Understanding the role of averaging in non-smooth stochastic gradient descent

by

Sikander Randhawa

B. Sc, University of British Columbia, 2018

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Computer Science)

The University of British Columbia
(Vancouver)

August 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Understanding the role of averaging in non-smooth stochastic gradient descent

submitted by **Sikander Randhawa** in partial fulfillment of the requirements for the degree of **Master of Science** in **Computer Science**.

Examining Committee:

Nicholas J. A. Harvey, Computer Science
Supervisor

Bruce Shepherd, Computer Science
Supervisory Committee Member

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 (SGD), 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 [42].

We prove analogous results for functions which are Lipschitz and convex, but not necessarily strongly convex or differentiable. After T steps of stochastic gradient descent, the error of the final iterate is $O(\log(T)/\sqrt{T})$ *with high probability*, and there exists a function for which the error of the final iterate of *deterministic* gradient descent is $\Omega(\log(T)/\sqrt{T})$.

In the strongly-convex setting, several forms of SGD, including suffix averaging, are known to achieve the optimal $O(1/T)$ convergence rate *in expectation*. An intermediate step of our high probability analysis for the error of the final iterate 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. [36] and Hazan and Kale [17], 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. This is the first known high-probability result which attains the optimal $O(1/T)$ rate.

We also consider a simple, non-uniform averaging strategy of Lacoste-Julien et al. [26] and prove that it too achieves the optimal $O(1/T)$ convergence rate *with high probability*. This provides a second algorithm which achieves the optimal $O(1/T)$ convergence rate with high-probability. Our high-probability results are proven using a generalization of Freedman's Inequality which we develop.

Lay Summary

Many machine learning tasks are represented as optimization problems where the objective is to produce a minimizer of some function. Often, what makes this task difficult is that it is very slow to compute the function value at any point. Nonetheless, there are simple procedures which provide non-trivial theoretical guarantees on finding an approximate minimizer. One such algorithm is *gradient descent*, which makes incremental progress towards a minimizer by taking small steps in the "steepest downhill direction." If the shape of the function does not change rapidly, gradient descent guarantees progress at every step. Otherwise, some steps can *increase* the function value; this is the setting we consider. We study the function value of different combinations of the points visited by gradient descent and compare them to the value of the minimizer. Understanding this is of practical significance because gradient descent is a fundamental algorithm for many machine learning tasks.

Preface

The main body of work presented in this thesis was largely conducted in collaboration with Nicholas Harvey, Christopher Liaw, and Yaniv Plan and was accepted for publication [15] at the Conference of Learning Theory 2019 [Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, Sikander Randhawa. Tight analyses of non-smooth stochastic gradient descent. In *Conference on Learning Theory*, 2019.]. The majority of the results and proofs found in Chapter 2, Chapter 3, Chapter 4, Section 5.2 can be found in [15]. The proofs of these results were developed as a result of collaborative effort and the writing was shared equally amongst authors. Some exceptions include the proofs in Section 3.4 and Section 3.5 which are currently unpublished and were developed and written by me. The statements of results from Chapter 5 and Chapter 6 can be found in [15] without proof. We provide full proofs in this thesis. I wrote Section 5.1 in Chapter 5, derived and wrote the results in Section 6.3, and wrote Section 6.2. The writing responsibilities in other sections of these chapters were shared amongst collaborators.

Table of Contents

Acknowledgments

First and foremost I would like to express a deep sense of gratitude to my supervisor, Nick Harvey. I am extremely grateful and consider myself lucky to have been able to learn from him. I would not be in the position I am in now without the patience, guidance, and optimism of Nick. I would also like to thank Hu Fu and Bruce Shepherd for being wonderful mentors to me.

To my Mom, Dad, Brother, Grandmother and my Uncle Sam: without your consistent love, support, sacrifice and encouragement there is not much that I would be able to accomplish. Thank you for raising me and giving me everything when you had nothing. To my closest friends, Angad Kalra and Harish Anand, thank you for always inspiring me to grow as a human and for the countless laughs and good times which we have shared.

Dedication

This thesis is dedicated to my Mom, Dad, and Brother.

Chapter 1

Introduction

1.1 Introduction

Stochastic gradient descent (SGD) is one of the oldest randomized algorithms, dating back to 1951 [37]. It is a very simple and widely used iterative method for minimizing a function. 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 both as an essential tool for applied machine learning [21, 40], and as a versatile framework for theoretical algorithm design.

In theoretical algorithms, SGD often appears in the guise of *coordinate descent*, an important special case in which each gradient estimate has a single non-zero coordinate. Some of the fast algorithms for Laplacian linear systems [23, 28] are based on coordinate descent (and the related Kaczmarz method [44]). Multi-armed Bandits were discovered years ago to be a perfect setting for coordinate descent [2]: the famous Exp3 algorithm combines coordinate descent and the multiplicative weight method. Recent work on the geometric median problem [9] gave a sublinear time algorithm based on SGD, and very recently a new privacy amplification technique [12] has been developed that injects noise to the subgradients while executing SGD. Surveys and monographs discussing gradient descent and aimed at a theoretical CS audience include Bansal and Gupta [4], Bubeck [7], Hazan [16], and Vishnoi [46].

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 [32] 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 thesis. In theoretical algorithms and discrete optimization, the convex functions that arise are often non-smooth. For example, the objective for the geometric median problem is a (sum of) 2-norms [9], so Lipschitz but not smooth. Similarly, formulating the minimum *s-t* cut problem as convex minimization [29], the objective is a 1-norm, so Lipschitz but not smooth. In machine learning, the

objective for regularized support vector machines [41] 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 [33], 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 [33], which is optimal, whereas for functions that are also strongly convex [18, 36] 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 [17]; 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 $O(1/T)$ [36], 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 [27, 43], although the solutions may be less interpretable.

Shamir [42] 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 [43], 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 [43] 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.

Our work shows that the $\log(T)$ factor is necessary, 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 [42]. 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. [20] 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 in expectation and with high probability. 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. A common misconception is that picking the best of several independent trials of SGD would yield high-probability bounds, but this approach is not as efficient as it might seem[1]. So it is both interesting and useful that high-probability bounds hold for a single execution of SGD. Some known high-probability bounds for the strongly convex setting include [22], for uniform averaging, and [17, 36], which give

[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.

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 for strongly convex functions: $O(1/T)$ 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. [36, §6], and the latter resolves the second question of Shamir [42]. For Lipschitz functions, we prove a high probability bound of $O(\log(T)/\sqrt{T})$ for the final iterate, which is also tight.

Our work can also be seen as extending a line of work on understanding the difference between an *average* of the iterates or the last iterate of an iterative process. For instance, an important result in game theory is that the multiplicative weights update algorithm converges to an equilibrium [14], i.e. the set of players are required to play some sort of "coordinated average" of their past strategies. Recently, [3] studied the convergence behaviour of players' individual strategies and found that the strategies *diverge* and hence, coordination (i.e. averaging) is needed to obtain an equilibrium. In a similar spirit, our work shows that the iterates of gradient descent have a sub-optimal convergence rate, at least for non-smooth convex functions, and thus, some form of averaging is needed to achieve the optimal rate. It is an interesting direction to see whether or not this is necessary in other iterative methods as well. For instance, the multiplicative weights update algorithm can be used to give an iterative algorithm for maximum flow [8], or linear programming in general [1, 34], but also requires some form of averaging. We hope that this thesis contributes to a better understanding on when averaging is necessary in iterative processes.

1.2 Preliminaries

Let $\mathscr{X}$ be a closed, convex subset of $\mathbb{R}^n$, $f\colon \mathscr{X} \to \mathbb{R}$ be a convex function, and $\partial f(x)$ the subdifferential of f at x. Our goal is to solve the convex program $\min_{x\in\mathscr{X}} f(x)$. (Our later assumptions will imply that $\min_{x\in\mathscr{X}} f(x)$ is realized by some point 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 $\mathscr{X}$ is represented by a projection oracle, which returns the point in $\mathscr{X}$ closest in Euclidean norm to a given point x. We say that f is α-strongly convex if

$$f(y) \;\geq\; f(x) + \langle g, y-x\rangle + \frac{\alpha}{2}\|y-x\|^2 \quad \forall y,x \in \mathscr{X}, g \in \partial f(x). \tag{1.1}$$

Throughout this thesis, $\|\cdot\|$ 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 $x \in \mathscr{X}$ and $g \in \partial f(x)$. For the remainder of this thesis, 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 Section 6.1). 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 subgaussian (see Section 6.1).

Let $\Pi_{\mathscr{X}}$ denote the projection operator on $\mathscr{X}$, which is defined by $\Pi_{\mathscr{X}}(y) = \arg\min_{x\in\mathscr{X}} \|x-y\|$. The (projected) stochastic gradient algorithm is given in Algorithm 1. Notice that 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 [35, 39] or averaging only a fraction of the final iterates [36]. 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 [32, 36], and $\eta_t = \Theta(1/\sqrt{t})$ for Lipschitz

functions. For our analyses, we will use a step size of $\eta_t = 1/t$ for strongly convex functions and $\eta_t = 1/\sqrt{t}$ for Lipschitz functions.

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

procedure STOCHASTICGRADIENTDESCENT($\mathscr{X} \subseteq \mathbb{R}^n$, $x_1 \in \mathscr{X}$, 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)
    $x_{t+1} \leftarrow \Pi_{\mathscr{X}}(y_{t+1})$ (project y_{t+1} onto the set $\mathscr{X}$)
  return either $\begin{cases} x_{T+1} & \text{(final iterate)} \\ \frac{1}{T+1}\sum_{t=1}^{T+1} x_t & \text{(uniform averaging)} \\ \frac{1}{T/2+1}\sum_{t=T/2+1}^{T+1} x_t & \text{(suffix averaging)} \end{cases}$

1.2.1 Preliminaries on martingales

Let $\{d_i, \mathscr{F}_i\}_{i=1}^n$ be such that d_i are $\mathscr{F}_i$ measurable random variables where $\{\mathscr{F}_i\}_{i=1}^n$ forms a filtration of a sigma field $\mathscr{F}$, $\mathrm{E}[|d_i|] < \infty$ and $\mathrm{E}[d_i \mid \mathscr{F}_i] = 0$. We call $\{d_i, \mathscr{F}_i\}_{i=1}^n$ a *martingale difference sequence.* The partial sums of a martingale difference sequence form a *martingale*.

In order to control the tail distribution of a martingale, one needs to make some assumptions on the behavior of the increments, d_i. A common assumption which is used in the classical Hoeffding's inequality is that $|d_i| \leq c_i$ almost surely where $c_i \in \mathbb{R}$. Then, Hoeffding's inequality yields

$$\Pr\left[\sum_{i=1}^n d_i \geq x\right] \leq \exp\left(-\frac{x^2}{2\sum_{i=1}^n c_i^2}\right), \qquad \forall x > 0.$$

Therefore, Hoeffding's inequality is useful if one can derive an *almost sure* bound on the sum of squared magnitudes by a scalar, $\sum_{i=1}^n c_i^2$. A weaker assumption which allows Hoeffding's inequality to be applicable for a much broader class of martingales is that the increments are conditionally subgaussian[2]. In this setting, one can use Hoeffding's inequality when the sum of squared subgaussian norms is *almost surely* bounded by a scalar.

The classical Freedman's inequality is a martingale concentration inequality which has the advantage that one need not require an almost sure bound on the sum of squared magnitudes. Freedman's inequality can roughly be stated as follows:

$$\Pr\left[\sum_{i=1}^n d_i \geq x \text{ and } B_n \leq y\right] \leq \exp\left(-\frac{Cx^2}{x+y}\right) \qquad \forall x, \beta > 0,$$

where B_t is some process derived from $\sum_{i=1}^t d_i$. Notice that a *high probability* bound of y on B_n yields a high probability bound of $\sqrt{y}$ on $\sum_{i=1}^n d_i$, whereas Hoeffdings inequality requires an *almost sure* bound of $\sum_{i=1}^n c_i^2$ on the sum of squared magnitudes in order to achieve a high probability bound of $\sqrt{\sum_{i=1}^n c_i^2}$ on $\sum_{i=1}^n d_i$. B_t can be defined in many ways and there are various forms of Freedman's inequality which use different selections of B_t.

[2] A random variable is subgaussian if its tail distribution resembles the gaussian tail distribution. For a more detailed description see Section 6.2.

Some of these processes, together with an accompanying version of Freedman's inequality using that process follow below:

Definition 1.1. *Let $\{d_i, \mathscr{F}_i\}_{i=1}^n$ be a martingale difference sequence. Let $var_{i-1} = \mathrm{E}\left[d_i^2 \mid \mathscr{F}_{i-1}\right]$. Let $Var_t = \sum_{i=1}^t var_{i-1}$. The predictable process $(Var_t)_{t=1}^n$ is called the* sum of conditional variances *(SCV).*

Theorem 1.2 ([13, Theorem 1.6])**.** *Let $\{d_i, \mathscr{F}_i\}_{i=1}^n$ be a martingale difference sequence. Assume that $|d_i| \leq c$ for every i almost surely. Then, for every $x, y > 0$*

$$\Pr\left[\bigcup_{i=1}^n \left\{\sum_{i=1}^t d_i \geq x \text{ and } Var_t \leq y\right\}\right] \leq \exp\left(-\frac{c^2x^2}{2(cx+y)}\right).$$

Remark 1.3. *de la Peña [10] later proved a similar result which relaxes the almost sure bound on the increments and instead assumes a Bernstein style conditional bound on the moments of d_i. Theorem 1.2 yields a high probability bound of $\sqrt{y}$ on $\sum_{i=1}^n d_i$ if one can prove a high probability bound of y on the SCV process.*

Definition 1.4. *Let $\{d_i, \mathscr{F}_i\}_{i=1}^n$ be a martingale difference sequence. Let v_{i-1} be $\mathscr{F}_{i-1}$ measureable and minimal such that for all $\lambda > 0$ we have $\mathrm{E}[\exp(\lambda d_i) \mid \mathscr{F}_{i-1}] \leq \exp\left(\frac{\lambda^2}{2} v_{i-1}\right)$. Let $V_t = \sum_{i=1}^t v_{i-1}$. The predictable process $(V_t)_{t=1}^n$ is called the* sum of squared conditional subgaussian norms *(SSCSN).*

Theorem 1.5 ([11, Theorem 2.6])**.** *Let $\{d_i, \mathscr{F}_i\}_{i=1}^n$ be a martingale difference sequence. Assume, for $1 \leq i \leq n$, there exist $\mathscr{F}_{i-1}$ measureable random variables v_{i-1} such that for all $\lambda > 0$ we have $\mathrm{E}[\exp(\lambda d_i) \mid \mathscr{F}_{i-1}] \leq \exp\left(\frac{\lambda^2}{2} v_{i-1}\right)$. Then, for every $x, y > 0$*

$$\Pr\left[\bigcup_{i=1}^n \left\{\sum_{i=1}^t d_i \geq x \text{ and } V_t \leq y\right\}\right] \leq \exp\left(-\frac{x^2}{2y}\right).$$

Remark 1.6. *Notice that Theorem 1.5 does not require an almost sure bound on the increments d_i. Instead, it assumes subgaussian increments. Theorem 1.5 yields a high probability bound of $\sqrt{y}$ if one can prove a high probability bound of y on the SSCSN process.*

It may be challenging to obtain an a high probability bound on the SCV or SSCSN processes. Instead, it is sometimes useful to consider the following process. Indeed, by Lemma 1.8 the *sum of squared magnitudes* (SSCM) process bounds both the SCV process and the SSCSN process. For a statement of a concentration inequality similar to Freedman's inequality which uses a process similar to the SSCM process, see [31, Theorem 3.14].

Definition 1.7. *Let $\{d_i, \mathscr{F}_i\}_{i=1}^n$ be a martingale difference sequence. Suppose that $|d_i| \leq m_{i-1}$ where m_{i-1} is $\mathscr{F}_{i-1}$ measureable. Let $M_t = \sum_{i=1}^t m_{i-1}^2$. The predictable process $(M_t)_{t=1}^n$ is called the* sum of squared conditional magntitudes *(SSCM).*

Lemma 1.8. *Suppose $|d_i| \leq m_{i-1}$ where m_{i-1} is $\mathscr{F}_{i-1}$ measurable. Then, $var_{i-1} \leq m_{i-1}$ and $v_{i-1} \leq m_{i-1}^2$. Consequently, $Var_t \leq M_t$ and $V_t \leq M_t$ for all $t \in [n]$.*

Proof (of Lemma 1.8). The first part of the claim follows because $\mathrm{E}[\mathrm{var}_{i-1} \mid \mathscr{F}_{i-1}] = \mathrm{E}\left[d_i^2 \mid \mathscr{F}_{i-1}\right] \leq m_{i-1}^2$, since $|d_i| \leq m_{i-1}$ which is $\mathscr{F}_{i-1}$ measurable. The second part of the claim follows because $\mathrm{E}[\exp(\lambda d_i) \mid \mathscr{F}_{i-1}] \leq \exp\left(\frac{\lambda^2}{2} m_{i-1}^2\right)$ by Hoeffding's Lemma (Lemma A.5). ■

In this thesis we derive a generalization of Freedman's inequality (as formulated in Theorem 1.5). While Theorem 1.5 is useful in the setting where one can obtain a high probability bound on the SSCSN process by a *scalar*, our Theorem 1.11 can be used with a high probability upper bound on the SSCSN by a *random* quantity with a particular structure. This random quantity contains a scaled and translated version of the original martingale – and we will refer to this entanglement between the SSCSN process and the original martingale as the *chicken and egg phenomenon.* Theorem 1.11 is used to derive a high probability bound on the error of the final iterate of Algorithm 1. While Theorem 1.11 requires one to control the SSCSN process, our application of Theorem 1.11 begins by bounding the SSCM process. Lemma 1.8 implies that this is sufficient.

1.3 Our contributions

Our main results are bounds on the error of the final iterate of stochastic gradient descent for non-smooth, convex functions.

Strongly convex and Lipschitz functions. We prove an $\Omega(\log(T)/T)$ lower bound, even in the non-stochastic case, and an $O(\log(T)\log(1/\delta)/T)$ upper bound with probability $1-\delta$.

Lipschitz functions. We prove an $\Omega(\log(T)/\sqrt{T})$ lower bound, even in the non-stochastic case, and an $O(\log(T)\log(1/\delta)/\sqrt{T})$ upper bound with probability $1-\delta$.

1.3.1 High probability upper bounds

Theorem 1.9. *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 $x^* = \arg\min_{x\in\mathscr{X}} f(x)$ and $\delta \in (0,1)$ be arbitrary. Then, with probability at least $1-\delta$,*

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

Theorem 1.10. *Suppose f is 1-Lipschitz and $\mathscr{X}$ has diameter 1. 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 $\eta_t = 1/\sqrt{t}$. Let $x^* = \arg\min_{x\in\mathscr{X}} f(x)$ and $\delta \in (0,1)$ be arbitrary. Then, with probability at least $1-\delta$,*

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

The assumptions on the strong convexity parameter, Lipschitz parameter, and diameter are without loss of generality; see Section 6.1. 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 subgaussian condition; see Section 6.3. We also remark that a linear dependence on $\log(1/\delta)$ is necessary for strongly convex functions; see Section 5.2.

Our main probabilistic tool to prove Theorem 1.9 and Theorem 1.10 is a new extension of the classic Freedman inequality [13] to a setting in which the martingale exhibits a curious phenomenon. Ordinarily a martingale is roughly bounded by the square root of the sum of squared magnitudes (SSCM) (this is the content of Freedman's inequality). We consider a setting in which the SSCM is itself bounded by (a linear transformation of) the martingale. We refer to this as a "chicken and egg" phenomenon.

Theorem 1.11 (Generalized Freedman). *Let $\{d_i, \mathscr{F}_i\}_{i=1}^n$ be a martingale difference sequence. Suppose v_{i-1}, for $i \in [n]$, are positive and $\mathscr{F}_{i-1}$-measurable random variables such that $\mathrm{E}[\exp(\lambda d_i) \mid \mathscr{F}_{i-1}] \le \exp\left(\frac{\lambda^2}{2} v_{i-1}\right)$ 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 $\alpha = \max_{i \in [n]} \alpha_i$. Then*

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

The proof of Theorem 1.11 appears in Section 4.1. Freedman's Inequality [13] (as formulated in Theorem 1.5, up to constants) simply omits the terms highlighted in yellow, i.e., it sets $\alpha = 0$. Observe that Theorem 1.11 assumes subgaussian increments as in [11] (as opposed to [10, 13] which assumes bounded increments).

1.3.2 Lower bounds

Theorem 1.12. *For any T and any constant $c > 0$, there exists a convex function $f_T : \mathscr{X} \to \mathbb{R}$, where $\mathscr{X}$ 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 $x^* = \arg\min_{x \in \mathscr{X}} f_T(x)$. Then*

$$f_T(x_T) - f_T(x^*) \;\ge\; \frac{\log T}{4c \cdot T} \tag{1.2}$$

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

$$f_T(\bar{x}) - f_T(x^*) \;\ge\; \frac{\ln(T) - \ln(k)}{4c \cdot T}. \tag{1.3}$$

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

Remark 1.13. *Let $L = (3/c)$ and $\alpha = (1/c)$. Then, the lower bound from Eq. (1.2) can be re-written as $\frac{L^2}{36\alpha}\frac{\log T}{T}$. This is within a constant factor of the guaranteed upper bound of $\frac{17L^2}{\alpha}\frac{1+\log T}{T}$ by Shamir and Zhang [43].*

Theorem 1.14. *For any T and any constant $c > 0$, there exists a convex function $f_T : \mathscr{X} \to \mathbb{R}$, where $\mathscr{X}$ is the unit Euclidean ball in $\mathbb{R}^T$, such that f_T is $(1/c)$-Lipschitz, and satisfies the following. Suppose that Algorithm 1 is executed from the initial point $x_1 = 0$ with step sizes $\eta_t = c/\sqrt{t}$ with $c > 0$. Let $x^* = \arg\min_{x \in \mathscr{X}} f_T(x)$. Then*

$$f_T(x_T) - f_T(x^*) \;\ge\; \frac{\log T}{32c\sqrt{T}}. \tag{1.4}$$

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

$$f_T(\bar{x}) - f_T(x^*) \;\ge\; \frac{\ln(T) - \ln(k)}{32c\sqrt{T}}. \tag{1.5}$$

Furthermore, the value of f strictly monotonically increases *for the first T iterations:*

$$f(x_{i+1}) \;\ge\; f(x_i) + \frac{1}{64c\sqrt{T}(T-i+1)} \qquad \forall i \in [T]. \tag{1.6}$$

Remark 1.15. *Let $L = (1/c)$ and $R = 1$. Then, the lower bound from Eq. (1.4) can be written as $(R/c +$*

$cL^2)\frac{\log T}{64\sqrt{T}}$. *This is within a constant factor of the guaranteed upper bound of* $(R/c+cL^2)\frac{2+\log T}{\sqrt{T}}$ *by Shamir and Zhang [43].*

Remark 1.16. *In order to incur a* $\log T$ *factor in the error of the* T *iterate, Theorem 1.12 and Theorem 1.14 constructs a function* f_T *parameterized by* T*. It is also possible to create a single function* f*,* independent *of* T*, which incurs an additional factor very slightly below* $\log T$ *for infinitely many* T*. This is described in Theorem 5.1 and Theorem 5.2.*

1.3.3 High probability upper bound for suffix averaging

Interestingly, our proof of Theorem 1.9 requires understanding the suffix average. (In fact this connection is implicit in [43]). 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 1.17. *Suppose* f *is* 1*-strongly convex and* 1*-Lipschitz. Consider running Algorithm 1 for* T *iterations with step size* $\eta_t = 1/t$ *and assume* T *is even. Let* $x^* = \arg\min_{x\in\mathscr{X}} f(x)$ *and* $\delta \in (0,1)$ *be arbitrary. Then, with probability at least* $1-\delta$*,*

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

Remark 1.18. *This upper bound is optimal. Indeed, Section 5.2 shows that the error is* $\Omega(\log(1/\delta)/T)$ *even for the one-dimensional function* $f(x) = x^2/2$*.*

Theorem 1.17 is an improvement over the $O\big(\log(\log(T)/\delta)/T\big)$ bounds independently proven by Rakhlin et al. [36] (for suffix averaging) and Hazan and Kale [17] (for EpochGD). Once again, we defer the statement of the theorem for general strongly-convex and Lipschitz parameters to Section 6.1.

1.4 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 [7, §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 $\mathrm{var}_{t-1} = \mathrm{E}\left[(X_t - X_{t-1})^2 \mid X_1,\dots,X_{t-1}\right]$ be the conditional variance at time t and let $\mathrm{Var}_t = \sum_{i=1}^{t}\mathrm{var}_{i-1}$ be the *sum of conditional variances* (SCV). Freedman's inequality roughly states that $X_T \lesssim \sqrt{\mathrm{Var}_T}$. A significant technical step of our analysis (Lemma 3.4) shows that the *sum of squared conditional magnitudes* (SSCM), $(M_t)_{t\le T}$ of the accumulated noise exhibits the "chicken and egg" phenomenon alluded to in the discussion of Theorem 1.11. Roughly speaking, we have $M_T \le \alpha X_{T-1} + \beta$ (where $\alpha,\beta > 0$ are scalars), which implies a bound on Var_T since Lemma 1.8 states that $\mathrm{Var}_T \le M_T$. Therefore, an inductive argument using Freedman's inequality shows 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 invokes 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 [6]: by upper-bounding the SSCM 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 $X_t \lesssim \sqrt{M_t}$ *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 [17, 22, 36] and incurs the $\log\log T$ factor. Rakhlin et al. [36] 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 1.11) 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 us to have 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 3.3). 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 $\|x_a - x_b\|^2 \le \frac{(b-a)^2}{\min\{a^2,b^2\}}$ (in the strongly convex case). We derive a much stronger bound, comparable to $\|x_a - x_b\|^2 \le \frac{|b-a|}{\min\{a^2,b^2\}}$. 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 [43], 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 squared conditional magnitude at step t satisfies $m_t \le \alpha_t m_{t-1} + \beta_t \hat{w}_t \sqrt{m_{t-1}} + \gamma_t$, where $\hat{w}_t$ is a conditionally mean-zero random variable and α_t, β_t and γ_t are constants. In [36], using Freedman's Inequality combined with the trick from [6], 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 $X_T \lesssim \sqrt{M_T}$, where M_T is the sum of squared conditional magnitudes at time T. So, our goal becomes to bound M_T. To do so, we develop a probabilistic tool to bound the t iterate of a stochastic process that satisfies a recursive dependence on the $(t-1)$ iterate similar to the one exhibited by M_t.

Theorem 1.19. *Let $(X_t)_{t=1}^T$ be a stochastic process and let $(\mathscr{F}_t)_{t=1}^T$ be a filtration such that X_t is $\mathscr{F}_t$ measurable and X_t is non-negative almost surely. Let $\alpha_t \in [0,1)$ and $\beta_t, \gamma_t \ge 0$ for every t. Assume that $\mathrm{E}[\exp(\lambda X_1)] \le \exp(\lambda K)$ for $\lambda \in (0, 1/K]$ where $K = \max_{1\le t\le T}\left(\frac{2\gamma_t}{1-\alpha_t}, \frac{2\beta_t^2}{1-\alpha_t}\right)$. Let $\hat{w}_t$ be a mean-zero random variable conditioned on $\mathscr{F}_t$ 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, $\Pr[X_t \ge K\log(1/\delta)] \le \mathrm{e}\delta$.*

[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.

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

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 1.19 can be found in Section 4.2. (The recursive MGF bound which removes the non-linear dependence is by Claim 4.6.)

Deterministic lower bound. As mentioned above, a challenge with non-smooth gradient is that the error of the T 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 error in the T iteration. We believe that this example illuminates some non-obvious behavior of gradient descent.

Chapter 2

Finite Dimensional Lower Bounds

2.1 Lower bound on error of final iterate, strongly convex case

In this section we prove that the final iterate of SGD for strongly convex functions has error that is suboptimal by a factor $\Omega(\log T)$, even in the non-stochastic case. We give the proof of Theorem 1.12 in the case where $c = 1$. Theorem 1.12 can be obtained in full generality from the analysis in this section by applying the following reduction, which is easily verifiable via induction.

Lemma 2.1. *Suppose that executing Algorithm 1 over the feasible region $\mathscr{X} \subset \mathbb{R}^n$, on the convex function $f : \mathbb{R}^n \mapsto \mathbb{R}$, using initial point x_1, step-sizes η_t, and map σ such that $\sigma(x) \in \partial f(x)$ as a subgradient oracle, yields the iterates $x_1, x_2, \ldots$. Then, executing Algorithm 1 over $\mathscr{X}$ on the function $(1/c) \cdot f$, using initial point x_1, step-sizes $c \cdot \eta_t$, and subgradient oracle $(1/c)\sigma$ also yields the iterates $x_1, x_2, \ldots$.*

When $c = 1$, we define a function $f = f_T$, depending on T, for which the final iterate produced by Algorithm 1 has $f(x_T) = \Omega(\log(T)/T)$ and $\min_{x \in \mathscr{X}} f(x) \leq 0$, thereby proving (1.2). Let $\mathscr{X}$ be the Euclidean unit ball in $\mathbb{R}^T$. Define $f : \mathscr{X} \to \mathbb{R}$ and $h_i \in \mathbb{R}^T$ for $i \in [T+1]$ by

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

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

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 $\mathscr{X}$ because $\|\nabla H_i(x)\| \leq \|h_i\| + 1$ and $\|h_i\|^2 \leq 1 + \frac{1}{4}\sum_{j=1}^{T} \frac{1}{(T+1-j)^2} < 1 + \frac{1}{2}$. Finally, the minimum value of f over $\mathscr{X}$ 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 [19, Theorem 4.4.2].

Claim 2.2. *$\partial f(x)$ is the convex hull of $\{ h_i + x : i \in \mathscr{I}(x) \}$, where $\mathscr{I}(x) = \{ i : H_i(x) = f(x) \}$.*

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

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 following claim is easy to verify from the definition of z_t.

Claim 2.3.

- *For $t \in [T+1]$, z_t is non-negative. In particular, $z_{t,j} \ge \frac{1}{2(t-1)}$ 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 $z_t \in \mathscr{X}$ 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 2.4. *$f(z_t) = H_t(z_t)$ for all $t \in [T+1]$. The subgradient oracle for f at z_t returns the vector $h_t + z_t$.*

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 subclaim.

Next we claim that $z_t^\mathsf{T} h_t > z_t^\mathsf{T} h_i$ for all $1 \le i < t$. This also follows from 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 $\mathscr{I}(z_t) = \{\, i : H_i(z_t) = f(z_t) \,\} = \{t, \ldots, T+1\}$. 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 2.4, 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. Recall that $\eta_t = 1/t$ since we assume $c = 1$.

Lemma 2.5. *For the function f constructed in this section, the vector x_t in Algorithm 1 equals z_t, for every $t \in [T+1]$.*

Proof. By definition, $z_1 = x_1 = 0$. By Claim 2.4, 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 $x_2 = \Pi_{\mathscr{X}}(y_2)$, which equals e_1 since $y_2 \in \mathscr{X}$. 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 2.4, 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

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

So $y_{t+1} = z_{t+1}$. Since $x_{t+1} = \Pi_{\mathscr{X}}(y_{t+1})$ is defined to be the projection onto $\mathscr{X}$, and $y_{t+1} \in \mathscr{X}$ by Claim 2.3, we have $x_{t+1} = y_{t+1} = z_{t+1}$. ■

The value of the final iterate is easy to determine from Lemma 2.5 and Claim 2.4:

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

(Here the second inequality uses Claim 2.3.) This proves (1.2). A small modification of the last calculation proves (1.3); details may be found in Claim 2.11 in the case where $c = 1$ (the general case follows from Lemma 2.1). This completes the proof of Theorem 1.12.

2.2 Lower bound on error of final iterate, Lipschitz case

In this section we prove a lower bound result for Lipschitz functions analogous to those in Section 2.1. Throughout this section we will assume that $\eta_t = \frac{c}{\sqrt{t}}$ where $c = 1$. Specifically, we define a function $f = f_T$, depending on T, for which the final iterate produced by Algorithm 1 has $f(x_T) = \Omega(\log(T)/\sqrt{T})$, thereby proving (1.4) when $c = 1$. Then, Eq. (1.4) for arbitrary $c > 0$ is a corollary of the result with $c = 1$ and Lemma 2.1.

The function f is defined as follows. As before, $\mathscr{X}$ denotes the Euclidean unit ball in $\mathbb{R}^T$. For $i \in [T]$, define the positive scalar parameters

$$a_i \;=\; \frac{1}{8(T-i+1)} \qquad\qquad b_i \;=\; \frac{\sqrt{i}}{2\sqrt{T}}.$$

Define $f : \mathscr{X} \to \mathbb{R}$ and $h_i \in \mathbb{R}^T$ for $i \in [T+1]$ by

$$f(x) \;=\; \max_{i\in[T+1]} h_i^{\mathsf{T}} x \qquad \text{where} \qquad h_{i,j} \;=\; \begin{cases} a_j & (\text{if } 1 \le j < i) \\ -b_i & (\text{if } i = j \le T) \\ 0 & (\text{if } i < j \le T) \end{cases}.$$

Note that f is 1-Lipschitz over $\mathscr{X}$ because

$$\|h_i\|^2 \;\le\; \sum_{j=1}^{T} a_j^2 + b_T^2 \;=\; \frac{1}{64}\sum_{j=1}^{T}\frac{1}{j^2} + \frac{1}{4} \;<\; \frac{1}{2}.$$

Also, the minimum value of f over $\mathscr{X}$ is non-positive because $f(0) = 0$.

Subgradient oracle. In order to execute Algorithm 1 on f we must specify a subgradient oracle. Similar to Claim 2.2, [19, Theorem 4.4.2] implies

Claim 2.6. *$\partial f(x)$ is the convex hull of $\{\, h_i : i \in \mathscr{I}(x) \,\}$, where $\mathscr{I}(x) = \{\, i : h_i^{\mathsf{T}} x = f(x) \,\}$.*

Our subgradient oracle is as follows: given x, it simply returns $h_{i'} + x$ where $i' = \min \mathscr{I}(x)$.

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

$$z_{t,j} \;=\; \begin{cases} \left(\dfrac{b_j}{\sqrt{j}} - a_j \displaystyle\sum_{k=j+1}^{t-1} \frac{1}{\sqrt{k}}\right) & (\text{if } 1 \le j < t) \\ 0 & (\text{if } t \le j \le T). \end{cases} \qquad (\text{for } t > 1).$$

We will show inductively that these are precisely the first T iterates produced by Algorithm 1 when using the subgradient oracle defined above.

Claim 2.7. *For $t \in [T+1]$, z_t is non-negative. In particular, $z_{t,j} \ge \frac{1}{4\sqrt{T}}$ for $j < t$ and $z_{t,j} = 0$ for $j \ge t$.*

Proof. By definition, $z_{t,j} = 0$ for all $j \ge t$. For $j < t$,

$$\begin{aligned} z_{t,j} \;&=\; \left(\frac{b_j}{\sqrt{j}} - a_j \sum_{k=j+1}^{t-1} \frac{1}{\sqrt{k}}\right) \\ &=\; \left(\frac{1}{2\sqrt{T}} - \frac{1}{8(T-j+1)} \sum_{k=j+1}^{t-1} \frac{1}{\sqrt{k}}\right) \quad (\text{by definition of } a_j \text{ and } b_j) \\ &\ge\; \frac{1}{2\sqrt{T}} - \frac{1}{4(T-j+1)} \frac{t-1-j}{\sqrt{t-1}} \quad (\text{by Claim A.10}) \\ &\ge\; \frac{1}{2\sqrt{T}} - \frac{1}{4\sqrt{T}} \quad (\text{by Claim A.11}), \text{ replacing } t \text{ with } t+1 \\ &=\; \frac{1}{4\sqrt{T}}. \end{aligned}$$

■

Claim 2.8. *$z_{t,j} \le 1/\sqrt{T}$ for all j. In particular, $z_t \in \mathscr{X}$ (the unit ball in $\mathbb{R}^T$).*

Proof. We have $z_{t,j} = 0$ for all $j \ge t$, and for $j < t$, we have

$$z_{t,j} \;=\; \left(\frac{b_j}{\sqrt{j}} - a_j \sum_{k=j+1}^{t} \frac{1}{\sqrt{k}}\right) \;\le\; \frac{b_j}{\sqrt{j}} \;=\; \frac{1}{2\sqrt{T}}.$$

Since Claim 2.7 shows that $z_t \ge 0$, we have $\|z_t\| \le 1$, and therefore $z_t \in \mathscr{X}$. ■

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

Claim 2.9. *$f(z_t) = h_t^T z_t$ for all $t \in [T+1]$. The subgradient oracle for f at z_t returns the vector h_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 subclaim.

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

$$\begin{aligned}
z_t^\mathsf{T}(h_t - h_i) \;&=\; \sum_{j=1}^{t-1} z_{t,j}(h_{t,j} - h_{i,j}) \quad (z_t \text{ is supported on first } t-1 \text{ coordinates}) \\
&=\; \sum_{j=i}^{t-1} z_{t,j}(h_{t,j} - h_{i,j}) \quad (h_i \text{ and } h_t \text{ agree on first } i-1 \text{ coordinates}) \\
&=\; z_{t,i}(a_i + b_i) + \sum_{j=i+1}^{t-1} z_{t,j} a_j \\
&>\; 0.
\end{aligned}$$

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

Since the subgradient returned at z_t is determined by Claim 2.9, 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.

Lemma 2.10. *For the function f constructed in this section, the vector x_t in Algorithm 1 equals z_t, for every $t \in [T+1]$.*

Proof. The proof is by induction. By definition $x_1 = 0$ and $z_1 = 0$, establishing the base case.

So assume $z_t = x_t$ for $t \le T$; we will prove that $z_{t+1} = x_{t+1}$. Recall that Algorithm 1 sets $y_{t+1} = x_t - \eta_t g_t$, and that $\eta_t = \frac{1}{\sqrt{t}}$. By the inductive hypothesis, $x_t = z_t$. By Claim 2.9, the algorithm uses the subgradient $g_t = h_t$.

Thus,

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

So $y_{t+1} = z_{t+1}$. Since $x_{t+1} = \Pi_{\mathscr{B}_T}(y_{t+1})$ by definition, and $y_{t+1} \in \mathscr{X}$ by Claim 2.8, we have $x_{t+1} = y_{t+1} = z_{t+1}$. ■

The value of the final iterate is easy to determine from Lemma 2.5 and Claim 2.4:

$$f(x_{T+1}) \;=\; f(z_{T+1}) \;=\; h_{T+1}^{\mathsf{T}} z_{T+1} \;=\; \sum_{j=1}^{T} h_{T+1,j} \cdot z_{T+1,j} \;\ge\; \sum_{j=1}^{T} \frac{1}{8(T+1-j)} \cdot \frac{1}{4\sqrt{T}} \;>\; \frac{\log T}{32\sqrt{T}}.$$

(Here the second inequality uses Claim 2.7.) This proves (1.4). A small modification of the last calculation proves (1.5) in the case where $c = 1$ (the general case follows from Lemma 2.1); details may be found in Claim 2.12. The proof of (1.6) may be found in Subsection 2.3.3. This completes the proof of Theorem 1.14.

2.3 Omitted proofs for the lower bounds

2.3.1 Strongly convex case

The following claim proves Eq. (1.3) when $c = 1$. The result in full generality follows then as a corollary of the case when $c = 1$ and Lemma 2.1.

Claim 2.11. *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 2.5, $x_t = z_t \; \forall t \in [T+1]$. By Claim 2.3, 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,

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

■

2.3.2 Lipschitz case

The following claim proves Eq. (1.5) in the case when $c = 1$. The result in full generality follows from the case $c = 1$ and Lemma 2.1.

Claim 2.12. *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*

$$f(\bar{x}) \ge \frac{\ln(T) - \ln(k+1)}{32\sqrt{T}}.$$

Proof. By Lemma 2.10, $x_t = z_t$ for all t. By Claim 2.7, every $z_t \ge 0$ so $\bar{x} \ge 0$. Moreover, $z_{t,j} \ge 1/4\sqrt{T}$ for all $T-k+2 \le t \le T+1$ and $1 \le j \le T-k+1$, and $z_{t,T} \ge 0$ for all $t \le T+1$. Consequently, $\bar{x}_j \ge 1/4\sqrt{T}$ for all

$1 \le j \le T-k+1$ and $\bar{x}_T \ge 0$. Thus,

$$\begin{aligned}
f(\bar{x}) &\ge h_{T+1}^{\mathsf{T}}\bar{x} \quad \text{(by definition of } f\text{)} \\
&= \sum_{j=1}^{T-k+1} h_{T+1,j}\bar{x}_j + \sum_{j=T-k+2}^{T} \underbrace{h_{T+1,j}\bar{x}_j}_{\ge 0} \\
&\ge \sum_{j=1}^{T-k+1} a_j \frac{1}{4\sqrt{T}} \\
&= \frac{1}{4\sqrt{T}} \sum_{j=1}^{T-k+1} \frac{1}{8(T-j+1)} \\
&\ge \frac{1}{32\sqrt{T}} \int_1^{T-k+1} \frac{1}{T-x+1}\,dx \\
&= \frac{\log(T)-\log(k)}{32\sqrt{T}}
\end{aligned}$$

■

2.3.3 Monotonicity

The following claim completes the proof of Eq. (1.6) when $c = 1$. The general result follows as a consequence of the case when $c = 1$ and Lemma 2.1.

Claim 2.13. *For any $i \le T$, we have $f(x_{i+1}) \ge f(x_i) + 1/32\sqrt{T}(T-i+1)$.*

Proof.

$$\begin{aligned}
f(x_{i+1}) - f(x_i) &= h_{i+1}^{\mathsf{T}} z_{i+1} - h_i^{\mathsf{T}} z_i \qquad \text{(by Claim 2.9)} \\
&= \sum_{j=1}^{i} (h_{i+1,j} z_{i+1,j} - h_{i,j} z_{i,j}) \\
&= \sum_{j=1}^{i-1} (h_{i+1,j} z_{i+1,j} - h_{i,j} z_{i,j}) + (h_{i+1,i} z_{i+1,i} - h_{i,i} \underbrace{z_{i,i}}_{=0}) \\
&= \sum_{j=1}^{i-1} a_j (z_{i+1,j} - z_{i,j}) + a_i z_{i+1,i} \\
&= \sum_{j=1}^{i-1} a_j \cdot \Big(\frac{-a_j}{\sqrt{i}}\Big) + \frac{1}{8(T-i+1)} z_{i+1,i} \\
&\ge -\frac{1}{64\sqrt{i}} \sum_{j=1}^{i-1} \Big(\frac{1}{T-j+1}\Big)^2 + \frac{1}{32\sqrt{T}(T-i+1)} \qquad \text{(by Claim 2.7)} \\
&\ge \frac{1}{64\sqrt{T}(T-i+1)} \qquad \text{(by Claim 2.14)}
\end{aligned}$$

■

Claim 2.14. *For any $i \le T$*

$$\frac{1}{\sqrt{i}} \sum_{j=1}^{i-1} \Big(\frac{1}{T-j+1} \Big)^2 \;\le\; \frac{1}{\sqrt{T}} \cdot \frac{1}{T-i+1}.$$

Proof. Applying Claim A.12 shows that

$$\sum_{j=1}^{i-1} \Big(\frac{1}{T-j+1} \Big)^2 \;=\; \sum_{\ell=T-i+2}^{T} \frac{1}{\ell^2} \;\le\; \frac{1}{T-i+1} - \frac{1}{T} \;=\; \frac{i-1}{T(T-i+1)} \;\le\; \;=\; \frac{i}{T(T-i+1)}.$$

So it suffices to prove that

$$\frac{\sqrt{i}}{T(T-i+1)} \;\le\; \frac{1}{\sqrt{T}} \cdot \frac{1}{T-i+1}.$$

This obviously holds as $i \le T$. ■

Chapter 3

High Probability Bounds

3.1 Upper bound on error of final iterate, strongly convex case

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 1.9). 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 $\mathscr{F}_t = \sigma(\hat{z}_1, \dots, \hat{z}_t)$ be the σ-algebra generated by the first t steps of SGD. Finally, recall that $\|\hat{z}_t\| \le 1$ and $\mathrm{E}[\hat{z}_t \mid \mathscr{F}_{t-1}] = 0$. Without loss of generality, we may assume that $T \ge 4$ since $f(x_T) - f(x^*) \le O(1)$ for $T < 4$. Moreover, we may assume without loss of generality that T is even because $f(x_T) - f(x_{T-1}) \le O(1/T)$.

We begin with the following lemma which can be inferred from the proof of Theorem 1 in Shamir and Zhang [43]. For completeness, we provide a proof in Section 3.3.

Lemma 3.1. *Let f be 1-strongly convex and 1-Lipschitz. Suppose that we run SGD (Algorithm 1) with step sizes $\eta_t = 1/t$. Assume that T is even. Then*

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

Lemma 3.1 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 1.17) and the noise terms, which we denote by Z_T, are small. We defer the proof of Theorem 1.17 to Subsection 3.1.3. By changing the order of summation, we can write

$$Z_T \;=\; \sum_{t=T/2}^{T}\langle \hat{z}_t, w_t\rangle \tag{3.1}$$

$$\text{where}\quad w_t \;=\; \sum_{j=T/2}^{t}\alpha_j(x_t - x_j) \qquad \text{and} \qquad \alpha_j = \frac{1}{(T-j)(T-j+1)} \qquad \text{for } j \in [T-1],$$

and set $\alpha_T = 0$ to ensure a summand of 0 when $j = t = T$. 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 3.1.1.

Lemma 3.2. *$Z_T \le O\left(\frac{\log(T)\log(1/\delta)}{T}\right)$ with probability at least $1-\delta$.*

Given Theorem 1.17 and Lemma 3.2, the proof of Theorem 1.9 is immediate when T is even. In the case where T is odd, we may apply the argument for $T-1$ and then use the fact that $f(x_T) \le f(x_{T-1}) + O(1/T)$ since $\eta_t = 1/T$ and $\|\hat{g}_t\| = O(1)$.

3.1.1 Bounding the noise

The main technical difficulty in the proof is to understand the noise term, which we have denoted by Z_T. Notice that $(Z_t)_{T/2 \le t \le T}$ is a martingale with respect to $(\mathscr{F}_t)_{T/2 \le t \le T}$ since $\mathrm{E}[\hat{z}_t \mid \mathscr{F}_{t-1}] = 0$, w_t is $\mathscr{F}_{t-1}$ measurable and $\mathrm{E}[|\langle \hat{z}_t, w_t \rangle|] < \infty$. The last property holds because we may use the fact that $\|x_t - x_j\| = O(1)$. The natural starting point is to better understand the sum of squared magnitudes (SSCM) of Z_T. Notice that $|\langle \hat{z}_t, w_t \rangle| \le \|w_t\|$ and w_t is $\mathscr{F}_{t-1}$ measurable. Therefore, we use $\sum_{t=T/2}^{T} \|w_t\|^2$ as the SSCM process of Z_T. We will see that $\sum_{t=T/2}^{T} \|w_t\|^2$ is bounded by a linear transformation of Z_T. This "chicken and egg" relationship inspires us to derive a new probabilistic tool (generalizing Freedman's Inequality) to disentangle the SSCM from the martingale.

The main challenge in analyzing $\|w_t\|$ is to precisely analyze the distance $\|x_t - x_j\|$ between SGD iterates. A loose bound of $\|x_t - x_j\|^2 \lesssim (t-j)\sum_{i=j}^{t-1} \frac{\|\hat{g}_i\|^2}{i^2}$ follows easily from Jensen's Inequality. We prove the following tighter bound, which may be of independent interest. The proof is in Section 3.3.

Lemma 3.3. *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 3.3 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 SSCM process). This is the main technical step that inspired our probabilistic tool (the generalized Freedman's Inequality).

Lemma 3.4 (Main Technical Lemma)**.** *There exists positive values $R_1 = O\left(\frac{\log^2 T}{T^2}\right)$, $R_2 = O\left(\frac{\log T}{T}\right)$, $C_t = \Theta(\log(T-t))$, $R_3 = O\left(\frac{\log T}{T^2}\right)$ such that*

$$\sum_{t=T/2}^{T} \|w_t\|^2 \le R_1 + R_2\|x_{T/2} - x^*\|^2 + \underbrace{\sum_{t=T/2}^{T-1} \frac{C_t}{t}\langle \hat{z}_t, w_t \rangle}_{\approx O(\log(T)/T) Z_T} + R_3 \sum_{t=T/2}^{T-1} \langle \hat{z}_t, x_t - x^* \rangle. \tag{3.2}$$

This bound is mysterious in that the left-hand side is an upper bound on the sum of squared magnitudes of Z_T, whereas the third term on the right-hand side is essentially a scaled version of Z_T itself. This is the "chicken and egg phenomenon" alluded to in Section 1.4, and it poses another one of the main challenges of bounding Z_T. This bound inspires our main probabilistic tool, which we restate for convenience here.

Theorem 1.11 (Generalized Freedman). Let $\{d_i, \mathscr{F}_i\}_{i=1}^n$ be a martingale difference sequence. Suppose v_{i-1}, for $i \in [n]$, are positive and $\mathscr{F}_{i-1}$-measurable random variables such that $\mathrm{E}[\exp(\lambda d_i) \mid \mathscr{F}_{i-1}] \leq \exp\left(\frac{\lambda^2}{2} v_{i-1}\right)$ 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 1.11, we need to refine Lemma 3.4 to replace the terms $\left\|x_{T/2} - x^*\right\|^2$ and $R_3 \sum_{t=T/2}^{T-1} \langle \hat{z}_t, (x_t - x^*) \rangle$ with sufficient high probability upper bounds. Rakhlin et al. [36] 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, which does not suffice for our purposes. In contrast, our analysis only needs a high probability bound on $\left\|x_{T/2} - x^*\right\|^2$ and $R_3 \sum_{t=T/2}^{T} \|x_t - x^*\|^2$; this allows us to avoid a $\log\log T$ factor here by foregoing a simultaneous analysis on all of the $\|x_t - x^*\|^2$ terms. Indeed, we have

Theorem 3.5. *Both of the following hold:*

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

The proof of Theorem 3.5, in Subsection 3.1.2, uses our tool for bounding recursive stochastic processes (Theorem 1.19). Therefore, we need to expose a recursive relationship between $\|x_{t+1} - x^*\|^2$ and $\|x_t - x^*\|^2$ that satisfies the conditions of Theorem 1.19. Interestingly, Theorem 3.5 is also the main ingredient in the analysis of the error of the suffix average (see Subsection 3.1.3). We now have enough to give our refined version of Lemma 3.4, which is now in a format usable by Freedman's Inequality.

Lemma 3.6. *For every $\delta > 0$ there exists positive values $R = O\left(\frac{\log^2 T \log(1/\delta)}{T^2}\right)$, $C_t = \Theta(\log(T - t))$ 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 3.5 with an easy corollary of Freedman's Inequality (Corollary 4.4) which states that a high probability bound of M on the SSCM of a martingale implies a high probability bound of $\sqrt{M}$ on the martingale.

Let R_1, R_2, R_3 and C_t be as in Lemma 3.4. Consider the resulting upper bound on $\sum_{t=T/2}^{T} \|w_t\|^2$. We already have a bound on the first term in Eq. (3.2), namely $R_1 = O(log^2(T)/T^2)$. We know proceed to give a high probability bound on the second term, and the fourth term. The first claim in Theorem 3.5 gives $R_2 \left\|x_{T/2} - x^*\right\|^2 = O\left(\frac{\log T \log(1/\delta)}{T^2}\right)$ with probability at least $1 - \delta$ because $R_2 = O(\log(T)/T)$.

By the second claim in Theorem 3.5, we have $R_3^2 \sum_{t=T/2}^{T-1} \|x_t - x^*\|^2 = O\left(\frac{\log^2 T}{T^4} \log(1/\delta)\right)$ with probability at least $1 - \delta$ because $R_3 = O\left(\frac{\log T}{T^2}\right)$. Hence, we have derived a high probability bound on the SSCM of $R_3 \sum_{t=T/2}^{T} \langle \hat{z}_t, x_t - x^* \rangle$. Therefore, we turn this into a high probability bound on the martingale itself by applying Corollary 4.4 and obtain $R_3 \sum_{t=T/2}^{T-1} \langle \hat{z}_t, x_t - x^* \rangle = O\left(\frac{\log T \log(1/\delta)}{T^2}\right)$ with probability at least $1 - \delta$. ■

Now that we have derived an upper bound on the sum of squared conditional magnitudes of Z_T in the form required by our Generalized Freedman Inequality (Theorem 1.11), we are finally ready to prove Lemma 3.2 (our high probability upper bound on the noise, Z_T).

Proof (of Lemma 3.2). 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 1.11. Indeed, if $(M_t)_{T/2\le t\le T}$ is a martingale with increments d_i and if its SSCM process at time step T is bounded by $R\log(1/\delta)+\sum_{t=T/2}^{T}\alpha_t d_t$ with $\max_t \alpha_t = O(\sqrt{R})$. Then Corollary 4.5 bounds the martingale at time step T by $\sqrt{R}\log(1/\delta)$ with high probability.

Recall that our martingale at time T is $Z_T = \sum_{t=T/2}^{T}\langle \hat{z}_t, w_t \rangle$. Let $a_t = \hat{z}_t$, $b_t = w_t$, and $d_t = \langle a_t, b_t \rangle$ for $t = T/2,\dots,T$ and $a_t = b_t = d_t = 0$ for $t < T/2$. Then, $Z_T = \sum_{t=T/2}^{T} d_t$ as in the conclusion of Corollary 4.5. To satisfy the hypothesis of Corollary 4.5, we will use Lemma 3.6 which shows

$$\sum_{t=1}^{T}\|b_t\|^2 \;=\; \sum_{t=T/2}^{T}\|w_t\|^2 \;\le\; R\log(1/\delta)+\sum_{t=T/2}^{T}\underbrace{\frac{C_t}{t}}_{:=\alpha_t}\underbrace{\langle \hat{z}_t, w_t \rangle}_{=d_t},$$

where $R = O(\log(T)^2/T^2)$. Notice that $\max_t \alpha_t = \max_t \frac{C_t}{t} = O(\log(T)/T) = O(\sqrt{R})$. Hence, applying Corollary 4.5 proves that $Z_T = O(\sqrt{R}\log(1/\delta)) = O((\log T\log(1/\delta))/T)$ with probability at least $1-\delta$, as required. ■

3.1.2 High probability bounds on squared distances to x^*

In this section, we prove Theorem 3.5. We begin with the following claim which can be extracted from [36, Equation (11)] by taking $G = 4$ (since $\|\hat{g}_t\|^2 \le (\|\hat{z}_t\|+\|g_t\|)^2 \le (1+1)^2 = 4$), $\lambda = 1$, and $\eta_t = 1/t$.

Claim 3.7 ([36, Proof of Lemma 6])**.** *Suppose f is 1-strongly-convex and 1-Lipschitz. Define $Y_t = t\,\|x_{t+1}-x^*\|^2$ and $U_t = \langle \hat{z}_{t+1}, x_{t+1}-x^* \rangle / \|x_{t+1}-x^*\|_2$. 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 $\|x_{t+1}-x^*\|^2$ and $\|x_t-x^*\|^2$ and inspires our probabilistic tool for recursive stochastic processes (Theorem 1.19). We prove Theorem 3.5 using this tool:

Proof (of Theorem 3.5). Consider the stochastic process $(Y_t)_{t=1}^{T-1}$ where Y_t is as defined by Claim 3.7. Note that Y_t satisfies the conditions of Theorem 1.19 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 $\mathscr{F}_{t+1}$ measurable random variable which is mean zero conditioned on $\mathscr{F}_t$. Furthermore, note that $|U_t| \le 1$ with probability 1 because $\|\hat{z}_{t+1}\| \le 1$ with probability 1. Furthermore, 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. Lastly, we observe that $\|x_2-x^*\|^2 \le 4$ with probability 1. Indeed by 1-strong-convexity and 1-Lipschitzness of f,

$$\|x_t-x^*\| \ge \langle g_t, x_t-x^* \rangle \ge \frac{1}{2}\|x_t-x^*\|^2.$$

Thus, taking $K = 8$, it follows that $\mathrm{E}[\exp(\lambda X_1)] \le \exp(\lambda K)$, as required by Theorem 1.19. So, we may apply Theorem 1.19 to obtain:

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

Recalling that $Y_t = t\|x_{t+1} - x^*\|^2$ and setting $\sigma'_t = \sigma_t/t$ and adjusting δ as required proves Theorem 3.5. ■

3.1.3 Upper bound on error of suffix averaging

To complete the proof of the final iterate upper bound (Theorem 1.9), it still remains to prove the suffix averaging upper bound (Theorem 1.17). In this section, we prove this result as a corollary of the high probability bounds on $\|x_t - x^*\|^2$ that we obtained in the previous subsection. See Section 3.4 for an alternative proof of Theorem 1.17 which does not involve the high probability bounds on $\|x_t - x^*\|^2$.

The following proof of Theorem 1.17 is quite short and involves several ingredients. Some of the ingredients are fairly standard: Lemma 3.14 is a standard analysis of Algorithm 1 and Corollary 4.4 is an easy corollary of Freedman's inequality which takes a high probability bound on the sum of squared magnitudes (SSCM) process of a martingale and converts it into a high probability bound on the martingale itself. The main novel ingredient is Theorem 3.5 which allows one to bound $\sum_{t=T/2}^{T}\|x_t - x^*\|^2$ by $O(\log(1/\delta))$ with probability at least $1-\delta$. This summation is the SSCM process of a martingale which arises from an application of Lemma 3.14. Then, Corollary 4.4 converts this high probability bound on the SSCM to a high probability bound on the martingale.

Proof (of Theorem 1.17). By Lemma 3.14 with $w = x^*$ we have

$$\sum_{t=T/2}^{T}[f(x_t) - f(x^*)] \;\le\; \underbrace{\frac{1}{2}\sum_{t=T/2}^{T}\eta_t\|\hat{g}_t\|^2}_{(a)} + \underbrace{\frac{1}{2\eta_{T/2}}\left\|x_{T/2} - x^*\right\|^2}_{(b)} + \underbrace{\sum_{t=T/2}^{T}\langle \hat{z}_t, x_t - x^*\rangle}_{(c)}. \tag{3.3}$$

It suffices to bound the right hand side of (3.3) by $O(\log(1/\delta))$ with probability at least $1-\delta$. Indeed, bounding $\|\hat{g}_t\|^2$ by 4, (a) in (3.3) is bounded by $O(1)$. Term (b) is bounded by $O(\log(1/\delta))$ by Theorem 3.5.

It remains to bound (c). Theorem 3.5 implies $\sum_{t=T/2}^{T}\|x_t - x^*\|^2 = O(\log(1/\delta))$ with probability at least $1-\delta$. Therefore, Corollary 4.4 shows that (c) is at most $O(\log(1/\delta))$ with probability at least $1-\delta$. ■

3.2 Upper bound on error of final iterate, Lipschitz case: proof sketch

In this section we provide a proof sketch of the upper bound of the final iterate of SGD, in the case where f is 1-Lipschitz but not necessarily strongly-convex (Theorem 1.10). The proof of Theorem 1.10 closely resembles the proof of Theorem 1.9 and we will highlight the main important differences. Perhaps the most notable difference is that the analysis in the Lipschitz case does not require a high probability bound on $\|x_t - x^*\|^2$.

Recall that the step size used by Algorithm 1 in this case is $\eta_t = 1/\sqrt{t}$. 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 $\mathscr{F}_t = \sigma(\hat{z}_1, \ldots, \hat{z}_t)$ be the σ-algebra generated by the first t steps of SGD. Finally, recall that $\|\hat{z}_t\| \le 1$ and $\mathrm{E}[\hat{z}_t \mid \mathscr{F}_{t-1}] = 0$.

As before, we begin with a lemma which can be obtained by modifying the proof of Lemma 3.1 to replace applications of strong convexity with the subgradient inequality.

Lemma 3.8. *Let f be 1-Lipschitz. Suppose that we run SGD (Algorithm 1) with step sizes $\eta_t = \frac{1}{\sqrt{t}}$. Then,*

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

Lemma 3.8 asserts that the error of the last iterate is bounded by the sum of the error of the average of the iterates and some noise terms (up to the additive $O(\log T/\sqrt{T})$ term). A standard analysis (similar to the proof of Lemma 3.14) reveals $\sum_{t=T/2}^{T}[f(x_t)-f(x^*)] \le O(\sqrt{T})+\sum_{t=T/2}^{T}\langle \hat{z}_t, x_t-x^* \rangle$. Applying Azuma's inequality on the summation (using the diameter bound to obtain $\langle \hat{z}_t, x_t-x^* \rangle^2 \le 1$) shows

Lemma 3.9. *For every $\delta \in (0,1)$,*

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

with probability at least $1-\delta$.

As a consequence of Lemma 3.9, it is enough to prove that the error due to the noise terms are small in order to complete the proof of Theorem 1.10. By changing the order of summation, we can write $Z_T = \sum_{t=T/2}^{T}\langle \hat{z}_t, w_t \rangle$ where

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

Just as in Section 3.1, 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 3.2.1.

Lemma 3.10. *For every $\delta \in (0,1)$, $Z_T \le O\left(\log(T)\log(1/\delta)/\sqrt{T}\right)$ with probability at least $1-\delta$.*

Given Lemma 3.9 and Lemma 3.10, the proof of Theorem 1.10 is straightforward. The next sub-section provides a proof sketch of Lemma 3.10.

3.2.1 Bounding the noise

The goal of this section is to prove Lemma 3.10. Just as in Section 3.1, the main technical difficulty is to understand the noise term, denoted Z_T. Observe that $(Z_t)_{T/2\le t\le T}$ a martingale, and $\sum_{t=T/2}^{T}\|w_t\|^2$ is the SSCM of Z_T. The SSCM of Z_T will be shown to exhibit the "chicken and egg" relationship which we have already seen explicitly exhibited by the SSCM of the noise terms in the strongly convex case. That is, we will see that the $\sum_{t=T/2}^{T}\|w_t\|^2$ is bounded by a linear transformation of Z_T. We will again use our Generalized Freedman to disentangle the sum of squared magnitudes from the martingale.

The distance $\|x_t - x_j\|$ between SGD iterates is again a crucial quantity to understand in order to bound $\sum_{t=T/2}^{T}\|w_t\|^2$ (see Subsection 3.1.1 to see why). Therefore, we develop a distance estimate analogous to Lemma 3.3

Lemma 3.11. *Suppose f is 1-Lipschitz. Suppose we run Algorithm 1 for T iterations with step sizes $\eta_t = 1/\sqrt{t}$. Let $a < b$. Then,*

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

We then use Lemma 3.11 to prove Lemma 3.12, our main upper bound on $\sum_{t=T/2}^{T}\|w_t\|^2$. This follows from some delicate calculations similar to those in Subsection 3.3.4, replacing the strongly-convex distance estimate (Lemma 3.3) with the Lipschitz distance estimate (Lemma 3.11), along with some other minor modifications.

This upper bound reveals the surprisingly intricate relationship between Z_T (the martingale) and $\sum_{t=T/2}^{T} \|w_t\|^2$ (its SSCM).

Lemma 3.12 (Main Technical Lemma (Lipschitz Case))**.** *There exists positive values* $R_1 = O\left(\frac{\log^2 T}{T}\right)$, $R_2 = O\left(\frac{\log T}{T^{1.5}}\right)$, *and* $C_t = O(\log T)$, *such that*

$$\sum_{t=T/2}^{T} \|w_t\|^2 \leq R_1 + R_2 \sum_{t=T/2}^{T} \langle \hat{z}_t, x_t - x^* \rangle + \underbrace{\sum_{t=T/2}^{T} \langle \hat{z}_t, \frac{C_t}{\sqrt{t}} w_t \rangle}_{\approx O(\log T/\sqrt{T}) Z_T}.$$

Just as in Lemma 3.4, the left-hand side is an upper bound on the sum of squared magnitudes of Z_T, whereas the right-hand side essentially contains a scaled version of Z_T itself. This is another instance of the "chicken and egg phenomenon" alluded to in Section 1.4, and it is the main challenge of bounding Z_T. For convenience, we restate our main probabilistic tool which allows us to deal with the chicken and egg phenomenon.

Theorem 1.11 (Generalized Freedman). Let $\{d_i, \mathscr{F}_i\}_{i=1}^n$ be a martingale difference sequence. Suppose v_{i-1}, for $i \in [n]$, are positive and $\mathscr{F}_{i-1}$-measurable random variables such that $\mathrm{E}[\exp(\lambda d_i) \mid \mathscr{F}_{i-1}] \leq \exp\left(\frac{\lambda^2}{2} v_{i-1}\right)$ 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 1.11, we need to refine Lemma 3.12 to replace $R_2 \sum_{t=T/2}^{T} \langle \hat{z}_t, x_t - x^* \rangle$ with a sufficient high probability upper bound. This is similar to the refinement of Lemma 3.4 from Subsection 3.1.1. However, unlike the refinement in Subsection 3.1.1 (which required a high probability bound on $\sum_{t=T/2}^{T} A_t \|x_t - x^*\|^2$ without any diameter bound), the refinement here is quite easy. Using the diameter bound, the almost sure bound of $\|\hat{z}_t\| \leq 1$, and Azuma's inequality, we can bound $\sum_{t=T/2}^{T} \langle \hat{z}_t, x_t - x^* \rangle$ by $\sqrt{T \log(1/\delta)}$ with probability at least $1 - \delta$. This yields the following lemma.

Lemma 3.13. *For every* $\delta \in (0,1)$, *there exists positive values* $R = O\left(\frac{\log^2 T \sqrt{\log(1/\delta)}}{T}\right)$, $C_t = O(\log T)$, *such that* $\sum_{t=T/2}^{T} \|w_t\|^2 \leq R + \sum_{t=T/2}^{T-1} \langle \hat{z}_t, \frac{C_t}{\sqrt{t}} w_t \rangle$, *with probability at least* $1 - \delta$.

Now that we have derived an upper bound on the sum of squared magnitutdes of Z_T in the form required by Generalized Freedman Inequality (Theorem 1.11), we are now finally ready to prove Lemma 3.10 (the high probability upper bound on the noise, Z_T).

Proof (of Lemma 3.10). 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 1.11.

Indeed, consider a filtration $\{\mathscr{F}_t\}_{t=T/2}^{T}$. Let $d_t = \langle a_t, b_t \rangle$ define a martingale difference sequence where $\|a_t\| \leq 1$ and $\mathrm{E}[a_t \mid \mathscr{F}_{t-1}] = 0$. Suppose there are positive values, R, α_t, such that $\max_{t=T/2}^{T} \{\alpha_t\} = O\left(\sqrt{R}\right)$ and $\sum_{t=T/2}^{T} \|b_t\|^2 \leq \sum_{t=T/2}^{T} \alpha_t d_t + R\sqrt{\log(1/\delta)}$ with probability at least $1 - \delta$. Then, Corollary 4.5 bounds the martingale at time step T by $\sqrt{R}\log(1/\delta)$ with high probability.

Observe that Lemma 3.13 allows us to apply Corollary 4.5 with $a_t = \hat{z}_t$, $b_t = w_t$, $\alpha_t = (C_t/\sqrt{t})$ for $t = T/2, \ldots, T-1$, $\alpha_T = 0$, $\max_{t=T/2}^{T}\{\alpha_t\} = O\left(\log T/\sqrt{T}\right)$, and $R = O\left(\log^2 T/T\right)$ to prove Lemma 3.10.

■

3.3 Omitted proofs from Section 3.1

3.3.1 Standard analysis of SGD

The following lemma is standard.

Lemma 3.14. *Let f be an 1-strongly convex and 1-Lipschitz function. Consider running Algorithm 1 for T iterations. Then, for every $w \in \mathscr{X}$ 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.

$$\begin{aligned}
f(x_t) - f(w) &\le \langle g_t, x_t - w\rangle - \frac{1}{2}\|x_t - w\|^2 \qquad \text{(by strong-convexity)}\\
&= \langle \hat{g}_t, x_t - w\rangle - \frac{1}{2}\|x_t - w\|^2 + \langle \hat{z}_t, x_t - w\rangle \qquad (\hat{g}_t = g_t - \hat{z}_t)\\
&= \frac{1}{\eta_t}\langle x_t - y_{t+1}, x_t - w\rangle - \frac{1}{2}\|x_t - w\|^2 + \langle \hat{z}_t, x_t - w\rangle \qquad (y_{t+1} = x_t - \eta_t\hat{g}_t)\\
&= \frac{1}{2\eta_t}\left(\|x_t - y_{t+1}\|^2 + \|x_t - w\|^2 - \|y_{t+1} - w\|^2\right) - \frac{1}{2}\|x_t - w\|^2 + \langle \hat{z}_t, x_t - w\rangle\\
&\le \frac{1}{2\eta_t}\left(\|\eta_t\hat{g}_t\|^2 + \|x_t - w\|^2 - \|x_{t+1} - w\|^2\right) - \frac{1}{2}\|x_t - w\|^2 + \langle \hat{z}_t, x_t - w\rangle \qquad \text{(by Claim A.8).}
\end{aligned}$$

Now, summing t from k to T,

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

as desired. ■

3.3.2 Proof of Lemma 3.1

Lemma 3.1. Let f be 1-strongly convex and 1-Lipschitz. Suppose that we run SGD (Algorithm 1) with step sizes $\eta_t = 1/t$. Assume that T is even. Then

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

Proof (of Lemma 3.1). Let $k \in [T-1]$. Apply Lemma 3.14, replacing k with $T-k$ and $w = x_{T-k}$ to obtain:

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

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

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

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

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

Dividing by k, we obtain:

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

Thus, by induction:

$$\begin{aligned} f(x_T) \;&=\; S_0 \\ &\le\; S_{T/2} + \sum_{k=1}^{T/2}\frac{1}{2k(k+1)}\sum_{t=T-k}^{T}\eta_t\,\|\hat{g}_t\|^2 + \sum_{k=1}^{T/2}\frac{1}{k(k+1)}\sum_{t=T-k}^{T}\langle \hat{z}_t, x_t - x_{T-k}\rangle \qquad (3.4)\\ &=\; \frac{1}{T/2+1}\sum_{t=T/2}^{T} f(x_t) + \sum_{k=1}^{T/2}\frac{1}{2k(k+1)}\sum_{t=T-k}^{T}\eta_t\,\|\hat{g}_t\|^2 + \sum_{k=1}^{T/2}\frac{1}{k(k+1)}\sum_{t=T-k}^{T}\langle \hat{z}_t, x_t - x_{T-k}\rangle. \end{aligned}$$

Since we assume that f is 1 Lipschitz, we have $\|g_t\| \le 1$. Recall that we assume that the noise is bounded,

i.e. $\|\hat{z}_t\| \le 1$. Thus $\|\hat{g}_t\|^2 \le 4$. Recall that $\eta_t = 1/t$. So we can bound the middle term as

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

This completes the proof. ■

3.3.3 Proof of Lemma 3.3

Lemma 3.3. 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,

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

Proof (of Lemma 3.3).

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

Repeating this argument iteratively on $\|x_a - x_{b-1}\|$, $\|x_a - x_{b-2}\|$, …, $\|x_a - x_{a+1}\|$, 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 subgradient 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. ■

3.3.4 Proof of Lemma 3.4

Proof (of Lemma 3.4). Recall from Eq. (3.1) that $w_t = \sum_{j=T/2}^{t-1} \alpha_j (x_t - x_j)$.

Definition 3.15. *Define* $B_T := \sum_{t=T/2}^{T} \frac{1}{T-t+1} \sum_{j=T/2}^{t-1} \alpha_j \left\|x_t - x_j\right\|^2$.

Claim 3.16. $\sum_{t=T/2}^{T} \|w_t\|^2 \le B_T$.

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

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

where the first inequality is due to the convexity of $\|\cdot\|^2$ and the second inequality is Claim A.13. ■

Using the definition of B_T and Lemma 3.3, let us write $B_T \le \Lambda_1 + \Lambda_2 + \Lambda_3$ where

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

Let us bound each of the terms separately. Recall from Eq. (3.1) that $\alpha_j = \frac{1}{(T-j)(T-j+1)}$ for $j \in [T-1]$.

Claim 3.17. $\Lambda_1 \le O\left(\frac{\log^2(T)}{T^2}\right)$.

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 3.18.

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

We will prove Claim 3.18 in the next section.

Claim 3.19.

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

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

Proof. Rearranging the order of summation in Λ_3 we get:

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

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 3.16) complete the proof of Lemma 3.4. ■

3.3.5 Proof of Claim 3.18

Let us rewrite

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

and determine the coefficients γ_a.

Claim 3.20. *For each* $a \in \{T/2, \ldots, T-1\}$, $\gamma_a \;=\; O\Big(\frac{\log(T)}{T^2}\Big)$.

Proof. In the definition of Λ_2, the indices providing a positive coefficient for F_a must satisfy $j = a$, $a \le i$, and

$a \le t-1$. Recall that we assume $T \ge 4$, so $a \ge 2$. Hence, the positive contribution to γ_a is:

$$\begin{aligned}
&\sum_{t=1+a}^{T} \frac{2}{T-t+1}\alpha_a \sum_{i=a}^{t-1}\frac{1}{i} \\
&\le \sum_{t=1+a}^{T}\left(\frac{2}{T-t+1}\alpha_a\right)\left(\log\left(T/(a-1)\right)\right) \qquad \text{(by Claim A.15)} \\
&\le \sum_{t=1+a}^{T}\left(\frac{2}{T-t+1}\alpha_a\right)\left(\frac{T-a+1}{a-1}\right) \qquad \text{(by Claim A.14)} \\
&= \sum_{t=1+a}^{T}\left(\frac{2}{T-t+1}\right)\left(\frac{1}{(T-a)(T-a+1)}\right)\left(\frac{T-a+1}{a-1}\right) \\
&= \frac{1}{T-a}\sum_{t=1+a}^{T}\frac{2}{(T-t+1)(a-1)}.
\end{aligned}$$

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

$$\begin{aligned}
&-\sum_{t=1+a}^{T}\frac{2}{T-t+1}\sum_{j=T/2}^{a}\alpha_j \cdot \frac{1}{a} \\
&= -\sum_{t=1+a}^{T}\left(\frac{2}{T-t+1}\right)\left(\frac{1}{a}\right)\left(\frac{1}{T-a}-\frac{1}{T/2+1}\right) \qquad \text{(by Claim A.13)} \\
&= -\sum_{t=1+a}^{T}\left(\frac{2}{T-t+1}\right)\left(\frac{1}{a}\right)\left(\frac{2a-T+2}{2(T/2+1)(T-a)}\right) \\
&= -\frac{1}{(T/2+1)(T-a)}\sum_{t=1+a}^{T}\left(\frac{2}{T-t+1}\right)\left(\frac{2a-T+2}{2a}\right) \\
&= -\frac{2}{(T+2)(T-a)}\sum_{t=1+a}^{T}\left(\frac{2}{T-t+1}\right)\left(1-\frac{T-2}{2a}\right).
\end{aligned}$$

Now, combining the positive and negative contribution we see:

$$\begin{aligned}
\gamma_a &\le \frac{1}{T-a}\sum_{t=1+a}^{T}\frac{2}{T-t+1}\left(\frac{1}{a-1}-\frac{2}{T+2}\left(1-\frac{T-2}{2a}\right)\right)\\
&= \frac{1}{T-a}\sum_{t=1+a}^{T}\frac{2}{T-t+1}\left(\frac{T+2-2\big(a-1\big)\big(1-\frac{T-2}{2a}\big)}{(a-1)(T+2)}\right)\\
&= \frac{1}{T-a}\sum_{t=1+a}^{T}\frac{2}{T-t+1}\left(\frac{T+2-2(a-1)+\frac{2(T-2)(a-1)}{2a}}{(a-1)(T+2)}\right)\\
&\le \frac{1}{T-a}\sum_{t=1+a}^{T}\frac{2}{T-t+1}\left(\frac{2\big(T-a\big)+2}{(T+2)(a-1)}\right)\\
&\le \frac{1}{T-a}\sum_{t=1+a}^{T}\frac{2}{T-t+1}\left(\frac{2\big(T-a\big)+2\big(T-a\big)}{(T+2)(a-1)}\right) \qquad (a\le T-1)\\
&= \frac{1}{(T+2)(a-1)}\sum_{t=1+a}^{T}\frac{8}{T-t+1}\\
&\le \frac{1}{(T+2)(T/2-1)}\sum_{t=1+a}^{T}\frac{8}{T-t+1} \qquad (a\ge T/2)\\
&= O\left(\frac{\log(T)}{T^2}\right),
\end{aligned}$$

as desired. ■

Proof (of Claim 3.18).

$$\begin{aligned}
\Lambda_2 &= \sum_{a=T/2}^{T-1}\gamma_a F_a\\
&\le O\left(\frac{\log(T)}{T^2}\right)\sum_{a=T/2}^{T-1}\big(f(x_a)-f(x^*)\big) \qquad \text{(by Claim 3.20 and the definition of } F_a\text{)}
\end{aligned}$$

Then by Lemma 3.14 with $k=T/2$ and $w=x^*$,

$$\begin{aligned}
&\le O\left(\frac{\log(T)}{T^2}\right)\left(\frac{1}{2}\sum_{t=T/2}^{T-1}\eta_t\left\|\hat{g}_t\right\|^2+\frac{1}{2\eta_{T/2}}\left\|x_{T/2}-x^*\right\|^2+\sum_{t=T/2}^{T-1}\left\langle\hat{z}_t,x_t-x^*\right\rangle\right) \qquad (3.5)\\
&\le O\left(\frac{\log(T)}{T^2}\right)\sum_{t=T/2}^{T-1}\frac{1}{t}+O\left(\frac{\log(T)}{T}\right)\left\|x_{T/2}-x^*\right\|^2+O\left(\frac{\log(T)}{T^2}\right)\sum_{t=T/2}^{T-1}\left\langle\hat{z}_t,x_t-x^*\right\rangle \quad (\|\hat{g}_t\|\le 2)\\
&\le O\left(\frac{\log(T)}{T^2}\right)+O\left(\frac{\log(T)}{T}\right)\left\|x_{T/2}-x^*\right\|^2+O\left(\frac{\log(T)}{T^2}\right)\sum_{t=T/2}^{T-1}\left\langle\hat{z}_t,x_t-x^*\right\rangle,
\end{aligned}$$

as desired. ■

3.3.6 Proof of Claim 3.7

Proof (of Claim 3.7). 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. $f(x_t) - f(x^*) \geq \frac{1}{2}\|x_t - x^*\|^2$ (since $0 \in \partial f(x^*)$).

The analysis proceeds as follows:

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

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. ■

3.4 Alternative proof of Theorem 1.17

In this section we demonstrate how to prove Theorem 1.17 using the Generalized Freedman inequality. In particular, we will not use Theorem 3.5 (the high probability bound on $\|x_t - x^*\|^2$), which was the main tool used to bound the error of suffix-averaging in Subsection 3.1.3. Here, we state a specialized form of the Generalized Freedman's Inequality which is a direct corollary from Lemma 4.3. Notice that Theorem 3.21 does not include V_T in the event which it bounds whereas Theorem 1.11 includes V_T. This is because Theorem 3.21 assumes that V_T satisfies the "chicken and egg" phenomenon *almost surely* whereas Theorem 1.11 does not make this assumption.

Theorem 3.21. *Let $\{d_t, \mathscr{F}_t\}_{t=1}^T$ be a martingale difference sequence. Suppose that, for $t \in [T]$, v_{t-1} are non-negative $\mathscr{F}_{t-1}$-measurable random variables satisfying $\mathrm{E}[\exp(\lambda d_t) \mid \mathscr{F}_{t-1}] \leq \exp\left(\frac{\lambda^2}{2} v_{t-1}\right)$ for all $\lambda > 0$. Let $S_T = \sum_{t=1}^T d_t$ and $V_T = \sum_{t=1}^T v_{t-1}$. Suppose there exists $\alpha_1, \ldots, \alpha_T, \beta \in \mathbb{R}_{\geq 0}$ such that $V_T \leq \sum_{t=1}^T \alpha_t d_t + \beta$. Let $\alpha \geq \max_{t \in [T]} \alpha_t$. Then*

$$\Pr[S_T \geq x] \;\leq\; \exp\left(-\frac{x^2}{4\alpha \cdot x + 8\beta}\right).$$

The main idea of this proof is to upper bound the error of the suffix average by a martingale whose sum of squared conditional magnitudes (SSCM) is bounded above by a linear combination of its own increments, where the coefficients in the linear combination are all $O(1)$. Informally, we will show

$$\sum_{t=T/2}^{T} [f(x_t) - f(x^*)] \;\leq\; \sum_{t=1}^{T} d_t + O(1) \qquad \text{and} \qquad \sum_{t=1}^{T} d_t^2 \;\leq\; \sum_{t=1}^{T} O(1) d_t + O(1).$$

Then, an application of Theorem 3.21 using $\alpha = O(1)$, $\beta = O(1)$ and $x = O(\log(1/\delta))$ proves Theorem 1.17. More formally, we prove the following lemmata.

Lemma 3.22. *Let f be a 1 strongly-convex and 1-Lipschitz function over $\mathscr{X}$. Consider running SGD (Algorithm 1) until time T with step size $\eta_t = 1/t$. Let $x^* = \arg\min_{x\in\mathscr{X}} f(x)$. Then,*

$$\sum_{t=T/2}^{T} (f(x_t) - f(x^*)) \;\le\; \zeta + \sum_{t=2}^{T/2-1} \xi_t \langle \hat{z}_t, x_t - x^* \rangle + \sum_{t=T/2}^{T} \langle \hat{z}_t, x_t - x^* \rangle, \tag{3.6}$$

where $\zeta = O(1)$ and $\xi_t = \Theta\left(\frac{t}{T}\right)$.

Lemma 3.23. *Let ξ_t be as in Lemma 3.22. Then,*

$$\sum_{t=2}^{T/2-1} \xi_t^2 \|x_t - x^*\|^2 + \sum_{t=T/2}^{T} \|x_t - x^*\|^2 \;\le\; \sigma + \sum_{t=2}^{T/2-1} \sigma_t \xi_t \langle \hat{z}_t, x_t - x^* \rangle + \sum_{t=T/2}^{T} \sigma_t \langle \hat{z}_t, x_t - x^* \rangle, \tag{3.7}$$

where σ and σ_t are $O(1)$ for all t.

Lemma 3.22 and Lemma 3.23 suffice for us to prove Theorem 1.17 by applying Theorem 3.21.

Proof (of Theorem 1.17). It suffices to bound $\sum_{t=T/2}^{T} (f(x_t) - f(x^*))$ by $O(\log(1/\delta))$ with probability at least $1-\delta$. Let $d_t = \xi_t \langle \hat{z}_t, x_t - x^* \rangle$ for $t = 2, \ldots, T/2-1$, $d_t = \langle \hat{z}_t, x_t - x^* \rangle$ for $t = T/2, \ldots, T$, and $d_1 = 0$. Applying Lemma 3.22, it suffices to bound $\sum_{t=1}^{T} d_t$ by $O(\log(1/\delta))$ with probability at least $1-\delta$. We check the conditions needed to apply Theorem 3.21.

Let $v_{t-1} = 0$ for $t = 1$, $v_{t-1} = \xi_t^2 \|x_t - x^*\|^2$ for $t = 2, \ldots, T/2-1$ and $v_{t-1} = \|x_t - x^*\|^2$ for $t = T/2, \ldots, T$. Using Cauchy-Schwarz and the almost sure bound on $\|\hat{z}_t\|$, we see $|\langle \hat{z}_t, x_t - x^* \rangle| \le \|x_t - x^*\|$. Because $\|x_t - x^*\|$ is $\mathscr{F}_{t-1}$-measurable, this implies (via Hoeffdings Lemma – Lemma A.5) that for every t and $\lambda > 0$ we have $\mathrm{E}[\exp(\lambda d_t) \mid \mathscr{F}_{t-1}] \le \exp\left(\frac{\lambda^2}{2} v_{t-1}\right)$.

Next, applying Lemma 3.23, we have $\sum_{t=1} v_{t-1} \le \sigma + \sum_{t=1}^{T} \sigma_t d_t$ with $\sigma_t, \sigma = O(1)$. Setting $\beta = \sigma$ and $\alpha_t = \sigma_t$, we have $\alpha = \max_{t\in[T]} \alpha_t = O(1)$, we may now apply Theorem 3.21 using $\beta = O(1)$, $\alpha = O(1)$, $x = O(\log(1/\delta))$ to obtain the desired result.

■

A tool which we will use repeatedly to prove the two lemmata above is the following result from [36].

Lemma 3.24 ([36, Lemma 6])**.** *Let f be λ strongly-convex. Consider running SGD (Algorithm 1) until time step T with step size $\eta_t = 1/(\lambda t)$. Then, for all $t \ge 3$*

$$\|x_t - x^*\|^2 \;\le\; \frac{2}{\lambda(t-2)(t-1)} \sum_{i=2}^{t-1} (i-1) \langle \hat{z}_i, x_i - x^* \rangle + \frac{\|\hat{g}_{t-1}\|^2}{\lambda^2 (t-1)}.$$

3.4.1 Proof of Lemma 3.22

Proof (of Lemma 3.22). By Lemma 3.14 with $w = x^*$ we have

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

Observe that (a) is bounded by $O(1)$ by bounding $\|\hat{g}_t\|^2$ by 4 using 1-Lipschitzness, that $\|\hat{z}_t\| \le 1$ almost surely, and that $\eta_t = 1/t$.

Now, we bound (b). Applying Lemma 3.24 and recalling $\eta_{T/2} = T/2$, we obtain

$$\left\|x_{T/2} - x^*\right\|^2 \;\le\; \frac{T}{(T/2-2)(T/2-1)} \sum_{t=2}^{T/2-1} (t-1)\langle \hat{z}_t, x_t - x^* \rangle + \frac{\left\|\hat{g}_{T/2-1}\right\|^2}{T/2-1}.$$

Now, setting $\zeta = (a) + \frac{\|\hat{g}_{t-1}\|^2}{(T/2-1)}$ and $\xi_t = \frac{T(t-1)}{(T/2-2)(T/2-1)}$ completes the proof of Lemma 3.22. ■

3.4.2 Proof of Lemma 3.23

Proof (of Lemma 3.23). Applying Lemma 3.24 and rearranging the sum on the first line, we have

$$\begin{aligned}
\sum_{t=2}^{T/2-1} \xi_t^2 \|x_t - x^*\|^2 \;&\le\; \xi_2^2 \|x_2 - x^*\|^2 + \sum_{t=3}^{T/2-1} \xi_t^2 \left(\frac{2}{(t-2)(t-1)} \sum_{i=2}^{t-1} (i-1)\langle \hat{z}_i, x_i - x^* \rangle + \frac{\|\hat{g}_{t-1}\|^2}{t-1} \right) \\
&= \underbrace{\xi_2^2 \|x_2 - x^*\|^2 + \sum_{t=3}^{T/2-1} \frac{\xi_t^2 \|\hat{g}_{t-1}\|^2}{(t-1)}}_{:=s_1} \\
&\quad + \sum_{i=2}^{T/2-2} (i-1) \left(\sum_{t=i+1}^{T/2-1} \frac{\xi_t^2}{(t-2)(t-1)} \right) \cdot \langle \hat{z}_i, x_i - x^* \rangle
\end{aligned}$$

One may use 1-Lipschitz and 1 strong convexity to bound $\|x_2 - x^*\|^2$ by $O(1)$. Using Lipschitzness and the assumption that $\|\hat{z}_t\| \le 1$ almost surely, and recalling that $\xi_t = \Theta(t/T)$, we can bound s_1 by $O(1)$.

Next, applying Lemma 3.24 and rearranging and splitting the sum on the first line below we have again we have

$$\begin{aligned}
\sum_{t=T/2}^{T} \|x_t - x^*\|^2 \;&\le\; \sum_{t=T/2}^{T} \frac{2}{(t-2)(t-1)} \sum_{i=2}^{t-1} (i-1)\langle \hat{z}_i, x_i - x^* \rangle + \sum_{t=T/2}^{T} \frac{\|\hat{g}_{t-1}\|^2}{t-1} \\
&= \underbrace{\sum_{t=T/2}^{T} \frac{\|\hat{g}_{t-1}\|^2}{t-1}}_{:=s_2} + \sum_{i=2}^{T/2-1} (i-1) \left(\sum_{t=T/2}^{T} \frac{2}{(t-2)(t-1)} \right) \cdot \langle \hat{z}_i, x_i - x^* \rangle \\
&\quad + \sum_{t=T/2}^{T} (i-1) \left(\sum_{t=i+1}^{T} \frac{2}{(t-1)(t-2)} \right) \cdot \langle \hat{z}_i, x_i - x^* \rangle.
\end{aligned}$$

Notice that $s_2 = O(1)$ by bounding $\|\hat{g}_{t-1}\|^2$ using Lipschitzness of f and the assumption that $\|\hat{z}_t\| \leq 1$ almost surely. Therefore we have,

$$\begin{aligned}
&\sum_{t=2}^{T/2-1} \xi_t^2 \|x_t - x^*\|^2 + \sum_{t=t/2}^{T} \|x_t - x^*\|^2 \\
&\leq\ s_1 + s_2 + \sum_{i=2}^{T/2-2} (i-1) \left(\sum_{t=i+1}^{T/2-1} \frac{\xi_t^2}{(t-2)(t-1)} \right) \cdot \langle \hat{z}_i, x_i - x^* \rangle \\
&\quad + \sum_{i=2}^{T/2-1} (i-1) \left(\sum_{t=T/2}^{T} \frac{2}{(t-2)(t-1)} \right) \cdot \langle \hat{z}_i, x_i - x^* \rangle \\
&\quad + \sum_{t=T/2}^{T} (i-1) \left(\sum_{t=i+1}^{T} \frac{2}{(t-1)(t-2)} \right) \cdot \langle \hat{z}_i, x_i - x^* \rangle.
\end{aligned}$$

We rewrite the right hand side of the above inequality as

$$\sigma + \sum_{i=2}^{T/2-1} \tilde{\sigma}_i \langle \hat{z}_i, x_i - x^* \rangle + \sum_{i=T/2}^{T} \sigma_i \langle \hat{z}_i, x_i - x^* \rangle$$

where $\sigma = s_1 + s_2$, $\sigma_i = (i-1) \left(\sum_{t=i+1}^{T} \frac{2}{(t-1)(t-2)} \right)$ for $i = T/2, \ldots, T$ and

$$\tilde{\sigma}_i := \begin{cases} (i-1) \left(\sum_{t=i+1}^{T/2-1} \frac{\xi_t^2}{(t-2)(t-1)} + \sum_{t=T/2}^{T} \frac{2}{(t-2)(t-1)} \right) & \text{for } i = 2, \ldots, T/2-2, \\ (i-1) \left(\sum_{t=T/2}^{T} \frac{2}{(t-2)(t-1)} \right) & \text{for } i = T/2-1. \end{cases}$$

Note that $\sigma_i = O(1)$ for all $i = T/2, \ldots, T$ and $\sigma = O(1)$ as well (recalling that s_1 and s_2 are $O(1)$). Next, using $\xi_i = \Theta(i/T)$ one can verify that $\tilde{\sigma}_i = O(i/T)$. Again, because $\xi_i = \Theta(i/T)$, this means that we can write $\tilde{\sigma}_i = \sigma_i \xi_i$ for some $\sigma_i = O(1)$, completing the proof of Lemma 3.23.

■

3.5 High probability bound on a non-uniform averaging scheme

In this section, we prove that a non-uniform averaging scheme attains the optimal $O(1/T)$ rate with high probability. Recall that we have already seen that the suffix average attains the optimal $O(1/T)$ with high probability in Section 3.4. The averaging scheme was first proposed and analyzed in expectation by Lacoste-Julien et al. [26]. The analysis of the final iterate, suffix-average, and this non-uniform averaging scheme can all be seen as applications of the Generalized Freedman inequality, Theorem 1.11.

Theorem 3.25. *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set. Suppose that $f : \mathcal{X} \to \mathbb{R}$ is μ-strongly convex (with respect to $\|\cdot\|_2$) and L-Lipschitz. Assume that:*

(a) *$g_t \in \partial f(x_t)$ for all t (with probability 1).*

(b) *$\|\hat{z}_t\| \leq L$ (with probability 1).*

Set $\eta_t = \frac{2}{\mu(t+1)}$ *and* $\gamma_t = \frac{t}{T(T+1)/2}$. *Then, for any* $\delta \in (0,1)$, *with probability at least* $1-\delta$,

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

3.5.1 Main idea of proof of Theorem 3.25

The main idea of the proof is to follow the in-expectation analysis by Lacoste-Julien et al. [26], but keep track of the mean zero "noise" terms which appear due to the stochastic nature of the subgradient oracle. In particular, Lacoste-Julien et al. [26] repeatedly use the subgradient inequality for μ-strongly-convex functions (see Eq. (1.1)):

$$f(x_t) - f(x^*) \;\leq\; \langle\, g_t, x_t - x^* \,\rangle - \frac{\mu}{2}\,\|x_t - x^*\|_2^2,$$

where x_t is the current iterate, x^* is the minimizer, and $g_t \in \partial f(x_t)$. The subgradient oracle used by SGD produces $\hat{g}_t$ which is not necessarily a member of $\partial f(x_t)$, and therefore the above inequality is not valid. However, since $\mathrm{E}[\hat{g}_t] \in \partial f(x_t)$, we may write $\hat{g}_t = g_t - \hat{z}_t$, where $\mathrm{E}[\hat{z}_t \mid \mathscr{F}_{t-1}] = 0$. Therefore, we obtain a "noisy" subgradient inequality:

$$\begin{aligned} &f(x_t) - f(x^*) \\ &\leq\; \langle\, g_t, x_t - x^* \,\rangle - \frac{\mu}{2}\,\|x_t - x^*\|_2^2 + \langle\, \hat{z}_t, x_t - x^* \,\rangle. \end{aligned} \tag{3.8}$$

Now, applying the analysis of Lacoste-Julien et al. [26] and replacing applications of the subgradient inequality with Eq. (3.8) we obtain the following inequality.

Claim 3.26. *Let* $Z_T = \sum_{t=1}^{T} t\langle\, \hat{z}_t, x_t - x^* \,\rangle$. *Then,*

$$f\left(\sum_{t=1}^{T} \gamma_t x_t\right) - f(x^*) \;\leq\; O\left(\frac{L^2}{\mu T}\right) + O\left(\frac{1}{T^2}\right) Z_T.$$

Therefore, the main challenge lies in bounding Z_T by $O(T)$ with high probability. Notice that Z_T is the sum of conditionally mean-zero increments, making it the value of a martingale at time T. The standard Azuma inequality would require an *almost sure* bound of $O(1/T)$ on its total variance, denoted $V_T = O\left(\sum_{t=1}^{T} t^2 \|x_t - x^*\|_2^2\right)$, in order to obtain the desired bound on Z_T.

On the other hand, the classic Freedman inequality would require a *high probability* bound of $O(1/T)$ on V_T. The most obvious way to do this is come up with tail bounds on $\|x_t - x^*\|^2$ directly and this is the approach taken by Rakhlin et al. [36], which led to the suboptimal $O(\log(\log(T))/T)$ convergence rate for the suffix-averaging strategy.

Instead, we turn to the Generalized Freedman Inequality by Harvey et al. [15], which can help us derive the desired high-probability bound if we are able to prove that V_T satisfies some specific recursive structure with Z_T – that is, $V_T \leq O(T)Z_T + O(T^2)$. The advantage of this approach is that, unlike the approach of applying the classic Freedman inequality, we no longer require a tail bound on $\|x_t - x^*\|^2$. Instead, we can achieve our goal by using elementary analysis. This results in an elegant proof of our main result using a technique which we

believe can be broadly applicable to other first order stochastic optimization procedures.

3.5.2 High probability upper bound analysis

Assuming Claim 3.26 holds, it remains to bound the quantity $Z_T = \sum_{t=1}^T t \langle \hat{z}_t, x_t - x^* \rangle$ with high probability. The strategy here will be to show that $V_T \coloneqq L^2 \sum_{t=1}^T t^2 \|x_t - x^*\|^2$ (an upper bound on the sum of the squared increments of Z_T), satisfies a recursive property involving Z_T again. In particular, we will show that V_T is bounded by a linear transformation of Z_T with probability one. Then, applying Theorem 3.21 gives

Lemma 3.27. *Let $Z_T = \sum_{t=1}^T t \cdot \langle \hat{z}_t, x_t - x^* \rangle$. Then for any $\delta \in (0,1)$, $Z_T \le O\left(\frac{L^2}{\mu} \cdot T \log(1/\delta)\right)$, with probability at least $1-\delta$.*

Lemma 3.27 and Claim 3.26 together prove Theorem 3.25. The proof of Claim 3.26 can be found in Subsection 3.5.3 and the proof of Lemma 3.27 can be found in Subsection 3.5.4.

3.5.3 Proof of Claim 3.26

The proof follows carefully the analysis of Lacoste-Julien et al. [26], but we must be careful with the noise terms as our goal is obtain a high probability bound.

Proof (of Claim 3.26). We write $\hat{z}_t = g_t - \hat{g}_t$. Since f is μ-strongly convex, we have

$$\begin{aligned} f(x_t) - f(x^*) &\le \langle g_t, x_t - x^* \rangle - \frac{\mu}{2} \|x_t - x^*\|_2^2 \\ &= \langle \hat{g}_t, x_t - x^* \rangle - \frac{\mu}{2} \|x_t - x^*\|_2^2 \\ &\quad + \langle \hat{z}_t, x_t - x^* \rangle. \end{aligned}$$

The first two terms can be bounded as follows.

$$\begin{aligned} &\langle \hat{g}_t, x_t - x^* \rangle - \frac{\mu}{2} \|x_t - x^*\|_2^2 \\ &= \frac{1}{\eta_t} \langle x_t - y_{t+1}, x_t - x^* \rangle - \frac{\mu}{2} \|x_t - x^*\|_2^2 \\ &= \frac{1}{2\eta_t} \Big(\|x_t - y_{t+1}\|_2^2 + \|x_t - x^*\|_2^2 - \|y_{t+1} - x^*\|_2^2 \Big) - \frac{\mu}{2} \|x_t - x^*\|_2^2 \\ &\le \frac{1}{2\eta_t} \Big(\|x_t - y_{t+1}\|_2^2 + \|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2 \Big) - \frac{\mu}{2} \|x_t - x^*\|_2^2 . \end{aligned}$$

The first line uses the definition of Algorithm 1 and the last line uses a property of Euclidean projections: since x_{t+1} is the projected point $\Pi_{\mathscr{X}}(y_{t+1})$ and $x^* \in \mathscr{X}$, we have $\|x_{t+1} - x^*\|_2^2 \le \|y_{t+1} - x^*\|_2^2$.

It is convenient to scale by t in order to later obtain a telescoping sum. Using the definition of the gradient

step, i.e. $x_t - y_{t+1} = \eta_t \hat{g}_t$, we have

$$\begin{aligned}
& t \cdot \Big(f(x_t) - f(x^*) - \langle\, \hat{z}_t, x_t - x^* \,\rangle \Big) \\
&\leq \frac{t \left\| \eta_t \hat{g}_t \right\|_2^2}{2\eta_t} + t\Big(\frac{1}{2\eta_t} - \frac{\mu}{2}\Big) \left\| x_t - x^* \right\|_2^2 - \frac{t}{2\eta_t} \left\| x_{t+1} - x^* \right\|_2^2 \\
&= \frac{t \left\| \hat{g}_t \right\|_2^2}{\mu(t+1)} + \Big(\frac{\mu t(t+1)}{4} - \frac{2\mu t}{4}\Big) \left\| x_t - x^* \right\|_2^2 - \frac{t(t+1)\mu}{4} \left\| x_{t+1} - x^* \right\|_2^2 \\
&\leq \frac{(2L)^2}{\mu} + \frac{\mu}{4} \cdot \Big(t(t-1) \left\| x_t - x^* \right\|_2^2 - t(t+1) \left\| x_{t+1} - x^* \right\|_2^2 \Big).
\end{aligned}$$

Now, summing over t, the right-hand side telescopes and we obtain $\sum_{t=1}^{T} t \cdot (f(x_t) - f(x^*))$ is bounded above by

$$\sum_{t=1}^{T} t \cdot \langle\, \hat{z}_t, x_t - x^* \,\rangle + \frac{4L^2 \cdot T}{\mu}.$$

Dividing by $T(T+1)/2$ and applying Jensen's inequality, we obtain that

$$\begin{aligned}
f\Big(\sum_{t=1}^{T} \gamma_t x_t \Big) - f(x^*) &\leq \sum_{t=1}^{T} \gamma_t \cdot \big(f(x_t) - f(x^*) \big) \\
&\leq \frac{2}{T(T+1)} \underbrace{\sum_{t=1}^{T} t \cdot \langle\, \hat{z}_t, x_t - x^* \,\rangle}_{=Z_T} + \frac{8L^2}{\mu(T+1)}.
\end{aligned}$$

■

3.5.4 Bounding Z_T

Observe that Z_T is a sum of a martingale difference sequence. Define $d_t = t \cdot \langle\, \hat{z}_t, x_t - x^* \,\rangle$, $v_{t-1} := t^2 \left\| x_t - x^* \right\|$, and $V_T = L^2 \sum_{t=1}^{T} v_{t-1}$. Note that v_{t-1} is $\mathscr{F}_{t-1}$-measurable. The next claim shows that v_{t-1} and d_t satisfy the assumptions of Generalized Freedman's inequality (Theorem 3.21).

Claim 3.28. *For all t and $\lambda > 0$, we have* $\mathrm{E}\left[\exp(\lambda d_t) \mid \mathscr{F}_{t-1} \right] \leq \exp\Big(\frac{\lambda^2}{2} L^2 \cdot v_{t-1} \Big)$.

Proof. First, we can apply Cauchy-Schwarz to get that $|t \langle\, \hat{z}_t, x_t - x^* \,\rangle| \leq t \cdot \left\| \hat{z}_t \right\| \cdot \left\| x_t - x^* \right\| \leq t \cdot L \cdot \left\| x_t - x^* \right\|$ because $\left\| \hat{z}_t \right\| \leq L$ a.s. Next, applying Hoeffding's Lemma ([30] - Lemma 2.6), we have $\mathrm{E}\left[\exp\left(\lambda t \langle\, \hat{z}_t, x_t - x^* \,\rangle \right) \mid \mathscr{F}_{t-1} \right] \leq \exp\Big(\frac{\lambda^2}{2} L^2 \cdot t^2 \left\| x_t - x^* \right\|^2 \Big)$. ■

To bound Z_T, we will show that we can bound its sum of squared magnitudes (SSCM) by a linear combination of the increments. This will allow us to use the Generalized Freedman Inequality (Theorem 3.21).

Lemma 3.29. *There exists non-negative constants $\alpha_1, \ldots, \alpha_T$ such that* $\max_{i \in [T]} \{\alpha_i\} = O\Big(\frac{L^2}{\mu} \cdot T \Big)$ *and* $\beta = O\Big(\frac{L^4}{\mu^2} \cdot T^2 \Big)$ *such that* $V_T \leq \sum_{t=1}^{T} \alpha_t d_t + \beta$.

Proof (of Lemma 3.27). By Claim 3.28, we have $\mathrm{E}\left[\exp(\lambda d_t) \mid \mathscr{F}_{t-1} \right] \leq \exp\Big(\frac{\lambda^2}{2} L^2 \cdot v_{t-1} \Big)$ for all $\lambda > 0$ and $V_T = \sum_{t=1}^{T} L^2 \cdot v_{t-1}$. By Lemma 3.29, we have $V_T \leq \sum_{t=1}^{T} \cdot \alpha_t d_t + \cdot \beta$. Plugging $\alpha = O\Big(\frac{L^2}{\mu} \cdot T \Big)$, $\beta = O\Big(\frac{L^4}{\mu^2} \cdot T^2 \Big)$, and $x = O\Big(\frac{L^2}{\mu} \cdot T \log(1/\delta) \Big)$ into Theorem 3.21 proves the lemma. ■

It remains to prove Lemma 3.29. To do so, we will need the following two lemmata, which are adapted from [36] to use the step sizes $\eta_t = \frac{2}{\mu(t+1)}$. For completeness, we provide a proof in the supplementary material.

Lemma 3.30 ([36] - Lemma 5)**.** *With probability* 1*, and for all t,* $\|x_t - x^*\| \leq \frac{2L}{\mu}$.

Lemma 3.31 ([36] - Lemma 6)**.** *For all* $t \geq 3$*, there exists non-negative numbers* $a_1(t), \ldots, a_t(t)$ *with* $a_i(t) = \Theta(i^3/t^4)$ *and* $b_1(t), \ldots, b_t(t)$ *with* $b_i(t) = \Theta(i^2/t^4)$*, such that with probability* 1 $\|x_{t+1} - x^*\|^2$ *is bounded above by*

$$\frac{4}{\mu}\sum_{i=3}^{t} a_i(t)\langle \hat{z}_i, x_i - x^* \rangle + \frac{4}{\mu^2}\sum_{i=3}^{t} b_i(t)\|\hat{g}_t\|^2.$$

Remark 3.32. *Lemma 3.30 and Lemma 3.31 are true regardless of the assumption we place on* $\hat{z}_t$*.*

Proof (of Lemma 3.29). Recall $\|\hat{g}_i\| \leq 2L$ because f is L-Lipschitz and $\|\hat{z}_i\| \leq L$ almost surely. By Lemma 3.30 and Lemma 3.31, we bound $\sum_{t=1}^{T} v_{t-1}$ and then multiply through by L^2 to obtain a bound on V_T:

$$\begin{aligned}
\sum_{t=1}^{T} v_{t-1} &= \sum_{t=1}^{T} t^2 \cdot \|x_t - x^*\|^2 \\
&\leq \frac{56L^2}{\mu^2} + \sum_{t=4}^{T} t^2 \left(\frac{4}{\mu}\sum_{i=3}^{t-1} a_i(t-1)\langle \hat{z}_i, x_i - x^* \rangle \right) + \frac{4}{\mu^2}\sum_{t=4}^{T}\left(\sum_{i=3}^{t-1} b_i(t-1)\|\hat{g}_i\|^2 \right) \\
&\leq \frac{56L^2}{\mu^2} + \sum_{t=4}^{T} t^2 \left(\frac{4}{\mu}\sum_{i=3}^{t-1} a_i(t-1)\langle \hat{z}_i, x_i - x^* \rangle \right) + \frac{4(2L)^2}{\mu^2}\sum_{t=4}^{T}\left(\sum_{i=3}^{t-1} b_i(t-1) \right) \\
&= \frac{4}{\mu}\sum_{t=4}^{T} t^2 \left(\sum_{i=3}^{t-1} a_i(t-1)\langle \hat{z}_i, x_i - x^* \rangle \right) + \frac{16 \cdot L+^2}{\mu^2}\sum_{t=4}^{T} t^2 \left(\sum_{i=3}^{t-1} b_i(t-1) \right) + \frac{56L^2}{\mu^2} \\
&= \sum_{i=3}^{T-1} \underbrace{\frac{4}{\mu}\left(\sum_{t=i+1}^{T} t^2 \cdot \frac{a_i(t-1)}{i} \right)}_{:=\alpha_i} \cdot i\langle \hat{z}_i, x_i - x^* \rangle + \underbrace{\frac{16 \cdot L^2}{\mu^2}\sum_{t=4}^{T} t^2 \left(\sum_{i=3}^{t-1} b_i(t-1) \right) + \frac{56L^2}{\mu^2}}_{:=\beta}
\end{aligned}$$

Define $\alpha_1, \alpha_2, \alpha_T = 0$. We have $V_T \leq \sum_{i=1}^{T} \alpha_i \cdot i \cdot \langle \hat{z}_i, x_i - x^* \rangle + \beta$. It remains to show $\max\{\alpha_i\} = O\left(\frac{T}{\mu}\right)$ and $\beta = O\left(\frac{L^2}{\mu^2}T^2\right)$ (multiplying through by L^2 yields the bound on V_T). To bound $\max\{\alpha_i\}$, observe that for $i \in \{3, \ldots, T-1\}$,

$$\begin{aligned}
\sum_{t=i+1}^{T} t^2 \cdot \frac{a_i(t-1)}{i} &= \sum_{t=i+1}^{T} t^2 O\left(\frac{i^2}{t^4}\right) \\
&= \sum_{t=i+1}^{T} t^2 O\left(\frac{1}{t^2}\right) = O(T-i).
\end{aligned}$$

To bound β, observe

$$\begin{aligned}\sum_{t=4}^{T} t^2 \left(\sum_{i=3}^{t-1} b_i (t-1) \right) &= \sum_{t=4}^{T} t^2 \sum_{i=3}^{t-1} O\left(\frac{i^2}{t^4}\right) \\ &= \sum_{t=4}^{T} t^2 \sum_{i=3}^{t-1} O\left(\frac{1}{t^2}\right) \\ &= \sum_{t=4}^{T} O(t) \;=\; O(T^2).\end{aligned}$$

■

3.5.5 Missing proofs from Subsection 3.5.4

Both of the proofs in this section are slight modifications of the proofs found by Rakhlin et al. [36].

Proof (of Lemma 3.30). Due to strong convexity and the fact that $f(x_t) - f(x^*) \geq 0$, we have

$$\begin{aligned}L \|x_t - x^*\| &\geq \|g_t\| \|x_t - x^*\| \\ &\geq \langle g_t, x_t - x^* \rangle \\ &\geq \frac{\mu}{2} \|x_t - x^*\|^2,\end{aligned}$$

where we used L-Lipschitzness of f to bound $\|g_t\|$ by L. ■

Proof (of Lemma 3.31). The definition of strong convexity yields

$$\langle g_t, x_t - x^* \rangle \geq f(x_t) - f(x^*) + \frac{\mu}{2} \|x_t - x^*\|^2.$$

Strong convexity and the fact that $0 \in \partial f(x^*)$ implies

$$f(x_t) - f(x^*) \geq \frac{\mu}{2} \|x_t - x^*\|^2.$$

Next, recall that for any $x \in \mathscr{X}$, and for any z, we have $\|\Pi_{\mathscr{X}}(z) - x\| \leq \|z - x\|$. Lastly, recall $\eta_t = \frac{2}{\mu(t+1)}$. Using these, we have

$$\begin{aligned}&\|x_{t+1} - x^*\|^2 \\ = \|\Pi_{\mathscr{X}}(x_t - \eta_t \hat{g}_t) - x^*\|^2 &\leq \|x_t - \eta_t \hat{g}_t - x^*\|^2 \\ &= \|x_t - x^*\|^2 - 2\eta_t \langle \hat{g}_t, x_t - x^* \rangle + \eta_t^2 \|\hat{g}_t\|^2 \\ &= \|x_t - x^*\|^2 - 2\eta_t \langle g_t, x_t - x^* \rangle + 2\eta_t \langle \hat{z}_t, x_t - x^* \rangle \\ &\quad + \eta_t^2 \|\hat{g}_t\|^2 \\ &\leq \|x_t - x^*\|^2 - 2\eta_t (f(x_t) - f(x^*)) - \eta_t \mu \|x_t - x^*\|^2 + 2\eta_t \langle \hat{z}_t, x_t - x^* \rangle + \eta_t^2 \|\hat{g}_t\|^2 \\ &\leq (1 - 2\eta_t \mu) \|x_t - x^*\|^2 + 2\eta_t \langle \hat{z}_t, x_t - x^* \rangle + \eta_t^2 \|\hat{g}_t\|^2 \\ &= \left(1 - \frac{4}{t+1}\right) \|x_t - x^*\|^2 + \frac{4}{\mu(t+1)} \langle \hat{z}_t, x_t - x^* \rangle + \frac{4}{\mu^2 (t+1)^2} \|\hat{g}_t\|^2. \qquad (3.9)\end{aligned}$$

Repeatedly applying Eq. (3.9) until $t = 4$, yields the following upper bound on $\|x_{t+1} - x^*\|^2$

$$\frac{4}{\mu}\sum_{i=4}^{t}\left[\frac{1}{i+1}\prod_{j=i+1}^{t}\left(1-\frac{4}{j+1}\right)\right]\cdot\langle\,\hat{z}_i,\,x_i-x^*\,\rangle+\frac{4}{\mu^2}\sum_{i=4}^{t}\left[\frac{1}{(i+1)^2}\prod_{j=i+1}^{t}\left(1-\frac{4}{j+1}\right)\right]\cdot\|\hat{g}_t\|\,.$$

Observing that

$$\begin{aligned}\prod_{j=i+1}^{t}\left(1-\frac{4}{j+1}\right) &= \prod_{j=i+1}^{t}\frac{j-3}{j+1}\\ &= \frac{(i-2)\cdot(i-1)\cdot i\cdot(i+1)}{(t-2)\cdot(t-1)\cdot t\cdot(t+1)},\end{aligned}$$

proves the lemma by taking $a_i(t) = \frac{1}{i+1}\cdot\frac{(i-2)\cdot(i-1)\cdot i\cdot(i+1)}{(t-2)\cdot(t-1)\cdot t\cdot(t+1)}$ and $b_i(t) = \frac{1}{(i+1)^2}\cdot\frac{(i-2)\cdot(i-1)\cdot i\cdot(i+1)}{(t-2)\cdot(t-1)\cdot t\cdot(t+1)}$ ■

Chapter 4

Probabilistic Tools

4.1 Proof of Theorem 1.11 and corollaries

In this section we prove Theorem 1.11 and derive some corollaries. We restate Theorem 1.11 here for convenience.

Theorem 1.11. Let $\{d_i, \mathscr{F}_i\}_{i=1}^n$ be a martingale difference sequence. Suppose v_{i-1}, for $i \in [n]$, are positive and $\mathscr{F}_{i-1}$-measurable random variables such that $\mathrm{E}[\exp(\lambda d_i) \mid \mathscr{F}_{i-1}] \leq \exp\left(\frac{\lambda^2}{2} v_{i-1}\right)$ 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

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

The proof of Theorem 1.11 is inspired by a standard proof of Freedman's inequality. The standard proof begins with the observation that $M_t(\lambda) \coloneqq \exp\left(\lambda S_t - \frac{\lambda^2}{2} V_t\right)$ defines a supermartingale for any $\lambda > 0$ (recall that S_t is the martingale at time t and V_t is the sum of squared conditional subgaussian norms (SSCSN) at time n). Then, noting that $\{S_n \geq x \text{ and } V_n \leq y\} \subset \left\{\lambda S_n - \frac{\lambda^2}{2} V_n \geq \lambda x - \frac{\lambda^2}{2} y\right\}$ for any $\lambda > 0$ allows one to apply the exponentiated Markov inequality (Lemma A.1) to the larger event to obtain $\Pr[S_n \geq x, V_n \leq y] \leq \exp\left(-\lambda x + \frac{\lambda^2}{2} y\right) \mathrm{E}[M_n]$. Since $(M_t)_{t \in \mathbb{N}}$ is a supermartingale with $M_0 = 1$, minimizing over λ yields a simplified version of Freedman's inequality. To obtain Freedman's inequality in full generality, one may apply a modification of the above argument involving the stopping time $\tau = \min\{\, t : S_t \geq x, V_t \leq y \,\}$.

In our setting, we also aim to design a supermartingale whose role is similar to the supermartingale $M_t(\lambda)$ plays in the proof above. However, due to the entanglement between our martingale and its SSCSN process in the event whose probability we aim to bound (i.e. the chicken and egg phenomenon), some special care is required in order to derive such a supermartingale. Indeed, as we will see in the following proof, we will need to carefully balance the parameters c and $\tilde{\lambda}$.

Proof (of Theorem 1.11). Suppose $0 \leq \lambda < 1/(2\alpha)$; the actually choice of λ will be optimized later. Define

$c = c(\lambda, \alpha)$ as in Claim 4.2. Let $\tilde{\lambda} = \lambda + c\lambda^2\alpha$. Define $\mathscr{U}_0 \coloneqq 1$ and for $t \in [n]$, define

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

Claim 4.1. *$\mathscr{U}_t(\lambda)$ is a supermartingale w.r.t. $\mathscr{F}_t$.*

Explanation of the role of $\mathscr{U}_t(\lambda)$ and why it differs from $M_t(\lambda)$. Notice that the supermartingale $\mathscr{U}_t(\lambda)$ differs from $M_t(\lambda)$. The scaling on the increments d_i in the definition of $\mathscr{U}_t(\lambda)$ are specifically designed with the goal of applying the exponentiated Markov inequality (Lemma A.1) to expose the expectation of a nonnegative supermartingale with bounded initial value. This was a key step in the proof of Freedman's inequality described above. In our setting, recalling that $S_t = \sum_{i=1}^{t} d_i$, we have

$$\underbrace{\left\{\sum_{i=1}^{n}(\lambda + c\lambda^2\alpha_i)d_i - c\lambda^2 V_t \geq \lambda x - c\lambda^2\beta\right\}}_{\coloneqq\mathscr{E}} \supset \underbrace{\left\{S_t \geq x, V_t \leq \sum_{i=1}^{n}\alpha_i d_i + \beta\right\}}_{\coloneqq\mathscr{G}},$$

where the scaling on the increments d_i in the event $\mathscr{E}$ matches the scaling in $\mathscr{U}_t(\lambda)$ and the event $\mathscr{G}$ is essentially the event we would like to analyze. Notice that $\mathscr{E}$ is an event which we may apply the exponentiated Markov inequality to. Ideally, an application of Lemma A.1 on the event $\mathscr{E}$ would yield an upper bound of $\exp(-\lambda x + c\lambda^2\beta)$ times the expectation of a nonnegative supermartingale with bounded initial value as was the case in the standard proof outlined above. Instead, Lemma A.1 yields the expectation of the following process $\exp\left(A_n - c\lambda^2 V_n\right)$, where $A_n = \sum_{i=1}^{n}(\lambda + c\lambda^2\alpha_i)d_i$, and it is unclear whether this is a supermartingale. Now, according to Claim 4.1 $\exp\left(A_n - \frac{\tilde{\lambda}^2}{2}V_n\right)$ is a supermartingale. In order to enforce that Lemma A.1 yields an upper bound on the probability of $\mathscr{E}$ involving the expectation of a supermartingale, we will enforce $c\lambda^2 = \frac{\tilde{\lambda}^2}{2}$ so that the expectation of the random process which results from applying Lemma A.1 on the event $\mathscr{E}$ is equal to $\mathrm{E}[\mathscr{U}_n(\lambda)]$. Formally,

$$\begin{aligned}
\Pr[\mathscr{G}] \;\leq\; \Pr[\mathscr{E}] \;&\leq\; \exp\left(-\lambda x - c\lambda^2\beta\right)\mathrm{E}\left[\exp\left(A_n - c\lambda^2 V_n\right)\right] \qquad \text{(by Lemma A.1)}\\
&=\; \exp\left(-\lambda x - c\lambda^2\beta\right)\mathrm{E}[\mathscr{U}_n(\lambda)] \qquad \text{(by choice of } c\text{)}\\
&\leq\; \exp\left(-\lambda x - c\lambda^2\beta\right) \qquad (\mathscr{U}_t(\lambda) \text{ is a supermartingale}).
\end{aligned}$$

Proof (of Claim 4.1). For all $t \in [n]$:

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

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

Define the stopping time $T = \min\{\, t : S_t \ge x \text{ and } V_t \le \sum_{i=1}^{t} \alpha_i d_i + \beta \,\}$ with the convention that $\min\emptyset = \infty$. Since $\mathscr{U}_t$ is a supermartingale w.r.t. $\mathscr{F}_t$, then $\mathscr{U}_{T\wedge t}$ is a supermartingale w.r.t. $\mathscr{F}_t$ ([25, Theorem 10.15]). Recall that $S_{T\wedge n} = \sum_{i=1}^{T\wedge n} d_i$.

$$\begin{aligned}
\Pr\left[\bigcup_{t=1}^{n}\left\{S_t \ge x \text{ and } V_t \le \sum_{i=1}^{t}\alpha_i d_i + \beta\right\}\right] &= \Pr\left[S_{T\wedge n} \ge x \text{ and } V_{T\wedge n} \le \sum_{i=1}^{T\wedge n}\alpha_i d_i + \beta\right] \\
&= \Pr\Bigg[\underbrace{\lambda S_{T\wedge n} \ge \lambda x \text{ and } c\lambda^2 V_{T\wedge n} \le c\lambda^2 \sum_{i=1}^{T\wedge n}\alpha_i d_i + c\lambda^2\beta}_{:=\mathscr{A}}\Bigg] \\
\text{(subtracting inequalities from } \mathscr{A}\text{)} \quad &\le \Pr\left[\sum_{i=1}^{T\wedge n}(\lambda + c\lambda^2\alpha_i)d_i - c\lambda^2 V_{T\wedge n} \ge \lambda x - c\lambda^2\beta\right] \\
\text{(by Lemma A.1)} \quad &\le \mathrm{E}\left[\exp\left(\sum_{i=1}^{T\wedge n}(\lambda + c\lambda^2\alpha_i)d_i - c\lambda^2 V_{T\wedge n}\right)\right]\cdot\exp(-\lambda x + c\lambda^2\beta).
\end{aligned}$$

Recall that c was chosen (via Claim 4.2) so that $c\lambda^2 = (\lambda + c\lambda^2\alpha)^2/2 = \tilde{\lambda}^2/2$. This selection of c is crucial to connect the result of the application of Lemma A.1 above to the supermartingale $\mathscr{U}_t(\lambda)$. Hence,

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

where the inequality is by the optional stopping theorem for bounded stopping times ([25, Theorem 10.11]). 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 4.2. *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 $c = 1/2$). 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. One root of p is located at

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

Set $\gamma = \alpha\lambda$, and observe that $2\gamma \in [0,1]$. Using the numeric inequality $\sqrt{1-x} \ge 1 - x/2 - x^2/2$ valid for all $x \le 1$, we have

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

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

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

■

4.1.1 Corollaries of Theorem 1.11

In this thesis, we often deal with martingales, M_n, where the sum of squared conditional magnitudes (SSCM) 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 SSCM of M_n involves M_n itself). Transforming these entangled high probability bounds on the SSCM into high probability bounds on the martingale itself are easy consequences of our Generalized Freedman inequality (Theorem 1.11).

Lemma 4.3. *Let $\{d_i, \mathscr{F}_i\}_{i=1}^n$ be a martingale difference sequence. Let v_{i-1} be a $\mathscr{F}_{i-1}$ measurable random variable such that $\mathrm{E}[\exp(\lambda d_i) \mid \mathscr{F}_{i-1}] \le \exp\left(\frac{\lambda^2}{2}v_{i-1}\right)$ for all $\lambda > 0$ and for all $i \in [n]$. Define $S_n = \sum_{i=1}^n d_i$*

and define $V_n = \sum_{i=1}^{n} v_{i-1}$. *Suppose that there exists* $\alpha_1, \ldots, \alpha_n \geq 0$ *and* $R(\delta) > 0$ *(for* $\delta \in (0,1)$*) such that* $\Pr[V_n \leq \sum_{i=1}^{n} \alpha_i d_i + R(\delta)] \geq 1 - \delta$. *Let* $\alpha = \max_{i \in [n]} \alpha_i$. *Then,*

$$\Pr[S_n \geq x] \leq \delta + \exp\left(-\frac{x^2}{4\alpha x + 8R(\delta)}\right) \qquad \forall x > 0.$$

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

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

where the final inequality is due to applying Theorem 1.11 to $\Pr[\mathscr{E}(x) \wedge \mathscr{G}]$. ■

In this thesis, we use Lemma 4.3 via the following two corollaries. The first Corollary can be proven by using the original Freedman's inequality.

Corollary 4.4. *Let* $\{\mathscr{F}_t\}_{t=1}^{T}$ *be a filtration and suppose that* a_t *are* $\mathscr{F}_t$*-measurable random vectors and* b_t *are* $\mathscr{F}_{t-1}$*-measurable random vectors. Further, suppose that*

1. $\|a_t\| \leq 1$ *almost surely and* $\mathrm{E}[a_t \mid \mathscr{F}_{t-1}] = 0$*; 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* $\sum_{t=1}^{T} d_t \leq O\left(\sqrt{R}\log(1/\delta)\right)$ *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, $\mathrm{E}\left[\exp\left(\lambda d_t\right) \mid \mathscr{F}_{t-1}\right] \leq \exp\left(\frac{\lambda^2}{2}\|b_t\|^2\right)$ for all λ by Lemma A.5. Next, applying Lemma 4.3 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 4.5. *Let* $\{\mathscr{F}_t\}_{t=1}^{T}$ *be a filtration and suppose that* a_t *are* $\mathscr{F}_t$*-measurable random vectors and* b_t *are* $\mathscr{F}_{t-1}$*-measurable random vectors. Define* $d_t = \langle a_t, b_t \rangle$. *Assume that* $\|a_t\| \leq 1$ *almost surely and* $\mathrm{E}[a_t \mid \mathscr{F}_{t-1}] = 0$. *Furthermore, suppose that there exists* $R > 0$ *and non-negative values* $\{\alpha_t\}_{t=1}^{T-1}$ *where* $\max\{\alpha_t\}_{t=1}^{T-1} = O\left(\sqrt{R}\right)$, *such that for every* δ *sufficiently small, with probability at least* $1 - O(\delta)$, *we have* $\sum_{t=1}^{T} \|b_t\|^2 \leq \sum_{t=1}^{T-1} \alpha_t d_t + R\log(1/\delta)$. *Then* $\sum_{t=1}^{T} d_t \leq O\left(\sqrt{R}\log(1/\delta)\right)$ *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 $\|a_t\| \leq 1$, by Cauchy-Schwarz we have that $|d_t| \leq \|b_t\|$. Therefore, $\mathrm{E}\left[\exp\left(\lambda d_t\right) \mid \mathscr{F}_{t-1}\right] \leq \exp\left(\frac{\lambda^2}{2}\|b_t\|^2\right)$ for all λ by Lemma A.5. Next, applying Lemma 4.3 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

$$\Pr\left[\sum_{t=1}^{T} d_t \geq x\right] \;\leq\; \delta + \exp\left(-\frac{x^2}{4\left(\max_{t=1}^{T-1}\{\alpha_t\}\right)x + 8R\log(1/\delta)}\right). \tag{4.1}$$

Recall that $\max_{t=1}^{T-1}\{\alpha_t\} = O(\sqrt{R})$. Thus, for some $x = \Theta\left(\sqrt{R}\log(1/\delta)\right)$, the last term in (4.1) is at most δ. Replacing δ with $\delta/2$ completes the proof. ■

4.2 Proof of Theorem 1.19

Theorem 1.19. Let $(X_t)_{t=1}^T$ be a stochastic process and let $(\mathscr{F}_t)_{t=1}^T$ be a filtration such that X_t is $\mathscr{F}_t$ measurable and X_t is non-negative almost surely. Let $\alpha_t \in [0,1)$ and $\beta_t, \gamma_t \geq 0$ for every t. Assume that $\mathrm{E}[\exp(\lambda X_1)] \leq \exp(\lambda K)$ for $\lambda \in (0, 1/K]$ where $K = \max_{1\leq t\leq T}\left(\frac{2\gamma_t}{1-\alpha_t}, \frac{2\beta_t^2}{1-\alpha_t}\right)$. Let $\hat{w}_t$ be a mean-zero random variable conditioned on $\mathscr{F}_t$ 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 $\Pr\left[\sum_{t=1}^T \sigma_t X_t \geq K\log(1/\delta)\sum_{t=1}^T \sigma_t\right] \leq \mathrm{e}\delta$.

Proof (of Theorem 1.19).

We begin by deriving a recursive MGF bound on X_t.

Claim 4.6. *Suppose* $0 \leq \lambda \leq \min_{1\leq t\leq T}\left(\frac{1-\alpha_t}{2\beta_t^2}\right)$. *Then for every* t,

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

Proof. Since X_t is non-negative almost surely, we may define the random variable $X = \beta_t\hat{w}_t\sqrt{X_t}$. Since $|\hat{w}_t| \leq 1$ almost surely and X_t is $\mathscr{F}_t$-measurable, $\mathrm{E}\left[\exp\left(\lambda^2 X^2\right) \mid \mathscr{F}_t\right] \leq \exp\left(\lambda^2\beta_t^2 X_t\right)$ for all λ. Since $\mathrm{E}[\hat{w}_t \mid \mathscr{F}_t] = 0$ we may apply Claim A.6 with $c^2 = \beta_t^2 X_t$ to obtain

$$\mathrm{E}\left[\exp(\lambda\beta_t\hat{w}_t\sqrt{X_t}) \mid \mathscr{F}_t\right] \;=\; \mathrm{E}[\exp(\lambda X) \mid \mathscr{F}_t] \;\leq\; \exp\left(\lambda^2\beta_t^2 X_t\right). \tag{4.2}$$

Hence,

$$\begin{aligned}
\mathrm{E}[\exp(\lambda X_{t+1})] &\leq \mathrm{E}\left[\exp\left(\lambda\alpha_t X_t + \lambda\beta_t\hat{w}_t\sqrt{X_t} + \lambda\gamma_t\right)\right] && \text{(by assumption)}\\
&= \mathrm{E}\left[\exp(\lambda\alpha_t X_t + \lambda\gamma_t)\,\mathrm{E}\left[\exp\left(\lambda\beta_t\hat{w}_t\sqrt{X_t}\right) \mid \mathscr{F}_t\right]\right]\\
&\leq \mathrm{E}\left[\exp\left(\lambda\alpha_t X_t + \lambda^2\beta_t^2 X_t + \lambda\gamma_t\right)\right] && \text{(by Eq. (4.2))}\\
&= \mathrm{E}\left[\exp\left(\lambda X_t\left(\alpha_t + \lambda\beta_t^2\right) + \lambda\gamma_t\right)\right]\\
&\leq \mathrm{E}\left[\exp\left(\lambda\gamma_t + \lambda X_t\left(\frac{1+\alpha_t}{2}\right)\right)\right] && \left(\text{because } \lambda \leq \frac{1-\alpha_t}{2\beta_t^2}\right).
\end{aligned}$$

■

Next, we prove an MGF bound on X_t.

Claim 4.7. *For every t and for all $0 \le \lambda \le 1/K$, $\mathrm{E}[\exp(\lambda X_t)] \le \exp(\lambda K)$.*

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

$$\begin{aligned}\mathrm{E}[\exp(\lambda X_{t+1})] &\le \mathrm{E}\left[\exp\left(\lambda\gamma_t + \lambda X_t\left(\frac{1+\alpha_t}{2}\right)\right)\right] \qquad \text{(by Claim 4.6)}\\ &\le \exp\left(\lambda\gamma_t + \lambda K\left(\frac{1+\alpha_t}{2}\right)\right),\end{aligned}$$

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

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

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

$$\mathrm{E}[\exp(\lambda X_{t+1})] \le \exp(\lambda K),$$

as desired. ■

Now we are ready to complete the proof of both claims in Theorem 1.19. The first claim in Theorem 1.19 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 (with $c = 1$ and $C = K$).

Next, we prove the second claim in Theorem 1.19. Claim 4.7 gives that for every t and for all $\lambda \le 1/(\sigma_t K)$, we have $\mathrm{E}[\exp(\lambda\sigma_t X_t)] \le \exp(\lambda\sigma_t K)$. Hence, we can combine these MGF bounds using Lemma A.4 to obtain $\mathrm{E}\left[\exp\left(\lambda\sum_{t=1}^T \sigma_t X_t\right)\right] \le \exp\left(\lambda K\sum_{t=1}^T \sigma_t\right)$ for all $\lambda \le \left(K\sum_{t=1}^T \sigma_t\right)^{-1}$. 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 in Theorem 1.19. ■

Chapter 5

Infinite Dimensional and Probabilistic Lower Bounds

5.1 Functions attaining large error infinitely often

In order to achieve large error for the T iterate, Theorem 1.12 constructs a strongly convex function parameterized by T. (Similarly, Theorem 1.14 constructs a Lipschitz function parameterized by T.) It is not possible for a *single* function to achieve error $\omega(1/T)$ for *every* $T \geq 1$ in the strongly-convex case (or $\omega(1/\sqrt{T})$ for every $T \geq 1$ in the Lipschitz case). The reason is that this would contradict the fact that suffix averaging achieves error $O(1/T)$ in the strongly-convex case (and that the average over all iterates achieves error $O(1/\sqrt{T})$ in the Lipschitz case). Nevertheless, for every function $g(T) = o(\log(T)/T)$ it possible to construct a strongly convex function which achieves error $C \cdot g(T)$, for every $C > 0$ for infinitely many T. (A similar statement is true for the Lipschitz case.) In the remainder of this section, we use ℓ_2 to denote the space of square-summable sequences in $\mathbb{R}^{\mathbb{N}}$.

Theorem 5.1. *For every non-negative $g(T) = o(\log(T)/T)$, and for any $c > 0$, there exists a convex function $f : \mathscr{X} \to \mathbb{R}$ where $\mathscr{X} \subset \ell_2$, such that f is $(3/c)$-Lipschitz and $(1/c)$-strongly convex, $\inf_{x \in \mathscr{X}} f(x) = 0$ and satisfies the following. Suppose that Algorithm 1 is executed from the initial point $x_1 = 0$ with step sizes $\eta_t = c/t$. Then,*

$$\limsup_{T \to \infty} \frac{f(x_T)}{g(T)} \;=\; +\infty.$$

Theorem 5.2. *For every non-negative $g(T) = o\big(\log(T)/\sqrt{T}\big)$ and for every $c > 0$, there exists a convex function $f : \mathscr{X} \to \mathbb{R}$ where $\mathscr{X} \subset \ell_2$, such that f is $(1/c)$-Lipschitz, $\inf_{x \in \mathscr{X}} f(x) = 0$ and satisfies the following. Suppose that Algorithm 1 is executed from the initial point $x_1 = 0$ with step sizes $\eta_t = c/\sqrt{t}$. Then,*

$$\limsup_{T \to \infty} \frac{f(x_T)}{g(T)} \;=\; +\infty.$$

Remark 5.3. *We provide the proof of Theorem 5.1 in Sub-subsection 5.1.1. The proof of Theorem 5.2 is omitted due to its similarity with the proof of Theorem 5.1. We prove Theorem 5.1 in the setting where $c = 1$. The result follows in full generality as a corollary of the $c = 1$ case and an analogous statement to Lemma 2.1 for Hilbert*

spaces.

Since we are now working with infinite dimensional functions, rigour requires us to state the definition of the subdifferential of a convex function applicable to functions defined on a Hilbert space E.

Definition 5.4 ([5, Definition 2.30])**.** *Let E be a Hilbert space and let f be a proper convex function defined on E. Then, $g \in E^*$ (the dual Hilbert space of E) is a subgradient of f at x if, for all $y \in E$, $f(y) \geq f(x) + \langle y - x, g \rangle$. The subdifferential of f at x, denoted $\partial f(x)$, is the collection of subgradients of f at x.*

The main tool we use to prove Theorem 5.1 and Theorem 5.2 is the following natural result. In a nutshell, Lemma 5.5 states that running Algorithm 1 on an infinite sum of convex functions is equivalent to running an instance of Algorithm 1 for each summand in parallel. While the statement of the following result is fairly intuitive, the proof of Lemma 5.5 is rather technical. Therefore, we defer its proof to Subsection 5.1.2 and focus on using Lemma 5.5 to prove Theorem 5.1 and Theorem 5.2.

Lemma 5.5. *Consider a family of non-negative, convex functions, $\{f_i\}_{i=1}^{\infty}$, where $f_i : \mathbb{R}^{T_i} \to \mathbb{R}$. Let $\mathscr{X}_i \subseteq \mathbb{R}^{T_i}$ be a closed and convex set such that $\mathscr{X}_i \subseteq \mathscr{B}_{T_i}(0,R)$, where $\mathscr{B}_{T_i}(0,R)$ is the Euclidean ball of radius R in $\mathbb{R}^{T_i}$. Assume that f_i is bounded by M and L Lipschitz on $\mathscr{X}_i$ for every i. Define $f : \prod_{i=1}^{\infty} \mathbb{R}^{T_i} \to \mathbb{R}$ and $\mathscr{X}$ as*

$$f\left(x^{(1)}, x^{(2)}, \ldots\right) \;=\; \sum_{i=1}^{\infty} \frac{1}{C_i^2} f_i\left(C_i x^{(i)}\right) \qquad \text{and} \qquad \mathscr{X} = \prod_{i=1}^{\infty} \frac{\mathscr{X}_i}{C_i}, \tag{5.1}$$

where $x^{(i)} \in \mathbb{R}^{T_i}$ and $\sum_{i=1}^{\infty} \frac{1}{C_i} \leq 1$, with $C_i \in [1, \infty)$. Then, the following hold:

1. *$f(x)$ is well-defined for every $x \in \mathscr{X}$ (i.e. $\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{C_i^2} f_i(C_i x^{(i)})$ exists for all $x \in \mathscr{X}$).*
2. *f is convex.*
3. *If f_i is α strongly convex on $\mathscr{X}_i$ for every i, then f is α strongly convex on $\mathscr{X}$.*
4. *If f_i is subdifferentiable on $\mathscr{X}_i$ ($\partial f_i(x) \neq \emptyset$ for every $x \in \mathscr{X}_i$), then f is subdifferentiable on $\mathscr{X}$. Therefore, if one could execute Algorithm 1 individually for each f_i, then one can execute Algorithm 1 on the function f.*
5. *f is L-Lipschitz on $\mathscr{X}$. That is, for every $x \in \mathscr{X}$ and $g \in \partial f(x)$, $\|g\| \leq L$.*
6. *Let σ_i be a map such that for every $x \in \mathscr{X}_i$, $\sigma_i(x) \in \partial f_i(x)$. Let $x_t^{(i)}$ denote the t-th iterate of Algorithm 1 on the function f_i using the feasible region $\mathscr{X}_i$, step sizes η_t, initial point $x_1^{(i)}$ and the subgradient oracle σ_i. Then, there is a map σ such that for every $x \in \mathscr{X}$, $\sigma(x) \in \partial f(x)$ and the t-th iterate of gradient descent on the function f using the feasible region $\mathscr{X}$, step sizes η_t, initial point $x_1 = (x_1^{(1)}/C_1, x_1^{(2)}/C_2, \ldots)$ and the subgradient oracle σ is given by*

$$x_t \;=\; \left(\frac{x_t^{(1)}}{C_1}, \frac{x_t^{(2)}}{C_2}, \ldots\right).$$

How Lemma 5.5 will be used. Lemma 5.5 allows one to carefully design a single infinite dimensional function from many finite dimensional functions (see Eq. (5.1)) while maintaining crucial properties such as convexity, Lipschitzness, and boundedness. Importantly, running Algorithm 1 on this infinite dimensional function is "equivalent" to running an instance of Algorithm 1 for each finite dimensional function in parallel: The value of the t-th iterate of the infinite dimensional instance can be obtained by a weighted sum of the values of t-th

iterates of each of the finite dimensional instances. Therefore, if we use $f_i = f_{T_i}$ where f_{T_i} are the T_i-dimensional functions from Section 2.1 which exhibit $\Omega(\log(T_i)/T_i)$ error after T_i iterations of Algorithm 1, then we could enforce error $\Omega(\log(T_i)/T_i)$ at iteration T_i as long as each f_i were non-negative. Letting $\lim_{i\to\infty} T_i = \infty$ and ignoring the scaling factors $1/C_i$, we have that $\limsup_{T\to\infty} \frac{f(x_T)}{\log(T)/T} = \Omega(1)$. The presence of the scaling factors $1/C_i$, which are necessary for maintaining Lipschitz-ness, weaken the result to $\limsup_{T\to\infty} \frac{f(x_T)}{g(T)} = +\infty$ for any $g(T) = o(\log(T)/T)$.

5.1.1 Proof of Theorem 5.1

We will prove Lemma 5.6, which is quite similar to Theorem 5.1, except that it appears to have a stronger hypothesis. However, Lemma 5.7 implies that the hypothesis is actually *not* stronger: For each $g = o\left(\log(T)/T\right)$ we also have $g = o\big(\log(T)/\left(T\cdot h(T)\right)\big)$ for some function h such that $\lim_{T\to\infty} h(T) = +\infty$. Thus Theorem 5.1 follows from Lemma 5.6.

Lemma 5.6. *For every non-negative $g(T) = o\big(\log(T)/\left(T\cdot h(T)\right)\big)$ where $\lim_{T\to\infty} h(T) = +\infty$, there exists a convex function $f : \mathscr{X} \to \mathbb{R}$ where $\mathscr{X} \subset \ell_2$ is convex and f is 3-Lipschitz, 1-strongly convex, $\inf_{x\in\mathscr{X}} f(x) = 0$ and satisfies the following. Suppose that Algorithm 1 is executed from the initial point $x_1 = 0$ with step sizes $\eta_t = 1/t$. Then,*

$$\limsup_{T\to\infty} \frac{f(x_T)}{g(T)} \;=\; +\infty.$$

Lemma 5.7. *Suppose $g(T)$ is a non-negative function such that $g(T) = o\left(\log(T)/T\right)$. Then we may write $g(T) = o\left(\log(T)/\left(T\cdot h(T)\right)\right)$ for some $h(T)$ satisfying $\lim_{T\to\infty} h(T) = +\infty$.*

Proof. Let $h(T) = \sqrt{\frac{\log(T)/T}{g(T)}}$. Then, $\lim_{T\to\infty} h(T) = +\infty$ because $g = o\left(\log(T)/T\right)$. We have

$$\lim_{T\to\infty}\left[\frac{g(T)}{\log(T)/\left(T\cdot h(T)\right)}\right] \;=\; \lim_{T\to\infty}\left[\sqrt{\frac{g(T)}{\log(T)/T}}\right] \;=\; 0,$$

because $g(T) = o\left(\log(T)/T\right)$. ■

Proof (of Lemma 5.6).

Showing that $\limsup_{T\to\infty} \frac{f(x_T)}{g(T)} = +\infty$ is equivalent to showing the following:

$$\forall M > 0, \forall N \in \mathbb{N}, \exists T \geq N \;\text{ s.t. }\; f(x_T) > Mg(T). \tag{5.2}$$

We would like to apply Lemma 5.5, and so we must define functions f_i which satisfy the required properties. Let $C_i = 2^i$. Let $T_i \geq i$ be such that

$$\forall t \geq T_i \geq i :\; g(t) \leq \frac{1}{4C_i^2}\left(\frac{\log t}{t\cdot h(t)}\right), \tag{5.3}$$

which is possible because $g(T) = o\left(\log(t)/(t\cdot h(t))\right)$. Our functions f_i are defined mainly using the T_i-dimensional version of the function defined in Section 2.1, which we will denote by f_{T_i} and provides the $\Omega(\log(T)/T)$ lower bound on the error of Algorithm 1. Recall some important properties of the function f_{T_i} which were derived in Section 2.1. Let $\mathscr{B}_T(0,R)$ be the Euclidean ball of radius R in $\mathbb{R}^T$.

- f_{T_i} is 3-Lipschitz and 1-strongly convex over $\mathscr{B}_{T_i}(0,1)$,
- There exists a subgradient oracle such that running Algorithm 1 on f_{T_i} using this oracle, feasible region $\mathscr{B}_{T_i}(0,1)$, and step sizes $\eta_t = 1/t$ satisfies $f_{T_i}(x_{T_i+1}) \geq \frac{\log T_i}{4T_i}$ (this is the content of Theorem 1.12).

Defining the functions f_i. Using f_{T_i} as the definition of f_i in an application of Lemma 5.5 is problematic because f_{T_i} is not guaranteed to be non-negative. As a consequence, we will not be able to assert that the value of the T_i-th iterate of Algorithm 1 on f is $\Omega\left(\frac{\log(T_i)}{C_i^2 T_i}\right)$, even though the proof[1] Theorem 1.12 shows that this true for the T_i-th iterate of Algorithm 1 on f_{T_i} and Lemma 5.5 guarantees that the value of the t-h iterate for f is a scaled sum of the values of the t-th iterates of the f_{T_i}'s. Instead, we augment the definition of f_{T_i} so that the function remains Lipschitz, strongly convex, attains the lower bound from Theorem 1.12, and is also non-negative. Note that simply taking the max of 0 and f_{T_i} is not a feasible approach since the resulting function is not strongly convex.

Let $f_i : \mathscr{B}_{T_i}(0,1) \to \mathbb{R}$ be defined as $f_i(x) = \max\{f_{T_i}(x), \frac{1}{2}\|x\|^2\}$. It is straightforward to check that f_i is 3-Lipschitz, bounded by 3 on $\mathscr{B}_{T_i}(0,1)$ and also 1-strongly convex (these properties follow easily from the fact that f_{T_i} is 3-Lipschitz and 1 strongly convex on $\mathscr{B}_{T_i}(0,1)$). Also, because $f_{T_i}(0) = 0$, we see that $\inf_{x \in \mathscr{B}_{T_i}(0,1)} f_i(x) = 0$. Finally, f_i is subdifferentiable on $\mathscr{X}_i$, and there is a subgradient oracle for which f_i can be made to achieve the same lower bound as f_{T_i} achieves in Theorem 1.12. This is formalized in the following claim, whose proof we defer to Subsection 5.1.2.

Claim 5.8. *There exists a subgradient oracle for the function f_i such that if $x_1^{(i)}, x_2^{(i)}, \ldots$ are the iterates produced by running Algorithm 1 with the function f_i, the feasible region $\mathscr{B}_{T_i}(0,1)$, the step sizes $\eta_t = 1/t$, and initial point $x_1^{(i)} = 0$ then,*

$$f_i(x_{T_i+1}^{(i)}) \;\geq\; \frac{\log T_i}{4T_i}. \tag{5.4}$$

Defining the function f. Recall that $C_i = 2^i$. We define $f : \prod_{i=1}^{\infty} \mathscr{B}_{T_i}(0,1)/C_i \to \mathbb{R}$ as follows. Let $x = (x^{(1)}, x^{(2)}, \ldots)$ be such that $x^{(i)} \in \mathscr{B}_{T_i}(0,1)/C_i$. Then, define

$$f\left(x^{(1)}, x^{(2)}, \ldots\right) \;:=\; \sum_{i=1}^{\infty} \frac{1}{C_i^2} f_i\left(C_i x^{(i)}\right).$$

Note that f is non-negative and $f(0) = 0$. Therefore, $\inf_{x \in \mathscr{X}} f(x) = 0$. Let σ_i be the subgradient oracle for f_i whose existence is guaranteed by Claim 5.8 and let $x_t^{(i)}$ denote the t-th iterate of Algorithm 1 on f_i using the subgradient oracle σ_i, initial point $x_1^{(i)} = 0$, and step size $\eta_t = 1/t$. Then, using Lemma 5.5, we get the following:

- f is well defined and subdifferentiable over $\mathscr{X}$,
- f is 3-Lipschitz over $\mathscr{X}$,
- f is 1-strongly convex over $\mathscr{X}$,

[1] The statement of Theorem 1.12 states that $f_{T_i}(x_{T_i+1}) - \inf_{x \in \mathscr{X}} f_{T_i}(x) = \Omega\left(\frac{\log T_i}{T_i}\right)$, however the proof argues that $f_{T_i}(x_{T_i+1}) = \Omega\left(\frac{\log T_i}{T_i}\right)$ and shows that $\inf_{x \mathscr{X}} f_{T_i}(x) \leq 0$

- There exists a subgradient oracle, σ, for f over $\mathscr{X}$ such that the t-th iterate, x_t, of Algorithm 1 on f using the subgradient oracle σ, initial point $x_1 = 0$, and step size $\eta_t = 1/t$ satisfies,

$$x_t \;=\; \left(\frac{x_t^{(1)}}{C_1}, \frac{x_t^{(2)}}{C_2}, \dots\right). \tag{5.5}$$

Showing Eq. (5.2) holds. Let $M > 0$ and let $N \in \mathbb{N}$. The following is true because $\lim_{T\to\infty} h(T) = \infty$.

$$\exists n \text{ s.t., } \forall T \geq n,\; h(T) > M. \tag{5.6}$$

Consider $N' = \max\{n, N\}$. Then, we have the following:

$$\begin{aligned}
f\left(x_{T_{N'}+1}\right) \;&=\; \sum_{i=1}^{\infty} \frac{1}{C_i^2} f_i\left(x^{(i)}_{T_{N'}+1}\right) && \text{(by definition of } f \text{ and Eq. (5.5))} \\
&\geq\; \frac{1}{C_{N'}^2} f_{N'}\left(x^{(N')}_{T_{N'}+1}\right) && \text{(each } f_i \text{ is non-negative)} \\
&\geq\; \frac{1}{4C_{N'}^2} \frac{\log T_{N'}}{T_{N'}} && \text{(by Eq. (5.4))} \\
&\geq\; g(T_{N'})h(T_{N'}) && \text{(by Eq. (5.3))} \\
&>\; M \cdot g(T_{N'}) && \text{(because } T_{N'} \geq N' \geq n \text{ and using Eq. (5.6)).}
\end{aligned}$$

Hence, we've shown for every $M > 0$ and for every $N \in \mathbb{N}$, there exists $t \geq N$ (namely, $T_{N'} \geq N' \geq N$) such that $f(x_t) > M \cdot g(t)$. This demonstrates that Eq. (5.2) holds and therefore the proof is complete. ■

5.1.2 Proof of Lemma 5.5

Throughout this subsection, we work under the assumptions of Lemma 5.5. That is, we assume we have a family of non-negative convex functions $\{f_i\}_{i=1}^{\infty}$ such that $f_i : \mathbb{R}^{T_i} \to \mathbb{R}$ and closed convex sets $\mathscr{X}_i \subset \mathbb{R}^{T_i}$ such $\mathscr{X}_i$ is contained in some Euclidean ball of radius R, denoted $\mathscr{B}_{T_i}(0,R)$. We set $\mathscr{X} = \prod_{i=1}^{\infty} \frac{\mathscr{X}_i}{C_i}$ where $\sum_{i=1}^{\infty} \frac{1}{C_i} \leq 1$ and $C_i \in [1,\infty)$ and define $f : \prod_{i=1}^{\infty} \mathbb{R}^{T_i} \to \mathbb{R}$ as $f(x^{(1)}, x^{(2)}, \dots) = \sum_{i=1}^{\infty} \frac{1}{C_i^2} f_i(C_i x^{(i)})$. We will prove Lemma 5.5 in pieces. Combining Claim 5.9, Claim 5.11, Claim 5.13, Claim 5.14, and Claim 5.15 proves Lemma 5.5.

First, let us check that f is well defined.

Claim 5.9. *For every $x \in \mathscr{X}$, $\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{C_i^2} f_i(C_i x^{(i)})$ exists and is finite.*

Indeed, this follows as a corollary of the following claim since a series with non-negative summands converges if and only if it is bounded. Recall that f_i are non-negative functions.

Claim 5.10. *f is bounded by on $\mathscr{X}$.*

Proof. Let $x = (x^{(1)}/C_1, x^{(2)}/C_2, \dots) \in \mathscr{X}$.

$$f(x) \;=\; \sum_{i=1}^{\infty} \frac{1}{C_i^2} f_i(x^{(i)}) \;\leq\; M \sum_{i=1}^{\infty} \frac{1}{C_i^2} \;\leq\; M,$$

because f_i is bounded by M on $\mathscr{X}_i$. ■

Next, we prove the implications involving convexity and strong convexity.

Claim 5.11. *f is a convex function. If f_i is α-strongly convex for every i, then f is α-strongly convex.*

Proof. Convexity follows easily by considering $\lambda x + (1-\lambda)y$ where $x, y \in \prod_{i=1}^{\infty} \mathbb{R}^{T_i}$ and $\lambda \in (0,1)$. Indeed, since each f_i is convex we have $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$, as desired.

To check strong convexity, it suffices to prove that the function $g(x) := f(x) - \frac{\alpha}{2}\|x\|^2$ is convex. Let $x = (x^{(1)}, x^{(2)}, \ldots)$ where $x^{(i)} \in \mathbb{R}^{T_i}$. Then,

$$g(x) \;=\; \sum_{i=1}^{\infty} \left[\frac{1}{C_i^2} f_i(C_i x^{(i)}) - \frac{\alpha}{2} \left\| x^{(i)} \right\|^2 \right] \;=\; \sum_{i=1}^{\infty} \left[\frac{1}{C_i^2} \left(f_i(C_i x^{(i)}) - \frac{\alpha}{2} \left\| C_i x^{(i)} \right\|^2 \right) \right],$$

which is convex because $f_i - \frac{\alpha}{2}\|\cdot\|^2$ is a convex function since f_i is α-strongly convex. ■

To prove the implications involving the subdifferentiability and Lipschitz-ness of f on $\mathscr{X}$, we will require a characterization of the subdifferential of f (Claim 5.17) which will also be useful in proving that running Algorithm 1 on f is equivalent to running an instance of Algorithm 1 for each f_i in parallel.

Claim 5.12. *Let $x = (x^{(1)}/C_1, x^{(2)}/C_2, \ldots) \in \mathscr{X}$ (that is $x^{(i)} \in \mathscr{X}_i$). Then, $\partial f(x) = \prod_{i=1}^{\infty} \frac{1}{C_i} \partial f_i(x^{(i)})$*

Claim 5.12 immediately proves the following claim regarding the subdifferentiability of f on $\mathscr{X}$ and Lipschitzness is an easy consquence of Claim 5.12 and our assumptions.

Claim 5.13. *Suppose for every i, for every $x \in \mathscr{X}_i$ that $\partial f_i(x) \neq \emptyset$. Then, $\partial f(x) \neq \emptyset$ for every $x \in \mathscr{X}$.*

Claim 5.14. *f is L-Lipschitz on $\mathscr{X}$.*

Proof. Let $x = (x^{(1)}/C_1, x^{(2)}/C_2, \ldots) \in \mathscr{X}$. By Claim 5.12, we may write any $g \in \partial f(x)$ as

$$g \;=\; \left(g^{(1)}/C_1, g^{(2)}/C_2, \ldots \right),$$

where $g^{(i)} \in \partial f_i(x^{(i)})$. Hence,

$$\|g\|^2 \;=\; \sum_{i=1}^{\infty} \frac{\left\| g^{(i)} \right\|^2}{C_i^2} \;\leq\; L^2 \sum_{i=1}^{\infty} \frac{1}{C_i^2} \;\leq\; L^2.$$

■

Next, we will use Claim 5.12 to prove that that running Algorithm 1 on f is equivalent to running an instance of Algorithm 1 for each f_i in parallel.

Claim 5.15. *Let σ_i be a map such that for every $x \in \mathscr{X}_i$, $\sigma_i(x) \in \partial f_i(x)$. Let $x_t^{(i)}$ denote the t-th iterate of Algorithm 1 on the function f_i, using the feasible region $\mathscr{X}_i$, step sizes η_t, initial point $x_1^{(i)}$ and the subgradient oracle σ_i. Then, there is a map σ such that for every $x \in \mathscr{X}$, $\sigma(x) \in \partial f(x)$ and the t-th iterate of gradient descent on the function f using the feasible region $\mathscr{X}$, step sizes η_t, initial point $x_1 = (x_1^{(1)}/C_1, x_1^{(2)}/C_2, \ldots)$ and the subgradient oracle σ is given by*

$$x_t \;=\; \left(\frac{x_t^{(1)}}{C_1}, \frac{x_t^{(2)}}{C_2}, \ldots \right).$$

Proof. We begin by specifying σ. By Claim 5.12, the following definition of σ is valid:

$$\sigma\left(\frac{x^{(1)}}{C_1}, \frac{x^{(2)}}{C_2}, \dots\right) \;=\; \left(\frac{1}{C_1}\sigma_1(x^{(1)}), \frac{1}{C_2}\sigma_2(x^{(2)}), \dots\right).$$

We proceed by induction, using the subgradient oracle σ. The base case is true by assumption, that is we assumed $x_1 = \left(\frac{x_1^{(1)}}{C_1}, \frac{x_1^{(2)}}{C_2}, \dots\right)$. Now, suppose $x_t = \left(\frac{x_t^{(1)}}{C_1}, \frac{x_t^{(2)}}{C_2}, \dots\right)$. Then,

$$\begin{aligned}
y_{t+1} &= x_t - \eta_t \sigma(x_t) \\
&= \left(\frac{x_t^{(1)}}{C_1}, \frac{x_t^{(2)}}{C_2}, \dots\right) - \eta_t \sigma\left(\frac{x_t^{(1)}}{C_1}, \frac{x_t^{(2)}}{C_2}, \dots\right) && \text{(by induction hypothesis)} \\
&= \left(\frac{x_t^{(1)}}{C_1}, \frac{x_t^{(2)}}{C_2}\right) - \eta_t \left(\frac{1}{C_1}\sigma_1(x_t^{(1)}), \frac{1}{C_2}\sigma_2(x_t^{(2)}), \dots\right) && \text{(by definition of } \sigma) \\
&= \left(\frac{1}{C_1}\left(x_t^{(1)} - \eta_t \sigma_1(x_t^{(1)})\right), \frac{1}{C_2}\left(x_t^{(2)} - \eta_t \sigma_2(x_t^{(2)})\right), \dots\right) \\
&= \left(\frac{1}{C_1} y_{t+1}^{(1)}, \frac{1}{C_2} y_{t+1}^{(2)}, \dots\right).
\end{aligned}$$

Then, $x_{t+1} = \Pi_{\mathscr{X}}(y_{t+1})$. Now, we will apply Claim 5.16. To do so, we must check $y_{t+1} \in \ell_2$. Indeed, the following shows that this is true.

$$\begin{aligned}
\|y_{t+1}\|^2 &= \sum_{i=1}^{\infty} \frac{1}{C_i^2} \left\| x_t^{(i)} - \eta_t \sigma_i(x_t^{(i)}) \right\|^2 \\
&\le \sum_{i=1}^{\infty} \frac{1}{C_i^2} \left(\left\| x_t^{(i)} \right\|^2 + \left\| \sigma_i(x_t^{(i)}) \right\|^2 + 2 \left\| x_t^{(i)} \right\| \left\| \sigma_i(x_t^{(i)}) \right\| \right) && \text{(by Cauchy-Schwarz)} \\
&= \sum_{i=1}^{\infty} \frac{1}{C_i^2} \left\| x_t^{(i)} \right\|^2 + \sum_{i=1}^{\infty} \frac{2}{C_i^2} \left\| x_t^{(i)} \right\| \left\| \sigma_i(x_t^{(i)}) \right\| + \sum_{i=1}^{\infty} \frac{1}{C_i^2} \left\| \sigma_i(x_t^{(i)}) \right\|^2 && \text{(each series converges absolutely)} \\
&\le R^2 \sum_{i=1}^{\infty} \frac{1}{C_i^2} + RL \sum_{i=1}^{\infty} \frac{2}{C_i^2} + L^2 \sum_{i=1}^{\infty} \frac{1}{C_i^2} && \text{(since } \mathscr{X}_i \subset \mathscr{B}_{T_i}(0,R) \text{ and } f_i \text{ is } L \text{ Lipschitz)} \\
&< \infty \qquad \text{(because } \sum_{i=1}^{\infty} \frac{1}{C_i^2} \le 1)
\end{aligned}$$

Therefore, by Claim 5.16, because each $\mathscr{X}_i$ is closed and convex we may compute this by computing the projection of $\frac{1}{C_i} y_{t+1}^{(i)}$ onto $\mathscr{X}_i / C_i$ and then concatenating the results. Recall that $\Pi_{\mathscr{X}_i}(y_{t+1}^{(i)}) = x_{t+1}^{(i)}$, by definition

of Algorithm 1. Hence, we have

$$\begin{aligned}\Pi_{\mathscr{X}_i/C_i}\left(\frac{1}{C_i}y_{t+1}^{(i)}\right) &= \operatorname*{arg\,min}_{z\in\mathscr{X}_i/C_i}\left\|z-\frac{1}{C_i}y_{t+1}^{(i)}\right\| \\ &= \frac{1}{C_i}\operatorname*{arg\,min}_{z\in\mathscr{X}_i}\left\|z-y_{t+1}^{(i)}\right\| \\ &= \frac{1}{C_i}\Pi_{\mathscr{X}_i}\left(y_{t+1}^{(i)}\right) \\ &= \frac{1}{C_i}x_{t+1}^{(i)}.\end{aligned}$$

Hence, we have

$$x_{t+1} = \Pi_{\mathscr{X}}(y_{t+1}) = \left(\Pi_{\mathscr{X}_1/C_1}\left(\frac{1}{C_1}y_{t+1}^{(1)}\right),\Pi_{\mathscr{X}_2/C_2}\left(\frac{1}{C_2}y_{t+1}^{(2)}\right),\dots\right) = \left(\frac{x_{t+1}^{(1)}}{C_1},\frac{x_{t+1}^{(2)}}{C_2},\dots\right),$$

as desired. ■

Claim 5.16. *Let $\mathscr{Y}=\prod_{i=1}^{\infty}\mathscr{Y}_i$ where each $Y_i\subseteq\mathbb{R}^{T_i}$ is a closed convex set. Let $z=\left(z^{(1)},z^{(2)},\dots\right)\in\ell_2$ where $z^{(i)}\in\mathbb{R}^{T_i}$. Then, $\Pi_{\mathscr{Y}}(z)=\left(\Pi_{\mathscr{Y}_1}(z^{(1)}),\Pi_{\mathscr{Y}_2}(z^{(2)}),\dots\right)$.*

Proof. Since $\prod_{i=1}^{\infty}\mathscr{Y}_i$ is closed and convex, then by [38, Theorem 4.10] we have that $\Pi_{\mathscr{Y}}(z)$ is achieved by a unique element for $z\in\ell_2$. Let $\Pi_{\mathscr{Y}}(z)=(y^{(1)},y^{(2)},\dots)=:y$ where $y^{(i)}\in\mathscr{Y}_i$. Suppose that for some i, $y^{(i)}\neq\Pi_{\mathscr{Y}_i}(z^{(i)})$. Then, by the uniqueness of the projection onto a convex set, we have that $\left\|y^{(i)}-z^{(i)}\right\|^2>\left\|\Pi_{\mathscr{Y}_i}(z^{(i)})-z^{(i)}\right\|^2$. Consider the point $y':=(y^{(1)},\dots y^{(i-1)},\Pi_{\mathscr{Y}_i}(z^{(i)}),y^{(i+1)},\dots)\in\mathscr{Y}$. We have

$$\|y-z\|^2 = \sum_{j=1}^{\infty}\left\|y^{(j)}-z^{(j)}\right\|^2 > \sum_{j\neq i}\left\|y^{(j)}-z^{(j)}\right\|^2+\left\|\Pi_{\mathscr{Y}_i}(z^{(i)})-z^{(i)}\right\|^2 = \left\|y'-z\right\|^2,$$

which contradicts the assumption that $y=\Pi_{\mathscr{Y}}(z)$.

■

Next, we must check Claim 5.12.

Proof of Claim 5.12

The proof of Claim 5.12 depends on the following lemma.

Claim 5.17. *Suppose $h:\ell_2\to\mathbb{R}$ is such that $h(y^{(1)},y^{(2)},\dots)=\sum_{n=1}^{\infty}h_n(y^{(n)})$ where each $h_n:\mathbb{R}^{T_n}\to\mathbb{R}$ is a convex function. Assume h converges on some set $\mathscr{Y}\subseteq\ell_2$. Suppose that for every $y=(y^{(1)},y^{(2)},\dots)\in\mathscr{Y}$ with $y^{(n)}\in\mathbb{R}^{T_n}$, we have that $\sum_{n=1}^{\infty}\left\|g^{(n)}\right\|<+\infty$ for every $g^{(n)}\in\partial h_n(y^{(n)})$. Then, for all $y\in\mathscr{X}$*

$$\partial h(y^{(1)},y^{(2)},\dots)=\partial h_1(y^{(1)})\times\partial h_2(y^{(2)})\times\dots.$$

See Sub-subsection 5.1.2 for a proof. To prove Claim 5.12 we will apply Claim 5.17, which requires us to check $\mathscr{X}\subset\ell_2$.

Claim 5.18. $\mathscr{X}\subset\ell_2$.

Proof. Let $x = (x^{(1)}/C_1, x^{(2)}/C_2, \ldots) \in \mathscr{X}$.

$$\|x\|^2 \;=\; \sum_{i=1}^{\infty} \frac{\left\|x^{(i)}\right\|^2}{C_i^2} \;\leq\; R^2 \sum_{i=1}^{\infty} \frac{1}{C_i^2} \;\leq\; R^2$$

because $\mathscr{X}_i \subseteq \mathscr{B}_{T_i}(0,R)$. ■

Proof (of Claim 5.12). We will apply Claim 5.17 with $h_n = \frac{1}{C_n^2} f_n \circ C_n I_n$ where I_n is the $n \times n$ identity matrix, $\mathscr{Y} = \mathscr{X} = \prod_{n=1}^{\infty} \frac{1}{C_n} \mathscr{X}_n$, and $h = f = \sum_{n=1}^{\infty} f_n$. To do so, we must check

- Each f_n is convex (which implies h_n is convex),
- $\mathscr{X} = \frac{1}{C_i} \prod_{i=1}^{\infty} \mathscr{X}_i \subset \ell_2$,
- f is bounded everywhere on $\mathscr{X}$ (since $f(x)$ is a sum of non-negative numbers, it converges if and only if it is bounded),
- For every $x^{(i)} \in \mathscr{X}_i$, $g^{(i)} \in \partial\left(\frac{1}{C_i^2} f_i \circ C_i I_{T_i}\right)(x^{(i)}/C_i)$, $\sum_{i=1}^{\infty} \left\|g^{(i)}\right\| < +\infty$, where I_{T_i} is the identity matrix in T_i dimensions.

The first three points are handled by assumption, Claim 5.18 and Claim 5.10. It remains to check the fourth point. By Claim A.9, $\partial\left(\frac{1}{C_i^2} f_i \circ C_i I_{T_i}\right)(x^{(i)}/C_i) = \frac{1}{C_i} \partial f_i(x^{(i)})$. Furthermore, because f_i is L-Lipschitz, then every subgradient of f_i on $\mathscr{X}_i$ has norm at most L. Hence, for every $g \in \frac{1}{C_i} \partial f_i(x^{(i)})$, $\|g\|^2 \leq (L/C_i)^2$. We have $\sum_{i=1}^{\infty} \frac{L}{C_i} \leq L$, and therefore the fourth point holds.

Hence, applying Claim 5.17, we have for every $x = \left(x^{(1)}/C_1, x^{(2)}/C_2, \ldots\right) \in \mathscr{X}$, we have

$$\partial f(x^{(1)}/C_1, x^{(2)}/C_2, \ldots) \;=\; \prod_{i=1}^{\infty} \partial\left(\frac{1}{C_i^2} f_i \circ C_i I_{T_i}\right)\left(\frac{x^{(i)}}{C_i}\right) \;=\; \prod_{i=1}^{\infty} \frac{1}{C_i} \partial f_i(x^{(i)}),$$

which completes the proof of the claim.

■

Proof of Claim 5.17

Claim 5.17. Suppose $h : \ell_2 \to \mathbb{R}$ is such that $h(y^{(1)}, y^{(2)}, \ldots) = \sum_{n=1}^{\infty} h_n(y^{(n)})$ where each $h_n : \mathbb{R}^{T_n} \to \mathbb{R}$ is a convex function. Assume h converges on some set $\mathscr{Y} \subseteq \ell_2$. Suppose that for every $y = (y^{(1)}, y^{(2)}, \ldots) \in \mathscr{Y}$ with $y^{(n)} \in \mathbb{R}^{T_n}$, we have that $\sum_{n=1}^{\infty} \left\|g^{(n)}\right\| < +\infty$ for every $g^{(n)} \in \partial h_n(y^{(n)})$. Then, for all $y \in \mathscr{X}$

$$\partial h(y^{(1)}, y^{(2)}, \ldots) = \partial h_1(y^{(1)}) \times \partial h_2(y^{(2)}) \times \ldots.$$

Proof. Let $y \in \mathscr{Y}$. We write y as $\oplus_{n=1}^{\infty} y^{(n)}$ where $y^{(n)} \in \mathbb{R}^{T_n}$ denote the components of y. We have $\sum_{n=1}^{\infty} h_n(y^{(n)})$ converges by assumption.

First we show that $\partial h_1(y^{(1)}) \times \partial h_2(y^{(2)}) \times \ldots \subseteq \partial h(y^{(1)}, y^{(2)}, \ldots)$. Let $z = \oplus_{n=1}^{\infty} z^{(n)} \in \ell_2$ be arbitrary. For each n, suppose $g^{(n)} \in \partial h_n(y^{(n)})$. Hence, for all $z^{(n)} \in \mathbb{R}^{T_n}$

$$h_n(z^{(n)}) \geq h_n(y^{(n)}) + (z^{(n)} - y^{(n)})^{\mathsf{T}} y^{(n)}.$$

Therefore, summing this inequality over n, we have:

$$\begin{aligned} h(z) &= \sum_{n=1}^{\infty} h_n(z^{(n)}) \\ &\geq \sum_{n=1}^{\infty} \left[h_n(y^{(n)}) + (z^{(n)} - y^{(n)})^{\mathsf{T}} g^{(n)} \right] \end{aligned}$$

We may write $\sum_{n=1}^{\infty} \left[h_n(y^{(n)}) + (z^{(n)} - y^{(n)})^{\mathsf{T}} g^{(n)} \right] = \sum_{n=1}^{\infty} h_n(y^{(n)}) + \sum_{n=1}^{\infty} (z^{(n)} - y^{(n)})^{\mathsf{T}} g^{(n)}$ if both series on the right hand side converge. Observe that $\sum_{n=1}^{\infty} h_n(y^{(n)})$ converges by assumption. We now show $\sum_{n=1}^{\infty} (z^{(n)} - y^{(n)})^{\mathsf{T}} g^{(n)}$ converges absolutely. Indeed,

$$\begin{aligned} \sum_{n=1}^{\infty} |(z^{(n)} - y^{(n)})^{\mathsf{T}} g^{(n)}| &\leq \sum_{n=1}^{\infty} \left\| z^{(n)} - y^{(n)} \right\| \left\| g^{(n)} \right\| \qquad \text{(by Cauchy-Schwarz)} \\ &\leq \sum_{n=1}^{\infty} \|z - y\| \left\| g^{(n)} \right\| \\ &= \|z - y\| \sum_{n=1}^{\infty} \left\| g^{(n)} \right\| \qquad \text{(since } x \text{ and } y \text{ are in } \ell_2) \\ &< +\infty \qquad (\textstyle\sum_{n=1}^{\infty} \|g^{(n)}\| < +\infty \text{ by assumption}) , \end{aligned}$$

as desired. Therefore,

$$h(z) \geq h(y) + \langle z - y, g \rangle,$$

where $z = \oplus_{n=1}^{\infty} z^{(n)}$, $y = \oplus_{n=1}^{\infty} y^{(n)}$, $g = \oplus_{n=1}^{\infty} g^{(n)}$. This demonstrates that $\oplus_{n=1}^{\infty} g^{(n)} \in \partial h(y^{(1)}, y^{(2)}, \ldots)$, since $Z \in \ell_2$ was arbitrary.

Now, we prove $\partial h(y^{(1)}, y^{(2)}, \ldots) \subseteq \partial h_1(y^{(1)}) \times \partial h_2(y^{(2)}) \times \ldots$. Suppose $g = \oplus_{n=1}^{\infty} g^{(n)}$ is a subgradient of h at $y \in \mathscr{Y}$. Since ℓ_2 is self-dual, then by Definition 5.4 the subgradients of h live in ℓ_2 and so $g \in \ell_2$. We verify that for every n, $g^{(n)} \in \partial h_n(y^{(n)})$. Let $z \in \mathbb{R}^{T_n}$ be arbitrary and let $\tilde{y} = (y^{(1)}, y^{(2)}, \ldots, y^{(n-1)}, z, y^{(n+1)}, \ldots)$. Observe that $\tilde{y} \in \ell_2$ because $y \in \ell_2$ and moreover notice that $h(\tilde{y})$ converges since $h(\tilde{y})$ differs from $h(y)$ in only a single summand and $h(y)$ converges. Since g is a subgradient of h at y, we have

$$\begin{aligned} h(\tilde{y}) - h(y) &\geq (\tilde{y} - y)^{\mathsf{T}} g \\ \Leftrightarrow \sum_{j=1}^{\infty} h_j(\tilde{y}^{(j)}) - \sum_{j=1}^{\infty} h_j(y^{(j)}) &\geq \sum_{j=1}^{\infty} (\tilde{y}^{(j)} - y^{(j)})^{\mathsf{T}} g^{(j)} \\ \Leftrightarrow \sum_{j=1}^{\infty} \left[h_j(\tilde{y}^{(j)}) - h_j(y^{(j)}) \right] &\geq \sum_{j=1}^{\infty} (\tilde{y}^{(j)} - y^{(j)})^{\mathsf{T}} g^{(j)} \qquad (h(y) \text{ and } h(\tilde{y}) \text{ converge}). \end{aligned}$$

By definition of y we have

$$\tilde{y}^{(j)} = \begin{cases} y^{(j)} & \text{if } j \neq n, \\ z & \text{otherwise.} \end{cases}$$

Therefore, $h(\tilde{y}) - h(y) \geq (\tilde{y} - y)^\mathsf{T} g$ is equivalent to

$$h_n(z) - h_n(y^{(n)}) \geq (z - y^{(n)})^\mathsf{T} g^{(n)},$$

which is equivalent to saying that $g^{(n)}$ is a subgradient of h_n at $y^{(n)}$ as desired. ■

5.1.3 Proof of Claim 5.8

Recall the definition of the augmented version of f_{T_i}, $f_i = \max\left\{f_{T_i}(x), \frac{1}{2}\|x\|^2\right\}$.The main idea of showing that f_i provides the same lower bound on the performance of Algorithm 1 as f_{T_i} itself is to show that under the appropriate choice of subgradient oracle, running Algorithm 1 using f_i is equivalent to running Algorithm 1 using f_{T_i} (at least for the first $T_i + 1$ iterations) in the sense that both executions produce the same iterates.

Proof (of Claim 5.8). We begin by recalling the definition of the function f_{T_i} and its subgradient oracle:

Let $\mathscr{B}_{T_i}(0,1)$ be the Euclidean unit ball in $\mathbb{R}^{T_i}$. Define $f : \mathscr{B}_{T_i}(0,1) \to \mathbb{R}$ and $h_\ell \in \mathbb{R}^{T_i}$ for $\ell \in [T+1]$ by

$$f_{T_i}(x) \;=\; \max_{\ell \in [T+1]} H_\ell(x) \qquad \text{where} \qquad H_\ell(x) = h_\ell^\mathsf{T} x + \frac{1}{2}\|x\|^2$$

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

Recall the definition of the subgradient oracle for f_{T_i} from Section 2.1: Given a point $x \in \mathscr{B}_{T_i}(0,1)$, define $\sigma_i'(x) = h_{\ell'} + x$ where $\ell' = \min \mathscr{I}(x)$ and $\mathscr{I}(x) = \{\, \ell : H_\ell(x) = f_{T_i}(x) \,\}$. Define

$$\sigma_i(x) \;=\; \begin{cases} \sigma_i'(x) & \text{if } \partial f_i(x) \subseteq \partial f_{T_i}(x), \\ \text{arbitrary } x \in \partial f_i(x) & \text{otherwise.} \end{cases}$$

Recall that by Lemma 2.5 the t-th (for $t = 1, \ldots, T_i + 1$) iterate produced by Algorithm 1 on f_{T_i} when using step size $\eta_t = 1/t$ and feasible region $\mathscr{B}_{T_i}(0,1)$ is given by the vector $z_t \in \mathbb{R}^{T_i}$ which is strictly positive on the first $t-1$ coordinates and is zero on the remaining coordinates (by Claim 2.3). Therefore, the subgradient oracles σ_i and σ_i' agree on the vectors $z_1, \ldots z_{T_i+1}$ because $f_i(z_t) = \max\left\{f_{T_i}(z_t), \frac{1}{2}\|z_t\|^2\right\} = f_{T_i}(z_t) > \frac{1}{2}\|z_t\|^2$ (which implies $\partial f_i(z_t) = \partial f_{T_i}(x)$).

Since the instance of Algorithm 1 on f_{T_i} and the instance of Algorithm 1 on f_i begin at the same starting point and because the subgradient oracles for both algorithms agree on $z_1, \ldots, z_{T_i+1}$ (which are the iterates produced by the instance running on f_{T_i}), we have that the t-th iterate produced by Algorithm 1 when executed on f_i is z_t. That is, $x_t^{(i)} = z_t$ for $t \in [T_i + 1]$. Observe that

$$f_i(x_{T_i+1}^{(i)}) \;=\; f_i(z_{T_i+1}) \;=\; f_{T_i}(z_{T_i+1}) \;\geq\; \frac{\log T_i}{4T_i},$$

where the final inequality is because f_{T_i} is the function which realizes the lower bound in Theorem 1.12. ■

5.2 Necessity of $\log(1/\delta)$

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

Lemma 5.19 ([24, Lemma 4])**.** *Let $X_1,\dots,X_T$ 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 $\sqrt{6} \le c < 2\sqrt{T}$,*

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

Consider the single-variable function $f(x) = \frac{1}{2}x^2$ and suppose that the domain is $\mathscr{X} = [-1,1]$. Then f is 1-strongly convex and 1-Lipschitz on $\mathscr{X}$. 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 5.20. *Let $\sqrt{6} \le \sqrt{\frac{2}{9}\log(1/\delta)} \le 2\sqrt{T}$. Then $f(x_{T+1}) \ge \frac{1}{9}\frac{\log(1/\delta)}{T}$ 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, $x_2 = \Pi_{\mathscr{X}}(y_2) = y_2$ since $|y_2| \le 1$. 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 $|y_{t+1}| \le 1$, we have $x_{t+1} = y_{t+1}$.

Hence, by Lemma 5.19 with $c = \sqrt{\frac{2}{9}\log(1/\delta)}$, we have $x_{T+1} \ge \sqrt{\frac{2}{9}\log(1/\delta)}/\sqrt{T}$ with probability at least δ (provided $T \ge \frac{1}{18}\log(1/\delta)$). We conclude that $f(x_{T+1}) \ge \frac{\log(1/\delta)}{9T}$ with probability at least $\Omega(\delta)$. ■

We can also show that Theorem 1.17 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 $1 \le t < T/2$ and $t > 3T/4$, $\hat{z}_t = 0$. For $T/2 \le t \le 3T/4$, 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 $A_t \ge 1/4$ for $T/2 \le t \le 3T/4$ so we still have $|\hat{z}_t| \le 1$ for all t.

Claim 5.21. *Let $\sqrt{6} \ge \sqrt{\frac{2}{9}\log(1/\delta)} \le 2\sqrt{T}$. Then $f\left(\frac{1}{T/2+1}\sum_{t=T/2+1}^{T+1} x_t\right) \ge \Omega\left(\frac{\log(1/\delta)}{T}\right)$ 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

$$\frac{1}{T/2+1}\sum_{t=T/2+1}^{T+1} x_t = \frac{1}{T/2+1}\sum_{t=T/2}^{3T/4} A_t\hat{z}_t. \tag{5.7}$$

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. (5.7) as

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

where X_t are random signs. Applying Lemma 5.19 with $c = \sqrt{\log(1/\delta)}$, we conclude that Eq. (5.7) is at least $\Omega(\sqrt{\log(1/\delta)}/\sqrt{T})$ with probability at least δ. 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 δ. ■

Chapter 6

Extensions and Generalizations

6.1 Generalizations

In this section, we discuss generalizations of our results. In Subsection 6.1.1, we explain that the scaling of the function (e.g., Lipschitzness) can be normalized without loss of generality. In Subsection 6.1.2, we state some results which demonstrate that the assumption of almost surely bounded noise can be relaxed to subgaussian noise in our upper bounds (Theorems 1.9, 1.10 and 1.17). The proofs of these results can be found in Section 6.3.

6.1.1 Scaling assumptions

For most of this thesis 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 and so we only sketch the proof.

Strongly convex setting

Theorem 6.1. *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 $x^* = \arg\min_{x \in \mathscr{X}} f(x)$. 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 6.2. *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 $x^* = \arg\min_{x \in \mathscr{X}} f(x)$. Then, with*

probability at least $1-\delta$*,*

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

We prove these theorems by reduction to Theorem 1.9 and Theorem 1.17, respectively. That is, suppose that f is a function that has strong convexity parameter α and Lipschitz parameter L. We construct a function h that is 1-Lipschitz and 1-strongly convex (using Claim 6.3) and a subgradient oracle such that running SGD on h with this subgradient oracle is equivalent to running SGD on f. Formally, we prove the following two claims:

Claim 6.3. *Let f be an α-strongly convex and L-Lipschitz function. Then, $h(x) := \frac{\alpha}{L^2} f(\frac{L}{\alpha}x)$ is 1-Lipschitz and 1-strongly convex.*

Claim 6.4. *Suppose f is α-strongly convex and L-Lipschitz on a domain $\mathscr{X} \subset \mathbb{R}^n$. Let the initial point $x_1 \in \mathscr{X}$ be given. Let h be as defined in Claim 6.3. 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 feasible region $\mathscr{X}$*
- *the execution of Algorithm 1 on input h with initial point $\tilde{x}_1 := (\alpha/L)x_1$, step size $\tilde{\eta}_t = 1/t$ and feasible region $(\alpha/L)\mathscr{X}$*

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 h and x_t for the execution on f, then $\tilde{x}_t = (\alpha/L)x_t$.

Theorem 6.1 and Theorem 6.2 follow easily now. We prove Theorem 6.1 by reducing to Theorem 1.9. The proof of Theorem 6.2 can be obtained by reducing similarly to Theorem 1.17.

Proof (of Theorem 6.1). Suppose f is α strongly-convex and L-Lipschitz and let $\mathscr{X}$ be the feasible region. Let $h(x) = \frac{\alpha}{L^2} f\left(\frac{L}{\alpha}x\right)$. By Claim 6.3, h is a 1 strongly-convex and 1-Lipschitz function. Note that $\min_{x\in(\alpha/L)\mathscr{X}} h(x) = \frac{\alpha}{L^2}\min_{x\in\mathscr{X}} f(x)$. Using the coupling provided in Claim 6.4, let x_t be the iterates of produced by running SGD when using f over the region $\mathscr{X}$ and $\tilde{x}_t$ be the iterates of SGD when using h over the region $(\alpha/L)\mathscr{X}$. Claim 6.4 shows that $\tilde{x}_t = (\alpha/L)x_t$. Therefore, we have

$$h(\tilde{x}_{T+1}) - \min_{x\in(\alpha/L)\mathscr{X}} h(x) \;=\; \frac{\alpha}{L^2}\left(f(x_{T+1}) - \min_{x\in\mathscr{X}} f(x)\right).$$

Hence, applying Theorem 1.9, we have our desired result by noting that

$$h(\tilde{x}_{T+1}) - \min_{x\in(\alpha/L)\mathscr{X}} h(x) = O\left(\frac{\log(T)\log(1/\delta)}{T}\right) \qquad \text{with probability } 1-\delta.$$

■

Now it remains to prove Claim 6.4 and Claim 6.3

Proof (of Claim 6.4). The coupling is given by constraining the algorithms to run in parallel and enforcing the execution of SGD on h 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 (i.e. $\mathrm{E}[\hat{g}_t] \in \partial f(x_t)$), then we set the output of the subgradient oracle of the execution of SGD on h at step t to be $\frac{1}{L}\hat{g}_t$.

The subgradient oracle for h is valid if $\tilde{x}_t = (\alpha/L)x_t$. Indeed, $\partial h(x) = \frac{1}{L} f\left(\frac{L}{\alpha}x\right)$ by Claim A.9. Therefore, $\partial h((\alpha/L)x_t) = \frac{1}{L}\partial f(x_t)$. Since $\mathrm{E}[\hat{g}_t] \in \partial f(x_t)$ then $\mathrm{E}[(1/L)\hat{g}_t] \in \partial h((\alpha/L)x_t)$. Next, we prove via induction that $\tilde{x}_t = (\alpha/L)x_t$.

By definition, $\tilde{x}_1 = (\alpha/L)x_1$, which handles the base case. 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 h at $\tilde{x}_t$ is $\frac{1}{L}\partial f(x_t)$. Therefore, the subgradient oracle for h is valid at this step. Now, $y_{t+1} = x_t - \frac{1}{\alpha t}\hat{g}_t$. Meanwhile, $\tilde{y}_{t+1} = \tilde{x}_t - \frac{1}{t}\frac{1}{L}\hat{g}_t = \frac{\alpha}{L}(x_t - \frac{1}{\alpha t}\tilde{g}_t) = \frac{\alpha}{L}y_{t+1}$. Therefore,

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

as desired. ■

Proof (of Claim 6.3). First we show that h is 1-Lipschitz:

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

The inequality holds since f is L-Lipschitz.

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

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

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

Lipschitz setting

Theorem 6.5. *Suppose f is an L-Lipschitz function on $\mathscr{X}$ and that $\hat{z}_t$ has norm at most L almost surely. Consider running Algorithm 1 for T iterations with step size $\eta_t = \frac{R}{L\sqrt{t}}$. Let $f^* = \min_{\mathscr{X}} f(x)$.Then, with probability at least $1-\delta$,*

$$f(x_{T+1}) - f^* \;=\; O\left(\frac{RL\cdot\log(T)\log(1/\delta)}{\sqrt{T}}\right).$$

Theorem 6.5 can be proved by a reduction to Theorem 1.10, similarl to the strongly-convex setting. We do not provide proofs, as the ideas are similar to the strongly-convex setting and can be easily adapted to prove the results in this setting. The main claims are as follows.

Claim 6.6. *Let f be an L-Lipschitz function on $\mathscr{X}$. Then, $h(x) := \frac{1}{RL}f(Rx)$ is 1-Lipschitz on $(1/R)\mathscr{X}$ and $\min_{x\in(1/R)\mathscr{X}} h(x) = \frac{1}{RL}\min_{x\in\mathscr{X}} f(x)$. Note that $\operatorname{diam}(\mathscr{X}) \le 1$.*

Claim 6.7. *Suppose f is L-Lipschitz on a domain $\mathscr{X} \subset \mathbb{R}^n$. Let the initial point $x_1 \in \mathscr{X}$ be given. Let h be as defined in Claim 6.6. Then, there is a coupling between the following two processes:*

- *the execution of Algorithm 1 on input h with initial point $\tilde{x}_1 := (1/R)x_1$, step size $\tilde{\eta}_t$ and feasible region $(1/R)\mathscr{X}$, and*

- *the execution of Algorithm 1 on input f with initial point x_1, step size $\eta_t = \frac{R}{L}\tilde{\eta}_t$ and feasible region $\mathscr{X}$*

such that the iterates of the first process correspond to the iterates of the second process scaled by $\frac{1}{R}$. That is, if we denote by $\tilde{x}_t$ the iterates of the execution of SGD using h and x_t for the execution on f, then $\tilde{x}_t = (1/R)x_t$.

6.1.2 Subgaussian noise

Theorems 1.9 and 1.10 assume that the stochastic gradient oracle produces noise at each step that almost surely has Euclidean norm at most 1. It is possible to relax this assumption to include stochastic gradient oracles which allow unbounded noise as long as the norm of the noise is sufficiently "concentrated" around some constant. The notion of concentration which we use is *subgaussian-ness*. For example, normally distributed random variables are subgaussian. See Section 6.2 for a formal introduction to subgaussian random variables. Below we informally state our high probability upper bounds generalized to handle subgaussian noise. We will prove a formal version of Theorem 6.8 in Section 6.3. A formal version of Theorem 6.9 can be obtained by applying similar modifications to the proof of Theorem 6.9 as was done to the proof of Theorem 1.9 to obtain Theorem 6.8.

Theorem 6.8 (Informal subgaussian extension, strongly convex case)**.** *Suppose f is 1-strongly convex and 1-Lipschitz. Assume $\|\hat{z}_t\|$ is a subgaussian random variable conditioned on $\mathscr{F}_{t-1}$. Consider running Algorithm 1 for T iterations with step size $\eta_t = 1/t$. Let $x^* = \arg\min_{x\in\mathscr{X}} f(x)$. Then, with probability at least $1-\delta$,*

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

Theorem 6.9 (Informal subgaussian extension, Lipschitz case)**.** *Suppose f is 1-Lipschitz and $\operatorname{diam}(\mathscr{X}) \leq 1$. Assume $\|\hat{z}_t\|$ is a subgaussian random variable conditioned on $\mathscr{F}_{t-1}$. Consider running Algorithm 1 for T iterations with step size $\eta_t = 1/\sqrt{t}$. Let $f^* = \min_{x\in\mathscr{X}} f(x)$. Then, with probability at least $1-\delta$,*

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

Remark 6.10. *Theorem 6.8 and Theorem 6.9 can be extended to handle functions with arbitrary Lipschitz and strong convexity constants using the reductions from Subsection 6.1.1.*

6.2 Subgaussian and subexponential random variables

In this section, we go over some basic preliminaries on subgaussian and subexponential random variables; the reader is referred to the book by Vershynin [45, Chapter 2] for further background. The results stated here will be useful in Section 6.3 where we extend our high probability convergence by relaxing the bounded noise assumption.

6.2.1 Subgaussian random variables

Definition 6.11 (Subgaussian random variable)**.**

- *For a random variable X, define*

$$\|X\|_{\psi_2} \;=\; \inf\left\{\, t>0 : \mathrm{E}\left[\exp\left(X^2/t^2\right)\right] \le 2 \,\right\}.$$

We call X subgaussian if $\|X\|_{\psi_2}$ is finite.

- *For a random variable X and a sigma-algebra $\mathscr{F}$, define*

$$\|X \mid \mathscr{F}\|_{\psi_2} \;=\; \inf\left\{\, t>0 : \mathrm{E}\left[\exp\left(X^2/t^2\right) \mid \mathscr{F}\right] \le 2 \,\right\}.$$

We call X subgaussian conditioned on $\mathscr{F}$ if $\|X \mid \mathscr{F}\|_{\psi_2}$ is finite.

Fact 6.12. *$\|\cdot\|_{\psi_2}$ is a norm. It is refered to as the ψ_2-norm.*

Claim 6.13. *Let X be a subgaussian random variable. Then* $\mathrm{E}\left[\exp\left(X^2/\|X\|_{\psi_2}^2\right)\right] \le 2$.

Proof. Let $S = \left\{\, t>0 : \mathrm{E}\left[\exp\left(X^2/t^2\right)\right] \le 2 \,\right\}$. For every $\varepsilon>0$, we have that $\|X\|_{\psi_2}+\varepsilon > t$ for some $t \in S$. Hence,

$$\mathrm{E}\left[\exp\left(X^2/(\|X\|_{\psi_2}+\varepsilon)^2\right)\right] \;<\; \mathrm{E}\left[\exp\left(X^2/t^2\right)\right] \le 2.$$

Therefore, for every $\varepsilon>0$ we have $\mathrm{E}\left[\exp\left(X^2/(\|X\|_{\psi_2}+\varepsilon)^2\right)\right] < 2$. Taking $\varepsilon \to 0$ and applying the monotone convergence theorem completes the proof. ■

Claim 6.14. *Let X be a subgaussian random variable. Then* $\mathrm{E}\left[\exp\left(\lambda^2 X^2\right)\right] \le \exp\left(\lambda^2\|X\|_{\psi_2}^2\right)$ *for all* $|\lambda| \le 1/\|X\|_{\psi_2}$.

Proof. Suppose $|\lambda| \le \frac{1}{\|X\|_{\psi_2}}$, which implies that the function $x \mapsto x^{\lambda^2\|X\|_{\psi_2}^2}$ is concave. Hence,

$$\begin{aligned}
\mathrm{E}\left[\exp\left(\lambda^2 X^2\right)\right] \;&=\; \mathrm{E}\left[\exp\left(\lambda^2\|X\|_{\psi_2}^2 X^2/\|X\|_{\psi_2}^2\right)\right] \\
&\le\; \mathrm{E}\left[\exp\left(X^2/\|X\|_{\psi_2}^2\right)\right]^{\lambda^2\|X\|_{\psi_2}^2} \qquad \text{(by Jensen's inequality)} \\
&\le\; 2^{\lambda^2\|X\|_{\psi_2}^2} \qquad \text{(by Claim 6.13)} \\
&\le\; \exp\left(\lambda^2\|X\|_{\psi_2}^2\right).
\end{aligned}$$

■

Claim 6.15. *Suppose that* $\mathrm{E}\left[\exp\left(\lambda^2 X^2\right)\right] \le \exp\left(\lambda^2\kappa^2\right)$ *for all* $|\lambda| \le 1/\kappa$. *Then* $\|X\|_{\psi_2} \le \kappa/\sqrt{\ln 2}$.

Proof. Let $\tilde{\lambda} = \sqrt{\ln 2}/\kappa \le 1/\kappa$. By Claim 6.14,

$$\mathrm{E}\left[\exp\left((\ln 2)X^2/\kappa^2\right)\right] \;=\; \mathrm{E}\left[\exp\left(\tilde{\lambda}^2 X^2\right)\right] \;\le\; \exp\left(\tilde{\lambda}^2\kappa^2\right) \;=\; 2.$$

Hence, by Definition 6.11, we have $\|X\|_{\psi_2} \le \kappa/\sqrt{\ln 2}$. ■

The combination of Claim 6.15 and Claim 6.14 establishes that the two properties $\mathrm{E}\left[\exp\left(X^2/t\right)\right] \le 2$ and $\mathrm{E}\left[\exp\left(\lambda^2 X^2\right)\right] \le \exp\left(\lambda^2 t^2\right)$ for all $|\lambda| \le 1/t$ are "equivalent". That is, if a random variable satisfies one of

the properties with value t, then it satisfies the other property with value t' which differs from t only by some constant factor independent of the random variable X.

Claim 6.16 ($\|\cdot\|_{\psi_2}$ bound to tail bound)**.** *Suppose* $\|X\|_{\psi_2} \le \kappa$*. Then,*

$$\Pr\left[X \le \kappa\sqrt{\log(e/\delta)}\right] \;\ge\; 1-\delta \qquad \forall \delta \in (0,1].$$

Proof. By the exponentiated Markov inequality

$$\Pr[X \ge t] \;\le\; \exp(-\lambda^2 t^2)\,\mathrm{E}\left[\exp\left(\lambda^2 X^2\right)\right] \;\le\; \exp\left(-\lambda^2 t^2 + \lambda^2\kappa^2\right) \qquad \text{(for all } |\lambda| \le 1/\kappa),$$

where the second inequality uses Claim 6.14. Setting $\lambda = 1/\kappa$ and $t = \kappa\sqrt{\log(e/\delta)}$ completes the proof. ■

The following Claim is an extension of the well known Hoeffding's Lemma (Lemma A.5) from bounded random variables to subgaussian random variables.

Claim 6.17. *Suppose* X *is mean-zero such that* $\|X\|_{\psi_2} \le \kappa$*. Then* $\mathrm{E}[\exp(\lambda X)] \le \exp\left(\lambda^2\kappa^2\right)$ *for all* $\lambda \in \mathbb{R}$*.*

Proof. Claim 6.14 implies that $\mathrm{E}\left[\exp\left(\lambda^2 X^2\right)\right] \le \exp\left(\lambda^2\kappa^2\right)$ for all $|\lambda| \le \frac{1}{\kappa}$. Hence, the result follows from Claim A.6. ■

6.2.2 Subexponential random variables

Definition 6.18 (Subexponential random variable)**.**

- *For a random variable* X*, define*

$$\|X\|_{\psi_1} \;=\; \inf\{\, t > 0 : \mathrm{E}[\exp(|X|/t)] \le 2 \,\}.$$

  We call X *subexponential if* $\|X\|_{\psi_1}$ *is finite.*

- *For a random variable* X *and a sigma-algebra* $\mathscr{F}$*, define*

$$\|X \mid \mathscr{F}\|_{\psi_1} \;=\; \inf\{\, t > 0 : \mathrm{E}[\exp(|X|/t) \mid \mathscr{F}] \le 2 \,\}.$$

  We call X *subexponential conditioned on* $\mathscr{F}$ *if* $\|X \mid \mathscr{F}\|_{\psi_1}$ *is finite.*

Fact 6.19. $\|\cdot\|_{\psi_1}$ *is a norm. It is referred to as the* ψ_1*-norm.*

Claim 6.20. *Let* X *be a subexponential random variable. Then* $\mathrm{E}\left[\exp\left(X/\|X\|_{\psi_1}\right)\right] \le 2$*.*

Proof. Let $S = \{\, t > 0 : \mathrm{E}[\exp(|X|/t)] \le 2 \,\}$. For every $\varepsilon > 0$, we have that $\|X\|_{\psi_1} + \varepsilon > t$ for some $t \in S$. Hence,

$$\mathrm{E}\left[\exp\left(|X|/(\|X\|_{\psi_1} + \varepsilon)\right)\right] \;<\; \mathrm{E}[\exp(|X|/t)] \le 2.$$

Therefore, for every $\varepsilon > 0$ we have $\mathrm{E}\left[\exp\left(|X|/(\|X\|_{\psi_1} + \varepsilon)\right)\right] < 2$. Taking $\varepsilon \to 0$ and applying the monotone convergence theorem completes the proof. ■

Claim 6.21. *Let X be a subexponential random variable. Then* $\mathrm{E}[\exp(\lambda|X|)] \le \exp\left(\lambda \|X\|_{\psi_1}\right)$ *for all* $\lambda \in [0, 1/\|X\|_{\psi_1}]$.

Proof. Suppose $\lambda \in [0, 1/\|X\|_{\psi_1}]$, which implies that the function $x \mapsto x^{\lambda\|X\|_{\psi_1}}$ is concave. Hence,

$$\begin{aligned}
\mathrm{E}[\exp(\lambda|X|)] &= \mathrm{E}\left[\exp\left(\lambda\|X\|_{\psi_1}|X|/\|X\|_{\psi_1}\right)\right] \\
&\le \mathrm{E}\left[\exp\left(|X|/\|X\|_{\psi_1}\right)\right]^{\lambda\|X\|_{\psi_1}} \qquad \text{(by Jensen's inequality)} \\
&\le 2^{\lambda\|X\|_{\psi_1}} \qquad \text{(by Claim 6.20)} \\
&\le \exp\left(\lambda\|X\|_{\psi_1}\right).
\end{aligned}$$

■

Claim 6.22. *Suppose that* $\mathrm{E}[\exp(\lambda|X|)] \le \exp(\lambda\kappa)$ *for all* $\lambda \in [0, 1/\kappa]$. *Then* $\|X\|_{\psi_1} \le \kappa/(\ln 2)$.

Proof. Let $\tilde{\lambda} = (\ln 2)/\kappa$, so clearly $\tilde{\lambda} \in [0, 1/\kappa]$. By Claim 6.21,

$$\mathrm{E}[\exp(\ln 2\,|X|/\kappa)] = \mathrm{E}\left[\exp\left(\tilde{\lambda}|X|\right)\right] \le \exp\left(\tilde{\lambda}\kappa\right) = 2.$$

Hence, by Definition 6.18, we have $\|X\|_{\psi_1} \le \kappa/(\ln 2)$. ■

The combination of Claim 6.22 and Claim 6.21 establishes that the two properties $\mathrm{E}[\exp(|X|/t)] \le 2$ and $\mathrm{E}[\exp(\lambda|X|)] \le \exp(\lambda t)$ for all $|\lambda| \le 1/t$ are "equivalent". That is, if a random variable satisfies one of the properties with value t, then it satisfies the other property with value t' which differs from t only by some constant factor independent of the random variable X.

Claim 6.23 ($\|\cdot\|_{\psi_1}$ bound to tail bound)**.** *Suppose* $\|X\|_{\psi_1} \le \kappa$. *Then, for every* $\delta \in (0,1)$, $X \le \kappa\log(e/\delta)$ *with probability at least* $1-\delta$.

Proof. By the exponentiated Markov inequality,

$$\Pr[X \ge t] \le \exp(-\lambda t)\,\mathrm{E}[\exp(\lambda X)] \le \exp(-\lambda t + \lambda\kappa) \qquad (\text{for all } \lambda \in (0, 1/\kappa]),$$

where the second inequality follows from Claim 6.21. Setting $\lambda = 1/\kappa$ and $t = \kappa\log(e/\delta)$ completes the proof. ■

6.2.3 Relationship between subgaussian and subexponential random variables

Subgaussian and subexponential random variables share a simple connection, as hilighted by the result below.

Claim 6.24. *Suppose X is a subgaussian random variable. Then, X^2 is subexponential with* $\left\|X^2\right\|_{\psi_1} = \|X\|_{\psi_2}^2$

Proof. This is easy to see by comparing the definition of each norm:

$$\left\|X^2\right\|_{\psi_1} = \inf\underbrace{\left\{\, t > 0 : \mathrm{E}\left[\exp\left(X^2/t\right)\right] \le 2 \,\right\}}_{:=A},$$

$$\|X\|_{\psi_2} \;=\; \inf\underbrace{\left\{\, t>0 : \mathrm{E}\left[\exp\left(X^2/t^2\right)\right] \le 2 \,\right\}}_{:=B}.$$

Clearly, $\left\{\, b^2 : b \in B \,\right\} = A$. Hence,

$$\left\|X^2\right\|_{\psi_1} \;=\; \inf A \;=\; (\inf B)^2 \;=\; \|X\|_{\psi_2}^2,$$

as desired. ■

6.3 Upper bound on error of final iterate: subgaussian noise

Theorems 1.9 and 1.10 assume that the stochastic gradient oracle produces noise at each step that almost surely has Euclidean norm at most 1. In this section, we work with the weaker assumption that the noise at each step is conditionally subgaussian. (A formal definition of this term appears in Section 6.2.)

As before, 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 $\mathscr{F}_t = \sigma(\hat{z}_1, \ldots, \hat{z}_t)$ be the σ-algebra generated by the first t steps of SGD. Finally, recall that $\mathrm{E}[\,\hat{z}_t \mid \mathscr{F}_{t-1}\,] = 0$. Instead of assuming that $\|\hat{z}_t\| \le 1$ almost surely, we will assume that $\|\hat{z}_t\|$ is a subgaussian random variable conditioned on $\mathscr{F}_{t-1}$.

We will prove an extension of Theorem 1.9 in the following subsection (see Theorem 6.25). A similar extension for Theorem 1.10 can also be obtained, however its proof is omitted since it relies mainly on the ideas introduced in the next section.

6.3.1 Upper bound on error of final iterate, strongly convex case with subgaussian noise

The goal of this subsection is to prove Theorem 6.25. The analysis will reuse many claims from the analysis of Theorem 1.9 and follow the general approach taken there as well.

Assumptions. The main difference between Theorem 6.25 and Theorem 1.9 is the assumption on the noise produced by the stochastic gradient oracle. In this subsection, we assume $\left\|\|\hat{z}_t\|_2 \mid \mathscr{F}_{t-1}\right\|_{\psi_2} \le \kappa$ for every t.

Theorem 6.25. *Suppose f is 1-strongly convex and 1-Lipschitz. Suppose that $\left\|\|\hat{z}_t\| \mid \mathscr{F}_{t-1}\right\|_{\psi_2} \le \kappa$ for all t. Consider running Algorithm 1 for T iterations with step size $\eta_t = 1/t$. Let $x^* = \arg\min_{x\in\mathscr{X}} f(x)$. Then, with probability at least $1-\delta$,*

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

One important result that we need is a bound on $\left\|\|\hat{g}_t\|^2\right\|_{\psi_1}$. In Section 3.1, this was not necessary, since $\|\hat{g}_t\|^2$ was almost-surely bounded by 4. This made for a clean analysis. Here, $\|\hat{g}_t\|^2$ can potentially be unbounded, and therefore we need to provide control on it's tail distribution. Indeed, this is accomplished by the following claim.

Claim 6.26. *For every t, $\left\|\|\hat{g}_t\| \mid \mathscr{F}_{t-1}\right\|_{\psi_2} = O(\kappa+1)$. Therefore, $\left\|\|\hat{g}_t\|^2 \mid \mathscr{F}_{t-1}\right\|_{\psi_1} = O\left((\kappa+1)^2\right)$. This implies $\left\|\|\hat{g}_t\|^2\right\|_{\psi_1} = O\left((\kappa+1)^2\right)$.*

Proof. Recall, $\hat{g}_t = g_t - \hat{z}_t$ where $g_t \in \partial f(x_t)$ and $\hat{z}_t$ is mean-zero conditioned on $\mathscr{F}_{t-1}$ such that $\left\|\|\hat{z}_t\| \mid \mathscr{F}_{t-1}\right\|_{\psi_2} \leq \kappa$. Now, because $\|\cdot\|_{\psi_1}$ is a norm, we have

$$\left\|\|\hat{g}_t\| \mid \mathscr{F}_{t-1}\right\|_{\psi_2} \;\leq\; \left\|\|g_t\| + \|\hat{z}_t\| \mid \mathscr{F}_{t-1}\right\|_{\psi_2} \;\leq\; \left\|\|g_t\| \mid \mathscr{F}_{t-1}\right\|_{\psi_2} + \left\|\|\hat{z}_t\| \mid \mathscr{F}_{t-1}\right\|_{\psi_2} \;\leq\; \kappa + 1/\sqrt{\ln 2}.$$

We have used that $\left\|\|g_t\| \mid \mathscr{F}_{t-1}\right\|_{\psi_2} \leq \left\|1 \mid \mathscr{F}_{t-1}\right\|_{\psi_2}$ by 1-Lipschitzness of f. By definition, we have $\left\|1 \mid \mathscr{F}_{t-1}\right\|_{\psi_2} \leq 1/\sqrt{\ln 2}$. Lastly, we have used the assumption that $\left\|\|\hat{z}_t\| \mid \mathscr{F}_{t-1}\right\|_{\psi_2} \leq \kappa$. ■

Lemma 6.27. *Let f be 1-strongly convex and 1-Lipschitz. Suppose that we run SGD (Algorithm 1) with step sizes $\eta_t = 1/t$ and that $\left\|\hat{z}_t \mid \mathscr{F}_{t-1}\right\|_{\psi_2} \leq \kappa$. Then*

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

where X is a random variable such that $\|X\|_{\psi_1} = O\left((\kappa+1)^2\,\frac{\log T}{T^2}\right)$.

Proof. We may proceed just as in the proof of Lemma 3.1. Applying Eq. (3.4), we obtain

$$f(x_T) \;\leq\; S_{T/2} + \sum_{k=1}^{T/2}\frac{1}{2k(k+1)}\sum_{t=T-k}^{T}\eta_t\,\|\hat{g}_t\|^2 + \sum_{k=1}^{T/2}\frac{1}{k(k+1)}\sum_{t=T-k}^{T}\langle\, \hat{z}_t, x_t - x_{T-k}\,\rangle,$$

where $S_{T/2} = \frac{1}{T/2+1}\sum_{t=T/2}^{T} f(x_t)$. From this point, the only difference is analyzing the following sum:

$$\sum_{k=1}^{T/2}\frac{1}{2k(k+1)}\sum_{t=T-k}^{T}\eta_t\,\|\hat{g}_t\|^2,$$

because we cannot bound $\|\hat{g}_t\|^2$ as was done in the proof of Lemma 3.1. Instead, we may bound it's ψ_1-norm using the fact that $\left\|\|\hat{g}_t\|^2\right\|_{\psi_1} = O\left((\kappa+1)^2\right)$ (Claim 6.26) and the triangle inequality for $\|\cdot\|_{\psi_1}$. This yields:

$$\left\|\sum_{k=1}^{T/2}\frac{1}{2k(k+1)}\sum_{t=T-k}^{T}\eta_t\,\|\hat{g}_t\|^2\right\|_{\psi_1} \leq O\left((\kappa+1)^2\,\frac{\log T}{T}\right),$$

as desired. ■

Since $\|X\|_{\psi_1} \leq O\left((\kappa+1)^2\,\frac{\log T}{T}\right)$, we have that by Claim 6.23, $X \leq O\left((\kappa+1)^2\,\frac{\log T}{T}\log(1/\delta)\right)$, with probability at least $1-\delta$. The suffix average is bounded by $O\left((\kappa+1)^2\,\frac{\log(1/\delta)}{T}\right)$ by Theorem 6.37 (we defer the proof to Subsection 6.3.3). Therefore, it remains to analyze Z_T. By changing the order of summation, we can write $Z_T = \sum_{t=T/2}^{T}\langle\, \hat{z}_t, w_t\,\rangle$ where

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

We will prove the following lemma, whose proof is outlined in Sub-subsection 6.3.1.

Lemma 6.28. *$Z_T = O\left(\kappa(\kappa+1)\frac{\log T \log(1/\delta)}{T}\right)$ with probability at least $1-\delta$.*

Theorem 6.25 follows from Theorem 6.37, Lemma 6.28 and Lemma 6.27.

Bounding the noise

The main idea is to follow the steps in Subsection 3.1.1. However, certain steps will require a slightly different analysis due to the absence of a bound on $\|\hat{g}_t\|^2$. Lemma 3.3 still holds in the subgaussian noise case. We restate it here:

Lemma 3.3. 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,

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

However, we prove a slight variant of Lemma 3.4.

Lemma 6.29. *There exists positive values $R = O\left(\frac{\log T}{T}\right)$, $C_t = \Theta(\log(T-t))$, $A_t = O\left(\frac{\log T}{T^2}\right)$ and a non-negative random variable X with $\|X\|_{\psi_1} = O\left(\frac{(\kappa+1)^2 \log^2(T)}{T^2}\right)$ such that*

$$\sum_{t=T/2}^{T} \|w_t\|^2 \leq X + R\left\|x_{T/2} - x^*\right\|^2 + \sum_{t=T/2}^{T-1} \langle\, \hat{z}_t, \frac{C_t}{t} w_t \,\rangle + \sum_{t=T/2}^{T-1} \langle\, \hat{z}_t, A_t(x_t - x^*) \,\rangle.$$

Proof. We may follow the proof of Lemma 3.4 from Subsection 3.3.4, with some minor differences. Here we will only provide a brief *outline* of the differences.

We redefine Λ_1 as follows, while keeping the definition of Λ_2 and Λ_3 as in the proof of Lemma 3.4.

$$\Lambda_1 := \sum_{t=T/2}^{T} \frac{1}{T-t+1} \sum_{j=T/2}^{t-1} \alpha_j \sum_{i=j}^{j-1} \frac{\|\hat{g}_i\|^2}{i^2}.$$

Recall, $\sum_{t=T/2}^{T} \|w_t\|^2 \leq \Lambda_1 + \Lambda_2 + \Lambda_3$, just as in Subsection 3.3.4.

Using the triangle inequality and the ψ_1 norm bound on $\|\hat{g}_i\|^2$ from Claim 6.26, we may replace Claim 3.17 with

Claim 6.30. $\|\Lambda_1\|_{\psi_1} = O\left((\kappa+1)^2 \frac{\log^2(T)}{T^2}\right)$.

We may keep Claim 3.19 as our bound on Λ_3:

Claim 3.19.

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

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

It remains to bound Λ_2. We provide a slight variant of Claim 3.18:

Claim 6.31. *There exists positive values R_1, R_2 such that $R_1 = O\left(\frac{\log T}{T}\right)$ and $R_2 = O\left(\frac{\log T}{T^2}\right)$ and a non-negative random variable, X, such that $\|X\|_{\psi_1} = O\left(\frac{(\kappa+1)^2\log(T)}{T^2}\right)$, where*

$$\Lambda_2 \;\le\; X + R_1\left\|x_{T/2} - x^*\right\|^2 + R_2\sum_{t=T/2}^{T-1}\langle\,\hat{z}_t, x_t - x^*\,\rangle.$$

Proof. We may follow the proof of Claim 3.18 (found in Subsection 3.3.5) up to Eq. (3.5). At this stage, we may continue the proof without bounding $\|\hat{g}_t\|^2$, and instead collect these terms to obtain:

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

At this step, we can take $X = O\left(\frac{\log T}{T^2}\right)\sum_{t=T/2}^{T-1}\frac{1}{t}\|\hat{g}_t\|^2$. It's ψ_1-norm is bounded as desired by using Claim 6.26 and the triangle inequality for the ψ_1-norm. ■

Lemma 6.29 follows from Claim 6.30, Claim 6.31, and Claim 3.19. ■

Just as in Section 3.1, we will use Theorem 1.11. A snag is that our Lemma 6.29 does not quite bound the SSCM of Z_T by a linear transformation of Z_T. However, Lemma 6.33 refines Lemma 6.29 so that we may use Theorem 1.11 (similar to how Lemma 3.6 refines Lemma 3.4) . The proof of Lemma 6.33 uses the following analog of Theorem 3.5 (see Subsection 6.3.2 for a proof).

Theorem 6.32. *Both of the following hold:*

- *For all $t \ge 2$, $\|x_t - x^*\|^2 = O\left((\kappa+1)^2\log(1/\delta)/t\right)$ with probability $1-\delta$, and*
- *Let $\sigma_t \ge 0$ for $t = 2,\ldots,T$. Then, $\sum_{t=1}^{T}\sigma_t\|x_t - x^*\|^2 = O\left((\kappa+1)^2\log(1/\delta)\sum_{t=2}^{T}\frac{\sigma_t}{t}\right)$ with probability $1-\delta$.*

Lemma 6.33. *For every $\delta \in (0,1)$, there exists positive values $R = \left((\kappa+1)^2\frac{\log^2(T)\log(1/\delta)}{T^2}\right)$, and $C_t = O(\log T)$ such that $\sum_{t=T/2}^{T}\|w_t\|^2 \le R + \sum_{t=T/2}^{T-1}\frac{C_t}{t}\langle\,\hat{z}_t, w_t\,\rangle$ with probability at least $1-\delta$.*

Proof.

From Lemma 6.29, we have

$$\sum_{t=T/2}^{T}\|w_t\|^2 \;\le\; X + R'\left\|x_{T/2} - x^*\right\|^2 + \sum_{t=T/2}^{T-1}\langle\,\hat{z}_t, \frac{C_t}{t}w_t\,\rangle + \sum_{t=T/2}^{T-1}\langle\,\hat{z}_t, A_t(x_t - x^*)\,\rangle,$$

where X, R', A_t, C_t are as promised by Lemma 6.29. Because $\|X\|_{\psi_1} = O\left(\frac{(\kappa+1)^2\log T}{T^2}\right)$, then by Claim 6.23, we have $X = O\left((\kappa+1)^2\frac{\log T\log(1/\delta)}{T^2}\right)$, with probability at least $1-\delta$.

Furthermore, Theorem 6.32 states that $\left\|x_{T/2} - x^*\right\|^2 = O\left(\frac{(\kappa+1)^2\log(1/\delta)}{T}\right)$ with probability at least $1-\delta$. Hence, $R'\left\|x_{T/2} - x^*\right\|^2 = O\left((\kappa+1)^2\frac{\log(1/\delta)\log(T)}{T^2}\right)$ with probability at least $1-\delta$.

Lastly, it remains to deal with $\sum_{t=T/2}^{T-1}\langle\,\hat{z}_t, A_t(x_t - x^*)\,\rangle$. Observe that this is a sum of a martingale difference sequence with SSCM bounded by $\sum_{t=T/2}^{T-1}A_t^2\|x_t - x^*\|^2$. Using Theorem 6.32 (the upper bound on squared distances of the iterates x_t to x^*), this is bounded above by $O\left((\kappa+1)^2\frac{\log(1/\delta)\log^2(T)}{T^4}\right)$. Now, we can translate

our SSCM bound to a high probability bound on the martingale itself by applying Corollary 6.40 (which is an application of Theorem 1.11 for bounding a martingale with subgaussian increments when it's SSCM is bounded) to obtain $\sum_{t=T/2}^{T-1}\langle \hat{z}_t, A_t(x_t - x^*)\rangle = O\left(\kappa(\kappa+1)\frac{\log(1/\delta)\log(T)}{T^2}\right)$ with probability at least $1-\delta$. ■

Now, we are ready to complete the proof of Lemma 6.28.

Proof (of Lemma 6.28). Let R, C_t, X be as promised from Lemma 6.33. These values provide a high probability upper bound of $R'\log(1/\delta) + \sum_{t=T/2}^{T-1}\frac{C_t}{t}\langle \hat{z}_t, w_t\rangle$ on the SSCM of Z_T, where $R'\log(1/\delta) = R$. Corollary 6.41 (an application of Theorem 1.11 for martingales with subgaussian increments) allows us to then bound Z_T by $\kappa\sqrt{R'}\log(1/\delta)$ with probability at least $1-\delta$. Indeed, recalling that $R' = O\left(\frac{(\kappa+1)^2\log^2(T)}{T^2}\right)$ yields our desired result. ■

6.3.2 High probability bounds on squared distances to x^*

The main goal of this subsection is to prove Theorem 6.32.

Claim 6.34. *Suppose f is 1-strongly-convex and 1-Lipschitz. Define $Y_t = t\|x_{t+1} - x^*\|^2$ and $U_t = \langle \hat{z}_{t+1}, x_{t+1} - x^*\rangle / \|x_{t+1} - x^*\|$. Then,*

$$Y_{t+1} \le \left(\frac{t-1}{t}\right)Y_t + 2\cdot U_t\sqrt{\frac{Y_t}{t}} + \frac{\|\hat{g}_{t+1}\|^2}{t+1}.$$

The proof of Claim 6.34 is identical to the proof of Claim 3.7, except that we cannot bound $\|\hat{g}_t\|^2$ by 4 due to the fact that the noise, $\hat{z}_t$, is not bounded as was assumed in Claim 3.7.

The main tool to prove this theorem is the following analog to Theorem 1.19. See Subsection 6.3.4 for a proof.

Theorem 6.35. *Let $(X_t)_{t=1}^T$ be a stochastic process and let $(\mathscr{F}_t)_{t=1}^T$ be a filtration such that X_t is $\mathscr{F}_t$-measurable and X_t is non-negative almost surely. Let $\alpha_t \in [0,1)$ and $\beta_t, \gamma_t \ge 0$ for every t. Let $\hat{w}_t$ be $\mathscr{F}_{t+1}$-measurable such that $\|\hat{w}_t \mid \mathscr{F}_t\|_{\psi_2} \le \tau$ for every t and $\mathrm{E}[\hat{w}_t \mid \mathscr{F}_t] = 0$. Let $\hat{y}_t$ be $\mathscr{F}_{t+1}$-measurable such that $\|\hat{y}_t \mid \mathscr{F}_t\|_{\psi_1} \le \rho$ for every t. Suppose $X_{t+1} \le \alpha_t X_t + \beta_t\hat{w}_t\sqrt{X_t} + \gamma_t\hat{y}_t$, for every t. Assume that $\mathrm{E}[\exp(\lambda X_1)] \le \exp(\lambda K)$, for $\lambda \in (0, 1/K]$. Then,*

- *For every t, $\Pr[X_t \ge K\log(1/\delta)] \le e\delta$,*
- *If $\sigma_1, \dots, \sigma_T \ge 0$, then $\Pr\left[\sum_{t=1}^T \sigma_t X_t \ge K\log(1/\delta)\sum_{t=1}^T \sigma_t\right] \le e\delta$,*

where $K = \max_{t=1}^T\left(\frac{2\gamma_t\rho}{1-\alpha_t}, \frac{4\beta_t^2\tau^2}{1-\alpha_t}\right)$.

Proof (of Theorem 6.32). Consider the stochastic process $(Y_t)_{t=1}^T$ where Y_t is as defined by Claim 6.34. Note that Y_t satisfied the conditions of Theorem 6.35 with $X_t = Y_t$, $\hat{w}_t = U_t$, $\alpha_t = \frac{t-1}{t}$, $\beta_t = \frac{2}{\sqrt{t}}$, $\gamma_t = \frac{1}{t+1}$, $\hat{y}_t = \|\hat{g}_t\|^2$, $\tau = \kappa$, $\rho = (\kappa+1)^2$. Indeed, Y_t is $\mathscr{F}_t$-measurable and non-negative almost surely. U_t is $\mathscr{F}_{t+1}$-measurable which is mean zero conditioned on $\mathscr{F}_t$. Furthermore, $\|U_t \mid \mathscr{F}_t\|_{\psi_2} \le \kappa$ because $\|\|\hat{z}_t\| \mid \mathscr{F}_t\|_{\psi_2} \le \kappa$. Furthermore, $\|\hat{g}_t\|^2$ is also $\mathscr{F}_{t+1}$-measurable and $\left\|\|\hat{g}_t\|^2 \mid \mathscr{F}_t\right\| \le (\kappa+1)^2$ by Claim 6.26. It's easy to check that $\max_{1\le t\le T}\left(\frac{2\gamma_t\rho}{1-\alpha_t}, \frac{4\beta_t^2\tau^2}{1-\alpha_t}\right) = O\left((\kappa+1)^2\right)$. We must check the following claim:

Claim 6.36. *For all $\lambda \in (0, \frac{1}{\Theta((\kappa+1)^2)}]$, $\mathrm{E}\left[\exp\left(\lambda\|x_2 - x^*\|^2\right)\right] \le \exp\left(\lambda\Theta\left((\kappa+1)^2\right)\right)$.*

Proof.

$$\begin{aligned}
\mathrm{E}\left[\exp\left(\lambda\left\|x_2-x^*\right\|^2\right)\right] &= \mathrm{E}\left[\exp\left(\lambda\left(\left\|\Pi_{\mathscr{X}}(x_1-\hat{g}_1)-x^*\right\|^2\right)\right)\right] \\
&\leq \mathrm{E}\left[\exp\left(\lambda\left(\left\|x_1-x^*-\hat{g}_1\right\|^2\right)\right)\right] \\
&= \mathrm{E}\left[\exp\left(\lambda\left(\left\|x_1-x^*\right\|^2+\left\|\hat{g}_1\right\|^2-2\langle x_1-x^*,\hat{g}_1\rangle\right)\right)\right].
\end{aligned}$$

By 1-strong-convexity and 1-Lipschitzness of f,

$$\left\|x_t-x^*\right\| \geq \langle g_t, x_t-x^*\rangle \geq \frac{1}{2}\left\|x_t-x^*\right\|^2,$$

for every t. Hence,

$$\begin{aligned}
\mathrm{E}\left[\exp\left(\lambda\left\|x_2-x^*\right\|^2\right)\right] &\leq \mathrm{E}\left[\exp\left(\lambda\left(4+\left\|\hat{g}_1\right\|^2+4\left\|\hat{g}_1\right\|\right)\right)\right] \\
&= \exp(4\lambda)\,\mathrm{E}\left[\exp\left(\lambda\left\|\hat{g}_1\right\|^2\right)\exp\left(4\lambda\left\|\hat{g}_1\right\|\right)\right] \\
&\leq \exp(4\lambda)\left(\mathrm{E}\left[\exp\left(2\lambda\left\|\hat{g}_1\right\|^2\right)\right]\right)^{1/2}\left(\mathrm{E}\left[\exp\left(8\lambda\left\|\hat{g}_1\right\|\right)\right]\right)^{1/2} \qquad \text{(by Theorem A.2).}
\end{aligned}$$

Recall that $\left\|\left\|\hat{g}_1\right\|\right\|_{\psi_2} = O(\kappa+1)$ and $\left\|\left\|\hat{g}_1\right\|^2\right\|_{\psi_1} = O(\kappa+1)^2$ (by Claim 6.26). Therefore, we may bound the MGF's of $\left\|\hat{g}_1\right\|$ and $\left\|\hat{g}_1\right\|^2$ using Claim 6.14 for the former and Claim 6.21 for the latter. That is we may write:

$$\mathrm{E}\left[\exp\left(2\lambda\left\|\hat{g}_1\right\|^2\right)\right] \leq \exp\left(2\lambda(\kappa+1)^2\right) \qquad \text{for all } \lambda \in [0, \frac{1}{2(\kappa+1)^2}],$$

and

$$\mathrm{E}\left[\exp\left(8\lambda\left\|\hat{g}_1\right\|\right)\right] \leq \exp\left(8\lambda(\kappa+1)^2\right) \qquad \text{for all } \lambda \in [0, \frac{1}{8(\kappa+1)^2}].$$

Hence, we may plug in these MGF bounds to obtain

$$\mathrm{E}\left[\exp\left(\lambda\left\|x_2-x^*\right\|^2\right)\right] \leq \exp(4\lambda)\exp\left(\lambda(\kappa+1)^2\right)\exp\left(4\lambda(\kappa+1)^2\right),$$

for all $\lambda \in (0, \frac{1}{8(\kappa+1)^2}]$. Hence,

$$\mathrm{E}\left[\exp\left(\lambda\left\|x_2-x^*\right\|^2\right)\right] \leq \exp\left(4\lambda+5\lambda(\kappa+1)^2\right) \leq \exp\left(9\lambda(\kappa+1)^2\right),$$

for all $\lambda \in (0, \frac{1}{8(\kappa+1)^2}]$. ■

■

6.3.3 Suffix averaging

The following is an extension of Theorem 1.17 to the case of subgaussian noise.

Theorem 6.37. *Suppose f is 1-strongly convex and 1-Lipschitz. Consider running Algorithm 1 for T iterations with step size $\eta_t = 1/t$. Let $x^* = \arg\min_{x\in\mathscr{X}} f(x)$. Suppose that there exist $\kappa > 0$, such that $\left\|\hat{z}_t \mid \mathscr{F}_{t-1}\right\|_{\psi_2} \leq \kappa$*

for every t. Then, with probability at least $1-\delta$,

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

Proof. By Lemma 3.14 with $w = x^*$ we have

$$\sum_{t=T/2}^{T}[f(x_t)-f(x^*)] \;\le\; \underbrace{\frac{1}{2}\sum_{t=T/2}^{T}\eta_t\left\|\hat{g}_t\right\|^2}_{(a)} + \underbrace{\frac{1}{2\eta_{T/2}}\left\|x_{T/2}-x^*\right\|^2}_{(b)} + \underbrace{\sum_{t=T/2}^{T}\langle\,\hat{z}_t, x_t-x^*\,\rangle}_{(c)}. \tag{6.1}$$

It suffices to bound each term of the right hand side of (6.1) by $O(\kappa^2\log(1/\delta))$ with probability at least $1-\delta$.

(a). By Claim 6.26, we have

$$\left\|\sum_{t=T/2}^{T}\eta_t\left\|\hat{g}_t\right\|^2\right\|_{\psi_1} \;\le\; \sum_{t=T/2}^{T}\eta_t\left\|\left\|\hat{g}_t\right\|^2\right\|_{\psi_1} \;\le\; O\left((\kappa+1)^2\right).$$

Hence, we apply Claim 6.23 to bound (a) by $O\left((\kappa+1)^2\log(1/\delta)\right)$ with probability at least $1-\delta$.

(b). We have already bounded (b) by $O\left((\kappa+1)^2\log(1/\delta)\right)$ in Theorem 6.32.

(c). To bound (c), we will come up with a high probability bound on the sum of squared conditional magnitudes of (c) (observe that (c) is a martingale with conditionally subgaussian increments) and then apply our transition from high probability bounds on the SSCM to high probability bounds on martingales (Corollary 6.40). The SSCM of (c) is given by $\sum_{t=T/2}^{T}\|x_t-x^*\|^2$. Indeed, by Theorem 6.32, this is bounded by $O\left((\kappa+1)^2\log(1/\delta)\right)$ with probability at least $1-\delta$. This translates into a high probability bound on (c) using Corollary 6.40. Indeed, take $a_t = \hat{z}_t$, $b_t = x_t - x^*$, and $R = \kappa^2$. This yields a bound of $O(\kappa(\kappa+1)\log(1/\delta))$ with probability $1-\delta$ on (c) as desired.

■

6.3.4 Proof of Theorem 6.35

Theorem 6.35. Let $(X_t)_{t=1}^T$ be a stochastic process and let $(\mathscr{F}_t)_{t=1}^T$ be a filtration such that X_t is $\mathscr{F}_t$-measurable and X_t is non-negative almost surely. Let $\alpha_t \in [0,1)$ and $\beta_t, \gamma_t \ge 0$ for every t. Let $\hat{w}_t$ be $\mathscr{F}_{t+1}$-measurable such that $\|\hat{w}_t \mid \mathscr{F}_t\|_{\psi_2} \le \tau$ for every t and $\mathrm{E}[\hat{w}_t \mid \mathscr{F}_t] = 0$. Let $\hat{y}_t$ be $\mathscr{F}_{t+1}$-measurable such that $\|\hat{y}_t \mid \mathscr{F}_t\|_{\psi_1} \le \rho$ for every t. Suppose $X_{t+1} \le \alpha_t X_t + \beta_t\hat{w}_t\sqrt{X_t} + \gamma_t\hat{y}_t$, for every t. Assume that $\mathrm{E}[\exp(\lambda X_1)] \le \exp(\lambda K)$, for $\lambda \in (0, 1/K]$. Then,

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

where $K = \max_{t=1}^{T} \left(\frac{2\gamma_t\rho}{1-\alpha_t}, \frac{4\beta_t^2\tau^2}{1-\alpha_t} \right)$.

Proof (of Theorem 6.35). We begin by deriving a recursive MGF bound on X_t. The proof of this recursive MGF bound differs from the proof of Claim 4.6 because in Claim 4.6 after conditioning on $\mathscr{F}_t$, we were left to bound the MGF of a subgaussian random variable. In this case however, due to the presence of $\hat{y}_t$, we will be left to bound the MGF of the sum of a subgaussian and subexponential random variable. For this reason, the MGF bound we obtain is valid in a smaller region than the MGF bound from Claim 4.6.

Claim 6.38. *Suppose* $\lambda \in (0, \min_{t=1}^{T} \left(\frac{1}{2\gamma_t\rho}, \frac{1-\alpha_t}{4\beta_t^2\tau^2} \right)]$. *Then, for every* t

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

Proof (of Claim 6.38). Because $\left\|\hat{y}_t \mid \mathscr{F}_t\right\|_{\psi_1} \leq \rho$, we have, by Claim 6.21:

$$\mathrm{E}\left[\exp\left(2\lambda\gamma_t\hat{y}_t\right) \mid \mathscr{F}_t\right] \;\leq\; \exp\left(2\lambda\gamma_t\rho\right) \qquad \text{for all } \lambda \in (0, \frac{1}{2\gamma_t\rho}]. \tag{6.2}$$

Furthermore, because $\left\|\hat{w}_t \mid \mathscr{F}_t\right\|_{\psi_2} \leq \tau$, $\mathrm{E}\left[\hat{w}_t \mid \mathscr{F}_t\right] = 0$ and $\sqrt{X_t}$ is $\mathscr{F}_t$-measurable, we have by Claim 6.17:

$$\mathrm{E}\left[\exp\left(2\lambda\beta_t\hat{w}_t\sqrt{X_t}\right) \mid \mathscr{F}_t\right] \;\leq\; \exp\left(4\lambda^2\beta_t^2\tau^2X_t\right) \qquad \text{for all } \lambda \in \mathbb{R}. \tag{6.3}$$

Now, we put these MGF bounds together (assuming $\lambda \in (0, \min_{t=1}^{T} \frac{1}{2\gamma_t\rho}]$:

$$\begin{aligned}
\mathrm{E}\left[\exp\left(\lambda X_{t+1}\right)\right] \;&\leq\; \mathrm{E}\left[\exp\left(\lambda\alpha_t X_t\right)\mathrm{E}\left[\exp\left(\lambda\beta_t\hat{w}_t\sqrt{X_t}\right)\exp\left(\lambda\gamma_t\hat{y}_t\right) \mid \mathscr{F}_t\right]\right] \\
&\leq\; \mathrm{E}\left[\exp\left(\lambda\alpha_t X_t\right)\mathrm{E}\left[\exp\left(2\lambda\beta_t\hat{w}_t\sqrt{X_t}\right) \mid \mathscr{F}_t\right]^{1/2}\mathrm{E}\left[\exp\left(2\lambda\gamma_t\hat{y}_t\right) \mid \mathscr{F}_t\right]^{1/2}\right] \\
&\leq\; \mathrm{E}\left[\exp\left(\lambda\alpha_t X_t\right)\exp\left(2\lambda^2\beta_t^2\tau^2X_t\right)\exp\left(\lambda\gamma_t\rho\right)\right] \qquad \text{(by Eq. (6.2) and Eq. (6.3))} \\
&=\; \exp\left(\lambda\gamma_t\rho\right)\mathrm{E}\left[\exp\left(\lambda X_t\left(\alpha_t + 2\lambda\beta_t^2\tau^2\right)\right)\right].
\end{aligned}$$

■

If we assume that $\lambda \in (0, \min_{t=1}^{T} \frac{1-\alpha_t}{4\beta_t^2\tau^2}]$, then we have

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

as desired. Next, we prove an MGF bound on X_t.

Claim 6.39. *For every* t*, and for every* $\lambda \in (0, 1/K]$*, we have* $\mathrm{E}\left[\exp\left(\lambda X_t\right)\right] \leq \exp\left(\lambda K\right)$. *That is,* $\left\|X_t\right\|_{\psi_1} \leq K/\ln 2$.

Proof. Let $\lambda \in (0, 1/K]$. We proceed by induction over t. The base case holds by assumption. Assume that

$\mathrm{E}[\exp(\lambda X_t)] \le \exp(\lambda K)$. Observe that λ satisfies the condition in Claim 6.38. Then,

$$\begin{aligned}
\mathrm{E}[\exp(\lambda X_{t+1})] &\le \exp(\lambda\gamma_t\rho)\,\mathrm{E}\left[\exp\left(\lambda X_t\left(\frac{1+\alpha_t}{2}\right)\right)\right] \qquad \text{(by Claim 6.38)}\\
&\le \exp(\lambda\gamma_t\rho)\exp\left(\lambda K\left(\frac{1+\alpha_t}{2}\right)\right) \qquad \text{(by induction hypothesis)}\\
&= \exp\left(\lambda\left(\gamma_t\rho + K\left(\frac{1+\alpha_t}{2}\right)\right)\right).
\end{aligned}$$

Hence, we need $K \ge K\left(\frac{1+\alpha_t}{2}\right) + \gamma_t\rho$. Indeed, by definition of K,

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

which yields the desired inequality. ■

Now, we can use Claim 6.39 to prove Theorem 6.35. The first claim in Theorem 6.35 follows from using Claim 6.39 and the MGF bound to tail bound transition given by Claim A.7.

Next, we prove the second claim from Theorem 6.35. Claim 6.39 gives that for every t and for all $\lambda \in (0, 1/(\sigma_t K)]$, we have $\mathrm{E}[\exp(\lambda\sigma_t X_t)] \le \exp(\lambda\sigma_t K)$. Hence, we can combine these MGF bounds using Lemma A.4 to obtain $\mathrm{E}\left[\exp\left(\lambda\sum_{t=1}^T \sigma_t X_t\right)\right] \le \exp\left(\lambda K\sum_{t=1}^T \sigma_t\right)$ for all $\lambda \in (0, \left(K\sum_{t=1}^T \sigma_t\right)^{-1}]$. 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 6.35. ■

6.3.5 Using Theorem 1.11 with conditionally subgaussian increments

Corollary 6.40. *Let $\{\mathscr{F}_t\}_{t=1}^T$ be a filtration and suppose that a_t are $\mathscr{F}_t$-measurable random variables and b_t are $\mathscr{F}_{t-1}$ measurable random variables. Further, suppose that*

1. *There exists $\kappa > 0$ such that $\|a_t \mid \mathscr{F}_{t-1}\|_{\psi_2} \le \kappa$ for every t,*
2. *$\mathrm{E}[a_t \mid \mathscr{F}_{t-1}] = 0$ for every t,*
3. *$\sum_{t=1}^T \|b_t\|^2 \le R\log(1/\delta)$ with probability at least $1 - O(\delta)$.*

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

Proof. The assumption that $\mathrm{E}[a_t \mid \mathscr{F}_{t-1}] = 0$ and $\|a_t \mid \mathscr{F}_{t-1}\|_{\psi_2} \le \kappa$ imply

$$\mathrm{E}[\exp(\lambda d_t) \mid \mathscr{F}_{t-1}] \le \exp\left(\lambda^2\kappa^2\|b_t\|^2\right),$$

by Claim 6.17 because $\|d_t \mid \mathscr{F}_{t-1}\|_{\psi_2} \le \kappa\|b_t\|$.

We may apply Lemma 4.3 with $d_t = \langle a_t, b_t \rangle$, $v_{t-1} = 2\kappa^2\|b_t\|^2$, $\alpha_t = 0$ for every i, and $R(\delta) = 2\kappa^2 R\log(1/\delta)$ to obtain

$$\Pr\left[\sum_{t=1}^T d_t \ge x\right] \le \delta + \exp\left(-\frac{x^2}{16\kappa^2 R\log(1/\delta)}.\right)$$

The final term above is at most δ if $x = 4\kappa\sqrt{R}\log(1/\delta)$. ■

Corollary 6.41. *Let $\{\mathscr{F}_t\}_{t=1}^T$ be a filtration and suppose that a_t are $\mathscr{F}_t$-measurable random variables and b_t are $\mathscr{F}_{t-1}$-measurable random variables. Define $d_t = \langle a_t, b_t \rangle$. Assume there exists $\kappa > 0$ such that $\|a_t \mid \mathscr{F}_{t-1}\|_{\psi_2} \le \kappa$*

and $\mathrm{E}[a_t \mid \mathscr{F}_{t-1}] = 0$. *Suppose that there exists* $R > 0$ *and non-negative values* $\{\alpha_t\}_{t=1}^{T-1}$ *where* $\kappa^2 \max\{\alpha_t\}_{t=1}^{T-1} = O\left(\kappa\sqrt{R}\right)$, *such that exactly one of the following holds for every* $\delta \in (0,1)$:

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

Then, $\sum_{t=1}^{T} d_t \leq O\left(\kappa\sqrt{R}\log(1/\delta)\right)$ *with probability at least* $1 - O(\delta)$.

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.

Proceeding as in the proof of Corollary 6.40, we have

$$\mathrm{E}[\exp(\lambda d_t) \mid \mathscr{F}_{t-1}] \;\leq\; \exp\left(\lambda^2\kappa^2 \|b_t\|^2\right).$$

We may now apply Lemma 4.3 with $d_t = \langle a_t, b_t \rangle$ and $v_{t-1} = 2\kappa^2 \|b_t\|^2$, with $\alpha_T = 0$ and $R(\delta) = R\log(1/\delta)$ to obtain:

$$\Pr\left[\sum_{t=1}^{T} d_t \geq x\right] \leq \delta + \exp\left(-\frac{x^2}{(8\kappa^2 \max_{t=1}^{T-1}(\alpha_t)\cdot x + 16\kappa^2 R)}\right).$$

Recalling that $\kappa^2 \max\{\alpha_t\}_{t=1}^{T-1} = O(\kappa\sqrt{R})$, we may set $x = \Theta\left(\kappa\sqrt{R}\log(1/\delta)\right)$ to bound the final term by δ.

■

Chapter 7

Conclusions and Future Work

In this thesis we identified some gaps in the theoretical understanding of the role that the averaging of iterates produced by SGD plays in the non-smooth setting. We fully characterized the performance of the final iterate in both the strongly-convex and non strongly-convex setting. That is, we have provided deterministic lower bounds matching the expected upper bounds of Shamir and Zhang [43], answering a COLT 2012 open question [42]. Moreover, we have extended the known expected convergence rates to hold with arbitrarily high probability.

Next, we identified a shortage of high-probability upper bounds in the strongly-convex setting of non-smooth SGD. The work in this thesis establishes two *optimal* and *tight* (even including $\log(1/\delta)$ factors) high-probability upper bounds, using suffix-averaging from [36] and non-uniform averaging from [26].

Along the way, we developed a new concentration inequality which extends the classical Freedman's inequality. This concentration inequality was a key probabilistic tool used on multiple occasions throughout this thesis, perhaps highlighting it's practical significance. The Generalized Freedman Inequality was used crucially to break the barrier of sub-optimal high probability results in the strongly-convex setting.

7.1 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. [20] 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 $\log(1/\delta)$ term, 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 $O\big((\log(T) + \sqrt{\log(1/\delta)})/\sqrt{T}\big)$.

Bibliography

[1] S. Arora, E. Hazan, and S. Kale. The multiplicative weights update method: A meta algorithm and its applications. *Theory of Computing*, 8(6):121–164, 2012. → page 3

[2] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-arm bandit problem. In *Proceedings of FOCS*, pages 322–331, 1995. → page 1

[3] J. P. Bailey and G. Piliouras. Multiplicative weights update in zero-sum games. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 321–338. ACM, 2018. → page 3

[4] N. Bansal and A. Gupta. Potential-function proofs for first-order methods. arXiv:1712.04581, 2017. → page 1

[5] V. Barbu and T. Precupanu. *Convexity and optimization in Banach spaces*. Springer Science & Business Media, 2012. → page 52

[6] P. L. Bartlett, V. Dani, T. Hayes, S. Kakade, A. Rakhlin, and A. Tewari. High-probability regret bounds for bandit online linear optimization. In *21th Annual Conference on Learning Theory (COLT 2008)*, July 2008. → page 9

[7] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3–4), 2015. → pages 1, 8

[8] P. Christiano, J. A. Kelner, A. Madry, D. A. Spielman, and S.-H. Teng. Electrical flows, laplacian systems, and faster approximation of maximum flow in undirected graphs. In *Proceedings of STOC*, pages 273–282. ACM, 2011. → page 3

[9] M. B. Cohen, Y. T. Lee, G. Miller, J. Pachocki, and A. Sidford. Geometric median in nearly linear time. In *Proceedings of STOC*, pages 9–21, 2016. → page 1

[10] V. de la Peña. A general class of exponential inequalities for martingales and ratios. *The Annals of Probability*, 27(1):537–564, 1999. → pages 5, 7

[11] X. Fan, I. Grama, and Q. Liu. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20, 2015. → pages 5, 7

[12] V. Feldman, I. Mironov, K. Talwar, and A. Thakurta. Privacy amplification by iteration. In *Proceedings of FOCS*, 2018. → page 1

[13] D. A. Freedman. On tail probabilities for martingales. *Annals of Probability*, 3(1):100–118, 1975. → pages 5, 6, 7

[14] Y. Freund, R. E. Schapire, et al. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999. → page 3

[15] N. J. Harvey, C. Liaw, Y. Plan, and S. Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613, 2019. → pages v, 38

[16] E. Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3–4), 2015. → page 1

[17] E. Hazan and S. Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014. → pages iii, 2, 8, 9

[18] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007. → page 2

[19] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*. Springer-Verlag, 1996. → pages 11, 14, 86

[20] P. Jain, D. Nagaraj, and P. Netrapalli. Making the last iterate of sgd information theoretically optimal. In *Conference on Learning Theory*, pages 1752–1755, 2019. → pages 2, 81

[21] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013. → page 1

[22] S. M. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithms. In *NIPS*, pages 801–808, 2008. → pages 2, 9

[23] J. A. Kelner, L. Orecchia, A. Sidford, and Z. A. Zhu. A simple, combinatorial algorithm for solving SDD systems in nearly-linear time. In *Proceedings of STOC*, 2013. → page 1

[24] P. Klein and N. 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. → page 62

[25] A. Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013. → page 46

[26] S. Lacoste-Julien, M. Schmidt, and F. Bach. A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012. → pages iii, 37, 38, 39, 81

[27] S. Lacoste-Julien, M. W. Schmidt, and F. R. Bach. A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method, Dec. 2012. arXiv:1212.2002. → page 2

[28] Y. T. Lee and A. Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. In *Proceedings of FOCS*, 2013. → page 1

[29] Y. T. Lee, S. Rao, and N. Srivastava. A new approach to computing maximum flows using electrical flows. In *Proceedings of STOC*, pages 755–764, 2013. → page 1

[30] P. Massart. *Concentration inequalities and model selection*. Springer, 2007. → page 40

[31] C. McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998. → page 5

[32] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. → pages 1, 3

[33] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley, 1983. → page 2

[34] S. A. Plotkin, D. B. Shmoys, and É. Tardos. Fast approximation algorithms for fractional packing and covering problems. *Mathematics of Operations Research*, 20(2):257–301, 1995. → page 3

[35] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. → page 3

[36] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of ICML*, 2012. → pages iii, 2, 3, 8, 9, 22, 23, 35, 38, 41, 42, 81

[37] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3): 400–407, Sept. 1951. → page 1

[38] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, 1987. → page 58

[39] D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988. → page 3

[40] M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017. → page 1

[41] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos: primal estimated sub-gradient solver for SVM. *Mathematical Programming*, 127(1):3–30, 2011. → page 2

[42] O. 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. → pages iii, 2, 3, 81

[43] O. Shamir and T. 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. → pages 2, 7, 8, 9, 20, 81

[44] T. Strohmer and R. Vershynin. A randomized solver for linear systems with exponential convergence. In *Proceedings of APPROX/RANDOM*, 2006. → page 1

[45] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018. → pages 67, 85

[46] N. Vishnoi. Algorithms for convex optimization, 2018. https://nisheethvishnoi.wordpress.com/convex-optimization/. → page 1

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

Proof. Follows by induction using Hölder's inequality. ■

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* $0 < \lambda \leq 1/K_i$. *Then* $\mathrm{E}[\exp(\lambda \sum_{i=1}^n X_i)] \leq \exp(\lambda \sum_{i=1}^n K_i)$ *for all* $0 < \lambda \leq 1/\sum_{i=1}^n K_i$.

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 ([45, Proposition 2.5.2])**.** *Suppose there is $c > 0$ such that for all* $0 < \lambda \leq \frac{1}{c}$, $\mathrm{E}\left[\exp\left(\lambda^2 X^2\right)\right] \leq \exp\left(\lambda^2 c^2\right)$. *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[1] $ab \leq a^2/2 + b^2/2$, valid for all $a, b \in \mathbb{R}$, to obtain

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

[1] Young's Inequality

■

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 $0 < \lambda \leq 1/C$. Then, $\Pr[X \geq C\log(1/\delta)] \leq ce\delta$.*

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 ([19, Eq. (3.1.6)])**.** *Let $\mathscr{X}$ be a convex set and $x \in \mathscr{X} \subseteq \mathbb{R}^n$. Then $\|\Pi_{\mathscr{X}}(y) - x\| \leq \|y - x\|$ for all $y \in \mathbb{R}^n$.*

Claim A.9 ([19, 4.2.1])**.** *Let $A : \mathbb{R}^n \to \mathbb{R}^m$ be a linear map and let g be a finite convex function on $\mathbb{R}^m$. Then $\partial(g \circ A)(x) = A^T \partial g(Ax)$ for all $x \in \mathbb{R}^m$.*

A.1 Useful scalar inequalities

Claim A.10. *For $1 \leq a \leq b$, $\sum_{k=a}^{b} \frac{1}{\sqrt{k}} \leq 2\frac{b-a+1}{\sqrt{b}}$.*

Proof.

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

■

Claim A.11. *For any $1 \leq j \leq t \leq T$, we have $\frac{t-j}{(T-j+1)\sqrt{t}} \leq \frac{1}{\sqrt{T}}$.*

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 $j \geq 0$, and so

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

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

Claim A.12. *Assume $0 \leq k$ and $k+1 \leq m$.*

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

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

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

■

Claim A.13. *Let* $\alpha_j = \frac{1}{(T-j)(T-j+1)}$. *Let* a, b *be such that* $a \leq b < T$. *Then,*

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

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.14. *Suppose* $a < b$. *Then,* $\log(b/a) \leq (b-a)/a$.

Claim A.15. *Let* $b \geq a > 1$. *Then,* $\sum_{i=a}^{b} \frac{1}{i} \leq \log\big(b/(a-1)\big)$.