NORTHEASTERN UNIVERSITY

DOCTORAL THESIS

Stochastic Optimization for Machine Learning: Stronger Convergence Guarantees and More Efficient Algorithms

---

Author:
Thien Hang NGUYEN

Supervisor:
Professor Huy Le NGUYEN

A thesis submitted in fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

Khoury College of Computer Sciences
Northeastern University
Boston, Massachusetts

March 12, 2025

Northeastern University
Khoury College of
Computer Sciences

PhD Thesis Approval

Thesis Title: Stochastic Optimization for Machine Learning: Stronger Convergence Guarantees and More Efficient Algorithms

Author: Thien Hang Nguyen

PhD Program: X Computer Science ____ Cybersecurity ____ Personal Health Informatics

PhD Thesis Approval to complete all degree requirements for the above PhD program.

Signed by: Huy Nguyen 313B45018FU2490...	3/14/2025
Thesis Advisor	*Date*
Signed by: Hongyang Zhang 06762527FDC1419...	3/13/2025
Thesis Reader	*Date*
Signed by: Paul Hand F782218FBFBE45A...	3/12/2025
Thesis Reader	*Date*
Signed by: Alina Ene AF2AF24E75734DF...	3/12/2025
Thesis Reader	*Date*
Thesis Reader	*Date*

KHOURY COLLEGE APPROVAL:

	3/17/25
Associate Dean for Graduate Programs	*Date*

COPY RECEIVED BY GRADUATE STUDENT SERVICES:

	18 March 2025
Recipient's Signature	*Date*

Distribution: Once completed, this form should be attached as page 2, immediately following the title page of the dissertation document. An electronic version of the document can then be uploaded to the Northeastern University-UMI Website.

NORTHEASTERN UNIVERSITY

Abstract

Stochastic Optimization for Machine Learning: Stronger Convergence Guarantees and More Efficient Algorithms

by Thien Hang NGUYEN

As deep neural networks scale to billions or even trillions of parameters, training costs have become a critical bottleneck. Stochastic optimizers such as Stochastic Gradient Descent (SGD), AdaGrad, and Adam form the algorithmic backbone of modern machine learning, making their theoretical understanding essential for efficient scaling and cost management. This thesis advances the foundations of stochastic optimization by establishing stronger convergence guarantees under relaxed assumptions for SGD and adaptive optimizers like AdaGrad. We then leverage these theoretical insights to design algorithms that are more memory efficient and more sample efficient for training modern large scale deep neural networks.

Specifically, we develop new general-purpose techniques to derive high probability convergence rates for a broad class of stochastic optimization algorithms – including SGD, Stochastic Mirror Descent (SMD), Accelerated SGD/SMD, AdaGrad-Norm, and AdaGrad-Coordinate – under sub-Gaussian gradient noise and other relaxed conditions like unbounded domains. Additionally, our techniques achieve optimal high-probability convergence rates for clipped gradient methods under heavy-tailed gradient noise.

Building on these theoretical insights, we introduce Subset-Norm (SN) and Subspace-Momentum (SM), two novel algorithms that compress the adaptive step-size and momentum states of optimizers like Adam. SN and SM improve both sample efficiency and memory efficiency in large-scale language model (LLM) training. Notably, combining SN and SM achieves Adam's validation perplexity for pre-training LLaMA-1B in approximately half the training tokens (6.8B vs. 13.1B) while reducing the optimizer-state memory footprint by over 80%.

Acknowledgements

This thesis would not have been possible without the guidance of my advisor, Huy Le Nguyen, to whom I owe my deepest gratitude. Huy's brilliance, wisdom, patience, and encouragement have been invaluable throughout this journey. I feel incredibly fortunate to have had my advisor's guidance over the past five years.

I am deeply grateful for the opportunity to grow and collaborate within the vibrant research communities at Northeastern University and Boston University, which have greatly enriched my graduate school experience. In particular, I sincerely appreciate my co-authors, Alina Ene and Duy Nguyen, for warmly hosting me at Boston University and for the many rewarding projects we have pursued together. I also extend heartfelt gratitude to my labmates – Anamay, Konstantina, Thy, Matthew, Lydia – and my colleagues – Fabian, Dat, Trung, Hai, Zijian, Themis, and Hieu – from whom I have learned and grown immensely. My sincere appreciation goes to Paul Hand, Alina Ene, Hongyang Zhang and my advisor for generously offering their time, support, and insights as members of my thesis committee. Finally, I must express my profound gratitude to Professor Vershynin, whose encouragement during my undergraduate years at UCI inspired me to pursue graduate studies in research. It has been an immense privilege to have been your student.

Friends are a great pleasure and fortune in life and I feel extremely blessed to have been surrounded by wonderful friends in Boston: Dat & Khanh, Hai & Van-Anh, Trung & Mai, Duy, Fabian, Ngu, Chau & Vy, Tuan-Anh & Nguyet-Anh, Minh, Alexa & William, Hardwood soccer team, coach Brandon, Vitor, Skyler, Kenny, and many other dear friends whose unwavering support and friendship have deeply enriched my time in Boston. Special shoutout to my friends in California and who have always been there throughout many years: Tony, Andy, Naoya, Yiming, Alex, Romie, Loc, Khoa, Takuya, Bi, Justin, Lucas, Taeyoung, Jungwoo, and many other dear friends whose unwavering support and friendship.

Finally, I wish to express my deepest gratitude to my parents and sister Thy for their unwavering support, guidance, and love, which have been my foundation and strength throughout my life. Above all, my heartfelt thanks go to my wife Thao, whose love, care, and support have given my life profound meaning and purpose. Thank you for being my greatest source of joy and inspiration.

Skip forward two hundred years into the Utopian future, and the scene is totally different. Hardly one of the things I have imagined will still be there. In that age when there is no manual labour and everyone is 'educated', it is hardly likely that Father will still be a rough man with enlarged hands who likes to sit in shirt-sleeves and says 'Ah wur coomin' oop street'. And there won't be a coal fire in the grate, only some kind of invisible heater. The furniture will be made of rubber, glass, and steel. If there are still such things as evening papers there will certainly be no racing news in them, for gambling will be meaningless in a world where there is no poverty and the horse will have vanished from the face of the earth. Dogs, too, will have been suppressed on grounds of hygiene. And there won't be so many children, either, if the birth-controllers have their way.

– George Orwell

Contents

List of Figures

List of Tables

To my parents, for your unending support and encouragement.
To my wife, for your love and patience.

Chapter 1

Introduction

Everyone in their intellectual life goes through a stage. It can happen in early graduate school, mid graduate school. It can also happen in later life, which is bad. It's not good to have it when you're an adult. Let me describe this stage of intellectual development. You read a couple of books and you wake up at 3:00 in the morning and say, "Oh my god, everything is an optimization problem." Actually, a lot of books start this way. My answer to that is – you have to go through this stage, so that's fine. But get over it quickly, please. Of course, everything is an optimization problem. What you'll find out quickly is it doesn't mean anything to say that. It says nothing.

—Stephen Boyd

1.1 Stochastic Optimization for Machine Learning: Then and Now

The central problem in machine learning can be broadly viewed as follows: given a dataset, design a model capable of accomplishing a "task" based on the information contained within the data. For instance, in image classification, the dataset comprises images of various categories such as cats, dogs, and cars, and the task involves assigning labels to objects within the images. Similarly, in language modeling, the dataset consists of a corpus of textual data, and the task is to predict the next word in a given sentence.

More formally, a task is represented via a loss function that quantifies the discrepancy between the model's predictions and a specified "ground truth." To achieve this, a parameterized function f_θ is selected. This is known as a *model* for the task and its parameters are fitted to the data to minimize the loss. This turns the learning problem into an optimization problem.

More concretely, let the data D be drawn *i.i.d.* from some underlying distribution $\mathcal{D}$ over a domain $\mathcal{X}$ that we want to estimate over. Given a loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that evaluates how well a prediction $f(x) \in \mathcal{Y}$ aligns with an input $x \in \mathcal{X}$, the goal is to design an algorithm $\mathcal{A}$ that outputs parameters θ for f_θ that minimizes the *population loss*:

$$L(f) := \mathbb{E}_{x \sim \mathcal{D}}[\ell(x, f(x))].$$

This process of designing and training machine learning models involves three deeply interconnected aspects: approximation, optimization, and generalization. Approximation pertains to choices such as model architecture, selection of loss functions, and the quality and scope of data collection. Optimization involves minimizing the loss function, often using iterative methods such as gradient descent and its variants. Generalization addresses the model's ability to perform well on unseen data, often involving techniques like regularization, the incorporation of additional

data, or ensuring stability in training. While this thesis focuses primarily on optimization, the three aspects deeply influence one another.

Remark 1. *The optimization formulation above only covers aspects like supervised learning and unsupervised learning in machine learning but not areas like reinforcement learning, where the expectation of the objective depends on the optimization parameters themselves i.e.* $L(f) := -\mathbb{E}_{\tau\sim f,P}\left[R(\tau)\right]$*, where* R *is some reward given by a trajectory of states and actions generated by our policy/model* f *and random environment* P*.*

1.1.1 Machine Learning Pre Scaling Laws

Prior to the advent of LLMs and scaling laws (Kaplan et al., 2020), much of machine learning was often constrained by limited data and computational resources, where reliant on supervised learning required more expensive labeled datasets. As a result, a key aspect during this period was managing the bias-variance tradeoff, where controlling model capacity via some form of regularization – such as weight-decay, dropout, early-stopping, etc. – was essential to avoid overfitting. Optimization methods typically operated over multiple epochs to maximize performance on limited datasets.

1.1.2 Machine Learning in the Scaling Laws Era: Big Models and Big Data

Recent advancements have redefined the landscape of machine learning. Advancements in self-supervised pretraining has dramatically expanded access to (relatively) inexpensive, large-scale datasets, removing many of the limitations imposed by reliance on labeled data, and ushering in the era of large-language models (LLMs) and large-vision models (LVMs) and many other large-X models.

Scaling laws, which suggest consistent improvements by scaling up model sizes and datasets simultaneously, have diminished concerns over model capacity. Increasing model size not only improves performance but also unlocks novel capabilities, fundamentally transforming what is achievable. Hence, this new era has transformed the algorithmic landscape for machine learning towards this era of ever larger models and more data. The computational cost of training and deploying such models has risen significantly, necessitating memory-efficient solutions and algorithms that operate in sublinear time and space. Addressing these challenges will require the development of innovative algorithms capable of scaling to unprecedented levels of complexity.

This thesis primarily focuses on optimization in the context of modern machine learning[1], where big models and big data demand new approaches to efficiency and scalability.

1.2 Contributions: From Theory to Practice

Modern machine learning with large models and datasets necessitates optimization methods that provide stronger guarantees due to the high cost of each training run. Consequently, theoretical interest in the convergence analysis of adaptive methods extends beyond asymptotic considerations. It now encompasses not only assumptions about the objective function (e.g., convexity, smoothness) and stochastic gradients (e.g., noise distribution), but also non-asymptotic dependencies on the total

[1]Our experiments focus on pre-training and supervised fine-tuning of LLMs.

TABLE 1.1: Summary of contributions in the **Light-Tailed Noise** setting. Note that T is the time horizon, δ is the failure probability, d is the model dimension, and σ is the sub-Gaussian noise parameter.

Setting	Method	Previous results	Our results
Convex	(Accelerated) Stochastic Mirror Descent	Bounded domain	*Unconstrained* domain
Non-convex	Stochastic Gradient Descent	$O\left(\log \frac{1}{\delta} \log T\right)$	$O\left(\log \frac{1}{\delta} + \log T\right)$
Non-convex	AdaGrad-Norm	Bounded stochastic grads	*Unbounded* gradients
Non-convex	AdaGrad-Coordinate	N/A	$\tilde{O}\left(\frac{d\sqrt{\sigma \log \frac{dT}{\delta}}}{\sqrt{T}}\right)$

TABLE 1.2: Summary of contributions in the **Heavy-Tailed Noise** setting where the gradient noise ξ_t is heavy-tailed i.e. $\mathbb{E}[\|\xi_t\|_*^p \mid x] \leq \sigma^p$ for $p \in (1,2]$. The bounds are for the optimal function gap $\frac{1}{T}\sum_{t=1}^T f(x_t) - f^*$ for convex functions and the average gradient norm $\frac{1}{T}\sum_{t=1}^T \|\nabla f(x_t)\|^2$ for non-convex functions.

	Assumptions	Convex (SMD)	Non-convex (SGD)
Previous results	Known T	$\widetilde{O}\left(T^{\frac{1-p}{p}}\right)$	$\widetilde{O}\left(T^{\frac{1-p}{p}}\right)$
Our results	Known T	$O\left(T^{\frac{1-p}{p}}\right)$	$O\left(T^{\frac{2-2p}{3p-2}}\right)$
	Unknown T (new)	$\widetilde{O}\left(T^{\frac{1-p}{p}}\right)$	$\widetilde{O}\left(T^{\frac{2-2p}{3p-2}}\right)$
Lower bound	$p \in (1,2]$	$\Omega\left(T^{\frac{1-p}{p}}\right)$	$\Omega\left(T^{\frac{2-2p}{3p-2}}\right)$

number of iterations, parameter count, and failure probability. Hence, this thesis focuses on not only obtaining *stronger convergence guarantee* – i.e. high probability – under *weaker assumptions* (removing assumptions such as bounded domain, bounded gradients, bounded noise, etc.) but also utilizes these theoretical insights to design more efficient algorithms.

1.2.1 Theory

In this thesis, we investigate the convergence properties of stochastic gradient descent (SGD) and adaptive optimization algorithms like AdaGrad under different noise models. While traditional analyses of stochastic gradient methods often provide convergence guarantees in expectation, such results typically fail to capture the behavior of algorithms in single-run scenarios. This is particularly important for modern applications, where high computational costs and hyperparameter-tuning demand more reliable performance on individual runs. High-probability analyses, in contrast, offer a better understanding of optimization dynamics and provide crucial insights for designing more robust algorithms.

Light-tailed noise. In the **light-tailed noise** setting, we propose a general framework to establish high-probability convergence guarantees for stochastic optimization methods under sub-Gaussian gradient noise. A summary of our contributions in this setting is provided in Table 1.1, where our techniques obtain stronger convergence guarantees under relaxed conditions compared to previous works.

Heavy-tailed noise. While the light-tailed noise assumption provides a natural framework for high-probability convergence, modern large-scale models, particularly transformers, often exhibit **heavy-tailed gradient noise** (Zhang et al., 2020). In this setting, the gradient noise has unbounded variance, making the analysis more complex since trading bias for reducing variance is often necessary. Traditional stochastic gradient descent (SGD) has been proven to fail under heavy-tailed gradient noise unless clipping is applied. This may explain why adaptive methods, which inherently normalize gradients by adjusting step sizes according to gradient norms, often outperform SGD in large-scale transformer models. This shift from light-tailed to heavy-tailed noise motivates the need for new high-probability convergence guarantees tailored to these more challenging conditions. We apply similar techniques developed in handling light-tailed gradient noise to tackle heavy-tailed noise as well, where we obtain *optimal* high-probability rates for clipped SGD in both convex and non-convex heavy-tailed noise settings, and the first results for versions with unknown time-horizon. A summary of our contributions in the heavy-tailed noise setting is provided in Table 1.2, where we compare our results against previous works and theoretical lower-bounds.

1.2.2 Practice

While theoretical insights are valuable, practical advancements are required to address the challenges posed by large-scale machine learning. As models and datasets grow in size, algorithms must be robust to varying noise models and scalable in terms of both computation and memory. Our research focuses on leveraging theoretical insights to develop algorithms that address these challenges. For instance, current adaptive optimizers like Adam, while effective, are memory-intensive and require storage proportional to twice the model size. By revisiting the theoretical foundations of adaptive methods, we propose new optimization algorithms that are not only faster but also significantly more memory-efficient. These contributions have the potential to reduce resource costs and enable training of even larger models in resource-constrained environments.

More specifically, we introduce two memory-efficient optimization algorithms for large-scale language model training: **Subset-Norm** (**SN**) for adaptive step-size memory reduction and **Subspace-Momentum** (**SM**) for momentum compression. While existing approaches trade performance for memory savings, our theoretically-grounded methods achieve both a reduced memory footprint and improved convergence. These methods are direct generalization of existing methods. Concretely, Subset-Norm adaptive step size generalizes AdaGrad-Norm and AdaGrad-Coordinate, and Subspace-Momentum (SM) generalizes SGD with Momentum. These methods build on top of our theoretical analysis and strong empirical results demonstrate their practical effectiveness in real-world large scale tasks.

1.3 Dissertation Overview

This thesis investigates the challenges and advancements in stochastic optimization for machine learning, focusing on both theoretical and practical contributions.

Chapter 2 establishes the theoretical framework of thesis. It formalizes the problem of stochastic optimization for machine learning and articulates the assumptions required for theoretical analysis.

Part I focuses on theoretical advancements, presenting key results and theorems under various settings. It highlights novel techniques for relaxing assumptions in prior work, offering insights that improve the rigor and applicability of existing theories. There, Chapter 3 provides an overview and literature review of the area, as well as provide important context for the contributions of our technical framework and results. Then, Chapter 4 presents high-probability convergence results under relaxed assumptions (unbounded domain and light-tailed gradient noise) for (Accelerated) Stochastic Mirror Descent on convex objectives and SGD, AdaGrad-Norm, and AdaGrad on non-convex objectives. Chapter 5 transitions to the more challenging setting of heavy-tailed noise. There, we provide optimal convergence rates for clipped methods in both convex and non-convex regimes. These theoretical insights serve as foundations for the development of principled practical algorithms.

Part II transitions to practical advancements, where Chapter 6 introduces the issues of existing optimizers and the problems we are trying to solve. Then we introduce our two novel algorithms, Subset Norm in Chapter 7 and Subspace Momentum in Chapter 8, for reducing memory while maintaining strong theoretical guarantees under relaxed assumptions. Then we demonstrate the practical effectiveness of our methods through a series of extensive empirical validations on a wide range of real-world LLMs training tasks in Chapter 9. The results showcase the algorithms speedup, reduced memory requirements, and robustness to gradient noise, validating the proposed theoretical benefits.

Chapter 2

Problem Statement and Notations

2.1 Problem Statement

We consider the problem $\min_{x\in\mathcal{X}} f(x)$ where $f : \mathbb{R}^d \to \mathbb{R}$ is the objective function and $\mathcal{X}$ is the domain of the problem. In the convex case, we consider the general setting where f is potentially not strongly convex and the domain $\mathcal{X}$ is convex but not necessarily compact. The distance between solutions in $\mathcal{X}$ is measured by a general norm $\|\cdot\|$. Let $\|\cdot\|_*$ denote the dual norm of $\|\cdot\|$. In the non-convex case, we consider the setting where $\mathcal{X}$ is $\mathbb{R}^d$ and $\|\cdot\|$ is the ℓ_2 norm.

2.1.1 Goals

In the **convex** case, the goal is to find iterates $x_t \in \mathcal{X}$ that approaches the global optimal solution i.e. $f(x_t) \to \min f(x)$ as $t \to \infty$. In the **non-convex** case, finding a global solution is NP-hard. Hence, we find solutions that approach an approximate stationary point i.e. $\|\nabla f(x_t)\| \to 0$ as $t \to \infty$.

We will present our probabilistic results in terms of the number of iterations T required in order for the optimal gap $f(x_T) - f^*$ where $f^* := \inf_{x\in\mathcal{X}} f(x)$ or gradient norm $\|\nabla f(x_T)\|$ to be within some small error range $\epsilon \in (0, 1)$ with probability at least $1 - \delta$ for some (small) failure probability $\delta \in (0, 1)$. As alluded to in Table 1.2, our results will be presented in the form of an average case result (as is common in stochastic optimization results): $\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f^*$ for convex functions and the average gradient norm $\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\|^2$ for non-convex functions. The next Section expands on how one can interpret these results in more practical terms.

2.1.2 Interpreting Average Results

In stochastic optimization, average-case convergence results—such as $\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f^*$ for convex objectives and $\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\|^2$ for non-convex objectives—guide practical algorithm design. For convex functions, if $\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f^* \leq \epsilon$, then by definition of the average, there exists some t such that $f(x_t) - f^* \leq \epsilon$ (since $\min_t(f(x_t) - f^*) \leq \frac{1}{T}\sum_{t=1}^{T} f(x_t) - f^*$), justifying the selection of the *best iterate* via validation (e.g., on a holdout set), while the *average iterate* $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$ often converges to the optimum due to Jensen's inequality: $f(\bar{x}_T) \leq \frac{1}{T}\sum_{t=1}^{T} f(x_t) \leq f^* + \epsilon$. For non-convex functions, a small average gradient norm $\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \leq \epsilon$ implies $\min_t \|\nabla f(x_t)\|^2 \leq \epsilon$ (same averaging argument), enabling selection of the best iterate with a small gradient via validation, though averaging iterates is less common due to non-convexity. These strategies transform theoretical averages into actionable solutions for practical algorithms.

2.2 Notations

We let v_i denote the i-th coordinate of a vector $v \in \mathbb{R}^d$. If a vector x_t is already indexed as part of a sequence of vectors (where x_t denotes the t-th update) then we use $x_{t,i}$ to denote x_t's i-th coordinate and $x_{t,\Psi} \in \mathbb{R}^k$ to denote the indexing with respect to an ordered subset $\Psi \subseteq [d]$ of size k where $(x_{t,\Psi})_k = x_{t,\Psi^{(k)}}$ with $\Psi^{(k)}$ denoting the k-th element of Ψ. For gradients, we let $\nabla_i f(x) := \frac{\partial f}{\partial x_i}$ denote the partial derivative with respect to the i-th coordinate. Similarly, for stochastic gradients $\widehat{\nabla} f(x)$, we let $\widehat{\nabla}_i f(x)$ denotes its i-th coordinate. If $a, b \in \mathbb{R}^d$, then ab and a/b denotes coordinate-wise multiplication and division, respectively i.e. $(ab)_i = a_i b_i$ and $(a/b)_i = a_i/b_i$.

2.3 Assumptions

We use the following standard assumptions:

(1) Existence of a minimizer: In the convex setting, we assume that there exists $x^* = \arg\min_{x \in \mathcal{X}} f(x)$.

(1′) Finite lowerbound: In the nonconvex setting, we assume that f admits a finite lower bound $\inf_{x \in \mathcal{X}} f(x) := f_* > -\infty$.

(2) Unbiased estimator: We assume to have access to a history independent, non-biased gradient estimator $\widehat{\nabla} f(x)$ for any $x \in \mathcal{X}$, that is $\mathbb{E}\left[\widehat{\nabla} f(x) \mid x\right] = \nabla f(x)$.

Light-tailed noise i.e. sub-Gaussian noise. There are several equivalent definitions of sub-Gaussian random variables up to an absolute constant scaling (see, e.g., Proposition 2.5.2 in (Vershynin, 2018)). For convenience, we use the following property as the definition.

Definition 2.3.1. A random variable X is σ-sub-Gaussian if

$$\mathbb{E}\left[\exp\left(\lambda^2 X^2\right)\right] \leq \exp\left(\lambda^2 \sigma^2\right) \text{ for all } \lambda \text{ such that } |\lambda| \leq \frac{1}{\sigma}.$$

Coordinate-wise sub-gaussian noise. If we denote the stochastic gradient noise as $\xi_t := \widehat{\nabla} f(x_t) - \nabla f(x_t)$ and $\xi_{t,i}$ as the i-th coordinate of ξ_t, then we assume the noise is per-coordinate subgaussian i.e. there exists $\sigma_i > 0$ for $i \in [d]$ such that ξ_t satisfies

$$\mathbb{E}\left[\exp\left(\lambda^2 \xi_{t,i}^2\right)\right] \leq \exp\left(\lambda^2 \sigma_i^2\right), \forall\, |\lambda| \leq \frac{1}{\sigma_i}, \forall i \in [d]\,. \tag{2.1}$$

Note that $\|\xi_t\|$ being σ-subgaussian implies that each $\xi_{t,i}$ is also σ-subgaussian, so coordinate-wise sub-gaussian is more general than standard scalar sub-gaussian noise assumption. Furthermore, when $\|\cdot\|$ is used without explicitly specifying the norm, we assume it is the ℓ_2 norm $\|\cdot\|_2$. We also use 0-indexing convention i.e. $[n] := \{0, 1, \ldots, n-1\}$ for integer $n \in \mathbb{N}$.

Heavy-tailed noise. There exists $\sigma > 0$ such that for some $1 < p \leq 2$, $\mathbb{E}[\|\widehat{\nabla} f(x) - \nabla f(x)\|_*^p \mid x] \leq \sigma^p$.

Part I

Theory

Chapter 3

Introduction

It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong.

– Richard Feynman

3.1 Introduction

Stochastic optimization is a fundamental area with extensive applications in many domains, ranging from machine learning to algorithm design and beyond. The design and analysis of iterative methods for stochastic optimization has been the focus of a long line of work, leading to a rich understanding of the convergence of paradigmatic iterative methods such as stochastic gradient descent, mirror descent, and accelerated methods for both convex and non-convex optimization. However, most existing works focus on establishing convergence guarantees that hold only *in expectation*. Although meaningful, these results do not fully capture the convergence behaviors of the algorithms on a small number of runs, as typical in modern ML applications where there are significant costs associated with performing multiple runs of the algorithm (Harvey et al., 2019; Madden et al., 2020; Davis et al., 2021).

Compared to the guarantees that hold in expectation, high probability guarantees are harder to obtain and hold in more limited settings. They often require stronger assumptions on the problem settings and the noise distribution. Existing high-probability results focus on the setting where the magnitude of the stochastic noise follows a light-tail (sub-Gaussian) distribution (Juditsky et al., 2011; Lan, 2012; Lan, 2020; Li and Orabona, 2020; Madden et al., 2020; Kavis et al., 2021). Recent works also study the more challenging heavy-tail setting, notably under a bounded variance (Nazin et al., 2019; Gorbunov et al., 2020; Cutkosky and Mehta, 2021) or bounded p-moment assumption (Cutkosky and Mehta, 2021) on the norm of the stochastic noise. Both settings are highly relevant in practice. For instance, Zhang et al. (2020) empirically studied the noise distribution for two common tasks, training a ResNet model for computer vision and a BERT transformer model for natural language processing. The authors observe that the noise distribution in the former task is well-approximated by a sub-Gaussian distribution, while the latter task appears to be heavy-tailed.

3.1.1 Challenges in the Light-Tailed Setting

Despite these progress, the convergence of cornerstone methods is not fully understood even in the more structured light-tailed noise setting. Specifically, the existing works for both convex and non-convex optimization rely on strong assumptions on the optimization domain and the gradients:

The problem domain is restricted to either the unconstrained domain or a constrained domain with bounded Bregman diameter. The convergence guarantees established depend on the Bregman diameter of the domain instead of the initial distance to the optimum. Even for compact domains, since the diameter can be much larger than the initial distance, these guarantees are pessimistic and diminish the benefits of good initializations. Thus an important direction remains to establish high probability guarantees for general optimization that scale only with the initial Bregman distance.

The gradients or stochastic gradients are assumed to be bounded even in the smooth setting. These additional assumptions are very restrictive and they significantly limit the applicability of the algorithm, e.g., they do not apply to important settings such as quadratic optimization. Moreover, the stochastic gradient assumption is more restrictive than other commonly studied assumptions, such as the gradients and the stochastic noise being bounded almost surely.

The above assumptions are not merely an artifact of the analysis, and they stem from important considerations and technical challenges. The high probability convergence guarantees are established via martingale concentration inequalities that impose necessary conditions on how much the martingale sequence can change in each step. However, the natural martingale sequences that arise in optimization depend on quantities such as the distance between the iterates and the optimum and the stochastic gradients, which are not a priori bounded. The aforementioned assumptions ensure that the concentration inequalities can be readily applied due to the relevant stochastic terms being all bounded almost surely. These difficulties are even more pronounced for adaptive algorithms in the AdaGrad family that set the step sizes based on the stochastic gradients. The adaptive step sizes introduce correlations between the step sizes and the update directions, and a crucial component is the analysis of the evolution of the adaptive step sizes and the cumulative stochastic noise. If the gradients are bounded, both of these challenges can be overcome by paying error terms proportional to the lengths of the gradients and stochastic gradients. Removing the bounded gradient assumptions requires new technical insights and tools.

In addition to requiring stronger assumptions, due to the technical challenges involved, several of the prior works are only able to establish convergence guarantees that are slower than the ideal sub-Gaussian rates. For example, a common approach is to control the relevant stochastic quantities across all T iterations of the algorithm via repeated applications of the concentration inequalities, leading to convergence rates that have additional factors that are poly-logarithmic in T. Additionally, achieving noise-adaptive rates that improve towards the deterministic rate as the amount of noise decreases is very challenging with existing techniques.

3.1.2 Challenges in the Heavy-Tailed Setting

In the heavy-tailed setting, recent works (Cutkosky and Mehta, 2021; Sadiev et al., 2023; Liu et al., 2023d) show that variants of Clipped-SGD in fact converge with high probability. This is a pleasing result, extending the earlier work by (Gorbunov et al., 2020) for $p = 2$. However, there are several shortcomings of these results when compared with the corresponding bounds in the light-tailed setting. First, the clipped algorithm uses a fixed step size and a fixed clipping parameter depending on the number of iterations, which precludes results with *unknown* time horizons. Second, the convergence guarantees are worse than the light-tailed bounds by a $\log T$ factor, even for fixed step sizes and clipping parameters.

3.2 Contributions

In the **light-tailed noise** setting (Liu et al., 2023c), we develop a general framework to establish high-probability convergence guarantees for stochastic optimization methods under sub-Gaussian gradient noise. For convex objectives, our results extend to stochastic mirror descent (SMD) and stochastic accelerated mirror descent, achieving rates that depend only on the Bregman distance between the initial point and the optimum (Juditsky et al., 2011; Lan, 2012; Lan, 2020). In the non-convex setting, we improve the time horizon and success probability dependencies for stochastic gradient descent (SGD) compared to prior works (Madden et al., 2020; Li and Orabona, 2020), and extend high-probability guarantees to AdaGrad-Norm (Ward et al., 2019), eliminating restrictive gradient assumptions made in earlier studies (Kavis et al., 2021). Furthermore, our analysis provides the first high-probability convergence results for standard AdaGrad (Duchi et al., 2011), broadening its theoretical guarantees.

In the **heavy-tailed noise** setting (Nguyen et al., 2023a), we analyze clipped gradient methods and demonstrate time-optimal high-probability convergence rates across convex and non-convex objectives, addressing key limitations in prior works. For convex optimization, we establish rates for clipped-SMD and clipped accelerated SMD that match lower bounds (Raginsky and Rakhlin, 2009; Vural et al., 2022), even for unbounded domains. In the non-convex setting, clipped-SGD achieves the optimal rate (Zhang et al., 2020), complementing in-expectation results. Notably, our approach removes dependency on the time horizon T and allows for unknown problem parameters such as the noise parameter σ, failure probability δ, and initial distance to the optimum, which were restrictive in prior analyses (Freedman, 1975; Dzhaparidze and Van Zanten, 2001). These results extend clipped gradient techniques to stochastic mirror descent and stochastic accelerated mirror descent, accommodating arbitrary norms and domains.

3.3 Main Techniques

Compared to prior works that rely on black-box applications of martingale concentration inequalities such as Freedman's inequality and its extensions (Freedman, 1975; Harvey et al., 2019; Madden et al., 2020), in this work we introduce a "white-box" concentration argument that leverages existing convergence analyses for first-order methods. The high-level approach is to define a novel martingale sequence derived from the standard convergence analyses and analyze its moment generating function from first principles. By leveraging the structure of the optimization problem, we are able to overcome a key difficulty associated with black-box applications of martingale concentration results: these results pose necessary conditions on how much the martingale sequence can change, which do not a priori hold for the natural martingales that arise in optimization. By seamlessly combining the optimization and probability toolkits, we obtain a flexible analysis template that allows us to handle general optimization domains with very large or even unbounded diameter, general objectives that are not globally Lipschitz, and adaptive step sizes.

Our technique is inspired by classical works in concentration inequalities, specifically a type of martingale inequalities where the variance of the martingale difference is bounded by a linear function of the previous value. This technique is first applied by Harvey et al. (2019) to show high probability convergence for SGD in the strongly convex setting. Our proof is inspired by the proof of Theorem 7.3 by Chung

and Lu (2006). In each time step with iterate x_t, let $\xi_t := \widehat{\nabla} f(x_t) - \nabla f(x_t)$ be the stochastic error in our gradient estimate. Classical proofs of convergence evolve around analyzing the sum of $\langle \xi_t, x^* - x_t \rangle$, which can be viewed as a martingale sequence. Assuming a bounded domain, the concentration of the sum can be shown via classical martingale inequalities. The key new insight is that instead of analyzing this sum, we analyze a related sum where the coefficients decrease over time to account for the fact that we have a looser grip on the distance to the optimal solution as time increases. Nonetheless, the coefficients are kept within a constant factor of each others and the same asymptotic convergence is attained with high probability.

3.4 Related Works

Convex optimization: Nemirovski et al. (2009) and Lan (2012) establish high probability bounds for stochastic mirror descent and accelerated stochastic mirror descent with sub-Gaussian noise. The rates shown in these works match the best rates known in expectation, but they depend on the Bregman diameter $\max_{x,y \in \mathcal{X}} \mathbf{D}_\psi(x, y)$ of the domain, which can be very large or even unbounded. Our work complements the analysis with a novel concentration argument that allows us to establish convergence with respect to the distance $\mathbf{D}_\psi(x^*, x_1)$ from the initial point to the optimum. Our analysis applies to the general setting considered in (Lan, 2020) and we use the same sub-Gaussian assumption on the noise.

Nazin et al. (2019) and Gorbunov et al. (2020) consider the more general setting of bounded variance noise. However, their problem settings are more restricted than ours. Specifically, Nazin et al. (2019) analyze stochastic mirror descent only in the setting where the optimization domain has bounded Bregman diameter. Gorbunov et al. (2020) analyze modifications of stochastic gradient descent and accelerated stochastic gradient descent, but only for unconstrained optimization with the ℓ_2 setup. In contrast, our work addresses the sub-Gaussian noise setting but it applies to general optimization, and we analyze the classical stochastic mirror descent and accelerated mirror descent without any modifications and with general Bregman distances and optimization domains.

The algorithm of Davis et al. (2021) is restricted to well-conditioned objectives that are both smooth and strongly convex, and do not apply to general convex optimization. Additionally, compared to classical methods such as SGD and stochastic mirror descent, the proposed algorithm solves an auxiliary optimization problem in each iteration and is thus more computationally expensive. The high-probability convergence of SGD is studied in the works (Kakade and Tewari, 2008; Rakhlin et al., 2011; Hazan and Kale, 2014; Harvey et al., 2019; Dvurechensky and Gasnikov, 2016). These works either assume that the function is strongly convex or the domain has bounded diameter. In contrast, our work applies to non-strongly convex optimization with a general domain.

Non-convex optimization: Li and Orabona (2020) demonstrate a high probability bound for an SGD algorithm with momentum while Madden et al. (2020) and Li and Liu (2022) show for the vanilla SGD and generalize to the family of sub-Weibull noise. However, the existing bounds are not optimal, which we improve in our work, using a very different approach. Convergence in high probability of algorithms with adaptive step size for non-convex problems has also been studied, for example, by Li and Orabona (2020) and Kavis et al. (2021). We note that the algorithm of Li and Orabona (2020) is not fully adaptive due to the dependence of the

initial step size on the problem parameters, whereas in Kavis et al. (2021) the gradients or stochastic gradients are required to be uniformly bounded almost surely. By contrast, using new techniques, we are able to establish convergence in high probability of the vanilla version of AdaGrad-Norm (Ward et al., 2019; Faw et al., 2022) without any of these additional assumptions. A key distinction from prior work is the analysis does not involve the division by the step size. This allows us to directly extend the analysis to the general AdaGrad Duchi et al. (2011), which is not possible previously. We provide a more detailed comparison with prior work in the subsequent sections.

High probability convergence for noises with bounded variance and heavy tails. The design of new gradient algorithms and their analysis in the presence of heavy-tailed noises has drawn significant recent interest. Starting from the work (Pascanu et al., 2012) which propose Clipped-SGD to handle exploding gradients in recurrent neural networks, the recent works (Simsekli et al., 2019; Şimşekli et al., 2019; Zhang et al., 2020; Gurbuzbalaban et al., 2021) give new motivation for clipped methods in the context of convolutional networks and attention deep networks that attempts to explain the dominance of adaptive methods over SGD in practical modern scenarios.

While the convergence in expectation of vanilla SGD has been extensively studied (Ghadimi and Lan, 2013; Nemirovski et al., 2009; Khaled and Richtárik, 2020; Liu et al., 2023c), only recently has the convergence of Clipped-SGD with heavy tailed noises been closely examined. There, (Zhang et al., 2020) first show the convergence in expectation of Clipped-SGD for nonconvex functions and provide a matching lower bound. In the convex regime, several works with different clipping strategies for the case of $p = 2$ have shown high probability convergence for smooth problems with bounded domain (Nazin et al., 2019; Parletta et al., 2022), smooth unconstrained problems (Gorbunov et al., 2020), and non-smooth problems (Gorbunov et al., 2021). A variant of Clipped-SGD that utilizes momentum (Cutkosky and Mehta, 2021) has also been shown to converge with high probability for bounded pth moments gradient noise. However, the analysis in (Cutkosky and Mehta, 2021) requires a strong assumption which implies that the true gradients are bounded, a restrictive assumption that excludes objectives like quadratic functions.

More recently, (Sadiev et al., 2023; Liu et al., 2023d; Zhang and Cutkosky, 2022) give nearly-optimal convergence rates for several Clipped-SGD variants. These works follow the recipe of using Freedman-type inequalities (Freedman, 1975; Dzhaparidze and Van Zanten, 2001) as a blackbox and bound the iterates inductively for all iterations, which incur an additional $\log T$ dependency in the final convergence rate. We show in our work that existing convergence rates can be tightened up and improved. Tight lower bounds for the optimal convergence rate have been shown by (Raginsky and Rakhlin, 2009; Vural et al., 2022) for convex objectives and by (Zhang et al., 2020) for nonconvex settings. In both cases, our paper provides optimal convergence guarantees.

In a related but different line of work, (Wang et al., 2021) show that vanilla SGD can converge with heavy tailed noise for a special type of strongly convex functions, and (Vural et al., 2022) show that stochastic mirror descent converges in expectation for a special choice of mirror maps, although only for strongly convex objectives with bounded domains.

Chapter 4

Light-Tailed Noise: (Accelerated) SMD, SGD, and AdaGrad

4.1 Convex Case: Stochastic Mirror Descent (SMD) and Accelerated SMD

In this section, we analyze the Stochastic Mirror Descent algorithm (Algorithm 1) and Accelerated Stochastic Mirror Descent algorithm (Algorithm 2) for convex optimization. We define the Bregman divergence $\mathbf{D}_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y\rangle$ where $\psi : \mathbb{R}^d \to \mathbb{R}$ is an 1-strongly convex mirror map with respect to $\|\cdot\|$ on $\mathcal{X}$. We remark that the domain of ψ is defined as $\mathbb{R}^d$ for simplicity, though which is not necessary.

4.1.1 Analysis of Stochastic Mirror Descent

Algorithm 1 Stochastic Mirror Descent Algorithm

Parameters: initial point $x_1 \in \mathcal{X}$, step sizes $\{\eta_t\}$, strongly convex mirror map ψ

for $t = 1$ to T:

$x_{t+1} = \arg\min_{x\in\mathcal{X}} \left\{ \eta_t \left\langle \widehat{\nabla} f(x_t), x \right\rangle + \mathbf{D}_\psi(x, x_t) \right\}$

return $\frac{1}{T}\sum_{t=1}^{T} x_t$

The end result of this section is the convergence guarantee of Algorithm 1 for constant step sizes (when the time horizon T is known) and time-varying step sizes (when T is unknown) presented in Theorem 4.1.1. However, we will emphasize more on presenting the core idea of our approach, which will serve as the basis for the analysis in subsequent sections. For simplicity, here we consider the non-smooth setting, and assume that f is G-Lipschitz continuous, i.e., we have $\|\nabla f(x)\|_* \le G$ for all $x \in \mathcal{X}$. However, this is not necessary. The analysis for the smooth setting follows via a simple modification to the analysis presented here as well as the analysis for the accelerated setting given in the next section.

Theorem 4.1.1. *Assume f is G-Lipschitz continuous, a minimizer x^* exists (assumption 1), access to unbiased gradient estimators (assumption 2), and σ-sub-gaussian gradient noise. Then, with probability at least $1 - \delta$, the iterate sequence $(x_t)_{t\ge1}$ output by Algorithm 1 satisfies*

(1) Setting $\eta_t = \sqrt{\frac{\mathbf{D}_\psi(x^*,x_1)}{6\left(G^2+\sigma^2\left(1+\log\left(\frac{1}{\delta}\right)\right)\right)T}}$, *then* $\mathbf{D}_\psi\left(x^*, x_{T+1}\right) \leq 4\mathbf{D}_\psi\left(x^*, x_1\right)$, *and*

$$\frac{1}{T}\sum_{t=1}^{T}\left(f\left(x_t\right)-f\left(x^*\right)\right) \leq \frac{4\sqrt{6}}{\sqrt{T}}\sqrt{\mathbf{D}_\psi\left(x^*, x_1\right)\left(G^2+\sigma^2\left(1+\log\left(\frac{1}{\delta}\right)\right)\right)}.$$

(2) Setting $\eta_t = \sqrt{\frac{\mathbf{D}_\psi(x^*,x_1)}{6\left(G^2+\sigma^2\left(1+\log\left(\frac{1}{\delta}\right)\right)\right)t}}$, *then* $\mathbf{D}_\psi\left(x^*, x_{T+1}\right) \leq 2(2+\log T)\mathbf{D}_\psi\left(x^*, x_1\right)$, *and*

$$\frac{1}{T}\sum_{t=1}^{T}\left(f\left(x_t\right)-f\left(x^*\right)\right) \leq \frac{2\sqrt{6}}{\sqrt{T}}(2+\log T)\sqrt{\mathbf{D}_\psi\left(x^*, x_1\right)\left(G^2+\sigma^2\left(1+\log\left(\frac{1}{\delta}\right)\right)\right)}.$$

We define $\xi_t := \widehat{\nabla} f\left(x_t\right) - \nabla f\left(x_t\right)$ and let $\mathcal{F}_t = \sigma\left(\xi_1, \ldots, \xi_{t-1}\right)$ denote the natural filtration. Note that x_t is $\mathcal{F}_t$-measurable. The starting point of our analysis is the following inequality that follows from the standard stochastic mirror descent analysis (see, e.g., Lan (2020)). We include the proof in Section 4.4 for completeness.

Lemma 4.1.2. *Lan (2020) For every iteration t, we have*

$$\begin{aligned} A_t &:= \eta_t\left(f\left(x_t\right)-f\left(x^*\right)\right) - \eta_t^2 G^2 + \mathbf{D}_\psi\left(x^*, x_{t+1}\right) - \mathbf{D}_\psi\left(x^*, x_t\right) \\ &\leq \eta_t\left\langle \xi_t, x^* - x_t\right\rangle + \eta_t^2\left\|\xi_t\right\|_*^2. \end{aligned}$$

We now turn our attention to our main concentration argument. Towards our goal of obtaining a high-probability convergence rate, we analyze the moment generating function for a random variable that is closely related to the left-hand side of the inequality above. We let $\{w_t\}$ be a sequence where $w_t \geq 0$ for all t. We define

$$\begin{aligned} Z_t &= w_t A_t - v_t \mathbf{D}_\psi\left(x^*, x_t\right), && \forall\, 1 \leq t \leq T \\ \text{where } v_t &= 6\sigma^2\eta_t^2 w_t^2 \\ \text{and } S_t &= \sum_{i=t}^{T} Z_i, && \forall\, 1 \leq t \leq T+1 \end{aligned}$$

Before proceeding with the analysis, we provide intuition for our approach. If we consider S_1, we see that it combines the gains in function value gaps with weights given by the sequence $\{w_t\}$ and the losses given by the Bregman divergence terms $\mathbf{D}_\psi\left(x^*, x_t\right)$ with coefficients v_t chosen based on the step size η_t and w_t. The intuition here is that we want to transfer the error from the stochastic error terms on the RHS of Lemma 4.1.2 into the loss term $v_t\mathbf{D}_\psi\left(x^*, x_t\right)$ then leverage the progression of the Bregman divergence $\mathbf{D}_\psi\left(x^*, x_{t+1}\right) - \mathbf{D}_\psi\left(x^*, x_t\right)$ to absorb this loss. For the first step, we can do that by setting the coefficient v_t to equalize coefficient of divergence term that will appear from the RHS of Lemma 4.1.2. For the second step, we can aim at making all the divergence terms telescope, by selecting v_t and w_t such that $w_t + v_t \leq w_{t-1}$ to have a telescoping sum of the terms $w_t\mathbf{D}_\psi\left(x^*, x_{t+1}\right) - w_{t-1}\mathbf{D}_\psi\left(x^*, x_t\right)$. In the end we will obtain a bound for the function value gaps in terms of only the deterministic quantities, namely η_t, w_t, G and the initial distance. In Theorem 4.1.3, we upper bound the moment generating function of S_1 and derive a set of conditions for the weights $\{w_t\}$ that allow us to absorb the stochastic errors. In Corollary 4.1.4, we show how to choose the weights $\{w_t\}$ and obtain a convergence rate that matches the standard rates that hold in expectation.

We now give our main concentration argument that bounds the moment generating function of S_t inspired by the proof of Theorem 7.3 in Chung and Lu (2006).

Theorem 4.1.3. *Suppose that $w_t\eta_t^2 \leq \frac{1}{4\sigma^2}$ for every $1 \leq t \leq T$. For every $1 \leq t \leq T+1$, we have*

$$\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_t\right] \leq \exp\left(3\sigma^2 \sum_{i=t}^{T} w_i\eta_i^2\right).$$

Proof. We proceed by induction on t. Consider the base case $t = T+1$. We have the inequality holds true trivially. Next, we consider $1 \leq t \leq T$. We have

$$\begin{aligned}\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_t\right] &= \mathbb{E}\left[\exp\left(Z_t + S_{t+1}\right) \mid \mathcal{F}_t\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(Z_t + S_{t+1}\right) \mid \mathcal{F}_{t+1}\right] \mid \mathcal{F}_t\right]. \end{aligned} \tag{4.1}$$

We now analyze the inner expectation. Conditioned on $\mathcal{F}_{t+1}$, Z_t is fixed. Using the inductive hypothesis, we obtain

$$\mathbb{E}\left[\exp\left(Z_t + S_{t+1}\right) \mid \mathcal{F}_{t+1}\right] \leq \exp\left(Z_t\right) \exp\left(3\sigma^2 \sum_{i=t+1}^{T} w_i\eta_i^2\right). \tag{4.2}$$

Plugging into (4.1), we obtain

$$\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_t\right] \leq \mathbb{E}\left[\exp\left(Z_t\right) \mid \mathcal{F}_t\right] \exp\left(3\sigma^2 \sum_{i=t+1}^{T} w_i\eta_i^2\right). \tag{4.3}$$

By Lemma 4.1.2

$$\begin{aligned}&\exp\left(Z_t\right) \\ &\quad= \exp\left(w_t\big(\eta_t\left(f\left(x_t\right) - f\left(x^*\right)\right) - \eta_t^2 G^2 + \mathbf{D}_\psi\left(x^*, x_{t+1}\right) - \mathbf{D}_\psi\left(x^*, x_t\right)\big) - v_t \mathbf{D}_\psi\left(x^*, x_t\right)\right) \\ &\quad\leq \exp\left(w_t\eta_t\left\langle \xi_t, x^* - x_t\right\rangle + w_t\eta_t^2\left\|\xi_t\right\|_*^2\right) \exp\left(-v_t \mathbf{D}_\psi\left(x^*, x_t\right)\right)\end{aligned}$$

Next, we analyze the first term in the last line of the above inequality in expectation. Since $\mathbb{E}\left[\left\langle \xi_t, x^* - x_t\right\rangle \mid \mathcal{F}_t\right] = 0$ we can use Lemma 4.3.2 to obtain

$$\begin{aligned}\mathbb{E}\left[\exp\left(w_t\eta_t\left\langle \xi_t, x^* - x_t\right\rangle + w_t\eta_t^2\left\|\xi_t\right\|_*^2\right) \mid \mathcal{F}_t\right] &\leq \exp\left(3\sigma^2\left(w_t^2\eta_t^2\left\|x^* - x_t\right\|^2 + w_t\eta_t^2\right)\right) \\ &\leq \exp\left(3\sigma^2\left(2w_t^2\eta_t^2 \mathbf{D}_\psi\left(x^*, x_t\right) + w_t\eta_t^2\right)\right)\end{aligned} \tag{4.4}$$

where in the last line we used that $\mathbf{D}_\psi\left(x^*, x_t\right) \geq \frac{1}{2}\left\|x^* - x_t\right\|^2$ from the strong convexity of ψ.

Plugging back into (4.3) and using that $v_t = 6\sigma^2\eta_t^2 w_t^2$, we obtain the desired inequality

$$\begin{aligned}\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_t\right] &\leq \exp\left(\left(6\sigma^2\eta_t^2 w_t^2 - v_t\right) \mathbf{D}_\psi\left(x^*, x_t\right) + 3\sigma^2 \sum_{i=t}^{T} w_i\eta_i^2\right) \\ &= \exp\left(3\sigma^2 \sum_{i=t}^{T} w_i\eta_i^2\right).\end{aligned}$$

□

Using Theorem 4.1.3 and Markov's inequality, we obtain the following convergence guarantee.

Corollary 4.1.4. *Suppose the sequence $\{w_t\}$ satisfies the conditions of Theorem 4.1.3 and that $w_t + 6\sigma^2\eta_t^2 w_t^2 \leq w_{t-1}$. For any $\delta > 0$, with probability at least $1-\delta$:*

$$\sum_{t=1}^{T} w_t \eta_t \left(f\left(x_t\right) - f\left(x^*\right)\right) + w_T \mathbf{D}_\psi\left(x^*, x_{T+1}\right) \leq w_0 \mathbf{D}_\psi\left(x^*, x_1\right) + \left(G^2 + 3\sigma^2\right) \sum_{t=1}^{T} w_t \eta_t^2 + \log\left(\frac{1}{\delta}\right)$$

With the above result in hand, we complete the convergence analysis by showing how to define the sequence $\{w_t\}$ with the desired properties. For the stochastic Mirror Descent algorithm with fixed step sizes $\eta_t = \frac{\eta}{\sqrt{T}}$, we set $w_T = \frac{1}{12\sigma^2\eta^2}$ and $w_{t-1} = w_t + \frac{6}{T}\sigma^2\eta^2 w_t^2$ for all $1 \leq t \leq T$. For Stochastic Mirror Descent algorithm with time-varying step sizes $\eta_t = \frac{\eta}{\sqrt{t}}$, we set $w_T = \frac{1}{12\sigma^2\eta^2\left(\sum_{t=1}^{T}\frac{1}{t}\right)}$ and $w_{t-1} = w_t + 6\sigma^2\eta_t^2 w_t^2$ for all $1 \leq t \leq T$. In Section 4.4, we show that these choices have the give us the results in Theorem 4.1.1.

4.1.2 Analysis of Accelerated Stochastic Mirror Descent

Algorithm 2 Accelerated Stochastic Mirror Descent Algorithm Lan (2020).

Parameters: initial point $x_0 = y_0 = z_0 \in \mathcal{X}$, step size η, strongly convex mirror map ψ

for $t = 1$ to T:

Set $\alpha_t = \frac{2}{t+1}$

$x_t = (1-\alpha_t)\, y_{t-1} + \alpha_t z_{t-1}$

$z_t = \arg\min_{x\in\mathcal{X}} \left(\eta_t \left\langle \widehat{\nabla} f(x_t), x \right\rangle + \mathbf{D}_\psi\left(x, z_{t-1}\right)\right)$

$y_t = (1-\alpha_t)\, y_{t-1} + \alpha_t z_t$

return y_T

In this section, we extend the analysis detailed in the previous section to analyze the Accelerated Stochastic Mirror Descent Algorithm (Algorithm (2)). We assume that f satisfies the following condition: for all $x, y \in \mathcal{X}$

$$f(y) \leq f(x) + \langle \nabla f\left(x\right), y - x\rangle + G\left\|y - x\right\| + \frac{L}{2}\left\|y - x\right\|^2 \tag{4.5}$$

L-smooth functions, G-Lipschitz functions, and their sums all satisfy the above condition. The full convergence guarantees are given in Theorem 4.4.3. We will only highlight the application of the previous analysis in this case. As before, we define $\xi_t := \widehat{\nabla} f\left(x_t\right) - \nabla f\left(x_t\right)$.

We also start with the inequalities shown in the standard analysis, e.g, from Lan (2020) (proof in Section 4.4).

Lemma 4.1.5. *Lan (2020) For every iteration t, we have*

$$\begin{aligned} B_t :=& \frac{\eta_t}{\alpha_t}\left(f\left(y_t\right)-f\left(x^*\right)\right)-\frac{\eta_t\left(1-\alpha_t\right)}{\alpha_t}\left(f\left(y_{t-1}\right)-f\left(x^*\right)\right) \\ &-\frac{\eta_t^2}{1-L\alpha_t\eta_t}G^2+\mathbf{D}_\psi\left(x^*,z_t\right)-\mathbf{D}_\psi\left(x^*,z_{t-1}\right) \\ \leq& \eta_t\left\langle\xi_t,x^*-z_{t-1}\right\rangle+\frac{\eta_t^2}{1-L\alpha_t\eta_t}\left\|\xi_t\right\|_*^2. \end{aligned}$$

We now turn our attention to our main concentration argument. Similar to the previous section, we define

$$\begin{aligned} Z_t &= w_tB_t-v_t\mathbf{D}_\psi\left(x^*,z_{t-1}\right), && \forall 1\leq t\leq T \\ \text{where } v_t &= 6\sigma^2w_t^2\eta_t^2 \\ \text{and } S_t &= \sum_{i=t}^{T}Z_i, && \forall 1\leq t\leq T+1 \end{aligned}$$

Notice that here we are following the exact same step as before. By transferring the error terms in the RHS of Lemma 4.1.5 into the Bregman divergence terms $\mathbf{D}_\psi\left(x^*,z_{t-1}\right)$, we can absorb them by setting the coefficients appropriately. In the same manner, we can show the following theorem.

Theorem 4.1.6. *Suppose that $\frac{w_t\eta_t^2}{1-L\alpha_t\eta_t}\leq\frac{1}{4\sigma^2}$ for every $0\leq t\leq T$. For every $1\leq t\leq T+1$, we have*

$$\mathbb{E}\left[\exp\left(S_t\right)\mid\mathcal{F}_t\right]\leq\exp\left(3\sigma^2\sum_{i=t}^{T}w_i\frac{\eta_i^2}{1-L\alpha_i\eta_i}\right).$$

Corollary 4.1.7. *Suppose the sequence $\{w_t\}$ satisfies the conditions of Theorem 4.1.6. For any $\delta>0$, the following event holds with probability at least $1-\delta$:*

$$\begin{aligned} &\sum_{t=1}^{T}w_t\left(\frac{\eta_t}{\alpha_t}\left(f\left(y_t\right)-f\left(x^*\right)\right)-\frac{\eta_t\left(1-\alpha_t\right)}{\alpha_t}\left(f\left(y_{t-1}\right)-f\left(x^*\right)\right)\right)+w_T\mathbf{D}_\psi\left(x^*,z_T\right) \\ \leq& w_0\mathbf{D}_\psi\left(x^*,z_0\right)+\left(G^2+3\sigma^2\right)\sum_{t=1}^{T}w_t\frac{\eta_t^2}{1-L\alpha_t\eta_t}+\log\left(\frac{1}{\delta}\right). \end{aligned}$$

With the above result in hand, we can complete the convergence analysis by showing how to define the sequence $\{w_t\}$ with the desired properties. Theorem 4.4.3 can be obtained from corollaries 4.4.4 and 4.4.5 provided in Section 4.4, for constant and time-varying step sizes.

4.2 Non-convex Case: Stochastic Gradient Descent and AdaGrad

In this section, we consider non-convex objectives and analyze the Stochastic Gradient Descent algorithm (Algorithm 3) along with two versions of AdaGrad: (1) AdaGrad-Norm Ward et al. (2019) (Algorithm 4), where the step-size is a scalar, and (2) the original AdaGrad algorithm Duchi et al. (2011) (Algorithm (5)), where

the step-size for each coordinates varies. Since AdaGrad-Norm is simpler to analyze, most results for AdaGrad have been for this scalar version either in-expectation Ward et al. (2019), Faw et al. (2022), Li and Orabona (2020), Li and Orabona (2019), Liu et al. (2022), and Ene et al. (2021) or high-probability Kavis et al. (2021). For the standard AdaGrad algorithm, to the best of our knowledge, Défossez et al. (2022) is the only work that has analyzed the standard version of AdaGrad in expectation, but their result does not adapt to noise and requires a strong assumption: the stochastic gradients are uniformly bounded. On the other hand, our high probability result for vanilla AdaGrad adapts to noise and holds under relatively mild assumptions.

Recall that, we assume that the optimization problem has domain $\mathcal{X} = \mathbb{R}^d$. As usual in non-convex analysis, we assume that f is an L-smooth function: for all $x, y \in \mathbb{R}^d$,

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\| .$$

Smoothness implies the following quadratic upperbound that we will utilize: for all $x, y \in \mathbb{R}^d$

$$f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 . \tag{4.6}$$

4.2.1 Analysis of Stochastic Gradient Descent

Algorithm 3 Stochastic Gradient Descent (SGD)

Parameters: initial point x_1, step sizes $\{\eta_t\}$
for $t = 1$ to T do
  $x_{t+1} = x_t - \eta_t \widehat{\nabla} f(x_t)$

We will prove the following convergence guarantee of Algorithm 3.

Theorem 4.2.1. *Assume f is L-smooth and satisfies Assumptions (1'), (2), and that the gradient noise is σ-sub-gaussian. Let $\Delta_1 := f(x_1) - f_*$. With probability at least $1 - \delta$, the iterate sequence $(x_t)_{t \geq 1}$ output by Algorithm 3 satisfies*

(1) Setting $\eta_t = \min\left\{\frac{1}{L}; \sqrt{\frac{\Delta_1}{\sigma^2 L T}}\right\}$,

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \leq \frac{2\Delta_1 L}{T} + 5\sigma\sqrt{\frac{\Delta_1 L}{T}} + \frac{12\sigma^2 \log\frac{1}{\delta}}{T};$$

(2) Setting $\eta_t = \frac{1}{L\sqrt{t}}$,

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \leq \frac{2\Delta_1 L + 3\sigma^2(1 + \log T) + 12\sigma^2 \log\frac{1}{\delta}}{\sqrt{T}}.$$

Comparison with prior works: When the time horizon T is known to the algorithm, by choosing the step size η in part (1) of Theorem 4.2.1, the bound is adaptive to noise, i.e, when $\sigma = 0$ we recover $O(\frac{1}{T})$ convergence rate of the (deterministic) gradient descent algorithm. Notice that the bound in this case does not have a $\log T$ term incurred. When T is unknown, the extra $\log T$ appears as a result of setting a time-varying step size $\eta_t = \frac{1}{L\sqrt{t}}$. This $\log T$ appears as an additive term to the $\log\frac{1}{\delta}$ term, as opposed to being multiplicative, i.e, $\log T \log\frac{1}{\delta}$ as in previous works Li and Orabona (2020), Madden et al. (2020), and Li and Liu (2022).

Analysis: To proceed, we define for $t \geq 1$

$$\Delta_t := f(x_t) - f_*; \quad \xi_t := \widehat{\nabla} f(x_t) - \nabla f(x_t).$$

We let $\mathcal{F}_t := \sigma(\xi_1, \ldots, \xi_{t-1})$ denote the natural filtration. Note that x_t is $\mathcal{F}_t$-measurable. The following lemma serves as a fundamental step of our analysis; the proof of which can be found in Section 4.5.

Lemma 4.2.2. *For $t \geq 1$, we have*

$$\begin{aligned} C_t &:= \eta_t \left(1 - \frac{L\eta_t}{2}\right) \|\nabla f(x_t)\|^2 + \Delta_{t+1} - \Delta_t \\ &\leq \left(L\eta_t^2 - \eta_t\right) \langle \nabla f(x_t), \xi_t \rangle + \frac{L\eta_t^2}{2} \|\xi_t\|^2. \end{aligned} \tag{4.7}$$

Now we can follow the similar concentration argument from the convex setting. The difference now is the error term in the RHS of (4.7) can be transferred into the gradient term $\|\nabla f(x_t)\|^2$ instead of a function value gap term. This actually makes things easier since this term can be readily absorbed by the gradient term in C_t, and we do not have to carefully impose an additional condition on w_t to make a telescoping sum. For $w_t \geq 0$, we define

$$\begin{aligned} Z_t &= w_t C_t - v_t \|\nabla f(x_t)\|^2, && \forall 1 \leq t \leq T \\ \text{where } v_t &= 3\sigma^2 w_t^2 \eta_t^2 (\eta_t L - 1)^2 \\ \text{and } S_t &= \sum_{i=t}^{T} Z_i. && \forall 1 \leq t \leq T+1 \end{aligned}$$

Using the same technique as in the previous Section, we can prove the following key inequality.

Theorem 4.2.3. *Suppose for all $1 \leq t \leq T$, η_t, w_t satisfying $0 \leq w_t \eta_t^2 L \leq \frac{1}{2\sigma^2}$ then*

$$\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_t\right] \leq \exp\left(3\sigma^2 \sum_{s=t}^{T} \frac{w_t \eta_t^2 L}{2}\right). \tag{4.8}$$

Markov's inequality gives us the following guarantee.

Corollary 4.2.4. For all $1 \leq t \leq T$, if $\eta_t L \leq 1$ and $0 \leq w_t \eta_t^2 L \leq \frac{1}{2\sigma^2}$ *then*

$$\begin{aligned} &\sum_{t=1}^{T} \left[w_t \eta_t \left(1 - \frac{\eta_t L}{2}\right) - v_t\right] \|\nabla f(x_t)\|^2 + w_T \Delta_{T+1} \\ \leq & w_1 \Delta_1 + \left(\sum_{t=2}^{T} (w_t - w_{t-1}) \Delta_t + 3\sigma^2 \sum_{t=1}^{T} \frac{w_t \eta_t^2 L}{2}\right) + \log \frac{1}{\delta}. \end{aligned} \tag{4.9}$$

Equipped with Lemmas 4.2.2 and 4.2.3, we are ready to prove Theorem 4.2.1 by specifying the choice of w_t that satisfy the condition of Lemma 4.2.3. In the first case, we choose $\eta_t = \eta$, $w_t = w = \frac{1}{6\sigma^2 \eta}$ where $\eta = \min\{\frac{1}{L}; \sqrt{\frac{\Delta_1}{\sigma^2 LT}}\}$. In the second case, we set $\eta_t = \frac{\eta}{\sqrt{t}}$, $w_t = w = \frac{1}{6\sigma^2 \eta}$ where $\eta = \frac{1}{L}$. We show the full proof in Section 4.5.

Algorithm 4 AdaGrad-Norm

Parameters: $x_1, \eta > 0$.
for $t = 1$ to T
$b_t = \sqrt{b_0^2 + \sum_{i=1}^{t} \|\widehat{\nabla} f(x_i)\|^2}$
$x_{t+1} = x_t - \frac{\eta}{b_t} \widehat{\nabla} f(x_t)$

Algorithm 5 AdaGrad

Parameters: $x_1, b_0 \in \mathbb{R}^d$ and $\eta \in \mathbb{R}$.
for $t = 1$ to T do
$b_{t,i} = \sqrt{b_{0,i}^2 + \sum_{j=1}^{t} \widehat{\nabla}_i f(x_j)^2}$, for $i \in [d]$.
$x_{t+1,i} = x_{t,i} - \frac{\eta}{b_{t,i}} \widehat{\nabla}_i f(x_t)$, for $i \in [d]$.

4.2.2 AdaGrad-Norm and AdaGrad-Coordinate

In this section, we present our main results for the high probability convergence for non-convex objectives of AdaGrad-Norm Ward et al. (2019) (Algorithm 4) as well as the standard AdaGrad Duchi et al. (2011) algorithm (Algorithm 5) that updates each coordinate separately. Here, $d \in \mathbb{N}$ denotes the dimension of the problem, v_i denotes the i-th coordinate of a vector v, and $\widehat{\nabla}_i f(x_t)$ denotes the i-th coordinate of the stochastic gradient at time t.

Comparison with prior works: Ward et al. (2019) and Faw et al. (2022) show the convergence of AdaGrad-Norm with polynomial dependency on poly $\left(\frac{1}{\delta}\right)$ where $1 - \delta$ is the success probability. The latter relaxes several assumptions made in the former, including the boundedness of the gradients and noise variance. When assuming a sub-Gaussian noise, Kavis et al. (2021) show a convergence in high probability, but still assume that the gradients are bounded which circumvents many of the difficulties due to the error term. We remove this assumption and establish the convergence of AdaGrad-Norm in the theorem 4.2.5. Unlike existing work, the technique employed to prove this theorem readily extends to the standard version of AdaGrad (Algorithm 5) with per-coordinate update.

For simplicity, we let $\Delta_t := f(x_t) - f_*$, where f_* is any valid lower bound for f.

Theorem 4.2.5. *If f is L-smooth and satisfies assumptions (1'), (2) and (3). With probability at least $1 - \delta$, the iterate sequence $(x_t)_{t \geq 1}$ output by AdaGrad-Norm (Algorithm 4) satisfies*

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \leq g(\delta) \cdot O\left(\frac{\sigma}{\sqrt{T}} + \frac{r(\delta)}{T}\right).$$

where

$$g(\delta) := O\left(\Delta_1 + c(\delta)\sqrt{\log \frac{T}{\delta}} + L \log\left(\sigma\sqrt{T} + r(\delta)\right)\right)$$

$$c(\delta) := O\left(\sigma^3 \log\left(\frac{1}{\delta}\right) + \sigma \log\left(1 + \sigma^2 T + \sigma^2 \log \frac{1}{\delta}\right) + \sigma \log\left(\sigma\sqrt{T} + r(\delta)\right)\right), \text{ and}$$

$$r(\delta) := O(\Delta_1 + \sigma^2 \log \frac{1}{\delta} + L \log L)$$

are polylog terms.

The next theorem show the first convergence result in high-probability for vanilla AdaGrad in the non-convex regime.

Theorem 4.2.6. *If f is L-smooth and satisfies assumptions (1'), (2) and (3). With probability at least $1-\delta$, the iterate sequence $(x_t)_{t\geq 1}$ output by AdaGrad (Algorithm 5) satisfies*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x_t)\|_1^2 \leq g(\delta)\cdot O\left(\frac{\|\sigma\|_1}{\sqrt{T}}+\frac{r(\delta)}{T}\right),$$

where

$$g(\delta) := O\left(\Delta_1 + \left(d\sigma_{\max}+\sum_{i=1}^{d}c_i(\delta)\right)\sqrt{\log\frac{dT}{\delta}} + dL\log\left(\|\sigma\|_1\sqrt{T}+r(\delta)\right)\right),$$

$$c_i(\delta) := O\left(\sigma_i^3\log\left(\frac{d}{\delta}\right)+\sigma_i\log\left(1+\sigma_i^2T+\sigma_i^2\log\frac{d}{\delta}\right)+\|\sigma\|_1\log\left(\|\sigma\|_1\sqrt{T}+r(\delta)\right)\right), \text{ and}$$

$$r(\delta) := O\left(\Delta_1+\left\|\sigma^2\right\|_1\log\left(\frac{d}{\delta}\right)+\|\sigma\|_1\sqrt{\log\frac{d}{\delta}}+Ld\log L\right),$$

are the polylog terms.

Both of these results are adaptive to noise: the rate $\tilde{O}\left(\frac{1}{\sqrt{T}}\right)$ will improve to $\tilde{O}\left(\frac{1}{T}\right)$ as the noise σ approaches 0. Furthermore, they hold regardless of how η and b_0 is set.

Analysis overview The first key new technique is unlike prior works: we do not use the division by the step size, which makes the analysis of AdaGrad-Norm and AdaGrad virtually the same. We can thus focus on AdaGrad-Norm. To obtain a high probability bound, our analysis of AdaGrad-Norm utilizes the same martingale concentration technique as presented throughout this paper to bound the error terms $\eta_t\langle\nabla f(x_t),\xi_t\rangle$. However, the step size $\eta_t=\frac{\eta}{b_t}$ now has a dependency on the randomness at time t due to b_t, preventing us from applying Lemma 4.3.2. To circumvent this, inspired by Ward et al. (2019), we introduce a proxy step size $a_t := b_{t-1}^2+\|\nabla f(x_t)\|^2$ that replaces the stochastic gradient with the true gradient at time t for analysis purposes. Using that along with standard smoothness analysis, we obtain

Lemma 4.2.7. *For $t\geq 1$, let $\xi_t=\widehat{\nabla} f x_t)-\nabla f(x_t)$, $a_t^2 := b_{t-1}^2+\|\nabla f(x_t)\|^2$, and $M_t=\max_{i\leq t}\|\xi_i\|$, then we have*

$$\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{b_t} \leq \frac{\Delta_1}{\eta}+\frac{M_T}{2}\left[\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{a_t^2}+\sum_{t=1}^{T}\frac{\|\xi_t\|^2}{b_t^2}\right] - \sum_{t=1}^{T}\frac{1}{a_t}\langle\nabla f(x_t),\xi_t\rangle+\sum_{t=1}^{T}\frac{L\eta}{2b_t^2}\left\|\widehat{\nabla} f x_t)\right\|^2.$$

Now, the randomness at time t of the error term $\frac{1}{a_t}\langle\nabla f(x_t),\xi_t\rangle$ only depends on ξ_t, which follows a sub-Gaussian distribution with mean 0. Hence, we can utilize our previous techniques to bound $-\sum_{t=1}^{T}\frac{1}{a_t}\langle\nabla f(x_t),\xi_t\rangle$ with high probability. Comparing to the analysis in expectation from Ward et al. (2019), terms like $\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{a_t^2}$ must be handled more carefully to obtain a high probability bound. A bound for M_T has also been derived in previous works by Li and Orabona (2020) and Liu et al. (2022). Combining with Lemma 4.2.7, we obtain the following lemma.

Lemma 4.2.8. *With probability at least* $1-2\delta$*, we have*

$$\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{b_t}\leq\frac{\Delta_1}{\eta}+\sigma\sqrt{\log\frac{T}{\delta}}\left[8\log\left(\frac{b_T}{b_0}\right)+5\sum_{t=1}^{T}\frac{\|\xi_t\|^2}{b_t^2}\right]+\sigma\sqrt{\log\frac{1}{\delta}}+L\eta\log\frac{b_T}{b_0}.$$

Since, $\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{b_t}\geq\frac{1}{b_T}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2$, it suffices to bound b_T and $\sum_{t=1}^{T}\frac{\|\xi_t\|^2}{b_t^2}$ from this point on (see Lemma 4.6.1 and Lemma 4.6.6). The analysis for these terms utilize similar martingale techniques throughout this paper, where the details are deferred to Section 4.6. For the coordinate version of AdaGrad, since our techniques only rely on addition and scalar multiplication, we can (with some effort) generalize our technique to the vanilla Adagrad version. The full proofs for vanilla AdaGrad are presented in Section 4.7.

4.3 Technical Tools

We will use the following technical tools throughout our analysis for light-tailed noise.

Lemma 4.3.1. *For any* $a\geq 0$, $0\leq b\leq\frac{1}{2\sigma}$ *and an* σ*-sub-Gaussian random variable* X*,*

$$\mathbb{E}\left[1+b^2X^2+\sum_{i=2}^{\infty}\frac{1}{i!}\left(aX+b^2X^2\right)^i\right]\leq\exp\left(3\left(a^2+b^2\right)\sigma^2\right).$$

Especially, when $b=0$*, we have*

$$\mathbb{E}\left[1+\sum_{i=2}^{\infty}\frac{1}{i!}\left(aX\right)^i\right]\leq\exp\left(2a^2\sigma^2\right).$$

Proof of Lemma 4.3.1. Consider two cases either $a\geq 1/(2\sigma)$ or $a\leq 1/(2\sigma)$. First suppose $a\geq 1/(2\sigma)$. We use the inequality $uv\leq\frac{u^2}{4}+v^2$ here to first obtain

$$\left(aX+b^2X^2\right)^i\leq\left|aX+b^2X^2\right|^i\leq\left(a\left|X\right|+b^2X^2\right)^i\leq\left(\frac{1}{4\sigma^2}X^2+a^2\sigma^2+b^2X^2\right)^i.$$

Thus, we have

$$\begin{aligned}
&\mathbb{E}\left[1+b^2X^2+\sum_{i=2}^{\infty}\frac{1}{i!}\left(aX+b^2X^2\right)^i\right]\\
&\qquad\leq\mathbb{E}\left[1+b^2X^2+\sum_{i=2}^{\infty}\frac{1}{i!}\left(\frac{1}{4\sigma^2}X^2+a^2\sigma^2+b^2X^2\right)^i\right]\\
&\qquad=\mathbb{E}\left[b^2X^2+\exp\left(\left(\frac{1}{4\sigma^2}+b^2\right)X^2+a^2\sigma^2\right)-\left(\frac{1}{4\sigma^2}+b^2\right)X^2-a^2\sigma^2\right]\\
&\qquad=\mathbb{E}\left[\exp\left(\left(\frac{1}{4\sigma^2}+b^2\right)X^2+a^2\sigma^2\right)-\frac{1}{4\sigma^2}X^2-a^2\sigma^2\right]\\
&\qquad\leq\exp\left(\left(\frac{1}{4\sigma^2}+b^2\right)\sigma^2+a^2\sigma^2\right)\\
&\qquad\leq\exp\left(2a^2\sigma^2+b^2\sigma^2\right)\\
&\qquad\leq\exp\left(3\left(a^2+b^2\right)\sigma^2\right)
\end{aligned}$$

Next, let $c = \max(a, b) \leq 1/(2\sigma)$. We have

$$\begin{aligned}
\mathbb{E}\left[1 + b^2X^2 + \sum_{i=2}^{\infty} \frac{1}{i!} \left(aX + b^2X^2\right)^i\right] &= \mathbb{E}\left[\exp\left(aX + b^2X^2\right) - aX\right] \\
&\leq \mathbb{E}\left[\left(aX + \exp\left(a^2X^2\right)\right) \exp\left(b^2X^2\right) - aX\right] \\
&= \mathbb{E}\left[\exp\left(\left(a^2 + b^2\right)X^2\right) + aX\left(\exp\left(b^2X^2\right) - 1\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(\left(a^2 + b^2\right)X^2\right) + c\,|X|\left(\exp\left(c^2X^2\right) - 1\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(\left(a^2 + b^2\right)X^2\right) + \exp\left(2c^2X^2\right) - 1\right] \\
&\leq \mathbb{E}\left[\exp\left(\left(a^2 + b^2 + 2c^2\right)X^2\right)\right] \\
&\leq \exp\left(\left(a^2 + b^2 + 2c^2\right)\sigma^2\right) \\
&\leq \exp\left(3\left(a^2 + b^2\right)\sigma^2\right).
\end{aligned}$$

In the first inequality, we use the inequality $e^x - x \leq e^{x^2} \forall x$. In the third inequality, we use $x\left(e^{x^2} - 1\right) \leq e^{2x^2} - 1\ \forall x$. This inequality can be proved with the Taylor expansion.

$$\begin{aligned}
x\left(e^{x^2} - 1\right) &= \sum_{i=1}^{\infty} \frac{1}{i!} x^{2i+1} \\
&\leq \sum_{i=1}^{\infty} \frac{1}{i!} \frac{x^{2i} + x^{2i+2}}{2} \\
&= \frac{x^2}{2} + \sum_{i=2}^{\infty} \left(\frac{1+i}{2i!}\right) x^{2i} \\
&\leq \frac{x^2}{2} + \sum_{i=2}^{\infty} \left(\frac{2^i}{i!}\right) x^{2i} \\
&\leq e^{2x^2} - 1
\end{aligned}$$

The case when $b = 0$ simply follows from the above proof. □

That implies the following results:

Lemma 4.3.2. *Suppose $X \in \mathbb{R}^d$ such that $\mathbb{E}[X] = 0$ and $\|X\|$ is a σ-sub-Gaussian random variable, then for any $a \in \mathbb{R}^d$, $0 \leq b \leq \frac{1}{2\sigma}$,*

$$\mathbb{E}\left[\exp\left(\langle a, X\rangle + b^2\|X\|^2\right)\right] \leq \exp\left(3\left(\|a\|_*^2 + b^2\right)\sigma^2\right).$$

Especially, when $b = 0$, we have

$$\mathbb{E}\left[\exp\left(\langle a, X\rangle\right)\right] \leq \exp\left(2\|a\|_*^2\,\sigma^2\right).$$

Proof of Lemma 4.3.2. Using Taylor expansion of e^x and the fact that $\mathbb{E}[X] = 0$ we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\langle a, X\rangle + b^2\|X\|^2\right)\right] &= \mathbb{E}\left[1 + \langle a, X\rangle + b^2\|X\|^2 + \sum_{i=2}^{\infty}\frac{1}{i!}\left(\langle a, X\rangle + b^2\|X\|^2\right)^i\right] \\ &= \mathbb{E}\left[1 + b^2\|X\|^2 + \sum_{i=2}^{\infty}\frac{1}{i!}\left(\langle a, X\rangle + b^2\|X\|^2\right)^i\right] \\ &\leq \mathbb{E}\left[1 + b^2\|X\|^2 + \sum_{i=2}^{\infty}\frac{1}{i!}\left(\|a\|_*\|X\| + b^2\|X\|^2\right)^i\right]\end{aligned}$$

where for the last line we use Cauchy-Schwartz to obtain $\langle a, X\rangle \leq \|a\|_*\|X\|$. Now applying Lemma 4.3.1, we obtain

$$\mathbb{E}\left[\exp\left(\langle a, X\rangle + b^2\|X\|^2\right)\right] \leq \exp\left(3\left(\|a\|_*^2 + b^2\right)\sigma^2\right)$$

Similarly, we obtain the corresponding bound for the case $b = 0$. □

We can generalize that to the vector version:

Corollary 4.3.3. *Suppose that X is a mean zero random vector in $\mathbb{R}^d$, where $\|X\|$ is σ-subgaussian. For $0 \leq a \leq \frac{1}{4\sigma^2}$ and $B \in \mathbb{R}^d$ then*

$$\mathbb{E}\left[\exp\left(a\|X\|^2 + \langle B, X\rangle\right)\right] \leq \exp\left(3\sigma^2(a + \|B\|^2)\right).$$

Proof. We have

$$\begin{aligned}\mathbb{E}\left[\exp\left(a\|X\|^2 + \langle B, X\rangle\right)\right] &= \mathbb{E}\left[1 + a^2\|X\|^2 + \langle B, X\rangle + \sum_{k=2}^{\infty}\frac{1}{k!}\left(a\|X\|^2 + \langle B, X\rangle\right)^k\right] \\ &= \mathbb{E}\left[1 + a\|X\|^2 + \sum_{k=2}^{\infty}\frac{1}{k!}\left(a\|X\|^2 + \langle B, X\rangle\right)^k\right] \\ &\leq \mathbb{E}\left[1 + a\|X\|^2 + \sum_{k=2}^{\infty}\frac{1}{k!}\left(a\|X\|^2 + \|B\|\|X\|\right)^k\right] \\ &\leq \exp\left(3\sigma^2(a + \|B\|^2)\right).\end{aligned}$$

□

We can now control martingale via:

Lemma 4.3.4. *If we have a sequence of random variable X_t with $\mathcal{F}_t = \sigma(X_1, X_2, \ldots, X_{t-1})$ for $t = 1, 2, \ldots, T$. If we can bound $\mathbb{E}[\exp(X_t) \mid \mathcal{F}_t] \leq \exp(Y_t)$, where Y_t is $\mathcal{F}_t$-measurable, then*

$$\sum_{t=1}^{T} X_t \leq \sum_{t=1}^{T} Y_t + \log(1/\delta)$$

holds with probability at least $1 - \delta$.

Proof. Define the $Z_t = X_t - Y_t$ and $S_t = \sum_{i=t}^{T} Z_i$. Then

$$\begin{aligned}\mathbb{E}\left[\exp\left(Z_t\right) \mid \mathcal{F}_t\right] &= \mathbb{E}\left[\exp\left(X_t - Y_t\right) \mid \mathcal{F}_t\right] \\ &= \exp\left(-Y_t\right) \mathbb{E}\left[\exp\left(X_t\right) \mid \mathcal{F}_t\right] \qquad (Y_t \text{ is } \mathcal{F}_t\text{-measurable}) \\ &\leq \exp(-Y_t)\exp(Y_t) = \exp(0) = 1.\end{aligned}$$

Then we show $\mathbb{E}\left[\exp\left(S_1\right)\right] \leq 1$ via an induction: we have $\mathbb{E}\left[\exp\left(S_T\right) \mid \mathcal{F}_T\right] = \mathbb{E}\left[\exp\left(Z_T\right) \mid \mathcal{F}_T\right] \leq 1$. Suppose that $\mathbb{E}\left[\exp\left(S_{t+1}\right) \mid \mathcal{F}_{t+1}\right]$

$$\begin{aligned}\mathbb{E}\left[\exp(S_t) \mid \mathcal{F}_t\right] &= \mathbb{E}\left[\exp(Z_t)\exp(S_{t+1}) \mid \mathcal{F}_t\right] \\ &= \mathbb{E}\left[\exp\left(Z_t\right) \mathbb{E}\left[\exp\left(S_{t+1}\right) \mid \mathcal{F}_{t+1}\right] \mid \mathcal{F}_t\right] \\ &\leq \mathbb{E}\left[\exp(Z_t) \mid \mathcal{F}_t\right] \leq 1.\end{aligned}$$

Hence, this implies that $\mathbb{E}\left[\exp(S_1)\right] \leq 1$. By Markov's inequality, this means that $S_1 \leq \log(\frac{1}{\delta})$ with probability at least $1 - \delta$:

$$\begin{aligned}S_1 = \sum_{t=1}^{T} Z_t = \sum_{t=1}^{T} X_t - Y_t &\leq \log\left(1/\delta\right) \\ \implies \sum_{t=1}^{T} X_t &\leq \sum_{t=1}^{T} Y_t + \log\left(1/\delta\right).\end{aligned}$$

□

4.4 Missing Proofs from Section 4.1

4.4.1 Stochastic Mirror Descent

Proof of Lemma (4.1.2). By the optimality condition, we have

$$\left\langle \eta_t \widehat{\nabla} f(x_t) + \nabla_x \mathbf{D}_\psi\left(x_{t+1}, x_t\right), x^* - x_{t+1} \right\rangle \geq 0$$

and thus

$$\left\langle \eta_t \widehat{\nabla} f(x_t), x_{t+1} - x^* \right\rangle \leq \left\langle \nabla_x \mathbf{D}_\psi\left(x_{t+1}, x_t\right), x^* - x_{t+1} \right\rangle$$

Note that

$$\begin{aligned}\left\langle \nabla_x \mathbf{D}_\psi\left(x_{t+1}, x_t\right), x^* - x_{t+1} \right\rangle &= \left\langle \nabla\psi\left(x_{t+1}\right) - \nabla\psi\left(x_t\right), x^* - x_{t+1} \right\rangle \\ &= \mathbf{D}_\psi\left(x^*, x_t\right) - \mathbf{D}_\psi\left(x_{t+1}, x_t\right) - \mathbf{D}_\psi\left(x^*, x_{t+1}\right)\end{aligned}$$

and thus

$$\begin{aligned}\eta_t \left\langle \widehat{\nabla} f(x_t), x_{t+1} - x^* \right\rangle &\leq \mathbf{D}_\psi\left(x^*, x_t\right) - \mathbf{D}_\psi\left(x^*, x_{t+1}\right) - \mathbf{D}_\psi\left(x_{t+1}, x_t\right) \\ &\leq \mathbf{D}_\psi\left(x^*, x_t\right) - \mathbf{D}_\psi\left(x^*, x_{t+1}\right) - \frac{1}{2}\left\|x_{t+1} - x_t\right\|^2\end{aligned}$$

where we have used that $\mathbf{D}_\psi\left(x_{t+1}, x_t\right) \geq \frac{1}{2}\left\|x_{t+1} - x_t\right\|^2$ by the strong convexity of ψ.

By convexity,

$$f\left(x_t\right) - f\left(x^*\right) \leq \left\langle \nabla f\left(x_t\right), x_t - x^* \right\rangle = \left\langle \xi_t, x^* - x_t \right\rangle + \left\langle \widehat{\nabla} f\left(x_t\right), x_t - x^* \right\rangle$$

Combining the two inequalities, we obtain

$$\begin{aligned}
&\eta_t \left(f\left(x_t\right) - f\left(x^*\right)\right) + \mathbf{D}_\psi\left(x^*, x_{t+1}\right) - \mathbf{D}_\psi\left(x^*, x_t\right) \\
&\leq \eta_t \left\langle \xi_t, x^* - x_t \right\rangle + \eta_t \left\langle \widehat{\nabla} f(x_t), x_t - x_{t+1} \right\rangle - \frac{1}{2}\left\|x_{t+1} - x_t\right\|^2 \\
&\leq \eta_t \left\langle \xi_t, x^* - x_t \right\rangle + \frac{\eta_t^2}{2}\left\|\widehat{\nabla} f(x_t)\right\|_*^2
\end{aligned}$$

Using the triangle inequality and the bounded gradient assumption $\|\nabla f(x)\|_* \leq G$, we obtain

$$\left\|\widehat{\nabla} f(x_t)\right\|_*^2 = \left\|\xi_t + \nabla f(x_t)\right\|_*^2 \leq 2\left\|\xi_t\right\|_*^2 + 2\left\|\nabla f(x_t)\right\|_*^2 \leq 2\left(\left\|\xi_t\right\|_*^2 + G^2\right).$$

Thus

$$\eta_t \left(f\left(x_t\right) - f\left(x^*\right)\right) + \mathbf{D}_\psi\left(x^*, x_{t+1}\right) - \mathbf{D}_\psi\left(x^*, x_t\right) \leq \eta_t \left\langle \xi_t, x^* - x_t \right\rangle + \eta_t^2\left(\left\|\xi_t\right\|_*^2 + G^2\right)$$

as needed. □

Proof of Corollary 4.1.4. Let

$$K = 3\sigma^2 \sum_{t=1}^T w_t \eta_t^2 + \log\left(\frac{1}{\delta}\right).$$

By Theorem 4.1.3 and Markov's inequality, we have

$$\begin{aligned}
\Pr\left[S_1 \geq K\right] &\leq \Pr\left[\exp\left(S_1\right) \geq \exp\left(K\right)\right] \\
&\leq \exp\left(-K\right)\mathbb{E}\left[\exp\left(S_1\right)\right] \\
&\leq \exp\left(-K\right)\exp\left(3\sigma^2\sum_{t=1}^T w_t\eta_t^2\right) \\
&= \delta.
\end{aligned}$$

Note that since $v_t + w_t \leq w_{t-1}$

$$\begin{aligned}
S_1 &= \sum_{t=1}^T Z_t \\
&= \sum_{t=1}^T w_t \eta_t \left(f\left(x_t\right) - f\left(x^*\right)\right) - G^2\sum_{t=1}^T w_t\eta_t^2 + \sum_{t=1}^T \left(w_t \mathbf{D}_\psi\left(x^*, x_{t+1}\right) - (v_t + w_t)\mathbf{D}_\psi\left(x^*, x_t\right)\right) \\
&\geq \sum_{t=1}^T w_t \eta_t \left(f\left(x_t\right) - f\left(x^*\right)\right) - G^2\sum_{t=1}^T w_t\eta_t^2 + \sum_{t=1}^T \left(w_t \mathbf{D}_\psi\left(x^*, x_{t+1}\right) - w_{t-1}\mathbf{D}_\psi\left(x^*, x_t\right)\right) \\
&= \sum_{t=1}^T w_t \eta_t \left(f\left(x_t\right) - f\left(x^*\right)\right) - G^2\sum_{t=1}^T w_t\eta_t^2 + w_T \mathbf{D}_\psi\left(x^*, x_{T+1}\right) - w_0\mathbf{D}_\psi\left(x^*, x_1\right).
\end{aligned}$$

Therefore, with probability at least $1-\delta$, we have

$$\sum_{t=1}^{T} w_t \eta_t \left(f\left(x_t\right)-f\left(x^*\right)\right)+w_T \mathbf{D}_\psi\left(x^*, x_{T+1}\right) \leq w_0 \mathbf{D}_\psi\left(x^*, x_1\right)+\left(G^2+3\sigma^2\right) \sum_{t=1}^{T} w_t \eta_t^2+\log\left(\frac{1}{\delta}\right)$$

□

With the above result in hand, we complete the convergence analysis by showing how to define the sequence $\{w_t\}$ with the desired properties. Theorem 4.1.1 can be obtained from the two following corollaries.

Corollary 4.4.1. *Suppose we run the Stochastic Mirror Descent algorithm with fixed step sizes $\eta_t = \frac{\eta}{\sqrt{T}}$. Let $w_T = \frac{1}{12\sigma^2\eta^2}$ and $w_{t-1} = w_t + \frac{6}{T}\sigma^2\eta^2 w_t^2$ for all $1 \leq t \leq T$. The sequence $\{w_t\}$ satisfies the conditions required by Corollary 4.1.4. By Corollary 4.1.4, for any $\delta > 0$, the following events hold with probability at least $1-\delta$: $\mathbf{D}_\psi\left(x^*, x_{T+1}\right) \leq 2\mathbf{D}_\psi\left(x^*, x_1\right) + 12\left(G^2+\sigma^2\left(1+\log\left(\frac{1}{\delta}\right)\right)\right)\eta^2$, and*

$$\frac{1}{T}\sum_{t=1}^{T}\left(f\left(x_t\right)-f\left(x^*\right)\right) \leq \frac{1}{\sqrt{T}}\frac{2\mathbf{D}_\psi\left(x^*, x_1\right)}{\eta}+\frac{12}{\sqrt{T}}\left(G^2+\sigma^2\left(1+\log\left(\frac{1}{\delta}\right)\right)\right)\eta.$$

In particular, setting $\eta_t = \sqrt{\frac{\mathbf{D}_\psi(x^*,x_1)}{6\left(G^2+\sigma^2\left(1+\log\left(\frac{1}{\delta}\right)\right)\right)T}}$ we obtain the first case of Theorem 4.1.1.

Proof of Corollary (4.4.1) . Recall from Corollary 4.1.4 that the sequence $\{w_t\}$ needs to satisfy the following conditions for all $1 \leq t \leq T$:

$$w_t + 6\sigma^2\eta_t^2 w_t^2 \leq w_{t-1}$$
$$w_t\eta_t^2 \leq \frac{1}{4\sigma^2}$$

Let $C = 6\sigma^2\eta^2$. We set $w_T = \frac{1}{C+6\sigma^2\eta^2} = \frac{1}{2C}$. For $1 \leq t \leq T$, we set w_t so that the first condition holds with equality

$$w_{t-1} = w_t + 6\sigma^2 w_t^2\eta_t^2 = w_t + \frac{6}{T}\sigma^2\eta^2 w_t^2.$$

We can show by induction that, for every $1 \leq t \leq T$, we have

$$w_t \leq \frac{1}{C+\frac{6}{T}\sigma^2\eta^2 t}.$$

The base case $t = T$ follows from the definition of w_T. Consider $1 \leq t \leq T$. Using the definition of w_{t-1} and the inductive hypothesis, we obtain

$$\begin{aligned} w_{t-1} &= w_t + \frac{6}{T}\sigma^2\eta^2 w_t^2 \\ &\leq \frac{1}{C + \frac{6}{T}\sigma^2\eta^2 t} + \frac{6\sigma^2\eta^2}{T\left(C + \frac{6}{T}\sigma^2\eta^2 t\right)^2} \\ &\leq \frac{1}{C + \frac{6}{T}\sigma^2\eta^2 t} + \frac{\left(C + \frac{6}{T}\sigma^2\eta^2 t\right) - \left(C + \frac{6}{T}\sigma^2\eta^2 (t-1)\right)}{\left(C + \frac{6}{T}\sigma^2\eta^2 (t-1)\right)\left(C + \frac{6}{T}\sigma^2\eta^2 t\right)} \\ &= \frac{1}{C + \frac{6}{T}\sigma^2\eta^2 (t-1)} \end{aligned}$$

as needed.

Using this fact, we now show that $\{w_t\}$ satisfies the second condition. Indeed, for every $1 \leq t \leq T$, we have

$$w_t\eta_t^2 = w_t\frac{\eta^2}{T} \leq \frac{\eta^2}{6\sigma^2\eta^2 t} = \frac{1}{6\sigma^2}.$$

Thus, by Corollary 4.1.4, with probability $\geq 1 - \delta$, we have

$$\sum_{t=1}^{T} w_t\eta_t \left(f(x_t) - f(x^*)\right) + w_T \mathbf{D}_\psi(x^*, x_{T+1}) \leq w_0 \mathbf{D}_\psi(x^*, x_1) + \left(G^2 + 3\sigma^2\right)\sum_{t=1}^{T} w_t\eta_t^2 + \log\left(\frac{1}{\delta}\right)$$

Note that $w_T = \frac{1}{2C}$ and $\frac{1}{2C} \leq w_t \leq \frac{1}{C}$ for all $0 \leq t \leq T$. Thus we obtain

$$\begin{aligned} \frac{\eta}{\sqrt{T}}\sum_{t=1}^{T}\left(f(x_t) - f(x^*)\right) + \mathbf{D}_\psi(x^*, x_{T+1}) &\leq 2\mathbf{D}_\psi(x^*, x_1) + 2\left(G^2 + 3\sigma^2\right)\eta^2 + 2C\log\left(\frac{1}{\delta}\right) \\ &= 2\mathbf{D}_\psi(x^*, x_1) + 2\left(G^2 + 3\sigma^2\right)\eta^2 + 12\sigma^2\log\left(\frac{1}{\delta}\right)\eta^2 \\ &\leq 2\mathbf{D}_\psi(x^*, x_1) + 12\left(G^2 + \sigma^2\left(1 + \log\left(\frac{1}{\delta}\right)\right)\right)\eta^2 \end{aligned}$$

Thus we have

$$\frac{1}{T}\sum_{t=1}^{T}\left(f(x_t) - f(x^*)\right) \leq \frac{1}{\sqrt{T}}\left(\frac{2\mathbf{D}_\psi(x^*, x_1)}{\eta} + 12\left(G^2 + \sigma^2\left(1 + \log\left(\frac{1}{\delta}\right)\right)\right)\eta\right)$$

and

$$\mathbf{D}_\psi(x^*, x_{T+1}) \leq 2\mathbf{D}_\psi(x^*, x_1) + 12\left(G^2 + \sigma^2\left(1 + \log\left(\frac{1}{\delta}\right)\right)\right)\eta^2.$$

□

The analysis extends to the setting where the T is not known and we use the step sizes $\eta_t = \frac{\eta}{\sqrt{t}}$.

Corollary 4.4.2. *Suppose we run the Stochastic Mirror Descent algorithm with time-varying step sizes $\eta_t = \frac{\eta}{\sqrt{t}}$. Let $w_T = \frac{1}{12\sigma^2\eta^2\left(\sum_{t=1}^{T}\frac{1}{t}\right)}$ and $w_{t-1} = w_t + 6\sigma^2\eta_t^2 w_t^2$ for all $1 \leq t \leq T$. The sequence $\{w_t\}$ satisfies the conditions required by Corollary 4.1.4. By Corollary 4.1.4, for any $\delta > 0$, the following events hold with probability at least $1 - \delta$: $\mathbf{D}_\psi(x^*, x_{T+1}) \leq$*

$2\mathbf{D}_\psi(x^*, x_1) + 12\left(G^2 + \sigma^2\left(1 + \log\left(\frac{1}{\delta}\right)\right)\right)\eta^2(1+\log T)$, *and*

$$\frac{1}{T}\sum_{t=1}^{T}\left(f(x_t) - f(x^*)\right) \leq \frac{1}{\sqrt{T}}\frac{2\mathbf{D}_\psi(x^*, x_1)}{\eta} + \frac{12}{\sqrt{T}}\left(G^2 + \sigma^2\left(1 + \log\left(\frac{1}{\delta}\right)\right)\right)\eta(1+\log T).$$

In particular, setting $\eta_t = \sqrt{\frac{\mathbf{D}_\psi(x^*,x_1)}{6\left(G^2+\sigma^2\left(1+\ln\left(\frac{1}{\delta}\right)\right)\right)t}}$ *we obtain the second case of Theorem 4.1.1.*

Proof of Corollary (4.4.2). Recall from Corollary 4.1.4 that the sequence $\{w_t\}$ needs to satisfy the following conditions for all $1 \leq t \leq T$:

$$w_t + 6\sigma^2\eta_t^2 w_t^2 \leq w_{t-1}$$
$$w_t\eta_t^2 \leq \frac{1}{4\sigma^2}$$

Let $M_t = 6\sigma^2\sum_{i=1}^{t}\eta_i^2$ and $C = M_T = 6\sigma^2\eta^2\left(\sum_{t=1}^{T}\frac{1}{t}\right)$. We set $w_T = \frac{1}{C+M_T}$. For $1 \leq t \leq T$, we set w_t so that the first condition holds with equality

$$w_{t-1} = w_t + 6\sigma^2\eta_t^2 w_t^2$$

We can show by induction that, for every $1 \leq t \leq T$, we have

$$w_t \leq \frac{1}{C + M_t}$$

The base case $t = T$ follows from the definition of w_T. Consider $1 \leq t \leq T$. Using the definition of w_t and the inductive hypothesis, we obtain

$$\begin{aligned} w_{t-1} &= w_t + 6\sigma^2\eta_t^2 w_t^2 \\ &\leq \frac{1}{C+M_t} + \frac{6\sigma^2\eta_t^2}{(C+M_t)^2} \\ &\leq \frac{1}{C+M_t} + \frac{(C+M_t)-(C+M_{t-1})}{(C+M_t)(C+M_{t-1})} \\ &= \frac{1}{C+M_{t-1}} \end{aligned}$$

as needed.

Using this fact, we now show that $\{w_t\}$ satisfies the second condition. For every $1 \leq t \leq T$, we have

$$w_t\eta_t^2 \leq \frac{\eta_t^2}{C} \leq \frac{\eta_t^2}{6\sigma^2\eta_t^2} = \frac{1}{6\sigma^2}$$

as needed.

Thus, by Corollary 4.1.4, with probability $\geq 1-\delta$, we have

$$\sum_{t=1}^{T} w_t\eta_t\left(f(x_t) - f(x^*)\right) + w_T\mathbf{D}_\psi(x^*, x_{T+1}) \leq w_0\mathbf{D}_\psi(x^*, x_1) + \left(G^2 + 3\sigma^2\right)\sum_{t=1}^{T} w_t\eta_t^2 + \log\left(\frac{1}{\delta}\right)$$

Note that $w_T = \frac{1}{2C}$ and $\frac{1}{2C} \leq w_t \leq \frac{1}{C}$ for all $1 \leq t \leq T$. Thus we obtain

$$\frac{1}{2C}\eta_T\sum_{t=1}^{T}\left(f(x_t) - f(x^*)\right) + \frac{1}{2C}\mathbf{D}_\psi(x^*, x_{T+1}) \leq \frac{1}{C}\mathbf{D}_\psi(x^*, x_1) + \left(G^2 + 3\sigma^2\right)\frac{1}{C}\sum_{t=1}^{T}\eta_t^2 + \log\left(\frac{1}{\delta}\right)$$

Plugging in $\eta_t = \frac{\eta}{\sqrt{t}}$ and simplifying, we obtain

$$\begin{aligned}\frac{\eta}{\sqrt{T}}\sum_{t=1}^{T}\left(f\left(x_t\right)-f\left(x^*\right)\right)+&\mathbf{D}_\psi\left(x^*,x_{T+1}\right)\\ &\leq 2\mathbf{D}_\psi\left(x^*,x_1\right)+\left(2G^2+6\sigma^2\right)\eta^2\left(\sum_{t=1}^{T}\frac{1}{t}\right)+2C\log\left(\frac{1}{\delta}\right)\\ &= 2\mathbf{D}_\psi\left(x^*,x_1\right)+\left(2G^2+6\sigma^2\left(1+2\log\left(\frac{1}{\delta}\right)\right)\right)\eta^2\left(\sum_{t=1}^{T}\frac{1}{t}\right)\end{aligned}$$

Thus we have

$$\frac{1}{T}\sum_{t=1}^{T}\left(f\left(x_t\right)-f\left(x^*\right)\right)\leq\frac{1}{\sqrt{T}}\left(\frac{2\mathbf{D}_\psi\left(x^*,x_1\right)}{\eta}+\left(2G^2+6\sigma^2\left(1+2\log\left(\frac{1}{\delta}\right)\right)\right)\eta\left(\sum_{t=1}^{T}\frac{1}{t}\right)\right)$$

and

$$\mathbf{D}_\psi\left(x^*,x_{T+1}\right)\leq 2\mathbf{D}_\psi\left(x^*,x_1\right)+\left(2G^2+6\sigma^2\left(1+2\log\left(\frac{1}{\delta}\right)\right)\right)\eta^2\left(\sum_{t=1}^{T}\frac{1}{t}\right)$$

□

4.4.2 Accelerated Stochastic Mirror Descent

The convergence of Algorithm 2 is given in the following Theorem.

Theorem 4.4.3. *Assume f satisfies Assumptions (1), (2), (3) and condition (4.5), with probability at least $1-\delta$,*

(1) Setting $\eta_t = \min\left\{\frac{t}{4L}, \frac{\sqrt{\mathbf{D}_\psi(x^*,z_0)}t}{\sqrt{6}\sqrt{G^2+\sigma^2\left(1+\log\left(\frac{1}{\delta}\right)\right)}T^{3/2}}\right\}$, *then* $\mathbf{D}_\psi\left(x^*,z_T\right)\leq 4\mathbf{D}_\psi\left(x^*,z_0\right)$ *and*

$$f\left(y_T\right)-f\left(x^*\right)\leq\frac{16L\mathbf{D}_\psi\left(x^*,z_0\right)}{T^2}+\frac{8\sqrt{6}}{\sqrt{T}}\sqrt{\mathbf{D}_\psi\left(x^*,z_0\right)\left(G^2+\left(1+\log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)}.$$

(2) Setting $\eta_t = \min\left\{\frac{t}{4L}, \frac{\sqrt{\mathbf{D}_\psi(x^*,z_0)}}{\sqrt{6}\sqrt{G^2+\sigma^2\left(1+\log\left(\frac{1}{\delta}\right)\right)}t^{1/2}}\right\}$, *then* $\mathbf{D}_\psi\left(x^*,z_T\right)\leq 2(2+\log T)\mathbf{D}_\psi\left(x^*,z_0\right)$ *and*

$$f\left(y_T\right)-f\left(x^*\right)\leq\frac{16L\mathbf{D}_\psi\left(x^*,z_0\right)}{T^2}+\frac{4\sqrt{6}(2+\log T)}{\sqrt{T}}\sqrt{\mathbf{D}_\psi\left(x^*,z_0\right)\left(G^2+\left(1+\log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)}.$$

Proof of Lemma 4.1.5. Starting with smoothness, we obtain

$$
\begin{aligned}
f(y_t) &\leq f(x_t) + \langle \nabla f(x_t), y_t - x_t \rangle + G\|y_t - x_t\| + \frac{\beta}{2}\|y_t - x_t\|^2 \ \forall x \in \mathcal{X} \\
&= f(x_t) + \langle \nabla f(x_t), y_{t-1} - x_t \rangle + \langle \nabla f(x_t), y_t - y_{t-1} \rangle + G\|y_t - x_t\| + \frac{\beta}{2}\|y_t - x_t\|^2 \\
&= (1-\alpha_t)\underbrace{\left(f(x_t) + \langle \nabla f(x_t), y_{t-1} - x_t \rangle\right)}_{\text{convexity}} + \alpha_t \underbrace{\left(f(x_t) + \langle \nabla f(x_t), y_{t-1} - x_t \rangle\right)}_{\text{convexity}} \\
&\quad + \alpha_t \langle \nabla f(x_t), z_t - y_{t-1} \rangle + G\|y_t - x_t\| + \frac{\beta}{2}\|y_t - x_t\|^2 \\
&\leq (1-\alpha_t) f(y_{t-1}) + \alpha_t f(x_t) + \alpha_t \langle \nabla f(x_t), z_t - x_t \rangle + G\underbrace{\|y_t - x_t\|}_{=\alpha_t\|z_t - z_{t-1}\|} + \frac{\beta}{2}\underbrace{\|y_t - x_t\|^2}_{=\alpha_t^2\|z_t - z_{t-1}\|^2} \\
&= (1-\alpha_t) f(y_{t-1}) + \alpha_t f(x_t) + \alpha_t \langle \nabla f(x_t), z_t - x_t \rangle + G\alpha_t\|z_t - z_{t-1}\| + \frac{\beta}{2}\alpha_t^2\|z_t - z_{t-1}\|^2
\end{aligned}
$$

By the optimality condition for z_t,

$$
\eta_t \left\langle \widehat{\nabla} f(x_t), z_t - x^* \right\rangle \leq \langle \nabla_x \mathbf{D}_\psi(z_t, z_{t-1}), x^* - z_t \rangle = \mathbf{D}_\psi(x^*, z_{t-1}) - \mathbf{D}_\psi(z_t, z_{t-1}) - \mathbf{D}_\psi(x^*, z_t)
$$

Rearranging, we obtain

$$
\mathbf{D}_\psi(x^*, z_t) - \mathbf{D}_\psi(x^*, z_{t-1}) + \mathbf{D}_\psi(z_t, z_{t-1}) \leq \eta_t \left\langle \widehat{\nabla} f(x_t), x^* - z_t \right\rangle = \eta_t \langle \nabla f(x_t) + \xi_t, x^* - z_t \rangle
$$

By combining the two inequalities, we obtain

$$
\begin{aligned}
&f(y_t) + \frac{\alpha_t}{\eta_t}\left(\mathbf{D}_\psi(x^*, z_t) - \mathbf{D}_\psi(x^*, z_{t-1}) + \mathbf{D}_\psi(z_t, z_{t-1})\right) \\
&\leq (1-\alpha_t) f(y_{t-1}) + \alpha_t \underbrace{\left(f(x_t) + \langle \nabla f(x_t), x^* - x_t \rangle\right)}_{\text{convexity}} \\
&\quad + G\alpha_t\|z_t - z_{t-1}\| + \frac{\beta}{2}\alpha_t^2\|z_t - z_{t-1}\|^2 + \alpha_t \langle \xi_t, x^* - z_t \rangle \\
&\leq (1-\alpha_t) f(y_{t-1}) + \alpha_t f(x^*) + G\alpha_t\|z_t - z_{t-1}\| + \frac{\beta}{2}\alpha_t^2\|z_t - z_{t-1}\|^2 + \alpha_t \langle \xi_t, x^* - z_t \rangle
\end{aligned}
$$

Subtracting $f(x^*)$ from both sides, rearranging, and using that $\mathbf{D}_\psi(z_t, z_{t-1}) \geq \frac{1}{2}\|z_t - z_{t-1}\|^2$, we obtain

$$
\begin{aligned}
&f(y_t) - f(x^*) + \frac{\alpha_t}{\eta_t}\left(\mathbf{D}_\psi(x^*, z_t) - \mathbf{D}_\psi(x^*, z_{t-1})\right) \\
&\leq (1-\alpha_t)\left(f(y_{t-1}) - f(x^*)\right) + \alpha_t \langle \xi_t, x^* - z_t \rangle + G\alpha_t\|z_t - z_{t-1}\| - \alpha_t \frac{1 - \beta\alpha_t\eta_t}{2\eta_t}\|z_t - z_{t-1}\|^2 \\
&= (1-\alpha_t)\left(f(y_{t-1}) - f(x^*)\right) + \alpha_t \langle \xi_t, x^* - z_{t-1} \rangle + \alpha_t \langle \xi_t, z_t - z_{t-1} \rangle + \\
&\quad G\alpha_t\|z_t - z_{t-1}\| - \alpha_t \frac{1 - \beta\alpha_t\eta_t}{2\eta_t}\|z_t - z_{t-1}\|^2 \\
&\leq (1-\alpha_t)\left(f(y_{t-1}) - f(x^*)\right) + \alpha_t \langle \xi_t, x^* - z_{t-1} \rangle + \alpha_t\|z_t - z_{t-1}\|\left(\|\xi_t\|_* + G\right) - \\
&\quad \alpha_t \frac{1 - \beta\alpha_t\eta_t}{2\eta_t}\|z_t - z_{t-1}\|^2 \\
&\leq (1-\alpha_t)\left(f(y_{t-1}) - f(x^*)\right) + \alpha_t \langle \xi_t, x^* - z_{t-1} \rangle + \frac{\alpha_t\eta_t}{2(1 - \beta\alpha_t\eta_t)}\left(\|\xi_t\|_* + G\right)^2
\end{aligned}
$$

Finally, we divide by $\frac{\alpha_t}{\eta_t}$, and obtain

$$\begin{aligned}
&\frac{\eta_t}{\alpha_t}\left(f\left(y_t\right)-f\left(x^*\right)\right)+\mathbf{D}_\psi\left(x^*, z_t\right)-\mathbf{D}_\psi\left(x^*, z_{t-1}\right)\\
&\leq \frac{\eta_t}{\alpha_t}\left(1-\alpha_t\right)\left(f\left(y_{t-1}\right)-f\left(x^*\right)\right)+\eta_t\left\langle \xi_t, x^*-z_{t-1}\right\rangle+\frac{\eta_t^2}{2\left(1-\beta\alpha_t\eta_t\right)}\left(\left\|\xi_t\right\|_*+G\right)^2\\
&\leq \frac{\eta_t}{\alpha_t}\left(1-\alpha_t\right)\left(f\left(y_{t-1}\right)-f\left(x^*\right)\right)+\eta_t\left\langle \xi_t, x^*-z_{t-1}\right\rangle+\frac{\eta_t^2}{1-\beta\alpha_t\eta_t}\left(\left\|\xi_t\right\|_*^2+G^2\right).
\end{aligned}$$

□

Proof of Theorem 4.1.6. We proceed by induction on t. Consider the base case $t = T+1$, the inequality trivially holds. Next, we consider $t \leq T$. We have

$$\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_t\right]=\mathbb{E}\left[\exp\left(Z_t+S_{t+1}\right) \mid \mathcal{F}_t\right]=\mathbb{E}\left[\mathbb{E}\left[\exp\left(Z_t+S_{t+1}\right) \mid \mathcal{F}_{t+1}\right] \mid \mathcal{F}_t\right] \tag{4.10}$$

We now analyze the inner expectation. Conditioned on $\mathcal{F}_{t+1}$, Z_t is fixed. Using the inductive hypothesis, we obtain

$$\mathbb{E}\left[\exp\left(Z_t+S_{t+1}\right) \mid \mathcal{F}_{t+1}\right] \leq \exp\left(Z_t\right)\exp\left(3\sigma^2\sum_{i=t+1}^{T} w_i\frac{\eta_i^2}{1-L\alpha_i\eta_i}\right) \tag{4.11}$$

Let $X_t = \eta_t\left\langle \xi_t, x^*-z_{t-1}\right\rangle$. By Lemma 4.1.5, we have

$$\begin{aligned}
&\frac{\eta_t}{\alpha_t}\left(f\left(y_t\right)-f\left(x^*\right)\right)-\frac{\eta_t}{\alpha_t}\left(1-\alpha_t\right)\left(f\left(y_{t-1}\right)-f\left(x^*\right)\right)-\frac{\eta_t^2}{1-L\alpha_t\eta_t}G^2\\
&\quad+\mathbf{D}_\psi\left(x^*, z_t\right)-\mathbf{D}_\psi\left(x^*, z_{t-1}\right)\\
&\leq X_t+\frac{\eta_t^2}{\left(1-L\alpha_t\eta_t\right)}\left\|\xi_t\right\|_*^2
\end{aligned}$$

and thus

$$Z_t \leq w_t X_t + w_t\frac{\eta_t^2}{1-L\alpha_t\eta_t}\left\|\xi_t\right\|_*^2 - v_t\mathbf{D}_\psi\left(x^*, z_{t-1}\right)$$

Plugging into (4.11), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(Z_t+S_{t+1}\right) \mid \mathcal{F}_{t+1}\right]\\
&\leq \exp\left(w_t X_t - v_t\mathbf{D}_\psi\left(x^*, z_{t-1}\right)+w_t\frac{\eta_t^2}{1-L\alpha_t\eta_t}\left\|\xi_t\right\|_*^2+3\sigma^2\sum_{i=t+1}^{T} w_i\frac{\eta_i^2}{1-L\alpha_i\eta_i}\right)
\end{aligned}$$

Plugging into (4.10), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_t\right]\\
&\leq \exp\left(-v_t\mathbf{D}_\psi\left(x^*, z_{t-1}\right)+3\sigma^2\sum_{i=t+1}^{T} w_i\frac{\eta_i^2}{1-L\alpha_i\eta_i}\right)\mathbb{E}\left[\exp\left(w_t X_t+w_t\frac{\eta_t^2}{1-L\alpha_t\eta_t}\left\|\xi_t\right\|_*^2\right) \mid \mathcal{F}_t\right]
\end{aligned} \tag{4.12}$$

Next, we analyze the expectation on the RHS of the above inequality. Note that $X_t = \eta_t \langle \xi_t, x^* - z_{t-1} \rangle$ and $\mathbb{E}[X_t \mid \mathcal{F}_t] = 0$. Applying Lemma 4.3.2, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(w_t X_t + w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t} \|\xi_t\|_*^2\right) \mid \mathcal{F}_t\right] \\
&\leq \exp\left(3\left(w_t^2\eta_t^2 \|x^* - z_{t-1}\|^2 + w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t}\right)\sigma^2\right) \\
&\leq \exp\left(3\left(2w_t^2\eta_t^2 \mathbf{D}_\psi(x^*, z_{t-1}) + w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t}\right)\sigma^2\right) \qquad (4.13)
\end{aligned}$$

On the last line we used that $\mathbf{D}_\psi(x^*, z_{t-1}) \geq \frac{1}{2}\|x^* - z_{t-1}\|^2$, which follows from the strong convexity of ψ.

Plugging in (4.13) into (4.12) and using that $v_t = 6\sigma^2 w_t^2 \eta_t^2$, we obtain

$$\mathbb{E}[\exp(S_t) \mid \mathcal{F}_t] \leq \exp\left(3\sigma^2 \sum_{i=t}^{T} w_i \frac{\eta_i^2}{1 - L\alpha_i\eta_i}\right)$$

as needed. □

Proof of Corollary 4.1.7. Let

$$K = 3\sigma^2 \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t} + \log\left(\frac{1}{\delta}\right)$$

By Theorem 4.1.6 and Markov's inequality, we have

$$\begin{aligned}
\Pr[S_1 \geq K] &\leq \Pr[\exp(S_1) \geq \exp(K)] \\
&\leq \exp(-K)\,\mathbb{E}[\exp(S_1)] \\
&\leq \exp(-K)\exp\left(3\sigma^2 \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t}\right) \\
&= \delta
\end{aligned}$$

Note that since $v_t + w_t \leq w_{t-1}$

$$\begin{aligned}
S_1 &= \sum_{t=1}^{T} Z_t \\
&= \sum_{t=1}^{T} w_t \left(\frac{\eta_t}{\alpha_t} \left(f\left(y_t\right) - f\left(x^*\right) \right) - \frac{\eta_t \left(1-\alpha_t\right)}{\alpha_t} \left(f\left(y_{t-1}\right) - f\left(x^*\right) \right) \right) \\
&\quad + \sum_{t=1}^{T} w_t \mathbf{D}_\psi\left(x^*, z_t\right) - \left(v_t + w_t\right) \mathbf{D}_\psi\left(x^*, z_{t-1}\right) - G^2 \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t} \\
&\geq \sum_{t=1}^{T} w_t \left(\frac{\eta_t}{\alpha_t} \left(f\left(y_t\right) - f\left(x^*\right) \right) - \frac{\eta_t \left(1-\alpha_t\right)}{\alpha_t} \left(f\left(y_{t-1}\right) - f\left(x^*\right) \right) \right) \\
&\quad + \sum_{t=1}^{T} w_t \mathbf{D}_\psi\left(x^*, z_t\right) - w_{t-1} \mathbf{D}_\psi\left(x^*, z_{t-1}\right) - G^2 \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t} \\
&= \sum_{t=1}^{T} w_t \left(\frac{\eta_t}{\alpha_t} \left(f\left(y_t\right) - f\left(x^*\right) \right) - \frac{\eta_t \left(1-\alpha_t\right)}{\alpha_t} \left(f\left(y_{t-1}\right) - f\left(x^*\right) \right) \right) \\
&\quad + w_T \mathbf{D}_\psi\left(x^*, z_T\right) - w_0 \mathbf{D}_\psi\left(x^*, z_0\right) - G^2 \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t}
\end{aligned}$$

Therefore, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\sum_{t=1}^{T} w_t \left(\frac{\eta_t}{\alpha_t} \left(f\left(y_t\right) - f\left(x^*\right) \right) - \frac{\eta_t \left(1-\alpha_t\right)}{\alpha_t} \left(f\left(y_{t-1}\right) - f\left(x^*\right) \right) \right) + w_T \mathbf{D}_\psi\left(x^*, z_T\right) \\
&\leq w_0 \mathbf{D}_\psi\left(x^*, z_0\right) + \left(G^2 + 3\sigma^2\right) \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t} + \log\left(\frac{1}{\delta}\right).
\end{aligned}$$

□

Corollary 4.4.4. *Suppose we run the Accelerated Stochastic Mirror Descent algorithm with the standard choices $\alpha_t = \frac{2}{t+1}$ and $\eta_t = \eta t$ with $\eta \leq \frac{1}{4L}$. Let $w_T = \frac{1}{3\sigma^2\eta^2 T(T+1)(2T+1)}$ and $w_{t-1} = w_t + 6\sigma^2\eta_t^2 w_t^2$ for all $1 \leq t \leq T$. The sequence $\{w_t\}_{0\leq t\leq T}$ satisfies the conditions required by Corollary 4.1.7. By Corollary 4.1.7, with probability at least $1-\delta$, $\mathbf{D}_\psi\left(x^*, z_T\right) \leq 2\mathbf{D}_\psi\left(x^*, z_0\right) + 12\left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2 T^3$ and*

$$f\left(y_T\right) - f\left(x^*\right) \leq \frac{4\mathbf{D}_\psi\left(x^*, z_0\right)}{\eta T^2} + 24\left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta T.$$

In particular, setting $\eta = \min\left\{\frac{1}{4L}, \frac{\sqrt{\mathbf{D}_\psi\left(x^*, z_0\right)}}{\sqrt{6}\sqrt{G^2 + \sigma^2\left(1 + \log\left(\frac{1}{\delta}\right)\right)}T^{3/2}}\right\}$, we obtain the first case of Theorem 4.4.3.

Proof of Corollary 4.4.4. Recall from Corollary 4.1.7 that the sequence $\{w_t\}$ needs to satisfy the following conditions:

$$w_t + 6\sigma^2\eta_t^2 w_t^2 \leq w_{t-1} \quad \forall 1 \leq t \leq T \tag{4.14}$$

$$\frac{w_t \eta_t^2}{1 - L\alpha_t\eta_t} \leq \frac{1}{4\sigma^2} \quad \forall 0 \leq t \leq T \tag{4.15}$$

We will set $\{w_t\}$ so that it satisfies the following additional condition, which will allow us to telescope the sum on the RHS of Corollary 4.1.7:

$$w_{t-1}\frac{\eta_{t-1}}{\alpha_{t-1}} \geq w_t\frac{\eta_t\left(1-\alpha_t\right)}{\alpha_t} \quad \forall 1 \leq t \leq T \tag{4.16}$$

Given w_T, we set w_{t-1} for every $1 \leq t \leq T$ so that the first condition (4.14) holds with equality:

$$w_{t-1} = w_t + 6\sigma^2\eta_t^2 w_t^2 = w_t + 6\sigma^2\eta^2 t^2 w_t^2$$

Let $C = \sigma^2\eta^2 T\left(T+1\right)\left(2T+1\right)$. We set

$$w_T = \frac{1}{C + 6\sigma^2\eta^2\sum_{i=1}^{T} i^2} = \frac{1}{C + \sigma^2\eta^2 T\left(T+1\right)\left(2T+1\right)} = \frac{1}{2\sigma^2\eta^2 T\left(T+1\right)\left(2T+1\right)}$$

Given this choice for w_T, we now verify that, for all $0 \leq t \leq T$, we have

$$w_t \leq \frac{1}{C + 6\sigma^2\eta^2\sum_{i=1}^{t} i^2} = \frac{1}{C + \sigma^2\eta^2 t\left(t+1\right)\left(2t+1\right)}$$

We proceed by induction on t. The base case $t = T$ follows from the definition of w_T. Consider $t \leq T$. Using the definition of w_{t-1} and the inductive hypothesis, we obtain

$$\begin{aligned}
w_{t-1} &= w_t + 6\sigma^2\eta^2 t^2 w_t^2 \\
&\leq \frac{1}{C + 6\sigma^2\eta^2\sum_{i=1}^{t} i^2} + \frac{6\sigma^2\eta^2 t^2}{\left(C + 6\sigma^2\eta^2\sum_{i=1}^{t} i^2\right)^2} \\
&\leq \frac{1}{C + 6\sigma^2\eta^2\sum_{i=1}^{t} i^2} + \frac{\left(C + 6\sigma^2\eta^2\sum_{i=1}^{t} i^2\right) - \left(C + 6\sigma^2\eta^2\sum_{i=1}^{t-1} i^2\right)}{\left(C + 6\sigma^2\eta^2\sum_{i=1}^{t} i^2\right)\left(C + 6\sigma^2\eta^2\sum_{i=1}^{t-1} i^2\right)} \\
&= \frac{1}{C + 6\sigma^2\eta^2\sum_{i=1}^{t-1} i^2}
\end{aligned}$$

as needed.

Let us now verify that the second condition (4.15) also holds. Using that $\frac{2t}{t+1} \leq 2$, $L\eta \leq \frac{1}{4}$, and $T \geq 2$, we obtain

$$\begin{aligned}
\frac{w_t\eta_t^2}{1 - L\alpha_t\eta_t} = \frac{w_t\eta^2 t^2}{1 - L\eta\frac{2t}{t+1}} &\leq 2w_t\eta^2 t^2 \leq \frac{2\eta^2 t^2}{C + 6\sigma^2\eta^2 t^2} \\
&= \frac{t^2}{\sigma^2 T\left(T+1\right)\left(2T+1\right) + 3\sigma^2 t^2} \\
&\leq \frac{1}{\sigma^2\left(2T+1\right) + 3\sigma^2} \leq \frac{1}{4\sigma^2}
\end{aligned}$$

as needed.

Let us now verify that the third condition (4.16) also holds. Since $\eta_t = \eta t$ and $\alpha_t = \frac{2}{t+1}$, we have $\frac{\eta_{t-1}}{\alpha_{t-1}} = \frac{\eta_t(1-\alpha_t)}{\alpha_t} = \frac{\eta t(t-1)}{2}$. Since $w_t \leq w_{t-1}$, it follows that condition (4.16) holds.

We now turn our attention to the convergence. By Corollary 4.1.7, with probability $\geq 1-\delta$, we have

$$\sum_{t=1}^{T} w_t \left(\frac{\eta_t}{\alpha_t} \left(f\left(y_t\right) - f\left(x^*\right) \right) - \frac{\eta_t \left(1-\alpha_t\right)}{\alpha_t} \left(f\left(y_{t-1}\right) - f\left(x^*\right) \right) \right) + w_T \mathbf{D}_\psi\left(x^*, z_T\right)$$
$$\leq w_0 \mathbf{D}_\psi\left(x^*, z_0\right) + \left(G^2 + 3\sigma^2\right) \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t} + \log\left(\frac{1}{\delta}\right)$$

Grouping terms on the LHS and using that $\alpha_1 = 1$, we obtain

$$\sum_{t=1}^{T-1} \left(w_t \frac{\eta_t}{\alpha_t} - w_{t+1} \frac{\eta_{t+1}\left(1-\alpha_{t+1}\right)}{\alpha_{t+1}} \right) \left(f\left(y_t\right) - f\left(x^*\right) \right) + w_T \frac{\eta_T}{\alpha_T} \left(f\left(y_T\right) - f\left(x^*\right) \right) + w_T \mathbf{D}_\psi\left(x^*, z_T\right)$$
$$\leq w_0 \mathbf{D}_\psi\left(x^*, z_0\right) + \left(G^2 + 3\sigma^2\right) \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t} + \log\left(\frac{1}{\delta}\right)$$

Since $\{w_t\}$ satisfies condition (4.16), the coefficient of $f\left(y_t\right) - f\left(x^*\right)$ is non-negative and thus we can drop the above sum. We obtain

$$w_T \frac{\eta_T}{\alpha_T} \left(f\left(y_T\right) - f\left(x^*\right) \right) + w_T \mathbf{D}_\psi\left(x^*, z_T\right)$$
$$\leq w_0 \mathbf{D}_\psi\left(x^*, z_0\right) + \left(G^2 + 3\sigma^2\right) \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t} + \log\left(\frac{1}{\delta}\right)$$

Using that $w_T = \frac{1}{2C}$ and $w_t \leq \frac{1}{C}$ for all $0 \leq t \leq T-1$, we obtain

$$\frac{1}{2C} \frac{\eta_T}{\alpha_T} \left(f\left(y_T\right) - f\left(x^*\right) \right) + \frac{1}{2C} \mathbf{D}_\psi\left(x^*, z_T\right)$$
$$\leq \frac{1}{C} \mathbf{D}_\psi\left(x^*, z_0\right) + \frac{1}{C} \left(G^2 + 3\sigma^2\right) \sum_{t=1}^{T} \frac{\eta_t^2}{1 - L\alpha_t\eta_t} + \log\left(\frac{1}{\delta}\right).$$

Thus

$$\frac{\eta_T}{\alpha_T} \left(f\left(y_T\right) - f\left(x^*\right) \right) + \mathbf{D}_\psi\left(x^*, z_T\right)$$
$$\leq 2\mathbf{D}_\psi\left(x^*, z_0\right) + 2\left(G^2 + 3\sigma^2\right) \sum_{t=1}^{T} \frac{\eta_t^2}{1 - L\alpha_t\eta_t} + 2C \log\left(\frac{1}{\delta}\right)$$
$$= 2\mathbf{D}_\psi\left(x^*, z_0\right) + 2\left(G^2 + 3\sigma^2\right) \sum_{t=1}^{T} \frac{\eta_t^2}{1 - L\alpha_t\eta_t} + 2\sigma^2 \log\left(\frac{1}{\delta}\right) \eta^2 T\left(T+1\right)\left(2T+1\right).$$

Using that $L\eta \leq \frac{1}{4}$ and $\frac{2t}{t+1} \leq 2$, we obtain

$$\sum_{t=1}^{T} \frac{\eta_t^2}{1 - L\alpha_t\eta_t} = \sum_{t=1}^{T} \frac{\eta^2 t^2}{1 - L\eta\frac{2t}{t+1}} \leq \sum_{t=1}^{T} 2\eta^2 t^2 = \frac{1}{3}\eta^2 T\left(T+1\right)\left(2T+1\right)$$

Plugging in and using that $\eta_T = \eta T$ and $\alpha_T = \frac{2}{T+1}$, we obtain

$$\begin{aligned}
&\eta \frac{T(T+1)}{2}\left(f(y_T) - f(x^*)\right) + \mathbf{D}_\psi(x^*, z_T) \\
&\leq 2\mathbf{D}_\psi(x^*, z_0) + \left(\frac{2}{3}G^2 + 2\left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2 T(T+1)(2T+1) \\
&\leq 2\mathbf{D}_\psi(x^*, z_0) + 2\left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2 T(T+1)(2T+1).
\end{aligned}$$

We can further simplify the bound by lower bounding $T(T+1) \geq T^2$ and upper bounding $T(T+1)(2T+1) \leq 6T^3$. We obtain

$$\eta T^2\left(f(y_T) - f(x^*)\right) + 2\mathbf{D}_\psi(x^*, z_T) \leq 4\mathbf{D}_\psi(x^*, z_0) + 24\left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2 T^3.$$

Thus we obtain

$$f(y_T) - f(x^*) \leq \frac{4\mathbf{D}_\psi(x^*, z_0)}{\eta T^2} + 24\left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta T,$$

and

$$\mathbf{D}_\psi(x^*, z_T) \leq 2\mathbf{D}_\psi(x^*, z_0) + 12\left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2 T^3.$$

□

Corollary 4.4.5. *Suppose we run the Accelerated Stochastic Mirror Descent algorithm with the standard choices* $\alpha_t = \frac{2}{t+1}$ *and* $\eta_t = \min\left\{\frac{t}{4L}, \frac{\eta}{\sqrt{t}}\right\}$. *Let* $w_T = \frac{1}{12\sigma^2 \sum_{i=1}^T \eta_t^2}$ *and* $w_{t-1} = w_t + 6\sigma^2\eta_t^2 w_t^2$ *for all* $1 \leq t \leq T$. *The sequence* $\{w_t\}_{0 \leq t \leq T}$ *satisfies the conditions required by Corollary 4.1.7. By Corollary 4.1.7, with probability at least* $1 - \delta$, $\mathbf{D}_\psi(x^*, z_T) \leq 2\mathbf{D}_\psi(x^*, z_0) + 12\left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2(1 + \log T)$ *and*

$$\begin{aligned}
&f(y_T) - f(x^*) \\
&\leq \frac{16L}{T^2}\mathbf{D}_\psi(x^*, z_0) \\
&\quad + \frac{2}{T^{1/2}\eta}\left(2\mathbf{D}_\psi(x^*, z_0) + 12\left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2(1 + \log T)\right).
\end{aligned}$$

In particular, setting $\eta_t = \min\left\{\frac{t}{4L}, \frac{\sqrt{\mathbf{D}_\psi(x^*, z_0)}}{\sqrt{6}\sqrt{G^2 + \sigma^2\left(1 + \log\left(\frac{1}{\delta}\right)\right)}\, t^{1/2}}\right\}$, *we obtain the second case of Theorem 4.4.3.*

Proof of Corollary 4.4.5. Recall from Corollary 4.1.7 that the sequence $\{w_t\}$ needs to satisfy the following conditions:

$$w_t + 6\sigma^2\eta_t^2 w_t^2 \leq w_{t-1} \quad \forall 1 \leq t \leq T \tag{4.17}$$

$$\frac{w_t \eta_t^2}{1 - L\alpha_t\eta_t} \leq \frac{1}{4\sigma^2} \quad \forall 0 \leq t \leq T \tag{4.18}$$

We will set $\{w_t\}$ so that it satisfies the following additional condition, which will allow us to telescope the sum on the RHS of Corollary 4.1.7:

$$w_{t-1}\frac{\eta_{t-1}}{\alpha_{t-1}} \geq w_t \frac{\eta_t\left(1-\alpha_t\right)}{\alpha_t} \quad \forall 1 \leq t \leq T-1 \tag{4.19}$$

Given w_T, we set w_{t-1} for every $1 \leq t \leq T$ so that the first condition (4.17) holds with equality:

$$w_{t-1} = w_t + 6\sigma^2\eta_t^2 w_t^2 = w_t + 6\sigma^2\eta^2 t^2 w_t^2$$

Let $C = 6\sigma^2\sum_{i=1}^T \eta_t^2$. We set

$$w_T = \frac{1}{12\sigma^2\sum_{i=1}^T \eta_t^2} = \frac{1}{2C}$$

Given this choice for w_T, we now verify that, for all $0 \leq t \leq T$, we have

$$w_t \leq \frac{1}{C + 6\sigma^2\sum_{i=1}^t \eta_i^2}$$

We proceed by induction on t. The base case $t = T$ follows from the definition of w_T. Consider $t \leq T$. Using the definition of w_{t-1} and the inductive hypothesis, we obtain

$$\begin{aligned}
w_{t-1} &= w_t + 6\sigma^2\eta_t^2 w_t^2 \\
&\leq \frac{1}{C + 6\sigma^2\sum_{i=1}^t \eta_i^2} + \frac{6\sigma^2\eta_t^2}{\left(C + 6\sigma^2\sum_{i=1}^t \eta_i^2\right)^2} \\
&\leq \frac{1}{C + 6\sigma^2\sum_{i=1}^t \eta_i^2} + \frac{\left(C + 6\sigma^2\sum_{i=1}^t \eta_i^2\right) - \left(C + 6\sigma^2\sum_{i=1}^{t-1} \eta_i^2\right)}{\left(C + 6\sigma^2\sum_{i=1}^t \eta_i^2\right)\left(C + 6\sigma^2\sum_{i=1}^{t-1} \eta_i^2\right)} \\
&= \frac{1}{C + 6\sigma^2\sum_{i=1}^{t-1} \eta_i^2}
\end{aligned}$$

as needed.

Let us now verify that the second condition (4.18) also holds. Using that $L\eta_t \leq \frac{t}{4}$, and $T \geq 2$, we obtain

$$\frac{w_t\eta_t^2}{1 - L\alpha_t\eta_t} \leq \frac{w_t\eta_t^2}{1 - \frac{t}{4}\frac{2}{t+1}} \leq 2w_t\eta_t^2 \leq \frac{2\eta_t^2}{6\sigma^2\sum_{i=1}^T \eta_t^2 + 6\sigma^2\sum_{i=1}^t \eta_i^2} \leq \frac{2\eta_t^2}{12\sigma^2\eta_t^2} \leq \frac{1}{4\sigma^2}$$

as needed.

Let us now verify that the third condition (4.19) also holds. Since $\alpha_t = \frac{2}{t+1}$, we have

$$\begin{aligned}
\frac{\eta_{t-1}}{\alpha_{t-1}} &= \frac{\eta_{t-1}t}{2} \\
\frac{\eta_t\left(1-\alpha_t\right)}{\alpha_t} &= \frac{\eta_t\left(t-1\right)}{2}
\end{aligned}$$

If $\eta_{t-1} = \frac{t-1}{4L}$ then we have $\eta_t \leq \frac{t}{4L}$ and $\frac{\eta_t(1-\alpha_t)}{\alpha_t} \leq \frac{\eta_{t-1}}{\alpha_{t-1}} = \frac{t(t-1)}{8L}$. If $\eta_{t-1} = \frac{\eta}{\sqrt{t-1}}$ then $\eta_t = \frac{\eta}{\sqrt{t}}$,we also have $\frac{\eta_t(1-\alpha_t)}{\alpha_t} \leq \frac{\eta_{t-1}}{\alpha_{t-1}}$. Since $w_t \leq w_{t-1}$, it follows that condition (4.19) holds.

We now turn our attention to the convergence. By Corollary 4.1.7, with probability $\geq 1-\delta$, we have

$$\sum_{t=1}^{T} w_t \left(\frac{\eta_t}{\alpha_t} \left(f\left(y_t\right) - f\left(x^*\right) \right) - \frac{\eta_t \left(1-\alpha_t\right)}{\alpha_t} \left(f\left(y_{t-1}\right) - f\left(x^*\right) \right) \right) + w_T \mathbf{D}_\psi\left(x^*, z_T\right)$$
$$\leq w_0 \mathbf{D}_\psi\left(x^*, z_0\right) + \left(G^2 + 3\sigma^2\right) \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t} + \log\left(\frac{1}{\delta}\right).$$

Grouping terms on the LHS and using that $\alpha_1 = 1$, we obtain

$$\sum_{t=1}^{T-1} \left(w_t \frac{\eta_t}{\alpha_t} - w_{t+1} \frac{\eta_{t+1}\left(1-\alpha_{t+1}\right)}{\alpha_{t+1}} \right) \left(f\left(y_t\right) - f\left(x^*\right) \right)$$
$$+ w_T \frac{\eta_T}{\alpha_T} \left(f\left(y_T\right) - f\left(x^*\right) \right) + w_T \mathbf{D}_\psi\left(x^*, z_T\right)$$
$$\leq w_0 \mathbf{D}_\psi\left(x^*, z_0\right) + \left(G^2 + 3\sigma^2\right) \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t} + \log\left(\frac{1}{\delta}\right).$$

Since $\{w_t\}$ satisfies condition (4.19), the coefficient of $f(y_t) - f(x^*)$ is non-negative and thus we can drop the above sum. We obtain

$$w_T \frac{\eta_T}{\alpha_T} \left(f\left(y_T\right) - f\left(x^*\right) \right) + w_T \mathbf{D}_\psi\left(x^*, z_T\right)$$
$$\leq w_0 \mathbf{D}_\psi\left(x^*, z_0\right) + \left(G^2 + 3\sigma^2\right) \sum_{t=1}^{T} w_t \frac{\eta_t^2}{1 - L\alpha_t\eta_t} + \log\left(\frac{1}{\delta}\right).$$

Using that $w_T = \frac{1}{2C}$ and $w_t \leq \frac{1}{C}$ for all $0 \leq t \leq T-1$, we obtain

$$\frac{1}{2C} \frac{\eta_T}{\alpha_T} \left(f\left(y_T\right) - f\left(x^*\right) \right) + \frac{1}{2C} \mathbf{D}_\psi\left(x^*, z_T\right)$$
$$\leq \frac{1}{C} \mathbf{D}_\psi\left(x^*, z_0\right) + \frac{1}{C} \left(G^2 + 3\sigma^2\right) \sum_{t=1}^{T} \frac{\eta_t^2}{1 - L\alpha_t\eta_t} + \log\left(\frac{1}{\delta}\right)$$

Thus

$$\frac{\eta_T}{\alpha_T} \left(f\left(y_T\right) - f\left(x^*\right) \right) + \mathbf{D}_\psi\left(x^*, z_T\right)$$
$$\leq 2\mathbf{D}_\psi\left(x^*, z_0\right) + 2\left(G^2 + 3\sigma^2\right) \sum_{t=1}^{T} \frac{\eta_t^2}{1 - L\alpha_t\eta_t} + 2C \log\left(\frac{1}{\delta}\right)$$

Using that $L\eta_t \leq \frac{t}{4}$, we obtain

$$\sum_{t=1}^{T} \frac{\eta_t^2}{1 - L\alpha_t\eta_t} = \sum_{t=1}^{T} \frac{\eta_t^2}{1 - \frac{t}{4}\frac{2}{t+1}} \leq \sum_{t=1}^{T} 2\eta_t^2 = \frac{C}{3\sigma^2}$$

Plugging in and using that $\eta_T = \eta T$ and $\alpha_T = \frac{2}{T+1}$, we obtain

$$\frac{\eta_T \left(T+1\right)}{2} \left(f\left(y_T\right) - f\left(x^*\right) \right) + \mathbf{D}_\psi\left(x^*, z_T\right)$$
$$\leq 2\mathbf{D}_\psi\left(x^*, z_0\right) + \left(2G^2 + 6\left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2 \right) \frac{C}{3\sigma^2}$$

If $\frac{T}{4L} \leq \frac{\eta}{\sqrt{T}}$ which means $T^{3/2} \leq 4L\eta$ then $\eta_T = \frac{T}{4L}$ we have

$$C = 6\sigma^2 \sum_{i=1}^{T} \eta_t^2 = \frac{6\sigma^2}{16L^2} \sum_{i=1}^{T} t^2 \leq \frac{3\sigma^2 T^3}{8L^2} \leq 6\sigma^2\eta^2$$

Hence

$$\begin{aligned}
&\frac{\eta_T (T+1)}{2} \left(f(y_T) - f(x^*)\right) + \mathbf{D}_\psi(x^*, z_T) \\
&\leq 2\mathbf{D}_\psi(x^*, z_0) + \left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right) \frac{3T^3}{4L^2}
\end{aligned}$$

which entails

$$\begin{aligned}
f(y_T) - f(x^*) &\leq \frac{16L}{T^2}\mathbf{D}_\psi(x^*, z_0) + \left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right) \frac{6T}{L} \\
&= \frac{16L}{T^2}\mathbf{D}_\psi(x^*, z_0) + \frac{6}{\sqrt{T}} \left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right) \frac{T^{3/2}}{L} \\
&\leq \frac{16L}{T^2}\mathbf{D}_\psi(x^*, z_0) + \frac{24}{\sqrt{T}} \left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right) \eta
\end{aligned}$$

and

$$\mathbf{D}_\psi(x^*, z_T) \leq 2\mathbf{D}_\psi(x^*, z_0) + 12\left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2$$

If $\frac{\eta}{\sqrt{T}} \leq \frac{T}{4L}$ then $\eta_T = \frac{\eta}{\sqrt{T}}$. Let T_0 be the largest t such that $\frac{\eta}{\sqrt{t}} \geq \frac{t}{4L}$, we have $T_0^3 \leq 16L^2\eta^2$

$$\begin{aligned}
C &= 6\sigma^2 \sum_{i=1}^{T} \eta_t^2 \\
&= 6\sigma^2 \sum_{i=1}^{T_0} \eta_t^2 + 6\sigma^2 \sum_{i=T_0+1}^{T} \eta_t^2 \\
&= \frac{6\sigma^2}{16L^2} \sum_{i=1}^{T_0} t^2 + 6\sigma^2\eta^2 \sum_{i=T_0+1}^{T} \frac{1}{t} \\
&\leq \frac{6\sigma^2}{16L^2} T_0^3 + 6\sigma^2\eta^2 \sum_{i=T_0+1}^{T} \frac{1}{t} \\
&\leq 6\sigma^2\eta^2 \sum_{i=1}^{T} \frac{1}{t} \leq 6\sigma^2\eta^2(1 + \log T)
\end{aligned}$$

Hence

$$f(y_T) - f(x^*) \leq \frac{2}{T^{1/2}\eta} \left(2\mathbf{D}_\psi(x^*, z_0) + 12\left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2(1 + \log T)\right)$$

and

$$\mathbf{D}_\psi(x^*, z_T) \leq 2\mathbf{D}_\psi(x^*, z_0) + 12\left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2(1 + \log T)$$

Overall we have

$$f(y_T) - f(x^*) \leq \frac{16L}{T^2}\mathbf{D}_\psi(x^*, z_0) + \frac{2}{T^{1/2}\eta}\left(2\mathbf{D}_\psi(x^*, z_0) + 12\left(G^2 + \left(1 + \log\left(\frac{1}{\delta}\right)\right)\sigma^2\right)\eta^2(1 + \log T)\right).$$

□

4.5 Missing Proofs from Section 4.2

Proof of Lemma 4.2.2. We start from the smoothness of f

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ &= -\eta_t\left\langle \nabla f(x_t), \widehat{\nabla} f(x_t)\right\rangle + \frac{L\eta_t^2}{2}\left\|\widehat{\nabla} f(x_t)\right\|^2. \end{aligned}$$

By writing $\widehat{\nabla} f(x_t) = \xi_t + \nabla f(x_t)$ we have

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq -\eta_t\langle \nabla f(x_t), \xi_t + \nabla f(x_t)\rangle + \frac{L\eta_t^2}{2}\|\xi_t + \nabla f(x_t)\|^2 \\ &= -\eta_t\|\nabla f(x_t)\|^2 - \eta_t\langle \nabla f(x_t), \xi_t\rangle \\ &\quad + \frac{L\eta_t^2}{2}\|\xi_t\|^2 + \frac{L\eta_t^2}{2}\|\nabla f(x_t)\|^2 + L\eta_t^2\langle \nabla f(x_t), \xi_t\rangle. \end{aligned}$$

We obtain the inequality (4.7) by rearranging the terms. □

Proof of Theorem 4.2.3. We prove by induction. The base case $t = T + 1$ trivially holds. Consider $1 \leq t \leq T$, we have

$$\begin{aligned} \mathbb{E}[\exp(S_t) \mid \mathcal{F}_t] &= \mathbb{E}[\mathbb{E}[\exp(Z_t + S_{t+1}) \mid \mathcal{F}_{t+1}] \mid \mathcal{F}_t] \\ &= \mathbb{E}[\exp(Z_t)\,\mathbb{E}[\exp(S_{t+1}) \mid \mathcal{F}_{t+1}] \mid \mathcal{F}_k] \end{aligned}$$

From the induction hypothesis we have $\mathbb{E}[\exp(S_{t+1}) \mid \mathcal{F}_{t+1}] \leq \exp\left(3\sigma^2\sum_{i=t+1}^T \frac{w_i\eta_i^2 L}{2}\right)$, hence

$$\mathbb{E}[\exp(S_t) \mid \mathcal{F}_t] \leq \exp\left(3\sigma^2\sum_{i=t+1}^T \frac{w_i\eta_i^2 L}{2}\right)\mathbb{E}[\exp(Z_t) \mid \mathcal{F}_t]$$

We have then

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(Z_t\right) \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\exp\left(w_t\left(\eta_t\left(1-\frac{\eta_t L}{2}\right)\left\|\nabla f(x_t)\right\|^2+\Delta_{t+1}-\Delta_t\right)-v_t\left\|\nabla f(x_T)\right\|^2\right) \mid \mathcal{F}_t\right] \\
&\leq \mathbb{E}\left[\exp\left(w_t\left(\eta_t(\eta_t L-1)\left\langle\nabla f(x_t), \xi_t\right\rangle+\frac{\eta_t^2 L}{2}\left\|\xi_t\right\|^2\right)-v_t\left\|\nabla f(x_t)\right\|^2\right) \mid \mathcal{F}_t\right] \\
&= \exp\left(-v_t\left\|\nabla f(x_t)\right\|^2\right) \mathbb{E}\left[\exp\left(w_t\left(\eta_t(\eta_t L-1)\left\langle\nabla f(x_t), \xi_t\right\rangle+\frac{\eta_t^2 L}{2}\left\|\xi_t\right\|^2\right)\right) \mid \mathcal{F}_t\right] \\
&\leq \exp\left(-v_t\left\|\nabla f(x_t)\right\|^2\right) \exp\left(3\sigma^2\left(w_t^2\eta_t^2(\eta_t L-1)^2\left\|\nabla f(x_t)\right\|^2+\frac{w_t\eta_t^2 L}{2}\right)\right) \\
&= \exp\left(3\sigma^2\frac{w_t\eta_t^2 L}{2}\right).
\end{aligned}$$

where the second line is due to (4.7) in Lemma 4.2.2 and the second to last line is due to Lemma 4.3.2.Therefore

$$\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_t\right] \leq \exp\left(3\sigma^2\sum_{i=t}^{T}\frac{w_i\eta_i^2 L}{2}\right)$$

which we what we need to show. □

Proof of Corollary 4.2.4. In Lemma 4.2.3, Let $t=1$ we obtain

$$\mathbb{E}\left[\exp\left(S_1\right)\right] \leq \exp\left(3\sigma^2\sum_{t=1}^{T}\frac{w_t\eta_t^2 L}{2}\right)$$

hence by Markov's inequality we have

$$\Pr\left[S_1 \geq \left(3\sigma^2\sum_{t=1}^{T}\frac{w_t\eta_t^2 L}{2}\right)+\log\frac{1}{\delta}\right] \leq \delta$$

In other words, with probability $\geq 1-\delta$ (once the condition in Lemma 4.2.3 is satisfied)

$$\begin{aligned}
&\sum_{t=1}^{T}\left[w_t\eta_t\left(1-\frac{\eta_t L}{2}\right)-v_t\right]\left\|\nabla f(x_t)\right\|^2+w_t\left(\Delta_{t+1}-\Delta_t\right) \\
&\leq 3\sigma^2\sum_{t=1}^{T}\frac{w_t\eta_t^2 L}{2}+\log\frac{1}{\delta}
\end{aligned}$$

This gives

$$\begin{aligned}
&\sum_{t=1}^{T}\left[w_t\eta_t\left(1-\frac{\eta_t L}{2}\right)-v_t\right]\left\|\nabla f(x_t)\right\|^2+w_T\Delta_{T+1} \\
&\qquad\leq w_1\Delta_1+\left(\sum_{t=2}^{T}(w_t-w_{t-1})\Delta_t+3\sigma^2\sum_{t=1}^{T}\frac{w_t\eta_t^2 L}{2}\right)+\log\frac{1}{\delta}
\end{aligned}$$

as needed. □

Proof of Theorem 4.2.1 . **First case.**

Starting from this inequality, we will specify the choice of η_t and w_t to obtain the bound. Consider $\eta_t = \eta$ with $\eta L \leq 1$, $w_t = w = \frac{1}{6\sigma^2\eta}$. Note that $w_t\eta_t^2 L = \frac{\eta L}{6\sigma^2} \leq \frac{1}{2\sigma^2}$ satisfies the condition of Lemma 4.2.3, we have

$$\begin{aligned}\text{LHS of (4.9)} &= w\Delta_{T+1} + \sum_{t=1}^{T}\left[w\eta\left(1-\frac{\eta L}{2}\right) - 3\sigma^2 w^2\eta^2(\eta L-1)^2\right]\|\nabla f(x_t)\|^2 \\ &= w\Delta_{T+1} + w\eta\sum_{t=1}^{T}\left[1-\frac{\eta L}{2}-\frac{1}{2}(\eta L-1)^2\right]\|\nabla f(x_t)\|^2 \\ &\geq w\Delta_{T+1} + \frac{w\eta}{2}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2,\end{aligned}$$

where the last inequality is due to $1-\frac{\eta L}{2}-\frac{(1-\eta L)^2}{2} \geq \frac{1}{2}$ when $0 \leq \eta L \leq 1$. Besides,

$$\text{RHS of (4.9)} = w\Delta_1 + \frac{3\sigma^2}{2}w\eta^2 LT + \log\frac{1}{\delta}.$$

Hence with probability $\geq 1-\delta$

$$\begin{aligned}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2 + \frac{2\Delta_{T+1}}{\eta} &\leq \frac{2\Delta_1}{\eta} + 3\sigma^2\eta LT + \frac{2}{w\eta}\log\frac{1}{\delta} \\ &= \frac{2\Delta_1}{\eta} + 3\sigma^2\eta LT + 12\sigma^2\log\frac{1}{\delta}.\end{aligned}$$

Finally by choosing $\eta = \min\left\{\frac{1}{L}; \sqrt{\frac{\Delta_1}{\sigma^2 LT}}\right\}$ and noticing $\Delta_{T+1} \geq 0$, we obtain the desired inequality.

Second case.

Consider $\eta_t = \frac{\eta}{\sqrt{t}}$ with $\eta L \leq 1$, $w_t = w = \frac{1}{6\sigma^2\eta}$. Again, we have $w_t\eta_t^2 L = \frac{\eta L}{6\sigma^2 t} \leq \frac{1}{2\sigma^2}$, then

$$\begin{aligned}\text{LHS of (4.9)} &= \sum_{t=1}^{T}\left[\frac{w\eta}{\sqrt{t}}\left(1-\frac{\eta L}{2\sqrt{t}}\right) - \frac{3\sigma^2 w^2\eta^2}{t}\left(1-\frac{\eta L}{\sqrt{t}}\right)^2\right]\|\nabla f(x_t)\|^2 + w\Delta_{T+1} \\ &= \sum_{t=1}^{T}\frac{w\eta}{\sqrt{t}}\left[1-\frac{\eta L}{2\sqrt{t}} - \frac{3\sigma^2 w\eta}{\sqrt{t}}\left(1-\frac{\eta L}{\sqrt{t}}\right)^2\right]\|\nabla f(x_t)\|^2 + w\Delta_{T+1} \\ &\geq \sum_{t=1}^{T}\frac{w\eta}{\sqrt{t}}\left[1-\frac{\eta L}{2\sqrt{t}} - 3\sigma^2 w\eta\left(1-\frac{\eta L}{\sqrt{t}}\right)^2\right]\|\nabla f(x_t)\|^2 + w\Delta_{T+1} \\ &= \sum_{t=1}^{T}\frac{w\eta}{\sqrt{t}}\left[1-\frac{\eta L}{2\sqrt{t}} - \frac{1}{2}\left(1-\frac{\eta L}{\sqrt{t}}\right)^2\right]\|\nabla f(x_t)\|^2 + w\Delta_{T+1} \\ &\geq \sum_{t=1}^{T}\frac{w\eta}{2\sqrt{t}}\|\nabla f(x_t)\|^2 + w\Delta_{T+1} \geq \frac{w\eta}{2\sqrt{T}}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2 + w\Delta_{T+1}\end{aligned}$$

where the second inequality is due to $1 - \frac{\eta L}{2\sqrt{t}} - \frac{1}{2}\left(1 - \frac{\eta L}{\sqrt{t}}\right)^2 \geq \frac{1}{2}$ when $0 \leq \frac{\eta L}{\sqrt{t}} \leq 1$. Besides,

$$\begin{aligned}\text{RHS of (4.9)} =& w\Delta_1 + \frac{3\sigma^2}{2} w\eta^2 L \sum_{t=1}^{T} \frac{1}{t} + \log\frac{1}{\delta} \\ \leq& w\Delta_1 + \frac{3\sigma^2}{2} w\eta^2 L(1+\log T) + \log\frac{1}{\delta}\end{aligned}$$

Therefore with probability $\geq 1-\delta$

$$\begin{aligned}&\sum_{t=1}^{T} \|\nabla f(x_t)\|^2 + \frac{2\sqrt{T}\Delta_{T+1}}{\eta} \\ \leq& \sqrt{T}\left(\frac{2\Delta_1}{\eta} + 3\sigma^2\eta L\,(1+\log T) + \frac{2}{w\eta}\log\frac{1}{\delta}\right) \\ =& \sqrt{T}\left(\frac{2\Delta_1}{\eta} + 3\sigma^2\eta L\,(1+\log T) + 12\sigma^2\log\frac{1}{\delta}\right)\end{aligned}$$

Choose $\eta = \frac{1}{L}$, and notice $\Delta_{T+1} \geq 0$, we obtain

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \leq \frac{2\Delta_1 L + 3\sigma^2\,(1+\log T) + 12\sigma^2\log\frac{1}{\delta}}{\sqrt{T}}.$$

□

4.6 AdaGrad-Norm Omitted Proofs

We first provide the proofs for some of the Lemmas in Section 4.2.2.

Proof of Lemma 4.2.7. We start by using the smoothness of f

$$\begin{aligned}&f(x_{t+1}) - f(x_t) \qquad (4.20)\\ &\leq \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ &= -\frac{\eta}{b_t}\left\langle \nabla f(x_t), \widehat{\nabla} f(x_t)\right\rangle + \frac{L\eta^2}{2b_t^2}\left\|\widehat{\nabla} f(x_t)\right\|^2 \\ &= -\frac{\eta}{b_t}\|\nabla f(x_t)\|^2 - \frac{\eta}{b_t}\langle \nabla f(x_t), \xi_t\rangle + \frac{L\eta^2}{2b_t^2}\left\|\widehat{\nabla} f(x_t)\right\|^2 \\ &= \eta\left(\frac{1}{a_t} - \frac{1}{b_t}\right)\langle \nabla f(x_t), \xi_t\rangle - \frac{\eta}{a_t}\langle \nabla f(x_t), \xi_t\rangle - \frac{\eta}{b_t}\|\nabla f(x_t)\|^2 + \frac{L\eta^2}{2b_t^2}\left\|\widehat{\nabla} f(x_t)\right\|^2 \qquad (4.21)\end{aligned}$$

First, by Lemma 4.6.2, we have

$$\left|\frac{1}{a_t} - \frac{1}{b_t}\right| \leq \frac{\|\xi_t\|}{a_t b_t}.$$

This gives

$$\begin{aligned}\left(\frac{1}{a_t}-\frac{1}{b_t}\right)\langle\nabla f(x_t),\xi_t\rangle &\le \left|\frac{1}{a_t}-\frac{1}{b_t}\right|\|\nabla f(x_t)\|\,\|\xi_t\| \\ &\le \frac{\|\xi_t\|}{a_t b_t}\|\nabla f(x_t)\|\,\|\xi_t\| \\ &\le \|\xi_t\|\left(\frac{\|\nabla f(x_t)\|^2}{2a_t^2}+\frac{\|\xi_t\|^2}{2b_t^2}\right).\end{aligned}$$

Plugging this back into 4.21, we have

$$f(x_{t+1})-f(x_t)\le \eta\|\xi_t\|\left(\frac{\|\nabla f(x_t)\|^2}{2a_t^2}+\frac{\|\xi_t\|^2}{2b_t^2}\right)-\frac{\eta\langle\nabla f(x_t),\xi_t\rangle}{a_t}-\frac{\eta}{b_t}\|\nabla f(x_t)\|^2+\frac{L\eta^2}{2b_t^2}\left\|\widehat{\nabla} f(x_t)\right\|^2$$

After summing up, rearranging the terms, using $\|\xi_t\|\le M_T$ and $f(x_1)-f(x_{T+1})\le \Delta_1$, we obtain

$$\sum_{t=1}^T\frac{\|\nabla f(x_t)\|^2}{b_t}\le\frac{\Delta_1}{\eta}+\frac{M_T}{2}\left[\sum_{t=1}^T\frac{\|\nabla f(x_t)\|^2}{a_t^2}+\sum_{t=1}^T\frac{\|\xi_t\|^2}{b_t^2}\right]-\sum_{t=1}^T\frac{\langle\nabla f(x_t),\xi_t\rangle}{a_t}+\sum_{t=1}^T\frac{L\eta}{2b_t^2}\left\|\widehat{\nabla} f(x_t)\right\|^2.$$

□

Proof of Lemma 4.2.8. By Lemma 4.3.2 with some $w>0$, we have

$$\mathbb{E}\left[\exp\left(\left\langle -w\frac{\nabla f(x_t),\xi_t}{a_t}\right\rangle-2\sigma^2w^2\frac{\|\nabla f(x_t)\|^2}{a_t^2}\right)\mid\mathcal{F}_t\right]\le 1.$$

Thus it is not difficult to verify that

$$\mathbb{E}\left[\exp\left(\sum_{t=1}^T\left\langle -w\frac{\nabla f(x_t),\xi_t}{a_t}\right\rangle-2\sigma^2w^2\frac{\|\nabla f(x_t)\|^2}{a_t^2}\right)\right]\le 1.$$

By Markov's inequality we obtain, with probability at least $1-\delta$,

$$\sum_{t=1}^T-\frac{\langle\nabla f(x_t),\xi_t\rangle}{a_t}\le 2\sigma^2w\sum_{t=1}^T\frac{\|\nabla f(x_t)\|^2}{a_t^2}+\frac{1}{w}\log\frac{1}{\delta}.$$

It is also known that with probability at least $1-\delta$, $M_T\le\sigma\sqrt{1+\log\frac{T}{\delta}}\le 2\sigma\sqrt{\log\frac{T}{\delta}}$ Li and Orabona (2020) and Liu et al. (2022) for $T\ge 1$ and $\delta\in(0,1)$. Thus by a union bound and setting $w:=\frac{\sqrt{\log\frac{1}{\delta}}}{\sigma}$, we can bound Lemma 4.2.7 with probability at least

$1-2\delta$

$$\begin{aligned}
&\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{b_t}\\
&\quad\le \frac{\Delta_1}{\eta}+M_T\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{2a_t^2}+M_T\sum_{t=1}^{T}\frac{\|\xi_t\|^2}{2b_t^2}-\sum_{t=1}^{T}\frac{\langle\nabla f(x_t),\xi_t\rangle}{a_t}+\sum_{t=1}^{T}\frac{L\eta}{2b_t^2}\left\|\widehat{\nabla} f(x_t)\right\|^2\\
&\quad\le \frac{\Delta_1}{\eta}+\sigma\sqrt{\log\frac{T}{\delta}}\left[\underbrace{2\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{a_t^2}}_{B}+\sum_{t=1}^{T}\frac{\|\xi_t\|^2}{b_t^2}\right]+\sigma\sqrt{\log\frac{1}{\delta}}+\frac{L\eta}{2}\underbrace{\sum_{t=1}^{T}\frac{\left\|\widehat{\nabla} f(x_t)\right\|^2}{b_t^2}}_{A}.
\end{aligned}$$

Let us consider the term A. We have

$$\begin{aligned}
\sum_{t=1}^{T}\frac{\left\|\widehat{\nabla} f(x_t)\right\|^2}{b_t^2}=\sum_{t=1}^{T}\frac{b_t^2-b_{t-1}^2}{b_t^2}&=\sum_{t=1}^{T}1-\frac{b_{t-1}^2}{b_t^2}\\
&\le 2\sum_{t=1}^{T}\log\frac{b_t}{b_{t-1}}=2\log\frac{b_T}{b_0}.
\end{aligned}$$

For B, note that since $\|\nabla f(x_t)\|^2\le 2\left\|\widehat{\nabla} f(x_t)\right\|^2+2\|\xi_t\|^2$, we have

$$\begin{aligned}
\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{a_t^2}&=\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{b_{t-1}^2+\|\nabla f(x_t)\|^2}\\
&\overset{(*)}{\le}\sum_{t=1}^{T}\frac{2\left\|\widehat{\nabla} f(x_t)\right\|^2+2\|\xi_t\|^2}{b_{t-1}^2+2\left\|\widehat{\nabla} f(x_t)\right\|^2+2\|\xi_t\|^2}\\
&=\sum_{t=1}^{T}\frac{2\left\|\widehat{\nabla} f(x_t)\right\|^2}{b_{t-1}^2+2\left\|\widehat{\nabla} f(x_t)\right\|^2+2\|\xi_t\|^2}+\sum_{t=1}^{T}\frac{2\|\xi_t\|^2}{b_{t-1}^2+2\left\|\widehat{\nabla} f(x_t)\right\|^2+2\|\xi_t\|^2}\\
&\le 2\sum_{t=1}^{T}\frac{\left\|\widehat{\nabla} f(x_t)\right\|^2}{b_t^2}+2\sum_{t=1}^{T}\frac{\|\xi_t\|^2}{b_t^2}\\
&\le 4\log\left(\frac{b_T}{b_0}\right)+2\sum_{t=1}^{T}\frac{\|\xi_t\|^2}{b_t^2}.
\end{aligned}$$

For $(*)$ we use the fact that $\frac{x}{c+x}$ is an increasing function. Combining the bound for A and B, we obtain, with probability at least $1-2\delta$,

$$\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{b_t}\le\frac{\Delta_1}{\eta}+\sigma\sqrt{\log\frac{T}{\delta}}\left[8\log\left(\frac{b_T}{b_0}\right)+5\sum_{t=1}^{T}\frac{\|\xi_t\|^2}{b_t^2}\right]+\sigma\sqrt{\log\frac{1}{\delta}}+L\eta\log\frac{b_T}{b_0}.$$

□

Lemma 4.6.1. *For AdaGrad-Norm stepsizes b_t, if f is L-smooth and the stochastic gradients have σ-subgaussian noise, then with probability at least $1-\delta$*

$$\begin{aligned} b_T &\leq 4b_0 + 4\frac{\Delta_1}{\eta} + \frac{32}{\eta^2 b_0}\sigma^2 \ln\left(\frac{2}{\delta}\right) + \frac{16\sigma}{\eta^2}\sqrt{T + \log\frac{2}{\delta}} + 4L\eta\log\frac{L\eta}{b_0} \\ &= O\left(\Delta_1 + \sigma\sqrt{T} + \sigma^2\log\frac{1}{\delta} + L\log L\right). \end{aligned}$$

Proof. We start from function value analysis

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ &= \frac{1}{b_t}\left\langle \widehat{\nabla} f(x_t), \xi_t\right\rangle + \eta\left(\frac{L\eta}{2b_t^2} - \frac{1}{2b_t}\right)\left\|\widehat{\nabla} f(x_t)\right\|^2 - \frac{\eta}{2b_t}\left\|\widehat{\nabla} f(x_t)\right\|^2. \end{aligned}$$

We can bound $\sum_{t=1}^T \left(\frac{L\eta}{2b_t^2} - \frac{1}{2b_t}\right)\left\|\widehat{\nabla} f(x_t)\right\|^2$ via a standard argument. Let $\tau = \max\{t \leq T \mid b_t \leq \eta L\}$ so that $t \geq \tau$ implies $b_t \geq \eta L \iff \frac{L\eta}{b_t^2} \leq \frac{1}{b_t}$. Then

$$\begin{aligned} \sum_{t=1}^T \left(\frac{L\eta}{2b_t^2} - \frac{1}{2b_t}\right)\left\|\widehat{\nabla} f(x_t)\right\|^2 &\leq \sum_{t=1}^\tau \left(\frac{L\eta}{2b_t^2} - \frac{1}{2b_t}\right)\left\|\widehat{\nabla} f(x_t)\right\|^2 \\ &\leq \frac{L\eta}{2}\sum_{t=1}^\tau \frac{1}{b_t^2}\left\|\widehat{\nabla} f(x_t)\right\|^2 \\ &= L\eta\log\frac{b_\tau}{b_0} \leq L\eta\log\frac{L\eta}{b_0}. \end{aligned}$$

Summing and plugging in the above gives

$$\begin{aligned} f(x_{T+1}) - f(x_1) &\leq \sum_{t=1}^T \frac{1}{b_t}\left\langle \widehat{\nabla} f(x_t), \xi_t\right\rangle + L\eta^2\log\frac{L\eta}{b_0} - \frac{\eta}{2}\sum_{t=1}^T \frac{\left\|\widehat{\nabla} f(x_t)\right\|^2}{b_t} \\ &\leq \frac{1}{\eta}\sum_{t=1}^T \frac{\|\xi_t\|^2}{b_t} - \frac{\eta}{4}\sum_{t=1}^T \frac{\left\|\widehat{\nabla} f(x_t)\right\|^2}{b_t} + L\eta^2\log\frac{L\eta}{b_0}, \end{aligned}$$

where we use $\frac{1}{b_t}\left\langle \widehat{\nabla} f(x_t), \xi_t\right\rangle \leq \frac{\|\xi_t\|^2}{\eta b_t} + \frac{\eta\left\|\widehat{\nabla} f(x_t)\right\|^2}{4b_t}$ in the second inequality. Rearranging and dividing by η, we get

$$\sum_{t=1}^T \frac{\left\|\widehat{\nabla} f(x_t)\right\|^2}{4b_t} \leq \frac{f(x_1) - f(x_{T+1})}{\eta} + \frac{1}{\eta^2}\sum_{t=1}^T \frac{\|\xi_t\|^2}{b_t} + L\eta\log\frac{L\eta}{b_0}.$$

On the LHS, we have

$$\sum_{t=1}^T \frac{\left\|\widehat{\nabla} f(x_t)\right\|^2}{b_t} = \sum_{t=1}^T \frac{b_t^2 - b_{t-1}^2}{b_t} \geq \sum_{t=1}^T b_t - \frac{b_{t-1}^2}{b_{t-1}} = b_T - b_0.$$

Combining this with Lemma 4.6.5 where $\sum_{t=1}^T \frac{\|\xi_t\|^2}{b_t} \leq \frac{8}{b_0}\sigma^2\ln\left(\frac{2}{\delta}\right) - b_0 + 4\sigma\sqrt{T + \log\frac{2}{\delta}}$, we get the result. □

Now, we can prove Theorem 4.2.5.

Proof of Theorem 4.2.5. From Lemma 4.2.8, we have with probability at least $1-2\delta$

$$\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{b_t}\leq\frac{\Delta_1}{\eta}+\sigma\sqrt{\log\frac{T}{\delta}}\left[8\log\left(\frac{b_T}{b_0}\right)+5\sum_{t=1}^{T}\frac{\|\xi_t\|^2}{b_t^2}\right]+\sigma\sqrt{\log\frac{1}{\delta}}+L\eta\log\frac{b_T}{b_0}.$$

Since b_t is increasing, we have $\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{b_t}\geq\sum_{t=1}^{T}\frac{\|\nabla f(x_t)\|^2}{b_T}$. That means

$$\sum_{t=1}^{T}\|\nabla f(x_t)\|^2\leq b_T\left[\frac{\Delta_1}{\eta}+\sigma\sqrt{\log\frac{T}{\delta}}\left[8\log\left(\frac{b_T}{b_0}\right)+5\sum_{t=1}^{T}\frac{\|\xi_t\|^2}{b_t^2}\right]+\sigma\sqrt{\log\frac{1}{\delta}}+L\eta\log\left(\frac{b_T}{b_0}\right)\right].$$

Combining this with the event from Lemma 4.6.6 that bounds $\sum_{t=1}^{T}\frac{\|\xi_t\|^2}{b_t^2}$ and Lemma 4.6.1 that bounds b_T gives us the Theorem. □

4.6.1 Additional Helper Lemmas

Lemma 4.6.2. *For $t\geq 1$ and a_t, ξ_t defined in Lemma 4.2.7, we have*

$$\left|\frac{1}{a_t}-\frac{1}{b_t}\right|\leq\frac{\|\xi_t\|}{a_t b_t}.$$

Proof. We have

$$\begin{aligned}\left|\frac{1}{a_t}-\frac{1}{b_t}\right|&=\left|\frac{b_t-a_t}{a_t b_t}\right|\\&=\left|\frac{b_t^2-a_t^2}{a_t b_t\left(b_t+a_t\right)}\right|\\&=\left|\frac{b_t^2-b_{t-1}^2-\|\nabla f(x_t)\|^2}{a_t b_t\left(b_t+a_t\right)}\right|\\&=\left|\frac{\left\|\widehat{\nabla} f(x_t)\right\|^2-\|\nabla f(x_t)\|^2}{a_t b_t\left(b_t+a_t\right)}\right|\\&\leq\left|\frac{\left(\left\|\widehat{\nabla} f(x_t)\right\|-\|\nabla f(x_t)\|\right)\left(\left\|\widehat{\nabla} f(x_t)\right\|+\|\nabla f(x_t)\|\right)}{a_t b_t\left(b_t+a_t\right)}\right|.\end{aligned}$$

Since $b_t=\sqrt{b_{t-1}^2+\left\|\widehat{\nabla} f(x_t)\right\|^2}\geq\left\|\widehat{\nabla} f(x_t)\right\|$ and $a_t=\sqrt{b_{t-1}^2+\|\nabla f(x_t)\|^2}\geq\|\nabla f(x_t)\|$, we have

$$\begin{aligned}\left|\frac{1}{a_t}-\frac{1}{b_t}\right|&\leq\left|\frac{\left\|\widehat{\nabla} f(x_t)\right\|-\|\nabla f(x_t)\|}{a_t b_t}\right|\\&\leq\frac{\left\|\widehat{\nabla} f(x_t)-\nabla f(x_t)\right\|}{a_t b_t}\\&=\frac{\|\xi_t\|}{a_t b_t}.\end{aligned}$$

□

Lemma 4.6.3. *With prob $\geq 1-\delta$, for any $0 \leq t \leq T$, we have*

$$\sum_{s=1}^{t} \|\xi_s\|^2 \leq \sum_{s=1}^{t} \|\widehat{\nabla} f(x_s)\|^2 + 4\sigma^2 \log \frac{1}{\delta}.$$

Proof. Note that

$$\begin{aligned} \|\widehat{\nabla} f(x_t)\|^2 &= \|\nabla f(x_t)\|^2 + 2\langle \xi_t, \nabla f(x_t)\rangle + \|\xi_t\|^2 \\ \Rightarrow \|\nabla f(x_t)\| - \|\widehat{\nabla} f(x_t)\|^2 + \|\xi_t\|^2 &= 2\langle \xi_t, \nabla f(x_t)\rangle. \end{aligned}$$

Define for $t \in \{0, 1, \cdots, T\}$

$$U_{t+1} = \exp\left(\sum_{s=1}^{t} w_s \left(\|\nabla f(x_s)\|^2 - \|\widehat{\nabla} f(x_s)\|^2 + \|\xi_s\|^2\right) - v_s \|\nabla f(x_s)\|^2\right); \quad v_s = 4\sigma^2 w_s^2.$$

Let $\mathcal{F}_t = \sigma(\xi_{i \leq t-1})$. We know $U_t \in \mathcal{F}_t$. Note that U_t is a supermartingale

$$\begin{aligned} \mathbb{E}\left[U_{t+1} \mid \mathcal{F}_t\right] &= U_t \exp\left(-v_t \|\nabla f(x_t)\|^2\right) \mathbb{E}\left[\exp\left(2w_t \langle \xi_t, \nabla f(x_t)\rangle\right) \mid \mathcal{F}_t\right] \\ &\leq U_t \exp\left(-v_t \|\nabla f(x_t)\|^2\right) \mathbb{E}\left[\exp\left(4\sigma^2 w_t^2 \|\nabla f(x_t)\|^2\right) \mid \mathcal{F}_t\right] \\ &= U_t \end{aligned}$$

By Doob's supermartingale inequality, there is

$$\Pr\left[\max_{t \in [T+1]} U_t \geq \delta^{-1}\right] \leq \delta \mathbb{E}\left[U_1\right] = \delta$$

which implies w.p. $\geq 1-\delta$, $\forall 0 \leq t \leq T$

$$\begin{aligned} &\sum_{s=1}^{t} w_s \left(\|\nabla f(x_s)\|^2 - \|\widehat{\nabla} f(x_s)\|^2 + \|\xi_s\|^2\right) - v_s \|\nabla f(x_s)\|^2 \leq \log \frac{1}{\delta} \\ &\sum_{s=1}^{t} \left(w_s - 4\sigma^2 w_s^2\right) \|\nabla f(x_s)\|^2 + w_s \|\xi_s\|^2 \leq \sum_{s=1}^{t} w_s \|\widehat{\nabla} f(x_s)\|^2 + \log \frac{1}{\delta}. \end{aligned}$$

Set $w_s = \frac{1}{4\sigma^2}$ to get

$$\sum_{s=1}^{t} \|\xi_s\|^2 \leq \sum_{s=1}^{t} \|\widehat{\nabla} f(x_s)\|^2 + 4\sigma^2 \log \frac{1}{\delta}.$$

□

Lemma 4.6.4. *With probability $\geq 1-\delta$, we have*

$$\sum_{t=1}^{T} \|\xi_t\|^2 \leq \sigma^2 T + \sigma^2 \log \frac{1}{\delta}.$$

Proof. Note that

$$\begin{aligned}\Pr\left[\sum_{t=1}^{T}\|\xi_t\|^2 \geq u\right] &= \Pr\left[\exp\left(\sum_{t=1}^{T}\frac{\|\xi_t\|^2}{\sigma^2}\right) \geq \exp\left(\frac{u}{\sigma^2}\right)\right] \\ &\leq \frac{\mathbb{E}\left[\exp\left(\sum_{t=1}^{T}\frac{\|\xi_t\|^2}{\sigma^2}\right)\right]}{\exp\left(\frac{u}{\sigma^2}\right)} \\ &\leq \frac{\exp(T)}{\exp\left(\frac{u}{\sigma^2}\right)}\end{aligned}$$

where we choose

$$u = \sigma^2 T + \sigma^2 \log\frac{1}{\delta}.$$

□

Lemma 4.6.5. *For AdaGrad stepsize b_t and σ-subgaussian noise $\|\xi_t\|$, with probability at least $1-\delta$*

$$\sum_{t=1}^{T}\frac{\|\xi_t\|^2}{b_t} \leq \frac{8}{b_0}\sigma^2\ln\left(\frac{2}{\delta}\right) - b_0 + 4\sigma\sqrt{T + \log\frac{2}{\delta}}.$$

Proof. First, Lemma 4.6.3 gives that with probability at least $1-\delta$, for all $t \leq T$

$$\begin{aligned}\sum_{i=1}^{t}\|\xi_i\|^2 &\leq \sum_{i=1}^{t}\left\|\widehat{\nabla} f(x_i)\right\|^2 + 4\sigma^2\ln\left(\frac{1}{\delta}\right) \\ &= b_t^2 - b_0^2 + 4\sigma^2\ln\left(\frac{1}{\delta}\right) \\ \implies b_t^2 &\geq \sum_{i=1}^{t}\|\xi_i\|^2 - \underbrace{\left[4\sigma^2\ln\left(\frac{1}{\delta}\right) - b_0^2\right]}_{=:C}.\end{aligned}$$

This means that

$$b_t \geq \max\left\{b_0, \sqrt{\left(\sum_{i=1}^{t}\|\xi_i\|^2 - C\right)^+}\right\}.$$

Let $\tau = \max\left(\{0\} \cup \left\{t \in \mathbb{N}_{\leq T} \mid \sum_{i=1}^{t} \|\xi_i\|^2 \leq 2C\right\}\right)$. Then

$$\begin{aligned}
\sum_{t=1}^{T} \frac{1}{b_t} \|\xi_t\|^2 &\leq \sum_{t=1}^{\tau} \frac{1}{b_t} \|\xi_t\|^2 + \sum_{t=\tau+1}^{T} \frac{1}{b_t} \|\xi_t\|^2 \\
&\leq \frac{1}{b_0} \sum_{t=1}^{\tau} \|\xi_t\|^2 + \sum_{t=\tau+1}^{T} \frac{\|\xi_t\|^2}{\max\left\{b_0, \sqrt{\sum_{i=1}^{t} \|\xi_i\|^2 - C}\right\}} \\
&\leq \frac{2C}{b_0} + \sum_{t=\tau+1}^{T} \frac{\|\xi_t\|^2}{\sqrt{\sum_{i=1}^{t} \|\xi_i\|^2 - C}} \\
&\leq \frac{2C}{b_0} + \sum_{t=\tau+1}^{T} \frac{\|\xi_t\|^2}{\sqrt{\frac{1}{2}\sum_{i=1}^{t} \|\xi_i\|^2}} \qquad \text{(since)} \qquad \left(\textstyle\sum_{i=1}^{t} \|\xi_i\|^2 > 2C \text{ for } t > \tau\right) \\
&\leq \frac{2C}{b_0} + 2\sum_{t=1}^{T} \frac{\|\xi_t\|^2}{\sqrt{\sum_{i=1}^{t} \|\xi_i\|^2}} \\
&\leq \frac{2C}{b_0} + 4\sqrt{\sum_{t=1}^{T} \|\xi_t\|^2}.
\end{aligned}$$

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

$$\sum_{t=1}^{T} \frac{\|\xi_t\|^2}{b_t} \leq \frac{8}{b_0}\sigma^2 \ln\left(\frac{1}{\delta}\right) - b_0 + 4\sqrt{\sum_{t=1}^{T} \|\xi_t\|^2}.$$

Combining with Lemma 4.6.4, we get with probability at least $1 - 2\delta$

$$\sum_{t=1}^{T} \frac{\|\xi_t\|^2}{b_t} \leq \frac{8}{b_0}\sigma^2 \ln\left(\frac{1}{\delta}\right) - b_0 + 4\sigma\sqrt{T + \log\frac{1}{\delta}}.$$

□

Lemma 4.6.6. *For AdaGrad-Norm stepsize b_t and σ-subgaussian noise $\|\xi_t\|$, with probability at least $1 - \delta$,*

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\|\xi_t\|^2}{b_t^2} &\leq \frac{4\sigma^2}{b_0^2} \log\left(\frac{2}{\delta}\right) + 2\log\left(1 + \sigma^2 T + \sigma^2 \log\frac{2}{\delta}\right) \\
&= O\left(\sigma^2 \log\left(\frac{1}{\delta}\right) + \log\left(1 + \sigma^2 T + \sigma^2 \log\frac{1}{\delta}\right)\right).
\end{aligned}$$

Proof. Lemma 4.6.3 gives that with probability at least $1 - \delta$

$$\begin{aligned}\sum_{i=1}^{t}\|\xi_i\|^2 &\le \sum_{i=1}^{t}\left\|\widehat{\nabla} f(x_i)\right\|^2 + 4\sigma^2\ln\left(\frac{1}{\delta}\right)\\ &= b_t^2 - b_0^2 + 4\sigma^2\ln\left(\frac{1}{\delta}\right)\\ \implies b_t^2 &\ge \sum_{i=1}^{t}\|\xi_i\|^2 - \underbrace{\left[4\sigma^2\ln\left(\frac{1}{\delta}\right) - b_0^2\right]}_{=:C}.\end{aligned}$$

Let $\tau = \max\left(\{0\}\cup\left\{t\in\mathbb{N}_{\le T} \mid \sum_{i=1}^{t}\|\xi_i\|^2 \le 2C\right\}\right)$. We have

$$\begin{aligned}\sum_{t=1}^{T}\frac{\|\xi_t\|^2}{b_t^2} &\le \sum_{t=1}^{\tau}\frac{\|\xi_t\|^2}{b_t^2} + \sum_{t=\tau+1}^{T}\frac{\|\xi_t\|^2}{b_t^2}\\ &\le \frac{1}{b_0^2}\sum_{t=1}^{\tau}\|\xi_t\|^2 + \sum_{t=\tau+1}^{T}\frac{\|\xi_t\|^2}{\sum_{i=1}^{t}\|\xi_i\|^2 - C}\\ &\qquad\qquad \left(\quad\left(\text{since } \textstyle\sum_{i=1}^{t}\|\xi_i\|^2 > 2C \text{ for } t>\tau\right)\right)\\ &\le \frac{2C}{b_0^2} + 2\sum_{t=\tau+1}^{T}\frac{\|\xi_t\|^2}{\sum_{i=1}^{t}\|\xi_i\|^2}\\ &\le \frac{2C}{b_0^2} + 2\sum_{t=1}^{T}\frac{\|\xi_t\|^2}{\sum_{i=1}^{t}\|\xi_i\|^2}\\ &\le \frac{2C}{b_0^2} + 2 + 2\log\left(1+\sum_{t=1}^{T}\|\xi_t\|^2\right)\\ &= \frac{4\sigma^2}{b_0^2}\log\left(\frac{1}{\delta}\right) + 2\log\left(1+\sum_{t=1}^{T}\|\xi_t\|^2\right).\end{aligned}$$

Then, we can combine this with Lemma 4.6.4 to get that with probability at least $1-2\delta$

$$\sum_{t=1}^{T}\frac{\|\xi_t\|^2}{b_t^2} \le \frac{4\sigma^2}{b_0^2}\log\left(\frac{1}{\delta}\right) + 2\log\left(1+\sigma^2 T + \sigma^2\log\frac{1}{\delta}\right).$$

Replacing δ with $\delta/2$ yields the result. □

4.7 AdaGrad (Coordinate) Analysis

In this section, we show that our same technique can be generalized to the standard (per-coordinate) version of AdaGrad. The analysis is analogous to our AdaGrad-norm analysis but with the coordinates taken into account.

4.7.1 Preliminaries and notations

Let $d\in\mathbb{N}$ be the dimension of the problem. We let v_i denote the i-th coordinate of a vector $v\in\mathbb{R}^d$. If a vector like x_t is already indexed as part of a sequence of vectors (where x_t denotes the t-th update) then we use $x_{t,i}$ to denote x_t's i-th coordinate. For gradients, we let $\nabla_i f(x) := \frac{\partial f}{\partial x_i}$ denote the partial derivative wrt the i-th coordinate. Similarly, for stochastic gradients $\widehat{\nabla} f(x)$, we let $\widehat{\nabla}_i f(x)$ denotes its i-th coordinate.

For simplicity, in our analysis, we will use $\widehat{\nabla}_{t,i} := \widehat{\nabla}_i f(x_t)$ and $\nabla_{t,i} := \nabla_i f(x_t)$ to denote the i-th coordinate of the stochastic gradients and gradients at iterate t, respectively. If $a, b \in \mathbb{R}^d$, then ab and a/b denotes coordinate-wise multiplication and division, respectively i.e. $(ab)_i = a_i b_i$ and $(a/b)_i = a_i/b_i$.

If we denote the noise as $\xi_t := \widehat{\nabla} f(x_t) - \nabla f(x_t)$ and $\xi_{t,i}$ as the i-th coordinate of ξ_t, then we assume the noise is per-coordinate sub-gaussian i.e. there exists $\sigma_i > 0$ for $i \in [d]$ such that ξ_t satisfies

$$\mathbb{E}\left[\exp\left(\lambda^2 \xi_{t,i}^2\right)\right] \leq \exp\left(\lambda^2 \sigma_i^2\right), \forall\, |\lambda| \leq \frac{1}{\sigma_i}, \forall i \in [d].$$

Note that $\|\xi_t\|$ being σ-subgaussian implies that each $\xi_{t,i}$ is also σ-subgaussian, thus the assumption above is more general.

4.7.2 Analysis

Similarly to our Adagrad-norm analysis in Section 4.6, we define a proxy step size a_t that replaces the stochastic gradient at time t with the true gradient: $a_{t,i}^2 := b_{t-1,i}^2 + \nabla_{t,i}^2$, for $i \in [d]$. First, we present an analogous starting point to Lemma 4.2.7 in the Lemma below:

Lemma 4.7.1. *For $t \geq 1$, let $\xi_{t,i} = \widehat{\nabla}_{t,i} - \nabla_{t,i}$, $a_{t,i}^2 := b_{t-1,i}^2 + \nabla_{t,i}^2$ and $M_{t,i} = \max_{j \leq t} |\xi_{j,i}|$, then we have*

$$\sum_{t=1}^{T}\sum_{i=1}^{d} \frac{\nabla_{t,i}^2}{b_{t,i}} \leq \frac{\Delta_1}{\eta} - \sum_{t=1}^{T}\sum_{i=1}^{d} \frac{\nabla_{t,i}\xi_{t,i}}{a_{t,i}} + \sum_{t=1}^{T}\sum_{i=1}^{d} |\xi_{t,i}| \left[\frac{\nabla_{t,i}^2}{2a_{t,i}^2} + \frac{\xi_{t,i}^2}{2b_{t,i}^2}\right] + \frac{\eta L}{2}\sum_{t=1}^{T}\sum_{i=1}^{d} \frac{\widehat{\nabla}_{t,i}^2}{b_{t,i}^2}.$$

Proof. We start with the smoothness of f

$$\begin{aligned}
\Delta_{t+1} - \Delta_t &\leq \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\
&= -\eta \sum_{i=1}^{d} \frac{\nabla_{t,i}\widehat{\nabla}_{t,i}}{b_{t,i}} + \frac{\eta^2 L}{2}\sum_{i=1}^{d} \frac{\widehat{\nabla}_{t,i}^2}{b_{t,i}^2} \\
&= -\eta \sum_{i=1}^{d} \frac{\nabla_{t,i}^2}{b_{t,i}} - \eta \sum_{i=1}^{d} \frac{\nabla_{t,i}\xi_{t,i}}{b_{t,i}} + \frac{\eta^2 L}{2}\sum_{i=1}^{d} \frac{\widehat{\nabla}_{t,i}^2}{b_{t,i}^2} \\
&= -\eta \sum_{i=1}^{d} \frac{\nabla_{t,i}^2}{b_{t,i}} - \eta \sum_{i=1}^{d} \frac{\nabla_{t,i}\xi_{t,i}}{a_{t,i}} + \eta \sum_{i=1}^{d} \left(\frac{1}{a_{t,i}} - \frac{1}{b_{t,i}}\right)\nabla_{t,i}\xi_{t,i} + \frac{\eta^2 L}{2}\sum_{i=1}^{d} \frac{\widehat{\nabla}_{t,i}^2}{b_{t,i}^2}.
\end{aligned}$$

Similarly to Lemma 4.6.2, we have:

$$\left|\frac{1}{a_{t,i}} - \frac{1}{b_{t,i}}\right| \leq \frac{|\xi_{t,i}|}{a_{t,i} b_{t,i}}.$$

Then

$$\begin{aligned}\Delta_{t+1}-\Delta_t &\leq -\eta\sum_{i=1}^d\frac{\nabla_{t,i}^2}{b_{t,i}}-\eta\sum_{i=1}^d\frac{\nabla_{t,i}\xi_{t,i}}{a_{t,i}}+\eta\sum_{i=1}^d\left(\frac{1}{a_{t,i}}-\frac{1}{b_{t,i}}\right)\nabla_{t,i}\xi_{t,i}+\frac{\eta^2L}{2}\sum_{i=1}^d\frac{\widehat{\nabla}_{t,i}^2}{b_{t,i}^2}\\ &\leq -\eta\sum_{i=1}^d\frac{\nabla_{t,i}^2}{b_{t,i}}-\eta\sum_{i=1}^d\frac{\nabla_{t,i}\xi_{t,i}}{a_{t,i}}+\eta\sum_{i=1}^d\frac{|\xi_{t,i}|}{a_{t,i}b_{t,i}}\left|\nabla_{t,i}\xi_{t,i}\right|+\frac{\eta^2L}{2}\sum_{i=1}^d\frac{\widehat{\nabla}_{t,i}^2}{b_{t,i}^2}\\ &\leq -\eta\sum_{i=1}^d\frac{\nabla_{t,i}^2}{b_{t,i}}-\eta\sum_{i=1}^d\frac{\nabla_{t,i}\xi_{t,i}}{a_{t,i}}+\eta\sum_{i=1}^d|\xi_{t,i}|\left[\frac{\nabla_{t,i}^2}{2a_{t,i}^2}+\frac{\xi_{t,i}^2}{2b_{t,i}^2}\right]+\frac{\eta^2L}{2}\sum_{i=1}^d\frac{\widehat{\nabla}_{t,i}^2}{b_{t,i}^2}.\end{aligned}$$

Rearranging and summing give us the Lemma. □

Next, we present an analogous per-coordinate result to Lemma 4.6.

Lemma 4.7.2. *With* $M_{T,i}=\max_{t\leq T}|\xi_{t,i}|$, $\sigma_{\max}=\max_{i\in[d]}\sigma_i$, *and for any* $w>0$, *we have with probability at least* $1-2d\delta$

$$\frac{1}{\|b_T\|_1}\sum_{t=1}^T\|\nabla f(x_t)\|_1^2\leq\frac{\Delta_1}{\eta}+d\sigma_{\max}\sqrt{\log\frac{1}{\delta}}+\left(8\|\sigma\|_1\sqrt{\log\frac{T}{\delta}}+d\eta L\right)\log\left(\frac{\|b_T\|_1}{\min b_{0,i}}\right)+\sum_{i=1}^d6\sigma_i\sqrt{\log\frac{T}{\delta}}\sum_{t=1}^T\frac{\xi_{t,i}^2}{b_{t,i}^2}.$$

Proof. We first take care of the term $\sum_{t=1}^T\sum_{i=1}^d\frac{\widehat{\nabla}_{t,i}^2}{b_{t,i}^2}$ from Lemma 4.7.1:

$$\sum_{t=1}^T\sum_{i=1}^d\frac{\widehat{\nabla}_{t,i}^2}{b_{t,i}^2}=\sum_{i=1}^d\sum_{t=1}^T\frac{\widehat{\nabla}_{t,i}^2}{b_{t,i}^2}=\sum_{i=1}^d\sum_{t=1}^T\frac{b_{t,i}^2-b_{t-1,i}^2}{b_{t,i}^2}\leq\sum_{i=1}^d2\log\frac{b_{T,i}}{b_{0,i}}.$$

Next, we deal with $-\sum_{t=1}^T\sum_{i=1}^d\frac{\nabla_{t,i}\xi_{t,i}}{a_{t,i}}$ via our martingale argument. For any $w>0$, we have for each $i\in[d]$:

$$\begin{aligned}\mathbb{E}\left[\exp\left(-w\frac{\nabla_{t,i}\xi_{t,i}}{a_{t,i}}-2w^2\frac{\sigma_i^2\nabla_{t,i}^2}{a_{t,i}^2}\right)\mid\mathcal{F}_t\right]&=\exp\left(-2w^2\frac{\sigma_i^2\nabla_{t,i}^2}{a_{t,i}^2}\right)\mathbb{E}\left[\exp\left(-w\frac{\nabla_{t,i}\xi_{t,i}}{a_{t,i}}\right)\mid\mathcal{F}_t\right]\\&\leq 1.\end{aligned}$$

Then a simple inductive argument gives with probability at least $1-\delta$:

$$-w\sum_{t=1}^T\frac{\nabla_{t,i}\xi_{t,i}}{a_{t,i}}\leq 2w^2\sum_{t=1}^T\frac{\sigma_i^2\nabla_{t,i}^2}{a_{t,i}^2}+\log\frac{1}{\delta}.$$

By a union bound across all coordinate d, we have w.p. at least $1-d\delta$:

$$-\sum_{t=1}^T\sum_{i=1}^d\frac{\nabla_{t,i}\xi_{t,i}}{a_{t,i}}\leq\sum_{t=1}^T\sum_{i=1}^d\frac{w\sigma_i^2\nabla_{t,i}^2}{a_{t,i}^2}+\frac{d}{w}\log\frac{1}{\delta}.\tag{4.22}$$

Let's call the event that (4.22) happens E_1. Now, we deal with $\sum_{t=1}^T\sum_{i=1}^d\frac{\nabla_{t,i}^2}{a_{t,i}^2}$. Note that

$$\frac{\nabla_{t,i}^2}{a_{t,i}^2}=\frac{\nabla_{t,i}^2}{b_{t-1,i}^2+\nabla_{t,i}^2+\sigma_i^2}\leq\frac{2\widehat{\nabla}_{t,i}^2+2\xi_{t,i}^2}{b_{t-1,i}^2+2\widehat{\nabla}_{t,i}^2+2\xi_{t,i}^2+\sigma_i^2}\leq 2\frac{\widehat{\nabla}_{t,i}^2}{b_{t,i}^2}+2\frac{\xi_{t,i}^2}{b_{t,i}^2}.$$

Under the even E_1 and the above result, we can bound Lemma 4.7.1 with probability at least $1 - d\delta$:

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\nabla_{t,i}^2}{b_{t,i}} \\
&\leq \frac{\Delta_1}{\eta} + w\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\sigma_i^2\nabla_{t,i}^2}{a_{t,i}^2} + \frac{d}{w}\log\frac{1}{\delta} + \sum_{t=1}^{T}\sum_{i=1}^{d}\frac{M_{T,i}}{2}\left[\frac{\nabla_{t,i}^2}{a_{t,i}^2} + \frac{\xi_{t,i}^2}{b_{t,i}^2}\right] + \frac{\eta L}{2}\sum_{i=1}^{d}2\log\frac{b_{T,i}}{b_{0,i}} \\
&= \frac{\Delta_1}{\eta} + \frac{d}{w}\log\frac{1}{\delta} + \sum_{t=1}^{T}\sum_{i=1}^{d}\left(2w\sigma_i^2 + M_{T,i}\right)\frac{\widehat{\nabla}_{t,i}^2}{b_{t,i}^2} + 2\sum_{t=1}^{T}\sum_{i=1}^{d}\left(M_{T,i} + w\sigma_i^2\right)\frac{\xi_{t,i}^2}{b_{t,i}^2} + \eta L\sum_{i=1}^{d}\log\frac{b_{T,i}}{b_{0,i}} \\
&= \frac{\Delta_1}{\eta} + \frac{d}{w}\log\frac{1}{\delta} + \sum_{i=1}^{d}\left(4w\sigma_i^2 + 2M_{T,i} + \eta L\right)\log\frac{b_{T,i}}{b_{0,i}} + 2\sum_{i=1}^{d}\left(M_{T,i} + w\sigma_i^2\right)\sum_{t=1}^{T}\frac{\xi_{t,i}^2}{b_{t,i}^2}.
\end{aligned}$$

Note that

$$\begin{aligned}
\left(\sum_{i=1}^{d}\frac{\nabla_{t,i}^2}{b_{t,i}}\right)\left(\sum_{i=1}^{d}b_{t,i}\right) \geq \left(\sum_{i=1}^{d}\|\nabla_{t,i}\|\right)^2 = \|\nabla f(x_t)\|_1^2 \\
\Rightarrow \left(\sum_{i=1}^{d}\frac{\nabla_{t,i}^2}{b_{t,i}}\right) \geq \frac{\|\nabla f(x_t)\|_1^2}{\|b_t\|_1} \geq \frac{\|\nabla f(x_t)\|_1^2}{\|b_T\|_1}.
\end{aligned}$$

Hence, we have

$$\frac{1}{\|b_T\|_1}\sum_{t=1}^{T}\|\nabla f(x_t)\|_1^2 \leq \sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\nabla_{t,i}^2}{b_{t,i}}.$$

Since it is known that with probability at least $1-\delta$, $\max_{t\in[T]}|\xi_{t,i}| \leq \sigma_i\sqrt{1+\log\frac{T}{\delta}}$ for each $i \in [d]$ Li and Orabona (2020) and Liu et al. (2022), a union bound over all d gives us that w.p.$\geq 1 - d\delta$

$$M_{T,i} \leq 2\sigma_i\sqrt{\log\frac{T}{\delta}}, \forall i \in [d]\,. \tag{4.23}$$

Condition under this event and choosing $\frac{1}{w} = \frac{\sigma_{\max}}{\sqrt{\log\frac{1}{\delta}}}$ gives us with probability at least $1 - 2d\delta$

$$\begin{aligned}
\frac{1}{\|b_T\|_1}\sum_{t=1}^{T}\|\nabla f(x_t)\|_1^2 &\leq \frac{\Delta_1}{\eta} + d\sigma_{\max}\sqrt{\log\frac{1}{\delta}} + \sum_{i=1}^{d}\left(\frac{4\sigma_i^2}{\sigma_{\max}}\sqrt{\log\frac{1}{\delta}} + 4\sigma_i\sqrt{\log\frac{T}{\delta}} + \eta L\right)\log\frac{b_{T,i}}{b_{0,i}} \\
&\quad + 2\sum_{i=1}^{d}\left(2\sigma_i\sqrt{\log\frac{T}{\delta}} + \frac{\sigma_i^2}{\sigma_{\max}}\sqrt{\log\frac{1}{\delta}}\right)\sum_{t=1}^{T}\frac{\xi_{t,i}^2}{b_{t,i}^2} \\
&\leq \frac{\Delta_1}{\eta} + d\sigma_{\max}\sqrt{\log\frac{1}{\delta}} + \left(8\|\sigma\|_1\sqrt{\log\frac{T}{\delta}} + d\eta L\right)\log\left(\frac{\|b_T\|_1}{\min b_{0,i}}\right) + \\
&\quad \sum_{i=1}^{d}6\sigma_i\sqrt{\log\frac{T}{\delta}}\sum_{t=1}^{T}\frac{\xi_{t,i}^2}{b_{t,i}^2}.
\end{aligned}$$

□

Finally, it remains to bound $\|b_T\|_1$ and $\sum_{t=1}^{T}\frac{\xi_{t,i}^2}{b_{t,i}^2}$. For this we use Lemma 4.7.6 to show the following bound on $\|b_T\|_1$:

Lemma 4.7.3. *With probability at least* $1 - 2d\delta$

$$\begin{aligned}\|b_T\|_1 &\leq 2\|b_0\|_1 + \frac{4\Delta_1}{\eta} + \log\left(\frac{2}{\delta}\right)\sum_{i=1}^{d}\frac{8\sigma_i^2}{b_{0,i}} + 4\sum_{i=1}^{d}\sqrt{\sigma_i^2 T + \sigma_i^2\log\frac{2}{\delta}} + 4\eta^2 L\sum_{i=1}^{d}\log\frac{4\eta^2 L}{b_{0,i}}\\ &= O\left(\|\sigma\|_1\sqrt{T} + \|b_0\|_1 + \frac{\Delta_1}{\eta} + \left\|\frac{\sigma^2}{b_0}\right\|_1\log\left(\frac{1}{\delta}\right) + \|\sigma\|_1\sqrt{\log\frac{1}{\delta}} + \eta^2 L\sum_{i=1}^{d}\log\frac{\eta^2 L}{b_{0,i}}\right).\end{aligned}$$

Proof. We start via the smoothness of f

$$\begin{aligned}f(x_{t+1}) - f(x_t) &\leq \langle\nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2\\ &= -\frac{\eta}{b_t}\left\langle\nabla f(x_t), \widehat{\nabla} f(x_t)\right\rangle + \frac{\eta^2 L}{2}\sum_{i=1}^{d}\frac{\widehat{\nabla}_{t,i}^2}{b_{t,i}^2}\\ &= -\eta\sum_{i=1}^{d}\frac{\widehat{\nabla}_{t,i}^2}{b_{t,i}} + \eta\sum_{i=1}^{d}\frac{\xi_{t,i}\widehat{\nabla}_{t,i}}{b_{t,i}} + \frac{\eta^2 L}{2}\sum_{i=1}^{d}\frac{\widehat{\nabla}_{t,i}^2}{b_{t,i}^2}\\ &\leq -\frac{\eta}{2}\sum_{i=1}^{d}\frac{\widehat{\nabla}_{t,i}^2}{b_{t,i}} + \frac{\eta}{2}\sum_{i=1}^{d}\frac{\xi_{t,i}^2}{b_{t,i}} + \frac{\eta^2 L}{2}\sum_{i=1}^{d}\frac{\widehat{\nabla}_{t,i}^2}{b_{t,i}^2}.\end{aligned}$$

Summing up over t we obtain

$$\begin{aligned}\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\widehat{\nabla}_{t,i}^2}{b_{t,i}} &\leq \frac{2\Delta_1}{\eta} + \sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\xi_{t,i}^2}{b_{t,i}} + \sum_{t=1}^{T}\sum_{i=1}^{d}\eta^2 L\frac{\widehat{\nabla}_{t,i}^2}{b_{t,i}^2}\\ &\leq \frac{2\Delta_1}{\eta} + \sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\xi_{t,i}^2}{b_{t,i}} + \sum_{i=1}^{d}2\eta^2 L\log\frac{b_{T,i}}{b_{0,i}}.\end{aligned}$$

Note that the LHS of the above inequality is lower-bounded by $\|b_T\|_1 - \|b_0\|_1$. Thus

$$\begin{aligned}\|b_T\|_1 - \|b_0\|_1 &\leq \frac{2\Delta_1}{\eta} + \sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\xi_{t,i}^2}{b_{t,i}} + \sum_{i=1}^{d}2\eta^2 L\log\frac{b_{T,i}}{b_{0,i}}\\ &\leq \frac{2\Delta_1}{\eta} + \sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\xi_{t,i}^2}{b_{t,i}} + \sum_{i=1}^{d}2\eta^2 L\left(\log\frac{b_{T,i}}{4\eta^2 L} + \log\frac{4\eta^2 L}{b_{0,i}}\right)\\ &\leq \frac{2\Delta_1}{\eta} + \sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\xi_{t,i}^2}{b_{t,i}} + \frac{\|b_T\|_1}{2} + \sum_{i=1}^{d}2\eta^2 L\log\frac{4\eta^2 L}{b_{0,i}};\\ \|b_T\|_1 &\leq 2\|b_0\|_1 + \frac{4\Delta_1}{\eta} + 2\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\xi_{t,i}^2}{b_{t,i}} + 4\eta^2 L\sum_{i=1}^{d}\log\frac{4\eta^2 L}{b_{0,i}}.\end{aligned}$$

Note that by Lemma 4.7.6, with probability at least $1 - 2d\delta$

$$\begin{aligned}\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\xi_{t,i}^2}{b_{t,i}} &\leq \sum_{i=1}^{d}\frac{8\sigma_i^2\log\frac{1}{\delta}}{b_{0,i}} + 4\sqrt{\sigma_i^2 T + \sigma_i^2\log\frac{1}{\delta}}\\ &= O\left(\left\|\frac{\sigma^2}{b_0}\right\|_1\log\frac{1}{\delta} + \|\sigma\|_1\left(\sqrt{\log\frac{1}{\delta}} + \sqrt{T}\right)\right).\end{aligned}$$

Hence, under this event, we have that with probability at least $1 - 2d\delta$

$$\|b_T\|_1 \leq 2\|b_0\|_1 + \frac{4\Delta_1}{\eta} + O\left(\left\|\frac{\sigma^2}{b_0}\right\|_1 \log\frac{1}{\delta} + \|\sigma\|_1\left(\sqrt{\log\frac{1}{\delta}} + \sqrt{T}\right)\right) + 4\eta^2 L\sum_i \log\frac{4\eta^2 L}{b_{0,i}}.$$

□

Now we are ready to prove Theorem 4.2.6.

Proof of Theorem 4.2.6. Combining Lemma 4.7.1 with Lemma 4.7.6, we get with probability at least $1 - 4d\delta$

$$\begin{aligned}\frac{1}{\|b_T\|_1}\sum_{t=1}^T \|\nabla f(x_t)\|_1^2 &\leq \frac{\Delta_1}{\eta} + d\sigma_{\max}\sqrt{\log\frac{1}{\delta}} + \left(8\|\sigma\|_1\sqrt{\log\frac{T}{\delta}} + d\eta L\right)\log\left(\frac{\|b_T\|_1}{\min b_{0,i}}\right) + \\ &\quad \sum_{i=1}^d 6\sigma_i\sqrt{\log\frac{T}{\delta}}\sum_{t=1}^T\frac{\xi_{t,i}^2}{b_{t,i}^2}. \\ &\leq \frac{\Delta_1}{\eta} + d\sigma_{\max}\sqrt{\log\frac{1}{\delta}} + \left(8\|\sigma\|_1\sqrt{\log\frac{T}{\delta}} + d\eta L\right)\log\left(\frac{\|b_T\|_1}{\min b_{0,i}}\right) \\ &\quad + 6\sqrt{\log\frac{T}{\delta}}\sum_{i=1}^d \sigma_i\left(\frac{8\sigma_i^2}{b_{0,i}^2}\log\frac{1}{\delta} + 2\log\left(1 + \frac{\sigma_i^2 T + \sigma_i^2\log\frac{1}{\delta}}{2b_{0,i}^2}\right)\right).\end{aligned}$$

Rearranging, combining this with the bound for $\|b_T\|_1$, and replacing δ with $\frac{\delta}{6d}$ yield the Theorem. □

4.7.3 Additional Helper Lemmas

Lemma 4.7.4. *We have w.p.* $\geq 1 - d\delta$

$$\sum_{t=1}^\tau \xi_{t,i}^2 \leq \sum_{t=1}^\tau \widehat{\nabla}_{t,i}^2 + 4\sigma_i^2\log\frac{1}{\delta}, \forall\tau\in[T], \forall i\in[d].$$

Proof. We apply Lemma 4.6.3 to each coordinate individually and then union bound over all the dimensions to get the result. □

Lemma 4.7.5. *We have w.p.*$\geq 1 - d\delta$

$$\sum_{t=1}^T \xi_{t,i}^2 \leq \sigma_i^2 T + \sigma_i^2\log\frac{1}{\delta}, \forall i\in[d].$$

Proof. We apply Lemma 4.6.4 to each coordinate individually and then union bound over all the dimensions to get the result. □

We can show a bound on $\sum_{t=1}^T \frac{\xi_{t,i}^2}{b_{t,i}}$ and $\sum_{t=1}^T \frac{\xi_{t,i}^2}{b_{t,i}^2}$ for each $i \in [d]$:

Lemma 4.7.6. *We have*

1. *With probability at least* $1 - 2d\delta$*, we have for all* $i \in [d]$

$$\sum_{t=1}^T \frac{\xi_{t,i}^2}{b_{t,i}} \leq \frac{8\sigma_i^2\log\frac{1}{\delta}}{b_{0,i}} + 4\sqrt{\sigma_i^2 T + \sigma_i^2\log\frac{1}{\delta}}.$$

2. *With probability at least* $1-2d\delta$*, we have for all* $i \in [d]$

$$\sum_{t=1}^{T} \frac{\xi_{t,i}^2}{b_{t,i}^2} \leq \frac{8\sigma_i^2}{b_{0,i}^2} \log \frac{1}{\delta} + 2\log\left(1 + \frac{\sigma_i^2 T + \sigma_i^2 \log \frac{1}{\delta}}{2b_{0,i}^2}\right).$$

Proof. For (1), we have with probability at least $1-2d\delta$

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\xi_{t,i}^2}{b_{t,i}} &= \sum_{t=1}^{T} \frac{\xi_{t,i}^2}{\sqrt{b_{0,i}^2 + \sum_{s=1}^{t} \widehat{\nabla}_{s,i}^2}} \\
&\overset{(1)}{\leq} \sum_{t=1}^{T} \frac{\xi_{t,i}^2}{\sqrt{b_{0,i}^2 + \left(\sum_{s=1}^{t} \xi_{s,i}^2 - 4\sigma_i^2 \log \frac{1}{\delta}\right)^+}} \\
&\leq \frac{8\sigma_i^2 \log \frac{1}{\delta}}{b_{0,i}} + 2\sqrt{2}\sqrt{\sum_{s=1}^{T} \xi_{s,i}^2} \\
&\overset{(2)}{\leq} \frac{8\sigma_i^2 \log \frac{1}{\delta}}{b_{0,i}} + 4\sqrt{\sigma_i^2 T + \sigma_i^2 \log \frac{1}{\delta}},
\end{aligned}$$

where (1) is due to Lemma 4.7.4 and (2) is due to Lemma 4.7.5.

For (2), we have with probability at least $1-2d\delta$

$$\begin{aligned}
\sum_{t} \frac{\xi_{t,i}^2}{b_{t,i}^2} &= \sum_{t} \frac{\xi_{t,i}^2}{b_{0,i}^2 + \sum_{s=1}^{t} \widehat{\nabla}_{s,i}^2} \\
&\overset{(1)}{\leq} \sum_{t} \frac{\xi_{t,i}^2}{b_{0,i}^2 + \left(\sum_{s=1}^{t} \xi_{s,i}^2 - 4\sigma_i^2 \log \frac{1}{\delta}\right)^+} \\
&\leq \frac{8\sigma_i^2 \log \frac{1}{\delta}}{b_{0,i}^2} + 2\log\left(1 + \frac{\sum_{t=1}^{T} \xi_{t,i}^2}{2b_{0,i}^2}\right) \\
&\overset{(2)}{\leq} \frac{8\sigma_i^2 \log \frac{1}{\delta}}{b_{0,i}^2} + 2\log\left(1 + \frac{\sigma_i^2 T + \sigma_i^2 \log \frac{1}{\delta}}{2b_{0,i}^2}\right),
\end{aligned}$$

where (1) is due to Lemma 4.7.4 and (2) is due to Lemma 4.7.5. □

4.8 Simplified Proof for High Probability Convergence of SGD under Convex Objectives

In this section, we present simplified proofs for SGD under a simplified convex setup over the proof in Section 4.1. This proof strategy can be generalized to mirror descent and accelerated variants but we present a simplified setting here to show the main ideas. This proof strategy presented here also applies to the non-convex case, where we utilize this strategy when proving the convergence of Subspace Momentum in Section 8.5 which is quite similar in spirit to the SGD for non-convex objective proof.

Proposition 4.8.1. *Suppose that* f *is convex and* L*-smooth. Suppose that the gradient noise* $\xi_t := \widehat{\nabla} f(x) - \nabla f(x)$ *is* σ*-sub-gaussian for all* x*. We have that the SGD update*

$x_{t+1} = x_t - \eta_t \widehat{\nabla} f(x_t)$ satisfies the following bound with probability at least $1-\delta$:

$$\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x_*) \leq \frac{\sigma}{\sqrt{T}} \|x_1 - x_*\| \sqrt{6 + 12 \log \frac{1}{\delta}}.$$

4.8.1 Simplified Proof

Proof. If $x_{t+1} = x_t - \eta_t \widehat{\nabla} f(x_t)$ then:

$$\begin{aligned}
\|x_{t+1} - x_*\|^2 &= \|x_{t+1} - x_t + x_t - x_*\|^2 \\
&= \|x_{t+1} - x_t\|^2 + \|x_t - x_*\|^2 + 2\langle x_{t+1} - x_t, x_t - x_*\rangle \\
&= \eta_t^2 \left\|\widehat{\nabla} f(x_t)\right\|^2 + \|x_t - x_*\|^2 + 2\eta_t \left\langle \widehat{\nabla} f(x_t), x_* - x_t \right\rangle \\
&= \eta_t^2 \left\|\widehat{\nabla} f(x_t)\right\|^2 + \|x_t - x_*\|^2 + 2\eta_t \langle \xi_t, x_* - x_t\rangle + 2\eta_t \langle \nabla f(x_t), x_* - x_t\rangle
\end{aligned}$$

Using smooth and convex, we have

$$\begin{aligned}
&\|x_{t+1} - x_*\|^2 \\
&= \eta_t^2 \left\|\widehat{\nabla}(x_t)\right\|^2 + \|x_t - x_*\|^2 + 2\eta_t \langle \xi_t, x_* - x_t\rangle + 2\eta_t \underbrace{\langle \nabla f(x_t), x_* - x_t\rangle}_{\text{smooth and convex}} \\
&\leq \eta_t^2 \left\|\widehat{\nabla}(x_t)\right\|^2 + \|x_t - x_*\|^2 + 2\eta_t \langle \xi_t, x_* - x_t\rangle + 2\eta_t (f(x_*) - f(x_t)) - \frac{\eta_t}{L}\|\nabla f(x_t)\|^2 .
\end{aligned}$$

Since $\left\|\widehat{\nabla}(x_t)\right\|^2 = \|\xi_t + \nabla f(x_t)\|^2 \leq 2\|\xi_t\|^2 + 2\|\nabla f(x_t)\|^2$, we have

$$\|x_{t+1} - x_*\|^2 - \|x_t - x_*\|^2 + 2\eta_t (f(x_t) - f(x_*)) \tag{4.24}$$

$$\leq \eta_t \left(2\eta_t - \frac{1}{L}\right)\|\nabla f(x_t)\|^2 + 2\eta_t^2 \|\xi_t\|^2 + 2\eta_t \langle \xi_t, x_* - x_t\rangle . \tag{4.25}$$

If $\eta_t < \frac{1}{2L}$, we have:

$$\|x_{t+1} - x_*\|^2 - \|x_t - x_*\|^2 + 2\eta_t (f(x_t) - f(x_*)) \leq 2\eta_t^2 \|\xi_t\|^2 + 2\eta_t \langle \xi_t, x_* - x_t\rangle$$

$$\frac{1}{2\eta_t}\left[\|x_{t+1} - x_*\|^2 - \|x_t - x_*\|^2\right] + [f(x_t) - f(x_*)] \leq \eta_t \|\xi_t\|^2 + \langle \xi_t, x_* - x_t\rangle .$$

We define additional weights to help with telescoping (that we pay somewhere else), where $w_t \geq 0$ that satisfies the following conditions (which will be obvious from the analysis):

- w_t is non-increasing i.e. $w_1 \geq w_2 \geq \cdots \geq w_T$.
- $w_t \eta_t \leq \frac{1}{4\sigma^2}$ and w_t is $\mathcal{F}_t$-measurable.
- $\frac{w_t}{2\eta_t} + 3\sigma^2 w_t^2 \leq \frac{w_{t-1}}{2\eta_{t-1}}$ or for fixed step size $w_t + 6\sigma^2 \eta w_t^2 \leq w_{t-1}$.

$$\frac{w_t}{2\eta_t}\left[\|x_{t+1} - x_*\|^2 - \|x_t - x_*\|^2\right] + w_t [f(x_t) - f(x_*)] \leq w_t \eta_t \|\xi_t\|^2 + w_t \langle \xi_t, x_* - x_t\rangle .$$

We can apply Corollary 4.3.3 to get that if $w_t \eta_t \leq \frac{1}{4\sigma^2}$ and w_t is $\mathcal{F}_t$-measurable then

$$\mathbb{E}\left[\exp\left(w_t \eta_t \|\xi_t\|^2 + w_t \langle \xi_t, x_* - x_t\rangle\right) \mid \mathcal{F}_t\right] \leq \exp\left(3\sigma^2 \left(\eta_t w_t + w_t^2 \|x_* - x_t\|^2\right)\right).$$

Then applying Lemma (4.3.4), we get that w.p. at least $1-\delta$

$$\sum_{t=1}^{T} w_t \eta_t \|\xi_t\|^2 + w_t \langle \xi_t, x_* - x_t \rangle \leq \sum_{t=1}^{T} 3\sigma^2 \left(\eta_t w_t + w_t^2 \|x_* - x_t\|^2 \right) + \log(1/\delta).$$

Combining things, we would get that w.p. at least $1-\delta$:

$$\begin{aligned}\sum_{t=1}^{T} w_t [f(x_t) - f(x_*)] \leq \sum_{t=1}^{T} 3\sigma^2 \eta_t w_t + \sum_{t=1}^{T} 3\sigma^2 w_t^2 \|x_* - x_t\|^2 + \log(1/\delta) \\ + \sum_{t=1}^{T} \frac{w_t}{2\eta_t} \left[\|x_t - x_*\|^2 - \|x_{t+1} - x_*\|^2 \right].\end{aligned}$$

Let us focus on the distance from optimal terms

$$\begin{aligned}&\sum_{t=1}^{T} \frac{w_t}{2\eta_t} \|x_t - x_*\|^2 + 3\sigma^2 w_t^2 \|x_* - x_t\|^2 - \frac{w_t}{2\eta_t} \|x_{t+1} - x_*\|^2 \\ =&\sum_{t=1}^{T} \left(\frac{w_t}{2\eta_t} + 3\sigma^2 w_t^2 \right) \|x_t - x_*\|^2 - \frac{w_t}{2\eta_t} \|x_{t+1} - x_*\|^2.\end{aligned}$$

To make this telescope, we need to set w_t so that $\frac{w_t}{2\eta_t} + 3\sigma^2 w_t^2 \leq \frac{w_{t-1}}{2\eta_{t-1}}$. Let's assume we are working with a fixed step size $\eta_t = \eta$. We need to set

$$w_t + 6\sigma^2 \eta w_t^2 \leq w_{t-1}.$$

Suppose that we can select w_t's that satisfy those requirements. We would then have

$$\begin{aligned}&\sum_{t=1}^{T} w_t [f(x_t) - f(x_*)] \\ &\leq \sum_{t=1}^{T} 3\sigma^2 \eta_t w_t + \log(1/\delta) + \sum_{t=1}^{T} \left(\frac{w_t}{2\eta_t} + 3\sigma^2 w_t^2 \right) \|x_t - x_*\|^2 - \frac{w_t}{2\eta_t} \|x_{t+1} - x_*\|^2 \\ &\leq \sum_{t=1}^{T} 3\sigma^2 \eta_t w_t + \log(1/\delta) + \sum_{t=1}^{T} \frac{w_{t-1}}{2\eta_{t-1}} \|x_t - x_*\|^2 - \frac{w_t}{2\eta_t} \|x_{t+1} - x_*\|^2 \\ &= \sum_{t=1}^{T} 3\sigma^2 \eta_t w_t + \log(1/\delta) + \frac{w_0}{2\eta_0} \|x_1 - x_*\|^2 - \frac{w_T}{2\eta_T} \|x_{T+1} - x_*\|^2 \\ &\leq \sum_{t=1}^{T} 3\sigma^2 \eta_t w_t + \log(1/\delta) + \frac{w_0}{2\eta_0} \|x_1 - x_*\|^2.\end{aligned}$$

Since w_t is non-increasing, we have $w_T \leq w_t$ for all t. That means

$$\sum_{t=1}^{T} f(x_t) - f(x_*) \leq 3\sigma^2 \sum_{t=1}^{T} \eta_t \frac{w_t}{w_T} + \frac{1}{w_T} \log(1/\delta) + \frac{w_0}{2\eta_0 w_T} \|x_1 - x_*\|^2.$$

Note that on the RHS we need an $O\left(\sqrt{T}\right)$ bound.

Setting w_t. We need $w_t \eta_t \leq \frac{1}{4\sigma^2} \iff w_t \leq \frac{1}{4\sigma^2 \eta}$. This is a recursion. If we can solve it, we might have a better idea on how to solve it. Solving recursion is somewhat

similar to ODE. Rearranging this, we get

$$\begin{aligned} w_t - w_{t-1} &= \underbrace{-6\sigma^2\eta}_{C} w_t^2 \\ &\sim \frac{dw_t}{dt} = Cw_t^2 \\ \implies \int w_t^{-2}\,\mathrm{d}w_t &= \int C\,\mathrm{d}t \\ -w_t^{-1} + B &= Ct \\ \implies w_t &= \frac{1}{6\sigma^2\eta t + B}. \end{aligned}$$

This suggests that we set $w_T = \frac{1}{6\sigma^2\eta T + B}$ for some B. Note that $w_T \leq \frac{1}{4\sigma^2\eta}$ if $B \geq 0$. We now have:

$$\begin{aligned} w_{t-1} &= w_t + 6\sigma^2\eta w_t^2 \\ &\leq \frac{1}{6\sigma^2\eta t + B} + \frac{6\sigma^2\eta}{(6\sigma^2\eta t + B)^2} \\ &\leq \frac{1}{6\sigma^2\eta t + B} + \frac{6\sigma^2\eta}{(6\sigma^2\eta t + B)(6\sigma^2\eta(t-1) + B)} \\ &= \frac{6\sigma^2\eta(t-1) + B + 6\sigma^2\eta}{(6\sigma^2\eta t + B)(6\sigma^2\eta(t-1) + B)} \\ &= \frac{1}{6\sigma^2\eta(t-1) + B}. \end{aligned}$$

So we can set B to be anything as long as $B \geq 0$.

Finishing. Since $w_t \leq \frac{1}{6\sigma^2\eta t + B}$, we have $w_t \leq \frac{1}{B}$. We have $\frac{1}{6\sigma^2\eta T + B} = w_T \leq w_t \leq \frac{1}{B}$. That means $\frac{w_t}{w_T} \leq \frac{6\sigma^2\eta T + B}{B}$. Setting $B = 6\sigma^2\eta T$ (since the upperbound cannot be smaller than $O(1)$) gives $\frac{w_t}{w_T} \leq 2$ and $w_T = \frac{1}{12\sigma^2\eta T}$.

$$\begin{aligned} \sum_{t=1}^{T} f(x_t) - f(x_*) &\leq 3\sigma^2\eta \sum_{t=1}^{T} \frac{w_t}{w_T} + \frac{1}{w_T}\log(1/\delta) + \frac{w_0}{2\eta_0 w_T}\|x_1 - x_*\|^2 \\ &\leq 6T\sigma^2\eta\left[1 + 2\log(1/\delta)\right] + \frac{1}{\eta}\|x_1 - x_*\|^2. \end{aligned}$$

Then setting η to balance the two terms finish the proof

$$\begin{aligned} 6T\sigma^2\eta\left[1 + 2\log(1/\delta)\right] &= \frac{1}{\eta}\|x_1 - x_*\|^2 \\ \implies \eta &= \frac{\|x_1 - x_*\|}{\sqrt{6T\sigma^2\left[1 + 2\log(1/\delta)\right]}}. \end{aligned}$$

Our final bound is

$$\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x_*) \leq \frac{\sigma}{\sqrt{T}}\|x_1 - x_*\|\sqrt{6 + 12\log\frac{1}{\delta}}.$$

□

Chapter 5

Heavy-Tailed Noise: Clipped SGD and Clipped (Accelerated) SMD

5.1 Overview

This section addresses several open questions posed by previous works including handling general domains and dealing with an unknown time horizon under heavy-tailed noise. Qualitatively, we close the logarithmic suboptimality gap and achieve the optimal rate in several settings. More specifically:

– We demonstrate a novel approach to analyze clipped gradient methods in high probability that is general and applies to various standard settings. In the convex setting, we analyze Clipped-SMD and clipped stochastic accelerated mirror descent. In the non-convex setting, we analyze Clipped-SGD. Using our new analysis, we show that clipped methods attain time-optimal convergence in high probability for both convex and nonconvex objectives under heavy-tailed gradient noise. In the convex setting, we obtain an $O\left(T^{\frac{1-p}{p}}\right)$ convergence rate for arbitrary (not necessarily compact) convex domains for Clipped-SMD and $O\left(T^{\frac{1-p}{p}}\sigma + T^{-2}\right)$ for accelerated Clipped-SMD, where σ is the noise parameter. These rates are time-optimal and match the lower bounds in (Raginsky and Rakhlin, 2009; Vural et al., 2022). In the nonconvex setting, we obtain the optimal convergence rate of $O\left(T^{\frac{2-2p}{3p-2}}\right)$ for clipped-SGD. This bound is also time-optimal and matches the lower bound in (Zhang et al., 2020); it also complements the in-expectation convergence of clipped-SGD provided by (Zhang et al., 2020).

– Previous works for heavy-tailed noises follow the recipe of using Freedman-type inequalities (Freedman, 1975; Dzhaparidze and Van Zanten, 2001) as a *blackbox* and bound the iterates inductively for all iterations. This process incurs an additional $\log T$ dependency in the final convergence rate; in other words, the success probability goes from $1-\delta$ to $1-T\delta$. The step sizes and clipping parameters of this approach depend on the time horizon T to enable the union bound and induction across all iterations in the analysis, excluding the important case when the time horizon is unknown. Our whitebox approach forgoes the aforementioned induction, not only circumventing the $\log T$ loss but also allowing for an unknown time horizon. We further show that our analysis allows for a choice of step size and clipping parameters that do not depend on generally unknown parameters like the noise-parameter σ, the failure probability δ, and the initial distance to the optimum, all of which appear impossible using only the techniques from prior works.

5.1.1 Assumptions

We reiterate the assumptions in this setting:

(1) Existence of a minimizer: In the convex setting, we assume that there exists $x^* \in \arg\min_{x\in\mathcal{X}} f(x)$. We let $f^* = f(x^*)$.

(1') Existence of a finite lower bound: In the nonconvex setting, we assume that f admits a finite lower bound, i.e., $f^* := \inf_{x\in\mathbb{R}^d} f(x) > -\infty$.

(2) Unbiased estimator: We assume that our algorithm is allowed to query a stochastic first-order oracle that returns a history-independent, unbiased gradient estimator $\widehat{\nabla} f(x)$ of $\nabla f(x)$ for any $x \in \mathcal{X}$. That is, conditioned on the history and the queried point x, we have $\mathbb{E}[\widehat{\nabla} f(x) \mid x] = \nabla f(x)$.

(3) Bounded pth moment noise: We assume that there exists $\sigma > 0$ such that for some $1 < p \leq 2$ and for any $x \in \mathcal{X}$, $\widehat{\nabla} f(x)$ satisfies $\mathbb{E}[\|\widehat{\nabla} f(x) - \nabla f(x)\|_*^p \mid x] \leq \sigma^p$.

(4) L-smoothness: We consider the class of L-smooth functions: for all $x, y \in \mathbb{R}^d$, $\|\nabla f(x) - \nabla f(y)\|_* \leq L \|x - y\|$.

5.2 Gradient Clipping Operator and Notations

We introduce the gradient clipping operator and its general properties used in Clipped-SMD (Algorithm 7) and Clipped-SGD (Algorithm 6). Let x_t be the output at iteration t of an algorithm of interest. We denote by $\widehat{\nabla} f(x_t)$ the stochastic gradient obtained by querying the gradient oracle. The clipped gradient estimate $\widetilde{\nabla} f(x_t)$ is taken as

$$\widetilde{\nabla} f(x_t) = \min\left\{1, \frac{\lambda_t}{\left\|\widehat{\nabla} f(x_t)\right\|_*}\right\} \widehat{\nabla} f(x_t), \tag{5.1}$$

where λ_t is the clipping parameter used in iteration t. In subsequent sections, we let $\Delta_t := f(x_t) - f^*$ denote the optimal function value gap at x_t. We let $\mathcal{F}_t = \sigma\left(\widehat{\nabla} f(x_1), \ldots, \widehat{\nabla} f(x_t)\right)$ be the natural filtration at time t and define the following notations for the stochastic error, the deviation, and the bias of the clipped gradient estimate at time t:

$$\begin{aligned}
\theta_t &= \widetilde{\nabla} f(x_t) - \nabla f(x_t); \\
\theta_t^u &= \widetilde{\nabla} f(x_t) - \mathbb{E}\left[\widetilde{\nabla} f(x_t) \mid \mathcal{F}_{t-1}\right] \\
\theta_t^b; &= \mathbb{E}\left[\widetilde{\nabla} f(x_t) \mid \mathcal{F}_{t-1}\right] - \nabla f(x_t).
\end{aligned}$$

Note that $\theta_t^u + \theta_t^b = \theta_t$. Regardless of the convexity of the function f, the following lemma provides upper bounds for these quantities. These bounds can be found in prior works (Gorbunov et al., 2020; Zhang et al., 2020; Liu et al., 2023d; Sadiev et al., 2023) for the special case of ℓ_2 norm. The extension to the general norm follows in the same manner, which we omit in this work.

Lemma 5.2.1. *For stochastic gradients $\widehat{\nabla} f(x_t)$ with bounded pth moment noise, the clipped gradients $\widetilde{\nabla} f(x_t)$ satisfy the following properties:*

$$\|\theta_t^u\|_* = \left\|\widetilde{\nabla} f(x_t) - \mathbb{E}\left[\widetilde{\nabla} f(x_t) \mid \mathcal{F}_{t-1}\right]\right\|_* \leq 2\lambda_t. \tag{5.2}$$

Furthermore, if $\|\nabla f(x_t)\|_* \leq \frac{\lambda_t}{2}$ *then*

$$\left\|\theta_t^b\right\|_* = \left\|\mathbb{E}\left[\widetilde{\nabla} f(x_t) \mid \mathcal{F}_{t-1}\right] - \nabla f(x_t)\right\|_* \leq 4\sigma^p \lambda_t^{1-p}; \tag{5.3}$$

$$\mathbb{E}\left[\|\theta_t^u\|_*^2\right] = \mathbb{E}\left[\left\|\widetilde{\nabla} f(x_t) - \mathbb{E}_t\left[\widetilde{\nabla} f(x_t)\right]\right\|_*^2 \mid \mathcal{F}_{t-1}\right] \leq 40\sigma^p \lambda_t^{2-p}. \tag{5.4}$$

Finally, we state a simple but important lemma that bounds the moment generating function of a zero-mean bounded random variable. The proof can be found in, for example, equation (3) of (Beygelzimer et al., 2011).

Lemma 5.2.2. *Let* X *be a random variable such that* $\mathbb{E}[X] = 0$ *and* $|X| \leq R$ *almost surely. Then for* $0 \leq \lambda \leq \frac{1}{R}$

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] \leq \exp\left(\frac{3}{4}\lambda^2 \mathbb{E}\left[X^2\right]\right).$$

5.3 Clipped Stochastic Gradient Descent for Nonconvex Functions

Algorithm 6 Clipped-SGD

Parameters: initial point x_1, step sizes $\{\eta_t\}$, clipping parameters $\{\lambda_t\}$
for $t = 1$ to T do
$\widetilde{\nabla} f(x_t) = \min\left\{1, \frac{\lambda_t}{\|\widehat{\nabla} f(x_t)\|}\right\} \widehat{\nabla} f(x_t)$
$x_{t+1} = x_t - \eta_t \widetilde{\nabla} f(x_t)$

In this section, we study the convergence of Clipped-SGD for nonconvex functions. Here, we consider the domain to be $\mathbb{R}^d$ equipped with the standard ℓ_2 norm. We first outline a blackbox concentration argument to show convergence in high probability of Algorithm 6 and then follow-up with a more powerful whitebox approach that allows for a tight high probability convergence analysis.

Comparison to previous works. In the simple setting of known time horizon and without momentum for Clipped-SGD, the $\widetilde{O}(T^{\frac{2-2p}{3p-2}})$ convergence rate has not been shown before to the best of our knowledge. The recent work by (Sadiev et al., 2023) study this case and only give a suboptimal rate of $\widetilde{O}(T^{\frac{1-p}{p}})$. Note that (Cutkosky and Mehta, 2021; Liu et al., 2023d) study other variants of Clipped-SGD with momentums incorporated. Although (Cutkosky and Mehta, 2021; Liu et al., 2023d) achieve the nearly-optimal time dependency of $\widetilde{O}(T^{\frac{2-2p}{3p-2}})$ in the non-convex settings, they rely on using blackbox concentration inequalities which result in a suboptimal convergence rate that also requires a known time horizon.

We first present the guarantee for known time horizon T via our whitebox approach in Theorem 5.3.1 and defer the statement for unknown T in Theorem 5.7.2 to Section 5.7.

Theorem 5.3.1. *Assume that f satisfies Assumption (1'), (2), (3), (4). Let $\gamma := \max\left\{\log\frac{1}{\delta}; 1\right\}$ and $\Delta_1 := f(x_1) - f^*$. For known time horizon T, we choose λ_t and η_t such that*

$$\lambda_t := \lambda := \max\left\{\left(\frac{8\gamma}{\sqrt{L\Delta_1}}\right)^{\frac{1}{p-1}} T^{\frac{1}{3p-2}}\sigma^{\frac{p}{p-1}}; 2\sqrt{90L\Delta_1}; 32^{\frac{1}{p}}\sigma T^{\frac{1}{3p-2}}\right\}$$

$$\eta_t := \eta := \frac{\sqrt{\Delta_1}T^{\frac{1-p}{3p-2}}}{8\lambda\sqrt{L\gamma}} = \frac{\sqrt{\Delta_1}}{8\sqrt{L\gamma}}\min\left\{\left(\frac{8\gamma}{\sqrt{L\Delta_1}}\right)^{\frac{-1}{p-1}} T^{\frac{-p}{3p-2}}\sigma^{\frac{-p}{p-1}}; \frac{T^{\frac{1-p}{3p-2}}}{2\sqrt{90L\Delta_1}}; \frac{T^{\frac{-p}{3p-2}}}{32^{1/p}\sigma}\right\}.$$

Then with probability at least $1-\delta$

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2 \leq 720\sqrt{\Delta_1 L}\gamma\max\left\{\left(\frac{8\gamma}{\sqrt{L\Delta_1}}\right)^{\frac{1}{p-1}} T^{\frac{2-2p}{3p-2}}\sigma^{\frac{p}{p-1}}; 2\sqrt{90L\Delta_1}T^{\frac{1-2p}{3p-2}}; 32^{1/p}\sigma T^{\frac{2-2p}{3p-2}}\right\} = O\left(T^{\frac{2-2p}{3p-2}}\right).$$

Remark 2. *In comparison to the corresponding results in (Sadiev et al., 2023) (Theorem E.2), while our result achieves a poly T factor better rate when $p < 2$, the dependency on $\log\frac{1}{\delta}$ in our result contains a dependency on p while the result in (Sadiev et al., 2023) does not. That term can dominate the convergence rate in the regime when δ is very small and p is very close to 1. Hence, an open question is to remove such dependency on p for the $\log\frac{1}{\delta}$ term while still maintain the optimal rate on T.*

Now, we turn to the analysis, starting with the key Lemma 5.3.2 (proof in Section 5.7).

Lemma 5.3.2. *Assume that f satisfies Assumption (1'), (2), (3), (4) and $\eta_t \leq \frac{1}{L}$ then for all $t \geq 1$,*

$$\frac{\eta_t}{2}\|\nabla f(x_t)\|^2 \leq \Delta_t - \Delta_{t+1} + \left(L\eta_t^2 - \eta_t\right)\langle\nabla f(x_t), \theta_t^u\rangle + \frac{3\eta_t}{2}\left\|\theta_t^b\right\|^2 + L\eta_t^2\left(\|\theta_t^u\|^2 - \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right]\right) + L\eta_t^2\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right]. \tag{5.5}$$

Remark 3. *In Lemma 5.3.2, we decompose the RHS into appropriate terms that allow us to define a martingale. This lemma helps us understand why we can achieve a better convergence rate $O(T^{\frac{2-2p}{3p-2}})$ (for minimizing the norm squared of the gradient) in comparison to the best rate of $O(T^{\frac{1-p}{p}})$ in the convex setting. We focus on the error term $\langle\nabla f(x_t), \theta_t\rangle = \langle\nabla f(x_t), \theta_t^u\rangle + \langle\nabla f(x_t), \theta_t^b\rangle$ on the RHS of (5.5). Since this error contains the gradient $\nabla f(x_t)$, we leverage some of the gain $\|\nabla f(x_t)\|^2$ on the LHS of 5.5: we use Cauchy-Schwarz to bound $\langle\nabla f(x_t), \theta_t^b\rangle \leq \frac{1}{2}\|\nabla f(x_t)\|^2 + \frac{1}{2}\|\theta_t^b\|^2$ and use the some of the gain to absorb the first term. Then setting our parameters λ_t, η_t appropriately to balance the remaining terms helps us achieve the $O(T^{\frac{2-2p}{3p-2}})$ rate. Contrast this to the convex setting in the next section: the mismatch between the error term that contains the distance term $\|x^* - x_t\|$ and the gain term that contains the function value gap $f(x_t) - f^*$ prevents us from using the gain to absorb some of the error. Thus, this explains the convergence rate discrepancy between the convex case and the non-convex setting (see also Remark 6).*

Before giving a sketch of our whitebox approach, we present a sketch of a black-box argument that gives a nearly time-optimal convergence rate. This approach has an additional $\log T$ factor in the final rate but will serve as a point of comparison for our new techniques, which will close the logarithmic gap.

Blackbox approach. The key lies in the following lemma, which yields the near optimal $\widetilde{O}(T^{\frac{2-2p}{3p-2}})$ convergence rate of Clipped-SGD. In this case, we assume that the clipping parameters λ_t and the step sizes η_t are fixed. Note that the success probability is only $1 - T\delta$. This result uses Lemma 5.3.2 and Freedman's inequality (Theorem 5.6.1) primarily as a *blackbox* to bound the error terms inductively by the initial function value gap to optimality.

Lemma 5.3.3. *For $1 \leq N \leq T+1$, let $\eta_t = \eta$, $\lambda_t = \lambda$ (the specific choices are omitted here for brevity) and E_N be the event that for all $k = 1, \dots N$,*

$$L\eta^2 \sum_{t=1}^{k-1} \|\theta_t^u\|^2 + \left(L\eta^2 - \eta\right) \sum_{t=1}^{k-1} \langle \nabla f(x_t), \theta_t^u \rangle + \frac{3\eta}{2} \sum_{t=1}^{k-1} \left\|\theta_t^b\right\|^2 \leq \Delta_1.$$

Then E_N happens with probability at least $1 - \frac{(N-1)\delta}{T}$ for each $N \in [T+1]$.

With the above lemma, we can obtain a near-optimal convergence rate. However, this rate is still suboptimal due to the use of T union bounds as part of the induction proof. We now discuss an improved analysis that closes the remaining gap.

Whitebox approach. Our whitebox approach defines a novel supermartingale difference sequence Z_t (shown below) and analyzes its moment generating function from first principles. The sequence is designed to leverage the structure of the problem and Clipped-SGD via carefully chosen decreasing weights z_t (shown below).

$$\begin{aligned} Z_t := z_t \Bigg(&\frac{\eta_t}{2} \|\nabla f(x_t)\|^2 + \Delta_{t+1} - \Delta_t - \frac{3\eta_t}{2} \left\|\theta_t^b\right\|^2 - L\eta_t^2 \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right] \Bigg) \\ &- \left(3z_t^2 L\eta_t^2 \Delta_t + 6L^2 z_t^2 \eta_t^4 \lambda_t^2\right) \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right] \\ \text{where } z_t := &\frac{1}{2P_t\eta_t\lambda_t \max_{i \leq t} \sqrt{2L\Delta_i} + 8Q_t L\eta_t^2\lambda_t^2} \end{aligned}$$

for $P_t, Q_t \in \mathcal{F}_{t-1} \geq 1$. We also define $S_t := \sum_{i=1}^t Z_i$. Note that by selecting $P_t, Q_t, \eta_t, \lambda_t$ appropriately so that $P_t\eta_t\lambda_t$ and $Q_t\eta_t^2\lambda_t^2$ are constants (see for example the proof of Proposition 5.3.5 in Section 5.7), we can ensure that the sequence z_t is decreasing.

We now present Lemma 5.3.4 which is the main result for controlling the above martingale, whose proof will offer insights into the main technique in this paper. The technique to prove Lemma 5.3.4 is similar to the standard way of bounding the moment generating function in proving concentration inequalities, such as Freedman's inequality (Freedman, 1975; Dzhaparidze and Van Zanten, 2001). The main challenge here is to find a way to leverage the structure of Clipped-SGD and choose the suitable coefficients z_t. Similarly to (Liu et al., 2023c) where the authors analyze SGD with sub-Gaussian noise, we analyze the martingale difference sequence in a "whitebox" manner. In (Liu et al., 2023c), however, thanks to the light-tailed noise, the weights z_t can be chosen depending only on the problem parameters and independently of the algorithm history. On the other hand, to use Lemma 5.2.2, we have to make sure that $z_t \leq \frac{1}{R}$, where R is an upper bound for the martingale elements. The key here is to choose z_t depending on the past iterates, and use the function value gaps Δ_t to absorb the error incurred during the analysis. We give a proof sketch and defer the full version to Section 5.7.

Lemma 5.3.4. *For any $\delta > 0$, let $E(\delta)$ be the event that for all $1 \leq k \leq T$*

$$\frac{1}{2}\sum_{t=1}^{k} z_t\eta_t \left\|\nabla f(x_t)\right\|^2 + z_k\Delta_{k+1} \leq z_1\Delta_1 + \log\frac{1}{\delta} + \sum_{t=1}^{k}\frac{3z_t\eta_t}{2}\left\|\theta_t^b\right\|^2 + \sum_{t=1}^{k}\left((3z_t^2L\eta_t^2\Delta_t + 6L^2z_t^2\eta_t^4\lambda_t^2 + z_tL\eta_t^2)\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right]\right).$$

Then $\Pr\left[E(\delta)\right] \geq 1-\delta$.

Proof Sketch. Using Lemmas 5.3.2, 5.2.2, and the condition for z_t, we can show that $\mathbb{E}\left[\exp\left(Z_t\right) \mid \mathcal{F}_{t-1}\right] \leq 1$. This then implies

$$\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_{t-1}\right] = \exp\left(S_{t-1}\right)\mathbb{E}\left[\exp\left(Z_t\right) \mid \mathcal{F}_{t-1}\right] \leq \exp\left(S_{t-1}\right),$$

which means $\left(\exp\left(S_t\right)\right)_{t\geq 1}$ is a supermartingale. By Ville's inequality, we have, for all $k \geq 1$, $\Pr\left[S_k \geq \log\frac{1}{\delta}\right] \leq \delta\mathbb{E}\left[\exp\left(S_1\right)\right] \leq \delta$. In other words, with probability at least $1-\delta$, for all $k \geq 1$, $\sum_{t=1}^{k} Z_t \leq \log\frac{1}{\delta}$. Plugging in the definition of Z_t we obtain the desired inequality. □

We now specify the choice of η_t and λ_t. The following lemma gives a general condition for the choice of η_t and λ_t that gives the right convergence rate in time T.

Proposition 5.3.5. *We assume that the event $E(\delta)$ from Lemma 5.3.4 happens. Suppose that for some $\ell \leq T$, there are constants C_1, C_2 and C_3 such that for all $t \leq \ell$*

1. $\lambda_t\eta_t\sqrt{2L} \leq C_1$; *2.* $\frac{1}{L\eta_t}\left(\frac{1}{\lambda_t}\right)^p \leq C_2$; *3.* $\sum_{t=1}^{T} L\left(\frac{1}{\lambda_t}\right)^p\lambda_t^2\eta_t^2 \leq C_3$; *4.* $\|\nabla f(x_t)\| \leq \frac{\lambda_t}{2}$.

Then for all $t \leq \ell+1$

$$\frac{1}{2}\sum_{i=1}^{t}\eta_i\left\|\nabla f(x_i)\right\|^2 + \Delta_{t+1} \leq \left(\sqrt{\Delta_1} + 2\sqrt{A}C_1\right)^2$$

for a constant $A \geq \max\left\{64\left(\log\frac{1}{\delta} + \frac{60\sigma^pC_3}{C_1^2}\right)^2 + \frac{48\sigma^{2p}C_2C_3+140\sigma^pC_3}{C_1^2}; 1\right\}$.

Finally, the proof for Theorem 5.3.1 is a direct consequence of Proposition 5.3.5 where we defer the details to Section 5.7.

5.4 Clipped Stochastic Mirror Descent for Convex Objectives

Algorithm 7 Clipped-SMD

Parameters: initial point x_1, step sizes $\{\eta_t\}$, clipping parameters $\{\lambda_t\}$, ψ is 1-strongly convex wrt $\|\cdot\|$
for $t = 1$ to T do
$\widetilde{\nabla} f(x_t) = \min\left\{1, \frac{\lambda_t}{\left\|\widehat{\nabla} f(x_t)\right\|_*}\right\}\widehat{\nabla} f(x_t)$
$x_{t+1} = \arg\min_{x\in\mathcal{X}}\left\{\eta_t\left\langle\widetilde{\nabla} f(x_t), x\right\rangle + \mathbf{D}_\psi\left(x, x_t\right)\right\}$

In this section, we present and analyze the Clipped Stochastic Mirror Descent algorithm (Algorithm 7) under heavy-tailed noise, with a general domain and arbitrary norm.

We define the Bregman divergence $\mathbf{D}_\psi(x,y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x-y\rangle$, where $\psi : \mathbb{R}^d \to \mathbb{R}$ is a 1-strongly convex differentiable function with respect to the norm $\|\cdot\|$ on $\mathcal{X}$. We assume for convenience that $\operatorname{dom}(\psi) = \mathbb{R}^d$. Algorithm 7 is a generalization of Clipped-SGD for convex functions to an arbitrary norm. The only difference from the standard Stochastic Mirror Descent algorithm is the use of the clipped gradient $\widetilde{\nabla} f(x_t)$ in place of the true stochastic gradient $\widehat{\nabla} f(x_t)$ when computing the new iterate x_{t+1}.

Prior works such as (Gorbunov et al., 2020) only consider the setting where the global minimizer lies in $\mathcal{X}$. Our algorithm and analysis does not require this restriction and instead only uses the following initial gradient estimate assumption from (Nazin et al., 2019):

(5) Initial gradient estimate: Let x_1 be the initial point. We assume that we have access to an upperbound ∇_1 of $\|\nabla f(x_1)\|_*$ i.e. $\|\nabla f(x_1)\|_* \le \nabla_1$. This assumption is justified as follows. If the noise parameter σ defined in assumption (3) is known, we can use the procedure of (Minsker, 2015) to estimate $\|\nabla f(x_1)\|_*$: we take $O(\ln(1/\delta))$ stochastic gradient samples at x_1, and let g_1 be the geometric median of these samples; we then set $\nabla_1 := \|g_1\|_* + 10\sigma$. It follows from (Minsker, 2015) that $\|\nabla f(x_1)\|_* \le \nabla_1$ holds with probability at least $1-\delta$. If the domain contains the global optimum x^* ($\nabla f(x^*) = 0$) and the initial distance $\|x_1 - x^*\|$ is known, we have the following alternative upper bound that follows from $\nabla f(x^*) = 0$ and smoothness:$\|\nabla f(x_1)\|_* = \|\nabla f(x_1) - \nabla f(x^*)\|_* \le L\|x_1 - x^*\|$.

Convergence guarantees. We first state the convergence guarantee for this algorithm in Theorem 5.4.1 which works for an arbitrary norm and a general domain which may not include the global optimum. In this theorem, we assume that we know several problem parameters to show the main idea of our analysis. In Theorem 5.4.2, we remove the knowledge of the problem parameters.

Theorem 5.4.1. *Assume that convex f satisfies Assumptions (1), (2), (3), (4) and (5). Let $\gamma = \max\{\log\frac{1}{\delta}; 1\}$; $R_1 = \sqrt{2\mathbf{D}_\psi(x^*, x_1)}$, and assume that ∇_1 is an upper bound of $\|\nabla f(x_1)\|_*$. For known T, we choose λ_t and η_t such that*

$$\lambda_t = \lambda = \max\left\{\left(\frac{26T}{\gamma}\right)^{1/p}\sigma; 2(3LR_1 + \nabla_1)\right\}, \text{ and}$$
$$\eta_t = \eta = \frac{R_1}{24\lambda_t\gamma} = \frac{R_1}{24\gamma}\min\left\{\left(\frac{26T}{\gamma}\right)^{-1/p}\sigma^{-1}; \frac{1}{2}(3LR_1 + \nabla_1)^{-1}\right\}.$$

Then with probability at least $1-\delta$

$$\frac{1}{T}\sum_{t=2}^{T+1}\Delta_t \le 48R_1\max\left\{26^{\frac{1}{p}}T^{\frac{1-p}{p}}\sigma\gamma^{\frac{p-1}{p}}; 2(3LR_1+\nabla_1)T^{-1}\gamma\right\} = O\left(T^{\frac{1-p}{p}}\right).$$

Remark 4. *This theorem shows that the convergence rate for the known time horizon case is $O(T^{\frac{1-p}{p}})$. This rate is known to be optimal, matching the lower bounds shown in (Raginsky and Rakhlin, 2009; Vural et al., 2022). The above guarantee is also adaptive to σ, i.e., when $\sigma \to 0$, we obtain the standard $O(T^{-1})$ convergence rate of deterministic mirror descent.*

Remark 5. *The term ∇_1 in the above theorem comes from the inexact estimation of $\|\nabla f(x_1)\|_*$. If we assume that the global optimum lies in the domain $\mathcal{X}$, we can simply select $\nabla_1 = LR_1$ without using the estimation procedure, as discussed in (5).*

In Theorem 5.4.1, we use the initial distance R_1 to the optimal solution to set the step sizes and clipping parameters. This information is generally not available, but can be avoided. For example, for constrained problems where the domain radius is bounded by R, we can replace R_1 in Theorem 5.4.1 by R without change in the dependency. However, for the general problem, we present Theorem 5.4.2, where we do not require knowledge of the constants T, σ, δ or R_1 to set the step sizes and clipping parameters. However, we still need the mild assumption of knowing an upper bound ∇_1 on $\|\nabla f(x_1)\|_*$. As discussed in (5), ∇_1 can be estimated with good accuracy when σ is known.

Theorem 5.4.2. *Assume that convex f satisfies Assumption (1), (2), (3), (4) and (5). Let $\gamma = \max\{\log\frac{1}{\delta}; 1\}$; $R_1 = \sqrt{2\mathbf{D}_\psi(x^*, x_1)}$, and assume that ∇_1 is an upper bound of $\|\nabla f(x_1)\|_*$. We choose λ_t and η_t such that*

$$\lambda_t = \max\left\{\left(52t(1+\log t)^2 c_2\right)^{1/p}; 2\left(L\max_{i\le t}\|x_i - x_1\| + \nabla_1\right); \frac{Lc_1}{6}\right\}, \text{ and}$$
$$\eta_t = \frac{c_1}{24\lambda_t} = \frac{c_1}{24}\min\left\{\left(52t(1+\log t)^2 c_2\right)^{-1/p}; \frac{1}{2\left(L\max_{i\le t}\|x_i - x_1\| + \nabla_1\right)}; \frac{6}{Lc_1}\right\},$$

where the absolute constants c_1 and c_2 are to ensure the correctness of the dimensions. Then, with probability at least $1-\delta$, we have

$$\frac{1}{T}\sum_{t=2}^{T+1}\Delta_t \le \frac{8}{Tc_1}\left(R_1 + \frac{c_1}{3}\left(\gamma + \frac{2\sigma^p}{c_2}\right)\right)^2 \max\left\{\left(52T(1+\log T)^2 c_2\right)^{1/p}; \right.$$
$$\left. 4R_1 L + \frac{2c_1}{3}L\left(\gamma + \frac{2\sigma^p}{c_2}\right) + 2\nabla_1; \frac{Lc_1}{6}\right\} = \widetilde{O}\left(T^{\frac{1-p}{p}}\right).$$

Sketch of the analysis. In the remainder of this section, we provide a sketch of the analysis for Theorem 5.4.1, which starts with the following lemma.

Lemma 5.4.3. *Assume that convex f satisfies Assumption (1), (2), (3), (4) and $\eta_t \le \frac{1}{4L}$, the iterate sequence $(x_t)_{t\ge1}$ output by Algorithm 7 satisfies the following:*

$$\eta_t\Delta_{t+1} \le \mathbf{D}_\psi(x^*, x_t) - \mathbf{D}_\psi(x^*, x_{t+1}) + \eta_t\langle x^* - x_t, \theta_t^u\rangle + \eta_t\left\langle x^* - x_t, \theta_t^b\right\rangle$$
$$+ 2\eta_t^2\left(\|\theta_t^u\|_*^2 - \mathbb{E}\left[\|\theta_t^u\|_*^2 \mid \mathcal{F}_{t-1}\right]\right) + 2\eta_t^2\mathbb{E}\left[\|\theta_t^u\|_*^2 \mid \mathcal{F}_{t-1}\right] + 2\eta_t^2\left\|\theta_t^b\right\|_*^2.$$

Remark 6. *In contrast to Remark 3, there is a mismatch between the gain Δ_{t+1} and the loss $\langle x^* - x_t, \theta_t\rangle$. Since the distance $\|x^* - x_t\|$ and the function value gap Δ_t cannot be related in the general convex case, we do not obtain the same rate as in the nonconvex case.*

We now define the following terms for $t \ge 1$:

$$Z_t := z_t\left(\eta_t\Delta_{t+1} + \mathbf{D}_\psi(x^*, x_{t+1}) - \mathbf{D}_\psi(x^*, x_t) - \eta_t\left\langle x^* - x_t, \theta_t^b\right\rangle - 2\eta_t^2\left\|\theta_t^b\right\|_*^2\right.$$
$$\left. - 2\eta_t^2\mathbb{E}\left[\|\theta_t^u\|_*^2 \mid \mathcal{F}_{t-1}\right]\right) - \left(\frac{3}{8\lambda_t^2} + 24z_t^2\eta_t^4\lambda_t^2\right)\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right],$$

where $z_t := \dfrac{1}{2\eta_t\lambda_t\max_{i\le t}\sqrt{2\mathbf{D}_\psi(x^*, x_i)} + 16Q\eta_t^2\lambda_t^2}$

for a constant $Q \geq 1$. We also define $S_t := \sum_{i=1}^{t} Z_i$. We have the following lemma, which is analogous to Lemma 5.3.4 in the nonconvex case.

Lemma 5.4.4. *For any $\delta > 0$, let $E(\delta)$ be the event that for all $1 \leq k \leq T$*

$$\sum_{t=1}^{k} z_t \eta_t \Delta_{t+1} + z_k \mathbf{D}_\psi\left(x^*, x_{k+1}\right) \leq z_1 \mathbf{D}_\psi\left(x^*, x_1\right) + \log\frac{1}{\delta} + \sum_{t=1}^{k} z_t \eta_t \left\langle x^* - x_t, \theta_t^b \right\rangle + 2\sum_{t=1}^{k} z_t \eta_t^2 \left\|\theta_t^b\right\|_*^2 + \sum_{t=1}^{k}\left(\left(2 z_t \eta_t^2 + \frac{3}{8\lambda_t^2} + 24 z_t^2 \eta_t^4 \lambda_t^2\right) \mathbb{E}\left[\|\theta_t^u\|_*^2 \mid \mathcal{F}_{t-1}\right]\right). \tag{5.6}$$

Then $\Pr\left[E(\delta)\right] \geq 1 - \delta$.

We now specify the choice of η_t and λ_t. The following proposition gives a general condition for the choice of η_t and λ_t that gives the right convergence rate in time T.

Proposition 5.4.5. *We assume that the event $E(\delta)$ from Lemma 5.4.4 happens. Suppose that for some $\ell \leq T$, there are constants C_1, C_2, C_3, and A such that for all $t \leq \ell$*

1. $\lambda_t \eta_t = C_1$; 2. $\sum_{t=1}^{\ell} \left(\frac{1}{\lambda_t}\right)^p \leq C_2$; 3. $\left(\frac{1}{\lambda_t}\right)^{2p} \leq C_3 \left(\frac{1}{\lambda_t}\right)^p$; 4. $\|\nabla f(x_t)\|_* \leq \frac{\lambda_t}{2}$. Then for all $t \leq \ell + 1$

$$\sum_{i=1}^{t} \eta_i \Delta_{i+1} + \mathbf{D}_\psi\left(x^*, x_{t+1}\right) \leq \frac{1}{2}\left(R_1 + 8AC_1\right)^2$$

for $A \geq \max\left\{\log\frac{1}{\delta} + 26\sigma^p C_2 + \frac{2\sigma^{2p} C_2 C_3}{A}; 1\right\}$.

Theorem 5.4.1 follows from Proposition 5.4.5. Both proofs can be found in Section 5.7.

5.5 Accelerated Stochastic Mirror Descent and Extensions

Algorithm 8 Clipped-ASMD

Parameters: initial point $y_1 = z_1$, step sizes $\{\eta_t\}$, clipping parameters $\{\lambda_t\}$, and mirror map ψ, where ψ is 1-strongly convex wrt $\|\cdot\|$.
For $t = 1$ to T do:
  Set $\alpha_t = \frac{2}{t+1}$.
  $x_t = (1 - \alpha_t) y_t + \alpha_t z_t$.
  $\widetilde{\nabla} f(x_t) = \min\left\{1, \frac{\lambda_t}{\left\|\widehat{\nabla} f(x_t)\right\|_*}\right\} \widehat{\nabla} f(x_t)$.
  $z_{t+1} = \arg\min_{x \in \mathcal{X}} \left\{\eta_t \left\langle \widetilde{\nabla} f(x_t), x \right\rangle + \mathbf{D}_\psi\left(x, z_t\right)\right\}$.
  $y_{t+1} = (1 - \alpha_t) y_t + \alpha_t z_{t+1}$.

In Section 5.9, we also show the convergence and its analysis for Clipped Accelerated Stochastic Mirror Descent (Algorithm 8). We require the following additional assumption:

(5′) Global minimizer: We assume that $\nabla f(x^*) = 0$.

In other words, we assume that the global minimizer lies in the domain of the problem. This assumption is consistent with the works of (Gorbunov et al., 2020;

Sadiev et al., 2023). Our analysis readily extends to non-smooth settings, and more generally to functions that satisfy

$$f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + G \|y - x\| + \frac{L}{2} \|y - x\|^2, \quad \forall y, x \in \mathcal{X}.$$

This condition is satisfied by both Lipschitz functions (when $L = 0$) and smooth functions (when $G = 0$). The key step is to extend Lemma 5.4.3. The proof follows from (Lan, 2020) and can be found in Section 5.7.

5.6 Freedman's Inequality

Lemma 5.6.1 (Freedman's inequality). *Let $(X_t)_{t \geq 1}$ be a martingale difference sequence. Assume that there exists a constant $c > 0$ such that $|X_t| \leq c$ almost surely for all $t \geq 1$ and define $\sigma_t^2 = \mathbb{E}\left[X_t^2 \mid X_{t-1}, \ldots, X_1\right]$. Then for all $b > 0$, $F > 0$ and $T \geq 1$*

$$\Pr\left[\left|\sum_{t=1}^{T} X_t\right| > b \text{ and } \sum_{t=1}^{T} \sigma_t^2 \leq F\right] \leq 2 \exp\left(-\frac{b^2}{2F + 2cb/3}\right).$$

5.7 Missing Proofs from Section 5.3

Proof of Lemma 5.3.2. By the smoothness of f and the update $x_{t+1} = x_t - \frac{1}{\eta_t} \widetilde{\nabla} f(x_t)$ we have

$$\begin{aligned}
& f(x_{t+1}) - f(x_t) \\
& \leq \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\
& = -\eta_t \left\langle \nabla f(x_t), \widetilde{\nabla} f(x_t) \right\rangle + \frac{L\eta_t^2}{2} \left\| \widetilde{\nabla} f(x_t) \right\|^2 \\
& = -\eta_t \langle \nabla f(x_t), \theta_t + \nabla f(x_t) \rangle + \frac{L\eta_t^2}{2} \|\theta_t + \nabla f(x_t)\|^2 \\
& = -\eta_t \|\nabla f(x_t)\|^2 - \eta_t \langle \nabla f(x_t), \theta_t \rangle + \frac{L\eta_t^2}{2} \|\theta_t\|^2 + \frac{L\eta_t^2}{2} \|\nabla f(x_t)\|^2 + L\eta_t^2 \langle \nabla f(x_t), \theta_t \rangle \\
& = -\left(\eta_t - \frac{L\eta_t^2}{2}\right) \|\nabla f(x_t)\|^2 + \frac{L\eta_t^2}{2} \|\theta_t\|^2 + \left(L\eta_t^2 - \eta_t\right) \langle \nabla f(x_t), \theta_t \rangle \\
& = -\left(\eta_t - \frac{L\eta_t^2}{2}\right) \|\nabla f(x_t)\|^2 + \frac{L\eta_t^2}{2} \|\theta_t\|^2 + \underbrace{\left(L\eta_t^2 - \eta_t\right)}_{\leq 0} \left\langle \nabla f(x_t), \theta_t^u + \theta_t^b \right\rangle.
\end{aligned}$$

Using Cauchy-Schwarz, we have $\langle \nabla f(x_t), \theta_t^b \rangle \leq \frac{1}{2} \|\nabla f(x_t)\|^2 + \frac{1}{2} \|\theta_t^b\|^2$. Thus, we derive

$$\begin{aligned}
\Delta_{t+1} - \Delta_t &\leq -\left(\frac{2\eta_t - L\eta_t^2}{2}\right) \|\nabla f(x_t)\|^2 + \frac{L\eta_t^2}{2} \|\theta_t\|^2 + \left(L\eta_t^2 - \eta_t\right) \langle \nabla f(x_t), \theta_t^u \rangle \\
&\quad + \frac{\eta_t - L\eta_t^2}{2} \|\nabla f(x_t)\|^2 + \frac{\eta_t - L\eta_t^2}{2} \left\|\theta_t^b\right\|^2 \\
&\leq -\frac{\eta_t}{2} \|\nabla f(x_t)\|^2 + \frac{L\eta_t^2}{2} \|\theta_t\|^2 + \left(L\eta_t^2 - \eta_t\right) \langle \nabla f(x_t), \theta_t^u \rangle + \frac{\eta_t}{2} \left\|\theta_t^b\right\|^2 \\
&\leq -\frac{\eta_t}{2} \|\nabla f(x_t)\|^2 + L\eta_t^2 \|\theta_t^u\|^2 + \left(L\eta_t^2 - \eta_t\right) \langle \nabla f(x_t), \theta_t^u \rangle + \left(L\eta_t^2 + \frac{\eta_t}{2}\right) \left\|\theta_t^b\right\|^2 \\
&\leq -\frac{\eta_t}{2} \|\nabla f(x_t)\|^2 + L\eta_t^2 \|\theta_t^u\|^2 + \left(L\eta_t^2 - \eta_t\right) \langle \nabla f(x_t), \theta_t^u \rangle + \frac{3\eta_t}{2} \left\|\theta_t^b\right\|^2,
\end{aligned}$$

where the third inequality is due to $\|\theta_t\|^2 \leq 2\|\theta_t^u\|^2 + 2\|\theta_t^b\|^2$, and the last inequality is due to $\eta_t \leq \frac{1}{L}$. Rearranging, adding, and subtracting $\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right]$, we obtain the lemma. □

Detailed proof of Lemma 5.3.3. We state the following simple properties of the choice of η and λ in Theorem 5.3.1. We have

$$\frac{1}{L}\left(\frac{\sigma}{\lambda}\right)^p \leq \eta \tag{5.7}$$

$$\eta \leq \frac{1}{L} \tag{5.8}$$

$$\left(\frac{\sigma}{\lambda}\right)^p T^{\frac{p}{3p-2}} \leq \frac{1}{32} \tag{5.9}$$

$$TL\left(\frac{\sigma}{\lambda}\right)^p \lambda^2\eta^2 \leq \frac{\Delta_1}{2048}. \tag{5.10}$$

We will now prove by induction on N that E_N happens with probability at least $1 - \frac{(N-1)\delta}{T}$. For $N = 1$, the event happens with probability 1. Suppose that for some $N \leq T$, $\Pr[E_N] \geq 1 - \frac{(N-1)\delta}{T}$. We will prove that $\Pr[E_{N+1}] \geq 1 - \frac{N\delta}{T}$.

Since the LHS of (5.5) is non-negative, for $k \leq N$, we have, under the event E_N,

$$\begin{aligned}
\Delta_k &\leq \Delta_1 + \left(L\eta^2 - \eta\right) \sum_{t=1}^{k-1} \langle \nabla f(x_t), \theta_t^u \rangle + L\eta^2 \sum_{t=1}^{k-1} \left(\|\theta_t^u\|^2 - \mathbb{E}_t\left[\|\theta_t^u\|^2\right]\right) \\
&\quad + \frac{3\eta}{2} \sum_{t=1}^{k-1} \left\|\theta_t^b\right\|^2 + L\eta^2 \sum_{t=1}^{k-1} \mathbb{E}_t\left[\|\theta_t^u\|^2\right] \leq 2\Delta_1.
\end{aligned}$$

From the induction hypothesis and Lemma 5.3.2, we have that for all $k \leq N$, $\Delta_k \leq 2\Delta_1$. Since the LHS of (5.5) is non-negative, by summing over t from 1 to N we have,

$$\begin{aligned}
\Delta_{N+1} \leq \underbrace{\left(\eta - L\eta^2\right) \sum_{t=1}^{N} \langle -\nabla f(x_t), \theta_t^u \rangle}_{A} + \underbrace{\frac{3\eta}{2} \sum_{t=1}^{N} \left\|\theta_t^b\right\|^2}_{B} \\
\underbrace{+ L\eta^2 \sum_{t=1}^{N} \left(\|\theta_t^u\|^2 - \mathbb{E}_t\left[\|\theta_t^u\|^2\right]\right)}_{C} + \underbrace{L\eta^2 \sum_{t=1}^{N} \mathbb{E}_t\left[\|\theta_t^u\|^2\right]}_{D}.
\end{aligned}$$

The bounds for B and D are straightforward from Lemma 5.2.1. First, with probability 1, we have $\|\theta_t^u\| \leq 2\lambda$. By the smoothness of f and the fact that f is bounded below, we have

$$\|\nabla f(x_t)\| \leq \sqrt{2L\Delta_t}.$$

Furthermore, when the event E_N happens, we have

$$\|\nabla f(x_t)\| \leq \sqrt{2L\Delta_t} \leq \sqrt{4L\Delta_1} \leq \frac{\lambda}{2}.$$

Thus, we can apply Lemma 5.2.1 and obtain $\left\|\theta_t^b\right\| \leq 4\sigma^p\lambda^{1-p}$ and $\mathbb{E}_t\left[\|\theta_t^u\|^2\right] \leq 40\sigma^p\lambda^{2-p}$.

Upperbound for B. By (5.3), when the event E_N happens,

$$\begin{aligned} B = \frac{3\eta}{2}\left\|\theta_t^b\right\|^2 &\leq \frac{3\eta}{2}\sum_{t=1}^{N} 16\sigma^{2p}\lambda^{2-2p} = 24\sigma^{2p}\lambda^{2-2p}\eta N \\ &\leq 24T\left(\frac{\sigma}{\lambda}\right)^{2p}\lambda^2\eta \leq 24TL\left(\frac{\sigma}{\lambda}\right)^{p}\lambda^2\eta^2 \leq \frac{3\Delta_1}{256}. \end{aligned}$$

Upperbound for D. By 5.4, when the event E_N happens,

$$\begin{aligned} D = L\eta^2\sum_{t=1}^{N}\mathbb{E}_t\left[\|\theta_t^u\|^2\right] &\leq L\eta^2\sum_{t=1}^{N} 40\sigma^p\lambda^{2-p} \\ &\leq 40\sigma^p\lambda^{2-p}L\eta^2 N \leq 40LT\left(\frac{\sigma}{\lambda}\right)^{p}(\lambda\eta)^2 \leq \frac{5\Delta_1}{256}. \end{aligned}$$

To bound A and C, we use Freedman's inequality (Theorem 5.6.1). We define, for $t \geq 1$, the following random variables

$$Z_t = \begin{cases} -\nabla f(x_t) & \text{if } \Delta_t \leq 2\Delta_1 \\ 0 & \text{otherwise.} \end{cases}$$

Thus $\|Z_t\| \leq \|\nabla f(x_t)\| \leq 2\sqrt{L\Delta_1}$ for all t.

Upperbound for A. Instead of bounding $A = \left(\eta - L\eta^2\right)\sum_{t=1}^{N}\langle -\nabla f(x_t), \theta_t^u\rangle$, we will bound $A' = \left(\eta - L\eta^2\right)\sum_{t=1}^{N}\langle Z_t, \theta_t^u\rangle$. We check the conditions to apply Freedman's inequality. First $\mathbb{E}_t\left[\left(\eta - L\eta^2\right)\langle Z_t, \theta_t^u\rangle\right] = 0$. Further, with probability 1, $\|\theta_t^u\|^2 \leq 2\lambda$, and $Z_t \leq 2\sqrt{L\Delta_1}$, thus $\left|\left(\eta - L\eta^2\right)\langle Z_t, \theta_t^u\rangle\right| \leq \left(\eta - L\eta^2\right)\|Z_t\|\,\|\theta_t^u\| \leq 4\sqrt{L\Delta_1}\left(\eta - L\eta^2\right)\lambda \leq 4\sqrt{L\Delta_1}\eta\lambda$. Hence, $\left\{\left(\eta - L\eta^2\right)\langle Z_t, \theta_t^u\rangle\right\}$ is a bounded martingale difference sequence. Therefore, for constant a and F to be chosen we have

$$\begin{aligned} &\Pr\left[\left|\sum_{t=1}^{N}\left(\eta - L\eta^2\right)\langle Z_t, \theta_t^u\rangle\right| > a \text{ and } \sum_{t=1}^{N}\mathbb{E}_t\left[\left(\left(\eta - L\eta^2\right)\langle Z_t, \theta_t^u\rangle\right)^2\right] \leq F\ln\frac{4T}{\delta}\right] \\ &\leq 2\exp\left(-\frac{a^2}{2F\ln\frac{4T}{\delta} + \frac{8}{3}\sqrt{L\Delta_1}\eta\lambda a}\right) \end{aligned}$$

We choose a such that

$$2\exp\left(-\frac{a^2}{2F\ln\frac{4T}{\delta}+\frac{8}{3}\sqrt{L\Delta_1}\eta\lambda a}\right)=\frac{\delta}{2T}$$

which gives

$$a=\left(\frac{4}{3}\sqrt{L\Delta_1}\eta\lambda+\sqrt{\frac{16L\Delta_1\eta^2\lambda^2}{9}+2F}\right)\ln\frac{4T}{\delta}.$$

If we choose $F=64L\Delta_1\sigma^p\lambda^{2-p}\eta^2T$, we can easily show that $a\leq\frac{7\Delta_1}{12}$. Therefore, with probability at least $1-\frac{\delta}{2T}$ we have

$$E_A=\Bigg\{\text{either } A'\leq\left|\sum_{t=1}^N\left(\eta-L\eta^2\right)\langle Z_t,\theta_t^u\rangle\right|\leq\frac{7\Delta_1}{12}$$
$$\text{or }\sum_{t=1}^N\mathbb{E}_t\left[\left(\left(\eta-L\eta^2\right)\langle Z_t,\theta_t^u\rangle\right)^2\right]>F\ln\frac{4T}{\delta}\Bigg\}.$$

Also notice that under the event E_N, we have

$$\begin{aligned}&\sum_{t=1}^N\mathbb{E}_t\left[\left(\left(\eta-L\eta^2\right)\langle Z_t,\theta_t^u\rangle\right)^2\right]\\ \leq&\eta^2\sum_{t=1}^N\mathbb{E}_t\left[\|Z_t\|^2\|\theta_t^u\|^2\right]\leq4\eta^2L\Delta_1\sum_{t=1}^N\mathbb{E}_t\left[\|\theta_t^u\|^2\right]\\ \leq&64L\Delta_1\sigma^p\lambda^{2-p}\eta^2N\leq64\Delta_1LT\left(\frac{\sigma}{\lambda}\right)^p\lambda^2\eta^2\leq F\leq F\ln\frac{4T}{\delta}.\end{aligned}\tag{5.11}$$

Under E_N, we have that $Z_t=-\nabla f(x_t)$ for all $t\leq N$. Therefore, when $E_N\cap E_A$ happens, we have $A=A'\leq a$.

Upperbound for C. We check the conditions to apply Freedman's inequality. First, $\mathbb{E}_t\left[L\eta^2\left(\|\theta_t^u\|^2-\mathbb{E}_t\left[\|\theta_t^u\|^2\right]\right)\right]=0$. Further, with probability 1, $\|\theta_t^u\|^2\leq2\lambda$, thus $\left|L\eta^2\left(\|\theta_t^u\|^2-\mathbb{E}_t\left[\|\theta_t^u\|^2\right]\right)\right|\leq L\eta^2\left(4\lambda^2+4\lambda^2\right)=8L\lambda^2\eta^2$. Hence, $\left\{L\eta^2\left(\|\theta_t^u\|^2-\mathbb{E}_t\left[\|\theta_t^u\|^2\right]\right)\right\}$ is a bounded martingale difference sequence. Applying Freedman's inequality for constants c and G to be chosen, we have

$$\begin{aligned}&\Pr\left[\left|L\eta^2\sum_{t=1}^N\left(\|\theta_t^u\|^2-\mathbb{E}_t\left[\|\theta_t^u\|^2\right]\right)\right|>c\text{ and }\sum_{t=1}^N\mathbb{E}_t\left[\left(L\eta^2\left(\|\theta_t^u\|^2-\mathbb{E}_t\left[\|\theta_t^u\|^2\right]\right)\right)^2\right]\leq G\ln\frac{4T}{\delta}\right]\\ &\leq2\exp\left(-\frac{c^2}{2G\ln\frac{4T}{\delta}+\frac{16}{3}L\lambda^2\eta^2c}\right).\end{aligned}$$

We choose c such that

$$2\exp\left(-\frac{c^2}{2G\ln\frac{4T}{\delta}+\frac{16}{3}L\lambda^2\eta^2c}\right)=\frac{\delta}{2T}$$

which gives

$$c = \left(\frac{8}{3}L\lambda^2\eta^2 + \sqrt{\frac{64L^2\lambda^4\eta^4}{9} + 2G}\right)\ln\frac{4T}{\delta}.$$

If we choose $G = 256L^2\sigma^p\lambda^{4-p}\eta^4 T$, a simple calculation shows that $c \le \frac{7\Delta_1}{48}$. we can show that with probability at least $1 - \frac{\delta}{2T}$, the following event happens

$$E_C = \left\{\text{either } C \le \left|L\eta^2\sum_{t=1}^{N}\left(\|\theta_t^u\|^2 - \mathbb{E}_t\left[\|\theta_t^u\|^2\right]\right)\right| \le \frac{7\Delta_1}{48}\right.$$
$$\left.\text{or } \sum_{t=1}^{N}\mathbb{E}_t\left[\left(L\eta^2\left(\|\theta_t^u\|^2 - \mathbb{E}_t\left[\|\theta_t^u\|^2\right]\right)\right)^2\right] \ge G\ln\frac{4T}{\delta}\right\}.$$

Notice that when $G = 256L^2\sigma^p\lambda^{4-p}\eta^4 T$, under E_N we have

$$\begin{aligned}
&\sum_{t=1}^{N}\mathbb{E}_t\left[\left(L\eta^2\left(\|\theta_t^u\|^2 - \mathbb{E}_t\left[\|\theta_t^u\|^2\right]\right)\right)^2\right]\\
\le& 8L\lambda^2\eta^2\sum_{t=1}^{N}\mathbb{E}_t\left[\left|L\eta^2\left(\|\theta_t^u\|^2 - \mathbb{E}_t\left[\|\theta_t^u\|^2\right]\right)\right|\right] \le 16L^2\lambda^2\eta^4\sum_{t=1}^{N}\mathbb{E}\left[\|\theta_t^u\|^2\right]\\
\le& 256L^2\sigma^p\lambda^{4-p}\eta^4 N \le G < G\ln\frac{4T}{\delta}.
\end{aligned} \tag{5.12}$$

Therefore, when $E_N \cap E_C$ happens, we have $C \le c$.

Finally, combining all the bounds for A, B, C, D using union bound and selecting λ and η appropriately to simplify the constants, we obtain the lemma. □

Proof of Lemma 5.3.4. We have

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(Z_t\right) \mid \mathcal{F}_{t-1}\right]\exp\left(\left(3z_t^2L\eta_t^2\Delta_t + 6L^2z_t^2\eta_t^4\lambda_t^2\right)\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right]\right)\\
\overset{(a)}{\le}& \mathbb{E}\left[\exp\left(z_t\left(\left(L\eta_t^2 - \eta_t\right)\langle\nabla f(x_t), \theta_t^u\rangle + L\eta_t^2\left(\|\theta_t^u\|^2 - \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right]\right)\right)\right) \mid \mathcal{F}_{t-1}\right]\\
\overset{(b)}{\le}& \exp\left(\mathbb{E}\left[\frac{3}{4}\left(z_t\left(\left(L\eta_t^2 - \eta_t\right)\langle\nabla f(x_t), \theta_t^u\rangle + L\eta_t^2\left(\|\theta_t^u\|^2 - \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right]\right)\right)\right)^2 \mid \mathcal{F}_{t-1}\right]\right)\\
\overset{(c)}{\le}& \exp\left(\mathbb{E}\left[\frac{3}{2}z_t^2\eta_t^2\|\nabla f(x_t)\|^2\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right] + \mathbb{E}\left[\frac{3}{2}L^2z_t^2\eta_t^4\|\theta_t^u\|^4 \mid \mathcal{F}_{t-1}\right]\right)\\
\overset{(d)}{\le}& \exp\left(3z_t^2L\eta_t^2\Delta_t\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right] + 6L^2z_t^2\eta_t^4\lambda_t^2\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right]\right)\\
=& \exp\left(\left(3z_t^2L\eta_t^2\Delta_t + 6L^2z_t^2\eta_t^4\lambda_t^2\right)\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right]\right).
\end{aligned}$$

For (a) we use Lemma 5.3.2. For (b) we use Lemma 5.2.2. Notice that

$$\mathbb{E}\left[\langle\nabla f(x_t), \theta_t^u\rangle\right] = \mathbb{E}\left[\|\theta_t^u\|_*^2 - \mathbb{E}\left[\|\theta_t^u\|_*^2 \mid \mathcal{F}_{t-1}\right]\right] = 0,$$

and since $\|\theta_t^u\| \leq 2\lambda_t$ and $\|\nabla f(x_t)\| \leq \sqrt{2L\Delta_t}$ for an L-smooth function, we have

$$\begin{aligned}
&\left|\left(L\eta_t^2 - \eta_t\right)\langle \nabla f(x_t), \theta_t^u\rangle + L\eta_t^2\left(\|\theta_t^u\|^2 - \mathbb{E}\left[\|\theta^u\|^2 \mid \mathcal{F}_{t-1}\right]\right)\right| \\
\leq& 2\eta_t\lambda_t\|\nabla f(x_t)\| + L\eta_t^2\left(\|\theta_t^u\|^2 + \mathbb{E}\left[\|\theta^u\|^2 \mid \mathcal{F}_{t-1}\right]\right) \\
\leq& 2\eta_t\lambda_t\|\nabla f(x_t)\| + 8L\eta_t^2\lambda_t^2 \\
\leq& 2\eta_t\lambda_t\sqrt{2L\Delta_t} + 8L\eta_t^2\lambda_t^2.
\end{aligned}$$

Thus $z_t \leq \frac{1}{2\eta_t\lambda_t\sqrt{2L\Delta_t}+8L\eta_t^2\lambda_t^2}$. For (c) we use $(a+b)^2 \leq 2a^2+2b^2$ and $\mathbb{E}\left[(X-\mathbb{E}[X])^2\right] \leq \mathbb{E}\left[X^2\right]$. For (d), we use $\|\nabla f(x_t)\|^2 \leq 2L\Delta_t$ and $\|\theta_t^u\| \leq 2\lambda_t$. We obtain

$$\mathbb{E}\left[\exp\left(Z_t\right) \mid \mathcal{F}_{t-1}\right] \leq 1.$$

Therefore

$$\begin{aligned}
\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_{t-1}\right] &= \exp\left(S_{t-1}\right)\mathbb{E}\left[\exp\left(Z_t\right) \mid \mathcal{F}_{t-1}\right] \\
&\leq \exp\left(S_{t-1}\right)
\end{aligned}$$

which means $\left(\exp\left(S_t\right)\right)_{t\geq 1}$ is a supermartingale. By Ville's inequality, we have, for all $k \geq 1$

$$\Pr\left[S_k \geq \log\frac{1}{\delta}\right] \leq \delta\mathbb{E}\left[\exp\left(S_1\right)\right] \leq \delta.$$

In other words, with probability at least $1-\delta$, for all $k \geq 1$

$$\sum_{t=1}^{k} Z_t \leq \log\frac{1}{\delta}.$$

Plugging in the definition of Z_t we have

$$\begin{aligned}
&\frac{1}{2}\sum_{t=1}^{k} z_t\eta_t\|\nabla f(x_t)\|^2 + \sum_{t=1}^{k}\left(z_t\Delta_{t+1} - z_t\Delta_t\right) \\
\leq& \log\frac{1}{\delta} + \sum_{t=1}^{k}\frac{3z_t\eta_t}{2}\left\|\theta_t^b\right\|^2 \\
&+ \sum_{t=1}^{k}\left(\left(3z_t^2L\eta_t^2\Delta_t + 6L^2z_t^2\eta_t^4\lambda_t^2 + z_tL\eta_t^2\right)\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right]\right).
\end{aligned}$$

Note that we have z_t is a decreasing sequence by construction (see the proof of Proposition 5.3.5 below). Hence, the LHS of the above inequality can be bounded by

$$\begin{aligned}
\text{LHS} &= \frac{1}{2}\sum_{t=1}^{k} z_t\eta_t\|\nabla f(x_t)\|^2 + z_k\Delta_{k+1} - z_1\Delta_1 + \sum_{t=2}^{k}\left(z_{k-1} - z_k\right)\Delta_k \\
&\geq \frac{1}{2}\sum_{t=1}^{k} z_t\eta_t\|\nabla f(x_t)\|^2 + z_k\Delta_{k+1} - z_1\Delta_1.
\end{aligned}$$

We obtain the desired inequality. □

Proof of Proposition 5.3.5. We will prove by induction on k that

$$\frac{1}{2}\sum_{i=1}^{k}\eta_i\left\|\nabla f(x_i)\right\|^2+\Delta_{k+1}\leq\left(\sqrt{\Delta_1}+2\sqrt{A}C_1\right)^2.$$

The base case $k=0$ is trivial. Suppose the statement is true for all $t\leq k\leq\ell$. Now we show for $k+1$. Recall that

$$z_t=\frac{1}{2P_t\eta_t\lambda_t\max_{i\leq t}\sqrt{2L\Delta_i}+8Q_tL\eta_t^2\lambda_t^2}.$$

Let us choose

$$P_t=\frac{C_1}{\lambda_t\eta_t\sqrt{2L}}\geq 1$$
$$Q_t=\frac{C_1^2\sqrt{A}}{2L\eta_t^2\lambda_t^2}\geq 1.$$

We have

$$z_t=\frac{1}{2C_1\max_{i\leq t}\sqrt{\Delta_i}+4C_1^2\sqrt{A}}.$$

Now, note that $(z_t)_{t\geq 1}$ is a decreasing sequence. By the induction hypothesis $\max_{i\leq k}\sqrt{\Delta_i}\leq\sqrt{\Delta_1}+2\sqrt{A}C_1$. Hence:

$$\begin{aligned}\frac{z_t}{z_k}&=\frac{2C_1\max_{i\leq k}\sqrt{\Delta_i}+4C_1^2\sqrt{A}}{2C_1\max_{i\leq t}\sqrt{\Delta_i}+4C_1^2\sqrt{A}}\\&\leq\frac{2C_1\left(\sqrt{\Delta_1}+2\sqrt{A}C_1\right)+4C_1^2\sqrt{A}}{2C_1\sqrt{\Delta_1}+4C_1^2\sqrt{A}}\\&=\frac{\sqrt{\Delta_1}+4\sqrt{A}C_1}{\sqrt{\Delta_1}+2\sqrt{A}C_1}\leq 2.\end{aligned}$$

By the choice of λ_t, for all $t\leq k$, $\|\nabla f(x_t)\|\leq\frac{\lambda_t}{2}$, we can apply the second part of Lemma 5.2.1 to obtain

$$\begin{aligned}\left\|\theta_t^b\right\|&\leq 4\sigma^p\lambda_t^{1-p};\\\mathbb{E}\left[\|\theta_t^u\|^2\mid\mathcal{F}_{t-1}\right]&\leq 40\sigma^p\lambda_t^{2-p}.\end{aligned}$$

Thus,

$$
\begin{aligned}
&\frac{1}{2}z_k\sum_{t=1}^{k}\eta_t\left\|\nabla f(x_t)\right\|^2+z_k\Delta_{k+1}\\
\leq& z_1\Delta_1+\log\frac{1}{\delta}+\sum_{t=1}^{k}\frac{3z_t\eta_t}{2}\left\|\theta_t^b\right\|^2\\
&+\sum_{t=1}^{k}\left(\left(3z_t^2L\eta_t^2\Delta_t+6L^2z_t^2\eta_t^4\lambda_t^2+z_tL\eta_t^2\right)\mathbb{E}\left[\left\|\theta_t^u\right\|^2\mid\mathcal{F}_{t-1}\right]\right)\\
\leq& z_1\Delta_1+\log\frac{1}{\delta}+24\sigma^{2p}\sum_{t=1}^{k}z_t\eta_t\lambda_t^2\left(\frac{1}{\lambda_t}\right)^{2p}\\
&+40\sigma^p\sum_{t=1}^{k}\left(\left(3z_t^2\Delta_t+6z_t^2L\eta_t^2\lambda_t^2+z_t\right)L\eta_t^2\lambda_t^2\left(\frac{1}{\lambda_t}\right)^p\right).
\end{aligned}
$$

Since $\frac{z_t}{z_k}\leq 2$, we have

$$
\begin{aligned}
&\frac{1}{2}\sum_{t=1}^{k}\eta_t\left\|\nabla f(x_t)\right\|^2+\Delta_{k+1}\\
\leq&\frac{z_1\Delta_1}{z_k}+\frac{1}{z_k}\log\frac{1}{\delta}+48\sigma^{2p}\sum_{t=1}^{k}\eta_t\lambda_t^2\left(\frac{1}{\lambda_t}\right)^{2p}\\
&+80\sigma^p\sum_{t=1}^{k}\left(\left(3z_t\Delta_t+6z_tL\eta_t^2\lambda_t^2+1\right)L\eta_t^2\lambda_t^2\left(\frac{1}{\lambda_t}\right)^p\right)\\
\overset{(a)}{\leq}&\frac{\sqrt{\Delta_1}+4\sqrt{A}C_1}{\sqrt{\Delta_1}+2\sqrt{A}C_1}\Delta_1+2C_1\left(\sqrt{\Delta_1}+4\sqrt{A}C_1\right)\log\frac{1}{\delta}+48\sigma^{2p}C_2\sum_{t=1}^{k}L\eta_t^2\lambda_t^2\left(\frac{1}{\lambda_t}\right)^p\\
&+80\sigma^p\sum_{t=1}^{k}\left(\left(\frac{3\left(\sqrt{\Delta_1}+2\sqrt{A}C_1\right)^2}{2C_1\left(\sqrt{\Delta_1}+2\sqrt{A}C_1\right)}+\frac{6}{8Q_t}+1\right)L\eta_t^2\lambda_t^2\left(\frac{1}{\lambda_t}\right)^p\right)\\
\overset{(b)}{\leq}&\Delta_1+2\sqrt{\Delta_1}\sqrt{A}C_1+2C_1\left(\sqrt{\Delta_1}+4\sqrt{A}C_1\right)\log\frac{1}{\delta}+48\sigma^{2p}C_2C_3\\
&+80\sigma^p\left(\frac{3\left(\sqrt{\Delta_1}+2\sqrt{A}C_1\right)}{2C_1}+\frac{7}{4}\right)C_3\\
\leq&\Delta_1+2\sqrt{\Delta_1}\sqrt{A}C_1+2C_1\left(\sqrt{\Delta_1}+4\sqrt{A}C_1\right)\left(\log\frac{1}{\delta}+\frac{60\sigma^pC_3}{C_1^2}\right)\\
&+48\sigma^{2p}C_2C_3+140\sigma^pC_3\\
\overset{(c)}{\leq}&\Delta_1+2\sqrt{\Delta_1}\sqrt{A}C_1+2C_1\left(\sqrt{\Delta_1}+4\sqrt{A}C_1\right)\frac{\sqrt{A}}{8}+AC_1^2\\
\leq&\left(\sqrt{\Delta_1}+2\sqrt{A}C_1\right)^2.
\end{aligned}
$$

For (a), we use $\left(\frac{1}{\lambda_t}\right)^p \leq C_2 L\eta_t$ and the induction hypothesis. For (b), we use $\sum_{t=1}^T L\left(\frac{1}{\lambda_t}\right)^p \lambda_t^2\eta_t^2 \leq C_3$ and $Q_t \geq 1$. For (c), we have

$$\log\frac{1}{\delta} + \frac{60\sigma^p C_3}{C_1^2} \leq \frac{\sqrt{A}}{8}$$
$$48\sigma^{2p}C_2C_3 + 140\sigma^p C_3 \leq AC_1^2,$$

since

$$A \geq 64\left(\log\frac{1}{\delta} + \frac{60\sigma^p C_3}{C_1^2}\right)^2 + \frac{48\sigma^{2p}C_2C_3 + 140\sigma^p C_3}{C_1^2}.$$

This concludes the proof. □

Lemma 5.7.1. *The choices of η_t and λ_t in Theorem 5.3.1 satisfy the condition (1)-(3) of Proposition 5.3.5 for*

$$C_1 = \frac{\sqrt{\Delta_1}}{4\sqrt{2}\gamma},$$
$$C_2 = \frac{1}{\sigma^p},$$
$$C_3 = \frac{\Delta_1}{2048\sigma^p\gamma}.$$

Proof. We verify for the first case. The second follows exactly the same. First, we have $p > 1$ hence

$$\eta_t\lambda_t\sqrt{2L} = \frac{\sqrt{\Delta_1}T^{\frac{1-p}{3p-2}}}{8\sqrt{L}\gamma}\sqrt{2L} \leq \frac{\sqrt{\Delta_1}}{4\sqrt{2}\gamma} = C_1.$$

Since $\eta_t = \frac{\sqrt{\Delta_1}T^{\frac{1-p}{3p-2}}}{8\lambda_t\sqrt{L}\gamma}$, $p > 1$ and $\lambda_t \geq \left(\frac{8\gamma}{\sqrt{L\Delta_1}}\right)^{\frac{1}{p-1}} T^{\frac{1}{3p-2}}\sigma^{\frac{p}{p-1}}$

$$\begin{aligned}\eta_t\lambda_t^p &= \frac{\sqrt{\Delta_1}T^{\frac{1-p}{3p-2}}}{8\sqrt{L}\gamma}\lambda_t^{p-1} \\ &\geq \frac{\sqrt{\Delta_1}T^{\frac{1-p}{3p-2}}}{8\sqrt{L}\gamma}\frac{8\gamma}{\sqrt{L\Delta_1}}T^{\frac{p-1}{3p-2}}\sigma^p \\ &= \frac{\sigma^p}{L},\end{aligned}$$

which gives

$$\frac{1}{L\eta_t}\left(\frac{1}{\lambda_t}\right)^p \leq \frac{1}{\sigma^p} = C_2.$$

Finally, we have $\lambda_t \geq 32^{1/p}\sigma T^{\frac{1}{3p-2}}$ hence

$$\left(\frac{1}{\lambda_t}\right)^p T^{\frac{p}{3p-2}} \leq \frac{1}{32\sigma^p}.$$

Therefore,

$$\begin{aligned}\sum_{t=1}^{T} L\left(\frac{1}{\lambda_t}\right)^p \lambda_t^2\eta_t^2 &= \sum_{t=1}^{T} L\left(\frac{1}{\lambda_t}\right)^p \left(\frac{\sqrt{\Delta_1}T^{\frac{1-p}{3p-2}}}{8\sqrt{L}\gamma}\right)^2 \\ &= \frac{1}{T}\sum_{t=1}^{T} L\left(\frac{1}{\lambda_t}\right)^p T\cdot T^{\frac{2-2p}{3p-2}}\frac{\Delta_1}{64L\gamma} \\ &= \frac{1}{T}\sum_{t=1}^{T} \left(\frac{1}{\lambda_t}\right)^p T^{\frac{p}{3p-2}}\frac{\Delta_1}{64\gamma^2} \\ &\leq \frac{1}{T}\sum_{t=1}^{T} \frac{1}{32\sigma^p}\frac{\Delta_1}{64\gamma^2} \\ &= \frac{1}{32\sigma^p}\frac{\Delta_1}{64\gamma^2} \leq \frac{\Delta_1}{2048\sigma^p\gamma}.\end{aligned}$$

□

Proof of Theorem 5.3.1. Note that $\eta \leq \frac{T^{\frac{1-p}{3p-2}}}{16\sqrt{90L}\gamma} \leq \frac{1}{L}$. We have that with probability at least $1-\delta$, event $E(\delta)$ happens. Conditioning on this event, we verify the conditions of Proposition 5.3.5. We select the following constants

$$C_1 = \frac{\sqrt{\Delta_1}}{4\sqrt{2}\gamma}; \quad C_2 = \frac{1}{\sigma^p}; \quad C_3 = \frac{\Delta_1}{2048\sigma^p\gamma}; \quad A = 256\gamma^2.$$

We verify in Lemma 5.7.1 that for these choice of constants, conditions (1)-(3) of Proposition 5.3.5 are satisfied. Furthermore, we have

$$\begin{aligned}&64\left(\log\frac{1}{\delta} + \frac{60\sigma^p C_3}{C_1^2}\right)^2 + \frac{48\sigma^{2p}C_2C_3 + 140\sigma^p C_3}{C_1^2} \\ &= 64\left(\log\frac{1}{\delta} + 60\log\frac{1}{\delta}\frac{32}{\Delta_1}\frac{\Delta_1}{2048}\right)^2 + \left(48\frac{\Delta_1}{2048} + 140\frac{\Delta_1}{2048}\right)\frac{32}{\Delta_1} \\ &\leq 256\gamma^2 = A.\end{aligned}$$

We only need to show that, for all t, $\|\nabla f(x_t)\| \leq \frac{\lambda_t}{2}$. We will show this by induction. Indeed, for the base case we have $\|\nabla f(x_1)\| \leq \sqrt{2L\Delta_1} \leq \frac{\lambda_1}{2}$. Suppose that it is true for all $t \leq k$. We will prove that $\|\nabla f(x_{k+1})\| \leq \frac{\lambda_{k+1}}{2}$. By Proposition 5.3.5 and the induction hypothesis

$$\Delta_{k+1} \leq \left(\sqrt{\Delta_1} + 2\sqrt{A}C_1\right)^2 \leq \left(\sqrt{\Delta_1} + \frac{\sqrt{\Delta_1}}{2\sqrt{2}\gamma}\times 16\gamma\right)^2 \leq 45\Delta_1.$$

Thus, we get

$$\|\nabla f(x_{k+1})\| \leq \sqrt{2L\Delta_{k+1}} \leq \sqrt{90L\Delta_1} \leq \frac{\lambda_{k+1}}{2}$$

as needed. From Proposition 5.3.5, we have

$$\frac{\eta}{2}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2 + \Delta_{k+1} \leq 45\Delta_1.$$

Therefore

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2 \leq \frac{90\Delta_1}{\eta T}$$
$$= 720\sqrt{\Delta_1 L}\gamma \max\left\{\left(\frac{8\gamma}{\sqrt{L\Delta_1}}\right)^{\frac{1}{p-1}} T^{\frac{2-2p}{3p-2}}\sigma^{\frac{p}{p-1}}; 2\sqrt{90L\Delta_1}T^{\frac{1-2p}{3p-2}}; 32^{\frac{1}{p}}\sigma T^{\frac{2-2p}{3p-2}}\right\}.$$

□

Theorem 5.7.2. *Assume that f satisfies Assumption (1'), (2), (3), (4). Let $\gamma = \max\left\{\log\frac{1}{\delta}; 1\right\}$ and $\Delta_1 = f(x_1) - f^*$. For unknown T, we choose λ_t and η_t such that*

$$\lambda_t = \max\left\{\left(\frac{8\gamma}{\sqrt{L\Delta_1}}\right)^{\frac{1}{p-1}}\left(2t\left(1+\log t\right)^2\right)^{\frac{1}{3p-2}}\sigma^{\frac{p}{p-1}}; 2\sqrt{90L\Delta_1}; 32^{\frac{1}{p}}\sigma\left(2t\left(1+\log t\right)^2\right)^{\frac{1}{3p-2}}\right\},$$
$$\eta_t = \frac{\sqrt{\Delta_1}\left(2t\left(1+\log t\right)^2\right)^{\frac{1-p}{3p-2}}}{8\lambda_t\sqrt{L}\gamma}.$$

Then with probability at least $1-\delta$

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2 \leq 720\sqrt{\Delta_1 L}\gamma\max\left\{\left(\frac{8\gamma}{\sqrt{L\Delta_1}}\right)^{\frac{1}{p-1}}\left(2\left(1+\log T\right)^2\right)^{\frac{p}{3p-2}}\sigma^{\frac{p}{p-1}}T^{\frac{2-2p}{3p-2}};\right.$$
$$\left. 2\sqrt{90L\Delta_1}\left(2\left(1+\log T\right)^2\right)^{\frac{p-1}{3p-2}}T^{\frac{1-2p}{3p-2}}; 32^{\frac{1}{p}}\sigma\left(2\left(1+\log T\right)^2\right)^{\frac{p}{3p-2}}T^{\frac{2-2p}{3p-2}}\right\}.$$

We again verify the conditions of Proposition 5.3.5 for the choices of η_t and λ_t in Theorem 5.7.2.

Lemma 5.7.3. *The choices of η_t and λ_t in Theorem 5.7.2 satisfy the condition (1)-(3) of Proposition 5.3.5 for*

$$C_1 = \frac{\sqrt{\Delta_1}}{4\sqrt{2}\gamma},$$
$$C_2 = \frac{1}{\sigma^p},$$
$$C_3 = \frac{\Delta_1}{2048\sigma^p\gamma}.$$

The proof utilizes the following fact:

Fact 5.7.4. *We have $\sum_{t=1}^{\infty}\frac{1}{2t(1+\log t)^2} < 1$.*

Proof. First, we have $p > 1$ hence

$$\eta_t\lambda_t\sqrt{2L} = \frac{\sqrt{\Delta_1}\left(2t\left(1+\log t\right)^2\right)^{\frac{1-p}{3p-2}}}{8\sqrt{L}\gamma}\sqrt{2L}$$
$$\leq \frac{\sqrt{\Delta_1}}{4\sqrt{2}\gamma} = C_1.$$

Since $\eta_t = \frac{\sqrt{\Delta_1} T^{\frac{1-p}{3p-2}}}{8\lambda_t \sqrt{L\gamma}}$, $p > 1$ and $\lambda_t \geq \left(\frac{8\gamma}{\sqrt{L\Delta_1}}\right)^{\frac{1}{p-1}} \left(2t\left(1+\log t\right)^2\right)^{\frac{1-p}{3p-2}} \sigma^{\frac{p}{p-1}}$

$$\begin{aligned}
\eta_t \lambda_t^p &= \frac{\sqrt{\Delta_1}\left(2t\left(1+\log t\right)^2\right)^{\frac{1-p}{3p-2}}}{8\sqrt{L\gamma}} \lambda_t^{p-1} \\
&\geq \frac{\sqrt{\Delta_1}\left(2t\left(1+\log t\right)^2\right)^{\frac{1-p}{3p-2}}}{8\sqrt{L\gamma}} \frac{8\gamma}{\sqrt{L\Delta_1}} \left(2t\left(1+\log t\right)^2\right)^{\frac{p-1}{3p-2}} \sigma^p \\
&= \frac{\sigma^p}{L},
\end{aligned}$$

which gives

$$\frac{1}{L\eta_t}\left(\frac{1}{\lambda_t}\right)^p \leq \frac{1}{\sigma^p} = C_2.$$

Finally, we have $\lambda_t \geq 32^{\frac{1}{p}} \sigma \left(2t\left(1+\log t\right)^2\right)^{\frac{1}{3p-2}}$, hence

$$\left(\frac{1}{\lambda_t}\right)^p \left(2t\left(1+\log t\right)^2\right)^{\frac{p}{3p-2}} \leq \frac{1}{32\sigma^p}. \tag{5.13}$$

Therefore,

$$\begin{aligned}
\sum_{t=1}^{T} L\left(\frac{1}{\lambda_t}\right)^p \lambda_t^2 \eta_t^2 &= \sum_{t=1}^{T} L\left(\frac{1}{\lambda_t}\right)^p \left(2t\left(1+\log t\right)^2\right)^{\frac{2-2p}{3p-2}} \left(\frac{\sqrt{\Delta_1}}{8\sqrt{L\gamma}}\right)^2 \\
&= \sum_{t=1}^{T} L \frac{1}{2t\left(1+\log t\right)^2} \left(\frac{1}{\lambda_t}\right)^p \left(2t\left(1+\log t\right)^2\right)^{\frac{p}{3p-2}} \frac{\Delta_1}{64\gamma^2} \\
&\leq \sum_{t=1}^{T} L \frac{1}{2t\left(1+\log t\right)^2} \frac{1}{32\sigma^p} \frac{\Delta_1}{64\gamma^2} && \text{(by (5.13))} \\
&= \frac{1}{32\sigma^p} \frac{\Delta_1}{64\gamma^2} \sum_{t=1}^{T} \frac{1}{2t\left(1+\log t\right)^2} \\
&\leq \frac{1}{32\sigma^p} \frac{\Delta_1}{64\gamma^2} \leq \frac{\Delta_1}{2048\sigma^p\gamma}. && \text{(by Fact 5.7.4)}
\end{aligned}$$

□

Proof of Theorem 5.7.2. Note that

$$\begin{aligned}
\eta_t &= \frac{\sqrt{\Delta_1}\left(2t\left(1+\log t\right)^2\right)^{\frac{1-p}{3p-2}}}{8\lambda_t\sqrt{L\gamma}} \\
&\leq \frac{\left(2t\left(1+\log t\right)^2\right)^{\frac{1-p}{3p-2}}}{16L\gamma\sqrt{90}} \\
&\leq \frac{1}{L}.
\end{aligned}$$

Note that with Lemma 5.7.3, verifying the conditions of Proposition 5.3.5 is identical to the proof of theorem 5.3.1. We have that with probability at least $1-\delta$, event $E(\delta)$

from 5.3.5 happens. We have with probability at least $1-\delta$:

$$\frac{1}{2}\sum_{t=1}^{T}\eta_t\left\|\nabla f(x_t)\right\|^2+\Delta_{k+1}\leq 45\Delta_1.$$

Since η_t is decreasing, we have

$$\frac{1}{T}\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2\leq\frac{90\Delta_1}{T\eta_T}.$$

This means that

$$\frac{1}{T}\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2\leq 720\sqrt{\Delta_1 L}\gamma\max\left\{\left(\frac{8\gamma}{\sqrt{L\Delta_1}}\right)^{\frac{1}{p-1}}\left(2\left(1+\log T\right)^2\right)^{\frac{p}{3p-2}}\sigma^{\frac{p}{p-1}}T^{\frac{2-2p}{3p-2}};\right.$$
$$\left.2\sqrt{90L\Delta_1}\left(2\left(1+\log T\right)^2\right)^{\frac{p-1}{3p-2}}T^{\frac{1-2p}{3p-2}};32^{\frac{1}{p}}\sigma\left(2\left(1+\log T\right)^2\right)^{\frac{p}{3p-2}}T^{\frac{2-2p}{3p-2}}\right\}.$$

□

5.8 Missing Proofs from Section 5.4

Lemma 5.8.1. *Suppose that $\eta_t\leq\frac{1}{4L}$ and assume f satisfies Assumption (1), (2), (3) as well as the following condition*

$$f(y)-f(x)\leq\langle\nabla f(x),y-x\rangle+G\left\|y-x\right\|+\frac{L}{2}\left\|y-x\right\|^2,\quad\forall y,x\in\mathcal{X}.\tag{5.14}$$

Then the iterate sequence $(x_t)_{t\geq 1}$ output by Algorithm 7 satisfies the following:

$$\eta_t\Delta_{t+1}\leq\mathbf{D}_\psi\left(x^*,x_t\right)-\mathbf{D}_\psi\left(x^*,x_{t+1}\right)+\eta_t\left\langle x^*-x_t,\theta_t^u\right\rangle+\eta_t\left\langle x^*-x_t,\theta_t^b\right\rangle$$
$$+2\eta_t^2\left(\left\|\theta_t^u\right\|_*^2-\mathbb{E}\left[\left\|\theta_t^u\right\|_*^2\mid\mathcal{F}_{t-1}\right]\right)+2\eta_t^2\mathbb{E}\left[\left\|\theta_t^u\right\|_*^2\mid\mathcal{F}_{t-1}\right]+2\eta_t^2\left\|\theta_t^b\right\|_*^2+2G^2\eta_t^2.$$

Proof. By condition (5.14) and convexity,

$$\begin{aligned}
&f\left(x_{t+1}\right)-f\left(x^*\right)\\
&\quad\leq\underbrace{f\left(x_{t+1}\right)-f\left(x_t\right)}_{\text{condition (5.14)}}+\underbrace{f\left(x_t\right)-f\left(x^*\right)}_{\text{convexity}}\\
&\quad\leq\left\langle\nabla f\left(x_t\right),x_{t+1}-x_t\right\rangle+\frac{L}{2}\left\|x_t-x_{t+1}\right\|^2+G\left\|x_t-x_{t+1}\right\|+\left\langle\nabla f\left(x_t\right),x_t-x^*\right\rangle\\
&\quad=\left\langle\nabla f\left(x_t\right),x_{t+1}-x^*\right\rangle+\frac{L}{2}\left\|x_t-x_{t+1}\right\|^2+G\left\|x_t-x_{t+1}\right\|\\
&\quad=\left\langle\theta_t,x^*-x_{t+1}\right\rangle+\left\langle\widetilde{\nabla}f(x_t),x_{t+1}-x^*\right\rangle+\frac{L}{2}\left\|x_t-x_{t+1}\right\|^2+G\left\|x_t-x_{t+1}\right\|.
\end{aligned}$$

By the optimality condition, we have

$$\left\langle\eta_t\widetilde{\nabla}f(x_t)+\nabla_x\mathbf{D}_\psi\left(x_{t+1},x_t\right),x^*-x_{t+1}\right\rangle\geq 0$$

and thus

$$\left\langle \eta_t \widetilde{\nabla} f(x_t), x_{t+1} - x^* \right\rangle \leq \left\langle \nabla_x \mathbf{D}_\psi \left(x_{t+1}, x_t\right), x^* - x_{t+1} \right\rangle.$$

Note that

$$\begin{aligned}\left\langle \nabla_x \mathbf{D}_\psi \left(x_{t+1}, x_t\right), x^* - x_{t+1} \right\rangle &= \left\langle \nabla \psi \left(x_{t+1}\right) - \nabla \psi \left(x_t\right), x^* - x_{t+1} \right\rangle \\ &= \mathbf{D}_\psi \left(x^*, x_t\right) - \mathbf{D}_\psi \left(x_{t+1}, x_t\right) - \mathbf{D}_\psi \left(x^*, x_{t+1}\right).\end{aligned}$$

Thus

$$\begin{aligned}\eta_t \left\langle \widetilde{\nabla} f(x_t), x_{t+1} - x^* \right\rangle &\leq \mathbf{D}_\psi \left(x^*, x_t\right) - \mathbf{D}_\psi \left(x^*, x_{t+1}\right) - \mathbf{D}_\psi \left(x_{t+1}, x_t\right) \\ &\leq \mathbf{D}_\psi \left(x^*, x_t\right) - \mathbf{D}_\psi \left(x^*, x_{t+1}\right) - \frac{1}{2} \left\|x_{t+1} - x_t\right\|^2,\end{aligned}$$

where we have used that $\mathbf{D}_\psi \left(x_{t+1}, x_t\right) \geq \frac{1}{2} \left\|x_{t+1} - x_t\right\|^2$ by the strong convexity of ψ.

Combining the two inequalities, and using the assumption that $L\eta_t \leq \frac{1}{4}$, we obtain

$$\begin{aligned}&\eta_t \Delta_{t+1} + \mathbf{D}_\psi \left(x^*, x_{t+1}\right) - \mathbf{D}_\psi \left(x^*, x_t\right) \\ &\leq \eta_t \left\langle \theta_t, x^* - x_{t+1} \right\rangle + \frac{L\eta_t}{2} \left\|x_t - x_{t+1}\right\|^2 + G\eta_t \left\|x_t - x_{t+1}\right\| - \frac{1}{2} \left\|x_{t+1} - x_t\right\|^2 \\ &\leq \eta_t \left\langle \theta_t, x^* - x_t \right\rangle + \eta_t \left\langle \theta_t, x_t - x_{t+1} \right\rangle - \frac{3}{8} \left\|x_{t+1} - x_t\right\|^2 + G\eta_t \left\|x_t - x_{t+1}\right\| \\ &\leq \eta_t \left\langle \theta_t, x^* - x_t \right\rangle + \eta_t^2 \left\|\theta_t\right\|_*^2 + 2G^2\eta_t^2 \\ &\leq \eta_t \left\langle \theta_t^u + \theta_t^b, x^* - x_t \right\rangle + 2\eta_t^2 \left\|\theta_t^u\right\|_*^2 + 2\eta_t^2 \left\|\theta_t^b\right\|_*^2 + 2G^2\eta_t^2.\end{aligned}$$

This is what we want to show. □

Proof of Lemma 5.4.4. We have

$$\begin{aligned}&\mathbb{E}\left[\exp\left(Z_t\right) \mid \mathcal{F}_{t-1}\right] \times \exp\left(\left(\frac{3}{8\lambda_t^2} + 24 z_t^2 \eta_t^4 \lambda_t^2\right) \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right) \\ &\overset{(a)}{\leq} \mathbb{E}\left[\exp\left(z_t \left(\eta_t \left\langle x^* - x_t, \theta_t^u \right\rangle + 2\eta_t^2 \left(\left\|\theta_t^u\right\|_*^2 - \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right)\right)\right) \mid \mathcal{F}_{t-1}\right] \\ &\overset{(b)}{\leq} \exp\left(\mathbb{E}\left[\frac{3}{4}\left(z_t \left(\eta_t \left\langle x^* - x_t, \theta_t^u \right\rangle + 2\eta_t^2 \left(\left\|\theta_t^u\right\|_*^2 - \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right)\right)\right)^2 \mid \mathcal{F}_{t-1}\right]\right) \\ &\overset{(c)}{\leq} \exp\left(\left(\frac{3}{2} z_t^2 \eta_t^2 \left\|x^* - x_t\right\|^2 \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right] + 6 z_t^2 \eta_t^4 \mathbb{E}\left[\left\|\theta_t^u\right\|_*^4 \mid \mathcal{F}_{t-1}\right]\right)\right) \\ &\overset{(d)}{\leq} \exp\left(\left(\frac{3}{2} z_t^2 \eta_t^2 \left\|x^* - x_t\right\|^2 + 24 z_t^2 \eta_t^4 \lambda_t^2\right) \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right) \\ &\overset{(e)}{\leq} \exp\left(\left(\frac{3}{8\lambda_t^2} + 24 z_t^2 \eta_t^4 \lambda_t^2\right) \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right).\end{aligned}$$

For (a), we use Lemma 5.4.3. For (b), we use Lemma 5.2.2. Notice that

$$\mathbb{E}\left[\left\langle x^* - x_t, \theta_t^u \right\rangle\right] = \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 - \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right] = 0,$$

and since $\|\theta_t^u\|_* \leq 2\lambda_t$, we have

$$\begin{aligned}
&\left|\eta_t \langle x^* - x_t, \theta_t^u \rangle + 2\eta_t^2 \left(\|\theta_t^u\|_*^2 - \mathbb{E}\left[\|\theta_t^u\|_*^2 \mid \mathcal{F}_{t-1}\right]\right)\right| \\
&\leq \eta_t \|x^* - x_t\| \|\theta_t^u\|_* + 2\eta_t^2 \left(\|\theta_t^u\|_*^2 + \mathbb{E}\left[\|\theta_t^u\|_*^2 \mid \mathcal{F}_{t-1}\right]\right) \\
&\leq 2\eta_t\lambda_t \|x^* - x_t\| + 16\eta_t^2\lambda_t^2 \\
&\leq 2\eta_t\lambda_t \sqrt{2\mathbf{D}_\psi(x^*, x_t)} + 16\eta_t^2\lambda_t^2.
\end{aligned}$$

Thus, $z_t \leq \frac{1}{2\eta_t\lambda_t\sqrt{2\mathbf{D}_\psi(x^*,x_t)}+16\eta_t^2\lambda_t^2}$. For (c), we use the inequalities $(a+b)^2 \leq 2a^2 + 2b^2$ and $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \leq \mathbb{E}\left[X^2\right]$. For (d), we use the fact $\|\theta_t^u\|_*^2 \leq 4\lambda_t^2$ to get $\mathbb{E}\left[\|\theta_t^u\|_*^4 \mid \mathcal{F}_{t-1}\right] \leq 4\lambda_t^2\mathbb{E}\left[\|\theta_t^u\|_*^2 \mid \mathcal{F}_{t-1}\right]$. For (e), we use the fact that $\|\theta_t^u\|_* \leq 2\lambda_t$ and

$$z_t\eta_t \|x^* - x_t\| \leq \frac{\eta_t \|x^* - x_t\|}{2\eta_t\lambda_t\sqrt{2\mathbf{D}_\psi(x^*, x_t)}} \leq \frac{1}{2\lambda_t}.$$

We obtain $\mathbb{E}[\exp(Z_t) \mid \mathcal{F}_{t-1}] \leq 1$. Therefore

$$\mathbb{E}[\exp(S_t) \mid \mathcal{F}_{t-1}] = \exp(S_{t-1})\,\mathbb{E}[\exp(Z_t) \mid \mathcal{F}_{t-1}] \leq \exp(S_{t-1}).$$

which means $(\exp(S_t))_{t\geq 1}$ is a supermartingale. By Ville's inequality, we have, for all $k \geq 1$

$$\Pr\left[S_k \geq \log\frac{1}{\delta}\right] \leq \delta\mathbb{E}[\exp(S_1)] \leq \delta.$$

In other words, with probability at least $1-\delta$, for all $k \geq 1$

$$\sum_{t=1}^k Z_t \leq \log\frac{1}{\delta}.$$

Plugging in the definition of Z_t we have

$$\begin{aligned}
&\sum_{t=1}^k z_t\eta_t\Delta_{t+1} + \sum_{t=1}^k \left(z_t\mathbf{D}_\psi(x^*, x_{t+1}) - z_t\mathbf{D}_\psi(x^*, x_t)\right) \\
\leq &\log\frac{1}{\delta} + \sum_{t=1}^k z_t\eta_t\left\langle x^* - x_t, \theta_t^b\right\rangle + 2\sum_{t=1}^k z_t\eta_t^2\left\|\theta_t^b\right\|_*^2 \\
&+ \sum_{t=1}^k \left(\left(2z_t\eta_t^2 + \frac{3}{8\lambda_t^2} + 24z_t^2\eta_t^4\lambda_t^2\right)\mathbb{E}\left[\|\theta_t^u\|_*^2 \mid \mathcal{F}_{t-1}\right]\right).
\end{aligned}$$

Note that we have z_t is a decreasing sequence, hence the LHS of the above inequality can be bounded by

$$\begin{aligned}
\text{LHS} &= \sum_{t=1}^k z_t\eta_t\Delta_{t+1} + z_k\mathbf{D}_\psi(x^*, x_{k+1}) - z_1\mathbf{D}_\psi(x^*, x_1) + \sum_{t=2}^k (z_{k-1} - z_k)\,\mathbf{D}_\psi(x^*, x_k) \\
&\geq \sum_{t=1}^k z_t\eta_t\Delta_{t+1} + z_k\mathbf{D}_\psi(x^*, x_{k+1}) - z_1\mathbf{D}_\psi(x^*, x_1).
\end{aligned}$$

We obtain from here the desired inequality. □

Proof of Proposition 5.4.5. We will prove by induction that on k

$$\sum_{i=1}^{k} \eta_i \Delta_{i+1} + \mathbf{D}_\psi\left(x^*, x_{k+1}\right) \leq \frac{1}{2}\left(R_1 + 8AC_1\right)^2.$$

The base case $k = 0$ is trivial. We have $\mathbf{D}_\psi\left(x^*, x_1\right) = \frac{R_1^2}{2}$. Suppose the statement is true for all $t \leq k \leq \ell$. Now, we show for $k+1$. Recall that

$$z_t = \frac{1}{2\eta_t\lambda_t \max_{i\leq t}\sqrt{2\mathbf{D}_\psi\left(x^*, x_i\right)} + 16Q\eta_t^2\lambda_t^2}.$$

Let us choose $Q = A > 1$. By the induction hypothesis, we have $\max_{i\leq t}\sqrt{2\mathbf{D}_\psi\left(x^*, x_i\right)} \leq R_1 + 8AC_1$, which implies

$$z_k \geq \frac{1}{2\eta_k\lambda_k\left(R_1 + 8AC_1\right) + 16A\eta_k^2\lambda_k^2} = \frac{1}{2C_1\left(R_1 + 16AC_1\right)}.$$

For an upperbound, since $\sqrt{2\mathbf{D}_\psi\left(x^*, x_1\right)} = R_1$, we have:

$$z_t \leq \frac{1}{2C_1\left(R_1 + 8AC_1\right)}.$$

Since z_k is a decreasing sequence, we have

$$z_k\sum_{t=1}^{k}\eta_t\Delta_{t+1} + z_k\mathbf{D}_\psi\left(x^*, x_{k+1}\right) \leq z_1\mathbf{D}_\psi\left(x^*, x_1\right) + \log\frac{1}{\delta} + \sum_{t=1}^{k} z_t\eta_t\left\langle x^* - x_t, \theta_t^b\right\rangle + 2\sum_{t=1}^{k} z_t\eta_t^2\left\|\theta_t^b\right\|_*^2$$
$$+ \sum_{t=1}^{k}\left(\left(2z_t\eta_t^2 + \frac{3}{8\lambda_t^2} + 24z_t^2\eta_t^4\lambda_t^2\right)\mathbb{E}\left[\|\theta_t^u\|_*^2 \mid \mathcal{F}_{t-1}\right]\right).$$

By the choice of λ_t, for all $t \leq k$, $\|\nabla f(x_t)\|_* \leq \frac{\lambda_t}{2}$, we can apply Lemma 5.2.1 and have

$$\left\|\theta_t^b\right\|_* \leq 4\sigma^p\lambda_t^{1-p};$$
$$\mathbb{E}\left[\|\theta_t^u\|_*^2 \mid \mathcal{F}_{t-1}\right] \leq 40\sigma^p\lambda_t^{2-p}.$$

Thus, we have

$$
\begin{aligned}
&z_k \sum_{t=1}^{k} \eta_t \Delta_{t+1} + z_k \mathbf{D}_\psi\left(x^*, x_{k+1}\right) \\
\leq & z_1 \mathbf{D}_\psi\left(x^*, x_1\right) + \log\frac{1}{\delta} + 4\sum_{t=1}^{k} z_t \eta_t \sigma^p \lambda_t^{1-p} \sqrt{2\mathbf{D}_\psi\left(x^*, x_t\right)} + 32 \sum_{t=1}^{k} z_t \eta_t^2 \sigma^{2p} \lambda_t^{2-2p} \\
&+ 40 \sum_{t=1}^{k} \left(\left(2 z_t \eta_t^2 + \frac{3}{8\lambda_t^2} + 24 z_t^2 \eta_t^4 \lambda_t^2 \right) \sigma^p \lambda_t^{2-p} \right) \\
\leq & z_1 \mathbf{D}_\psi\left(x^*, x_1\right) + \log\frac{1}{\delta} + \frac{2C_1\left(R_1 + 8AC_1\right)\sigma^p}{C_1\left(R_1 + 8AC_1\right)} \sum_{t=1}^{k} \left(\frac{1}{\lambda_t}\right)^p + \frac{16 C_1^2 \sigma^{2p}}{C_1\left(R_1 + 8AC_1\right)} \sum_{t=1}^{k} \left(\frac{1}{\lambda_t}\right)^{2p} \\
&+ 40 \left(\frac{C_1^2}{C_1\left(R_1 + 8AC_1\right)} + \frac{3}{8} + \frac{6C_1^4}{C_1^2\left(R_1 + 8AC_1\right)^2} \right) \sigma^p \sum_{t=1}^{k} \left(\frac{1}{\lambda_t}\right)^p \\
\leq & \frac{R_1^2}{4\left(C_1 R_1 + 8AC_1^2\right)} + \log\frac{1}{\delta} + 2\sigma^p C_2 + \frac{2\sigma^{2p} C_2 C_3}{A} + 24\sigma^p C_2 \\
\leq & \frac{R_1^2}{4\left(C_1 R_1 + 8AC_1^2\right)} + A,
\end{aligned}
$$

where for the last inequality we use $\sum_{t=1}^{k} \left(\frac{1}{\lambda_t}\right)^p \leq C_2$ and $\left(\frac{1}{\lambda_t}\right)^{2p} \leq C_3 \left(\frac{1}{\lambda_t}\right)^p$. We obtain

$$
\begin{aligned}
\sum_{t=1}^{k} \eta_t \Delta_{t+1} + \mathbf{D}_\psi\left(x^*, x_{k+1}\right) &\leq 2C_1\left(R_1 + 16AC_1\right)\left(\frac{R_1^2}{4\left(C_1 R_1 + 8AC_1^2\right)} + A \right) \\
&= \frac{1}{2} R_1^2 + \frac{4AC_1^2 R_1^2}{C_1 R_1 + 8AC_1^2} + 2A\left(C_1 R_1 + 16AC_1^2\right) \\
&\leq \frac{1}{2} R_1^2 + 6AC_1 R_1 + 32A^2 C_1^2 \\
&\leq \frac{1}{2}\left(R_1 + 8AC_1\right)^2.
\end{aligned}
$$

□

Proof of Theorem 5.4.1. Note that our choice of η ensures $\eta \leq \frac{R_1}{16}\frac{1}{4LR_1} \leq \frac{1}{4L}$. We have that with probability at least $1-\delta$, event $E(\delta)$ happens. Conditioning on this event, in 5.4.5 we choose

$$
C_1 = \frac{R_1}{24\gamma}; \quad C_2 = \frac{\gamma}{26\sigma^p}; \quad C_3 = \frac{\gamma}{26T\sigma^p}; \quad A = 3\gamma.
$$

We have

$$\begin{aligned}\lambda_t\eta_t &= C_1\\ \sum_{t=1}^{T}\left(\frac{1}{\lambda_t}\right)^p &\le \sum_{t=1}^{T}\left(\frac{\gamma}{26T}\right)\frac{1}{\sigma^p} = C_2\\ \left(\frac{1}{\lambda_t}\right)^{2p} &\le \frac{1}{\sigma^p}\left(\frac{\gamma}{26T}\right)\left(\frac{1}{\lambda_t}\right)^p = C_3\left(\frac{1}{\lambda_t}\right)^p\\ \max\left\{\log\frac{1}{\delta} + 26\sigma^p C_2 + \frac{2\sigma^{2p}C_2C_3}{A}; 1\right\} &\le 3\gamma = A.\end{aligned}$$

We only need to show that for all t

$$\|\nabla f(x_t)\|_* \le \frac{\lambda_t}{2}.$$

We will show this by induction. Indeed, we have

$$\|\nabla f(x_1)\|_* \le \nabla_1 \le \frac{\lambda_1}{2}.$$

Suppose that it is true for all $t \le k$. We prove that

$$\|\nabla f(x_{k+1})\|_* \le \frac{\lambda_{k+1}}{2}.$$

By 5.4.5 we have

$$\|x_{k+1} - x^*\| \le \sqrt{2\mathbf{D}_\psi\left(x^*, x_{k+1}\right)} \le R_1 + 8AC_1 = 2R_1.$$

Thus

$$\begin{aligned}\|\nabla f(x_{k+1})\|_* &\le \|\nabla f(x_{k+1}) - \nabla f(x^*)\|_* + \|\nabla f(x_1) - \nabla f(x^*)\|_* + \|\nabla f(x_1)\|_*\\ &\le L\|x_{k+1} - x^*\| + L\|x_1 - x^*\| + \nabla_1\\ &\le 3LR_1 + \nabla_1 \le \frac{\lambda_{k+1}}{2}\end{aligned}$$

as needed. Therefore from Lemma 5.4.4 we have

$$\eta\sum_{t=1}^{T}\Delta_{t+1} + \mathbf{D}_\psi\left(x^*, x_{T+1}\right) \le 2R_1^2,$$

which gives

$$\frac{1}{T}\sum_{t=2}^{T+1}\Delta_t \le \frac{2R_1^2}{\eta} = 48R_1\max\left\{26^{\frac{1}{p}}T^{\frac{1-p}{p}}\sigma\gamma^{\frac{p-1}{p}}; 2\left(3LR_1 + \nabla_1\right)T^{-1}\gamma\right\}.$$

□

Theorem 5.8.2. *Assume that f satisfies Assumption (1), (2), (3), (4) and (5). Let $\gamma = \max\left\{\log\frac{1}{\delta}; 1\right\}$; $R_1 = \sqrt{2\mathbf{D}_\psi\left(x^*, x_1\right)}$ assume that ∇_1 is an upper bound of $\|\nabla f(x_1)\|_*$.*

For unknown T, we choose

$$\lambda_t = \max\left\{\left(\frac{52t\left(1+\log t\right)^2}{\gamma}\right)^{1/p}\sigma; 2\left(3LR_1+\nabla_1\right)\right\}, \text{ and}$$
$$\eta_t = \frac{R_1}{24\lambda_t\gamma} = \frac{R_1}{24\gamma}\min\left\{\left(\frac{52t\left(1+\log t\right)^2}{\gamma}\right)^{-1/p}\sigma^{-1}; \frac{1}{2}\left(3LR_1+\nabla_1\right)^{-1}\right\}.$$

Then with probability at least $1-\delta$

$$\frac{1}{T}\sum_{t=2}^{T+1}\Delta_t \leq 48R_1\max\left\{52^{\frac{1}{p}}T^{\frac{1-p}{p}}\left(1+\log T\right)^{\frac{2}{p}}\sigma\gamma^{\frac{p-1}{p}}; 2\left(3LR_1+\nabla_1\right)T^{-1}\gamma\right\} = \widetilde{O}\left(T^{\frac{1-p}{p}}\right).$$

Proof. We can follow the similar steps. Notice that (η_t) is a decreasing sequence. We also use Fact 5.7.4 to verify the second condition of Proposition 5.4.5. The proof is omitted. □

Proof of Theorem 5.4.2. Note that $\eta_t \leq \frac{1}{4L}$. We have that with probability at least $1-\delta$, event $E(\delta)$ happens. Conditioning on this event, in 5.4.5. We choose

$$C_1 = \frac{c_1}{24}; \quad C_2 = \frac{1}{26c_2}; \quad C_3 = \frac{1}{52c_2}; \quad A = \gamma + \frac{2\sigma^p}{c_2}.$$

We verify the conditions of Proposition 5.4.5

$$\begin{aligned}
\lambda_t\eta_t &= C_1\\
\sum_{t=1}^{T}\left(\frac{1}{\lambda_t}\right)^p &\leq \sum_{t=1}^{T}\frac{1}{52t(1+\log t)^2c_2} \leq \frac{1}{26c_2} = C_2\\
\left(\frac{1}{\lambda_t}\right)^{2p} &\leq \frac{1}{52tc_2}\left(\frac{1}{\lambda_t}\right)^p \leq C_3\left(\frac{1}{\lambda_t}\right)^p\\
\max\left\{\log\frac{1}{\delta} + 26\sigma^pC_2 + \frac{2\sigma^{2p}C_2C_3}{A}; 1\right\} &= \max\left\{\log\frac{1}{\delta} + \frac{\sigma^p}{c_2} + \frac{\sigma^p}{c_2}; 1\right\} \leq A,
\end{aligned}$$

where we have $\frac{2\sigma^{2p}C_2C_3}{A} \leq 2\sigma^{2p}C_2C_3 \times \frac{c_2}{2\sigma^p} \leq \frac{\sigma^p}{c_2}$. Also, note that

$$\begin{aligned}
\left\|\nabla f(x_t)\right\|_* &\leq \left\|\nabla f(x_t) - \nabla f(x_1)\right\|_* + \left\|\nabla f(x_1)\right\|_*\\
&\leq L\left\|x_t - x_1\right\|_* + \left\|\nabla f(x_1)\right\|_* \leq \frac{\lambda_t}{2}.
\end{aligned}$$

Therefore, from Lemma 5.4.4, we have

$$\begin{aligned}
\eta_T\sum_{t=1}^{T}\Delta_{t+1} + \mathbf{D}_\psi\left(x^*, x_{T+1}\right) &\leq \frac{1}{2}\left(R_1 + 8AC_1\right)^2\\
&= \frac{1}{2}\left(R_1 + \frac{c_1}{3}\left(\gamma + \frac{2\sigma^p}{c_2}\right)\right)^2
\end{aligned}$$

which gives

$$\begin{aligned}\frac{1}{T}\sum_{t=2}^{T+1}\Delta_t &\leq \frac{1}{2T\eta_T}\left(R_1+\frac{c_1}{3}\left(\gamma+\frac{2\sigma^p}{c_2}\right)\right)^2\\ &= \frac{8}{Tc_1}\left(R_1+\frac{c_1}{3}\left(\gamma+\frac{2\sigma^p}{c_2}\right)\right)^2\\ &\quad\cdot\max\left\{\left(52T(1+\log T)^2c_2\right)^{1/p};2\left(L\max_{i\leq T}\|x_i-x_1\|+\nabla_1\right);\frac{L}{8}\right\}.\end{aligned}$$

Note that

$$\begin{aligned}\|x_i-x_1\| &\leq \|x_i-x^*\|+\|x_1-x^*\|\\ &\leq 2R_1+\frac{c_1}{3}\left(\gamma+\frac{2\sigma^p}{c_2}\right)\end{aligned}$$

which gives us the final convergence rate. □

5.9 Clipped Accelerated Stochastic Mirror Descent

In this section, we extend the analysis of Clipped-SMD to the case of Clipped Accelerated Stochastic Mirror Descent (Algorithm 8). We will see that the analysis is basically the same with little modification. We present in Algorithm 8 the clipped version of accelerated stochastic mirror descent (see (Lan, 2020)), where the clipped gradient $\widetilde{\nabla}f(x_t)$ is used to update the iterates in place of the stochastic gradient $\widehat{\nabla}f(x_t)$.

We use the following additional assumption:

(5') Global minimizer: We assume that $\nabla f(x^*)=0$.

Theorem 5.9.1. *Assume that f satisfies Assumption (1), (2), (3), (4) and (5'). Let $\gamma = \max\left\{\log\frac{1}{\delta};1\right\}$; and $R_1=\sqrt{2\mathbf{D}_\psi\left(x^*,x_1\right)}$.*

1. For known T, we choose a constant c and λ_t and η_t such that

$$\begin{aligned}c &= \max\left\{10^4;\frac{4\left(T+1\right)\left(\frac{26T}{\gamma}\right)^{\frac{1}{p}}\sigma}{\gamma LR_1}\right\},\\ \lambda_t &= \frac{cR_1\gamma L\alpha_t}{8}=\max\left\{\frac{10^4R_1\gamma L}{6(t+1)};\frac{T+1}{t+1}\left(\frac{26T}{\gamma}\right)^{1/p}\sigma\right\},\\ \eta_t &= \frac{1}{3c\gamma^2L\alpha_t}=\frac{R_1}{24\gamma}\min\left\{\frac{4(t+1)}{10^4R_1\gamma L};\frac{t+1}{T+1}\left(\frac{26T}{\gamma}\right)^{-1/p}\sigma^{-1}\right\}.\end{aligned}$$

Then with probability at least $1-\delta$

$$f\left(y_{T+1}\right)-f\left(x^*\right)\leq 6\max\left\{10^4L\gamma^2R_1^2(T+1)^{-2};4R_1\left(T+1\right)^{-1}\left(26T\right)^{\frac{1}{p}}\gamma^{\frac{p-1}{p}}\sigma\right\}.$$

2. For unknown T, we choose c_t, λ_t and η_t such that

$$c_t = \max\left\{10^4; \frac{4\left(t+1\right)\left(\frac{52t(1+\log t)^2}{\gamma}\right)^{\frac{1}{p}}\sigma}{\gamma L R_1}\right\},$$

$$\lambda_t = \frac{c_t R_1 \gamma L \alpha_t}{8} = \max\left\{\frac{10^4 R_1 \gamma L}{4(t+1)}; \left(\frac{52t\left(1+\log t\right)^2}{\gamma}\right)^{1/p}\sigma\right\},$$

$$\eta_t = \frac{1}{3c_t\gamma^2 L\alpha_t} = \frac{R_1}{24\gamma}\min\left\{\frac{4(t+1)}{10^4 R_1\gamma L}; \left(\frac{52t\left(1+\log t\right)^2}{\gamma}\right)^{-1/p}\sigma^{-1}\right\}.$$

Then with probability at least $1-\delta$

$$f\left(y_{T+1}\right) - f\left(x^*\right) \leq 6\max\left\{10^4 L\gamma^2 R_1^2 (T+1)^{-2}; 4R_1\left(T+1\right)^{-1}\left(52T\left(1+\log T\right)^2\right)^{\frac{1}{p}}\gamma^{\frac{p-1}{p}}\sigma\right\}.$$

Remark 7. *One feature of the accelerated algorithm is the interpolation between the two regimes: When σ is large, the algorithm achieves the $O\left(T^{\frac{1-p}{p}}\right)$ convergence rate, which is the same as unaccelerated algorithms; however, when σ is sufficiently small, the algorithm achieves the accelerated $O\left(T^{-2}\right)$ rate.*

We also start the analysis of accelerated stochastic mirror descent with the following lemma.

Lemma 5.9.2. *Assume that f satisfies Assumption (1), (2), (3), (4) and $\eta_t \leq \frac{1}{2L\alpha_t}$, the iterate sequence $(x_t)_{t\geq 1}$ output by Algorithm 7 satisfies the following*

$$\begin{aligned}
&\frac{\eta_t}{\alpha_t}\left(f\left(y_{t+1}\right) - f\left(x^*\right)\right) - \frac{\eta_t\left(1-\alpha_t\right)}{\alpha_t}\left(f\left(y_t\right) - f\left(x^*\right)\right) + \mathbf{D}_\psi\left(x^*, z_{t+1}\right) - \mathbf{D}_\psi\left(x^*, z_t\right)\\
\leq&\eta_t\left\langle\theta_t^u, x^* - z_t\right\rangle + \eta_t\left\langle\theta_t^b, x^* - z_t\right\rangle + 2\eta_t^2\left(\left\|\theta_t^u\right\|_*^2 - \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right)\\
&+ 2\eta_t^2\left\|\theta_t^b\right\|_*^2 + 2\eta_t^2\mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right].
\end{aligned}$$

Proof of Lemma 5.9.2. We have

$$
\begin{aligned}
f(y_{t+1}) - f(x^*) &= \underbrace{f(y_{t+1}) - f(x_t)}_{\text{smoothness}} + \underbrace{f(x_t) - f(x^*)}_{\text{convexity}} \\
&\leq \langle \nabla f(x_t), y_{t+1} - x_t \rangle + \frac{L}{2} \|y_{t+1} - x_t\|^2 \\
&\quad + \alpha_t \langle \nabla f(x_t), x_t - x^* \rangle + (1-\alpha_t)(f(x_t) - f(x^*)) \\
&= \underbrace{(1-\alpha_t) \langle \nabla f(x_t), y_t - x_t \rangle}_{\text{convexity}} + \alpha_t \langle \nabla f(x_t), z_{t+1} - x^* \rangle \\
&\quad + \frac{L\alpha_t^2}{2} \|z_{t+1} - z_t\|^2 + (1-\alpha_t)(f(x_t) - f(x^*)) \\
&\leq (1-\alpha_t)(f(y_t) - f(x_t)) + (1-\alpha_t)(f(x_t) - f(x^*)) \\
&\quad + \alpha_t \langle \theta_t, x^* - z_{t+1} \rangle + \alpha_t \left\langle \widetilde{\nabla} f(x_t), z_{t+1} - x^* \right\rangle + \frac{L\alpha_t^2}{2} \|z_{t+1} - z_t\|^2 \\
&\leq (1-\alpha_t)(f(y_t) - f(x^*)) + \alpha_t \langle \theta_t, x^* - z_{t+1} \rangle \\
&\quad + \alpha_t \left\langle \widetilde{\nabla} f(x_t), z_{t+1} - x^* \right\rangle + \frac{L\alpha_t^2}{2} \|z_{t+1} - z_t\|^2 .
\end{aligned}
$$

By the optimality condition, we have

$$
\left\langle \eta_t \widetilde{\nabla} f(x_t) + \nabla_x \mathbf{D}_\psi(z_{t+1}, z_t), x^* - z_{t+1} \right\rangle \geq 0
$$

and thus

$$
\left\langle \eta_t \widetilde{\nabla} f(x_t), z_{t+1} - x^* \right\rangle \leq \langle \nabla_x \mathbf{D}_\psi(z_{t+1}, z_t), x^* - z_{t+1} \rangle .
$$

Note that

$$
\begin{aligned}
\langle \nabla_x \mathbf{D}_\psi(z_{t+1}, z_t), x^* - z_{t+1} \rangle &= \langle \nabla \psi(z_{t+1}) - \nabla \psi(z_t), x^* - z_{t+1} \rangle \\
&= \mathbf{D}_\psi(x^*, z_t) - \mathbf{D}_\psi(z_{t+1}, z_t) - \mathbf{D}_\psi(x^*, z_{t+1}) .
\end{aligned}
$$

Thus

$$
\begin{aligned}
\eta_t \left\langle \widetilde{\nabla} f(x_t), z_{t+1} - x^* \right\rangle &\leq \mathbf{D}_\psi(x^*, z_t) - \mathbf{D}_\psi(x^*, z_{t+1}) - \mathbf{D}_\psi(z_{t+1}, z_t) \\
&\leq \mathbf{D}_\psi(x^*, z_t) - \mathbf{D}_\psi(x^*, z_{t+1}) - \frac{1}{2} \|z_{t+1} - z_t\|^2
\end{aligned}
$$

where we have used that $\mathbf{D}_\psi(z_{t+1}, z_t) \geq \frac{1}{2} \|z_{t+1} - z_t\|^2$ by the strong convexity of ψ. We have

$$
\begin{aligned}
f(y_{t+1}) - f(x^*) &\leq (1-\alpha_t)(f(y_t) - f(x^*)) + \alpha_t \langle \theta_t, x^* - z_{t+1} \rangle \\
&\quad + \frac{\alpha_t}{\eta_t} \mathbf{D}_\psi(x^*, z_t) - \frac{\alpha_t}{\eta_t} \mathbf{D}_\psi(x^*, z_{t+1}) + \left(\frac{L\alpha_t^2}{2} - \frac{\alpha_t}{2\eta_t} \right) \|z_{t+1} - z_t\|^2 .
\end{aligned}
$$

Dividing both sides by $\frac{\alpha_t}{\eta_t}$ and using the condition $L\eta_t\alpha_t \leq \frac{1}{2}$, we have

$$\begin{aligned}
&\frac{\eta_t}{\alpha_t}\left(f\left(y_{t+1}\right)-f\left(x^*\right)\right)+\mathbf{D}_\psi\left(x^*, z_{t+1}\right)-\mathbf{D}_\psi\left(x^*, z_t\right)\\
\leq&\frac{\eta_t\left(1-\alpha_t\right)}{\alpha_t}\left(f\left(y_t\right)-f\left(x^*\right)\right)+\eta_t\left\langle\theta_t, x^*-z_t\right\rangle\\
&+\eta_t\left\langle\theta_t, z_t-z_{t+1}\right\rangle-\frac{1-L\eta_t\alpha_t}{2}\left\|z_{t+1}-z_t\right\|^2\\
\leq&\frac{\eta_t\left(1-\alpha_t\right)}{\alpha_t}\left(f\left(y_t\right)-f\left(x^*\right)\right)+\eta_t\left\langle\theta_t, x^*-z_t\right\rangle\\
&+\frac{\eta_t^2\left\|\theta_t\right\|_*^2}{2\left(1-L\eta_t\alpha_t\right)}\\
\leq&\frac{\eta_t\left(1-\alpha_t\right)}{\alpha_t}\left(f\left(y_t\right)-f\left(x^*\right)\right)+\eta_t\left\langle\theta_t^u+\theta_t^b, x^*-z_t\right\rangle\\
&+2\eta_t^2\left\|\theta_t^u\right\|_*^2+2\eta_t^2\left\|\theta_t^b\right\|_*^2
\end{aligned}$$

as needed. □

Similarly to the previous section, we define the following variables

$$\begin{aligned}
Z_t=z_t\Bigg(&\frac{\eta_t}{\alpha_t}\left(f\left(y_{t+1}\right)-f\left(x^*\right)\right)-\frac{\eta_t\left(1-\alpha_t\right)}{\alpha_t}\left(f\left(y_t\right)-f\left(x^*\right)\right)\\
&+\mathbf{D}_\psi\left(x^*, z_{t+1}\right)-\mathbf{D}_\psi\left(x^*, z_t\right)\\
&-\eta_t\left\langle\theta_t^b, x^*-z_t\right\rangle-2\eta_t^2\left\|\theta_t^b\right\|_*^2-2\eta_t^2\mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\Bigg)\\
&-\left(\frac{3}{8\lambda_t^2}+24z_t^2\eta_t^4\lambda_t^2\right)\mathbb{E}\left[\left\|\theta_t^u\right\|^2 \mid \mathcal{F}_{t-1}\right],\\
\text{where } z_t=&\frac{1}{2\eta_t\lambda_t\max_{i\leq t}\sqrt{2\mathbf{D}_\psi\left(x^*, x_i\right)}+16Q\eta_t^2\lambda_t^2}
\end{aligned}$$

for a constant $Q \geq 1$. We also let $S_t = \sum_{i=1}^t Z_i$. Following the same analysis as in previous sections, we can obtain Lemma 5.9.3 and Proposition 5.9.4, for which we will omit the proofs here. The only step we need to pay attention to when showing Lemma 5.9.3 is when we bound the sum

$$\sum_{t=1}^k \frac{z_t\eta_t}{\alpha_t}\left(f\left(y_{t+1}\right)-f\left(x^*\right)\right)-\frac{z_t\eta_t\left(1-\alpha_t\right)}{\alpha_t}\left(f\left(y_t\right)-f\left(x^*\right)\right).$$

If we assume $\frac{\eta_{t-1}}{\alpha_{t-1}} \geq \frac{\eta_t(1-\alpha_t)}{\alpha_t}$, since z_t is a decreasing sequence and $\alpha_1 = 0$, we can lower bound the above sum by the last term $\frac{z_k\eta_k}{\alpha_k}\left(f\left(y_{k+1}\right)-f\left(x^*\right)\right)$, which gives us the desired inequality.

Lemma 5.9.3. *Assume that for all $t \geq 1$, η_t satisfies $\frac{\eta_{t-1}}{\alpha_{t-1}} \geq \frac{\eta_t(1-\alpha_t)}{\alpha_t}$. For any $\delta > 0$, let $E(\delta)$ be the event that for all $1 \leq k \leq T$*

$$\begin{aligned}
&\frac{z_k \eta_k}{\alpha_k}\left(f\left(y_{k+1}\right)-f\left(x^*\right)\right)+z_k \mathbf{D}_\psi\left(x^*, x_{k+1}\right) \\
\leq & z_1 \mathbf{D}_\psi\left(x^*, x_1\right)+\log \frac{1}{\delta}+\sum_{t=1}^k z_t \eta_t\left\langle x^*-x_t, \theta_t^b\right\rangle+2 \sum_{t=1}^k z_t \eta_t^2\left\|\theta_t^b\right\|_*^2 \\
&+\sum_{t=1}^k\left(\left(2 z_t \eta_t^2+\frac{3}{8 \lambda_t^2}+24 z_t^2 \eta_t^4 \lambda_t^2\right) \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right).
\end{aligned}$$

Then $\Pr[E(\delta)] \geq 1-\delta$.

Finally, we state a general condition for the choice of η_t and λ_t, which follows exactly the same as in Proposition 5.4.5. The proof for Theorem 5.9.1 is a direct consequence of this.

Proposition 5.9.4. *We assume that the event $E(\delta)$ from Lemma 5.9.3 happens. Suppose that for some $\ell \leq T$, there are constants C_1 and C_2 such that for all $t \leq \ell$*

1. $\lambda_t \eta_t = C_1$; *2.* $\sum_{t=1}^{\ell}\left(\frac{1}{\lambda_t}\right)^p \leq C_2$; *3.* $\left(\frac{1}{\lambda_t}\right)^{2p} \leq C_3\left(\frac{1}{\lambda_t}\right)^p$; *4.* $\|\nabla f(x_t)\|_* \leq \frac{\lambda_t}{2}$.

Then for all $t \leq \ell+1$

$$\frac{\eta_t}{\alpha_t}\left(f\left(y_{t+1}\right)-f\left(x^*\right)\right)+\mathbf{D}_\psi\left(x^*, z_{t+1}\right) \leq \frac{1}{2}\left(R_1+8 A C_1\right)^2$$

for $A \geq \max\left\{\log \frac{1}{\delta}+26 \sigma^p C_2+\frac{2 \sigma^{2 p} C_2 C_3}{A} ; 1\right\}$.

Proof of Theorem 5.9.1. 1. Note that $\eta_t \leq \frac{1}{2c\gamma^2 L \alpha_t} \leq \frac{1}{2L\alpha_t}$ and

$$\begin{aligned}
\frac{\eta_{t-1}}{\alpha_{t-1}} &= \frac{t^2}{8c\gamma^2 L} \\
\frac{\eta_t\left(1-\alpha_t\right)}{\alpha_t} &= \frac{(t+1)(t-1)}{8c\gamma^2 L}
\end{aligned}$$

thus $\frac{\eta_{t-1}}{\alpha_{t-1}} \geq \frac{\eta_t(1-\alpha_t)}{\alpha_t}$. We have that with probability at least $1-\delta$, event $E(\delta)$ happens. Conditioning on this event, in 5.4.5 We choose

$$C_1 = \frac{R_1}{24\gamma}; \quad C_2 = \frac{\gamma}{26\sigma^p}; \quad C_3 = \frac{\gamma}{26T\sigma^p}; \quad A = 3\gamma.$$

We can verify the conditions of Proposition 5.9.4 similarly as in previous section for these choices of C_1, C_2, and C_3.

We will show by induction that for all $t \geq 1$, $\|\nabla f(x_t)\|_* \leq \frac{\lambda_t}{2}$ and

$$\max\left\{\left\|x_t-x^*\right\|,\left\|y_t-x^*\right\|,\left\|z_t-x^*\right\|\right\} \leq 2R_1.$$

For $t=1$, notice that $x_1 = y_1 = z_1$. Thus, we have

$$\|\nabla f(x_1)\|_* = \|\nabla f(x_1) - \nabla f(x^*)\|_* \leq LR_1 \leq \frac{\lambda_1}{2}.$$

Now assume that the claim holds for $1 \leq t \leq k$. By Proposition 5.9.4, we know that

$$\frac{2\eta_k}{\alpha_k} f\left(y_{k+1}\right) - f\left(x^*\right) + \left\|z_{k+1} - x^*\right\|^2 \leq 4R_1^2.$$

Furthermore

$$\begin{aligned}
\left\|y_{k+1} - x^*\right\| &\leq (1-\alpha_k)\left\|y_k - x^*\right\| + \alpha_k \left\|z_{k+1} - x^*\right\| \leq 2R_1 \\
\left\|x_{k+1} - x^*\right\| &\leq (1-\alpha_k)\left\|y_{k+1} - x^*\right\| + \alpha_k \left\|z_{k+1} - x^*\right\| \leq 2R_1
\end{aligned}$$

For $k \geq 1$ we have $\alpha_{k+1} = \frac{2}{k+2} < 1$; $\frac{\alpha_{k+1}}{1-\alpha_{k+1}} = \frac{2}{k} \leq \frac{4}{k+2} \leq 2\alpha_{t+1}$ and $\alpha_t \leq \frac{3}{2}\alpha_{t+1}$. Hence,

$$\begin{aligned}
\left\|\nabla f(x_{k+1})\right\|_* &\leq \left\|\nabla f(x_{k+1}) - \nabla f(y_{k+1})\right\|_* + \left\|\nabla f(y_{k+1}) - \nabla f(x^*)\right\|_* \\
&\leq L\left\|x_{k+1} - y_{k+1}\right\| + \sqrt{2L\left(f\left(y_{k+1}\right) - f\left(x^*\right)\right)} \\
&\leq \frac{L\alpha_{k+1}\left\|x_{k+1} - z_{k+1}\right\|}{1-\alpha_{k+1}} + 2R_1\sqrt{\frac{L\alpha_t}{2\eta_t}} \\
&\leq 4LR_1\frac{\alpha_{k+1}}{1-\alpha_{k+1}} + 2\sqrt{\frac{3}{2}c}\gamma R_1 L\alpha_t \\
&\leq 8\gamma L R_1\alpha_{t+1} + 3\sqrt{\frac{3}{2}c}\gamma L R_1\alpha_{t+1} \\
&\leq (8 + 3\sqrt{\frac{3}{2}c})R_1\gamma L\alpha_{t+1} \\
&= \frac{16(8 + 3\sqrt{\frac{3}{2}c})\lambda_{t+1}}{2c} \leq \frac{\lambda_{t+1}}{2}
\end{aligned}$$

as needed. Therefore, we have

$$\frac{\eta_T}{\alpha_T}\left(f\left(y_{T+1}\right) - f\left(x^*\right)\right) + \mathbf{D}_\psi\left(x^*, x_{T+1}\right) \leq 2R_1^2$$

which gives

$$\begin{aligned}
f\left(y_{T+1}\right) - f\left(x^*\right) &\leq \frac{2R_1^2\alpha_T}{\eta_T} = 6R_1^2 c\gamma^2 L\alpha_T^2 \\
&= 6\max\left\{10^4 L\gamma^2 R_1^2 (T+1)^{-2}; 6R_1\left(T+1\right)^{-1}\left(26T\right)^{\frac{1}{p}}\gamma^{\frac{p-1}{p}}\sigma\right\}.
\end{aligned}$$

2. Following the similar steps to the proof of Theorem 5.9.1, and noticing that (c_t) is a increasing sequence, we obtain the convergence rate. □

Part II

Practice

Chapter 6

Introduction

6.1 Introduction

Adaptive optimizers like Adam (Kingma and Ba, 2014), AdaGrad (Duchi et al., 2011), and RMSProp (Tieleman, Hinton, et al., 2012) are widely used for training large-scale deep neural networks but require significant memory for storing momentum and adaptive step size states, often doubling the model's memory footprint. As the size of deep neural networks continues to grow, especially with large-language models (LLMs), reducing the memory consumption of optimizer states has become crucial. Recent approaches, including quantization (Li et al., 2024a; Dettmers et al., 2021; Dettmers et al., 2024), low-rank decomposition (Hu et al., 2021; Lialin et al., 2023; Zhao et al., 2024; Shazeer and Stern, 2018), and sketching-based dimensionality reduction (Muhamed et al., 2024; Hao et al., 2024), aim to address this issue. However, these methods often lack theoretical guarantees, compromise performance, or require extensive tuning, especially in pretraining tasks. Our work addresses these challenges by developing methods that attempt to reduce the theory-practice gap and advance the cost-performance trade-off of algorithms for training DNNs.

6.1.1 The Anatomy of Common Optimizers

We provide a generic template for adaptive optimizers in Algorithm 9, which captures a broad range of first-order optimizers that leverage either momentum or adaptive step sizes. As detailed in Table 6.1, many standard optimizers can be represented within this framework by varying the choices of momentum and adaptive step-size terms.

Algorithm 9 Generic Template for Stochastic Adaptive Optimizers with Momentum

Require: Initial point $x_1 \in \mathbb{R}^d$, base step size $\eta > 0$, and constant $\epsilon > 0$.

for $t = 1$ to T **do**
  Obtain stochastic gradient $\widehat{\nabla} f(x_t)$
  $m_t = \text{update\_momentum}\left(\widehat{\nabla} f(x_t); m_{t-1}\right)$ ▷ Update momentum
  $v_t^2 = \text{update\_adaptive\_stepsize}\left(\widehat{\nabla} f(x_t); v_{t-1}^2\right)$ ▷ Update adaptive step size.
  $x_{t+1} = x_t - \eta \cdot \frac{m_t}{v_t + \epsilon}$ ▷ Update step. Division is element-wise.
end for

TABLE 6.1: Update rules for common optimizers in the framework of Algorithm 9. We omit bias correction terms and numerical stabilizer ϵ for simplicity. Memory for optimizer state is shown for model of size d.

Optimizer	Memory	update_adaptive_stepsize	update_momentum
Adam	$2d$	$\beta_2 v_{t-1}^2 + (1-\beta_2)\cdot \widehat{\nabla} f(x_t)^2$	$\beta_1 m_{t-1} + (1-\beta_1)\widehat{\nabla} f(x_t)$
SGDm	d	N/A	$\beta m_{t-1} + (1-\beta)\widehat{\nabla} f(x_t)$
AdaGrad	d	$v_{t-1}^2 + \widehat{\nabla} f(x_t)^2$	$\widehat{\nabla} f(x_t)$
AdaGrad-Norm	1	$v_{t-1}^2 + \left\|\widehat{\nabla} f(x_t)\right\|^2$	$\widehat{\nabla} f(x_t)$
RMSProp	d	$\sqrt{\beta_1 v_{t-1}^2 + (1-\beta_1)\cdot \widehat{\nabla} f(x_t)^2}$	$\widehat{\nabla} f(x_t)$
SGD	1	N/A	$\widehat{\nabla} f(x_t)$

6.2 Contributions and Overview

We aim to reduce memory consumption while maintaining strong performance and theoretical guarantees. To this end, we introduce two memory-efficient optimization algorithms for large-scale DNN training: **Subset-Norm** (**SN**) for adaptive step-size memory reduction (Chapter 7) and **Subspace-Momentum** (**SM**) for momentum compression (Chapter 8). We first present the algorithms, motivations, and theoretical analysis, then we present our extensive experimental results for both algorithms in Chapter 9. While existing approaches trade performance for memory savings, our theoretically-grounded methods achieve both a reduced memory footprint and faster training:

- **Subset-Norm (SN)**: A memory-efficient adaptive step-size algorithm with high-probability convergence guarantees for non-convex objectives under coordinate-wise sub-gaussian noise. By unifying AdaGrad-Coordinate's and AdaGrad-Norm's analysis, we show that the SN adaptive step size (Algorithm 10) achieves improved dimensional dependence, while reducing the memory footprint from $O(d)$ to roughly $O(\sqrt{d})$. On LLaMA models' pretraining tasks, SN step sizes achieves better perplexity than coordinate-wise step size across a range of optimizers and model sizes, while using significantly less memory and introducing minimal additional hyperparameters.[1]
- **Subspace-Momentum (SM)**: A momentum compression method that applies momentum in a chosen subspace and SGD in the orthogonal complement with high-probability convergence guarantees under sub-gaussian noise for non-convex smooth objectives. When combined with SN, our method (SNSM) reduces the memory footprint of Adam and AdaGrad+m from $2d$ to $k + \sqrt{d}$ (see Table 9.2) while delivers improved training speed and performance.[2]

Empirical evaluations on LLaMA models from 60M to 1B parameters demonstrate that our algorithms scale effectively and attain better performances than existing optimizers.

[1] Although the subset size can be tuned (Section 9.4.1), we provide a heuristic in Section 7.5 that works effectively across model sizes, eliminating the need for additional tuning.

[2] Typically, k is chosen to be around $d/4$.

6.3 Related Works

As model sizes grow, memory-efficient training techniques have become crucial. Following up on AdaFactor (Shazeer and Stern, 2018), low-rank methods like Galore (Zhao et al., 2024), LoRA (Hao et al., 2024), and ReLORA (Lialin et al., 2023) reduce memory usage by approximating large weight matrices with low-rank representations. Projection-based approaches, such as GRASS (Muhamed et al., 2024) and FLORA (Hao et al., 2024), compress gradients or combine low-rank ideas with projections to reduce memory requirements. Recently, AdaMeM (Vyas et al., 2024a) proposes to incorporate the orthogonal subspace to the AdaFactor optimizer; this is related to but different from our simpler SM algorithms, where we use subspace decompositions to decouple the momentum and SGD. BAdam (Luo et al., 2024), a block coordinate descent method that utilizes Adam as an inner solver, has been proposed for fine-tuning large language models. Very recently, Adam-mini (Zhang et al., 2024) also uses shared step sizes as Subset-Norm; however, the partition strategy is quite different and mostly empirical. In contrast to our proposed methods, these methods are largely heuristic-driven and often lack convergence guarantees under standard assumptions. On the other hand, methods like SM3 (Anil et al., 2019), which uses subset (cover) statistics to show convergence in online learning, and MicroAdam (Modoranu et al., 2024), which provides convergence guarantees for a gradient compression scheme with error correction, offer theoretical guarantees.

Additional approaches to reducing memory during training include optimizer quantization (Li et al., 2024a; Dettmers et al., 2021; Dettmers et al., 2024), attention computation compression/optimization (Wu et al., 2022; Dao et al., 2022; Dao, 2023; Shah et al., 2024), activation checkpointing (Chen et al., 2016), and distributed training (Rajbhandari et al., 2020). For inference, compression techniques are also actively being explored (Sakr and Khailany, 2024; Dettmers et al., 2022; Xiao et al., 2024; Lin et al., 2024; Frantar et al., 2023). These are orthogonal directions to our work and can be combined.

Another orthogonal direction is approximated second-order optimization, where one aims to approximate the Hessian preconditioner using only first-order information in order to achieve faster convergence. Some works in this area include (Gupta et al., 2018; Liu et al., 2023a; Vyas et al., 2024b). These methods typically demonstrate faster training but at the cost of super-linear memory and additional computational overhead.

Convergence analysis of non-convex optimization methods has seen significant progress, with recent works providing convergence proofs for adaptive algorithms like Adam (Li et al., 2024b; Défossez et al., 2022). Numerous studies have explored convergence properties of various adaptive and stochastic gradient methods (Chen et al., 2018; Défossez et al., 2022; Ene and Nguyen, 2021; Liu et al., 2023c; Liu et al., 2023b; Ward et al., 2019; Zou et al., 2019; Reddi et al., 2018; Nesterov, 1983), while lower bound analyses (Arjevani et al., 2023) have highlighted fundamental limits in non-convex optimization. Here, obtaining convergence results for EMA updates (Adam style) for subset-norm and under further relaxed assumptions like affine smoothness (Wang et al., 2023; Attia and Koren, 2023), affine noise (Hong and Lin, 2024; Faw et al., 2022), heavy-tailed noise (Zhang et al., 2019; Zhang et al., 2020; Nguyen et al., 2023a; Nguyen et al., 2023b) are of great interest.

Chapter 7

Subset-Norm

7.1 Introduction

Insights from high-probability convergence analysis of AdaGrad and AdaGrad-Norm reveal the importance of interactions between gradient noise and the adaptive step-size state. Specifically, parameter grouping in AdaGrad-Norm demonstrates a better dependency on the noise parameter when the gradient noise is dense. Building on this idea, we propose Subset-Norm, which introduces flexible parameter-grouping schemes for adaptive learning rates. Instead of using a single scalar learning rate for all coordinates (memory $O(1)$) as in AdaGrad-Norm, or separate learning rates for each coordinate (memory $O(d)$) as in AdaGrad-Coordinate, Subset-Norm uses separate adaptive learning rates for different parameters *groups* or *subsets* (memory $O(d/k)$ where $k =$ #subsets). Our existing flexible analysis scheme for AdaGrad and AdaGrad-Norm generalizes to obtain a high-probability convergence guarantee for Subset-Norm adaptive step size for *any partition*. Our analysis shows that under a vast range of coordinate noise density, simple but general partitioning schemes (memory $O(\sqrt{d})$) can yield *improved dimensional dependence* of the convergence rate for Subset-Norm over AdaGrad-Norm and AdaGrad-Coordinate. This is important as models continue to increase in size.

7.2 Subset-Norm Adaptive Step Size

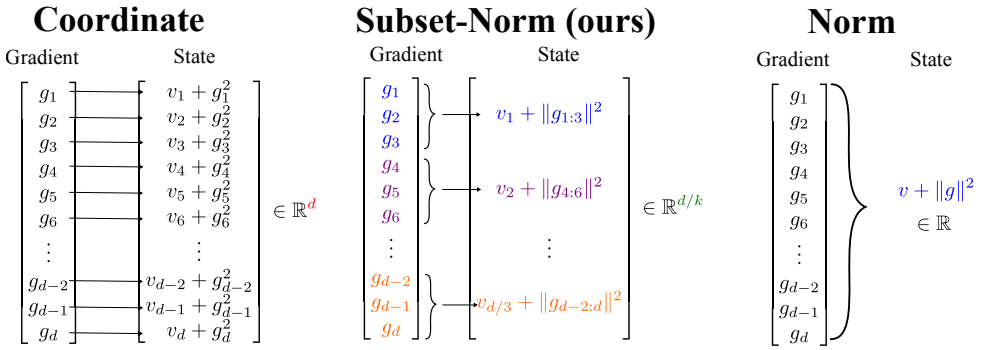

FIGURE 7.1: AdaGrad variants: Coordinate, Subset-Norm, and Norm. Subset-Norm generalizes Coordinate ($k = 1$) and Norm ($k = d$).

We compress the second moment adaptive step size by partitioning parameters into subsets for which they share the same adaptive step size as AdaGrad-Norm (McMahan and Streeter, 2010; Ward et al., 2019). Formally, we need to specify a partition function $\psi : [d] \twoheadrightarrow [c]$ that partitions the d coordinates into c non-empty subsets

Algorithm 10 SGD with Subset-Norm Adaptive Step Size

Require: Initial point $x_1 \in \mathbb{R}^d$, base step size $\eta > 0$, function $\psi : [d] \to [c]$ that partitions the coordinates into c subsets $\Psi_i = \psi^{-1}(i) \subset [d]$, where $\coprod_{i=1}^{c} \Psi_i = [d]$, and $b_{0,i} > 0$ for $i \in [c]$.

1: **for** $t = 1$ to T **do**
2: Obtain stochastic gradient $\widehat{\nabla} f(x_t)$
3: $b_{t,i}^2 = b_{t-1,i}^2 + \left\| \widehat{\nabla}_{\Psi_i} f(x_t) \right\|^2$, for $i \in [c]$ ▷ Update accumulated gradient norms
4: $x_{t+1,k} = x_{t,k} - \frac{\eta}{b_{t,\psi(k)}} \widehat{\nabla}_j f(x_t)$, for $k \in [d]$ ▷ Update coordinates
5: **end for**

$\Psi_i = \psi^{-1}(i) \subset [d]$, where $\coprod_{i=1}^{c} \Psi_i = [d]$. For example, one can pick $\psi(j) = (j/c) \mod k$ to get consecutive equipartitioned subsets $\Psi_i = \{ik, ik+1, \ldots, ik+(k-1)\}$ for some subset-size $k \in \mathbb{N}$ so that $kc = d$.[1]

Given a stochastic gradient $\widehat{\nabla} f(x_t) \in \mathbb{R}^d$ at time t for parameter x_t, we denote $\widehat{\nabla}_{\Psi_i} f(x_t) \in \mathbb{R}^k$ to be the subset of the coordinates of the stochastic gradient with respect to the subset Ψ_i. For example, given $\psi(j) = (j/c) \mod k$ as above, we have $\left(\widehat{\nabla}_{\Psi_i} f(x_t)\right)_j = \widehat{\nabla}_{ik+j-1} f(x_t)$. Similarly, we can define $\nabla_{\Psi_i} f(x_t) \in \mathbb{R}^{|\Psi_i|}$ to be $\frac{\partial f(x_t)}{\partial x_{\Psi_i}}$. We define the *subset-norm adaptive step size* $b_{t,i}$ for subset Ψ_i and the update rule for x_{t+1}:

$$b_{t,i}^2 = b_{t-1,i}^2 + \left\| \widehat{\nabla}_{\Psi_i} f(x_t) \right\|^2 = b_0^2 + \sum_{j=1}^{t} \left\| \widehat{\nabla}_{\Psi_i} f(x_t) \right\|^2, \; i = 0, 1, \ldots, c-1$$

$$x_{t+1,j} = x_{t,j} - \frac{\eta}{b_{t,\psi(j)}} \widehat{\nabla}_j f(x_t), \text{ for } j = 0, 1, \ldots, d-1. \tag{7.1}$$

Note that choosing $c = d$ and $c = 1$ recovers AdaGrad-Coordinate and AdaGrad-Norm, respectively.

7.3 High Probability Convergence of Subset-Norm

We have the following high probability convergence result for the subset-norm adaptive step size:

Theorem 7.3.1. *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and lower bounded by f_*. Given unbiased stochastic gradients $\widehat{\nabla} f(x_t)$ with stochastic gradient noise $\xi_t := \widehat{\nabla} f(x_t) - \nabla f(x_t)$ that is σ_i-per-coordinate subgaussian for $i \in [d]$. For partitions of the parameters into disjoint subsets $[d] = \bigcup_{i=0}^{c-1} \Psi_i$ with $\Psi_i \cap \Psi_j = \emptyset$, for $i \neq j$, the iterates x_t given by Algorithm 10 satisfies the following inequality with probability at least $1 - O(c\delta)$ (for failure probability $\delta > 0$)*

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla f(x_t)\|_2^2 \leq G(\delta) \cdot \tilde{O}\left(\frac{\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\|}{\sqrt{T}} + \frac{\|\sigma\|_2^2 + \sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| + Lc}{T} \right), \text{ where}$$

$$G(\delta) := \tilde{O}\left(\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\|^4 + \sigma_{\max} \|\sigma\|_2^2 + cL + c^{3/2} \sigma_{\max} \right).$$

[1] We use this strategy in all our implementations for simplicity.

Polylog terms are hidden in Theorem 7.3.1 for simplicity. The full result, Theorem 7.6.1, and proofs are presented in Section 7.6. Theorem 7.3.1 provides guarantee for all partitions of the parameters into arbitrary disjoint subsets and generalizes AdaGrad-Norm ($c = 1$) and AdaGrad-Coordinate ($c = d$) results. The result is noise-adapted: if $\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\|$ is small enough, the rate becomes the optimal deterministic rate of $O(\frac{1}{T})$. The next section explores implications of Theorem 7.3.1.

7.4 Coordinate-Noise Density and Dimensional Dependency

Theorem 7.3.1 presents trade-offs between the number of subsets c, and stochastic gradient noise. Intuitively, if few coordinates contribute to the total noise, the scalar version is more useful as $\|\sigma_{\Psi_i}\|^2$ is small for most subsets. However, when many coordinates contribute to the noise, $\|\sigma_{\Psi_i}\|^2$ can be large for many subsets and become the dominating term.

7.4.1 Coordinate-Noise Density

To make the intuition above concrete, consider a scenario with various coordinate-noise density rate: fix a rate $\beta \in [0, 1]$, some d^β coordinates have noise $\alpha > 0$ while the rest are 0. The rate β controls the density of coordinate noise. When $\beta = 0$, only 1 coordinate have noise. When $\beta = 1$, all coordinates have noise. To get a feel for β's relationship to the fraction of coordinates containing noise, half the coordinates contain noise when $\beta \approx 0.96$ when $d = 60\text{M}$ and $\beta \approx 0.97$ when $d = 10\text{B}$ and $\beta \approx 0.98$ when $d = 10^{15}$ (see also Figure 7.3). Furthermore, α upper bounds all coordinate noise, i.e. $\|\sigma\|_\infty \leq \alpha$, which is common in coordinate-wise analysis (Défossez et al., 2022).

7.4.2 Coordinate Noise Density's Convergence Rate's Derivation

Given $\beta \in [0, 1]$, we can obtain a concrete expression for the convergence rates of various methods (different subset sizes) from Theorem 7.3.1. For SGD with Subset-Norm, we consider an *equal partition strategy*, where we divide the coordinates into $c = d^{1-\beta}k$ subsets of size d^β/k each with the d^β noisy coordinates into just k subsets so that the rest of the $c - k$ subsets have no noisy coordinate. We defer the derivation details to Section 7.4.6 and summarize the results in the first row of Table 7.1.

7.4.3 Discussions

In Table 7.1, the equal subset-size partition strategy for Subset-Norm has better dependency on the dimension d when the noise is not completely sparse i.e. $\beta = 0$. Hence, if we expect the actual noise density β to be around[2] 0.75 to 0.90, then compressing with a subset size of around $d^{0.45}$ to $d^{0.66}$ is optimal. The dependency on d is important for modern neural network, since the number of parameters d is typically much greater than the total number of iterations T.

[2]Figure 7.2 shows that overall noise is quite sparse but varies more when limited to a particular layer as in Figure 7.3. See Section 9.4.1 for more details.

TABLE 7.1: Algorithms comparison between dimensional dependencies and convergence rates under different coordinate-noise density settings. Given a density rate β, convergence rates' dimensional dependency are highlighted in red and green to denote the worst and best dependency on the dimension. Note that memory usage of AdaGrad-Coordinate is $O(d)$ while SGD with Subset-Norm (with the partition strategy presented here) is $O(d/k)$, where $k = d^{1.4\beta-0.6}$ is chosen as an optimal noise dependent subset size.

Density rate	AdaGrad-Coordinate	AdaGrad-Norm	Subset-Norm (equipartition subsets)
$\beta \in [0,1]$	$\tilde{O}\left(d^{1.5+\beta}/\sqrt{T} + d^{2.5}/T\right)$	$\tilde{O}\left(d^{2.5\beta}/\sqrt{T} + d^{3\beta}/T\right)$	$\tilde{O}\left(d^{0.3+1.8\beta}/\sqrt{T} + d^{\beta+1}/T\right)$ if $\beta \in [0, 2/3]$ $\tilde{O}\left(d^{0.3+1.8\beta}/\sqrt{T} + d^{1.6\beta+0.6}/T\right)$ if $\beta \in [2/3, 1]$
$\beta = 0$	$\tilde{O}\left(d^{1.5}/\sqrt{T} + d^{2.5}/T\right)$	$\tilde{O}\left(1/\sqrt{T} + 1/T\right)$	$\tilde{O}\left(d^{0.3}/\sqrt{T} + d/T\right)$
$\beta = 0.5$	$\tilde{O}\left(d^{2}/\sqrt{T} + d^{2.5}/T\right)$	$\tilde{O}\left(d^{1.25}/\sqrt{T} + d^{1.5}/T\right)$	$\tilde{O}\left(d^{1.2}/\sqrt{T} + d^{1.5}/T\right)$
$\beta = 0.9$	$\tilde{O}\left(d^{2.4}/\sqrt{T} + d^{2.5}/T\right)$	$\tilde{O}\left(d^{2.25}/\sqrt{T} + d^{2.7}/T\right)$	$\tilde{O}\left(d^{1.92}/\sqrt{T} + d^{2.04}/T\right)$
$\beta = 1$	$\tilde{O}\left(d^{2.5}/\sqrt{T} + d^{2.5}/T\right)$	$\tilde{O}\left(d^{2.5}/\sqrt{T} + d^{3}/T\right)$	$\tilde{O}\left(d^{2.1}/\sqrt{T} + d^{2.2}/T\right)$

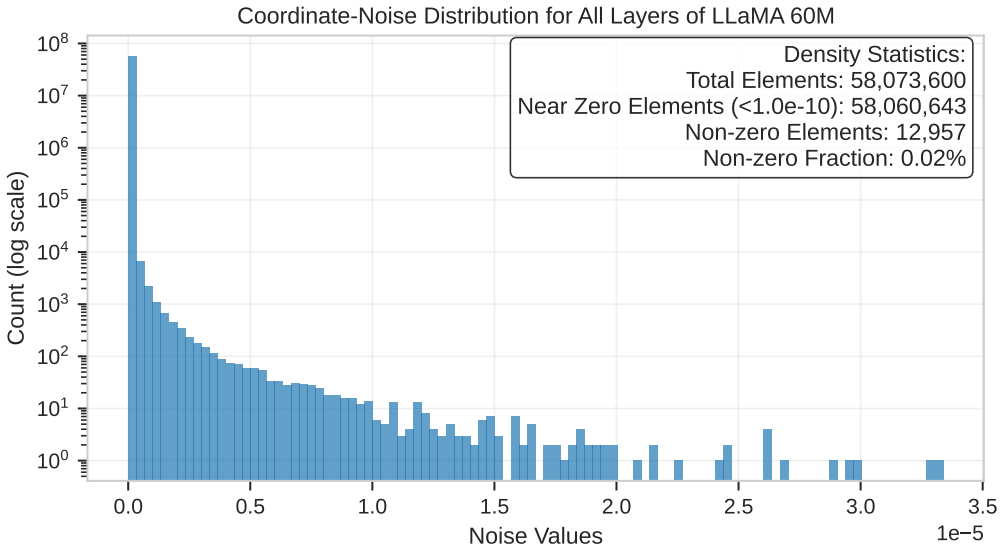

FIGURE 7.2: Aggregated noise distribution across *all* parameters after 100 steps of training.

7.4.4 Coordinate-Noise Density Experiments

To validate the coordinate-noise density model, we sample stochastic gradients repeatedly (via different mini batches) to obtain a sample variance estimate for the true sub-gaussian parameter σ_i for each coordinate: if $g_1, \ldots, g_n \in \mathbb{R}^d$ are independent stochastic gradient samples, we can calculate the sample variance S^2 as an estimator for σ^2 as $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(g_i - \bar{g})^2$, where $\bar{g} = \frac{1}{n}\sum_{i=1}^{n} g_i$ is the sample mean. We pick $n = 200$ samples (with batch size equals 128) for estimating coordinate-noise on LLaMA 60M across various steps during the training process. Figure 7.2 shows the aggregated noise distribution across *all* parameters for LLaMA 60M after 100 training steps. There, the noise is quite low for the vast majority of coordinates except for some outliers. While the noise seems sparse in aggragate, a more fine-grained analysis, presented in Figure 7.3, shows that noises are dense per parameter, except for the Q and K attention projections in the deeper layers. Figures 7.4 to 7.9 in Section 7.4.5 present more noise density rates across various parameters throughout different points of the training progress.

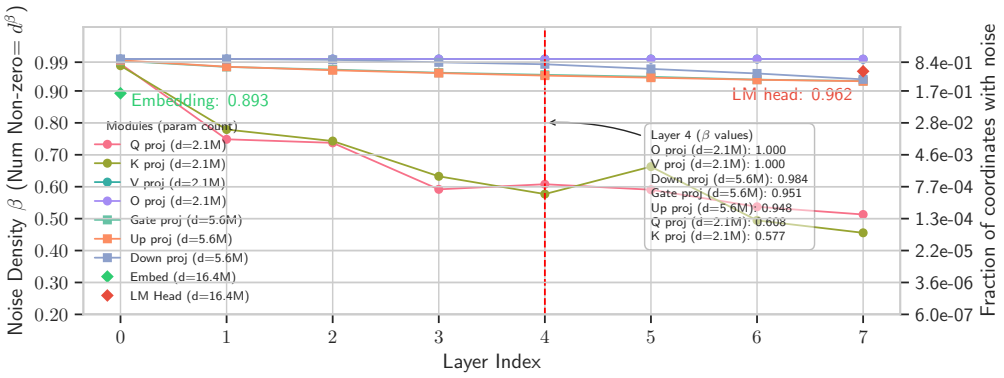

FIGURE 7.3: Noise density per parameter across layers for LLaMA 60M after 100 steps of training.

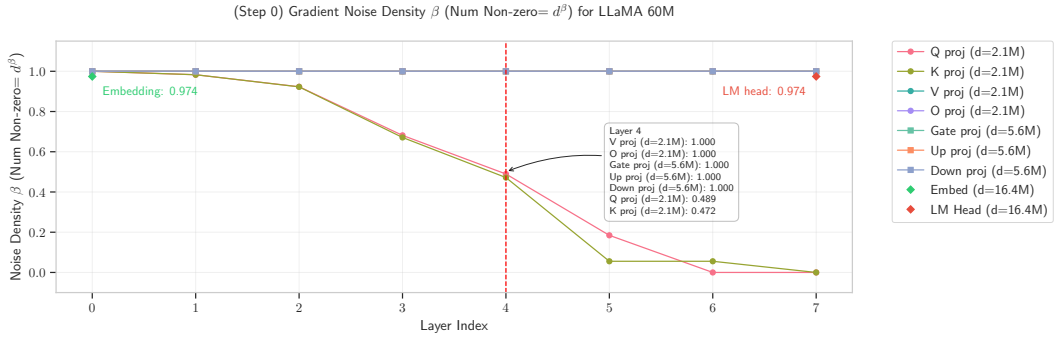

FIGURE 7.4: Noise density for different parameters of LLaMA 60M at Step 0.

7.4.5 Empirical Validation

Figure 7.4 to 7.8 show the normalized noise density ratio for different parameters of LLaMA 60M as described in Section 7.4. The noise patterns show a clear layer-dependent structure, where early layers (like layer 0) maintain consistently high density (close to 1.0) throughout training, while deeper layers start very sparse and gradually become denser as training progresses. Notably, the embedding layer shows an opposite trend, starting relatively dense and becoming increasingly sparse by step 5000, suggesting different dynamics for embedding updates compared to attention layers. The middle layers show an interesting transition pattern, starting sparse but rapidly becoming dense after about 1000 steps, indicating a potential critical phase in training where these layers become more actively involved in learning.

7.4.6 Convergence Rate Derivation

We derive the dimensional dependency of convergence rates for different AdaGrad variants below.

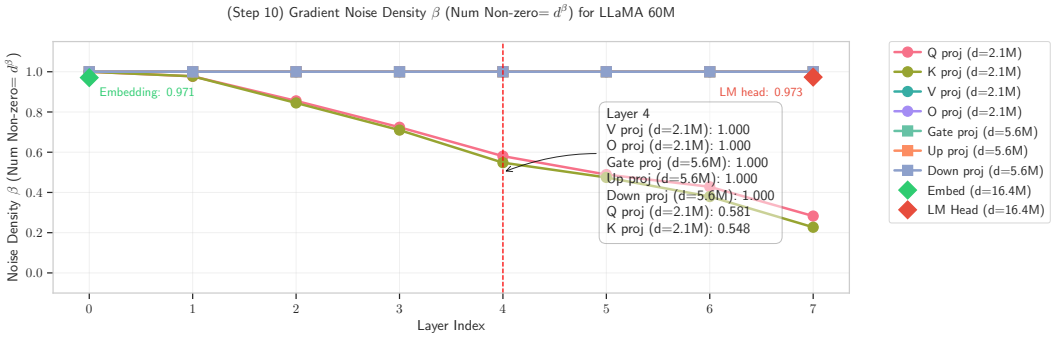

FIGURE 7.5: Noise density for different parameters of LLaMA 60M at Step 10.

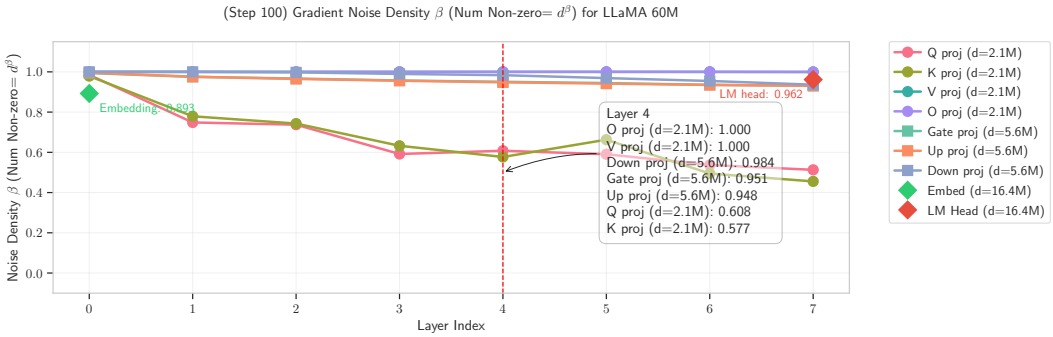

FIGURE 7.6: Noise density for different parameters of LLaMA 60M at Step 100.

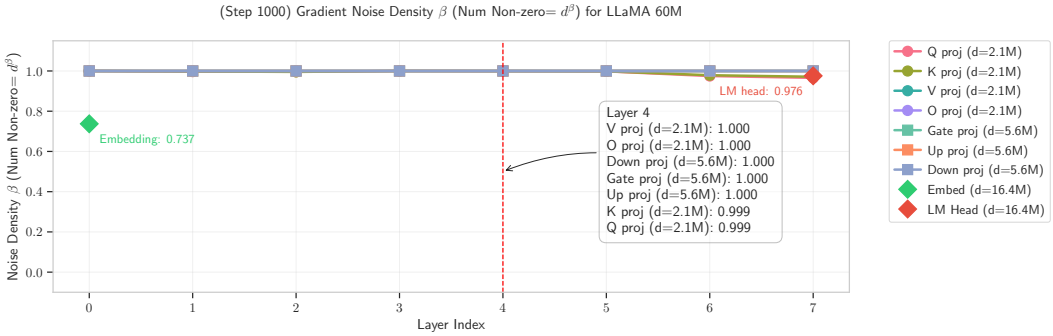

FIGURE 7.7: Noise density for different parameters of LLaMA 60M at Step 1000.

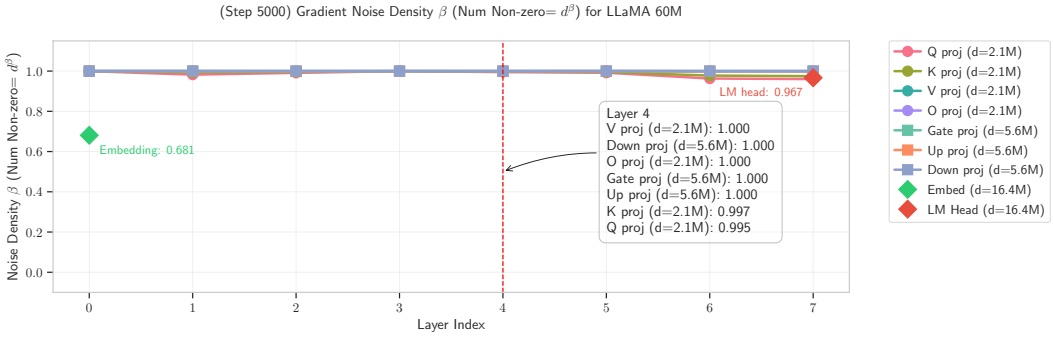

FIGURE 7.8: Noise density for different parameters of LLaMA 60M at Step 5000.

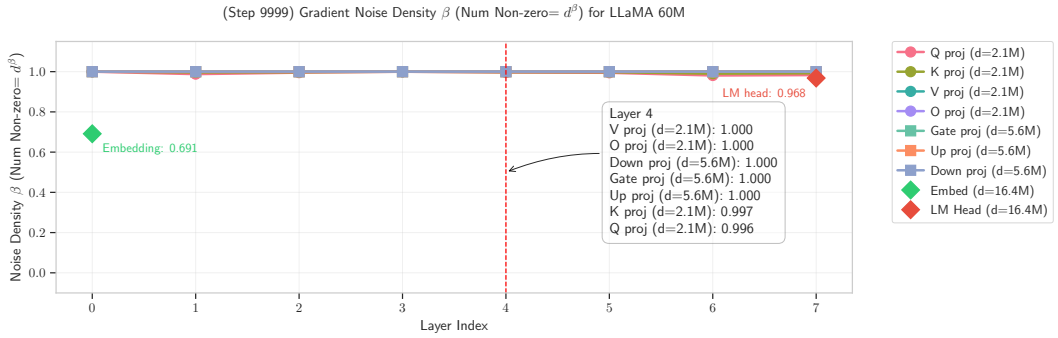

FIGURE 7.9: Noise density for different parameters of LLaMA 60M at Step 9999.

AdaGrad-Coordinate. For $c = d$ (AdaGrad-Coordinate), we get $\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| = \alpha d^\beta$, $\|\sigma\|_2^2 = \alpha^2 d^\beta$, and $\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\|^4 = \alpha^4 d^\beta$, so the bound from Theorem 7.3.1 becomes

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\|_2^2 \leq \tilde{O}\left(\alpha^4 d^\beta + \alpha^3 d^\beta + dL + d^{1.5}\alpha\right) \cdot \tilde{O}\left(\frac{\alpha d^\beta}{\sqrt{T}} + \frac{\alpha^2 d^\beta + \alpha d^\beta + Ld}{T}\right).$$

The dependency on d for the slow term $O(1/\sqrt{T})$ is $d^{1.5} d^\beta = d^{1.5+\beta}$. The dependency on d for the fast term $O(1/T)$ is $d^{1.5} d = d^{2.5}$. Note that there is an inherent $d^{1.5}$ dependency for the slow term that does not reduce as the coordinate-noise density decrease.

AdaGrad-Norm For $c = 1$ (AdaGrad-Norm), we get $\|\sigma\|_2^2 = \sum_{i=0}^{d} \|\sigma_i\|^2 = \alpha^2 d^\beta$, $\|\sigma\|_2 = \alpha d^{\beta/2}$, and $\|\sigma\|^4 = \alpha^4 d^{2\beta}$. This means that our bound from Theorem 7.3.1 becomes

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\|_2^2 \leq \tilde{O}\left(\alpha^4 d^{2\beta} + \alpha^3 d^\beta + L + \alpha\right) \cdot \tilde{O}\left(\frac{\alpha d^{\beta/2}}{\sqrt{T}} + \frac{\alpha^2 d^\beta + \alpha d^{\beta/2} + L}{T}\right).$$

The dependency on d for the slow term $O(1/\sqrt{T})$ is $d^{2\beta} \cdot d^{\beta/2} = d^{2.5\beta}$. The dependency on d for the fast term $O(1/T)$ is $d^{2\beta} \cdot d^\beta = d^{3\beta}$. Note that when $\beta = 0$, or when all the noise is on a single coordinate, we recover the dimension-free results of previous works.

AdaGrad-Subset-Norm. Now, consider the following partition strategy, where we divide the coordinates into $c = d^{1-\beta}k$ subsets of size d^β/k each with the d^β noisy coordinates into just k subsets so that the rest of the $c - k$ subsets do not contain any noisy coordinate. This is a reasonable choice due to the empirical validation from Section 7.4.5: The noisy parameters seem to cluster in groups corresponding to the architecture.

With this strategy, we have $\left\|\sigma_{\Psi_j}\right\|_2^2 = \alpha^2 d^\beta/k \implies \left\|\sigma_{\Psi_j}\right\|_2 = \alpha d^{\beta/2}/k^{0.5}$ if j is a noisy subset. We can compute $\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| = \alpha d^{\beta/2} k^{0.5}$, $\|\sigma\|_2^2 = \sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\|_2^2 = \alpha^2 d^\beta$,

and $\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\|^4 = \alpha^4 d^{2\beta}/k$. From Theorem 7.3.1, we get a bound of

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\|_2^2 \leq \tilde{O}\left(\alpha^4 d^{2\beta}/k + \alpha^3 d^{\beta} + d^{1-\beta}kL + \left(d^{1-\beta}k\right)^{3/2}\alpha\right) \cdot \tilde{O}\left(\frac{\alpha d^{\beta/2}k^{0.5}}{\sqrt{T}} + \frac{\alpha^2 d^{\beta} + \alpha d^{\beta/2}k^{0.5} + Ld^{1-\beta}k}{T}\right).$$

Set $k = d^{7\beta/5-3/5}$ so that $\left(d^{1-\beta}k\right)^{3/2} = d^{2\beta}/k = d^{3\beta/5+3/5}$. Then we can simplify

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\|_2^2 \leq \tilde{O}\left(\alpha^4 d^{3(\beta+1)/5} + \alpha^3 d^{\beta} + d^{2(\beta+1)/5}L + d^{3(\beta+1)/5}\alpha\right) \cdot \tilde{O}\left(\frac{\alpha d^{(12\beta-3)/10}}{\sqrt{T}} + \frac{\alpha^2 d^{\beta} + \alpha d^{(12\beta-3)/10} + Ld^{2(\beta+1)/5}}{T}\right).$$

The dependency on d for the slow term $O(1/\sqrt{T})$ is $d^{3(\beta+1)/5} \cdot d^{(12\beta-3)/10} = d^{3(1+6\beta)/10} = d^{0.3+1.8\beta}$. The dependency on d for the fast term $O(1/T)$ is a bit more complicated: For $\beta \in [0, \frac{2}{3}]$, we have the dependency on d is $d^{3(\beta+1)/5} \cdot d^{2(\beta+1)/5} = d^{\beta+1}$. For $\beta \in [\frac{2}{3}, 1]$, we have the dependency on d is $d^{3(\beta+1)/5} \cdot d^{\beta} = d^{3(\beta+1)/5+\beta} = d^{1.6\beta+0.6}$. Note that this is only a possible partition strategy where the subset sizes are of equal size (which is probably the most natural and easiest to implement). There, the optimal subset size is $k = d^{1.4\beta-0.6}$, for which if we plug in $\beta \in [0, 1]$ we get a range from 1 to $d^{0.8}$.

7.5 Implementation

We provide pseudocode for a general version of Algorithm 10 in Section 9.6.4. Different choices of subset sizes are explored in Section 9.4.1. Furthermore, in contrast to methods like AdaFactor or GaLore that are limited to 2D parameters, subset-norm is a coordinate-wise algorithm and admits an easy implementation to FSDP, where parameters are flattened to 1D tensors for efficient communication.

Subset-size heuristics to avoid additional hyperparameters. In our experiments, to avoid additional hyperparameters, we implement a simple partitioning scheme: for $p \in \mathbb{R}^{m \times n}$, the adaptive step size state is set to $\max(m, n)$, where the subsets are either the rows or the columns. This is a natural grouping scheme that maintains the norm of the larger dimension and aims for the rough $d^{0.45}$ subset size discussed in Section 7.4. Another simplification is that subset-norm is applied only on *linear* modules, since 2D linear modules makes up the vast majority of parameters in transformers. This means we compress all the attention, MLP, and final LM head weights. This implementation is presented in more details in Section 9.6.3. Section 9.4.1 shows that this heuristic grouping is not optimal and can further be improved by tuning the subset size. However, a method with minimal additional tuning is preferred to avoid overfitting, so experiments in Section 9 use this heuristic unless stated otherwise.

7.6 Full Theorem and Proof

We show the full result in Theorem 7.6.1 with all the polylog terms omitted from Theorem 7.3.1.

Theorem 7.6.1. *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and lower bounded by f_*. Given unbiased stochastic gradients $\widehat{\nabla} f(x_t)$ with stochastic gradient noise $\xi_t := \widehat{\nabla} f(x_t) - \nabla f(x_t)$ being σ_i-per-coordinate subgaussian for $i \in [d]$. For partitions of the parameters into disjoint subsets $[d] = \bigcup_{i=0}^{c-1} \Psi_i$ with $\Psi_i \cap \Psi_j = \emptyset$, if $i \neq j$, the iterates x_t given by (7.1) satisfies the following inequality with probability at least $1 - 6c\delta$ (for failure probability $\delta > 0$):*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla_t\|_2^2 \leq G(\delta) \cdot \left(\frac{4\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\|}{\sqrt{T}} + \frac{I(\delta)}{T} \right), \text{ where } G(\delta) \text{ and } I(\delta) \text{ are polylog terms:}$$

$$G(\delta) := \frac{\Delta_1}{\eta} + H(\delta) + \left(\ln T/\delta \, \|\sigma\|_2^2 + c\eta L + 4c^{3/2}\sigma_{\max}\sqrt{\log\frac{1}{\delta}} \right) \log\left(\frac{4\sqrt{T}\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| + I(\delta)}{b_{0,\min}} \right)$$

$$I(\delta) := \|b_0\|_1 + \frac{2\Delta_1}{\eta} + \frac{8\log\frac{1}{\delta}}{b_{0,\min}} \|\sigma\|_2^2 + \sqrt{\log\frac{1}{\delta}} \sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| + 8\eta L c \log\frac{4\eta L}{b_{0,\min}}$$

$$H(\delta) := \sum_{i=0}^{c-1} \left(\ln(T/\delta) \, \|\sigma_{\Psi_i}\|^2 + 2\alpha \right) \left(\frac{8\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}}{b_{0,i}^2} + 2\log\left(1 + \|\sigma_{\Psi_i}\|^2 T + \|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta} \right) \right).$$

where $\|\sigma\|_2^2 = \sum_{i=1}^{d} \sigma_i^2$, $\|\sigma_{\Psi_i}\|^2 = \sum_{j\in\Psi_i} \sigma_j^2$, $\sigma_{\max} = \max_{i\in[d]} \sigma_i$, $\Delta_1 = f(x_1) - f_*$, $b_{0,\min} = \min_{i\in[d]} b_{0,i} > 0$.

7.6.1 Proof of Theorem 7.6.1

For simplicity, in our analysis, we will use $\widehat{\nabla} f_{t,i} := \widehat{\nabla}_i f(x_t)$ and $\nabla_{t,i} := \nabla_i f(x_t)$ to denote the i-th coordinate of the stochastic gradients and gradients at iterate t, respectively. The proof utilizes techniques and follows the strategies (Liu et al., 2023c), where the main effort is to adapt the techniques for handling subsets from the AdaGrad-Norm and AdaGrad-Coordinate proofs in (Liu et al., 2023c).

Proof. We write $\frac{\widehat{\nabla} f_t}{b_t}$ to denote $\left(\frac{\widehat{\nabla} f_t}{b_t}\right)_k = \frac{\widehat{\nabla}_k f(x_t)}{b_{t,i}}$ for $k \in \Psi_i$ (we will use this notation briefly to show some steps and will not be crucial in the main analysis). We start with the smoothness of f and $\Delta_t := f(x_t) - f_*$.

$$\Delta_{t+1} - \Delta_t \tag{7.2}$$

$$\leq \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2$$

$$= -\eta \left\langle \nabla_t, \frac{\widehat{\nabla} f_t}{b_t} \right\rangle + \frac{\eta^2 L}{2} \left\| \frac{\widehat{\nabla} f_t}{b_t} \right\|^2 \tag{7.3}$$

$$= -\eta \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\nabla_{t,j} \widehat{\nabla} f_{t,j}}{b_{t,i}} + \frac{\eta^2 L}{2} \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\widehat{\nabla} f_{t,j}^2}{b_{t,i}^2}$$

$$= -\eta \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\nabla_{t,j} \left(\xi_{t,j} + \nabla_{t,j} \right)}{b_{t,i}} + \frac{\eta^2 L}{2} \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\widehat{\nabla} f_{t,j}^2}{b_{t,i}^2} \qquad (\xi_{t,i} = \widehat{\nabla} f_{t,i} - \nabla_{t,i})$$

$$= -\eta \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\nabla_{t,j}^2}{b_{t,i}} - \eta \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\nabla_{t,j} \xi_{t,j}}{b_{t,i}} + \frac{\eta^2 L}{2} \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\widehat{\nabla} f_{t,j}^2}{b_{t,i}^2}$$

$$= -\eta \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\nabla_{t,j}^2}{b_{t,i}} - \eta \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\nabla_{t,j} \xi_{t,j}}{a_{t,i}} + \eta \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \left(\frac{1}{a_{t,i}} - \frac{1}{b_{t,i}} \right) \nabla_{t,j} \xi_{t,j}$$

$$+ \frac{\eta^2 L}{2} \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\widehat{\nabla} f_{t,j}^2}{b_{t,i}^2}. \tag{7.4}$$

Now, we analyze $\frac{1}{a_{t,i}} - \frac{1}{b_{t,i}}$ for $i = 0, 1, \ldots, c-1$:

$$\left| \frac{1}{a_{t,i}} - \frac{1}{b_{t,i}} \right| = \left| \frac{b_{t,i} - a_{t,i}}{a_{t,i} b_{t,i}} \right|$$

$$= \left| \frac{b_{t,i}^2 - a_{t,i}^2}{a_{t,i} b_{t,i} \left(b_{t,i} + a_{t,i} \right)} \right|$$

$$= \left| \frac{b_{t-1,i}^2 + \left\| \widehat{\nabla} f_{\Psi_i} f(x_t) \right\|^2 - b_{t-1,i}^2 - \left\| \nabla_{\Psi_i} f(x_t) \right\|^2}{a_{t,i} b_{t,i} \left(b_{t,i} + a_{t,i} \right)} \right|$$

$$= \left| \frac{\left\| \widehat{\nabla} f_{\Psi_i} f(x_t) \right\|^2 - \left\| \nabla_{\Psi_i} f(x_t) \right\|^2}{a_{t,i} b_{t,i} \left(b_{t,i} + a_{t,i} \right)} \right|$$

$$= \left| \frac{\left(\left\| \widehat{\nabla} f_{\Psi_i} f(x_t) \right\| - \left\| \nabla_{\Psi_i} f(x_t) \right\| \right) \left(\left\| \widehat{\nabla} f_{\Psi_i} f(x_t) \right\| + \left\| \nabla_{\Psi_i} f(x_t) \right\| \right)}{a_{t,i} b_{t,i} \left(b_{t,i} + a_{t,i} \right)} \right|.$$

Since $b_{t,i} = \sqrt{b_{t-1,i}^2 + \left\|\widehat{\nabla} f_{\Psi_i} f(x_t)\right\|^2} \geq \left\|\widehat{\nabla} f_{\Psi_i} f(x_t)\right\|$ and $a_{t,i} = \sqrt{b_{t-1,i}^2 + \|\nabla_{\Psi_i} f(x_t)\|^2} \geq \|\nabla_{\Psi_i} f(x_t)\|$, we have

$$\begin{aligned}
\left|\frac{1}{a_{t,i}} - \frac{1}{b_{t,i}}\right| &\leq \left|\frac{\left(\left\|\widehat{\nabla} f_{\Psi_i} f(x_t)\right\| - \|\nabla_{\Psi_i} f(x_t)\|\right)\left(\left\|\widehat{\nabla} f_{\Psi_i} f(x_t)\right\| + \|\nabla_{\Psi_i} f(x_t)\|\right)}{a_{t,i} b_{t,i}\left(\left\|\widehat{\nabla} f_{\Psi_i} f(x_t)\right\| + \|\nabla_{\Psi_i} f(x_t)\|\right)}\right| \\
&\leq \left|\frac{\left\|\widehat{\nabla} f_{\Psi_i} f(x_t)\right\| - \|\nabla_{\Psi_i} f(x_t)\|}{a_{t,i} b_{t,i}}\right| \\
&\leq \frac{\left\|\widehat{\nabla} f_{\Psi_i} f(x_t) - \nabla_{\Psi_i} f(x_t)\right\|}{a_{t,i} b_{t,i}} \\
&= \frac{\|\xi_{t,\Psi_i}\|}{a_{t,i} b_{t,i}}.
\end{aligned}$$

Hence, we have

$$\left|\frac{1}{a_{t,i}} - \frac{1}{b_{t,i}}\right| \leq \frac{\|\xi_{t,\Psi_i}\|}{a_{t,i} b_{t,i}}.$$

Then from 7.4, taking the absolute value of $\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\left(\frac{1}{a_{t,i}} - \frac{1}{b_{t,i}}\right)\nabla_{t,j}\xi_{t,j}$, we can bound:

$$\begin{aligned}
&\Delta_{t+1} - \Delta_t \\
&\quad\leq -\eta \sum_{i=0}^{c-1}\sum_{j\in\Psi_i} \frac{\nabla_{t,j}^2}{b_{t,i}} - \eta \sum_{i=0}^{c-1}\sum_{j\in\Psi_i} \frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} + \eta \sum_{i=0}^{c-1}\sum_{j\in\Psi_i} \left|\frac{1}{a_{t,i}} - \frac{1}{b_{t,i}}\right| \left|\nabla_{t,j}\xi_{t,j}\right| + \frac{\eta^2 L}{2}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i} \frac{\widehat{\nabla} f_{t,j}^2}{b_{t,i}^2} \\
&\quad\leq -\eta \sum_{i=0}^{c-1}\sum_{j\in\Psi_i} \frac{\nabla_{t,j}^2}{b_{t,i}} - \eta \sum_{i=0}^{c-1}\sum_{j\in\Psi_i} \frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} + \eta \sum_{i=0}^{c-1} \frac{\|\xi_{t,\Psi_i}\|}{a_{t,i} b_{t,i}} \sum_{j\in\Psi_i} \left|\nabla_{t,j}\xi_{t,j}\right| + \frac{\eta^2 L}{2}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i} \frac{\widehat{\nabla} f_{t,j}^2}{b_{t,i}^2} \\
&\quad\overset{(1)}{\leq} -\eta \sum_{i=0}^{c-1}\sum_{j\in\Psi_i} \frac{\nabla_{t,j}^2}{b_{t,i}} - \eta \sum_{i=0}^{c-1}\sum_{j\in\Psi_i} \frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} + \eta \sum_{i=0}^{c-1} \frac{\|\xi_{t,\Psi_i}\|}{a_{t,i} b_{t,i}} \|\nabla_{t,\Psi_i}\| \|\xi_{t,\Psi_i}\| + \frac{\eta^2 L}{2}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i} \frac{\widehat{\nabla} f_{t,j}^2}{b_{t,i}^2} \\
&\quad\leq -\eta \sum_{i=0}^{c-1}\sum_{j\in\Psi_i} \frac{\nabla_{t,j}^2}{b_{t,i}} - \eta \sum_{i=0}^{c-1}\sum_{j\in\Psi_i} \frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} \\
&\qquad + \eta \sum_{i=0}^{c-1} \|\xi_{t,\Psi_i}\| \left(\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2} + \frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2}\right) + \frac{\eta^2 L}{2}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i} \frac{\widehat{\nabla} f_{t,j}^2}{b_{t,i}^2},
\end{aligned}$$

where (1) is due to $\sum_{j\in\Psi_i}\left|\nabla_{t,j}\xi_{t,j}\right| = \left\langle\left|\nabla_{t,\Psi_i}\right|, \left|\xi_{t,\Psi_i}\right|\right\rangle \leq \|\nabla_{t,\Psi_i}\| \|\xi_{t,\Psi_i}\|$ and $|\cdot|$ denotes coordinate-wise absolute value when we apply to vectors. The last inequality is due to $2ab \leq a^2 + b^2$. Now, we can sum both sides for $t = 1, \ldots, T$ to telescope the LHS:

$$\begin{aligned}
\Delta_{T+1} - \Delta_1 \leq \sum_{t=1}^{T}\Bigg(&-\eta \sum_{i=0}^{c-1}\sum_{j\in\Psi_i} \frac{\nabla_{t,j}^2}{b_{t,i}} - \eta \sum_{i=0}^{c-1}\sum_{j\in\Psi_i} \frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} \\
&+ \eta \sum_{i=0}^{c-1} \|\xi_{t,\Psi_i}\| \left(\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2} + \frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2}\right) + \frac{\eta^2 L}{2}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i} \frac{\widehat{\nabla} f_{t,j}^2}{b_{t,i}^2}\Bigg).
\end{aligned}$$

Rearranging gives

$$\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}^2}{b_{t,i}} \le \frac{\Delta_1-\Delta_{T+1}}{\eta} - \underbrace{\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}}}_{A}$$
$$+\underbrace{\sum_{t=1}^{T}\sum_{i=0}^{c-1}\|\xi_{t,\Psi_i}\|\left(\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2}+\frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2}\right)}_{B}+\frac{\eta L}{2}\underbrace{\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\widehat{\nabla}f_{t,j}^2}{b_{t,i}^2}}_{C}.$$

On the LHS, we note that

$$\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}^2}{b_{t,i}}=\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\nabla_{t,\Psi_i}\|^2}{b_{t,i}}.$$

We now bound each term separately. It's easiest to bound C: $\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\widehat{\nabla}f_{t,j}^2}{b_{t,i}^2}$:

$$\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\widehat{\nabla}f_{t,j}^2}{b_{t,i}^2}=\sum_{i=0}^{c-1}\sum_{t=1}^{T}\sum_{j\in\Psi_i}\frac{\widehat{\nabla}f_{t,j}^2}{b_{t,i}^2}=\sum_{i=1}^{d}\sum_{t=1}^{T}\frac{b_{t,i}^2-b_{t-1,i}^2}{b_{t,i}^2}\le\sum_{i=1}^{d}2\log\frac{b_{T,i}}{b_{0,i}}.$$
$$\begin{aligned}
&=\sum_{i=0}^{c-1}\sum_{t=1}^{T}\frac{\left\|\widehat{\nabla}f_{t,\Psi_i}\right\|^2}{b_{t,i}^2}\\
&=\sum_{i=0}^{c-1}\sum_{t=1}^{T}\frac{b_{t,i}^2-b_{t-1,i}^2}{b_{t,i}^2}\\
&=\sum_{i=0}^{c-1}\sum_{t=1}^{T}1-\frac{b_{t-1,i}^2}{b_{t,i}^2}\\
&\le\sum_{i=0}^{c-1}\sum_{t=1}^{T}\log\frac{b_{t,i}^2}{b_{t-1,i}^2}\\
&=2\sum_{i=0}^{c-1}\log\prod_{t=1}^{T}\frac{b_{t,i}}{b_{t-1,i}}\\
&=2\sum_{i=0}^{c-1}\log\frac{b_{T,i}}{b_{0,i}}.
\end{aligned}$$

We now have a useful inequality

$$\sum_{t=1}^{T}\frac{\left\|\widehat{\nabla}f_{t,\Psi_i}\right\|^2}{b_{t,i}^2}\le 2\log\frac{b_{T,i}}{b_{0,i}},\ \forall i=0,\dots,c-1. \tag{7.5}$$

Next, we deal with $-\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}}$ via a martingale argument. Let $\mathcal{F}_t := \sigma(\xi_1,\dots,\xi_{t-1})$ denote the natural filtration. Note that x_t is $\mathcal{F}_t$-measurable. For any

$w > 0$, we have for each $i \in [c]$:

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(-w\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} - 2w^2\sum_{j\in\Psi_i}\frac{\sigma_j^2\nabla_{t,j}^2}{a_{t,i}^2}\right) \mid \mathcal{F}_t\right] \\
&= \exp\left(-2w^2\sum_{j\in\Psi_i}\frac{\sigma_j^2\nabla_{t,j}^2}{a_{t,i}^2}\right)\mathbb{E}\left[\exp\left(-w\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}}\right) \mid \mathcal{F}_t\right] \\
&\leq 1.
\end{aligned}$$

Then a simple inductive argument and using Markov's inequality gives with probability at least $1-\delta$:

$$-w\sum_{t=1}^{T}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} \leq 2w^2\sum_{t=1}^{T}\sum_{j\in\Psi_i}\frac{\sigma_j^2\nabla_{t,j}^2}{a_{t,i}^2} + \log\frac{1}{\delta}.$$

By a union bound across all c subsets, we have w.p. at least $1-c\delta$:

$$-\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} \leq \sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{w\sigma_j^2\nabla_{t,j}^2}{a_{t,i}^2} + \frac{c}{w}\log\frac{1}{\delta}. \tag{7.6}$$

Let's call the event that (7.6) happens E_1. Now, consider $\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}^2}{a_{t,i}^2}$. We have

$$\begin{aligned}
\sum_{j\in\Psi_i}\frac{\nabla_{t,j}^2}{a_{t,i}^2} = \frac{\|\nabla_{t,\Psi_i}\|^2}{a_{t,i}^2} &= \frac{\|\nabla_{t,\Psi_i}\|^2}{b_{t-1,i}^2 + \|\nabla_{t,\Psi_i}\|^2} \\
&\overset{(*)}{\leq} \frac{2\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2 + 2\|\xi_{t,\Psi_i}\|^2}{b_{t-1,i}^2 + 2\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2 + 2\|\xi_{t,\Psi_i}\|^2} \\
\frac{\|\nabla_{t,\Psi_i}\|^2}{a_{t,i}^2} &\leq 2\frac{\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2}{b_{t,i}^2} + 2\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2}.
\end{aligned}$$

For $(*)$ we use the fact that $\frac{x}{c+x}$ is an increasing function and $\|\nabla_{t,\Psi_i}\|^2 = \left\|\widehat{\nabla} f_{t,\Psi_i} + \xi_{t,\Psi_i}\right\|^2 \leq 2\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2 + 2\|\xi_{t,\Psi_i}\|^2$. Let $\sigma_{\max} := \max_{i\in[d]}\sigma_i$, then under event E_1, we have with probability at least $1-c\delta$:

$$
\begin{aligned}
-\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} &\le \sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{w\sigma_j^2\nabla_{t,j}^2}{a_{t,i}^2}+\frac{c}{w}\log\frac{1}{\delta}\\
&\le w\sigma_{\max}^2\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}^2}{a_{t,i}^2}+\frac{c}{w}\log\frac{1}{\delta}\\
&\le w\sigma_{\max}^2\sum_{t=1}^{T}\sum_{i=0}^{c-1}\left(2\frac{\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2}{b_{t,i}^2}+2\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2}\right)+\frac{c}{w}\log\frac{1}{\delta}\\
&=\underbrace{\sigma_{\max}\sqrt{c\log\frac{1}{\delta}}}_{=:\alpha}\sum_{t=1}^{T}\sum_{i=0}^{c-1}\left(2\frac{\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2}{b_{t,i}^2}+2\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2}\right)+\sigma_{\max}\sqrt{c\log\frac{1}{\delta}}\\
&\qquad\qquad \left(\text{set } w:=\frac{\sqrt{c\log\frac{1}{\delta}}}{\sigma_{\max}}\right)\\
&=2\alpha\sum_{t=1}^{T}\sum_{i=0}^{c-1}\left(\frac{\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2}{b_{t,i}^2}+\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2}\right)+\alpha.
\end{aligned}
$$

where the second to last equality is due to choosing $w=\frac{\sqrt{c\log\frac{1}{\delta}}}{\sigma_{\max}}$ and the last equality is letting $\alpha:=\sigma_{\max}\sqrt{c\log\frac{1}{\delta}}$ for readability.

Let $M_{T,i}=\max_{t\le T}|\xi_{t,i}|$. Using our notation, we can define $M_{T,\Psi_i}:=\max_{t\le T}\|\xi_{t,\Psi_i}\|$. Under event E_1 (and our new bound for C), we have that with probability at least

$1 - c\delta$:

$$\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\nabla_{t,\Psi_i}\|^2}{b_{t,i}} \tag{7.7}$$

$$\overset{\text{(C)}}{\leq} \frac{\Delta_1}{\eta} - \sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} + \sum_{t=1}^{T}\sum_{i=0}^{c-1}\|\xi_{t,\Psi_i}\|\left(\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2} + \frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2}\right) \tag{7.8}$$

$$+ \eta L\sum_{i=0}^{c-1}\log\frac{b_{T,i}}{b_{0,i}}$$

$$\leq \frac{\Delta_1}{\eta} - \sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} \tag{7.9}$$

$$+ \sum_{t=1}^{T}\sum_{i=0}^{c-1}M_{T,\Psi_i}\left(\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2} + \frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2}\right) + \eta L\sum_{i=0}^{c-1}\log\frac{b_{T,i}}{b_{0,i}} \qquad \text{(def of } M_{T,\Psi_i}\text{)}$$

$$\overset{(E_1)}{\leq} \frac{\Delta_1}{\eta} + 2\alpha\sum_{t=1}^{T}\sum_{i=0}^{c-1}\left(\underbrace{\frac{\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2}{b_{t,i}^2}}_{\text{bound with (C)}} + \frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2}\right) + \alpha +$$

$$\sum_{t=1}^{T}\sum_{i=0}^{c-1}M_{T,\Psi_i}\left(\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2} + \frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2}\right) + \eta L\sum_{i=0}^{c-1}\log\frac{b_{T,i}}{b_{0,i}} \tag{7.10}$$

$$\overset{\text{(C)}}{\leq} \frac{\Delta_1}{\eta} + 2\alpha\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} + \alpha +$$

$$\sum_{t=1}^{T}\sum_{i=0}^{c-1}M_{T,\Psi_i}\left(\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2} + \frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2}\right) + (\eta L + 4\alpha)\sum_{i=0}^{c-1}\log\frac{b_{T,i}}{b_{0,i}} \tag{7.11}$$

$$\leq \frac{\Delta_1}{\eta} + 2\alpha\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} + \alpha + \tag{7.12}$$

$$\sum_{t=1}^{T}\sum_{i=0}^{c-1}M_{T,\Psi_i}\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2} + \sum_{t=1}^{T}\sum_{i=0}^{c-1}M_{T,\Psi_i}\frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2} + (\eta L + 4\alpha)\sum_{i=0}^{c-1}\log\frac{b_{T,i}}{b_{0,i}}. \tag{7.13}$$

Let us turn our attention to $M_{T,\Psi_i} := \max_{t\leq T}\|\xi_{t,\Psi_i}\|$. Note that

$$\begin{aligned}
\Pr\left[\max_{t\in[T]}\|\xi_{t,\Psi_i}\|^2 \geq A\right] &= \Pr\left[\exp\left(\frac{\max_{t\in[T]}\|\xi_{t,\Psi_i}\|^2}{w}\right) \geq \exp\left(\frac{A}{w}\right)\right] \qquad \text{(for } w>0\text{)} \\
&\leq \exp\left(-\frac{A}{w}\right)\mathbb{E}\left[\exp\left(\frac{\max_{t\in[T]}\|\xi_{t,\Psi_i}\|^2}{w}\right)\right] \qquad \text{(Markov)} \\
&= \exp\left(-\frac{A}{w}\right)\mathbb{E}\left[\max_{t\in[T]}\exp\left(\frac{\|\xi_{t,\Psi_i}\|^2}{w}\right)\right] \\
&\leq \exp\left(-\frac{A}{w}\right)\sum_{t\in[T]}\mathbb{E}\left[\exp\left(\frac{\|\xi_{t,\Psi_i}\|^2}{w}\right)\right].
\end{aligned}$$

We have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{\|\xi_{t,\Psi_i}\|^2}{w}\right)\right] &= \mathbb{E}\left[\exp\left(\frac{\sum_{j\in\Psi_i}\xi_{t,j}^2}{w}\right)\right] \\ &= \mathbb{E}\left[\exp\left(\frac{\sum_{j\in\Psi_i}\xi_{t,j}^2}{w}\right)\right] \\ &= \mathbb{E}\left[\prod_{j\in\Psi_i}\exp\left(\frac{\xi_{t,j}^2}{w}\right)\right] \\ &= \prod_{j\in\Psi_i}\mathbb{E}\left[\exp\left(\frac{\xi_{t,j}^2}{w}\right)\right]. \qquad \text{(independence)}\end{aligned}$$

Since sub-gaussianity give us

$$\mathbb{E}\left[\exp\left(\lambda^2\xi_{t,i}^2\right)\right] \leq \exp\left(\lambda^2\sigma_i^2\right), \forall\,|\lambda| \leq \frac{1}{\sigma_i}, \forall i \in [d],$$

we have $\mathbb{E}\left[\exp\left(\frac{\xi_{t,j}^2}{w}\right)\right] \leq \exp\left(\frac{\sigma_j^2}{w}\right)$ if $\sqrt{\frac{1}{w}} \leq \frac{1}{\sigma_j}$. We pick $w := \|\sigma_{\Psi_i}\|^2 = \sum_{j\in\Psi_i}\sigma_j^2 \geq \sigma_j^2,\ \forall j \in \Psi_i$. Hence, we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{\|\xi_{t,\Psi_i}\|^2}{\|\sigma_{\Psi_i}\|^2}\right)\right] &\leq \prod_{j\in\Psi_i}\exp\left(\frac{\sigma_j^2}{\|\sigma_{\Psi_i}\|^2}\right) \\ &= \exp\left(\frac{\|\sigma_{\Psi_i}\|^2}{\|\sigma_{\Psi_i}\|^2}\right) = 1. \qquad (7.14)\end{aligned}$$

We have actually shown that ξ_{t,Ψ_i} is a $\|\sigma_{\Psi_i}\|^2$-subgaussian random variable in $\mathbb{R}^k$ (see Proposition 2.5.2 in (Vershynin, 2018)). This fact will come in handy later. Now, we have

$$\begin{aligned}\Pr\left[\max_{t\in[T]}\|\xi_{t,\Psi_i}\|^2 \geq A\right] &\leq \exp\left(-\frac{A}{\|\sigma_{\Psi_i}\|^2}\right)\sum_{t\in[T]}\mathbb{E}\left[\exp\left(\frac{\|\xi_{t,\Psi_i}\|^2}{\|\sigma_{\Psi_i}\|^2}\right)\right] \\ &= \exp\left(-\frac{A}{\|\sigma_{\Psi_i}\|^2}\right)T.\end{aligned}$$

Setting $\exp\left(-\frac{A}{\|\sigma_{\Psi_i}\|^2}\right)T = \delta$ gives $A = \|\sigma_{\Psi_i}\|^2 \ln T/\delta$. Hence, we have with probability at least $1-\delta$,

$$M_{T,\Psi_i} = \max_{t\in[T]}\|\xi_{t,\Psi_i}\|^2 \leq \|\sigma_{\Psi_i}\|^2 \ln T/\delta. \qquad (7.15)$$

Union bounding across all $i = 0, 1, \ldots, c-1$, we have that with probability at least $1 - c\delta$,

$$M_{T,\Psi_i} \leq \|\sigma_{\Psi_i}\|^2 \ln T/\delta,\ \forall i = 0, 1, \ldots, c-1. \qquad (7.16)$$

Let us denote the event in (7.16) by E_2. Combining it with event E_1 and starting from 7.12, we have that with probability $1 - c\delta$:

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\nabla_{t,\Psi_i}\|^2}{b_{t,i}} \\
&\quad\le \frac{\Delta_1}{\eta} + 2\alpha\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} + \alpha + \sum_{t=1}^{T}\sum_{i=0}^{c-1} M_{T,\Psi_i}\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2} + \\
&\qquad \sum_{t=1}^{T}\sum_{i=0}^{c-1} M_{T,\Psi_i}\frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2} + (\eta L + 4\alpha)\sum_{i=0}^{c-1}\log\frac{b_{T,i}}{b_{0,i}} \\
&\quad\le \frac{\Delta_1}{\eta} + 2\alpha\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} + \ln T/\delta\sum_{t=1}^{T}\sum_{i=0}^{c-1}\|\sigma_{\Psi_i}\|^2\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2} + \alpha + \\
&\qquad \ln T/\delta\sum_{t=1}^{T}\sum_{i=0}^{c-1}\|\sigma_{\Psi_i}\|^2\frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2} + (\eta L + 4\alpha)\sum_{i=0}^{c-1}\log\frac{b_{T,i}}{b_{0,i}} \\
&\quad= \frac{\Delta_1}{\eta} + \sum_{i=0}^{c-1}\left(\ln T/\delta\frac{\|\sigma_{\Psi_i}\|^2}{2} + 2\alpha\right)\sum_{t=1}^{T}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} + \alpha + \\
&\qquad \ln T/\delta\sum_{i=0}^{c-1}\frac{\|\sigma_{\Psi_i}\|^2}{2}\sum_{t=1}^{T}\frac{\|\nabla_{t,\Psi_i}\|^2}{a_{t,i}^2} + (\eta L + 4\alpha)\sum_{i=0}^{c-1}\log\frac{b_{T,i}}{b_{0,i}}.
\end{aligned}$$

Recall that $\frac{\|\nabla_{t,\Psi_i}\|^2}{a_{t,i}^2} \le 2\frac{\|\widehat{\nabla} f_{t,\Psi_i}\|^2}{b_{t,i}^2} + 2\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2}$, we then have

$$\begin{aligned}
&\ln T/\delta\sum_{i=0}^{c-1}\frac{\|\sigma_{\Psi_i}\|^2}{2}\sum_{t=1}^{T}\frac{\|\nabla_{t,\Psi_i}\|^2}{a_{t,i}^2} \\
&\quad\le \ln T/\delta\sum_{i=0}^{c-1}\frac{\|\sigma_{\Psi_i}\|^2}{2}\sum_{t=1}^{T}\left(2\frac{\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2}{b_{t,i}^2} + 2\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2}\right) \\
&\quad= \ln T/\delta\sum_{i=0}^{c-1}\|\sigma_{\Psi_i}\|^2\sum_{t=1}^{T}\frac{\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2}{b_{t,i}^2} + \ln T/\delta\sum_{i=0}^{c-1}\|\sigma_{\Psi_i}\|^2\sum_{t=1}^{T}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} \\
&\quad\le \ln T/\delta\sum_{i=0}^{c-1}\|\sigma_{\Psi_i}\|^2\log\frac{b_{T,i}}{b_{0,i}} + \ln T/\delta\sum_{i=0}^{c-1}\|\sigma_{\Psi_i}\|^2\sum_{t=1}^{T}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2}. \quad \text{(from 7.5)}
\end{aligned}$$

Hence, we have with probability at least $1 - 2c\delta$:

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\nabla_{t,\Psi_i}\|^2}{b_{t,i}} &\le \frac{\Delta_1}{\eta} + \sum_{i=0}^{c-1}\left(\ln T/\delta\,\|\sigma_{\Psi_i}\|^2 + 2\alpha\right)\sum_{t=1}^{T}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} && (7.17)\\
&\quad + \alpha + \sum_{i=0}^{c-1}\ln T/\delta\,\|\sigma_{\Psi_i}\|^2\log\frac{b_{T,i}}{b_{0,i}} + \sum_{i=0}^{c-1}(\eta L + 4\alpha)\log\frac{b_{T,i}}{b_{0,i}} \\
&= \frac{\Delta_1}{\eta} + \sum_{i=0}^{c-1}\left(\ln T/\delta\,\|\sigma_{\Psi_i}\|^2 + 2\alpha\right)\sum_{t=1}^{T}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} && (7.18)\\
&\quad + \alpha + \sum_{i=0}^{c-1}\left(\ln T/\delta\,\|\sigma_{\Psi_i}\|^2 + \eta L + 4\alpha\right)\log\frac{b_{T,i}}{b_{0,i}}. && (7.19)
\end{aligned}$$

Now, we bound $\sum_{t=1}^{T} \frac{\left\|\xi_{t,\Psi_i}\right\|^2}{b_{t,i}^2}$ and $\log \frac{b_{T,i}}{b_{0,i}}$. We need to first lower bound $\sum_{s=1}^{t} \left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2$. We proceed by noting that

$$\begin{aligned} \|\widehat{\nabla} f_{t,\Psi_i}\|^2 &= \|\nabla_{t,\Psi_i} + \xi_{t,\Psi_i}\|^2 \\ &= \|\nabla_{t,\Psi_i}\|^2 + 2\langle \xi_{t,\Psi_i}, \nabla_{t,\Psi_i}\rangle + \|\xi_{t,\Psi_i}\|^2 \\ \Rightarrow \|\nabla_{t,\Psi_i}\| - \|\widehat{\nabla} f_{t,\Psi_i}\|^2 + \|\xi_{t,\Psi_i}\|^2 &= 2\langle \xi_{t,\Psi_i}, \nabla_{t,\Psi_i}\rangle. \end{aligned}$$

Define for $t \in \{0, 1, \cdots, T\}$ and some constant v_s to be specified later:

$$\begin{aligned} U_{t+1} &= \exp\left(\sum_{s=1}^{t} w_s \left(\|\nabla_{s,\Psi_i}\| - \|\widehat{\nabla} f_{s,\Psi_i}\|^2 + \|\xi_{s,\Psi_i}\|^2\right) - v_s \|\nabla_{s,\Psi_i}\|^2\right) \\ &= U_t \cdot \exp\left(w_t \left(\|\nabla_{t,\Psi_i}\| - \|\widehat{\nabla} f_{t,\Psi_i}\|^2 + \|\xi_{t,\Psi_i}\|^2\right) - v_t \|\nabla_{t,\Psi_i}\|^2\right) \\ &= U_t \cdot \exp\left(w_t \left(2\langle \xi_{t,\Psi_i}, \nabla_{t,\Psi_i}\rangle\right) - v_t \|\nabla_{t,\Psi_i}\|^2\right). \end{aligned}$$

First, note that $U_t \in \mathcal{F}_t$. We show that U_t is a supermartingale

$$\begin{aligned} \mathbb{E}\left[U_{t+1} \mid \mathcal{F}_t\right] &= \mathbb{E}\left[U_t \cdot \exp\left(w_t \left(2\langle \xi_{t,\Psi_i}, \nabla_{t,\Psi_i}\rangle\right) - v_t \|\nabla_{t,\Psi_i}\|^2\right) \mid \mathcal{F}_t\right] \\ &= U_t \exp\left(-v_t \|\nabla_{t,\Psi_i}\|^2\right) \mathbb{E}\left[\exp\left(2w_t \langle \xi_{t,\Psi_i}, \nabla_{t,\Psi_i}\rangle\right) \mid \mathcal{F}_t\right] \\ &\overset{(*)}{\leq} U_t \exp\left(-v_t \|\nabla_{t,\Psi_i}\|^2\right) \mathbb{E}\left[\exp\left(4w_t^2 \|\sigma_{\Psi_i}\|^2 \|\nabla_{t,\Psi_i}\|^2\right) \mid \mathcal{F}_t\right] \\ &= U_t, \qquad (v_t{=}4w_t^2 \|\sigma_{\Psi_i}\|^2) \end{aligned}$$

where $(*)$ is due to Lemma 2.2 of (Liu et al., 2023c) and the fact that ξ_{t,Ψ_i} is $\|\sigma_{\Psi_i}\|^2$-subgaussian from (7.14). Hence, by Ville's supermartingale inequality, we have

$$\Pr\left[\max_{t\in[T+1]} U_t \geq \delta^{-1}\right] \leq \delta \mathbb{E}\left[U_1\right] = \delta.$$

This implies w.p. $\geq 1-\delta$, $\forall 0 \leq t \leq T$:

$$\begin{aligned} &\sum_{s=1}^{t} w_s \left(\|\nabla_{s,\Psi_i}\| - \|\widehat{\nabla} f_{s,\Psi_i}\|^2 + \|\xi_{s,\Psi_i}\|^2\right) - v_s \|\nabla_{s,\Psi_i}\|^2 \leq \log \frac{1}{\delta} \\ \Longrightarrow &\sum_{s=1}^{t} \left(w_s - 4w_s^2 \|\sigma_{\Psi_i}\|^2\right) \|\nabla_{s,\Psi_i}\|^2 + \sum_{s=1}^{t} w_s \|\xi_{s,\Psi_i}\|^2 \leq \sum_{s=1}^{t} w_s \|\widehat{\nabla} f_{s,\Psi_i}\|^2 + \log \frac{1}{\delta} \\ \Longleftrightarrow &\sum_{s=1}^{t} \left(1 - 4w_s \|\sigma_{\Psi_i}\|^2\right) \|\nabla_{s,\Psi_i}\|^2 + \sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2 \leq \sum_{s=1}^{t} \|\widehat{\nabla} f_{s,\Psi_i}\|^2 + \frac{1}{w_s} \log \frac{1}{\delta}. \end{aligned}$$

Set $w_s = \frac{1}{4\|\sigma_{\Psi_i}\|^2}$ to get

$$\sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2 \leq \sum_{s=1}^{t} \|\widehat{\nabla} f_{s,\Psi_i}\|^2 + 4\|\sigma_{\Psi_i}\|^2 \log \frac{1}{\delta}, \ \forall t \leq T. \tag{7.20}$$

We are now ready to bound $\sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2}$. Starting by applying (7.20), we have that with probability at least $1-\delta$

$$\begin{aligned}\sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} &= \sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{0,i}^2 + \sum_{s=1}^{t} \left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2} \\ &\leq \sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{0,i}^2 + \left(\sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2 - 4\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}\right)^+}\end{aligned}$$

where $(x)^+ = \max\{x, 0\}$. Let $\tau = \max\left(\{0\} \cup \left\{t \in \mathbb{N}_{\leq T} \mid \sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2 \leq 2C\right\}\right)$ for some $C \geq 0$. We have

$$\begin{aligned}\sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} &= \sum_{t=1}^{\tau} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} + \sum_{t=\tau+1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{0,i}^2 + \sum_{s=1}^{t} \left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2} \\ &\leq \frac{1}{b_{0,i}^2} \sum_{t=1}^{\tau} \|\xi_{t,\Psi_i}\|^2 + \sum_{t=\tau+1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{0,i}^2 + \sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2 - 4\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}} \\ &\leq \frac{2C}{b_{0,i}^2} + \sum_{t=\tau+1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{0,i}^2 + \sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2 - 4\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}}.\end{aligned}$$

Now, since $\frac{\sum_{s=1}^{t}\|\xi_{s,\Psi_i}\|^2}{2} \geq C$ for $t > \tau$, we have $b_{0,i}^2 + \sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2 - 4\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta} \geq b_{0,i}^2 - 4\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta} + C + \frac{1}{2}\sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2$. If $b_{0,i}^2 - 4\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta} \geq 0$, then we pick $C = 0$ and $b_{0,i}^2 - 4\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta} + C + \frac{1}{2}\sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2 \geq \frac{1}{2}\sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2$. If $b_{0,i}^2 - 4\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta} < 0$, we pick $C = 4\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta} - b_{0,i}^2 > 0$, which gives $b_{0,i}^2 - 4\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta} + C + \frac{1}{2}\sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2 \geq \frac{1}{2}\sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2$. In either case, we have $b_{0,i}^2 - 4\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta} + C + \frac{1}{2}\sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2 \geq \frac{1}{2}\sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2$. Hence, letting $C = \max\left(0, 4\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta} - b_{0,i}^2\right) \leq 4\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}$, we have with probability at least $1-\delta$:

$$\begin{aligned}\sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} &\leq \frac{2C}{b_{0,i}^2} + 2\sum_{t=\tau+1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{\sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2} \\ &\leq \frac{2C}{b_{0,i}^2} + 2\sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{\sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2} \\ &\leq \frac{8\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}}{b_{0,i}^2} + 2\sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{\sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2}.\end{aligned}$$

Let $X_t = 1 + \sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2 = X_{t-1} + \|\xi_{t,\Psi_i}\|^2$, where $X_0 = 1$. Then,

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{\sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2} &= \sum_{t=1}^{T} \frac{X_t - X_{t-1}}{X_t} = \sum_{t=1}^{T} 1 - \frac{X_{t-1}}{X_t} \\
&\leq \sum_{t=1}^{T} \log\left(\frac{X_t}{X_{t-1}}\right) \\
&= \log\left(\prod_{t=1}^{T} \frac{X_t}{X_{t-1}}\right) \\
&= \log\left(\frac{X_T}{X_0}\right) = \log\left(1 + \sum_{t=1}^{T} \|\xi_{s,\Psi_i}\|^2\right).
\end{aligned}$$

Hence, with probability at least $1 - \delta$:

$$\sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} \leq \frac{8\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}}{b_{0,i}^2} + 2\log\left(1 + \sum_{t=1}^{T} \|\xi_{s,\Psi_i}\|^2\right). \tag{7.21}$$

It remains to bound $\sum_{t=1}^{T} \|\xi_{s,\Psi_i}\|^2$. Note that

$$\begin{aligned}
\Pr\left[\sum_{t=1}^{T} \|\xi_{s,\Psi_i}\|^2 \geq u\right] &= \Pr\left[\exp\left(\sum_{t=1}^{T} \frac{\|\xi_{s,\Psi_i}\|^2}{\|\sigma_{\Psi_i}\|^2}\right) \geq \exp\left(\frac{u}{\|\sigma_{\Psi_i}\|^2}\right)\right] \\
&\leq \frac{\mathbb{E}\left[\exp\left(\sum_{t=1}^{T} \frac{\|\xi_{s,\Psi_i}\|^2}{\|\sigma_{\Psi_i}\|^2}\right)\right]}{\exp\left(\frac{u}{\|\sigma_{\Psi_i}\|^2}\right)} \\
&\leq \frac{\exp(T)}{\exp\left(\frac{u}{\|\sigma_{\Psi_i}\|^2}\right)} \qquad (\xi_{s,\Psi_i} \text{is } \|\sigma_{\Psi_i}\|^2\text{-subgaussian})
\end{aligned}$$

Choosing $u = \|\sigma_{\Psi_i}\|^2 T + \|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}$ gives that with probability at least $1 - \delta$, we have

$$\sum_{t=1}^{T} \|\xi_{s,\Psi_i}\|^2 \leq \|\sigma_{\Psi_i}\|^2 T + \|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}. \tag{7.22}$$

Having a high probability bound on the sum of the stochastic error of the subset-norm, we can combine both events from (7.21) and (7.22) to get that with probability at least $1 - 2\delta$:

$$\sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} \leq \frac{8\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}}{b_{0,i}^2} + 2\log\left(1 + \|\sigma_{\Psi_i}\|^2 T + \|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}\right). \tag{7.23}$$

Then we can also condition on the event that (7.23) happens and combine it with the event in (7.19) to get that with probability at least $1 - 2c\delta$ (assuming $c \geq 2$), we have

$$\sum_{t=1}^{T} \sum_{i=0}^{c-1} \frac{\|\nabla_{t,\Psi_i}\|_2^2}{b_{t,i}} \tag{7.24}$$

$$\leq \frac{\Delta_1}{\eta} + \sum_{i=0}^{c-1} \left(\ln T/\delta \, \|\sigma_{\Psi_i}\|^2 + 2\alpha \right) \sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} \tag{7.25}$$

$$+ \alpha + \sum_{i=0}^{c-1} \left(\ln T/\delta \, \|\sigma_{\Psi_i}\|^2 + \eta L + 4\alpha \right) \log \frac{b_{T,i}}{b_{0,i}} \tag{7.26}$$

$$\leq \frac{\Delta_1}{\eta} + \underbrace{\sum_{i=0}^{c-1} \left(\ln T/\delta \, \|\sigma_{\Psi_i}\|^2 + 2\alpha \right) \left(\frac{8 \, \|\sigma_{\Psi_i}\|^2 \log \frac{1}{\delta}}{b_{0,i}^2} + 2 \log \left(1 + \|\sigma_{\Psi_i}\|^2 T + \|\sigma_{\Psi_i}\|^2 \log \frac{1}{\delta} \right) \right)}_{=:H(\delta)} \tag{7.27}$$

$$+ \alpha + \sum_{i=0}^{c-1} \left(\ln T/\delta \, \|\sigma_{\Psi_i}\|^2 + \eta L + 4\alpha \right) \log \frac{b_{T,i}}{b_{0,i}}$$

$$= \frac{\Delta_1}{\eta} + H(\delta) + \alpha + \sum_{i=0}^{c-1} \left(\ln T/\delta \, \|\sigma_{\Psi_i}\|^2 + \eta L + 4\alpha \right) \log \frac{b_{T,i}}{b_{0,i}}. \tag{7.28}$$

First, note that $b_{T,i} \leq \|b_T\|_1 = \sum_{i=0}^{c-1} b_{T,i}$. Letting $b_{0,\min} := \min_i b_{0,i}$, we then have

$$\begin{aligned} \sum_{i=0}^{c-1} \left(\ln T/\delta \, \|\sigma_{\Psi_i}\|^2 + \eta L + 4\alpha \right) \log \frac{b_{T,i}}{b_{0,i}} &\leq \log \frac{\|b_T\|_1}{b_{0,\min}} \sum_{i=0}^{c-1} \left(\ln T/\delta \, \|\sigma_{\Psi_i}\|^2 + \eta L + 4\alpha \right) \\ &= \log \frac{\|b_T\|_1}{b_{0,\min}} \left(\ln T/\delta \, \|\sigma\|_2^2 + c\eta L + 4c\alpha \right). \end{aligned}$$

Now, note the LHS term $\sum_{t=1}^{T} \sum_{i=0}^{c-1} \frac{\|\nabla_{t,\Psi_i}\|_2^2}{b_{t,i}}$ of (7.26):

$$\left(\sum_{i=0}^{c-1} \frac{\|\nabla_{t,\Psi_i}\|_2^2}{b_{t,i}} \right) \left(\sum_{i=0}^{c-1} b_{t,i} \right) \geq \left(\sum_{i=0}^{c-1} \|\nabla_{t,\Psi_i}\|_2 \right)^2 \geq \sum_{i=0}^{c-1} \|\nabla_{t,\Psi_i}\|_2^2 = \|\nabla_t\|_2^2$$

$$\implies \frac{\|\nabla_t\|_2^2}{\left(\sum_{i=0}^{c-1} b_{t,i} \right)} \leq \sum_{i=0}^{c-1} \frac{\|\nabla_{t,\Psi_i}\|_2^2}{b_{t,i}}.$$

Now, $\sum_{i=0}^{c-1} b_{t,i} = \sum_{i=0}^{c-1} |b_{t,i}| = \|b_t\|_1$, so with probability $1 - 2c\delta$:

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{\|\nabla_t\|_2^2}{\|b_T\|_1} \leq \sum_{t=1}^{T} \frac{\|\nabla_t\|_2^2}{\|b_t\|_1} \leq \sum_{t=1}^{T} \sum_{i=0}^{c-1} \frac{\|\nabla_{t,\Psi_i}\|_2^2}{b_{t,i}} \\
\implies &\sum_{t=1}^{T} \|\nabla_t\|_2^2 \leq \|b_T\|_1 \sum_{t=1}^{T} \sum_{i=0}^{c-1} \frac{\|\nabla_{t,\Psi_i}\|_2^2}{b_{t,i}} \\
&\leq \|b_T\|_1 \left(\frac{\Delta_1}{\eta} + cH(\delta) + \left(\ln T/\delta \, \|\sigma\|_2^2 + c\eta L + 4c\alpha \right) \log \frac{\|b_T\|_1}{b_{0,\min}} \right) \qquad (7.29) \\
&\leq \|b_T\|_1 \left(\frac{\Delta_1}{\eta} + cH(\delta) + \left(\ln T/\delta \, \|\sigma\|_2^2 + c\eta L + 4c\alpha \right) \log \frac{\|b_T\|_1}{b_{0,\min}} \right). \qquad (7.30)
\end{aligned}$$

It remains to bound $\|b_T\|_1$. We start again from smoothness of f:

$$\begin{aligned}
\Delta_{t+1} - \Delta_t &\leq \langle \nabla_t, x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\
&= -\eta \left\langle \nabla_t, \frac{\widehat{\nabla} f_t}{b_t} \right\rangle + \frac{\eta^2 L}{2} \left\| \frac{\widehat{\nabla} f_t}{b_t} \right\|^2 \\
&= -\eta \left\langle \widehat{\nabla} f_t - \xi_t, \frac{\widehat{\nabla} f_t}{b_t} \right\rangle + \frac{\eta^2 L}{2} \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\widehat{\nabla} f_{t,\Psi_j}^2}{b_{t,i}^2} \\
&= -\eta \left\langle \widehat{\nabla} f_t, \frac{\widehat{\nabla} f_t}{b_t} \right\rangle + \eta \left\langle \xi_t, \frac{\widehat{\nabla} f_t}{b_t} \right\rangle + \frac{\eta^2 L}{2} \sum_{i=0}^{c-1} \frac{\left\| \widehat{\nabla} f_{t,\Psi_i} \right\|^2}{b_{t,i}^2} \\
&= -\eta \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\widehat{\nabla} f_{t,j}^2}{b_{t,i}} + \eta \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\xi_{t,j} \widehat{\nabla} f_{t,j}}{b_{t,i}} + \frac{\eta^2 L}{2} \sum_{i=0}^{c-1} \frac{\left\| \widehat{\nabla} f_{t,\Psi_i} \right\|^2}{b_{t,i}^2} \\
&= -\eta \sum_{i=0}^{c-1} \frac{\left\| \widehat{\nabla} f_{t,\Psi_i} \right\|^2}{b_{t,i}} + \frac{\eta^2 L}{2} \sum_{i=0}^{c-1} \frac{\left\| \widehat{\nabla} f_{t,\Psi_i} \right\|^2}{b_{t,i}^2} + \eta \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\xi_{t,j} \widehat{\nabla} f_{t,j}}{b_{t,i}}. \qquad (7.31)
\end{aligned}$$

Note that

$$\begin{aligned}
\sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\xi_{t,j} \widehat{\nabla} f_{t,j}}{b_{t,i}} &\leq \frac{1}{2} \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\xi_{t,j}^2}{b_{t,i}} + \frac{1}{2} \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\widehat{\nabla} f_{t,j}^2}{b_{t,i}} \\
&= \frac{1}{2} \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\xi_{t,j}^2}{b_{t,i}} + \frac{1}{2} \sum_{i=0}^{c-1} \frac{\left\| \widehat{\nabla} f_{t,\Psi_i} \right\|^2}{b_{t,i}}.
\end{aligned}$$

Plugging back in, we have

$$\Delta_{t+1} - \Delta_t \leq -\frac{\eta}{2} \sum_{i=0}^{c-1} \frac{\left\| \widehat{\nabla} f_{t,\Psi_i} \right\|^2}{b_{t,i}} + \eta^2 L \sum_{i=0}^{c-1} \frac{\left\| \widehat{\nabla} f_{t,\Psi_i} \right\|^2}{b_{t,i}^2} + \frac{\eta}{2} \sum_{i=0}^{c-1} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}}.$$

Summing over T and rearranging, we get

$$\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2}{b_{t,i}} \leq \frac{2\Delta_1}{\eta} + \sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}} + 2\eta L\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2}{b_{t,i}^2}$$

$$\implies \sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2}{b_{t,i}} \leq \frac{4\Delta_1}{\eta} + 2\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}} + \sum_{t=1}^{T}\sum_{i=0}^{c-1}\left(\frac{4\eta L}{b_{t,i}^2} - \frac{1}{b_{t,i}}\right)\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2.$$

We can bound $\sum_{t=1}^{T}\sum_{i=0}^{c-1}\left(\frac{4\eta L}{b_{t,i}^2} - \frac{1}{b_{t,i}}\right)\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2$ as follows. Consider $i \in [c]$. Let $\tau_i = \max\{t \leq T \mid b_{t,i} \leq 4\eta L\}$ so that $t \geq \tau_i$ implies $b_{t,i} > 4\eta L \iff \frac{4\eta L}{b_{t,i}^2} < \frac{1}{b_{t,i}}$:

$$\begin{aligned}
\sum_{t=1}^{T}\left(\frac{4\eta L}{b_{t,i}^2} - \frac{1}{b_{t,i}}\right)\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2 &= \sum_{t=1}^{\tau_i}\left(\frac{4\eta L}{b_{t,i}^2} - \frac{1}{b_{t,i}}\right)\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2 + \sum_{t=\tau_i+1}^{T}\left(\underbrace{\frac{4\eta L}{b_{t,i}^2} - \frac{1}{b_{t,i}}}_{<0}\right)\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2 \\
&\leq \sum_{t=1}^{\tau_i}\left(\frac{4\eta L}{b_{t,i}^2} - \frac{1}{b_{t,i}}\right)\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2 \\
&\leq 4\eta L\sum_{t=1}^{\tau_i}\frac{\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2}{b_{t,i}^2} \\
&\leq 8\eta L\log\frac{b_{\tau_i,i}}{b_{0,i}} \leq 8\eta L\log\frac{4\eta L}{b_{0,i}}.
\end{aligned}$$

Hence, we have

$$\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2}{b_{t,i}} \leq \frac{4\Delta_1}{\eta} + 2\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}} + 8\eta L\sum_{i=0}^{c-1}\log\frac{4\eta L}{b_{0,i}}.$$

Consider the LHS

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\left\|\widehat{\nabla} f_{t,\Psi_i}\right\|^2}{b_{t,i}} &= \sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{b_{t,i}^2 - b_{t-1,i}^2}{b_{t,i}} = \sum_{t=1}^{T}\sum_{i=0}^{c-1} b_{t,i} - \frac{b_{t-1,i}^2}{b_{t,i}} \\
&\geq \sum_{t=1}^{T}\sum_{i=0}^{c-1} b_{t,i} - \frac{b_{t-1,i}^2}{b_{t-1,i}} = \sum_{t=1}^{T}\sum_{i=0}^{c-1} b_{t,i} - b_{t-1,i} \\
&= \sum_{i=0}^{c-1}\sum_{t=1}^{T} b_{t,i} - b_{t-1,i} = \sum_{i=0}^{c-1} b_{T,i} - b_{0,i} \\
&= \|b_T\|_1 - \|b_0\|_1 .
\end{aligned}$$

Hence, we have

$$\|b_T\|_1 \leq \|b_0\|_1 + \frac{2\Delta_1}{\eta} + \sum_{i=0}^{c-1}\sum_{t=1}^{T}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}} + 8\eta Lc\log\frac{4\eta L}{b_{0,\min}}.$$

It remains to bound $\sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}}$ for each $i \in [c]$. Recall from (7.23), with probability at least $1-\delta$

$$\sum_{s=1}^{t} \|\xi_{t,\Psi_i}\|^2 \leq \sum_{s=1}^{t} \|\widehat{\nabla} f_{t,\Psi_i}\|^2 + 4\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta},\ \forall t \leq T.$$

We have with probability at least $1-2c\delta$,

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}} &= \sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{\sqrt{b_{0,i}^2 + \sum_{s=1}^{t} \|\widehat{\nabla} f_{s,\Psi_i}\|^2}} \\
&\overset{(1)}{\leq} \sum_{t=1}^{T} \frac{\xi_{t,i}^2}{\sqrt{b_{0,i}^2 + \left(\sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2 - 4\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}\right)^+}} \\
&\leq \frac{8\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}}{b_{0,i}} + 2\sqrt{2}\sqrt{\sum_{s=1}^{T} \|\xi_{s,\Psi_i}\|^2} \\
&\overset{(2)}{\leq} \frac{8\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}}{b_{0,i}} + 4\sqrt{\|\sigma_{\Psi_i}\|^2 T + \|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}},
\end{aligned}$$

where (1) is due to (7.20) and (2) is due to Lemma (7.22). Hence, we have that with probability at least $1-2c\delta$,

$$\begin{aligned}
\|b_T\|_1 &\leq \|b_0\|_1 + \frac{2\Delta_1}{\eta} + \sum_{i=0}^{c-1} \frac{8\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}}{b_{0,i}} + \sum_{i=0}^{c-1} 4\sqrt{\|\sigma_{\Psi_i}\|^2 T + \|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}} + 8\eta L c \log\frac{4\eta L}{b_{0,\min}} \\
&\leq \|b_0\|_1 + \frac{2\Delta_1}{\eta} + \frac{8\log\frac{1}{\delta}}{b_{0,\min}} \sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\|^2 + 4\sqrt{T} \sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| + \sqrt{\log\frac{1}{\delta}} \sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| + 8\eta L c \log\frac{4\eta L}{b_{0,\min}} \\
&= 4\sqrt{T} \sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| + \underbrace{\|b_0\|_1 + \frac{2\Delta_1}{\eta} + \frac{8\log\frac{1}{\delta}}{b_{0,\min}} \|\sigma\|_2^2 + \sqrt{\log\frac{1}{\delta}} \sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| + 8\eta L c \log\frac{4\eta L}{b_{0,\min}}}_{=:I(\delta)}.
\end{aligned}$$

Hence, we can combine (7.30) with the bound for $\|b_T\|_1$ to get that with probability $1-6c\delta$:

$$\begin{aligned}
&\sum_{t=1}^{T} \|\nabla_t\|_2^2 \\
&\leq \|b_T\|_1 \left(\frac{\Delta_1}{\eta} + H(\delta) + \left(\ln T/\delta \|\sigma\|_2^2 + c\eta L + 4c\sigma_{\max}\sqrt{c\log\frac{1}{\delta}}\right) \log\frac{\|b_T\|_1}{b_{0,\min}}\right) \\
&\leq \left(4\sqrt{T} \sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| + I(\delta)\right) \cdot \\
&\quad \left(\frac{\Delta_1}{\eta} + H(\delta) + \left(\ln T/\delta \|\sigma\|_2^2 + c\eta L + 4c^{3/2}\sigma_{\max}\sqrt{\log\frac{1}{\delta}}\right) \log\left(\frac{4\sqrt{T}\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| + I(\delta)}{b_{0,\min}}\right)\right).
\end{aligned}$$

Dividing both sides by T, we get the theorem that with probability $1 - 6c\delta$:

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla_t\|_2^2 \le G(\delta)\cdot\left(\frac{4\sum_{i=0}^{c-1}\|\sigma_{\Psi_i}\|}{\sqrt{T}}+\frac{I(\delta)}{T}\right), \text{ where } G(\delta) \text{ and } I(\delta) \text{ are polylog terms:}$$

$$G(\delta) := \frac{\Delta_1}{\eta} + H(\delta) + \left(\ln T/\delta\,\|\sigma\|_2^2 + c\eta L + 4c^{3/2}\sigma_{\max}\sqrt{\log\frac{1}{\delta}}\right)\log\left(\frac{4\sqrt{T}\sum_{i=0}^{c-1}\|\sigma_{\Psi_i}\| + I(\delta)}{b_{0,\min}}\right)$$

$$I(\delta) := \|b_0\|_1 + \frac{2\Delta_1}{\eta} + \frac{8\log\frac{1}{\delta}}{b_{0,\min}}\|\sigma\|_2^2 + \sqrt{\log\frac{1}{\delta}}\sum_{i=0}^{c-1}\|\sigma_{\Psi_i}\| + 8\eta Lc\log\frac{4\eta L}{b_{0,\min}}$$

$$H(\delta) := \sum_{i=0}^{c-1}\left(\ln(T/\delta)\,\|\sigma_{\Psi_i}\|^2 + 2\alpha\right)\left(\frac{8\|\sigma_{\Psi_i}\|^2\log\frac{1}{\delta}}{b_{0,i}^2} + 2\log\left(1+\|\sigma_{\Psi_i}\|^2 T + \|\sigma_{\Psi_i}\|^2\log\frac{1}{\delta}\right)\right).$$

We are done. □

Chapter 8

Subspace-Momentum

8.1 Introduction

Low rank methods like LoRA (Hu et al., 2021), and more recently, GaLore (Zhao et al., 2024) are popular methods for reducing memory during training. However, these methods often lack theoretical guarantees unless stronger conditions are assumed. We propose Subspace-Momentum (SM) that ensures convergence under standard assumptions by *incorporating the orthogonal complement* of the stochastic gradient to the optimization (Figure 8.1), where the convergence analysis can be decoupled between the two orthogonal subspaces: one for SGD and another for SGD with momentum. For rank r, Subspace-Momentum uses only $O(r)$ memory for the optimization state due to only maintaining momentum state in the low rank space and the use of SGD in the orthogonal subspace.

8.2 Subspace-Momentum

Existing algorithmic compression approaches like GaLore (Zhao et al., 2024), GRASS (Muhamed et al., 2024), and FLORA (Hao et al., 2024) project the gradient to a lower dimensional space $\mathbb{R}^k$ for updating the optimizer state via some linear operator $P : \mathbb{R}^d \to \mathbb{R}^k$ such that $P^*P : \mathbb{R}^d \to \mathbb{R}^d$ is a projection, where $P^* : \mathbb{R}^k \to \mathbb{R}^d$ is the adjoint operator of P. More concretely, given a stochastic gradient $\widehat{\nabla} f(x_t) \in \mathbb{R}^d$ at time t, a low-dimensional version $c_t := P\widehat{\nabla} f(x_t) \in \mathbb{R}^k$ is computed that is used to update the states before projecting back to $\mathbb{R}^d$ for update:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\, c_t;\; v_t^2 = \beta_2 v_{t-1}^2 + (1-\beta_2)\, c_t^2;\; x_{t+1} = x_t - P^*\left(m_t / v_t\right). \quad (8.1)$$

This update essentially performs adaptive optimization in the row span $U \subseteq \mathbb{R}^d$ of P when viewed as a linear operator, with $\dim(U) = k$. For example, GaLore (Zhao et al., 2024) utilizes the top k singular vectors of snapshots of stochastic gradients, and FLORA (Hao et al., 2024) simply projects to a random subspace using dense Gaussian matrices. Due to the optimization happening only in a low rank subspace, convergence is not guaranteed unless stronger conditions are assumed.

Subspace momentum guarantees convergence by incorporating the orthogonal complement of $P^*P\widehat{\nabla} f(x_t) \in U$ that lives in the orthogonal complement $U^\perp$ of U (with $U \oplus U^\perp = \mathbb{R}^d$), of which we can compute via $\left(\widehat{\nabla} f(x_t) - P^*P\widehat{\nabla} f(x_t)\right) \in U^\perp$. This gives rise to Subspace Momentum (SM) presented in Algorithm 11.

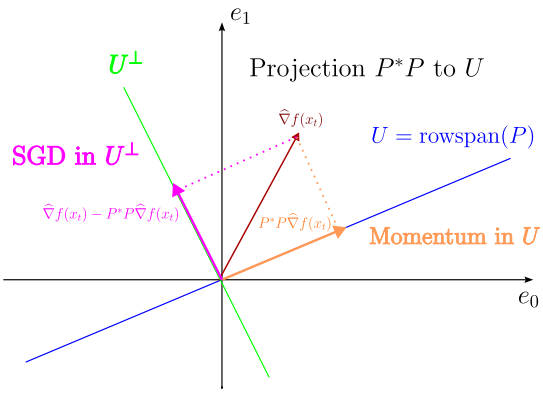

FIGURE 8.1: Subspace Momentum.

Algorithm 11 SGD with Subspace Momentum (SM).

Require: Projection $P : \mathbb{R}^d \to \mathbb{R}^k$ and its adjoint P^*

1: **for** $t = 1, 2, \ldots, T$ **do**
2: Obtain stochastic gradient $\widehat{\nabla} f(x_t)$
3: $m_t = \beta_1 m_{t-1} + (1-\beta_1) P\widehat{\nabla} f(x_t)$ ▷ Momentum in subspace
4: $r_t = \widehat{\nabla} f(x_t) - P^*P\widehat{\nabla} f(x_t)$ ▷ Orthogonal complement
5: $x_{t+1} = x_t - \eta(P^* m_t + r_t)$ ▷ Step in both spaces
6: **end for**

8.3 High-probability Convergence of Subspace-Momentum

We show that Subspace-Momentum, Algorithm 11, converges with high-probability under the standard assumption of smoothness and σ-subgaussian stochastic gradient noise in Theorem 8.3.1.

Theorem 8.3.1. *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and lower bounded by f_*. Assume unbiased stochastic gradients $\widehat{\nabla} f(x_t)$ with σ-subgaussian stochastic gradient noise. Then, the iterates x_t given by SGD with Subspace-Momentum (Algorithm 11) with step size $\eta := \frac{1}{\alpha\sqrt{T}}$ for $\alpha := \frac{(3-\beta)L}{2(1-\beta)}$ satisfies the following with probability at least $1-\delta$*

$$T \sum_{i=1}^{T} \|\nabla_t\|^2 \leq \frac{8\Delta_1 \alpha}{T} + \frac{7\sigma\sqrt{\alpha\Delta_1}}{\sqrt{T}} + \frac{48\sigma^2 \log(1/\delta)}{T},$$

where $\Delta_1 := f(x_1) - f_*$ is the initial function gap.

We observe that Theorem 8.3.1 has a similar convergence rate to vanilla SGD. The proof is presented in Section 8.5.2, where we provide some intuition for the algorithm and the proof.

8.4 Implementation: SN+SM and Choice of Projection

The projection P in Algorithm 11 can be a dense random projection as in FLORA, projection to top k singular vectors as in GaLore, or projection to random standard bases (sampling coordinates) as in GRASS. Note that Subspace Momentum maintains the same memory footprint, $O(k)$, as existing low-rank optimizers. However,

the update step of SM is full rank: it performs momentum only in $U := \text{rowspan}(P)$ while performs SGD in $U^\perp$. Unlike joint compression techniques such as GaLore (8.1), SM only affects the momentum term, so it is modular and hence fits into the framework of Algorithm 9 for which we can combine it with different adaptive step sizes such as subset-norm.[1]

8.4.1 Projection Selection

Using a subspace from SVD can be expensive for larger models and consumes additional memory to store the projection. Gradient-independent projections like random gaussian as in FLORA (Hao et al., 2024) avoids the expensive SVD computation and can save memory by storing the pseudorandom seed (at the cost of recomputating the projection at every step). One can further speed up the random projection by using a faster (sparse) random projection like the Subsampled-Randomized Hadamard Transform (SRHT) used in the Fast-JL transform (Ailon and Chazelle, 2009). Random projections like SRHT can also be used to approximate SVD (Appx-SVD) computation (Halko et al., 2011) that can be much faster than full SVD. Finally, the cheapest projection is just selecting a subspace of random standard bases. Recently, GRASS (Muhamed et al., 2024) explores this idea and tested sampling random rows and columns with large norms. We examine different choices for the subspace projection and compare their time, space, and performance tradeoffs in Table 9.4.

Note that the choice of the projection is important as some projections are more computationally and memory expensive than other, although trading other qualities for given the cost. Simple projections like selecting a subset of coordinates for momentum (Subset-Momentum) are not only faster but enables simple distributed training like FSDP unlike more complex subspace selection mechanism that requires additional priors about the parameters (shape, low-rank, etc.) that might not always satisfied.

8.4.2 Subspace Switching and Projection Updates

Algorithm 11 and the accompanying theory in Section 8.3 are only for a fixed projection. However, from our experiments, we find that performing subspace switching every G steps (as in GaLore) to be beneficial, especially for smaller ranks. Section 9.4.2 contains more details on ranks and updating projections. We incorporate projection updates in our main algorithms by picking a projection update gap and then fully resetting the momentum term to *zero* when we switch (in contrast to GaLore's accumulated statistics when switching subspace).

8.5 Subspace-Momentum Convergence Proofs

In this section, we provide a high-probability convergence proof for SGD with Subspace-Momentum for non-convex smooth objective under sub-gaussian gradient noise.

[1] Section 9.4.3 shows the different combinations of momentum and step sizes.

8.5.1 Setup and Intuition

Notations. Given a linear operator $P : \mathbb{R}^d \to \mathbb{R}^k$, we have $P^* : \mathbb{R}^k \to \mathbb{R}^d$ is P's adjoint[2], and we consider $P^*P : \mathbb{R}^d \to \mathbb{R}^d$ is a projection operator i.e. P^*P is a bounded linear operator such that $(P^*P)^2 = P^*P$. Given a space $V \subseteq \mathbb{R}^d$, we denote its orthogonal subspace by $V^\perp := \{v \in \mathbb{R}^d : \langle v, u\rangle = 0,\ \forall u \in V\}$.

Let $U = \text{row}\,(P) \subseteq \mathbb{R}^d$ be the row span of P. Let $\Psi : \mathbb{R}^d \to U$ be $\Psi(x) = P^*Px$ and $\Psi^\perp : \mathbb{R}^d \to U^\perp$ be $\Psi^\perp(x) = x - P^*Px$. Then for any vector x in $\mathbb{R}^d$, have the orthogonal decomposition

$$x = \Psi(x) + \Psi^\perp(x).$$

SGD with Subspace Momentum. Let $g_t := \widehat{\nabla} f f(x_t)$ denotes the stochastic gradient at time t. Let $\hat{c}_t = Pg_t$, $g_t^U = \Psi g_t = P^*Pg_t \in U$, and $g_t^\perp = g_t - g_t^U \in U^\perp$. Let $\nabla_t := \nabla f(x_t)$ be a short hand for the gradient at time t and let $\nabla_t^U := \Psi\left(\nabla f(x_t)\right) \in U$ and $\nabla_t^\perp := \Psi^\perp\left(\nabla f(x_t)\right) \in U^\perp$ be the orthogonal decomposition of $\nabla f(x_t)$ with respect to U and $U^\perp$, so that $\nabla_t = \nabla_t^U + \nabla_t^\perp$. Note that the superscript of a variable tries to suggest the space that it lives in (either U or $U^\perp$). We have the following update rule for subspace momentum:

$$\begin{aligned}
\hat{m}_t &= \beta \hat{m}_{t-1} + (1-\beta) P g_t \\
g_t^\perp &= g_t - P^*P g_t \\
m_t &= P^* \hat{m}_t \\
x_{t+1} &= x_t - \eta \left(m_t + g_t^\perp\right).
\end{aligned}$$

Note that

$$\begin{aligned}
m_t &= \beta P^* \hat{m}_{t-1} + (1-\beta) P^*P g_t \\
&= \beta m_{t-1} + (1-\beta) g_t^U.
\end{aligned}$$

Expanding the terms, we see that this is just momentum in U

$$\begin{aligned}
m_t &= \beta P^* \hat{m}_{t-1} + (1-\beta) P^*P g_t \\
&= \beta P^* \hat{m}_{t-1} + (1-\beta) g_t^U \\
&= \beta^2 P^* \hat{m}_{t-2} + (1-\beta)\beta g_{t-1}^U + (1-\beta) g_t^U \\
&= (1-\beta) \sum_{i=0}^{t} \beta^i g_{t-i}^U. \qquad (8.2)
\end{aligned}$$

Hence, we can think of the update of SGD-SM as performing two separate algorithms in the orthogonal subspaces: momentum in the subspace U and SGD in the subspace $U^\perp$ (see also Figure 8.1) i.e. if we decompose x_t into its orthogonal components $x_t = x_t^U + x_t^\perp$, then

$$\begin{aligned}
x_{t+1}^U &= x_t^U - \eta m_t \\
&= x_t^U - \eta(1-\beta) \sum_{i=0}^{t} \beta^i g_{t-i}^U \\
x_{t+1}^\perp &= x_t^\perp - \eta g_t^\perp.
\end{aligned}$$

[2] In $\mathbb{R}^d$, the adjoint P^* of a linear operator P is the linear operator given by the *transpose* of the matrix representation of P. We can also generalize Subspace-Momentum to general Hilbert spaces.

For our analysis, let $\xi_t := g_t - \nabla_t$ denote the stochastic gradient error at time t. We can further decompose the error into its subspace components:

$$\begin{aligned}\xi_t &= \xi_t^U + \xi_t^\perp \\ &= \left(g_t^U - \nabla_t^U\right) + \left(g_t^\perp - \nabla_t^\perp\right).\end{aligned}$$

Basic facts. We establish some facts for subspace momentum.

1. Pythagorean: $\|g_t\|^2 = \left\|g_t^U\right\|^2 + \left\|g_t^\perp\right\|^2$ and $\|\nabla_t\|^2 = \left\|\nabla_t^U\right\|^2 + \left\|\nabla_t^\perp\right\|^2$ and so on for these decompositions.

2. Subspace smoothness: If f is smooth $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$, then due to contraction property of the projection operator, we have that the projected gradients of f are also L-Lipschitz:

$$\begin{aligned}\|P^*P\nabla f(x) - P^*P\nabla f(y)\|^2 = \|\nabla f(x) - \nabla f(y)\| \\ \leq L\|x - y\|.\end{aligned} \tag{8.3}$$

3. Subspace non-bias:

$$\begin{aligned}\mathbb{E}\left[g_t^U - \nabla_t^U\right] &= \mathbb{E}\left[\xi_t^U\right] \\ &= \mathbb{E}\left[P^*P\left(g_t - \nabla_t\right)\right] \\ &= 0,\end{aligned}$$

   and similarly for the orthogonal subspace

$$\begin{aligned}\mathbb{E}\left[g_t^\perp - \nabla_t^\perp\right] &= \mathbb{E}\left[\xi_t^\perp\right] \\ &= \mathbb{E}\left[\xi_t - \xi_t^U\right] \\ &= 0.\end{aligned}$$

4. Subspace bounded variance: if the stochastic gradient's variance is bounded, then its subspace components are also bounded $\mathbb{E}\left[\|\xi_t\|^2\right]$:

$$\begin{aligned}\mathbb{E}\left[\left\|g_t^U - \nabla_t^U\right\|^2\right] &= \mathbb{E}\left[\left\|\xi_t^U\right\|^2\right] \\ &= \mathbb{E}\left[\|\xi_t\|^2 - \left\|\xi_t^\perp\right\|^2\right] \\ &\leq \sigma^2 - \mathbb{E}\left[\left\|\xi_t^\perp\right\|^2\right],\end{aligned}$$

   and similarly,

$$\mathbb{E}\left[\left\|\xi_t^\perp\right\|^2\right] \leq \sigma^2 - \mathbb{E}\left[\left\|\xi_t^U\right\|^2\right].$$

8.5.2 Subspace-Momentum convergence proof

Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and stochastic gradients $\widehat{\nabla} f f(x_t) = g_t$ is unbiased, i.e. $\mathbb{E}[g_t] = \nabla f(x_t)$, and has σ-sub-gaussian noise, i.e. $\mathbb{E}[\exp(\lambda^2\|g_t - \nabla f(x_t)\|^2)] \leq$

$\exp(\lambda^2\sigma^2)$ for all λ s.t. $|\lambda| \leq 1/\sigma$. First, we will show an error bound that is a starting point for the high-probability convergence results.

Lemma 8.5.1. *If f is L-smooth, then SGD with Subspace-Momentum (Algorithm 11) yields*

$$f(x_{T+1}) - f(x_1) \leq -\eta \sum_{t=1}^{T} \|\nabla_t\|^2 - \eta \sum_{t=1}^{T} \langle \nabla_t, \xi_t \rangle + \frac{(3-\beta)L\eta^2}{2(1-\beta)} \sum_{t=1}^{T} \|g_t\|^2 .$$

Remark 8. *Lemma 8.5.1 shows that the optimization error of SGD-SM is quite similar to SGD-M.*

Proof. Note that $m_t \in U$ and $r_t \in U^\perp$. Starting with smoothness, we have

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\ &= f(x_t) - \eta \left\langle \nabla f(x_t), m_t + g_t^\perp \right\rangle + \frac{\eta^2 L}{2} \left\| m_t + g_t^\perp \right\|^2 \\ &= f(x_t) - \eta \langle \nabla f(x_t), m_t \rangle - \eta \left\langle \nabla f(x_t), g_t^\perp \right\rangle + \frac{\eta^2 L}{2} \left\| m_t + g_t^\perp \right\|^2 . \end{aligned}$$

We have

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq -\eta \left\langle \nabla_t^U, m_t \right\rangle - \eta \left\langle \nabla_t^\perp, g_t^\perp \right\rangle + \frac{\eta^2 L}{2} \left\| m_t + g_t^\perp \right\|^2 \\ &= -\eta \left\langle \nabla_t^U, m_t \right\rangle - \eta \left\langle \nabla_t^\perp, g_t^\perp \right\rangle + \frac{\eta^2 L}{2} \|m_t\|^2 + \frac{\eta^2 L}{2} \left\| g_t^\perp \right\|^2 . \\ &\qquad \text{(Pythagorean)} \end{aligned}$$

Summing it up, we get

$$f(x_{T+1}) - f(x_1) \leq \underbrace{-\eta \sum_{t=1}^{T} \left\langle \nabla_t^U, m_t \right\rangle + \frac{\eta^2 L}{2} \sum_{t=1}^{T} \|m_t\|^2}_{\text{SGD with momentum error in } U} \underbrace{-\eta \sum_{t=1}^{T} \left\langle \nabla_t^\perp, g_t^\perp \right\rangle + \frac{\eta^2 L}{2} \sum_{t=1}^{T} \left\| g_t^\perp \right\|^2}_{\text{vanilla SGD error in } U^\perp} . \tag{8.4}$$

We analyze $-\eta \left\langle \nabla_t^U, m_t \right\rangle + \frac{\eta^2 L}{2} \|m_t\|^2$ and $-\eta \left\langle \nabla_t^\perp, g_t^\perp \right\rangle + \frac{\eta^2 L}{2} \left\| g_t^\perp \right\|^2$ separately. Intuitively, the error within each subspace is controlled by their respective algorithm. Investigating the momentum term, we have

$$\begin{aligned} -\left\langle \nabla_t^U, m_t \right\rangle &= -\left\langle \nabla_t^U, \beta m_{t-1} + (1-\beta) g_t^U \right\rangle \\ &= -\beta \left\langle \nabla_t^U, m_{t-1} \right\rangle - (1-\beta) \left\langle \nabla_t^U, g_t^U \right\rangle \\ &= -\beta \left\langle \nabla_t^U - \nabla_{t-1}^U, m_{t-1} \right\rangle - \beta \left\langle \nabla_{t-1}^U, m_{t-1} \right\rangle - (1-\beta) \left\langle \nabla_t^U, g_t^U \right\rangle . \end{aligned}$$

We examine $-\left\langle \nabla_t^U - \nabla_{t-1}^U, m_{t-1} \right\rangle$:

$$\begin{aligned}
-\left\langle \nabla_t^U - \nabla_{t-1}^U, m_{t-1} \right\rangle &= -\left\langle \nabla_t - \nabla_{t-1}, m_{t-1} \right\rangle + \left\langle \nabla_t^\perp - \nabla_{t-1}^\perp, m_{t-1} \right\rangle \\
&= -\left\langle \nabla_t - \nabla_{t-1}, m_{t-1} \right\rangle \\
&\le \left\| \nabla_t - \nabla_{t-1} \right\| \left\| m_{t-1} \right\| \\
&\le L \left\| x_t - x_{t-1} \right\| \left\| m_{t-1} \right\| \\
&= \eta L \left\| m_{t-1} + g_{t-1}^\perp \right\| \left\| m_{t-1} \right\| \\
&\le \eta L \left\| m_{t-1} + g_{t-1}^\perp \right\|^2 .
\end{aligned}$$

Now we have

$$\begin{aligned}
-\left\langle \nabla_t^U, m_t \right\rangle &\le \eta L \beta \left\| m_{t-1} + g_{t-1}^\perp \right\|^2 - \beta \left\langle \nabla_{t-1}^U, m_{t-1} \right\rangle - (1-\beta) \left\langle \nabla_t^U, g_t^U \right\rangle \\
&\le \eta L \sum_{i=1}^{t-1} \beta^{t-i} \left\| m_i + g_i^\perp \right\|^2 - (1-\beta) \sum_{i=1}^{t} \beta^{t-i} \left\langle \nabla_i^U, g_i^U \right\rangle .
\end{aligned}$$

Summing over t, we have

$$\begin{aligned}
-\eta \sum_{t=1}^{T} \left\langle \nabla_t^U, m_t \right\rangle &\le \eta^2 L \sum_{t=1}^{T} \sum_{i=1}^{t-1} \beta^{t-i} \left\| m_i + g_i^\perp \right\|^2 - (1-\beta)\eta \sum_{t=1}^{T} \sum_{i=1}^{t} \beta^{t-i} \left\langle \nabla_i^U, g_i^U \right\rangle \\
&= L\eta^2 \sum_{i=1}^{T} \sum_{t=i}^{T} \beta^{t-i} \left\| m_i + g_i^\perp \right\|^2 - (1-\beta)\eta \sum_{i=1}^{T} \sum_{t=i}^{T} \beta^{t-i} \left\langle \nabla_i^U, g_i^U \right\rangle \\
&\qquad\qquad \text{(swap the sum)} \\
&\le L\eta^2 \sum_{i=1}^{T} \left\| m_i + g_i^\perp \right\|^2 \sum_{t=i}^{T} \beta^t - (1-\beta)\eta \sum_{i=1}^{T} \left\langle \nabla_i^U, g_i^U \right\rangle \sum_{t=i}^{T} \beta^t \\
&\le \frac{L\eta^2}{1-\beta} \sum_{i=1}^{T} \left\| m_i + g_i^\perp \right\|^2 - \eta \sum_{i=1}^{T} \left\langle \nabla_i^U, g_i^U \right\rangle \\
&= \frac{L\eta^2}{1-\beta} \sum_{i=1}^{T} \left(\left\| m_i \right\|^2 + \left\| g_i^\perp \right\|^2 \right) - \eta \sum_{i=1}^{T} \left\langle \nabla_i^U, \xi_i^U \right\rangle - \eta \sum_{i=1}^{T} \left\| \nabla_i^U \right\|^2 .
\end{aligned}$$

Now, we look at $\sum_{i=1}^{T} \left\| m_i \right\|^2$:

$$\begin{aligned}
\sum_{t=1}^{T} \left\| m_t \right\|^2 &= \sum_{t=1}^{T} \left\| \beta m_{t-1} + (1-\beta) g_t^U \right\|^2 \\
&\le \sum_{t=1}^{T} \beta \left\| m_{t-1} \right\|^2 + (1-\beta) \left\| g_t^U \right\|^2 \qquad \text{(convexity of } \left\| \cdot \right\|^2\text{)} \\
&\le \sum_{t=1}^{T} \beta \left\| m_t \right\|^2 + (1-\beta) \sum_{t=1}^{T} \left\| g_t^U \right\|^2 \\
\implies \sum_{t=1}^{T} \left\| m_t \right\|^2 &\le \sum_{t=1}^{T} \left\| g_t^U \right\|^2 .
\end{aligned}$$

Examining the momentum error terms, we get

$$\begin{aligned}
&-\eta \sum_{t=1}^{T} \left\langle \nabla_t^U, m_t \right\rangle + \frac{\eta^2 L}{2} \sum_{t=1}^{T} \|m_t\|^2 \\
&\leq \left(\frac{L\eta^2}{1-\beta} + \frac{\eta^2 L}{2} \right) \sum_{i=1}^{T} \|m_i\|^2 + \frac{L\eta^2}{1-\beta} \sum_{i=1}^{T} \left\| g_i^{\perp} \right\|^2 - \eta \sum_{i=1}^{T} \left\langle \nabla_i^U, \xi_i^U \right\rangle - \eta \sum_{i=1}^{T} \left\| \nabla_i^U \right\|^2 \\
&\leq \left(\frac{(3-\beta) L\eta^2}{2(1-\beta)} \right) \sum_{i=1}^{T} \left\| g_i^U \right\|^2 + \frac{L\eta^2}{1-\beta} \sum_{i=1}^{T} \left\| g_i^{\perp} \right\|^2 - \eta \sum_{i=1}^{T} \left\langle \nabla_i^U, \xi_i^U \right\rangle - \eta \sum_{i=1}^{T} \left\| \nabla_i^U \right\|^2 .
\end{aligned} \tag{8.5}$$

We consider the SGD error terms in the orthogonal subspace:

$$\begin{aligned}
-\eta \left\langle \nabla_t^{\perp}, g_t^{\perp} \right\rangle + \frac{\eta^2 L}{2} \left\| g_t^{\perp} \right\|^2 &= -\eta \left\langle \nabla_t^{\perp}, \nabla_t^{\perp} - g_t^{\perp} \right\rangle - \eta \left\| \nabla_t^{\perp} \right\|^2 + \frac{\eta^2 L}{2} \left\| g_t^{\perp} \right\|^2 \\
&= -\eta \left\langle \nabla_t^{\perp}, \xi_t^{\perp} \right\rangle - \eta \left\| \nabla_t^{\perp} \right\|^2 + \frac{\eta^2 L}{2} \left\| g_t^{\perp} \right\|^2 .
\end{aligned} \tag{8.6}$$

Now we are ready to combine (8.5) and (8.6). First note the common terms $\left\| g_i^{\perp} \right\|^2$ in both equations combine to a sum similarly to $\left\| g_t^U \right\|^2$:

$$\underbrace{\frac{L\eta^2}{1-\beta} \sum_{t=1}^{T} \left\| g_t^{\perp} \right\|^2}_{\text{momentum}} + \underbrace{\frac{\eta^2 L}{2} \sum_{t=1}^{T} \left\| g_t^{\perp} \right\|^2}_{\text{SGD}} = \left(\frac{(3-\beta) L\eta^2}{2(1-\beta)} \right) \sum_{t=1}^{T} \left\| g_t^{\perp} \right\|^2 .$$

Combining both terms, we see that the terms are combined from both subspaces (red from (8.5) and blue from (8.6)):

$$\begin{aligned}
-\eta \sum_{t=1}^{T} \left\| \nabla_t^U \right\|^2 - \eta \sum_{t=1}^{T} \left\| \nabla_t^{\perp} \right\|^2 &= -\eta \sum_{t=1}^{T} \|\nabla_t\|^2 \\
-\eta \sum_{t=1}^{T} \left\langle \nabla_t^U, \xi_t^U \right\rangle - \eta \sum_{t=1}^{T} \left\langle \nabla_t^{\perp}, \xi_t^{\perp} \right\rangle &= -\eta \sum_{t=1}^{T} \langle \nabla_t, \xi_t \rangle \\
\frac{(3-\beta) L\eta^2}{2(1-\beta)} \sum_{t=1}^{T} \left\| g_t^U \right\|^2 + \left(\frac{L\eta^2}{1-\beta} + \frac{\eta^2 L}{2} \right) \sum_{t=1}^{T} \left\| g_t^{\perp} \right\|^2 &= \frac{(3-\beta) L\eta^2}{2(1-\beta)} \sum_{t=1}^{T} \|g_t\|^2 .
\end{aligned}$$

Plugging everything back into (8.4), we have

$$f(x_{T+1}) - f(x_1) \leq -\eta \sum_{t=1}^{T} \|\nabla_t\|^2 - \eta \sum_{t=1}^{T} \langle \nabla_t, \xi_t \rangle + \frac{(3-\beta) L\eta^2}{2(1-\beta)} \sum_{t=1}^{T} \|g_t\|^2 . \tag{8.7}$$

□

8.5.3 Proof of Theorem 8.3.1.

Proof of Theorem 8.3.1. Starting from Lemma 8.5.1 and letting $\alpha = \frac{(3-\beta)L}{2(1-\beta)}$ and $\Delta_1 := f(x_1) - f_*$ for simplicity, we have

$$\begin{aligned}
&\Delta_{T+1} - \Delta_1 \\
\leq& -\eta \sum_{i=1}^{T} \|\nabla_t\|^2 - \eta \sum_{i=1}^{T} \langle \nabla_t, \xi_t \rangle + \alpha\eta^2 \sum_{t=1}^{T} \|g_t\|^2 \\
=& -\eta \sum_{i=1}^{T} \|\nabla_t\|^2 - \eta \sum_{i=1}^{T} \langle \nabla_t, \xi_t \rangle + \alpha\eta^2 \sum_{t=1}^{T} \|\xi_t + \nabla_t\|^2 \\
=& \eta (\alpha\eta - 1) \sum_{i=1}^{T} \|\nabla_t\|^2 + \eta (\alpha\eta - 1) \sum_{i=1}^{T} \langle \nabla_t, \xi_t \rangle + \alpha\eta^2 \sum_{t=1}^{T} \|\xi_t\|^2 .
\end{aligned}$$

Rearranging and defining some weight $w > 0$, we have

$$w (\Delta_{T+1} - \Delta_1) + \eta w (1 - \alpha\eta) \sum_{i=1}^{T} \|\nabla_t\|^2 \leq \eta w (\alpha\eta - 1) \sum_{i=1}^{T} \langle \nabla_t, \xi_t \rangle + \alpha w \eta^2 \sum_{t=1}^{T} \|\xi_t\|^2 .$$

Let $\mathcal{F}_t := \sigma(\xi_1, \ldots, \xi_{t-1})$ denote the natural filtration. Now, since $\mathbb{E}\left[\sum_{t=1}^{T} \langle \nabla_t, \xi_t \rangle\right] = 0$ and ξ_t is σ-sub-gaussian, we have that $\nabla_t \in \mathcal{F}_t$ and so if $0 \leq w\alpha\eta^2 \leq \frac{1}{4\sigma^2}$, Corollary 4.3.3 implies

$$\mathbb{E}\left[\exp\left(w\eta (\alpha\eta - 1) \langle \nabla_t, \xi_t \rangle + w\alpha\eta^2 \|\xi_t\|^2\right) \mid \mathcal{F}_t\right] \leq \exp\left(3\sigma^2 \left(w\alpha\eta^2 + w^2\eta^2 (\alpha\eta - 1)^2 \|\nabla_t\|^2\right)\right),$$

Then Lemma 4.3.4 implies that with probability at least $1 - \delta$, we have

$$\begin{aligned}
w\eta (\alpha\eta - 1) \sum_{t=1}^{T} \langle \nabla_t, \xi_t \rangle + w\alpha\eta^2 \sum_{t=1}^{T} \|\xi_t\|^2 &\leq 3\sigma^2 \sum_{t=1}^{T} \left(w\alpha\eta^2 + w^2\eta^2 (\alpha\eta - 1)^2 \|\nabla_t\|^2\right) + \log(1/\delta) \\
&= 3\sigma^2 w\eta^2\alpha T + 3\sigma^2 w^2\eta^2 (\alpha\eta - 1)^2 \sum_{t=1}^{T} \|\nabla_t\|^2 + \log(1/\delta) .
\end{aligned}$$

Then with probability at least $1 - \delta$, we have

$$\begin{aligned}
&w (\Delta_{T+1} - \Delta_1) + \eta w (1 - \alpha\eta) \sum_{i=1}^{T} \|\nabla_t\|^2 \\
&\qquad \leq \eta w (\alpha\eta - 1) \sum_{i=1}^{T} \langle \nabla_t, \xi_t \rangle + \alpha w \eta^2 \sum_{t=1}^{T} \|\xi_t\|^2 \\
&\qquad \leq 3\sigma^2 w\eta^2\alpha T + 3\sigma^2 w^2\eta^2 (\alpha\eta - 1)^2 \sum_{t=1}^{T} \|\nabla_t\|^2 + \log(1/\delta) \\
\implies \eta w (1 - \alpha\eta) \sum_{i=1}^{T} \|\nabla_t\|^2 &\leq w\Delta_1 + 3\sigma^2 w\eta^2\alpha T + 3\sigma^2 w^2\eta^2 (\alpha\eta - 1)^2 \sum_{t=1}^{T} \|\nabla_t\|^2 + \log(1/\delta) .
\end{aligned}$$

Combining the $\|\nabla_t\|^2$ terms, we get

$$\left(\eta w (1 - \alpha\eta) - 3\sigma^2 w^2\eta^2 (\alpha\eta - 1)^2\right) \sum_{i=1}^{T} \|\nabla_t\|^2 \leq w\Delta_1 + 3\sigma^2 w\eta^2\alpha T + \log(1/\delta) . \tag{8.8}$$

Setting $w = \frac{1}{12\sigma^2\eta}$, then

$$\begin{aligned}\eta w\left(1-\alpha\eta\right) - 3\sigma^2 w^2\eta^2\left(\alpha\eta-1\right)^2 &= \eta w\left(1-\alpha\eta - 3\sigma^2 w\eta\left(\alpha\eta-1\right)^2\right)\\ &= \eta w\left(1-\alpha\eta-\frac{1}{4}\left(\alpha\eta-1\right)^2\right)\\ &\geq \eta w\frac{1}{4}.\end{aligned}$$

if $1-\alpha\eta \geq 1/2$. Furthermore, we have that $w\alpha\eta^2 = \frac{\alpha\eta}{12\sigma^2} \leq \frac{1}{4\sigma^2}$ if $\eta \leq \frac{3}{\alpha}$, as required for Corr 4.3.3. Hence, if $\eta \leq \frac{1}{2\alpha}$ then both requirements are satisfied. Consider the LHS of 8.8, we can bound

$$\left(\eta w\left(1-\alpha\eta\right) - 3\sigma^2 w^2\eta^2\left(\alpha\eta-1\right)^2\right)\sum_{i=1}^{T}\left\|\nabla_t\right\|^2 \geq \eta w\frac{1}{4}\sum_{i=1}^{T}\left\|\nabla_t\right\|^2$$

Finally, we have

$$\begin{aligned}\frac{\eta w}{4}\sum_{i=1}^{T}\left\|\nabla_t\right\|^2 &\leq w\Delta_1 + 3\sigma^2 w\eta^2\alpha T + \log\left(1/\delta\right)\\ \sum_{i=1}^{T}\left\|\nabla_t\right\|^2 &\leq \frac{4}{\eta}\Delta_1 + 3\sigma^2\eta\alpha T + 48\sigma^2\log\left(1/\delta\right).\end{aligned}$$

Setting $\eta = \min\left\{\frac{1}{2\alpha}; \sqrt{\frac{\Delta_1}{\sigma^2\alpha T}}\right\}$, we have that with probability at least $1-\delta$

$$\begin{aligned}\sum_{i=1}^{T}\left\|\nabla_t\right\|^2 &\leq \frac{4}{\eta}\Delta_1 + 3\sigma^2\eta\alpha T + 48\sigma^2\log\left(1/\delta\right)\\ &= \frac{4}{\min\left\{\frac{1}{2\alpha}; \sqrt{\frac{\Delta_1}{\sigma^2\alpha T}}\right\}}\Delta_1 + 3\sigma^2\min\left\{\frac{1}{2\alpha}; \sqrt{\frac{\Delta_1}{\sigma^2\alpha T}}\right\}\alpha T + 48\sigma^2\log\left(1/\delta\right)\\ &\leq 4\left(2\alpha + \sqrt{\frac{\sigma^2\alpha T}{\Delta_1}}\right)\Delta_1 + 3\sigma^2\sqrt{\frac{\Delta_1}{\sigma^2\alpha T}}\alpha T + 48\sigma^2\log\left(1/\delta\right)\\ &= 8\Delta_1\alpha + 4\sigma\sqrt{\alpha T\Delta_1} + 3\sigma\sqrt{\Delta_1\alpha T} + 48\sigma^2\log\left(1/\delta\right)\\ &= 8\Delta_1\alpha + 7\sigma\sqrt{\alpha T\Delta_1} + 48\sigma^2\log\left(1/\delta\right)\\ \implies \frac{1}{T}\sum_{i=1}^{T}\left\|\nabla_t\right\|^2 &\leq \frac{8\Delta_1\alpha}{T} + \frac{7\sigma\sqrt{\alpha\Delta_1}}{\sqrt{T}} + \frac{48\sigma^2\log\left(1/\delta\right)}{T}.\end{aligned}$$

We are done. □

Chapter 9

Subset-Norm and Subspace-Momentum Experiments

9.1 Overview

We perform extensive experiments on LLM pre-training tasks that demonstrate SN's and SM's faster convergence, for both training and validation, than Adam while significantly reducing the optimizer's memory footprint. Our methods, Adam with Subset-Norm step size (AdamSN) and Subspace Momentum (AdamSNSM), achieve Adam's perplexity in *half* the training steps (and tokens) while using 80% less memory for the optimizer state. Furthermore, our algorithms incur minimal additional hyperparameter, exhibit less sensitivity to smaller batch sizes (i.e. gradient noise), and show better learning rate stability across model scales. Notably, we demonstrate that AdaGrad-Subset-Norm and its Subspace-Momentum variant close the performance gap with AdamSN/SNSM or even outperform it, further closing the theory-practice gap. This raises a question on whether the use of exponential moving average as in Adam is necessary or optimal for obtaining strong optimization performance for training DNNs.

9.1.1 Experimental Setup

We evaluate Subset-Norm (SN) and Subspace-Momentum (SM) on LLM pretraining and supervised fine-tuning tasks, where memory is often a bottleneck. We compare against several baselines, with memory estimates given for parameters of size $m \times n$, where we assume WLOG $m \geq n$.

Baselines. We consider **AdaGrad** (Duchi et al., 2011), **AdaGradm** where we incorporate momentum 0.9 to AdaGrad, **Adam** (Kingma and Ba, 2014), and **RMSProp** (Tieleman, Hinton, et al., 2012) as standard optimizers. We also consider **GaLore** (Zhao et al., 2024) as a recent memory-efficient method that projects the optimizer states into a low-rank subspace (typically rank $n/4$), using $2(mn/4)$ memory but requiring 6 hyperparameters including subspace rank, projection update frequency, and scaling parameters.

Our methods. We incorporate SN and SM to AdaGrad, AdaGradm, Adam and RMSProp. **SN** reduces the adaptive step size (e.g. Adam's second moment term) memory from mn to m for a parameter of size $m \times n$. **SNSM** further compresses the momentum term of momentum methods like Adam and AdaGradm by adding SM with SVD at the cost of additional hyperparameters (See Algorithm 14 for the full implementation used in our experiments). **RMSPropSN** and **AdaGradSN** achieves minimal memory footprint of just m while requiring only 2 hyperparameters.

9.2 LLMs Pre-Training Experiments

We test our method on the task of pre-training LLaMA models (Dubey et al., 2024; Touvron et al., 2023) on the C4 dataset (Raffel et al., 2023) with a standard setup – details in Section 9.6.1. Table 9.1 presents the main pre-training results and Table 9.2 shows the memory footprint[1] of different optimizers across a range of model sizes.

TABLE 9.1: Final perplexity ("Perpl.") along with the number of tokens in parentheses of different optimizers on pretraining LLaMA models task. **Bolded methods** are ours. Columns LR and HP denote the learning rate and the number of parameters of the corresponding method, respectively. We only tune for the base learning and set other parameters as in previous implementations. The memory column shows the optimizer's states memory consumption given a parameter of shape $m \times n$ with $m \geq n$. Red LR highlights instability.

Methods	Memory (for $m \times n$)	HP	**60M** (1.38B)		**130M** (2.62B)		**350M** (7.86B)		**1B** (13.1B)	
			Perpl.	LR	Perpl.	LR	Perpl.	LR	Perpl.	LR
Adam	$2mn$	3	30.46	0.005	24.60	0.005	18.67	0.001	16.00	0.0005
AdamSN	$mn + m$	3	29.75	0.05	22.90	0.05	17.49	0.05	14.96	0.05
AdamSNSM	$rn + m$	5	29.74	0.05	**22.43**	0.05	**16.91**	0.05	14.05	0.05
AdaGradm	$2mn$	2	30.40	0.10	24.86	0.10	18.30	0.10	17.42	0.10
AdaGradmSN	$mn + m$	2	**29.73**	2.00	22.58	2.00	17.14	2.00	14.48	2.00
AdaGradSNSM	$rn + m$	4	29.81	1.00	**22.43**	1.00	16.99	1.00	**13.96**	1.00
AdaGrad	mn	1	37.12	0.05	25.76	0.05	18.14	0.05	15.25	0.01
AdaGradSN	m	1	29.85	2.00	24.19	1.00	17.72	1.00	14.82	1.00
RMSProp	mn	2	35.51	0.001	25.94	0.001	20.01	0.001	17.03	0.001
RMSPropSN	m	2	34.57	0.01	25.67	0.01	18.72	0.01	15.97	0.001
GaLore	$2rn$	6	34.73	0.01	25.31	0.01	18.95	0.01	16.76	0.001
Rank r / Dimension m			128/512		256/768		256/1024		512/2048	

9.2.1 Discussions

Subset-Norm (SN) improves upon all existing adaptive methods while reducing memory. Modifying Adam, AdaGradm, AdaGrad, and RMSProp with the SN adaptive step size not only reduces memory footprint but improves their performance across different scales. Notably, AdaGrad and AdaGradm benefit the most from the SN step size, providing empirical support for the theoretical benefits of SN presented in Section 7.2.

Combining Subspace-Momentum (SM) with SN further improves performance while saving additional memory. Perhaps surprisingly, limiting the use of momentum to a subspace *improves* performance in SN-adaptive step sizes rather than degrading it. Our experiments show that SNSM, combining SN and SM, gives the best performance for the least amount of memory across model sizes. While adding SM introduces additional hyperparameters, Section 9.5.4 suggests that these parameters are not too sensitive.

Furthermore, Section 9.4.2 shows that the choice of the subspace matters i.e. the subspace spanned by a top-k singular vectors of a snapshot of a stochastic gradient

[1]The memory footprint is the total parameters in the optimizer states multiplied by 16 bits. See Listing 9.1 for more details.

TABLE 9.2: Optimizer states memory footprint (in GB for BF16 dtype) for different LLaMA models. Our methods, AdamSN, AdamSNSM, and RMSPropSN (RMSPSN), are modifications of Adam and RMSProp (RMSP) to utilize Subset-Norm (SN) and Subspace-Momentum (SM). For GaLore and AdamSNSM, the subspace is of dimension[2] d/r, where the memory accounts for additional space for storing the projection matrices.

Opt.	**AdamW**	**AdamSN**	**RMSP**	**GaLore**	**AdamSNSM**	**RMSPSN**
Mem.	$2d$	$d+\sqrt{d}$	d	$4d/r$	$2^d/r+\sqrt{d}$	$\sqrt{d}$
60M	0.22	0.14	0.11	0.15	0.08	0.03
130M	0.50	0.30	0.25	0.29	0.16	0.05
350M	1.37	0.75	0.69	0.53	0.28	0.06
1B	4.99	2.62	2.49	1.61	0.84	0.12
3B	10.01	5.16	5.00	2.96	1.52	0.15
7B	25.10	13.04	12.55	7.01	2.73	0.49

seems to be the most beneficial for momentum as opposed to simpler choices like a random subspace. Our current guarantee for SM, presented in Section 8.5, does not yet explain why or when subspace momentum is useful, and theoretical understanding of (EMA style) momentum in stochastic optimization is still limited (Kidambi et al., 2018). We believe this could be related to how momentum is beneficial when noise is low (and harmful when noise is high) and the choice of the subspace could correlate to the amount of gradient noise or optimization landscape that harm or benefit momentum (Wang et al., 2024; Gitman et al., 2019).

Hyperparameter robustness. In Table 9.1, the best learning rate (LR) found via grid search is displayed and is highlighted in red as the best LR changes across scales. This indicates potential sensitivity to tuning for each respective algorithm. We see that Adam requires smaller LR for larger models, but using SN and SNSM does not. AdaGradm seems less sensitive to the base LR overall.

Closing the theory-practice gap. While there is a non-trivial performance gap between Adam and AdaGrad(m) for larger models, using the SN step size closes this gap across scales. This shows that AdaGrad style algorithms can be competitive to Adam when using the SN step size. Interestingly, vanilla AdaGrad seems to perform well as model size increases. This is important because AdaGrad enjoys stronger theoretical understanding than Adam and has one fewer parameter – β_2 – to tune.

9.3 LLMs Supervised Fine-Tuning (SFT) Experiments

We further evaluate on a supervised-fine-tuning task, where we fine-tune a pre-trained LLaMA 7B model on the UltraFeedback dataset (Cui et al., 2024) using the chosen responses with max sequence length of 1024. We train for 1 epoch with linear decay and gradient clipping of 1. Table 9.3 contains the result with the time and memory of one training epoch on a single A100-80GB GPU. Note SNSM's r denotes the dimension of SM but the optimization is full-rank.

Discussion. We observe similar improvement over Adam as in pre-training tasks. Surprisingly, the smaller rank (for momentum) is more beneficial than the larger rank. In contrast to LoRA, since we report peak-memory here, due to the full parameter training of SNSM, the primary memory bottlenecks are gradients and activations. Furthermore, we note that the primary contributor to SNSM's slower wall

TABLE 9.3: Last and minimum validation perplexity for SFT of LLaMA 7B on the UltraFeedback dataset between Adam, LoRA, and AdamSNSM for 2 different ranks. We also show the wall-clock time and peak memory for batchsize 1 for these optimizers.

	Adam	**LoRA (r=64)**	**AdamSNSM (r=64)**	**SNSM (r=32)**
Last	2.622	2.632	2.584	**2.580**
Min.	2.401	2.410	2.392	**2.390**
Time (min.)	266	249	303	301
Memory (GB)	77.11	20.75	42.89	42.89

clock time is the SVD computation on large dimension. We try larger projection update gaps in Table 9.6 which reduce this cost while maintaining good performance for our methods. Furthermore, we discuss potential more efficient alternatives in Section 8.4.1 and leave further exploration to future works.

GLUE Fine-tuning. Additional results on fine-tuning on GLUE tasks with BERT models are in Section 9.5.1.

9.4 Ablation Studies

In this section, we present ablation studies on various parameters of SN and SM.

9.4.1 Subset-Norm's Subset Size Ablation

While we use a simple scheme to compress the adaptive step size of linear modules in the previous experiments, Table 7.1 suggests that there is an optimal subset size that depends on the noise. Figure 9.1 shows performance for various subset-size selection. Since the step size scales with the subset size, the optimal base LR should be decreased as we decrease the subset size closer towards Adam. We include additional results for 130M model in Figure 9.3.

While one can use the heuristics discussed on models where linear modules make up the vast majority, for arbitrary models with weights of d elements, we found that a subset size of $\sqrt{d}/2$ is probably a reasonable choice. If more resources are available, the subset size can also be tuned.

9.4.2 Subspace-Momentum Projection Choice Ablations

Projection types. Table 9.4 tests different choices for projection in SM discussed in Section 8.4.1. Note that for memory storage, SVD, Random Projection via dense Gaussian projection (Gaussian), and Approximated-SVD (Appx-SVD) need to store the $r \times n$ projection matrix (unless we recompute at every step). The remaining methods only need to store the indices for sampling and/or the random seed to regenerate any random choices.

Note that the choice of the projection is important as some projections are more computationally and memory expensive than other, although trading other qualities for given the cost. Simple projections like selecting a subset of coordinates for momentum (Subset-Momentum) are not only faster but enables simple distributed training like FSDP unlike more complex subspace selection mechanism that requires additional priors about the parameters (shape, low-rank, etc.) that might not always satisfied.

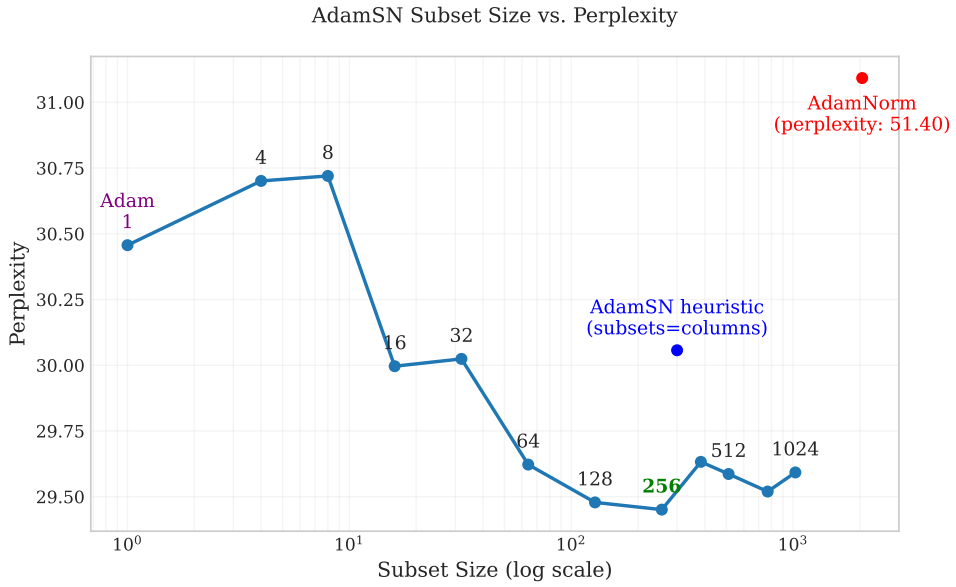

FIGURE 9.1: Subset size ablation for AdamSN on LLaMA 60M trained for 1.38B tokens (batch size of 512 of max length 256 for 10,000 steps). The higher the subset size, the smaller the memory footprint of the second moment optimizer state.

TABLE 9.4: Different projections selection for Subspace-Momentum and validation perplexity. All methods are evaluated on LLaMA 60M with rank 128/512 and a projection update gap of 200. Time and space rows denote time and space to compute and store the projection.

AdamSNSM's projection	SVD	Approx-SVD	Gaussian	SRHT	Top-k	Random-rows	OPCA	Oja
Time (for $m \times n$)	$O(mn^2)$	$O(mn \log k + kn^2)$	$O(kn)$	$O(\max(m,n))$	$O(mn)$	$O(k)$	$O(kn)$	$O(kn)$
Space (for rank k)	$O(kn)$	$O(kn)$	$O(kn)$	$O(k)$	$O(k)$	$O(k)$	$O(kn)$	$O(kn)$
Validation Perplexity	29.74	31.51	42.48	33.33	31.42	33.17	29.63	30.69

Online k-PCA and Streaming k-PCA for Up-to-date Subspace. Computing subspace from stochastic gradient snapshots can be noisy. Recently, (Liang et al., 2024) proposes a formulation of online-PCA to handle the problem of staled top-k components as the stochastic gradients evolve. We test this algorithm in the OPCA column. Another natural algorithm to ensure the top-k components stay up-to-date is Oja's algorithm for streaming k-PCA (Huang et al., 2021). We also test this algorithm in Table 9.4. While we can maintain up to date projection using these schemes, more frequent updates suffer from the same issue of transferring optimization statistics from one subspace to another. We only test for not resetting the statistics in this setting and leave additional investigation for future works. Furthermore, these schemes are more expensive computationally due to additional computation requirement at every step. OPCA further uses Adam for inner optimization which incurs additional memory.

9.4.3 Step Sizes and Momentum Choices Full Ablations

We investigate various combinations of momentum and adaptive step size approaches in Table 9.5. For adaptive methods, we compare EMA, which uses exponential moving average to accumulate the second moment ($v_t^2 = \beta v_{t-1}^2 + (1-\beta) g_t^2$), with AdaGrad's cumulative accumulation approach ($b_t^2 = b_{t-1}^2 + g_t^2$). Methods with the SN suffix utilize subset norm for parameter grouping, contrasting with per-coordinate approaches that are standard. While EMA momentum follows the standard momentum implementation, subspace momentum employs a reduced rank approximation with rank 128 for this model size.

TABLE 9.5: Different combinations of momentum (columns) and adaptive step-size (rows) and the effect of the learning rate schedule on each combination (cosine learning rate decay schedule with warmup "coslr" or constant learning rate "lr."). Memory footprint for each adaptive step size and/or momentum are shown. Green and red highlight runs with perplexity below 30 and above 50 respectively.

Final eval perplexity (lr) LLaMA 60M for 1.31B tokens	**No momentum** Mem = 0	**EMA momentum** Mem = $m \cdot n$	**Subspace momentum** Mem = $\max(m,n) \cdot \text{rank}$
No Adaptive Step-size Mem = 0	SGD 86.60 (coslr=1e-3) 100.04 (lr=1.0)	SGDm 55.76 (coslr=1e-3) 56.07 (lr=1.0)	SGD+SM 89.97 (coslr=1e-3) 213.21 (lr=5e-4)
EMA Coordinate Mem = $m \cdot n$	RMSProp 35.01 (coslr=1e-3) 36.46 (lr=5e-4)	Adam 30.46 (coslr=5e-3) 33.47 (lr=1e-2)	AdamSM 32.34 (coslr=1e-3) 32.25 (lr=5e-4)
EMA Subset-Norm Mem = $\max(m,n)$	RMSPropSN 34.86 (coslr=1e-2) 34.57 (lr=1e-2)	AdamSN 29.75 (coslr=5e-2) 33.69 (lr=1e-2)	AdamSNSM 29.74 (coslr=5e-2) 32.49 (lr=1e-2)
AdaGrad Coordinate Mem = $m \cdot n$	AdaGrad 37.12 (coslr=5e-3) 46.47 (lr=5e-4)	AdaGradm 31.48 (coslr=5e-2) 43.99 (lr=1e-2)	AdaGradSM 30.99 (coslr=5e-2) 41.32 (lr=5e-4)
AdaGrad Subset-Norm Mem = $\max(m,n)$	AdaGradSN 33.19 (coslr=5e-3) 41.23 (lr=0.1)	AdaGradSNm 29.73 (coslr=5e-3) 44.98 (lr=0.1)	AdaGradSNSM 29.81 (coslr=5e-3) 40.11 (lr=0.1)

Discussions. From Table 9.5, Subset norm (SN) step sizes consistently outperform coordinate-wise implementations while requiring less memory. Adaptivity proves crucial for optimization effectiveness, where the first row without adaptivity perform consistently poorly. The addition of momentum is beneficial in all configurations while SM is more beneficial for adaptive step sizes. The impact of learning rate scheduling is also evident across configurations, with cosine decay consistently outperforming constant learning rates. Notably, we observe varying degrees of learning rate sensitivity: adaptive methods demonstrate greater robustness to learning rate selection, while non-adaptive methods require more precise tuning.

9.4.4 Larger Projection Update Gaps

Frequently updating the projection map using SVD can be expensive, especially for larger models. Furthermore, updating the projection every 200 steps can be arbitrary. In Table 9.6, we examine more structured schedules: (1) updating every 5% of the total training steps (corresponding to 200/10K steps for the 60M model) and (2) only using a fixed subspace at the start. Compared to Table, 9.1 where a fixed gap of 200 is used across scales, we see SNSM's performance stay relatively similar

TABLE 9.6: Effects of less frequent subspace update schedule (gap). Compared to Table 9.1 where the gap is fixed to 200 across all scales.

Model Size	**60M**	**130M**	**350M**	**1B**
Gap/Steps (5%)	200/10K	1K/20K	3K/60K	5K/100K
AdamSNSM	29.84	22.71	18.43	15.28
AdaGradSNSM	30.28	22.76	17.02	13.90
GaLore	36.69	29.37	21.27	19.14
Fixed Subspace	10K/10K	20K/20K	60K/60K	100K/100K
AdamSNSM	30.65	23.65	18.94	15.16
AdaGradSNSM	31.43	24.85	18.04	14.62
GaLore	37.95	26.63	21.49	27.11

when we increase the update gap to 5% of the total training steps, whereas GaLore's performance suffers more.

TABLE 9.7: Fixed Subspace Choices on LLaMA 60M. We examine GaLore and SNSM with top-k singular vectors projections (SVD) and random subspaces (Random) using dense gaussian projections.

	GaloreSVD	**GaloreRandom**	**SNSM+SVD**	**SNSM+Random**
Perplexity	37.95	38.23	30.65	40.15

Interestingly, for fixed subspace (100% gap), GaLore still achieves decent performance even though the optimization only happens in a small subspace up until the 1B model, where the training stops improving after 50K/100K steps. In Table 9.7, we see that a random subspace seems to work decently well too. This suggests that a majority of progress can be made in a small subspace in smaller models. In contrast, this is not the same for restricting momentum to a subspace. Furthermore, we notice that there are training loss spikes at the times when we switch subspace for GaLore that impacts training with 5% gap, most likely due to incompatible optimizers' statistics between subspaces. This could explain why GaLore's 100% gap performs similarly or even better than 5% gap for certain run. Finally, we note that AdaGrad-SNSM performs the best here with the larger gaps as the dimension increases.

9.5 Additional Experiments and Ablation Studies

9.5.1 Fine-tuning on GLUE Tasks

Table 9.8 presents results for fine-tuning on GLUE dataset for various methods. The SN step size maintains good performance while reducing the memory footprint.

9.5.2 AdaGrad, AdaGrad-Norm, and AdaGrad-Subset-Norm

We examine the subset-norm step size for AdaGrad in Figure 9.2. We again see that subset-norm is slightly better than the full coordinate version while using a lot less memory. This is consistent with our observations for Adam and RMSProp when we replace the standard coordinate-wise step size with the subset-norm adaptive step size.

TABLE 9.8: Performance metrics across GLUE tasks. QQP, RTE, SST-2, MRPC, STSB, QNLI, and MNLI use accuracy as the metric, while CoLA uses the Matthews correlation coefficient. The **best** and runner-up results for each task and the average score are highlighted.

Method	QQP	RTE	SST2	MRPC	STSB	QNLI	MNLI	COLA	Avg
Adam	**92.0**	77.9	94.9	89.2	90.5	93.0	**87.6**	**65.4**	86.3
GaLore ($r = 4$)	90.9	79.4	**95.2**	88.7	**90.8**	92.4	86.9	61.9	85.8
RMSProp	91.9	79.4	**95.2**	**91.4**	90.3	92.8	**87.6**	65.1	**86.7**
RMSPropSN	91.9	**80.1**	95.1	90.0	90.7	**93.1**	87.5	63.8	86.5
AdamSN	91.2	74.4	94.5	89.5	90.4	92.0	86.7	64.4	85.4

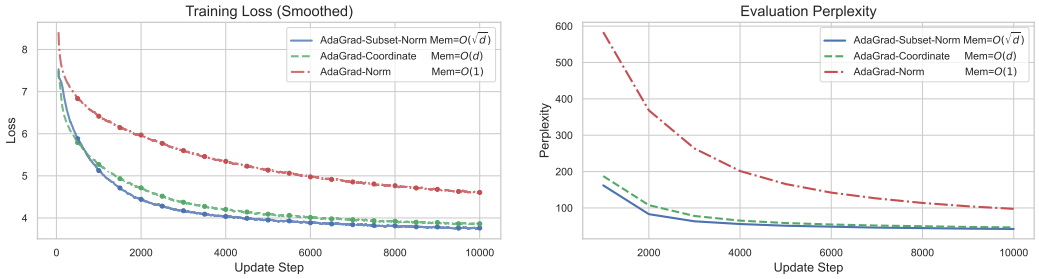

FIGURE 9.2: Pretraining LLaMA 60M on the C4 dataset for AdaGrad variants. Memory consumption estimate as a function of parameter count d is shown in the legend.

9.5.3 Additional Subset-Size Experiments for 130M model

We provide additional subset-size experiments similar to the ones in Section 9.4.1 for LLaMA 130M in Figure 9.3.

9.5.4 Subspace-Momentum Rank and Gap Ablations

Rank and gap ablations. We examine the impact of varying rank and update gap of subspace momentum, similarly to (Zhao et al., 2024), in Figure 9.4. There, we see that the higher the rank, the better the results. For the update gap, it seems like there is an optimal choice. However, due to the SVD computation, a larger gap will be cheaper than a more frequent gap.

9.5.5 Gradient Clipping

Gradient clipping is standard in training LLMs for many open source models like LLaMA, DeepSeek, OPT, etc. (DeepSeek-AI et al., 2024; Touvron et al., 2023; Workshop et al., 2022; Zhang et al., 2022; Chowdhery et al., 2023; Ding et al., 2023). Clipping has a strong connection to stochastic gradient noise being *heavy-tailed* (Zhang et al., 2019) and many theoretical results have been shown to suggest some form of clipping is beneficial when the noise could follow a heavy-tail distribution(Cutkosky and Mehta, 2021; Gorbunov et al., 2020; Li and Liu, 2022; Nguyen et al., 2023b; Nguyen et al., 2023a). We present the results with clipping equal to 1.0 for each method in Table 9.9.

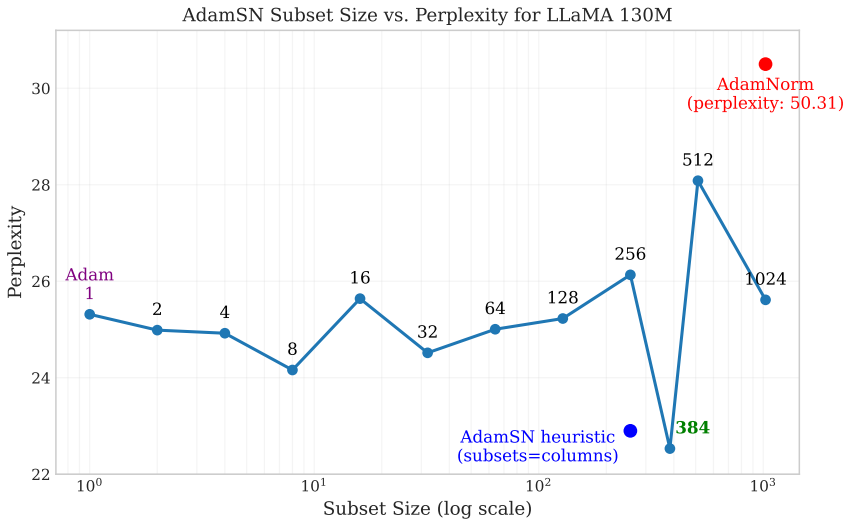

FIGURE 9.3: Subset size ablation for AdamSN on LLaMA 130M trained for 2.62B tokens (batch size of 512 of max length 256 for 20,000 steps). The higher the subset size, the smaller the memory footprint of the second moment optimizer state.

Method	60M (no clipping)	60M (with clipping)	130M (no clipping)	130M (with clipping)
Adam	30.58	30.46	25.07	25.07
AdamSN	**30.06**	**29.75**	23.54	22.89
GaLore	34.91	34.73	25.43	25.31

TABLE 9.9: Pre-training LLMs ablation experiments for gradient clipping. We compare validation perplexity between LLaMA 60M and 130M with and without clipping. We use the same hyperparameters as in Section 9.6.2 but just add clipping.

In Table 9.9, we see that gradient clipping indeed helps most of the methods achieve slightly better perplexity. In our experiments, we notice that adding some form of gradient clipping produces more stable training.

9.5.6 Batch Sizes and Random Seeds

Fixed number of steps. We measure the impact of different batch sizes on pre-training LLaMA 60M for 10,000 steps in Table 9.10.[2] We use the same configuration as in other experiments. Typically, smaller batch sizes require smaller learning rates, but curiously, AdamSNSM seems to be stable with the choice of learning rates. Even more interestingly, AdamSNSM's final performance seems to be affected less by the smaller batch size as opposed to other methods, especially GaLore.

Fixed data quantity. In the previous section, we compare the performances on different batch sizes fixing the same number of steps. In this section, we fix the amount of data to 1.3B tokens for pre-training LLaMA 60M. Hence, adjusting the batch size

[2]This reduces the amount of total tokens trained. However, we only compare optimizers against one another. To compare the same optimizer against different batch sizes, one should train for the same amount of tokens.

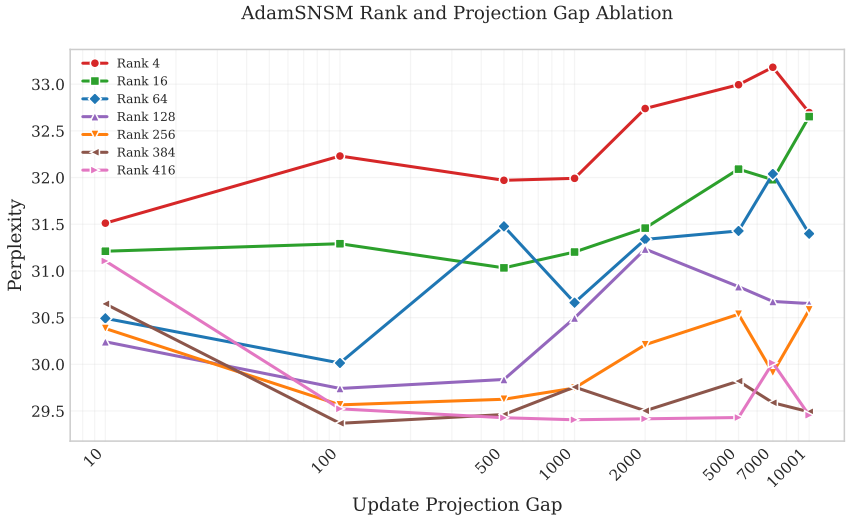

FIGURE 9.4: Rank and gap ablation for AdamSNSM on LLaMA 60M for 10,000 steps. The lower the rank, the less memory consumption used by the momentum state. The higher the projection gap, the less SVD computation is performed which is cheaper.

TABLE 9.10: Batch size ablation for various optimizers along with optimal learning rate.

Batch size	**Adam**		**GaLore**		**AdamSN**		**AdamSNSM**	
	Perpl.	LR	Perpl.	LR	Perpl.	LR	Perpl.	LR
1024	27.94	0.005	32.75	0.01	**27.68**	0.05	28.02	0.05
512	30.46	0.005	34.73	0.01	29.75	0.05	**29.74**	0.05
256	36.65	0.001	44.71	0.001	37.03	0.001	**32.82**	0.05
128	41.72	0.001	49.75	0.001	42.04	0.001	**36.82**	0.05

would also adjust the number of steps. Table 9.11 contains the result where SNSM shows consistently better performance than Adam across different batch sizes.

Random seeds. Throughout our experiments, we fix the random seed for all runs within a same table. In Table 9.11, we investigate the effects of random seeds by running each batch size on 3 random seeds and report the mean and standard deviation. We see that SNSM has better variance than Adam for many batch sizes overall. We also examine the random variation on the 130M model in Table 9.12.

TABLE 9.11: Mean and standard deviation (in parentheses) evaluation perplexities of Adam and AdamSNSM optimizers when pretraining LLaMA 60M for 1.3B tokens over 3 random seeds. SNSM rank = 128 and gap = 200. Learning rates were tuned over a grid for each batch size.

Batch size	1024	512	256	128	64	32	16	8	4
Adam	31.80 (1.87)	30.46 (0.29)	32.11 (1.32)	34.57 (0.16)	36.34 (0.16)	38.91 (0.12)	43.12 (0.26)	48.88 (0.17)	57.28 (0.80)
AdamwSN	30.11 (0.15)	29.81 (0.12)	30.32 (0.07)	31.30 (0.02)	32.72 (0.11)	35.38 (0.11)	40.46 (0.97)	45.81 (0.11)	51.01 (0.25)
AdamSNSM	31.39 (0.17)	29.93 (0.07)	30.08 (0.19)	30.57 (0.08)	32.35 (0.14)	34.51 (0.14)	37.05 (0.20)	39.39 (0.02)	44.27 (0.10)

	Adam	AdamSN	Adagrad	AdaGradSN
Mean	24.69	22.98	25.95	24.57
Stdev	0.07	0.07	0.16	0.37

TABLE 9.12: Mean and standard deviation across 3 runs for different optimizers on pretraining LLaMA 130M task.

9.6 Experimental and Implementation Details

In this section, we provide hyperparameters details, implementation details (pseudocode), and other practical considerations.

9.6.1 LLM Pre-training Experiment Setup

All of our pre-training experiments are conducted on NVIDIA RTX4090/3090 GPUs. We follow the experimental setup as in GaLore (Zhao et al., 2024), where we use a batch size of 512 and max sequence length of 256 for all models. We employ a standard training setup as in the LLaMA paper (Dubey et al., 2024; Touvron et al., 2023) with cosine decay and linear warmup as well as gradient clipping for all methods.[3] For all our experiments, we use the default $(\beta_1, \beta_2) = (0.9, 0.999)$ and only tune for the base learning rate within a grid[4] of

$$\{0.5, 0.1, 0.05, 0.01, 0.005, 0.001\}.$$

We train for 1.38B, 2.62B, 7.86B, and 13.1B tokens for models of sizes 60M, 130M, 350M, and 1B parameters, respectively, following (Zhao et al., 2024) and matches roughly the scaling laws in (Hoffmann et al., 2022). Additional details are in Section 9.6.2.

For GaLore, we use the same hyperparameters as in (Zhao et al., 2024), where we use rank 128/512, 256/768, 256/1024, and 512/2048 for the 60M, 130M, 350M, and 1B models, respectively (Table 2 of (Zhao et al., 2024)).[5] For AdamSNSM, we use the same ranks and projection update gap (of 200) as GaLore for all models.[6] However, we do not tune for an additional scaling parameter unlike GaLore, and we compresses the LM head (final linear layer) with SN and SM also.[7]

9.6.2 Hyperparameter Details

In Table 9.1, we run all experiments on BF16 format, weight decay of 0, gradient clipping of 1.0, cosine learning rate decay to 10% of the max learning rate with 10% linear warmup steps, and batch size of 512 (similarly to (Zhao et al., 2024) and (Touvron et al., 2023; Dubey et al., 2024)). We only tune for the learning rate across a grid

[3] Note that these addition improve the performance for all baselines. See Section 9.5.5.

[4] Except AdaGradSNm where we find higher learning rates in $\{0.5, 1, 2, 5\}$ to be better. We tune the lr on the 60M model and use the same learning rate for the larger model, where the base learning rate is only reduced if the method fails to converge.

[5] Note that our reproduced results for GaLore and baselines are similar to (Zhao et al., 2024).

[6] Note that a smaller gap is more expensive than a larger gap. Our experiments below show that we can increase the projection update gap without much performance loss. If data is not limited, one could use a larger gap to speed up training. However, if data is limited, then a smaller gap to converge in fewer tokens is potentially more desirable.

[7] Existing methods typically do *not* compress the embedding layer and final LM head, while our methods seem robust to this choice. Compressing these layers save additional memory.

of $\{0.1, 0.05, 0.01, 0.005, 0.001\}$ (except for AdaGrad with momentum where larger learning rates are better). We train for 10,000 steps and 20,000 steps for the 60M and 130M models, respectively.

9.6.3 Adam-Subset-Norm Implementation

Algorithm 12 presents the pseudocode for Adam-Subset-Norm as mentioned in Section 7.5 where we partition the coordinates (for each parameter) into subsets of equal sizes.

Algorithm 12 Adam-Subset-Norm with a simple partitioning scheme

Require: Learning rate η, EMA parameters β_1 and β_2, $\epsilon > 0$, optional weight decay $wd \geq 0$

for each $p \in \mathbb{R}^{m \times n}$ in params **do**
  grad $\leftarrow p$.grad
  $r \leftarrow 0$ if $m \geq n$ else 1
  $k \leftarrow p$.shape$[r]$ ▷ where $k = m$ if $r = 0$ else $k = n$
  gradN $\leftarrow$ grad.norm(dim=$1 - r$) $\in \mathbb{R}^k$ ▷ subset norm
  $m \leftarrow \beta_1 m + (1 - \beta_1) \cdot \text{grad} \in \mathbb{R}^{m \times n}$
  $v \leftarrow \beta_2 v + (1 - \beta_2) \cdot \text{gradN}^2 \in \mathbb{R}^k$ ▷ omitting bias correction terms
  $p \leftarrow p + \eta \frac{m}{\sqrt{v} + \epsilon}$ ▷ broadcast division
  $p \leftarrow p - \eta \cdot wd$ ▷ weight decay
end for

9.6.4 Generic Subset-Norm Adaptive Step Size Implementation

The heuristic implementation in Section 9.6.3 is simple and does not require any tuning. However, to modify existing algorithms to work with arbitrary subsets, one could utilize reshape as in Algorithm 13 as an example.

Algorithm 13 Generic Subset-Norm Adaptive Step Size Update Rule (PyTorch-y notation)

Require: Parameter $P \in \mathbb{R}^d$, step size $\eta > 0$, β, $\epsilon > 0$, and partition size k such that k divides d.

$R \leftarrow (\nabla P)$.reshape$(d/k, k)$ ▷ Reshape gradient into shape $\frac{d}{k} \times k$
$V \leftarrow \beta V + (1 - \beta) \cdot ((R\text{**}2).\text{sum(dim=1)}) \in \mathbb{R}^{d/k}$ ▷ Update state V via subset norm reduction on dim 1
$U \leftarrow \frac{R}{\sqrt{V} + \epsilon} \in \mathbb{R}^{\frac{d}{k} \times k}$ ▷ Broadcast addition and division for update step
$P \leftarrow P - \eta \cdot U$.view$(d)$ ▷ Reshape U back to $\mathbb{R}^d$ and update P

9.6.5 AdamSNSM Implementation Details

Algorithm 14 provides the pseudocode and implementation details for the version of AdamSNSM with SVD subspace momentum and heuristics subset-norm (as described in Section 7.5) used in our experiments.

Algorithm 14 AdamSNSM with Subspace Momentum via top-k singular vectors from SVD used in our experiments. Note that we only apply the compression to linear modules while performing vanilla Adam on the rest of the modules (all 1D params).

Require: Learning rate η, rank k, update gap G, momentum parameters $\beta_1, \beta_2 \in (0, 1)$, and stability parameter ϵ

for $t = 1, \ldots, T$ **do**
  Obtain stochastic gradient $g_t \in \mathbb{R}^{m \times n}$ ▷ WLOG, we assume $m \geq n$
  if $t \mod G = 0$ **then**
    $U, S, V = \text{SVD}(g_t)$ ▷ Compute singular value decomposition
    $P = U[:, :k] \in \mathbb{R}^{m \times k}$ ▷ Extract top k singular vectors
  end if
  $m = \beta_1 m + (1 - \beta_1) P^T g_t \in \mathbb{R}^{k \times n}$ ▷ Update subspace momentum
  $r = g_t - P P^T g_t$ ▷ Compute orthogonal SGD component
  $s = \text{sum}(g_t, \text{dim} = 1) \in \mathbb{R}^n$ ▷ Sum all columns for subset-norm heuristic
  $v = \beta_2 v + (1 - \beta_2) s^2 \in \mathbb{R}^n$ ▷ EMA of subset-norm
  $x_t = x_{t-1} + \eta \frac{Pm + r}{\sqrt{v} + \epsilon}$ ▷ Update with subspace momentum and subset-norm step size
end for

9.6.6 Measuring Memory Footprint of Optimizers

In PyTorch, we can obtain the number of parameters in optimizer states using the code in Listing 9.1.

9.6.7 Peak memory measurement during training for different optimizers

We measure peak memory consumption directly via running nvidia-smi in Figure 9.5 while training as oppose to controlled measurement as in Table 9.2. Note that these peak measurements incur additional memory from gradient computation and algorithms' overhead.

```python
def get_optimizer_state_size(optimizer) -> Tuple[int, Dict[str, int]]:
    total_state_size = 0
    state_size_breakdown = {}
    for group in optimizer.param_groups:
        for p in group['params']:
            state = optimizer.state[p]
            for state_key, state_value in state.items():
                if torch.is_tensor(state_value):
                    if state_value.numel() == 1:
                        # we do not count singleton
                        continue
                    total_state_size += state_value.numel()
                    if state_key not in state_size_breakdown:
                        state_size_breakdown[state_key] = 0
                    state_size_breakdown[state_key] += state_value.
                        numel()
    return total_state_size, state_size_breakdown
```

LISTING 9.1: PyTorch function to calculate optimizer state size

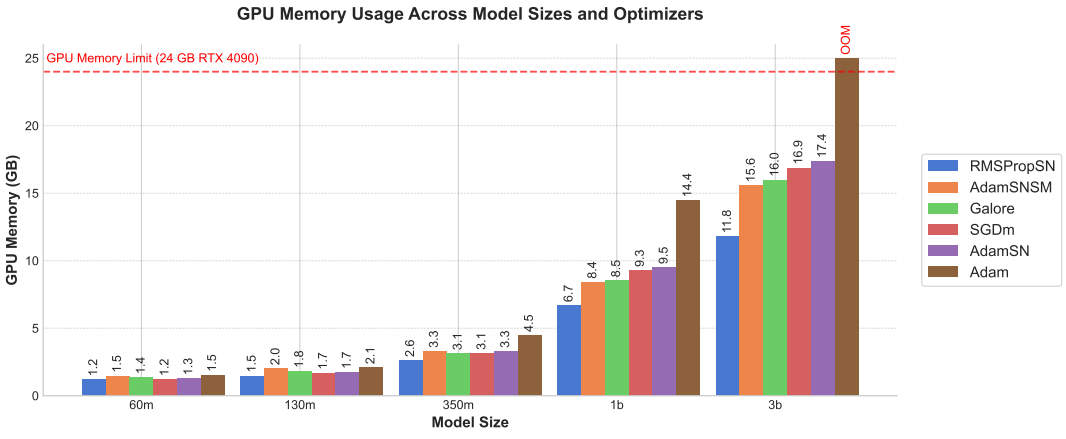

FIGURE 9.5: Peak GPU Memory Usage (Gb) for various model sizes, obtained with batch size 1 and activation checkpointing to measure the optimizer state footprint.

Part III

Conclusion and Future Directions

Chapter 10

Conclusion and Future Directions

All things shall pass.

– Traditional Proverb

10.1 Conclusion

This thesis has advanced the theory and practice of stochastic optimization for large-scale deep learning. We established robust high-probability convergence guarantees under relaxed assumptions, strengthening the theoretical underpinnings of widely used optimizers like SGD, SMD, and AdaGrad variants. These insights deepen our understanding of optimization dynamics and pave the way for more reliable, efficient training algorithms. We then introduced Subset Norm (SN) and Subspace Momentum (SM), two novel techniques that boost the memory and sample efficiency of adaptive optimizers. By compressing optimizer states without sacrificing convergence, these methods cut the training cost of LLaMA-1B in half (in tokens) while reducing optimizer memory use by over 80%. Such gains promise to make large-scale training more accessible and cost-effective for academia and industry alike. Though challenges persist in optimizing ever-complex architectures and dynamics, our work offers a step toward more efficient, scalable algorithms. As deep learning scales up in model size and data demands, optimization efficiency will remain key to progress. We hope the theoretical advances and practical innovations here inspire further research and real-world impact in large-scale AI systems.

10.2 Future Directions

10.2.1 Theoretical Directions

Convergence of adaptive methods under heavy-tailed noise. While optimal high-probability convergence rates have been established for clipped-SMD and clipped-SGD under heavy-tailed noise, understanding the conditions for the convergence of adaptive methods like AdaGrad and Adam under heavy-tailed noise is crucial for bridging the theory-practice gap. Theoretical insights into practical algorithms under heavy-tailed noise can inform stability, generalization, and the design of more effective and robust algorithms for DNNs under more pragmatic assumptions. Recent progress in proving the convergence of normalized-SGD without gradient clipping under heavy-tailed noise (Hübler et al., 2024) shows promising potential for establishing convergence guarantees for adaptive methods in this relevant setting.

Theoretical benefits of momentum in stochastic non-convex optimization. While momentum accelerates optimization in deterministic smooth convex settings, its

role in stochastic non-convex optimization – such as the training of DNNs – remains not entirely understood. In contrast, the theoretical understanding of adaptive step sizes in such settings is more advanced. Although many works on variance-reduction algorithms show theoretical acceleration for non-convex stochastic optimization through various momentum schemes (albeit under additional assumptions), these methods are often complex, computationally expensive, and/or less effective in real-world tasks. A deeper understanding of existing practical momentum and acceleration schemes could provide valuable insights into the convergence of these algorithms in the heavy-tailed noise setting (Hübler et al., 2024) or in the development of effective learning rate schedules under unknown time-horizon (Defazio et al., 2024).

10.2.2 Experimental Directions

Subspace optimization for general architectures and domains. Current subspace optimization implementations, such as GaLore and our Subspace-Momentum, are implicitly limited to linear modules and transformer architectures, i.e., rank-2 tensors, due to their reliance on the singular value decomposition (SVD) for finding useful subspaces. Effectively extending these methods to higher-order tensor modules – such as convolutional weights in vision architectures or higher-order tensors in neural operators – is highly desirable due to their broad applicability in a wide range of important domains. Successful generalization efforts would necessitate empirical and theoretical investigation into tensor decomposition schemes and the subspaces' effects on the optimization's preconditioning and interactions with gradient noise.

Deeper understanding of subspace optimization. Ablation studies indicate that the choice of subspace – e.g., top-k singular vectors versus random subspaces – has important effects on the optimization performance. Consequently, further exploration of effective subspaces could enhance practical aspects while providing insights into understanding subspace optimization dynamics in non-convex stochastic optimization problems. Subspace-Momentum, in particular, offers an analytical framework for restricting momentum to a controlled subspace while preserving standard gradient descent in the complement, enabling more fine-grained analysis. Recent progress in full-preconditioner algorithms like the *Shampoo* optimizer could provide useful hints in this investigation. Finally, an equally interesting question is whether adversarial subspaces can be identified where the use of momentum *negatively impacts* optimization. Addressing these questions could lead to the principled design of more robust and effective algorithms for modern AI systems.

Noise robustness and applications to various domains. The potential of Subset-Norm to stabilize training by effectively managing gradient noise warrants further exploration and broader application across various domains. For example, Reinforcement Learning offers promising opportunities for Subset-Norm's adaptive step size, particularly when gradient noise is exacerbated by sampling-based reward estimations. A closely related area is generative modeling through denoising diffusion probabilistic models (DDPMs), where substantial efforts have been dedicated to mitigating gradient noise to achieve more stable optimization. Similarly, although generative adversarial methods have been established for some time, their notorious instability during training presents an opportunity to investigate noise-robust algorithms. Finally, noise plays a critical role in several important applications like

robustness and privacy. Effectively managing noise during training could unlock new possibilities in these fields and beyond.

10.3 Final Remark

Foundation models (e.g., LLMs) are to the AI revolution what the steam engine was to the Industrial Revolution – except in the cognitive space rather than the physical one. Just as cars revolutionized travel by making it faster and more efficient, foundation models serve as cognitive engines that accelerate how we traverse across the information landscape. As we refine these engines, we won't just get to our destinations faster—we'll unlock entirely new frontiers, much like how airplanes enabled global travel and rockets took us beyond Earth.

However, a car is more than just an engine. To truly understand this new technology that is AI, we must build the full system – transmissions, wheels, suspension, and safety systems – to ensure reliability, efficiency, and control. The road ahead requires engineering not just more powerful models (engines) but also robust infrastructure, ethical frameworks, and thoughtful applications.

Hence today, AGI can then be thought of as teleportation. It's certainly a dream, but we should probably focus on building better cars, faster planes, and more reliable systems first. And most importantly, we need skilled drivers to navigate wisely, taking us to meaningful and beautiful destinations. Then we can be explorers and perhaps take a leisure road-trip across the beautiful landscape of the mind.

Bibliography

Ailon, Nir and Bernard Chazelle (2009). "The fast Johnson–Lindenstrauss transform and approximate nearest neighbors". In: *SIAM Journal on computing* 39.1, pp. 302–322.

Anil, Rohan, Vineet Gupta, Tomer Koren, and Yoram Singer (2019). "Memory efficient adaptive optimization". In: *Advances in Neural Information Processing Systems* 32.

Arjevani, Yossi, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth (2023). "Lower bounds for non-convex stochastic optimization". In: *Mathematical Programming* 199.1, pp. 165–214.

Attia, Amit and Tomer Koren (2023). "SGD with AdaGrad stepsizes: Full adaptivity with high probability to unknown parameters, unbounded gradients and affine variance". In: *International Conference on Machine Learning*. PMLR, pp. 1147–1171.

Beygelzimer, Alina, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire (2011). "Contextual bandit algorithms with supervised learning guarantees". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, pp. 19–26.

Chen, Tianqi, Bing Xu, Chiyuan Zhang, and Carlos Guestrin (2016). "Training deep nets with sublinear memory cost". In: *arXiv preprint arXiv:1604.06174*.

Chen, Xiangyi, Sijia Liu, Ruoyu Sun, and Mingyi Hong (2018). "On the Convergence of A Class of Adam-Type Algorithms for Non-Convex Optimization". In: *International Conference on Learning Representations*.

Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. (2023). "Palm: Scaling language modeling with pathways". In: *Journal of Machine Learning Research* 24.240, pp. 1–113.

Chung, Fan and Linyuan Lu (2006). "Concentration inequalities and martingale inequalities: a survey". In: *Internet mathematics* 3.1, pp. 79–127.

Cui, Ganqu et al. (2024). *UltraFeedback: Boosting Language Models with Scaled AI Feedback*. arXiv: 2310.01377 [cs.CL]. URL: https://arxiv.org/abs/2310.01377.

Cutkosky, Ashok and Harsh Mehta (2021). "High-probability bounds for non-convex stochastic optimization with heavy tails". In: *Advances in Neural Information Processing Systems* 34, pp. 4883–4895.

Dao, Tri (2023). "Flashattention-2: Faster attention with better parallelism and work partitioning". In: *arXiv preprint arXiv:2307.08691*.

Dao, Tri, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré (2022). "Flashattention: Fast and memory-efficient exact attention with io-awareness". In: *Advances in Neural Information Processing Systems* 35, pp. 16344–16359.

Davis, Damek, Dmitriy Drusvyatskiy, Lin Xiao, and Junyu Zhang (2021). "From low probability to high confidence in stochastic convex optimization". In: *Journal of machine learning research* 22.49.

DeepSeek-AI et al. (2024). *DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model*. arXiv: 2405.04434 [cs.CL]. URL: https://arxiv.org/abs/2405.04434.

Defazio, Aaron, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, Ashok Cutkosky, et al. (2024). "The Road Less Scheduled". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=0XeNkkENuI.

Défossez, Alexandre, Léon Bottou, Francis Bach, and Nicolas Usunier (2022). "A simple convergence proof of adam and adagrad". In: *Transactions on Machine Learning Research*.

Dettmers, Tim, Mike Lewis, Younes Belkada, and Luke Zettlemoyer (2022). "GPT3.int8(): 8-bit Matrix Multiplication for Transformers at Scale". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 30318–30332. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/c3ba4962c05c49636d4c6206a97e9c8a-Paper-Conference.pdf.

Dettmers, Tim, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer (2021). "8-bit optimizers via block-wise quantization". In: *arXiv preprint arXiv:2110.02861*.

Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer (2024). "Qlora: Efficient finetuning of quantized llms". In: *Advances in Neural Information Processing Systems* 36.

Ding, Jiayu, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei (2023). "Longnet: Scaling transformers to 1,000,000,000 tokens". In: *arXiv preprint arXiv:2307.02486*.

Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. (2024). "The llama 3 herd of models". In: *arXiv preprint arXiv:2407.21783*.

Duchi, John, Elad Hazan, and Yoram Singer (2011). "Adaptive subgradient methods for online learning and stochastic optimization." In: *Journal of machine learning research* 12.7.

Dvurechensky, Pavel and Alexander Gasnikov (2016). "Stochastic intermediate gradient method for convex problems with stochastic inexact oracle". In: *Journal of Optimization Theory and Applications* 171.1, pp. 121–145.

Dzhaparidze, Kacha and JH Van Zanten (2001). "On Bernstein-type inequalities for martingales". In: *Stochastic processes and their applications* 93.1, pp. 109–117.

Ene, Alina and Huy L Nguyen (2021). "Adaptive and Universal Algorithms for Variational Inequalities with Optimal Convergence s". In: *arXiv preprint arXiv:2010.07799*.

Ene, Alina, Huy L Nguyen, and Adrian Vladu (2021). "Adaptive Gradient Methods for Constrained Convex Optimization and Variational Inequalities". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 8, pp. 7314–7321.

Faw, Matthew, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward (2022). "The Power of Adaptivity in SGD: Self-Tuning Step Sizes with Unbounded Gradients and Affine Variance". In: *arXiv preprint arXiv:2202.05791*.

Frantar, Elias, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh (2023). *GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers*. arXiv: 2210.17323 [cs.LG]. URL: https://arxiv.org/abs/2210.17323.

Freedman, David A (1975). "On tail probabilities for martingales". In: *the Annals of Probability*, pp. 100–118.

Ghadimi, Saeed and Guanghui Lan (2013). "Stochastic first-and zeroth-order methods for nonconvex stochastic programming". In: *SIAM Journal on Optimization* 23.4, pp. 2341–2368.

Gitman, Igor, Hunter Lang, Pengchuan Zhang, and Lin Xiao (2019). "Understanding the role of momentum in stochastic gradient methods". In: *Advances in Neural Information Processing Systems* 32.

Gorbunov, Eduard, Marina Danilova, and Alexander Gasnikov (2020). "Stochastic optimization with heavy-tailed noise via accelerated gradient clipping". In: *Advances in Neural Information Processing Systems* 33, pp. 15042–15053.

Gorbunov, Eduard, Marina Danilova, Innokentiy Shibaev, Pavel Dvurechensky, and Alexander Gasnikov (2021). "Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise". In: *arXiv preprint arXiv:2106.05958*.

Gupta, Vineet, Tomer Koren, and Yoram Singer (2018). "Shampoo: Preconditioned stochastic tensor optimization". In: *International Conference on Machine Learning*. PMLR, pp. 1842–1850.

Gurbuzbalaban, Mert, Umut Simsekli, and Lingjiong Zhu (2021). "The heavy-tail phenomenon in SGD". In: *International Conference on Machine Learning*. PMLR, pp. 3964–3975.

Halko, Nathan, Per-Gunnar Martinsson, and Joel A Tropp (2011). "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions". In: *SIAM review* 53.2, pp. 217–288.

Hao, Yongchang, Yanshuai Cao, and Lili Mou (2024). "Flora: Low-Rank Adapters Are Secretly Gradient Compressors". In: *arXiv preprint arXiv:2402.03293*.

Harvey, Nicholas JA, Christopher Liaw, Yaniv Plan, and Sikander Randhawa (2019). "Tight analyses for non-smooth stochastic gradient descent". In: *Conference on Learning Theory*. PMLR, pp. 1579–1613.

Hazan, Elad and Satyen Kale (2014). "Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization". In: *The Journal of Machine Learning Research* 15.1, pp. 2489–2512.

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. (2022). "Training compute-optimal large language models". In: *arXiv preprint arXiv:2203.15556*.

Hong, Yusu and Junhong Lin (2024). "Revisiting Convergence of AdaGrad with Relaxed Assumptions". In: *arXiv preprint arXiv:2402.13794*.

Hu, Edward J, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen (2021). "Lora: Low-rank adaptation of large language models". In: *arXiv preprint arXiv:2106.09685*.

Huang, De, Jonathan Niles-Weed, and Rachel Ward (2021). "Streaming k-PCA: Efficient guarantees for Oja's algorithm, beyond rank-one updates". In: *Conference on Learning Theory*. PMLR, pp. 2463–2498.

Hübler, Florian, Ilyas Fatkhullin, and Niao He (2024). "From Gradient Clipping to Normalization for Heavy Tailed SGD". In: *arXiv preprint arXiv:2410.13849*.

Juditsky, Anatoli, Arkadi Nemirovski, and Claire Tauvel (2011). "Solving variational inequalities with stochastic mirror-prox algorithm". In: *Stochastic Systems* 1.1, pp. 17–58.

Kakade, Sham M and Ambuj Tewari (2008). "On the generalization ability of online strongly convex programming algorithms". In: *Advances in Neural Information Processing Systems* 21.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei (2020). "Scaling laws for neural language models". In: *arXiv preprint arXiv:2001.08361*.

Kavis, Ali, Kfir Yehuda Levy, and Volkan Cevher (2021). "High Probability Bounds for a Class of Nonconvex Algorithms with AdaGrad Stepsize". In: *International Conference on Learning Representations*.

Khaled, Ahmed and Peter Richtárik (2020). "Better theory for SGD in the nonconvex world". In: *arXiv preprint arXiv:2002.03329*.

Kidambi, Rahul, Praneeth Netrapalli, Prateek Jain, and Sham M. Kakade (2018). *On the insufficiency of existing momentum schemes for Stochastic Optimization*. arXiv: 1803.05591 [cs.LG]. URL: https://arxiv.org/abs/1803.05591.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Lan, Guanghui (2012). "An optimal method for stochastic composite optimization". In: *Mathematical Programming* 133.1, pp. 365–397.

— (2020). *First-order and stochastic optimization methods for machine learning*. Springer.

Li, Bingrui, Jianfei Chen, and Jun Zhu (2024a). "Memory efficient optimizers with 4-bit states". In: *Advances in Neural Information Processing Systems* 36.

Li, Haochuan, Alexander Rakhlin, and Ali Jadbabaie (2024b). "Convergence of adam under relaxed assumptions". In: *Advances in Neural Information Processing Systems* 36.

Li, Shaojie and Yong Liu (2022). "High Probability Guarantees for Nonconvex Stochastic Gradient Descent with Heavy Tails". In: *International Conference on Machine Learning*. PMLR, pp. 12931–12963.

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

— (2020). "A high probability analysis of adaptive SGD with momentum". In: *arXiv preprint arXiv:2007.14294*.

Lialin, Vladislav, Sherin Muckatira, Namrata Shivagunde, and Anna Rumshisky (2023). "Relora: High-rank training through low-rank updates". In: *The Twelfth International Conference on Learning Representations*.

Liang, Kaizhao, Bo Liu, Lizhang Chen, and Qiang Liu (2024). "Memory-efficient llm training with online subspace descent". In: *arXiv preprint arXiv:2408.12857*.

Lin, Ji, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han (2024). *AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration*. arXiv: 2306.00978 [cs.CL]. URL: https://arxiv.org/abs/2306.00978.

Liu, Hong, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma (2023a). "Sophia: A scalable stochastic second-order optimizer for language model pre-training". In: *arXiv preprint arXiv:2305.14342*.

Liu, Zijian, Ta Duy Nguyen, Alina Ene, and Huy Nguyen (2023b). "On the Convergence of AdaGrad(Norm) on $\mathbb{R}^d$: Beyond Convexity, Non-Asymptotic Rate and Acceleration". In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=ULnHxczCBaE.

Liu, Zijian, Ta Duy Nguyen, Alina Ene, and Huy L Nguyen (2022). "On the Convergence of AdaGrad on $\mathbb{R}^d$: Beyond Convexity, Non-Asymptotic Rate and Acceleration". In: *arXiv preprint arXiv:2209.14827*.

Liu, Zijian, Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Nguyen (2023c). "High probability convergence of stochastic gradient methods". In: *International Conference on Machine Learning*. PMLR, pp. 21884–21914.

Liu, Zijian, Jiawei Zhang, and Zhengyuan Zhou (2023d). "Breaking the Lower Bound with (Little) Structure: Acceleration in Non-Convex Stochastic Optimization with Heavy-Tailed Noise". In: *arXiv preprint arXiv:2302.06763*.

Luo, Qijun, Hengxu Yu, and Xiao Li (2024). *BAdam: A Memory Efficient Full Parameter Optimization Method for Large Language Models*. arXiv: 2404.02827 [cs.LG]. URL: https://arxiv.org/abs/2404.02827.

Madden, Liam, Emiliano Dall'Anese, and Stephen Becker (2020). "High probability convergence and uniform stability bounds for nonconvex stochastic gradient descent". In: *arXiv preprint arXiv:2006.05610*.

McMahan, H. Brendan and Matthew J. Streeter (2010). "Adaptive Bound Optimization for Online Convex Optimization". In: *Conference on Learning Theory (COLT)*. Omnipress, pp. 244–256.

Minsker, Stanislav (2015). "Geometric median and robust estimation in Banach spaces". In: *Bernoulli*, pp. 2308–2335.

Modoranu, Ionut-Vlad, Mher Safaryan, Grigory Malinovsky, Eldar Kurtic, Thomas Robert, Peter Richtarik, and Dan Alistarh (2024). "MicroAdam: Accurate Adaptive Optimization with Low Space Overhead and Provable Convergence". In: *arXiv preprint arXiv:2405.15593*.

Muhamed, Aashiq, Oscar Li, David Woodruff, Mona Diab, and Virginia Smith (2024). "GRASS: Compute Efficient Low-Memory LLM Training with Structured Sparse Gradients". In: *arXiv preprint arXiv:2406.17660*.

Nazin, Alexander V, Arkadi S Nemirovsky, Alexandre B Tsybakov, and Anatoli B Juditsky (2019). "Algorithms of robust stochastic optimization based on mirror descent method". In: *Automation and Remote Control* 80.9, pp. 1607–1627.

Nemirovski, Arkadi, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro (2009). "Robust stochastic approximation approach to stochastic programming". In: *SIAM Journal on optimization* 19.4, pp. 1574–1609.

Nesterov, Yurii (1983). "A method for unconstrained convex minimization problem with the rate of convergence O (1/kˆ 2)". In: *Doklady an ussr*. Vol. 269, pp. 543–547.

Nguyen, Ta Duy, Thien H Nguyen, Alina Ene, and Huy Nguyen (2023a). "Improved convergence in high probability of clipped gradient methods with heavy tailed noise". In: *Advances in Neural Information Processing Systems* 36, pp. 24191–24222.

Nguyen, Ta Duy, Thien Hang Nguyen, Alina Ene, and Huy Le Nguyen (2023b). "High probability convergence of clipped-sgd under heavy-tailed noise". In: *arXiv preprint arXiv:2302.05437*.

Parletta, Daniela A, Andrea Paudice, Massimiliano Pontil, and Saverio Salzo (2022). "High Probability Bounds for Stochastic Subgradient Schemes with Heavy Tailed Noise". In: *arXiv preprint arXiv:2208.08567*.

Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio (2012). "Understanding the exploding gradient problem". In: *CoRR, abs/1211.5063* 2.417, p. 1.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2023). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. arXiv: 1910.10683 [cs.LG]. URL: https://arxiv.org/abs/1910.10683.

Raginsky, Maxim and Alexander Rakhlin (2009). "Information complexity of black-box convex optimization: A new look via feedback information theory". In: *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, pp. 803–510.

Rajbhandari, Samyam, Jeff Rasley, Olatunji Ruwase, and Yuxiong He (2020). "Zero: Memory optimizations toward training trillion parameter models". In: *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, pp. 1–16.

Rakhlin, Alexander, Ohad Shamir, and Karthik Sridharan (2011). "Making gradient descent optimal for strongly convex stochastic optimization". In: *arXiv preprint arXiv:1109.5647*.

Reddi, Sashank J, Satyen Kale, and Sanjiv Kumar (2018). "On the Convergence of Adam and Beyond". In: *International Conference on Learning Representations*.

Sadiev, Abdurakhmon, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik (2023). "High-Probability Bounds for Stochastic Optimization and Variational Inequalities: the Case of Unbounded Variance". In: *arXiv preprint arXiv:2302.00999*.

Sakr, Charbel and Brucek Khailany (2024). "ESPACE: Dimensionality Reduction of Activations for Model Compression". In: *arXiv preprint arXiv:2410.05437*.

Shah, Jay, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao (2024). "Flashattention-3: Fast and accurate attention with asynchrony and low-precision". In: *arXiv preprint arXiv:2407.08608*.

Shazeer, Noam and Mitchell Stern (2018). "Adafactor: Adaptive learning rates with sublinear memory cost". In: *International Conference on Machine Learning*. PMLR, pp. 4596–4604.

Şimşekli, Umut, Mert Gürbüzbalaban, Thanh Huy Nguyen, Gaël Richard, and Levent Sagun (2019). "On the heavy-tailed theory of stochastic gradient descent for deep neural networks". In: *arXiv preprint arXiv:1912.00018*.

Simsekli, Umut, Levent Sagun, and Mert Gurbuzbalaban (2019). "A tail-index analysis of stochastic gradient noise in deep neural networks". In: pp. 5827–5837.

Tieleman, Tijmen, Geoffrey Hinton, et al. (2012). "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude". In: *COURSERA: Neural networks for machine learning* 4.2, pp. 26–31.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. (2023). "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971*.

Vershynin, Roman (2018). *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press.

Vural, Nuri Mert, Lu Yu, Krishna Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu (2022). "Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance". In: *Conference on Learning Theory*. PMLR, pp. 65–102.

Vyas, Nikhil, Depen Morwani, and Sham M. Kakade (2024a). "AdaMeM: Memory Efficient Momentum for Adafactor". In: *2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ICML 2024)*. URL: https://openreview.net/forum?id=fZqMVTz7K5.

Vyas, Nikhil, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade (2024b). "Soap: Improving and stabilizing shampoo using adam". In: *arXiv preprint arXiv:2409.11321*.

Wang, Bohan, Huishuai Zhang, Zhiming Ma, and Wei Chen (2023). "Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions". In: *The Thirty Sixth Annual Conference on Learning Theory*. PMLR, pp. 161–190.

Wang, Hongjian, Mert Gurbuzbalaban, Lingjiong Zhu, Umut Simsekli, and Murat A Erdogdu (2021). "Convergence rates of stochastic gradient descent under infinite noise variance". In: *Advances in Neural Information Processing Systems* 34, pp. 18866–18877.

Wang, Runzhe, Sadhika Malladi, Tianhao Wang, Kaifeng Lyu, and Zhiyuan Li (2024). "The Marginal Value of Momentum for Small Learning Rate SGD". In: *The Twelfth*

International Conference on Learning Representations. URL: https://openreview.net/forum?id=3JjJezzVkT.

Ward, Rachel, Xiaoxia Wu, and Leon Bottou (2019). "AdaGrad stepsizes: Sharp convergence over nonconvex landscapes". In: *International Conference on Machine Learning*. PMLR, pp. 6677–6686.

Workshop, BigScience, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. (2022). "Bloom: A 176b-parameter open-access multilingual language model". In: *arXiv preprint arXiv:2211.05100*.

Wu, Yuhuai, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy (2022). "Memorizing transformers". In: *arXiv preprint arXiv:2203.08913*.

Xiao, Guangxuan, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han (2024). *SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models*. arXiv: 2211.10438 [cs.CL]. URL: https://arxiv.org/abs/2211.10438.

Zhang, Jingzhao, Tianxing He, Suvrit Sra, and Ali Jadbabaie (2019). "Why gradient clipping accelerates training: A theoretical justification for adaptivity". In: *arXiv preprint arXiv:1905.11881*.

Zhang, Jingzhao, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra (2020). "Why are adaptive methods good for attention models?" In: *Advances in Neural Information Processing Systems (NeurIPS)* 33, pp. 15383–15393.

Zhang, Jiujia and Ashok Cutkosky (2022). "Parameter-free Regret in High Probability with Heavy Tails". In: *Advances in Neural Information Processing Systems*.

Zhang, Susan, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. (2022). "Opt: Open pre-trained transformer language models". In: *arXiv preprint arXiv:2205.01068*.

Zhang, Yushun, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Diederik P. Kingma, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun (2024). *Adam-mini: Use Fewer Learning Rates To Gain More*. arXiv: 2406.16793 [cs.LG]. URL: https://arxiv.org/abs/2406.16793.

Zhao, Jiawei, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian (2024). "Galore: Memory-efficient llm training by gradient low-rank projection". In: *arXiv preprint arXiv:2403.03507*.

Zou, Fangyu, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu (2019). "A sufficient condition for convergences of adam and rmsprop". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11127–11135.