Making SGD Parameter-Free

Yair Carmon YCARMON@TAUEX.TAU.AC.IL
Tel Aviv University

Oliver Hinder OHINDER@PITT.EDU
University of Pittsburgh

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We develop an algorithm for parameter-free stochastic convex optimization (SCO) whose rate of convergence is only a double-logarithmic factor larger than the optimal rate for the corresponding known-parameter setting. In contrast, the best previously known rates for parameter-free SCO are based on online parameter-free regret bounds, which contain unavoidable excess logarithmic terms compared to their known-parameter counterparts. Our algorithm is conceptually simple, has high-probability guarantees, and is also partially adaptive to unknown gradient norms, smoothness, and strong convexity. At the heart of our results is a novel parameter-free certificate for SGD step size choice, and a time-uniform concentration result that assumes no a-priori bounds on SGD iterates.

1. Introduction

Stochastic convex optimization (SCO) is a cornerstone of both the theory and practice of machine learning. Consequently, there is intense interest in developing SCO algorithms that require little to no prior knowledge of the problem parameters, and hence little to no tuning [27, 23, 20, 2, 22, 39]. In this work we consider the fundamental problem of non-smooth SCO (in a potentially unbounded domain) and seek methods that are adaptive to a key problem parameter: the initial distance to optimality.

Current approaches for tackling this problem focus on the more general online learning problem of *parameter-free regret minimization* [8, 10–12, 21, 24, 25, 30, 32, 37], where the goal is to to obtain regret guarantees that are valid for comparators with arbitrary norms. Research on parameter-free regret minimization has lead to practical algorithms for stochastic optimization [9, 27, 32], methods that are able to adapt to many problem parameters simultaneously [37] and methods that can work with any norm [12]. In the basic Euclidean setting with 1-Lipschitz losses where only the initial distance to optimality is unknown, there are essentially matching upper [24] and lower bounds [26], showing that the best achievable parameter-free *average* regret scales as

$$O\left(\|\mathring{x}\|\sqrt{\frac{1}{T}\log\left(\frac{T\|\mathring{x}\|^2}{\varepsilon^2}+1\right)}+\frac{\varepsilon}{T}\right) \tag{1}$$

where T is the number of steps, $\|\mathring{x}\|$ is the (Euclidean) comparator norm, and $\varepsilon > 0$ represents the (user-chosen) regret we will incur even if the comparator norm is zero. This is larger by a logarithmic factor than the optimal average-regret when the comparator norm is known in advance.

Parameter-free regret bounds immediately translate into parameter-free SCO algorithms using online-to-batch conversion [15]. The expected optimality gap bound of the resulting algorithm is identical to (1) when we replace $\mathring{x}$ by $x_\star - x_0$, i.e., the difference between the optimum and the

initial point. This bound is a logarithmic factor worse than what stochastic gradient descent (SGD) can achieve when we know the distance to optimality and use it to compute step sizes. While this logarithmic factor is unavoidable for regret minimization, it is unclear if it is necessary for SCO.

In this paper we show it is possible to obtain stronger parameter-free rates for SCO by moving beyond the regret minimization abstraction. In particular, for any $\varepsilon > 0$ and $\delta \in (0, 1)$, we a obtain probability $1 - \delta$ optimality gap bounds of

$$O\left(\left(\frac{\|x_0 - x_\star\|}{\sqrt{T}} + \frac{\varepsilon}{T}\right) \log^2\left(\frac{1}{\delta} \log_+ \frac{T\|x_0 - x_\star\|}{\varepsilon}\right)\right),$$

which is better than any bound achievable by online-to-batch conversion. While replacing the logarithmic factor by a double-logarithmic factor may appear a small improvement, we consider it important due to the fundamental nature of the problem as well as the theoretical separation it establishes between parameter-free SCO and OCO. Such separations are rare in the literature; we are only aware of one prior example [17].

Our method also provides high probability guarantees on the suboptimality gap. This resolves an open problem in parameter-free optimization; see [27] and [29, §7]. We are able to form high probability bounds because, unlike other parameter-free SCO algorithms, we prove a strong localization guarantee: our output $\bar{x}$ satisfies $\|\bar{x} - x_\star\| = O(\|x_0 - x_\star\|)$, and key intermediate points satisfy a similar bound as well. We suspect that such localization is difficult to establish with online-to-batch conversion, since online parameter-free algorithms may need to let their iterates fluctuate wildly in order to handle difficult adversaries.

In addition to independence of $\|x_0 - x_\star\|$, our algorithm exhibits three additional forms of adaptivity. First, our algorithm has adaptivity to gradient norms on par with the best existing parameter-free result [12]: the leading term of our bounds scales with a sum of squared observed gradient norms, and an a-priori gradient norm bound only affects low-order terms. Second, as a consequence, in the smooth and noiseless case our algorithm exhibits a $\frac{\log \log T}{T}$ rate of convergence. Finally, via a simple restart scheme we obtain the optimal rate for strongly-convex stochastic problems (up to double-logarithmic factors), without knowledge of the strong-convexity parameter.

On a technical level, our development differs significantly from prior parameter-free optimization methods. While online methods rely on advanced tools such as coin betting [30], and online Newton steps [12], our approach is essentially a careful scheme for correctly setting the step size of SGD. Underlying our algorithm is a parameter-free certificate for SGD, which implies both localization and optimality gap bounds. The certificate takes the form of an implicit equation over the SGD step size, which we solve via bisection on the logarithm of the step size. To obtain high-probability bounds, we develop a time-uniform empirical-Berstein-type concentration bound independent of any a-priori assumptions on the iterate norms. Given the ubiquity of SGD in practice and in the classroom, our insights on how choose its step size may be of independent interest.

Paper organization. In the following subsections we review additional related work, as well as the problem setup and notation. Section 2 develops our parameter-free step size certificate. Section 3 presents our algorithm and its analysis in the noiseless regime. Section 4 lifts the analysis to the stochastic setting, proving our main result on parameter-free SCO. Finally, Section 5 shows how our method adapts to smoothness and (via restarts) to strong convexity.

1.1. Additional related work

Parameter-free methods from deterministic optimization. The literature on noiseless optimization also offers a rich variety of parameter-free algorithms. In the smooth setting, the Armijo rule [1] is a standard technique for choosing step sizes for gradient descent. Using variants of this idea combined with acceleration, achieves essentially optimal and parameter-free rates of convergence [3]. The Polyak step size rule [34] simultaneously achieves optimal rates for smooth, non-smooth and strongly-convex optimization [16], but requires knowledge of the optimal function value. This requirement can be relaxed, making the Polyak method parameter-free, but at the cost of a multiplicative logarithmic factor to its bound [16]. Consequently, non-smooth parameter-free deterministic optimization appears to be as hard as SCO. Multiple works generalize line-search and the Polyak method to the stochastic setting [35, 4, 2, 22, 39, 40, 7], but do not obtain parameter-free rates in the sense we consider here.

Limitations of online-to-batch conversion. To the best of our knowledge, the only previous example of an SCO rate that is provably unachievable by online to batch conversion of a (uniform) regret bound occurs for strongly-convex optimization. Specifically, any online strongly-convex optimization algorithm must have logarithmic regret (implying suboptimality $(\log T)/T$ via online to batch conversion) [38, 17], while Hazan and Kale [17] and others [19, 14, 36] have achieved the optimal $1/T$ rate for stochastic strongly-convex optimization. The variant of our algorithm in Section 5.2 is based on the Epoch-SGD algorithm of [17], and simultantiously breaks both regret minization barriers, achieving optimality gap $(\log\log T)/T$ for parameter-free strongly-convex stochastic optimization with high probability.

Grid search. In practice, the standard technique for selecting the step size of SGD (and hyperparameters more broadly) is grid search [33, 13]. This typically consists of testing all step sizes on a geometrically spaced grid and choosing the one with the best performance on a held out set. Compared to our method, such grid search is computationally wasteful, as it tests exponentially more steps sizes than we do. Moreover, in the context of parameter-free SCO, proving guarantees for grid search is surprisingly difficult, since it is unclear how to bound the objective value estimation error for points that may be arbitrarily far apart.

1.2. Problem setup and notation

Let us briefly review the standard SCO setup, building up our notation along the way. Our goal is to minimize a convex objective function $f : \mathcal{X} \to \mathbb{R}$ defined on a closed an convex set $\mathcal{X} \subseteq \mathbb{R}^d$ (our results hold for $\mathcal{X} = \mathbb{R}^d$ as well). We let $x_\star$ denote a fixed minimizer of f, i.e., such that $f(x_\star) \leq f(x)$ for all $x \in \mathcal{X}$, implicitly assuming such point exists (see Appendix D.1 for further discussion of this assumption). We assume that our only access to f is via a stochastic gradient oracle $\mathcal{O}$ that, upon receiving query point x, returns a vector $\mathcal{O}(x) \in \mathbb{R}^d$ that is a subgradient of f in expectation, i.e., $\mathbb{E}[\mathcal{O}(x) \mid x] \in \partial f(x)$. With slight abuse of notation, we write $\nabla f(x) \coloneqq \mathbb{E}[\mathcal{O}(x) \mid x]$, corresponding to the gradient of f when it is differentiable and a particular subgradient otherwise. We interchangeably use *exact gradients*, *noiseless*, and *deterministic* to refer to the regime where $\mathcal{O}(x) = \nabla f(x)$ with probability 1.

Our development revolves around the classical fixed step size stochastic gradient descent (SGD) algorithm. Given step size η and initialization x_0, SGD iterates

$$x_{i+1}(\eta) = \Pi_{\mathcal{X}}\big(x_i(\eta) - \eta g_i(\eta)\big) \;\text{ where }\; g_i(\eta) \coloneqq \mathcal{O}(x_i(\eta)),$$

and $\Pi_{\mathcal{X}}$ is the Euclidean projection onto $\mathcal{X}$; we intentionally feature the η dependence of $x_i(\eta)$ and $g_i(\eta)$ prominently. We define the following quantities associated with the SGD iterates. First, we write the distance to $x_\star$ and its running maximum as

$$d_t(\eta) \coloneqq \|x_t(\eta) - x_\star\| \;\text{ and }\; \bar{d}_t(\eta) \coloneqq \max_{i \le t} \|x_i(\eta) - x_\star\|.$$

Replacing $x_\star$ with x_0 in the above definitions, we write

$$r_t(\eta) \coloneqq \|x_0 - x_t(\eta)\| \;\text{ and }\; \bar{r}_t(\eta) \coloneqq \max_{i \le t} \|x_0 - x_i(\eta)\|. \tag{2}$$

Finally, we denote the running sum of squared gradient norms and gradient oracle error by

$$G_t(\eta) \coloneqq \sum_{i<t} \|g_i(\eta)\|^2 \;\text{ and }\; \Delta_i \coloneqq g_i(\eta) - \nabla f(x_i(\eta)). \tag{3}$$

Additional notational conventions. Throughout, $\|\cdot\|$ denotes the Euclidean norm. We use $\log$ to denote the base 2 logarithm, and write $\log_+(x) \coloneqq \max\{2, \log(x)\}$ to simplify $O(\cdot)$ notation. For any particular value of η, the quantities $x_i(\eta)$, $g_i(\eta)$, etc. always refer to a *single* realization of the random process they represent.

2. A parameter-free step-size selection criterion for SGD

In this section we present the key component of our development: a computable certificate for the efficiency of a candidate SGD step size. For ease of exposition, in this section we restrict some of our arguments to the exact gradient setting, but emphasize that they ultimately translate to high-probability bounds in the stochastic setting.

Consider the noiseless setting with step size η, iterates $x_0, x_1(\eta), \dots, x_T(\eta)$ and gradients g_0, $g_1(\eta), \dots, g_{T-1}(\eta)$. It is well-known [29] that if η satisfies

$$\eta = \phi_{\text{ideal}}(\eta) \;\text{ where }\; \phi_{\text{ideal}}(\eta) \coloneqq \frac{\|x_0 - x_\star\|}{\sqrt{\sum_{i<T} \|g_i(\eta)\|^2}} = \frac{d_0}{\sqrt{G_T(\eta)}}$$

then the iterate average $\bar{x}(\eta) = \frac{1}{T}\sum_{i<T} x_i(\eta)$ satisfies the optimal error bound $f(\bar{x}) - f(x_\star) \le d_0\sqrt{G_T(\eta)}/T$, scaling as $O(d_0 L/\sqrt{T})$ when $\|g_i(\eta)\| \le L$ for all i. However, the quantity ϕ_{ideal}, which we call the "ideal step size" is not computable even in hindsight (when $G_T(\eta)$ is available), since the parameter $d_0 = \|x_0 - x_\star\|$ is unknown.

Our key proposal is to approximate the distance to the optimum d_0 with a computable proxy: the maximum distance traveled by the algorithm, $\bar{r}_T(\eta) \coloneqq \max_{i \le T} \|x_0 - x_i(\eta)\|$. We consider step sizes that (approximately) satisfy

$$\eta = \phi(\eta) \;\text{ where }\; \phi(\eta) \coloneqq \frac{\bar{r}_T(\eta)}{\sqrt{\alpha G_T(\eta) + \beta}} \tag{4}$$

for nonnegative damping parameters α and β; in the exact gradient setting we can set α to any number > 1 and $\beta = 0$, while the stochastic setting requires scaling α and β roughly as $\text{poly}(\log\log T)$. Intuitively, $\bar{r}_T$ approximates d_0 since the SGD iterates should converge to $x_\star$ and therefore $\|x_0 - x_T(\eta)\|$ should be similar to $\|x_0 - x_\star\|$. However, in non-smooth optimization, convergence to $x_\star$ can be arbitrarily slow. We nevertheless prove that, when $\eta \le \phi(\eta)$, we have $\bar{r}_T(\eta) = O(d_0)$ (Lemma 3 below). With this result and a refined SGD error bound (Lemma 2 below), we show that (with exact gradients) any η satisfying criterion (4) recovers the optimal error bound.

Proposition 1 *In the noiseless setting, any step size $\eta > 0$ satisfying (4) with $\alpha > 1$ and $\beta = 0$ produces $\bar{x} \coloneqq \frac{1}{T}\sum_{i<T} x_i(\eta)$ such that $\|x_\star - \bar{x}\| \leq \frac{2\alpha}{\alpha-1}\|x_\star - x_0\|$ and*

$$f(\bar{x}) - f(x_\star) \leq \frac{\alpha^{3/2}}{\alpha - 1} \cdot \frac{d_0\sqrt{G_T(\eta)}}{T} = O\left(\frac{d_0\sqrt{G_T(\eta)}}{T}\right).$$

Before proving Proposition 1, let us briefly discuss its algorithmic implications. Since the function $\phi(\cdot)$ is computable (at the cost of T gradient queries) without a-priori assumptions on d_0, we have reduced parameter-free optimization to solving the one-dimensional implicit equation (4). However, the function ϕ might be discontinuous and an exact solution to the implicit equation might not even exist. Nevertheless, in the next section we show that finding an interval $[\eta, 2\eta]$ in which $h \mapsto \phi(h) - h$ changes sign, produces nearly the same error certificates at an interval edge. Since such interval is readily found via bisection, this forms the basis of a working parameter-free step size tuner. We leave the details to Section 3 and for the remainder of this section prove Proposition 1.

2.1. Proof of Proposition 1

The proof of Proposition 1 hinges on two lemmas. The first is a variant of the standard SGD error bound (recall that $\Delta_i(\eta) = g_i(\eta) - \nabla f(x_i(\eta))$ is zero in the noiseless setting).

Lemma 2 *For $T \in \mathbb{N}$, $\bar{x} \coloneqq \frac{1}{T}\sum_{i<T} x_i(\eta)$, and $\eta > 0$, we have*

$$f(\bar{x}) - f(x_\star) \leq \frac{1}{T}\sum_{i<T} f(x_i(\eta)) - f(x_\star) \leq \frac{\bar{r}_T(\eta)d_0}{\eta T} + \frac{\eta G_T(\eta)}{2T} + \frac{1}{T}\sum_{i<T}\langle \Delta_i(\eta), x_\star - x_i(\eta)\rangle\,. \quad (5)$$

Proof Since η is fixed throughout this proof, we streamline notation by dropping it from x_i, g_i, Δ_i and G_i. By convexity and the definition of Δ_i,

$$\frac{1}{T}\sum_{i<T} f(x_i) - f(x_\star) \leq \frac{1}{T}\sum_{i<T}\langle \nabla f(x_i), x_i - x_\star\rangle = \frac{1}{T}\sum_{i<T}\langle g_i, x_i - x_\star\rangle + \langle \Delta_i, x_\star - x_i\rangle. \quad (6)$$

From $x_{i+1} = \Pi_{\mathcal{X}}(x_i - \eta g_i)$ we can derive the standard subgradient method inequality

$$d_{i+1}^2 \leq \|x_i - \eta g_i - x_\star\|^2 \leq 2\langle x_i - x_{i+1}, x_i - x_\star\rangle = d_i^2 - 2\eta\langle g_i, x_i - x_\star\rangle + \eta^2\|g_i\|^2 \quad (7)$$

for all $\eta \geq 0$ and $i = 0, \ldots, T-1$. Rearranging and summing over $i < T$ gives

$$\sum_{i<T}\langle g_i, x_i - x_\star\rangle = \frac{d_0^2 - d_T^2}{2\eta} + \frac{\eta G_T}{2} = \frac{(d_0 - d_T)(d_0 + d_T)}{2\eta} + \frac{\eta G_T}{2} \overset{(\star)}{\leq} \frac{\bar{r}_T \cdot 2d_0}{2\eta} + \frac{\eta G_T}{2}.$$

The inequality $(\star)$ is where our proof deviates from the textbook derivation [28, Theorem 2.13.]; it holds because $d_0 - d_T \leq r_T$ due to the triangle inequality, and either $d_T \leq d_0$ holds or $d_0^2 - d_T^2 \leq 0 \leq 2d_0\bar{r}_T$. Substituting into (6) and applying $f(\bar{x}) \leq \frac{1}{T}\sum_{i<T} f(x_i)$ by Jensens's inequality gives (5). ∎

The second lemma shows that for η satisfying $\eta \leq \phi(\eta)$ is guaranteed to produce iterates that do not wander too far from $x_\star$. This is our basic localization guarantee.

Lemma 3 *In the noiseless setting with $\alpha > 1$ and $\eta > 0$, if $\eta \le \phi(\eta)$ then we have $\bar{d}_T(\eta) \le \frac{\alpha+1}{\alpha-1} d_0$ and $\bar{r}_T(\eta) \le \frac{2\alpha}{\alpha-1} d_0$.*

Proof This proof once more drops the explicit dependence on η. Summing the inequality (7) over $i < t$ and noting that that $\langle g_i, x_i - x_\star \rangle \ge f(x_i) - f(x_\star) \ge 0$ due to convexity and the noiseless setting, we have $d_t^2 \le d_0^2 + \eta^2 G_t$ for every t. Maximizing over $t \le T$ and substituting $\eta \le \phi(\eta) \le \bar{r}_T / \sqrt{\alpha G_T}$ yields

$$\bar{d}_T^2 \le d_0^2 + \eta^2 G_T \le d_0^2 + \frac{1}{\alpha} \bar{r}_T^2 \overset{(\star)}{\le} d_0^2 + \frac{1}{\alpha} (\bar{d}_T + d_0)^2, \tag{8}$$

where $(\star)$ follows from the triangle inequality: $\bar{r}_T = r_t \le d_t + d_0 \le \bar{d}_T + d_0$ for some $t \le T$. Rearranging yields $\left(\bar{d}_T - \frac{1}{\alpha-1} d_0 \right)^2 \le \frac{\alpha^2}{(\alpha-1)^2} d_0^2$, and therefore $\bar{d}_T \le \frac{\alpha+1}{\alpha-1} d_0$ as required. The bound $\bar{r}_T \le \frac{2\alpha}{\alpha-1} d_0$ follows from substituting $\bar{d}_T \le \frac{\alpha+1}{\alpha-1} d_0$ into $\bar{r}_T \le \bar{d}_T + d_0$. ■

Proposition 1 follows from substituting $\eta = \bar{r}_T(\eta) / \sqrt{\alpha G_T}$ into bound (5) yielding $f(\bar{x}) - f(x_\star) \le \frac{d_0 \sqrt{\alpha G_T(\eta)}}{T} + \frac{\bar{r}_T(\eta) \sqrt{G_T(\eta)}}{2\sqrt{\alpha} T}$, and using $\bar{r}_T(\eta) \le \frac{2\alpha}{\alpha-1} d_0$ from Lemma 3. ■

3. Algorithm description, and analysis for exact gradients

In this section we turn the step-size selection criterion presented in the previous section into a complete algorithm (Algorithm 1)—valid for stochastic as well as exact gradients—and analyze it in the simpler setting of exact gradients, deferring the stochastic case to the following section. Our algorithm consists of a core log-scale bisection subroutine (ROOTFINDINGBISECTION) coupled with an outer loop that acts as an aggressive doubling scheme on the upper limit of the bisection. We describe and analyze the two components Sections 3.1 and 3.2, respectively. Then, in Section 3.3, we put these results together and obtain parameter-free rates in the exact gradient setting.

3.1. Bisection subroutine

Let us describe the ROOTFINDINGBISECTION subroutine of Algorithm 1. Its input is an initial interval $[\eta_{\text{lo}}, \eta_{\text{hi}}]$, SGD iteration number T and damping parameters (α, β) for defining the bisection target $\phi(\eta) = \bar{r}_T(\eta) / \sqrt{\alpha G_T(\eta) + \beta}$. After testing that $\eta_{\text{lo}} \le \phi(\eta_{\text{hi}})$ and $\eta_{\text{hi}} > \phi(\eta_{\text{hi}})$ (and handling the edge cases where this does not hold), we iteratively shrink the interval $[\eta_{\text{lo}}, \eta_{\text{hi}}]$ by replacing one of its edges with $\sqrt{\eta_{\text{lo}} \eta_{\text{hi}}}$ while maintaining the invariant $\eta_{\text{lo}} \le \phi(\eta_{\text{hi}})$ and $\eta_{\text{hi}} > \phi(\eta_{\text{hi}})$.[1] The iterations stop when $\eta_{\text{hi}} / \eta_{\text{lo}} \le 2$. Since each iteration halves $\log \frac{\eta_{\text{hi}}}{\eta_{\text{lo}}}$, the overall iteration number is double-logarithmic in the ratio of the input η_{hi} and η_{lo}. Specifically, if the input interval satisfies $\eta_{\text{hi}} / \eta_{\text{lo}} = 2^{2^k}$ for $k \in \mathbb{N}$ (which it does in Algorithm 1), then ROOTFINDINGBISECTION performs exactly $k = \log\log \frac{\eta_{\text{hi}}}{\eta_{\text{lo}}}$ bisection steps. Consequently, the overall oracle complexity of the the subroutine is $O(T \log\log_+ \frac{\eta_{\text{hi}}}{\eta_{\text{lo}}})$.

We now focus on the end of the bisection procedure and explain the choice of output in line 15. When the bisection loop is complete, we obtain a relatively narrow interval $[\eta_{\text{lo}}^\star, \eta_{\text{hi}}^\star]$ in which $\phi(\eta) - \eta$ is guaranteed to change its sign. When ϕ is continuous, this implies that some $\eta \in [\eta_{\text{lo}}^\star, \eta_{\text{hi}}^\star]$ solves

1. Our choice of $\eta_{\text{mid}} = \sqrt{\eta_{\text{lo}} \eta_{\text{hi}}}$, which corresponds to standard bisection on a *log-scale*, is crucial: the standard choice $\eta_{\text{mid}} = \frac{1}{2}(\eta_{\text{lo}} + \eta_{\text{hi}})$ would result in a logarithmic rather than double-logarithmic number of bisection steps.

Algorithm 1: Parameter-free SGD step size tuning

Input: Initial step size $\eta_\varepsilon > 0$, total gradient budget $B \in \mathbb{N}$, constants $\{\alpha^{(k)}, \beta^{(k)}\}$

▷ In the deterministic case, $\alpha^{(k)} = 3$ (or any constant > 1) and $\beta^{(k)} = 0$; in the stochastic case see eq. (13)

for $k = 2, 4, 8, 16, \ldots$ **do**
  if $k > B/4$ **then return** x_0 ▷ Only happens in the edge case $B = O\left(\log\log \frac{\|x_\star - x_0\|}{\eta_\varepsilon \sqrt{\alpha^{(0)}\|g_0\|^2 + \beta^{(0)}}}\right)$
  $T_k \leftarrow \left\lfloor \frac{B}{2k} \right\rfloor$
  $\eta_o \leftarrow \text{RootFindingBisection}(\eta_\varepsilon, 2^{2^k}\eta_\varepsilon; T_k, \alpha^{(k)}, \beta^{(k)})$
  if $\eta_o < \infty$ **then return** $\frac{1}{T_k}\sum_{i<T_k} x_i(\eta_o)$
end
function $\text{RootFindingBisection}(\eta_{lo}, \eta_{hi}; T, \alpha, \beta)$
  $\phi := \eta \mapsto \bar{r}_T(\eta)/\sqrt{\alpha G_T(\eta) + \beta}$ ▷ Bisection target ($\bar{r}_T$ and G_T defined in eqs. (2) and (3))
  if $\eta_{hi} \le \phi(\eta_{hi})$ **then return** ∞ ▷ η_{hi} is too low and should be increased
  if $\eta_{lo} > \phi(\eta_{lo})$ **then return** η_{lo} ▷ η_{lo} is sufficient (assuming it is very small)
  while $\eta_{hi} > 2\eta_{lo}$ **do** ▷ Invariant: $\eta_{lo} < \eta_{hi}$, $\eta_{lo} \le \phi(\eta_{lo})$, $\eta_{hi} > \phi(\eta_{hi})$
    $\eta_{mid} \leftarrow \sqrt{\eta_{lo}\eta_{hi}}$
    if $\eta_{mid} \le \phi(\eta_{mid})$ **then** $\eta_{lo} \leftarrow \eta_{mid}$ **else** $\eta_{hi} \leftarrow \eta_{mid}$
  end
  if $\bar{r}_T(\eta_{hi}) \le \bar{r}_T(\eta_{lo})\frac{\phi(\eta_{hi})}{\eta_{hi}}$ **then return** η_{hi} **else return** η_{lo}

$\eta = \phi(\eta)$ and therefore has a good error bound by Proposition 1. However, ϕ is not necessarily continuous. To explain why the bisection still outputs a good value of η, first note that Proposition 1 continues to hold (with a slightly worse constant factor) even when η only approximately solves $\eta = \phi(\eta)$, e.g., when $\frac{\bar{r}_T(\eta)}{2\sqrt{\alpha G_T(\eta)+\beta}} = \frac{1}{2}\phi(\eta) \le \eta \le \phi(\eta) = \frac{\bar{r}_T(\eta)}{\sqrt{\alpha G_T(\eta)+\beta}}$. The following lemma shows that the output of RootFindingBisection in fact satisfies a similar bound. See Appendix A.1 for the (easy) proof, and note the lemma also holds in the stochastic case.

Lemma 4 *Let $\eta_{lo}, \eta_{hi}, \alpha, \beta > 0$, and $T \in \mathbb{N}$. If $\text{RootFindingBisection}(\eta_{lo}, \eta_{hi}; T, \alpha, \beta)$ terminates in line 15 with final interval $[\eta_{lo}^\star, \eta_{hi}^\star]$ and returns η_o, then*

$$\frac{\bar{r}_T(\eta_o)}{2\sqrt{\alpha G_T(\eta_{hi}^\star) + \beta}} \le \eta_o \le \frac{\bar{r}_T(\eta_{lo}^\star)}{\sqrt{\alpha G_T(\eta_o) + \beta}}. \tag{9}$$

Moreover, we have $\bar{r}_T(\eta_o) \le \bar{r}_T(\eta_{lo}^\star)$ and $\sqrt{\alpha G_T(\eta_o) + \beta} \le 2\sqrt{\alpha G_T(\eta_{hi}^\star) + \beta}$.

We now combine Lemmas 2, 3 and 4 to show an error bound for GD with the η selected by RootFindingBisection. The proof of Proposition 5 appears in Appendix A.2.

Proposition 5 *In the noiseless setting, let $\eta_o = \text{RootFindingBisection}(\eta_{lo}, \eta_{hi}; T, \alpha, \beta)$ for $\alpha > 1$, $\eta_{lo}, \eta_{hi}, \beta > 0$, and $T \in \mathbb{N}$, assume that $\eta_{hi} > \phi(\eta_{hi})$, and let $\bar{x} = \frac{1}{T}\sum_{i<T} x_i(\eta_o)$. If $\eta_{lo} \le \phi(\eta_{lo})$, then*

$$\|\bar{x} - x_0\| \le \frac{2\alpha}{\alpha - 1} d_0 \quad \text{and} \quad f(\bar{x}(\eta_o)) - f(x_\star) \le \frac{2\alpha}{\alpha - 1} \cdot \frac{d_0\sqrt{\alpha G_T(\eta') + \beta}}{T}$$

for some $\eta' \in [\eta_o, 2\eta_o]$. *If instead* $\eta_{lo} > \phi(\eta_{lo})$, *then* $\eta_o = \eta_{lo}$ *and*

$$\|\bar{x} - x_0\| \le \eta_{lo}\sqrt{\alpha G_T(\eta_{lo}) + \beta} \;\; \textit{and} \;\; f(\bar{x}) - f(x_\star) \le \frac{d_0\sqrt{\alpha G_T(\eta_{lo}) + \beta} + \eta_{lo} G_T(\eta_{lo})}{T}.$$

Let us briefly summarize our findings so far. When ROOTFINDINGBISECTION returns at line 15, Proposition 5 provide a bound similar to the best achievable when d_0 is known, with only a factor $O(\log\log \frac{\eta_{hi}}{\eta_{lo}})$ complexity increase. If instead the lower limit of the bisection is invalid, i.e., $\eta_{lo} > \phi(\eta_{lo})$, our bound becomes the optimal rate plus a term proportional to η_{lo}. Therefore, by picking a very small value of η_{lo} we can ensure a good error bound in that case as well.[2] However, if the upper limit of the bisection is invalid, i.e., $\eta_{hi} \le \phi(\eta_{hi})$, then ROOTFINDINGBISECTION fails, returning $\eta_o = \infty$. We address this issue next.

3.2. Doubling scheme for upper bisection limit

Algorithm 1 iteratively calls ROOTFINDINGBISECTION with upper bisection limits η_{hi} of the form $2^{2^k}\eta_\varepsilon$ (for doubling values of k) until the bisection returns $\eta_o < \infty$, i.e., until $\eta_{hi} > \phi(\eta_{hi})$. To ensure the overall number of gradient queries never exceeds the budget B, for every k the algorithm also adjusts the SGD complexity T. In the stochastic case, the parameters α and β also increase with k in order to enforce a union bound over an increasing number of SGD sample paths.

Intuitively, the bisection should succeed once $\eta_{hi} > d_0/\|g_0\|$ since this is always an upper bound on the ideal step size ϕ_{ideal}. Even though we do not know d_0 and therefore cannot set η_{hi} a-priori,[3] Algorithm 1 will reach such η_{hi} when k is roughly $\log\log \frac{d_0}{\eta_\varepsilon\|g_0\|}$. Lemma 6, whose proof appears in Appendix A.3, provides a rigorous version of our intuitive reasoning.

Lemma 6 *In the noiseless setting with* $\alpha > 1$, $\beta \ge 0$ *and any* $T \in \mathbb{N}$, *if* $\eta > \eta_{max} := \frac{2\alpha}{\alpha-1} \cdot \frac{d_0}{\sqrt{\alpha\|g_0\|^2+\beta}}$ *then* $\eta > \phi(\eta)$. *Consequently, when Algorithm 1 terminates,* $k \le 2\log\log_+ \frac{\eta_{max}}{\eta_\varepsilon}$.

Note that we do not perform a similar doubling scheme to search for an η_{lo} value satisfying $\eta_{lo} \le \phi(\eta_{lo})$, since there is no η_{lo} for which such bound is guaranteed to hold: when $d_0 = 0$ it is possible to have $\phi(\eta) = 0$ for all η. More broadly, parameter-free methods must suffer some non-zero error when $d_0 = 0$ [26], and for our method such term only appears under the condition $\eta_{lo} > \phi(\eta_{lo})$ of Proposition 5, strongly suggesting that we cannot always force $\eta_{lo} \le \phi(\eta_{lo})$ to hold.

3.3. Error guarantees for exact gradients

With Algorithm 1 explained and Proposition 1 and Lemma 6 in place, we are ready to state the parameter-free convergence guarantee in the exact gradient setting. For simplicity of exposition, we fix $\alpha = 3$ and $\beta = 0$, but note that any $\alpha > 1$ yields a similar guarantee.

Theorem 7 *In the noiseless setting Algorithm 1, with parameters* $\alpha^{(k)} = 3$, $\beta^{(k)} = 0$, $\eta_\varepsilon > 0$, $B \in \mathbb{N}$, *and* $x_0 \in \mathbb{R}^d$, *performs at most* B *subgradient queries and returns* $\bar{x} = \frac{1}{T}\sum_{i<T} x_i(\eta) \in \mathbb{R}^d$

2. Moreover, when η_{lo} is very small we expect $\eta_{lo} \le \phi(\eta_{lo})$ to hold, since then (intuitively) $g_i(\eta_{lo}) \approx g_0$ for all i, which implies $\phi(\eta_{lo}) \approx \eta_{lo}\sqrt{T/\alpha} > \eta_{lo}$ for sufficiently large T.
3. If an upper bound $D \ge d_0$ is available (e.g., the domain diameter) then we may use it instead of a doubling scheme by directly fixing k to be $\log\log \frac{D}{\eta_\varepsilon\|g_0\|}$. However, this can improve our error bounds by at most a constant factor.

for some $\eta \geq \eta_\varepsilon$ and integer T satisfying

$$T \geq \max\left\{\frac{B}{12\log\log_+ \frac{\|x_0-x_\star\|}{\eta_\varepsilon\|g_0\|}}, 1\right\} \tag{10}$$

such that either

$$\|\bar{x}-x_\star\| \leq 4\|x_0-x_\star\| \;\; \textit{and} \;\; f(\bar{x})-f(x_\star) \leq \sqrt{27}\,\frac{\|x_0-x_\star\|\sqrt{G_T(\eta')}}{T} \tag{11}$$

for some $\eta' \in [\eta, 2\eta]$, or $\eta = \eta_\varepsilon$ and

$$\|\bar{x}-x_\star\| \leq \eta_\varepsilon\sqrt{3G_T(\eta_\varepsilon)} \;\; \textit{and} \;\; f(\bar{x})-f(x_\star) \leq 2\,\frac{\eta_\varepsilon G_T(\eta_\varepsilon)}{T}. \tag{12}$$

The proof of Theorem 7 appears in Section A.4. Let us briefly compare the bounds in Theorem 7 to our guarantees for a solution to $\eta = \phi(\eta)$ shown in Proposition 1. The "typical case" bound (11) is similar to the error bound of Proposition 1 with only two notable differences beyond a slightly larger constant factor. First, by eq. (10), the value of T in Theorem 7 is smaller than the total complexity budget by a double-logarithmic factor; this is the cost of performing a bisection instead of assuming we start with a solution to the implicit equation. Second, the term $G_T(\eta')$ in the RHS of (11) is computed at η' that is possibly different than the η for which we prove the error bound.

While bounding the error of SGD with step size η using the gradients observed by SGD with a different step size η' is unconventional, our resulting bounds appear to be as useful as their more conventional counterparts. First, note that η and η' are within a factor of 2 of each other, and we can bring this factor arbitrarily close to 1 by running more bisection steps. Second, despite the difference in η, we can still use our lower bounds to obtain (up to double-logarithmic factors) a $1/T$ rate of convergence for smooth problems with unknown smoothness (see Section 5.1); this is the hallmark of error bounds that scale with $\sqrt{G_T}$. Finally, the different η' issue disappears if we assume f is L-Lipschitz. In particular, from Theorem 7 with $\eta_\varepsilon = \frac{\varepsilon}{2L^2}$, using $G_T(\eta) \leq L^2 T$ for all η and $\|g_0\| \geq (f(x_0)-f(x_\star))/d_0$ (due to convexity of f) we get that $f(\bar{x})-f(x_\star) \leq \varepsilon + \lambda\frac{d_0 L}{\sqrt{B}}$ and $\lambda = 18\sqrt{\log\log_+ \frac{d_0^2 L^2}{\varepsilon(f(x_0)-f(x_\star))}}$.

4. Analysis for stochastic gradients

In this section, we extend the analysis of Algorithm 1 to the stochastic setting, using the following simple strategy: we define a "good event" under which the noiseless analysis goes through essentially unchanged (Section 4.1), and show that this event occurs with high probability (Section 4.2), obtaining a stochastic, high-probability, analog of our exact gradient result (Section 4.3).

4.1. Analysis in a "good event"

A careful inspection of our development thus far reveals that we only use the exact gradient assumption by substituting $\sum_{i<T}\langle \Delta_i(\eta), x_i(\eta)-x_\star\rangle \geq 0$ into Lemma 2. Therefore, we consider

the event where this inequality is approximately true. In particular, for $T \in \mathbb{N}$, and $\alpha, \beta, \eta > 0$ define

$$\mathfrak{E}_{T,\alpha,\beta}(\eta) := \left\{ \forall t \le T : \sum_{i<t} \langle \Delta_i(\eta), x_i(\eta) - x_\star \rangle \ge -\frac{1}{4} \max\{\bar{d}_t(\eta), \eta\sqrt{\beta}\} \sqrt{\alpha G_t(\eta) + \beta} \right\}.$$

With this definition in hand, slightly modified versions of our key lemmas from the deterministic analysis (Lemma 3, Proposition 5, Lemma 6) continue to hold. See Appendices B.1 to B.3 for proofs of these results, which follow very similarly to their exact-gradient counterparts.

Lemma 8 *For any $T \in \mathbb{N}$, $\alpha > 2$ and $\beta, \eta > 0$, if event $\mathfrak{E}_{T,\alpha,\beta}(\eta)$ holds and $\eta \le \phi(\eta)$ then $\bar{d}_T(\eta) \le \frac{3\alpha+2}{\alpha-2} d_0$ and $\bar{r}_T(\eta) \le \frac{4\alpha}{\alpha-2} d_0$.*

Proposition 9 *Let $T \in \mathbb{N}$, $\alpha > 2$, $\beta > 0$, $\eta_{\mathrm{lo}} > 0$ and $\eta_{\mathrm{hi}} = 2^{2^k} \eta_{\mathrm{lo}}$ for some $k \ge 1$. Let $\eta_{\mathrm{o}} = \textsc{RootFindingBisection}(\eta_{\mathrm{lo}}, \eta_{\mathrm{hi}}; T, \alpha, \beta)$ and $\bar{x} = \frac{1}{T}\sum_{i<T} x_i(\eta_{\mathrm{o}})$ where $\eta_{\mathrm{lo}} \le \eta_{\mathrm{hi}}$, and $\eta_{\mathrm{hi}} > \phi(\eta_{\mathrm{hi}})$. If the event $\bigcap_{j=0}^{2^k} \mathfrak{E}_{T,\alpha,\beta}(2^j \eta_{\mathrm{lo}})$ holds and $\eta_{\mathrm{lo}} \le \phi(\eta_{\mathrm{lo}})$ then*

$$\|\bar{x} - x_0\| \le \frac{4\alpha}{\alpha-2} d_0 \quad \text{and} \quad f(\bar{x}) - f(x_\star) \le \frac{9\alpha - 2}{2(\alpha-2)} \cdot \frac{d_0 \sqrt{\alpha G_T(\eta') + \beta}}{T}$$

for some $\eta' \in [\eta, 2\eta]$. Moreover, if $\mathfrak{E}_{T,\alpha,\beta}(\eta_{\mathrm{lo}})$ holds and $\eta_{\mathrm{lo}} > \phi(\eta_{\mathrm{lo}})$ then $\eta_{\mathrm{o}} = \eta_{\mathrm{lo}}$ and

$$\|\bar{x} - x_0\| \le \eta_{\mathrm{lo}} \sqrt{\alpha G_T(\eta_{\mathrm{lo}}) + \beta} \quad \text{and} \quad f(\bar{x}) - f(x_\star) \le \frac{5}{4} \frac{d_0 \sqrt{\alpha G_T(\eta_{\mathrm{lo}}) + \beta} + \eta_{\mathrm{lo}}(\alpha G_T(\eta_{\mathrm{lo}}) + \beta)}{T}.$$

Lemma 10 *For any $T \in \mathbb{N}$, $\alpha > 2$ and $\beta, \eta > 0$, if event $\mathfrak{E}_{T,\alpha,\beta}(\eta)$ holds then $\eta > \eta_{\max}(\alpha, \beta) := \frac{4\alpha}{\alpha-2} \cdot \frac{d_0}{\sqrt{\alpha\|g_0\|^2 + \beta}}$ implies that $\eta > \phi(\eta)$. Consequently, if the event $\bigcap_{k=2,4,8,\ldots} \mathfrak{E}_{T_k,\alpha^{(k)},\beta^{(k)}}(2^{2^k}\eta_\varepsilon)$ holds, then Algorithm 1 returns with $k \le 2\log\log_+ \frac{\eta_{\max}(\alpha^{(k)},\beta^{(k)})}{\eta_\varepsilon} \le 2\log\log_+ \frac{\eta_{\max}(\alpha^{(0)},\beta^{(0)})}{\eta_\varepsilon}$.*

4.2. The good event is likely

We now arrive at the challenging part of the stochastic analysis: showing that the good event we defined occurs with high probability. For this, we require the following standard assumption.

Assumption 11 *The stochastic gradient oracle satisfies $\|\mathcal{O}(\eta)\| \le L$ with probability 1.*

In online parameter-free optimization such assumption is unavoidable if one seeks regret scaling linearly in the comparator norm [11]. However, similarly to the best prior results, our bounds depend on L only via a low-order term.

The following result shows that, for appropriate choices of α and β and any fixed $\eta \ge 0$ the event $\mathfrak{E}_{T,\alpha,\beta}(\eta)$ has high probability.

Proposition 12 *Under Assumption 11, for any $T \in \mathbb{N}$, $\eta > 0$ and $\delta \in (0,1)$, if $\alpha \ge 32^2 C$ and $\beta \ge (32CL)^2$ for $C = \log\left(\frac{60}{\delta}\log^2(6T)\right)$ then $\mathbb{P}(\mathfrak{E}_{T,\alpha,\beta}(\eta)) \ge 1 - \delta$.*

Proposition 12 makes no a-priori assumption on the size of $x_i(\eta) - x_\star$, instead controlling it empirically via $\bar{d}_t(\eta)$; this is unusual in the literature and crucial for our purposes. Our proof (given in Appendix B.5) relies on a time-uniform empirical-Bernstein-type martingale concentration bound [18]. However, since this result requires martingale differences that are bounded with probability 1, we cannot apply it on $\langle \Delta_i(\eta), x_i(\eta) - x_\star \rangle$ (which is not bounded), nor can we apply it on $\langle \Delta_i(\eta), x_i(\eta) - x_\star \rangle / \bar{d}_t(\eta)$ (which is bounded but is not adapted to any filtration). Instead, we consider processes of the form $\langle \Delta_i(\eta), \Pi_1([x_i(\eta) - x_\star]/s) \rangle$, where $\Pi_1(\cdot)$ is the projection to the unit ball and s is a fixed scalar. By carefully union bounding over a set of $O(\log T)$ values of s, we are able to control the probability of $\mathfrak{E}_{T,\alpha,\beta}(\eta)$.

Having shown that the good event occurs with high probably for any fixed η, our next step is to show that, for proper choices of $\alpha^{(k)}$ and $\beta^{(k)}$, good events hold with high probability for each and every single value of η Algorithm 1 might try. Noting that for, each value of k, Algorithm 1 only tests step size values of the form $2^j \eta_\varepsilon$ for $j \in \{0, \dots, 2^k\}$, the following lemma (which is a direct application of union bounds) provides the required guarantee; see proof in Appendix B.6.

Lemma 13 *For budget $B \in \mathbb{N}$, initial step size $\eta_\varepsilon > 0$, and failure probability $\delta \in (0,1)$, let*

$$\alpha^{(k)} := 32^2 C_k \;\; \text{and} \;\; \beta^{(k)} := (32 C_k L)^2, \;\; \text{where} \;\; C_k = 2k + \log\left(\frac{60 \log^2(6B)}{\delta}\right). \tag{13}$$

Then, under Assumption 11, we have $\mathbb{P}\Big(\bigcap_{k=2,4,8,\dots} \bigcap_{j=0,1,\dots,2^k} \mathfrak{E}_{B,\alpha^{(k)},\beta^{(k)}}(2^j \eta_\varepsilon)\Big) \geq 1 - \delta$.

4.3. Parameter-free rates for stochastic convex optimization

We are ready to state our main result; see proof in Appendix B.7.

Theorem 14 *Under Assumption 11, for any $\delta \in (0,1)$ consider Algorithm 1 with parameters $\alpha^{(k)}$, $\beta^{(k)}$ given by eq. (13), $\eta_\varepsilon > 0$, $B \in \mathbb{N}$, and $x_0 \in \mathbb{R}^d$. The algorithm makes at most B gradient queries and returns $\bar{x} = \frac{1}{T}\sum_{i<T} x_i(\eta) \in \mathbb{R}^d$ for $\eta \geq \eta_\varepsilon$ such that, with probability at least $1 - \delta$,*

$$T \geq \max\left\{ \frac{B}{8 \log\log_+ \frac{\|x_0 - x_\star - x_0\|}{\eta_\varepsilon L}}, 1 \right\} \tag{14}$$

and for $C = \log\frac{1}{\delta} + \log\log_+ \frac{B\|x_\star - x_0\|}{\eta_\varepsilon L}$, either

$$\|\bar{x} - x_\star\| \leq 6\|x_0 - x_\star\| \;\; \text{and} \;\; f(\bar{x}) - f(x_\star) \leq O\left(\frac{\|x_0 - x_\star\|\sqrt{C G_T(\eta') + C^2 L^2}}{T}\right) \tag{15}$$

for some $\eta' \in [\eta, 2\eta]$, or

$$\|\bar{x} - x_\star\| = O\Big(\eta_\varepsilon \sqrt{C G_T(\eta_\varepsilon) + C^2 L^2}\Big) \;\; \text{and} \;\; f(\bar{x}) - f(x_\star) \leq O\left(\frac{\eta_\varepsilon\big(C G_T(\eta_\varepsilon) + C^2 L^2\big)}{T}\right). \tag{16}$$

Let us compare our bounds to the best known prior bounds, which follow from from online to batch conversion of parameter-free regret bounds. McMahan and Orabona [24] achieve an

optimal parameter-free regret bound for algorithms that are not adaptive to gradient norms: For any user-specified ε and gradient budget B, their result guarantees an *expected* optimality gap of $O\big(\frac{\varepsilon+d_0 L\sqrt{\Lambda}}{\sqrt{B}}\big)$ where $\Lambda = \log\big(1+\frac{d_0 L}{\varepsilon}\big)$ is logarithmic in $\frac{1}{\varepsilon}$. In comparison, by taking $\eta_\varepsilon = O(\frac{\varepsilon}{L^2 B})$ we guarantee a probability $1-\delta$ optimality gap of $O\big(\frac{(\varepsilon+d_0 L)\lambda^2}{\sqrt{B}}\big)$, where $\lambda = \log\Big(\frac{1}{\delta}\log_+\frac{B d_0 L}{\varepsilon}\Big)$ is only double-logarithmic in $\frac{1}{\varepsilon}$; see Appendix B.8 for a slightly tighter bound in this setting.

Cutkosky and Orabona [12] provide the best known parameter-free regret bound for algorithms adaptive to gradient norms. Letting $G = \sum_{i<B}\|\mathcal{O}(x_i)\|^2$ denote the sum of all squared gradient norms observed during optimization, they guarantee expected optimality gap $O\big(\frac{\varepsilon+d_0\mathbb{E}\sqrt{\Lambda G+\Lambda^2 L^2}}{B}\big)$ for $\Lambda = \log\big(1+\frac{d_0\sqrt{G}}{\varepsilon}\big)$. In comparison, substituting the somewhat crude bound $G_T(\eta') \le G$, we guarantee a probability $1-\delta$ optimality gap of $O\big(\frac{\varepsilon\lambda^2+d_0\sqrt{\lambda^2 G+\lambda^3 L^2}}{B}\big)$ for the double-logarithmic λ defined above.

For small values of ε, our bounds show a clear asymptotic improvement over prior art. However, we note that for a hypothetical optimally-tuned ε (which depends on the unknown problem parameter d_0), the logarithmic factor Λ of prior work becomes $O(1)$, while our double-logarithmic factor λ remains $O(\log(\frac{1}{\delta}\log B))$. This occurs because Lemma 10 only provides a somewhat loose bound on $\eta_{\max}$, and because of the union bound in the proof of Proposition 12. We can mitigate this issue at a cost of adaptivity to gradient norm; see Appendix D.2 for further discussion.

Our results give the the first high-probability parameter-free rates. However, while high-probability bounds are generally considered stronger than expectation bounds, it is not clear how to deduce an expectation bound from our results without increasing the error by a $\mathsf{poly}(\log B)$ factor, due to the need to set $\delta = \mathsf{poly}(1/B)$. Finally, we note that Theorem 14 also guarantees that the output of the algorithm is at most a multiplicative factor further from $x_\star$ than x_0 was; we believe that this type of guarantee is new in the parameter-free setting.[4]

5. Adaptivity to problem structure

5.1. Adaptivity to smoothness with exact gradients

Let us assume that f is S-smooth (i.e., has S-Lipschitz gradient), and consider for simplicity the exact gradient setting; we believe that similar results extend to the stochastic setting as well. Under these assumptions, we show that Algorithm 1, *without any changes*, achieves (up to double-logarithmic factors) the Sd_0^2/T suboptimality bound of optimally-tuned GD, as long as $\eta_\varepsilon < \frac{1}{2S}$. See Appendix C.1 for proof.

Theorem 15 *Consider the noiseless regime and assume f is S-smooth. Then Algorithm 1, with gradient budget $B \in \mathbb{N}$, parameters $\alpha^{(k)} = 3$, $\beta^{(k)} = 0$ (as in Theorem 7), and $\eta_\varepsilon \le \frac{1}{2S}$, returns $\bar{x} \in \mathbb{R}^d$ that satisfies $f(\bar{x}) - f(x_\star) \le O\Big(\frac{S\|x_0-x_\star\|^2}{B}\log\log_+\frac{\|x_\star-x_0\|}{\eta_\varepsilon\|g_0\|}\Big)$.*

We remark that UniXGrad [20] features optimal (accelerated) rates for smooth problems without dependence on the parameter S. However, unlike our method, it requires knowledge of d_0.

4. Orabona and Pál [31, Lemma 25] bound the distance moved by Follow the Regularized Leader iterates, but not by a multiple of $\|x_\star - x_0\|$.

5.2. Adaptivity strong convexity using restarts

We now consider a standard strongly-convex stochastic setup [e.g., 17, 36] in which we assume f to be μ-strongly-convex in $\mathcal{X}$ and admit a stochastic gradient oracle bounded by L. (Note that this implies a bound of L/μ on the diameter of $\mathcal{X}$). Hazan and Kale [17] propose to run SGD for epochs of doubling length and halving step sizes. For a total gradient budget of B, they obtain the optimal bound $O(L^2/(\mu B))$ on the expected optimality gap. However, their scheme requires the initial step size to be proportional to $1/\mu$, and hence requires knowledge of μ.

We show that restarting Algorithm 1 with doubling gradient budgets (and no step size to tune) recovers (up to double-logarithmic factors) the optimal $1/B$ rate of convergence. To describe the procedure formally, let $\textsc{ParameterFreeTuner}(x_0, B, \delta, \eta_\varepsilon)$ denote the output of Algorithm 1 with initial point x_0, gradient budget B, failure probability δ, minimal step size η_ε and $\alpha^{(k)}, \beta^{(k)}$ as in eq. (13). For user-specified $\varepsilon > 0$ and $\delta \in (0,1)$ and $x^{(0)} = x_0$, our doubling procedure is

$$x^{(m)} \leftarrow \textsc{ParameterFreeTuner}\Big(x^{(m-1)}, B^{(m)} := 2^m, \delta^{(m)} := \tfrac{1}{m(m+1)}\delta, \eta_\varepsilon^{(m)} := \frac{\varepsilon}{L^2 B^{(m)}}\Big).$$

Theorem 16 *For any $\varepsilon > 0$, $\delta \in (0,1)$, and $M \in \mathbb{N}$, computing $x^{(M)}$ in the procedure described above requires at most $B = 2^{M+1}$ gradient queries. If f is μ-strongly convex and has a stochastic gradient oracle bounded by L, then with probability at least $1-\delta$ we have $f(x^{(M)}) - f(x_\star) \le O\Big(\frac{L^2/\mu+\varepsilon}{B} \log^3\Big(\frac{1}{\delta}\log_+ \frac{B\|x_0 - x_\star\| L}{\varepsilon}\Big)\Big)$.*

See proof in Appendix C.2. Compared to results obtained via parameter-free strongly-convex regret bounds [12, Thm. 7], we remove a squared logarithmic factor, breaking two regret minimization barriers at once.

Acknowledgments

YC was partially supported by the Len Blavatnik and the Blavatnik Family Foundation and an Alon Fellowship. OH was partially supported by the Pitt Momentum Funds.

References

[1] Larry Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966.

[2] Hilal Asi and John C Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

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

[4] Leonard Berrada, Andrew Zisserman, and M Pawan Kumar. Training neural networks for and by interpolation. In *International Conference on Machine Learning*, pages 799–809, 2020.

[5] D Blackwell. Large deviations for martingales. In *Festschrift for Lucien Le Cam*, pages 89–91. Springer, 1997.

[6] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and trends in machine learning*, 2015.

[7] Richard H Byrd, Gillian M Chin, Jorge Nocedal, and Yuchen Wu. Sample size selection in optimization methods for machine learning. *Mathematical programming*, 134(1):127–155, 2012.

[8] Kamalika Chaudhuri, Yoav Freund, and Daniel J Hsu. A parameter-free hedging algorithm. *Advances in neural information processing systems*, 2009.

[9] Keyi Chen, John Langford, and Francesco Orabona. Better parameter-free stochastic optimization with ODE updates for coin-betting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[10] Ashok Cutkosky. Artificial constraints and hints for unbounded online learning. In *Conference on Learning Theory*, pages 874–894, 2019.

[11] Ashok Cutkosky and Kwabena Boahen. Online learning without prior information. In *Conference on Learning Theory*, pages 643–677, 2017.

[12] Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in Banach spaces. In *Conference On Learning Theory*, pages 1493–1529, 2018.

[13] Matthias Feurer and Frank Hutter. Hyperparameter optimization. In *Automated machine learning*, pages 3–33. Springer, 2019.

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

[15] Elad Hazan. *Introduction to online convex optimization*. Foundations and Trends in Optimization, 2016.

[16] Elad Hazan and Sham Kakade. Revisiting the Polyak step size. *arXiv:1905.00313*, 2019.

[17] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1): 2489–2512, 2014.

[18] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

[19] Anatoli Juditsky and Yuri Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. *arXiv:1401.1792*, 2014.

[20] Ali Kavis, Kfir Y Levy, Francis Bach, and Volkan Cevher. UniXGrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. *Advances in Neural Information Processing Systems*, 2019.

[21] Michal Kempka, Wojciech Kotlowski, and Manfred K Warmuth. Adaptive scale-invariant online algorithms for learning linear models. In *International Conference on Machine Learning*, pages 3321–3330, 2019.

[22] Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic Polyak step-size for SGD: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 1306–1314, 2021.

[23] Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *The Journal of Machine Learning Research*, 18(1):3117–3166, 2017.

[24] H Brendan McMahan and Francesco Orabona. Unconstrained online linear learning in Hilbert spaces: Minimax algorithms and normal approximations. In *Conference on Learning Theory*, pages 1020–1039, 2014.

[25] Zakaria Mhammedi and Wouter M Koolen. Lipschitz and comparator-norm adaptivity in online learning. In *Conference on Learning Theory*, pages 2858–2887, 2020.

[26] Francesco Orabona. Dimension-free exponentiated gradient. *Advances in Neural Information Processing Systems*, 26, 2013.

[27] Francesco Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. *Advances in Neural Information Processing Systems*, 2014.

[28] Francesco Orabona. A modern introduction to online learning. *arXiv:1912.13213*, 2021.

[29] Francesco Orabona and Ashok Cutkosky. ICML tutorial on parameter-free stochastic optimization, 2020.

[30] Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems*, 29:577–585, 2016.

[31] Francesco Orabona and Dávid Pál. Parameter-free stochastic optimization of variationally coherent functions. *arXiv:2102.00236*, 2021.

[32] Francesco Orabona and Tatiana Tommasi. Training deep networks without learning rates through coin betting. *Advances in Neural Information Processing Systems*, 30:2160–2170, 2017.

[33] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[34] Boris T. Polyak. *Introduction to Optimization*. 1987.

[35] Michal Rolinek and Georg Martius. L4: Practical loss-based stepsize adaptation for deep learning. *Advances in neural information processing systems*, 2018.

[36] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.

[37] Matthew Streeter and H Brendan McMahan. No-regret algorithms for unconstrained online convex optimization. *Advances in Neural Information Processing Systems*, pages 2402–2410, 2012.

[38] Eiji Takimoto and Manfred Warmuth. The minimax strategy for Gaussian density estimation. In *Conference on Learning Theory*, 2000.

[39] Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. *Advances in neural information processing systems*, 2019.

[40] Pengchuan Zhang, Hunter Lang, Qiang Liu, and Lin Xiao. Statistical adaptive stochastic gradient methods. *arXiv:2002.10597*, 2020.

Appendix A. Proofs for Section 3

A.1. Proof of Lemma 4

Proof Consider the case when ROOTFINDINGBISECTION returns $\eta_o = \eta^\star_{lo}$, i.e., when

$$\frac{\bar{r}_T(\eta^\star_{lo})}{2\bar{r}_T(\eta^\star_{hi})}\phi(\eta^\star_{hi}) \overset{(i)}{\leq} \frac{1}{2}\eta^\star_{hi} \overset{(ii)}{\leq} \eta^\star_{lo} \overset{(iii)}{\leq} \phi(\eta^\star_{lo})$$

due to (i) the condition for returning $\eta_o = \eta^\star_{lo}$, (ii) the bisection termination condition $\eta^\star_{hi} \leq 2\eta^\star_{lo}$, and (iii) the bisection invariant. Substituting $\phi(\eta_o) = \bar{r}_T(\eta_o)/\sqrt{\alpha G_T(\eta_o) + \beta}$ into the above display yields

$$\frac{\bar{r}_T(\eta^\star_{lo})}{2\sqrt{\alpha G_T(\eta^\star_{hi}) + \beta}} \leq \eta^\star_{lo} \leq \frac{\bar{r}_T(\eta^\star_{lo})}{\sqrt{\alpha G_T(\eta^\star_{lo}) + \beta}}. \tag{17}$$

Next, consider the case when ROOTFINDINGBISECTION returns $\eta_o = \eta^\star_{hi}$, i.e., when

$$\frac{\bar{r}_T(\eta^\star_{hi})}{\sqrt{\alpha G_T(\eta^\star_{hi}) + \beta}} = \phi(\eta^\star_{hi}) \overset{(i)}{\leq} \eta^\star_{hi} \overset{(ii)}{\leq} \frac{\bar{r}_T(\eta^\star_{lo})}{\bar{r}_T(\eta^\star_{hi})}\phi(\eta^\star_{hi}) = \frac{\bar{r}_T(\eta^\star_{lo})}{\sqrt{\alpha G_T(\eta^\star_{hi}) + \beta}}, \tag{18}$$

due to (i) the bisection invariant, and (ii) the condition for returning $\eta_o = \eta^\star_{hi}$.

Finally, $\bar{r}_T(\eta_o) \leq \bar{r}_T(\eta^\star_{lo})$ holds trivially if $\eta_o = \eta^\star_{lo}$, and if $\eta_o = \eta^\star_{hi}$ then it holds by (18); similarly $\sqrt{\alpha G_T(\eta_o) + \beta} \leq 2\sqrt{\alpha G_T(\eta^\star_{hi}) + \beta}$ holds either trivially or from (17). ■

A.2. Proof of Proposition 5

Proof We begin with the first case of the proposition, assuming that the initial η_{lo}, η_{hi} satisfy $\eta_{lo} \leq \phi(\eta_{lo})$ and $\eta_{hi} > \phi(\eta_{hi})$ so that ROOTFINDINGBISECTION terminates in line 15 with final interval $[\eta^\star_{lo}, \eta^\star_{hi}]$. In this case, we have the following error bound

$$\begin{aligned}
f(\bar{x}) - f(x_\star) &\overset{(i)}{\leq} \frac{\bar{r}_T(\eta_o)d_0}{\eta_o T} + \frac{\eta_o G_T(\eta_o)}{2T} \\
&\overset{(ii)}{\leq} \frac{2d_0\sqrt{\alpha G_T(\eta^\star_{hi}) + \beta}}{T} + \frac{\bar{r}_T(\eta^\star_{lo})}{2\alpha T}\sqrt{\alpha G_T(\eta_o) + \beta} \\
&\overset{(iii)}{\leq} \frac{2d_0\sqrt{\alpha G_T(\eta^\star_{hi}) + \beta}}{T} + \frac{d_0}{T(\alpha - 1)}\sqrt{\alpha G_T(\eta_o) + \beta} \\
&\overset{(iv)}{\leq} \frac{2d_0\sqrt{\alpha G_T(\eta^\star_{hi}) + \beta}}{T} + \frac{2d_0}{T(\alpha - 1)}\sqrt{\alpha G_T(\eta^\star_{hi}) + \beta} \\
&= \frac{2\alpha}{\alpha - 1} \cdot \frac{d_0\sqrt{\alpha G_T(\eta^\star_{hi}) + \beta}}{T},
\end{aligned} \tag{19}$$

due to (i) Lemma 2, (ii) Lemma 4, (iii) Lemma 3 and (iv) Lemma 4 again. Moreover, we have $\|\bar{x} - x_0\| \leq \bar{r}_T(\eta_o) \leq \bar{r}_T(\eta^\star_{lo}) \leq \frac{2\alpha}{\alpha-1}d_0$ from Lemmas 3 and 4.

Next, consider that case where ROOTFINDINGBISECTION returns the initial η_{lo}, i.e., where $\eta_{lo} > \phi(\eta_{lo}) = \bar{r}_T(\eta_{lo})/\sqrt{\alpha G_T(\eta_{lo}) + \beta}$. Rearranging this condition immediately yields the claimed bound on $\|\bar{x}(\eta_{lo}) - x_0\| \leq \bar{r}_T(\eta_{lo})$. Substituting the lower bound on η_o into the error bound (19) completes the proof. ■

A.3. Proof of Lemma 6

Proof Assume by contradiction that $\eta > \eta_{\max}$ but $\eta \le \phi(\eta)$. On the one hand,

$$\frac{\bar{r}_T(\eta)}{\sqrt{\alpha\|g_0\|^2+\beta}} \ge \frac{\bar{r}_T(\eta)}{\sqrt{\alpha G_T(\eta)+\beta}} = \phi(\eta) \ge \eta > \eta_{\max} = \frac{2\alpha}{\alpha-1}\cdot\frac{d_0}{\sqrt{\alpha\|g_0\|^2+\beta}}.$$

which implies $\bar{r}_T(\eta) > \frac{2\alpha}{\alpha-1}d_0$. On the other hand, by Lemma 3, we have $\bar{r}_T(\eta) \le \frac{2\alpha}{\alpha-1}d_0$ which yields a contradiction.

By the discussion so far, the doubling scheme in Algorithm 1 must stop when $2^{2^k}\eta_\varepsilon > \eta_{\max}$. Therefore, if it stops with $k = k_{\mathrm{o}}$ we may conclude that either $k_{\mathrm{o}} = 2$ or $2^{2^{k_{\mathrm{o}}/2}}\eta_\varepsilon \le \eta_{\max}$, as otherwise it would have stopped with $k_{\mathrm{o}}/2$ instead. Rearranging yields $k_{\mathrm{o}} \le 2\log\log_+\frac{\eta_{\max}}{\eta_\varepsilon}$ as claimed. ■

A.4. Proof of Theorem 7

Proof Let us first show that Algorithm 1 never makes more than B gradient queries. The algorithm repeatedly calls RootFindingBisection, with parameters $T = T_k = \lfloor\frac{B}{2k}\rfloor \le \frac{B}{2k}$ and $\eta_{\mathrm{hi}}/\eta_{\mathrm{lo}} = 2^{2^k}$ for $k = 2, 4, 8, \ldots$, until it returns $\eta_{\mathrm{o}} < \infty$ (or k passes $B/8$). In the iterations where RootFindingBisection returns $\eta_{\mathrm{o}} = \infty$ it uses only T_k subgradient queries (a single evaluation of ϕ). In the final iteration where RootFindingBisection returns $\eta_{\mathrm{o}} < \infty$, it evaluates ϕ at most $k+2$ times: once at line 9, once at line 10, and $\log\log(\eta_{\mathrm{hi}}/\eta_{\mathrm{lo}}) = k$ times at line 13 during the bisection. Therefore, letting $k' = 2^{j'}$ denote the index of the final iteration, the total query complexity of the final bisection call is

$$(k'+2)T_{k'} \le (2+k')\frac{B}{2k'} \le \frac{B}{2} + \frac{B}{2^{j'}},$$

and the complexity of all preceding bisection calls is at most

$$\sum_{j=1}^{j'-1} T_{2^j} \le \frac{B}{2}\sum_{j=1}^{j'-1} 2^{-j} = \frac{B}{2} - \frac{B}{2^{j'}},$$

giving a total complexity bound of B.

Next, we establish (10). Note that Algorithm 1 indeed always returns a point of the form $\bar{x} = \frac{1}{T}\sum_{i<T} x_i(\eta)$ with some $T \ge 1$; the edge case of returning in line 2 corresponds to $T = 1$. Moreover, in the typical case of returning in line 5, we have $T = \lfloor\frac{B}{2k}\rfloor \ge \frac{B}{4k}$ for some $k \le B/4$. By Lemma 6, we have

$$k \le 2\log\log_+\frac{\eta_{\max}}{\eta_\varepsilon} = 2\log\log_+\frac{\sqrt{3}d_0}{\eta_\varepsilon\|g_0\|} \le 3\log\log_+\frac{d_0}{\eta_\varepsilon\|g_0\|},$$

giving the claimed lower bound (10) on T.

It remains to show that one of the conclusions (11) and (12) must hold. When Algorithm 1 returns in line 5, this follows immediately from Proposition 5 (if $\eta_\varepsilon \le \phi(\eta_\varepsilon)$ then conclusion (11) holds; if $\eta_\varepsilon > \phi(\eta_\varepsilon)$ then either $\eta_\varepsilon\sqrt{3G_T(\eta_\varepsilon)} \le 3d_0$ and conclusion (11) holds, or $d_0\sqrt{3G_T(\eta_\varepsilon)} \le \eta_\varepsilon G_T(\eta_\varepsilon)$ and conclusion (12) holds). In the edge case of returning $\bar{x} = x_0$ in line 2, corresponding

to $T = 1$, conclusion (11) clearly holds, as $\|x_0 - x_\star\| \leq 4\|x_0 - x_\star\|$ trivially and $f(x_0) - f(x_\star) \leq \langle g_0, x_0 - x_\star\rangle \leq \|x_0 - x_\star\|\|g_0\|$ due to convexity of f. We remark that due to inequality (10), the $T = 1$ edge case is only possible for a very small iteration budget $B = O(\log\log_+ \frac{\|x_0 - x_\star\|}{\eta_\varepsilon \|g_0\|})$. ■

Appendix B. Proofs for Section 4

B.1. Proof of Lemma 8

Proof The proof proceeds similarly to the proof of Lemma 3, except that instead of assuming exact subgradients we make use of the definition of $\mathfrak{E}_{T,\alpha,\beta}$. As usual in proofs where η is fixed, we drop the explicit dependence on it from $x_t, g_t, d_t, \bar{d}_t, r_t, \bar{r}_t$ and G_t.

We start with the following inequality, which holds for any $t \in \mathbb{N}$ by summing Eq. (7) and recalling that $\Delta_i = g_i - \nabla f(x_i)$,

$$d_t^2 \leq d_0^2 + \eta^2 G_t - 2\eta \sum_{i<t} \langle g_i, x_i - x_\star\rangle = d_0^2 + \eta^2 G_t - 2\eta \sum_{i<t} \langle \nabla f(x_i), x_i - x_\star\rangle - 2\eta \sum_{i<t} \langle \Delta_i, x_i - x_\star\rangle .$$

Noting that $\langle \nabla f(x_i), x_i - x_\star\rangle \geq f(x_i) - f(x_\star) \geq 0$ due to convexity and that $\sum_{i<t} \langle \Delta_i, x_i - x_\star\rangle \geq -\frac{1}{4}\max\{\bar{d}_t, \eta\sqrt{\beta}\}\sqrt{\alpha G_t + \beta}$ for all $t \leq T$ due to $\mathfrak{E}_{T,\alpha,\beta}$ holding, we have

$$d_t^2 \leq d_0^2 + \eta^2 G_t + \frac{1}{2}\max\{\eta\bar{d}_t, \eta^2\sqrt{\beta}\}\sqrt{\alpha G_t + \beta}$$

for all $t \leq T$.

Maximizing both sides of the inequality over $t \leq T$ and recalling that $\eta \leq \phi(\eta) = \bar{r}_T/\sqrt{\alpha G_T + \beta}$, we get

$$\bar{d}_T^2 \leq d_0^2 + \frac{1}{\alpha}\bar{r}_T^2 + \max\left\{\frac{\bar{r}_T \bar{d}_T}{2}, \frac{\bar{r}_T^2}{2\sqrt{\alpha}}\right\}.$$

Substituting $\bar{r}_T \leq \bar{d}_T + d_0$ (which holds due to the triangle inequality), we get

$$\bar{d}_T^2 \leq d_0^2 + \frac{1}{\alpha}(\bar{d}_T + d_0)^2 + \max\left\{\frac{(\bar{d}_T + d_0)^2}{2}, \frac{(\bar{d}_T + d_0)^2}{2\sqrt{\alpha}}\right\} = d_0^2 + \underbrace{\left(\frac{1}{\alpha} + \frac{1}{2}\right)}_{1/\alpha'}(\bar{d}_T + d_0)^2,$$

where the final equality is due to $\alpha > 1$. Thus, we arrive again at inequality (8) from the proof of Lemma 3, but with α replaced by $\alpha' = 2\alpha/(\alpha + 2)$. We consequently find that

$$\bar{d}_T \leq \frac{\alpha' + 1}{\alpha' - 1} d_0 = \frac{3\alpha + 2}{\alpha - 2} d_0 \quad \text{and} \quad \bar{r}_T \leq \bar{d}_T + d_0 \leq \frac{4\alpha}{\alpha - 2} d_0.$$

■

B.2. Proof of Proposition 9

Proof In both cases of the proposition, the event $\mathfrak{E}_{T,\alpha,\beta}(\eta_o)$ holds (in the first case because $2^j\eta_{lo}$ for $j = 1, \dots, 2^k$ represents all possible values for η_o; in the second case because $\eta_o = \eta_{lo}$ by Line 10 of ROOTFINDINGBISECTION) and therefore,

$$\sum_{i<T} \langle \Delta_i(\eta_o), x_\star - x_i(\eta_o) \rangle \le \frac{1}{4} \max\left\{ \bar{d}_T(\eta_o), \eta_o \sqrt{\beta} \right\} \sqrt{\alpha G_T(\eta_o) + \beta}. \tag{20}$$

We begin with the first case of the lemma (assuming $\eta_{lo} \le \eta_{hi}$ satisfy $\eta_{lo} \le \phi(\eta_{lo})$ and $\eta_{hi} > \phi(\eta_{hi})$). In this case, ROOTFINDINGBISECTION terminates at line 15. Let $\eta_{lo}^\star$ and $\eta_{hi}^\star$ denote the values of η_{lo} and η_{hi} at line 15 of ROOTFINDINGBISECTION, respectively. First, note that, by (20) and Lemma 4 we have

$$\begin{aligned}\sum_{i<T} \langle \Delta_i(\eta_o), x_\star - x_i(\eta_o) \rangle &\le \frac{1}{4} \max\left\{ \bar{d}_T(\eta_o), \bar{r}_T(\eta_{lo}^\star) \right\} \sqrt{\alpha G_T(\eta_o) + \beta} \\ &\le \frac{d_0 + \bar{r}_T(\eta_{lo}^\star)}{4} \sqrt{\alpha G_T(\eta_o) + \beta}. \end{aligned} \tag{21}$$

where the first inequality uses (20), $\eta_o \le \frac{\bar{r}_T(\eta_{lo}^\star)}{\sqrt{\alpha G_T(\eta_o)+\beta}}$ (Lemma 4) and $\sqrt{\beta} \le \sqrt{\alpha G_T(\eta_o) + \beta}$, the second inequality is by $\bar{d}_T(\eta_o) \le d_0 + \bar{r}_T(\eta_o)$ and $\bar{r}_T(\eta_o) \le \bar{r}_T(\eta_{lo}^\star)$ (Lemma 4).

The remainder of the proof is very similar to the proof of Proposition 5. Combining (5), (9) and (21) yields:

$$\begin{aligned} f(\bar{x}) - f(x_\star) &\le \frac{2d_0\sqrt{\alpha G_T(\eta_{hi}^\star) + \beta}}{T} + \frac{2\alpha^{-1}\bar{r}_T(\eta_{lo}^\star) + d_0 + \bar{r}_T(\eta_{lo}^\star)}{4T} \sqrt{\alpha G_T(\eta_o) + \beta} \\ &\le \frac{2d_0\sqrt{\alpha G_T(\eta_{hi}^\star) + \beta}}{T} + \frac{(2\alpha^{-1}+1)\frac{4\alpha}{\alpha-2} + 1}{4T} d_0 \sqrt{\alpha G_T(\eta_o) + \beta} \\ &\le \frac{2d_0\sqrt{\alpha G_T(\eta_{hi}^\star) + \beta}}{T} + \frac{(2\alpha^{-1}+1)\frac{4\alpha}{\alpha-2} + 1}{2T} d_0 \sqrt{\alpha G_T(\eta_{hi}^\star) + \beta} \\ &= \frac{9\alpha - 2}{2(\alpha-2)} \cdot \frac{d_0\sqrt{\alpha G_T(\eta_{hi}^\star) + \beta}}{T} \end{aligned}$$

where the first inequality substitutes (9) and (21) into (5), the second inequality uses that $\bar{r}_T(\eta_{lo}^\star) \le \frac{4\alpha}{\alpha-2} d_0$ by $\eta_{lo}^\star \le \phi(\eta_{lo}^\star)$ and Lemma 8, the third inequality uses $\sqrt{\alpha G_T(\eta_o) + \beta} \le 2\sqrt{\alpha G_T(\eta_{hi}^\star) + \beta}$ by Lemma 4, and the final equality is algebra.

Finally, consider that case where ROOTFINDINGBISECTION returns the initial η_{lo}, i.e., where $\eta_{lo} > \phi(\eta_{lo}) = \bar{r}_T(\eta_{lo})/\sqrt{\alpha G_T(\eta_{lo}) + \beta}$. Rearranging this condition immediately yields the claimed bound on $\|\bar{x}(\eta_{lo}) - x_0\| \le \bar{r}_T(\eta_{lo})$. Substituting the lower bound on η_{lo} into the error bound (5) and applying (20) using $\bar{d}_T(\eta_{lo}) \le \bar{r}_T(\eta_{lo}) + d_0$ completes the proof. ■

B.3. Proof of Lemma 10

Proof The proof is essentially identical to the proof of Lemma 6, with Lemma 8 used instead of Lemma 3; we give it here for completeness.

To show the first part of the lemma, assume that $\mathfrak{E}_{T,\alpha,\beta}(\eta)$ holds and assume by contradiction that $\eta > \eta_{\max}(\alpha,\beta)$ but $\eta \le \phi(\eta)$. On the one hand,

$$\frac{\bar{r}_T(\eta)}{\sqrt{\alpha\|g_0\|^2+\beta}} \ge \frac{\bar{r}_T(\eta)}{\sqrt{\alpha G_T(\eta)+\beta}} = \phi(\eta) \ge \eta > \eta_{\max}(\alpha,\beta) = \frac{4\alpha}{\alpha-2}\cdot\frac{d_0}{\sqrt{\alpha\|g_0\|^2+\beta}}.$$

which implies $\bar{r}_T(\eta) > \frac{4\alpha}{\alpha-2}d_0$. On the other hand, by Lemma 8, we have $\bar{r}_T(\eta) \le \frac{4\alpha}{\alpha-2}d_0$ which yields a contradiction.

Algorithm 1 repeatedly invokes ROOTFINDINGBISECTION with η_{hi} values of the form $2^{2^k}\eta_\varepsilon$, $\alpha = \alpha^{(k)}$, and $\beta = \beta^{(k)}$ (for $k = 2, 4, \ldots$) until the bisection returns $\eta_{\text{o}} < \infty$. This happens as soon as $\eta_{\text{hi}} > \phi(\eta_{\text{hi}})$, which holds (by the discussion above and since we assume $\mathfrak{E}_{T_k,\beta^{(k)},\beta^{(k)}}(2^{2^k}\eta_\varepsilon)$ holds for all $k = 2, 4, \ldots$) whenever $\eta_{\text{hi}} > \eta_{\max}(\alpha^{(k)},\beta^{(k)})$. Therefore, if the algorithm returns with $k = k_{\text{o}}$ we may conclude that either $k_{\text{o}} = 2$ or $2^{2^{k_{\text{o}}/2}}\eta_\varepsilon \le \eta_{\max}(\alpha^{(k_{\text{o}})},\beta^{(k_{\text{o}})})$, as otherwise we would have returned with $k_{\text{o}}/2$ instead. Rearranging yields $k_{\text{o}} \le 2\log\log_+ \frac{\eta_{\max}(\alpha^{(k_{\text{o}})},\beta^{(k_{\text{o}})})}{\eta_\varepsilon}$ as claimed. Finally, we note that $\eta_{\max}(\alpha,\beta)$ is decreasing in both α and β, and we may therefore replace $\alpha^{(k)},\beta^{(k)}$ by the larger values $\alpha^{(0)},\beta^{(0)}$ as defined in eq. (13). ■

B.4. A martingale concentration bound

The following corollary is a translation of [18, Theorem 4] which simplifies notation at the cost of looser constants. We remark that it holds even when $\log$ denotes the natural logarithm (as is the convention in [18]).

Corollary 17 (of Howard et al. [18, Theorem 4]) *Let X_t be adapted to $\mathcal{F}_t$ such that $|X_t| \le 1$ with probability 1 for all t. Then, for every $\delta \in (0,1)$ and any $\hat{X}_t \in \mathcal{F}_{t-1}$ such that $|\hat{X}_t| \le 1$ with probability 1,*

$$\mathbb{P}\left(\exists t < \infty : \left|\sum_{s\le t}(X_s - \mathbb{E}[X_s \mid \mathcal{F}_{s-1}])\right| \ge \sqrt{A_t(\delta)\sum_{s\le t}\left(X_t - \hat{X}_t\right)^2 + B_t(\delta)}\right) \le \delta,$$

where $A_t(\delta) = 16\log\left(\frac{60\log(6t)}{\delta}\right)$ and $B_t(\delta) = 16\log^2\left(\frac{60\log(6t)}{\delta}\right)$.

Proof Throughout we proof we use the binary maximization notation $a \vee b := \max\{a,b\}$.

We apply [18, Theorem 4] with $a = -b = 1$ and the polynomial stitched boundary [18, Eq. (10)] with parameters $m, \eta, s \ge 1$ to be specified below. This yields

$$\mathbb{P}\left(\exists t < \infty : \left|\sum_{s\le t}(X_s - \mathbb{E}[X_s \mid \mathcal{F}_{s-1}])\right| \ge S_{\delta/2}\left(m \vee \sum_{s\le t}\left(X_s - \hat{X}_s\right)^2\right)\right) \le \delta,$$

where, for every $v' > 0$

$$S_{\delta/2}\left(v'\right) := k_1\sqrt{v'\left(s\log\log\left(\frac{\eta v'}{m}\right) + \log\frac{2\zeta\left(s\right)}{\delta\log^s\eta}\right)} + 2k_2\left(s\log\log\left(\frac{\eta v'}{m}\right) + \log\frac{2\zeta\left(s\right)}{\delta\log^s\eta}\right)$$

with ζ denoting the Riemann zeta function,

$$k_1 = \left(\eta^{1/4} + \eta^{-1/4}\right)/\sqrt{2} \text{ and } k_2 = \left(\sqrt{\eta}+1\right)/2.$$

Let us first simplify $S_{\delta/2}(m \vee v)$ and then choose the parameters m, η, s to yield decent constants. Writing $Z = s \log\log\left(\frac{\eta(m\vee v)}{m}\right) + \log\frac{2\zeta(s)}{\delta \log^s \eta}$, we have

$$S_{\delta/2}(m \vee v) \le (k_1 + 2k_2)\sqrt{(m \vee v \vee Z)Z}.$$

Note that $\log\log\left(\frac{\eta(m\vee v)}{m}\right) \ge \log\log\eta$ and therefore $Z \ge \log\frac{2\zeta(s)}{\delta} \ge \log(2\zeta(s))$. Therefore, if $m \le \log(2\zeta(s))$, we have the slightly simplifies bound

$$S_{\delta/2}(m \vee v) \le (k_1 + 2k_2)\sqrt{(v+Z)Z}.$$

Moreover, for $m \ge \eta$ we may upper bound Z by

$$Z \le s \log\left(\frac{[2\zeta(s)]^{1/s}}{\delta \log\eta}\log(m+v)\right)$$

Taking $\eta = m = 1.8$ and $s = 1.05$, one easily confirms that $m \le 3 \le \log(2\zeta(s))$ and, substituting back k_1 and k_2 as defined above, we have

$$S_{\delta/2}(m \vee v) \le \sqrt{16\log\left(\frac{60\log(m+v)}{\delta}\right)v + 16\log^2\left(\frac{60\log(m+v)}{\delta}\right)}$$

Finally, noting that $\left(X_s - \hat{X}_s\right)^2 \le 4$ for all s, we may substitute $m + v \le 6t$ in the bound above, concluding the proof. ■

B.5. Proof of Proposition 12

Proof Since η is fixed throughout this proof, we drop the explicit dependence on it to simplify notation. Furthermore, we define the normalized/shorthand quantities:

$$\Delta_i' := \Delta_i/L\ ,\ \bar{d}_t' := \max\{\bar{d}_t, \eta\sqrt{\beta}\}\ ,\ \alpha' := \alpha/64\ ,\ G_t' := G_t/L^2 \text{ and } \beta' := \beta/(64L^2).$$

With these definitions, the failure probability we wish to bound is

$$\mathbb{P}\left(\mathfrak{E}^c_{T,\alpha,\beta}\right) = \mathbb{P}\left(\exists t \le T : \sum_{i<t}\langle \Delta_i', x_i - x_\star\rangle < -2\bar{d}_t'\sqrt{\alpha' G_t' + \beta'}\right).$$

Now, define for any $k \ge 0$:

$$s_k := 2^k \bar{d}_0' = 2^k \max\{d_0, \eta\sqrt{\beta}\}$$

and in addition define

$$k_t := \left\lceil \log\frac{\bar{d}_t'}{\bar{d}_0'}\right\rceil.$$

Note that k_t satisfies the following

$$\bar{d}_t \leq \bar{d}'_t \leq s_{k_t} \leq 2\bar{d}'_t \text{ and } 0 \leq k_t \leq \left\lceil \log\left(\frac{t}{4}+1\right)\right\rceil \leq \log(6t).$$

The first set of inequalities follows from the definition of $\bar{d}'_t$, k_t and s_k, while the latter inequality is due to the fact that $\bar{d}_t \leq d_0 + t\eta L \leq d_0 + \frac{t}{4}\eta\sqrt{\beta}$, which follows from the definition of SGD, the triangle inequality, the assumption $\|g_i\| \leq L$ w.p. 1, and $\beta \geq 16L^2$.

Writing $\Pi_1(x) = x/\max\{1, \|x\|\}$ for the projection to the Euclidean unit ball, we now bound the failure probability as follows

$$\begin{aligned}
\mathbb{P}\big(\mathfrak{E}^c_{T,\alpha,\beta}\big) &\overset{(i)}{=} \mathbb{P}\left(\exists t \leq T : \sum_{i<t}\left\langle \Delta'_i, \Pi_1\left(\frac{x_i - x_\star}{s_{k_t}}\right)\right\rangle < -\frac{2\bar{d}'_t}{s_{k_t}}\sqrt{\alpha' G'_t + \beta'}\right) \\
&\overset{(ii)}{\leq} \mathbb{P}\left(\exists t \leq T : \sum_{i<t}\left\langle \Delta'_i, \Pi_1\left(\frac{x_i - x_\star}{s_{k_t}}\right)\right\rangle < -\sqrt{\alpha' G'_t + \beta'}\right) \\
&\overset{(iii)}{\leq} \sum_{k=0}^{\lfloor \log(6T)\rfloor} \mathbb{P}\left(\exists t \leq T : \sum_{i<t}\left\langle \Delta'_i, \Pi_1\left(\frac{x_i - x_\star}{s_{k}}\right)\right\rangle < -\sqrt{\alpha' G'_t + \beta'}\right),
\end{aligned} \tag{22}$$

where (i) follows from $\|x_i - x_\star\| = d_i \leq \bar{d}_t \leq s_{k_t}$ (which means that the projection does nothing), (ii) follows from $s_{k_t} \leq 2\bar{d}'_t$, and (iii) follows from $0 \leq k_t \leq \log(6t) \leq \log(6T)$ and a union bound. We can now define a nicely behaved stochastic process: for every i and k let

$$X_i^{(k)} := \left\langle \frac{g_i}{L}, \Pi_1\left(\frac{x_i - x_\star}{s_k}\right)\right\rangle$$

and note that $X_i^{(k)}$ is adapted to the filtration $\mathcal{F}_t = \sigma(g_1, \dots, g_t)$ (i.e., $X_i^{(k)} \in \mathcal{F}_t$) and satisfies $\left|X_i^{(k)}\right| \leq \frac{\|g_i\|}{L} \leq 1$ by Cauchy-Schwarz and $\|g_i\| \leq L$. Applying Corollary 17 with $\hat{X} = 0$ as the predictable sequence, we obtain, for any k and $\delta' \in (0,1)$,

$$\mathbb{P}\left(\exists t \leq T : \left|\sum_{i<t}\left(X_i^{(k)} - \mathbb{E}\left[X_i^{(k)} \;\middle|\; \mathcal{F}_{i-1}\right]\right)\right| \geq \sqrt{A_{t-1}(\delta')\sum_{i<t}(X_i^{(k)})^2 + B_{t-1}(\delta')}\right) \leq \delta', \tag{23}$$

where $A_t(\delta') = 16\log\left(\frac{60\log(6t)}{\delta'}\right)$ and $B_t(\delta') = 16\log^2\left(\frac{60\log(6t)}{\delta'}\right)$. Note that

$$X_i^{(k)} - \mathbb{E}\left[X_i^{(k)} \;\middle|\; \mathcal{F}_{i-1}\right] = \left\langle \Delta'_i, \Pi_1\left(\frac{x_i - x_\star}{s_k}\right)\right\rangle$$

and that

$$\sum_{i<t}(X_i^{(k)})^2 \leq \sum_{i<t}\|g_i\|^2/L^2 = G'_t.$$

Furthermore, note that, for $\delta' = \frac{\delta}{\log(6T)}$ we have that $A_t(\delta') \leq A_T(\delta') = 16C \leq (16/32^2)\alpha = \alpha'$ and that $B_t(\delta') \leq B_T(\delta') = 16C^2 \leq (16/(32L))^2\beta = \beta'$. Substituting to inequality (23) we conclude that

$$\mathbb{P}\left(\exists t \leq T : \sum_{i<t}\left\langle \Delta'_i, \Pi_1\left(\frac{x_i - x_\star}{s_k}\right)\right\rangle < -\sqrt{\alpha' G'_t + \beta'}\right) \leq \frac{\delta}{\log(6T)},$$

for all k, and therefore $\mathbb{P}\Big(\mathfrak{E}^c_{T,\alpha,\beta}\Big) \le \delta$ by the bound (22). ■

B.6. Proof of Lemma 13

Proof Fixing some $k \in \{2, 4, 8, \ldots\}$ and noting that $C_k = \log\Big(\frac{60\log^2(6B)}{2^{-2k}\delta}\Big)$, we may apply Proposition 12 with $T = B$, $\alpha = \alpha^{(k)}$, $\beta = \beta^{(k)}$ and failure probability $2^{-2k}\delta$, giving $1 - \mathbb{P}\Big(\mathfrak{E}_{B,\alpha^{(k)},\beta^{(k)}}(\eta)\Big) \le 2^{-2k}\delta$ for any η. Therefore, by the union bound

$$1 - \mathbb{P}\left(\bigcap_{j=0,1,\ldots,2^k} \mathfrak{E}_{B,\alpha^{(k)},\beta^{(k)}}(2^j\eta_\varepsilon)\right) \le (2^k+1)2^{-2k}\delta \le 2^{-(k-1)}\delta.$$

Applying the union bound once more, we have

$$1 - \mathbb{P}\left(\bigcap_{k=2,4,8,\ldots}\ \bigcap_{j=0,1,\ldots,2^k} \mathfrak{E}_{B,\alpha^{(k)},\beta^{(k)}}(2^j\eta_\varepsilon)\right) \le \sum_{k=2,4,8,\ldots} 2^{-(k-1)}\delta \le \sum_{k\ge 1} 2^{-k}\delta = \delta.$$

■

B.7. Proof of Theorem 14

Proof The bound B on the algorithm's query number is deterministic and therefore follows exactly as in the proof of Theorem 7. For the remainder of the analysis we assume the event $\bigcap_{k=2,4,8,\ldots}\bigcap_{j=0,1,\ldots,2^k} \mathfrak{E}_{B,\alpha^{(k)},\beta^{(k)}}(2^j\eta_\varepsilon)$ holds, which by Lemma 13 happens with probability at least $1-\delta$; we will show that, conditional on this event holding, the conclusions of the theorem hold deterministically. (Note that $\mathfrak{E}_{B,\alpha^{(k)},\beta^{(k)}}$ implies $\mathfrak{E}_{T_k,\alpha^{(k)},\beta^{(k)}}$ for all $T_k \le B$).

Next, we establish the lower bound (14) on T. Note that Algorithm 1 indeed always returns a point of the form $\bar{x} = \frac{1}{T}\sum_{i<T} x_i(\eta)$ with some $T \ge 1$; the edge case of returning in line 2 corresponds to $T = 1$. Moreover, in the typical case of returning in line 5, we have $T = T_{k_\mathrm{o}} = \left\lfloor \frac{B}{2k_\mathrm{o}} \right\rfloor \ge \frac{B}{4k_\mathrm{o}}$ for some $k_\mathrm{o} \le B/4$. By Lemma 10, we have

$$k_\mathrm{o} \le 2\log\log_+ \frac{\eta_{\max}(\alpha^{(0)},\beta^{(0)})}{\eta_\varepsilon} \overset{(\star)}{\le} 2\log\log_+\left(\frac{4\cdot 32}{32^2-2}\cdot\frac{d_0}{\eta_\varepsilon L}\right) \le 2\log\log_+\frac{d_0}{\eta_\varepsilon L}, \tag{24}$$

where $(\star)$ follows the facts that $\eta_{\max}(\alpha,\beta)$ is decreasing in α,β, and $\alpha^{(0)} \ge 32^2$ and $\beta^{(0)} \ge 32^2L^2$ by the setting (13). Thus we obtain the claimed lower bound (14) on $T = T_{k_\mathrm{o}}$.

It remains to show that one of the conclusions (15) and (16) must hold. When Algorithm 1 returns in line 5, we apply Proposition 9 on the final bisection performed by the algorithm, with parameters $T = T_{k_\mathrm{o}}$, $\alpha = \alpha^{(k_\mathrm{o})}$ and $\beta = \beta^{(k_\mathrm{o})}$. In the typical case that $\eta_\varepsilon \le \phi(\eta_\varepsilon)$, the proposition gives

$$\|\bar{x} - x_\star\| \le d_0 + \|\bar{x} - x_0\| \le \frac{5\alpha-2}{\alpha-2}d_0 \le \frac{5\cdot 32^2-2}{32^2-2}d_0 \le 6d_0$$

(where we have used $\alpha \geq \alpha^{(0)} \geq 32^2$), and, for some $\eta' \in [\eta, 2\eta]$,

$$f(\bar{x}) - f(x_\star) \leq \frac{9\alpha - 2}{2(\alpha - 2)} \cdot \frac{d_0\sqrt{\alpha G_T(\eta') + \beta}}{T} = O\left(\frac{d_0\sqrt{\alpha G_T(\eta') + \beta}}{T}\right).$$

Letting $C = \log \frac{\log_+ B}{\delta} + \log\log_+ \frac{\|x_\star - x_0\|}{\eta_\varepsilon L}$, we note that $\alpha = O(k_\mathrm{o} + \log \frac{\log_+ B}{\delta}) = O(C)$ and similarly $\beta = O(C^2L^2)$ due to the setting (13) and upper bound (24) on k_o. Therefore, $f(\bar{x}) - f(x_\star) \leq O\left(\frac{d_0\sqrt{CG_T(\eta') + C^2L^2}}{T}\right)$ and conclusion (15) holds.

In the edge case that $\eta_\varepsilon > \phi(\eta_\varepsilon)$ in the final bisection, we separately consider the cases $\eta_\varepsilon\sqrt{\alpha G_T(\eta_\varepsilon) + \beta} \leq 5d_0$ and $\eta_\varepsilon\sqrt{\alpha G_T(\eta_\varepsilon) + \beta} > 5d_0$. In the former case, Proposition 9 gives

$$\|\bar{x} - x_\star\| \leq d_0 + \|\bar{x} - x_0\| \leq d_0 + \eta_\varepsilon\sqrt{\alpha G_T(\eta_\varepsilon) + \beta} \leq 6d_0$$

and

$$f(\bar{x}) - f(x_\star) = O\left(\frac{(d_0 + \eta_\varepsilon\sqrt{\alpha G_T(\eta_\varepsilon) + \beta})\sqrt{\alpha G_T(\eta_\varepsilon) + \beta}}{T}\right) = O\left(\frac{d_0\sqrt{\alpha G_T(\eta_\varepsilon) + \beta}}{T}\right),$$

so conclusion (15) holds as before. In the second case, where $5d_0 < \eta_\varepsilon\sqrt{\alpha G_T(\eta_\varepsilon) + \beta}$, we have

$$\|\bar{x} - x_\star\| \leq d_0 + \|\bar{x} - x_0\| \leq d_0 + \eta_\varepsilon\sqrt{\alpha G_T(\eta_\varepsilon) + \beta} = O\left(\eta_\varepsilon\sqrt{\alpha G_T(\eta_\varepsilon) + \beta}\right)$$

and

$$f(\bar{x}) - f(x_\star) = O\left(\frac{(d_0 + \eta_\varepsilon\sqrt{\alpha G_T(\eta_\varepsilon) + \beta})\sqrt{\alpha G_T(\eta_\varepsilon) + \beta}}{T}\right) = O\left(\frac{\eta_\varepsilon(\alpha G_T(\eta_\varepsilon) + \beta)}{T}\right).$$

Recalling that $\alpha = O(C)$ and $\beta = O(C^2L^2)$, we see that conclusion (16) holds.

Finally, if Algorithm 1 returns in line 2 instead of line 5, we have $\bar{x} = x_0$ and $T = 1$, and so conclusion (15) holds trivially, since $\|x_0 - x_\star\| \leq 6\|x_0 - x_\star\|$ and $f(x_0) - f(x_\star) \leq \langle \nabla f(x_0), x_0 - x_\star\rangle \leq \|x_0 - x_\star\|L = O(d_0\sqrt{C^2L^2})$ due to convexity of f and Assumption 11. ■

B.8. A corollary for uniform gradient bounds

The following corollary translates Theorem 14 to the setting where we replace all observed gradient norms by L. In it λ represents a double-logarithmic factor and we use $\iota < 1$ to indicate low order terms which can be ignored as soon as $B = \Omega(\lambda^2)$.

Corollary 18 *Under Assumption 11, for any $\delta \in (0, 1)$ and $\varepsilon > 0$, Algorithm 1 with parameters $\alpha^{(k)}, \beta^{(k)}$ given by Equation (13), $B \in \mathbb{N}$ and $\eta_\varepsilon = \frac{\varepsilon}{L^2}$ makes at most B gradient queries and returns $\bar{x}$ such that, with probability at least $1 - \delta$,*

$$\|\bar{x} - x_\star\| \leq O\left(\max\left\{\|x_0 - x_\star\|, \frac{\sqrt{B(\lambda + \lambda^2\iota^2)}}{L}\varepsilon\right\}\right)$$

and

$$f(\bar{x}) - f(x_\star) \leq O\left(\frac{\|x_0 - x_\star\| L(\lambda + \lambda^{3/2}\iota)}{\sqrt{B}} + (\lambda + \lambda^2\iota^2)\varepsilon\right).$$

where

$$\lambda := \log\frac{1}{\delta}\log_+\frac{B\|x_0 - x_\star\|L}{\varepsilon} \quad \textit{and} \quad \iota = \sqrt{\frac{\lambda}{B+\lambda}}.$$

Proof The corollary follows by substitution of $\eta_\varepsilon = \frac{\varepsilon}{L^2}$ into Theorem 14. In particular, the bound (14) becomes

$$T = \Omega\left(\frac{B}{\lambda} + 1\right),$$

the quantity C in Theorem 14 satisfies

$$C = \log\frac{1}{\delta} + \log\log_+\frac{Bd_0}{\eta_\varepsilon L} = \lambda,$$

and the upper bound on $\|\bar{x} - x_\star\|$ in (16) is

$$O\left(\eta_\varepsilon\sqrt{CG_T(\eta_\varepsilon) + C^2L^2}\right) = O\left(\frac{\varepsilon}{L}\sqrt{\lambda T + \lambda^2}\right) = O\left(\frac{\sqrt{T}\sqrt{\lambda + \lambda^2/(1+B/\lambda)}}{L}\varepsilon\right),$$

which, when substituting $T \leq B$ and $\iota^2 = \lambda(\lambda + B)$ and combining with the bound on $\|\bar{x} - x_\star\|$ in (15) yields the claimed distance bound.

Finally, recalling that $T \geq 1$ and $T = \Omega(B/m)$, the suboptimality bound in (15) reads

$$f(\bar{x}) - f(x_\star) \leq O\left(\frac{d_0\sqrt{CG_T(\eta') + C^2L^2}}{T}\right) = O\left(\frac{d_0L}{\sqrt{B}}\sqrt{\lambda^2 + \frac{\lambda^3}{T}}\right) = O\left(\frac{d_0L(\lambda + \lambda^{3/2}\iota)}{\sqrt{B}}\right)$$

and the suboptimality bound in (16) reads

$$f(\bar{x}) - f(x_\star) \leq O\left(\frac{\eta_\varepsilon\left(CG_T(\eta_\varepsilon) + C^2L^2\right)}{T}\right) = O\left(\varepsilon\left(\lambda + \frac{\lambda^2}{T}\right)\right) = O((\lambda + \lambda^2\iota^2)\varepsilon);$$

combined, these yield the claimed suboptimality bound. ■

Appendix C. Proofs for Section 5

C.1. Proof of Theorem 15

Proof Recall the following basic property of any S-smooth functions [6, Lemma 3.4],

$$f(u) - f(v) - \langle\nabla f(v), u - v\rangle \leq \frac{S}{2}\|u - v\|^2 \quad \text{for all } u, v \in \mathbb{R}^d. \tag{25}$$

Using this fact we establish two useful inequalities for our proof. First, for any $\eta \leq \frac{1}{S}$ substituting $u = x_{i+1}(\eta)$ and $v = x_i(\eta)$ into (25) gives $\frac{\eta}{2}\|g_i(\eta)\|^2 \leq f(x_i(\eta)) - f(x_{i+1}(\eta))$. Summing over $i < T$ we obtain

$$\eta \leq \frac{1}{S} \implies \frac{\eta}{2}G_T(\eta) \leq f(x_0) - f(x_T(\eta)) \leq f(x_0) - f(x_\star) \overset{(\star)}{\leq} \frac{1}{2}Sd_0^2, \tag{26}$$

where $(\star)$ follows from (25) with $u = x_0$ and $v = x_\star$. Second, for any $\eta \geq 0$, substituting $x_i(\eta) - \frac{1}{S} g_i(\eta)$ and $v = x_t(\eta)$ into (25) yields $\|g_i(\eta)\|^2 \leq 2S\big[f(x_i(\eta)) - f\big(x_i(\eta) - \frac{1}{S} g_i(\eta)\big)\big] \leq 2S[f(x_i(\eta)) - f(x_\star)]$. Summing for $i < T$ yields

$$G_T(\eta) \leq 2S \sum_{i<T} [f(x_i(\eta)) - f(x_\star)] \text{ for all } \eta \geq 0. \tag{27}$$

We now split the proof into three cases based on the value of η_o. In each case, we show that $f(\bar{x}) - f(x_\star) = O\left(\frac{Sd_0^2}{T}\right)$ which, when combined with the lower bound (10) on T, gives the result.

First, consider the case that $\eta_o \geq \frac{1}{2S}$ (and hence also $\eta_o \neq \eta_\varepsilon$, so that final the call to ROOTFINDINGBISECTION returns at line 15). Then, by Lemma 2,

$$\begin{aligned}
\frac{1}{T} \sum_{i<T} f(x_i(\eta_o)) - f(x_\star) &\leq \frac{\bar{r}_T(\eta_o) d_0}{\eta_o T} + \frac{\eta_o G_T(\eta_o)}{2T} \\
&\overset{(i)}{\leq} \frac{2S\bar{r}_T(\eta_{lo}^\star) d_0}{T} + \frac{\bar{r}_T(\eta_{lo}^\star)\sqrt{G_T(\eta_o)}}{2T\sqrt{\alpha}} \overset{(ii)}{=} O\left(d_0 \frac{Sd_0 + \sqrt{G_T(\eta_o)}}{T}\right) \\
&\overset{(iii)}{=} O\left(d_0 \frac{Sd_0 + \sqrt{S\sum_{i<T}[f(x_i(\eta_o)) - f(x_\star)]}}{T}\right).
\end{aligned}$$

Above, transition (i) follows from substituting $\eta_o \geq \frac{1}{2S}$, $\eta_o \leq \frac{\bar{r}_T(\eta_{lo}^\star)}{\sqrt{\alpha G_T(\eta_o)}}$ and $\bar{r}_T(\eta_o) \leq \bar{r}_T(\eta_{lo}^\star)$ (the latter two bounds due to Lemma 4). Transition (ii) is from Lemma 3, where we recall that $\eta_{lo}^\star \leq \phi(\eta_{lo}^\star)$ by the bisection invariant. Transition (iii) follows from substituting the bound (27). If $\sqrt{\sum_{i=1}^T f(x_i(\eta)) - f(x_\star)} \leq \sqrt{S} d_0$ then the desired bound holds by squaring this inequality and recalling that $f(\bar{x}) \leq \frac{1}{T}\sum_{i<T} f(x_i(\eta_o))$ by convexity. Conversely, if $Sd_0 \leq \sqrt{S\sum_{i=1}^T f(x_i(\eta)) - f(x_\star)}$ then, due to the previous display,

$$\frac{1}{T}\sum_{i=1}^T f(x_i(\eta)) - f(x_\star) = O\left(d_0 \frac{\sqrt{S\sum_{i=1}^T f(x_i(\eta)) - f(x_\star)}}{T}\right)$$

and rearranging gives $\sqrt{\sum_{i=1}^T f(x_i(\eta)) - f(x_\star)} = O(\sqrt{S} d_0)$ again.

Second, consider the case that $\eta_\varepsilon < \eta_o \leq \frac{1}{2S}$ (so that the final call to ROOTFINDINGBISECTION still returns at line 15). Let $\eta_{lo}^\star$ and $\eta_{hi}^\star$ denote the final values of η_{lo} and η_{hi}, respectively, so that $\eta_{hi}^\star \leq 2\eta_o = \frac{1}{S}$. Let us bound the suboptimality of the $\eta_{hi}^\star$ iterates: beginning with Lemma 2, we have

$$\begin{aligned}
\frac{1}{T}\sum_{i<T} f(x_i(\eta_{hi}^\star)) - f(x_\star) &\leq \frac{\bar{r}_T(\eta_{hi}^\star) d_0}{\eta_{hi}^\star T} + \frac{\eta_{hi}^\star G_T(\eta_{hi}^\star)}{2T} \overset{(i)}{=} O\left(\frac{d_0\sqrt{G_T(\eta_{hi}^\star)} + Sd_0^2}{T}\right) \\
&\overset{(ii)}{=} O\left(d_0 \frac{Sd_0 + \sqrt{S\sum_{i<T}[f(x_i(\eta_{hi}^\star)) - f(x_\star)]}}{T}\right)
\end{aligned}$$

where (i) follows from the bisection invariant $\eta_{\text{hi}}^\star \geq \phi(\eta_{\text{hi}}^\star) = \frac{\bar{r}_T(\eta_{\text{hi}}^\star)}{\sqrt{\alpha G_T(\eta_{\text{hi}}^\star)}}$ and the bound (26), while (ii) follows from the bound (27). Therefore, by that same considerations as in the $\eta_{\text{o}} \geq \frac{1}{2S}$ case, we conclude that

$$\sum_{i<T}[f(x_i(\eta_{\text{hi}}^\star)) - f(x_\star)] = O(Sd_0^2). \tag{28}$$

If $\eta_{\text{o}} = \eta_{\text{hi}}^\star$ then we are done since $f(\bar{x}) \leq \frac{1}{T}\sum_{i<T} f(x_i(\eta_{\text{hi}}^\star))$ by convexity. Otherwise, applying Lemma 2 a final time gives

$$\begin{aligned}\sum_{i<T} f(\bar{x}) - f(x_\star) \leq \frac{\bar{r}_T(\eta_{\text{o}})d_0}{\eta_{\text{o}}T} + \frac{\eta_{\text{o}}G_T(\eta_{\text{o}})}{2T} &\overset{(i)}{=} O\left(\frac{d_0\sqrt{G_T(\eta_{\text{hi}}^\star)} + Sd_0^2}{T}\right) \\ &\overset{(ii)}{=} O\left(d_0\frac{Sd_0 + \sqrt{S\sum_{i<T}[f(x_i(\eta_{\text{hi}}^\star)) - f(x_\star)]}}{T}\right) \overset{(iii)}{=} O\left(\frac{Sd_0^2}{T}\right),\end{aligned}$$

where (i) follows form $\eta_{\text{o}} \geq \frac{\bar{r}_T(\eta_{\text{o}})}{\sqrt{\alpha G_T(\eta_{\text{hi}}^\star)}}$ (via Lemma 4) and the bound (26), (ii) is due to (27), and for (iii) we substituted our bound (28) on the error of the $\eta_{\text{hi}}^\star$ iterates.

Finally, the case where $\eta_{\text{o}} = \eta_\varepsilon < \frac{1}{2S}$ follows identically to the bound (28) since we have $\eta_{\text{o}} \geq \phi(\eta_{\text{o}})$ and $\eta_{\text{o}} \leq \frac{1}{S}$ in that case as well. ■

C.2. Proof of Theorem 16

Proof First, note that by Theorem 14, computing $x^{(M)}$ requires $\sum_{m=1}^M B^{(m)} = 2+4+\cdots+2^M = 2^{M+1} - 2 \leq B$ gradients queries as claimed. Next, note that

$$\sum_{m=1}^M \delta^{(m)} = \delta\sum_{m=1}^M \frac{1}{m(m+1)} = \delta\left(1 - \frac{1}{M+1}\right) \leq \delta,$$

and therefore by the union bound, with probability at least $1-\delta$ the conclusions of Theorem 14 hold for all applications of PARAMETERFREETUNER; we proceed with our analysis conditional on that event.

Note that $\delta^{(m)} \geq \delta^{(M)} = \Omega(\delta/\log^2 B)$ and $\eta_\varepsilon^{(m)} \geq \eta_\varepsilon^{(M)} = \Omega(\varepsilon/(L^2 B))$. Consequently, we have

$$\lambda := \log\left(\frac{1}{\delta^{(M)}}\log_+\frac{Bd_0}{\eta_\varepsilon^{(M)}L}\right) = O\left(\log\left(\frac{1}{\delta}\log_+\frac{Bd_0}{\varepsilon}\right)\right).$$

With this, we apply Theorem 14 on to bound the suboptimality of $x^{(m)}$ for $m \leq M$. Let $T^{(m)}$ be the corresponding T from Theorem 14, and note that $T^{(m)} \geq \max\{1, B^{(m)}/\lambda\}$. Noting also that $G_{T^{(m)}}(\eta')/T^{(m)} \leq L^2$ for all η', we have

$$\begin{aligned}f(x^{(m)}) - f(x_\star) &= O\left(\max\left\{\frac{\varepsilon(\lambda + \lambda^2/T^{(m)})}{B^{(m)}}, \frac{\|x^{(m-1)} - x_\star\|L\sqrt{\lambda^2 + \lambda^3/T^{(m)}}}{\sqrt{B^{(m)}}}\right\}\right) \\ &\leq \max\left\{\frac{\widetilde{\varepsilon}}{B^{(m)}}, \frac{\widetilde{L}\|x^{(m-1)} - x_\star\|}{\sqrt{B^{(m)}}}\right\},\end{aligned}$$

for some $\widetilde{\varepsilon} = O(\varepsilon\lambda^2)$ and $\widetilde{L} = O(L\lambda^{3/2})$, and all $m \leq M$. Applying strong convexity, we have that $\frac{\mu}{2}\|x^{(m-1)} - x_\star\|^2 \leq f(x^{(m-1)}) - f(x_\star)$ which implies

$$\begin{aligned} f(x^{(m)}) - f(x_\star) &\leq \max\left\{ \frac{\widetilde{\varepsilon}}{B^{(m)}}, \sqrt{\frac{2\widetilde{L}^2}{\mu B^{(m)}}(f(x^{(m-1)}) - f(x_\star))} \right\} \\ &\overset{(\star)}{\leq} \max\left\{ \frac{\widetilde{\varepsilon}}{B^{(m)}}, \frac{4\widetilde{L}^2}{\mu B^{(m)}}, \frac{f(x^{(m-1)}) - f(x_\star)}{2} \right\} \end{aligned}$$

where $(\star)$ follows from $\sqrt{ab} \leq \max\{2a, b/2\}$. Iterating this bound and noting that $2B^{(m-1)} = B^{(m)} = 2^m$ we conclude that

$$f(x^{(m)}) - f(x_\star) \leq \max\left\{ \frac{\widetilde{\varepsilon}}{B^{(m)}}, \frac{4\widetilde{L}^2}{\mu B^{(m)}}, \frac{f(x^{(0)}) - f(x_\star)}{B^{(m)}} \right\}. \tag{29}$$

Finally, the strong convexity and Lipschitz continuity assumptions imply that

$$\frac{\mu}{2}\|x^{(0)} - x_\star\|^2 \leq f(x^{(0)}) - f(x_\star) \leq L\|x^{(0)} - x_\star\|$$

and therefore $\|x^{(0)} - x_\star\| \leq \frac{2L}{\mu}$ and, using Lipschitz continuity again, $f(x^{(0)}) - f(x_\star) \leq \frac{2L^2}{\mu} \leq \frac{4\widetilde{L}^2}{\mu}$. Substituting back into (29), the second and third terms merge. Recalling that $B^{(M)} = 2^M = B/2$, and that $\widetilde{\varepsilon} = O(\varepsilon\lambda^2)$ and $\widetilde{L} = O(L\lambda^{3/2})$, we have

$$f(x^{(M)}) - f(x_\star) \leq \max\left\{ \frac{2\widetilde{\varepsilon}}{B}, \frac{8\widetilde{L}^2}{\mu B} \right\} = O\left(\frac{(L^2/\mu + \varepsilon)\lambda^3}{B} \right)$$

as required. ■

Appendix D. Additional discussion

D.1. Relaxing the assumption that $x_\star$ is optimal

Optimality gap bounds obtained via online-to-batch conversion have the appealing property of holding for any comparator $\mathring{x} \in \mathcal{X}$ and not necessarily a minimizer of f [15]. Consequently, the parameter-free regret minimization algorithm of McMahan and Orabona [24] outputs a point $\bar{x}$ with an error bound of the form

$$\mathbb{E}f(\bar{x}) \leq f(x') + \varepsilon + O\left(\frac{L\|x_0 - x'\|}{\sqrt{T}} \sqrt{\log\left(1 + \frac{L\|x_0 - x'\|}{\varepsilon\sqrt{T}}\right)} \right) \text{ for all } x' \in \mathcal{X}.$$

In contrast, we only provide guarantees for $x' = x_\star$, a minimizer of f. This can be restrictive in settings where $\|x_0 - x_\star\|$ is very large or possibly infinite, i.e., when the minimum of f is not attained, as is the case in logistic regression on separable data.

However, the assumption that $x_\star$ is optimal can be relaxed. In particular, our only real requirement from $x_\star$ is that, for every SGD iterate $x_t(\eta)$ evaluated in Algorithm 1, we have $f(x_t(\eta)) -$

$f(x_\star) \geq 0$. In the noiseless setting, we may modify Algorithm 1 to return the GD iterate with lowest objective value (from all the GD executions combined). Such modified algorithm would satisfy the error bounds in Theorem 7 with respect to an *arbitrary* point $x_\star$: if the algorithm's output has function value smaller than $f(x_\star)$, we are done; otherwise, we have $f(x_t(\eta)) - f(x_\star) \geq 0$ for every GD iterate, and our analysis goes through. (Note, however, that we lose the guarantee on the distance between $x_\star$ and the algorithm's output). Extension to the stochastic case is more involved since we do not have the privilege of choosing the best SGD iterate; we leave it to future work.

D.2. An alternative bisection target without gradient norm adaptivity

Algorithm 1 is fairly adaptive to stochastic gradient norms, with performance guarantees that depend mainly on observed norms, featuring an a-priori gradient norm bound in low-order terms. Moreover, in the noiseless setting our method requires no a-priori bound on gradient norms and our bounds depend solely on observed norms.

It is possible, however, to slightly simplify our method and sharpen some of our bounds by forgoing adaptivity to gradient norms. Specifically, if we only seek guarantees that depend on an a-priori gradient norm bound L, then it is possible to replace the bisection target ϕ defined in line 8 of Algorithm 1 with

$$\phi(\eta) = \frac{\bar{r}_T(\eta)}{\sqrt{\alpha L^2 T}}.$$

Our analysis applies to this modified bisection target as well, but with $G_T(\eta)$ replaced by $L^2 T$ throughout. Moreover, this modification allows us to slightly improve two parts of the analysis.

First, we may sharpen the bound on $\eta_{\max}$ in Lemmas 6 and 10 to $\eta_{\max} = O\left(\frac{d_0}{L\sqrt{T}}\right)$, improving our bound on the maximum value of k used in Algorithm 1. In the deterministic case, this allows us to establish optimality gap bounds scaling as $\varepsilon + \lambda' \frac{d_0 L}{\sqrt{B}}$ for $\lambda' = O\left(\sqrt{\log\log_+ \frac{d_0 L}{\varepsilon\sqrt{B}}}\right)$ which satisfies $\lambda' = O(1)$ for $\varepsilon = \frac{d_0 L}{\sqrt{B}}$, similarly to the bounds of previous works, as discussed at the end of Section 4.

Second, in the stochastic setting, we may use Blackwell's inequality [5] instead of the time uniform empirical Bernstein. This allows us to take $\alpha^{(k)} = 2k + O\big(\log\big(\frac{1}{\delta}\log B\big)\big)$, eliminating the additive square logarithmic term stemming from $\beta^{(k)}$ in (13). Consequently, in the stochastic setting we obtain a probability $1-\delta$ optimality gap bound of $\varepsilon + \lambda'' \frac{d_0 L}{\sqrt{B}}$ for $\lambda'' = O\left(\lambda'\big[\lambda' + \sqrt{\log\big(\frac{1}{\delta}\log B\big)}\big]\right)$, with λ' defined above. Therefore, in the stochastic case we do not remove $\log\log B$ term entirely, even for the optimal ε. The source of the remaining $\log\log B$ is the union bound we use in the proof of Proposition 12, which might be removable via a more careful probabilistic argument.