How Free is Parameter-Free Stochastic Optimization?

Amit Attia [1]   **Tomer Koren** [1 2]

Abstract

We study the problem of parameter-free stochastic optimization, inquiring whether, and under what conditions, do fully parameter-free methods exist: these are methods that achieve convergence rates competitive with optimally tuned methods, without requiring significant knowledge of the true problem parameters. Existing parameter-free methods can only be considered "partially" parameter-free, as they require some non-trivial knowledge of the true problem parameters, such as a bound on the stochastic gradient norms, a bound on the distance to a minimizer, etc. In the non-convex setting, we demonstrate that a simple hyperparameter search technique results in a fully parameter-free method that outperforms more sophisticated state-of-the-art algorithms. We also provide a similar result in the convex setting with access to noisy function values under mild noise assumptions. Finally, assuming only access to stochastic gradients, we establish a lower bound that renders fully parameter-free stochastic convex optimization infeasible, and provide a method which is (partially) parameter-free up to the limit indicated by our lower bound.

1. Introduction

Stochastic first-order optimization is a cornerstone of modern machine learning, with stochastic gradient descent (SGD, Robbins & Monro, 1951) as the go-to method for addressing statistical learning problems. The tuning of algorithmic parameters, particularly those associated with SGD such as the step-size, proves to be a challenging task, both theoretically and in practice (e.g., Bottou, 2012; Schaul et al., 2013), especially when the problem parameters are unknown. The present paper focuses on theoretical aspects of this challenge.

Addressing the challenge, a variety of methods emerged over the years, with the goal of minimizing knowledge about problem parameters required for tuning, while maintaining performance competitive with carefully tuned algorithms. These include the so-called adaptive methods, such as AdaGrad and Adam (e.g., Duchi et al., 2011; Kingma & Ba, 2015) and their recent theoretical advancements (Reddi et al., 2018; Tran et al., 2019; Kavis et al., 2019; Alacaoglu et al., 2020; Faw et al., 2022; Kavis et al., 2022; Attia & Koren, 2023; Liu et al., 2023), that self-tune learning rates based on gradient statistics; parameter-free methods (e.g., Chaudhuri et al., 2009; Streeter & McMahan, 2012; Luo & Schapire, 2015; Orabona & Pál, 2021; Carmon & Hinder, 2022) that focus primarily on automatically adapting to the complexity of (i.e., distance to) an optimal solution; and advanced techniques from the literature on online learning (Orabona & Pál, 2016; Cutkosky & Orabona, 2018) that can be used in a stochastic optimization setting via an online-to-batch reduction. More recently, several works have achieved advancements in improving practical performance, narrowing the gap to finely-tuned methods (Orabona & Tommasi, 2017; Chen et al., 2020; Ivgi et al., 2023; Defazio & Mishchenko, 2023; Mishchenko & Defazio, 2023).

While adaptive and parameter-free methods enjoy strong theoretical guarantees, each of the aforementioned approaches still demands some level of non-trivial knowledge regarding problem parameters. This includes knowledge about the smoothness parameter and the initial suboptimality for non-convex adaptive and self-tuning methods, or a bound on either the stochastic gradient norms or on the distance to an optimal solution for parameter-free convex (including online) optimization algorithms. To the best of our knowledge, in the absence of such non-trivial assumptions, no method in the literature is fully parameter-free, in either the non-convex or convex, smooth or non-smooth cases.

This paper explores the following question: in what stochastic optimization scenarios, and under what conditions, do fully parameter-free optimization algorithms exist? In this context, we note that direct hyperparameter tuning, which parameter-free self-tuning methods aim to eschew, could be considered a valid approach for designing parameter-free algorithms. Consequently, we also examine the question: can self-tuning optimization algorithms actually outperform direct tuning methods based on simple hyperparam-

---

[1]Blavatnik School of Computer Science, Tel Aviv University [2]Google Research Tel Aviv. Correspondence to: Amit Attia <amitattia@mail.tau.ac.il>.

eter search?

We begin our inquiry in the general setting of non-convex (smooth) optimization. We observe that a simple hyperparameter search technique for tuning of fixed stepsize SGD results in a fully parameter-free method with convergence rate that matches *optimally tuned* SGD up to poly-logarithmic factors. While extremely straightforward, this approach outperforms the state-of-the-art bounds for adaptive methods in this context (Faw et al., 2022; Kavis et al., 2022; Attia & Koren, 2023; Liu et al., 2023) that require non-trivial knowledge on the smoothness parameter and the initial suboptimality for optimal tuning.

Moving on to the convex optimization setting, we observe that a similar hyperparameter search technique can be used to design a simple and efficient parameter-free algorithm, provided access to noisy function value queries to the objective. Under very mild assumptions on the scale of noise in the value queries, the obtained algorithm is fully parameter-free with convergence rate matching that of perfectly tuned SGD (up to poly-log factors), in both the smooth and non-smooth convex cases. This algorithm thus outperforms existing results in the literature, albeit under an additional assumption of a reasonably-behaved stochastic function value oracle. Indeed, an inspection of previous studies in this context reveals that they neglect access to function values and only rely on stochastic gradient queries (even though function value access is often readily available in practical applications of these methods).

Our last area of focus is therefore the convex optimization setting with access solely to stochastic gradients (and no access whatsoever to function values), which emerges as the only general stochastic optimization setting where achieving parameter-freeness is potentially non-trivial. Indeed, and perhaps somewhat surprisingly, we identify in this setting a limitation to obtaining a fully parameter-free algorithm: we establish a lower bound showing that any optimization method cannot be parameter-free unless it is provided with either the gradient-noise magnitude or the distance to a minimizer, each up to a (multiplicative) factor of $O(\sqrt{T})$ where T is the number of optimization steps. The bound sheds light on why prior methods require non-trivial knowledge of problem parameters, yet still leaves room for improvement: state-of-the-art results in this context (e.g., Carmon & Hinder, 2022; Ivgi et al., 2023) assume knowledge on the magnitude of the stochastic gradients, whereas our lower bound only necessitates knowledge on the magnitude of the *noise component* in the gradient estimates.

Our final contribution complements the lower bound and introduces a parameter-free method in the convex non-smooth (Lipschitz) setting, achieving the same rate as of optimally-tuned SGD, up to poly-log factors and an unavoidable term prescribed by our lower bound. When given a bound on

the noise in the stochastic gradients accurate up to a $O(\sqrt{T})$ factor, our method becomes fully parameter-free. The same method achieves a similar parameter-free guarantee in the setting of convex smooth optimization, nearly matching the rate of tuned SGD in this case.

1.1. Summary of contributions

To summarize our results in some more detail, let us first specify the setup of parameter-free optimization slightly more concretely. In all cases, we consider unconstrained optimization of an objective function $f : \mathbb{R}^d \to \mathbb{R}$. Rather than directly receiving the ground-true problem parameters (e.g., smoothness parameter, Lipschitz constant, etc.), the algorithms are provided with a range (lower and upper bounds) containing each parameter. We refer to a method as parameter-free if its convergence rate matches a benchmark rate (e.g., that of optimally-tuned SGD), up to poly-logarithmic factors in the ranges parameters (as well as in the number of steps T and the probability margin δ).

Non-convex setting: fully parameter-free algorithm. We give a fully parameter-free method with the same convergence rate as *tuned* SGD up to poly-log factors. Given an initialization $w_1 \in \mathbb{R}^d$, number of queries T, probability margin δ and a range $[\eta_{\min}, \eta_{\max}]$ containing the theoretically-tuned SGD stepsize (the latter can be computed from the problem parameter ranges), the algorithm produces $\overline{w} \in \mathbb{R}^d$ such that with probability at least $1 - \delta$,

$$\|\nabla f(\overline{w})\|^2 \le \left(\frac{\beta_\star F_\star + \sigma_\star^2}{T} + \frac{\sigma_\star \sqrt{\beta_\star F_\star}}{\sqrt{T}} \right) \text{poly} \log \left(\frac{\eta_{\max}}{\eta_{\min}}, \frac{1}{\delta} \right),$$

where $\beta_\star$ is the smoothness parameter, $F_\star = f(w_1) - \min f(w)$, and $\sigma_\star$ is a noise bound of the stochastic gradients. While the algorithm is based on simple observations, it is fully parameter-free and provides stronger guarantees with a simpler analysis compared to existing adaptive or self-tuning methods in this setting.

Convex setting: fully parameter-free algorithm with noisy function values. Moving to the convex case, we first consider the canonical setting with access to noisy function values. We devise a simple parameter-free method, that given w_1, T, δ and a range $[\eta_{\min}, \eta_{\max}]$ containing the theoretically-tuned SGD stepsize, produces $\overline{w} \in \mathbb{R}^d$ such that with probability $\ge 1 - \delta$,

$$f(\overline{w}) - f(w^\star) \le \left(\frac{D_\star \sqrt{G_\star^2 + \sigma_\star^2}}{\sqrt{T}} + \frac{\sigma_0}{\sqrt{T}} \right) \text{poly} \log \left(\frac{\eta_{\max}}{\eta_{\min}}, \frac{1}{\delta} \right),$$

where $G_\star$ is the Lipschitz constant, $D_\star = \|w_1 - w^\star\|$ ($w^\star$ is a minimizer of f), and σ_0 and $\sigma_\star$ are noise bounds of the function and gradient oracles respectively. Under very mild

2

assumptions on the function values noise σ_0, the method is fully parameter-free and achieves the same rate as *tuned* SGD up to poly-log factors.

Convex setting: impossibility without function values. Next we consider the convex setting without any access to function values. We establish that without further assumptions, no convex optimization method can be fully parameter-free while nearly matching the rate of *tuned* SGD. We show that for any $\alpha \in [1, \frac{3}{4}\sqrt{T}]$ and every T-queries deterministic algorithm receiving ranges $[\frac{1}{a\sqrt{T}}D_{\max}, D_{\max}]$, $[\frac{1}{a\sqrt{T}}\sigma_{\max}, \sigma_{\max}]$ and known $G_\star = \frac{\sigma_{\max}}{(2T-1)}$, there exist $D_\star$ and $\sigma_\star$ that belong to the ranges (resp.), a convex and $G_\star$-Lipschitz function f with a minimizer $w^\star$ such that $\|w^\star\| = D_\star$, and a $\sigma_\star$-bounded first-order oracle such that with constant probability, the output of the algorithm, $\overline{w}$, satisfy

$$f(\overline{w}) - f(w^\star) \geq \frac{D_\star(G_\star + \sigma_\star)\alpha}{6\sqrt{T}} \geq \frac{D_\star \sigma_{\max}}{6T}.$$

Thus, without non-trivial prior knowledge on the parameter ranges, no algorithm can match the performance of *tuned* SGD and must include a term linear in $\sigma_{\max}$.

Parameter-free algorithm for the convex non-smooth setting. Assuming only gradient access, we propose a method which requires knowledge of $D_{\min}, D_{\max}, G_{\max}$ and $\sigma_{\max}$ such that $D_{\min} \leq D_\star = \|w_1 - w^\star\| \leq D_{\max}$, $G_\star \leq G_{\max}$ and $\sigma_\star \leq \sigma_{\max}$ and produces $\overline{w}$ such that with probability at least $1 - \delta$,

$$f(\overline{w}) - f^\star \leq \widetilde{C}\left(\frac{D_\star(G_\star + \sigma_\star)}{\sqrt{T}} + \frac{D_\star \sigma_{\max}}{T}\right),$$

where $\widetilde{C} = \text{poly}\log(\frac{D_{\max}}{D_{\min}}, G_{\max}, \sigma_{\max}, \frac{T}{\delta})$. The method is parameter-free if $\sigma_{\max}$ is provided up to a tolerance of $O(\sqrt{T})$, in which case it achieves the same rate of convergence as *tuned* SGD up to logarithmic factors. This is the maximal tolerance possible in light of our lower bound, which shows that the additional $\sigma_{\max}/T$ term is unavoidable without further assumptions.

Parameter-free algorithm for the convex smooth setting. Assuming that the objective is $\beta_\star$-smooth rather than Lipschitz, the same convex optimization method with parameters $D_{\min}, D_{\max}, \beta_{\max}$ and $\sigma_{\max}$ such that $D_{\min} \leq D_\star \leq D_{\max}$, $\beta_\star \leq \beta_{\max}$ and $\sigma_\star \leq \sigma_{\max}$, produce $\overline{w} \in \mathbb{R}^d$ such that with probability at least $1 - \delta$,

$$f(\overline{w}) - f^\star \leq \widetilde{C}\left(\frac{\beta_\star D_\star^2}{T} + \frac{D_\star \sigma_\star}{\sqrt{T}} + \frac{D_\star \sigma_{\max}}{T}\right),$$

where $\widetilde{C} = \text{poly}\log(D_{\max}, \frac{1}{D_{\min}}, \beta_{\max}, \sigma_{\max}, \frac{T}{\delta})$. To the best of our knowledge, this is the first parameter-free method

for stochastic convex and smooth optimization (in the case where $\sigma_\star$ is provided up to $O(\sqrt{T})$ tolerance, which is again required to account for the unavoidable term prescribed by our lower bound).

1.2. Additional related work

Adaptive stochastic non-convex optimization. A long line of work focus on the analysis of SGD with AdaGrad-type stepsizes (Ward et al., 2019; Li & Orabona, 2019; Kavis et al., 2022; Faw et al., 2022; Attia & Koren, 2023; Liu et al., 2023), which we refer to as AdaSGD (also known as AdaGrad-norm, e.g., Ward et al. 2019; Faw et al. 2022). AdaSGD enjoys high-probability guarantees and rate interpolation for small noise (Attia & Koren, 2023; Liu et al., 2023). Compared to our tuning-based fully parameter-free method, the analysis of AdaSGD is cumbersome and suboptimal without knowing the smoothness and function suboptimality.

Parameter-free and adaptive stochastic convex optimization methods. Existing methods assume either known diameter or known stochastic gradient norm bound. Assuming a known diameter, (Kavis et al., 2019) presented a method with an (almost) optimal rate for convex optimization and an accelerated rate for convex smooth optimization. It is an open question whether parameter-free acceleration is possible. Given a a bound of the stochastic gradient norm, Carmon & Hinder (2022) provided the first method which achieve the optimal rate up to a double-logarithmic factor in the diameter range by utilizing an efficient bisection procedure to find a good SGD stepsize. Ivgi et al. (2023) introduced a parameter-free method which use dynamic stepsizes for SGD, and demonstrated strong practical performance. Both methods are parameter-free if the bound of the stochastic gradients norm is given up to a tolerance of $O(\sqrt{T})$, while our method and analysis require only such bound for the noise of the stochastic gradients; this distinction is crucial to our guarantee in the convex and smooth case, where the gradients are unbounded.

Parameter-free deterministic optimization methods. While our focus here is on parameter-free stochastic optimization, in the context of *deterministic* optimization several fully parameter-free methods were previously suggested. The adaptive Polyak method of Hazan & Kakade (2019) achieves almost the same rate as tuned gradient descent for convex Lipschitz and convex smooth problems. For convex smooth problems, Beck & Teboulle (2009) and Nesterov (2015) use line-search techniques to obtain the optimal (accelerated) rate. It is also worth mentioning the tuning-free methods of Defazio & Mishchenko (2023); Mishchenko & Defazio (2023), which are not fully parameter-free (they need knowledge of the Lipschitz constant) yet demonstrate

strong practical performance without tuning.

Parameter-free online convex optimization. In the setting of online convex optimization (OCO), several works concerned themselves with parameter-free (e.g., McMahan & Orabona, 2014; Orabona & Pál, 2016; Cutkosky & Orabona, 2018; Mhammedi & Koolen, 2020) and scale-free (e.g., Orabona & Pál, 2018; Mhammedi & Koolen, 2020) methods. The parameter-free OCO literature assume known gradient norm bound and achieve regret of $\widetilde{O}(\|u\|G_{\max}\sqrt{T\log\|u\|})$ where u, $G_{\max}$ and T are an arbitrary comparator, gradient norm bound and number of rounds respectively (McMahan & Orabona, 2014). Cutkosky & Boahen (2016; 2017) established lower bounds which rule out such regret bounds if both the comparator and gradient norm bounds are unknown. Sidestepping the lower bound, several works designed parameter-free methods (with unknown Lipschitz bound and comparator norm), suffering an additional $\widetilde{O}(\|u\|^3 G_{\max})$ term (Cutkosky, 2019; Mhammedi & Koolen, 2020). Considering online-to-batch conversions, online parameter-free methods with known gradient norm bound still require a non-trivial bound of the stochastic gradient norms, not only the noise bound. Jun & Orabona (2019) relaxed the bounded gradients assumption using stochastic gradients, but requires a known bound of the expected gradient norms instead. We note that the method of Cutkosky (2019) can use crude range bounds for a guarantee of $\widetilde{O}(D_\star G_\star/\sqrt{T} + D_{\max}G_\star/T^{3/4})$, which is parameter-free when $D_{\max}/D_{\min} = O(T^{1/4})$. Diameter tolerance is a promising future research direction, as most current studies focus on gradient norm/noise tolerance.

Classical analyses of stochastic gradient descent. Ghadimi & Lan (2013) were the first to examine stochastic gradient descent in the smooth non-convex setting and obtain tight convergence. They demonstrated that a *properly tuned* SGD with T steps achieves a rate of $O(1/T + \sigma/\sqrt{T})$ under the assumption of uniformly bounded variance σ^2; Arjevani et al. (2022) provided a tight lower bound. Assuming sub-Gaussian noise, Ghadimi & Lan (2013) provided a high-probability convergence rate of $\widetilde{O}(1/T + \sigma^2/T + \sigma/\sqrt{T})$ by amplifying the success probability using multiple runs of SGD, while Liu et al. (2023) established a similar rate for a single run of SGD. For the convex setting, Lan (2012) provided convergence guarantees for stochastic gradient descent in composite stochastic optimization, establishing a convergence rate of $O((G + \sigma)/\sqrt{T})$ for convex and G-Lipschitz objectives and $O(1/T + \sigma/\sqrt{T})$ for convex smooth objectives, assuming uniformly bounded variance of σ^2. Classical non-asymptotic analyses of SGD and minimax lower bounds in stochastic optimization trace back to Nemirovskij & Yudin (1983).

Addendum: concurrent work on parameter-free stochastic optimization. Shortly after the present manuscript first appeared on arXiv (Attia & Koren, 2024a), other works appeared which explore similar questions in parameter-free stochastic optimization. Carmon & Hinder (2024) study the "price of adaptivity" to uncertainty in problem parameters in stochastic convex problems, with focus on characterizing the tight penalty (i.e., poly-logarithmic factors in the bound) one must pay for being parameter-free. They provide several lower bounds for different uncertainty scenarios, including stochastic gradients with bounded norms and stochastic gradients with a bounded second moment, as well as a result similar to our Theorem 6 that quantifies the price of uncertainty in both diameter and stochastic gradient norms. In contrast, our primary focus is on studying conditions under which parameter-free algorithms are possible in various optimization scenarios, modulo poly-logarithmic factors in the convergence bounds.

In an independent work, Khaled & Jin (2024) study parameter-free optimization in the non-convex and convex settings and establish a set of results closely related to ours. For non-convex smooth optimization, they provide an upper bound similar to our Theorem 1 (they also give a lower bound for the in-expectation rate of SGD, showing that it is unattainable by parameter-free methods). For convex optimization, they prove that full parameter-freeness is in general impossible, while showing a convergence result that inversely depends (polynomially) on a certain SNR parameter, that for favorable noise distributions is bounded away from zero, in which case the algorithm is parameter-free. Comparing to our lower bound in Theorem 6, we provide a detailed characterization of the cost of being parameter-free in this setting, showing how convergence of parameter-free methods degrades with the degree of uncertainty in the gradient noise parameter. In terms of upper bounds, we establish feasibility of parameter-free convex optimization with no dependence on SNR (that can be arbitrarily small even for bounded noise distributions), with rates matching our lower bound up to poly-log factors. In addition, we establish an analogous upper upper bound for convex smooth problems, which is, to our knowledge, the first parameter-free result in this setting.

2. Preliminaries

Notation. Throughout, we use log to denote the base 2 logarithm and $\log_+(\cdot) \triangleq 1 + \log(\cdot)$. $\widetilde{O}$ and $\widehat{O}$ are asymptotic notations that suppress logarithmic and doubly-logarithmic factors respectively (in addition to constant factors).

2.1. Optimization setup

Our focus is unconstrained stochastic optimization in d-dimensional Euclidean space $\mathbb{R}^d$ with the ℓ_2 norm, denoted

||·||. We assume a stochastic first-order access to a differentiable objective function $f : \mathbb{R}^d \to \mathbb{R}$ through a randomized oracle O. We will consider two variants of such access, common to the literature on stochastic optimization:

(i) **Stochastic first-order oracle:** The oracle O generates, given any $w \in \mathbb{R}^d$, unbiased value and gradient samples for f at the point w, namely $O(w) = (\widetilde{f}(w), \widetilde{g}(w))$ such that for all $w \in \mathbb{R}^d$, $\mathbb{E}[\widetilde{f}(w)] = f(w)$ and $\mathbb{E}[\widetilde{g}(w)] = \nabla f(w)$. We will make the standard assumption that $\widetilde{f}$ and $\widetilde{g}$ have uniformly bounded noise,[1] that is, for some parameters $\sigma_0 > 0$ and $\sigma_\star > 0$,

$$\forall \, w \in \mathbb{R}^d : \quad \Pr(|\widetilde{f}(w) - f(w)| \leq \sigma_0) = 1$$
$$\text{and} \quad \Pr(\|\widetilde{g}(w) - \nabla f(w)\| \leq \sigma_\star) = 1.$$

(ii) **Stochastic gradient oracle:** The oracle O only generates a stochastic gradient sample with $\sigma_\star$-bounded noise for f given a point w; namely $O(w) = (\widetilde{g}(w))$ such that for all $w \in \mathbb{R}^d$, $\mathbb{E}[\widetilde{g}(w)] = \nabla f(w)$ and

$$\Pr(\|\widetilde{g}(w) - \nabla f(w)\| \leq \sigma_\star) = 1.$$

In this framework, we distinguish between several different optimization scenarios:

(i) **Non-convex setting.** In this case we will assume the objective is not necessarily convex but is $\beta_\star$-smooth.[2] We further assume that the function is lower bounded by some $f^\star \in \mathbb{R}$ and denote $F_\star \triangleq f(w_1) - f^\star$ for a given reference point $w_1 \in \mathbb{R}^d$. The goal in the setting is, given T queries to the stochastic oracle and $\delta > 0$, is to find a point $w \in \mathbb{R}^d$ such that $\|\nabla f(w)\|^2 \leq \epsilon$ with probability at least $1 - \delta$, for ϵ as small as possible (as a function of the problem parameters $\beta_\star, F_\star, \sigma_\star, \sigma_0$ and T, δ).

(ii) **Convex non-smooth setting.** Here we assume f is convex, $G_\star$-Lipschitz and admits a minimizer $w^\star \in \arg\min_{w \in \mathbb{R}^d} f(w)$ with value $f^\star \triangleq f(w^\star)$. We denote $D_\star \triangleq \|w_1 - w^\star\|$ for a reference point $w_1 \in \mathbb{R}^d$. The goal is, given T queries to the stochastic oracle and $\delta > 0$, is to produce a point $w \in \mathbb{R}^d$ such that $f(w) - f^\star \leq \epsilon$ with probability at least $1 - \delta$, for ϵ as small as possible (in terms of the problem parameters $G_\star, D_\star, \sigma_\star, \sigma_0$ and T, δ).

(iii) **Convex and smooth setting.** This scenario is identical to the former, but here we assume that f is $\beta_\star$-smooth rather than $G_\star$-Lipschitz. The convergence rate ϵ is then given in terms of the parameters $\beta_\star, D_\star, \sigma_\star, \sigma_0$ and T, δ.

2.2. Parameter-free algorithmic setup

Since our focus is on parameter-free optimization, it is crucial to specify what algorithms in this setting are allowed to receive as input. First, we will always assume that the number of steps T, the failure probability δ and the reference point w_1 are given as inputs.[3] The remaining problem parameters are not known ahead of time and thus are not received as inputs; however, we will assume that algorithms do receive, for each ground-truth parameter, a crude range (i.e., lower and upper bounds) in which it lies. Namely, for any of the relevant parameters among $\beta_\star, F_\star, G_\star, D_\star, \sigma_\star$, algorithms receive respective ranges such that $\beta_\star \in [\beta_{\min}, \beta_{\max}]$, $F_\star \in [F_{\min}, F_{\max}]$, $G_\star \in [G_{\min}, G_{\max}]$, $D_\star \in [D_{\min}, D_{\max}]$, $\sigma_\star \in [\sigma_{\min}, \sigma_{\max}]$ (a range for the parameter σ_0 is not used in any of our methods). We refer to a method as parameter-free, with respect to a well-tuned benchmark algorithm, if its convergence rate matches that of the benchmark, up to poly-logarithmic factors in the range parameters as well as in T and $1/\delta$. (We do not include a more formal definition here, since it will not be required for the statement of any of our results.)

2.3. Tuned benchmarks

Our benchmark rates, for both convex and non-convex optimization, are those of *tuned* (fixed stepsize, possibly projected) Stochastic Gradient Descent (SGD). Starting at $w_1 \in \mathbb{R}^d$, the update step of SGD is $w_{t+1} = w_t - \eta g_t$, where η is the stepsize parameter, $g_t = \widetilde{g}(w_t)$ is the stochastic gradient at step t, evaluated at w_t. The update of projected SGD is $w_{t+1} = \Pi_\Omega(w_t - \eta g_t)$, where $\Pi_\Omega(\cdot)$ is the Euclidean projection onto a convex domain $\Omega \subseteq \mathbb{R}^d$. With a tuned stepsize parameter, this method achieves the best known rate in the non-convex setting and the optimal rate in convex non-smooth setting (accelerated methods achieve a slightly better rate in the convex and smooth setting). The benchmark rates of SGD in the optimization scenarios described above are detailed in Table 1. Note that SGD is not parameter-free with respect to *tuned* SGD, unless the range parameters are such that the ground-true parameters are known up to a constant factor.

---

[1]The bounded noise assumption is made here for simplicity and can be relaxed to a sub-Gaussian assumption on the noise; see, for example, Attia & Koren (2024b) or Appendix C of Attia & Koren (2023) for further details.

[2]A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be β-smooth if $\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$ for all $x, y \in \mathbb{R}^d$. In particular, this implies that $|f(y) - f(w) - \nabla f(x) \cdot (y - x)| \leq \frac{\beta}{2}\|y - x\|^2$ for all $x, y \in \mathbb{R}^d$.

[3]A standard doubling trick can be used to handle an unknown T while keeping the same rate of convergence up to log-factors.

Table 1. Benchmark high-probability rates of tuned SGD, ignoring log-factors.

SETTING	BENCHMARK RATE	
Non-convex	$\frac{\beta_\star F_\star}{T} + \frac{\sigma_\star^2}{T} + \sqrt{\frac{\beta_\star F_\star \sigma_\star^2}{T}}$	(Ghadimi & Lan, 2013)
Convex, Lipschitz	$\frac{D_\star(G_\star + \sigma_\star)}{\sqrt{T}}$	(Lan, 2012)
Convex, smooth	$\frac{\beta_\star D_\star^2}{T} + \frac{D_\star \sigma_\star}{\sqrt{T}}$	(Lan, 2012)

3. Parameter-Free Non-convex Optimization

We begin with the general non-convex smooth case, providing a fully parameter-free method which achieves the same rate of *tuned* SGD up to poly-logarithmic factors. Our approach is based on a simple observation that has also been used in the two-phase SGD approach of Ghadimi & Lan (2013). Our parameter-free method performs grid search over multiple SGD stepsizes and for each sequence of SGD iterations randomly samples a small portion of the iterations. Then the method picks the solution with the minimal approximate gradient norm based on stochastic gradients. For the full algorithm statement see Appendix A. Here we only states its guarantees.

Theorem 1. *Assume that f is $\beta_\star$-smooth and lower bounded by some $f^\star$ and $\widetilde{g}$ is a $\sigma_\star$-bounded unbiased gradient oracle of f. Let $\eta_{\min}, \eta_{\max} > 0$ such that*

$$\eta_{\min} \leq \eta_\star \triangleq \min\left\{\frac{1}{2\beta_\star}, \sqrt{\frac{F_\star}{\beta_\star \sigma_\star^2 T}}\right\} \leq \eta_{\max}$$

where $F_\star = f(w_1) - f^\star$. Then for any $\delta \in (0, \frac{1}{3})$, given $w_1, T, \delta, \eta_{\min}$ and $\eta_{\max}$, Algorithm 2 performs T gradient queries and produce $\overline{w}$ such that with probability $\geq 1 - \delta$,

$$\|\nabla f(\overline{w})\|^2 = \widehat{O}\left(\frac{\log_+(\frac{\eta_{\max}}{\eta_{\min}})\log_+(\frac{1}{\delta})(\beta_\star F_\star + \sigma_\star^2 \log\frac{1}{\delta})}{T}\right.$$
$$\left. + \frac{\sigma_\star\sqrt{\beta_\star F_\star \log_+(\frac{\eta_{\max}}{\eta_{\min}})\log_+(\frac{1}{\delta})}}{\sqrt{T}}\right).$$

First, we note that the rate of Algorithm 2 match the rate of *fully-tuned* SGD up to logarithmic factors. Note that valid $\eta_{\min}, \eta_{\max}$ can be determined using ranges of the problem parameters (details at Appendix A). Additionally, the method produce a single point instead of the obscure average norm guarantee of SGD, a product of comparing multiple iterations of SGD. Comparing the result to recent advances in the analysis of adaptive SGD with AdaGrad-like stepsizes (Faw et al., 2022; Attia & Koren, 2023; Liu et al., 2023), Algorithm 2 has only logarithmic dependency in the specification of the parameters $\eta_{\min}$ and $\eta_{\max}$ while adaptive SGD require tight knowledge of $F_\star/\beta_\star$ for tuning or suffer a suboptimal polynomial dependency on problem parameters. In addition, the proof technique of Theorem 1 is simple and

straightforward compared to the cumbersome analysis of adaptive SGD with AdaGrad-like stepsizes.

Due to the simplicity of the analysis, we defer the full proof to Appendix A and provide here the general guidelines. The proof uses standard techniques (which are also used in the original analysis of Ghadimi & Lan 2013). As $\eta_\star \in [\eta_{\min}, \eta_{\max}]$, the algorithm's grid search execute SGD with some $\eta_\star' \approx \eta_\star$, and with high probability, obtains (among other candidates) at least one point with a similar guarantee to *tuned* SGD. A concentration of the stochastic gradient norms ensures that we can select a good candidate with high probability.

4. Parameter-Free Convex Optimization

We proceed to consider convex problems, where the objective is either convex Lipschitz or convex smooth. The first result in this section assumes access via a first-order oracle (which includes access to noisy function values), while later results deal with the case of stochastic gradient oracle access.

4.1. Optimization with a stochastic first-order oracle

In this setting we assume access using a first-order oracle $O(w) = (\widetilde{f}(w), \widetilde{g}(w))$ with bounded noise. Stochastic function value access (with a reasonable level of noise variance) is often a valid and natural assumption since (i) it is often the case that stochastic gradients are obtained by random samples and those can be used for both function and gradient evaluation; and (ii) the practice of tuning using a validation set implicitly assume this exact assumption. We present a fully parameter-free method which utilize the function value oracle to perform model selection, similar to the use of the gradient oracle for model selection in Algorithm 2. The algorithm is detailed in Appendix B and following is its convergence result. Due to the similarities between the method and Algorithm 2, the proof is also deferred to Appendix B.

Theorem 2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable, $G_\star$-Lipschitz and convex function, admitting a minimizer $w^\star \in \arg\min_{w \in \mathbb{R}^d} f(w)$. Let $\widetilde{f}$ be a σ_0-bounded unbiased zero-oracle of f and $\widetilde{g}$ be a $\sigma_\star$-bounded unbiased gradient oracle of f. Let $\eta_{\min}, \eta_{\max} > 0$ such that $\eta_{\min} \leq \|w_1 - w^\star\|/\sqrt{(G_\star^2 + \sigma_\star^2)T} \leq \eta_{\max}$ for some $w_1 \in \mathbb{R}^d$. Then Algorithm 3 performs T queries and produce $\overline{w}$ such that with probability at least $1 - 2\delta$,*

$$f(\overline{w}) - f(w^\star) = \widehat{O}\left(\frac{\sigma_0 \log(\frac{1}{\delta})\sqrt{\log(\frac{\eta_{\max}}{\eta_{\min}})}}{\sqrt{T}}\right.$$
$$\left. + \frac{\|w_1 - w^\star\|\sqrt{(G_\star^2 + \sigma_\star^2)\log(\frac{\eta_{\max}}{\eta_{\min}})\log(\frac{1}{\delta})}}{\sqrt{T}}\right).$$

Algorithm 3 obtains the same rate of convergence as *tuned* SGD up to logarithmic factors and a term of order $\sigma_0/\sqrt{T}$, being parameter-free given the mild assumption that $\sigma_0 = O(D_\star(G_\star + \sigma_\star))$. Similarly to Algorithm 2, the inputs $\eta_{\min}, \eta_{\max}$ may be replaced by problem parameters ranges (details at Appendix B). We note that Algorithm 3 achieves an analogous guarantee for convex smooth objectives (via a proper setting of $\eta_{\min}, \eta_{\max}$ using $D_{\min}, D_{\max}, \beta_{\min}, \beta_{\max}$ and $\sigma_{\min}, \sigma_{\max}$).

4.2. Optimization with a stochastic gradient oracle

While function value access is a reasonable assumption which leads to a simple parameter-free method, much of the existing work on adaptive, self-tuning optimization only assume access to stochastic gradient queries. In this section we present a parameter-free method, using only stochastic gradient access, which match the convergence rate of *tuned* SGD up to a lower-order term $O(\sigma_{\max}/T)$ and poly-logarithmic factors. Our proposed method (Algorithm 1) is based on tuning the diameter of the projected variant of Stochastic Gradient Descent with AdaGrad-like step-sizes method (AdaPSGD). For a given diameter, the update step of AdaSGD is $w_{t+1} = \text{Proj}_{w_1, D}(w_t - \eta_t g_t)$, where $\text{Proj}_{w_1, D}(\cdot)$ is the Euclidean projection onto the D-ball centered at the initialization w_1, $g_t = \widetilde{g}(w_t)$ is the stochastic gradient at step t, and for some $\alpha > 0$ and $\gamma \geq 0$,

$$\eta_t = \frac{\alpha D}{\sqrt{\gamma^2 + \Sigma_{s=1}^t \|g_s\|^2}}.$$

Our method performs multiple runs of AdaPSGD with an exponential grid of diameters and stops when all iterations are fully within the D-ball. The method may be considered as a middle ground between Carmon & Hinder (2022) and Ivgi et al. (2023) as it combines tuning and dynamic stepsizes.

Following is the convergence result of our method.

Theorem 3. *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a differentiable, $G_\star$-Lipschitz and convex function, admitting a minimizer $w^\star \in \arg\min_{w \in \mathbb{R}^d} f(w)$. Let $\widetilde{g}$ be an unbiased gradient oracle of f with $\sigma_\star$-bounded noise, and let $w_1 \in \mathbb{R}^d$. Then Algorithm 1 with parameters $D_{\min} \leq \|w_1 - w^\star\| < D_{\max}$, $G_{\max} \geq G_\star$ and $\sigma_{\max} \geq \sigma_\star$ performs at most $T \log \frac{D_{\max}}{D_{\min}}$ gradient queries and produce $\overline{w} \in \mathbb{R}^d$ such that with probability at least $1 - \delta \log \frac{D_{\max}}{D_{\min}}$,*

$$f(\overline{w}) - f(w^\star) \leq \left(\frac{\|w_1 - w^\star\|(G_\star + \sigma_\star)}{\sqrt{T}} \right.$$
$$\left. + \frac{\|w_1 - w^\star\|\sigma_{\max}}{T} \right) \text{poly}\log\left(\frac{G_{\max}}{\sigma_{\max}}, T, \frac{1}{\delta}\right).$$

Observing the convergence rate in Theorem 3, Algorithm 1 achieve the same rate of convergence as *fully-tuned* SGD

Algorithm 1: Adaptive projected SGD tuning

Input: $D_{\min}, D_{\max}, G_{\max}, \sigma_{\max}, T$ and δ

$K \leftarrow \log \frac{D_{\max}}{D_{\min}}$

$\gamma \leftarrow \sqrt{5\sigma_{\max}^2 \log \frac{T}{\delta}}$

$\theta_{T,\delta} \leftarrow \log \frac{60 \log(6T)}{\delta}$

$\alpha \leftarrow \left(48 \log\left(1 + \frac{2G_{\max}^2 T + 2\sigma_{\max}^2 T}{5\sigma_{\max}^2 \log \frac{T}{\delta}} \right) + 32 \log(1 + T) + \right.$

$\left. 128\sqrt{\theta_{T,\delta} \log(1 + T) + \frac{\theta_{T,\delta}^2}{5 \log \frac{T}{\delta}}} \right)^{-1}$

for $k \leftarrow 1, 2, \ldots, K$ **do**

$\quad D_k \leftarrow 2^{k+2} D_{\min}$

$\quad w_1^{(k)}, \ldots, w_T^{(k)} \leftarrow \text{AdaPSGD}(D_k, \alpha, \gamma, w_1, T)$

$\quad$ **if** $\max_{t \leq T} \|w_{t+1}^{(k)} - w_1\| < D_k$ **then**

$\quad\quad$ **return** $\bar{w}^{(k)} \leftarrow \frac{1}{T} \Sigma_{t=1}^T w_t^{(k)}$

$\quad$ **end**

end

return w_1 ▷ Failure case

---

up to logarithmic factors and an additional lower order term of $\widetilde{O}(\sigma_\star/T)$.[4] Hence, in case the ratio $\sigma_{\max}/(\sigma_{\min} + G_{\min})$ is $O(\sqrt{T})$, Algorithm 1 is parameter-free compared to the rate of *tuned* SGD. Previous parameter-free results for SCO (Carmon & Hinder, 2022; Ivgi et al., 2023) suffer a similar lower order term, while our result depends only on the noise bound and not a bound of the norm of stochastic gradients which encapsulate both a noise and gradient norm bound, a property that stems from a careful combination of "decorrelated" stepsizes and martingale analysis instead of using the crude bound of stochastic gradients norm. (This point will subsequently prove critical in the convex smooth case where gradients are not directly bounded.) We also note that the bisection technique of Carmon & Hinder (2022) can improve the poly-logarithmic dependency; we abstain from using such a technique in favour of making the tuning procedure straightforward and leave improvements of the logarithmic factors to future work.

The convergence result of Algorithm 1 holds in fact holds more generally, assuming the gradient and noise bounds are satisfied only inside a ball of radius $8\|w_1 - w^\star\|$. We will prove a more general version and Theorem 3 will follow as a corollary at Appendix C. The refined result will be relevant later when we use the convergence guarantee for convex smooth objectives where a global gradient norm bound does not hold. To that end we define $G_{8\star}, \sigma_{8\star} \in \mathbb{R}_+$ as the minimal parameters satisfying, for all $w \in \mathbb{R}^d$ such

---

[4]In Theorem 6 we show that such a term is unavoidable without further assumptions and in Appendix D we provide a further assumption which mitigate the excess term.

that $\|w - w_1\| \leq 8\|w_1 - w^\star\|$,

$$\|\nabla f(w)\| \leq G_{8\star}, \ \Pr(\|\widetilde{g}(w) - \nabla f(w)\| \leq \sigma_{8\star}) = 1. \quad (1)$$

Following is the general version of Theorem 3.

Theorem 4. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable and convex function, admitting a minimizer $w^\star \in \arg\min f(w)$. Let $\widetilde{g}$ be an unbiased gradient oracle of f such that $\sigma_{8\star} < \infty$ (see Equation (1)), and let $w_1 \in \mathbb{R}^d$. Then Algorithm 1 with parameters $D_{\min} \leq \|w_1 - w^\star\| < D_{\max}$, $G_{\max} \geq G_{8\star}$ and $\sigma_{\max} \geq \sigma_{8\star}$, where $G_{8\star}$ and $\sigma_{8\star}$ are defined by Equation (1), performs at most $T \log \frac{D_{\max}}{D_{\min}}$ gradient queries and produce $\overline{w}_k = \frac{1}{T} \sum_{t=1}^{T} w_t^{(k)}$ for some $k \in [\log \frac{D_{\max}}{D_{\min}}]$ such that with probability at least $1 - \delta \log \frac{D_{\max}}{D_{\min}}$, $D_k \leq 8\|w_1 - w^\star\|$ and*

$$f(\overline{w}_k) - f^\star \leq \frac{1}{T} \sum_{t=1}^{T} f(w_t^{(k)}) - f^\star$$

$$= O\left(\frac{\widetilde{C}\|w_1 - w^\star\|}{T}\left(\sqrt{\sum_{t=1}^{T}\|g_t^{(k)}\|^2} + \sigma_{\max}\sqrt{\log\frac{T}{\delta}}\right)\right)$$

where $\widetilde{C} = \log(\frac{T}{\delta} + \frac{G_{\max}}{\sigma_{\max}})$ and $g_t^{(k)}$ is the observed gradient at $w_t^{(k)}$.

The bound of the general version uses the norms of the observed stochastic gradients instead of the gradient norm and noise bounds, and a simple translation appearing in Appendix C yields Theorem 3.

In order to prove Theorem 4, we first provide several intermediate results for T-steps AdaPSGD for some $D \leq 8\|w_1 - w^\star\|$ at Lemmas 1 to 3 (in which case the bounds $G_{8\star}$ and $\sigma_{8\star}$ hold for all the projected domain). We defer their proofs to Appendix C. Differently from common analyses of projected methods, we are interested in both the local minimizer inside the D-ball and the global minimizer in $\mathbb{R}^d$. To that end, let $w_D^\star = \arg\min_{w:\|w - w_1\| \leq D} f(w)$ such that $\|w_1 - w_D^\star\| \leq \|w_1 - w^\star\|$.[5] The first lemma establishes a convergence guarantee with respect to the *global minimizer* given the following two events. The first is that all the iterates are contained (fully) within the D-ball, i.e., $E_1 \triangleq \{\max_{t \leq T}\|w_{t+1} - w_1\| < D\}$. The second,

$$E_2 \triangleq \left\{\left|\sum_{t=1}^{T}(\nabla f(w_t) - g_t) \cdot (w_t - w^\star)\right|\right.$$

$$\left. \leq 4(D + \|w_1 - w^\star\|)\sqrt{\theta_{T,\delta}\sum_{t=1}^{T}\|g_t\|^2 + 4\theta_{T,\delta}^2\sigma_{8\star}^2}\right\},$$

is the concentration needed to modify the regret analysis to a high-probability analysis.

---

[5]Such $w_D^\star$ exists since either $\|w_1 - w^\star\| \leq D$ and we may use $w_D^\star = w^\star$, or $\|w_1 - w_D^\star\| \leq D < \|w_1 - w^\star\|$.

Lemma 1. *Under the event $E_1 \cap E_2$,*

$$\frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w^\star) \leq \left(\frac{2\gamma}{\alpha} + 8\theta_{T,\delta}\sigma_{8\star}\right)\frac{\|w_1 - w^\star\| + D}{T}$$

$$+ \left((\|w_1 - w^\star\| + D)(\frac{2}{\alpha} + 4\sqrt{\theta_{T,\delta}}) + \alpha D\right)\frac{\sqrt{\sum_{t=1}^{T}\|g_t\|^2}}{T}.$$

The second lemma bounds the distance of the iterates from the initialization. This lemma will prove useful to show that E_1 holds with high probability for a large enough D.

Lemma 2. *Let α, γ as defined by Algorithm 1. Then with probability at least $1 - 2\delta$, for all $t \in [T]$,*

$$\|w_{t+1} - w_1\| \leq 2\|w_1 - w^\star\| + \frac{1}{2}D.$$

The third lemma lower bounds the probability of E_2. Note that the concentration bound does not depends on the gradient norm bound, only the observed (noisy) gradients and the noise bound.

Lemma 3. $\Pr(E_2) \geq 1 - 2\delta$.

We are ready to prove the main result.

Proof of Theorem 4. Let $\overline{w}^{(\tau)}$ for $\tau \in [K+1]$ be the output of Algorithm 1 (treating w_1 as $\overline{w}^{(K+1)}$). Let

$$i = \min\{k \in [K] : D_k > 4\|w_1 - w^\star\|\}.$$

Note that i is well defined as $D_K = 4D_{\max} > 4\|w_1 - w^\star\|$ and that by minimality of i, $D_i \leq 8\|w_1 - w^\star\|$. We define the events $E_1^{(k)}$ and $E_2^{(k)}$ as the respective E_1 and E_2 events for the k'th sequence of AdaPSGD, $w_1^{(k)}, \ldots, w_T^{(k)}$. From Lemma 2, with probability at least $1 - 2\delta$, for all $t \in [T]$,

$$\|w_{t+1}^{(i)} - w_1\| \leq 2\|w_1 - w^\star\| + \frac{1}{2}D_i < D_i,$$

where the last inequality follows by $D_i > 4\|w_1 - w^\star\|$. Hence, $\Pr\{\tau \leq i\} \geq \Pr\{\tau = i \mid \tau \geq i\} \geq 1 - 2\delta$. By Lemma 3, with probability at least $1 - 2\delta K$ (union bound over all $k \in [i]$, where $i \leq K$), $E_2^{(k)}$ holds for all $k \in [i]$. Again using union bound with $\{\tau \leq i\}$, with probability at least $1 - 2\delta(K+1)$, we also have that $\{\tau \leq i\}$, which means that both $E_1^{(\tau)}$ and $E_2^{(\tau)}$ hold and we can invoke Lemma 1. Hence, with probability $1 - 2\delta(K+1)$, Algorithm 1 returns $\overline{w}^{(\tau)}$ for some $\tau \leq i$ for which

$$\frac{1}{T}\sum_{t=1}^{T} f(w_t^{(\tau)}) - f^\star \leq \left(\frac{2\gamma}{\alpha} + 8\theta_{T,\delta}\sigma_{8\star}\right)\frac{\|w_1 - w^\star\| + D_\tau}{T}$$

$$+ \left((\|w_1 - w^\star\| + D_\tau)(\frac{2}{\alpha} + 4\sqrt{\theta_{T,\delta}}) + \alpha D_\tau\right)\frac{\sqrt{\sum_{t=1}^{T}\|g_t\|^2}}{T}.$$

Using the inequalities $D_\tau \leq D_i \leq 8\|w_1 - w^\star\|$, $\sigma_{8\star} \leq \sigma_{\max}$, and a standard application of Jensen's inequality,

$$f(\overline{w}^{(\tau)}) - f^\star = O\left(\left(\frac{\gamma}{\alpha\sigma_{\max}} + \theta_{T,\delta}\right)\frac{\|w_1 - w^\star\|\sigma_{\max}}{T} \right.$$

$$\left. + \left(\left(\frac{1}{\alpha} + \sqrt{\theta_{T,\delta}}\right) + \alpha\right)\frac{\|w_1 - w^\star\|\sqrt{\Sigma_{t=1}^T \|g_t^{(\tau)}\|^2}}{T}\right).$$

We conclude by substituting γ, noting that $\alpha \leq 1$, $\sqrt{\theta_{T,\delta}}, \theta_{T,\delta}, \alpha^{-1} = O\left(\log\left(\frac{T}{\delta} + \frac{G_{\max}}{\sigma_{\max}}\right)\right)$. $\square$

4.3. Convex and smooth stochastic optimization

In this setting we assume that the objective is smooth rather then Lipschitz. We prove that Algorithm 1 also achieve the same rate of convergence as *tuned* SGD for smooth objectives, up to logarithmic factors and a lower-order term.

Theorem 5. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a $\beta_\star$-smooth and convex function, admitting a minimizer $w^\star \in \arg\min f(w)$. Let $\widetilde{g}$ be an unbiased gradient oracle of f and let $w_1 \in \mathbb{R}^d$. Then Algorithm 1 with parameters $D_{\min} \leq \|w_1 - w^\star\| < D_{\max}$, $G_{\max} \geq 9\beta_\star\|w_1 - w^\star\|$ and $\sigma_{\max} \geq \sigma_{8\star}$, where $\sigma_{8\star}$ is defined by Equation (1), performs at most $T \log\frac{D_{\max}}{D_{\min}}$ gradient queries and produce $\overline{w}$ such that with probability at least $1 - \delta\log\frac{D_{\max}}{D_{\min}}$, $D_k \leq 8\|w_1 - w^\star\|$ and*

$$f(\overline{w}) - f^\star = O\left(\frac{\widetilde{C}^2\beta_\star\|w_1 - w^\star\|^2}{T} + \frac{\widetilde{C}\|w_1 - w^\star\|\sigma_{8\star}}{\sqrt{T}} \right.$$

$$\left. + \sigma_{\max}\frac{\widetilde{C}\|w_1 - w^\star\|\sqrt{\log\frac{T}{\delta}}}{T}\right),$$

where $\widetilde{C} = \log(\frac{T}{\delta} + \frac{G_{\max}}{\sigma_{\max}})$.

The method is parameter-free compared to *tuned* SGD, assuming $\sigma_{\max}/\sigma_{\min} = O(\sqrt{T})$. Previous parameter-free methods for stochastic convex optimization (Carmon & Hinder, 2022; Ivgi et al., 2023) provided only a noiseless guarantee for smooth objectives. Attia & Koren (2023) provided a convergence result for adaptive SGD with AdaGrad-like stepsizes for convex and smooth objectives, but similarly to the non-convex case, the method is not parameter-free compared to *tuned* SGD.

Theorem 4 requires only a local gradients norm bound, inside a ball of radius $8\|w_1 - w^\star\|$. An application of smoothness with the triangle inequality yields that for all w such that $\|w - w_1\| \leq 8\|w_1 - w^\star\|$, $\|\nabla f(w)\| \leq 9\beta_\star\|w_1 - w^\star\|$. Hence, we can apply Theorem 4 and the error incurred by the bound is merely logarithmic. Using a standard analysis we can translate the convex rate guarantee to a convex and smooth rate guarantee. For the full proof see Appendix C.

4.4. Information-theoretic limits to parameter freeness

Both Theorems 4 and 5 suffer an additional term of $\widetilde{O}(\|w_1 - w^\star\|\sigma_{\max}/T)$ compared to tuned SGD, that prevents them from being fully parameter-free compared to SGD due to the polynomial dependence on $\sigma_{\max}$. The following theorem shows that without additional assumptions, this term is in fact unavoidable and the results of Theorems 4 and 5 are essentially the best one can hope for.

Theorem 6. *Let $T \geq 4$, $D_{\max} > 0$, $\sigma_{\max} > 0$, $\alpha \in [1, \frac{3}{4}\sqrt{T}]$ and $G_\star = \frac{\sigma_{\max}}{2T-1}$. Let $\mathcal{A}$ be any deterministic algorithm which performs T gradient queries and outputs $\overline{w} \in \mathbb{R}$. Then there exist some $D_\star \in [\frac{D_{\max}}{\alpha\sqrt{T}}, D_{\max}]$, $\sigma_\star \in [\frac{\sigma_{\max}}{\alpha\sqrt{T}}, \sigma_{\max}]$, a convex and $G_\star$-Lipschitz function $f : \mathbb{R} \to \mathbb{R}$ which admits a minimizer $w^\star \in \arg\min_{w\in\mathbb{R}} f(w)$ such that $\|w^\star\| = D_\star$, and a $\sigma_\star$-bounded unbiased sub-gradient[6] oracle of f, $\widetilde{g}$ such that with probability at least $\frac{1}{4}$,*

$$f(\overline{w}) - f(w^\star) \geq \frac{D_\star(G_\star + \sigma_\star)\alpha}{6\sqrt{T}} \geq \frac{D_\star\sigma_{\max}}{6T}.$$

We remark that the deterministic requirement is made for simplicity and the argument can be adapted to randomized algorithms. To understand the lower bound consider the convergence guarantee of tuned SGD. A standard in-expectation analysis of SGD with a *tuned* stepsize of $\eta = D/(G + \sigma)\sqrt{T}$ and an application of Markov's inequality shows that with probability at least $\frac{3}{4}$, SGD produce $\overline{w}$ with $f(\overline{w}) - f(w^\star) = O(D_\star(G_\star + \sigma_\star)/\sqrt{T})$. Hence, when $\alpha = \omega(1)$, the worst case analysis of SGD is strictly better than any algorithm with the limited knowledge prescribed in Theorem 6. In particular, a method cannot be parameter-free and expect a comparable rate of convergence to SGD unless one of $[D_{\min}, D_{\max}]$, $[\sigma_{\min}, \sigma_{\max}]$ has ratio of $\widetilde{O}(\sqrt{T})$ or the $G_\star$ value is outside the range (for example if the noise bound is small). This lower bound of Theorem 6 indeed appears in the guarantees of Carmon & Hinder (2022); Ivgi et al. (2023) in the form of the bound of stochastic gradients norm and in our Theorem 4. While the term $\Omega(\|w_1 - w^\star\|\sigma_{\max}/T)$ is unavoidable, in Appendix D we show that assuming the initialization is a "bad enough" approximate minimizer, we can infer a reasonable bound of $\sigma_\star$ which reduce the lower bound to a term proportional to the guarantee of tuned SGD.

We defer the proof to Appendix C and provide the general structure of the proof. The proof is based on constructing two problem instances which are indistinguishable with constant probability using only T queries, where one instance pretends to be the other using a large $\sigma_\star$ value. The lower bound is established by showing that no point has good sub-optimality for both problems.

---

[6]The function is of the form $G_\star|w - D_\star|$, and smoothing can be used to support smooth objectives. With a slight abuse of notation we treat $\nabla f(w)$ as some sub-gradient of f such that $\mathbb{E}[\widetilde{g}(w)] = \nabla f(w)$.

Acknowledgements

Impact Statement

This paper presents work whose goal is to advance the field of theoretical Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Alacaoglu, A., Malitsky, Y., Mertikopoulos, P., and Cevher, V. A new regret analysis for adam-type algorithms. In *International conference on machine learning*, pp. 202–210. PMLR, 2020.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, pp. 1–50, 2022.

Attia, A. and Koren, T. Sgd with adagrad stepsizes: Full adaptivity with high probability to unknown parameters, unbounded gradients and affine variance. In *International Conference on Machine Learning*, 2023.

Attia, A. and Koren, T. How free is parameter-free stochastic optimization? *arXiv preprint arXiv:2402.03126*, 2024a.

Attia, A. and Koren, T. A note on high-probability analysis of algorithms with exponential, sub-gaussian, and general light tails. *arXiv preprint arXiv:2403.02873*, 2024b.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Bottou, L. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pp. 421–436. Springer, 2012.

Carmon, Y. and Hinder, O. Making sgd parameter-free. In *Conference on Learning Theory*, pp. 2360–2389. PMLR, 2022.

Carmon, Y. and Hinder, O. The price of adaptivity in stochastic convex optimization. *arXiv preprint arXiv:2402.10898*, 2024.

Chaudhuri, K., Freund, Y., and Hsu, D. J. A parameter-free hedging algorithm. *Advances in neural information processing systems*, 22, 2009.

Chen, K., Langford, J., and Orabona, F. Better parameter-free stochastic optimization with ode updates for coin-betting. In *AAAI Conference on Artificial Intelligence*, 2020.

Cutkosky, A. Artificial constraints and hints for unbounded online learning. In *Conference on Learning Theory*, pp. 874–894. PMLR, 2019.

Cutkosky, A. and Boahen, K. Online learning without prior information. In *Conference on learning theory*, pp. 643–677. PMLR, 2017.

Cutkosky, A. and Boahen, K. A. Online convex optimization with unconstrained domains and losses. *Advances in neural information processing systems*, 29, 2016.

Cutkosky, A. and Orabona, F. Black-box reductions for parameter-free online learning in banach spaces. In *Conference On Learning Theory*, pp. 1493–1529. PMLR, 2018.

Defazio, A. and Mishchenko, K. Learning-rate-free learning by d-adaptation. In *International Conference on Machine Learning*, 2023.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Faw, M., Tziotis, I., Caramanis, C., Mokhtari, A., Shakkottai, S., and Ward, R. A. The power of adaptivity in sgd: Self-tuning step sizes with unbounded gradients and affine variance. In *COLT*, 2022.

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

Hazan, E. and Kakade, S. Revisiting the polyak step size. *arXiv preprint arXiv:1905.00313*, 2019.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

Ivgi, M., Hinder, O., and Carmon, Y. Dog is sgd's best friend: A parameter-free dynamic step size schedule. In *International Conference on Machine Learning*, 2023.

Jun, K.-S. and Orabona, F. Parameter-free online convex optimization with sub-exponential noise. In *Conference on Learning Theory*, pp. 1802–1823. PMLR, 2019.

Kavis, A., Levy, K. Y., Bach, F., and Cevher, V. Unixgrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. *Advances in neural information processing systems*, 32, 2019.

Kavis, A., Levy, K. Y., and Cevher, V. High probability bounds for a class of nonconvex algorithms with ada-grad stepsize. In *International Conference on Learning Representations*, 2022.

Khaled, A. and Jin, C. Tuning-free stochastic optimization. *arXiv preprint arXiv:2402.07793*, 2024.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Lan, G. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

Li, X. and Orabona, F. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 983–992. PMLR, 2019.

Li, X. and Orabona, F. A high probability analysis of adaptive sgd with momentum. In *Workshop on Beyond First Order Methods in ML Systems at ICML'20*, 2020.

Liu, Z., Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. L. High probability convergence of stochastic gradient methods. *arXiv preprint arXiv:2302.14843*, 2023.

Luo, H. and Schapire, R. E. Achieving all with no parameters: Adanormalhedge. In *Conference on Learning Theory*, pp. 1286–1304. PMLR, 2015.

McMahan, H. B. and Orabona, F. Unconstrained online linear learning in hilbert spaces: Minimax algorithms and normal approximations. In *Conference on Learning Theory*, pp. 1020–1039. PMLR, 2014.

Mhammedi, Z. and Koolen, W. M. Lipschitz and comparator-norm adaptivity in online learning. In *Conference on Learning Theory*, pp. 2858–2887. PMLR, 2020.

Mishchenko, K. and Defazio, A. Prodigy: An expeditiously adaptive parameter-free learner. *arXiv preprint arXiv:2306.06101*, 2023.

Nemirovskij, A. S. and Yudin, D. B. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

Nesterov, Y. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152 (1-2):381–404, 2015.

Orabona, F. and Pál, D. Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Orabona, F. and Pál, D. Scale-free online learning. *Theoretical Computer Science*, 716:50–69, 2018.

Orabona, F. and Pál, D. Parameter-free stochastic optimization of variationally coherent functions. *arXiv preprint arXiv:2102.00236*, 2021.

Orabona, F. and Tommasi, T. Training deep networks without learning rates through coin betting. *Advances in Neural Information Processing Systems*, 30, 2017.

Reddi, S. J., Kale, S., and Kumar, S. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

Robbins, H. and Monro, S. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Schaul, T., Zhang, S., and LeCun, Y. No more pesky learning rates. In *International conference on machine learning*, pp. 343–351. PMLR, 2013.

Streeter, M. J. and McMahan, H. B. No-regret algorithms for unconstrained online convex optimization. In *Neural Information Processing Systems*, 2012.

Tran, P. T. et al. On the convergence proof of amsgrad and a new version. *IEEE Access*, 7:61706–61716, 2019.

Ward, R., Wu, X., and Bottou, L. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. In *International Conference on Machine Learning*, pp. 6677–6686. PMLR, 2019.

Algorithm 2: Non-convex SGD tuning

Input: $w_1, T, \delta, \eta_{min}, \eta_{max}$ (see Equation (2) for replacing η_{min}, η_{max} with parameter ranges)

$S \leftarrow \{\}$

$K \leftarrow \log\left(\frac{\eta_{max}}{\eta_{min}} \sqrt{\log_+(\frac{\eta_{max}}{\eta_{min}}) \log_+(\frac{1}{\delta})}\right)$

$N \leftarrow \log \frac{1}{\delta}$

$T' \leftarrow \frac{T}{K(1+N)}$

for $k \leftarrow 1, 2, \ldots, K$ **do**

    $\eta_k \leftarrow 2^{k-1} \eta_{min}$

    $w_1^{(k)}, \ldots, w_T^{(k)} \leftarrow \mathsf{SGD}(w_1, \eta_k, T')$

    Uniformly at random select $\overline{w}_1^{(k)}, \ldots, \overline{w}_N^{(k)}$ from $w_1^{(k)}, \ldots, w_T^{(k)}$

    $S \leftarrow S \cup \{\overline{w}_1^{(k)}, \ldots, \overline{w}_n^{(k)}\}$

end

return $\arg\min_{w \in S} \|\sum_{t=1}^{T'} \widetilde{g}_t(w)\|$ ▷ `g̃_t(w) is an independent stochastic gradient at w`

A. Additional Details for Section 3

In this section we provide the algorithmic statement, additional details and proofs related to Theorem 1. For completeness we re-state the theorem.

Theorem 1. *Assume that f is $\beta_\star$-smooth and lower bounded by some $f^\star$ and $\widetilde{g}$ is a $\sigma_\star$-bounded unbiased gradient oracle of f. Let $\eta_{min}, \eta_{max} > 0$ such that*

$$\eta_{min} \leq \eta_\star \triangleq \min\left\{\frac{1}{2\beta_\star}, \sqrt{\frac{F_\star}{\beta_\star \sigma_\star^2 T}}\right\} \leq \eta_{max}$$

where $F_\star = f(w_1) - f^\star$. Then for any $\delta \in (0, \frac{1}{3})$, given $w_1, T, \delta, \eta_{min}$ and η_{max}, Algorithm 2 performs T gradient queries and produce $\overline{w}$ such that with probability $\geq 1 - \delta$,

$$\|\nabla f(\overline{w})\|^2 = \widehat{O}\left(\frac{\log_+(\frac{\eta_{max}}{\eta_{min}}) \log_+(\frac{1}{\delta})(\beta_\star F_\star + \sigma_\star^2 \log \frac{1}{\delta})}{T}\right.$$
$$\left. + \frac{\sigma_\star \sqrt{\beta_\star F_\star \log_+(\frac{\eta_{max}}{\eta_{min}}) \log_+(\frac{1}{\delta})}}{\sqrt{T}}\right).$$

Note that Algorithm 2 receive η_{min}, η_{max} instead of ranges $[F_{min}, F_{max}], [\beta_{min}, \beta_{max}], [\sigma_{min}, \sigma_{max}]$. This is done for framing the algorithm as a form of stepsize tuning (which in practice requires only two hyperparameters instead of two hyperparameters per problem parameter) and given the ranges the algorithm will use

$$\eta_{min} = \min\left\{\frac{1}{2\beta_{max}}, \sqrt{\frac{F_{min}}{\beta_{max}\sigma_{max}^2 T}}\right\} \quad \text{and} \quad \eta_{max} = \min\left\{\frac{1}{2\beta_{min}}, \sqrt{\frac{F_{max}}{\beta_{min}\sigma_{min}^2 T}}\right\}. \tag{2}$$

In order to prove the theorem we need the following lemmas , their proofs follow . The first is a standard convergence result of constant stepsize SGD. Alternatively, one can use the in-expectation analysis of SGD as in the two-phase SGD method of Ghadimi & Lan (2013).

Lemma 4 (SGD convergence with high probability)**.** *Assume that f is $\beta_\star$-smooth and lower bounded by $f^\star$. Then for the iterates of SGD with $\eta \leq 1/2\beta_\star$, it holds with probability at least $1 - \delta$ that*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|^2 \leq \left(\frac{4F_\star}{\eta} + 12\sigma_\star^2 \log \frac{1}{\delta}\right)\frac{1}{T} + 4\beta_\star\sigma_\star^2\eta,$$

where $F_\star = f(w_1) - f^\star$.

Next is a standard success amplification technique for choosing between candidate solutions SGD; the precise version below is due to Ghadimi & Lan (2013).

Lemma 5. *Given candidates $w_1, \ldots, w_N$, let $\overline{w} = \arg\min_{n \in [N]} \| \sum_{t=1}^T \widetilde{g}_t(w_n) \|$ where $g_t(w)$ for $t \in [T]$ are independent stochastic gradients at w. Then assuming $\mathbb{E}[\widetilde{g}_t(w)] = \nabla f(w)$ and $\Pr(\|\nabla f(w) - \widetilde{g}_t(w)\| \le \sigma) = 1$ for any fixed w, for any $\delta \in (0, 1)$, we have with probability at least $1 - \delta$ that*

$$\|\nabla f(\overline{w})\|^2 \le 4 \min_{n \in [N]} \|\nabla f(w_n)\|^2 + \frac{24(1 + 3\log\frac{N}{\delta})\sigma^2}{T}.$$

We proceed to proving the theorem.

Proof of Theorem 1. To fix the discrepancy between T and T', we denote

$$\eta'_\star \triangleq \min\left\{ \frac{1}{2\beta_\star}, \sqrt{\frac{F_\star \log_+(\frac{\eta_{\max}}{\eta_{\min}})\log_+(\frac{1}{\delta})}{\beta_\star \sigma_\star^2 T}} \right\} \ge \eta_\star,$$

Note that

$$\eta_1 \le \eta_\star \le \eta'_\star \le \eta_\star \sqrt{\log_+(\tfrac{\eta_{\max}}{\eta_{\min}})\log_+(\tfrac{1}{\delta})} \le \eta_{\max}\sqrt{\log_+(\tfrac{\eta_{\max}}{\eta_{\min}})\log_+(\tfrac{1}{\delta})} = 2\eta_K.$$

Let $\tau = \max\{k \in [K] : \eta_k \le \eta'_\star\}$. τ is well-defined as $\eta_1 \le \eta'_\star$. Hence, since $2\eta_K \ge \eta'_\star$, $\eta_\tau \in (\eta'_\star/2, \eta'_\star]$. Thus, by Lemma 4, with probability at least $1 - \delta$,

$$\frac{1}{T'}\sum_{t=1}^{T'} \|\nabla f(w_t^{(\tau)})\|^2 \le \left(\frac{4F_\star}{\eta_\tau} + 12\sigma_\star^2 \log\frac{1}{\delta}\right)\frac{1}{T'} + 4\beta_\star \sigma_\star^2 \eta_\tau$$

$$\le \left(\frac{8F_\star}{\eta'_\star} + 12\sigma_\star^2 \log\frac{1}{\delta}\right)\frac{1}{T'} + 4\beta_\star \sigma_\star^2 \eta'_\star$$

$$\le \frac{K(1+N)(16\beta_\star F_\star + 12\sigma_\star^2 \log\frac{1}{\delta})}{T} + \frac{12K(1+N)\sigma_\star\sqrt{\beta_\star F_\star}}{\sqrt{T\log_+(\frac{\eta_{\max}}{\eta_{\min}})\log_+(\frac{1}{\delta})}}.$$

By Markov's inequality, with probability at least $\frac{1}{2}$, a uniformly at random index $k \in [T']$ satisfy

$$\|\nabla f(w_k^{(\tau)})\|^2 \le 2\mathbb{E}_k[\|\nabla f(w_k^{(\tau)})\|^2] = \frac{2}{T'}\sum_{t=1}^{T} \|\nabla f(w_t^{(\tau)})\|^2,$$

and as we sample $\log\frac{1}{\delta}$ random indices, with probability at least $1 - \delta$ we add at least one point with the above guarantee to S. Hence, by a union bound with the convergence guarantee, with probability at least $1 - 2\delta$ we add to S some $w_k^{(\tau)}$ with

$$\|\nabla f(w_k^{(\tau)})\|^2 \le \frac{K(1+N)(32\beta_\star F_\star + 24\sigma_\star^2 \log\frac{1}{\delta})}{T} + \frac{24K(1+N)\sigma_\star\sqrt{\beta_\star F_\star}}{\sqrt{T\log_+(\frac{\eta_{\max}}{\eta_{\min}})\log_+(\frac{1}{\delta})}}.$$

The last step of the algorithm is an application of Lemma 5 (the bounded noise assumption satisfy the sub-Gaussian requirement of Lemma 5) with the set of KN candidates, and we conclude with a union bound that with probability at least $1 - 3\delta$, $\overline{w}$, the output of Algorithm 2 satisfy

$$\|\nabla f(\overline{w})\|^2 \le \frac{K(1+N)(128\beta_\star F_\star + 192\sigma_\star^2 \log\frac{KN}{\delta})}{T} + \frac{96K(1+N)\sigma_\star\sqrt{\beta_\star F_\star}}{\sqrt{T\log_+(\frac{\eta_{\max}}{\eta_{\min}})\log_+(\frac{1}{\delta})}}$$

$$= \widehat{O}\left(\frac{\log_+(\frac{\eta_{\max}}{\eta_{\min}})\log_+(\frac{1}{\delta})(\beta_\star F_\star + \sigma_\star^2 \log\frac{1}{\delta})}{T} + \frac{\sigma_\star\sqrt{\beta_\star F_\star \log_+(\frac{\eta_{\max}}{\eta_{\min}})\log_+(\frac{1}{\delta})}}{\sqrt{T}}\right).$$

For each $k \in [K]$ we perform a single T'-steps SGD execution and $T'N$ gradient queries to approximate the norm of the randomly selected points, to a total of $T = T'K(1+N)$ queries. We translate the probability of $1 - 3\delta$ to $1 - \delta$ for the $\widehat{O}$ notation, limiting $\delta \in (0, \frac{1}{3})$. $\qquad\square$

A.1. Proof of Lemma 4

In order to prove the lemma we use the following martingale concentration inequality due to Li & Orabona (2020).

Lemma 6 (Lemma 1 of Li & Orabona (2020)). *Assume that $Z_1, Z_2, ..., Z_T$ is a martingale difference sequence with respect to $\xi_1, \xi_2, ..., \xi_T$ and $\mathbb{E}_t \left[\exp(Z_t^2/\sigma_t^2) \right] \leq \exp(1)$ for all t, where $\sigma_1, \ldots, \sigma_T$ is a sequence of random variables such that σ_t is measurable with respect to $\xi_1, \xi_2, \ldots, \xi_{t-1}$. Then, for any fixed $\lambda > 0$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$, it holds that*

$$\sum_{t=1}^T Z_t \leq \frac{3}{4}\lambda \sum_{t=1}^T \sigma_t^2 + \frac{1}{\lambda} \log \frac{1}{\delta} .$$

Proof of Lemma 4. From smoothness,

$$f(w_{t+1}) \leq f(w_t) - \eta \nabla f(w_t) \cdot g_t + \frac{\beta_\star \eta^2}{2} \|g_t\|^2.$$

Summing for $t = 1, \ldots, T$ and rearranging,

$$\sum_{t=1}^T \nabla f(w_t) \cdot g_t \leq \frac{f(w_1) - f(w_{T+1})}{\eta} + \frac{\beta_\star \eta}{2} \sum_{t=1}^T \|g_t\|^2.$$

By Lemma 6 with $Z_t = \nabla f(w_t) \cdot (\nabla f(w_t) - g_t)$, where $|Z_t| \leq \sigma_\star \|\nabla f(w_t)\|$, and $\lambda = \frac{1}{3\sigma_\star^2}$, with probability at least $1 - \delta$,

$$\sum_{t=1}^T \nabla f(w_t) \cdot (\nabla f(w_t) - g_t) \leq \frac{1}{4} \sum_{t=1}^T \|\nabla f(w_t)\|^2 + 3\sigma_\star^2 \log \frac{1}{\delta}.$$

Summing both inequalities and rearranging,

$$\frac{3}{4} \sum_{t=1}^T \|\nabla f(w_t)\|^2 \leq \frac{f(w_1) - f(w_{T+1})}{\eta} + \frac{\beta_\star \eta}{2} \sum_{t=1}^T \|g_t\|^2 + 3\sigma_\star^2 \log \frac{1}{\delta}.$$

Since $\|g_t\|^2 \leq 2\|\nabla f(w_t)\|^2 + 2\sigma_\star^2$,

$$\frac{3}{4} \sum_{t=1}^T \|\nabla f(w_t)\|^2 \leq \frac{f(w_1) - f(w_{T+1})}{\eta} + \beta_\star \eta \sum_{t=1}^T \|\nabla f(w_t)\|^2 + \beta_\star \sigma_\star^2 T\eta + 3\sigma_\star^2 \log \frac{1}{\delta}.$$

As $\eta \leq \frac{1}{2\beta_\star}$, substituting and rearranging,

$$\sum_{t=1}^T \|\nabla f(w_t)\|^2 \leq \frac{4(f(w_1) - f(w_{T+1}))}{\eta} + 4\beta_\star \sigma_\star^2 T\eta + 12\sigma_\star^2 \log \frac{1}{\delta}.$$

Dividing by $\frac{1}{T}$ and replacing $f(w_{T+1}) \geq f^\star$ we conclude the proof. $\qquad\square$

A.2. Proof of Lemma 5

In order to prove Lemma 5 we use the following martingale lemma.

Lemma 7 (Lemma 2.3 of Ghadimi & Lan (2013)). *Let $Z_1, \ldots, Z_n \in \mathbb{R}^d$ be a martingale difference sequence with respect to $\xi_1, \ldots, \xi_n$. Assuming $\mathbb{E}[\exp(\|Z_i\|^2/\sigma^2) \mid \xi_1, \ldots, \xi_{i-1}] \leq \exp(1)$, for any $\lambda > 0$,*

$$\Pr\left(\left\| \sum_{i=1}^n Z_i \right\| \geq \sqrt{2}(1 + \lambda)\sigma\sqrt{n} \right) \leq \exp(-\lambda^2/3).$$

Proof of Lemma 5. Denote $\bar{g}(w) \triangleq \frac{1}{T} \sum_{t=1}^T \widetilde{g}_t(w)$. Let $k = \arg\min_{n \in [N]} \|\nabla f(w_n)\|$. Using $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and the minimality of $\overline{w}$,

$$\|\nabla f(\overline{w})\|^2 \leq 2\|\bar{g}(\overline{w})\|^2 + 2\|g(\overline{w}) - \nabla f(\overline{w})\|^2 \leq 2\|\bar{g}(w_k)\|^2 + 2\|\bar{g}(\overline{w}) - \nabla f(\overline{w})\|^2$$

$$\leq 4\|\nabla f(w_k)\|^2 + 4\|\bar{g}(w_k) - \nabla f(w_k)\|^2 + 2\|\bar{g}(\overline{w}) - \nabla f(\overline{w})\|^2.$$

Algorithm 3: Convex SGD tuning

Input: $w_1, T, \delta, \eta_{\min}, \eta_{\max}$ (see Eq. treating $\frac{0}{0} \triangleq 0$ for replacing $\eta_{\min}, \eta_{\max}$ with parameter ranges)

$S \leftarrow \{\}$

$K \leftarrow \log \frac{\eta_{\max}}{\eta_{\min}}$

$N \leftarrow \log \frac{1}{\delta}$

$T' \leftarrow \frac{T}{2KN}$

for $k \leftarrow 1, 2, \ldots, K$ **do**

    $\eta_k \leftarrow 2^{k-1}\eta_{\min}\sqrt{2KN}$ ▷ The $\sqrt{2KN}$ factor fix discrepancy between T and T'

    for $n \leftarrow 1, 2, \ldots, N$ **do**

        $\overline{w}_k^{(n)} \leftarrow \mathsf{SGD}(w_1, \eta_k, T')$ ▷ $\overline{w}_k^{(n)}$ is the average of the iterates of SGD

        $S \leftarrow S \cup \{\overline{w}_k^{(n)}\}$

    end

end

return $\arg\min_{w \in S} \sum_{t=1}^{T'} \widetilde{f_t}(w)$ ▷ $\widetilde{f_t}(w)$ is an independent evaluation of $\widetilde{f}(w)$

Let $n \in [N]$. By Lemma 7 (note that the bounded noise assumption satisfy the sub-Gaussian condition), for any $\lambda > 0$,

$$\Pr\left(\|\bar{g}(w_n) - \nabla f(w_n)\|^2 \geq \frac{2(1+\lambda)^2\sigma^2}{T}\right) = \Pr\left(\|\sum_{t=1}^{T}(\widetilde{g}_t(w_n) - \nabla f(w_n))\| \geq \sqrt{2}(1+\lambda)\sigma\sqrt{T}\right) \leq e^{-\lambda^2/3}.$$

Thus, under a union bound,

$$\Pr\left(\exists n \in [N], \|\bar{g}(w_n) - \nabla f(w_n)\|^2 \geq \frac{2(1+\lambda)^2\sigma^2}{T}\right) \leq S\exp(-\lambda^2/3).$$

Setting $\lambda = \sqrt{3\log\frac{N}{\delta}}$, with probability at least $1 - \delta$, for all $n \in [N]$,

$$\|\bar{g}(w_n) - \nabla f(w_n)\|^2 \leq \frac{4(1+\lambda^2)\sigma^2}{T} \leq \frac{4(1+3\log\frac{N}{\delta})\sigma^2}{T}.$$

Thus, with probability at least $1 - \delta$.

$$\|\nabla f(\overline{w})\|^2 \leq 4\|\nabla f(w_k)\|^2 + 4\|\bar{g}(w_k) - \nabla f(w_k)\|^2 + 2\|\bar{g}(\overline{w}) - \nabla f(\overline{w})\|^2$$

$$\leq 4\|\nabla f(w_k)\|^2 + \frac{24(1+3\log\frac{N}{\delta})\sigma^2}{T}. \qquad \qquad \square$$

B. Additional Details for Section 4.1

In this section we provide the algorithmic statement, additional details and proofs related to Theorem 2. For completeness we re-state the theorem.

Theorem 2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable, $G_\star$-Lipschitz and convex function, admitting a minimizer $w^\star \in \arg\min_{w \in \mathbb{R}^d} f(w)$. Let $\widetilde{f}$ be a σ_0-bounded unbiased zero-oracle of f and $\widetilde{g}$ be a $\sigma_\star$-bounded unbiased gradient oracle of f. Let $\eta_{\min}, \eta_{\max} > 0$ such that $\eta_{\min} \leq \|w_1 - w^\star\|/\sqrt{(G_\star^2 + \sigma_\star^2)T} \leq \eta_{\max}$ for some $w_1 \in \mathbb{R}^d$. Then Algorithm 3 performs T queries and produce $\overline{w}$ such that with probability at least $1 - 2\delta$,*

$$f(\overline{w}) - f(w^\star) = \widehat{O}\left(\frac{\sigma_0\log(\frac{1}{\delta})\sqrt{\log(\frac{\eta_{\max}}{\eta_{\min}})}}{\sqrt{T}}\right.$$

$$\left. + \frac{\|w_1 - w^\star\|\sqrt{(G_\star^2 + \sigma_\star^2)\log(\frac{\eta_{\max}}{\eta_{\min}})\log(\frac{1}{\delta})}}{\sqrt{T}}\right).$$

Similarly to Algorithm 2, the inputs $\eta_{\min}, \eta_{\max}$ may be replaced by problem parameters ranges by setting

$$\eta_{\min} = \frac{D_{\min}}{\sqrt{(G_{\max}^2 + \sigma_{\max}^2)T}} \qquad \text{and} \qquad \eta_{\max} = \frac{D_{\max}}{\sqrt{(G_{\min}^2 + \sigma_{\min}^2)T}}.$$

Before proving Theorem 2, we need the following standard lemmas , their proofs follow. First is an in-expectation convergence of SGD.

Lemma 8. *Let* $f : \mathbb{R}^d \to \mathbb{R}$ *be a differentiable,* $G_\star$*-Lipschitz and convex function, admitting a minimizer* $w^\star \in$ $\arg\min_{w \in \mathbb{R}^d} f(w)$ *and let* $\widetilde{g} : \mathbb{R}^d \to \mathbb{R}^d$ *be a* $\sigma_\star$*-bounded unbiased gradient oracle of* f*. Given some* $w_1 \in \mathbb{R}^d$*, the iterates of* T*-steps SGD with stepsize* $\eta > 0$ *satisfy*

$$\mathbb{E}\left[f\left(\frac{1}{T} \sum_{t=1}^{T} w_t \right) - f(w^\star) \right] \le \frac{\|w_1 - w^\star\|^2}{2\eta T} + \frac{\eta(G_\star^2 + \sigma_\star^2)}{2}.$$

The second is a candidate selection lemma based on stochastic function value access.

Lemma 9. *Let* $f : \mathbb{R}^d \to \mathbb{R}$ *and* $\widetilde{f}$ *a zero-order oracle of* f *such that*

$$\forall w \in \mathbb{R}^d : \mathbb{E}[\widetilde{f}(w)] = f(w) \quad \text{and} \quad \Pr(|\widetilde{f}(w) - f(w)| \le \sigma_0) = 1$$

for some $\sigma_0 > 0$*. Given candidates* $w_1, \dots, w_N$*, let*

$$\overline{w} = \arg\min_{n \in [N]} \frac{1}{T} \sum_{t=1}^{T} \widetilde{f}_t(w_n),$$

where $\widetilde{f}_t(w_n)$ *for* $t \in [T]$ *are independent function evaluations at* w_n*. Then for any* $\delta \in (0, 1)$*, with probability at least* $1 - \delta$*,*

$$f(\overline{w}) \le \min_{n \in [N]} f(w_s) + \sqrt{\frac{8\sigma_0^2 \log \frac{2N}{\delta}}{T}}.$$

We proceed to prove the convergence result.

Proof of Theorem 2. For each pair (k, n), T' gradient queries are made for the SGD and additional T' function queries for the selection step. Hence, $T = 2KNT'$ queries are performed. By the condition of $\eta_{\min}$ and $\eta_{\max}$, there is some $\tau \in [K]$ such that

$$\eta_\tau \in \left(\frac{\|w_1 - w^\star\|}{2\sqrt{(G_\star^2 + \sigma_\star^2)T'}}, \frac{\|w_1 - w^\star\|}{\sqrt{(G_\star^2 + \sigma_\star^2)T'}} \right].$$

Applying Markov's inequality to Lemma 8 with η_τ, for all $n \in [N]$, with probability at least $\frac{1}{2}$,

$$f\left(\overline{w}_\tau^{(n)} \right) - f(w^\star) \le \frac{\|w_1 - w^\star\|^2}{\eta_\tau T'} + \eta_\tau (G_\star^2 + \sigma_\star^2) \le \frac{3\|w_1 - w^\star\| \sqrt{G_\star^2 + \sigma_\star^2}}{\sqrt{T'}}.$$

Hence, by a union bound, with probability at least $1 - \delta$, S contains some $\overline{w}_\tau^{(n)}$ with the above guarantee. We conclude by a union bound of the above with Lemma 9 and substituting $T' = \frac{T}{2\log(D_{\max}/D_{\min})\log(1/\delta)}$, such that with probability at least $1 - 2\delta$,

$$f(\overline{w}) - f(w^\star) \le \frac{3\|w_1 - w^\star\| \sqrt{(G_\star^2 + \sigma_\star^2) \log(\frac{\eta_{\max}}{\eta_{\min}}) \log(\frac{1}{\delta})}}{\sqrt{T}} + \frac{\sigma_0 \sqrt{8\log(\frac{2KN}{\delta}) \log(\frac{1}{\delta}) \log(\frac{\eta_{\max}}{\eta_{\min}})}}{\sqrt{T}}$$

$$= \widehat{O}\left(\frac{\|w_1 - w^\star\| \sqrt{(G_\star^2 + \sigma_\star^2) \log(\frac{\eta_{\max}}{\eta_{\min}}) \log(\frac{1}{\delta})} + \sigma_0 \log(\frac{1}{\delta}) \sqrt{\log(\frac{\eta_{\max}}{\eta_{\min}})}}{\sqrt{T}} \right). \qquad \square$$

B.1. Proof of Lemma 8

By the update step of SGD,

$$\|w_{t+1} - w^\star\|^2 = \|w_t - w^\star\|^2 - 2\eta g_t \cdot (w_t - w^\star) + \eta^2 \|g_t\|^2.$$

Summing for $t = 1, \ldots, T$ and rearranging,

$$\sum_{t=1}^T g_t \cdot (w_t - w^\star) = \frac{\|w_1 - w^\star\|^2 - \|w_{T+1} - w^\star\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|g_t\|^2.$$

Note that by the Lipschitz and noise assumptions, $\mathbb{E}[\|g_t\|^2] = \mathbb{E}[\|g_t - \nabla f(w_t)\|^2 - 2\nabla f(w_t) \cdot (\nabla f(w_t) - g_t) + \|\nabla f(w_t)\|^2] \leq G_\star^2 + \sigma_\star^2$. Taking expectation,

$$\mathbb{E}\left[\sum_{t=1}^T \nabla f(w_t) \cdot (w_t - w^\star)\right] \leq \frac{\|w_1 - w^\star\|^2}{2\eta} + \frac{\eta(G^2 + \sigma^2)T}{2}.$$

We conclude by an application of convexity and Jensen's inequality. $\qquad\square$

B.2. Proof of Lemma 9

Let $k = \arg\min_{n \in [N]} f(w_n)$. By the minimality of $\overline{w}$,

$$f(\overline{w}) = \frac{1}{T} \sum_{t=1}^T \widetilde{f}_t(\overline{w}) + \left(f(\overline{w}) - \frac{1}{T} \sum_{t=1}^T \widetilde{f}_t(\overline{w})\right) \leq \frac{1}{T} \sum_{t=1}^T \widetilde{f}_t(w_k) + \left(f(\overline{w}) - \frac{1}{T} \sum_{t=1}^T \widetilde{f}_t(\overline{w})\right)$$

$$= f(w_k) + \left(\frac{1}{T} \sum_{t=1}^T \widetilde{f}_t(w_k) - f(w_k)\right) + \left(f(\overline{w}) - \frac{1}{T} \sum_{t=1}^T \widetilde{f}_t(\overline{w})\right).$$

Let $n \in [N]$. By Hoeffding's inequality, for any $\epsilon > 0$,

$$\Pr\left(\left|f(w_n) - \frac{1}{T} \sum_{t=1}^T \widetilde{f}_t(w_n)\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2 T}{2\sigma_0^2}\right).$$

Thus, under a union bound,

$$\Pr\left(\exists n \in [N] : \left|f(w_n) - \frac{1}{T} \sum_{t=1}^T \widetilde{f}_t(w_n)\right| \geq \epsilon\right) \leq 2N\exp\left(-\frac{\epsilon^2 T}{2\sigma_0^2}\right).$$

Setting $\epsilon = \sqrt{2\sigma_0^2 \log \frac{2N}{\delta} / T}$, with probability at least $1 - \delta$, for all $n \in [N]$,

$$\left|f(w_n) - \frac{1}{T} \sum_{t=1}^T \widetilde{f}_t(w_n)\right| \leq \sqrt{\frac{2\sigma_0^2 \log \frac{2N}{\delta}}{T}}.$$

Thus, with probability at least $1 - \delta$,

$$f(\overline{w}) \leq f(w_k) + \sqrt{\frac{8\sigma_0^2 \log \frac{2N}{\delta}}{T}}. \qquad\square$$

C. Proofs of Section 4

C.1. Proof of Theorem 3

Using the identity $\|u + v\|^2 \leq 2\|u\|^2 + 2\|v\|^2$ and the global bounds, for all $w \in \mathbb{R}^d$,

$$\|\widetilde{g}(w)\| \leq 2\|\nabla f(w)\|^2 + 2\|\widetilde{g}(w) - \nabla f(w)\|^2 \leq 2G_\star^2 + 2\sigma_\star^2.$$

By the minimality of $G_{8\star}$ and $\sigma_{8\star}$, $G_{\max} \geq G_\star \geq G_{8\star}$ and $\sigma_{\max} \geq \sigma_\star \geq \sigma_{8\star}$, which means that Theorem 4 is applicable. Combining Theorem 4 with the above inequality, Algorithm 1 produce $\overline{w}_k$ such that with probability at least $1 - \delta \log \frac{D_{\max}}{D_{\min}}$,

$$f(\overline{w}_k) - f^\star = O\left(\frac{\widetilde{C}\|w_1 - w^\star\|}{T} \left(\sqrt{(G_\star^2 + \sigma_\star^2)T} + \sigma_{\max}\sqrt{\log \frac{T}{\delta}} \right) \right).$$

We conclude by moving from big-O notation to poly-log notation, replacing the term $\widetilde{C}\sqrt{\log \frac{T}{\delta}}$ with poly $\log(\frac{G_{\max}}{\sigma_{\max}}, T, \frac{1}{\delta})$. $\quad\square$

C.2. Proof of Lemma 1

In order to prove the lemma we use the following standard inequality.

Lemma 10. *Let $a_1, \ldots, a_n \geq 0$. Then*

$$\sum_{i=1}^{n} \frac{a_i}{\sqrt{\sum_{j=1}^{i} a_j}} \leq 2\sqrt{\sum_{i=1}^{n} a_i}. \qquad \text{(treating $\frac{0}{0} \triangleq 0$)}$$

Proof. Using the identity $a^2 - b^2 = (a - b)(a + b)$,

$$\sum_{i=1}^{n} \frac{a_i}{\sqrt{\sum_{j=1}^{i} a_j}} = \sum_{i=1}^{n} \frac{\left(\sqrt{\sum_{j=1}^{i} a_j} - \sqrt{\sum_{j=1}^{i-1} a_j}\right)\left(\sqrt{\sum_{j=1}^{i} a_j} + \sqrt{\sum_{j=1}^{i-1} a_j}\right)}{\sqrt{\sum_{j=1}^{i} a_j}}$$

$$\leq 2\sum_{i=1}^{n}\left(\sqrt{\sum_{j=1}^{i} a_j} - \sqrt{\sum_{j=1}^{i-1} a_j}\right) = 2\sqrt{\sum_{i=1}^{n} a_i}. \qquad\qquad \square$$

We proceed to prove the lemma.

Proof of Lemma 1. As E_1 holds, $\max_{t \leq T}\|w_{t+1} - w_1\| < D$ and $w_{t+1} = w_t - \eta_t g_t$ for all $t \in [T]$ (if a projection is performed the norm will equal D). Thus,

$$\|w_{t+1} - w^\star\|^2 = \|w_t - w^\star\|^2 - 2\eta_t g_t \cdot (w_t - w^\star) + \eta_t^2\|g_t\|^2.$$

Rearranging and summing for $t = 1, \ldots, T$,

$$\sum_{t=1}^{T} g_t \cdot (w_t - w^\star) = \frac{\|w_1 - w^\star\|^2}{2\eta_1} - \frac{\|w_{T+1} - w^\star\|^2}{2\eta_T} + \frac{1}{2}\sum_{t=2}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\|w_t - w^\star\|^2 + \frac{1}{2}\sum_{t=1}^{T} \eta_t\|g_t\|^2.$$

Let $t' \in \arg\max_{t \leq T+1}\|w_t - w^\star\|$. Hence,

$$\sum_{t=1}^{T} g_t \cdot (w_t - w^\star) \leq \frac{\|w_1 - w^\star\|^2}{2\eta_1} - \frac{\|w_{T+1} - w^\star\|^2}{2\eta_T} + \frac{\|w_{t'} - w^\star\|^2}{2}\sum_{t=2}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{1}{2}\sum_{t=1}^{T} \eta_t\|g_t\|^2$$

$$= \frac{\|w_1 - w^\star\|^2}{2\eta_1} - \frac{\|w_{T+1} - w^\star\|^2}{2\eta_T} + \frac{\|w_{t'} - w^\star\|^2}{2}\left(\frac{1}{\eta_T} - \frac{1}{\eta_1}\right) + \frac{1}{2}\sum_{t=1}^{T} \eta_t\|g_t\|^2$$

$$\leq \frac{\|w_{t'} - w^\star\|^2 - \|w_{T+1} - w^\star\|^2}{2\eta_T} + \frac{1}{2}\sum_{t=1}^{T} \eta_t\|g_t\|^2. \qquad (\|w_1 - w^\star\| \leq \|w_{t'} - w^\star\|)$$

Note that by the triangle inequality,

$$\|w_{t'} - w^\star\|^2 - \|w_{T+1} - w^\star\|^2 = (\|w_{t'} - w^\star\| - \|w_{T+1} - w^\star\|)(\|w_{t'} - w^\star\| + \|w_{T+1} - w^\star\|)$$

$$\leq 2\|w_{t'} - w_{T+1}\|\|w_{t'} - w^\star\| \qquad\qquad \text{(triangle inequality and def of t')}$$

$$\leq 2\|w_{t'} - w_{T+1}\|(D + \|w_1 - w^\star\|) \qquad\qquad (\|w_{t'} - w^\star\| \leq \|w_{t'} - w_1\| + \|w_1 - w^\star\|)$$

$$\leq 4D(D + \|w_1 - w^\star\|). \qquad\qquad (\|w_{t'} - w_{T+1}\| \leq \|w_{t'} - w_1\| + \|w_1 - w_{T+1}\| \leq 2D)$$

Plugging the above inequality,

$$\sum_{t=1}^{T} g_t \cdot (w_t - w^\star) \le \frac{2D(D + \|w_1 - w^\star\|)}{\eta_T} + \frac{1}{2} \sum_{t=1}^{T} \eta_t \|g_t\|^2.$$

By Lemma 10 with $a_i = \|g_i\|^2$,

$$\sum_{t=1}^{T} \eta_t \|g_t\|^2 \le 2\alpha D \sqrt{\sum_{t=1}^{T} \|g_t\|^2}.$$

Combining the two inequalities with the inequality of E_2,

$$\sum_{t=1}^{T} \nabla f(w_t) \cdot (w_t - w^\star) \le \frac{2D(D + \|w_1 - w^\star\|)}{\eta_T} + \alpha D \sqrt{\sum_{t=1}^{T} \|g_t\|^2}$$

$$+ 4(D + \|w_1 - w^\star\|) \sqrt{\theta_{T,\delta} \sum_{t=1}^{T} \|g_t\|^2 + 4\sigma_{8\star}^2 \theta_{t,\delta}^2}$$

$$\le \left(\frac{2(D + \|w_1 - w^\star\|)}{\alpha} + \alpha D + 4(D + \|w_1 - w^\star\|)\sqrt{\theta_{T,\delta}} \right) \sqrt{\sum_{t=1}^{T} \|g_t\|^2}$$

$$+ \frac{2(D + \|w_1 - w^\star\|)\gamma}{\alpha} + 8\theta_{T,\delta}(D + \|w_1 - w^\star\|)\sigma_{8\star} \qquad (\sqrt{a+b} \le \sqrt{a} + \sqrt{b})$$

$$= \left((\|w_1 - w^\star\| + D)(\tfrac{2}{\alpha} + 4\sqrt{\theta_{T,\delta}}) + \alpha D \right) \sqrt{\sum_{t=1}^{T} \|g_t\|^2} + \left(\tfrac{2\gamma}{\alpha} + 8\theta_{T,\delta}\sigma_{8\star} \right)(\|w_1 - w^\star\| + D).$$

Dividing by T, we conclude by applying $\nabla f(w_t) \cdot (w_t - w^\star) \ge f(w_t) - f(w^\star)$ due to convexity. $\qquad \square$

C.3. Proof of Lemma 2

In order to prove lemma Lemma 2 we use the following lemmas. Their proofs will be presented subsequently. The first lemma bounds the errors that arise from potential correlation between η_t and g_t. To that end we use the following "decorrelated" stepsize (Attia & Koren, 2023),

$$\tilde{\eta}_t \triangleq \frac{\alpha D}{\sqrt{\gamma^2 + \|\nabla f(w_t)\|^2 + \sum_{s=1}^{t-1} \|g_s\|^2}}.$$

In the noiseless case we could simply replace the lemma with $\nabla f(w_s) \cdot (w_s - w_D^\star) \ge 0$ due to convexity.

Lemma 11. *With probability at least* $1 - \delta$, *for any* $t \in [T]$,

$$-\sum_{s=1}^{t} \eta_s g_s \cdot (w_s - w_D^\star) \le \frac{1}{\alpha} \sum_{s=1}^{t} \eta_s^2 \|g_s\|^2 + \frac{1}{\alpha} \sum_{s=1}^{t} \tilde{\eta}_s^2 \|g_s - \nabla f(w_s)\|^2$$

$$+ 8D \sqrt{\theta_{t,\delta} \sum_{s=1}^{t} \tilde{\eta}_s^2 \|\nabla f(w_s) - g_s\|^2 + \frac{\alpha^2 D^2 \sigma_{8\star}^2 \theta_{t,\delta}^2}{\gamma^2}},$$

where $\theta_{t,\delta} = \log \frac{60 \log(6t)}{\delta}$.

Following is a standard summation lemma of methods with AdaGrad-like stepsizes.

Lemma 12. *For any* $t \in [T]$,

$$\sum_{s=1}^{t} \eta_s^2 \|g_s\|^2 \le \alpha^2 D^2 \log \left(1 + \frac{2G_{8\star}^2 T + 2\sigma_{8\star}^2 T}{\gamma^2} \right).$$

The next lemma is similar to Lemma 12, while using decorrelated stepsizes. Here, knowledge of the noise bound is required to achieve a logarithmic summation.

Lemma 13. *Assuming* $\gamma^2 \geq 5\sigma_{8\star}^2 \log \frac{T}{\delta}$, *with probability at least* $1 - \delta$, *for any* $t \in [T]$,

$$\sum_{s=1}^{t} \tilde{\eta}_s^2 \|g_s - \nabla f(w_s)\|^2 \leq \alpha^2 D^2 \log(1 + T).$$

Proof of Lemma 2. As w_{s+1} is the projection of $w_s - \eta_s g_s$ onto the convex set containing $w_D^\star$,

$$\|w_{s+1} - w_D^\star\|^2 \leq \|w_s - w_D^\star\|^2 - 2\eta_s g_s \cdot (w_s - w_D^\star) + \eta_s^2 \|g_s\|^2.$$

Summing for $s = 1, \ldots, t$,

$$\|w_{t+1} - w_D^\star\|^2 \leq \|w_1 - w_D^\star\|^2 - 2\sum_{s=1}^{t} \eta_s g_s \cdot (w_s - w_D^\star) + \sum_{s=1}^{t} \eta_s^2 \|g_s\|^2.$$

Using lemma Lemma 11, with probability at least $1 - \delta$, for any $t \in [T]$,

$$\|w_{t+1} - w_D^\star\|^2 \leq \|w_1 - w_D^\star\|^2 + 16D\sqrt{\theta_{t,\delta} \sum_{s=1}^{t} \tilde{\eta}_s^2 \|\nabla f(w_s) - g_s\|^2 + \frac{\alpha^2 D^2 \sigma_{8\star}^2 \theta_{t,\delta}^2}{\gamma^2}}$$

$$+ \frac{2}{\alpha} \sum_{s=1}^{t} \tilde{\eta}_s^2 \|g_s - \nabla f(w_s)\|^2 + \left(1 + \frac{2}{\alpha}\right) \sum_{s=1}^{t} \eta_s^2 \|g_s\|^2.$$

Applying Lemmas 12 and 13 under a union bound, with probability at least $1 - 2\delta$, for any $t \in [T]$,

$$\|w_{t+1} - w_D^\star\|^2 \leq \|w_1 - w_D^\star\|^2 + 16\alpha D^2 \sqrt{\theta_{t,\delta} \log(1 + T) + \frac{\sigma_{8\star}^2 \theta_{t,\delta}^2}{\gamma^2}}$$

$$+ 2\alpha D^2 \log(1 + T) + \alpha(\alpha + 2)D^2 \log\left(1 + \frac{2G_{8\star}^2 T + 2\sigma_{8\star}^2 T}{\gamma^2}\right)$$

$$\leq \|w_1 - w_D^\star\|^2 + \frac{D^2}{8} + \frac{D^2}{16} + \frac{D^2}{16} \qquad \text{(Defs of } \alpha, \gamma \text{ at Algorithm 1 and } \alpha \leq 1)$$

$$= \|w_1 - w_D^\star\|^2 + \frac{1}{4}D^2.$$

Thus, using the triangle inequality and the fact that $\|w_1 - w^\star\| \geq \|w_1 - w_D^\star\|$,

$$\|w_{t+1} - w_1\| \leq \|w_{t+1} - w_D^\star\| + \|w_D^\star - w_1\| \leq 2\|w_1 - w_D^\star\| + \frac{1}{2}D \leq 2\|w_1 - w^\star\| + \frac{1}{2}D.$$

In addition,

$$\|w_{t+1} - w^\star\| \leq \|w_{t+1} - w_1\| + \|w_1 - w^\star\| \leq 3\|w_1 - w^\star\| + \frac{1}{2}D. \qquad \square$$

C.4. Proof of Lemma 11

Using $\tilde{\eta}_t$ we can decompose the term $-\eta_s g_t \cdot (w_s - w_D^\star)$,

$$-\eta_s g_s \cdot (w_s - w_D^\star) = -\tilde{\eta}_s g_s \cdot (w_s - w_D^\star) + (\tilde{\eta}_s - \eta_s)g_s \cdot (w_s - w_D^\star)$$

$$= -\tilde{\eta}_s \nabla f(w_s) \cdot (w_s - w_D^\star) + \tilde{\eta}_s(\nabla f(w_s) - g_s) \cdot (w_s - w_D^\star) + (\tilde{\eta}_s - \eta_s)g_s \cdot (w_s - w_D^\star)$$

$$\leq \tilde{\eta}_s(\nabla f(w_s) - g_s) \cdot (w_s - w_D^\star) + (\tilde{\eta}_s - \eta_s)g_s \cdot (w_s - w_D^\star),$$

where the last inequality follows from convexity. Thus,

$$-\sum_{s=1}^{t} \eta_s g_s \cdot (w_s - w_D^\star) \leq \sum_{s=1}^{t} \tilde{\eta}_s(\nabla f(w_s) - g_s) \cdot (w_s - w_D^\star) + \sum_{s=1}^{t} (\tilde{\eta}_s - \eta_s)g_s \cdot (w_s - w_D^\star). \qquad (3)$$

The next two lemmas bounds the two sums and concludes the proof of Lemma 11. Their proofs follows.

Lemma 14. *With probability at least $1 - \delta$, for any $t \in [T]$,*

$$\sum_{s=1}^{t} \tilde{\eta}_s (\nabla f(w_s) - g_s) \cdot (w_s - w_D^\star) \le 8D \sqrt{\theta_{t,\delta} \sum_{s=1}^{t} \tilde{\eta}_s^2 \|\nabla f(w_s) - g_s\|^2 + \frac{\alpha^2 D^2 \sigma_{8\star}^2 \theta_{t,\delta}^2}{\gamma^2}}.$$

Lemma 15. *For any $t \in [T]$,*

$$\sum_{s=1}^{t} (\tilde{\eta}_s - \eta_s) g_s \cdot (w_s - w_D^\star) \le \frac{1}{\alpha} \sum_{s=1}^{t} \eta_s^2 \|g_s\|^2 + \frac{1}{\alpha} \sum_{s=1}^{t} \tilde{\eta}_s^2 \|g_s - \nabla f(w_s)\|^2.$$

Combining both into Equation (3) yields the desired inequality. $\qquad\square$

C.5. Proof of Lemma 14

While $\|\nabla f(w_s) - g_s\| \le \sigma_{8\star}$, this bound is too crude for the required martingale analysis. To this end we use the following empirical Bernstein inequality due to Howard et al. (2021) in order to exploit the relationship between $\tilde{\eta}_t$ and $\nabla f(w_s) - g_s$.

Lemma 16 (Corollary 2 of Ivgi et al. (2023)). *Let $c > 0$ and X_t be a martingale difference sequence adapted to $\mathcal{F}_t$ such that $|X_t| \le c$ with probability 1 for all t. Then, for all $\delta \in (0, 1)$, and $\hat{X}_t \in \mathcal{F}_{t-1}$ such that $|\hat{X}_t| \le c$ with probability 1,*

$$\Pr\left(\exists t \in [T] : \left| \sum_{s=1}^{t} X_s \right| \ge 4 \sqrt{\theta_{t,\delta} \sum_{s=1}^{t} (X_s - \hat{X}_s)^2 + c^2 \theta_{t,\delta}^2} \right) \le \delta,$$

where $\theta_{t,\delta} = \log \frac{60 \log(6t)}{\delta}$.

Invoking Lemma 16 with

$$X_s = \tilde{\eta}_s (\nabla f(w_s) - g_s) \cdot (w_s - w_D^\star), \qquad \hat{X}_s = 0 \qquad \text{and} \qquad c = \frac{2\alpha D^2 \sigma_{8\star}}{\gamma},$$

(note the use of $D \le 8\|w_1 - w^\star\|$ to bound $\|\nabla f(w_s) - g_s\| \le \sigma_{8\star}$) with probability at least $1 - \delta$, for any $t \in [T]$,

$$\left| \sum_{s=1}^{t} \tilde{\eta}_s (\nabla f(w_s) - g_s) \cdot (w_s - w_D^\star) \right| \le 4 \sqrt{\theta_{t,\delta} \sum_{s=1}^{t} \left(\tilde{\eta}_s (\nabla f(w_s) - g_s) \cdot (w_s - w_D^\star) \right)^2 + \frac{4\alpha^2 D^4 \sigma_{8\star}^2 \theta_{t,\delta}^2}{\gamma^2}}.$$

By Cauchy-Schwarz inequality and the triangle inequality ($\|w_s - w_D^\star\| \le \|w_s - w_1\| + \|w_1 - w_D^\star\| \le 2D$),

$$\left(\tilde{\eta}_s (\nabla f(w_s) - g_s) \cdot (w_s - w_D^\star) \right)^2 \le \tilde{\eta}_s^2 \|\nabla f(w_s) - g_s\|^2 \|w_s - w_D^\star\|^2 \le 4\tilde{\eta}_s^2 \|\nabla f(w_s) - g_s\|^2 D^2.$$

Hence, with probability at least $1 - \delta$, for any $t \in [T]$,

$$\left| \sum_{s=1}^{t} \tilde{\eta}_s (\nabla f(w_s) - g_s) \cdot (w_s - w_D^\star) \right| \le 8D \sqrt{\theta_{t,\delta} \sum_{s=1}^{t} \tilde{\eta}_s^2 \|\nabla f(w_s) - g_s\|^2 + \frac{\alpha^2 D^2 \sigma_{8\star}^2 \theta_{t,\delta}^2}{\gamma^2}}. \qquad\square$$

C.6. Proof of Lemma 15

Let $G_t = \sqrt{\sum_{s=1}^{t} \|g_s\|^2}$. Thus,

$$|\tilde{\eta}_s - \eta_s| = \frac{\alpha D |\|g_s\|^2 - \|\nabla f(w_s)\|^2|}{\sqrt{\sigma_{8\star}^2 + G_s^2} \sqrt{\sigma_{8\star}^2 + \|\nabla f(w_s)\|^2 + G_{s-1}^2} \left(\sqrt{\sigma_{8\star}^2 + G_s^2} + \sqrt{\sigma_{8\star}^2 + \|\nabla f(w_s)\|^2 + G_{s-1}^2} \right)}.$$

By the triangle inequality,

$$\frac{|\|g_s\|^2 - \|\nabla f(w_s)\|^2|}{\sqrt{\sigma_{8\star}^2 + G_s^2} + \sqrt{\sigma_{8\star}^2 + \|\nabla f(w_s)\|^2 + G_{s-1}^2}} \le \frac{\|g_s - \nabla f(w_s)\| (\|g_s\| + \|\nabla f(w_s)\|)}{\sqrt{\sigma_{8\star}^2 + G_s^2} + \sqrt{\sigma_{8\star}^2 + \|\nabla f(w_s)\|^2 + G_{s-1}^2}}$$

$$\le \|g_s - \nabla f(w_s)\|.$$

21

Hence, by Cauchy-Schwarz inequality,

$$\sum_{s=1}^{t}(\tilde{\eta}_s - \eta_s)g_s \cdot (w_s - w_D^\star) \le \alpha D \sum_{s=1}^{t} \frac{\|g_s - \nabla f(w_s)\|\|g_s\|\|w_s - w_D^\star\|}{\sqrt{\sigma_{8\star}^2 + G_s^2}\sqrt{\sigma_{8\star}^2 + \|\nabla f(w_s)\|^2 + G_{s-1}^2}}$$

$$\le 2\alpha D^2 \sum_{s=1}^{t} \frac{\|g_s - \nabla f(w_s)\|\|g_s\|}{\sqrt{\sigma_{8\star}^2 + G_s^2}\sqrt{\sigma_{8\star}^2 + \|\nabla f(w_s)\|^2 + G_{s-1}^2}}$$

$$= \frac{2}{\alpha} \sum_{s=1}^{t} \eta_s\tilde{\eta}_s\|g_s - \nabla f(w_s)\|\|g_s\|,$$

where the second inequality follows by $\|w_s - w_D^\star\| \le 2D$. Using the identity $2ab \le a^2 + b^2$,

$$\sum_{s=1}^{t}(\tilde{\eta}_s - \eta_s)g_s \cdot (w_s - w_D^\star) \le \frac{1}{\alpha} \sum_{s=1}^{t} \eta_s^2\|g_s\|^2 + \frac{1}{\alpha} \sum_{s=1}^{t} \tilde{\eta}_s^2\|g_s - \nabla f(w_s)\|^2. \qquad \square$$

C.7. Proof of Lemmas 12 and 13

In order to prove the lemmas we need the following standard lemma used in the analysis of AdaGrad-like methods.

Lemma 17. *Let $a_0 > 0$ and $a_1, a_2, \ldots, a_n \ge 0$. Then*

$$\sum_{i=1}^{n} \frac{a_i}{\sum_{j=0}^{i} a_j} \le \ln \frac{\sum_{i=0}^{n} a_i}{a_0} \le \log \frac{\sum_{i=0}^{n} a_i}{a_0}.$$

Proof. Using the inequality $1 - x \le \ln(1/x)$ for $x > 0$,

$$\sum_{i=1}^{n} \frac{a_i}{\sum_{j=0}^{i} a_j} = \sum_{i=1}^{n}\left(1 - \frac{\sum_{j=0}^{i-1} a_j}{\sum_{j=0}^{i} a_j}\right) \le \sum_{i=1}^{n} \ln\left(\frac{\sum_{j=0}^{i} a_j}{\sum_{j=0}^{i-1} a_j}\right) = \ln\left(\frac{\sum_{j=0}^{n} a_j}{a_0}\right) \le \log\left(\frac{\sum_{j=0}^{n} a_j}{a_0}\right),$$

where the last inequality holds since $\ln(\sum_{j=0}^{n} a_j/a_0) \ge 0$. $\qquad \square$

Proof of Lemma 12. Using Lemma 17 with $a_0 = \gamma^2$ and $a_i = \|g_i\|^2$,

$$\sum_{s=1}^{t} \eta_s^2\|g_s\|^2 = \alpha^2 D^2 \sum_{s=1}^{t} \frac{\|g_s\|^2}{\gamma^2 + \sum_{k=1}^{s}\|g_k\|^2} \le \alpha^2 D^2 \log\left(1 + \frac{\sum_{s=1}^{t}\|g_s\|^2}{\gamma^2}\right).$$

By the inequality $\|u + v\|^2 \le 2\|u\|^2 + \|v\|^2$, the Lipschitz assumption and the noise assumption, $\|g_s\|^2 \le 2G_{8\star}^2 + 2\sigma_{8\star}^2$. Thus,

$$\sum_{s=1}^{t} \eta_s^2\|g_s\|^2 \le \alpha^2 D^2 \log\left(1 + \frac{2G_{8\star}^2 T + 2\sigma_{8\star}^2 T}{\gamma^2}\right). \qquad \square$$

Proof of Lemma 13. Using lemma Lemma 6, with probability at least $1 - \frac{\delta}{T}$, for $\lambda = \frac{2}{3\sigma_{8\star}^2}$,

$$\sum_{s=1}^{t-1} \nabla f(w_s) \cdot (\nabla f(w_s) - g_s) \le \frac{1}{2}\sum_{s=1}^{t-1}\|\nabla f(w_s)\|^2 + \frac{3}{2}\sigma_{8\star}^2 \log \frac{T}{\delta}.$$

Thus, with probability at least $1 - \frac{\delta}{T}$,

$$\sum_{s=1}^{t-1}\|g_s\|^2 = \sum_{s=1}^{t-1}(\|\nabla f(w_s)\|^2 + \|g_s - \nabla f(w_s)\|^2 - 2\nabla f(w_s) \cdot (\nabla f(w_s) - g_s))$$

$$\ge \sum_{s=1}^{t-1}\|g_s - \nabla f(w_s)\|^2 - 3\sigma_{8\star}^2 \log \frac{T}{\delta},$$

and since $\gamma^2 \geq 5\sigma_{8\star}^2 \log \frac{T}{\delta}$ and $\|g_t - \nabla f(w_t)\| \leq \sigma_{8\star}$,

$$\tilde{\eta}_s \leq \frac{\alpha D}{\sqrt{5\sigma_{8\star}^2 \log \frac{T}{\delta} + \Sigma_{s=1}^{t-1}\|g_s\|^2}} \leq \frac{\alpha D}{\sqrt{2\sigma_{8\star}^2 \log \frac{T}{\delta} + \Sigma_{s=1}^{t-1}\|g_s - \nabla f(w_s)\|^2}}$$

$$\leq \frac{\alpha D}{\sqrt{\sigma_{8\star}^2 + \Sigma_{s=1}^{t}\|g_s - \nabla f(w_s)\|^2}}.$$

Hence, under a union bound, with probability at least $1 - \delta$, for any $t \in [T]$,

$$\tilde{\eta}_s \leq \frac{\alpha D}{\sqrt{\sigma_{8\star}^2 + \Sigma_{s=1}^{t}\|g_s - \nabla f(w_s)\|^2}}$$

and using Lemma 17 with $a_0 = \sigma_{8\star}^2$ and $a_i = \|g_i - \nabla f(w_i)\|^2$,

$$\sum_{s=1}^{t} \tilde{\eta}_s^2\|g_s - \nabla f(w_s)\|^2 \leq \alpha^2 D^2 \sum_{s=1}^{t} \frac{\|g_s - \nabla f(w_s)\|^2}{\sigma_{8\star}^2 + \Sigma_{s=1}^{t}\|g_s - \nabla f(w_s)\|^2}$$

$$\leq \alpha^2 D^2 \log\left(\frac{\sigma_{8\star}^2 + \Sigma_{s=1}^{t}\|g_s - \nabla f(w_s)\|^2}{\sigma_{8\star}^2}\right) \leq \alpha^2 D^2 \log(1 + T). \qquad \square$$

C.8. Proof of Lemma 3

Let

$$X_s = (\nabla f(w_s) - g_s) \cdot (w_s - w^\star), \qquad \hat{X}_s = 0, \qquad \text{and} \qquad c = \sigma_{8\star}(D + \|w_1 - w^\star\|).$$

Note that by Cauchy-Schwarz inequality and the triangle inequality,

$$(\nabla f(w_s) - g_s) \cdot (w_s - w^\star) \leq \|\nabla f(w_s) - g_s\|\|w_s - w^\star\| \leq \|\nabla f(w_s) - g_s\|(\|w_s - w_1\| + \|w_1 - w^\star\|)$$

$$\leq \sigma_{8\star}(D + \|w_1 - w^\star\|) = c,$$

where we used the assumption that $D \leq 8\|w_1 - w^\star\|$ to bound $\|\nabla f(w_s) - g_s\|$. Applying Lemma 16, with probability at least $1 - \delta$,

$$\left|\sum_{t=1}^{T}(\nabla f(w_t) - g_t) \cdot (w_s - w^\star)\right| \leq 4(D + \|w_1 - w^\star\|)\sqrt{\theta_{T,\delta}\sum_{t=1}^{T}\|\nabla f(w_t) - g_t\|^2 + \sigma_{8\star}^2\theta_{T,\delta}^2}.$$

Applying Lemma 6 with $Z_t = \nabla f(w_t) \cdot (\nabla f(w_t) - g_t)$ and $\lambda = \frac{2}{3\sigma_{8\star}^2}$, with probability at least $1 - \delta$

$$\sum_{t=1}^{T} \nabla f(w_t) \cdot (\nabla f(w_t) - g_t) \leq \frac{3\sigma_{8\star}^2}{4}\lambda \sum_{t=1}^{T}\|\nabla f(w_t)\| + \frac{\log \frac{1}{\delta}}{\lambda} = \frac{1}{2}\sum_{t=1}^{T}\|\nabla f(w_t)\| + \frac{3\sigma_{8\star}^2 \log \frac{1}{\delta}}{2}.$$

Hence,

$$\sum_{t=1}^{T}\|\nabla f(w_t) - g_t\|^2 = \sum_{t=1}^{T}\left(\|g_t\|^2 + 2\nabla f(w_t) \cdot (\nabla f(w_t) - g_t) - \|\nabla f(w_t)\|^2\right)$$

$$\leq 3\sigma_{8\star}^2 \log \frac{1}{\delta} + \sum_{t=1}^{T}\|g_t\|^2.$$

Returning to the previous inequality, with probability at least $1 - 2\delta$ (union bound),

$$\left|\sum_{t=1}^{T}(\nabla f(w_t) - g_t) \cdot (w_s - w^\star)\right| \leq 4(D + \|w_1 - w^\star\|)\sqrt{\theta_{T,\delta}\sum_{t=1}^{T}\|g_t\|^2 + \theta_{T,\delta}\left(3\log \frac{1}{\delta} + \theta_{T,\delta}\right)\sigma_{8\star}^2}$$

$$\leq 4(D + \|w_1 - w^\star\|)\sqrt{\theta_{T,\delta}\sum_{t=1}^{T}\|g_t\|^2 + 4\theta_{T,\delta}^2\sigma_{8\star}^2}$$

where the last inequality follows by $\theta_{T,\delta} \geq \log \frac{1}{\delta}$. Thus, $\Pr(E_2) \geq 1 - 2\delta$. $\qquad \square$

C.9. Proof of Theorem 5

In order to show that Theorem 4 is applicable we need to show that $G_{\max} \geq G_{8\star}$. Let $w \in \mathbb{R}^d$ such that $\|w - w_1\| \leq 8\|w_1 - w^\star\|$. Thus, by the smoothness of f and the triangle inequality,

$$\|\nabla f(w)\| \leq \beta_\star \|w - w^\star\| \leq \beta_\star(\|w - w_1\| + \|w_1 - w^\star\|) \leq 9\beta_\star \|w_1 - w^\star\| \leq G_{\max}.$$

Hence, by Equation (1), $G_{\max} \geq G_{8\star}$. Using Theorem 4, with probability at least $1 - \delta \log \frac{D_{\max}}{D_{\min}}$, we have a sequence $w_1, \ldots, w_T$ with

$$\frac{1}{T} \sum_{t=1}^{T} f(w_t) - f^\star \leq C\left(\frac{\log(\frac{T}{\delta} + \frac{G_{\max}}{\sigma_{\max}})\|w_1 - w^\star\|}{T}\left(\sqrt{\sum_{t=1}^{T}\|g_t\|^2} + \sigma_{\max}\sqrt{\log \frac{T}{\delta}}\right)\right) \tag{4}$$

for some $C > 0$, and for all t, $\|w_t - w_1\| \leq 8\|w_1 - w^\star\|$. By the identity $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and Equation (1),

$$\|g_t\|^2 \leq 2\|\nabla f(w_t)\|^2 + 2\|g_t - \nabla f(w_t)\|^2 \leq 2\|\nabla f(w_t)\|^2 + 2\sigma_{8\star}^2 \leq 4\beta_\star(f(w_t) - f^\star) + 2\sigma_{8\star}^2.$$

where the last inequality is due to the following standard smoothness argument. For any $w \in \mathbb{R}^d$ and $w^+ = w - \frac{1}{\beta_\star}\nabla f(w)$, by the smoothness of f,

$$2\beta_\star(f(w) - f^\star) \geq 2\beta_\star(f(w) - f(w^+)) \geq 2\beta_\star(-\nabla f(w) \cdot (w^+ - w) - \frac{\beta_\star}{2}\|w^+ - w\|^2) = \|\nabla f(w)\|^2.$$

Hence, by the identities $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$ and $ab \leq \frac{1}{2\lambda}a^2 + \frac{\lambda}{2}b^2$ for $\lambda > 0$,

$$\sqrt{\sum_{t=1}^{T}\|g_t\|^2} \leq \sqrt{4\beta_\star \sum_{t=1}^{T}(f(w_t) - f^\star) + \sigma_{8\star}\sqrt{2T}} \leq \frac{1}{2\lambda}\sum_{t=1}^{T}(f(w_t) - f^\star) + 2\lambda\beta_\star + \sigma_{8\star}\sqrt{2T}.$$

Setting $\lambda = C\|w_1 - w^\star\|\log(\frac{T}{\delta} + \frac{G_{\max}}{\sigma_{\max}})$, returning to Equation (4) and rearranging,

$$\frac{1}{T}\sum_{t=1}^{T}f(w_t) - f^\star \leq \frac{4C^2\beta_\star\|w_1 - w^\star\|^2\log^2(\frac{T}{\delta} + \frac{G_{\max}}{\sigma_{\max}})}{T} + \frac{\sqrt{8}C\|w_1 - w^\star\|\sigma_{8\star}\log(\frac{T}{\delta} + \frac{G_{\max}}{\sigma_{\max}})}{\sqrt{T}}$$

$$+ \sigma_{\max}\frac{2C\|w_1 - w^\star\|\log(\frac{T}{\delta} + \frac{G_{\max}}{\sigma_{\max}})\sqrt{\log\frac{T}{\delta}}}{T}.$$

We conclude by a standard application of Jensen's inequality. □

C.10. Proof of Theorem 6

We will define two problem instances $(f_1, \widetilde{g}_1, D_1, G_1), (f_2, \widetilde{g}_2, D_2, G_2)$, which will be indistinguishable with constant probability using only T gradient queries, where no solution $\overline{w} \in \mathbb{R}$ can satisfy the convergence bound for both. Let $\sigma_1 = \frac{\sigma_{\max}}{\alpha\sqrt{T}}$, $D_1 = D_{\max}$, $\sigma_2 = \sigma_{\max}$, $D_2 = \frac{D_{\max}}{\alpha\sqrt{T}}$. We define the two functions,

$$f_1(w) = G_\star|w - D_1| \quad \text{and} \quad f_2(w) = G_\star|w + D_2|.$$

Note that both functions are $G_\star$-Lipschitz and their minimizers satisfy $\|\arg\min_{w \in \mathbb{R}} f_i(w)\| \leq D_i$ for $i \in \{1, 2\}$. In addition, we define the following two first-order oracles,

$$\widetilde{g}_1(w) = \begin{cases} -G_\star & \text{if } w \leq -D_2; \\ G_\star & \text{if } w \geq D_1; \\ -\sigma_1 + G_\star & \text{if } w \in (-D_2, D_1), \text{ w.p. } \frac{1}{2} + \frac{G_\star}{2(\sigma_1 - G_\star)}; \\ \sigma_1 - G_\star & \text{if } w \in (-D_2, D_1), \text{ w.p. } \frac{1}{2} - \frac{G_\star}{2(\sigma_1 - G_\star)} \end{cases}$$

and

$$\widetilde{g}_2(w) = \begin{cases} -G_\star & \text{if } w \leq -D_2; \\ G_\star & \text{if } w \geq D_1; \\ \sigma_2 + G_\star & \text{if } w \in (-D_2, D_1), \text{ w.p. } \frac{1}{T}; \\ -\sigma_1 + G_\star & \text{if } w \in (-D_2, D_1), \text{ w.p. } \frac{1}{2} + \frac{G_\star}{2(\sigma_1 - G_\star)} - \frac{1}{2T}; \\ \sigma_1 - G_\star & \text{if } w \in (-D_2, D_1), \text{ w.p. } \frac{1}{2} - \frac{G_\star}{2(\sigma_1 - G_\star)} - \frac{1}{2T}. \end{cases}$$

Note that the probabilities are within $[0, 1]$ (and sum to 1) since $\frac{1}{2T} \leq \frac{1}{8}$ as $T \geq 4$, and for $\alpha \in [1, \frac{3}{4}\sqrt{T}]$,

$$\frac{G_\star}{2(\sigma_1 - G_\star)} = \frac{\sigma_{\max}/(2T-1)}{2(\sigma_{\max}/\alpha\sqrt{T} - \sigma_{\max}/(2T-1))} = \frac{\alpha\sqrt{T}}{2(2T-1-\alpha\sqrt{T})} \leq \frac{3T}{10T-8} \leq \frac{3}{8}.$$

Additionally, for $w \in (-D_2, D_1)$,

$$\mathbb{E}[\widetilde{g}_1(w)] = (\sigma_1 - G_\star)\left(\frac{1}{2} - \frac{G_\star}{2(\sigma_1 - G_\star)} - \frac{1}{2} - \frac{G_\star}{2(\sigma_1 - G_\star)}\right) = -\frac{G_\star(\sigma_1 - G_\star)}{\sigma_1 - G_\star} = -G_\star = \nabla f_1(w)$$

and similarly

$$\mathbb{E}[\widetilde{g}_2(w)] = \frac{\sigma_2 + G_\star}{T} - G_\star = \frac{(2T-1)G_\star + G_\star}{T} - G_\star = G_\star = \nabla f_2(w).$$

Hence, for all $w \in \mathbb{R}$, $\mathbb{E}[\widetilde{g}_1(w)] = \nabla f_1(w)$, $\mathbb{E}[\widetilde{g}_2(w)] = \nabla f_2(w)$, and with probability 1, as $G_\star < \sigma_1 < \sigma_2$,

$$\|\widetilde{g}_1(w) - \nabla f_1(w)\| \leq \sigma_1 \quad \text{and} \quad \|\widetilde{g}_2(w) - \nabla f_2(w)\| \leq \sigma_2.$$

We concluded the properties of f and $\widetilde{g}$ and we move to the lower bound. For $n \geq 4$,

$$\left(1 - \frac{1}{n}\right)^n = \left(\frac{n}{n-1}\right)^{-n} \geq \frac{3}{4}\left(\frac{n}{n-1}\right)^{-(n-1)} = \frac{3}{4}\frac{1}{\left(1 + \frac{1}{n-1}\right)^{n-1}} \geq \frac{1}{4},$$

where the last inequality follows by

$$\left(1 + \frac{1}{n-1}\right)^{n-1} = 2 + \sum_{k=2}^{n-1}\binom{n-1}{k}(n-1)^{-k} \leq 2 + \sum_{k=2}^{n-1}\frac{1}{k(k-1)} = 2 + \sum_{k=2}^{n-1}\left(\frac{1}{k-1} - \frac{1}{k}\right) \leq 3.$$

Hence, with probability at least $\frac{1}{4}$, the $1/T$ event of $\widetilde{g}_2$ do not occur in any of the T queries and the two gradient oracles will be indistinguishable from each other. Let $\overline{w}$ be the output of $\mathcal{A}(T)$ given that the two oracles return the exact same gradients. We will assume by contradiction that $\overline{w}$ satisfy a convergence rate better than $D_i(G_\star + \sigma_i)\alpha/6\sqrt{T}$ for each f_i ($i \in \{1, 2\}$ respectively). Thus, as $\sigma_1 = \frac{G_\star(2T-1)}{\alpha\sqrt{T}}$ and $\alpha < \sqrt{T}$,

$$f_1(\overline{w}) - f_1(w^\star) = G_\star|\overline{w} - D_1| = \frac{D_1(G_\star + \sigma_1)\alpha}{\sqrt{T}} \cdot \frac{G_\star\sqrt{T}|\overline{w} - D_1|}{D_1(G_\star + \sigma_1)\alpha}$$

$$= \frac{D_1(G_\star + \sigma_1)\alpha}{\sqrt{T}} \cdot \frac{T|\overline{w} - D_1|}{D_1(\alpha\sqrt{T} + 2T - 1)} \geq \frac{D_1(G_\star + \sigma_1)\alpha}{\sqrt{T}} \cdot \frac{|\overline{w} - D_1|}{3D_1}$$

which implies that $\overline{w} > D_1/2$ by the convergence assumption. Similarly, as $\sigma_2 = G_\star(2T-1)$,

$$f_2(\overline{w}) - f_2(w^\star) = G_\star|\overline{w} + D_2| = \frac{D_2(G_\star + \sigma_2)\alpha}{\sqrt{T}} \cdot \frac{G_\star\sqrt{T}|\overline{w} + D_2|}{D_2(G_\star + \sigma_2)\alpha} = \frac{D_2(G_\star + \sigma_2)\alpha}{\sqrt{T}} \cdot \frac{|\overline{w} + D_2|}{2D_2\sqrt{T}\alpha}$$

which implies that $\overline{w} < (\alpha\sqrt{T}/3 - 1)D_2$ by the convergence assumption. But

$$\overline{w} > \frac{D_1}{2} = \frac{\alpha\sqrt{T}D_2}{2} > (\alpha\sqrt{T}/3 - 1)D_2 > \overline{w}$$

and we obtain a contradiction. The second inequality of the convergence lower bound follows by $(G_\star + \sigma)\alpha \geq \sigma_{\max}(\frac{\alpha}{2T-1} + \frac{1}{\sqrt{T}}) \geq \frac{\sigma_{\max}}{\sqrt{T}}$ for $\sigma \in \{\sigma_1, \sigma_2\}$. $\qquad \square$

D. Assumption for Noise Upper Bound

In the convex non-smooth setting, Theorem 6 implies that a stochastic convex optimization method cannot be parameter-free with respect to all three parameters, diameter, Lipschitz bound and noise bound, while being competitive with tuned SGD. Next, we observe that while such a term is unavoidable without further assumptions, if a success probability of $1 - O(\delta)$ is desired, the following reasonable assumption is sufficient.

Assumption: for some $\delta \in (0, 1)$ and $c = 4\sqrt{1 + 3 \log \frac{1}{\delta}}$,

$$f(w_1) - f^\star > \frac{c\sigma_\star \|w_1 - w^\star\|}{\sqrt{T}}.$$

In other words, w_1 is a "bad enough" approximate minimizer.

Let us first explain why this is a reasonable assumption. Consider the convergence rate of tuned SGD, which ensures with probability at least $1 - \delta$ that

$$f(\overline{w}) - f^\star \le \frac{C\|w_1 - w^\star\|(G_\star + \sigma_\star \sqrt{\log \frac{1}{\delta}})}{\sqrt{T}}$$

for some constant $C > 0$. In that case, the assumption is weaker then assuming that the above convergence bound is sufficient to ensure that with probability at least $1 - \delta$, SGD produce a solution better than w_1 by some constant factor. If we aim to compete with tuned SGD, it is reasonable to focus on the regime where SGD produce better solutions than the initialization.

We move to analyze what the assumption above yields. Using the assumption and the convexity of the objective,

$$\frac{c\sigma_\star \|w_1 - w^\star\|}{\sqrt{T}} < f(w_1) - f^\star \le \nabla f(w_1) \cdot (w_1 - w^\star) \le \|\nabla f(w_1)\|\|w_1 - w^\star\|.$$

Thus,

$$\sigma_\star < \frac{\|\nabla f(w_1)\|\sqrt{T}}{c}.$$

So upper bounding $\|\nabla f(w_1)\|$ produce an upper bound of $\sigma_\star$. The simplest approach to estimate the gradient at $w_1 \in \mathbb{R}^d$ is to average n stochastic samples. Let $Z = \frac{1}{n}\sum_{i=1}^n \widetilde{g}_i(w_1)$ where $\widetilde{g}_i(w_1)$ for $i \in [n]$ are independent stochastic gradients. Thus, $\mathbb{E}[Z] = \nabla f(w_1)$. By Lemma 7,

$$\Pr\left(\|Z - \nabla f(w_1)\| \ge \sqrt{2}(1 + \lambda)\sigma/\sqrt{n}\right) \le \exp(-\lambda^2/3).$$

Let $\lambda = \sqrt{3 \log \frac{1}{\delta}}$. Hence, with probability at least $1 - \delta$,

$$\|Z - \nabla f(w_1)\| < \sqrt{2}(1 + \sqrt{3 \log \frac{1}{\delta}}\sigma)/\sqrt{n} < \sqrt{2}(1 + \sqrt{3 \log \frac{1}{\delta}})\|\nabla f(w_1)\|\sqrt{T}/c\sqrt{n}.$$

Setting $n = 8(1 + 3 \log \frac{1}{\delta})T/c^2 = \frac{T}{2}$, with probability at least $1 - \delta$,

$$\frac{\|Z\|}{\|\nabla f(w_1)\|} \in [\frac{1}{2}, \frac{3}{2}].$$

Hence, using $\frac{T}{2}$ steps we can estimate $\|\nabla f(w_1)\|$ and obtain a bound

$$\sigma_\star < \frac{\|\nabla f(w_1)\|\sqrt{T}}{c} \le \frac{2\|Z\|\sqrt{T}}{4\sqrt{1 + 3 \log \frac{1}{\delta}}}$$

which holds with probability at least $1 - \delta$, and such bound is tight enough to make Theorem 6 inapplicable. Further note that having a bound $\sigma_\star \le \sigma_{\max} = O(G_\star \sqrt{T})$, which we obtain with high probability using this technique, is sufficient to reduce the unavoidable lower order term presented at Theorem 4 to a term proportional to the (almost) optimal rate of tuned SGD.