On the Oracle Complexity of Higher-Order Smooth Non-Convex Finite-Sum Optimization

Nicolas Emmenegger
ETH Zurich

Rasmus Kyng
ETH Zurich

Ahad N. Zehmakan
ETH Zurich

Abstract

We prove lower bounds for higher-order methods in smooth non-convex finite-sum optimization. Our contribution is threefold: We first show that a deterministic algorithm cannot profit from the finite-sum structure of the objective, and that simulating a pth-order regularized method on the whole function by constructing exact gradient information is optimal up to constant factors. We further show lower bounds for randomized algorithms and compare them with the best known upper bounds. To address some gaps between the bounds, we propose a new second-order smoothness assumption that can be seen as an analogue of the first-order mean-squared smoothness assumption. We prove that it is sufficient to ensure state-of-the-art convergence guarantees, while allowing for a sharper lower bound.

1 INTRODUCTION

Many problems in machine learning can be formulated as empirical risk minimization, viewing the loss function on each data point as a component function f_i in a sum. This problem can then be cast as minimizing an objective function $F : \mathbb{R}^d \to \mathbb{R}$, $F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$ under a variety of smoothness assumptions, where in each iteration, the derivatives of the loss on a single data point can be queried. The ultimate goal would be to find

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}).$$

Without relying on convexity, which is an unrealistic assumption in many machine learning applications (e.g. training neural networks (LeCun et al., 2015) or robust linear regression (Yu and Yao, 2017)), we consider the case where the objective is possibly non-convex. In this setting, finding such a global minimum is in general NP-Complete (Murty and Kabadi, 1987), so theoretical guarantees are expressed in terms of weaker requirements. Inspired by necessary conditions for minima, customary guarantees are approximate first-order or second-order stationary points (FOSP, SOSP). We will focus here on the oracle complexity of finding an ε-approximate first-order stationary point of F, that is a point $\mathbf{x}$, such that $\|\nabla F(\mathbf{x})\| \le \varepsilon$. This is a standard problem formulation commonly studied in the literature (Carmon et al., 2019a,b; Arjevani et al., 2019; Fang et al., 2018; Zhou and Gu, 2019; Zhou et al., 2019; Zhou and Gu, 2020).

When machine learning data sets are large, gradients are often approximated by evaluating only a subset of all training examples (Bottou et al., 2018). This leads to a model where in each iteration of an algorithm, one f_i's derivative information can be queried. In this model, the most prevalent algorithms today are stochastic gradient descent (SGD) and variants thereof. More query efficient algorithms have been explored too. Variance reduction techniques – first introduced for convex optimization (Johnson and Zhang, 2013) – have been successfully applied in the non-convex setting (Allen-Zhu and Hazan, 2016; Reddi et al., 2016; Lei et al., 2017; Fang et al., 2018). These algorithms draw their speedup from cleverly constructed low-variance gradient estimators. On the other hand, higher-order information provably

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

helps speed up the convergence, and can potentially be harnessed to get guarantees in terms of SOSPs (Nesterov and Polyak, 2006; Birgin et al., 2017; Carmon et al., 2019a). Motivated by this fact there have been successful attempts to apply variance reduction techniques to higher-order algorithms, in order to make practically relevant subsampled variants of classical procedures like the cubic regularized Newton's method (Zhou et al., 2019; Zhou and Gu, 2020).

These results naturally raise complexity-theoretic questions and warrant the study of corresponding lower bounds. In their seminal work Carmon et al. (2019a) show that the optimal rate for the $n = 1$, non-convex case is $\Theta(\varepsilon^{-(p+1)/p})$, by giving a non-convex variant of Nesterov's "worst function in the world" (Nesterov, 2004). However, as mentioned before, modern machine learning systems are typically working with datasets so large that computing the full derivative information in each step of the algorithm is impractical. It is therefore imperative to understand the complexity of higher-order methods in the finite-sum case.

We give the first lower bound results for the problem of finding an approximate stationary point of a sum of pth-order individually smooth non-convex functions, in a model where the algorithm queries the derivatives of individual functions at each time-step. We provide lower bounds for both deterministic and randomized algorithms. An overview of our results is given in Table 1.

First we consider deterministic algorithms and show that a pth-order regularized method that constructs the full derivative at each iteration is optimal up to constant factors. We use an adversarial construction that forces the algorithm to spend a large number of queries to discover useful information. To the best of our knowledge, this result is also new for the widely studied case of first-order smooth non-convex finite-sum optimization and implies that gradient descent on the full function is optimal up to constant factors. With this result we demonstrates a clear separation between deterministic and randomized algorithms.

1.1 Our Contributions

Further, we give the first lower bounds for randomized algorithms in this setting, which allow for evaluation of the new line of research of higher-order variance-reduction. We derive the bounds with a probabilistic construction, building on the family of zero-chain functions first introduced by Carmon et al. (2019a). In contrast to the first-order case studied by Zhou and Gu (2019), we show a non-trivial dependence on n for the $p > 1$ regime. Our bounds indicate that variance reduction indeed gets harder for higher-orders of smoothness, which is consistent with practical findings (Goodfellow et al., 2016). Contrary to some prior work (Zhou and Gu, 2019; Han et al., 2021; Zhang et al., 2021), we do not need the assumption that the points queried by the algorithm must lie in the span of previously queried points and oracle responses. Under this assumption, lower bounds would not apply to algorithms that add noise at each step (e.g. stochastic gradient langevin dynamics (Welling and Teh, 2011)) or purposefully break with the span assumption to get faster convergence rates (Hannah et al., 2018).

Table 1: A comprehensive overview of the upper and lower bounds for first-, second- and higher-order oracle models. We assume each f_i has Lipschitz pth-order derivative tensor. The first row refers to deterministic algorithms while the three below concern the randomized setting. Our contributions are highlighted in grey.

		Upper bound	Lower bound
Deter.		$\mathcal{O}(n\varepsilon^{-\frac{p+1}{p}})$ [a]	$\Omega(n\varepsilon^{-\frac{p+1}{p}})$ [b]
Rand.	$p = 1$	$\mathcal{O}(n^{\frac{1}{2}}\varepsilon^{-2})$ [c]	$\Omega(\varepsilon^{-2})$ [d;f]
	$p = 2$	$\tilde{\mathcal{O}}(n^{\frac{4}{5}}\varepsilon^{-\frac{3}{2}})$ [e]	$\Omega(n^{\frac{1}{4}}\varepsilon^{-\frac{3}{2}})$ [f]
	$p > 2$	$\mathcal{O}(n\varepsilon^{-\frac{p+1}{p}})$ [a]	$\Omega(n^{\frac{p-1}{2p}}\varepsilon^{-\frac{p+1}{p}})$ [f]

[a] Birgin et al. (2017) [b] Theorem 3.5 [c] Fang et al. (2018) [d] Zhou and Gu (2019) [e] Zhou et al. (2019) [f] Theorem 4.7

All our bounds are tight in terms of ε dependence, but some gaps with respect to the dependence on n remain. To alleviate this gap in the second-order case, we introduce a new, weaker notion of second-order smoothness and show that it is sufficient to guarantee state-of-the-art oracle complexities for variance-reduced methods, while allowing for a tighter lower bound. To upper bound the oracle complexity in this new setting, we show that the variance of SVRC's (Zhou et al., 2019) Hessian and gradient estimators can be controlled via the second-order mean-cubed smoothness of the finite-sum function. Table 2 shows

Table 2: Lower and upper bounds for randomized second-order methods under smoothness of individual functions f_i (Indiv.) and the third-moment smoothness of Assumption 4.8 (Avg.). Our contributions are again highlighted in grey.

	Upper bound	Lower bound
Indiv.	$\tilde{\mathcal{O}}(n^{\frac{4}{5}}\varepsilon^{-\frac{3}{2}})$ [a]	$\Omega(n^{\frac{1}{4}}\varepsilon^{-\frac{3}{2}})$ [b]
Avg.	$\tilde{\mathcal{O}}(n^{\frac{4}{5}}\varepsilon^{-\frac{3}{2}})$ [c]	$\Omega(n^{\frac{5}{12}}\varepsilon^{-\frac{3}{2}})$ [d]

[a] Zhou et al. (2019) [b] Theorem 4.7 [c] Theorem 4.9 [d] Theorem 4.10

our bounds and contrasts them with our results for individually smooth functions.

1.2 Related Work

1.2.1 Stochastic Second-Order Methods

While there exist approaches exploiting third-order derivatives (Lucchi and Kohler, 2019), most work has focused on Hessian based algorithms: Zhou et al. (2019) give a method (SVRC) that uses only $\tilde{\mathcal{O}}(n^{4/5}\varepsilon^{-3/2})$ [1] second-order oracle queries to find a SOSP under a second-order smoothness assumption on each of the f_is. Shen et al. (2019) provide an even faster trust-region method (STR2) that achieves the second-order oracle complexity of $\tilde{\mathcal{O}}(n^{3/4}\varepsilon^{-3/2})$, but under the stronger assumption that the gradient is Lipschitz continuous as well. Finally, we point out that there is a line of research which tries to minimize Hessian complexity at the cost of additional gradient queries (Shen et al., 2019; Zhou and Gu, 2020). For the higher-order oracle complexity measure that we focus on here, SVRC and STR2 represent the best known upper bounds for second-order randomized algorithms. As we only assume pth-order smoothness, we take SVRC (Zhou et al., 2019) as reference for second-order methods.

1.2.2 Related Work on Lower Bounds

Lower bounds for smooth non-convex optimization have all built on the works of Carmon et al. (2019a,b). This line of work focuses on the case where the objective is composed of a single smooth function (i.e., $n = 1$) and full derivative information is available at each iteration. Among other things, they establish the optimal rate of $\Theta(\varepsilon^{-(p+1)/p})$ to find ε-approximate FOSPs for algorithms having access to as much derivative information as needed under the assumption that the function is pth-order smooth. We build on this by generalizing to the finite-sum case. On the other hand, our work also draws on that of Fang et al. (2018), which show a lower bound for the first-order finite-sum case under the mean-squared smoothness assumption. We generalize this to arbitrary orders of smoothness, deriving bounds which suggest that reducing the variance is harder for higher-orders of smoothness. We also propose a third-moment smoothness assumption on the Hessian that can be seen as the higher-order analogue of the mean-squared smoothness assumption.

Furthermore, Zhou and Gu (2019) prove lower bounds on first-order algorithms for a variety of regimes in finite-sum optimization, including the non-convex case. A shortcoming of their results is that they place a linear-span restriction on the algorithms in question. This assumption may be violated, and our work does not rely on it, which makes our bounds more future proof. Indeed, there are indications that breaking with the span assumption yields better rates (Hannah et al., 2018). Another lower bounds paper that shares some settings in common with ours is a recent survey from Han et al. (2021).

It is worth noting that Arjevani et al. (2019) and Arjevani et al. (2020) prove lower bounds for a different stochastic setting. In this model one does not assume a finite-sum structure, but typically places variance assumptions on the queried derivative information. The first paper focuses on first-order stationary points, while the second is considering approximate local minima and higher-order algorithms. Interestingly, they show that information beyond second-order is not useful in stochastic non-convex optimization (Arjevani et al., 2020). This is very opposed to the deterministic finite-sum setting we cover, where higher-order smoothness results in substantially more query efficient algorithms. Our results suggest that the finite-sum randomized setting lies in between these two: higher-order information helps to get a better dependency on the accuracy parameter ε, but the higher-order smoothness alone makes the variance of an unbiased derivative estimator harder to control, yielding an increasing dependence in n.

[1] We use $\tilde{\mathcal{O}}$ to hide polylogarithmic factors in d, n and $1/\varepsilon$

2 PRELIMINARIES

2.1 Problem Description

We focus on finding ε-approximate first-order stationary points. We assume that a problem instance is a function $F = \frac{1}{n}\sum_{i=1}^{n} f_i$, which satisfies the following assumption.

Assumption 2.1. *We say $F \in \mathcal{F}_p^n(\Delta, L_p)$ if for some d, $F : \mathbb{R}^d \to \mathbb{R}$, $\mathbf{x} \mapsto \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$ satisfies the following properties*

i) Each function f_i is pth-order smooth, i.e. its pth-order derivative tensor is Lipschitz continuous w.r.t. to the tensor operator norm:

$$\|\nabla^p f_i(\mathbf{x}) - \nabla^p f_i(\mathbf{y})\| \leq L_p \|\mathbf{x} - \mathbf{y}\|.$$

ii) Assuming that an algorithm starts at iterate $\mathbf{x}_0 = \mathbf{0}$, the initial gap to optimality is bounded by

$$\frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}_0) - \inf_{\mathbf{x}} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \leq \Delta.$$

Whenever n, p, L_p and Δ are obvious from context, we say that F satisfies Assumption 2.1 if $F \in \mathcal{F}_p^n(\Delta, L_p)$.

2.2 Algorithm and Oracle Models

Usually, when pth-order smoothness is assumed, one works with derivatives up to the pth order. Therefore, in the interest of deriving lower bounds, it is even stronger to let the algorithm have access to as many derivatives as it could possibly require. It turns out that this actually will not change the bounds, and that they depend only on the order of smoothness p of the considered function. We assume that an algorithm queries iterates according to the following definition, and we will lower bound the number of such queries it needs to do to reach its objective.

Assumption 2.2. *In the incremental higher-order oracle model (IHO), an oracle for a function $F = \frac{1}{n}\sum f_i$ consists of a mapping*[2]

$$\mathsf{O}_F^{(q)} : [n] \times \mathbb{R}^d \to \left(\mathbb{R}, \mathbb{R}^d, ..., \mathbb{R}^{\otimes^q d}\right)$$
$$(i, \mathbf{x}) \mapsto \nabla^{(0:q)} f_i(\mathbf{x}).$$

We condense the notation by letting $\mathsf{O}_F^{(q)}(i^{0:t-1}, \mathbf{x}^{(0:t-1)})$ correspond to the union of all oracle responses before iteration t, i.e.

At this point we would like to mention that this is a widely used model, but it is not the only one. There are methods like Lite-SVRC (Zhou et al., 2018; Wang et al., 2019; Zhou et al., 2019) and STR1 (Shen et al., 2019) that minimize the number of Hessian queries at the cost of slightly more gradient queries, since gradient computations are less expensive. Thus, a model might allow for queries to different oracles. Finding lower bounds in a model where multiple query complexities are distinguished between has been done in distributed convex optimization (e.g. Woodworth et al. (2021)), where gradient oracle queries and communication between workers are assumed to incur separate costs. We leave exploring related ideas in our setting to future work.

We will now think of an algorithm as generating a sequence of indices and iterates, namely those it queries the IHO on.

Assumption 2.3. *We will assume that an algorithm A has access to an infinite sequence of random bits $\xi \sim \mathcal{U}([0,1])$ drawn at the beginning of the procedure.* [3] *Then, A consists of a sequence of mappings $\{A^{(t)}\}_{t\in\mathbb{N}}$ which produce indices and iterates $[i^t, \mathbf{x}^{(t)}]$ based on previous oracle responses:*

$$A^{(t)}\left\{\xi, i^{0:t-1}, \mathbf{x}^{(0:t-1)}, \mathsf{O}_F^{(q)}(i^{0:t-1}, \mathbf{x}^{(0:t-1)})\right\}.$$

Without loss of generality, we set $\mathbf{x}^{(0)} = \mathbf{0}$ because if a function f is difficult to optimize for starting point $\mathbf{0}$, then $\mathbf{x} \mapsto f(\mathbf{x} - \mathbf{x}^{(0)})$ is difficult to optimize for starting point $\mathbf{x}^{(0)}$. Finally, we set no restrictions on how i^0 is chosen.

Note that this is quite a general assumption, merely capturing the fact that the algorithm performs some arbitrary computation between different queries. It is worth mentioning that in our setting, any potential randomness is inside the algorithm and not the oracle.

2.3 Complexity Measure

Finally, we need a proper measure to characterize the complexity of an algorithm. We choose the following.

Definition 2.4. *We define the oracle complexity $T_\varepsilon(\mathsf{A}, F)$ of an algorithm A on F as the infimum*

[2] We write $i : j$ or $[i : j]$ for the set of integers $\{i, \ldots, j\}$ and let $[m] := [1 : m]$. Furthermore, we define $\mathbb{R}^{\otimes^k d}$ to be the space of k-dimensional tensors over $\mathbb{R}^d$. We denote by $\nabla^{(0:q)}$ the union of derivative tensors up to the order q.

[3] For a deterministic algorithm, we simply assume the sequence is fixed.

over all $t \in \mathbb{N}$ such that the following holds with probability at most $\frac{1}{2}$:

$$\forall s \leq t \, : \, \|\nabla F(\mathbf{x}^{(s)})\| > \varepsilon.$$

In other words, this corresponds to t such that for all larger t', with probability $1/2$ the algorithm will encounter an iterate $s \leq t'$ with sufficiently small gradient.

We note that $T_\varepsilon(\mathsf{A}, F) \geq t$ implies that for all $s \leq t$, $P(\|\nabla F(\mathbf{x}^{(s)})\| > \varepsilon) \geq 1/2$, and so by Markov's inequality, $\varepsilon/2 \leq \varepsilon P(\|\nabla F(\mathbf{x}^{(s)})\| > \varepsilon) \leq \mathbb{E}\|\nabla F(\mathbf{x}^{(s)})\|$ for all $s \leq t$. This implies that we can also compare our lower bounds to the methods which give guarantees in terms of a complexity that ensures an output with a small gradient *in expectation.*

3 DETERMINISTIC METHODS

In this section, we show that any algorithm that can not resort to randomness can outperform only by a constant factor one that simulates a higher-order regularized method (Birgin et al., 2017). By the latter, we mean a procedure which constructs the full derivative information at each step by querying all n functions.

Inspired by Carmon et al. (2019a) and Woodworth and Srebro (2016), we define a family of hard instances that we will later instantiate depending on the algorithm's behaviour. The main intuition is to utilize an underlying function which has a large gradient as long as there are coordinates left which are very close to zero. Depending on the queries of the algorithm, we will adversarially and incrementally choose a rotation of the input space in such a way that these coordinates indeed stay close to zero for a long time.

Definition 3.1. *Let $K \in \mathbb{N}$ and for $k \in [K]$ let $\delta_k \in \{0, 1\}$ be arbitrary. We define the function $f_{K,\boldsymbol{\delta}} : \mathbb{R}^K \to \mathbb{R}$ as*

$$\begin{aligned} f_{K,\delta}(\mathbf{x}) &:= -\delta_1 \Psi(1)\Phi(x_1) \\ &\quad + \sum_{k=2}^{K} \delta_k \left[\Psi(-x_{k-1})\Phi(-x_k) - \Psi(x_{k-1})\Phi(x_k)\right], \end{aligned}$$

where the functions Φ and Ψ are given by

$$\Psi(x) := \begin{cases} 0 & x \leq 1/2 \\ \exp\left(1 - \frac{1}{(2x-1)^2}\right) & \textit{otherwise} \end{cases}$$

and

$$\Phi(x) := \sqrt{e} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} \mathrm{d}t.$$

We should emphasize that the function $\bar{f}_K$ defined by Carmon et al. (2019a) can be represented by $f_{K,\mathbf{1}}$.

For the remaining parts of this section, assume the algorithm A, the number of functions n and parameters Δ and L_p to be fixed. The idea is to construct n functions of the above family, where each function $f_i(\mathbf{x})$ will be given (modulo rescaling) by $f_{K+1,\boldsymbol{\delta}_i}(\mathbf{V}^T\mathbf{x})$ for some suitable $\boldsymbol{\delta}_i \in \{0, 1\}^{K+1}$ and shared $\mathbf{V} \in \mathbb{R}^{d \times K+1}$. This can be seen as a non-convex instantiation of the ideas introduced by Woodworth and Srebro (2016) for first-order algorithms. We will split up the iterates of the algorithm in rounds, starting at $k = 2$ and ending at $k = K + 1$. Thus after round k, in total $k - 1$ rounds will have elapsed. We define a round to span queries to $\lceil n/2 \rceil$ *different* functions. With those concepts in hand, we define the hard instance as:

Definition 3.2. *For $i \in [n]$ let $\delta_{i,1} = \mathbf{1}[i \leq \lceil n/2 \rceil]$. For $k \in [2 : K + 1]$ let $\delta_{i,k} = 1$ iff A does not query function i during round k. Further, let $d \geq K+1$ and let $\mathbf{V} \in \mathsf{Ortho}(d, K+1)$ be a matrix with orthonormal columns. Let $\lambda, \sigma > 0$ be parameters we will fix later. Then, we define*

$$f_i(\mathbf{x}) = \lambda\sigma^{p+1} f_{K+1,\boldsymbol{\delta}_i}\left(\mathbf{V}^T\mathbf{x}/\sigma\right),$$

and consequently $F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$.

We now prove that there exists an adversarial rotation with the following property:

Lemma 3.3. *In Definition 3.2, $\mathbf{V}$ can be chosen such that for the sequence of indices and iterates $\{[i^t, \mathbf{x}^{(t)}]\}$ that algorithm A produces up to the end of round $K + 1$, we have $\langle \mathbf{v}_{K+1}, \mathbf{x}^{(t)} \rangle = 0$ for all t.*

A key property of the function $\bar{f}_K = f_{K,\mathbf{1}}$ is that as long as the last coordinate in its input is zero, the gradient of the function will be lower bounded by a constant. This property can be transferred to $F = \frac{1}{n}\sum f_i$:

Lemma 3.4. *For all iterates up to the end of round $K + 1$, we have $\left(\mathbf{V}^T\mathbf{x}^{(t)}\right)_{K+1} = 0$, and consequently*

$$\|\nabla F(\mathbf{x}^{(t)})\| > \lambda\sigma^p/4.$$

To show the main result, we have to set the scaling parameters such that our function respects Assumption 2.1. Note that λ controls the smoothness parameter, σ controls the gradient norm lower bound and K needs to be chosen as large as possible, but in a way that makes F respect the initial optimality gap Δ. Together, they can be chosen to imply the theorem below.

Theorem 3.5. *For any p and deterministic algorithm A satisfying Assumption 2.3, for any n, Δ and L_p and ε there exists a function $F \in \mathcal{F}_p^n(\Delta, L_p)$ such that*

$$T_\varepsilon(\mathsf{A}, F) \geq \Omega\left((L_p/\ell_p)^{1/p}\,\Delta n \varepsilon^{-(p+1)/p}\right), \qquad (1)$$

where the constant factors hidden by Ω do not depend on n, ε or p and $\ell_p \leq \exp\left(\frac{5}{2}p\log p + cp\right)$ for some constant $c < \infty$. Moreover, the dimension of this function merely needs to be of the same order as (1).

To summarize, we get a lower bound of $\Omega\left(n\varepsilon^{-(p+1)/p}\right)$. In the noiseless $n = 1$ setting, the optimal complexity is characterized by $\Theta(\varepsilon^{-(p+1)/p})$ (Carmon et al., 2019a). Indeed, Birgin et al. (2017) prove this to be achievable with higher-order regularized methods, subsuming results known for gradient descent and cubic regularization. These methods also imply that Theorem 3.5 characterizes the optimal oracle complexity, as we can simulate a higher-order regularized method by spending n queries at each iterate.

4 RANDOMIZED METHODS

When constructing hard instances for randomized algorithms, one does not have the luxury of reacting to the algorithm's queries, because we cannot anticipate the random seed ξ. The approach taken here is to draw orthogonal vectors from some high-dimensional space and show that with some fixed probability, the iterates of the algorithm will be close to orthogonal to this set of important directions.

The analysis of those constructions is quite intricate, and the dimension of the domain required is larger, even more so when dropping the assumption that the iterates stay in the span of the queried derivatives (as assumed, e.g. by Zhou and Gu (2019)). We will first provide a sketch of the main argument, and then go on to state the key results.

To reason about the construction, a very useful notion is that of a higher-order (robust) zero-chain (Carmon et al., 2019a).

Definition 4.1 [Robust zero-chain]**.** *A function $f : \mathbb{R}^d \to \mathbb{R}$ is a robust zero-chain if for all $i \in [d]$ and $\mathbf{x} \in \mathbb{R}^d$ the following implication holds: If $|x_j| < \frac{1}{2}$ for all $j \geq i$, then*

$$\forall \mathbf{y} \in N(\mathbf{x}) \,:\, f(\mathbf{y}) = f(y_1, \ldots, y_i, 0, \ldots, 0),$$

where $N(\mathbf{x})$ denotes an open neighborhood of $\mathbf{x}$.

One can observe that the partial derivatives of such a function f at $\mathbf{x}$ are zero for all indices $j > i$, which is the key to ensure that the oracle responses give away information slowly.

Recall the function $\bar{f}_K = f_{K,\mathbf{1}}$ from Definition 3.1. This function is a robust zero-chain for any $K \geq 1$ (Carmon et al., 2019a). Since it also has the desirable property that its gradient is large as long as there remain coordinates which are close to zero, we can exploit copies of it in a lower bound construction. Instead of using a single matrix $\mathbf{V}$ to rotate the input adversarially, we will follow Fang et al. (2018) and use n different matrices $\mathbf{B}_i$ with orthogonal columns drawn at random and prove that with some fixed probability for a large number of iterates

$$\langle \mathbf{b}_{i,K}, \mathbf{x}^{(t)} \rangle < 1/2.$$

This will (similarly as in the deterministic case) imply that the gradient of function i is bounded from below. We will refer to the process of the algorithm finding inputs that make these inner products large as "discovering" coordinates.

Definition 4.2 [Finite-sum hard distribution]**.** *Let $n, K \in \mathbb{N}$. Let $d \in \mathbb{N}$ be divisible by n and let $R = 230\sqrt{K}$. Then draw $\mathbf{B} = \left[\mathbf{B}_1 \,|\, \cdots \,|\, \mathbf{B}_n\right] \in \mathsf{Ortho}(d/n, nK)$ uniformly at random and let $\mathbf{C} = \left[\mathbf{C}_1 \,|\, \cdots \,|\, \mathbf{C}_n\right] \in \mathsf{Ortho}(d, d)$ be arbitrary. We define our unscaled hard instance as $F^* = \frac{1}{n}\sum_{i=1}^n f_i^*$ with $f_i^*(\mathbf{x}) = \hat{f}_{K;\mathbf{B}_i}(\mathbf{C}_i^T\mathbf{x})$ where*

$$\hat{f}_{K;\mathbf{B}_i}(\mathbf{y}) := \bar{f}_K\left(\mathbf{B}_i^T \rho\left(\mathbf{y}\right)\right) + \frac{1}{10}\|\mathbf{y}\|^2$$

and

$$\rho(\mathbf{y}) := \frac{\mathbf{y}}{\sqrt{1 + \|\mathbf{y}\|^2/R^2}}.$$

Because of the random choice of matrix $\mathbf{B}$, this induces a distribution.

This construction has been used in Fang et al. (2018) to show a lower bound for the first-order mean-squared setting and the last two definitions are originally due to Carmon et al. (2019a). In the following, we generalize their result to arbitrary orders of smoothness. A brief discussion is in order: the purpose of $\mathbf{B}_i$ is as discussed above, namely using the zero-chain property of $\bar{f}_K$ to make sure that any algorithm has a hard time discovering coordinates. The composition with ρ ensures that an algorithm cannot simply make the iterates large to learn coordinates, and $\mathbf{C}_i$ will be useful to bound the gradient norm of F in terms of the $\|\nabla f_i\|$'s: exactly what we need for a lower bound.

Our goal is to derive lower bounds for any possible Lipschitz constants and optimality gaps. This means that we will scale F^* to meet the various requirements. The notion of a function-informed process (Carmon et al., 2019a) will enable us to reason about a scaled version of our function F^* while thinking about what another algorithm would do on the unscaled F^*.

Definition 4.3 [Function-informed process]**.** *We call a sequence of indices and iterates $\{[i^t, x^{(t)}]\}_{t\in\mathbb{N}}$ informed by a function F if it follows the same distribution as $\mathsf{A}[F]$ for some randomized A.*

Lemma 4.4. *Let F be an instance of a finite-sum optimization problem. Let $a, b > 0$. Consider the function $G(\mathbf{x}) = aF(\mathbf{x}/b)$ and assume $\{[i^t, \mathbf{x}^{(t)}]\}_{t\in\mathbb{N}}$ is produced by A on function G, i.e. $\mathsf{A}[G] = \{[i^t, \mathbf{x}^{(t)}]\}_{t\in\mathbb{N}}$. Then $\{[i^t, \mathbf{x}^{(t)}/b]\}_{t\in\mathbb{N}}$ is informed by F.*

Above, we hinted at the importance of small inner products of the iterates with the columns of the matrices $\mathbf{B}_i$. This intuition is formalized in the lemma below, that generalizes results from Fang et al. (2018), Woodworth and Srebro (2016) and Carmon et al. (2019a).

Lemma 4.5. *Let $\{[i^t, \mathbf{x}^{(t)}]\}_{t\in\mathbb{N}}$ be informed by F^* drawn from the distribution in Definition 4.2, let $\delta \in (0, 1)$ and $T = \frac{nK}{2}$. For any $t \in [T]$ and $i \in [n]$, let I_i^{t-1} be the number of occurrences of index i in $i^{0:t-1}$, i.e. the number of queries with index i up to iteration t (the iteration producing $\mathbf{x}^{(t)}$). Let $I_i^{-1} = 0$ by default. For any $t \in \{0, \ldots, T\}$ define $\mathcal{U}_i^{(t)}$ to be the set of the last $K - I_i^{(t-1)}$ columns of $\mathbf{B}_i$ (provided $K - I_i^{t-1} \geq 1$, otherwise the set is empty). More formally*

$$\mathcal{U}_i^{(t)} := \left\{\mathbf{b}_{i,I_i^{t-1}+1}, \ldots, \mathbf{b}_{i,K}\right\}.$$

Then the following holds for some constant $c_0 < \infty$: if $d \geq c_0 n^3 K^2 \log(\frac{n^2K^2}{\delta})$, then with probability at least $1 - \delta$ we have $\forall t \in \{0, \ldots, T\}$, $\forall i \in [n]$, $\forall \mathbf{u} \in \mathcal{U}_i^{(t)}$

$$|\langle \mathbf{u}, \rho(\mathbf{C}_i^T \mathbf{x}^{(t)})\rangle| < \frac{1}{2}.$$

The key takeaway from Lemma 4.5 is that for each index $i \in [n]$ the algorithm needs K queries to that index to learn all columns of $\mathbf{B}_i$. Consequently, the input of the zero-chain $\bar{f}_K$ stays small in absolute terms for the coordinates corresponding to columns in $\mathcal{U}_i^{(t)}$ with high probability. This is good because $\bar{f}_K$'s large-gradient property (see discussion preceding Lemma 3.4) then makes the gradient of F^* large as well:

Lemma 4.6. *Let $\{[i^t, \mathbf{x}^{(t)}]\}_{t\in\mathbb{N}}$ be informed by F^* drawn from the distribution in Definition 4.2 and let $\delta \in (0, 1)$. Then the following holds for some numerical constant $c_0 < \infty$: if $d \geq c_0 n^3 K^2 \log(\frac{n^2K^2}{\delta})$, with probability at least $1 - \delta$ we have for all $t \in \{0, \ldots, T\}$*

$$\|\nabla F^*(\mathbf{x}^{(t)})\| > 1/(4\sqrt{n}).$$

4.1 Lower Bound for the Individual Smooth Setting

To derive results for any incarnation of the function classes in Assumption 2.1, one can rescale the function and the inputs and use the above lemmas, exploiting the fact that they hold for function-informed processes. The analysis yields:

Theorem 4.7. *For any randomized algorithm A satisfying Assumption 2.3, $p \in \mathbb{N}$, Δ, L_p, ε and $n \leq c_p \Delta^{\frac{2p}{p+1}} L_p^{\frac{2}{p+1}} \varepsilon^{-2}$, there exists a dimension $d \leq \tilde{\mathcal{O}}(n^{\frac{2p-1}{p}} \Delta L_p^{2/p} \varepsilon^{-\frac{2(p+1)}{p}}) \leq \tilde{\mathcal{O}}(n^2 \Delta L_p^2 \varepsilon^{-4})$ and a function $F \in \mathcal{F}_p^n(\Delta, L_p)$ such that*

$$T_\varepsilon(\mathsf{A}, F) \geq \Omega\left(L_p{}^{\frac{1}{p}} \hat{\ell}_p{}^{-\frac{1}{p}} \Delta n^{\frac{p-1}{2p}} \varepsilon^{-\frac{p+1}{p}}\right),$$

where $\hat{\ell}_p \leq \exp(cp \log p + c)$ for some constant $c < \infty$. For fixed p, c_p is also a universal constant.

Our result is essentially a lower bound of $\Omega\left(n^{\frac{p-1}{2p}} \varepsilon^{-\frac{p+1}{p}}\right)$ for fixed p, up to constant factors. The increasing dependence on n is consistent with the empirical observation that higher-order methods typically need to employ larger batch sizes (see Section 8.1.3 in Goodfellow et al. (2016)).

For second-order algorithms, the best rate with our individual smoothness assumption is achieved by Zhou et al. (2019). Their algorithm finds an approximate local minimum in $\tilde{\mathcal{O}}(n^{4/5}\varepsilon^{-3/2})$ oracle calls. Our lower bound reads as $\Omega(n^{1/4}\varepsilon^{-3/2})$ for Assumption 2.1 with $p = 2$, which implies it exhibits a $\tilde{\mathcal{O}}(n^{11/20})$ gap.

4.2 A new Assumption for Second-Order Smoothness

We point out that a similar $n^{1/2}$ gap is present in the case of $p = 1$ (Zhou and Gu, 2019; Han et al., 2021), which remains an open problem. For the first-order setting, a way to get matching bounds is to use the first-order mean-squared smoothness assumption, yielding the optimal $\Theta(\sqrt{n}/\varepsilon^{-2})$ oracle complexity (Fang et al., 2018). It has been observed by Zhou and Gu (2019) that this assumption is sufficient for a variety of first-order methods. This raises a natural question: is there a second-order analogue to mean-squared smoothness? The mean-squared assumption effectively controls the second moment of the random variable that arises when fixing $\mathbf{x}, \mathbf{y}$, drawing f_i at random and considering $\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})$. For cubic regularization methods, a natural analogue is the *third* moment of the Hessian difference.

In the following, we will show that one can indeed weaken the assumption of the SVRC algorithm from Zhou et al. (2019) to Assumption 4.8.

Assumption 4.8. *We say a function $F = \sum_{i=1}^n f_i$ with $f_i : \mathbb{R}^d \to \mathbb{R}$ respects the third-moment smoothness assumption with constant L_2 if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$*

$$\left(\mathbb{E}_i \|\nabla^2 f_i(\mathbf{x}) - \nabla^2 f_i(\mathbf{y})\|^3\right)^{\frac{1}{3}} \le L_2 \|\mathbf{x} - \mathbf{y}\|.$$

The expected value is taken w.r.t. a uniform distribution on $[n]$. We continue to assume F satisfies Assumption 2.1 ii).

Note that this assumption is weaker than the usual second-order smoothness, but it is stronger than a second moment assumption, due to $\mathbb{E}[|X|^s]^{1/s} \le \mathbb{E}[|X|^t]^{1/t}$ for $s < t$. Furthermore, by Jensen's inequality, F has Lipschitz continuous Hessian, which is one reason why the assumption turns out to be useful. The second one is that error terms for cubic regularization are third powers, so this assumption provides a more natural fit than, say, a mean-squared Lipschitz assumption on the Hessian.

With some changes to the convergence analysis, the guarantees of SVRC (actually even to second-order stationarity) can essentially be retained.

Theorem 4.9. *SVRC under Assumption 4.8 needs $\tilde{\mathcal{O}}\left(n + \Delta\sqrt{L_2} n^{4/5}\varepsilon^{-3/2}\right)$ oracle queries to find a point $\mathbf{x}_{\text{out}}$ such that $\mathbb{E}\|\nabla F(\mathbf{x}_{\text{out}})\| \le \varepsilon$.*

What is now left to do is to provide a tighter lower bound. Indeed, the following holds:

Theorem 4.10. *For any randomized algorithm A satisfying Assumption 2.3, Δ, L_2, ε, and $n \le \frac{c\Delta^{12/7} L_2^{6/7}}{\varepsilon^{18/7}}$ there exists a dimension $d \le \tilde{\mathcal{O}}(n^2 \Delta L_2 \varepsilon^{-3})$ and a function $F = \frac{1}{n}\sum_{i=1}^n f_i$ that satisfies Assumption 4.8 such that*

$$T_\varepsilon(\mathsf{A}, F) \ge \Omega\left(L_2^{1/2} \Delta n^{5/12} \varepsilon^{-\frac{3}{2}}\right),$$

where the constants hidden by Ω do not depend on ε or n. c is also a universal constant.

Note the $n^{1/6}$ difference when compared to Theorem 4.7. So – to conclude – under Assumption 4.8 and $p = 2$, one can find an ε-approximate local minimum in $\tilde{\mathcal{O}}(n^{4/5}\varepsilon^{-3/2})$ oracle queries while the lower bound lies at $\Omega(n^{5/12}\varepsilon^{-3/2})$. While the gap remains at $\Omega(n^{23/60})$, this is a notable improvement over the results for Assumption 2.1, which means that the third-moment smoothness assumption gets us closer to understanding the fundamental limits for higher-order variance-reduced methods.

5 DISCUSSION

In this work, we analyzed the oracle complexity of higher-order smooth non-convex finite-sum optimization. We showed that speedup (e.g. through variance reduction) requires randomization. In the randomized case, our bounds indicate that variance reduction is harder for stochastic estimators of a higher-order derivative tensor than for the gradient. We introduced a new, weaker notion of higher-order smoothness that may prove to be a viable choice for the further study of the complexity of non-convex finite-sum optimization and be of independent interest.

To the best of our knowledge, we are the first to show a non-trivial dependency on n in a non-convex finite-sum randomized setting with individually smooth functions. Still, in this regime our bounds are not tight. This gap also appears in a variety of other non-convex finite-sum lower bound constructions (Zhou and Gu, 2019; Han et al., 2021). Interestingly, the gap is not present when working with convex functions (Woodworth and Srebro, 2016). We conjecture that this difference is due to the definition of success typically used in both scenarios and the fact that all constructions we know of (including ours) force some form of orthogonality between derivatives of *different* functions (up to shared terms stemming from shared regularizers). In the convex case, success is to provably bound suboptimality by ε. If one constructs a problem instance $F(\mathbf{x}) = \frac{1}{n}\sum_i \tilde{f}_i(\mathbf{x})$ that satisfies $\langle \nabla \tilde{f}_i(\mathbf{x}), \nabla \tilde{f}_j(\mathbf{x}) \rangle = 0$ for $i \neq j$ and is able to guarantee that a constant fraction $0 < c \le 1$ of the $\tilde{f}_i$s is larger than a constant c', the lower bound

$c \cdot c'$ on the value of F follows. When bounding the gradient norm – which is the quantity of interest in the non-convex case – this approach is less fruitful, as $\|\nabla \tilde{f}_i(\mathbf{x})\| > c'$ for a c-fraction of indices only implies $\|\nabla F(\mathbf{x})\| > \frac{c \times c'}{\sqrt{n}}$, i.e. we lose a factor of $\sqrt{n}$. This may suggest that getting sharper lower bound is indeed difficult, and that progress could be made on the algorithmic side. Whether this is possible hinges on the question of whether individual smoothness of functions is stronger – for algorithmic purposes – than its moment-based counterpart.

Finally, we mention that our deterministic construction does not suffer from this lost factor. Unfortunately, the adversarial nature of the construction renders it unusable for the randomized setting. Nonetheless, this feature could make it of theoretical interest for future work.

Acknowledgements

We thank the anonymous reviewers for their helpful and constructive feedback. Additionally, we are grateful to Parnian Kassraie for valuable comments on the organisation of this manuscript.

Bibliography

Allen-Zhu, Z. and Hazan, E. (2016). Variance reduction for faster non-convex optimization. In *International conference on machine learning*, pages 699–707.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Sekhari, A., and Sridharan, K. (2020). Second-order information in non-convex stochastic optimization: Power and limitations. In Abernethy, J. and Agarwal, S., editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 242–299. PMLR.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. (2019). Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*.

Ball, K. (1997). An elementary introduction to modern convex geometry. In *Flavors of geometry*, pages 1–58.

Birgin, E., Gardenghi, J., Martínez, J. M., Santos, S., and Toint, P. (2017). Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Mathematical Programming*, 163:359–368.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. (2019a). Lower bounds for finding stationary points i. *Mathematical Programming*, pages 1–50.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. (2019b). Lower bounds for finding stationary points ii: first-order methods. *Mathematical Programming*, pages 1–41.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. (2018). Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 689–699.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Han, Y., Xie, G., and Zhang, Z. (2021). Lower complexity bounds of finite-sum optimization problems: The results and construction. *ArXiv*, abs/2103.08280.

Hannah, R., Liu, Y., O'Connor, D., and Yin, W. (2018). Breaking the span assumption yields fast finite-sum minimization. In *NeurIPS*.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Lei, L., Ju, C., Chen, J., and Jordan, M. I. (2017). Non-convex finite-sum optimization via scsg methods. In *Advances in Neural Information Processing Systems*, pages 2348–2358.

Lucchi, A. and Kohler, J. (2019). A stochastic tensor method for non-convex optimization. *arXiv preprint arXiv:1911.10367*.

Murty, K. G. and Kabadi, S. N. (1987). Some np-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 39(2):117–129.

Nesterov, Y. (2004). Introductory lectures on convex optimization - a basic course. In *Applied Optimization*.

Nesterov, Y. and Polyak, B. T. (2006). Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205.

Reddi, S. J., Hefny, A., Sra, S., Poczos, B., and Smola, A. (2016). Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323.

Shen, Z., Zhou, P., Fang, C., and Ribeiro, A. (2019). A stochastic trust region method for non-convex minimization. *arXiv preprint arXiv:1903.01540*.

Wang, Z., Zhou, Y., Liang, Y., and Lan, G. (2019). Sample complexity of stochastic variance-reduced cubic regularization for nonconvex optimization. In *AISTATS*.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *ICML*.

Woodworth, B. E., Bullins, B., Shamir, O., and Srebro, N. (2021). The min-max complexity of distributed stochastic convex optimization with intermittent communication. In *COLT*.

Woodworth, B. E. and Srebro, N. (2016). Tight complexity bounds for optimizing composite objectives. In *Advances in neural information processing systems*, pages 3639–3647.

Yu, C. and Yao, W. (2017). Robust linear regression: A review and comparison. *Communications in Statistics - Simulation and Computation*, 46:6261 – 6282.

Zhang, S., Yang, J., Guzm'an, C., Kiyavash, N., and He, N. (2021). The complexity of nonconvex-strongly-concave minimax optimization. *ArXiv*, abs/2103.15888.

Zhou, D. and Gu, Q. (2019). Lower bounds for smooth nonconvex finite-sum optimization. In *ICML*.

Zhou, D. and Gu, Q. (2020). Stochastic recursive variance-reduced cubic regularization methods. In *International Conference on Artificial Intelligence and Statistics*, pages 3980–3990.

Zhou, D., Xu, P., and Gu, Q. (2018). Sample efficient stochastic variance-reduced cubic regularization method. *ArXiv*, abs/1811.11989.

Zhou, D., Xu, P., and Gu, Q. (2019). Stochastic variance-reduced cubic regularization methods. *Journal of Machine Learning Research*, 20(134):1–47.

Supplementary Material: On the Oracle Complexity of Higher-Order Smooth Non-Convex Finite-Sum Optimization

The appendix is structured in 4 parts. Appendix A provides all omitted proofs for Section 3, while Appendix B provides the same for Section 4, up to the end of Section 4.1. In Appendix C we give the proofs for Theorems 4.9 and 4.10. Finally Appendix D contains the proof of a simple observation that is needed for all constructions.

A LOWER BOUNDS FOR DETERMINISTIC ALGORITHMS

We first prove Lemmas 3.3 and 3.4 in Appendices A.1 and A.2 respectively, and then prove the main Theorem 3.5 in Appendix A.3.

A.1 Proof of Lemma 3.3

Proof [of Lemma 3.3]**.** We will omit the scaling parameters as they do not influence the proof in any way and define for $k \in [K]$ the shorthand $y_k = y_k(\mathbf{x}) = \langle \mathbf{v}_k, \mathbf{x} \rangle$. We will construct the oracle such that during round $r \in [2 : K+1]$, its responses are based on the function:

$$\begin{aligned} f_i^r(\mathbf{x}) = &-\delta_{i,1}\Psi(1)\Phi(y_1) \\ &+ \sum_{k=2}^{r-1} \delta_{i,k}\left[\Psi(-y_{k-1})\Phi(-y_k) - \Psi(y_{k-1})\Phi(y_k)\right]. \end{aligned}$$

We will show that $\mathbf{V}$ can be chosen such that these responses are consistent with Definition 3.2. By consistence, we mean equality of the function values and derivatives at the queried indices and points.

By construction, the answers for round r, only depend on $\mathbf{v}_k$ and $\delta_{i,k}$ for $k < r$. This allows us to determine $\delta_{i,r}$ and $\mathbf{v}_r$ at the *end* of round r. Specifically, we will choose $\mathbf{v}_r$ such that $\langle \mathbf{v}_r, \mathbf{x}^{(t)} \rangle = 0$ for all iterates occurring before the end of round r (i.e. all queries made so far). Further, $\mathbf{v}_r$ needs to be orthogonal to $\mathbf{v}_k$ for all $k < r$. These orthogonality constraints imply a requirement on the dimension of the domain of F. This dimension d must therefore be linear in the sum of K and of the final lower bound, to ensure orthogonality to both iterates and between the columns of $\mathbf{V}$ is possible. As mentioned above, we will also choose $\delta_{i,r} = 1$ iff function i was not queried during round r.

We must now prove that for all $q \geq 0$ and iterates t queried during round r, we have $\nabla^q f_{i^t}^r(\mathbf{x}^{(t)}) = \nabla^q f_{i^t}(\mathbf{x}^{(t)})$, guaranteeing that our oracle is aligned with the function from Definition 3.2. For simplicity, we define $\mathbf{x} = \mathbf{x}^{(t)}$ and $i = i^t$. Then, we can write $f_i(\mathbf{x})$ as

$$f_i^r(\mathbf{x}) + \delta_{i,r}[\Psi(-y_{r-1})\Phi(-y_r) - \Psi(y_{r-1})\Phi(y_r)] + g_i^r(\mathbf{x})$$

for $g_i^r(\mathbf{x}) = f_i(\mathbf{x}) - f_i^r(\mathbf{x})$. Since function i was queried during round r, we have $\delta_{i,r} = 0$, and so $f_i(\mathbf{x}) = f_i^r(\mathbf{x}) + g_i^r(\mathbf{x})$. Hence, it suffices that $\nabla^q g_i^r(\mathbf{x}) = \mathbf{0} \in \mathbb{R}^{\otimes^q d}$. Indeed, $\Psi(z) = 0$ for all $|z| \leq 1/2$. By our choice of $\mathbf{V}$, we have $\langle \mathbf{v}_k, \mathbf{x} \rangle = 0$ for all $k \geq r$. Since all terms in g_i^r have a multiplicative factor $\Psi(\pm\langle \mathbf{v}_{k-1}, \mathbf{x} \rangle)$ for some $k \geq r+1$, the function g_i^r is indeed constant 0 inside a neighbourhood of $\mathbf{x}$, and so all its derivative tensors are $\mathbf{0}$ at $\mathbf{x}$. □

A.2 Proof of Lemma 3.4

Before showing Lemma 3.4, we should stress that a key property of the function $\bar{f}_K = f_{K,\mathbf{1}}$ is that as long as the last coordinate in its input is zero, the gradient of the function will be lower bounded by a constant.

Lemma A.1 [Lemma 2 in Carmon et al. (2019a)]**.** *Let $\mathbf{x} \in \mathbb{R}^K$ with $|x_k| < 1$ for some $k \in [K]$. Then, there exists $l \leq k$ with $|x_l| < 1$ and*

$$\left| \frac{\partial \bar{f}_K}{\partial x_l}(\mathbf{x}) \right| > 1.$$

Proof [of Lemma 3.4]**.** By Lemma 3.3, we have $(\mathbf{V}^T \mathbf{x}^{(t)})_{K+1} = 0$ for all iterates up until the end of round $K+1$ and will therefore be able to apply Lemma A.1. We use $\tilde{\nabla}$ to denote the gradient with respect to $\mathbf{V}^T\mathbf{x}/\sigma$ and write

$$\lambda \sigma^{p+1} \tilde{\nabla} \left[\frac{1}{n} \sum_{i=1}^{n} f_{K+1,\boldsymbol{\delta}_i}(\mathbf{V}^T\mathbf{x}/\sigma) \right] = \frac{1}{n} \left\lceil \frac{n}{2} \right\rceil \lambda \sigma^{p+1} \tilde{\nabla} \left[\bar{f}_{K+1}(\mathbf{V}^T\mathbf{x}/\sigma) \right].$$

Using the chain rule, we see that

$$\nabla F(\mathbf{x}) = \frac{1}{n} \left\lceil \frac{n}{2} \right\rceil \lambda \sigma^{p} \mathbf{V} \tilde{\nabla} \left[\bar{f}_{K+1}(\mathbf{V}^T\mathbf{x}/\sigma) \right],$$

and thus by Lemma A.1 and by the fact that $\mathbf{V}^T\mathbf{V} = \mathbf{I}_{K+1}$

$$\|\nabla F(\mathbf{x})\| \geq \frac{\lambda \sigma^p}{4}.$$

□

A.3 Proof of Theorem 3.5

Along with Lemma 3.4, we need the following result that will allow us to ensure F satisfies Assumption 2.1.

Lemma A.2. *For all K and $\boldsymbol{\delta} \in \{0,1\}^K$, the function $f_{K,\boldsymbol{\delta}}$ from Definition 3.1 satisfies*

i) The initial sub-optimality can be bounded by $f_{K,\boldsymbol{\delta}}(\mathbf{0}) - \inf_{\mathbf{x}\in\mathbb{R}^K} f_{K,\boldsymbol{\delta}}(\mathbf{x}) \leq 12K$.

ii) The function is p-th order ℓ_p-smooth with $\ell_p \leq \exp(\frac{5}{2} p \log p + cp)$ for some numerical constant $c < \infty$.

To show Lemma A.2, we need the following technical result, which is a subset of Lemma 1 in Carmon et al. (2019a).

Lemma A.3. *For the functions from Definition 3.1 we have*

i) Both Ψ and Φ are infinitely differentiable, and for all $q \in \mathbb{N}$ we have

$$\sup_x |\Psi^{(q)}(x)| \leq \exp\left(\frac{5q}{2} \log(4q) \right) \quad \text{and} \quad \sup_x |\Phi^{(q)}(x)| \leq \exp\left(\frac{3q}{2} \log \frac{3q}{2} \right).$$

ii) The functions and derivatives Ψ, Ψ', Φ, Φ' are non-negative and bounded, with

$$0 \leq \Psi < e, \quad 0 \leq \Psi' \leq \sqrt{54/e}, \quad 0 < \Phi < \sqrt{2\pi e} \quad \text{and} \quad 0 < \Phi' \leq \sqrt{e}.$$

Now we present our proof, closely following Carmon et al. (2019a), Appendix B.2. We account for the indicators δ_k used in our construction, validating that they do not affect the aforementioned properties.

Proof [of Lemma A.2]**.** Fix $K \in \mathbb{N}, \boldsymbol{\delta} \in \{0,1\}^K$. We first bound the suboptimality gap. We have $f_{K,\boldsymbol{\delta}}(\mathbf{0}) \leq 0$ because $-\delta_1 \Psi(1)\Phi(0) \leq 0$ by Lemma A.3 ii). By the same arguments, for any $\mathbf{x}$, we have

$$\begin{aligned} f_{K,\boldsymbol{\delta}}(\mathbf{x}) &= -\delta_1 \Psi(1)\Phi(x_1) + \sum_{k=2}^{K} \delta_k [\Psi(-x_{k-1})\Phi(-x_k) - \Psi(x_{k-1})\Phi(x_k)] \\ &\geq -\delta_1 \Psi(1)\Phi(x_1) - \sum_{k=2}^{K} \delta_k [\Psi(x_{k-1})\Phi(x_k)] \geq -\delta_1 (e \cdot \sqrt{2\pi e}) - \sum_{k=2}^{K} \delta_k [e\sqrt{2\pi e}] > -K \cdot e \cdot \sqrt{2\pi e} \geq -12K. \end{aligned}$$

Thus, we get our bound on suboptimality.
For the second part, let $\mathbf{x} \in \mathbb{R}^K$. For a unit vector $\mathbf{v} \in \mathbb{R}^K$ we define the directional projection $h_{\mathbf{x},\mathbf{v}}(\theta) = f_{K,\boldsymbol{\delta}}(\mathbf{x} + \theta \mathbf{v})$. It suffices to show that $|h^{(p+1)}_{\mathbf{x},\mathbf{v}}(0)| \leq \ell_p$ for any $\mathbf{x}, \mathbf{v}$, because the directional projection is infinitely differentiable, by Lemma A.3. Fix $\mathbf{x}, \mathbf{v} \in \mathbb{R}^K$. We can write

$$h^{(p+1)}_{\mathbf{x},\mathbf{v}}(0) = \sum_{j_1,\ldots,j_{p+1}=1}^{K} \partial_{j_1} \cdots \partial_{j_{p+1}} f_{K,\boldsymbol{\delta}}(\mathbf{x}) v_{j_1} \cdots v_{j_{p+1}}.$$

All multiplicative terms in $f_{K,\boldsymbol{\delta}}$ have zero derivatives unless all derivatives are w.r.t. adjacent indices. Defining for convenience $v_0 = v_{K+1} = 0$ we can express the above as

$$h^{(p+1)}_{\mathbf{x},\mathbf{v}}(0) = \sum_{\boldsymbol{\gamma} \in \{0,1\}^p \cup \{0,-1\}^p} \sum_{j=1}^{K} \partial_{j+\gamma_1} \cdots \partial_{j+\gamma_p} \partial_j f_{K,\boldsymbol{\delta}}(\mathbf{x}) v_{j+\gamma_1} \cdots v_{j+\gamma_p} v_j.$$

We can bound

$$\max_{j \in [K]} \max_{\gamma \in \{0,1\}^p \cup \{0,-1\}^p} |\partial_{j+\gamma_1} \cdots \partial_{j+\gamma_p} \partial_j f_{K,\boldsymbol{\delta}}(\mathbf{x})| \leq \max_{k \in [0:k+1]} \left\{ 4 \sup_{y \in \mathbb{R}} \left| \Psi^{(k)}(y) \right| \sup_{y' \in \mathbb{R}} \left| \Phi^{(p+1-k)}(y') \right| \right\}.$$

Here, we have used that δ can only (potentially) suppress terms and that there are only 4 terms which may involve partial derivatives with respect to either x_j and x_{j+1} or x_j and x_{j-1}. Note that if $\boldsymbol{\gamma} \neq \mathbf{0}$, there are only 2 such terms.
With Lemma A.3, the above can be further bounded by

$$4\sqrt{2\pi e} \cdot e^{2.5(p+1)\log(4(p+1))} \leq e^{2.5p + \log p + 4p + 10}.$$

We define $\ell_p = 2^{p+1} e^{2.5p + \log p + 4p + 10} \leq e^{2.5p + \log p + 5p + 11}$. Finally, we can bound the quantity of interest

$$|h^{(p+1)}_{\mathbf{x},\mathbf{v}}(0)| \leq \sum_{\boldsymbol{\gamma} \in \{0,1\}^p \cup \{0,-1\}^p} 2^{-(p+1)} \ell_p \left| \sum_{j=1}^{K} v_{j+\gamma_1} \cdots v_{j+\gamma_p} v_j \right| \leq \ell_p,$$

because $\left| \sum_{j=1}^{K} v_{j+\gamma_1} \cdots v_{j+\gamma_p} v_j \right| \leq 1$, which follows from $\mathbf{v}$ being a unit vector (see Carmon et al. (2019a), B.2). This concludes the proof. □
Now we are ready to prove the main Theorem of this section.

Proof [of Theorem 3.5]**.** At this point, we are ready to proceed with our argument. Recall Definition 3.2 of the hard instance $F = \frac{1}{n} \sum f_i$ with $\mathbf{V} \in \mathsf{Ortho}(d, K+1)$:

$$f_i(\mathbf{x}) = \lambda \sigma^{p+1} f_{K+1,\boldsymbol{\delta}_i}(\mathbf{V}^T \mathbf{x} / \sigma).$$

We will guarantee smoothness through λ, bound the gradient norm from below through σ and finally control the distance to optimality with K. By Lemma D.1 and Lemma A.2ii), we can write for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$\|\nabla^p f_i(\mathbf{x}) - \nabla^p f_i(\mathbf{y})\| \leq \lambda \ell_p \|\mathbf{V}^T(\mathbf{x}-\mathbf{y})\| \leq \lambda \ell_p \|\mathbf{x}-\mathbf{y}\|,$$

where the second inequality follows because $d \geq K+1$ and because we can complete $\mathbf{V}$ to be a square orthogonal matrix. We see that the choice $\lambda = L_p/\ell_p$ guarantees pth-order smoothness with constant L_p.

Next, we will turn to bounding the gradient from below. By Lemma 3.4, we can lower bound $\|\nabla F(\mathbf{x}^{(t)})\| > \frac{\lambda \sigma^p}{4}$ for all iterates up to the end of round $K+1$. We desire a lower bound for ε-stationarity, so we will choose $\sigma = \left(\frac{4\varepsilon \ell_p}{L_p}\right)^{\frac{1}{p}}$.

As a last step, we will choose K such that the initial gap on suboptimality is bounded by Δ. For that, we use Lemma A.2i). We want

$$\begin{aligned} F(0) - \inf_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) &\leq \frac{1}{n}\sum_{i=1}^n \left[f_i(\mathbf{0}) - \inf_{\mathbf{x}\in\mathbb{R}^d} f_i(\mathbf{x}) \right] \\ &\leq \frac{\lambda \sigma^{p+1}}{n} \sum_{i=1}^n \left[f_{K+1,\boldsymbol{\delta}_i}(\mathbf{0}) - \inf_{\mathbf{y}\in\mathbb{R}^{K+1}} f_{K+1,\boldsymbol{\delta}_i}(\mathbf{y}) \right] \\ &\leq 12\lambda\sigma^{p+1}(K+1) \leq \Delta. \end{aligned}$$

As a larger value of K yields a better bound, we can choose

$$K+1 = \left\lfloor \frac{\Delta}{192}\left(\frac{L_p}{\ell_p}\right)^{\frac{1}{p}} \frac{1}{\varepsilon^{\frac{p+1}{p}}} \right\rfloor \leq \frac{\Delta}{12}\frac{\ell_p}{L_p}\left(\frac{L_p}{4\varepsilon\ell_p}\right)^{\frac{p+1}{p}}.$$

Because K is the number of rounds and each round consists of $\Omega(n)$ queries, this yields a $\Omega\left(\left(\frac{L_p}{\ell_p}\right)^{1/p} \frac{\Delta n}{\varepsilon^{(p+1)/p}}\right)$ lower bound, as desired. As explained in the proof of Lemma 3.3, the dimension d must merely be larger than the sum of the lower bound and the number of rounds, i.e. linear in the lower bound. This completes the proof. □

B LOWER BOUNDS FOR RANDOMIZED ALGORITHMS

We first prove Lemmas 4.4, 4.5 and 4.6 in Appendices B.1, B.2 and B.3 respectively, and then prove the main Theorem 4.7 in Appendix B.4.

B.1 Proof of Lemma 4.4

Proof [of Lemma 4.4]**.** We have $\nabla^p G(\mathbf{x}) = \frac{a}{b^p}\nabla^p F(\mathbf{x}/b)$. We have to exhibit an algorithm B such that $\mathsf{B}[F]$ follows the same distribution as $\{[i^t, \mathbf{x}^{(t)}/b]\}_{t\in\mathbb{N}}$.

Let $\{A^{(t)}\}_{t\in\mathbb{N}}$ be the sequence of mappings that produce the iterates of A. With some mild abuse of notation we may write [4]

$$\begin{aligned} \mathsf{A}_\xi[G]^{(t)} &= A^{(t)}\Big\{ \quad \xi, i^{0:t-1}, \mathbf{x}^{(0:t-1)}, \\ &\qquad\qquad \nabla^{(0:q)} g_{i^0}(\mathbf{x}^{(0)}), \dots, \nabla^{(0:q)} g_{i^{t-1}}\left(\mathbf{x}^{(t-1)}\right) \quad \Big\} \\ &= A^{(t)}\Big\{ \quad \xi, i^{0:t-1}, \mathbf{x}^{(0:t-1)}, g_{i^{0:t-1}}\left(\mathbf{x}^{(0:t-1)}\right), \\ &\qquad\qquad \nabla g_{i^{0:t-1}}\left(\mathbf{x}^{(0:t-1)}\right), \dots, \nabla^q g_{i^{0:t-1}}\left(\mathbf{x}^{(0:t-1)}\right) \quad \Big\}. \end{aligned}$$

B shall choose i^0 exactly like A does. We define the sequence of mappings $\{B^{(t)}\}_{t\in\mathbb{N}}$ underlying B on arbitrary input $H = \frac{1}{n}\sum_{i=1}^n h_i$ as

$$\begin{aligned} &\mathsf{B}_\xi[H]^{(t)} \\ &= B^{(t)}\Big\{ \quad \xi, i^{0:t-1}, \mathbf{y}^{(0:t-1)}, h_{i^{0:t-1}}\left(\mathbf{y}^{(0:t-1)}\right), \\ &\qquad\qquad \nabla h_{i^{0:t-1}}\left(\mathbf{y}^{(0:t-1)}\right), \dots, \nabla^q h_{i^{0:t-1}}\left(\mathbf{y}^{(0:t-1)}\right) \quad \Big\} \\ &= \frac{1}{b}A^{(t)}\Big\{ \quad \xi, i^{0:t-1}, b\cdot\mathbf{y}^{(0:t-1)}, a\cdot h_{i^{0:t-1}}\left(\mathbf{y}^{(0:t-1)}\right), \\ &\qquad\qquad \frac{a}{b}\nabla h_{i^{0:t-1}}\left(\mathbf{y}^{(0:t-1)}\right), \dots, \frac{a}{b^q}\nabla^q h_{i^{0:t-1}}\left(\mathbf{y}^{(0:t-1)}\right) \quad \Big\}, \end{aligned}$$

where we apply the outer division only on the iterates and not the indices. We can check by induction that for a fixed random seed ξ, $\mathsf{B}_\xi[F]^{(t)} = \frac{\mathsf{A}_\xi[G]^{(t)}}{b}$ for all $t \in \mathbb{N}$: The base case is clear as i^0 does not depend on any oracle queries and $\mathbf{x}^{(0)} = \mathbf{0}$ is deterministic. Now assume that the equality holds for all $t' < t$. Then

$$\begin{aligned} &\mathsf{B}_\xi[F]^{(t)} \\ &\overset{\text{I.H.}}{=} B^{(t)}\Big\{ \quad \xi, i^{0:t-1}, \frac{\mathbf{x}^{(0:t-1)}}{b}, f_{i^{0:t-1}}\left(\frac{\mathbf{x}^{(0:t-1)}}{b}\right), \\ &\qquad\qquad \nabla f_{i^{0:t-1}}\left(\frac{\mathbf{x}^{(0:t-1)}}{b}\right), \dots, \nabla^q f_{i^{0:t-1}}\left(\frac{\mathbf{x}^{(0:t-1)}}{b}\right) \quad \Big\} \\ &= \frac{1}{b}A^{(t)}\Big\{ \quad \xi, i^{0:t-1}, b\cdot\frac{\mathbf{x}^{(0:t-1)}}{b}, a\cdot f_{i^{0:t-1}}\left(\frac{\mathbf{x}^{(0:t-1)}}{b}\right), \\ &\qquad\qquad \frac{a}{b}\nabla f_{i^{0:t-1}}\left(\frac{\mathbf{x}^{(0:t-1)}}{b}\right), \dots, \frac{a}{b^q}\nabla^q f_{i^{0:t-1}}\left(\frac{\mathbf{x}^{(0:t-1)}}{b}\right) \quad \Big\} \\ &= \frac{1}{b}A^{(t)}\Big\{ \quad \xi, i^{0:t-1}, \mathbf{x}^{(0:t-1)}, g_{i^{0:t-1}}\left(\mathbf{x}^{(0:t-1)}\right), \\ &\qquad\qquad \nabla g_{i^{0:t-1}}\left(\mathbf{x}^{(0:t-1)}\right), \dots, \nabla^q g_{i^{0:t-1}}\left(\mathbf{x}^{(0:t-1)}\right) \quad \Big\} \\ &= \frac{\mathsf{A}_\xi[G]^{(t)}}{b}. \end{aligned}$$

Therefore $\mathsf{B}[F]$ follows the same distribution as $\{[i_t, \mathbf{x}^{(t)}/b]\}_{t\in\mathbb{N}}$ and so the sequence is informed by F, as desired. □

[4] We use $\nabla^k g_{i^{0:t-1}}(\mathbf{x}^{(0:t-1)})$ to denote the sequence of all queried kth-order derivatives to produce iterate t.

B.2 Proof of Lemma 4.5

The proof of Lemma 4.5 follows that of Lemma 12 in Fang et al. (2018) [5] and is similar to Lemma 4 in Carmon et al. (2019a). The reader accustomed to lower bounds for convex optimization will be familiar with the ideas involved (e.g. Lemma 6 and 7 in Woodworth and Srebro (2016)).

Proof [of Lemma 4.5]**.** First, we define quantities that we will use throughout the proof.

Define $\mathbf{y}_i^{(t)} = \rho(\mathbf{C}_i^T\mathbf{x}^{(t)}) = \frac{\mathbf{C}_i^T\mathbf{x}}{\sqrt{1+\|\mathbf{C}_i^T\mathbf{x}\|^2/R^2}}$. Then $\mathbf{y}_i^{(t)} \in \mathbb{R}^{d/n}$ satisfies $\|\mathbf{y}_i^{(t)}\| \leq R$. Let $\mathcal{V}_i^{(t)}$ be the set of previous transformed iterates at index i along with the discovered columns of $\mathbf{B}$ of after iteration t:

$$\mathcal{V}_i^{(t)} = \{\mathbf{y}_i^{(0)}, \dots, \mathbf{y}_i^{(t)}\} \cup \bigcup_{j=1}^{n} \{\mathbf{b}_{j,1}, \dots, \mathbf{b}_{j,\min(K,I_j^t)}\}.$$

Let $\mathcal{U}_i^{(t)}$ be defined as in the premise of Lemma 4.5 and denote by $\tilde{\mathcal{U}}_i^{(t)}$ its "complement" (all other columns):

$$\tilde{\mathcal{U}}_i^{(t)} = \left\{\mathbf{b}_{i,1}, \dots, \mathbf{b}_{i,\min(K,I_i^{t-1})}\right\}.$$

Define $\mathcal{U}^{(t)} = \bigcup_{i=1}^n \mathcal{U}_i^{(t)}$ and $\tilde{\mathcal{U}}^{(t)} = \bigcup_{i=1}^n \tilde{\mathcal{U}}_i^{(t)}$ and let $\mathbf{P}_i^{(t)}$ denote the orthogonal projection onto the span of $\mathcal{V}_i^{(t)}$. Let $\mathbf{P}_i^{(t)\perp} = \mathbf{I} - \mathbf{P}_i^{(t)}$ be its orthogonal complement. Both of these are mappings from $\mathbb{R}^{d/n} \to \mathbb{R}^{d/n}$.

Recall that our ultimate goal is to show that $\{[i^t, \mathbf{x}^{(t)}]\}_{t\in\mathbb{N}}$ being informed by F^* implies that with probability $1-\delta$, for all $t \in \{0, \dots, T\}$, all $i \in [n]$ and all corresponding $\mathbf{u} \in \mathcal{U}_i^{(t)}$ the inequality

$$|\langle \mathbf{u}, \mathbf{y}_i^{(t)}\rangle| < \frac{1}{2} \tag{2}$$

holds. The case $t = 0$ is obviously true, so from now on we focus on showing (2) for $t \geq 1$. We will first define an auxiliary event, show that it implies our result and then bound its probability. For any $t \in [T]$ define the event

$$G^t = \bigcup_{i\in[n]} \bigcup_{\mathbf{u}\in\mathcal{U}^{(t)}} \left\{ \left|\langle \mathbf{u}, \mathbf{P}_i^{(t-1)\perp}\mathbf{y}_i^{(t)}\rangle\right| < a\|\mathbf{P}_i^{(t-1)\perp}\mathbf{y}_i^{(t)}\| \right\},$$

where $a = \min\left(\frac{1}{3(T+1)}, \frac{1}{2(1+\sqrt{3T})R}\right)$. Note that the union is over $\mathcal{U}^{(t)}$ and not $\mathcal{U}_i^{(t)}$. Let $G^{\leq t} = \cap_{j=1}^t G^j$. We first show that $G^{\leq T}$ implies (2).

Assume $\mathcal{U}_i^{(t)} \neq \emptyset$, otherwise (2) holds trivially. For any $i \in [n]$, $t \in [T]$ and $\mathbf{u} \in \mathcal{U}_i^{(t)}$ we have

$$\begin{aligned}
|\langle \mathbf{u}, \mathbf{y}_i^{(t)}\rangle| &\leq \left|\langle \mathbf{u}, \mathbf{P}_i^{(t-1)}\mathbf{y}_i^{(t)} + \mathbf{P}_i^{(t-1)\perp}\mathbf{y}_i^{(t)}\rangle\right| \\
&< \left|\langle \mathbf{u}, \mathbf{P}_i^{(t-1)}\mathbf{y}_i^{(t)}\rangle\right| + a\|\mathbf{P}_i^{(t-1)\perp}\mathbf{y}_i^{(t)}\| \\
&\leq R\|\mathbf{P}_i^{(t-1)}\mathbf{u}\| + aR.
\end{aligned}$$

In the second step we used $G^{\leq T}$ and in the third step we used Cauchy-Schwarz and the fact that $\mathbf{P}_i^{(t-1)}$ and $\mathbf{P}_i^{(t-1)\perp}$ are orthogonal projectors and therefore self-adjoint. If we manage to show $\|\mathbf{P}_i^{(t-1)}\mathbf{u}\| \leq \sqrt{3T}a =: b$ we are done, because the choice of a then implies that $aR + R\|\mathbf{P}_i^{(t-1)}\mathbf{u}\| \leq \frac{1}{2}$.

We will show this by induction over $t \in [T]$: Consider $t = 1$ and let $i \in [n]$ be arbitrary. We have $\mathcal{V}_i^{(t-1)} = \mathcal{V}_i^{(0)} = \{\mathbf{y}_i^{(0)}, \mathbf{b}_{i^0,1}\} = \{\mathbf{0}, \mathbf{b}_{i^0,1}\}$. Because $\mathbf{u}$ can be any column of $\mathbf{B}$ except $\mathbf{b}_{i^0,1}$ we have

[5] The main difference is that we make clear (thanks to the formalism of a robust zero-chain), that it does not matter how many derivatives the algorithm has access to, hence the identical statement.

$\mathbf{P}_i^{(t-1)}\mathbf{u} = 0$. For the induction step, another way to write the vectors in $\mathcal{V}_i^{(t-1)}$ is in the order they are discovered. That is, add to the set each iterate at i and an additional column of $\mathbf{B}_{i^j}$ for the queried index i^j at iteration j. We get the sequence

$$\mathbf{y}_i^{(0)},\ \mathbf{b}_{i^0,\min(I^0_{i^0},K)},\ \mathbf{y}_i^{(1)},\ \mathbf{b}_{i^1,\min(I^1_{i^1},K)},\ \ldots,\ \mathbf{y}_i^{(t-1)},\ \mathbf{b}_{i^{t-1},\,\min(I^{t-1}_{i^{t-1}},K)}.$$

We will now apply the Gram-Schmidt procedure on these vectors. Remember that for a sequence of vectors $\mathbf{v}_i$ the Gram-Schmidt procedure (without normalization) constructs vectors

$$\begin{aligned}
\mathbf{u}_1 &= \mathbf{v}_1 \\
\mathbf{u}_2 &= \mathbf{v}_2 - \operatorname{proj}_{\mathbf{u}_1}(\mathbf{v}_2) \\
&\ \vdots \\
\mathbf{u}_k &= \mathbf{v}_k - \operatorname{proj}_{\mathbf{u}_1,\ldots,\mathbf{u}_{k-1}}(\mathbf{v}_k),
\end{aligned}$$

where proj_S shall denote the projection on a set of vectors S. Applying this scheme to our sequence above, we get vectors [6]

$$\left\{\mathbf{y}_i^{(z)} - \mathbf{P}_i^{(z-1)}\mathbf{y}_i^{(z)}\right\}_{z=0}^{t-1} = \left\{\mathbf{P}_i^{(z-1)\perp}\mathbf{y}_i^{(z)}\right\}_{z=0}^{t-1}$$

and

$$\begin{aligned}
&\left\{\mathbf{b}_{i^z,\min(I^z_{i^z},K)} - \mathbf{P}_i^{(z-1)}\mathbf{b}_{i^z,\min(I^z_{i^z},K)} - \operatorname{proj}_{\mathbf{P}_i^{(z-1)\perp}\mathbf{y}_i^{(z)}}\mathbf{b}_{i^z,\min(I^z_{i^z},K)}\right\}_{z=0}^{t-1} \\
=: &\left\{\hat{\mathbf{P}}_i^{(z-1)\perp}\mathbf{b}_{i^z,\min(I^z_{i^z},K)}\right\}_{z=0}^{t-1}.
\end{aligned}$$

We have $\operatorname{proj}_{\mathbf{P}_i^{(z-1)\perp}\mathbf{y}_i^{(z)}} = \frac{(\mathbf{P}_i^{(z-1)\perp}\mathbf{y}_i^{(z)})(\mathbf{P}_i^{(z-1)\perp}\mathbf{y}_i^{(z)})^T}{\|\mathbf{P}_i^{(z-1)\perp}\mathbf{y}_i^{(z)}\|^2}$ and therefore write the projection $\hat{\mathbf{P}}_i^{(z-1)}$ onto $\mathcal{V}_i^{(z-1)} \cup \{\mathbf{y}_i^{(z)}\}$ as

$$\hat{\mathbf{P}}_i^{(z-1)} = \mathbf{P}_i^{(z-1)} + \frac{(\mathbf{P}_i^{(z-1)\perp}\mathbf{y}_i^{(z)})(\mathbf{P}_i^{(z-1)\perp}\mathbf{y}_i^{(z)})^T}{\|\mathbf{P}_i^{(z-1)\perp}\mathbf{y}_i^{(z)}\|^2}.$$

The orthogonalized vectors give us a basis in which we can write the norm $\|\mathbf{P}_i^{(t-1)}\mathbf{u}\|^2$ as

$$\sum_{z=0}^{t-1}\left|\left\langle \frac{\mathbf{P}_i^{(z-1)\perp}\mathbf{y}_i^{(z)}}{\|\mathbf{P}_i^{(z-1)\perp}\mathbf{y}_i^{(z)}\|}, \mathbf{u}\right\rangle\right|^2 + \sum_{z=0,\, I^z_{i^z}\le K}^{t-1}\left|\left\langle \frac{\hat{\mathbf{P}}_i^{(z-1)\perp}\mathbf{b}_{i^z,I^z_{i^z}}}{\|\hat{\mathbf{P}}_i^{(z-1)\perp}\mathbf{b}_{i^z,I^z_{i^z}}\|}, \mathbf{u}\right\rangle\right|^2. \tag{3}$$

Note that the set we applied Gram-Schmidt on was not linearly independent so we may get $\mathbf{0}$-vectors. These do not influence the calculations, so we simply assume they are not present in (3) from now on. The first term in (3) is bounded by ta^2 by the induction hypothesis. Let z be arbitrary but fixed and assume $I^z_{i^z} \le K$. Recall the definition of $\mathcal{U}^{(t)}$. Then $\mathbf{u} = \mathbf{b}_{i,j}$ for some $j > I_i^{t-1} \ge I_i^z$. $\mathbf{B}$ has orthonormal columns and so $\mathbf{u} \perp \mathbf{b}_{i^z,I^z_{i^z}}$. We will bound the second term in (3) now:

$$\begin{aligned}
&\left|\left\langle \hat{\mathbf{P}}_i^{(z-1)\perp}\mathbf{b}_{i^z,I^z_{i^z}}, \mathbf{u}\right\rangle\right| \\
= &\left|\left\langle \mathbf{b}_{i^z,I^z_{i^z}} - \hat{\mathbf{P}}_i^{(z-1)}\mathbf{b}_{i^z,I^z_{i^z}}, \mathbf{u}\right\rangle\right| \\
= &\left|\left\langle \mathbf{b}_{i^z,I^z_{i^z}} - \mathbf{P}_i^{(z-1)}\mathbf{b}_{i^z,I^z_{i^z}} - \frac{(\mathbf{P}_i^{(z-1)\perp}\mathbf{y}_i^{(z)})(\mathbf{P}_i^{(z-1)\perp}\mathbf{y}_i^{(z)})^T}{\|\mathbf{P}_i^{(z-1)\perp}\mathbf{y}_i^{(z)}\|^2}\mathbf{b}_{i^z,I^z_{i^z}}, \mathbf{u}\right\rangle\right| \\
\le &\left|\left\langle \mathbf{P}_i^{(z-1)}\mathbf{b}_{i^z,I^z_{i^z}}, \mathbf{u}\right\rangle\right| + \left|\left\langle \frac{\mathbf{P}_i^{(z-1)\perp}\mathbf{y}_i^{(z)}}{\|\mathbf{P}_i^{(z-1)\perp}\mathbf{y}_i^{(z)}\|}, \mathbf{u}\right\rangle\left\langle \frac{\mathbf{P}_i^{(z-1)\perp}\mathbf{y}_i^{(z)}}{\|\mathbf{P}_i^{(z-1)\perp}\mathbf{y}_i^{(z)}\|}, \mathbf{b}_{i^z,I^z_{i^z}}\right\rangle\right|,
\end{aligned} \tag{4}$$

[6] Where $\mathbf{P}_i^{(-1)} = \mathbf{0}_{d/n,d/n}$ is the zero matrix for convenience.

where in the last step we used $\mathbf{u} \perp \mathbf{b}_{i^z, I^z_{i^z}}$ and the triangle inequality. For an orthonormal projector $\mathbf{P}$ and any vectors $\mathbf{v}, \mathbf{u}$ we have $\langle \mathbf{P}\mathbf{v}, \mathbf{u} \rangle = \langle \mathbf{P}\mathbf{v}, \mathbf{P}\mathbf{u} \rangle$. Therefore the left term in (4) can be bounded by b^2 as follows:

$$\begin{aligned} \left| \left\langle \mathbf{P}_i^{(z-1)} \mathbf{b}_{i^z, I^z_{i^z}}, \mathbf{u} \right\rangle \right| &= \left| \left\langle \mathbf{P}_i^{(z-1)} \mathbf{b}_{i^z, I^z_{i^z}}, \mathbf{P}_i^{(z-1)} \mathbf{u} \right\rangle \right| \\ &\leq \| \mathbf{P}_i^{(z-1)} \mathbf{b}_{i^z, I^z_{i^z}} \| \| \mathbf{P}_i^{(z-1)} \mathbf{u} \| \\ &\leq b^2. \end{aligned} \tag{5}$$

The last step holds because of the induction hypothesis. Indeed, we have $\mathbf{u} \in \mathcal{U}^{(t)} \subset \mathcal{U}^{(z)}$ and $\mathbf{b}_{i^z, I^z_{i^z}} = \mathbf{b}_{i^z, I^{z-1}_{i^z}+1} \in \mathcal{U}^{(z)}_{i^z} \subset \mathcal{U}^{(z)}$.

Next, our assumption is that $G^{\leq T}$ happens and therefore G^z as well. Using its definition twice on the right term in (4) yields

$$\left| \left\langle \frac{\mathbf{P}_i^{(z-1)\perp} \mathbf{y}_i^{(z)}}{\| \mathbf{P}_i^{(z-1)\perp} \mathbf{y}_i^{(z)} \|}, \mathbf{u} \right\rangle \left\langle \frac{\mathbf{P}_i^{(z-1)\perp} \mathbf{y}_i^{(z)}}{\| \mathbf{P}_i^{(z-1)\perp} \mathbf{y}_i^{(z)} \|}, \mathbf{b}_{i^z, I^z_{i^z}} \right\rangle \right| \leq a^2. \tag{6}$$

We bound the norm in the denominator of the right term in (3) by

$$\begin{aligned} \| \hat{\mathbf{P}}_i^{(z-1)\perp} \mathbf{b}_{i^z, I^z_{i^z}} \|^2 &= \| \mathbf{b}_{i^z, I^z_{i^z}} \|^2 - \| \hat{\mathbf{P}}_i^{(z-1)} \mathbf{b}_{i^z, I^z_{i^z}} \|^2 \\ &= \| \mathbf{b}_{i^z, I^z_{i^z}} \|^2 - \| \mathbf{P}_i^{(z-1)} \mathbf{b}_{i^z, I^z_{i^z}} \|^2 - \left| \left\langle \frac{\mathbf{P}_i^{(z-1)\perp} \mathbf{y}_i^{(z)}}{\| \mathbf{P}_i^{(z-1)\perp} \mathbf{y}_i^{(z)} \|}, \mathbf{b}_{i^z, I^z_{i^z}} \right\rangle \right|^2 \\ &\geq 1 - b^2 - a^2. \end{aligned}$$

The first step is justified by the Pythagorean theorem because $\hat{\mathbf{P}}_i^{(z-1)\perp} \mathbf{b}_{i^z, I^z_{i^z}}$ and $\hat{\mathbf{P}}_i^{(z-1)} \mathbf{b}_{i^z, I^z_{i^z}}$ are orthogonal. The second follows by the Pythagorean theorem and the definition of $\hat{\mathbf{P}}^{(z-1)}$. For the inequality, we use the same arguments as in (5) and (6).

We can return to (3). Recall that $b = \sqrt{3T}a$ and thus $a^2 + b^2 = 3Ta^2 + a^2 = (3T+1)a^2 \leq a$ by definition of a. We use this in step $(*)$ below:

$$\begin{aligned} \| \mathbf{P}_i^{(t-1)} \mathbf{u} \|^2 &= \sum_{z=0}^{t-1} \left| \left\langle \frac{\mathbf{P}_i^{(z-1)\perp} \mathbf{y}_i^{(z)}}{\| \mathbf{P}_i^{(z-1)\perp} \mathbf{y}_i^{(z)} \|}, \mathbf{u} \right\rangle \right|^2 + \sum_{z=0,\, I^z_{i^z} \leq K}^{t-1} \left| \left\langle \frac{\hat{\mathbf{P}}_i^{(z-1)\perp} \mathbf{b}_{i^z, I^z_{i^z}}}{\| \hat{\mathbf{P}}_i^{(z-1)\perp} \mathbf{b}_{i^z, I^z_{i^z}} \|}, \mathbf{u} \right\rangle \right|^2 \\ &\leq ta^2 + t \frac{(a^2+b^2)^2}{1-(a^2+b^2)} \\ &\overset{(*)}{\leq} ta^2 + t \frac{a^2}{1-a} \\ &\leq 3Ta^2 \\ &= b^2, \end{aligned}$$

where the last inequality holds because $a \leq 1/2$ and $t \leq T$. This concludes the induction. We have thus proven that $G^{\leq T}$ implies our result, namely that equation (2) holds for all $t \in \{0, \dots, T\}$, all $i \in [n]$ and all corresponding $\mathbf{u} \in \mathcal{U}_i^{(t)}$.

We now derive an upper bound for the probability of the complement event $(G^{\leq T})^c$. Note that if $G^{\leq T}$ does not happen, then there is a smallest t for which it fails. For convenience, let $G^{<1}$ be an event that always happens. Using a union bound, this argumentation is reflected by

$$\mathbb{P}((G^{\leq T})^c \leq \sum_{t=1}^{T} \mathbb{P}((G^{\leq t})^c \,|\, G^{<t}). \tag{7}$$

We will bound the probability $\mathbb{P}((G^{\leq t})^c \,|\, G^{<t})$. For the remainder of the proof, we need matrices analogous to the sets $\mathcal{U}^{(t)}$ and $\tilde{\mathcal{U}}^{(t)}$. First define $\hat{i}^t$ to be the sequence $i^{0:t-1}$. Then let

$$\tilde{\mathbf{U}}_j^{\hat{i}^t} = \left[\mathbf{b}_{j,1} \,|\, \cdots \,|\, \mathbf{b}_{j,\min(K, I_j^{t-1})}\right],$$

where I_j^{t-1} is according to the sequence $\hat{i}^t$. Then define $\tilde{\mathbf{U}}^{\hat{i}^t} = [\tilde{\mathbf{U}}_1^{\hat{i}^t} \cdots \tilde{\mathbf{U}}_n^{\hat{i}^t}]$. Similarly, we define the "complement" matrices

$$\mathbf{U}_j^{\hat{i}^t} = \left[\mathbf{b}_{j, I_j^{t-1}+1} \,|\, \cdots \,|\, \mathbf{b}_{j,K}\right].$$

Note that for any j, one of $\mathbf{U}_j^{\hat{i}^t}$ or $\tilde{\mathbf{U}}_j^{\hat{i}^t}$ could potentially be empty. This will not be problematic in what follows. Analogous to before $\mathbf{U}^{\hat{i}^t} = [\mathbf{U}_1^{\hat{i}^t} \cdots \mathbf{U}_n^{\hat{i}^t}]$. Finally $\bar{\mathbf{U}}^{\hat{i}^t} = [\tilde{\mathbf{U}}^{\hat{i}^t}, \mathbf{U}^{\hat{i}^t}]$ is a matrix with all columns of $\mathbf{B}$, but in different order. For our event, by the law of total probability we have

$$\begin{aligned} &\mathbb{P}((G^{\leq t})^c \,|\, G^{<t}) \\ &= \sum_{\hat{i}_0^t \in \hat{S}^t} \mathbb{E}_{\xi, \tilde{\mathbf{U}}^{\hat{i}_0^t}} \left[\mathbb{P}((G^{\leq t})^c \,|\, G^{<t}, \hat{i}^t = \hat{i}_0^t, \xi, \tilde{\mathbf{U}}^{\hat{i}_0^t})\, \mathbb{P}(\hat{i}^t = \hat{i}_0^t \,|\, G^{<t}, \xi, \tilde{\mathbf{U}}^{\hat{i}_0^t})\right]. \end{aligned} \tag{8}$$

In the rest, we show for all (fixed) t, $\xi_0, \tilde{\mathbf{U}}_0, \hat{i}_0^t$ a bound on the probability

$$\begin{aligned} &\mathbb{P}((G^{\leq t})^c \,|\, G^{<t}, \hat{i}^t = \hat{i}_0^t, \xi = \xi_0, \tilde{\mathbf{U}}^{\hat{i}_0^t} = \tilde{\mathbf{U}}_0) \\ &\leq \sum_{\substack{i \in [n] \\ \mathbf{u} \in \mathcal{U}^{(t)}}} \mathbb{P}\left(\left|\langle \mathbf{u}, \mathbf{P}_i^{(t-1)\perp} \mathbf{y}_i^{(t)} \rangle\right| \geq a \|\mathbf{P}_i^{(t-1)\perp} \mathbf{y}_i^{(t)}\| \,|\, G^{<t}, \hat{i}^t = \hat{i}_0^t, \xi = \xi_0, \tilde{\mathbf{U}}^{\hat{i}_0^t} = \tilde{\mathbf{U}}_0\right). \end{aligned} \tag{9}$$

A bound on (9) is also a bound for (8), because

$$\sum_{\hat{i}_0^t \in \hat{S}^t} \mathbb{E}_{\xi, \tilde{\mathbf{U}}^{\hat{i}_0^t}} \mathbb{P}(\hat{i}^t = \hat{i}_0^t \,|\, G^{<t}, \xi, \tilde{\mathbf{U}}^{\hat{i}_0^t}) = 1.$$

First, we show that given $G^{<t}$, the next iterate $[i^t, \mathbf{x}^{(t)}]$ produced by A only depends on $\tilde{\mathbf{U}}^{\hat{i}^t}$ and not the full draw of $\bar{\mathbf{U}}^{\hat{i}^t}$, because $\bar{f}_K$ is a robust zero-chain. This is formalized below:

Lemma B.1. *For every $t \in [T]$, there exist measurable functions $A_+^{(t)}$ and $A_-^{(t)}$ such that*

$$[i^t, \mathbf{x}^{(t)}] = A_+^{(t)}(\xi, \tilde{\mathbf{U}}^{\hat{i}^t}, \hat{i}^t)\mathbb{1}_{G^{<t}} + A_-^{(t)}(\xi, \bar{\mathbf{U}}^{\hat{i}^t}, \hat{i}^t)\mathbb{1}_{(G^{<t})^c}.$$

Proof [of Lemma B.1]**.** Recall the definition $f_i^*(\mathbf{x}) = \hat{f}_{K;\mathbf{B}_i}(\mathbf{C}_i^T\mathbf{x}) = \bar{f}_K(\mathbf{B}_i^T \rho(\mathbf{C}_i^T \mathbf{x})) + \frac{1}{10}\|\mathbf{C}_i^T\mathbf{x}\|^2$ for convenience. The sequence $\{[i^t, \mathbf{x}^{(t)}]\}_{t \in \mathbb{N}}$ is informed by F^*. Therefore, for any $t \in \mathbb{N}$, there exists a measurable mapping $A^{(t)}$ such that:

$$[i^t, \mathbf{x}^{(t)}] = A^{(t)}\left\{\xi, \hat{i}^t, \mathbf{x}^{(0:t-1)}, \nabla^{(0:q)} f_{i^0}^*(\mathbf{x}^{(0)}), \ldots, \nabla^{(0:q)} f_{i^{t-1}}^*(\mathbf{x}^{(t-1)})\right\}.$$

We show our result by induction on $t \in [T]$. The base case is clear, as the first iterate is $\mathbf{x}^{(0)} = \mathbf{0}$. For the step, assume $G^{<t+1}$ happens and that the result holds for any $s \leq t$. By the derivation on the previous pages we have $|\langle \mathbf{b}_{i^t,j}, \mathbf{y}_{i^t}^{(t)} \rangle| < 1/2$ for all $j \geq I_{i^t}^{t-1} + 1 = I_{i^t}^t$. Then because $\bar{f}_K$ is a robust zero-chain and $\mathbf{C}$ is fixed, $\nabla^{(0:q)} f_i^*(\mathbf{x}^{(t)})$ only depends on $\mathbf{x}^{(t)}$ and columns of $\mathbf{B}_{i^t}$ with indices up to $\min(K, I_{i^t}^t)$. Note that $\tilde{\mathbf{U}}^{\hat{i}^{t+1}}$ contains all of those columns of $\mathbf{B}_{i^t}$. Therefore the computation of the pair $[i^t, \mathbf{x}^{(t)}]$ only depends on $\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(t)}$, $\hat{i}^{t+1}$ and $\tilde{\mathbf{U}}^{\hat{i}^{t+1}}$ in case $G^{<t+1}$ happens. In that case, we may write

$$[i^{t+1}, \mathbf{x}^{(t+1)}] = A_+^{(t+1)}(\xi, \tilde{\mathbf{U}}^{\hat{i}^{t+1}}, \hat{i}^{t+1}),$$

with the dependence on the previous iterates being implicit (justified by the induction hypothesis). This leads to the statement of this sub-lemma. □

For $t \in [T]$, condition on $G^{<t}$, $\hat{i}^t = \hat{i}_0^t$, $\xi = \xi_0$ and $\tilde{\mathbf{U}}^{\hat{i}_0^t} = \tilde{\mathbf{U}}_0$. Consequently, the iterates $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(t)}$ are deterministic and so are the $\mathbf{y}_i$'s. Thus for all $i \in [n]$, the quantity $\mathbf{P}_i^{(t-1)\perp}\mathbf{y}_i^{(t)}$ is deterministic as well (recall the definition of $\mathcal{V}_i^{(t-1)}$).

For any (still random) $\mathbf{u} \in \mathcal{U}_i^{(t)}$, we are interested in (recall (9)):

$$\begin{aligned}
&\mathbb{P}\left(\left|\langle \mathbf{u}, \mathbf{P}_i^{(t-1)\perp}\mathbf{y}_i^{(t)}\rangle\right| \geq a\|\mathbf{P}_i^{(t-1)\perp}\mathbf{y}_i^{(t)}\| \,|\, G^{<t}, \hat{i}^t = \hat{i}_0^t, \xi = \xi_0, \tilde{\mathbf{U}}^{\hat{i}_0^t} = \tilde{\mathbf{U}}_0\right) \\
\leq\ &\mathbb{P}\left(\left|\left\langle \frac{\mathbf{P}_i^{(t-1)\perp}\mathbf{u}}{\|\mathbf{P}_i^{(t-1)\perp}\mathbf{u}\|}, \frac{\mathbf{P}_i^{(t-1)\perp}\mathbf{y}_i^{(t)}}{\|\mathbf{P}_i^{(t-1)\perp}\mathbf{y}_i^{(t)}\|}\right\rangle\right| \geq a \,|\, G^{<t}, \hat{i}^t = \hat{i}_0^t, \xi = \xi_0, \tilde{\mathbf{U}}^{\hat{i}_0^t} = \tilde{\mathbf{U}}_0\right).
\end{aligned}$$

The inequality follows because $\|\mathbf{P}_i^{(t-1)\perp}\mathbf{u}\| \leq \|\mathbf{u}\|$, which holds as $\mathbf{P}_i^{(t-1)\perp}$ is an orthogonal projector. By the previous discussion, we know the second term in this scalar product is a deterministic unit vector in the space orthogonal to $\mathcal{V}_i^{(t-1)}$ [7]. What remains to study is the distribution of $\frac{\mathbf{P}_i^{(t-1)\perp}\mathbf{u}}{\|\mathbf{P}_i^{(t-1)\perp}\mathbf{u}\|}$. We wish to show that $\frac{\mathbf{P}_i^{(t-1)\perp}\mathbf{u}}{\|\mathbf{P}_i^{(t-1)\perp}\mathbf{u}\|}$ is a uniformly distributed unit vector in the space orthogonal to $\mathcal{V}_i^{(t-1)}$. Let $\mathbf{Z}_i \in \mathbb{R}^{d/n \times d/n}$ be a rotation that lets the span of $\mathcal{V}_i^{(t-1)}$ invariant, i.e. $\mathbf{Z}_i^T\mathbf{u} = \mathbf{u} = \mathbf{Z}_i\mathbf{u}$ for any $\mathbf{u} \in \mathcal{V}_i^{(t-1)}$. For a random variable X, let p_X denote its density. We want to show the equality:

$$\begin{aligned}
&p_{\mathbf{U}^{\hat{i}_0^t}}(\mathbf{U}_0 \,|\, G^{<t}, \hat{i}^t = \hat{i}_0^t, \xi = \xi_0, \tilde{\mathbf{U}}^{\hat{i}_0^t} = \tilde{\mathbf{U}}_0) \\
=\ &p_{\mathbf{U}^{\hat{i}_0^t}}(\mathbf{Z}_i\mathbf{U}_0 \,|\, G^{<t}, \hat{i}^t = \hat{i}_0^t, \xi = \xi_0, \tilde{\mathbf{U}}^{\hat{i}_0^t} = \tilde{\mathbf{U}}_0),
\end{aligned}$$

to show the distribution of $\frac{\mathbf{P}_i^{(t-1)\perp}\mathbf{u}}{\|\mathbf{P}_i^{(t-1)\perp}\mathbf{u}\|}$ is indeed uniform.

Let $\bar{\mathbf{U}}_0 = [\tilde{\mathbf{U}}_0, \mathbf{U}_0]$. We lighten the notation up a bit by omitting the random variables where they are clear from context. Using conditional densities:

$$\begin{aligned}
p_{\mathbf{U}^{\hat{i}_0^t}}(\mathbf{U}_0 \,|\, \hat{i}_0^t, G^{<t}, \xi_0, \tilde{\mathbf{U}}_0) &= \frac{\mathbb{P}(G^{<t}, \hat{i}^t = \hat{i}_0^t \,|\, \xi_0, \mathbf{U}_0, \tilde{\mathbf{U}}_0) p_{\xi, \bar{\mathbf{U}}^{\hat{i}_0^t}}(\xi_0, \bar{\mathbf{U}}_0)}{\mathbb{P}(G^{<t}, \hat{i}^t = \hat{i}_0^t \,|\, \xi_0, \tilde{\mathbf{U}}_0) p_{\xi, \tilde{\mathbf{U}}^{\hat{i}_0^t}}(\xi_0, \tilde{\mathbf{U}}_0)} \\
&= \frac{\mathbb{P}(G^{<t}, \hat{i}^t = \hat{i}_0^t \,|\, \xi_0, \bar{\mathbf{U}}_0) p_{\bar{\mathbf{U}}^{\hat{i}_0^t}}(\bar{\mathbf{U}}_0)}{\mathbb{P}(G^{<t}, \hat{i}^t = \hat{i}_0^t \,|\, \xi_0, \tilde{\mathbf{U}}_0) p_{\tilde{\mathbf{U}}^{\hat{i}_0^t}}(\tilde{\mathbf{U}}_0)}.
\end{aligned}$$

Plugging in $\mathbf{Z}_i\mathbf{U}_0$ and using $\mathbf{Z}_i\tilde{\mathbf{U}}_0 = \tilde{\mathbf{U}}_0$ we obtain

$$p_{\mathbf{U}^{\hat{i}_0^t}}(\mathbf{Z}_i\mathbf{U}_0 \,|\, G^{<t}, \hat{i}_0^t, \xi_0, \tilde{\mathbf{U}}_0) = \frac{\mathbb{P}(G^{<t}, \hat{i}^t = \hat{i}_0^t \,|\, \xi_0, \mathbf{Z}_i\bar{\mathbf{U}}_0) p_{\bar{\mathbf{U}}^{\hat{i}_0^t}}(\mathbf{Z}_i\bar{\mathbf{U}}_0)}{\mathbb{P}(G^{<t}, \hat{i}^t = \hat{i}_0^t \,|\, \xi_0, \tilde{\mathbf{U}}_0) p_{\tilde{\mathbf{U}}^{\hat{i}_0^t}}(\tilde{\mathbf{U}}_0)}.$$

Because of the uniform distribution of $\mathbf{B}$ and thus also of $\bar{\mathbf{U}}^{\hat{i}_0^t}$, it suffices to show that

$$\mathbb{P}(G^{<t}, \hat{i}^t = \hat{i}_0^t \,|\, \xi_0, \bar{\mathbf{U}}_0) = \mathbb{P}(G^{<t}, \hat{i}^t = \hat{i}_0^t \,|\, \xi_0, \mathbf{Z}_i\bar{\mathbf{U}}_0).$$

This probability is either 0 or 1, because we condition on all the randomness involved. We show by induction on $s \in [t]$ that $\mathbb{P}(G^{<t}, \hat{i}^t = \hat{i}_0^t \,|\, \xi_0, \bar{\mathbf{U}}_0) = 1$ implies $\mathbb{P}(G^{<t}, \hat{i}^s = \hat{i}_0^s \,|\, \xi_0, \mathbf{Z}_i\bar{\mathbf{U}}_0) = 1$. The other direction is analogous.

[7] This set is also deterministic as a consequence of the conditioned variables.

Therefore assume $\hat{i}^t = \hat{i}_0^t$ and that $G^{<t}$ happens, conditioned on $\xi = \xi_0$ and $\bar{\mathbf{U}}^{\hat{i}_0^t} = \bar{\mathbf{U}}_0$. The base case is trivial, because $G^{<1}$ always happens. For the inductive step, let $s \geq 2$ and assume that $\hat{i}^{s-1} = \hat{i}_0^{s-1}$ and $G^{<s-1}$ happen, conditioned on $\xi = \xi_0$ and $\bar{\mathbf{U}}^{\hat{i}_0^t} = \mathbf{Z}_i \bar{\mathbf{U}}_0$ (induction hypothesis).

Let i'^{s-1}, $\mathbf{x}'^{(s-1)}$ denote the next index and iterate the algorithm produces, given $\bar{\mathbf{U}}^{\hat{i}_0^t} = \mathbf{Z}_i \bar{\mathbf{U}}_0$. By Lemma B.1, the induction hypothesis allows us to write for some $A_+^{(s-1)}$

$$\begin{aligned}
[i'^{s-1}, \mathbf{x}'^{(s-1)}] &= A_+^{(s-1)}(\xi_0, \mathbf{Z}_i \tilde{\mathbf{U}}_0, \hat{i}_0^{s-1}) \\
&= A_+^{(s-1)}(\xi_0, \tilde{\mathbf{U}}_0, \hat{i}_0^{s-1}) \\
&= [i^{s-1}, \mathbf{x}^{(s-1)}], \qquad (10)
\end{aligned}$$

where we also used $\mathbf{Z}_i \tilde{\mathbf{U}}_0 = \tilde{\mathbf{U}}_0$. This means that $\hat{i}^s = \hat{i}_0^s$ iff $\hat{i}'^s = \hat{i}_0^s$, which gets us halfway there. We just have to show that $G^{<s}$ happens as well, given $\bar{\mathbf{U}}^{\hat{i}_0^t} = \mathbf{Z}_i \bar{\mathbf{U}}_0$. Of course, showing G^{s-1} suffices, by the induction hypothesis. For this, let $\mathbf{u} \in \mathcal{U}^{(s-1)}$ and $i \in [n]$. We have

$$\begin{aligned}
\left\langle \mathbf{Z}_i \mathbf{u}, \frac{\mathbf{P}_i^{(s-2)\perp} \mathbf{y}_i'^{(s-1)}}{\|\mathbf{P}_i^{(s-2)\perp} \mathbf{y}_i'^{(s-1)}\|} \right\rangle &= \left\langle \mathbf{u}, \mathbf{Z}_i^T \frac{\mathbf{P}_i^{(s-2)\perp} \mathbf{y}_i^{(s-1)}}{\|\mathbf{P}_i^{(s-2)\perp} \mathbf{y}_i^{(s-1)}\|} \right\rangle \\
&= \left\langle \mathbf{u}, \frac{\mathbf{P}_i^{(s-2)\perp} \mathbf{y}_i^{(s-1)}}{\|\mathbf{P}_i^{(s-2)\perp} \mathbf{y}_i^{(s-1)}\|} \right\rangle .
\end{aligned}$$

The first equality follows because $\mathbf{P}_i^{(s-2)\perp} \mathbf{y}_i'^{(s-1)} = \mathbf{P}_i^{(s-2)\perp} \mathbf{y}_i^{(s-1)}$ by (10) and the second step follows because $\mathbf{P}_i^{(s-2)\perp} \mathbf{y}_i^{(s-1)} = \mathbf{y}_i^{(s-1)} - \mathbf{P}_i^{(s-2)} \mathbf{y}_i^{(s-1)}$ is in the span of $\mathcal{V}_i^{(s-1)} \subset \mathcal{V}_i^{(t)}$ and left invariant by $\mathbf{Z}_i^T$. Thus G^{s-1} holds as well, conditioned on $\bar{\mathbf{U}}^{\hat{i}_0^t} = \mathbf{Z}_i \bar{\mathbf{U}}_0$.

This concludes the inductive step and therefore our proof that $\frac{\mathbf{P}_i^{(t-1)\perp} \mathbf{u}}{\|\mathbf{P}_i^{(t-1)\perp} \mathbf{u}\|}$ is a uniformly distributed unit vector in a subspace of $\mathbb{R}^{d/n}$ of dimension at least

$$d' \geq d/n - |\mathcal{V}_i^{(t-1)}| \geq d/n - (t-1) - \sum_{j=1}^{n} \min(I_j^{(t-1)}, K) \geq d/n - 2t.$$

We may write our probability to bound

$$\mathbb{P}\left(\left| \left\langle \frac{\mathbf{P}_i^{(t-1)\perp} \mathbf{u}}{\|\mathbf{P}_i^{(t-1)\perp} \mathbf{u}\|}, \frac{\mathbf{P}_i^{(t-1)\perp} \mathbf{y}_i^{(t)}}{\|\mathbf{P}_i^{(t-1)\perp} \mathbf{y}_i^{(t)}\|} \right\rangle \right| \geq a \,\middle|\, G^{<t}, \hat{i}^t = \hat{i}_0^t, \xi = \xi_0, \tilde{\mathbf{U}}^{\hat{i}_0^t} = \tilde{\mathbf{U}}_0 \right)$$

as

$$\mathbb{P}(|v_1| \geq a),$$

where $\mathbf{v}$ is a uniformly distributed unit vector in $\mathbb{R}^{d'}$. This is because for the dot product, only the angle between the two vectors matters and with all conditioned variables, $\frac{\mathbf{P}_i^{(t-1)\perp} \mathbf{y}_i^{(t)}}{\|\mathbf{P}_i^{(t-1)\perp} \mathbf{y}_i^{(t)}\|}$ is fixed so we may assume w.l.o.g. that it is equal to $\mathbf{e}_1$. By a standard concentration of measure bound on the sphere (see Lecture 8 in Ball (1997)) we get

$$\mathbb{P}(|v_1| \geq a) = \mathbb{P}(|v_1| > a) \leq 2e^{-d'a^2/2} \leq 2e^{-\frac{a^2}{2}(d/n - 2t)} \leq 2e^{-\frac{a^2}{2}(d/n - 2T)}.$$

Returning to (9) we get for all $t \in [T]$ a bound for (8) of

$$\mathbb{P}((G^{\leq t})^c \,|\, G^{<t}) \leq n \cdot nK \cdot 2e^{-\frac{a^2}{2}(d/n - 2T)},$$

and therefore by (7)

$$\mathbb{P}((G^{\leq T})^c) \leq T \cdot n \cdot nK \cdot 2e^{-\frac{a^2}{2}(d/n-2T)} \leq 2(nK)(n^2K)e^{-\frac{a^2}{2}(d/n-2T)}.$$

Setting

$$d/n \geq \frac{2}{a^2}\log\left(\frac{2n^3K^2}{\delta}\right) + 2T$$

gives us a probability δ bound. By the definitions of a and T, the choice

$$d/n \geq 2\max(9n^2K^2, 12nKR^2)\log\left(\frac{2n^3K^2}{\delta}\right) + nK$$

suffices. This concludes the proof. □

B.3 Proof of Lemma 4.6

The proof is related to Carmon et al. (2019a), Lemma 5.

Proof [of Lemma 4.6]**.** Fix $t \in \{0, \ldots, T\}$. For any $i \in [n]$, define $\mathbf{y}_i^{(t)} = \rho(\mathbf{C}_i^T\mathbf{x}^{(t)})$. Then Lemma 4.5 gives for all $\mathbf{u} \in \mathcal{U}_i^{(t)}$ that $|\langle \mathbf{u}, \mathbf{y}_i^{(t)}\rangle| < 1/2$. Therefore for each i with $\mathcal{U}_i^{(t)} \neq \emptyset$ we have some $k \in [K]$ with

$$|\langle \mathbf{b}_{i,k}, \mathbf{y}_i^{(t)}\rangle| < \frac{1}{2} < 1.$$

With $\mathbf{z} = \mathbf{B}_i^T\mathbf{y}_i^{(t)}$, by Lemma A.1 there exists an index $j \leq k$ with $|z_j| = |\langle \mathbf{b}_{i,j}, \mathbf{y}_i^{(t)}\rangle| < 1$ and

$$\left|\frac{\partial \bar{f}_K}{\partial z_j}(\mathbf{B}_i^T\mathbf{y}^{(t)})\right| = \left|\frac{\partial \bar{f}_K}{\partial z_j}(\mathbf{z})\right| > 1.$$

Define $\tilde{f}_{K;B_i}(\mathbf{y}_i^{(t)}) = \bar{f}_K(\mathbf{B}_i^T\mathbf{y}_i^{(t)})$ and recall the definitions of $\bar{f}_K$ and $\hat{f}_{K;\mathbf{B}_i}$. They give

$$\hat{f}_{K;\mathbf{B}_i}(\mathbf{C}_i^T\mathbf{x}) = \tilde{f}_{K;\mathbf{B}_i}(\rho(\mathbf{C}_i^T\mathbf{x})) + \frac{1}{10}\|\mathbf{C}_i^T\mathbf{x}\|^2 = \bar{f}_K(\mathbf{B}_i^T\rho(\mathbf{C}_i^T\mathbf{x})) + \frac{1}{10}\|\mathbf{C}_i^T\mathbf{x}\|^2.$$

By the chain rule we have

$$\mathbf{B}_i^T(\nabla\tilde{f}_{K;\mathbf{B}_i}(\mathbf{y}_i^{(t)})) = \mathbf{B}_i^T(\mathbf{B}_i\nabla\bar{f}_K(\mathbf{B}_i^T\mathbf{y}_i^{(t)})) = \nabla\bar{f}_K(\mathbf{B}_i^T\mathbf{y}_i^{(t)}).$$

Combining this with the above we deduce that

$$\left|\langle \mathbf{b}_{i,j}, \nabla\tilde{f}_{K;\mathbf{B}_i}(\mathbf{y}_i^{(t)})\rangle\right| = \left|\frac{\partial \bar{f}_K}{\partial z_j}(\mathbf{B}_i^T\mathbf{y}^{(t)})\right| > 1.$$

Carmon et al. (2019a) show that $|\langle \mathbf{b}_{i,j}, \mathbf{y}_i^{(t)}\rangle| < 1$ and $\left|\langle \mathbf{b}_{i,j}, \nabla\tilde{f}_{K;\mathbf{B}_i}(\mathbf{y}_i^{(t)})\rangle\right| > 1$ imply

$$\|\nabla\hat{f}_{K;\mathbf{B}_i}(\mathbf{C}_i^T\mathbf{x}^{(t)})\| > \frac{1}{2},$$

where the gradient is w.r.t. the function argument, i.e. $\mathbf{C}_i^T\mathbf{x}^{(t)}$. They show this in the proof of Lemma 5, in the calculations following Equation (14) [8].

[8] With slightly different naming. Replace U with $\mathbf{B}_i$ and $u^{(j)}$ with $\mathbf{b}_{i,j}$, T with K and $x^{(t)}$ with $\mathbf{C}_i^T\mathbf{x}^{(t)}$. Also note that this is the part where the added regularization term in $\hat{f}$ is needed.

The only thing that remains to show is that this indeed guarantees $\nabla F^*(\mathbf{x}^{(t)})$ to be large. Note that in each iteration, one of the $\mathcal{U}_i$'s shrinks in size by at most 1, while the others do not change. That means that after $t \leq T = \frac{nK}{2}$ iterations, at most $\lfloor n/2 \rfloor$ indices i can have $\mathcal{U}_i^{(t)} = \emptyset$. Let $J \subset [K]$ be the set of those indices i with $\mathcal{U}_i^{(t)} \neq \emptyset$. Then $|J| \geq n/2$ and

$$\begin{aligned}
\|\nabla F^*(\mathbf{x}^{(t)})\|^2 &= \|\frac{1}{n}\sum_{i=1}^{n} \nabla f_i^*(\mathbf{x}^{(t)})\|^2 \\
&= \|\frac{1}{n}\sum_{i=1}^{n} \nabla \left[\hat{f}_{K;\mathbf{B}_i}\left(\mathbf{C}_i^T \mathbf{x}^{(t)}\right)\right]\|^2 \\
&= \|\frac{1}{n}\sum_{i=1}^{n} \mathbf{C}_i \nabla \hat{f}_{K;\mathbf{B}_i}\left(\mathbf{C}_i^T \mathbf{x}^{(t)}\right)\|^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \left(\nabla \hat{f}_{K;\mathbf{B}_i}\left(\mathbf{C}_i^T \mathbf{x}^{(t)}\right)\right)^T \mathbf{C}_i^T \mathbf{C}_j \nabla \hat{f}_{K;\mathbf{B}_j}\left(\mathbf{C}_j^T \mathbf{x}^{(t)}\right) \\
&\overset{(*)}{=} \frac{1}{n^2}\sum_{i=1}^{n} \left(\nabla \hat{f}_{K;\mathbf{B}_i}\left(\mathbf{C}_i^T \mathbf{x}^{(t)}\right)\right)^T \mathbf{C}_i^T \mathbf{C}_i \nabla \hat{f}_{K;\mathbf{B}_i}\left(\mathbf{C}_i^T \mathbf{x}^{(t)}\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \|\nabla \hat{f}_{K;\mathbf{B}_i}\left(\mathbf{C}_i^T \mathbf{x}^{(t)}\right)\|^2 \\
&\geq \frac{1}{n^2}\sum_{i\in J} \|\nabla \hat{f}_{K;\mathbf{B}_i}\left(\mathbf{C}_i^T \mathbf{x}^{(t)}\right)\|^2 \\
&\geq \frac{1}{n^2}\frac{n}{2}\frac{1}{4} \\
&\geq \frac{1}{16n}.
\end{aligned}$$

where $(*)$ is because of the definition of $\mathbf{C} \in \mathsf{Ortho}(d, d)$. □

B.4 Proof of Theorem 4.7

The function $\hat{f}_{K;\mathbf{B}_i}$ from Definition 4.2 has some very useful properties regarding its Lipschitz constants and its gap to optimality.

Lemma B.2 [Lemma 6 in Carmon et al. (2019a)]**.** *The function $\hat{f}_{K;\mathbf{B}_i}$ satisfies the following properties:*

i) $\hat{f}_{K;\mathbf{B}_i}(\mathbf{0}) - \inf_{\mathbf{y}\in\mathbb{R}^{d/n}} \hat{f}_{K;\mathbf{B}_i}(\mathbf{y}) \leq 12K$.

ii) For every $p \geq 1$, the pth-order derivatives of $\hat{f}_{K;\mathbf{B}_i}$ are $\hat{\ell}_p$-Lipschitz continuous, where $\hat{\ell}_p \leq \exp(cp\log p + c)$ for a numerical constant $c < \infty$.

With this, we can proceed with the proof of the main lower bound theorem.

Proof [of Theorem 4.7]**.** We define the functions

$$f_i(\mathbf{x}) = \lambda\sigma^{p+1} f_i^*\left(\frac{\mathbf{x}}{\sigma}\right) = \lambda\sigma^{p+1}\hat{f}_{K;\mathbf{B}_i}\left(\frac{\mathbf{C}_i^T\mathbf{x}}{\sigma}\right),$$

giving us

$$F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}).$$

We will choose the scaling parameters to ensure that our instance belongs to the desired function class. We have

$$\begin{aligned}\|\nabla^p f_i(\mathbf{x}) - \nabla^p f_i(\mathbf{y})\| &\leq \lambda\hat{\ell}_p \|\mathbf{C}_i^T\mathbf{x} - \mathbf{C}_i^T\mathbf{y}\| \\ &\leq \lambda\hat{\ell}_p\|\mathbf{x}-\mathbf{y}\|. \end{aligned} \tag{11}$$

The first inequality follows from Lemmas D.1 and B.2 and the second holds because $\mathbf{C}_i$ can be extended to the orthonormal matrix $\mathbf{C}$. The choice $\lambda = \frac{L_p}{\hat{\ell}_p}$ accomplishes our goal of smoothness with parameter L_p.

Now fix an algorithm A and assume $\{[i^t, \mathbf{x}^{(t)}]\}_{t\in\mathbb{N}}$ are the iterates produced by A on F. Consequently, by Lemma 4.4 $\{[i^t, \mathbf{x}^{(t)}/\sigma]\}_{t\in\mathbb{N}}$ is informed by F^*. Therefore, we can apply Lemma 4.6 on the sequence $\{[i^t, \mathbf{x}^{(t)}/\sigma]\}_{t\in\mathbb{N}}$ to bound

$$\begin{aligned}\|\nabla F(\mathbf{x}^{(t)})\|^2 &= \|\lambda\sigma^p\nabla F^*(\mathbf{x}^{(t)}/\sigma)\|^2 \\ &= \lambda^2\sigma^{2p}\|\nabla F^*(\mathbf{x}^{(t)}/\sigma)\|^2 \\ &\overset{4.6}{\geq} \frac{\sigma^{2p}\lambda^2}{16n},\end{aligned}$$

for all $t \in [0:T]$ with probability $1-\delta$ for a sufficiently large dimension d (that depends on δ). We will fix this dimension at the end. To get a lower bound for an ε precision requirement we can choose σ to be

$$\frac{\sigma^p\lambda}{4\sqrt{n}} = \varepsilon \iff \sigma = \left(\frac{4\sqrt{n}\varepsilon\hat{\ell}_p}{L_p}\right)^{\frac{1}{p}}.$$

As a last step, we will guarantee the optimality gap requirement. From Lemma B.2, we immediately have

$$F(\mathbf{0}) - \inf_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \leq 12\lambda\sigma^{p+1}K.$$

We require

$$12\lambda\sigma^{p+1}K = 12\frac{L_p}{\hat{\ell}_p}\left(\frac{4\sqrt{n}\varepsilon\hat{\ell}_p}{L_p}\right)^{\frac{p+1}{p}} K \leq \Delta.$$

To get the best possible bound, we choose

$$K = \left\lfloor \frac{\Delta}{192}\left(\frac{L_p}{\hat{\ell}_p}\right)^{\frac{1}{p}} \frac{1}{n^{\frac{p+1}{2p}}\varepsilon^{\frac{p+1}{p}}}\right\rfloor.$$

We will need that this K is at least 1 in order to get a sensible bound, as becomes clear in the subsequent steps. To enforce this, we may require that

$$\tilde{c}_p\Delta\left(L_p\right)^{\frac{1}{p}}\frac{1}{\varepsilon^{\frac{p+1}{p}}} \geq n^{\frac{p+1}{2p}},$$

or in other words,

$$n \leq c_p\frac{\Delta^{\frac{2p}{p+1}}L_p^{\frac{2}{p+1}}}{\varepsilon^2}$$

for some constants $c_p, \tilde{c}_p$ that depend on p. As Lemma 4.6 yields the lower bound $T = \frac{nK}{2}$ we get a lower bound of

$$\Omega\left(\left(\frac{L_p}{\hat{\ell}_p}\right)^{\frac{1}{p}}\frac{\Delta n^{\frac{p-1}{2p}}}{\varepsilon^{\frac{p+1}{p}}}\right)$$

with probability at least 1/2 for large enough dimension d (see below). Thus there must be a fixed function F such that for this many iterations – with probability 1/2 depending only on ξ – the iterates A produces on F all have gradient larger than ε. For the dimension requirement, one can plug in the values of K and $\delta = 1/2$ into the dimension requirement of Lemma 4.6, to see that some $d \in \tilde{\mathcal{O}}(n^{\frac{2p-1}{p}} \Delta L_p^{2/p} \varepsilon^{-\frac{2(p+1)}{p}}) \leq \tilde{\mathcal{O}}(n^2 \Delta L_p^2 \varepsilon^{-4})$ suffices. □

C PROOFS OF RESULTS UNDER ASSUMPTION 4.8

In Appendix C.1 we will first state a stronger statement than 4.9, namely Theorem C.1, and then show that Theorem 4.9 follows from that. The subsequent sections will be dedicated to proving Theorem C.1: in Appendix C.2 we will state and prove all lemmas necessary and then in Appendix C.3, we will prove Theorem C.1. Finally, in Appendix C.4 we will prove the lower bound Theorem 4.10.

C.1 Proof of Theorem 4.9

The convergence analysis in this section follows that of SVRC (Zhou et al., 2019). This supports our claim that Assumption 4.8 is a natural smoothness assumption. We refer the reader to Algorithm 1 for a description of SVRC.

Algorithm 1 SVRC (Zhou et al., 2019)

```
Input: Gradient and Hessian batch sizes $b_g$, $b_h$, cubic penalty parameter $M$, number of epochs $S$ and steps per epoch $T$. Starting point $\mathbf{x}_0$
$\widehat{\mathbf{x}}^1 = \mathbf{x}_0$
for $s = 1$ to $S$ do
  $\mathbf{x}_0^s = \widehat{\mathbf{x}}^s$
  $\mathbf{g}^s = \nabla F(\widehat{\mathbf{x}}^s)$, $\mathbf{H}^s = \nabla^2 F(\widehat{\mathbf{x}}^s)$
  for $t = 0$ to $T - 1$ do
    Sample index sets $I_g, I_h$, with $|I_h| = b_h, |I_g| = b_g$
    $\mathbf{v}_t^s = \frac{1}{b_g} \sum_{i_t \in I_g} [\nabla f_{i_t}(\mathbf{x}_t^s) - \nabla f_{i_t}(\widehat{\mathbf{x}}^s)] + \mathbf{g}^s - (\frac{1}{b_g} \sum_{i_t \in I_g} \nabla^2 f_{i_t}(\widehat{\mathbf{x}}^s) - \mathbf{H}^s)(\mathbf{x}_t^s - \widehat{\mathbf{x}}^s)$
    $\mathbf{U}_t^s = \frac{1}{b_h} \sum_{j_t \in I_h} [\nabla^2 f_{j_t}(\mathbf{x}_t^s) - \nabla^2 f_{j_t}(\widehat{\mathbf{x}}^s)] + \mathbf{H}^s$
    $\mathbf{h}_t^s = \arg\min_{\mathbf{h}} [\langle \mathbf{v}_t^s, \mathbf{h} \rangle + \frac{1}{2} \langle \mathbf{U}_t^s \mathbf{h}, \mathbf{h} \rangle + \frac{M}{6} \|\mathbf{h}\|^3]$
    $\mathbf{x}_{t+1}^s = \mathbf{x}_t^s + \mathbf{h}_t^s$
  end for
  $\widehat{\mathbf{x}}^{s+1} = \mathbf{x}_T^s$
end for
Input: $\mathbf{x}_{\text{out}} = \mathbf{x}_t^s$, where $s \in [S]$, $t \in [T]$ are chosen uniformly at random.
```

Recall the terminology in Algorithm 1. We will commonly call $\mathbf{v}_t^s$ and $\mathbf{U}_t^s$ the gradient and Hessian estimators respectively, we will refer to $\widehat{\mathbf{x}}^s$ as the snapshot point, and to $\mathbf{h}_t^s$ as the step. Finally, we will define

$$m_t^s(\mathbf{h}) = \langle \mathbf{v}_t^s, \mathbf{h} \rangle + \frac{1}{2} \langle \mathbf{U}_t^s \mathbf{h}, \mathbf{h} \rangle + \frac{M}{6} \|\mathbf{h}\|^3,$$

so that $\mathbf{h}_t^s = \arg\min_{\mathbf{h}} m_t^s(\mathbf{h})$. To aid in the analysis, we define the following quantity also introduced in Zhou et al. (2019):

$$\mu(\mathbf{x}) = \max \left\{ \|\nabla F(\mathbf{x})\|^{3/2}, -\frac{\lambda_{\min}^3(\nabla^2 F(\mathbf{x}))}{L_2^{3/2}} \right\}.$$

Whenever $\mu(\mathbf{x}) \leq \varepsilon^{3/2}$, $\mathbf{x}$ is an ε-approximate local minimum (Zhou et al., 2019). In Section C.3, we will show that we can bound the expected value of this quantity as follows (see also Theorem 6 in Zhou et al. (2019)):

Theorem C.1. *Let $M = C_M L_2$ for $C_M = 150$. Let $T \geq 2$ and choose $b_g \geq 5T^4$ and $b_h \geq 3000T^2 \log^3 d$. Then*

$$\mathbb{E}[\mu(\mathbf{x}_{\text{out}})] \leq \frac{240 C_M^2 L_2^{1/2} \Delta}{ST}.$$

Using this, we proceed with the proof of the main upper bound result.

Proof [of Theorem 4.9]**.** We first check that in the setting of Theorem 4.9, the assumptions of Theorem C.1 hold. It is clear that $T \geq 2$ and that $b_g = 5 \max\{n^{4/5}, 2^4\} = 5T^4$. Further, $b_h = 3000 \max\{4, n^{2/5}\} \log^3 d = 3000T^2 \log^3 d$. Plugging in the choices of S and T into the result of Theorem C.1, one gets

$$\mathbb{E}[\mu(\mathbf{x}_{\text{out}})] \leq \frac{240 C_M^2 L_2^{1/2} \Delta}{ST} \leq \frac{240 C_M^2 L_2^{1/2} \Delta}{\max\{1, 240 C_M^2 L_2^{1/2} \Delta n^{-1/5} \varepsilon^{-3/2}\} \max\{2, n^{1/5}\}} \leq \varepsilon^{3/2},$$

as desired. In particular, we have

$$\mathbb{E}[\|\nabla F(\mathbf{x}_{\text{out}})\|]^{3/2} \leq \mathbb{E}[\|\nabla F(\mathbf{x}_{\text{out}})\|^{3/2}] \leq \varepsilon^{3/2},$$

allowing comparison with our lower bound from Theorem 4.10.

During each epoch, n oracle calls are needed to construct $\mathbf{g}^s$ and $\mathbf{H}^s$, requiring Sn calls overall. To compute $\mathbf{v}_t^s$ and $\mathbf{U}_t^s$, we need

$$b_g + b_h = 5 \max\{n^{4/5}, 2^4\} + 3000 \max\{4, n^{2/5}\} \log^3$$

oracle queries at each iteration, requiring $ST(b_g + b_h)$ calls over all epochs and iterations. The total number of oracle queries is therefore at most

$$\begin{aligned}
&Sn + ST(b_g + b_h) \\
&= \max\{1, 240 C_M^2 L_2^{1/2} \Delta n^{-1/5} \varepsilon^{-3/2}\} n \\
&\quad + (\max\{1, 240 C_M^2 L_2^{1/2} \Delta n^{-1/5} \varepsilon^{-3/2}\})(\max\{2, n^{1/5}\})(5 \max\{n^{4/5}, 2^4\} + 3000 \max\{4, n^{2/5}\} \log^3 d) \\
&\leq \tilde{\mathcal{O}}\left(n + \frac{\Delta L_2^{1/2} n^{4/5}}{\varepsilon^{3/2}}\right).
\end{aligned}$$

□

C.2 Lemmas Required in the Proof of Theorem C.1

We will need some auxiliary lemmas to conduct the proof. The first is a version of Lemma 1 from Nesterov and Polyak (2006), but tailored to our finite-sum setting.

Lemma C.2. *Let $F = \frac{1}{n} \sum f_i$ satisfy Assumption 4.8. Then we have for any $\mathbf{x}$ and $\mathbf{y}$:*

$$\mathbb{E}_i \left[\|\nabla f_i(\mathbf{y}) - \nabla f_i(\mathbf{x}) - \nabla^2 f_i(\mathbf{x})(\mathbf{y} - \mathbf{x})\|^2\right] \leq \frac{1}{3} L_2^2 \|\mathbf{x} - \mathbf{y}\|^4.$$

and for any $\mathbf{h}$:

$$F(\mathbf{x} + \mathbf{h}) \leq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{h} \rangle + \frac{1}{2} \langle \nabla^2 F(\mathbf{x}) \mathbf{h}, \mathbf{h} \rangle + \frac{L_2}{6} \|\mathbf{h}\|^3.$$

Proof [of Lemma C.2]**.** We have

$$\begin{aligned}\mathbb{E}_i\|\nabla f_i(\mathbf{y})-\nabla f_i(\mathbf{x})-\nabla^2 f_i(\mathbf{x})(\mathbf{y}-\mathbf{x})\|^2 &= \mathbb{E}_i\|\int_0^1[\nabla^2 f_i(\mathbf{x}+\tau(\mathbf{y}-\mathbf{x}))-\nabla^2 f_i(\mathbf{y})](\mathbf{y}-\mathbf{x})d\tau\|^2\\ &\le \mathbb{E}_i\int_0^1\|\nabla^2 f_i(\mathbf{x}+\tau(\mathbf{y}-\mathbf{x}))-\nabla^2 f_i(\mathbf{y})\|^2\|\mathbf{x}-\mathbf{y}\|^2 d\tau\\ &= \int_0^1\mathbb{E}_i\|\nabla^2 f_i(\mathbf{x}+\tau(\mathbf{y}-\mathbf{x}))-\nabla^2 f_i(\mathbf{y})\|^2\|\mathbf{x}-\mathbf{y}\|^2 d\tau\\ &\le \int_0^1 L_2^2\|\mathbf{x}+\tau(\mathbf{y}-\mathbf{x})-\mathbf{y}\|^2\|\mathbf{x}-\mathbf{y}\|^2 d\tau\\ &= \frac{L_2^2}{3}\|\mathbf{x}-\mathbf{y}\|^4,\end{aligned}$$

where the first inequality is because of $\|\int_0^1\mathbf{v}d\tau\|^2\le\left(\int_0^1\|\mathbf{v}\|d\tau\right)^2\le\int_0^1\|\mathbf{v}\|^2 d\tau$ and the second inequality follows because of Assumption 4.8 and $\mathbb{E}[|X|^s]^{1/s}\le\mathbb{E}[|X|^t]^{1/t}$ for $s\le t$. □

We also take the following lemma directly from Zhou et al. (2019). Its proof exploits the optimality of $\mathbf{h}_t^s$.

Lemma C.3 [Lemma 24 in Zhou et al. (2019)]**.** *For the iterates in Algorithm 1 under the assumptions of Theorem 4.9 we have*

$$\begin{gathered}\mathbf{v}_t^s+\mathbf{U}_t^s\mathbf{h}_t^s+\frac{M}{2}\|\mathbf{h}_t^s\|\mathbf{h}_t^s=0,\\ \mathbf{U}_t^s+\frac{M}{2}\|\mathbf{h}_t^s\|\mathbf{I}\succeq 0,\\ \langle\mathbf{v}_t^s,\mathbf{h}_t^s\rangle+\frac{1}{2}\langle\mathbf{U}_t^s\mathbf{h}_t^s,\mathbf{h}_t^s\rangle+\frac{M}{6}\|\mathbf{h}_t^s\|^3\le-\frac{M}{12}\|\mathbf{h}_t^s\|^3.\end{gathered}$$

Lemmas C.4 and C.6 resemble Lemmas 25 and 26 in Zhou and Gu (2019) and bound the variances of the gradient and Hessian estimators of SVRC. Under the new smoothness Assumption 4.8, some constant factors change and the batch size for the Hessian estimator must comply to some stronger requirements, but other than that, third-moment smoothness is a viable alternative to an individual smoothness assumption. The proofs are analogous to the proofs of their respective counterparts.

The first lemma bounds the variance of $\mathbf{v}_t^s$:

Lemma C.4. *The gradient estimator $\mathbf{v}_t^s$ in Algorithm 1 satisfies*

$$\mathbb{E}_{i_t}\|\nabla F(\mathbf{x}_t^s)-\mathbf{v}_t^s\|^{3/2}\le\frac{2L_2^{3/2}}{b_g^{3/4}}\|\mathbf{x}_t^s-\widehat{\mathbf{x}}^s\|^3,$$

where $\mathbb{E}_{i_t}$ is the expectation over the batch indices $i_t\in I_g$.

To prove Lemma C.4 we will need the following technical result:

Lemma C.5 [Lemma 31 in Zhou et al. (2019)]**.** *Suppose $\mathbf{a}_1,\ldots,\mathbf{a}_N$ are i.i.d. and $\mathbb{E}\mathbf{a}_i=0$ for all i. Then*

$$\mathbb{E}\|\frac{1}{N}\sum_{i=1}^N\mathbf{a}_i\|^{3/2}\le\frac{1}{N^{3/4}}(\mathbb{E}\|\mathbf{a}_i\|^2)^{3/4}.$$

Proof [of Lemma C.4]**.** Using the definition of $\mathbf{v}_t^s$, we can write

$$
\begin{aligned}
&\mathbb{E}_{i_t}\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} \\
&= \mathbb{E}_{i_t}\|\frac{1}{b_g}\sum[\nabla f_{i_t}(\mathbf{x}_t^s) - \nabla f_{i_t}(\widehat{\mathbf{x}}^s)] + \mathbf{g}^s - \left[\frac{1}{b_g}\sum\nabla^2 f_{i_t}(\widehat{\mathbf{x}}^s) - \mathbf{H}^s\right](\mathbf{x}_t^s - \widehat{\mathbf{x}}^s) - \nabla F(\mathbf{x}_t^s)\|^{3/2} \\
&= \mathbb{E}_{i_t}\|\frac{1}{b_g}\sum[\nabla f_{i_t}(\mathbf{x}_t^s) - \nabla f_{i_t}(\widehat{\mathbf{x}}^s) - \nabla^2 f_{i_t}(\widehat{\mathbf{x}}^s)(\mathbf{x}_t^s - \widehat{\mathbf{x}}^s) - (\nabla F(\mathbf{x}_t^s) - \nabla F(\widehat{\mathbf{x}}^s) - \nabla^2 F(\widehat{\mathbf{x}}^s)(\mathbf{x}_t^s - \widehat{\mathbf{x}}^s))]\|^{3/2} \\
&\le \frac{1}{b_g^{3/4}}\left(\mathbb{E}_{i_t}\|\nabla f_{i_t}(\mathbf{x}_t^s) - \nabla f_{i_t}(\widehat{\mathbf{x}}^s) - \nabla^2 f_{i_t}(\widehat{\mathbf{x}}^s)(\mathbf{x}_t^s - \widehat{\mathbf{x}}^s) - (\nabla F(\mathbf{x}_t^s) - \nabla F(\widehat{\mathbf{x}}^s) - \nabla^2 F(\widehat{\mathbf{x}}^s)(\mathbf{x}_t^s - \widehat{\mathbf{x}}^s))\|^2\right)^{3/4} \\
&\le \frac{3^{3/4}}{b_g^{3/4}}\big(\mathbb{E}_{i_t}\|\nabla f_{i_t}(\mathbf{x}_t^s) - \nabla f_{i_t}(\widehat{\mathbf{x}}^s) - \nabla^2 f_{i_t}(\widehat{\mathbf{x}}^s)(\mathbf{x}_t^s - \widehat{\mathbf{x}}^s)\|^2 \\
&\qquad + \mathbb{E}_{i_t}\|(\nabla F(\mathbf{x}_t^s) - \nabla F(\widehat{\mathbf{x}}^s) - \nabla^2 F(\widehat{\mathbf{x}}^s)(\mathbf{x}_t^s - \widehat{\mathbf{x}}^s))\|^2\big)^{3/4} \\
&\le \frac{3^{3/4}}{b_g^{3/4}}\left(\frac{L_2^2}{3}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^4 + \frac{L_2^2}{3}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^4\right)^{3/4} \\
&= \frac{2L_2^{3/2}}{b_g^{3/4}}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^3.
\end{aligned}
$$

The first inequality is because of Lemma C.5. Indeed, as the different indices are independent, and the expectation is taken over the batch indices, we can apply Lemma C.5. The second holds due to the basic inequality $\|\mathbf{u}+\mathbf{v}\|^2 \le 3(\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2)$. The third inequality is because of Lemma C.2. □

This lemma bounds the variance of $\mathbf{U}_t^s$:

Lemma C.6. *If $b_h \ge 12000\log^3 d$, the Hessian estimator $\mathbf{U}_t^s$ satisfies*

$$
\mathbb{E}_{j_t}\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|^3 \le 15000 L_2^3\left(\frac{\log d}{b_h}\right)^{3/2}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^3,
$$

where $\mathbb{E}_{j_t}$ is the expectation over the batch indices $j_t \in I_h$.

In the proof of Lemma C.6, we will need the following matrix-moment inequality.

Lemma C.7 [Lemma 32 in Zhou et al. (2019)]**.** *Suppose that $q \ge 2, p \ge 2$, and fix $r \ge \max\{q, 2\log p\}$. Consider i.i.d. random self-adjoint matrices $\mathbf{Y}_1, \dots, \mathbf{Y}_N$ with dimension $p \times p$, $\mathbb{E}\mathbf{Y}_i = \mathbf{0}$. It holds that*

$$
\left[\mathbb{E}\|\sum_{i=1}^N \mathbf{Y}_i\|^q\right]^{1/q} \le 2\sqrt{er}\|\left(\sum_{i=1}^N \mathbb{E}\mathbf{Y}_i^2\right)^{1/2}\| + 4er(\mathbb{E}\max_i\|\mathbf{Y}_i\|^q)^{1/q}.
$$

Proof [of Lemma C.6]**.** We can rewrite

$$
\begin{aligned}
\mathbb{E}_{j_t}\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|^3 &= \mathbb{E}_{j_t}\|\nabla^2 F(\mathbf{x}_t^s) - \frac{1}{b_h}\left[\sum[\nabla^2 f_{j_t}(\mathbf{x}_t^s) - \nabla^2 f_{j_t}(\widehat{\mathbf{x}}^s) + \mathbf{H}^s]\right]\|^3 \\
&= \mathbb{E}_{j_t}\|\frac{1}{b_h}\left[\sum[\nabla^2 f_{j_t}(\mathbf{x}_t^s) - \nabla^2 f_{j_t}(\widehat{\mathbf{x}}^s) + \mathbf{H}^s - \nabla^2 F(\mathbf{x}_t^s)]\right]\|^3.
\end{aligned}
$$

Applying Lemma C.7, and using our third-moment assumption, we can bound this further. The Lemma controls the third moment of a sum with a sum of second moments and an additive term of the third moment of the maximum matrix. While Assumption 4.8 is not ideal for bounding maximum terms, we may replace the maximum with a sum over the whole batch, which is sufficient in this case. This only makes the batch size requirement grow polylogarithmically in the dimension of the domain. We proceed with the proof. Define $\mathbf{Y}_{j_t} = \nabla^2 f_{j_t}(\mathbf{x}_t^s) - \nabla^2 f_{j_t}(\widehat{\mathbf{x}}^s) + \mathbf{H}^s - \nabla^2 F(\mathbf{x}_t^s)$ and set $N = b_h$, $q = 3$, $p = d$ and $r = 2\log p$. Then

$$\left(\mathbb{E}_{j_t}\|\sum \mathbf{Y}_{j_t}\|^3\right)^{1/3} \leq 2\sqrt{er}\|\left(\sum \mathbb{E}_{j_t}\mathbf{Y}_{j_t}^2\right)^{1/2}\| + 4er(\mathbb{E}_{j_t}\max_{j_t}\|\mathbf{Y}_{j_t}\|^3)^{1/3}. \tag{12}$$

We bound both terms separately. For the first, we follow the original proof and get

$$\begin{aligned}
2\sqrt{er}\|\left(\sum \mathbb{E}_{j_t}\mathbf{Y}_{j_t}^2\right)^{1/2}\| &= 2\sqrt{er}\|\sum \mathbb{E}_{j_t}\mathbf{Y}_{j_t}^2\|^{1/2} \\
&= 2\sqrt{b_h er}\|\mathbb{E}_{j_t}\mathbf{Y}_{j_t}^2\|^{1/2} \\
&\leq 2\sqrt{b_h er}\left(\mathbb{E}_{j_t}\|\mathbf{Y}_{j_t}^2\|\right)^{1/2} \\
&\leq 2\sqrt{b_h er}\left(\mathbb{E}_{j_t}\|\mathbf{Y}_{j_t}\|^2\right)^{1/2}.
\end{aligned}$$

Plugging back the definition of $\mathbf{Y}_{j_t}$, and using Assumption 4.8 along with $\mathbb{E}[|X|^s]^{1/s} \leq \mathbb{E}[|X|^t]^{1/t}$ for $s \leq t$ allows us to bound

$$\begin{aligned}
2\sqrt{b_h er}\left(\mathbb{E}_{j_t}\|\mathbf{Y}_{j_t}\|^2\right)^{1/2} &= 2\sqrt{b_h er}\left(\mathbb{E}_{j_t}\|\nabla^2 f_{j_t}(\mathbf{x}_t^s) - \nabla^2 f_{j_t}(\widehat{\mathbf{x}}^s) + \mathbf{H}^s - \nabla^2 F(\mathbf{x}_t^s)\|^2\right)^{1/2} \\
&\leq 2\sqrt{b_h er}\left(3\,\mathbb{E}_{j_t}\|\nabla^2 f_{j_t}(\mathbf{x}_t^s) - \nabla^2 f_{j_t}(\widehat{\mathbf{x}}^s)\|^2 + 3\,\mathbb{E}_{j_t}\|\mathbf{H}^s - \nabla^2 F(\mathbf{x}_t^s)\|^2\right)^{1/2} \\
&\leq 2\sqrt{b_h er}\left(6\,L_2^2\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^2\right)^{1/2} \\
&\leq 5L_2\sqrt{b_h er}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|.
\end{aligned} \tag{13}$$

For the second term in Equation (12) we write

$$\begin{aligned}
4er\left(\mathbb{E}_{j_t}\max_{j_t}\|\mathbf{Y}_{j_t}\|^3\right)^{1/3} &\leq 4er\left(\mathbb{E}_{j_t}\sum\|\mathbf{Y}_{j_t}\|^3\right)^{1/3} \\
&\leq 4b_h^{1/3}er\left(\mathbb{E}_{j_t}\|\mathbf{Y}_{j_t}\|^3\right)^{1/3} \\
&= 4b_h^{1/3}er\left(\mathbb{E}_{j_t}\|\nabla^2 f_{j_t}(\mathbf{x}_t^s) - \nabla^2 f_{j_t}(\widehat{\mathbf{x}}^s) + \mathbf{H}^s - \nabla^2 F(\mathbf{x}_t^s)\|^3\right)^{1/3} \\
&\leq 4(7b_h)^{1/3}er\left(\mathbb{E}_{j_t}\|\nabla^2 f_{j_t}(\mathbf{x}_t^s) - \nabla^2 f_{j_t}(\widehat{\mathbf{x}}^s)\| + \|\mathbf{H}^s - \nabla^2 F(\mathbf{x}_t^s)\|^3\right)^{1/3} \\
&\leq 4(7b_h)^{1/3}er\left(2L_2^3\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^3\right)^{1/3} \\
&\leq 4(7b_h)^{1/3}er\left(2L_2^3\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^3\right)^{1/3} \\
&\leq 10L_2 b_h^{1/3}er\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|.
\end{aligned} \tag{14}$$

Plugging in Equations (13) and (14) into (12) we get

$$\left(\mathbb{E}_{j_t}\|\sum \mathbf{Y}_{j_t}\|^3\right)^{1/3} \leq 5L_2\sqrt{b_h er}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\| + 10L_2 b_h^{1/3}er\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|,$$

and therefore for the quantity we are interested in:

$$\begin{aligned}
\mathbb{E}_{j_t}\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|^3 &\leq 125L_2^3\left(\sqrt{\frac{er}{b_h}} + \frac{2er}{b_h^{2/3}}\right)^3\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^3 \\
&\leq 125L_2^3\left(\sqrt{\frac{2e\log d}{b_h}} + \frac{4e\log d}{b_h^{2/3}}\right)^3\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^3 \\
&\leq 15000L_2^3\left(\frac{\log d}{b_h}\right)^{3/2}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^3.
\end{aligned} \tag{15}$$

Because in (15) the first term in the parentheses dominates if $b_h \geq \sqrt{8e\log d}^6$, for which $b_h \geq 12000\log^3 d$ is sufficient. □

For completeness, we provide the rest of the lemmas from Zhou et al. (2019) that are needed in the analysis. We change the wording a bit, to make their applicability explicit, but all the proofs in the original paper can be applied *unchanged*, as is easily checked.

Lemma C.8 can be derived using the Cauchy-Schwarz and Young inequalities.

Lemma C.8 [Lemma 27 in Zhou et al. (2019)]**.** *For the iterates in Algorithm 1 under the assumptions of Theorem 4.9 and for any* $\mathbf{h}$*, we have*

$$\begin{aligned}\langle \nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s, \mathbf{h}\rangle &\leq \frac{M}{27}\|\mathbf{h}\|^3 + \frac{2\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2}}{M^{1/2}},\\ \langle \nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s, \mathbf{h}\rangle &\leq \frac{2M}{27}\|\mathbf{h}\|^3 + \frac{27}{M^2}\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|^3.\end{aligned}$$

Lemma C.9 [Lemma 28 in Zhou et al. (2019)]**.** *For the iterates in Algorithm 1 under the assumptions of Theorem 4.9 and for any* $\mathbf{h}$*, we have*

$$\begin{aligned}\mu(\mathbf{x}_t^s + \mathbf{h}) \leq 9C_M^{3/2}\Big[&M^{3/2}\|\mathbf{h}\|^3 + \|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} + M^{-3/2}\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|^3\\ &+ \|\nabla m_t^s(\mathbf{h})\|^{3/2} + M^{3/2}\big|\|\mathbf{h}\| - \|\mathbf{h}_t^s\|\big|^3\Big].\end{aligned}$$

Lemma C.10 [Lemma 29 in Zhou et al. (2019)]**.** *For any* $\mathbf{x}, \mathbf{y}, \mathbf{h}$ *and* $C \geq 3/2$ *we have*

$$\|\mathbf{x} + \mathbf{h} - \mathbf{y}\|^3 \leq 2C^2\|\mathbf{h}\|^3 + (1 + 3/C)\|\mathbf{x} - \mathbf{y}\|^3.$$

Lemma C.11 [Lemma 30 in Zhou et al. (2019)]**.** *Define* $c_T = 0$ *and for* $t \in [0 : T-1]$ *define* $c_t = c_{t+1}(1 + 3/T) + M(500T^3)^{-1}$*. Then for any* $t \in [1 : T]$ *we have:*

$$M/24 - 2c_tT^2 \geq 0.$$

C.3 Proof of Theorem C.1

Proof [of Theorem C.1]**.** This proof is very close to identical to the one of Theorem 6 in Zhou et al. (2019), but we give it again for completeness, with the changes coming from the slightly modified lemmas. We can bound the function value at the next iterate $F(\mathbf{x}_{t+1})$ as follows:

$$\begin{aligned}F(\mathbf{x}_{t+1}^s) &\leq F(\mathbf{x}_t^s) + \langle \nabla F(\mathbf{x}_t^s), \mathbf{h}_t^s\rangle + \frac{1}{2}\langle \nabla^2 F(\mathbf{x}_t^s)\mathbf{h}_t^s, \mathbf{h}_t^s\rangle + \frac{L_2}{6}\|\mathbf{h}_t^s\|^3 \qquad (16)\\ &= F(\mathbf{x}_t^s) + \langle \mathbf{v}_t^s, \mathbf{h}_t^s\rangle + \frac{1}{2}\langle \mathbf{U}_t^s\mathbf{h}_t^s, \mathbf{h}_t^s\rangle + \frac{M}{6}\|\mathbf{h}_t^s\|^3 + \langle \nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s, \mathbf{h}_t^s\rangle\\ &\quad + \frac{1}{2}\langle\left(\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\right)\mathbf{h}_t^s, \mathbf{h}_t^s\rangle + \frac{M - L_2}{6}\|\mathbf{h}_t^s\|^3\\ &\leq F(\mathbf{x}_t^s) - \frac{M}{2}\|\mathbf{h}_t^s\|^3 + \left(\frac{M}{27}\|\mathbf{h}_t^s\|^3 + \frac{2\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2}}{M^{1/2}}\right)\\ &\quad + \frac{1}{2}\left(\frac{2M}{27}\|\mathbf{h}_t^s\|^3 + \frac{27}{M^2}\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|^3\right) - \frac{M - L_2}{6}\|\mathbf{h}_t^s\|^3 \qquad (17)\\ &\leq F(\mathbf{x}_t^s) - \frac{M}{12}\|\mathbf{h}_t^s\|^3 + \frac{2}{M^{1/2}}\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} + \frac{27}{M^2}\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|^3. \qquad (18)\end{aligned}$$

(16) holds due to Lemma C.2 and (17) is valid because of Lemmas C.3 and C.8.

Define

$$R_t^s = \mathbb{E}\left[F(\mathbf{x}_t^s) + c_t\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^3\right],$$

where $c_T = 0$ and $c_t = c_{t+1}(1+3/T) + M(500T^3)^{-1}$ for $t \in [0 : T-1]$. We use Lemma C.10 with $T \geq 2 \geq 3/2$ to get a recurrence – involving the step – for the cubed distance from an iterate to the snapshot point:

$$c_{t+1}\|\mathbf{x}_{t+1}^s - \widehat{\mathbf{x}}^s\|^3 \leq 2c_{t+1}T^2\|\mathbf{h}_t^s\|^3 + c_{t+1}(1+3/T)\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^3. \tag{19}$$

We can make use of Lemma C.9 with $\mathbf{h} = \mathbf{h}_t^s$ followed by Lemma C.3

$$\begin{aligned}(240C_M^2L_2^{1/2})^{-1}\mu(\mathbf{x}_{t+1}^s) &\leq \frac{M}{24}\|\mathbf{h}_t^s\|^3 + \frac{\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2}}{24M^{1/2}} + \frac{\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|^3}{24M^2} \\ &\quad + \frac{\|\nabla m_t^s(\mathbf{h}_t^s)\|^{3/2}}{24M^{1/2}} + \frac{M}{24}\big|\|\mathbf{h}_t^s\| - \|\mathbf{h}_t^s\|\big|^3 \\ &= \frac{M}{24}\|\mathbf{h}_t^s\|^3 + \frac{\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2}}{24M^{1/2}} + \frac{\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|^3}{24M^2},.\end{aligned} \tag{20}$$

In the first step we used $C_M = 150$ and $M = C_M L_2$ and in the second we used the optimality of $\mathbf{h}_t^s$ as an argument of m_t^s. Our aim is to get a telescoping sum for the R_t's. For that, we start by combining (18), (19) and (20) (this time the expectation is over all the randomness involved in the algorithm):

$$\begin{aligned}R_{t+1}^s + (240C_M^2L_2^{1/2})^{-1}\mathbb{E}[\mu(\mathbf{x}_{t+1})] &= \mathbb{E}\left[F(\mathbf{x}_{t+1}^s) + c_{t+1}\|\mathbf{x}_{t+1}^s - \widehat{\mathbf{x}}^s\|^3 + (240C_M^2L_2^{1/2})^{-1}\mu(\mathbf{x}_{t+1}^s)\right] \\ &\leq \mathbb{E}\left[F(\mathbf{x}_t^s) + c_{t+1}(1+3/T)\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^3 - (M/24 - 2c_{t+1}T^2)\|\mathbf{h}_t^s\|^3\right] \\ &\quad + \mathbb{E}\left[3M^{-1/2}\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} + 28M^{-2}\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|^3\right] \\ &\leq \mathbb{E}\left[F(\mathbf{x}_t^s) + c_{t+1}(1+3/T)\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^3\right] \\ &\quad + \mathbb{E}\left[3M^{-1/2}\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} + 28M^{-2}\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|^3\right],\end{aligned} \tag{21}$$

because by Lemma C.11 we have $M/24 - 2c_{t+1}T^2 \geq 0$ for any $t \in [T]$. In the second term of (21), we recover the gradient and Hessian estimator variances that Lemmas C.4 and C.6 control. Indeed, taking iterated expectations yields

$$3M^{-1/2}\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} \leq \frac{6L_2^{3/2}}{M^{1/2}b_g^{3/4}}\mathbb{E}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^3 \leq \frac{M}{1000T^3}\mathbb{E}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^3.$$

Here we have used that $M = 150L_2$ and $b_g \geq 5T^4$. For the Hessian estimator, we get

$$\begin{aligned}28M^{-2}\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|^3 &\leq \frac{28 \cdot 15000L_2^3}{M^2(b_h/\log d)^{3/2}}\mathbb{E}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^3 \\ &\leq \frac{28 \cdot 15000M}{150^3(3000)^{3/2}T^3}\mathbb{E}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^3 \\ &\leq \frac{M}{1000T^3}\mathbb{E}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^3,\end{aligned}$$

where we additionally use $b_h \geq 3000T^2\log^3 d$. Note that our larger b_h actually gives us better constant factors than we derive, but we do not need this and therefore keep the same as in the original proof. From here, we exactly follow said original proof from Zhou et al. (2019). We can plug those 2 bounds back into (21) and use the definition of c_t to get the recurrence

$$\begin{aligned}R_{t+1}^s + (240C_M^2L_2^{1/2})^{-1}\mathbb{E}[\mu(\mathbf{x}_{t+1})] &\leq \mathbb{E}\left[F(\mathbf{x}_t^s) + \|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^3\left(c_{t+1}(1+3/T) + \frac{M}{500T^3}\right)\right] \\ &= \mathbb{E}[F(\mathbf{x}_t^s) + c_t\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^3] = R_t^s.\end{aligned}$$

We will now do 2 steps of telescoping. First, let $s \in [S]$ be arbitrary. As $c_T = 0$ and $x_T^s = \widehat{\mathbf{x}}^{s+1}$ by definition, we have $R_T^s = \mathbb{E}[F(\mathbf{x}_T^s) + c_T\|\mathbf{x}_T^s - \widehat{\mathbf{x}}^s\|^3] = \mathbb{E}F(\mathbf{x}_T^s) = \mathbb{E}F(\widehat{\mathbf{x}}^{s+1})$. As $\mathbf{x}_0^s = \widehat{\mathbf{x}}^s$, we have $R_0^s = \mathbb{E}[F(\mathbf{x}_0^s) + c_0\|\mathbf{x}_0^s - \widehat{\mathbf{x}}^s\|^3] = \mathbb{E}F(\widehat{\mathbf{x}}^s)$. Thus, rearranging and telescoping the above from $t = 0$ to $T-1$ yields

$$\mathbb{E}F(\widehat{\mathbf{x}}^s) - \mathbb{E}F(\widehat{\mathbf{x}}^{s+1}) = R_0^s - R_T^s \geq \sum_{t=1}^{T}(240C_M^2 L_2^{1/2})^{-1}\mathbb{E}[\mu(\mathbf{x}_t^s)].$$

Further, we can telescope this from $s = 1$ to S and obtain

$$\mathit{\Delta} \geq F(\widehat{\mathbf{x}}^1) - F(\widehat{\mathbf{x}}^S) = \sum_{s=1}^{S}\left[\mathbb{E}F(\widehat{\mathbf{x}}^s) - \mathbb{E}F(\widehat{\mathbf{x}}^{s+1})\right] \geq (240C_M^2 L_2^{1/2})^{-1}\sum_{s=1}^{S}\sum_{t=1}^{T}\mathbb{E}[\mu(\mathbf{x}_t^s)].$$

The first inequality holds because of the definition of $\mathbf{x}_0 = \widehat{\mathbf{x}}^1$ and because the choice of $\mathbf{h}_t^s$ guarantees the iterates do not yield increases in function value over time. Therefore, picking a random iterate $\mathbf{x}_t^s$, we will have

$$\mathbb{E}[\mu(\mathbf{x}_t^s)] \leq \frac{240C_M^2 L_2^{1/2}\mathit{\Delta}}{ST},$$

as desired. □

C.4 Proof of Theorem 4.10

Proof [of Theorem 4.10]**.** Let $\sigma, \lambda > 0$ be parameters yet to be chosen. The same is true for d and K. According to Definition 4.2, we define the scaled functions

$$f_i(\mathbf{x}) = \sqrt[3]{n}\lambda\sigma^3 f_i^*\left(\frac{\mathbf{x}}{\sigma}\right) = \sqrt[3]{n}\lambda\sigma^3 \hat{f}_{K;\mathbf{B}_i}\left(\frac{\mathbf{C}_i^T\mathbf{x}}{\sigma}\right),$$

giving us

$$F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}).$$

We will choose the scaling parameters to ensure that our instance satisfies Assumption 4.8, deriving the lower bound as we go along. We first guarantee smoothness: for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we have

$$\begin{aligned}
\mathbb{E}_i\|\nabla^2 f_i(\mathbf{x}) - \nabla^2 f_i(\mathbf{y})\|^3 &= \frac{1}{n}\sum_{i=1}^{n}\|\nabla^2 f_i(\mathbf{x}) - \nabla^2 f_i(\mathbf{y})\|^3 \\
&\leq \frac{1}{n}\sum_{i=1}^{n}(\sqrt[3]{n}\lambda\hat{\ell}_2)^3\|\mathbf{C}_i^T\mathbf{x} - \mathbf{C}_i^T\mathbf{y}\|^3 \qquad (22) \\
&= \lambda^3\hat{\ell}_2^3\sum_{i=1}^{n}\|\mathbf{C}_i^T(\mathbf{x} - \mathbf{y})\|^2\|\mathbf{C}^T(\mathbf{x} - \mathbf{y})\| \\
&= \lambda^3\hat{\ell}_2^3\|\mathbf{C}^T(\mathbf{x} - \mathbf{y})\|^3 \\
&= \lambda^3\hat{\ell}_2^3\|\mathbf{x} - \mathbf{y}\|^3,
\end{aligned}$$

where (22) follows from Lemmas D.1 and B.2. So, the choice $\lambda = \frac{L_2}{\hat{\ell}_2}$ therefore accomplishes third-moment smoothness with parameter L_2.

Now fix an algorithm A and assume $\{[i^t, \mathbf{x}^{(t)}]\}_{t\in\mathbb{N}}$ are the iterates produced by A on F. Consequently, by Lemma 4.4 $\{[i^t, \mathbf{x}^{(t)}/\sigma]\}_{t\in\mathbb{N}}$ is informed by F^*. Therefore we can apply Lemma 4.6 on the sequence

$\{[i^t, \mathbf{x}^{(t)}/\sigma]\}_{t\in\mathbb{N}}$ to get

$$\begin{aligned}\|\nabla F(\mathbf{x}^{(t)})\|^2 &= \|\sqrt[3]{n}\lambda\sigma^2\nabla F^*(\mathbf{x}^{(t)}/\sigma)\|^2\\ &= n^{2/3}\lambda^2\sigma^4\|\nabla F^*(\mathbf{x}^{(t)}/\sigma)\|^2\\ &\geq n^{2/3}\lambda^2\sigma^4\frac{1}{16n}\\ &= \frac{\sigma^4\lambda^2}{16n^{1/3}}.\end{aligned}$$

To get a lower bound for an ε precision requirement we can choose σ to be

$$\frac{\sigma^2\lambda}{4n^{1/6}} = \varepsilon \iff \sigma = \left(\frac{4\varepsilon\hat{\ell}_2 n^{1/6}}{L_2}\right)^{1/2}.$$

Next, we will guarantee the optimality gap requirement. We have

$$\begin{aligned}F(\mathbf{0}) - \inf_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) &\leq \sqrt[3]{n}\lambda\sigma^3\left[\frac{1}{n}\sum_{i=1}^n \hat{f}_{K;\mathbf{B}_i}\left(\frac{\mathbf{C}_i^T\mathbf{0}}{\sigma}\right) - \frac{1}{n}\sum_{i=1}^n \inf_{\mathbf{x}\in\mathbb{R}^d}\hat{f}_{K;\mathbf{B}_i}\left(\frac{\mathbf{C}_i^T\mathbf{x}}{\sigma}\right)\right]\\ &\leq \sqrt[3]{n}\lambda\sigma^3\frac{1}{n}\sum_{i=1}^n\left[\hat{f}_{K;\mathbf{B}_i}(\mathbf{0}) - \inf_{\mathbf{y}\in\mathbb{R}^{d/n}}\hat{f}_{K;\mathbf{B}_i}(\mathbf{y})\right]\\ &\leq 12\sqrt[3]{n}\lambda\sigma^3 K,\end{aligned}$$

where the last step uses Lemma B.2 i). We require

$$12\sqrt[3]{n}\lambda\sigma^3K = 12\sqrt[3]{n}\frac{L_2}{\hat{\ell}_2}\left(\frac{4\varepsilon\hat{\ell}_2n^{1/6}}{L_2}\right)^{3/2}K = 96n^{7/12}\left(\frac{\hat{\ell}_2}{L_2}\right)^{1/2}\varepsilon^{3/2}K \leq \mathit{\Delta}.$$

Our bounds get better with larger values of K, so we want to choose K as

$$K = \left\lfloor \frac{\mathit{\Delta}}{96n^{7/12}}\left(\frac{L_2}{\hat{\ell}_2}\right)^{1/2}\frac{1}{\varepsilon^{3/2}}\right\rfloor.$$

We need $K \geq 1$ to have a sensible bound as becomes apparent below, and so we require

$$\tilde{c}\mathit{\Delta}L_2^{1/2}\frac{1}{\varepsilon^{3/2}} \geq n^{7/12},$$

or more concisely

$$n \leq \frac{c\mathit{\Delta}^{12/7}L_2^{6/7}}{\varepsilon^{18/7}},$$

for some universal constants $c, \tilde{c}$. As Lemma 4.6 yields the lower bound $T = \frac{nK}{2}$, we get a lower bound of

$$\mathit{\Omega}\left(\left(\frac{L_2}{\hat{\ell}_2}\right)^{1/2}\frac{\mathit{\Delta}n^{5/12}}{\varepsilon^{3/2}}\right)$$

with probability at least 1/2 for large enough dimension d (see below). Thus there must be a fixed function F such that for this many iterations – with probability 1/2 depending only on ξ – the iterates A produces on F all have gradient larger than ε. This means that

$$T_\varepsilon(\mathsf{A}, F) \geq \mathit{\Omega}\left(\frac{\sqrt{L_2}\mathit{\Delta}n^{5/12}}{\varepsilon^{3/2}}\right).$$

For the requirement on the dimension d for the bound from Lemma 4.6 to hold, we can plug in our values of K and $\delta = 1/2$ to see that some $d \in \tilde{\mathcal{O}}(n^2\mathit{\Delta}L_2\varepsilon^{-3})$ suffices. This concludes the proof. □

D SHARED TECHNICAL LEMMA

We need the following result to guarantee the smoothness of our constructions.

Lemma D.1. *Assume $m_1 \geq m_2$. Let $f : \mathbb{R}^{m_2} \to \mathbb{R}$ and for $\mathbf{C} \in \mathsf{Ortho}(m_1, m_2)$ let $g : \mathbb{R}^{m_1} \to \mathbb{R}^{m_2}$, $\mathbf{x} \mapsto \mathbf{C}^T\mathbf{x}$. We will show that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{m_1}$:*

$$\|\nabla^p[f(\mathbf{C}^T\mathbf{x})] - \nabla^p[f(\mathbf{C}^T\mathbf{y})]\| \leq \|\tilde{\nabla}^p f(\mathbf{C}^T\mathbf{x}) - \tilde{\nabla}^p f(\mathbf{C}^T\mathbf{y})\|,$$

where the gradient operator ∇ is with respect to $\mathbf{x}$ while $\tilde{\nabla}$ is with respect to $g(\mathbf{x}) = \mathbf{C}^T\mathbf{x}$. Further, if f is pth-order smooth with constant L, then for any $\sigma > 0$

$$\|\nabla^p[\sigma^{p+1} f(\mathbf{C}^T\mathbf{x}/\sigma)] - \nabla^p[\sigma^{p+1} f(\mathbf{C}^T\mathbf{y}/\sigma)]\| \leq L\|\mathbf{C}^T(\mathbf{x} - \mathbf{y})\|.$$

Proof [of Lemma D.1]**.** We are interested in the tensor $\nabla^p[f(\mathbf{C}^T\mathbf{x})]$. Fix indices $i_1, \ldots, i_p$ and let Ξ be the set of partitions of $[p]$. For a set $S \subset [p]$ let $i_S = \{i_j \mid j \in S\}$. Define $\nabla^{|S|}_{i_S}$ to be the order $|S|$ partial derivative operator with respect to the coordinates with indices in i_S. Applying the higher-order chain rule we obtain

$$\nabla^p_{i_1,\ldots,i_p}[f(\mathbf{C}^T\mathbf{x})] = \sum_{(S_1,\ldots,S_L)\in\Xi} \sum_{j_1,\ldots,j_L=1}^{m_2} \left(\prod_{l=1}^{L} \nabla^{|S_l|}_{i_{S_l}} g_{j_l}(\mathbf{x})\right) \tilde{\nabla}^L_{j_1,\ldots,j_L} f(\mathbf{C}^T\mathbf{x}).$$

Now we use that g_{j_l}'s second and higher-order derivatives are zero, and that $\nabla_i g_{j_l}(\mathbf{x}) = \nabla_i[\langle \mathbf{c}_{j_l}, \mathbf{x}\rangle] = c_{i,j_l}$. This means that in the above sum, the only partition that matters has $L = p$ and $|S_1|, \ldots, |S_p| = 1$. W.l.o.g. we may take $S_l = \{l\}$ and consequently $i_{S_l} = \{i_l\}$. Then our expression simplifies to

$$\begin{aligned}
\nabla^p_{i_1,\ldots,i_p}[f(\mathbf{C}^T\mathbf{x})] &= \sum_{(S_1,\ldots,S_L)\in\Xi} \sum_{j_1,\ldots,j_L=1}^{m_2} \left(\prod_{l=1}^{L} \nabla^{|S_l|}_{i_{S_l}} g_{j_l}(\mathbf{x})\right) \tilde{\nabla}^L_{j_1,\ldots,j_L} f(\mathbf{C}^T\mathbf{x}) \\
&= \sum_{j_1,\ldots,j_p=1}^{m_2} \left(\prod_{l=1}^{p} \nabla_{i_l} g_{j_l}(\mathbf{x})\right) \tilde{\nabla}^p_{j_1,\ldots,j_p} f(\mathbf{C}^T\mathbf{x}) \\
&= \sum_{j_1,\ldots,j_p=1}^{m_2} \left(\prod_{l=1}^{p} c_{i_l,j_l}\right) \tilde{\nabla}^p_{j_1,\ldots,j_p} f(\mathbf{C}^T\mathbf{x}).
\end{aligned}$$

We now bound the tensor operator norm from the Lemma statement: let $\mathbf{v}^{(1)}, ..., \mathbf{v}^{(p)} \in \mathbb{R}^{m_1}$ be arbitrary unit vectors. Then we have

$$\begin{aligned}
&\left\langle \nabla^p[f(\mathbf{C}^T\mathbf{x})] - \nabla^p[f(\mathbf{C}^T\mathbf{y})],\, \mathbf{v}^{(1)} \otimes \cdots \otimes \mathbf{v}^{(p)} \right\rangle \\
&= \sum_{i_1,\ldots,i_p=1}^{m_1} v_{i_1}^{(1)} \cdots v_{i_p}^{(p)} \sum_{j_1,\ldots,j_p=1}^{m_2} \left(\prod_{l=1}^{p} c_{i_l,j_l} \right) \tilde{\nabla}^p_{j_1,\ldots,j_p} (f(\mathbf{C}^T\mathbf{x}) - f(\mathbf{C}^T\mathbf{y})) \\
&= \sum_{j_1,\ldots,j_p=1}^{m_2} \sum_{i_1,\ldots,i_p=1}^{m_1} v_{i_1}^{(1)} \cdots v_{i_p}^{(p)} \left(\prod_{l=1}^{p} c_{i_l,j_l} \right) \tilde{\nabla}^p_{j_1,\ldots,j_p} (f(\mathbf{C}^T\mathbf{x}) - f(\mathbf{C}^T\mathbf{y})) \\
&= \sum_{j_1,\ldots,j_p=1}^{m_2} \sum_{i_1,\ldots,i_p=1}^{m_1} \left(\prod_{l=1}^{p} v_{i_l}^{(l)} c_{i_l,j_l} \right) \tilde{\nabla}^p_{j_1,\ldots,j_p} (f(\mathbf{C}^T\mathbf{x}) - f(\mathbf{C}^T\mathbf{y})) \\
&= \sum_{j_1,\ldots,j_p=1}^{m_2} \left(\prod_{l=1}^{p} \left(\sum_{i_l=1}^{m_1} v_{i_l}^{(l)} c_{i_l,j_l} \right) \right) \tilde{\nabla}^p_{j_1,\ldots,j_p} (f(\mathbf{C}^T\mathbf{x}) - f(\mathbf{C}^T\mathbf{y})) \\
&= \sum_{j_1,\ldots,j_p=1}^{m_2} \left(\left(\langle \mathbf{v}^{(1)}, \mathbf{c}_{j_1} \rangle \right) \cdots \left(\langle \mathbf{v}^{(p)}, \mathbf{c}_{j_p} \rangle \right) \right) \tilde{\nabla}^p_{j_1,\ldots,j_p} (f(\mathbf{C}^T\mathbf{x}) - f(\mathbf{C}^T\mathbf{y})) \\
&= \sum_{j_1,\ldots,j_p=1}^{m_2} \left(\left(\mathbf{C}^T\mathbf{v}^{(1)} \right)_{j_1} \cdots \left(\mathbf{C}^T\mathbf{v}^{(p)} \right)_{j_p} \right) \tilde{\nabla}^p_{j_1,\ldots,j_p} (f(\mathbf{C}^T\mathbf{x}) - f(\mathbf{C}^T\mathbf{y})) \\
&= \left\langle \tilde{\nabla}^p f(\mathbf{C}^T\mathbf{x}) - \tilde{\nabla}^p f(\mathbf{C}^T\mathbf{y}),\, \mathbf{C}^T\mathbf{v}^{(1)} \otimes \cdots \otimes \mathbf{C}^T\mathbf{v}^{(p)} \right\rangle \\
&\leq \|\tilde{\nabla}^p f(\mathbf{C}^T\mathbf{x}) - \tilde{\nabla}^p f(\mathbf{C}^T\mathbf{y})\|.
\end{aligned}$$

The first statement follows because $\mathbf{C}$ has orthonormal columns and can be extended to an $\mathbb{R}^{m_1 \times m_1}$ matrix $\tilde{\mathbf{C}}$. Then $\|\mathbf{C}^T\mathbf{v}^{(k)}\| \leq \|\tilde{\mathbf{C}}^T\mathbf{v}^{(k)}\| = \|\mathbf{v}^{(k)}\| = 1$ for all $k \in [p]$, which justifies the application of the operator norm definition. Because $\mathbf{v}^{(1)}, ..., \mathbf{v}^{(p)}$ were arbitrary, we obtain the desired inequality.

The second statement follows from p applications of the chain rule. □