The Complexity of Finding Stationary Points with Stochastic Gradient Descent

Yoel Drori [*1] **Ohad Shamir** [*1 2]

Abstract

We study the iteration complexity of stochastic gradient descent (SGD) for minimizing the gradient norm of smooth, possibly nonconvex functions. We provide several results, implying that the classical $\mathcal{O}(\epsilon^{-4})$ upper bound (for making the average gradient norm less than ϵ) cannot be improved upon, unless a combination of additional assumptions is made. Notably, this holds even if we limit ourselves to convex quadratic functions. We also show that for nonconvex functions, the feasibility of minimizing gradients with SGD is surprisingly sensitive to the choice of optimality criteria.

1. Introduction

Stochastic gradient descent (SGD) is today one of the main workhorses for solving large-scale supervised learning and optimization problems. Much of its popularity is due to its extreme simplicity: Given a function f and an initialization point $\mathbf{x}$, we perform iterations of the form $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \mathbf{g}_t$, where $\eta_t > 0$ is a step-size parameter and $\mathbf{g}_t$ is a stochastic vector which satisfies $\mathbb{E}[\mathbf{g}_t|\mathbf{x}_t] = \nabla f(\mathbf{x}_t)$. For example, in the context of machine learning, $f(\mathbf{x})$ might be the expected loss of some predictor parameterized by $\mathbf{x}$ (over some underlying data distribution) and $\mathbf{g}_t$ is the gradient of the loss w.r.t. a single data sample. For convex problems, the convergence rate of SGD to a global minimum of f has been very well studied (for example, (Kushner & Yin, 2003; Nemirovski et al., 2009; Moulines & Bach, 2011; Bertsekas, 2011; Rakhlin et al., 2012; Bottou et al., 2018)), however, for nonconvex problems, convergence to a global minimum cannot in general be guaranteed. A reasonable substitute is to study the convergence to local minima, or at the very least, to stationary points. This can also be quantified as an optimization problem where the goal is not to minimize $f(\mathbf{x})$ over $\mathbf{x}$, but rather $\|\nabla f(\mathbf{x})\|$. This question of finding stationary points has gained more attention in recent years, with the rise of deep learning and other large-scale nonconvex optimization methods.

Compared to optimizing function values, the convergence of SGD in terms of minimizing the gradient norm is relatively less well-understood. A folklore result (see e.g., (Ghadimi & Lan, 2013), which we repeat in Appendix B for completeness, as well as (Allen-Zhu, 2018)) states that for smooth (Lipschitz gradient) functions, $\mathcal{O}(\epsilon^{-4})$ iterations are sufficient to make the average expected gradient $\mathbb{E}[\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\|]$ less than ϵ, and it was widely conjectured that this is the best complexity achievable with SGD. However, this bound was recently improved in Fang et al. (Fang et al., 2019), which showed a complexity bound of $\mathcal{O}(\epsilon^{-3.5})$ for SGD, under the following additional assumptions/algorithmic modifications:

1. **(Complex) aggregation.** Rather than considering the average or minimal gradient norm of the iterates, the algorithm considers the norm of a certain adaptive average of a suffix of the iterates (those which do not deviate too much from the final iterate).

2. **Lipschitz Hessian.** The function is twice differentiable, with a Lipschitz Hessian as well as a Lipschitz gradient.

3. **"Dispersive" noise.** The stochastic noise satisfies a "dispersive" property, which intuitively implies that it is well-spread (it is satisfied, for example, for Gaussian or uniform noise in some ball).

4. **Bounded dimension.** The dimension is bounded, in the sense that there is an explicit logarithmic dependence on it in the iteration complexity bound (in contrast, the folklore $\mathcal{O}(\epsilon^{-4})$ result is dimension-free).

The result of Fang et al. is even stronger, as it shows convergence to a *second-order* stationary point (where the Hessian is nearly positive definite), however, this will not be our focus here. Note that in this setting it is known that some dimension dependence is difficult to avoid (see (Simchowitz et al., 2017))

[1]Google Research [2]Weizmann Institute of Science. Correspondence to: Yoel Drori <dyoel@google.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

In this paper, we study the performance limits of SGD for minimizing gradients through several variants of lower bounds under different assumptions. In particular, we wish to understand which of the assumptions/modifications above are necessary to break the ϵ^{-4} barrier. Our main take-home message is that most of these appear to be needed in order to attain an iteration complexity better than $\mathcal{O}(\epsilon^{-4})$, in some cases even if we limit ourselves just to convex quadratic functions. In a bit more detail:

- If we drop Assumption 4 (bounded dimension), and consider the norm of the gradient at the output of some fixed, deterministic aggregation scheme (as opposed to returning, for example, an iterate with a minimal gradient norm), then perhaps surprisingly, we show that it is impossible to provide *any* finite complexity bound. This holds under mild algorithmic conditions, which extend far beyond SGD. This implies that for dimension-free bounds, we must either consider rather complicated aggregation schemes, apply randomization, or use optimality criteria which do not depend on a single point (e.g., consider the average gradient $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\|$ or $\min_t \|\nabla f(x_t)\|$, as is often done in the literature). This result is formalized as Thm. 1 in Subsection 3.1.
- Without Assumption 2 (Lipschitz Hessian) and Assumption 3 (dispersive noise), then even with rather arbitrary aggregation schemes, the iteration complexity of SGD is $\Omega(\epsilon^{-4})$. This result is formalized as Thm. 2 in Subsection 3.2.
- Without Assumption 1 (aggregation) and Assumption 3 (dispersive noise), the iteration complexity of SGD required to satisfy $\mathbb{E}[\min_t \|\nabla f(\mathbf{x}_t)\|] \leq \epsilon$ is $\Omega(\epsilon^{-3})$. This result is formalized as Thm. 3 in Subsection 3.2.
- Without aggregation, the iteration complexity of SGD with "reasonable" step sizes to attain $\mathbb{E}[\min_t \|\nabla f(\mathbf{x}_t)\|] \leq \epsilon$ is $\Omega(\epsilon^{-4})$, even for quadratic *convex* functions in moderate dimension and Gaussian noise (namely, all other assumptions are satisfied as well as convexity). This result is formalized as Thm. 4 in Section 4.

It is important to note that the SGD algorithm, which is the main focus of this paper, is not necessarily an optimal algorithm (in terms of iteration complexity) for minimizing gradient norms in our stochastic optimization setting. For example, for convex problems, it is known that it is possible to achieve an iteration complexity of $\tilde{O}(\epsilon^{-2})$, which strictly smaller than our $\Omega(\epsilon^{-4})$ lower bound (see (Foster et al., 2019), and for a related result in the deterministic setting see (Nesterov, 2012)). These algorithms are more complicated and less natural than plain SGD, a price that our results indicate might be necessary in order to achieve optimal iteration complexity in some cases.

We conclude this section by noting that following the initial dissemination of our paper, a recent arXiv preprint (Arjevani et al., 2019) studied a similar question of lower complexity bounds for finding stationary points, focusing on algorithm-independent $\Omega(\epsilon^{-4})$ or $\Omega(\epsilon^{-3})$ lower bounds for functions with Lipschitz-continuous gradients. Their results are mostly incomparable to ours. In particular, our Thm. 1 studies conditions under which no finite lower bound is possible, Thm. 3 considers the case where the Hessian (and not just the gradient) is Lipschitz-continuous, and Thm. 4 shows an $\Omega(\epsilon^{-4})$ lower bound for SGD, which holds *even* if the functions are convex and the noise is simply Gaussian (in contrast, the constructions in Arjevani et al. (2019) crucially depend on intricate non-convex functions and carefully tailored, location-dependent noise, using a considerably more involved proof). The result most similar to those in Arjevani et al. (2019) is Thm. 2, which is specific to SGD, but admits a simpler proof and significantly better constants.

2. Setting and Notation

We let bold-face letters denote vectors, use $\mathbf{e}_i$ to denote the canonical unit vector, and use $[T]$ as shorthand for $\{1, 2, \ldots, T\}$.

We assume throughout that the objective f maps $\mathbb{R}^d$ to $\mathbb{R}$, and either has an L-Lipschitz gradient for some fixed parameter $L > 0$ or a ρ-Lipschitz Hessian for some $\rho > 0$.

We consider algorithms which use a standard stochastic first-order oracle (Nemirovski & Yudin, 1983; Agarwal et al., 2009) in order to minimize some optimality criteria: This oracle, given a point $\mathbf{x}_t$, returns $\nabla f(\mathbf{x}_t) + \boldsymbol{\xi}_t$, where $\boldsymbol{\xi}_t$ is a random variable satisfying

$$\mathbb{E}[\boldsymbol{\xi}_t|\mathbf{x}_t] = 0 \quad \text{and} \quad \mathbb{E}[\|\boldsymbol{\xi}_t\|^2|\mathbf{x}_t] \leq \sigma^2$$

almost surely for some fixed σ^2. In this paper, we focus on optimality criteria involving minimizing gradient norms, using the Stochastic Gradient Descent (SGD) algorithm. This algorithm, given a budget of T iterations and an initialization point $\mathbf{x}_1$, produces T stochastic iterates $\mathbf{x}_1, \ldots, \mathbf{x}_T$ according to

$$\mathbf{x}_{t+1} \;=\; \mathbf{x}_t - \eta_t \cdot (\nabla f(\mathbf{x}_t) + \boldsymbol{\xi}_t)\,, \tag{1}$$

where η_t is a fixed step-size parameter. In some cases, we will also allow the algorithm to perform an additional aggregation step, generating a point $\mathbf{x}_{\text{out}}$ which is some function of $\mathbf{x}_1, \ldots, \mathbf{x}_T$ (for example, the average $\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t$). Additionally, in some of our results, we will allow the step

size to be adaptive, and depend on the previous iterates (under appropriate assumptions), in which case we will use the notation

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_{\mathbf{x}_1,\ldots,\mathbf{x}_t} \cdot (\nabla f(\mathbf{x}_t) + \boldsymbol{\xi}_t). \tag{2}$$

Regarding the initial conditions, we make the standard assumption[1] that $\mathbf{x}_1$ has bounded suboptimality, i.e., $f(\mathbf{x}_1) - f(\mathbf{x}_*) \leq \Delta$ for some fixed $\Delta > 0$, where in the convex case, we assume $\mathbf{x}_*$ is some point $\mathbf{x}_* \in \arg\min_{\mathbf{x}} f(\mathbf{x})$, and in the non-convex case, we assume $\mathbf{x}_*$ is a stationary point with $f(\mathbf{x}_*) \leq f(\mathbf{x}_t)$ for all $t \in [T]$. We note that some analyses (see for example (Allen-Zhu, 2018; Foster et al., 2019)) replace the assumption $f(\mathbf{x}_1) - f(\mathbf{x}_*) \leq \Delta$ with the assumption $\|\mathbf{x}_1 - \mathbf{x}_*\| \leq R$, but we do not consider this variant in this paper (in fact, some of our constructions rely on the fact that even if $f(\mathbf{x}_1) - f(\mathbf{x}_*)$ is small, $\|\mathbf{x}_1 - \mathbf{x}_*\|$ might be very large). It should also be pointed out that in the non-convex setting, $\mathbf{x}_*$ might not be uniquely defined or even belong to a single connected set, which makes $\|\mathbf{x}_1 - \mathbf{x}_*\|$ somewhat ambiguous.

3. Lower Bounds in the Non-Convex Case

In this section, we present several lower bounds relating to first-order methods in the non-convex stochastic setting. We start by considering a wide range of first-order methods, showing that if we consider any point which is a *fixed* function of the iterates, then *no* meaningful, dimension-free worst-case bound can be attained on its expected gradient norm. We conclude that it is necessary for any useful optimality criterion to relate to more than one iterate in some way, as is indeed the case with the standard optimality criteria, which considers the average expected norm of the gradients ($\frac{1}{T}\sum_t \mathbb{E}\|\nabla f(\mathbf{x}_t)\|$) or the minimal expected norm of the gradients ($\min_t \mathbb{E}\|\nabla f(\mathbf{x}_t)\|$).

We then turn our focus to the SGD method under the standard set of assumptions (see Sec. 2), and show that it requires $\Omega(\epsilon^{-4})$ iterations (or $\Omega(\epsilon^{-3})$ with Lipschitz Hessians) to attain a value of ϵ for any of the standard optimality criteria mentioned above.

3.1. Impossibility of minimizing the gradient at any fixed point

In this subsection, we show that in the nonconvex setting, perhaps surprisingly, *no* meaningful iteration complexity bound can be provided on $\|\nabla f(\mathbf{x}_{\text{out}})\|$, where $\mathbf{x}_{\text{out}}$ is the point returned by any fixed, deterministic aggregation scheme which depends continuously on the iterates and stochastic gradients (for example, some fixed weighted combination of the iterates).

To state the result, recall that SGD can be phrased in an oracle-based setting, where we model an optimization algorithm as interacting with a stochastic first-order oracle: Given an initial point $\mathbf{x}_1$, at every iteration $t = 2, \ldots, T$, the algorithm chooses a point $\mathbf{x}_t$, and the oracle returns a stochastic gradient estimate $\mathbf{g}_t := \nabla f(\mathbf{x}_t) + \boldsymbol{\xi}_t$, where $\mathbb{E}[\boldsymbol{\xi}_t|\mathbf{x}_t] = 0$ and $\mathbb{E}[\|\boldsymbol{\xi}_t\|^2|\mathbf{x}_t] \leq \sigma^2$ for some known σ^2. The algorithm then uses $\mathbf{g}_t$ (as well as $\mathbf{g}_1, \ldots, \mathbf{g}_{t-1}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_t$) to select a new point $\mathbf{x}_{t+1}$. After T iterations, the algorithm returns a final point $\mathbf{x}_{\text{out}}$, which depends on $\mathbf{g}_1, \ldots, \mathbf{g}_T$ and $\mathbf{x}_1, \ldots, \mathbf{x}_T$.

Theorem 1. *Consider any deterministic algorithm as above, which satisfies the following:*

- *There exists a finite C_T (dependent only on T) such that for any initialization $\mathbf{x}_1$ and any $t \in [T]$, if $\mathbf{g}_1 = \ldots = \mathbf{g}_t = \mathbf{0}$, then $\|\mathbf{x}_{t+1} - \mathbf{x}_1\| \leq C_T$. Moreover, if this holds for $t = T$, then $\|\mathbf{x}_{\text{out}} - \mathbf{x}_1\| \leq C_T$.*
- *For any $t \in [T]$, $\mathbf{x}_{t+1}$ is a fixed continuous function of $\mathbf{x}_1, \mathbf{g}_1, \ldots, \mathbf{x}_t, \mathbf{g}_t$, and $\mathbf{x}_{\text{out}}$ is a fixed continuous function of $\mathbf{x}_1, \mathbf{g}_1, \ldots, \mathbf{x}_T, \mathbf{g}_T$.*

Then for any $\delta \in (0, 1)$, and any choice of random variables $\boldsymbol{\xi}_t$ satisfying the assumptions above, there exists a dimension d, a twice-differentiable function $f : \mathbb{R}^d \mapsto \mathbb{R}$ with 2-Lipschitz gradients and 4-Lipschitz Hessians, and an initialization point $\mathbf{x}_1$ satisfying $f(\mathbf{x}_1) - \inf_{\mathbf{x}} f(\mathbf{x}) \leq 1$, such that $\|\nabla f(\mathbf{x}_{\text{out}})\| \geq \frac{1}{2}$ holds with probability at least $1 - \delta$. Moreover, if there is no stochastic noise ($\boldsymbol{\xi}_t \equiv \mathbf{0}$), then the result holds for $d = 1$.

Intuitively, the first condition in the theorem requires that the algorithm does not "move" too much from the initialization point $\mathbf{x}_1$ if all stochastic gradients are zero (this is trivially satisfied for SGD, and any other reasonable algorithm we are aware of), while the second condition requires the iterates produced by the algorithm to depend continuously on the previous iterates and stochastic gradients (again, this is satisfied by SGD). By constructing a one-dimensional function whose gradient is zero over two disjoint regions, these two conditions allow the application of the intermediate value theorem to find an initialization point such that function value at $\mathbf{x}_{\text{out}}$ attains any value in between the function values at the regions.

The theorem suggests that to get non-trivial results, we must either use a dimension-dependent analysis, use a non-continuous/adaptive/randomized scheme to compute $\mathbf{x}_{\text{out}}$, or measure the performance of the generated sequence using an optimality criterion that does not depend on a fixed point (e.g., the average gradient $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\|$ or $\min_t \|\nabla f(\mathbf{x}_t)\|$). We note that the positive result of (Fang et al., 2019) assumes both finite dimension, and computes $\mathbf{x}_{\text{out}}$ according to an adaptive non-continuous decision rule (involving branching depending on how far the

[1] See e.g. (Nesterov, 2004) and references mentioned earlier.

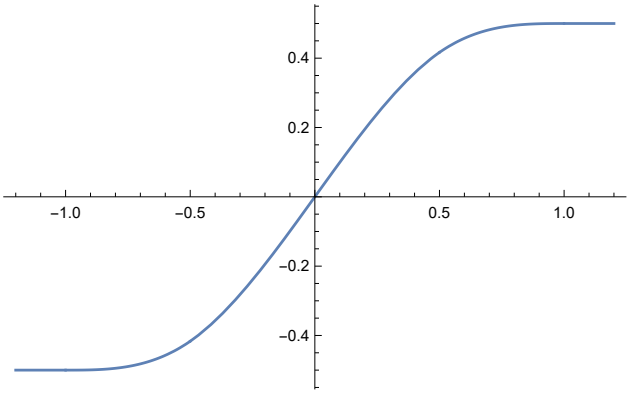

Figure 1. The function $s(x)$.

iterates have moved), hence there is no contradiction to the alluded theorem.

Proof of Thm. 1. We will first prove the result in the case where there is no noise, i.e. $\mathbf{g}_t = \nabla f(\mathbf{x}_t)$ deterministically, in which case $\mathbf{x}_2, \ldots, \mathbf{x}_T$ and $\mathbf{x}_{\text{out}}$ are deterministic functions of $\mathbf{x}_1$. To that end, let $d = 1$ and let $f(x) = s(x)$, where s is the sigmoid-like function (see Fig. 1)

$$s(x) = \begin{cases} -\frac{1}{2} & x \leq -1, \\ \frac{2}{3}(x+1)^3 - \frac{1}{2} & x \in [-1, -\frac{1}{2}], \\ -\frac{2}{3}x^3 + x & x \in [-\frac{1}{2}, \frac{1}{2}], \\ \frac{2}{3}(x-1)^3 + \frac{1}{2} & x \in [\frac{1}{2}, 1], \\ \frac{1}{2} & x \geq 1. \end{cases}$$

This function smoothly and monotonically interpolates between $-1/2$ at $x = -1$ at $1/2$ at $x = 1$. It can be easily verified to have 2-Lipschitz gradients and 4-Lipschitz Hessians, and for any x, satisfies $f(x) - \inf_x f(x) \leq 1$.

Let us consider the iterates generated by the algorithm, $x_1, \ldots, x_T$ and x_{out}, as we make $x_1 \to \infty$. Our function is such that $\nabla f(x) = 0$ for all $x \geq 1$, so at every iteration, the algorithm gets $g_t = 0$ as long as $x_t \geq 1$. Moreover, by the assumptions, as long as the gradients are zero, $|x_t - x_1|$ is bounded. As a result, by induction and our assumption that $|x_{\text{out}} - x_1|$ is bounded, we get that $x_{\text{out}} \to \infty$. A similar argument shows that when $x_1 \to -\infty$, we also have $x_{\text{out}} \to -\infty$.

Next, we argue that x_{out} is a continuous function of x_1. Indeed, x_2 is a continuous function of x_1, since it is a continuous function of $g_1 = \nabla f(x_1)$ by assumption, and $\nabla f(x_1)$ is Lipschitz (hence continuous) in x_1, and compositions of continuous functions is continuous. By induction, a similar argument holds for x_t for any t, and hence also to x_{out}.

Overall, we showed that x_{out} is a continuous function of x_1, that $x_{\text{out}} \to \infty$ when $x_1 \to \infty$, and that $x_{\text{out}} \to -\infty$ when $x_1 \to -\infty$. Therefore, by the mean value theorem, there exists some x_1 for which x_{out} is precisely zero, in which case $|f'(x_{\text{out}})| = |f'(0)| = |s'(0)| = 1$, satisfying the Theorem statement.

It remains to prove the theorem in the noisy case, where $\boldsymbol{\xi}_i$ are non-zero random variables. In that case, instead of choosing $f(x) = s(x)$, we let $f(\mathbf{x}) = s(\langle \mathbf{x}, \mathbf{e}_r \rangle)$, where the coordinate r is defined as

$$r := \arg\min_{j \in [d]} \max_{t \in [T]} \mathbb{E}[\langle \boldsymbol{\xi}_t, \mathbf{e}_j \rangle^2].$$

Since $\max_t \mathbb{E}[\|\boldsymbol{\xi}_t\|^2] = \max_t \sum_{j=1}^d \mathbb{E}[\langle \boldsymbol{\xi}_t, \mathbf{e}_j \rangle^2]$ is bounded by σ^2 independently of d, it follows that the variance of $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_t$ along coordinate r goes to zero as $d \to \infty$. Therefore, by making d large enough and using Chebyshev's inequality, we can ensure that $\max_t |\langle \boldsymbol{\xi}_t, \mathbf{e}_j \rangle|$ is arbitrarily small with arbitrarily high probability. Since the gradients of f are Lipschitz, and we assume each $\mathbf{x}_{t+1}$ is a continuous function of the noisy gradients $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_t$, it follows that the trajectory of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ and $\mathbf{x}_{\text{out}}$ on the j-th coordinate can be made arbitrarily close to the noiseless case analyzed earlier (where $\boldsymbol{\xi}_t \equiv \mathbf{0}$), with arbitrarily high probability. In particular, we can find an initialization point $\mathbf{x}_1$ such that the j-th coordinate of $\mathbf{x}_{\text{out}}$ is arbitrarily close to 0, hence the gradient is arbitrarily close to 1 (and in particular, larger than $1/2$). □

Remark 1 (Randomized Algorithms)**.** The theorem considers deterministic algorithms for simplicity, but the same proof idea holds for larger families of randomized algorithms, where the randomness is used "obliviously". For example, consider the popular technique of adding random perturbations to the iterates: If the perturbations have a fixed distribution with finite variance, then we can always embed our construction in a high enough dimension, so that the effective variance of the perturbations is arbitrarily small, and we are back to the deterministic setting.

3.2. Lower bounds on SGD

In this subsection, we focus on the analysis of SGD in the nonconvex setting. We present two main results: A lower bound on the performance of SGD with an aggregation step for objectives with L-Lipschitz gradient, followed by a lower bound in the case where the objective has ρ-Lipschitz Hessian that applies to "plain" SGD methods that do not perform an aggregation step. In both cases, the step sizes chosen by the method are allowed to be adaptive, in the sense that they are allowed to depend on past iterates and gradients. This dependence is *not* allowed to be completely general, but rather we assume that the dependence on the past iterates and gradients is done through a function of their norm and the dot-products between them (in the Lipschitz Hessian case, we also allow the step size to depend

on the Hessians). Note, that all commonly used adaptive schemes (including Adagrad (Duchi et al., 2011), normalized gradient (Nesterov, 1984; Kiwiel, 2001), among others) follow this type of adaptive scheme.

We start the analysis with a technical lemma.

Lemma 1. *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a function with L-Lipschitz gradient, and assume that the vectors $\mathbf{y}_1, \ldots, \mathbf{y}_m, \mathbf{z}_1, \ldots, \mathbf{z}_n, \boldsymbol{\gamma} \in \mathbb{R}^d$ ($n, m \in \mathbb{N}$) are such that*

1. $\nabla f(\mathbf{y}_1) = \cdots = \nabla f(\mathbf{y}_m) = \nabla f(\mathbf{z}_1) = \cdots = \nabla f(\mathbf{z}_n) = \boldsymbol{\gamma}$,
2. $f(\mathbf{y}_1) = \cdots = f(\mathbf{y}_m)$, *and*
3. $\langle \boldsymbol{\gamma}, \mathbf{y}_1 \rangle = \cdots = \langle \boldsymbol{\gamma}, \mathbf{y}_m \rangle$.

Then there exists a function $\hat{f}$ with L-Lipschitz gradient that has the same first-order information as f at $\{\mathbf{z}_i\}_{i\in[n]}$, i.e., for all $i \in [n]$

$$\hat{f}(\mathbf{z}_i) = f(\mathbf{z}_i),$$
$$\nabla \hat{f}(\mathbf{z}_i) = \nabla f(\mathbf{z}_i) = \boldsymbol{\gamma},$$

has the same gradient as f at $\{\mathbf{y}_i\}_{i\in[n]}$, i.e., for all $j \in [m]$

$$\nabla \hat{f}(\mathbf{y}_j) = \nabla f(\mathbf{y}_j) = \boldsymbol{\gamma},$$

and is bounded from below:

$$\inf_{x\in\mathbb{R}^d} \hat{f}(x) \geq \min_{k\in[n]} f(\mathbf{z}_k) - \frac{3}{2L}\|\boldsymbol{\gamma}\|^2.$$

We postpone the proof of this lemma to the appendix and turn to present the first main result of this subsection.

Theorem 2. *Consider a first-order method that given a function $f : \mathbb{R}^d \to \mathbb{R}$ and an initial point $\mathbf{x}_1 \in \mathbb{R}^d$ generates a sequence of $T \in \mathbb{N}$ points $\{\mathbf{x}_i\}$ satisfying*

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \eta_{\mathbf{x}_1,\ldots,\mathbf{x}_t} \cdot (\nabla f(\mathbf{x}_t) + \boldsymbol{\xi}_t), \quad t \in [T-1],$$

where $\boldsymbol{\xi}_i$ are some random noise vectors, and returns a point $\mathbf{x}_{\text{out}} \in \mathbb{R}^d$ as a non-negative linear combination of the iterates:

$$\mathbf{x}_{\text{out}} = \sum_{t=1}^{T} \zeta^{(t)}_{\mathbf{x}_1,\ldots,\mathbf{x}_T} \mathbf{x}_t.$$

We further assume that the step sizes $\eta_{\mathbf{x}_1,\ldots,\mathbf{x}_t}$ and aggregation coefficients $\zeta^{(t)}_{\mathbf{x}_1,\ldots,\mathbf{x}_T}$ are deterministic functions of the norms and inner products between the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_t, \nabla f(\mathbf{x}_1) + \boldsymbol{\xi}_1, \ldots, \nabla f(\mathbf{x}_t) + \boldsymbol{\xi}_t$. Then for any $L, \Delta, \sigma \in \mathbb{R}_{++}$ there exists a function $f : \mathbb{R}^T \mapsto \mathbb{R}$ with L-Lipschitz gradient, a point $\mathbf{x}_1 \in \mathbb{R}^T$ and independent random variables $\boldsymbol{\xi}_t$ with $\mathbb{E}[\boldsymbol{\xi}_t] = 0$ and $\mathbb{E}[\|\boldsymbol{\xi}_t\|^2] = \sigma^2$ such that

$$f(\mathbf{x}_1) - \inf_{\mathbf{x}} f(\mathbf{x}) \overset{a.s.}{\leq} \Delta,$$
$$\nabla f(\mathbf{x}_t) \overset{a.s.}{=} \boldsymbol{\gamma}, \qquad \forall t \in [T],$$
$$\nabla f(\mathbf{x}_{\text{out}}) \overset{a.s.}{=} \boldsymbol{\gamma},$$

where $\boldsymbol{\gamma} \in \mathbb{R}^T$ is a vector such that

$$\|\boldsymbol{\gamma}\|^2 = \frac{\sigma}{16(T-1)}\left(\sqrt{64L\Delta(T-1)+9\sigma^2} - 3\sigma\right) \underset{T\gg 1}{\approx} \frac{\sigma}{2}\sqrt{\frac{L\Delta}{T-1}}.$$

Proof. We will assume that the algorithm performs gradient steps with fixed step-size η_t and aggregation coefficients ζ_i, i.e., the algorithm is defined by the rule

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t(\nabla f(\mathbf{x}_t) + \boldsymbol{\xi}_t), \quad t \in [T-1],$$
$$\mathbf{x}_{\text{out}} = \sum_{t=1}^{T} \zeta_i \mathbf{x}_i.$$

The analysis of the general case appears in the appendix.

Under this assumption, the proof proceeds by 1. defining an adversarial objective and noise distribution, 2. showing that the gradients of the objective at the iterates posses the claimed properties, then 3. using Lemma 1, modifying the objective so that the claimed lower bound on the function is attained, while keeping the behavior of the function at the iterates unaffected.

We start by defining an adversarial example, choosing the noise vectors $\{\boldsymbol{\xi}_t\}$ to be independent random variables distributed such that

$$P(\boldsymbol{\xi}_t = \pm\sigma\mathbf{e}_{t+1}) = \frac{1}{2}, \quad t \in [T-1],$$

where $\mathbf{e}_i$ stands for the canonical unit vector, and defining the objective $f := f_{\{\eta_t\},\{\zeta_t\}} : \mathbb{R}^T \mapsto \mathbb{R}$ by

$$f_{\{\eta_t\},\{\zeta_t\}}(\mathbf{x}) := G \cdot \langle \mathbf{x}, \mathbf{e}_1 \rangle + \sum_{t=1}^{T-1} h_t(\langle \mathbf{x}, \mathbf{e}_{t+1} \rangle),$$

where $G \geq 0$ is a number chosen such that

$$G^2 = \frac{\sigma}{16(T-1)}\left(\sqrt{64L\Delta(T-1)+9\sigma^2} - 3\sigma\right),$$

and the functions h_t are defined as follows: First denote by

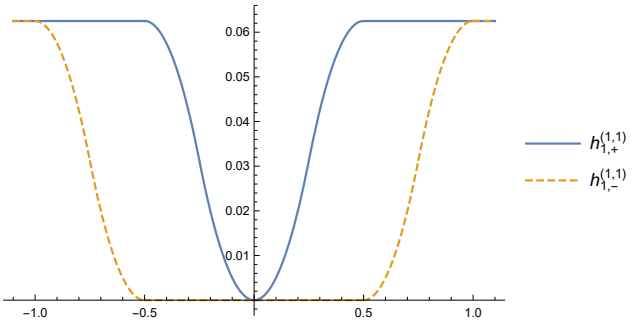

Figure 2. $h_{1,-}^{(1,1)}$ and $h_{1,+}^{(1,1)}$.

$h_{b,-}^{(1,L)}(x)$ and $h_{b,+}^{(1,L)}(x)$ the functions (see Fig. 2)

$$h_{b,+}^{(1,L)}(x) := \begin{cases} \frac{L}{2}x^2 & |x| \le b/4, \\ \frac{L}{16}b^2 - \frac{L}{2}(|x| - b/2)^2 & b/4 < |x| < b/2, \\ \frac{L}{16}b^2 & |x| \ge b/2, \end{cases}$$

$$h_{b,-}^{(1,L)}(x) := \begin{cases} 0 & |x| \le b/2, \\ \frac{L}{2}(|x| - \frac{b}{2})^2 & \frac{b}{2} \le |x| \le \frac{3b}{4}, \\ \frac{L}{16}b^2 - \frac{L}{2}(|x| - b)^2 & \frac{3b}{4} < |x| < b, \\ \frac{L}{16}b^2 & |x| \ge b, \end{cases}$$

then at indices t where the aggregation coefficient $|\zeta_{t+1}| \le \frac{1}{2}$, take h_t to be $h_t = h_{|\eta_t|\sigma,-}^{(1,L)}(x)$, and otherwise take $h_t = h_{|\eta_t|\sigma,+}^{(1,L)}$. Note that for all $t \in [T-1]$,

$$\begin{aligned} &h_t(0) = 0, \\ &h_t(x) = h_t(-x), \quad \forall x \in \mathbb{R}, \\ &h_t(\pm\eta_t\sigma) = \frac{L}{16}\eta_t^2\sigma^2, \\ &h_t'(0) = h_t'(\pm\eta_t\sigma) = h_t'(\pm\zeta_{t+1}\eta_t\sigma) = 0, \end{aligned}$$

and that h_t has L-Lipschitz gradient. The purpose of the functions h_t is to increase the value of $f(x_i)$ without affecting the gradient information available to the algorithm.

From the definition of f we conclude that f also shares the L-Lipschitz gradient of the functions h_t (being a separable sum of functions with L-Lipschitz gradient).

We now turn to analyze the dynamics of SGD when applied on the function f defined above. Given the objective f and the starting point

$$\mathbf{x}_1 = 0,$$

the algorithm at the first iteration sets

$$\mathbf{x}_2 = \mathbf{x}_1 - \eta_1(\nabla f(\mathbf{x}_1) + \boldsymbol{\xi}_1) = (-\eta_1 G, \pm\eta_1\sigma, 0, \ldots, 0)^\top,$$

hence, from the properties of h_1, we get

$$\begin{aligned} f(\mathbf{x}_2) &= -G^2\eta_1 + h_1(\eta_t\sigma), \\ \nabla f(\mathbf{x}_2) &= G\mathbf{e}_1 + h_1'(\pm\eta_1\sigma)\mathbf{e}_2 = G\mathbf{e}_1. \end{aligned}$$

Similarly, at the t-th iteration, $t \in [T-1]$, the algorithm sets

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{x}_t - \eta_t(\nabla f(\mathbf{x}_t) + \boldsymbol{\xi}_t)) \\ &= (-\sum_{k=1}^{t} \eta_k G, \pm\eta_1\sigma, \ldots, \pm\eta_t\sigma, 0, \ldots, 0)^\top, \end{aligned} \tag{3}$$

which leads to

$$f(\mathbf{x}_t) = -G^2 \sum_{k=1}^{t-1} \eta_k + \sum_{k=1}^{t-1} h_k(\eta_k\sigma), \tag{4}$$

$$\nabla f(\mathbf{x}_t) = G\mathbf{e}_1. \tag{5}$$

At the aggregation step, the algorithm sets

$$\begin{aligned} \mathbf{x}_{\text{out}} &= \sum_{t=1}^{T} \zeta_t \mathbf{x}_t \\ &= (-\sum_{t=1}^{T} \zeta_t \sum_{k=1}^{t-1} \eta_k G, \pm\zeta_2\eta_1\sigma, \ldots, \pm\zeta_T\eta_{T-1}\sigma)^\top, \end{aligned} \tag{6}$$

then by the properties of h_t, we get

$$f(\mathbf{x}_{\text{out}}) = -G^2 \sum_{t=1}^{T} \zeta_t \sum_{k=1}^{t-1} \eta_k + \sum_{k=1}^{t-1} h_k(\zeta_{k+1}\eta_k\sigma), \tag{7}$$

$$\nabla f(\mathbf{x}_{\text{out}}) = G\mathbf{e}_1, \tag{8}$$

where the first equality follows since h_t is even, and second equality follows from $h_t'(\pm\zeta_{t+1}\eta_t\sigma) = 0$.

To complete our treatment of the fixed-step case, we turn to show that it is possible to make f bounded from below without affecting the first-order information at the iterates and the gradient at $\mathbf{x}_{\text{out}}$. For this purpose, we continue to show that Lemma 1 can be applied when taking for $\mathbf{y}_1, \ldots, \mathbf{y}_m$ all the possible values the random variable $\mathbf{x}_{\text{out}}$ (6) can attain, and for $\mathbf{z}_1, \ldots, \mathbf{z}_n$ all possible values the random variables $\{\mathbf{x}_t\}_{t\in[T]}$ can attain (3).

Indeed, in view of (5) and (8), the first condition of Lemma 1 holds with $\boldsymbol{\gamma} = G\mathbf{e}_1$, the second requirement follows from (7), and the third requirement follows since

$$\langle \nabla f(\mathbf{x}_{\text{out}}), \mathbf{x}_{\text{out}} \rangle = -\sum_{t=1}^{T} \zeta_t \sum_{k=1}^{t-1} \eta_k G^2$$

does not depend on the sign of the noise vectors $\boldsymbol{\xi}_t$. As all the requirements of Lemma 1 hold, we conclude that there exists a function $\hat{f}$ that shares the the same first-order

information as f at $\{\mathbf{x}_t\}_{t\in[T]}$, the same gradient at $\mathbf{x}_{\text{out}}$ and in addition

$$\inf_{\mathbf{x}} \hat{f}(\mathbf{x}) \geq \min_{t\in[T]} f(\mathbf{x}_t) - \frac{3}{2L}G^2.$$

We get

$$\begin{aligned}
&\hat{f}(\mathbf{x}_1) - \inf_{\mathbf{x}} \hat{f}(\mathbf{x}) \\
&\leq 0 - \min_{t\in[T]} f(\mathbf{x}_t) + \frac{3}{2L}G^2 \\
&= \max_{t\in[T]} \sum_{k=1}^{t-1} \left(\eta_k G^2 - \frac{L}{16}\eta_k^2\sigma^2 + \frac{3}{2L(t-1)}G^2 \right) \\
&\leq (T-1)\frac{G^2}{2L}\left(\frac{8G^2}{\sigma^2} + \frac{3}{T-1}\right) = \Delta,
\end{aligned}$$

where the second inequality follows by maximizing the concave quadratic form over η_k and the last inequality follows from the definition of G^2 by basic algebra.

As SGD does not have access to the objective beyond the first-order information at the iterates, we conclude that the algorithm proceeds on $\hat{f}$ in exactly the same dynamics as it does on f, maintaining its behavior as derived above. □

The example provided by Thm. 2 comes with a guarantee that the gradient of the objective at all iterates is almost surely a constant; as a result, the theorem is applicable for forming lower bounds for all first-order optimality criteria used in the literature, including the best expected gradient norm $\min_t \mathbb{E}\|\nabla f(\mathbf{x}_t)\|$, average expected gradient norm $\frac{1}{T}\sum_t \mathbb{E}\|\nabla f(\mathbf{x}_t)\|$, and expected norm of the average gradient $\mathbb{E}\|\frac{1}{T}\sum_t \nabla f(\mathbf{x}_t)\|$, both when taking the actual gradient and when taking the noisy version of the gradient.

Note that although the theorem does not directly consider randomized sampling schemes for computing $\mathbf{x}_{\text{out}}$, the performance of any scheme that samples $\mathbf{x}_{\text{out}}$ out of $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ is bounded from below by the optimality criterion $\min_t \|\nabla f(\mathbf{x}_t)\|$, making the guarantees by the theorem applicable.

Remark 2 (Tightness results)**.** Consider the upper bound by Ghadimi and Lan (see Thm. 8 in the appendix) and set the step size by $\eta_t \equiv \eta := \sqrt{\frac{2\Delta}{(T-1)L\sigma^2}}$, where Δ is an upper bound on $f(\mathbf{x}_1) - f(\mathbf{x}_*)$. We obtain

$$\begin{aligned}
\min_{t\in[T]} \|\nabla f(\mathbf{x}_t)\|^2 \leq& \frac{2\Delta + L(T-1)\eta^2\sigma^2}{(T-1)\eta(2-L\eta)} \\
\underset{T\gg 1}{\approx}& \frac{2\Delta + L(T-1)\eta^2\sigma^2}{2(T-1)\eta} = \sigma\sqrt{\frac{2L\Delta}{T-1}},
\end{aligned}$$

which establishes on one hand, that the lower bound obtained in Thm. 2 on the iterates $\mathbf{x}_t$ is tight up to the constant factor $2\sqrt{2}$, and on the other hand, establishes that the constant step-size scheme defined above is optimal up to the same constant.

The second main result of this subsection gives an $\Omega(\epsilon^{-3})$ lower bound on the performance of "plain" SGD methods (i.e., methods that do not perform an aggregation step) acting on objectives with Lipchitz Hessians.

Theorem 3. *Consider a method that given a function $f : \mathbb{R}^d \to \mathbb{R}$ and an initial point $\mathbf{x}_1 \in \mathbb{R}^d$ generates a sequence of $T \in \mathbb{N}$ points $\{\mathbf{x}_t\}$ satisfying*

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \eta_{\mathbf{x}_1,\ldots,\mathbf{x}_t} \cdot (\nabla f(\mathbf{x}_t) + \boldsymbol{\xi}_t), \quad t \in [T-1],$$

where $\boldsymbol{\xi}_t$ are some random noise vectors. We further assume that the step sizes $\eta_{\mathbf{x}_1,\ldots,\mathbf{x}_t}$ are deterministic functions of the norms and inner products between $\mathbf{x}_1, \ldots, \mathbf{x}_t, \nabla f(\mathbf{x}_1) + \boldsymbol{\xi}_1, \ldots, \nabla f(\mathbf{x}_t) + \boldsymbol{\xi}_t$ and may also depend on the exact second-order information $\nabla^2 f(\mathbf{x}_1), \ldots, \nabla^2 f(\mathbf{x}_t)$. Then for any $\rho, \Delta, \sigma \in \mathbb{R}_{++}$ there exists a function $f : \mathbb{R}^T \mapsto \mathbb{R}$ with ρ-Lipschitz Hessian, $\mathbf{x}_1 \in \mathbb{R}^T$, and independent random variables $\boldsymbol{\xi}_t$ with $\mathbb{E}[\boldsymbol{\xi}_t] = 0$ and $\mathbb{E}[\|\boldsymbol{\xi}_t\|^2] = \sigma^2$ such that $\forall t \in [T]$

$$f(\mathbf{x}_1) - f(\mathbf{x}_t) \overset{a.s.}{\leq} \Delta, \qquad \|\nabla f(\mathbf{x}_t)\|^2 \overset{a.s.}{=} \boldsymbol{\gamma},$$

where $\boldsymbol{\gamma} \in \mathbb{R}^T$ is a vector that satisfies $\|\boldsymbol{\gamma}\|^2 = \frac{\sigma}{2}\left(\frac{\rho\Delta^2}{(T-1)^2}\right)^{1/3}$.

Proof. We proceed as in the proof of Thm. 2, taking for G the positive value that satisfies

$$G^2 = \frac{3\sigma}{32}\left(\frac{16^2\rho\Delta^2}{(T-1)^2}\right)^{1/3} \geq \frac{\sigma}{2}\left(\frac{\rho\Delta^2}{(T-1)^2}\right)^{1/3},$$

and set $h_t := h^{(2,\rho)}_{|\eta_t|\sigma}$, with $h^{(2,\rho)}_b$ defined by

$$h^{(2,\rho)}_b(x) := \begin{cases} \frac{\rho}{6}|x|^3 & |x| \leq b/4, \\ \frac{\rho}{2}\left(\frac{b^3}{96} - \frac{b^2}{8}|x| + \frac{b}{2}x^2 - \frac{1}{3}|x|^3\right) & \frac{b}{4} \leq |x| < \frac{3b}{4}, \\ \frac{\rho}{32}b^3 - \frac{\rho}{6}(b-|x|)^3 & \frac{3b}{4} \leq |x| < b, \\ \frac{\rho}{32}b^3 & |x| \geq b. \end{cases}$$

It is straightforward to verify that h_t has ρ-Lipschitz Hessian, and as in the Lipschitz gradient case, we have

$$\begin{aligned}
&h_t(0) = 0, \\
&h_t(x) = h_t(-x), \quad \forall x \in \mathbb{R}, \\
\text{and} \quad &h_t'(0) = h_t'(\eta_t\sigma) = h_t'(-\eta_t\sigma) = 0.
\end{aligned}$$

Proceeding with the new values, we reach

$$
\begin{aligned}
f(\mathbf{x}_1) - f(\mathbf{x}_t) &= \sum_{k=1}^{t-1} \left(G^2 \eta_k - h^{(2)}_{|\eta_k|\sigma}(|\eta_k|\sigma) \right) \\
&= \sum_{k=1}^{t-1} \left(\frac{3\sigma}{32} \left(\frac{16^2 \rho \Delta^2}{(T-1)^2} \right)^{1/3} \eta_k - \frac{\rho}{32} |\eta_k|^3 \sigma^3 \right) \\
&= \frac{1}{32} \sum_{k=1}^{t-1} \left(3 \left(\frac{16^2 \Delta^2}{(T-1)^2} \right)^{1/3} (\rho^{1/3} \sigma \eta_k) - |\rho^{1/3} \eta_k \sigma|^3 \right) \\
&\leq (t-1) \frac{\Delta}{T-1} \leq \Delta,
\end{aligned}
$$

where the one before last inequality follows from the inequality $3ax - |x|^3 \leq 2a^{\frac{3}{2}}$.

Finally, note that $h_t''(\eta_t \sigma) = h_t''(-\eta_t \sigma) = 0$, and as a result, the second-order information of f at all iterates is identically zero, thus the proof in the adaptive step-size case can proceed without change. □

Note that the main missing component needed for establishing a result bounding $f(\mathbf{x}_1) - \inf_{\mathbf{x}} f(\mathbf{x})$ in the Lipschitz-Hessian case is a set of necessary and sufficient interpolation conditions for Lipschitz-Hessian functions (as in the case of Lipschitz-gradient, Thm. 6 in the appendix). The existence of such conditions remains an open question.

4. Lower Bounds in the Convex Quadratic Case

In this section, we continue our analysis of the SGD method, showing that even for convex, quadratic functions in moderate dimensions and a standard Gaussian noise, SGD cannot achieve an iteration complexity better than $\mathcal{O}(\epsilon^{-4})$ in order for any of its iterates to have gradient norm less than ϵ. Note that for quadratic functions, the Hessian is constant, so the result still holds under additional Lipschitz assumptions on the Hessian and higher-order derivatives. We emphasize that the lower bounds only hold for the iterates themselves, without any aggregation step. Formally, we have the following:

Theorem 4. *Consider the SGD method defined by*

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \eta_t \cdot (\nabla f(\mathbf{x}_t) + \boldsymbol{\xi}_t), \quad t \in [T-1],$$

for some $T > 1$ and suppose that the step sizes $\eta_1, \ldots, \eta_{T-1}$ are non-negative and satisfy at least one of the following conditions:

1. (Small step sizes) $\max_{t \in [T-1]} \eta_t \leq 1/L$, *and* $\sum_{t=1}^{T-1} \eta_t \leq c\sqrt{T}/L$ *for some constant c (independent of the problem parameters).*
2. (Fixed step sizes) η_t *is the same for all t.*
3. (Polynomial decay schedule) $\eta_t = \frac{a}{b+t^\theta}$ *for some non-negative constants a, b, θ (independent of the problem parameters).*

Then for any $\delta \in (0,1)$, there exists a quadratic *function f on $\mathbb{R}^d$ (for any $d \geq d_0$ with $d_0 = \mathcal{O}(\log(T/\delta)\sigma^2 T/(L^2\Delta))$ with L-Lipschitz gradients, and $\mathbf{x}_1$ for which $f(\mathbf{x}_1) - \inf_{\mathbf{x}} f(\mathbf{x}) \leq \Delta$, such that if $\boldsymbol{\xi}_t$ has a Gaussian distribution $\mathcal{N}(\mathbf{0}, \frac{\sigma^2}{d} I_d)$, with probability at least $1-\delta$*

$$\min_{t \in [T-1]} \|\nabla f(\mathbf{x}_t)\|^2 \geq c_0 \frac{\min\{L\Delta, \sigma^2\}}{\sqrt{T}},$$

where c_0 is a positive constant depending only on the constants in the conditions stated above.

We note that all standard analyses for (non-adaptive) SGD methods rely on one of these step-size strategies. Moreover, the proof technique can plausibly be extended to other step sizes. Thus, the theorem provides a strong indication that SGD (without an aggregation step) cannot achieve a better iteration complexity, at least when the optimality criterion is $\min_t \|\nabla f(\mathbf{x}_t)\|$, even for convex quadratic functions.

The proof is based on the following two more technical propositions, which provide lower bounds depending on the step sizes and the problem parameters:

Proposition 1. *For any $L > 0, \Delta > 0, T > 1$ and $\delta \in (0,1)$, there exists a convex quadratic function f on $\mathbb{R}^d$ (for any $d \geq d_0$ where $d_0 = \mathcal{O}(\log(T/\delta)\sigma^2 T/(L^2\Delta))$) with L-Lipschitz gradient, and an $\mathbf{x}_1$ such that $f(\mathbf{x}_1) - \inf_{\mathbf{x}} f(\mathbf{x}) \leq \Delta$, such that if we initialize SGD at $\mathbf{x}_1$ with Gaussian noise $\mathcal{N}(\mathbf{0}, \frac{\sigma^2}{d} I_d)$ and use step sizes $\eta_1, \ldots, \eta_{T-1}$ in $[0, 1/L]$, then with probability at least $1-\delta$,*

$$\min_{t \in [T]} \|\nabla f(\mathbf{x}_t)\|^2 \geq \frac{\Delta}{25 \max\left\{1/L, \sum_{t=1}^{T-1} \eta_t\right\}}.$$

Proposition 2. *For any $L > 0, \Delta > 0, T > 1$ and $\delta \in (0,1)$, there exists a convex quadratic function f on $\mathbb{R}^d$ (for any $d \geq d_0$ where $d_0 = \mathcal{O}(\log(T/\delta))$) with L-Lipschitz gradient, and a vector $\mathbf{x}_1$ such that $f(\mathbf{x}_1) - \inf_{\mathbf{x}} f(\mathbf{x}) \leq \Delta$, such that if we initialize SGD at $\mathbf{x}_1$ with Gaussian noise $\mathcal{N}(\mathbf{0}, \frac{\sigma^2}{d} I_d)$, then the following holds with probability at least $1-\delta$:*

- *If for all t, $\eta_t = \eta$ with $\eta \in [0, 1/L)$, then $\min_{t \in [T]} \|\nabla f(\mathbf{x}_t)\|^2 \geq \frac{L}{2} \min\{\Delta, \frac{\eta\sigma^2}{2 - L\eta}\}$.*
- *If for all t, $\eta_t \geq c/L$ for some constant $c > 0$ then $\min_{t \in [T]} \|\nabla f(\mathbf{x}_t)\|^2 \geq \frac{\sigma^2 c^2}{2}$.*

- *If* $\eta_t = \frac{a}{L(b+t^\theta)}$ *for some positive constants* $a > 0, b \geq 0$ *and* $\theta \in (0,1)$*, then*

$$\min_{t\in[T]} \|\nabla f(\mathbf{x}_t)\|^2 \geq c_{a,b,\theta}\sigma^2 \min\{1, L\eta_T\},$$

where $c_{a,b,\theta}$ *is a constant dependent only on* a, b, θ*.*

The proofs of these propositions appear in Appendix A. Together, Propositions 1 and 2 imply the theorem:

Proof of Thm. 4. The theorem, under the first condition, is an immediate corollary of Proposition 1. Indeed,

$$\min_{t\in[T]} \|\nabla f(\mathbf{x}_t)\|^2 \geq \frac{\Delta}{25\max\left\{1/L, \sum_{t=1}^{T-1}\eta_t\right\}} \geq \frac{L\Delta}{25\max\left\{1, c\sqrt{T}\right\}} \geq \frac{\min\{L\Delta, \sigma^2\}}{25\max\{1,c\}\sqrt{T}}.$$

As to the second condition, let us consider three cases. First, if $\eta_t = \eta$ is at most $T^{-1/2}/L$, then $\sum_{t=1}^{T-1}\eta_t < \sqrt{T}/L$ and the result again follows from Proposition 1 as in the previous case. Next, suppose $T^{-1/2}/L \leq \eta < 1/L$, then the result follows from Proposition 2:

$$\min_{t\in[T]} \|\nabla f(\mathbf{x}_t)\|^2 \geq \frac{L}{2}\min\{\Delta, \frac{\eta\sigma^2}{2-L\eta}\} \geq \frac{1}{2}\min\{L\Delta, \frac{\sigma^2 T^{-1/2}}{2-T^{-1/2}}\} \geq \frac{1}{2}\frac{\min\{L\Delta,\sigma^2\}}{2\sqrt{T}-1}.$$

The last case for this condition is $1/L \leq \eta$, which does not converge due to the second case in Proposition 2.

As to the third condition (namely $\eta_t = \frac{a}{b+t^\theta}$), we can assume without loss of generality that $a > 0$ (otherwise we are back to the first condition in the theorem, and nothing is left to prove) and that $\theta \in (0,1)$ (since if $\theta = 0$, we are back to the second condition in the theorem and if $\theta \geq 1$ we are back to the first condition in the theorem). The result then follows from the third case of Proposition 2. □

5. Conclusion

In this paper, we studied the necessity of various assumptions that appear in classical and recent analyses of SGD. Firstly, we focused on the criteria used to measure the performance of the method, establishing that in order to obtain meaningful bounds in dimension-free nonconvex settings one must either utilize a criterion that incorporates multiple iterates or employ a (deterministic or randomized) aggregation scheme on the iterates. This is in contrast to similar results in the convex setting, where bounds on the last iterates are well-known. We then turned our attention to the assumptions required in order to break the $\mathcal{O}(\epsilon^{-4})$ iteration complexity barrier. We demonstrated that to improve upon this bound, additional assumptions are required either on the objective (beyond the standard Lipschitz-continuity of the gradient assumption) or on the noise (beyond independence and bounded variance), even when a complex aggregation of the iterates is allowed. Similarly, when the objective has Lipschitz-continuous Hessian, we showed that an $\mathcal{O}(\epsilon^{-3})$ bound cannot be improved without further assumptions. Finally, in the apparently "easy" convex quadratic case, we showed that when using "reasonable" step sizes, iterate aggregation schemes (or other modifications to SGD) are necessary for improving upon the $\mathcal{O}(\epsilon^{-4})$ bound.

References

Agarwal, A., Wainwright, M. J., Bartlett, P. L., and Ravikumar, P. K. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, pp. 1–9, 2009.

Allen-Zhu, Z. How to make the gradients small stochastically: Even faster convex and nonconvex SGD. In *Advances in Neural Information Processing Systems*, pp. 1157–1167, 2018.

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

Azagra, D. and Mudarra, C. An extension theorem for convex functions of class c1, 1 on hilbert spaces. *Journal of Mathematical Analysis and Applications*, 446(2):1167–1182, 2017.

Bertsekas, D. P. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*, 2010(1-38): 3, 2011.

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

Drori, Y. The exact information-based complexity of smooth convex minimization. *Journal of Complexity*, 39: 1–16, 2017.

Drori, Y. and Teboulle, M. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1-2):451–482, 2014. ISSN 0025-5610.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

Fang, C., Lin, Z., and Zhang, T. Sharp analysis for nonconvex SGD escaping from saddle points. *arXiv preprint arXiv:1902.00247*, 2019.

Fefferman, C., Israel, A., and Luli, G. K. Interpolation of data by smooth nonnegative functions. *Revista Matemática Iberoamericana*, 33(1):305–324, 2017.

Foster, D., Sekhari, A., Shamir, O., Srebro, N., Sridharan, K., and Woodworth, B. The complexity of making the gradient small in stochastic convex optimization. *arXiv preprint arXiv:1902.04686*, 2019.

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

Kiwiel, K. C. Convergence and efficiency of subgradient methods for quasiconvex minimization. *Mathematical programming*, 90(1):1–25, 2001.

Kushner, H. and Yin, G. G. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Moulines, E. and Bach, F. R. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pp. 451–459, 2011.

Nemirovski, A. and Yudin, D. Problem complexity and method efficiency in optimization. *Willey-Interscience, New York*, 1983.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4): 1574–1609, 2009.

Nesterov, Y. *Introductory lectures on convex optimization : a basic course*. Kluwer Academic Publ., 2004.

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

Nesterov, Y. E. Minimization methods for nonsmooth convex and quasiconvex functions. *Matekon*, 29:519–531, 1984.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1571–1578. Omnipress, 2012.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Simchowitz, M., Alaoui, A. E., and Recht, B. On the gap between strict-saddles and true convexity: An omega (log d) lower bound for eigenvector approximation. *arXiv preprint arXiv:1704.04548*, 2017.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization*, 27(3):1283–1313, 2017a.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1-2):307–345, 2017b.

Whitney, H. Analytic extensions of differentiable functions defined in closed sets. *Transactions of the American Mathematical Society*, 36(1):63–89, 1934.