left[\hat{\nabla}_{coord} f_i(x_k) - \hat{\nabla}_{coord} f_i(x_{k-1}) + \left(\hat{\nabla}_{coord} f(x_k) - \hat{\nabla}_{coord} f(x_{k-1}) \right) \right] = 0$$
$$\mathbb{E}_i \left[\nabla f_i(x_k) - \nabla f_i(x_{k-1}) + (\nabla f(x_k) - \nabla f(x_{k-1})) \right] = 0$$

Because x_k is generated by one of the three ways:

- when $w_1 = \bot$, we have $\|x_k - x_{k-1}\| = \|\eta(v_{k-1}/\|v_{k-1}\|)\| = \eta = \frac{\epsilon}{\ell n_0}$;

- when $w_1 \neq \bot$, we have $\|x_k - x_{k-1}\| = \|\eta w_1\| = \eta = \frac{\epsilon}{\ell n_0}$;

- when Algorithm 8 has already stopped, we have $\|x_{k+1} - x_k\| = 0 \leq \frac{\epsilon}{\ell n_0}$

So $v_k - \nabla f(x_k)$ is martingale. If we set $|\mathcal{S}_1|, |\mathcal{S}_1|, \eta, \mu$ and $\tilde{\epsilon} = 10\epsilon \log(128(K_0 + 1))$ as the same in Lemma 25. Then using the same technique of Lemma 25 with $p = \frac{1}{32}$, we have $\Pr(\tilde{H}_k) \geq 1 - \frac{1}{16(K_0+1)}, 0 \leq k \leq K_0$.

$\square$

Lemma 28. *If* $\left(\mathcal{H}_{K_0}^1\right)^c \cap \mathcal{H}_{K_0}^3$ *happens, Algorithm 8 outputs* x_k *satisfying*

$$\|\nabla f(x_k)\| \leq 3\tilde{\epsilon}, \quad \lambda_{min}(\nabla^2 f(x_k)) \geq -2\delta,$$

before K_0 *iterations.*

Proof. If $\left(\mathcal{H}_{K_0}^1\right)^c$ happens, then Algorithm 8 has already stopped before iteration K_0 and output x_k with $\|v_k\| \leq 2\tilde{\epsilon}$. If $\mathcal{H}_{K_0}^3$ happens, we have

- $\|\nabla f(x_k) - v_k\| \leq \sqrt{\epsilon \cdot \tilde{\epsilon}} \leq \tilde{\epsilon}$. So we have

$$\|\nabla f(x_k)\| \leq \|v_k\| + \|\nabla f(x_k) - v_k\| \leq 3\tilde{\epsilon}.$$

- Each time the Zeroth-Order Curvature Finding runs successfully, so from Theorem 1, we have

$$\lambda_{min}(\nabla^2 f(x_{k_0})) \geq -\delta,$$

where $k_0 = \lfloor k/\mathcal{K} \rfloor \mathcal{K}$.

From Assumption 2, we have

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq \left\| \frac{1}{n} \sum_{i=1}^n \left(\nabla^2 f_i(x) - \nabla^2 f_i(y) \right) \right\| \leq \frac{1}{n} \sum_{i=1}^n \|\nabla^2 f_i(x) - \nabla^2 f_i(y)\| \leq \rho \|x - y\|.$$

This means that f has ρ-Lipschitz Hessian. We have

$$\left\| \nabla^2 f(x_k) - \nabla^2 f(x_{k_0}) \right\| \leq \left\| \sum_{i=k_0}^{k-1} (\nabla^2 f(x_{i+1}) - \nabla^2 f(x_i)) \right\|$$

$$\leq \sum_{i=k_0}^{k-1} \rho \|x_{i+1} - x_i\| \leq \mathcal{K} \frac{\epsilon}{\ell n_0} \quad \mathcal{K} = \frac{\delta \ell n_0}{\rho \epsilon} = \delta$$

Finally we get $\lambda_{min}(\nabla^2 f(x_k)) \geq -2\delta$.

$\square$

49

Proof of Theorem 6. Denote $\mathcal{H}_k^4 = \mathcal{H}_k^1 \cap \mathcal{H}_k^3$. For all iteration $\mathcal{K}$ with $\mathrm{mod}(\mathcal{K}, \mathscr{K}) = 0$, given $\mathcal{F}_\mathcal{K}$, we consider:

❶ When $\mathcal{H}_k^4 \bigcap (\mathcal{I}_\mathcal{K} = 2)$ happens. From Lemma 4 and the fact that f is ρ-Hessian Lipschitz, we have

$$f(x_{\mathcal{K}+\mathscr{K}})$$
$$\leq f(x_\mathcal{K}) + \langle \nabla f(x_\mathcal{K}), x_{\mathcal{K}+\mathscr{K}} - x_\mathcal{K} \rangle + \frac{1}{2}(x_{\mathcal{K}+\mathscr{K}} - x_\mathcal{K})^\mathsf{T} \nabla^2 f(x_\mathcal{K})(x_{\mathcal{K}+\mathscr{K}} - x_\mathcal{K}) + \frac{\rho}{6}\|x_{\mathcal{K}+\mathscr{K}} - x_\mathcal{K}\|^3.$$

From Theorem 1, we have $w_1^\mathsf{T} \nabla^2 f(x_\mathcal{K}) w_1 \leq -\frac{\delta}{2}$. Take expectation on the random number of the sign we have

$$\mathbb{E}\langle \nabla f(x_\mathcal{K}), x_{\mathcal{K}+\mathscr{K}} - x_\mathcal{K} \rangle = 0$$

Thus we have

$$f(x_{\mathcal{K}+\mathscr{K}}) \leq f(x_\mathcal{K}) + \frac{1}{2}(\eta \cdot \mathscr{K} w_1)^\mathsf{T} \nabla^2 f(x_\mathcal{K})(\eta \cdot \mathscr{K} w_1) + \frac{\rho}{6}\|\eta \cdot \mathscr{K} w_1\|^3$$

$$\overset{\eta = \frac{\epsilon}{\ell n_0}, \mathscr{K} = \frac{\delta \ell n_0}{\rho \epsilon}}{=} f(x_\mathcal{K}) - \frac{\delta^3}{4\rho^2} + \frac{\delta^3}{6\rho^2} = f(x_\mathcal{K}) - \frac{\delta^3}{12\rho^2}$$

Then we analysis the difference of $(f(x_\mathcal{K} - f^*))\mathbb{I}_{\mathcal{H}_\mathcal{K}^4}$, where $\mathbb{I}_{\mathcal{H}_\mathcal{K}^4}$ is the indication function of $\mathcal{H}_\mathcal{K}^4$, then we have

$$\mathbb{E}\left[(f(x_{\mathcal{K}+\mathscr{K}} - f^*))\mathbb{I}_{\mathcal{H}_{\mathcal{K}+\mathscr{K}}^4} \mid \mathcal{F}_\mathcal{K} - (f(x_\mathcal{K} - f^*))\mathbb{I}_{\mathcal{H}_\mathcal{K}^4} \mid \mathcal{F}_\mathcal{K}\right]$$

$$= \mathbb{E}\left[(f(x_{\mathcal{K}+\mathscr{K}} - f^*))\left(\mathbb{I}_{\mathcal{H}_{\mathcal{K}+\mathscr{K}}^4} - \mathbb{I}_{\mathcal{H}_\mathcal{K}^4}\right) \mid \mathcal{F}_\mathcal{K}\right] + \mathbb{E}\left[(f(x_{\mathcal{K}+\mathscr{K}}) - f(x_\mathcal{K}))\mathbb{I}_{\mathcal{H}_\mathcal{K}^4} \mid \mathcal{F}_\mathcal{K}\right]$$

$$\leq \Pr(\mathcal{H}_\mathcal{K}^4 \mid \mathcal{F}_\mathcal{K})\mathbb{E}\left[(f(x_{\mathcal{K}+\mathscr{K}}) - f(x_\mathcal{K}))\mathbb{I}_{\mathcal{H}_\mathcal{K}^4} \mid \mathcal{F}_\mathcal{K} \cap \mathcal{H}_\mathcal{K}^4\right]$$

$$\leq - \Pr(\mathcal{H}_\mathcal{K}^4 \mid \mathcal{F}_\mathcal{K})\frac{\delta^3}{12\rho^2} \tag{16}$$

where the last second inequality uses the fact that $\mathbb{I}_{\mathcal{H}_{\mathcal{K}+\mathscr{K}}^4} - \mathbb{I}_{\mathcal{H}_\mathcal{K}^4} \leq 0$.

❷ When $\mathcal{H}_k^4 \bigcap (\mathcal{I}_\mathcal{K} = 1)$ happens, then for $\mathcal{K} \leq k < \mathcal{K} + \mathscr{K}$, from Eq.(15), we have

$$f(x_{k+1}) \leq f(x_k) + \frac{\eta_k}{2}\|v_k - \nabla f(x_k)\|^2 - \eta_k\left(\frac{1}{2} - \frac{\eta_k \ell}{2}\right)\|v_k\|^2$$

where $\eta_k = \eta/\|v_k\|$. If $\mathcal{H}_k^4$ happens we have $\|v_k\| \geq 2\tilde{\epsilon}$ and $\|v_k - \nabla f(x_k)\| \leq \epsilon \cdot \tilde{\epsilon}$, then from the proof of Lemma 26, we also have

$$f(x_{k+1}) \leq f(x_k) - \frac{\epsilon \cdot \tilde{\epsilon}}{4\ell n_0}$$

Taking expectation up to $\mathcal{F}_\mathcal{K}$, we have

$$\mathbb{E}\left[f(x_{k+1}) - f(x_k) \mid \mathcal{F}_\mathcal{K} \cap \mathcal{H}_k^4\right] \leq -\frac{\epsilon \cdot \tilde{\epsilon}}{4\ell n_0}$$

By analyzing the difference of $(f(x_k) - f^*)\mathbb{I}_{\mathcal{H}_k}^4$, we have

$$\mathbb{E}\left[(f(x_{k+1}) - f^*)\mathbb{I}_{\mathcal{H}_{k+1}^4} \mid \mathcal{F}_\mathcal{K} - (f(x_k) - f^*)\mathbb{I}_{\mathcal{H}_k^4} \mid \mathcal{F}_\mathcal{K}\right]$$

$$= \mathbb{E}\left[(f(x_{k+1}) - f^*)\left(\mathbb{I}_{\mathcal{H}_{k+1}^4} - \mathbb{I}_{\mathcal{H}_k^4}\right) \mid \mathcal{F}_\mathcal{K}\right] + \mathbb{E}\left[(f(x_{k+1})) - f(x_k)\mathbb{I}_{\mathcal{H}_k^4} \mid \mathcal{F}_\mathcal{K}\right]$$

$$\leq \Pr(\mathcal{H}_k^4 \mid \mathcal{F}_\mathcal{K})\mathbb{E}\left[(f(x_{k+1}) - f(x_k))\mathbb{I}_{\mathcal{H}_k^4} \mid \mathcal{F}_\mathcal{K} \cap \mathcal{H}_k^4\right]$$

$$\leq - \Pr(\mathcal{H}_k^4 \mid \mathcal{F}_\mathcal{K})\frac{\epsilon \cdot \tilde{\epsilon}}{4\ell n_0}$$

where the last second inequality uses the fact that $\mathbb{I}_{\mathcal{H}_{\mathcal{K}+\mathscr{K}}^4} - \mathbb{I}_{\mathcal{H}_\mathcal{K}^4} \leq 0$. By telescoping the above inequality with k from $\mathcal{K}$ to $\mathcal{K} + \mathscr{K} - 1$, we have

$$\mathbb{E}\left[(f(x_{\mathcal{K}+\mathscr{K}}) - f^*)\mathbb{I}_{\mathcal{H}_{\mathcal{K}+\mathscr{K}}^4} \mid \mathcal{F}_\mathcal{K} - (f(x_\mathcal{K}) - f^*)\mathbb{I}_{\mathcal{H}_\mathcal{K}^4} \mid \mathcal{F}_\mathcal{K}\right]$$

50

$$\leq -\frac{\epsilon \cdot \tilde{\epsilon}}{4\ell n_0} \sum_{i=\mathcal{K}}^{\mathcal{K}+\mathscr{K}} \Pr(\mathcal{H}_i^4 \mid \mathcal{F}_\mathcal{K}) \leq -\frac{\epsilon \cdot \tilde{\epsilon}}{4\ell n_0} \mathscr{K} \Pr(\mathcal{H}_{\mathcal{K}+\mathscr{K}}^4 \mid \mathcal{F}_\mathcal{K})$$

$$\overset{\mathscr{K} = \frac{\delta \ell n_0}{\rho \epsilon}}{=} -\Pr(\mathcal{H}_{\mathcal{K}+\mathscr{K}}^4 \mid \mathcal{F}_\mathcal{K})\frac{\delta\tilde{\epsilon}}{4\rho} \tag{17}$$

where the second inequality uses the fact that $\mathcal{H}_i^4 \subseteq \mathcal{H}_{\mathcal{K}+\mathscr{K}}^4$, thus $\Pr(\mathcal{H}_i^4 \mid \mathcal{F}_\mathcal{K}) \geq \Pr(\mathcal{H}_{\mathcal{K}+\mathscr{K}}^4 \mid \mathcal{F}_\mathcal{K})$. Combining (16) and (17) we have

$$\mathbb{E}\left[(f(x_{\mathcal{K}+\mathscr{K}}) - f^*)\,\mathbb{I}_{\mathcal{H}_{\mathcal{K}+\mathscr{K}}^4} \mid \mathcal{F}_\mathcal{K} - (f(x_\mathcal{K}) - f^*)\,\mathbb{I}_{\mathcal{H}_\mathcal{K}^4} \mid \mathcal{F}_\mathcal{K}\right]$$

$$\leq \min\left(-\Pr(\mathcal{H}_\mathcal{K}^4 \mid \mathcal{F}_\mathcal{K})\frac{\delta^3}{12\rho^2}, -\Pr(\mathcal{H}_{\mathcal{K}+\mathscr{K}}^4 \mid \mathcal{F}_\mathcal{K})\frac{\delta\tilde{\epsilon}}{4\rho}\right)$$

$$\leq -\Pr(\mathcal{H}_{\mathcal{K}+\mathscr{K}}^4 \mid \mathcal{F}_\mathcal{K})\min\left(\frac{\delta^3}{12\rho^2}, \frac{\delta\tilde{\epsilon}}{4\rho}\right)$$

By taking full expectation on the above inequality, and telescoping the results with $\mathcal{K} = 0, \mathscr{K}, \dots, (J-1)\mathscr{K}$, we have

$$\mathbb{E}\left[(f(x_{J\mathscr{K}}) - f^*)\,\mathbb{I}_{\mathcal{H}_{J\mathscr{K}}^4} - (f(x_0) - f^*)\,\mathbb{I}_{\mathcal{H}_0^4}\right]$$

$$\leq -\min\left(\frac{\delta^3}{12\rho^2}, \frac{\delta\tilde{\epsilon}}{4\rho}\right)\sum_{j=0}^{J-1}\Pr(\mathcal{H}_{j\mathscr{K}}^4)$$

$$\leq -\min\left(\frac{\delta^3}{12\rho^2}, \frac{\delta\tilde{\epsilon}}{4\rho}\right)J\Pr(\mathcal{H}_{J\mathscr{K}}^4)$$

Then we use the fact that $f(x_{J\mathscr{K}}) - f^* \geq 0, f(x_0) - f^* \leq \Delta_f$ and choose $J = 8\left(\left\lfloor\max\left(\frac{12\rho^2\Delta_f}{\delta^3}, \frac{4\rho\Delta_f}{\delta\epsilon}\right)\right\rfloor + 1\right) \geq \frac{8\Delta_f}{\min\left(\frac{\delta^3}{12\rho^2}, \frac{\delta\tilde{\epsilon}}{4\rho}\right)}$, we have

$$\Pr\left(\mathcal{H}_{J\mathscr{K}}^4\right) \overset{K_0 = J\mathscr{K}}{=} \Pr\left(\mathcal{H}_{K_0}^4\right) \leq \frac{1}{8}$$

Using the union bound and Lemma 27, we have

$$\Pr\left(\mathcal{H}_{K_0}^1\right) = \Pr\left(\mathcal{H}_{K_0}^1 \bigcap \mathcal{H}_{K_0}^3\right) + \Pr\left(\mathcal{H}_{K_0}^1 \bigcap \left(\mathcal{H}_{K_0}^3\right)^c\right) = \Pr\left(\mathcal{H}_{K_0}^4\right) + \Pr\left(\mathcal{H}_{K_0}^1 \bigcap \left(\mathcal{H}_{K_0}^3\right)^c\right)$$

$$\leq \Pr\left(\mathcal{H}_{K_0}^4\right) + \Pr\left(\left(\mathcal{H}_{K_0}^3\right)^c\right) = \Pr\left(\mathcal{H}_{K_0}^4\right) + 1 - \Pr\left(\mathcal{H}_{K_0}^3\right)$$

$$\leq \frac{1}{8} + 1 - \frac{15}{16} = \frac{3}{16}$$

Then we have

$$\Pr\left(\left(\mathcal{H}_{K_0}^1\right)^c \bigcap \mathcal{H}_{K_0}^3\right) = 1 - \Pr\left(\mathcal{H}_{K_0}^1 \bigcup \left(\mathcal{H}_{K_0}^3\right)^c\right) \geq 1 - \Pr\left(\mathcal{H}_{K_0}^1\right) - \Pr\left(\left(\mathcal{H}_{K_0}^3\right)^c\right)$$

$$= \Pr\left(\mathcal{H}_{K_0}^3\right) - \Pr\left(\mathcal{H}_{K_0}^1\right) \geq \frac{15}{16} - \frac{3}{16} = \frac{3}{4}$$

From Lemma 28, we have with probability at least $\frac{3}{4}$, Algorithm 8 will be terminated and output x_k before K_0 iterations satisfying

$$\|\nabla f(x_k)\| \leq 3\tilde{\epsilon}, \quad \lambda_{min}(\nabla^2 f(x_k)) \geq -2\delta,$$

Then we compute the total function query complexity:

- On the one hand, with probability at least $3/4$, the algorithm stops with no more than K_0 iterations, thus the function query complexity of computing the deterministic coordinate-wise gradient in Line 6 and Line 9 of Algorithm 8 can be bounded by

$$d\left(\lceil K_0/q\rceil |\mathcal{S}_1| + K_0|\mathcal{S}_2|\right) \leq d\left((K_0/q + 1)|\mathcal{S}_1| + K_0|\mathcal{S}_2|\right) \overset{|\mathcal{S}_1| = q|\mathcal{S}_2|}{\leq} d\left(|\mathcal{S}_1| + 2K_0|\mathcal{S}_2|\right)$$

$$\leq d\left(|\mathcal{S}_1| + \mathscr{K}\left(8\left\lfloor\max\left(\frac{12\rho^2\Delta_f}{\delta^3}, \frac{4\rho\Delta_f}{\delta\epsilon}\right)\right\rfloor + 8\right) \cdot 2|\mathcal{S}_2|\right)$$

51

- On the other hand, with probability at least $3/4$, the algorithm stops with no more than K_0 iterations, thus there are at most $K_0/\mathscr{K} = J$ times of zeroth-order negative curvature search. The total function query complexity for zeroth-order negative curvature search can be bounded by

$$\tilde{O}\left(d\left(8\left\lfloor \max\left(\frac{12\rho^2\Delta_f}{\delta^3}, \frac{4\rho\Delta_f}{\delta\epsilon}\right)\right\rfloor + 8\right)\frac{\ell^2}{\delta^2}\right)$$

  where $\tilde{O}$ hides a polylogarithmic factor of d.

Then by substituting $\mathscr{K} = \frac{\delta\ell n_0}{\rho\epsilon}, |\mathcal{S}_1| = \frac{16\sigma^2}{\epsilon^2}, |\mathcal{S}_2| = \frac{16\sigma}{\epsilon n_0}$ the total function query complexity can be bounded by

$$\tilde{\mathcal{O}}\left(d\left(\frac{16\sigma^2}{\epsilon^2} + \left(2\frac{16\sigma}{\epsilon n_0}\frac{\delta\ell n_0}{\rho\epsilon} + \frac{\ell^2}{\delta^2}\right) \cdot \left(8\left(\frac{12\rho^2\Delta_f}{\delta^3} + \frac{4\rho\Delta_f}{\delta\epsilon}\right) + 8\right)\right)\right)$$

$$=\tilde{\mathcal{O}}\left(d\frac{\sigma^2}{\epsilon^2} + d\left(\frac{\sigma\delta\ell}{\rho\epsilon^2} + \frac{\ell^2}{\delta^2}\right)\left(\frac{\rho^2\Delta_f}{\delta^3} + \frac{\rho\Delta_f}{\delta\epsilon} + 1\right)\right)$$

$$=\tilde{\mathcal{O}}\left(d\left(\frac{\sigma\ell\Delta_f}{\epsilon^3} + \frac{\sigma\ell\rho\Delta_f}{\epsilon^2\delta^2} + \frac{\ell^2\rho\Delta_f}{\delta^3\epsilon} + \frac{\ell^2\rho^2\Delta_f}{\delta^5} + \frac{\sigma^2}{\epsilon^2} + \frac{\sigma\delta\ell}{\rho\epsilon^2} + \frac{\ell^2}{\delta^2}\right)\right)$$

above, $\tilde{\mathcal{O}}$ hides the polylogarithmic factor of d and the constant factor. $\qquad\square$

G   Additional Experiments

In this section, we conduct several experiments to verify the effectiveness of our methods for both deterministic setting and stochastic setting. Specifically, for the deterministic setting, we compare our ZO-GD-NCF against three ZO algorithms for escaping saddle points, which are ZPSGD, PAGD, and RSPI. For the stochastic setting, we compare the three algorithms proposed in the paper, which are ZO-SGD-NCF, ZO-SCSG-NCF, and ZO-SPIDER-NCF. We don't compare our methods against ZO-SCRN because each iteration of ZO-SCRN needs to solve a cubic minimization subproblem: $x_{k+1} = \operatorname{argmin}_{x\in\mathbb{R}^d} g^\mathsf{T}(x-x_k) + \frac{1}{2}(x-x_k)^\mathsf{T} H(x-x_k) + \frac{\alpha}{6}\|x-x_k\|^3$, where g and H are inexact estimations of the full gradient $\nabla f(x_k)$ and $\nabla^2 f(x_k)$ by ZO oracle, respectively. Although many efficient inexact solvers of the cubic minimization subproblem have been proposed, most of them are second-order or first-order methods [1, 7, 10, 11], which is out of the scope of this paper.

G.1   Algorithms Description

Algorithm 10 Zero-th order Perturbed Stochastic Gradient Descent (ZPSGD)

Input: x_0, learning rate η, noise radius r, mini-batch size m.
    for $t = 0, 1, \ldots, T$ **do**
        sample $(z_t^{(1)}, \cdots, z_t^{(m)}) \sim \mathcal{N}(0, \sigma^2\mathbf{I})$
        $g_t(x_t) \leftarrow \sum_{i=1}^{m} z_t^{(i)}[f(x_t + z_t^{(i)}) - f(x_t)]/(m\sigma^2)$
        $x_{t+1} \leftarrow x_t - \eta(g_t(x_t) + \xi_t), \qquad \xi_t$ uniformly $\sim \mathbb{B}_0(r)$
    return x_T

Algorithm   Initialization: $(\ell, \rho, \epsilon, c, \delta, \Delta_f)$

1: $\chi \leftarrow 3\max\{\log(\frac{d\ell\Delta_f}{c\epsilon^2\delta}), 4\}$, $\eta \leftarrow \frac{c}{\ell}$, $r \leftarrow \frac{\sqrt{c}}{\chi^2}\cdot\frac{\epsilon}{\ell}$, $g_{\text{thres}} \leftarrow \frac{\sqrt{c}}{\chi^2}\cdot\epsilon$, $f_{\text{thres}} \leftarrow \frac{c}{\chi^3}\cdot\sqrt{\frac{\epsilon^3}{\rho}}$

2: $t_{\text{thres}} \leftarrow \frac{\chi}{c^2}\cdot\frac{\ell}{\sqrt{\rho\epsilon}}$, $S \leftarrow \frac{\sqrt{c}}{\chi}\frac{\sqrt{\rho\epsilon}}{\rho}$, $h_{low} \leftarrow \frac{1}{c_h}\min\{g_{\text{thres}}, \frac{r\rho\delta S}{2\sqrt{d}}\}$

52

Algorithm 12 PAGD($\mathbf{x}_0$)
1: **for** $t = 0, 1, \ldots$ **do**
2:      $\mathbf{z}_t \leftarrow q\left(\mathbf{x}_t, \frac{g_{\text{thres}}}{4c_h}\right)$
3:      **if** $\|\mathbf{z}_t\| \geq \frac{3}{4} g_{\text{thres}}$ **then**
4:          $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta \mathbf{z}_t$
5:      **else**
6:          $\mathbf{x}_{t+1} \leftarrow$ EscapeSaddle $(\mathbf{x}_t)$
7:          **if** $\mathbf{x}_{t+1} = \mathbf{x}_t$ **then return** $\mathbf{x}_t$

Algorithm 13 EscapeSaddle ($\hat{\mathbf{x}}$)
1: $\boldsymbol{\xi} \sim \text{Unif}(B_{\mathbf{0}}(r))$
2: $\tilde{\mathbf{x}}_0 \leftarrow \hat{\mathbf{x}} + \boldsymbol{\xi}$
3: **for** $i = 0, 1, \ldots t_{\text{thres}}$ **do**
4:      **if** $f(\hat{\mathbf{x}}) - f(\tilde{\mathbf{x}}_i) \geq f_{\text{thres}}$ **then**
5:          **return** $\tilde{\mathbf{x}}_i$
6:      $\tilde{\mathbf{x}}_{i+1} \leftarrow \tilde{\mathbf{x}}_i - \eta q(\tilde{\mathbf{x}}_i, h_{low})$
7: **return** $\hat{\mathbf{x}}$

Algorithm 14 Random search PI (RSPI).
1: **Parameters** $\sigma_1, \sigma_2 > 0$
2: Initialize x_0 at random
3: **for** $k = 0, 2, 4, \cdots 2K$ **do**
4:      $s_1 \sim S^{d-1}$ (uniformly)
5:      $x_{k+1} = \arg\min\{f(x_k), f(x_k + \sigma_1 s_1), f(x_k - \sigma_1 s_1)\}$
6:      $s_2 = \text{DFPI}(x_k)$
7:      $x_{k+2} = \arg\min\{f(x_{k+1}), f(x_{k+1} + \sigma_2 s_2), f(x_{k+1} - \sigma_2 s_2)\}$
8:      Optional: Update σ_1 and σ_2

Algorithm 15 Derivative-Free Power Iteration (DFPI)
1: **Parameters** $c, r, \eta > 0$ and $T_{\text{DFPI}} \in \mathbb{Z}^+$
2: **Inputs :** $x \in \mathbb{R}^d$,
3: $s_2^{(0)} \sim S^{d-1}$ (uniformly)
4: **for** $t = 0 \ldots T_{\text{DFPI}} - 1$ **do**
5:      $g_+ = \sum\limits_{i=1}^{d} \frac{f(x + r \cdot s_2^{(t)} + c \cdot e_i) - f(x + r \cdot s_2^{(t)} - c \cdot e_i)}{2c} e_i$
6:      $g_- = \sum\limits_{i=1}^{d} \frac{f(x - r \cdot s_2^{(t)} + c \cdot e_i) - f(x - r \cdot s_2^{(t)} - c \cdot e_i)}{2c} e_i$
7:      **Update**: $s_2^{(t+1)} = s_2^{(t)} - \eta \frac{g_+ - g_-}{2r}$
8:      Normalize $s_2^{(t+1)} = s_2^{(t+1)} / \|s_2^{(t+1)}\|$
9: **Return :** $s_2^{(T_{\text{DFPI}})}$

G.2 Parameter Settings of the Octopus Function Experiment

The detailed parameter settings of the octopus function experiment are stated in the following table.

Table 2: Choices of parameters for the experiment of the octopus function.

Algorithm	Parameters
d = 10, 30, 50, 100	
ZPSGD	$\ell = e, \rho = 1e - 4, r = \epsilon, \eta = \frac{1}{2\ell}, m = d$
PAGD	$\ell = e, \rho = e, \eta = \frac{1}{4\ell}, r = \frac{e}{100}, t_{\text{thresh}} = 1, g_{\text{thresh}} = \frac{e\gamma}{100}$
ZO-GD-NCF	$\ell = e, \rho = e, \eta = \frac{1}{4\ell}, p = 0.01$
d = 10	
RSPI	$\ell = e, \rho = e, \sigma_1 = 1, \sigma_2 = 1.25, \rho_{\sigma_1} = 0.95, T_{\sigma_1} = 20, T_{\text{DFPI}} = 20$
d = 30	
RSPI	$\ell = e, \rho = e, \sigma_1 = 1, \sigma_2 = 1.25, \rho_{\sigma_1} = 0.95, T_{\sigma_1} = 20, T_{\text{DFPI}} = 20$
d = 50	
RSPI	$\ell = e, \rho = e, \sigma_1 = 1, \sigma_2 = 1.25, \rho_{\sigma_1} = 0.85, T_{\sigma_1} = 20, T_{\text{DFPI}} = 20$
d = 100	
RSPI	$\ell = e, \rho = e, \sigma_1 = 0.75, \sigma_2 = 0.5, \rho_{\sigma_1} = 0.9, T_{\sigma_1} = 20, T_{\text{DFPI}} = 20$

G.3 Comparison between ZO-GD-NCF and Neon2+GD on Octopus Function

In order to show that our ZO-GD-NCF method will not significantly increase the iteration complexity compared to Neon2+GD [4], we compare the iteration performance between ZO-GD-NCF and Neon2+GD on octopus function. The parameters corresponding to the octopus function are set with $\tau = e, L = e, \gamma = 1$. We initialize the two algorithms at point $(0, \ldots, 0)^\mathsf{T}$.

We set $\epsilon = 1e - 4, \delta = \sqrt{\rho\epsilon}$ for all experiments and report the function value v.s. number of iterations in Figure 2.

53

(a) d=30   (b) d=50
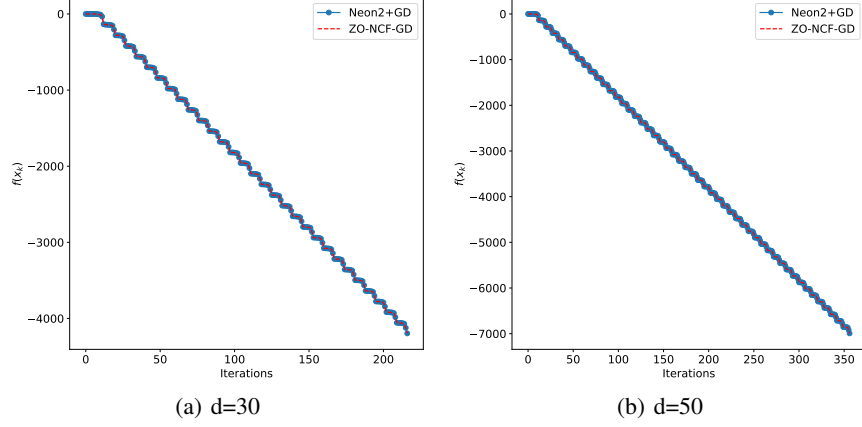
Figure 2: Comparison of ZO-GD-NCF and Neon2+GD for solving octopus function problem.

The results in Figure 2 clearly shows that ZO-GD-NCF have almost the same iteration performance with Neon2+GD. The detailed parameter settings are stated in the following table.

Table 3: Choices of parameters of ZO-GD-NCF and Neon2+GD.

Algorithm	Parameters
ZO-GD-NCF	$\ell = e, \rho = e, \eta = \frac{1}{2\ell}, p = 0.01$
Neon2+GD	$\ell = e, \rho = e, \eta = \frac{1}{2\ell}, p = 0.01$

G.4 Cubic Regularization Problem

To test performance of the proposed methods for both deterministic setting and stochastic setting. We consider the cubic regularization problem [34], which is defined as:

$$\min_{w \in \mathbb{R}^d} \frac{1}{2} w^\mathsf{T} A w + b^\mathsf{T} w + \frac{\alpha}{3} \|w\|^3. \tag{18}$$

For the deterministic setting, we generate a diagonal A such that 10% randomly selected diagonal entries are -1 and the rest diagonal entries are uniformly randomly chosen from [1, 2], and set b to a zero vector. For the stochastic setting, we let $A = A' + \mathbb{E}[\mathrm{diag}(\xi)]$ and $b = \mathbb{E}[\xi']$, where A' is generated the same way as that in the deterministic setting, ξ are uniformly randomly chosen from [-0.1, 0.1] and ξ' are uniformly randomly chosen from $[-1, 1]$. The parameter α in Eq. (18) is set to 0.5 for both deterministic setting and stochastic setting. We set $\epsilon = 10^{-2}, \delta = \sqrt{\rho\epsilon}$ for all experiments. To test the ability of different algorithms to escape from saddle points, we initialize all algorithms at a saddle point $(0, \ldots, 0)^\mathsf{T}$.

(a) Deterministic, d=100   (b) Deterministic, d=200   (c) Stochastic, d=20   (d) Stochastic, d=100
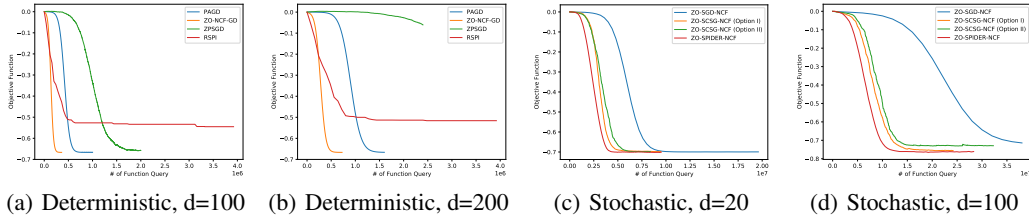
Figure 3: Comparison of different algorithms for solving cubic regularization problem in deterministic setting and stochastic setting.

In deterministic setting, the results in Fig. 3 (a) and (b) illustrate that, 1) the negative curvature finding based algorithms (ZO-GD-NCF, RSPI) can escape saddle points more efficient than the random

54

perturbation based algorithms (ZPSGD, PAGD). 2) On the other hand, the gradient estimation based algorithms (ZO-GD-NCF, ZPSGD, PAGD) converge faster than the random search based algorithm (RSPI). In the stochastic setting, ZO-SPIDER-NCF converges faster than other three algorithms.

Table 4: Choices of parameters for the experiment of the cubic regularization problem.

Algorithm	Parameters
Deterministic, d = 100	
ZPSGD	$\ell = 10^2, \rho = 1, r = \epsilon, \eta = \frac{1}{2\ell}, m = d$
PAGD	$\ell = 10^2, \rho = 1, \eta = \frac{1}{4\ell}, r = 10^{-2}, t_{\text{thresh}} = 1, g_{\text{thresh}} = 10^{-2}$
RSPI	$\ell = 10^2, \rho = 1, \sigma_1 = 0.4, \sigma_2 = 0.4, \rho_{\sigma_1} = 0.98, T_{\sigma_1} = 10, T_{\text{DFPI}} = 100$
ZO-GD-NCF	$\ell = 10^2, \rho = 1, \eta = \frac{1}{4\ell}, p = 0.01$
Deterministic, d = 200	
ZPSGD	$\ell = 10^2, \rho = 1, r = \epsilon, \eta = \frac{1}{2\ell}, m = d$
PAGD	$\ell = 10^2, \rho = 1, \eta = \frac{1}{4\ell}, r = 10^{-2}, t_{\text{thresh}} = 1, g_{\text{thresh}} = 10^{-2}$
RSPI	$\ell = 10^2, \rho = 1, \sigma_1 = 0.4, \sigma_2 = 0.4, \rho_{\sigma_1} = 0.98, T_{\sigma_1} = 10, T_{\text{DFPI}} = 100$
ZO-GD-NCF	$\ell = 10^2, \rho = 1, \eta = \frac{1}{4\ell}, p = 0.01$
Stochastic, d = 20, d=100	
ZO-SGD-NCF	$\ell = 10^2, \rho = 1, \eta = \frac{1}{3\ell}, |S| = 128, p = 0.01$
ZO-SCSG-NCF	**Option I**: $\ell = 10^2, \rho = 1, \eta = \frac{1}{4\ell}, B = 128, b = 10, p = 0.01$
ZO-SCSG-NCF	**Option II**: $\ell = 10^2, \rho = 1, \eta = \frac{1}{10\ell}, B = 128, b = 10, p = 0.01$
ZO-SPIDER-NCF	$\ell = 10^2, \rho = 1, \eta = \frac{1}{15\ell}, |\mathcal{S}_1| = 128, |\mathcal{S}_2| = 10, p = 0.01$

G.5 Regularized Non-Linear Least-Square Problem

We next consider the regularized non-linear least-square problem [34], which is defined as:

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (y_i - \sigma(w^\top x_i))^2 + \sum_{i=1}^{d} \frac{\lambda w_i^2}{1 + \alpha w_i^2}, \tag{19}$$

where $x_i \in \mathbb{R}^d$, $y_i \in \{0, 1\}$, $\sigma(s) = \frac{1}{1 + \exp(-s)}$, and the second term is a non-convex regularizer. We use the w1a data (n=2477, d=300) which can be downloaded from the LIBSVM website [12]. We set $\lambda = 1, \alpha = 1$ in Eq.(19) and $\epsilon = 1e - 2, \delta = \sqrt{\rho\epsilon}$ for both deterministic setting and stochastic setting. We report the objective function value v.s. the number of function queries in Figure 4 and can draw similar conclusions to the previous experiment.
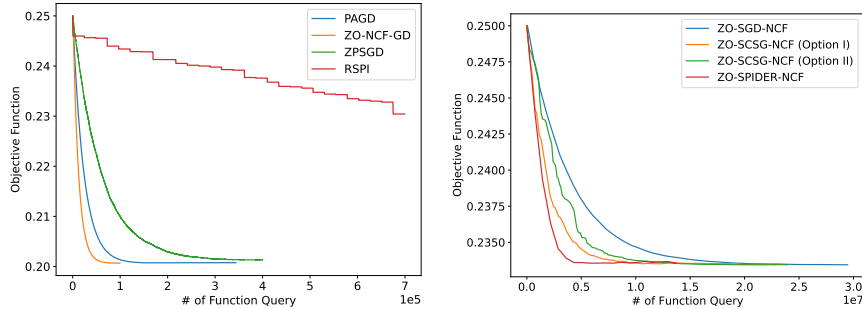
Figure 4: Comparison of different algorithms for solving regularized non-linear least-square problem in deterministic setting and stochastic setting

55

Table 5: Choices of parameters for the experiment of the regularized non-linear least-square problem.

Algorithm	Parameters
Deterministic	
ZPSGD	$\ell = 10^2, \rho = 1, r = \epsilon, \eta = \frac{1}{2\ell}, m = d$
PAGD	$\ell = 10^2, \rho = 1, \eta = \frac{1}{4\ell}, r = 10^{-2}, t_{\text{thresh}} = 1, g_{\text{thresh}} = 10^{-2}$
RSPI	$\ell = 10^2, \rho = 1, \sigma_1 = 0.4, \sigma_2 = 0.4, \rho_{\sigma_1} = 0.98, T_{\sigma_1} = 10, T_{\text{DFPI}} = 100$
ZO-GD-NCF	$\ell = 10^2, \rho = 1, \eta = \frac{1}{4\ell}, p = 0.01$
Stochastic	
ZO-SGD-NCF	$\ell = 10^2, \rho = 1, \eta = \frac{1}{3\ell}, |S| = 128, p = 0.01$
ZO-SCSG-NCF	**Option I** : $\ell = 10^2, \rho = 1, \eta = \frac{1}{4\ell}, B = 128, b = 10, p = 0.01$
ZO-SCSG-NCF	**Option II** : $\ell = 10^2, \rho = 1, \eta = \frac{1}{10\ell}, B = 128, b = 10, p = 0.01$
ZO-SPIDER-NCF	$\ell = 10^2, \rho = 1, \eta = \frac{1}{15\ell}, |\mathcal{S}_1| = 128, |\mathcal{S}_2| = 10, p = 0.01$

56