Proceedings of Machine Learning Research vol 99:1–27, 2019 32nd Annual Conference on Learning Theory

The Complexity of Making the Gradient Small in Stochastic Convex Optimization

Dylan J. Foster DYLANF@MIT.EDU
Massachusetts Institute of Technology

Ayush Sekhari SEKHARI@CS.CORNELL.EDU
Cornell University

Ohad Shamir OHAD.SHAMIR@WEIZMANN.AC.IL
Weizmann Institute of Science

Nathan Srebro NATI@TTIC.EDU
Toyota Technological Institute at Chicago

Karthik Sridharan SRIDHARAN@CS.CORNELL.EDU
Cornell University

Blake Woodworth BLAKE@TTIC.EDU
Toyota Technological Institute at Chicago

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We give nearly matching upper and lower bounds on the oracle complexity of finding ϵ-stationary points ($\|\nabla F(x)\| \le \epsilon$) in stochastic convex optimization. We jointly analyze the oracle complexity in both the local stochastic oracle model and the global oracle (or, statistical learning) model. This allows us to decompose the complexity of finding near-stationary points into *optimization complexity* and *sample complexity*, and reveals some surprising differences between the complexity of stochastic optimization versus learning. Notably, we show that in the global oracle/statistical learning model, only *logarithmic dependence on smoothness* is required to find a near-stationary point, whereas polynomial dependence on smoothness is necessary in the local stochastic oracle model. In other words, the separation in complexity between the two models can be exponential, and the folklore understanding that smoothness is required to find stationary points is only weakly true for statistical learning.

Our upper bounds are based on extensions of a recent "recursive regularization" technique proposed by Allen-Zhu (2018). We show how to extend the technique to achieve near-optimal rates, and in particular show how to leverage the extra information available in the global oracle model. Our algorithm for the global model can be implemented efficiently through finite sum methods, and suggests an interesting new computational-statistical tradeoff.

Keywords: stationary point, sample complexity, oracle complexity, stochastic optimization, non-convex optimization.

1. Introduction

Success in convex optimization is typically defined as finding a point whose value is close to the minimum possible value. *Information-based complexity* of optimization attempts to understand the minimal amount of effort required to reach a desired level of suboptimality under different

oracle models for access to the function (Nemirovski and Yudin, 1983; Traub et al., 1988). This complexity—for both deterministic and stochastic convex optimization—is tightly understood across a wide variety of settings (Nemirovski and Yudin, 1983; Traub et al., 1988; Agarwal et al., 2009; Braun et al., 2017), and efficient algorithms that achieve optimal complexity are well known.

Recently, there has been a surge of interest in optimization for *non-convex* functions. In this case, finding a point with near-optimal function value is typically intractable under standard assumptions, both computationally and information-theoretically. For this reason, a standard task in non-convex optimization is to find an ϵ-stationary point, i.e., a point where the gradient is small ($\|\nabla F(x)\| \leq \epsilon$). In stochastic non-convex optimization, a number of recent results provide provable guarantees for finding near-stationary points (Ghadimi and Lan, 2013, 2016; Reddi et al., 2016; Allen-Zhu, 2017; Lei et al., 2017; Jin et al., 2017; Zhou et al., 2018; Fang et al., 2018). However, the stochastic oracle complexity of finding near-stationary points is not yet well understood, so we do not know whether existing algorithms are optimal, or how we hope to improve upon them.

Carmon et al. (2017a,b) have established tight bounds on the *deterministic* first-order oracle complexity of finding near-stationary points of smooth functions, both convex and non-convex. For convex problems, they prove that accelerated gradient descent is optimal both for finding approximate minimizers and approximate stationary points, while for non-convex problems, gradient descent is optimal for finding approximate stationary points. The picture is simple and complete: the same deterministic first-order methods that are good at finding approximate minimizers are also good at finding approximate stationary points, even for non-convex functions.

However, when one turns their attention to the *stochastic* oracle complexity of finding near-stationary points, the picture is far from clear. Even for *stochastic convex optimization*, the oracle complexity is not yet well understood. This paper takes a first step toward resolving the general case by providing nearly tight upper and lower bounds on the oracle complexity of finding near-stationary points in stochastic convex optimization, both for first-order methods and for global (i.e., statistical learning) methods. At first glance, this might seem trivial, since exact minimizers are equivalent to exact stationary points for convex functions. When it comes to finding *approximate* stationary points the situation is considerably more complex, and the equivalence does not yield quantitatively optimal rates. For example, while the stochastic gradient descent (SGD) is (worst-case) optimal for stochastic convex optimization with a first-order oracle, it appears to be far from optimal for finding near-stationary points.

1.1. Contributions

We present a nearly tight analysis of the local stochastic oracle complexity and global stochastic oracle complexity ("sample complexity") of finding approximate stationary points in stochastic convex optimization. Briefly, the highlights are as follows:

- We give upper and lower bounds on the local and global stochastic oracle complexity that match up to log factors. In particular, we show that the local stochastic complexity of finding stationary points is (up to log factors) characterized as the sum of the deterministic oracle complexity and the sample complexity.
- As a consequence of this two-pronged approach, we show that the gap between local stochastic complexity and sample complexity of finding near-stationary points is at least *exponential* in the smoothness parameter.

		Deterministic First-Order Oracle	Sample Complexity	Stochastic First-Order Oracle
$\|x_0 - x^*\| \le R$	Upper:	$\tilde{O}\left(\sqrt{\frac{HR}{\epsilon}}\right)$ Nesterov (2012)	$O\left(\frac{\sigma^2}{\epsilon^2}\log^3\left(\frac{HR}{\epsilon}\right)\right)$ (Corollary 2)	$\tilde{O}\left(\sqrt{\frac{HR}{\epsilon}} + \frac{\sigma^2}{\epsilon^2}\right)$ (Corollary 1)
	Lower:	$\Omega\left(\sqrt{\frac{HR}{\epsilon}}\right)$ Carmon et al. (2017b)	$\Omega\left(\frac{\sigma^2}{\epsilon^2}\right)$ (Theorem 4)	$\Omega\left(\sqrt{\frac{HR}{\epsilon}} + \frac{\sigma^2}{\epsilon^2}\log\left(\frac{HR}{\epsilon}\right)\right)$ (Theorem 2)
$F(x_0) - F(x^*) \le \Delta$	Upper:	$\tilde{O}\left(\frac{\sqrt{H\Delta}}{\epsilon}\right)$ Carmon et al. (2017b)	$O\left(\frac{\sigma^2}{\epsilon^2}\log^3\left(\frac{H\Delta}{\epsilon^2}\right)\right)$ (Corollary 2)	$\tilde{O}\left(\frac{\sqrt{H\Delta}}{\epsilon} + \frac{\sigma^2}{\epsilon^2}\right)$ (Corollary 1)
	Lower:	$\Omega\left(\frac{\sqrt{H\Delta}}{\epsilon}\right)$ Carmon et al. (2017b)	$\Omega\left(\frac{\sigma^2}{\epsilon^2}\right)$ (Theorem 4)	$\Omega\left(\frac{\sqrt{H\Delta}}{\epsilon} + \frac{\sigma^2}{\epsilon^2}\log\left(\frac{H\Delta}{\epsilon^2}\right)\right)$ (Theorem 2)

Table 1: Upper and lower bounds on the complexity of finding x such that $\|\nabla F(x)\| \le \epsilon$ for convex problems with H-Lipschitz gradients, where σ^2 is a bound on the variance of gradient estimates.

- We obtain the above results through new algorithmic improvements. We show that the recursive regularization technique introduced by Allen-Zhu (2018) for local stochastic optimization can be combined with empirical risk minimization to obtain *logarithmic* dependence on smoothness in the global model, and that the resulting algorithms can be implemented efficiently.

Complexity results are summarized in Table 1. Here we discuss the conceptual contributions in more detail.

Decomposition of stochastic first-order complexity. For stochastic optimization of convex functions, there is a simple and powerful connection between three oracle complexities: first-order deterministic, first-order stochastic, and global stochastic. For many well-known problem classes, the stochastic first-order complexity is equal to the sum (equivalently, maximum) of the deterministic first-order complexity and the sample complexity. This decomposition of the local stochastic complexity into an "optimization term" plus a "statistical term" inspires optimization methods, guides analysis, and facilitates comparison of different algorithms. It indicates that "one pass" stochastic approximation algorithms like SGD are optimal for stochastic optimization in certain parameter regimes, so that we do not have to resort to sample average approximation or methods that require multiple passes over data.

We establish that the same decomposition holds for the task of finding approximate stationary points. Such a characterization should not be taken for granted, and it is not clear a priori that it should hold for finding stationary points. Establishing the result requires both developing new algorithms with near-optimal sample complexity in the global model, and improving previous local stochastic methods (Allen-Zhu, 2018) to match the optimal deterministic complexity.

Gap between sample complexity and stochastic first-order complexity. For non-smooth convex objectives, finding an approximate stationary point can require finding an *exact* minimizer of the function (consider the absolute value function). Therefore, as one would expect, the deterministic and stochastic first-order oracle complexities for finding near-stationary points scale polynomially with the smoothness constant, even in low dimensions. Surprisingly, we show that the sample complexity depends at most logarithmically on the smoothness. In fact, in one dimension the dependence on smoothness can be removed entirely.

Improved methods. Our improved sample complexity results for the global stochastic oracle/statistical learning model are based on a new algorithm which uses the *recursive regularization* (or, "SGD3") approach introduced by Allen-Zhu (2018). The new method iteratively solves a sequence of subproblems via regularized empirical risk minimization (RERM). Solving subproblems through RERM allows the method to exploit global access to the stochastic samples. Since the method enjoys only logarithmic dependence on smoothness (as well as initial suboptimality or distance to the optimum), it provides a better alternative to *any* stochastic first-order method whenever the smoothness parameter is large relative to the variance in the gradient estimates. Since RERM is a finite-sum optimization problem, standard finite-sum optimization methods can be used to implement the method efficiently; the result is that we can beat the sample complexity of stochastic first-order methods with only modest computational overhead.

For the local stochastic model, we improve the SGD3 method of Allen-Zhu (2018) so that the "optimization" term matches the optimal deterministic oracle complexity. This leads to a quadratic improvement in terms of the initial distance to the optimum (the "radius" of the problem), $\|x_0 - x^*\|$. We also extend the analysis to the setting where initial sub-optimality $F(x_0) - F(x^*)$ is bounded but not the radius–a common setting in the analysis of non-convex optimization algorithms and a setting in which recursive regularization was not previously analyzed.

2. Setup

We consider the problem of finding an $\epsilon-$stationary point in the stochastic convex optimization setting. That is, for a convex function $F : \mathbb{R}^d \mapsto \mathbb{R}$, our goal is to find a point $x \in \mathbb{R}^d$ such that

$$\|\nabla F(x)\| \leq \epsilon, \tag{1}$$

given access to F only through an *oracle*.[1] Formally, the problem is specified by a class of functions to which F belongs, and through the type of oracle through which we access F. We outline these now.

Function classes. Recall that $F : \mathbb{R}^d \to \mathbb{R}$ is is said to H-smooth if

$$F(y) \leq F(x) + \langle \nabla F(x), y - x \rangle + \frac{H}{2}\|y - x\|^2 \quad \forall x, y \in \mathbb{R}^d, \tag{2}$$

and is said to be λ-strongly-convex if

$$F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle + \frac{\lambda}{2}\|y - x\|^2 \quad \forall x, y \in \mathbb{R}^d. \tag{3}$$

We focus on two classes of objectives, both of which are defined relative to an arbitrary initial point x_0 provided to the optimization algorithm.

1. Here, and for the rest of the paper, $\|\cdot\|$ is taken to be the Euclidean norm.

1. *Domain-bounded functions.*

$$\mathcal{F}^d_{\mathrm{DB}}[H,\lambda;R] = \left\{ F:\mathbb{R}^d\to\mathbb{R} \;\middle|\; \begin{array}{l} F \text{ is } H\text{-smooth and } \lambda\text{-strongly convex} \\ \arg\min_x F(x) \neq \varnothing \\ \exists x^* \in \arg\min_x F(x) \text{ s.t. } \|x_0 - x^*\| \le R \end{array} \right\}. \tag{4}$$

2. *Range-bounded functions.*

$$\mathcal{F}^d_{\mathrm{RB}}[H,\lambda;\Delta] = \left\{ F:\mathbb{R}^d\to\mathbb{R} \;\middle|\; \begin{array}{l} F \text{ is } H\text{-smooth and } \lambda\text{-strongly convex} \\ \arg\min_x F(x) \neq \varnothing \\ F(x_0) - \min_x F(x) \le \Delta \end{array} \right\}. \tag{5}$$

We emphasize that while the classes are defined in terms of a strong convexity parameter, our main complexity results concern the non-strongly convex case where $\lambda = 0$. The strongly convex classes are used for intermediate results. We also note that our main results hold in arbitrary dimension, and so we drop the superscript d except when it is pertinent to discussion.

Oracle classes. An oracle accepts an argument $x \in \mathbb{R}^d$ and provides (possibly noisy/stochastic) information about the objective F around the point x. The oracle's output belongs to an *information space* $\mathcal{I}$. We consider three distinct types of oracles:

1. **Deterministic first-order oracle.** Denoted $\mathcal{O}_{\nabla F}$, with $\mathcal{I} \subseteq \mathbb{R}^d \times (\mathbb{R}^d)^*$. When queried at a point $x \in \mathbb{R}^d$, the oracle returns

$$\mathcal{O}_{\nabla F}(x) = (F(x), \nabla F(x)). \tag{6}$$

2. **Stochastic first-order oracle.** Denoted $\mathcal{O}^\sigma_{\nabla f}$, with $\mathcal{I} \subseteq \mathbb{R}^d \times \mathbb{R}^d$. The oracle is specified by a function $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ and a distribution $\mathcal{D}$ over $\mathcal{Z}$ with the property that $F(x) = \mathbb{E}_{z\sim\mathcal{D}}[f(x;z)]$ and $\sup_x \left[\mathbb{E}_{z\sim\mathcal{D}} \|\nabla f(x;z) - F(x)\|^2\right] \le \sigma^2$. When queried at a point $x \in \mathbb{R}^d$, the oracle draws an independent $z \sim \mathcal{D}$ and returns

$$\mathcal{O}^\sigma_{\nabla f}(x) = (f(x;z), \nabla f(x;z))_{z\sim\mathcal{D}}. \tag{7}$$

3. **Stochastic global oracle.** Denoted $\mathcal{O}^\sigma_f$, with $\mathcal{I} \subseteq (\mathbb{R}^d \mapsto \mathbb{R})$. The oracle is specified by a function $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ and a distribution $\mathcal{D}$ over $\mathcal{Z}$ with the property that $F(x) = \mathbb{E}_{z\sim\mathcal{D}}[f(x;z)]$ and $\sup_x \left[\mathbb{E}_{z\sim\mathcal{D}} \|\nabla f(x;z) - F(x)\|^2\right] \le \sigma^2$. When queried, the oracle draws an independent $z \in \mathcal{D}$ and returns the complete specification of the function $f(\cdot, z)$, specifically,

$$\mathcal{O}^\sigma_f(x) = (f(\cdot, z))_{z\sim\mathcal{D}}. \tag{8}$$

   For consistency with the other oracles, we say that $\mathcal{O}^\sigma_f$ accepts an argument x, even though this argument is ignored. The global oracle captures the *statistical learning* problem, in which $f(\cdot; z)$ is the loss of a model evaluated on an instance $z \sim \mathcal{D}$, and this component function is fully known to the optimizer. Consequently, we use the terms "global stochastic complexity" and "sample complexity" interchangeably.

For the stochastic oracles, while F itself may need to have properties such as convexity or smoothness, $f(\cdot; z)$ as defined need not have these properties unless stated otherwise.

Minimax oracle complexity. Given a function class $\mathcal{F}$ and an oracle $\mathcal{O}$ with information space $\mathcal{I}$, we define the minimax oracle complexity of finding an ϵ-stationary point as

$$\mathsf{m}_\epsilon(\mathcal{F}, \mathcal{O}) = \inf\left\{ m \in \mathbb{N} \;\middle|\; \inf_{A:\bigcup_{t\geq 0}\mathcal{I}^t \mapsto \mathbb{R}^d} \sup_{F\in\mathcal{F}} \mathbb{E}\,\|\nabla F(x_m)\| \leq \epsilon \right\}, \tag{9}$$

where $x_t \in \mathbb{R}^d$ is defined recursively as $x_t := A(O(x_0), \ldots, O(x_{t-1}))$ and the expectation is over the stochasticity of the oracle $\mathcal{O}$.[2]

Recap: Deterministic first-order oracle complexity. To position our new results on stochastic optimization we must first recall what is known about the *deterministic* first-order oracle complexity of finding near-stationary pointst. This complexity is tightly understood, with

$$\mathsf{m}_\epsilon(\mathcal{F}_{\mathrm{DB}}[H, \lambda = 0; R], \mathcal{O}_{\nabla F}) = \tilde{\Theta}(\sqrt{HR}/\sqrt{\epsilon}), \text{ and } \mathsf{m}_\epsilon(\mathcal{F}_{\mathrm{RB}}[H, \lambda = 0; \Delta], \mathcal{O}_{\nabla F}) = \tilde{\Theta}\left(\sqrt{H\Delta}/\epsilon\right),$$

up to logarithmic factors (Nesterov, 2012; Carmon et al., 2017b). The algorithm that achieves these rates is accelerated gradient descent (AGD).

3. Stochastic First-Order Complexity of Finding Stationary Points

Interestingly, the usual variants of stochastic gradient descent do not appear to be optimal in the stochastic model. A first concern is that they do not yield the correct dependence on desired stationarity ϵ. As an illustrative example, let $F \in \mathcal{F}_{\mathrm{DB}}[H, \lambda = 0; R]$ and let any stochastic first-order oracle $\mathcal{O}^\sigma_{\nabla f}$ be given. We adopt the naive approach of bounding stationarity by function value suboptimality. In this case the standard analysis of stochastic gradient descent (e.g., Dekel et al. (2012)) implies that after m iterations, $\mathbb{E}\,\|\nabla F(x_m)\| \leq O(\sqrt{H(\mathbb{E}\,F(x_m) - F(x^*))}) \leq O\left(\sqrt{H\left(HR^2/m + \sigma R/\sqrt{m}\right)}\right)$, and thus

$$\mathsf{m}_\epsilon(\mathcal{F}_{\mathrm{DB}}[H, \lambda = 0; R], \mathcal{O}^\sigma_{\nabla f}) \leq O\left(\frac{H^2R^2}{\epsilon^2} + \frac{H^2R^2\sigma^2}{\epsilon^4}\right).$$

The dependence on ϵ^{-4} is considerably worse than the ϵ^{-2} dependence enjoyed for function suboptimality.

In recent work, Allen-Zhu (2018) proposed a new *recursive regularization* approach and used this in an algorithm called SGD3 that obtains the correct ϵ^{-2} dependence.[3] For any $F \in \mathcal{F}_{\mathrm{DB}}[H, \lambda = 0; R]$ and $\mathcal{O}^\sigma_{\nabla f}$, SGD3 iteratively augments the objective with increasingly strong regularizers, "zooming in" on an approximate stationary point. Specifically, in the first iteration, SGD is used to find $\hat{x}_1$, an approximate minimizer of $F^{(0)}(x) = F(x)$. The objective is then augmented with a strongly-convex regularizer so $F^{(1)}(x) = F^{(0)}(x) + \lambda\|x - \hat{x}_1\|^2$. In the second round, SGD is initialized at $\hat{x}_1$, and used to find $\hat{x}_2$, an approximate minimizer of $F^{(1)}$. This process is repeated, with $F^{(t)}(x) := F^{(t-1)}(x) + 2^{t-1}\lambda\|x - \hat{x}_t\|^2$ for each $t \in [T]$. Allen-Zhu (2018) shows that SGD3 find an ϵ-stationary points using at most

$$m \leq \tilde{O}\left(\frac{HR}{\epsilon} + \frac{\sigma^2}{\epsilon^2}\right) \tag{10}$$

2. See Section 3 for discussion of randomized algorithms.
3. Allen-Zhu (2018) also show that some simple variants of SGD are able to reduce the poor ϵ^{-4} dependence to, e.g., $\epsilon^{-5/2}$, but they fall short of the ϵ^{-2} dependence one should hope for. Similar remarks apply for $F \in \mathcal{F}_{\mathrm{RB}}[H, \lambda = 0; \Delta]$.

local stochastic oracle queries. This oracle complexity has a familiar structure: it resembles the sum of an "optimization term" (HR/ϵ) and a "statistical term" (σ^2/ϵ^2). While we show that the statistical term is tight up to logarithmic factors (Theorem 2), the optimization term does not match the $\Omega(\sqrt{HR/\epsilon})$ lower bound for the deterministic setting (Carmon et al., 2017b).

Algorithm 1 Recursive Regularization Meta-Algorithm

Input: A function $F \in \mathcal{F}[H, \lambda]$, an oracle $\mathcal{O}$ and an alloted number of oracle accesses m, an initial point x_0, and an optimization sub-routine $\mathcal{A}$, with $A = \mathcal{A}[\mathcal{O}, \frac{m}{\lfloor \log_2(\frac{H}{\lambda}) \rfloor}]$.

$F^{(0)} := F$, $\hat{x}_0 := x_0$, $T \leftarrow \lfloor \log_2\left(\frac{H}{\lambda}\right) \rfloor$.

for $t = 1$ to T **do**

$\hat{x}_t$ is output of A used to optimize $F^{(t-1)}$ intitialized at $\hat{x}_{t-1}$

$F^{(t)}(x) := F(x) + \lambda \sum_{k=1}^{t} 2^{k-1} \|x - \hat{x}_k\|^2$

end for

return $\hat{x}_T$

Our first result is to close this gap. The key idea is to view SGD3 as a template algorithm, where the inner loop of SGD used in Allen-Zhu (2018) can be swapped out for an arbitrary optimization method $\mathcal{A}$. This template, Algorithm 1, forms the basis for all the new methods in this paper.[4] To obtain optimal complexity for the local stochastic oracle model we use a variant of the accelerated stochastic approximation method ("AC-SA") due to Ghadimi and Lan (2012) as the subroutine. Pseudocode for AC-SA is provided in Algorithm 2. We use a variant called AC-SA^2, see Algorithm 3. The AC-SA^2 algorithm is equivalent to AC-SA, except the stepsize parameter is reset halfway through. This leads to slightly different dependence on the smoothness and domain size parameters, which is important to control the final rate when invoked within Algorithm 1.

Toward proving the tight upper bound in Table 1, we first show that Algorithm 1 with AC-SA^2 as its subroutine guarantees fast convergence for strongly-convex domain-bounded objectives.

Theorem 1 *For any $F \in \mathcal{F}_{\mathrm{DB}}[H, \lambda; R]$ and any $\mathcal{O}^{\sigma}_{\nabla f}$, Algorithm 1 using AC-SA^2 as its subroutine finds a point $\hat{x}$ with $\mathbb{E}\,\|\nabla F(\hat{x})\| \le \epsilon$ using*

$$m \le O\left(\sqrt{\frac{H}{\lambda}}\log\left(\frac{H}{\lambda}\right) + \sqrt{\frac{HR}{\epsilon}}\log\left(\frac{H}{\lambda}\right) + \left(\frac{\sqrt{H}\sigma}{\sqrt{\lambda}\epsilon}\right)^{\frac{2}{3}}\log\left(\frac{H}{\lambda}\right) + \frac{\sigma^2}{\epsilon^2}\log^3\left(\frac{H}{\lambda}\right)\right)$$

total stochastic first-order oracle accesses.

The analysis of this algorithm is detailed in Appendix A and carefully matches the original analysis of SGD3 (Allen-Zhu, 2018). The essential component of the analysis is Lemma 8, which provides a bound on $\|\nabla F(\hat{x})\|$ in terms of the optimization error of each invocation of AC-SA^2 on the increasingly strongly convex subproblems $F^{(t)}$.

Our final result for non-strongly convex objectives uses Algorithm 1 with AC-SA^2 on the regularized objective $\tilde{F}(x) = F(x) + \frac{\lambda}{2}\|x - x_0\|^2$. The performance guarantee is as follows, and concerns both domain-bounded and range-bounded functions.

4. The idea of replacing the sub-algorithm in SGD3 was also used by Davis and Drusvyatskiy (2018), who showed that recursive regularization with a projected subgradient method can be used to find near-stationary points for the Moreau envelope of any Lipschitz function.

Algorithm 2 AC-SA

Input: A function $F \in \mathcal{F}_{\mathrm{DB}}[H, \lambda; R]$, a stochastic first-order oracle $\mathcal{O}^{\sigma}_{\nabla f}$, and an alloted number of oracle accesses m

$x_0^{ag} = x_0$

for $t = 1, 2, \ldots, m$ **do**

$\alpha_t \leftarrow \frac{2}{t+1}$

$\gamma_t \leftarrow \frac{4H}{t(t+1)}$

$x_t^{md} \leftarrow \frac{(1-\alpha_t)(\lambda+\gamma_t)}{\gamma_t+(1-\alpha_t^2)\lambda} x_{t-1}^{ag} + \frac{\alpha_t((1-\alpha_t)\lambda+\gamma_t)}{\gamma_t+(1-\alpha_t^2)\lambda} x_{t-1}$

$\nabla f(x_t^{md}; z_t) \leftarrow \mathcal{O}^{\sigma}_{\nabla f}(x_t^{md})$

$x_t \leftarrow \frac{\alpha_t \lambda}{\lambda+\gamma_t} x_t^{md} + \frac{(1-\alpha_t)\lambda+\gamma_t}{\lambda+\gamma_t} x_{t-1} - \frac{\alpha_t}{\lambda+\gamma_t} \nabla f(x_t^{md}; z_t)$

$x_t^{ag} \leftarrow \alpha_t x_t + (1-\alpha_t) x_{t-1}^{ag}$

end for

return x_m^{ag}

Corollary 1 *For any $F \in \mathcal{F}_{\mathrm{DB}}[H, \lambda = 0; R]$ and any $\mathcal{O}^{\sigma}_{\nabla f}$, Algorithm 1 with AC-SA^2 as its subroutine applied to $F(x) + \frac{\lambda}{2}\|x - x_0\|^2$ for $\lambda = \Theta\left(\min\left\{\frac{\epsilon}{R}, \frac{H\epsilon^4}{\sigma^4 \log^4(\sigma/\epsilon)}\right\}\right)$ yields a point $\hat{x}$ such that $\mathbb{E}\|\nabla F(\hat{x})\| \le \epsilon$ using*

$$m \le O\left(\sqrt{\frac{HR}{\epsilon}} \log\left(\frac{HR}{\epsilon}\right) + \frac{\sigma^2}{\epsilon^2} \log^3\left(\frac{\sigma}{\epsilon}\right)\right)$$

total stochastic first-order oracle accesses.
For any $F \in \mathcal{F}_{\mathrm{RB}}[H, \lambda = 0; \Delta]$ and any $\mathcal{O}^{\sigma}_{\nabla f}$, the same algorithm with $\lambda = \Theta\left(\min\left\{\frac{\epsilon^2}{\Delta}, \frac{H\epsilon^4}{\sigma^4 \log^4(\sigma/\epsilon)}\right\}\right)$ yields a point $\hat{x}$ with $\mathbb{E}\|\nabla F(\hat{x})\| \le \epsilon$ using

$$m \le O\left(\frac{\sqrt{H\Delta}}{\epsilon} \log\left(\frac{\sqrt{H\Delta}}{\epsilon}\right) + \frac{\sigma^2}{\epsilon^2} \log^3\left(\frac{\sigma}{\epsilon}\right)\right)$$

total stochastic first-order oracle accesses.

This result follows easily from Theorem 1 and is proven in Appendix A. Intuitively, when λ is chosen appropriately, the gradient of the regularized objective $\tilde{F}$ does not significantly deviate from the gradient of F, but the number of iterations required to find an $O(\epsilon)$-stationary point of $\tilde{F}$ is still controlled.

We now provide nearly-tight lower bounds for the stochastic first-order oracle complexity. A notable feature of the lower bound is to show that some of the logarithmic terms in the upper bound—which are not present in the optimal oracle complexity for function value suboptimality—are necessary.

Theorem 2 *For any $H, \Delta, R, \sigma > 0$, any $\epsilon \le \frac{HR}{8}$, the stochastic first-order oracle complexity for range-bounded functions is lower bounded as*

$$\mathsf{m}_\epsilon(\mathcal{F}_{\mathrm{DB}}[H, \lambda = 0; R], \mathcal{O}^{\sigma}_{\nabla f}) \ge \Omega\left(\sqrt{\frac{HR}{\epsilon}} + \frac{\sigma^2}{\epsilon^2} \log\left(\frac{HR}{\epsilon}\right)\right).$$

Algorithm 3 AC-SA^2

Input: A function $F \in \mathcal{F}_{\mathrm{DB}}[H, \lambda; R]$, a stochastic first-order oracle $\mathcal{O}^{\sigma}_{\nabla f}$, and an alloted number of oracle accesses m

$x_1 \leftarrow \text{AC-SA}\left(F, x_0, \frac{m}{2}\right)$

$x_2 \leftarrow \text{AC-SA}\left(F, x_1, \frac{m}{2}\right)$

return x_2

For any $\epsilon \le \sqrt{\frac{H\Delta}{8}}$, the stochastic first-order complexity for domain-bounded functions is lower bounded as

$$\mathsf{m}_\epsilon(\mathcal{F}_{\mathrm{RB}}[H, \lambda = 0; \Delta], \mathcal{O}^{\sigma}_{\nabla f}) \ge \Omega\left(\frac{\sqrt{H\Delta}}{\epsilon} + \frac{\sigma^2}{\epsilon^2}\log\left(\frac{H\Delta}{\epsilon^2}\right)\right).$$

The proof, detailed in Appendix C, combines the existing lower bound on the deterministic first-order oracle complexity (Carmon et al., 2017b) with a new lower bound for the statistical term. The approach is to show that any algorithm for finding near-stationary points can be used to solve noisy binary search (NBS), and then apply a known lower bound for NBS (Feige et al., 1994; Karp and Kleinberg, 2007). It is possible to extend the lower bound to randomized algorithms; see discussion in Carmon et al. (2017b).

4. Sample Complexity of Finding Stationary Points

Having tightly bounded the stochastic first-order oracle complexity of finding approximate stationary points, we now turn to sample complexity. If the heuristic reasoning that stochastic first-order complexity should decompose into sample complexity and deterministic first-order complexity ($\mathsf{m}_\epsilon(\mathcal{F}, \mathcal{O}^{\sigma}_{\nabla f}) \approx \mathsf{m}_\epsilon(\mathcal{F}, \mathcal{O}_{\nabla F}) + \mathsf{m}_\epsilon(\mathcal{F}, \mathcal{O}^{\sigma}_{f})$) is correct, then one would expect that the sample complexity should be $\tilde{O}(\sigma^2/\epsilon^2)$ for both domain-bounded and range-bounded function. A curious feature of this putative sample complexity is that it does not depend on the smoothness of the function. This is somewhat surprising since if the function is non-smooth in the vicinity of its minimizer, there may only be a single ϵ-stationary point, and an algorithm would need to return *exactly* that point using only a finite sample. We show that the sample complexity is in fact almost independent of the smoothness constant, with a mild logarithmic dependence. We also provide nearly tight lower bounds.

For the global setting, a natural algorithm to try is regularized empirical risk minimization (RERM), which returns $\hat{x} = \arg\min_x \frac{1}{m}\sum_{i=1}^m f(x; z_i) + \frac{\lambda}{2}\|x - x_0\|^2$.[5] For any domain-bounded function $F \in \mathcal{F}_{\mathrm{DB}}[H, \lambda = 0; R]$, a standard analysis of ERM based on stability (Shalev-Shwartz et al., 2009) shows that $\mathbb{E}\|\nabla F(\hat{x})\| \le \mathbb{E}\sqrt{2H(F(\hat{x}) - F^*)} + \lambda R \le O(\sqrt{H^3R^2/\lambda m} + \lambda R)$. Choosing $m = \Omega((HR)^3/\epsilon^3)$ and $\lambda = \Theta(\epsilon/R)$ yields an ϵ-stationary point. This upper bound, however, has two shortcomings. First, it scales with ϵ^{-3} rather than ϵ^{-2} that we hoped for and, second, it does not approach 1 as $\sigma \to 0$, which one should expect in the noise-free case. The stochastic first-order algorithm from the previous section has better sample complexity, but the number of samples still does not approach one when $\sigma \to 0$.

5. While it is also tempting to try constrained ERM, this does not succeed even for function value suboptimality (Shalev-Shwartz et al., 2009).

We fix both issues by combining regularized ERM with the recursive regularization approach, giving an upper bound that nearly matches the sample complexity lower bound $\Omega(\sigma^2/\epsilon^2)$. The key tool here is a sharp analysis of regularized ERM—stated in the appendix as Theorem 10—that obtains the correct dependence on the variance σ^2. As in the previous section, we first prove an intermediate result for the strongly convex case. Unlike Section 3, where F was required to be convex but the components $f(\cdot; z)$ were not required to be, we must assume here either that $f(\cdot; z)$ is convex for all z.[6]

Theorem 3 *For any $F \in \mathcal{F}[H, \lambda]$ and any global stochastic oracle $\mathcal{O}_f^\sigma$ with the restriction that $f(\cdot; z)$ is convex for all z, Algorithm 1 with ERM as its subroutine finds $\hat{x}$ with $\mathbb{E}\|\nabla F(\hat{x})\| \le \epsilon$ using at most*

$$m \le O\left(\frac{\sigma^2}{\epsilon^2}\log^3\left(\frac{H}{\lambda}\right)\right)$$

total samples.

The proof is given in Appendix B. As before, we handle the non-strongly convex case by applying the algorithm to $\tilde{F}(x) = F(x) + \frac{\lambda}{2}\|x - x_0\|^2$.

Corollary 2 *For any $F \in \mathcal{F}_{\mathrm{DB}}[H, \lambda = 0; R]$ and any global stochastic oracle $\mathcal{O}_f^\sigma$ with the restriction that $f(\cdot; z)$ is convex for all z, Algorithm 1 with ERM as its subroutine, when applied to $\tilde{F}(x) = F(x) + \frac{\lambda}{2}\|x - x_0\|^2$ with $\lambda = \Theta(\epsilon/R)$, finds a point $\hat{x}$ with $\mathbb{E}\|\nabla F(\hat{x})\| \le \epsilon$ using at most*

$$m \le O\left(\frac{\sigma^2}{\epsilon^2}\log^3\left(\frac{HR}{\epsilon}\right)\right)$$

total samples.
For any $F \in \mathcal{F}_{\mathrm{RB}}[H, \lambda = 0; \Delta]$ and any global stochastic oracle $\mathcal{O}_f^\sigma$ with the restriction that $f(\cdot; z)$ is convex for all z, the same approach with $\lambda = \Theta(\epsilon^2/\Delta)$ finds an ϵ-stationary point using at most

$$m \le O\left(\frac{\sigma^2}{\epsilon^2}\log^3\left(\frac{\sqrt{H\Delta}}{\epsilon}\right)\right)$$

total samples.

This result follows immediately from Theorem 3 by choosing λ small enough such that $\|\nabla F(x)\| \approx \|\nabla \tilde{F}(x)\|$. Details are deferred to Appendix B. With the new sample complexity upper bound, we proceed to provide an almost-tight lower bound.

Theorem 4 *For any $H, \Delta, R, \sigma > 0$, $\epsilon \le \min\{\frac{HR}{8}, \sqrt{\frac{H\Delta}{8}}, \frac{\sigma}{4}\}$, the sample complexity to find a ϵ-stationary point[7] is lower bounded as*

$$\mathsf{m}_\epsilon(\mathcal{F}_{\mathrm{DB}}[H, \lambda = 0; R] \cap \mathcal{F}_{\mathrm{RB}}[H, \lambda = 0; \Delta], \mathcal{O}_f^\sigma) \ge \Omega\left(\frac{\sigma^2}{\epsilon^2}\right).$$

6. We are not aware of any analysis of ERM for strongly convex losses that does not make such an assumption. It is interesting to see whether this can be removed.
7. This lower bound applies both to deterministic and randomized optimization algorithms.

This lower bound is similar to constructions used to prove lower bounds in the case of finding an approximate minimizer (Nemirovski and Yudin, 1983; Nesterov, 2004; Woodworth and Srebro, 2016). However, our lower bound applies for functions with simultaneously bounded domain and range, so extra care must be taken to ensure that these properties hold. The lower bound also ensures that $f(\cdot; z)$ is convex for all z. The proof is located in Appendix C.

Discussion: Efficient implementation. Corollary 2 provides a bound on the number of *samples* needed to find a near-stationary point. However, a convenient property of the method is that the ERM objective $F^{(t)}$ solved in each iteration is convex, $(H + 2^t\lambda)$-smooth, $(2^t\lambda)$-strongly convex, and has *finite sum* structure with m/T components. These subproblems can therefore be solved using at most $O\left(\left(\frac{m}{T} + \sqrt{\frac{m(H+\lambda 2^t)}{T\lambda 2^t}}\right)\log\frac{HR}{\epsilon}\right)$ gradient computations via a first-order optimization algorithm such as Katyusha (Allen-Zhu, 2017). This implies that the method can be implemented with a total gradient complexity of $O\left(\left(\frac{\sigma^2}{\epsilon^2} + \frac{\sigma^{3/2}\sqrt{H}}{\epsilon^{3/2}}\right)\log^4\left(\frac{HR}{\epsilon}\right)\right) \le \tilde{O}\left(\frac{\sigma^2}{\epsilon^2} + \frac{H^2R}{\epsilon}\right)$ over all T iterations, and similarly for the bounded-range case. Thus, the algorithm is not just sample-efficient, but also computationally efficient, albeit slightly less so than the algorithm from Section 3.

Removing smoothness entirely in one dimension. The gap between the upper and lower bounds for the statistical complexity is quite interesting. We conclude from Corollary 2 that the sample complexity depends at most logarithmically upon the smoothness constant, which raises the question of whether it must depend on the smoothness at all. We now show that for the special case of functions in one dimension, smoothness is not necessary. In other words, all that is required to find an ϵ-stationary point is Lipschitzness.

Theorem 5 *Consider any convex, L-Lipschitz function $F : \mathbb{R} \to \mathbb{R}$ that is bounded from below,*[8] *and any global stochastic oracle $\mathcal{O}_f^\sigma$ with the restriction that $f(\cdot; z)$ is convex for all z. There exists an algorithm which uses $O\left(\frac{\sigma^2\log\left(\frac{L}{\epsilon}\right)}{\epsilon^2}\right)$ samples and outputs a point $\hat{x}$ such that $\mathbb{E}\left[\inf_{g\in\partial F(\hat{x})} |g|\right] \le \epsilon$.*

The algorithm calculates the empirical risk minimizer on several independent samples, and then returns the point that has the smallest empirical gradient norm on a validation sample. The proof uses the fact that any function F as in the theorem statement has a single left-most and a single right-most ϵ-stationary point. As long as the empirical function's derivative is close to F's at those two points, we argue that the ERM lies between them with constant probability, and is thus an ϵ-stationary point of F. We are able to boost the confidence by repeating this a logarithmic number of times. A rigorous argument is included in Appendix B. Unfortunately, arguments of this type does not appear to extend to more than one dimension, as the boundary of the set of ϵ-stationary points will generally be uncountable, and thus it is not apparent that the empirical gradient will be uniformly close to the population gradient. It remains open whether smoothness is needed in two dimensions or more.

The algorithm succeeds even for non-differentiable functions, and requires neither strong convexity nor knowledge of a point x_0 for which $\|x_0 - x^*\|$ or $F(x_0) - F^*$ is bounded. In fact, the assumption of Lipschitzness (more generally, L-subgaussianity of the gradients) is only required to get an in-expectation statement. Without this assumption, it can still be shown that ERM finds an ϵ-stationary point with constant probability using $m \le O\left(\frac{\sigma^2}{\epsilon^2}\right)$ samples.

8. This lower bound does not enter the sample complexity quantitatively.

5. Discussion

We have proven nearly tight bounds on the oracle complexity of finding near-stationary points in stochastic convex optimization, both for local stochastic oracles and global stochastic oracles. We hope that the approach of jointly studying stochastic first-order complexity and sample complexity will find use more broadly in non-convex optimization. To this end, we close with a few remarks and open questions.

1. *Is smoothness necessary for finding ϵ-stationary points?* While the logarithmic factor separating the upper and lower bound we provide for stochastic first-order oracle complexity is fairly inconsequential, the gap between the upper and lower bound on the *sample complexity* is quite interesting. In particular, we show through Theorem 4 and Corollary 2 that

$$\Omega\left(\frac{\sigma^2}{\epsilon^2}\right) \le \mathsf{m}_\epsilon\left(\mathcal{F}_{\mathrm{RB}}[H, \lambda = 0; \Delta], \mathcal{O}_f^\sigma\right) \le O\left(\frac{\sigma^2}{\epsilon^2}\log^3\left(\frac{\sqrt{H\Delta}}{\epsilon}\right)\right),$$

   and similarly for the domain-bounded case. Can the $\mathrm{polylog}(H)$ factor on the right-hand side be removed entirely? Or in other words, is it possible to find near-stationary points in the statistical learning model without smoothness?[9] By Theorem 5, we know that this is possible in one dimension.

2. *Tradeoff between computational complexity and sample complexity.* Suppose our end goal is to find a near-stationary point in the statistical learning setting, but we wish to do so efficiently. For range-bounded functions, if we use Algorithm 1 with AC-SA^2 as a subroutine we require $\tilde{O}\left(\frac{\sqrt{H\Delta}}{\epsilon} + \frac{\sigma^2}{\epsilon^2}\right)$ samples, and the total computational effort (measured by number of gradient operations) is also $\tilde{O}\left(\frac{\sqrt{H\Delta}}{\epsilon} + \frac{\sigma^2}{\epsilon^2}\right)$. On the other hand, if we use Algorithm 1 with RERM as a subroutine and implement RERM with Katyusha, then we obtain an improved sample complexity of $\tilde{O}\left(\frac{\sigma^2}{\epsilon^2}\right)$, but at the cost of a larger number of gradient operations: $\tilde{O}\left(\frac{\sigma^2}{\epsilon^2} + \frac{\sqrt{H^3\Delta}}{\epsilon}\right)$. In summary, when faced with functions with poor smoothness, the latter algorithm is superior statistically, with similar but larger computation. Tightly characterizing such computational-statistical tradeoffs in this and related settings is an interesting direction for future work.

3. *Complexity of finding stationary points for smooth non-convex functions.* An important open problem is to characterize the minimax oracle complexity of finding near-stationary points for smooth non-convex functions, both for local and global stochastic oracles. For a deterministic first-order oracle, the optimal rate is $\tilde{\Theta}\left(\frac{H\Delta}{\epsilon^2}\right)$. In the stochastic setting, a simple sample complexity lower bound follows from the convex case, but this is not known to be tight.

Acknowledgements We thank Srinadh Bhojanapalli and Robert D. Kleinberg for helpful discussions. Part of this work was completed while DF was at Cornell University and supported by the Facebook Ph.D. fellowship. OS is partially supported by a European Research Council (ERC) grant. OS and NS are partially supported by an NSF/BSF grant. BW is supported by the NSF Graduate Research Fellowship under award 1754881.

9. For a general non-smooth function F, a point x is said to be an ϵ-stationary point if there exists $v \in \partial F(x)$ such that $\|v\|_2 \le \epsilon$.

References

Alekh Agarwal, Martin J Wainwright, Peter L Bartlett, and Pradeep K Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, pages 1–9, 2009.

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1200–1205. ACM, 2017.

Zeyuan Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex sgd. In *Advances in Neural Information Processing Systems*, pages 1165–1175. 2018.

Gábor Braun, Cristóbal Guzmán, and Sebastian Pokutta. Lower bounds on the oracle complexity of nonsmooth convex optimization via information theory. *IEEE Transactions on Information Theory*, 63(7):4709–4724, 2017.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *arXiv preprint arXiv:1710.11606*, 2017a.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points ii: First-order methods. *arXiv preprint arXiv:1711.00841*, 2017b.

Damek Davis and Dmitriy Drusvyatskiy. Complexity of finding near-stationary points of convex functions stochastically. *arXiv preprint arXiv:1802.08556*, 2018.

Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(Jan):165–202, 2012.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 687–697, 2018.

Uriel Feige, Prabhakar Raghavan, David Peleg, and Eli Upfal. Computing with noisy information. *SIAM Journal on Computing*, 23(5):1001–1018, 1994.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

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

Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Program.*, 156(1-2):59–99, 2016. doi: 10.1007/s10107-015-0871-8.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1724–1732, 2017.

Richard M Karp and Robert Kleinberg. Noisy binary search and its applications. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 881–890. Society for Industrial and Applied Mathematics, 2007.

Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via scsg methods. In *Advances in Neural Information Processing Systems*, pages 2348–2358, 2017.

Arkadii Semenovich Nemirovski and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Yurii Nesterov. Introductory lectures on convex optimization: a basic course. 2004.

Yurii Nesterov. How to make the gradients small. *Optima*, 88:10–11, 2012.

Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczós, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, pages 314–323, 2016.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *Conference on Learning Theory*, 2009.

Joseph F Traub, Grzegorz W Wasilkowski, and Henryk Woźniakowski. Information-based complexity. 1988.

Blake Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems 29*, pages 3639–3647. 2016.

Blake Woodworth, Jialei Wang, Brendan McMahan, and Nathan Srebro. Graph oracle models, lower bounds, and gaps for parallel stochastic optimization. In *Advances in Neural Information Processing Systems 31*, pages 8505–8515, 2018.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduced gradient descent for nonconvex optimization. In *Advances in Neural Information Processing Systems 31*, pages 3925–3936. 2018.

Appendix A. Proofs from Section 3: Upper Bounds

Theorem 6 (Proposition 9 of Ghadimi and Lan (2012)) *For any $F \in \mathcal{F}_{\mathrm{DB}}[H, \lambda; R]$ and any $\mathcal{O}^{\sigma}_{\nabla f}$, the AC-SA algorithm returns a point $\hat{x}_T$ after making T oracle accesses such that*

$$\mathbb{E}[F(\hat{x}_T) - F(x^*)] \le \frac{2HR^2}{T^2} + \frac{8\sigma^2}{\lambda T}.$$

Lemma 7 *For any $F \in \mathcal{F}_{\mathrm{DB}}[H, \lambda; R]$ and any $\mathcal{O}^{\sigma}_{\nabla f}$, the AC-SA^2 algorithm returns a point $\hat{x}$ after making T oracle accesses such that*

$$\mathbb{E}[F(\hat{x}) - F(x^*)] \le \frac{128H^2R^2}{\lambda T^4} + \frac{256H\sigma^2}{\lambda^2 T^3} + \frac{16\sigma^2}{\lambda T}.$$

Proof By Theorem 6, the first instance of AC-SA outputs $\hat{x}_1$ such that

$$\mathbb{E}[F(\hat{x}_1) - F(x^*)] \le \frac{8HR^2}{T^2} + \frac{16\sigma^2}{\lambda T}, \tag{11}$$

and since F is λ-strongly convex,

$$\frac{\lambda}{2}\mathbb{E}\|\hat{x}_1 - x^*\|^2 \le \mathbb{E}[F(\hat{x}_1) - F(x^*)] \le \frac{8HR^2}{T^2} + \frac{16\sigma^2}{\lambda T}. \tag{12}$$

Also by Theorem 6, the second instance of AC-SA outputs $\hat{x}_2$ such that

$$\mathbb{E}[F(\hat{x}_2) - F(x^*)] = \mathbb{E}[\mathbb{E}[F(\hat{x}_2) - F(x^*) \mid \hat{x}_1]] \tag{13}$$

$$\le \mathbb{E}\left[\frac{8H\|\hat{x}_1 - x^*\|^2}{T^2} + \frac{16\sigma^2}{\lambda T}\right] \tag{14}$$

$$\le \frac{128H^2R^2}{\lambda T^4} + \frac{256H\sigma^2}{\lambda^2 T^3} + \frac{16\sigma^2}{\lambda T}. \tag{15}$$

■

Lemma 8 (Claim 6.2 of Allen-Zhu (2018)) *Suppose that for every $t = 1, \dots, T$ the iterates of Algorithm 1 satisfy $\mathbb{E}\big[F^{(t-1)}(\hat{x}_t) - F^{(t-1)}(x^*_{t-1})\big] \le \delta_t$ where $x^*_{t-1} = \arg\min_x F^{(t-1)}(x)$, then*

1. *For all $t \ge 1$, $\mathbb{E}[\|\hat{x}_t - x^*_{t-1}\|]^2 \le \mathbb{E}\big[\|\hat{x}_t - x^*_{t-1}\|^2\big] \le \dfrac{\delta_t}{2^{t-2}\lambda}$.*
2. *For every $t \ge 1$, $\mathbb{E}[\|\hat{x}_t - x^*_t\|]^2 \le \mathbb{E}\big[\|\hat{x}_t - x^*_t\|^2\big] \le \dfrac{\delta_t}{2^t\lambda}$.*
3. *For all $t \ge 1$, $\mathbb{E}\big[\sum_{t=1}^T 2^t\lambda\|\hat{x}_t - x^*_T\|\big] \le 4\sum_{t=1}^T \sqrt{2^t\lambda\delta_t}$.*

Theorem 1 *For any $F \in \mathcal{F}_{\mathrm{DB}}[H, \lambda; R]$ and any $\mathcal{O}^{\sigma}_{\nabla f}$, Algorithm 1 using AC-SA^2 as its subroutine finds a point $\hat{x}$ with $\mathbb{E}\,\|\nabla F(\hat{x})\| \le \epsilon$ using*

$$m \le O\left(\sqrt{\frac{H}{\lambda}}\log\left(\frac{H}{\lambda}\right) + \sqrt{\frac{HR}{\epsilon}}\log\left(\frac{H}{\lambda}\right) + \left(\frac{\sqrt{H}\sigma}{\sqrt{\lambda}\epsilon}\right)^{\frac{2}{3}}\log\left(\frac{H}{\lambda}\right) + \frac{\sigma^2}{\epsilon^2}\log^3\left(\frac{H}{\lambda}\right)\right)$$

total stochastic first-order oracle accesses.

Proof As in Lemma 8, let $\mathbb{E}[F^{(t-1)}(\hat{x}_t) - F^{(t-1)}(x^*_{t-1})] \le \delta_t$ for each $t \ge 1$. The objective in the final iteration, $F^{(T-1)}(x) = F(x) + \lambda\sum_{t=1}^{T-1} 2^{t-1}\|x - \hat{x}_t\|^2$, so

$$\mathbb{E}[\|\nabla F(\hat{x}_T)\|] = \mathbb{E}\left[\left\|\nabla F^{(T-1)}(\hat{x}_T) + \lambda\sum_{t=1}^{T-1} 2^t(\hat{x}_t - \hat{x}_T)\right\|\right] \tag{16}$$

$$\leq \mathbb{E}\Big[\Big\|\nabla F^{(T-1)}(\hat{x}_T)\Big\|\Big] + \lambda \sum_{t=1}^{T-1} 2^t \,\mathbb{E}[\|\hat{x}_t - \hat{x}_T\|] \tag{17}$$

$$\leq \mathbb{E}\Big[\Big\|\nabla F^{(T-1)}(\hat{x}_T)\Big\|\Big] + \lambda \sum_{t=1}^{T-1} 2^t \,\mathbb{E}[\|\hat{x}_t - x^*_{T-1}\| + \|\hat{x}_T - x^*_{T-1}\|] \tag{18}$$

$$\leq 2\,\mathbb{E}\Big[\Big\|\nabla F^{(T-1)}(\hat{x}_T)\Big\|\Big] + \lambda \sum_{t=1}^{T-1} 2^t \,\mathbb{E}[\|\hat{x}_t - x^*_{T-1}\|] \tag{19}$$

$$\leq 2\,\mathbb{E}\Big[\Big\|\nabla F^{(T-1)}(\hat{x}_T)\Big\|\Big] + 4 \sum_{t=1}^{T-1} \sqrt{\lambda 2^t \delta_t} \tag{20}$$

$$\leq 4\sqrt{H\,\mathbb{E}\big[F^{(T-1)}(\hat{x}_T) - F^{(T-1)}(x^*_{T-1})\big]} + 4 \sum_{t=1}^{T-1} \sqrt{\lambda 2^t \delta_t} \tag{21}$$

$$\leq 4\sqrt{H\delta_T} + 4 \sum_{t=1}^{T-1} \sqrt{\lambda 2^t \delta_t} \tag{22}$$

$$\leq 4 \sum_{t=1}^{T} \sqrt{\lambda 2^{t+1} \delta_t}. \tag{23}$$

Above, (17) and (18) rely on the triangle inequality; (19) follows from the $\left(\lambda \sum_{t=1}^{T-1} 2^t\right)$-strong convexity of $F^{(T-1)}$; (20) applies the third conclusion of Lemma 8; (22) uses the fact that $F^{(t-1)}$ is $H + \lambda \sum_{t=1}^{T-1} 2^t < H + \lambda 2^T = H + \lambda 2^{\lfloor \log H/\lambda \rfloor} \leq 2H$-smooth; and finally (23) uses that $H \leq \lambda 2^{T+1}$.

We chose $\mathcal{A}(F^{(t-1)}, \hat{x}_{t-1})$ to be AC-SA^2 applied to $F^{(t-1)}$ initialized at $\hat{x}_{t-1}$ using m/T stochastic gradients. Therefore,

$$\delta_t \leq \frac{128H^2 \mathbb{E}\,\|\hat{x}_{t-1} - x^*_{t-1}\|^2}{2^{t-1}\lambda(m/T)^4} + \frac{256H\sigma^2}{2^{2t-2}\lambda^2(m/T)^3} + \frac{16\sigma^2}{2^{t-1}\lambda(m/T)}. \tag{24}$$

Using part two of Lemma 8, for $t > 1$ we can bound $\mathbb{E}\,\|\hat{x}_{t-1} - x^*_{t-1}\|^2 \leq \frac{\delta_{t-1}}{2^{t-1}\lambda}$, thus

$$\delta_t \leq \frac{128H^2\delta_{t-1}}{2^{2t-2}\lambda^2(m/T)^4} + \frac{256H\sigma^2}{2^{2t-2}\lambda^2(m/T)^3} + \frac{16\sigma^2}{2^{t-1}\lambda(m/T)}. \tag{25}$$

We can therefore bound

$$8 \sum_{t=1}^{T} \sqrt{\lambda 2^{t-1} \delta_t} \leq 8\sqrt{\frac{128H^2\,\|x_0 - x^*\|^2}{(m/T)^4} + \frac{256H\sigma^2}{\lambda(m/T)^3} + \frac{16\sigma^2}{(m/T)}} + 8 \sum_{t=2}^{T} \sqrt{\frac{128H^2\delta_{t-1}}{2^{t-1}\lambda(m/T)^4} + \frac{256H\sigma^2}{2^{t-1}\lambda(m/T)^3} + \frac{16\sigma^2}{(m/T)}} \tag{26}$$

$$\leq 8\sqrt{\frac{128H^2\,\|x_0 - x^*\|^2}{(m/T)^4}} + 8\sqrt{\frac{256H\sigma^2}{\lambda(m/T)^3}} + 8\sqrt{\frac{16\sigma^2}{(m/T)}} + 8 \sum_{t=2}^{T} \sqrt{\frac{128H^2\delta_{t-1}}{2^{t-1}\lambda(m/T)^4}} + \sqrt{\frac{256H\sigma^2}{2^{t-1}\lambda(m/T)^3}} + \sqrt{\frac{16\sigma^2}{(m/T)}} \tag{27}$$

$$= \frac{64\sqrt{2}H\left\|x_0 - x^*\right\|T^2}{m^2} + \frac{128\sqrt{H}\sigma T^{3/2}}{\sqrt{\lambda}m^{3/2}}\sum_{t=1}^{T}\frac{1}{\sqrt{2^{t-1}}} + \frac{32\sigma T^{3/2}}{\sqrt{m}} + \frac{128HT^2}{m^2}\sum_{t=2}^{T}\sqrt{\frac{\delta_{t-1}}{2^{t-2}\lambda}} \tag{28}$$

$$\leq \frac{64\sqrt{2}H\left\|x_0 - x^*\right\|T^2}{m^2} + \frac{512\sqrt{H}\sigma T^{3/2}}{\sqrt{\lambda}m^{3/2}} + \frac{32\sigma T^{3/2}}{\sqrt{m}} + \frac{128HT^2}{m^2}\sum_{t=1}^{T}\sqrt{\frac{\delta_t}{2^{t-1}\lambda}} \tag{29}$$

$$\leq \frac{64\sqrt{2}H\left\|x_0 - x^*\right\|T^2}{m^2} + \frac{512\sqrt{H}\sigma T^{3/2}}{\sqrt{\lambda}m^{3/2}} + \frac{32\sigma T^{3/2}}{\sqrt{m}} + \frac{128HT^2}{\lambda m^2}\sum_{t=1}^{T}\sqrt{\lambda 2^{t-1}\delta_t}. \tag{30}$$

Above, we arrive at (26) by upper bounding each δ_t via (25); (27) follows from the fact that for $a, b \geq 0$, $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$; (29) uses the fact that $\sum_{t=1}^{T}\frac{1}{\sqrt{2^{t-1}}} \leq 4$ and $\sqrt{\frac{\delta_T}{2^{T-1}\lambda}} \geq 0$; and finally, (30) follows by multiplying each non-negative term in the sum by 2^{t-1}. Rearranging inequality (30) and combining with (23) yields

$$\mathbb{E}\left\|\nabla F(\hat{x}_T)\right\| \leq \left(\frac{1}{1 - \frac{16HT^2}{\lambda m^2}}\right)\left(\frac{64\sqrt{2}H\left\|x_0 - x^*\right\|T^2}{m^2} + \frac{512\sqrt{H}\sigma T^{3/2}}{\sqrt{\lambda}m^{3/2}} + \frac{32\sigma T^{3/2}}{\sqrt{m}}\right). \tag{31}$$

Choosing $m > 8T\sqrt{\frac{H}{\lambda}}$ ensures that the first term is at most 2, and then solving for m such that the second term is $O(\epsilon)$ completes the proof. ■

Lemma 9 *For any F, define $\tilde{F}(x) = F(x) + \frac{\lambda}{2}\left\|x - x_0\right\|$. Then*

1. $F \in \mathcal{F}_{\mathrm{DB}}[H, \lambda = 0; R] \implies \tilde{F} \in \mathcal{F}_{\mathrm{DB}}[H + \lambda, \lambda; R]$ *and* $\forall x \left\|\nabla F(x)\right\| \leq 2\left\|\nabla \tilde{F}(x)\right\| + \lambda R$.
2. $F \in \mathcal{F}_{\mathrm{RB}}[H, \lambda = 0; \Delta] \implies$ $\tilde{F} \in \mathcal{F}_{\mathrm{DB}}[H + \lambda, \lambda; R = \sqrt{2\Delta/\lambda}]$ *and* $\forall x \left\|\nabla F(x)\right\| \leq 2\left\|\nabla \tilde{F}(x)\right\| + \sqrt{2\lambda\Delta}$.

Proof Let $\tilde{x}^* \in \arg\min_x \tilde{F}(x)$. Since $\nabla\tilde{F}(x) = \nabla F(x) + \lambda(x - x_0)$,

$$\left\|\nabla F(x)\right\| \leq \left\|\nabla \tilde{F}(x)\right\| + \lambda\left\|x - x_0\right\| \tag{32}$$

$$\leq \left\|\nabla \tilde{F}(x)\right\| + \lambda\left\|x_0 - \tilde{x}^*\right\| + \lambda\left\|x - \tilde{x}^*\right\| \tag{33}$$

$$\leq 2\left\|\nabla \tilde{F}(x)\right\| + \lambda\left\|x_0 - \tilde{x}^*\right\|, \tag{34}$$

where we used the λ-strong convexity of $\tilde{F}$ for the last inequality. Similarly, $0 = \nabla\tilde{F}(\tilde{x}^*) = \nabla F(\tilde{x}^*) + \lambda(\tilde{x}^* - x_0)$. Therefore,

$$\begin{aligned}
\lambda \left\|x_0 - \tilde{x}^*\right\|^2 &= \left\langle \nabla F(\tilde{x}^*), x_0 - \tilde{x}^* \right\rangle && (35)\\
&= \left\langle \nabla F(\tilde{x}^*), x_0 - x^* \right\rangle + \left\langle \nabla F(\tilde{x}^*), x^* - \tilde{x}^* \right\rangle && (36)\\
&\le \left\langle \nabla F(\tilde{x}^*), x_0 - x^* \right\rangle && (37)\\
&= \left\langle \lambda(x_0 - \tilde{x}^*), x_0 - x^* \right\rangle && (38)\\
&\le \lambda \left\|x_0 - \tilde{x}^*\right\| \left\|x_0 - x^*\right\|. && (39)
\end{aligned}$$

The first inequality follows from the convexity of F and the second from the Cauchy-Schwarz inequality. When $F \in \mathcal{F}_{\mathrm{DB}}[H, \lambda = 0; R]$, then $\|x_0 - \tilde{x}^*\| \le R$, which, combined with (34) proves the first claim.

Alternatively, when $F \in \mathcal{F}_{\mathrm{RB}}[H, \lambda = 0; \Delta]$

$$F(x_0) = \tilde{F}(x_0) \ge \tilde{F}(\tilde{x}^*) = F(\tilde{x}^*) + \frac{\lambda}{2}\left\|x_0 - \tilde{x}^*\right\|^2. \tag{40}$$

Rearranging,

$$\left\|x_0 - \tilde{x}^*\right\| \le \sqrt{\frac{2(F(x_0) - F(\tilde{x}^*))}{\lambda}} \le \sqrt{\frac{2(F(x_0) - F(x^*))}{\lambda}} \le \sqrt{\frac{2\Delta}{\lambda}}. \tag{41}$$

This, combined with (34), completes the proof. ■

Corollary 1 *For any $F \in \mathcal{F}_{\mathrm{DB}}[H, \lambda = 0; R]$ and any $\mathcal{O}^{\sigma}_{\nabla f}$, Algorithm 1 with AC-SA^2 as its subroutine applied to $F(x) + \frac{\lambda}{2}\|x - x_0\|^2$ for $\lambda = \Theta\left(\min\left\{\frac{\epsilon}{R}, \frac{H\epsilon^4}{\sigma^4 \log^4(\sigma/\epsilon)}\right\}\right)$ yields a point $\hat{x}$ such that $\mathbb{E}\|\nabla F(\hat{x})\| \le \epsilon$ using*

$$m \le O\left(\sqrt{\frac{HR}{\epsilon}} \log\left(\frac{HR}{\epsilon}\right) + \frac{\sigma^2}{\epsilon^2}\log^3\left(\frac{\sigma}{\epsilon}\right)\right)$$

total stochastic first-order oracle accesses.
For any $F \in \mathcal{F}_{\mathrm{RB}}[H, \lambda = 0; \Delta]$ and any $\mathcal{O}^{\sigma}_{\nabla f}$, the same algorithm with $\lambda = \Theta\left(\min\left\{\frac{\epsilon^2}{\Delta}, \frac{H\epsilon^4}{\sigma^4 \log^4(\sigma/\epsilon)}\right\}\right)$ yields a point $\hat{x}$ with $\mathbb{E}\|\nabla F(\hat{x})\| \le \epsilon$ using

$$m \le O\left(\frac{\sqrt{H\Delta}}{\epsilon} \log\left(\frac{\sqrt{H\Delta}}{\epsilon}\right) + \frac{\sigma^2}{\epsilon^2}\log^3\left(\frac{\sigma}{\epsilon}\right)\right)$$

total stochastic first-order oracle accesses.

Proof We use Algorithm 1 with AC-SA^2 as its subroutine to optimize $\tilde{F}(x) = F(x) + \frac{\lambda}{2}\|x - x_0\|^2$. Our choice of $\lambda = \frac{256H\log^2(m^2)}{m^2} \le O(H)$ ensures that $\tilde{F}$ is $H + \lambda \le O(H)$-smooth and λ-strongly convex; that $\frac{16(H+\lambda)\log^2\left(\frac{H+\lambda}{\lambda}\right)}{\lambda m^2} \le \frac{1}{2}$; and finally that $\frac{H}{\lambda} \le m^2$. Therefore, by Theorem 1, in particular, (31), the output satisfies

$$\mathbb{E}\left\|\nabla\tilde{F}(\hat{x})\right\| \le O\left(\frac{H\left\|x_0 - \tilde{x}^*\right\|\log^2(H/\lambda)}{m^2} + \frac{\sqrt{H}\sigma\log^{3/2}(H/\lambda)}{\sqrt{\lambda}m^{3/2}} + \frac{\sigma\log^{3/2}(H/\lambda)}{\sqrt{m}}\right) \tag{42}$$

$$\leq O\left(\frac{H\left\|x_0-\tilde{x}^*\right\|\log^2(m)}{m^2}+\frac{\sigma\log^{3/2}(m)}{\sqrt{m}}\right), \tag{43}$$

where $\tilde{x}^* = \arg\min_x \tilde{F}(x)$. For $F \in \mathcal{F}_{\mathrm{DB}}[H, \lambda = 0; R]$, by part one of Lemma 9, $\|x_0 - \tilde{x}^*\| \leq R$ and

$$\mathbb{E}\left\|\nabla F(\hat{x})\right\| \leq O\left(\frac{HR\log^2(m)}{m^2}+\frac{\sigma\log^{3/2}(m)}{\sqrt{m}}\right). \tag{44}$$

Solving for m such that this expression is $O(\epsilon)$ completes the first part of the proof. For this m,

$$\lambda = \Theta\left(\min\left\{\frac{\epsilon}{R}, \frac{H\epsilon^4}{\sigma^4\log^4(\sigma/\epsilon)}\right\}\right). \tag{45}$$

For $F \in \mathcal{F}_{\mathrm{RB}}[H, \lambda = 0; \Delta]$, by part two of Lemma 9, $\|x_0 - \tilde{x}^*\| \leq \sqrt{2\Delta/\lambda}$ and

$$\mathbb{E}\left\|\nabla F(\hat{x})\right\| \leq O\left(\frac{\sqrt{H\Delta}\log(m)}{m}+\frac{\sigma\log^{3/2}(m)}{\sqrt{m}}\right). \tag{46}$$

Solving for m such that this expression is $O(\epsilon)$ completes the the proof. For this m,

$$\lambda = \Theta\left(\min\left\{\frac{\epsilon^2}{\Delta}, \frac{H\epsilon^4}{\sigma^4\log^4(\sigma/\epsilon)}\right\}\right). \tag{47}$$

■

Appendix B. Proofs from Section 4: Upper Bounds

Theorem 10 *For any $F \in \mathcal{F}[H, \lambda]$ and any $\mathcal{O}_f^\sigma$ with the restriction that $f(x; z)$ is λ-strongly convex with respect to x for all z, define the empirical risk minimizer via*

$$\hat{x} = \arg\min_{x\in\mathbb{R}^d} \frac{1}{m}\sum_{t=1}^{m} f(x; z_t).$$

Then the empirical risk minimizer enjoys the guarantee

$$\mathbb{E}\|\hat{x} - x^*\|^2 \leq \frac{4\sigma^2}{\lambda^2 m}. \tag{48}$$

Proof Let $\widehat{F}_m(x) = \frac{1}{m}\sum_{t=1}^m f(x; z_t)$ be the empirical objective. Since $f(x; z_t)$ is λ-strongly convex for each z_t, $\widehat{F}_m$ is itself λ-strongly convex, and so we have

$$\left\langle\nabla\widehat{F}_m(x^*), \hat{x} - x^*\right\rangle + \frac{\lambda}{2}\|\hat{x} - x^*\|^2 \leq \widehat{F}_m(\hat{x}) - \widehat{F}_m(x^*).$$

Since, $\hat{x}$ is the empirical risk minimizer, we have $\widehat{F}_m(\hat{x}) - \widehat{F}_m(x^*) \leq 0$, and so, rearranging,

$$\frac{\lambda}{2}\|\hat{x} - x^*\|^2 \leq \left\langle\nabla\widehat{F}_m(x^*), \hat{x} - x^*\right\rangle \leq \left\|\nabla\widehat{F}_m(x^*)\right\|\|\hat{x} - x^*\|.$$

If $\hat{x} - x^* = 0$, then we are done. Otherwise,

$$\|\hat{x} - x^*\| \le \frac{2}{\lambda}\left\|\nabla \widehat{F}_m(x^*)\right\|.$$

Now square both sides and take the expectation, which gives

$$\mathbb{E}\|\hat{x} - x^*\|_2^2 \le \frac{4}{\lambda^2}\,\mathbb{E}\left\|\nabla \widehat{F}_m(x^*)\right\|^2.$$

The final result follows by observing that $\mathbb{E}\left\|\nabla \widehat{F}_m(x^*)\right\|^2 \le \frac{\sigma^2}{m}$. ■

Theorem 3 *For any $F \in \mathcal{F}[H, \lambda]$ and any global stochastic oracle $\mathcal{O}_f^\sigma$ with the restriction that $f(\cdot; z)$ is convex for all z, Algorithm 1 with ERM as its subroutine finds $\hat{x}$ with $\mathbb{E}\,\|\nabla F(\hat{x})\| \le \epsilon$ using at most*

$$m \le O\left(\frac{\sigma^2}{\epsilon^2}\log^3\left(\frac{H}{\lambda}\right)\right)$$

total samples.

Proof Consider the function $F^{(T)}(x) = F(x) + \lambda \sum_{t=1}^{T} 2^{t-1}\|x - \hat{x}_t\|^2$. Then

$$\|\nabla F(\hat{x}_T)\| = \left\|\nabla F^{(T)}(\hat{x}_T) + \lambda \sum_{t=1}^{T} 2^t(\hat{x}_t - \hat{x}_T)\right\| \tag{49}$$

$$\le \left\|\nabla F^{(T)}(\hat{x}_T)\right\| + \lambda \sum_{t=1}^{T-1} 2^t \|\hat{x}_t - \hat{x}_T\| \tag{50}$$

$$\le \left\|\nabla F^{(T)}(\hat{x}_T)\right\| + \lambda \sum_{t=1}^{T-1} 2^t \left(\|\hat{x}_t - x_T^*\| + \|\hat{x}_T - x_T^*\|\right) \tag{51}$$

$$\le 2\left\|\nabla F^{(T)}(\hat{x}_T)\right\| + \lambda \sum_{t=1}^{T-1} 2^t \|\hat{x}_t - x_T^*\| \tag{52}$$

$$\le 6H\|\hat{x}_T - x_T^*\| + \lambda \sum_{t=1}^{T-1} 2^t \|\hat{x}_t - x_T^*\| \tag{53}$$

$$\le 12\lambda \sum_{t=1}^{T} 2^t \|\hat{x}_t - x_T^*\|. \tag{54}$$

Above, (50) and (51) rely on the triangle inequality; (52) follows from the $\left(\lambda \sum_{t=1}^{T} 2^t\right)$-strong convexity of $F^{(T)}$; (53) uses the fact that $F^{(T)}$ is $H + \lambda \sum_{t=1}^{T} 2^t < H + \lambda 2^{T+1} = H + 2\lambda 2^{\lfloor \log H/\lambda \rfloor} \le 3H$-smooth.

Define $P_k = \sum_{t=1}^{k} 2^t \left\|\hat{x}_t - x_k^*\right\|$ for $1 \le k \le T$ with $P_0 = 0$. Note that our upper bound (54) is equal to $12\lambda P_T = 12\lambda \sum_{k=1}^{T}(P_k - P_{k-1})$, so we will estimate the terms of this sum.

$$P_k - P_{k-1} = 2^k \|\hat{x}_k - x_k^*\| + \sum_{t=1}^{k-1} 2^t \left(\|\hat{x}_t - x_k^*\| - \|\hat{x}_t - x_{k-1}^*\|\right) \tag{55}$$

$$\le 2^k \|\hat{x}_k - x_k^*\| + \sum_{t=1}^{k-1} 2^t \|x_k^* - x_{k-1}^*\| \tag{56}$$

$$\leq 2^k \left(\|\hat{x}_k - x_k^*\| + \|x_k^* - x_{k-1}^*\| \right) \tag{57}$$

$$\leq 2^k \left(2\|\hat{x}_k - x_k^*\| + \|\hat{x}_k - x_{k-1}^*\| \right). \tag{58}$$

Above, we used the reverse triangle inequality to derive (56). By optimality of x_{k-1}^* and x_k^*,

$$\|\hat{x}_k - x_{k-1}^*\|^2 - \|\hat{x}_k - x_k^*\|^2 = \frac{F^{(k)}(x_{k-1}^*) - F^{(k)}(x_k^*) + F^{(k-1)}(x_k^*) - F^{(k-1)}(x_{k-1}^*)}{2^{k-1}\lambda} \geq 0. \tag{59}$$

Thus $\|\hat{x}_k - x_k^*\| \leq \|\hat{x}_k - x_{k-1}^*\|$ and, combining (54) and (58) yields

$$\|\nabla F(\hat{x}_T)\| \leq 36\lambda \sum_{t=1}^{T} 2^t \|\hat{x}_t - x_{t-1}^*\|. \tag{60}$$

Since $\hat{x}_t$ is the output of ERM on the $2^{t-1}\lambda$-strongly convex function F^{t-1} using m/T samples, by Theorem 10, $\mathbb{E}\|\hat{x}_t - x_{t-1}^*\| \leq \frac{2\sigma\sqrt{T}}{2^{t-1}\lambda\sqrt{m}}$ and

$$\mathbb{E}\|\nabla F(\hat{x}_T)\| \leq 36\lambda \sum_{t=1}^{T} 2^t \mathbb{E}\|\hat{x}_t - x_{t-1}^*\| \tag{61}$$

$$\leq 36\lambda \sum_{t=1}^{T} 2^t \frac{\sigma\sqrt{T}}{2^{t-2}\lambda\sqrt{m}} \tag{62}$$

$$= \frac{144\sigma T^{3/2}}{\sqrt{m}}. \tag{63}$$

Solving for m such that the expression is less than ϵ completes the proof. ■

Corollary 2 *For any $F \in \mathcal{F}_{\mathrm{DB}}[H, \lambda = 0; R]$ and any global stochastic oracle $\mathcal{O}_f^\sigma$ with the restriction that $f(\cdot; z)$ is convex for all z, Algorithm 1 with ERM as its subroutine, when applied to $\tilde{F}(x) = F(x) + \frac{\lambda}{2}\|x - x_0\|^2$ with $\lambda = \Theta(\epsilon/R)$, finds a point $\hat{x}$ with $\mathbb{E}\|\nabla F(\hat{x})\| \leq \epsilon$ using at most*

$$m \leq O\left(\frac{\sigma^2}{\epsilon^2} \log^3\left(\frac{HR}{\epsilon} \right) \right)$$

total samples.
For any $F \in \mathcal{F}_{\mathrm{RB}}[H, \lambda = 0; \Delta]$ and any global stochastic oracle $\mathcal{O}_f^\sigma$ with the restriction that $f(\cdot; z)$ is convex for all z, the same approach with $\lambda = \Theta(\epsilon^2/\Delta)$ finds an ϵ-stationary point using at most

$$m \leq O\left(\frac{\sigma^2}{\epsilon^2} \log^3\left(\frac{\sqrt{H\Delta}}{\epsilon} \right) \right)$$

total samples.

Proof The objective function $\tilde{F}(x) = F(x) + \frac{\lambda}{2}\|x - x_0\|^2$ is $(H + \lambda)$-smooth and λ-strongly convex. Thus by Theorem 3, in particular (63), the output of the algorithm satisfies

$$\mathbb{E}\|\nabla \tilde{F}(\hat{x})\| \leq \frac{144\sigma \log^{3/2}\left(\frac{H+\lambda}{\lambda}\right)}{\sqrt{m}}. \tag{64}$$

For $F \in \mathcal{F}_{\mathrm{DB}}[H, \lambda = 0; R]$, with $\lambda = \Theta(\epsilon/R)$ and $m = \Omega\left(\frac{\sigma^2}{\epsilon^2}\log^3\left(\frac{HR}{\epsilon}\right)\right)$ and using part one of Lemma 9 we conclude

$$\mathbb{E}\,\|\nabla F(\hat{x})\| \le O(\epsilon + \lambda R) \le O(\epsilon), \tag{65}$$

which completes the first part of the proof.

Similarly, for $F \in \mathcal{F}_{\mathrm{RB}}[H, \lambda = 0; \Delta]$, with $\lambda = \Theta(\epsilon^2/\Delta)$ and $m = \Omega\left(\frac{\sigma^2}{\epsilon^2}\log^3\left(\frac{\sqrt{H\Delta}}{\epsilon}\right)\right)$, by part two of Lemma 9 we conclude

$$\mathbb{E}\,\|\nabla F(\hat{x})\| \le O\left(\epsilon + \sqrt{\lambda\Delta}\right) \le O(\epsilon), \tag{66}$$

which completes the proof. ■

Theorem 5 *Consider any convex, L-Lipschitz function $F : \mathbb{R} \to \mathbb{R}$ that is bounded from below,*[10] *and any global stochastic oracle $\mathcal{O}_f^\sigma$ with the restriction that $f(\cdot; z)$ is convex for all z. There exists an algorithm which uses $O\left(\frac{\sigma^2 \log\left(\frac{L}{\epsilon}\right)}{\epsilon^2}\right)$ samples and outputs a point $\hat{x}$ such that $\mathbb{E}\left[\inf_{g\in\partial F(\hat{x})} |g|\right] \le \epsilon$.*

Proof Our algorithm involves calculating the ERM on several independent samples, evaluating the gradient norm at these ERMs on a held-out sample, and returning the point with the smallest gradient norm.

Let $\nabla_- F(x)$ denote the left-derivative of F at x, and let $\nabla_+ F(x)$ denote the right-derivative. Since F is bounded from below, $\lim_{x\to-\infty} \nabla_- F(x) \le 0$ and $\lim_{x\to\infty} \nabla_+ F(x) \ge 0$, thus there exists at least one ϵ-stationary point for F. Consequently, there is a unique $a \in \mathbb{R} \cup \{-\infty\}$ for which $\nabla_+ F(a) \ge -\epsilon$ and $\forall x < a\ \nabla_+ F(x) < -\epsilon$. The point a is the left-most ϵ-stationary point. It is possible that $a = -\infty$, in which case there are no $x < a$. Similarly, there is a unique $b \in \mathbb{R} \cup \{\infty\}$ for which $\nabla_- F(b) \le \epsilon$ and $\forall x > b\ \nabla_- F(x) > \epsilon$. The point b is the right-most ϵ-stationary point. It is possible that $b = \infty$, in which case there are no $x > b$.

By convexity, $\forall x < y\ \nabla_- F(x) \le \nabla_+ F(x) \le \nabla_- F(y) \le \nabla_+ F(y)$. Therefore, $x < a \implies \inf_{g\in\partial F(x)} |g| \ge |\nabla_+ F(x)| > \epsilon$ and $x > b \implies \inf_{g\in\partial F(x)} |g| \ge |\nabla_- F(x)| > \epsilon$. Therefore, $[a, b] \equiv \{x : \inf_{g\in\partial F(x)} |g| \le \epsilon\}$. Consequently, all that we need to show is that our algorithm returns a point within the interval $[a, b]$.

Let $\hat{F}(x) = \frac{1}{m}\sum_{i=1}^m f(x; z_i)$ be the empirical objective function and let $\hat{x}$ be any minimizer of $\hat{F}$. Consider first the case that $a > -\infty$, we will argue that $\hat{x} \ge a$. Observe that if $\nabla_- \hat{F}(a) < 0$, then since $\hat{F}$ is convex, it is decreasing on $[-\infty, a]$ and thus $\hat{x} \ge a$. Since $a > -\infty$, $\nabla_- F(a) \le -\epsilon$, so the value $\nabla_- \hat{F}(a) = \frac{1}{m}\sum_{i=1}^m \nabla_- f(a; z_i)$ is the sum of i.i.d. random variables that have mean $\nabla_- F(a) \le -\epsilon$ and variance σ^2. By Chebyshev's inequality, the random variable $\nabla_- \hat{F}(a)$ will not deviate too far from its mean:

$$\mathbb{P}\left[\nabla_- \hat{F}(a) \ge 0\right] \le \frac{\sigma^2}{m\epsilon^2}. \tag{67}$$

Similarly,

$$\mathbb{P}\left[\nabla_+ \hat{F}(b) \le 0\right] \le \frac{\sigma^2}{m\epsilon^2}. \tag{68}$$

10. This lower bound does not enter the sample complexity quantitatively.

Therefore, with probability at least $1 - \frac{2\sigma^2}{m\epsilon^2}$, the minimum of $\hat{F}$ lies in the range $[a, b]$ and thus the ERM $\hat{x}$ is an ϵ-stationary point of F.

Consider calculating k ERMs $\hat{x}_1, \ldots, \hat{x}_k$ on k independent samples of size m. Then with probability at least $1 - \left(\frac{2\sigma^2}{m\epsilon^2}\right)^k$, at least one of these points is an ϵ-stationary point of F.

Now, suppose we have km additional heldout samples which constitute an empirical objective $\hat{F}$. Since the ERMs $\hat{x}_i$ are independent of these samples,

$$\mathbb{E}\left[\max_{i\in[k]} \left\|\nabla \hat{F}(\hat{x}_i) - \nabla F(\hat{x}_i)\right\|^2\right] \le \sum_{i=1}^{k} \mathbb{E}\left[\left\|\nabla \hat{F}(\hat{x}_i) - \nabla F(\hat{x}_i)\right\|^2\right] \le \frac{k\sigma^2}{km} = \frac{\sigma^2}{m}. \tag{69}$$

Condition on the event that at least one of the ERMs is an ϵ-stationary point of F and denote one of those ERMs as $\hat{x}_{i^*}$. Denote this event E. Let $\hat{i} \in \arg\min_i \left\|\nabla \hat{F}(\hat{x}_i)\right\|$ where we abuse notation and say $\left\|\nabla \hat{F}(\hat{x}_i)\right\| := \inf_{g\in\partial \hat{F}(\hat{x}_i)} |g|$ for cases where $\hat{F}$ is not differentiable at $\hat{x}_i$. Then

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla F(\hat{x}_{\hat{i}})\right\| \middle| E\right] &\le \mathbb{E}\left[\left\|\nabla \hat{F}(\hat{x}_{\hat{i}})\right\| \middle| E\right] + \mathbb{E}\left[\max_{i\in[k]} \left\|\nabla \hat{F}(\hat{x}_i) - \nabla F(\hat{x}_i)\right\| \middle| E\right] && (70)\\
&\le \mathbb{E}\left[\left\|\nabla \hat{F}(\hat{x}_{\hat{i}})\right\| \middle| E\right] + \sqrt{\frac{\sigma^2}{m}} && (71)\\
&\le \mathbb{E}\left[\left\|\nabla \hat{F}(\hat{x}_{i^*})\right\| \middle| E\right] + \sqrt{\frac{\sigma^2}{m}} && (72)\\
&\le \mathbb{E}[\|\nabla F(\hat{x}_{i^*})\| | E] + \mathbb{E}\left[\max_{i\in[k]} \left\|\nabla \hat{F}(\hat{x}_i) - \nabla F(\hat{x}_i)\right\| \middle| E\right] + \sqrt{\frac{\sigma^2}{m}} && (73)\\
&\le \epsilon + 2\sqrt{\frac{\sigma^2}{m}}. && (74)
\end{aligned}$$

The event that one of the ERMs is an ϵ-stationary point happens with probability at least $1 - \left(\frac{2\sigma^2}{m\epsilon^2}\right)^k$. Choosing $m = \Omega\left(\frac{\sigma^2}{\epsilon^2}\right)$ and $k = \Omega\left(\log \frac{L}{\epsilon}\right)$ ensures $1 - \left(\frac{2\sigma^2}{m\epsilon^2}\right)^k \ge 1 - \frac{\epsilon}{L}$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla F(\hat{x}_{\hat{i}})\right\|\right] &= \mathbb{P}[E]\mathbb{E}\left[\left\|\nabla F(\hat{x}_{\hat{i}})\right\| \middle| E\right] + \mathbb{P}[E^c]\mathbb{E}\left[\left\|\nabla F(\hat{x}_{\hat{i}})\right\| \middle| E^c\right] && (75)\\
&\le \left(1 - \frac{\epsilon}{L}\right)\left(\epsilon + 2\sqrt{\frac{\sigma^2}{m}}\right) + \left(\frac{\epsilon}{L}\right)(L) && (76)\\
&\le O(\epsilon). && (77)
\end{aligned}$$

This entire algorithm required $O(km) = O\left(\frac{\sigma^2 \log\left(\frac{L}{\epsilon}\right)}{\epsilon^2}\right)$ samples in total, completing the proof. ■

Appendix C. Proofs for the Lower Bounds

Theorem 4 *For any $H, \Delta, R, \sigma > 0$, $\epsilon \leq \min\{\frac{HR}{8}, \sqrt{\frac{H\Delta}{8}}, \frac{\sigma}{4}\}$, the sample complexity to find a ϵ-stationary point*[11] *is lower bounded as*

$$\mathsf{m}_\epsilon(\mathcal{F}_{\mathrm{DB}}[H, \lambda = 0; R] \cap \mathcal{F}_{\mathrm{RB}}[H, \lambda = 0; \Delta], \mathcal{O}_f^\sigma) \geq \Omega\left(\frac{\sigma^2}{\epsilon^2}\right).$$

Proof For a constant $b \in \mathbb{R}$ to be chosen later, let

$$f(x; z) = \sigma \langle x, z \rangle + \frac{b}{2} \|x\|^2. \tag{78}$$

The distribution $\mathcal{D}$ of the random variable z is the uniform distribution over $\{z_1, \ldots, z_m\}$ where the vectors $z_i \in \mathbb{R}^d$ are orthonormal ($d \geq m$). Therefore,

$$F(x) = \mathbb{E}[f(x; z)] = \sigma \left\langle x, \frac{1}{m} \sum_{i=1}^m z_i \right\rangle + \frac{b}{2} \|x\|^2. \tag{79}$$

This function is clearly convex, b-smooth, and attains its unique minimum at $x^* = -\frac{\sigma}{bm} \sum_{i=1}^m z_i$ which has norm $\|x^*\|^2 = \frac{\sigma^2}{b^2 m}$, so choosing $b \geq \frac{\sigma}{R\sqrt{m}}$ ensures $\|x^*\|^2 \leq R^2$. Furthermore, $F(0) - F(x^*) = \frac{\sigma^2}{2bm}$, so choosing $b \geq \frac{\sigma^2}{2\Delta m}$ ensures $F(0) - F(x^*) \leq \Delta$. Choosing $b = \max\{\frac{\sigma}{R\sqrt{m}}, \frac{\sigma^2}{2\Delta m}\}$ ensures both simultaneously. Finally, $\mathbb{E}\|\nabla f(x; z) - \nabla F(x)\| = \frac{1}{m} \sum_{i=1}^m \left\|\sigma z_i - \frac{\sigma}{m} \sum_{j=1}^m z_j\right\|^2 = \sigma^2 \left(1 - \frac{1}{m}\right) \leq \sigma^2$. Therefore, $F \in \mathcal{F}_{\mathrm{DB}}[H, \lambda = 0; R] \cap \mathcal{F}_{\mathrm{RB}}[H, \lambda = 0; \Delta]$ and $f, \mathcal{D}$ properly define a $\mathcal{O}_{\nabla f}^\sigma$. Suppose, for now, that x is a point such that $\langle x, v_i \rangle \geq -\frac{\sigma}{8bm}$ for all $i \geq m/2$. Then

$$\|\nabla F(x)\|^2 = \frac{\sigma^2}{m} + b^2 \|x\|^2 + \frac{2\sigma b}{m} \sum_{i=1}^m \langle x, v_i \rangle \tag{80}$$

$$\geq \frac{\sigma^2}{m} + b^2 \sum_{i<m/2} \langle x, v_i \rangle^2 + \frac{2\sigma b}{m} \sum_{i<m/2} \langle x, v_i \rangle - \frac{\sigma^2}{4m} \tag{81}$$

$$\geq \frac{3\sigma^2}{4m} + \min_{y \in \mathbb{R}} \frac{b^2 m}{2} y^2 + \sigma b y \tag{82}$$

$$= \frac{\sigma^2}{4m}. \tag{83}$$

Therefore, for all such vectors x, $\|\nabla F(x)\| \geq \frac{\sigma}{2\sqrt{m}}$. This holds for any $b \geq 0$ and set $\{z_1, \ldots, z_m\}$. From here, we will argue that any randomized algorithm with access to less than $m/2$ samples from $\mathcal{D}$ is likely to output such an x. We consider a random function instance determined by drawing the orthonormal set $\{z_1, \ldots, z_m\}$ uniformly at random from the set of orthonormal vectors in $\mathbb{R}^d$. We will argue that with moderate probability over the randomness in the algorithm and in the draw of $z_1, \ldots, z_m$, the output of the algorithm has small inner product with $z_{m/2}, \ldots, z_m$. This approach closely resembles previous work (Woodworth and Srebro, 2016, Lemma 7).

Less than $m/2$ samples fix less than $m/2$ of the vectors z_i; assume w.l.o.g. that the algorithm's sample $S = \{z_1, \ldots, z_{m/2-1}\}$. The vectors z_i are a uniformly random orthonormal set, therefore

11. This lower bound applies both to deterministic and randomized optimization algorithms.

for any $i \geq m/2$, $z_i|S$ is distributed uniformly on the $(d - m/2 + 1)$-dimensional unit sphere in the subspace orthogonal to span$\left(z_1, \ldots, z_{m/2-1}\right)$. Let $\hat{x}$ be the output of any randomized algorithm whose input is S. If $\|\hat{x}\| \geq \frac{2\sigma}{b\sqrt{m}}$ then it is easily confirmed that $\|\nabla F(\hat{x})\| \geq \frac{\sigma^2}{m}$. Otherwise, we analyze

$$\mathbb{P}\left[\langle \hat{x}, v_i \rangle < -\frac{\sigma}{8bm} \;\middle|\; S, \hat{x}\right] \leq \mathbb{P}\left[|\langle \hat{x}, v_i \rangle| \geq \frac{\sigma}{8bm} \;\middle|\; S, \hat{x}\right] \tag{84}$$

$$\leq \mathbb{P}\left[\frac{2\sigma}{b\sqrt{m}} \left|\left\langle \frac{\hat{x}}{\|\hat{x}\|}, v_i \right\rangle\right| \geq \frac{\sigma}{8bm} \;\middle|\; S, \hat{x}\right] \tag{85}$$

$$= \mathbb{P}\left[\left|\left\langle \frac{\hat{x}}{\|\hat{x}\|}, v_i \right\rangle\right| \geq \frac{1}{16\sqrt{m}} \;\middle|\; S, \hat{x}\right]. \tag{86}$$

This probability only increases if we assume that $\hat{x}$ is orthogonal to span$(z_1, \ldots, z_{m/2-1})$, in which case we are considering the inner product between a fixed unit vector and a uniformly random unit vector. The probability of the inner product being large is proportional to the surface area of the "cap" of a unit sphere in $(d - m/2 + 1)$-dimensions lying above and below circles of radius $\sqrt{1 - \frac{1}{256m}}$. These end caps, in total, have surface area less than that of a sphere with that same radius. Therefore,

$$\mathbb{P}\left[\langle \hat{x}, v_i \rangle < -\frac{\sigma}{8bm} \;\middle|\; S, \hat{x}\right] \leq \sqrt{\left(1 - \frac{1}{256m}\right)^{d - \frac{m}{2}}} \tag{87}$$

$$= \left(1 - \frac{\frac{d}{512m} - \frac{1}{1024}}{\frac{d}{2} - \frac{m}{4}}\right)^{\frac{d}{2} - \frac{m}{4}} \tag{88}$$

$$\leq \exp\left(\frac{1}{1024} - \frac{d}{512m}\right). \tag{89}$$

This did not require anything but the norm of $\hat{x}$ being small, so for $d \geq \frac{m}{2} + 512m\log(2m)$, this ensures that

$$\mathbb{P}\left[\langle \hat{x}, v_i \rangle < -\frac{\sigma}{8bm} \;\middle|\; S, \|\hat{x}\| < \frac{2\sigma}{b\sqrt{m}}\right] \leq \frac{1}{2m}. \tag{90}$$

A union bound ensures that either $\|\hat{x}\| \geq \frac{2\sigma}{b\sqrt{m}}$ or $\langle \hat{x}, v_i \rangle \geq -\frac{\sigma}{8bm}$ for all $i \geq m/2$ with probability at least $1/2$ over the randomness in the algorithm and draw of $z_1, \ldots, z_m$, and consequently, that $\mathbb{E}_{\hat{x}} \|\nabla F(\hat{x})\|^2 \geq \frac{\sigma^2}{8m}$. Setting $m = \left\lfloor \frac{\sigma^2}{8\epsilon^2} \right\rfloor$ ensures this is at least ϵ. For this m, $b = \max\{\frac{\sigma}{R\sqrt{m}}, \frac{\sigma^2}{2\Delta m}\} \leq \frac{4\epsilon}{R} + \frac{4\epsilon^2}{\Delta}$ which must be less than H, so this lower bound applies for $\epsilon \leq \min\{\frac{HR}{8}, \sqrt{\frac{H\Delta}{8}}\}$. ■

Theorem 11 *For any $H, R, \sigma > 0$ and any $\epsilon \in (0, \sigma/2)$, there exists a $F : \mathbb{R} \to \mathbb{R} \in \mathcal{F}_{\text{DB}}[H, \lambda = 0; R]$ and a $\mathcal{O}^{\sigma}_{\nabla f}$, such that for any algorithm interacting with the stochastic first-order oracle, and returning an ϵ-approximate stationary point with some fixed constant probability, the expected number of queries is at least $\Omega\left(\frac{\sigma^2}{\epsilon^2} \cdot \log\left(\frac{HR}{\epsilon}\right)\right)$. Moreover, a similar lower bound of $\Omega\left(\frac{\sigma^2}{\epsilon^2} \cdot \log\left(\frac{H\Delta}{\epsilon^2}\right)\right)$ holds if the radius constraint R is replaced by a suboptimality constraint Δ.*

Proof We prove the lower bound by reduction from the noisy binary search (NBS) problem: In this classical problem, we have N sorted elements $\{a_1, \ldots, a_N\}$, and we wish to insert a new element

e using only queries of the form "is $e > a_j$?" for some j. Rather than getting the true answer, an independent coin is flipped and we get the correct answer only with probability $\frac{1}{2} + p$ for some fixed parameter p. Moreover, let j^* be the unique index such that $a_{j^*} < e < a_{j^*+1}$[12]. It is well-known (see for example Feige et al. (1994); Karp and Kleinberg (2007)) that in order to identify j^* with any fixed constant probability, at least $\Omega(\log(N)/p^2)$ queries are required.

Let us first consider the case where the radius constraint R is fixed. We will construct a convex stochastic optimization problem with the given parameters, such that if there is an algorithm solving it (with constant probability) after T local stochastic oracle queries, then it can be used to solve an NBS problem (with the same probability) using $2T$ queries, where $p = \epsilon/\sigma$ and[13] $N = HR/4\epsilon$. Employing the lower bound above for NBS, this immediately implies the $\Omega\left(\frac{\sigma^2}{\epsilon^2} \cdot \log\left(\frac{HR}{\epsilon}\right)\right)$ lower bound in our theorem.

To describe the reduction, let us first restate the NBS problem in a slightly different manner. For a fixed query budget T, let Z be a $T \times N$ matrix, with entries in $\{-1, +1\}$ drawn independently according to the following distribution:

$$\Pr(Z_{t,j} = 1) = \begin{cases} \frac{1}{2} - p & j \le j^* \\ \frac{1}{2} + p & j > j* \end{cases} .$$

Each $Z_{t,j}$ can be considered as the noisy answer provided in the NBS problem to the t-th query, of the form "is $e > a_j$" (where -1 corresponds to "true" and 1 corresponds to "false"). Thus, an algorithm for the NBS problem can be seen as an algorithm which can query T entries from the matrix Z (one query from each row), and needs to find j^* based on this information. Moreover, it is easy to see that the NBS lower bound also holds for an algorithm which can query *any* T entries from the matrix: Since the entries are independent, this does not provide additional information, and can only "waste" queries if the algorithm queries the same entry twice.

We now turn to the reduction. Given an NBS problem on $N = HR/4\epsilon$ elements with $p = \epsilon/\sigma$ and a randomly-drawn matrix Z, we first divide the interval $[0, R]$ into N equal sub-intervals of length R/N each, and w.l.o.g. identify each element a_j with the smallest point in the interval. Then, for every (statistically independent) row Z_t of Z, we define a function $f(x, Z_t)$ on $\mathbb{R}$ by $f(0, Z_t) = 0$, and the rest is defined via its derivative as follows:

$$f'(x, Z_t) = \begin{cases} -2\epsilon & x < 0 \\ 2\epsilon & x \ge R \\ \frac{x-a_j}{R/N}\sigma Z_{t,j+1} + \left(1 - \frac{x-a_j}{R/N}\right)\sigma Z_{t,j} & x \in [a_j, a_{j+1}) \text{ for some } j < N \end{cases} .$$

Note that by construction $\frac{x-a_j}{R/N} \in [0, 1]$ and $Z_{t,j} \in \{-1, +1\}$, so $|f'(x, Z_t)| \le \max\{2\epsilon, \sigma\} \le \sigma$. Moreover, since the expected value of $\sigma Z_{t,j}$ is $\sigma \cdot (-2p) = -2\epsilon$ if $j \le j^*$, and $\sigma \cdot 2p = 2\epsilon$ if $j > j^*$, it is easily verified that

$$\mathbb{E}_{Z_t}[f'(x, Z_t)] = \begin{cases} -2\epsilon & x < a_{j^*} \\ 2\epsilon & x \ge a_{j^*+1} \\ -2\epsilon + 4\epsilon\frac{x-a_j}{R/N} & x \in [a_{j^*}, a_{j^*+1}) \end{cases} .$$

12. This is w.l.o.g., since if $e < a_1$ or $e > a_N$, we can just add two dummy elements smaller and larger than all other elements and e, increasing N by at most 2, hence not affecting the lower bound.
13. For simplicity we assume that $HR/4\epsilon$ is a whole number – otherwise, it can be rounded and this will only affect constant factors in the lower bound.

Noting that $4\epsilon\frac{x-a_j}{R/N} = H(x-a_j) \in [0,4\epsilon]$ in the above, we get that $F(x) := \mathbb{E}_{Z_t}[f(x,Z_t)]$ is a convex function with H-Lipschitz gradients, with a unique minimum at some $x : |x| < R$, and with $|F'(x)| \le \epsilon$ only when $x \in [a_{j^*}, a_{j^*+1})$. Overall, we get a valid convex stochastic optimization problem (with parameters H, R, σ as required), such that if we can identify x such that $|F'(x)| \le \epsilon$, then we can uniquely identify j^*. Moreover, given an algorithm to the optimization problem, we can simulate a query to a local stochastic oracle (specifying an iterate t and a point x) by returning $f'(x, Z_t)$ as defined above, which requires querying at most 2 entries $Z_{t,j}$ and $Z_{t,j+1}$ from the matrix Z. So, given an oracle query budget T to the stochastic problem, we can simulate it with at most $2T$ queries to the matrix Z in the NBS problem.

To complete the proof of the theorem, it remains to handle the case where there is a suboptimality constraint Δ rather than a radius constraint R. To that end, we simply use the same construction as above, with $R = \frac{\Delta}{2\epsilon}$. Since the derivative of F has magnitude at most 2ϵ, and its global minimum satisfies $|x^*| \le R$, it follows that $F(0) - F(x^*t) \le 2\epsilon R = \Delta$. Plugging in $R = \frac{\Delta}{2\epsilon}$ in the lower bound, the result follows. ■

Theorem 2 *For any $H, \Delta, R, \sigma > 0$, any $\epsilon \le \frac{HR}{8}$, the stochastic first-order oracle complexity for range-bounded functions is lower bounded as*

$$\mathsf{m}_\epsilon(\mathcal{F}_{\mathrm{DB}}[H, \lambda = 0; R], \mathcal{O}^\sigma_{\nabla f}) \ge \Omega\left(\sqrt{\frac{HR}{\epsilon}} + \frac{\sigma^2}{\epsilon^2}\log\left(\frac{HR}{\epsilon}\right)\right).$$

For any $\epsilon \le \sqrt{\frac{H\Delta}{8}}$, the stochastic first-order complexity for domain-bounded functions is lower bounded as

$$\mathsf{m}_\epsilon(\mathcal{F}_{\mathrm{RB}}[H, \lambda = 0; \Delta], \mathcal{O}^\sigma_{\nabla f}) \ge \Omega\left(\frac{\sqrt{H\Delta}}{\epsilon} + \frac{\sigma^2}{\epsilon^2}\log\left(\frac{H\Delta}{\epsilon^2}\right)\right).$$

Proof By Theorem 11, $\Omega(\sigma^2/\epsilon^2)\log(HR/\epsilon)$ and $\Omega(\sigma^2/\epsilon^2)\log\left(H\Delta/\epsilon^2\right)$ oracle calls (samples) are needed to find an ϵ-stationary point. Furthermore, a deterministic first-order oracle is a special case of a stochastic first-order oracle (corresponding to the case $\sigma = 0$). Therefore, lower bounds for deterministic first-order optimization apply also to stochastic first-order optimization. Therefore, the lower bound of Carmon et al. (2017b) (Theorem 1) completes the proof. ■