Proceedings of Machine Learning Research vol 99:1–35, 2019 32nd Annual Conference on Learning Theory

Tight analyses for non-smooth stochastic gradient descent

Nicholas J. A. Harvey NICKHAR@CS.UBC.COM
Christopher Liaw CVLIAW@CS.UBC.CA
University of British Columbia, Department of Computer Science

Yaniv Plan YANIV@MATH.UBC.CA
University of British Columbia, Department of Mathematics

Sikander Randhawa SRAND.CS.UBC.CA
University of British Columbia, Department of Computer Science

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

Consider the problem of minimizing functions that are Lipschitz and strongly convex, but not necessarily differentiable. We prove that after T steps of stochastic gradient descent, the error of the final iterate is $O(\log(T)/T)$ *with high probability*. We also construct a function from this class for which the error of the final iterate of *deterministic* gradient descent is $\Omega(\log(T)/T)$. This shows that the upper bound is tight and that, in this setting, the last iterate of stochastic gradient descent has the same general error rate (with high probability) as deterministic gradient descent. This resolves both open questions posed by Shamir (2012).

An intermediate step of our analysis proves that the suffix averaging method achieves error $O(1/T)$ *with high probability*, which is optimal (for any first-order optimization method). This improves results of Rakhlin et al. (2012) and Hazan and Kale (2014), both of which achieved error $O(1/T)$, but only in expectation, and achieved a high probability error bound of $O(\log\log(T)/T)$, which is suboptimal.

Keywords: Gradient Descent, Lipschitz Functions, Strong Convexity, Martingales, High Probability Analysis, Lower Bounds.

1. Introduction

Stochastic gradient descent (SGD) is a popular first order method which dates back to 1951 (Robbins and Monro, 1951). It is a very simple and widely used iterative method for minimizing convex loss functions. In a nutshell, the method works by querying an oracle for a noisy estimate of a subgradient, then taking a small step in the opposite direction. The simplicity and effectiveness of this algorithm has established it as an essential tool for applied machine learning (Schmidt et al., 2017; Johnson and Zhang, 2013). See (Bubeck, 2015) or (Hazan, 2015) for more details about SGD.

The efficiency of SGD is usually measured by the rate of decrease of the *error* — the difference in value between the algorithm's output and the true minimum. The optimal error rate is known under various assumptions on f, the function to be minimized. In addition to convexity, common assumptions are that f is *smooth* (gradient is Lipschitz) or *strongly convex* (locally lower-bounded by a quadratic). Strongly convex functions often arise due to regularization, whereas smooth functions can sometimes be obtained by smoothening approximations (e.g., convolution). Existing analyses (Nemirovski et al., 2009) show that, after T steps of SGD, the expected error of the final iterate

is $O(1/\sqrt{T})$ for *smooth* functions, and $O(1/T)$ for functions that are both *smooth* and strongly convex; furthermore, both of these error rates are optimal without further assumptions.

The *non-smooth* setting is the focus of this paper. One example of this setting is with ℓ_1 regularized learning problems. As another example, the objective for ℓ_2^2 regularized support vector machines (Shalev-Shwartz et al., 2011) is strongly convex but not smooth.

A trouble with the non-smooth setting is that the error of (even deterministic) gradient descent need not decrease monotonically with T, so it is not obvious how to analyze the error of the final iterate. A workaround, known as early as Nemirovsky and Yudin (1983), is to output the *average* of the iterates. Existing analyses of SGD show that the *expected* error of the average is $\Theta(1/\sqrt{T})$ for Lipschitz functions (Nemirovsky and Yudin, 1983), which is optimal, whereas for functions that are also strongly convex (Hazan et al., 2007; Rakhlin et al., 2012) the average has error $\Theta(\log(T)/T)$ with high probability, which is not the optimal rate. An alternative algorithm, more complicated than SGD, was discovered by Hazan and Kale (2014); it achieves the optimal *expected* error rate of $O(1/T)$. *Suffix averaging*, a simpler approach in which the last *half* of the SGD iterates are averaged, was also shown to achieve *expected* error $O(1/T)$ (Rakhlin et al., 2012), although implementations can be tricky or memory intensive if the number of iterations T is unknown a priori. Non-uniform averaging schemes with optimal expected error rate and simple implementations are also known (Lacoste-Julien et al., 2012; Shamir and Zhang, 2013).

Shamir (2012) asked the very natural question of whether the *final* iterate of SGD achieves the optimal rate in the non-smooth scenario, as it does in the smooth scenario. If true, this would yield a very simple, implementable and interpretable form of SGD. Substantial progress on this question was made by Shamir and Zhang (2013), who showed that the final iterate has *expected* error $O(\log(T)/\sqrt{T})$ for Lipschitz f, and $O(\log(T)/T)$ for strongly convex f. Both of these bounds are a $\log(T)$ factor worse than the optimal rate, so Shamir and Zhang (2013) write

> An important open question is whether the $O(\log(T)/T)$ [*expected*] rate we obtained on [the last iterate], for strongly-convex problems, is tight. This question is important, because running SGD for T iterations, and returning the last iterate, is a very common heuristic. In fact, even for the simpler case of (non-stochastic) gradient descent, we do not know whether the behavior of the last iterate... is tight.

Nesterov and Shikhman (2015) take an alternative approach and study a similar question (only in the non strongly-convex case, however). They develop a different first-order algorithm for which the individual iterates converge in value at the *optimal rate* to the true minimum. Their result makes the question of Shamir (2012) even more interesting; we now know that such algorithms actually exist and it was conceivable that SGD is one of them. Earlier, Tang and Monteleoni (2013) gave positive results on Shamir's question for restricted classes of functions.

Our work shows that the $\log(T)$ factor is necessary for the standard step sizes, both for Lipschitz functions and for strongly convex functions, even for *non*-stochastic gradient descent. So both of the expected upper bounds due to Shamir and Zhang are actually tight. This resolves the first question of Shamir (2012). In fact, we show a much stronger statement: *any convex combination* of the last k iterates must incur a $\log(T/k)$ factor. Thus, suffix averaging must average a constant fraction of the iterates to achieve the optimal rate.

Recently, Jain et al. (2019) consider the setting where the time horizon, T, is *fixed ahead of time*. They show that in both the strongly-convex case and the Lipschitz case, a suitable choice of step size gives the final iterate the optimal convergence rates of $O(1/T)$ and $O(1/\sqrt{T})$, respectively. On

the other hand, for the strongly-convex and *stochastic* case, when T is unknown, they show that no choice of step size gives the individual iterates of SGD the $O(1/T)$ rate for every T.

High probability bounds on SGD are somewhat scarce; most of the literature proves bounds in expectation, which is of course easier. The inefficiency[1] of selecting the best of many independent trials of SGD makes the existence of high probability bounds for a single execution of SGD more interesting and useful. Some known high-probability bounds for the strongly convex setting include (Kakade and Tewari, 2008), for uniform averaging, and (Hazan and Kale, 2014; Rakhlin et al., 2012), which give a suboptimal bound of $O(\log\log(T)/T)$ for suffix averaging (and a variant thereof). In this work, we give two high probability bounds on the error of SGD: $O(1/T)$ for suffix averaging and $O(\log(T)/T)$ for the final iterate. Both of these are tight. (Interestingly, the former is used as an ingredient for the latter.) The former answers a question of Rakhlin et al. (2012, §6), and the latter resolves the second question of Shamir (2012).

2. Preliminaries

Let $\mathcal{X}$ be a closed, convex subset of $\mathbb{R}^n$, $f\colon \mathcal{X} \to \mathbb{R}$ be a convex function, and $\partial f(x)$ the subdifferential of f at x. Our goal is to solve the convex program $\min_{x\in\mathcal{X}} f(x)$. We assume that f is not explicitly represented. Instead, the algorithm is allowed to query f via a stochastic gradient oracle, i.e., if the oracle is queried at x then it returns $\hat{g} = g - \hat{z}$ where $g \in \partial f(x)$ and $\mathrm{E}[\,\hat{z}\,] = 0$ conditioned on all past calls to the oracle. The set $\mathcal{X}$ is represented by a projection oracle, which returns the point in $\mathcal{X}$ closest in Euclidean norm to a given point x. We say that f is α-strongly convex if

$$f(y) \;\geq\; f(x) + \langle\, g,\, y - x \,\rangle + \frac{\alpha}{2}\,\|y - x\|^2 \quad \forall y, x \in \mathcal{X}, g \in \partial f(x). \tag{1}$$

Throughout this paper, $\|\cdot\|$ denotes the *Euclidean* norm in $\mathbb{R}^n$ and $[T]$ denotes the set $\{1, \ldots, T\}$.

We say that f is L-Lipschitz if $\|g\| \leq L$ for all $x \in \mathcal{X}$ and $g \in \partial f(x)$. For the remainder of this paper, unless otherwise stated, we make the assumption that $\alpha = 1$ and $L = 1$; this is only a normalization assumption and is without loss of generality (see Appendix F). For the sake of simplicity, we also assume that $\|\hat{z}\| \leq 1$ a.s. although our arguments generalize to the setting when $\hat{z}$ are sub-Gaussian (see Appendix F or Harvey et al. (2018)).

Let $\Pi_{\mathcal{X}}$ denote the projection operator on $\mathcal{X}$. The (projected) stochastic gradient algorithm is given in Algorithm 1. Notice that there the algorithm maintains a sequence of points and there are several strategies to output a single point. The simplest strategy is to simply output x_{T+1}. However, one can also consider averaging all the iterates (Polyak and Juditsky, 1992; Ruppert, 1988) or averaging only a fraction of the final iterates (Rakhlin et al., 2012). Notice that the algorithm also requires the user to specify a sequence of step sizes. The optimal choice of step size is known to be $\eta_t = \Theta(1/t)$ for strongly convex functions (Nemirovski et al., 2009; Rakhlin et al., 2012). For our analyses, we will use a step size of $\eta_t = 1/t$.

1. It is usually the case that selecting the best of many independent trials is very inefficient. Such a scenario, which is very common in uses of SGD, arises if f is defined as $\sum_{i=1}^m f_i$ or $\mathrm{E}_\omega[\,f_\omega\,]$. In such scenarios, evaluating f exactly could be inefficient, and even estimating it to within error $1/T$ requires $\Theta(T^2)$ samples via a Hoeffding bound, whereas SGD uses only $O(T)$ samples.

Algorithm 1 Projected stochastic gradient descent for minimizing a non-smooth, convex function.

procedure STOCHASTICGRADIENTDESCENT($\mathcal{X} \subseteq \mathbb{R}^n$, $x_1 \in \mathcal{X}$, step sizes $\eta_1, \eta_2, \ldots$)
  for $t \leftarrow 1, \ldots, T$ **do**
    Query stochastic gradient oracle at x_t for $\hat{g}_t$ such that $\mathrm{E}[\hat{g}_t \mid \hat{g}_1, \ldots, \hat{g}_{t-1}] \in \partial f(x_t)$
    $y_{t+1} \leftarrow x_t - \eta_t \hat{g}_t$ (take a step in the opposite direction of the subgradient)
    $x_{t+1} \leftarrow \Pi_{\mathcal{X}}(y_{t+1})$ (project y_{t+1} onto the set $\mathcal{X}$)
  return either $\begin{cases} x_{T+1} & \text{(final iterate)} \\ \frac{1}{T+1}\sum_{t=1}^{T+1} x_t & \text{(uniform averaging)} \\ \frac{1}{T/2+1}\sum_{t=T/2+1}^{T+1} x_t & \text{(suffix averaging)} \end{cases}$

3. Our Contributions

Our main results are bounds on the error of the final iterate of SGD: an $\Omega(\log(T)/T)$ lower bound (even in the non-stochastic case) and a $O(\log(T)\log(1/\delta)/T)$ upper bound with probability $1-\delta$. These results resolve both open questions of Shamir (2012).

Theorem 3.1 *Suppose f is 1-strongly convex and 1-Lipschitz. Suppose that $\hat{z}_t$ (i.e., $\mathrm{E}[\hat{g}_t] - \hat{g}_t$, the noise of the stochastic gradient oracle) has norm at most 1 almost surely. Consider running Algorithm 1 for T iterations with step size $\eta_t = 1/t$. Let $x^* = \operatorname{argmin}_{x\in\mathcal{X}} f(x)$. Then, with probability at least $1-\delta$,*

$$f(x_{T+1}) - f(x^*) \;\le\; O\left(\frac{\log(T)\log(1/\delta)}{T}\right).$$

The assumptions on the strong convexity parameter, Lipschitz parameter, and diameter are without loss of generality; see Appendix F. The bounded noise assumption for the stochastic gradient oracle is made only for simplicity; our analysis can be made to go through if one relaxes the a.s. bounded condition to a sub-Gaussian condition (Harvey et al., 2018). We also remark that a linear dependence on $\log(1/\delta)$ is necessary for strongly convex functions; see Appendix G.

Our main probabilistic tool to prove Theorem 3.1 is a new extension of the classic Freedman inequality (Freedman, 1975) to a setting in which the martingale exhibits a curious phenomenon. Ordinarily a martingale is roughly bounded by the square root of its total conditional variance (this is the content of Freedman's inequality). We consider a setting in which the total conditional variance[2] is itself bounded by (a linear transformation of) the martingale. We refer to this as a "chicken and egg" phenomenon.

Theorem 3.2 (Generalized Freedman) *Let $\{d_i, \mathcal{F}_i\}_{i=1}^n$ be a martingale difference sequence. Suppose $v_{i-1} \ge 0$, $i \in [n]$ are $\mathcal{F}_{i-1}$-measurable random variables such that $\mathrm{E}[\exp(\lambda d_i) \mid \mathcal{F}_{i-1}] \le \exp\left(\frac{\lambda^2}{2} v_{i-1}\right)$ for all $i \in [n]$, $\lambda > 0$. Let $S_t = \sum_{i=1}^t d_i$ and $V_t = \sum_{i=1}^t v_{i-1}$. Let $\alpha_i \ge 0$ and set*

2. As stated, Theorem 3.2 assumes a conditional sub-Gaussian bound on the martingale difference sequence, whereas Freedman assumes both a conditional variance bound and an almost-sure bound. These assumptions are easily interchangeable in both our proof and Freedman's proof. For example, Freedman's inequality with the sub-Gaussian assumption appears in (Fan et al., 2015, Theorem 2.6).

$\alpha = \max_{i\in[n]} \alpha_i$. *Then*

$$\Pr\left[\bigcup_{t=1}^{n}\left\{S_t \geq x \text{ and } V_t \leq \sum_{i=1}^{t}\alpha_i d_i + \beta\right\}\right] \leq \exp\left(-\frac{x}{4\alpha + 8\beta/x}\right) \qquad \forall x, \beta > 0.$$

The proof of Theorem 3.2 appears in Appendix C. Freedman's Inequality (Freedman, 1975) (as formulated in (Fan et al., 2015, Theorem 2.6), up to constants) simply omits the terms highlighted in yellow, i.e., it sets $\alpha = 0$.

Next we give a matching lower bound on the last iterate's error in deterministic gradient descent.

Theorem 3.3 *For any T and any constant $c > 0$, there exists a convex function $f_T : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is the unit Euclidean ball in $\mathbb{R}^T$, such that f_T is $(3/c)$-Lipschitz and $(1/c)$-strongly convex, and satisfies the following. Suppose that Algorithm 1 is executed from the initial point $x_1 = 0$ with step sizes $\eta_t = c/t$. Let $x^* = \operatorname{argmin}_{x\in\mathcal{X}} f_T(x)$. Then*

$$f_T(x_T) - f_T(x^*) \geq \frac{\log T}{4c \cdot T}. \tag{2}$$

More generally, any weighted average $\bar{x}$ of the last k iterates has

$$f_T(\bar{x}) - f_T(x^*) \geq \frac{\ln(T) - \ln(k)}{4c \cdot T}. \tag{3}$$

So suffix averaging must average a constant fraction of iterates to achieve the optimal $O(1/T)$ error.

Remark 3.4 *In order to incur a $\log T$ factor in the error of the T-th iterate, Theorem 3.3 constructs a function f_T parameterized by T. It is also possible to create a single function f,* independent *of T, which incurs the $\log T$ factor for infinitely many T. This is described in Remark B.2. The details can be found in the full version of the paper (see Harvey et al. (2018)).*

Interestingly, our proof of Theorem 3.1 requires understanding the suffix average. (In fact this connection is implicit in Shamir and Zhang (2013)). Hence, en route, we prove the following high probability bound on the error of the average of the last half of the iterates of SGD.

Theorem 3.5 *Suppose f is 1-strongly convex and 1-Lipschitz. Consider running Algorithm 1 for T iterations with step size $\eta_t = 1/t$. Let $x^* = \operatorname{argmin}_{x\in\mathcal{X}} f(x)$. With probability at least $1 - \delta$,*

$$f\left(\frac{1}{T/2+1}\sum_{t=T/2}^{T} x_t\right) - f(x^*) \leq O\left(\frac{\log(1/\delta)}{T}\right).$$

Remark 3.6 *This upper bound is optimal. Indeed, Appendix G shows that the error is $\Omega(\log(1/\delta)/T)$ even for the one-dimensional function $f(x) = x^2/2$.*

Theorem 3.5 is improves the $O\big(\log(\log(T)/\delta)/T\big)$ bounds independently proven by Rakhlin et al. (2012) (for suffix averaging) and Hazan and Kale (2014) (for EpochGD). Once again, we defer the statement of the theorem for general strongly-convex and Lipschitz parameters to Appendix F.

Remark 3.7 (The Lipschitz case) *Theorem 3.1 and Theorem 3.3 also have analogues in case of functions that are Lipschitz but not strongly convex: $f(x_{T+1}) - f(x^*) = O(\log(T)\log(1/\delta)/\sqrt{T})$*

with probability at least $1-\delta$, and there exists f_T with $f_T(x_{T+1}) - f_T(x^*) = \Omega(\log(T)/\sqrt{T})$. The formal statement and the proof appear in the full version of the paper (Harvey et al., 2018). Note that suffix averaging is less interesting in the Lipschitz setting because uniform averaging is already optimal. Furthermore, the high probability bound for uniform averaging follows via a standard application of Azuma's inequality.

4. Techniques

Final iterate. When analyzing gradient descent, it simplifies matters greatly to consider the *expected* error. This is because the effect of a gradient step is usually bounded by the subgradient inequality; so by linearity of expectation, one can plug in the *expected* subgradient, thus eliminating the noise (Bubeck, 2015, §6.1).

High probability bounds are more difficult. (Indeed, it is not a priori obvious that the error of the final iterate is tightly concentrated.) A high probability analysis must somehow control the total noise that accumulates from each noisy subgradient step. Fortunately, the accumulated noise forms a zero-mean martingale but unfortunately, the martingale depends on previous iterates in a highly nontrivial manner. Indeed, suppose (X_t) is the martingale of the accumulated noise and let $V_{t-1} = \mathrm{E}\left[(X_t - X_{t-1})^2 \mid X_1, \dots, X_{t-1} \right]$ be the conditional variance at time t. A significant technical step of our analysis (Lemma 6.4) shows that the total conditional variance (TCV) of the accumulated noise exhibits the "chicken and egg" phenomenon alluded to in the discussion of Theorem 3.2. Roughly speaking, we have $\sum_{t=1}^{T} V_{t-1} \le \alpha X_{T-1} + \beta$ where $\alpha, \beta > 0$ are scalars. Since Freedman's inequality shows that $X_T \lesssim \sqrt{\sum_{t=1}^{T} V_T}$, an inductive argument gives that $X_T \lesssim \sqrt{\alpha X_{T-1} + \beta} \lesssim \sqrt{\alpha\sqrt{\alpha X_{T-2} + \beta} + \beta} \lesssim \cdots$. This naive analysis involves invoking Freedman's inequality T times, so a union bound incurs an extra factor $\log T$ in the bound on X_T. This can be improved via a trick (Bartlett et al., 2008): by upper-bounding the TCV by a power-of-two (and by T), it suffices to invoke Freedman's inequality $\log T$ times, which only incurs an extra factor $\log\log T$ in the bound on X_T.

Notice that this analysis actually shows that $X_t \lesssim \sqrt{\sum_{i=1}^{t} V_i}$ *for all* $t \le T$, whereas the original goal was only to control X_T. Any analysis that simultaneously controls all X_t, $t \le T$, must necessarily incur an extra factor $\log\log T$. This is a consequence of the Law of the Iterated Logarithm[3]. Previous work employs exactly such an analysis (Hazan and Kale, 2014; Kakade and Tewari, 2008; Rakhlin et al., 2012) and incurs the $\log\log T$ factor. Rakhlin et al. (2012) explicitly raise the question of whether this $\log\log T$ factor is necessary.

Our work circumvents this issue by developing a generalization of Freedman's Inequality (Theorem 3.2) to handle martingales of the above form, which ultimately yields optimal high-probability bounds. We are no longer hindered by the Law of the Iterated Logarithm because our variant of Freedman's Inequality does not require fine grained control over the martingale over all times.

Another important tool that we employ is a new bound on the Euclidean distance between the iterates computed by SGD (Lemma 6.3). This is useful because, by the subgradient inequality, the change in the error at different iterations can be bounded using the distance between iterates. Various naive approaches yield a bound of the form $\|x_a - x_b\|^2 \le \frac{(b-a)^2}{\min\{a^2, b^2\}}$. We derive a much

3. Let $X_t \in \{-1, +1\}$ be uniform and i.i.d. and $S_T = \sum_{t=1}^{T} X_t$. The Law of the Iterated Logarithm states that $\limsup_T \frac{S_T}{\sqrt{2T\log\log T}} = 1$ a.s.

stronger bound, comparable to $\|x_a - x_b\|^2 \leq \frac{|b-a|}{\min\{a^2,b^2\}}$. Naturally, in the stochastic case, there are additional noise terms that contribute to the technical challenge of our analysis. Nevertheless, this new distance bound could be useful in further understanding non-smooth gradient descent (even in the non-stochastic setting).

As in previous work on the strongly convex case (Shamir and Zhang, 2013), the error of the suffix average plays a critical role in bounding the error of the final iterate. Therefore, we also need a tight high probability bound on the error of the suffix average.

Suffix averaging. To complete the optimal high probability analysis on the final iterate, we need a high probability bound on the suffix average that avoids the $\log\log T$ factor. As in the final iterate setting, the accumulated noise for the suffix average forms a zero-mean martingale, $(X_t)_{T/2}^T$, but now the conditional variance at step t satisfies $V_t \leq \alpha_t V_{t-1} + \beta_t \hat{w}_t \sqrt{V_{t-1}} + \gamma_t$, where $\hat{w}_t$ is a mean-zero random variable and α_t, β_t and γ_t are constants. In Rakhlin et al. (2012), using Freedman's Inequality combined with the trick from Bartlett et al. (2008), they obtain a bound on a similar martingale but do so over all time steps and incur a $\log\log T$ factor. However, our goal is only to bound X_T and according to Freedman's Inequality $X_T \lesssim \sqrt{\sum_{t=T/2}^T V_t}$. So, we aim to bound $\sum_{t=T/2}^T V_t$. To do so, we develop a probabilistic tool to bound the T-th iterate of a stochastic process that satisfies a recursive dependence on the $(t-1)$-th iterate similar to the one exhibited by V_t.

Theorem 4.1 *Let $(X_t)_{t=1}^T$ be a stochastic process and let $(\mathcal{F}_t)_{t=1}^T$ be a filtration such that X_t is $\mathcal{F}_t$ measurable and X_t is non-negative almost surely. Assume that $\mathrm{E}[\exp(\lambda X_1)] \leq \exp(\lambda K)$ with probability 1, for $\lambda \in (0, 1/K]$. Let $\alpha_t \in [0,1)$ and $\beta_t, \gamma_t \geq 0$ for every t. Let $\hat{w}_t$ be a mean-zero random variable conditioned on $\mathcal{F}_t$ such that $|\hat{w}_t| \leq 1$ almost surely for every t. Suppose that $X_{t+1} \leq \alpha_t X_t + \beta_t \hat{w}_t \sqrt{X_t} + \gamma_t$ for every t. Then, the following hold.*

- *For every t, $\Pr[X_t \geq K \log(1/\delta)] \leq \mathrm{e}\delta$.*
- *More generally, if $\sigma_1, \ldots, \sigma_T \geq 0$, then $\Pr\left[\sum_{t=1}^T \sigma_t X_t \geq K \log(1/\delta) \sum_{t=1}^T \sigma_t\right] \leq \mathrm{e}\delta$,*

where $K = \max_{1\leq t\leq T}\left(\frac{2\gamma_t}{1-\alpha_t}, \frac{2\beta_t^2}{1-\alpha_t}\right)$.

The recursion $X_{t+1} \leq \alpha_t + \beta_t \hat{w}_t \sqrt{X_t} + \gamma_t$ presents two challenges that make it difficult to analyze. Firstly, the fact that it is a non-linear recurrence makes it unclear how one should unwind X_{t+1}. Furthermore, unraveling the recurrence introduces many $\hat{w}_t$ terms in a non-trivial way. Interestingly, if we instead consider the moment generating function (MGF) of X_{t+1}, then we can derive an analogous recursive MGF relationship which removes this non-linear dependence and removes the $\hat{w}_t$ term. This greatly simplifies the recursion and leads to a surprisingly clean analysis. The proof of Theorem 4.1 can be found in Appendix D. (The recursive MGF bound which removes the non-linear dependence is by Claim D.1.)

Deterministic lower bound. As mentioned above, a challenge with non-smooth gradient descent is that the error of the T-th iterate may not monotonically decrease with T, even in the deterministic setting. The full extent of this non-decreasing behavior seems not to have been previously understood. We develop a technique that forces the error to be monotonically *increasing* for $\Omega(T)$ consecutive iterations. The idea is as follows. If GD takes a step in a certain direction, a non-differentiable point can allow the function to suddenly increase in that direction. If the function were one-dimensional, the next iteration of GD would then be guaranteed to step in the opposite direction,

thereby decreasing the function. However, in higher dimensions, the second gradient step could be nearly orthogonal to the first step, and the function could have yet another non-differentiable point in this second direction. In sufficiently high dimensions, this behavior can be repeated for many iterations. The tricky aspect is designing the function to have this behavior while also being convex. We show that this is possible, leading to the unexpectedly large $\Omega(\log(T)/T)$ error in the T-th iteration. We believe that this example illuminates some non-obvious behavior of gradient descent.

5. Lower bound on error of final iterate

In this section we prove that the final iterate of SGD has error that is suboptimal by a factor $\Omega(\log T)$, even in the non-stochastic case. Specifically, we define a function $f = f_T$, depending on T, for which the final iterate produced by Algorithm 1 has $f(x_T) = \Omega(\log(T)/T)$, thereby proving (2). We give the proof of Theorem 3.3 in the case where $c = 1$. Theorem 3.3 can be obtained in full generality from the analysis in this section by replacing f with $\frac{1}{c}f$ and using the step-sizes $\eta_t = \frac{c}{t}$.

Let $\mathcal{X}$ be the Euclidean unit ball in $\mathbb{R}^T$. lDefine $f : \mathcal{X} \to \mathbb{R}$ and $h_i \in \mathbb{R}^T$ for $i \in [T+1]$ by

$$f(x) = \max_{i \in [T+1]} H_i(x) \qquad \text{where} \qquad H_i(x) = h_i^\intercal x + \frac{1}{2}\|x\|^2$$

$$h_{i,j} = \begin{cases} a_j & (\text{if } 1 \le j < i) \\ -1 & (\text{if } i = j \le T) \\ 0 & (\text{if } i < j \le T) \end{cases} \qquad \text{and} \qquad a_j = \frac{1}{2(T+1-j)} \qquad (\text{for } j \in [T]).$$

It is easy to see that f is 1-strongly convex due to the $\frac{1}{2}\|x\|^2$ term. Furthermore f is 3-Lipschitz over $\mathcal{X}$ because $\|\nabla H_i(x)\| \le \|h_i\| + 1$ and $\|h_i\|^2 \le 1 + \frac{1}{4}\sum_{j=1}^{T} \frac{1}{(T-j)^2} < 1 + \frac{1}{2}$. Finally, the minimum value of f over $\mathcal{X}$ is non-positive because $f(0) = 0$.

Subgradient oracle. In order to execute Algorithm 1 on f we must specify a subgradient oracle. First, we require the following claim, which follows from standard facts in convex analysis (Hiriart-Urruty and Lemaréchal, 1996, Theorem 4.4.2).

Claim 5.1 *$\partial f(x)$ is the convex hull of $\{\ h_i + x\ :\ i \in \mathcal{I}(x)\ \}$, where $\mathcal{I}(x) = \{\ i\ :\ H_i(x) = f(x)\ \}$.*

Our subgradient oracle is non-stochastic: given x, it simply returns $h_{i'} + x$ where $i' = \min \mathcal{I}(x)$.

Explicit description of iterates. Next we will explicitly describe the iterates produced by executing Algorithm 1 on f. Define the points $z_t \in \mathbb{R}^T$ for $t \in [T+1]$ by $z_1 = 0$ and

$$z_{t,j} = \begin{cases} \dfrac{1 - (t-j-1)a_j}{t-1} & (\text{if } 1 \le j < t) \\ 0 & (\text{if } t \le j \le T). \end{cases} \qquad (\text{for } t > 1).$$

We will show inductively that these are precisely the first T iterates produced by Algorithm 1 when using the subgradient oracle defined above. The next claim follows easily from the definition of z_t.

Claim 5.2

- *$z_{t,j} \ge \frac{1}{2(t-1)}$ for $j < t$ and $z_{t,j} = 0$ for $j \ge t$.*
- *$\|z_1\| = 0$ and $\|z_t\|^2 \le \frac{1}{t-1}$ for $t > 1$. Thus $z_t \in \mathcal{X}$ for all $t \in [T+1]$.*

The "triangular shape" of the h_i vectors allows us to determine the value and subdifferential at z_t.

Claim 5.3 *$f(z_t) = H_t(z_t)$ for all $t \in [T+1]$. The subgradient oracle for f at z_t returns $h_t + z_t$.*

Proof We claim that $h_t^\mathsf{T} z_t = h_i^\mathsf{T} z_t$ for all $i > t$. By definition, z_t is supported on its first $t-1$ coordinates. However, h_t and h_i agree on the first $t-1$ coordinates (for $i > t$). This proves the first part of the claim.

Next we claim that $z_t^\mathsf{T} h_t > z_t^\mathsf{T} h_i$ for all $1 \le i < t$. By the definition of z_t and h_i:

$$z_t^\mathsf{T}(h_t - h_i) \;=\; \sum_{j=1}^{t-1} z_{t,j}(h_{t,j} - h_{i,j}) \;=\; \sum_{j=i}^{t-1} z_{t,j}(h_{t,j} - h_{i,j}) \;=\; z_{t,i}(a_i + 1) + \sum_{j=i+1}^{t-1} z_{t,j} a_j \;>\; 0.$$

These two claims imply that $H_t(z_t) \ge H_i(z_t)$ for all $i \in [T+1]$, and therefore $f(z_t) = H_t(z_t)$. Moreover $\mathcal{I}(z_t) = \{\, i \,:\, H_i(z_t) = f(z_t) \,\} = \{t, \dots, T+1\}$. Thus, when evaluating the subgradient oracle at the vector z_t, it returns the vector $h_t + z_t$. ■

Since the subgradient returned at z_t is determined by Claim 5.3, and the next iterate of SGD arises from a step in the opposite direction, a straightforward induction proof allows us to show the following lemma. A detailed proof is in Appendix B.

Lemma 5.4 *The vector x_t in Algorithm 1 equals z_t, for every $t \in [T+1]$.*

The value of the final iterate is easy to determine from Lemma 5.4 and Claim 5.3:

$$f(x_{T+1}) \;=\; f(z_{T+1}) \;=\; H_{T+1}(z_{T+1}) \;\ge\; \sum_{j=1}^{T} h_{T+1,j} \cdot z_{T+1,j} \;\ge\; \sum_{j=1}^{T} \frac{1}{2(T+1-j)} \cdot \frac{1}{2T} \;>\; \frac{\log T}{4T}.$$

(Here the second inequality uses Claim 5.2.) This proves (2). A small modification of the last calculation proves (3); details may be found in Claim B.1. This completes the proof of Theorem 3.3.

6. Upper bound on error of final iterate

We now turn to the proof of the upper bound on the error of the final iterate of SGD, in the case where f is 1-strongly convex and 1-Lipschitz (Theorem 3.1). Recall that the step size used by Algorithm 1 in this case is $\eta_t = 1/t$. We will write $\hat{g}_t = g_t - \hat{z}_t$, where $\hat{g}_t$ is the vector returned by the oracle at the point x_t, $g_t \in \partial f(x_t)$, and $\hat{z}_t$ is the noise. Let $\mathcal{F}_t = \sigma(\hat{z}_1, \dots, \hat{z}_t)$ be the σ-algebra generated by the first t steps of SGD. Finally, recall that $\|\hat{z}_t\| \le 1$ and $\mathrm{E}\,[\,\hat{z}_t \mid \mathcal{F}_{t-1}\,] = 0$.

We begin with the following lemma which can be inferred from the proof of Theorem 1 in Shamir and Zhang (2013). For completeness, we provide a proof in Appendix E.

Lemma 6.1 *Let f be 1-strongly convex and 1-Lipschitz. Suppose that we run SGD (Algorithm 1) with step sizes $\eta_t = 1/t$. Then*

$$f(x_T) \;\le\; \underbrace{\frac{1}{T/2+1} \sum_{t=T/2}^{T} f(x_t)}_{\textit{suffix average}} \;+\; \underbrace{\sum_{k=1}^{T/2} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} \langle\, \hat{z}_t,\, x_t - x_{T-k} \,\rangle}_{Z_T,\ \textit{the noise term}} \;+\; O\left(\frac{\log T}{T}\right).$$

Lemma 6.1 asserts that the error of the last iterate is upper bounded by the sum of the error of the suffix average and some noise terms (up to the additive $O(\log T/T)$ term). Thus, it remains to show that the error due to the suffix average is small with high probability (Theorem 3.5) and the noise

terms are small. We defer the proof of Theorem 3.5 to Subsection E.1. By changing the order of summation, we can write $Z_T = \sum_{t=T/2}^{T} \langle\, \hat{z}_t,\, w_t \,\rangle$ where

$$w_t = \sum_{j=T/2}^{t} \alpha_j (x_t - x_j) \qquad \text{and} \qquad \alpha_j = \frac{1}{(T-j)(T-j+1)}.$$

The main technical difficulty is to show that Z_T is small with high probability. Formally, we prove the following lemma, whose proof is outlined in Subsection 6.1.

Lemma 6.2 *$Z_T \le O\left(\frac{\log(T)\log(1/\delta)}{T}\right)$ with probability at least $1-\delta$.*

Given Theorem 3.5 and Lemma 6.2, the proof of Theorem 3.1 is immediate.

6.1. Bounding the noise

The main technical difficulty in the proof is to understand the noise term, which is denoted Z_T. Notice that Z_T is a sum of a martingale difference sequence. The natural starting point is to better understand the TCV of Z_T, which is at most $\sum_{t=T/2}^{T} \|w_t\|^2$. We we will see that this expression is bounded by a linear transformation of Z_T. This "chicken and egg" relationship inspires a new probabilistic tool (generalized Freedman's Inequality) which disentangles the TCV from the martingale.

The main challenge in analyzing $\|w_t\|$ is precisely analyzing the distance $\|x_t - x_j\|$ between SGD iterates. A loose bound of $\|x_t - x_j\|^2 \lesssim (t-j)\sum_{i=j}^{t} \frac{\|\hat{g}_i\|^2}{i^2}$ follows easily from Jensen's inequality. In Appendix E, we prove the following tighter bound, potentially of independent interest.

Lemma 6.3 *Suppose f is 1-Lipschitz and 1-strongly convex. Suppose we run Algorithm 1 for T iterations with step sizes $\eta_t = 1/t$. Let $a < b$. Then,*

$$\|x_a - x_b\|^2 \le \sum_{i=a}^{b-1} \frac{\|\hat{g}_i\|^2}{i^2} + 2\sum_{i=a}^{b-1} \frac{\big(f(x_a) - f(x_i)\big)}{i} + 2\sum_{i=a}^{b-1} \frac{\langle\, \hat{z}_i,\, x_i - x_a \,\rangle}{i}.$$

Using Lemma 6.3 and some delicate calculations we obtain the following upper bound on $\sum_{t=T/2}^{T} \|w_t\|^2$, revealing the surprisingly intricate relationship between Z_T (the martingale) and $\sum_{t=T/2}^{T} \|w_t\|^2$ (its TCV). This is the main technical step that inspired our probabilistic tool (the generalized Freedman's Inequality).

Lemma 6.4 (Main Technical Lemma) *There exists positive values $R_1 = O\left(\frac{\log^2 T}{T^2}\right)$, $R_2 = O\left(\frac{\log T}{T}\right)$, and $C_t = O(\log T)$, $A_t = O\left(\frac{\log T}{T^2}\right)$ for all t such that*

$$\sum_{t=T/2}^{T} \|w_t\|^2 \le R_1 + R_2 \left\|x_{T/2} - x^*\right\|^2 + \underbrace{\sum_{t=T/2}^{T-1} \frac{C_t}{t} \langle\, \hat{z}_t,\, w_t \,\rangle}_{\approx O(\log T/T) Z_T} + \sum_{t=T/2}^{T-1} \langle\, \hat{z}_t,\, A_t(x_t - x^*) \,\rangle. \tag{4}$$

This bound is mysterious in that the left-hand side is an upper bound on the total conditional variance of Z_T, whereas the right-hand side essentially contains a scaled version of Z_T itself. This is the "chicken and egg phenomenon" alluded to in Section 4, and it poses another one of the main challenges of bounding Z_T. This bound inspires our main probabilistic tool, Theorem 3.2.

Theorem 3.2 (Generalized Freedman). Let $\{d_i, \mathcal{F}_i\}_{i=1}^n$ be a martingale difference sequence. Suppose $v_{i-1} \geq 0$, $i \in [n]$ are $\mathcal{F}_{i-1}$-measurable random variables such that $\mathrm{E}\left[\exp(\lambda d_i) \mid \mathcal{F}_{i-1}\right] \leq \exp\left(\frac{\lambda^2}{2} v_{i-1}\right)$ for all $i \in [n]$, $\lambda > 0$. Let $S_t = \sum_{i=1}^t d_i$ and $V_t = \sum_{i=1}^t v_{i-1}$. Let $\alpha_i \geq 0$ and set $\alpha = \max_{i \in [n]} \alpha_i$. Then

$$\Pr\left[\bigcup_{t=1}^n \left\{S_t \geq x \text{ and } V_t \leq \sum_{i=1}^t \alpha_i d_i + \beta\right\}\right] \leq \exp\left(-\frac{x}{4\alpha + 8\beta/x}\right) \qquad \forall x, \beta > 0.$$

In order to apply Theorem 3.2, we need to refine Lemma 6.4 to replace the terms $\left\|x_{T/2} - x^*\right\|^2$ and $\sum_{t=T/2}^{T-1} \langle\, \hat{z}_t,\, A_t(x_t - x^*) \,\rangle$ with sufficient high probability upper bounds. In Rakhlin et al. (2012), they showed that $\|x_t - x^*\|^2 \leq O(\log\log(T)/T)$ for all $\frac{T}{2} \leq t \leq T$ simultaneously, with high probability, so using that would give a slightly suboptimal result. In contrast, our analysis only needs a high probability bound on $\left\|x_{T/2} - x^*\right\|^2$ and $\sum_{t=T/2}^T A_t \|x_t - x^*\|^2$; this allows us to avoid a $\log\log T$ factor here. Indeed, we have

Theorem 6.5 *Both of the following hold:*

- *For all $t \geq 2$, $\|x_t - x^*\|^2 \leq O\left(\log(1/\delta)/t\right)$ with probability $1 - \delta$, and*
- *$\sum_{t=2}^T \sigma_t \|x_t - x^*\|^2 = O\left(\sum_{t=2}^T \frac{\sigma_t}{t} \log(1/\delta)\right)$ w.p. $1 - \delta$, for all $\sigma_t \geq 0$.*

The proof of Theorem 6.5 (Subsection 6.2) uses our tool for bounding recursive stochastic processes (Theorem 4.1). So, we need to expose a recursive relationship between $\|x_{t+1} - x^*\|^2$ and $\|x_t - x^*\|^2$ that satisfies the conditions of Theorem 4.1. Interestingly, Theorem 6.5 is also the main ingredient in the suffix averaging analysis (Subsection E.1). We now have enough to give our refined version of Lemma 6.4, which is now in a format usable by Freedman's Inequality.

Lemma 6.6 *For every $\delta > 0$ there exists positive values $R = O\left(\frac{\log^2 T \log(1/\delta)}{T^2}\right)$, $C_t = O\left(\log T\right)$ such that $\sum_{t=T/2}^T \|w_t\|^2 \leq R + \sum_{t=T/2}^{T-1} \frac{C_t}{t} \langle\, \hat{z}_t,\, w_t \,\rangle$, with probability at least $1 - \delta$.*

Proof The lemma essentially follows from combining our bounds in Theorem 6.5 with an easy corollary of Freedman's Inequality (Corollary C.4) which states that a high probability bound of M on the TCV of a martingale implies a high probability bound of $\sqrt{M}$ on the martingale.

Let R_1, R_2, C_t, and A_t be as in Lemma 6.4, and consider the resulting upper bound on $\sum_{t=T/2}^T \|w_t\|^2$. The first claim in Theorem 6.5 gives $R_2 \left\|x_{T/2} - x^*\right\|^2 = O\left(\frac{\log^2 T \log(1/\delta)}{T^2}\right)$ because $R_2 = O\left(\log T/T\right)$.

By the second claim in Theorem 6.5, we have $\sum_{t=T/2}^{T-1} A_t^2 \|x_t - x^*\|^2 = O\left(\frac{\log^2 T}{T^4} \log(1/\delta)\right)$ with probability at least $1 - \delta$ because each $A_t = O\left(\frac{\log T}{T^2}\right)$. Hence, we have derived a high probability bound on the total conditional variance of $\sum_{t=T/2}^T \langle\, \hat{z}_t,\, A_t(x_t - x^*) \,\rangle$. Therefore, we turn this into a high probability bound on the martingale itself by applying Corollary C.4 and obtain $\sum_{t=T/2}^{T-1} \langle\, \hat{z}_t,\, A_t(x_t - x^*) \,\rangle = O\left(\frac{\log^2 T \log(1/\delta)}{T^2}\right)$ with probability at least $1 - \delta$. ■

Now that we have derived an upper bound on the total conditional variance of Z_T in the form required by our Generalized Freedman Inequality (Theorem 3.2), we are finally ready to prove Lemma 6.2 (our high probability upper bound on the noise, Z_T).

Proof (of Lemma 6.2). We have demonstrated that Z_T satisfies the "Chicken and Egg" phenomenon with high probability. Translating this into a high probability upper bound on the martingale Z_T itself is a corollary of Theorem 3.2.

Indeed, consider a filtration $\{\mathcal{F}_t\}_{t=T/2}^T$. Let $d_t = \langle\, a_t,\, b_t \,\rangle$ define a martingale difference sequence where $\|a_t\| \le 1$ and $\mathrm{E}\left[\, a_t \mid \mathcal{F}_{t-1} \,\right] = 0$. Suppose there are positive values, R, α_t, such that $\max_{t=T/2}^T\{\alpha_t\} = O\left(\sqrt{R}\right)$ and $\sum_{t=T/2}^T \|b_t\|^2 \le \sum_{t=T/2}^T \alpha_t d_t + R\log(1/\delta)$ with probability at least $1-\delta$. Corollary C.5 bounds the martingale at step T by $\sqrt{R}\log(1/\delta)$ with high probability.

To conclude, Lemma 6.6 allows us to apply Corollary C.5 with $a_t = \hat{z}_t$, $b_t = w_t$, $\alpha_t = (C_t/t)$ for $t = T/2, \ldots, T-1$, $\alpha_T = 0$, $\max_{t=T/2}^T\{\alpha_t\} = O\left(\log T/T\right)$, and $R = O\left(\log^2 T/T^2\right)$. ■

6.2. High Probability Bounds on Squared Distances to x^*

We prove Theorem 6.5. The following claim can be extracted from (Rakhlin et al., 2012).

Claim 6.7 (Proof of Lemma 6 in Rakhlin et al. (2012)) *Suppose f is 1-strongly-convex and 1-Lipschitz. Define $Y_t = t\left\|x_{t+1} - x^*\right\|^2$ and $U_t = \langle\, \hat{z}_{t+1},\, x_{t+1} - x^* \,\rangle / \left\|x_{t+1} - x^*\right\|_2$. Then $Y_{t+1} \le \left(\frac{t-1}{t}\right) Y_t + 2 \cdot U_t\sqrt{\frac{Y_t}{t}} + \frac{4}{t+1}$.*

This claim exposes a recursive relationship between $\left\|x_{t+1} - x^*\right\|^2$ and $\left\|x_t - x^*\right\|^2$ and inspires our probabilistic tool for recursive stochastic processes (Theorem 4.1), used to prove Theorem 6.5:

Proof (of Theorem 6.5). Consider the stochastic process $(Y_t)_{t=1}^{T-1}$ where Y_t is as defined by Claim 6.7. Note that Y_t satisfies the conditions of Theorem 4.1 with $X_t = Y_t$, $\hat{w}_t = U_t$, $\alpha_t = \frac{t-1}{t} = 1 - 1/t$, $\beta_t = 2/\sqrt{t}$, and $\gamma_t = 4/(t+1)$. Observe that U_t is a $\mathcal{F}_{t+1}$ measurable random variable which is mean zero conditioned on $\mathcal{F}_t$ Furthermore, note that $|U_t| \le 1$ with probability 1 because $\|\hat{z}_{t+1}\| \le 1$ with probability 1. It is easy to check that $\max_{1\le t\le T}\left(\frac{2\gamma_t}{1-\alpha_t}, \frac{2\beta^2}{1-\alpha_t}\right) = 8$ with the above setup. We claim that $\left\|x_2 - x^*\right\|^2 \le 8$ with probability 1. Indeed by 1-strong-convexity and 1-Lipschitzness of f, we have $\left\|x_t - x^*\right\| \ge \langle\, g_t,\, x_t - x^* \,\rangle \ge \frac{1}{2}\left\|x_t - x^*\right\|^2$. Apply Theorem 4.1 to obtain:

- For every $t = 1, \ldots T-1$, $\Pr\left[\, Y_t \ge 8\log(1/\delta) \,\right] \le \mathrm{e}\delta$.
- Let $\sigma'_t \ge 0$ for $t = 1, \ldots, T-1$. Then, $\Pr\left[\, \sum_{t=1}^{T-1} \sigma'_t Y_t \ge 8\sum_{t=1}^{T-1} \sigma'_t \,\right] \le \mathrm{e}\delta$.

Recalling that $Y_t = t\left\|x_{t+1} - x^*\right\|^2$ and setting $\sigma'_t = \sigma_t/t$ proves Theorem 6.5. ■

7. Open questions

There remain some interesting open questions. The first is whether or not there exists a sequence of step sizes for which the individual iterates obtain, for all t, error $o(\log(t)/t)$ in the strongly-convex cases and $o(\log(t)/\sqrt{t})$ in the Lipschitz case. Note that in the strongly convex case, (Jain et al., 2019) showed that for a fixed T, one can obtain a rate of $O(1/T)$ for the last iterate and that in the *stochastic* setting, *no* choice of step sizes yields expected error $O(1/t)$ for all $t > 0$.

Another question is to determine the exact dependence on δ of our high probability upper bound on the error of the final iterate. In the strongly-convex case, our best lower bound has an additive $\log(1/\delta)$, factor whereas our upper bound has a multiplicative factor of $\log(1/\delta)$. In contrast, for the final iterate in the Lipschitz case, we do not know a $\log(1/\delta)$ lower bound on the error; conceivably the upper bound could be improved to $O\big((\log(T) + \sqrt{\log(1/\delta)})/\sqrt{T}\big)$.

References

Peter L. Bartlett, Varsha Dani, Thomas Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. In *21th Annual Conference on Learning Theory (COLT 2008)*, July 2008.

Sebastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3–4), 2015.

Xiequan Fan, Ion Grama, and Quansheng Liu. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20, 2015.

David A. Freedman. On tail probabilities for martingales. *Annals of Probability*, 3(1):100–118, 1975.

Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. *CoRR*, abs/1812.05217, 2018. URL http://arxiv.org/abs/1812.05217.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3–4), 2015.

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.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms I*. Springer-Verlag, 1996.

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of SGD information theoretically optimal. *arXiv preprint arXiv:1904.12443*, 2019.

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

Sham M. Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In *NIPS*, pages 801–808, 2008.

Philip Klein and Neal E Young. On the number of iterations for Dantzig–Wolfe optimization and packing-covering approximation algorithms. *SIAM Journal on Computing*, 44(4):1154–1172, 2015.

Simon Lacoste-Julien, Mark W. Schmidt, and Francis R. Bach. A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method, December 2012. arXiv:1212.2002.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

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

Yu. Nesterov and V. Shikhman. Quasi-monotone subgradient methods for nonsmooth convex minimization. *Journal of Optimization Theory and Applications*, 165(3):917–940, Jun 2015.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of ICML*, 2012.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, September 1951.

David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: primal estimated sub-gradient solver for SVM. *Mathematical Programming*, 127(1):3–30, 2011.

Ohad Shamir. Open problem: Is averaging needed for strongly convex stochastic gradient descent? *Proceedings of the 25th Annual Conference on Learning Theory, PMLR*, 23:47.1–47.3, 2012.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. *Proceedings of the 30th International Conference on Machine Learning, PMLR*, 28(1):71–79, 2013.

Cheng Tang and Claire Monteleoni. Convergence analysis of stochastic gradient descent on strongly convex objective functions. In *Proceedings of ROKS*, pages 111–112, 2013.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

Appendix A. Standard results

Lemma A.1 (Exponentiated Markov) *Let X be a random variable and $\lambda > 0$. Then* $\Pr[X > t] \leq \exp(-\lambda t)\,\mathrm{E}[\exp(\lambda X)]$.

Theorem A.2 (Cauchy-Schwarz) *Let X and Y be random variables. Then* $|\mathrm{E}[XY]|^2 \leq \mathrm{E}\left[X^2\right]\mathrm{E}\left[Y^2\right]$.

Theorem A.3 (Hölder's Inequality) *Let $X_1, \ldots, X_n$ be random variables and $p_1, \ldots, p_n > 0$ be such that $\sum_i 1/p_i = 1$. Then* $\mathrm{E}[\prod_{i=1}^n |X_i|] \leq \prod_{i=1}^n (\mathrm{E}[|X_i|^{p_i}])^{1/p_i}$

Lemma A.4 *Let $X_1, \ldots, X_n$ be random variables and $K_1, \ldots, K_n > 0$ be such that* $\mathrm{E}[\exp(\lambda X_i)] \leq \exp(\lambda K_i)$ *for all $\lambda \leq 1/K_i$. Then* $\mathrm{E}[\exp(\lambda \sum_{i=1}^n X_i)] \leq \exp(\lambda \sum_{i=1}^n K_i)$ *for all $\lambda \leq 1/\sum_{i=1}^n K_i$.*

Proof Let $p_i = \sum_{j=1}^n K_j/K_i$ and observe that $p_i K_i = \sum_{j=1}^n K_j$. By assumption, if $\lambda p_i \leq 1/K_i$ (i.e. $\lambda \leq 1/\sum_{j=1}^n K_j$) then $\mathrm{E}[\exp(\lambda p_i X_i)] \leq \exp(\lambda p_i K_i)$. Applying Theorem A.3, we conclude that

$$\mathrm{E}\left[\exp(\lambda \sum_{i=1}^n X_i)\right] \leq \prod_{i=1}^n \mathrm{E}[\exp(\lambda p_i X_i)]^{1/p_i} \leq \prod_{i=1}^n \exp(\lambda p_i K_i)^{1/p_i} = \exp(\lambda \sum_{i=1}^n K_i).$$

■

Lemma A.5 (Hoeffding's Lemma) *Let X be any real valued random variable with expected value* $\mathrm{E}[X] = 0$ *and such that $a \leq X \leq b$ almost surely. Then, for all $\lambda \in \mathbb{R}$,* $\mathrm{E}[\exp(\lambda X)] \leq \exp\left(\lambda^2 (b-a)^2/8\right)$.

Claim A.6 ((Vershynin, 2018, Proposition 2.5.2)) *Suppose there is $c > 0$ such that for all $0 < \lambda \leq \frac{1}{c}$,* $\mathrm{E}\left[\exp\left(\lambda^2 X^2\right)\right] \leq \exp\left(\lambda^2 c^2\right)$ *for some constant c. Then, if X is mean zero it holds that*

$$\mathrm{E}\left[\exp\left(\lambda X\right)\right] \leq \exp\left(\lambda^2 c^2\right),$$

for all $\lambda \in \mathbb{R}$.

Proof Without loss of generality, assume $c = 1$; otherwise replace X with X/c. Using the numeric inequality $e^x \leq x + e^{x^2}$ which is valid for all $x \in \mathbb{R}$, if $|\lambda| \leq 1$ then $\mathrm{E}[\exp(\lambda X)] \leq \mathrm{E}[\lambda X] + \mathrm{E}\left[\exp(\lambda^2 X^2)\right] \leq \exp(\lambda^2)$. On the other hand, if $|\lambda| \geq 1$, we may use the numeric inequality[4] $ab \leq a^2/2 + b^2/2$, valid for all $a, b \in \mathbb{R}$, to obtain

$$\mathrm{E}[\exp(\lambda X)] \leq \mathrm{E}\left[\exp(\lambda^2/2 + X^2/2)\right] \leq \exp(\lambda^2/2)\exp(\lambda^2/2) = \exp(\lambda^2).$$

■

Claim A.7 *Suppose X is a random variable such that there exists constants c and C such that* $\mathrm{E}[\exp(\lambda X)] \leq c\exp(\lambda C)$ *for all $\lambda \leq 1/C$. Then,* $\Pr[X \geq C\log(1/\delta)] \leq c\epsilon\delta$.

Proof Apply Lemma A.1 to $\Pr[X \geq t]$ to get $\Pr[X \geq t] \leq c\exp(-\lambda t + \lambda C)$. Set $\lambda = 1/C$ and $t = C\log(1/\delta)$ to complete the proof. ■

Claim A.8 ((Hiriart-Urruty and Lemaréchal, 1996, Eq. (3.1.6))) *Let $\mathcal{X}$ be a convex set and $x \in \mathcal{X} \subseteq \mathbb{R}^n$. Then $\|\Pi_{\mathcal{X}}(y) - x\| \leq \|y - x\|$ for all $y \in \mathbb{R}^n$.*

4. Young's Inequality

A.1. Useful Scalar Inequalities

Claim A.9 *For* $1 \le a \le b$, $\sum_{k=a}^{b} \frac{1}{\sqrt{k}} \le 2\frac{b-a+1}{\sqrt{b}}$.

Proof

$$\sum_{k=a}^{b} \frac{1}{\sqrt{k}} \;\le\; \int_{a-1}^{b} \frac{1}{\sqrt{x}}\,dx \;=\; 2(\sqrt{b} - \sqrt{a-1}) \;=\; 2\frac{b-a+1}{\sqrt{b}+\sqrt{a-1}}.$$

■

Claim A.10 *For any* $1 \le j \le t \le T$, *we have* $\frac{t-j}{(T-j+1)\sqrt{t}} \le \frac{1}{\sqrt{T}}$.

Proof The function $g(x) = \frac{x-j}{\sqrt{x}}$ has derivative

$$g'(x) = \frac{1}{\sqrt{x}}\Big(1 - \frac{x-j}{2x}\Big) = \frac{1}{\sqrt{x}}\Big(\frac{1}{2} + \frac{j}{2x}\Big).$$

This is positive for all $x > 0$ and $j \ge 0$, and so

$$\frac{t-j}{\sqrt{t}} \;\le\; \frac{T-j}{\sqrt{T}},$$

for all $0 < t \le T$. This implies the claim. ■

Claim A.11

$$\sum_{\ell=k+1}^{m} \frac{1}{\ell^2} \;\le\; \frac{1}{k} - \frac{1}{m}.$$

Proof The sum may be upper-bounded by an integral as follows:

$$\sum_{\ell=k+1}^{m} \frac{1}{\ell^2} \;\le\; \int_{k}^{m} \frac{1}{x^2}\,dx \;=\; \frac{1}{k} - \frac{1}{m}.$$

■

Claim A.12 *Let* $\alpha_j = \frac{1}{(T-j)(T-j+1)}$. *Let* a, b *be such that* $a < b \le T$. *Then,*

$$\sum_{j=a}^{b} \alpha_j \;=\; \frac{1}{T-b} - \frac{1}{T-a+1} \;\le\; \frac{1}{T-b}.$$

Proof

$$\sum_{j=a}^{b} \alpha_j \;=\; \sum_{j=a}^{b} \frac{1}{(T-j)(T-j+1)} \;=\; \sum_{j=a}^{b} \left(\frac{1}{T-j} - \frac{1}{T-(j-1)}\right),$$

which is a telescoping sum. ■

Claim A.13 *Suppose* $a < b$. *Then,* $\log(b/a) \le (b-a)/a$.

Claim A.14 *Let* $b \ge a > 1$. *Then,* $\sum_{i=a}^{b} \frac{1}{i} \le \log\big(b/(a-1)\big)$.

Appendix B. Omitted proofs from Section 5

Proof (of Lemma 5.4). By definition, $z_1 = x_1 = 0$. By Claim 5.3, the subgradient returned at x_1 is $h_1 + x_1 = h_1$, so Algorithm 1 sets $y_2 = x_1 - \eta_1 h_1 = e_1$, the first standard basis vector. Then Algorithm 1 projects onto the feasible region, obtaining $x_2 = \Pi_{\mathcal{X}}(y_2)$, which equals e_1 since $y_2 \in \mathcal{X}$. Since z_2 also equals e_1, the base case is proven.

So assume $z_t = x_t$ for $2 \le t < T$; we will prove that $z_{t+1} = x_{t+1}$. By Claim 5.3, the subgradient returned at x_t is $\hat{g}_t = h_t + z_t$. Then Algorithm 1 sets $y_{t+1} = x_t - \eta_t \hat{g}_t$. Since $x_t = z_t$ and $\eta_t = 1/t$, we obtain

$$\begin{aligned}
y_{t+1,j} &= z_{t,j} - \frac{1}{t}(h_{t,j} + z_{t,j}) \\
&= \frac{t-1}{t} z_{t,j} - \frac{1}{t} h_{t,j} \\
&= \frac{t-1}{t} \left\{ \begin{array}{ll} \frac{1-(t-j-1)a_j}{t-1} & (\text{for } j < t) \\ 0 & (\text{for } j \ge t) \end{array} \right\} - \frac{1}{t} \left\{ \begin{array}{ll} a_j & (\text{for } j < t) \\ -1 & (\text{for } j = t) \\ 0 & (\text{for } j > t) \end{array} \right\} \\
&= \frac{1}{t} \left\{ \begin{array}{ll} 1-(t-j-1)a_j & (\text{for } j < t) \\ 0 & (\text{for } j \ge t) \end{array} \right\} - \frac{1}{t} \left\{ \begin{array}{ll} a_j & (\text{for } j < t) \\ -1 & (\text{for } j = t) \\ 0 & (\text{for } j > t) \end{array} \right\} \\
&= \frac{1}{t} \left\{ \begin{array}{ll} 1-(t-j)a_j & (\text{for } j < t) \\ 1 & (\text{for } j = t) \\ 0 & (\text{for } j \ge t+1) \end{array} \right\}
\end{aligned}$$

So $y_{t+1} = z_{t+1}$. Since $x_{t+1} = \Pi_{\mathcal{X}}(y_{t+1})$ is defined to be the projection onto $\mathcal{X}$, and $y_{t+1} \in \mathcal{X}$ by Claim 5.2, we have $x_{t+1} = y_{t+1} = z_{t+1}$. ■

Claim B.1 *For any $k \in [T]$, let $\bar{x} = \sum_{t=T-k+2}^{T+1} \lambda_t x_t$ be any convex combination of the last k iterates. Then*

$$f(\bar{x}) \ge \frac{\ln(T) - \ln(k)}{4T}.$$

Proof By Lemma 5.4, $x_t = z_t \; \forall t \in [T+1]$. By Claim 5.2, every $z_t \ge 0$ so $\bar{x} \ge 0$. Moreover, $z_{t,j} \ge 1/2T$ for all $T - k + 2 \le t \le T + 1$ and $1 \le j \le T - k + 1$. Consequently, $\bar{x}_j \ge 1/2T$ for

all $1 \le j \le T-k+1$. Thus,

$$\begin{aligned} f(\bar{x}) &\ge h_{T+1}^{\mathsf{T}} \bar{x} \qquad \text{(by definition of } f\text{)} \\ &= \sum_{j=1}^{T-k+1} h_{T+1,j} \underbrace{\bar{x}_j}_{\ge 1/2T} + \sum_{j=T-k+2}^{T} \underbrace{h_{T+1,j}\, \bar{x}_j}_{\ge 0} \\ &\ge \sum_{j=1}^{T-k+1} a_j \cdot \frac{1}{2T} \\ &= \frac{1}{2T} \sum_{j=1}^{T-k+1} \frac{1}{2(T+1-j)} \\ &\ge \frac{1}{4T} \sum_{j=1}^{T-k+1} \frac{1}{T+1-j} \\ &\ge \frac{1}{4T} \int_1^{T-k+1} \frac{1}{T+1-x}\, dx \\ &= \frac{\log(T) - \log(k)}{4T} \end{aligned}$$

■

Remark B.2 In order to achieve large error for the T-th iterate, Theorem 3.3 constructs a function parameterized by T. It is not possible for a *single* function to achieve error $\omega(1/T)$ for the T-th iterate simultaneously for *all* T, because that would contradict the fact that suffix averaging achieves error $O(1/T)$. Nevertheless, it is possible to construct a single function achieving error $g(T)$, for infinitely many T, for any function $g(T) = o(\log(T)/T)$, e.g., $g(T) = \log(T)/(T \log^*(T))$ where $\log^*(T)$ is the iterated logarithm. Formally, we can construct a function $f \in \ell_2$ such that $\inf_x f(x) = 0$ but $\limsup_T \frac{f(x_T)}{g(T)} = +\infty$. The main idea is to define a sequence $T_1 \ll T_2 \ll T_3 \ll \ldots$ and consider the "concatenation" of $c_1 f_{T_1}, c_2 f_{T_2}, \ldots$ into a single function f (here, c_i are appropriate constants chosen to ensure that f remains Lipschitz). Essentially, one can imagine running multiple instances of gradient descent in parallel where each instance corresponds to a bad instance given by Theorem 3.3, albeit at different scales. However, this construction has a slight loss (i.e., the $\log^*(T)$) to ensure that f remains Lipschitz. The details are discussed in the full version of this paper (Harvey et al., 2018).

Appendix C. Proof of Theorem 3.2 and Corollaries

In this section we prove Theorem 3.2 and derive some corollaries. We restate Theorem 3.2 here for convenience.

Theorem 3.2. Let $\{d_i, \mathcal{F}_i\}_{i=1}^n$ be a martingale difference sequence. Suppose $v_{i-1} \ge 0$, $i \in [n]$ are $\mathcal{F}_{i-1}$-measurable random variables such that $\mathrm{E}\left[\exp(\lambda d_i) \mid \mathcal{F}_{i-1}\right] \le \exp\left(\frac{\lambda^2}{2} v_{i-1}\right)$ for all $i \in [n]$, $\lambda > 0$. Let $S_t = \sum_{i=1}^t d_i$ and $V_t = \sum_{i=1}^t v_{i-1}$. Let $\alpha_i \ge 0$ and set $\alpha = \max_{i \in [n]} \alpha_i$. Then

$$\Pr\left[\bigcup_{t=1}^{n} \left\{ S_t \ge x \text{ and } V_t \le \sum_{i=1}^{t} \alpha_i d_i + \beta \right\}\right] \le \exp\left(-\frac{x}{4\alpha + 8\beta/x}\right) \qquad \forall x, \beta > 0.$$

Proof (of Theorem 3.2). Fix $\lambda < 1/(2\alpha)$ and define $c = c(\lambda, \alpha)$ as in Claim C.2. Let $\tilde{\lambda} = \lambda + c\lambda^2\alpha$. Define $\mathcal{U}_0 \coloneqq 1$ and for $t \in [n]$, define

$$\mathcal{U}_t(\lambda) \coloneqq \exp\left(\sum_{i=1}^{t}(\lambda + c\lambda^2\alpha_i)d_i - \sum_{i=1}^{t}\frac{\tilde{\lambda}^2}{2}v_{i-1}\right).$$

Claim C.1 *$\mathcal{U}_t(\lambda)$ is a supermartingale w.r.t. $\mathcal{F}_t$.*

Proof For all $t \in [n]$:

$$\begin{aligned}
\mathrm{E}\left[\mathcal{U}_t(\lambda) \mid \mathcal{F}_{t-1}\right] &= \mathcal{U}_{t-1}(\lambda)\exp\left(-\frac{\tilde{\lambda}^2}{2}v_{t-1}\right)\mathrm{E}\left[\exp\left((\lambda + c\lambda^2\alpha_t)d_t\right) \mid \mathcal{F}_{t-1}\right] \\
&\le \mathcal{U}_{t-1}(\lambda)\exp\left(-\frac{\tilde{\lambda}^2}{2}v_{t-1}\right)\exp\left(\frac{(\lambda + c\lambda^2\alpha_i)^2}{2}v_{t-1}\right) \\
&\le \mathcal{U}_{t-1}(\lambda)\exp\left(-\frac{\tilde{\lambda}^2}{2}v_{t-1}\right)\exp\left(\frac{\tilde{\lambda}^2}{2}v_{t-1}\right) \\
&= \mathcal{U}_{t-1}(\lambda),
\end{aligned}$$

where the second line follows from the assumption that $\mathrm{E}\left[\exp(\lambda d_t) \mid \mathcal{F}_{t-1}\right] \le \exp\left(\frac{\lambda^2}{2}v_{t-1}\right)$ for all $\lambda > 0$ and the third line is because $\lambda + c\lambda^2\alpha_i \le \tilde{\lambda}$ (since $c \ge 0$ and $\alpha_i \le \alpha$). We conclude that $\mathcal{U}_t(\lambda)$ is a martingale w.r.t. $\mathcal{F}_t$. ■

Define the stopping time $T = \min\left\{\, t : S_t \ge x \text{ and } V_t \le \sum_{i=1}^{t}\alpha_i d_i + \beta \,\right\}$ with the convention that $\min\emptyset = \infty$. Since $\mathcal{U}_t$ is a supermartingale w.r.t. $\mathcal{F}_t$, $\mathcal{U}_{T\wedge t}$ is a supermartingale w.r.t. $\mathcal{F}_t$. Hence,

$$\begin{aligned}
&\Pr\left[\bigcup_{t=1}^{n}\left\{S_t \ge x \text{ and } V_t \le \sum_{i=1}^{t}\alpha_i d_i + \beta\right\}\right] \\
&= \Pr\left[S_{T\wedge n} \ge x \text{ and } V_{T\wedge n} \le \sum_{i=1}^{T\wedge n}\alpha_i d_i + \beta\right] \\
&= \Pr\left[\lambda S_{T\wedge n} \ge \lambda x \text{ and } c\lambda^2 V_{T\wedge n} \le c\lambda^2\sum_{i=1}^{T\wedge n}\alpha_i d_i + c\lambda^2\beta\right] \\
&\le \Pr\left[\sum_{i=1}^{T\wedge n}(\lambda + \alpha_i\lambda^2)d_i - c\lambda^2 V_{T\wedge n} \ge \lambda x - c\lambda^2\beta\right] \\
&\le \mathrm{E}\left[\exp\left(\sum_{i=1}^{T\wedge n}(\lambda + \alpha_i\lambda^2)d_i - c\lambda^2 V_{T\wedge n}\right)\right]\cdot\exp(-\lambda x + c\lambda^2\beta).
\end{aligned}$$

Recall that c was chosen (via Claim C.2) so that $c\lambda^2 = \frac{\tilde{\lambda}^2}{2}$. Hence,

$$\begin{aligned}\mathrm{E}\left[\exp\left(\sum_{i=1}^{T\wedge n}(\lambda+\alpha_i\lambda^2)d_i - c\lambda^2 V_{T\wedge n}\right)\right] &= \mathrm{E}\left[\exp\left(\sum_{i=1}^{T\wedge n}(\lambda+\alpha_i\lambda^2)d_i - \frac{\tilde{\lambda}^2}{2}V_{T\wedge n}\right)\right]\\ &= \mathrm{E}\left[\mathcal{U}_{T\wedge n}(\lambda)\right] \le 1.\end{aligned}$$

Since $\lambda < 1/(2\alpha)$ was arbitrary, we conclude that

$$\begin{aligned}\Pr\left[\bigcup_{t=1}^{n}\left\{S_t \ge x \text{ and } V_t \le \sum_{i=1}^{t}\alpha_i d_i + \beta\right\}\right] &\le \exp(-\lambda x + c\lambda^2\beta)\\ &\le \exp(-\lambda x + 2\lambda^2\beta),\end{aligned}$$

where the inequality is because $c \le 2$. Now, we can pick $\lambda = \frac{1}{2\alpha+4\beta/x} < \frac{1}{2\alpha}$ to conclude that

$$\begin{aligned}\Pr\left[\bigcup_{t=1}^{n}\left\{S_t \ge x \text{ and } V_t \le \sum_{i=1}^{t}\alpha_i d_i + \beta\right\}\right] &\le \exp(-\lambda(x-2\lambda\beta))\\ &\le \exp\left(-\lambda\left(x - \frac{2\beta}{2\alpha+4\beta/x}\right)\right)\\ &\le \exp\left(-\lambda\left(x - \frac{2\beta}{4\beta/x}\right)\right)\\ &= \exp\left(-\frac{\lambda x}{2}\right)\\ &= \exp\left(-\frac{x}{4\alpha+8\beta/x}\right).\end{aligned}$$

■

Claim C.2 *Let $\alpha \ge 0$ and $\lambda \in [0, 1/2\alpha)$. Then there exists $c = c(\lambda,\alpha) \in [0,2]$ such that $2c\lambda^2 = (\lambda + c\lambda^2\alpha)^2$.*

Proof If $\lambda = 0$ or $\alpha = 0$ then the claim is trivial (just take $c = 0$). So assume $\alpha, \lambda > 0$.

The equality $2c\lambda^2 = (\lambda + c\lambda^2\alpha)^2$ holds if and only if $p(c) \coloneqq \alpha^2\lambda^2c^2 + (2\lambda\alpha - 2)c + 1 = 0$. The discriminant of p is $(2\lambda\alpha-2)^2 - 4\alpha^2\lambda^2 = 4 - 8\lambda\alpha$. Since $\lambda\alpha \le 1/2$, the discriminant of p is non-negative so the roots of p are real. The smallest root of p is located at

$$\begin{aligned}c &= \frac{2 - 2\alpha\lambda - \sqrt{(2\alpha\lambda-2)^2 - 4\lambda^2\alpha^2}}{2\lambda^2\alpha^2}\\ &= \frac{1-\alpha\lambda-\sqrt{1-2\alpha\lambda}}{\alpha^2\lambda^2}.\end{aligned}$$

Set $\gamma = \alpha\lambda$. Using the numeric inequality $\sqrt{1-x} \ge 1 - x/2 - x^2/2$ valid for all $x \le 1$, we have

$$c \le \frac{1-\gamma-(1-\gamma-2\gamma^2)}{\gamma^2} = 2.$$

On the other hand, using the numeric inequality $\sqrt{1-x} \le 1 - x/2 - x^2/8$ valid for all $0 \le x \le 1$, we have

$$c \ge \frac{1-\gamma-(1-\gamma-\gamma^2/2)}{\gamma^2} = \frac{1}{2} \ge 0.$$

■

C.1. Corollaries of Theorem 3.2

In this paper, we often deal with martingales, M_n, where the total conditional variance of the martingale is bounded by a linear transformation of the martingale, *with high probability* (which is what we often refer to as the "chicken and egg" phenomenon — the bound on the total conditional variance of M_n involves M_n itself). Transforming these entangled high probability bounds on the total conditional variance into high probability bounds on the martingale itself are easy consequences of our Generalized Freedman inequality (Theorem 3.2).

Lemma C.3 *Let $\{d_i, \mathcal{F}_i\}_{i=1}^n$ be a martingale difference sequence. Let v_{i-1} be a $\mathcal{F}_{i-1}$ measurable random variable such that $\mathrm{E}\left[\exp\left(\lambda d_i\right) \mid \mathcal{F}_{i-1}\right] \leq \exp\left(\frac{\lambda^2}{2} v_{i-1}\right)$ for all $\lambda > 0$ and for all $i \in [n]$. Define $S_n = \sum_{i=1}^n d_i$ and define $V_n = \sum_{i=1}^n v_{i-1}$. Let $\delta \in (0,1)$ and suppose there are positive values $R(\delta) > 0$, and non-negative values $\{\alpha_i\}_{i=1}^n$ such that $\Pr\left[V_n \leq \sum_{i=1}^n \alpha_i d_i + R(\delta)\right] \geq 1-\delta$. Then,*

$$\Pr\left[S_n \geq x\right] \leq \delta + \exp\left(-\frac{x^2}{4\left(\max_{i=1}^n\{\alpha_i\}\right)x + 8R(\delta)}\right).$$

Proof Fix $\delta \in (0,1)$. Define the following events: $\mathcal{E}(x) = \{S_n \geq x\}$, $\mathcal{G} = \{V_n \leq \sum_{i=1}^n \alpha_i d_i + R(\delta)\}$.

$$\begin{aligned}
\Pr\left[S_n \geq x\right] &= \Pr\left[\mathcal{E}(x) \wedge \mathcal{G}\right] + \Pr\left[\mathcal{E}(x) \wedge \mathcal{G}^c\right] \\
&\leq \Pr\left[\mathcal{E}(x) \wedge \mathcal{G}\right] + \underbrace{\Pr\left[\mathcal{G}^c\right]}_{\leq \delta} \\
&\leq \delta + \exp\left(-\frac{x^2}{4\left(\max_{i=1}^n\{\alpha_i\}\right)x + 8R(\delta)}\right),
\end{aligned}$$

where the final inequality is due to applying Theorem 3.2 to $\Pr\left[\mathcal{E}(x) \wedge \mathcal{G}\right]$. ■

In this paper, we use Lemma C.3 in the following ways:

Corollary C.4 *Let $\{\mathcal{F}_t\}_{t=1}^T$ be a filtration and suppose that a_t are $\mathcal{F}_t$-measurable random variables and b_t are $\mathcal{F}_{t-1}$-measurable random variables. Further, suppose that*

1. *$\|a_t\| \leq 1$ almost surely and $\mathrm{E}\left[a_t \mid \mathcal{F}_{t-1}\right] = 0$; and*
2. *$\sum_{t=1}^T \|b_t\|^2 \leq R\log(1/\delta)$ with probability at least $1 - O(\delta)$.*

Define $d_t = \langle\, a_t,\, b_t \,\rangle$. Then $\sum_{t=1}^T d_t \;\leq\; O\big(\sqrt{R}\log(1/\delta)\big)$ with probability at least $1 - O(\delta)$.

Proof Since $\|a_t\| \leq 1$, by Cauchy-Schwarz we have that $|d_t| \leq \|b_t\|$. Therefore, $\mathrm{E}\left[\exp\left(\lambda d_t\right) \mid \mathcal{F}_{t-1}\right] \leq \exp\left(\frac{\lambda^2}{2}\|b_t\|^2\right)$ for all λ by Lemma A.5. Next, applying Lemma C.3 with $d_t = \langle\, a_t,\, b_t \,\rangle$ and $v_{t-1} = \|b_t\|^2$, $\alpha_i = 0$ for all i, and $R(\delta) = R\log(1/\delta)$ yields

$$\Pr\left[\sum_{t=1}^T d_t \geq x\right] \;\leq\; \delta + \exp\left(-\frac{x^2}{8R\log(1/\delta)}\right).$$

The last term is at most δ by taking $x = \sqrt{8R}\log(1/\delta)$. ■

Corollary C.5 *Let $\{\mathcal{F}_t\}_{t=1}^T$ be a filtration and suppose that a_t are $\mathcal{F}_t$-measurable random variables and b_t are $\mathcal{F}_{t-1}$-measurable random variables. Define $d_t = \langle\, a_t,\, b_t \,\rangle$. Assume that $\|a_t\| \le 1$ almost surely and $\mathrm{E}\left[\, a_t \mid \mathcal{F}_{t-1} \,\right] = 0$. Furthermore, suppose that there exists $R > 0$ and non-negative values $\{\alpha_t\}_{i=1}^{T-1}$ where $\max\{\alpha_t\}_{t=1}^{T-1} = O\left(\sqrt{R}\right)$, such that exactly one of the following holds for every $\delta \in (0,1)$*

1. $\sum_{t=1}^T \|b_t\|^2 \le \sum_{t=1}^{T-1} \alpha_t d_t + R\log(1/\delta)$ *with probability at least* $1 - O(\delta)$.
2. $\sum_{t=1}^T \|b_t\|^2 \le \sum_{t=1}^{T-1} \alpha_t d_t + R\sqrt{\log(1/\delta)}$ *with probability at least* $1 - O(\delta)$.

Then $\sum_{t=1}^T d_t \;\le\; O\big(\sqrt{R}\log(1/\delta)\big)$ with probability at least $1-\delta$.

Proof We prove only the first case, the second case can be proved by bounding $\sqrt{\log(1/\delta)}$ by $\log(1/\delta)$ and using the proof of the first case.

Since $\|a_t\| \le 1$, by Cauchy-Schwarz we have that $|d_t| \le \|b_t\|$. Therefore, $\mathrm{E}\left[\exp\left(\lambda d_t\right) \mid \mathcal{F}_{t-1}\right] \le \exp\left(\frac{\lambda^2}{2}\|b_t\|^2\right)$ for all λ by Lemma A.5. Next, applying Lemma C.3 with $d_t = \langle\, a_t,\, b_t \,\rangle$ and $v_{t-1} = \|b_t\|^2$, with $\alpha_T = 0$, and $R(\delta) = R\log(1/\delta)$ yields

$$\Pr\left[\sum_{t=1}^T d_t \ge x\right] \;\le\; \delta + \exp\left(-\frac{x^2}{4\left(\max_{t=1}^{T-1}\{\alpha_t\}\right)x + 8R\log(1/\delta)}\right).$$

The last term is at most δ by taking $x = \Theta\left(\sqrt{R}\log(1/\delta)\right)$ because $\max_{t=1}^{T-1}\{\alpha_t\} = O\left(\sqrt{R}\right)$. ■

Appendix D. Proof of Theorem 4.1

Theorem 4.1. Let $(X_t)_{t=1}^T$ be a stochastic process and let $(\mathcal{F}_t)_{t=1}^T$ be a filtration such that X_t is $\mathcal{F}_t$ measurable and X_t is non-negative almost surely. Assume that $\mathrm{E}\left[\exp(\lambda X_1)\right] \le \exp(\lambda K)$ with probability 1, for $\lambda \in (0, 1/K]$. Let $\alpha_t \in [0,1)$ and $\beta_t, \gamma_t \ge 0$ for every t. Let $\hat{w}_t$ be a mean-zero random variable conditioned on $\mathcal{F}_t$ such that $|\hat{w}_t| \le 1$ almost surely for every t. Suppose that $X_{t+1} \le \alpha_t X_t + \beta_t \hat{w}_t\sqrt{X_t} + \gamma_t$ for every t. Then, the following hold.

- For every t, $\Pr\left[\, X_t \ge K\log(1/\delta)\,\right] \le \mathrm{e}\delta$.
- More generally, if $\sigma_1, \ldots, \sigma_T \ge 0$, then $\Pr\left[\sum_{t=1}^T \sigma_t X_t \ge K\log(1/\delta)\sum_{t=1}^T \sigma_t\right] \le \mathrm{e}\delta$,

where $K = \max_{1\le t\le T}\left(\frac{2\gamma_t}{1-\alpha_t}, \frac{2\beta_t^2}{1-\alpha_t}\right)$.

Proof (of Theorem 4.1).

We begin by deriving a recursive MGF bound on X_t.

Claim D.1 *Suppose $0 \le \lambda \le \min_{1\le t\le T}\left(\frac{1-\alpha_t}{2\beta_t^2}\right)$. Then for every t,*

$$\mathrm{E}\left[\exp\left(\lambda X_{t+1}\right)\right] \;\le\; \exp\left(\lambda\gamma_t\right)\mathrm{E}\left[\exp\left(\lambda X_t\left(\frac{1+\alpha_t}{2}\right)\right)\right].$$

Proof

Observe that $\beta_t^2 \hat{w}_t^2 \sqrt{X_t}^2 \leq \beta_t^2 X_t$ because $|\hat{w}_t| \leq 1$ almost surely. Since $\beta_t^2 X_t$ is $\mathcal{F}_t$-measurable, we have $\mathrm{E}\left[\exp\left(\lambda^2 \beta_t^2 \hat{w}_t^2 \sqrt{X_t}^2\right) \mid \mathcal{F}_t\right] \leq \exp\left(\lambda^2 \beta_t^2 X_t\right)$ for all λ. Hence, we may apply Claim A.6 to obtain

$$\mathrm{E}\left[\exp\left(\lambda \beta_t \hat{w}_t \sqrt{X_t}\right) \mid \mathcal{F}_t\right] \;\leq\; \exp\left(\lambda^2 \beta_t^2 X_t\right). \tag{5}$$

Hence,

$$\begin{aligned}
\mathrm{E}\left[\exp\left(\lambda X_{t+1}\right)\right] &\leq \mathrm{E}\left[\exp\left(\lambda \alpha_t X_t + \lambda \beta_t \hat{w}_t \sqrt{X_t} + \lambda \gamma_t\right)\right] && \text{(by assumption)} \\
&= \mathrm{E}\left[\exp\left(\lambda \alpha_t X_t + \lambda \gamma_t\right) \mathrm{E}\left[\exp\left(\lambda \beta_t \hat{w}_t \sqrt{X_t}\right) \mid \mathcal{F}_t\right]\right] \\
&\leq \mathrm{E}\left[\exp\left(\lambda \alpha_t X_t + \lambda^2 \beta_t^2 X_t + \lambda \gamma_t\right)\right] && \text{(by Eq. (5))} \\
&= \mathrm{E}\left[\exp\left(\lambda X_t \left(\alpha_t + \lambda \beta_t^2\right) + \lambda \gamma_t\right)\right] \\
&\leq \mathrm{E}\left[\exp\left(\lambda \gamma_t + \lambda X_t \left(\frac{1+\alpha_t}{2}\right)\right)\right] && \text{(because } \lambda \leq \frac{1-\alpha_t}{2\beta_t^2}\text{)}.
\end{aligned}$$

■

Next, we prove an MGF bound on X_t.

Claim D.2 *For every t and for all $0 \leq \lambda \leq 1/K$, $\mathrm{E}\left[\exp\left(\lambda X_t\right)\right] \leq \exp\left(\lambda K\right)$.*

Proof Let $\lambda \leq 1/K$. We proceed by induction over t. The base case holds by assumption. Assume that $\mathrm{E}\left[\exp\left(\lambda X_t\right)\right] \leq \exp\left(\lambda K\right)$. Now, consider the MGF of X_{t+1}:

$$\begin{aligned}
\mathrm{E}\left[\exp\left(\lambda X_{t+1}\right)\right] &\leq \mathrm{E}\left[\exp\left(\lambda \gamma_t + \lambda X_t \left(\frac{1+\alpha_t}{2}\right)\right)\right] && \text{(by Claim D.1)} \\
&\leq \exp\left(\lambda \gamma_t + \lambda K \left(\frac{1+\alpha_t}{2}\right)\right),
\end{aligned}$$

where the first inequality is valid because $\lambda \leq 1/K \leq \min_{1 \leq t \leq T}\left(\frac{1-\alpha_t}{2\beta_t^2}\right)$ and the second inequality follows because $(1+\alpha_t)/2 < 1$ and so we can use the induction hypothesis since $\lambda(1+\alpha_t)/2 < \lambda \leq 1/K$. Furthermore, because $K \geq 2\gamma_t/\left(1-\alpha_t\right)$ we have

$$K \;\geq\; \frac{2\gamma_t}{1-\alpha_t} \;=\; \frac{\gamma_t}{1-\left(\frac{1+\alpha_t}{2}\right)},$$

which shows that $\gamma_t + K\left(\frac{1+\alpha_t}{2}\right) \leq K$. Hence,

$$\mathrm{E}\left[\exp\left(\lambda X_{t+1}\right)\right] \;\leq\; \exp\left(\lambda K\right),$$

as desired. ■

Now we are ready to complete the proof of both claims in Theorem 4.1.The first claim from Theorem 4.1 follows by observing our MGF bound on X_t and then applying the transition from MGF bounds to tail bounds given by Claim A.7.

Next, we prove the second claim from Theorem 4.1. Claim D.2 gives that for every t and for all $\lambda \leq 1/(\sigma_t K)$, we have $\mathrm{E}\left[\exp\left(\lambda \sigma_t X_t\right)\right] \leq \exp\left(\lambda \sigma_t K\right)$. Hence, we can combine these MGF bounds using Lemma A.4 to obtain $\mathrm{E}\left[\exp\left(\lambda \sum_{t=1}^T \sigma_t X_t\right)\right] \leq \exp\left(\lambda K \sum_{t=1}^T \sigma_t\right)$ for all $\lambda \leq$

$\left(K\sum_{t=1}^{T}\sigma_t\right)^{-1}$. With this MGF bound in hand, we may apply the transition from MGF bounds to tail bounds given by Claim A.7 to complete the proof of the second claim from Theorem 4.1. ■

Appendix E. Omitted proofs from Section 6

The following lemma is standard.

Lemma E.1 *Let f be an 1-strongly convex and 1-Lipschitz function. Consider running Algorithm 1 for T iterations. Then, for every $w \in \mathcal{X}$ and every $k \in [T]$,*

$$\sum_{t=k}^{T}\left[f(x_t)-f(w)\right] \le \frac{1}{2}\sum_{t=k}^{T}\eta_t\|\hat{g}_t\|^2+\frac{1}{2\eta_k}\|x_k-w\|^2+\sum_{t=k}^{T}\langle\,\hat{z}_t,\,x_t-w\,\rangle.$$

Proof

$$\begin{aligned}
f(x_t)-f(w) &\le \langle\,g_t,\,x_t-w\,\rangle-\frac{1}{2}\|x_t-w\|^2 && \text{(by strong-convexity)}\\
&= \langle\,\hat{g}_t,\,x_t-w\,\rangle-\frac{1}{2}\|x_t-w\|^2+\langle\,\hat{z}_t,\,x_t-w\,\rangle && (\hat{g}_t=g_t-\hat{z}_t)\\
&= \frac{1}{\eta_t}\langle\,x_t-y_{t+1},\,x_t-w\,\rangle-\frac{1}{2}\|x_t-w\|^2+\langle\,\hat{z}_t,\,x_t-w\,\rangle && (y_{t+1}=x_t-\eta_t\hat{g}_t)\\
&= \frac{1}{2\eta_t}\left(\|x_t-y_{t+1}\|^2+\|x_t-w\|^2-\|y_{t+1}-w\|^2\right)-\frac{1}{2}\|x_t-w\|^2+\langle\,\hat{z}_t,\,x_t-w\,\rangle\\
&\le \frac{1}{2\eta_t}\left(\|\eta_t\hat{g}_t\|^2+\|x_t-w\|^2-\|x_{t+1}-w\|^2\right)-\frac{1}{2}\|x_t-w\|^2+\langle\,\hat{z}_t,\,x_t-w\,\rangle.
\end{aligned}$$

Now, summing t from k to T,

$$\begin{aligned}
&\sum_{t=k}^{T}\left[f(x_t)-f(w)\right]\\
&\le \frac{1}{2}\sum_{t=k}^{T}\eta_t\|\hat{g}_t\|^2+\frac{1}{2}\sum_{t=k+1}^{T}\underbrace{\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}-1\right)}_{=0}\|x_t-w\|^2+\left(\frac{1}{2\eta_k}-\frac{1}{2}\right)\|x_k-w\|^2+\sum_{t=k}^{T}\langle\,\hat{z}_t,\,x_t-w\,\rangle\\
&\le \frac{1}{2}\sum_{t=k}^{T}\eta_t\|\hat{g}_t\|^2+\frac{1}{2\eta_k}\|x_k-w\|^2+\sum_{t=k}^{T}\langle\,\hat{z}_t,\,x_t-w\,\rangle \qquad (\eta_t=1/t),
\end{aligned}$$

as desired. ■

Proof (of Lemma 6.1). Let $k \in [T-1]$. Apply Lemma E.1, replacing k with $T-k$ and $w=x_{T-k}$ to obtain:

$$\sum_{t=T-k}^{T}\left[f(x_t)-f(x_{T-k})\right] \le \frac{1}{2}\sum_{t=T-k}^{T}\eta_t\|\hat{g}_t\|^2+\sum_{t=T-k}^{T}\langle\,\hat{z}_t,\,x_t-x_{T-k}\,\rangle.$$

Now, divide this by $k+1$ and define $S_k = \frac{1}{k+1}\sum_{t=T-k}^{T} f(x_t)$ to obtain

$$S_k - f(x_{T-k}) \;\le\; \frac{1}{2(k+1)}\sum_{t=T-k}^{T}\eta_t\left\|\hat{g}_t\right\|^2 + \frac{1}{k+1}\sum_{t=T-k}^{T}\left\langle\,\hat{z}_t,\, x_t - x_{T-k}\,\right\rangle$$

Observe that $kS_{k-1} = (k+1)S_k - f(x_{T-k})$. Combining this with the previous inequality yields

$$kS_{k-1} \;=\; kS_k + \big(S_k - f(x_{T-k})\big) \;\le\; kS_k + \frac{1}{2(k+1)}\sum_{t=T-k}^{T}\eta_t\left\|\hat{g}_t\right\|^2 + \frac{1}{k+1}\sum_{t=T-k}^{T}\langle\hat{z}_t,\, x_t - x_{T-k}\rangle.$$

Dividing by k, we obtain:

$$S_{k-1} \;\le\; S_k + \frac{1}{2k(k+1)}\sum_{t=T-k}^{T}\eta_t\left\|\hat{g}_t\right\|^2 + \frac{1}{k(k+1)}\sum_{t=T-k}^{T}\left\langle\,\hat{z}_t,\, x_t - x_{T-k}\,\right\rangle.$$

Thus, by induction:

$$\begin{aligned}
f(x_T) \;&=\; S_0\\
&\le\; S_{T/2} + \sum_{k=1}^{T/2}\frac{1}{2k(k+1)}\sum_{t=T-k}^{T}\eta_t\left\|\hat{g}_t\right\|^2 + \sum_{k=1}^{T/2}\frac{1}{k(k+1)}\sum_{t=T-k}^{T}\left\langle\,\hat{z}_t,\, x_t - x_{T-k}\,\right\rangle\\
&=\; \frac{1}{T/2+1}\sum_{t=T/2}^{T} f(x_t) + \sum_{k=1}^{T/2}\frac{1}{2k(k+1)}\sum_{t=T-k}^{T}\eta_t\left\|\hat{g}_t\right\|^2 + \sum_{k=1}^{T/2}\frac{1}{k(k+1)}\sum_{t=T-k}^{T}\left\langle\,\hat{z}_t,\, x_t - x_{T-k}\,\right\rangle.
\end{aligned}$$

Note that $\left\|\hat{g}_t\right\|^2 \le 4$ and $\eta_t = 1/t$. So we can bound the middle term as

$$\begin{aligned}
\sum_{k=1}^{T/2}\frac{1}{2k(k+1)}\sum_{t=T-k}^{T}\eta_t\left\|\hat{g}_t\right\|^2 &\le 2\sum_{k=1}^{T/2}\frac{1}{k(k+1)}\sum_{t=T-k}^{T}\frac{1}{t}\\
&\le 2\sum_{k=1}^{T/2}\frac{1}{k(T-k)}\\
&\le \frac{4}{T}\sum_{k=1}^{T/2}\frac{1}{k}\\
&= O\left(\frac{\log T}{T}\right).
\end{aligned}$$

This completes the proof. ■

E.1. Upper Bound on Error of Suffix Averaging

To complete the proof of the final iterate upper bound (Theorem 3.1), it still remains to prove the suffix averaging upper bound (Theorem 3.5). In this section, we prove this result as a corollary of the high probability bounds on $\left\|x_t - x^*\right\|^2$ that we obtained in the previous subsection.

Proof (of Theorem 3.5). By Lemma E.1 with $w = x^*$ we have

$$\sum_{t=T/2}^{T} [f(x_t) - f(x^*)] \;\le\; \underbrace{\frac{1}{2}\sum_{t=T/2}^{T} \eta_t \left\|\hat{g}_t\right\|^2}_{(a)} + \underbrace{\frac{1}{2\eta_{T/2}} \left\|x_{T/2} - x^*\right\|^2}_{(b)} + \underbrace{\sum_{t=T/2}^{T} \langle\, \hat{z}_t,\, x_t - x^* \,\rangle}_{(c)}. \quad (6)$$

It suffices to bound the right hand side of (6) by $O(\log(1/\delta))$ with probability at least $1-\delta$. Indeed, bounding $\left\|\hat{g}_t\right\|^2$ by 4, (a) in (6) is bounded by $O(1)$. Theorem 6.5 bounds (b) by $O(\log(1/\delta))$.

Theorem 6.5 implies $\sum_{t=T/2}^{T} \left\|x_t - x^*\right\|^2 = O(\log(1/\delta))$ with probability at least $1-\delta$. So, Corollary C.4 bounds (c) by $O(\log(1/\delta))$ with probability at least $1-\delta$. ■

E.2. Proof of Lemma 6.4

Proof (of Lemma 6.4). Recall from Section 6 that $\alpha_j = \frac{1}{(T-j)(T-j+1)}$ and $w_t = \sum_{j=T/2}^{t-1} \alpha_j (x_t - x_j)$.

Definition E.2 *Define* $B_T \coloneqq \sum_{t=T/2}^{T} \frac{1}{T-t+1} \sum_{j=T/2}^{t-1} \alpha_j \left\|x_t - x_j\right\|^2$.

Claim E.3 $\sum_{t=T/2}^{T} \left\|w_t\right\|^2 \le B_T$.

Proof Let $A_t = \sum_{j=T/2}^{t-1} \alpha_j$. Then

$$\begin{aligned}
\left\|w_t\right\|^2 &= A^2 \left\| \sum_{j=T/2}^{T-1} \frac{\alpha_j}{A} (x_t - x_j) \right\|^2 \\
&\le A^2 \sum_{j=T/2}^{T-1} \frac{\alpha_j}{A} \left\|x_t - x_j\right\|^2 \\
&\le \frac{1}{T-t+1} \sum_{j=T/2}^{t-1} \alpha_j \left\|x_t - x_j\right\|^2,
\end{aligned}$$

where the first inequality is due to the convexity of $\left\|\cdot\right\|^2$ and the second inequality is Claim A.12. ■

Lemma 6.3. Suppose f is 1-Lipschitz and 1-strongly convex. Suppose we run Algorithm 1 for T iterations with step sizes $\eta_t = 1/t$. Let $a < b$. Then,

$$\left\|x_a - x_b\right\|^2 \le \sum_{i=a}^{b-1} \frac{\left\|\hat{g}_i\right\|^2}{i^2} + 2\sum_{i=a}^{b-1} \frac{\big(f(x_a) - f(x_i)\big)}{i} + 2\sum_{i=a}^{b-1} \frac{\langle\, \hat{z}_i,\, x_i - x_a \,\rangle}{i}.$$

Proof (of Lemma 6.3).

$$\begin{aligned}
\|x_a - x_b\|^2 &= \|x_a - \Pi_{\mathcal{X}}(y_b)\|_2^2 \\
&\le \|x_a - y_b\|_2^2 \quad \text{(Claim A.8)} \\
&= \|x_a - x_{b-1} + x_{b-1} - y_b\|_2^2 \\
&= \|x_a - x_{b-1}\|_2^2 + \|x_{b-1} - y_b\|_2^2 + 2\langle\, \eta_{b-1}\hat{g}_{b-1},\, x_a - x_{b-1}\,\rangle \\
&= \|x_a - x_{b-1}\|_2^2 + \eta_{b-1}^2\,\|\hat{g}_{b-1}\|_2^2 + 2\langle\, \eta_{b-1}\hat{g}_{b-1},\, x_a - x_{b-1}\,\rangle \\
&= \|x_a - x_{b-1}\|_2^2 + \eta_{b-1}^2\,\|\hat{g}_{b-1}\|_2^2 + 2\langle\, \eta_{b-1}g_{b-1},\, x_a - x_{b-1}\,\rangle + 2\langle\, \eta_{b-1}\hat{z}_{b-1},\, x_{b-1} - x_a\,\rangle
\end{aligned}$$

Repeating this argument iteratively on $\|x_a - x_{b-1}\|, \|x_a - x_{b-2}\|, \ldots, \|x_a - x_{a+1}\|$, we obtain:

$$\|x_a - x_b\|^2 \le \sum_{i=a}^{b-1} \frac{\|\hat{g}_i\|_2^2}{i^2} + 2\sum_{i=a}^{b-1} \frac{\langle\, g_i,\, x_a - x_i\,\rangle}{i} + 2\sum_{i=a}^{b-1} \frac{\langle\, \hat{z}_i,\, x_i - x_a\,\rangle}{i}.$$

Applying the inequality $\langle\, g_i,\, x_a - x_i\,\rangle \le f(x_a) - f(x_i)$ to each term of the second summation gives the desired result. ∎

Using Lemma 6.3 and the bound $\|\hat{g}_t\|^2 \le 4$ for all t, let us write $B_T \le \Lambda_1 + \Lambda_2 + \Lambda_3$ where

$$\begin{aligned}
\Lambda_1 &:= 4\sum_{t=T/2}^{T} \frac{1}{T-t+1} \sum_{j=T/2}^{t-1} \alpha_j \sum_{i=j}^{t-1} \frac{1}{i^2}, \\
\Lambda_2 &:= 2\sum_{t=T/2}^{T} \frac{1}{T-t+1} \sum_{j=T/2}^{t-1} \alpha_j \sum_{i=j}^{t-1} \frac{(F_j - F_i)}{i} \qquad \text{(where } F_a := f(x_a) - f(x^*)\text{)}, \\
\Lambda_3 &:= 2\sum_{t=T/2}^{T} \frac{1}{T-t+1} \sum_{j=T/2}^{t-1} \alpha_j \sum_{i=j}^{t-1} \frac{\langle\, \hat{z}_i,\, x_i - x_j\,\rangle}{i}.
\end{aligned}$$

Let us bound each of the terms separately.

Claim E.4 $\Lambda_1 \le O\left(\frac{\log^2(T)}{T^2}\right)$.

Proof This follows from some straightforward calculations. Indeed,

$$\begin{aligned}
\Lambda_1 &= 4\sum_{t=T/2}^{T} \frac{1}{T-t+1} \sum_{j=T/2}^{t-1} \alpha_j \sum_{i=j}^{t-1} \frac{1}{i^2} \\
&\le 4\sum_{t=T/2}^{T} \frac{1}{T-t+1} \sum_{j=T/2}^{t-1} \frac{1}{(T-j)(T-j+1)} \frac{(T-j)}{(T/2)^2} \\
&\le \frac{4}{(T/2)^2} \sum_{t=T/2}^{T} \frac{1}{T-t+1} \sum_{j=T/2}^{t-1} \frac{1}{T-j+1} \\
&\le O\left(\frac{\log^2(T)}{T^2}\right).
\end{aligned}$$

∎

Claim E.5

$$\Lambda_2 \;\le\; O\Big(\frac{\log(T)}{T^2}\Big) + O\Big(\frac{\log(T)}{T}\Big)\,\|x_{T/2} - x^*\|_2^2 + O\Big(\frac{\log(T)}{T^2}\Big)\sum_{t=T/2}^{T-1} \langle\, \hat{z}_t,\, x_t - x^* \,\rangle.$$

We will prove Claim E.5 in the next section.

Claim E.6

$$\Lambda_3 \;=\; \sum_{i=T/2}^{T-1} \langle\, \hat{z}_i,\, \frac{C_i}{i} w_i \,\rangle,$$

where $C_i \coloneqq \sum_{\ell=i+1}^{T} \frac{2}{T-i+1} = O\big(\log(T)\big)$.

Proof Rearranging the order of summation in Λ_3 we get:

$$\begin{aligned}
\Lambda_3 \;&=\; \sum_{t=T/2}^{T} \frac{2}{T-t+1} \sum_{j=T/2}^{t-1} \alpha_j \sum_{i=j}^{t-1} \frac{\langle\, \hat{z}_i,\, x_i - x_j \,\rangle}{i} \\
&=\; \sum_{t=T/2}^{T} \frac{2}{T-t+1} \sum_{i=T/2}^{t-1} \frac{\langle\, \hat{z}_i,\, \sum_{j=T/2}^{i-1} \alpha_j (x_i - x_j) \,\rangle}{i} \\
&=\; \sum_{t=T/2}^{T} \frac{2}{T-t+1} \sum_{i=T/2}^{t-1} \frac{\langle\, \hat{z}_i,\, w_i \,\rangle}{i} \\
&=\; \sum_{i=T/2}^{T-1} \langle\, \hat{z}_i,\, \frac{\left(\sum_{t=i+1}^{T} \frac{2}{T-t+1}\right)}{i} w_i \,\rangle \\
&=\; \sum_{i=T/2}^{T-1} \langle\, \hat{z}_i,\, \frac{C_i}{i} w_i \,\rangle,
\end{aligned}$$

as desired. ■

The previous three claims and the fact that B_T is an upper bound on $\sum_{t=T/2}^{T} \|w_t\|^2$ (Claim E.3) complete the proof of Lemma 6.4. ■

E.3. Proof of Claim E.5

Let us rewrite

$$\Lambda_2 \;=\; \sum_{a=T/2}^{T-1} \gamma_a F_a$$

and determine the coefficients γ_a.

Claim E.7 *For each* $a \in \{\lfloor T/2 \rfloor, \ldots, T-1\}$, $\gamma_a \;=\; O\Big(\frac{\log(T)}{T^2}\Big)$.

Proof In the definition of Λ_2, the indices providing a positive coefficient for F_a must satisfy $j = a$, $i \leq a$, and $a \leq t-1$. Hence, the positive contribution to γ_a is:

$$\begin{aligned}
&\sum_{t=1+a}^{T} \frac{2}{T-t+1}\alpha_a \sum_{i=a}^{t-1} \frac{1}{i} \\
&\leq \sum_{t=1+a}^{T} \left(\frac{2}{T-t+1}\alpha_a\right)\left(\log\left(T/(a-1)\right)\right) \qquad \text{(by Claim A.14)} \\
&\leq \sum_{t=1+a}^{T} \left(\frac{2}{T-t+1}\alpha_a\right)\left(\frac{T-a+1}{a-1}\right) \qquad \text{(by Claim A.13)} \\
&= \sum_{t=1+a}^{T} \left(\frac{2}{T-t+1}\right)\left(\frac{1}{(T-a)(T-a+1)}\right)\left(\frac{T-a+1}{a-1}\right) \\
&= \frac{1}{T-a}\sum_{t=1+a}^{T} \frac{2}{(T-t+1)(a-1)}
\end{aligned}$$

The terms contributing to the negative portion of γ_a satisfy, $i = a$, $j \leq a$, and $a \leq t-1$. The negative contribution can be written as

$$\begin{aligned}
&-\sum_{t=1+a}^{T} \frac{2}{T-t+1} \sum_{j=T/2}^{a} \alpha_j \frac{1}{a} \\
&= -\sum_{t=1+a}^{T} \left(\frac{2}{T-t+1}\right)\left(\frac{1}{a}\right)\left(\frac{1}{T-a} - \frac{1}{T/2+1}\right) \\
&= -\sum_{t=1+a}^{T} \left(\frac{2}{T-t+1}\right)\left(\frac{1}{a}\right)\left(\frac{2a-T+2}{2(T/2+1)(T-a)}\right) \\
&= -\frac{1}{(T/2+1)(T-a)}\sum_{t=1+a}^{T} \left(\frac{2}{T-t+1}\right)\left(\frac{2a-T+2}{2a}\right) \\
&= -\frac{2}{(T+2)(T-a)}\sum_{t=1+a}^{T} \left(\frac{2}{T-t+1}\right)\left(1 - \frac{T-2}{2a}\right)
\end{aligned}$$

where on the last line we used $T-1 \le 2\lfloor T/2 \rfloor \le T$. Now, combining the positive and negative contribution we see:

$$\begin{aligned}
\gamma_a &\le \frac{1}{T-a}\sum_{t=1+a}^{T}\frac{2}{T-t+1}\left(\frac{1}{a-1}-\frac{2}{T+2}\left(1-\frac{T-2}{2a}\right)\right)\\
&= \frac{1}{T-a}\sum_{t=1+a}^{T}\frac{2}{T-t+1}\left(\frac{T+2-2\big(a-1\big)\big(1-\frac{T-2}{2a}\big)}{(a-1)(T+2)}\right)\\
&= \frac{1}{T-a}\sum_{t=1+a}^{T}\frac{2}{T-t+1}\left(\frac{T+2-2(a-1)+\frac{2(T-2)(a-1)}{2a}}{(a-1)(T+2)}\right)\\
&\le \frac{1}{T-a}\sum_{t=1+a}^{T}\frac{2}{T-t+1}\left(\frac{2\big(T-a\big)+2}{(T+2)(a-1)}\right)\\
&\le \frac{1}{T-a}\sum_{t=1+a}^{T}\frac{2}{T-t+1}\left(\frac{2\big(T-a\big)+2\big(T-a\big)}{(T+2)(a-1)}\right) \qquad (a \le T-1)\\
&= \frac{1}{(T+2)(a-1)}\sum_{t=1+a}^{T}\frac{4}{T-t+1}\\
&\le \frac{2}{(T+2)(T-2)}\sum_{t=1+a}^{T}\frac{4}{T-t+1} \qquad (a \ge T/2)\\
&= O\left(\frac{\log(T)}{T^2}\right),
\end{aligned}$$

as desired. ∎

Proof (of Claim E.5).

$$\begin{aligned}
\Lambda_2 &= \sum_{a=T/2}^{T-1}\gamma_a F_a\\
&\le O\left(\frac{\log(T)}{T^2}\right)\sum_{a=T/2}^{T-1} f(x_a)-f(x^*) \qquad \text{(by Claim E.7)}\\
&\le O\left(\frac{\log(T)}{T^2}\right)\left(\frac{1}{2}\sum_{t=T/2}^{T-1}\eta_t\left\|\hat{g}_t\right\|_2^2+\frac{1}{2\eta_{T/2}}\left\|x_{T/2}-x^*\right\|_2^2+\sum_{t=T/2}^{T-1}\left\langle\,\hat{z}_t,\,x_t-x^*\,\right\rangle\right) \qquad \text{(by Lemma E.1)}\\
&\le O\left(\frac{\log(T)}{T^2}\right)\sum_{t=T/2}^{T-1}\frac{1}{t}+O\left(\frac{\log(T)}{T}\right)\left\|x_{T/2}-x^*\right\|_2^2+O\left(\frac{\log(T)}{T^2}\right)\sum_{t=T/2}^{T-1}\left\langle\,\hat{z}_t,\,x_t-x^*\,\right\rangle \quad (\|\hat{g}_t\|_2 \le 2)\\
&\le O\left(\frac{\log(T)}{T^2}\right)+O\left(\frac{\log(T)}{T}\right)\left\|x_{T/2}-x^*\right\|_2^2+O\left(\frac{\log(T)}{T^2}\right)\sum_{t=T/2}^{T-1}\left\langle\,\hat{z}_t,\,x_t-x^*\,\right\rangle,
\end{aligned}$$

as desired. ∎

E.4. Proof of Claim 6.7

Proof (of Claim 6.7). We begin by stating two consequences of strong convexity:

1. $\langle g_t, x_t - x^* \rangle \geq f(x_t) - f(x^*) + \frac{1}{2}\|x_t - x^*\|^2$,
2. $f(x_t) - f(x^*) \geq \frac{1}{2}\|x_t - x^*\|^2 \qquad$ (since $0 \in \partial f(x^*)$).

The analysis proceeds as follows:

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &= \|\Pi_{\mathcal{X}}(x_t - \eta_t \hat{g}_t) - x^*\|^2 \\
&\leq \|x_t - \eta_t \hat{g}_t - x^*\|^2 \quad \text{(Claim A.8)} \\
&= \|x_t - x^*\|^2 - 2\eta_t \langle \hat{g}_t, x_t - x^* \rangle + \eta_t^2 \|\hat{g}_t\|^2 \\
&= \|x_t - x^*\|^2 - 2\eta_t \langle g_t, x_t - x^* \rangle + 2\eta_t \langle \hat{z}_t, x_t - x^* \rangle + \eta_t^2 \|\hat{g}_t\|^2 \\
&\leq \|x_t - x^*\|^2 - 2\eta_t \Big(f(x_t) - f(x^*) \Big) - \frac{1}{t}\|x_t - x^*\|^2 + 2\eta_t \langle \hat{z}_t, x_t - x^* \rangle + \eta_t^2 \|\hat{g}_t\|^2 \\
&\leq \left(1 - \frac{2}{t}\right)\|x_t - x^*\|^2 + 2\eta_t \langle \hat{z}_t, x_t - x^* \rangle + \eta_t^2 \|\hat{g}_t\|^2 \\
&= \left(\frac{t-2}{t}\right)\frac{Y_{t-1}}{t-1} + \frac{2}{t} U_{t-1}\sqrt{\frac{Y_{t-1}}{t-1}} + \frac{\|\hat{g}_t\|^2}{t^2}.
\end{aligned}$$

Recall that $\|\hat{g}_t\|^2 \leq 4$ because $\hat{z}_t \leq 1$ and f is 1-Lipschitz. Multiplying through by t and bounding $\|\hat{g}_t\|^2$ by 4 yields the desired result. ■

Appendix F. Generalizations

In this section, we discuss generalizations of our results. In Subsection F.1, we explain that the scaling of the function (e.g., Lipschitzness) can be normalized without loss of generality. In Subsection F.2, we explain how the assumption of almost surely bounded noise can be relaxed to sub-Gaussian noise in our upper bounds (Theorems 3.1 and 3.5).

F.1. Scaling assumptions

For most of this paper we consider only convex functions that have been appropriately normalized, due to the following facts.

- **Strongly convex case.** The case of an α-strongly convex and L-Lipschitz function can be reduced to the case of a 1-strongly convex and 1-Lipschitz function.
- **Lipschitz case.** The case of an L-Lipschitz function on a domain of diameter R can be reduced to the case of a 1-Lipschitz function on a domain of diameter 1.

We will discuss only the first of these in detail. The second is proven with similar ideas.

The main results from this section are as follows.

Theorem F.1 *Suppose f is α-strongly convex and L-Lipschitz, and that $\hat{z}_t$ has norm at most L almost surely. Consider running Algorithm 1 for T iterations with step size $\eta_t = \frac{1}{\alpha t}$. Let $x^* = \operatorname{argmin}_{x \in \mathcal{X}} f(x)$. Then, with probability at least $1 - \delta$,*

$$f(x_{T+1}) - f(x^*) \leq O\left(\frac{L^2}{\alpha}\frac{\log(T)\log(1/\delta)}{T}\right).$$

Theorem F.2 *Suppose f is α-strongly convex and L-Lipschitz, and that $\hat{z}_t$ has norm at most L almost surely. Consider running Algorithm 1 for T iterations with step size $\eta_t = \frac{1}{\alpha t}$. Let $x^* = \operatorname{argmin}_{x \in \mathcal{X}} f(x)$. Then, with probability at least $1 - \delta$,*

$$f\left(\frac{1}{T/2+1}\sum_{t=T/2}^{T} x_t\right) - f(x^*) \leq O\left(\frac{L^2}{\alpha}\frac{\log(1/\delta)}{T}\right).$$

We prove these theorems by reduction to Theorem 3.1 and Theorem 3.5, respectively. That is, suppose that f is a function that has strong convexity parameter α and Lipschitz parameter L. We construct a function g that is 1-Lipschitz and 1-strongly convex (using Claim F.4) and a subgradient oracle such that running SGD on g with this subgradient oracle is equivalent to running SGD on f. Formally, we show the following:

Claim F.3 *Suppose f is α-strongly convex and L-Lipschitz on a domain $\mathcal{X} \subset \mathbb{R}^n$. Let the initial point $x_1 \in \mathcal{X}$ be given. Let g be as defined in Claim F.4. Then, there is a coupling between the following two processes:*

- *the execution of Algorithm 1 on input f with initial point x_1, step size $\eta_t = 1/(\alpha t)$ and convex set $\mathcal{X}$*
- *the execution of Algorithm 1 on input g with initial point $\tilde{x}_1 := (\alpha/L)x_1$, step size $\tilde{\eta}_t = 1/t$ and convex set $(\alpha/L)\mathcal{X}$*

such that the iterates of the second process correspond to the iterates of the first process scaled by α/L. That is, if we denote by $\tilde{x}_t$ the iterates of the execution of SGD using g and x_t for the execution on f, then $\tilde{x}_t = (\alpha/L)x_t$.

Now, suppose we are given an α-strongly convex and L-Lipschitz function, f, an initial point x_1 and a convex set $\mathcal{X}$. We obtain Theorem F.1 and Theorem F.2 by performing the above coupling and executing SGD on the 1-Lipschitz and 1-strongly convex function. We may apply our high probability upper bounds to this execution of SGD because it satisfies the assumptions of Theorem 3.1 and Theorem 3.5. Finally, because of Claim F.3, we can reinterpret the iterates of the execution of SGD on g as a scaled version of the iterates of the execution of SGD on f. This immediately proves Theorem F.1 and Theorem F.2. Now, let us prove Claim F.3.

Proof (of Claim F.3). The coupling is given by constraining the algorithms to run in parallel and enforcing the execution of SGD on g to use a scaled version of the outputs of the subgradient oracle used by the execution of SGD on f. That is, at step t, if $\hat{g}_t$ is the output of the subgradient oracle of the execution of SGD on f, then we set the output of the subgradient oracle of the execution of SGD on g at step t to be $\frac{1}{L}\hat{g}_t$.

In order for this coupling to make sense, we have to ensure that this subgradient oracle for g is valid. That is, we must show that at each step, the subgradient oracle we define for g returns a true subgradient in expectation, and that the noise of this subgradient oracle is at most 1 with probability 1. We show by induction, that at each step $\tilde{x}_t = (\alpha/L)x_t$.

By definition, $\tilde{x}_1 = (\alpha/L)x_1$. Now, assume $\tilde{x}_t = (\alpha/L)x_t$. Let $\hat{g}_t$ be the output of the subgradient oracle for SGD running on f. The subdifferential for g at $\tilde{x}_t$ is $\frac{1}{L}\partial f(x_t)$ using the chain rule for subdifferentials. Therefore, the subgradient oracle for g is certainly valid at this step. Now, $y_{t+1} = x_t - \frac{1}{\alpha t}\hat{g}_t$. Meanwhile, $\tilde{y}_{t+1} = \tilde{x}_t - \frac{1}{t}\frac{1}{L}\hat{g}_t = \frac{\alpha}{L}(x_t - \frac{1}{\alpha t}\hat{g}_t) = \frac{\alpha}{L}y_{t+1}$. Therefore,

$$\tilde{x}_{t+1} = \Pi_{(\alpha/L)\mathcal{X}}(\tilde{y}_{t+1}) = \Pi_{(\alpha/L)\mathcal{X}}(y_{t+1}(\alpha/L)) = (\alpha/L)\Pi_{\mathcal{X}}(y_{t+1}) = (\alpha/L)x_{t+1}$$

as desired. ■

Claim F.4 *Let f be an α-strongly convex and L-Lipschitz function. Then, $g(x) \coloneqq \frac{\alpha}{L^2} f(\frac{L}{\alpha}x)$ is 1-Lipschitz and 1-strongly convex.*

Proof First we show that g is 1-Lipschitz:

$$|g(x) - g(y)| \;=\; \frac{\alpha}{L^2}\left|f\left(\frac{L}{\alpha}x\right) - f\left(\frac{L}{\alpha}y\right)\right| \;\leq\; \frac{\alpha}{L^2}L\left\|\frac{L}{\alpha}(x-y)\right\| \;=\; \|x-y\|\,.$$

The inequality holds since f is L-Lipschitz.

Now we show that g is 1-strongly convex. A function h is α strongly convex, if and only if the function $x \mapsto h(x) - \frac{\alpha}{2}\|x\|^2$ is convex. Indeed, for g:

$$g(x) - \frac{1}{2}\|x\|^2 \;=\; \frac{\alpha}{L^2}f\left(\frac{L}{\alpha}x\right) - \frac{1}{2}\|x\|^2 \;=\; \frac{\alpha}{L^2}\left(f\left(\frac{L}{\alpha}x\right) - \frac{L^2}{2\alpha}\|x\|^2\right) \;=\; \frac{\alpha}{L^2}\left(f\left(\frac{L}{\alpha}x\right) - \frac{\alpha}{2}\left\|\frac{L}{\alpha}x\right\|^2\right).$$

The function on the right is convex because f is α-strongly convex. This implies that $x \mapsto g(x) - \frac{1}{2}\|x\|^2$ is convex, meaning that g is 1-strongly convex. ■

F.2. Sub-Gaussian Noise

In this section, we relax the assumption that $\|\hat{z}_t\| \leq 1$ with probability 1 and instead assume that for each t, $\|\hat{z}_t\|_2$ is sub-Gaussian conditioned on $\mathcal{F}_{t-1}$. The proof of the extensions are quite easy, given the current analyses. See the full version (Harvey et al., 2018) of our paper for statements and proofs of this extension.

Main ideas. Most of our analyses can remain unchanged. The main task at hand is identifying the places where we use the upper bound $\|\hat{z}_t\| \leq 1$ outside of the MGF analyses (using this bound inside an MGF is morally the same using the fact that $\|\hat{z}_t\|$ is sub-Gaussian). The main culprit is that we often bound $\|\hat{g}_t\|^2$ by 4. Instead we must carry these terms forward and handle them using MGFs. The consequences of this are two-fold. Firstly, this introduces new MGFs to bound, but intuitively these are easy to bound because the terms they were involved in in the original analysis were sufficiently bounded and therefore their MGFs should now also be sufficiently bounded. Furthermore, removing these constant bounds results in many of our MGF expressions to include more random terms which we previously ignored and pulled out of our MGF arguments because they were constant. But again, these terms can be dealt with by first isolating them by applying an MGF triangle inequality (using Hölder or Cauchy-Schwarz) and then bounding their MGF.

Appendix G. Necessity of $\log(1/\delta)$

In this section, we show that the error of the last iterate and suffix average of SGD is $\Omega(\log(1/\delta)/T)$ with probability at least δ.

Lemma G.1 ((Klein and Young, 2015, Lemma 4)) *Let $X_1, \ldots, X_T$ be independent random variables taking value $\{-1, +1\}$ uniformly at random and $X = \frac{1}{T}\sum_{t=1}^{T} X_i$. Then for any $0 < c < O(\sqrt{T})$,*

$$\Pr\left[X \geq \frac{c}{\sqrt{T}}\right] \geq \exp(-9c^2/2).$$

Consider the single-variable function $f(x) = \frac{1}{2}x^2$ and suppose that the domain is $\mathcal{X} = [-1, 1]$. Then f is 1-strongly convex and 1-Lipschitz on $\mathcal{X}$. Moreover, suppose that the subgradient oracle returns $x - \hat{z}$ where $\hat{z}$ is -1 or $+1$ with probability $1/2$ (independently from all previous calls to the oracle). Finally, suppose we run Algorithm 1 with step sizes $\eta_t = 1/t$ with an initial point $x_1 = 0$.

Claim G.2 *If $T \geq O(\log(1/\delta))$ then $f(x_{T+1}) \geq \Omega(\log(1/\delta)/T)$ with probability at least δ.*

Proof We claim that $x_{t+1} = \frac{1}{t}\sum_{i=1}^{t} \hat{z}_i$ for all $t \in [T]$ where $\hat{z}_i$ is the random sign returned by the subgradient oracle at iteration i. Indeed, for $t = 1$, we have $y_2 = x_1 - \eta_1(x_1 - \hat{z}_1) = \hat{z}_1$ since $\eta_1 = 1$. Moreover, $x_2 = \Pi_{\mathcal{X}}(y_2) = y_2$ since $|y_2| \leq 1$. Now, suppose that $x_t = \frac{1}{t-1}\sum_{i=1}^{t-1} \hat{z}_i$. Then $y_{t+1} = x_t - \eta_t(x_t - \hat{z}_t) = \frac{1}{t}\sum_{i=1}^{t} \hat{z}_i$. Since $|y_{t+1}| \leq 1$, we have $x_{t+1} = y_{t+1}$.

Hence, by Lemma G.1 with $c = \sqrt{\log(1/\delta)}$, we have $x_{T+1} \geq \sqrt{\log(1/\delta)}/\sqrt{T}$ with probability at least $\Omega(\delta)$ (provided $T \geq O(\log(1/\delta))$). We conclude that $f(x_{T+1}) \geq \frac{\log(1/\delta)}{2T}$ with probability at least $\Omega(\delta)$. ■

We can also show that Theorem 3.5 is tight. To make the calculations simpler, first assume T is a multiple of 4. We further assume that the noise introduced by the stochastic subgradient oracle is generated as follows. For $1 \leq t < T/2$ and $t > 3T/4$, $\hat{z}_t = 0$. For $T/2 \leq t \leq 3T/4$, first define $A_t = \sum_{i=t}^{T} \frac{1}{i}$. Then we set $\hat{z}_t$ to be $\pm\frac{1}{4A_t}$ with probability $1/2$. Note that $A_t \geq 1/4$ for $T/2 \leq t \leq 3T/4$ so we still have $|\hat{z}_t| \leq 1$ for all t.

Claim G.3 *If $T \geq O(\log(1/\delta))$ then $f\left(\frac{1}{T/2+1}\sum_{t=T/2+1}^{T+1} x_t\right) \geq \Omega\left(\frac{\log(1/\delta)}{T}\right)$ with probability at least δ.*

Proof Proceeding as in the above claim, we have $x_{t+1} = \frac{1}{t}\sum_{i=1}^{t} \hat{z}_i$. We claim that

$$\frac{1}{T/2+1}\sum_{t=T/2+1}^{T+1} x_t = \frac{1}{T/2+1}\sum_{t=T/2}^{3T/4} A_t\hat{z}_t. \tag{7}$$

To see this, we have

$$\begin{aligned}
\frac{1}{T/2+1}\sum_{t=T/2}^{T} x_{t+1} &= \frac{1}{T/2+1}\sum_{t=T/2}^{T}\frac{1}{t}\sum_{i=1}^{T}\hat{z}_i \\
&= \frac{1}{T/2+1}\sum_{i=1}^{T}\hat{z}_i\sum_{t=\max\{i,T/2\}}^{T}\frac{1}{t} \\
&= \frac{1}{T/2+1}\sum_{t=T/2}^{3T/4} A_t\hat{z}_t,
\end{aligned}$$

where the last equality uses the assumption that $\hat{z}_t \neq 0$ only if $T/2 \leq t \leq 3T/4$ and changes the name of the index. Notice that $A_t\hat{z}_t$ is $\pm\frac{1}{4}$ with probability $1/2$ so we can write Eq. (7) as

$$\frac{1}{4(T/2+1)}\sum_{t=1}^{T/4+1} X_t$$

where X_t are random signs. Applying Lemma G.1 with $c = \sqrt{\log(1/\delta)}$, we conclude that Eq. (7) is at least $\Omega(\sqrt{\log(1/\delta)}/\sqrt{T})$ with probability at least $\Omega(\delta)$ (provided $T \geq O(\log(1/\delta))$). So we conclude that $f\left(\frac{1}{T/2+1}\sum_{t=T/2+1}^{T+1} x_t\right) \geq \Omega\left(\frac{\log(1/\delta)}{T}\right)$ with probability at least $\Omega(\delta)$. ■