Smoothness and Adaptivity in Nonlinear Optimization for Machine Learning Applications

by

Haochuan Li

B.S., Peking University (2019)
M.S., Massachusetts Institute of Technology (2021)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Authored by:	Haochuan Li Department of Electrical Engineering and Computer Science May 17, 2024
Certified by:	Ali Jadbabaie Professor of Civil and Environmental Engineering Thesis Supervisor
Certified by:	Alexander Rakhlin Professor of Brain and Cognitive Sciences Thesis Supervisor
Accepted by:	Leslie A. Kolodziejski Professor of Electrical Engineering and Computer Science Chair, Department Committee on Graduate Students

Smoothness and Adaptivity in Nonlinear Optimization for Machine Learning Applications

by

Haochuan Li

Submitted to the Department of Electrical Engineering and Computer Science
on May 17, 2024 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

ABSTRACT

Nonlinear optimization has become the workhorse of machine learning. However, our theoretical understanding of optimization in machine learning is still limited. For example, classical optimization theory relies on assumptions like bounded Lipschitz smoothness of the loss function which are rarely met in machine learning. Besides, existing theory cannot well explain why adaptive methods outperform gradient descent in certain machine learning tasks like training transformers. In this thesis, to bridge this gap, we propose more general smoothness conditions that are closer to machine learning practice and study the convergence of popular classical and adaptive methods under such conditions. Our convergence results improve over existing ones and also provide new insights into understanding the role of adaptivity in optimization for machine learning applications.

First, inspired by some recent works and insights from deep neural network training, we propose a generalized non-uniform smoothness condition with the Hessian norm bounded by a function of the gradient norm almost everywhere. We develop a simple, yet powerful analysis technique that bounds the gradients along the trajectory, thereby leading to stronger results for both convex and non-convex optimization problems. In particular, we obtain the classical convergence rates for gradient descent (GD), stochastic gradient descent (SGD), and Nesterov's accelerated gradient method (NAG) in the convex or non-convex settings under this general smoothness condition.

In addition, the new analysis technique also allows us to obtain an improved convergence result for the Adaptive Moment Estimation (Adam) method. Despite the popularity and efficiency of Adam in training deep neural networks, its theoretical properties are not yet fully understood, and existing convergence proofs require unrealistically strong assumptions, such as globally bounded gradients, to show the convergence to stationary points. In this thesis, we show that Adam provably converges to stationary points under far more realistic conditions. In particular, we do not require the strong assumptions made in previous works and also consider the generalized smoothness condition.

However, the above results can not explain why adaptive methods like Adam significantly outperform SGD in machine learning applications like training transformers, as the convergence rate we have obtained for Adam is not faster than that of SGD. Previous research has empirically observed that adaptive methods tend to exhibit much smaller directional smoothness along the training trajectory compared to SGD. In this thesis, we formalize this observation into a more rigorous theoretical explanation. Specifically, we propose a directional smoothness condition, under which we prove faster convergence of memoryless Adam and RMSProp in the deterministic setting. Notably, our convergence rate is faster than the typical rate of gradient descent, providing new insights into the benefits of adaptivity in training transformers.

Thesis supervisor: Ali Jadbabaie
Title: Professor of Civil and Environmental Engineering

Thesis supervisor: Alexander Rakhlin
Title: Professor of Brain and Cognitive Sciences

Acknowledgments

As I bring this chapter of my academic journey to a close, I am filled with a sense of accomplishment and gratitude. I am thankful for the support, encouragement, and resources I received from numerous people and organizations, which have enabled me to achieve my goals and pursue my passions.

First of all, I would like to express my sincere gratitude to the following funding sources for their generous support during my PhD studies: ONR N00014-23-1-2299, MIT-IBM Watson AI Lab, and NSF DMS-1953181, which allowed me to focus on my research and pursue my goals without financial burden.

I am deeply grateful to my advisors, Ali and Sasha, for their exceptional guidance, support, and mentorship throughout my academic journey. Their expertise, feedback, and mentorship have been invaluable, and I appreciate the time and effort they invested in helping me grow as a mature researcher. They have always been available to meet and discuss my technical and non-technical challenges, offering valuable insights and solutions. I also greatly appreciate the freedom they have provided me to explore topics that ignite my passion, gain industry experience through internships, and work remotely during the COVID-19 pandemic. Their trust and flexibility have enabled me to thrive and reach my full potential.

I would like to extend my heartfelt gratitude to my thesis committee members, Suvrit and Devavrat. Despite their busy schedules, they have offered constructive feedback and insightful suggestions regarding my thesis and defense. I am particularly grateful to Suvrit for introducing me to the field of optimization through his course, Optimization for Machine Learning, where I gained a systematic understanding of the subject for the first time. Furthermore, I was fortunate to serve as his teaching assistant in the Nonlinear Optimization course, which provided me the opportunity to reinforce my knowledge, improve my presentation skills, and gain valuable teaching experience.

I would also like to thank my undergraduate advisor, Liwei, for teaching me the ropes of research and sparking my interest in the field. I am also grateful for the guidance and mentorship I received from Jason during my internship at USC. The invaluable research experiences and references from both Liwei and Jason played a significant role in my acceptance to MIT, laying a strong foundation for my future academic pursuits.

I would like to express my sincere appreciation to all my collaborators and friends who have supported me throughout my PhD and undergraduate journey. In particular, I would like to thank Yi, Jian, Jingzhao, Molin, Yuyang, Amir, Farzan, Xiyu, Simon, Kaiqing, Kwangjun, Zeyu, Chen-Yu, Ayush, Dylan, Tiancheng, Haidong, Songtao, Subhro, Tianle, Ruiqi, Mengxiao, and Xiaoyu. The contributions of my collaborators, whether through brainstorming sessions, feedback on my work, or simply being a sounding board, have had a profound impact on my research and overall experience. I am grateful for the opportunity to have worked alongside such a talented and supportive group of individuals, and I cherish the memories and experiences we have shared.

I am grateful to the mentors, managers, and friends I had during my internships at Amazon and Two Sigma, who provided valuable industry insights and helped me apply my research skills in real-world settings. In particular, I want to thank Lingxue, Jin, and Keyulu at Two Sigma, and Sujay, Mutasem, Sriram, and Kiran at Amazon. They offered guidance on how to navigate the industry, provided feedback on my work and presentation, and helped me develop a deeper understanding of the practical applications of my research.

On a personal note, I would like to express my deepest and most heartfelt appreciation to my family, who have supported me unconditionally throughout my academic journey. My parents, parents-in-law, brother, and wife have been a constant source of encouragement, providing unwavering love, support, and guidance throughout my life. I would like to dedicate my most sincere and profound gratitude to my wife, Fuyang, who has been my rock, partner, and best friend throughout this journey. Her unwavering belief in me has made me more confident, outgoing, and a better version of myself. I am forever grateful for her presence in my life and for the unconditional love and support she has provided, and I look forward to continuing our journey together, exploring new horizons, and creating a bright future filled with love, laughter, and happiness.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the past few decades, machine learning, particularly deep learning, has revolutionized numerous application fields, including computer vision, natural language processing, and sequential decision making. A crucial step in developing a successful machine learning model is optimization, which is essential for achieving high performance. However, despite its importance, our theoretical understanding of optimization in the context of machine learning remains limited.

Classical optimization problems have been extensively studied, with well-established upper and lower bounds on convergence rates. Nevertheless, these theoretical analyses rely on certain assumptions, such as Lipschitz smoothness, that are rarely satisfied in machine learning applications. Moreover, empirical optimization techniques like batch normalization, adaptive stepsizes, and momentum significantly enhance optimization speed, yet these improvements cannot be fully explained by existing theory.

This thesis makes a step towards bridging the gap between optimization theory and machine learning practice by delving into the smoothness condition and adaptivity in non-linear optimization for machine learning applications. In particular, we study the following *unconstrained* optimization problem

$$\min_{x\in\mathcal{X}} f(x), \tag{1.1}$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is the domain of f.

Classical textbook analyses [Nemirovskij and Yudin, 1983, Nesterov, 2003] of (1.1) often require the Lipschitz smoothness condition, which assumes $\|\nabla^2 f(x)\| \le L$ almost everywhere for some $L \ge 0$ called the smoothness constant. This condition, however, is rather restrictive and only satisfied by functions that are both upper and lower bounded by quadratic functions. Recently, Zhang et al. [2019] proposed the more general (L_0, L_1)-smoothness condition, which assumes $\|\nabla^2 f(x)\| \le L_0 + L_1 \|\nabla f(x)\|$ for some constants $L_0, L_1 \ge 0$, motivated by their extensive language model experiments. This notion generalizes the standard Lipschitz smoothness condition and also contains e.g. univariate polynomial and exponential functions. For *non-convex* and (L_0, L_1)-smooth functions, they prove convergence of gradient descent (GD) and stochastic gradient descent (SGD) *with gradient clipping* and also provide a complexity lower bound for *constant-stepsize* GD/SGD without clipping. Based on these results, they claim gradient clipping or other forms of adaptivity *provably* accelerate the convergence for (L_0, L_1)-smooth functions. Perhaps due to the lower bound, all the follow-up works under this condition that we are aware of limit their analyses to adaptive methods. Most of these focus on non-convex functions.

In this thesis, we significantly generalize the (L_0, L_1)-smoothness condition to the ℓ-smoothness condition which assumes $\|\nabla^2 f(x)\| \le \ell(\|\nabla f(x)\|)$ for some non-decreasing continuous function ℓ. We develop a simple, yet powerful technique, which allows us to obtain stronger results for *both convex and non-convex* optimization problems when ℓ is sub-quadratic (i.e., $\lim_{u\to\infty} \ell(u)/u^2 = 0$) or even more general. The ℓ-smooth function class with a sub-quadratic ℓ also contains e.g. univariate rational and double exponential functions. In particular, we prove the convergence of classical non-adaptive methods including *constant-stepsize* GD/SGD and Nesterov's accelerated gradient method (NAG) in the convex or non-convex settings. For each method and setting, we obtain the classical convergence rate, under a certain requirement of ℓ. In addition, we relax the assumption of bounded noise to the weaker one of bounded variance with the simple SGD method. See Table 1.1 for a summary of our results and assumptions for each method and setting.

Our approach analyzes boundedness of gradients along the optimization trajectory. The idea behind it can be informally illustrated by the following "circular" reasoning. On the one hand, if gradients along the trajectory are bounded by a constant G, then the Hessian norms are bounded by the constant $\ell(G)$. Informally speaking, we essentially have the standard Lipschitz smoothness condition and can apply classical textbook analyses to prove convergence, which implies that gradients converge to zero. On the other hand, if gradients converge, they must be bounded, since any convergent sequence is bounded. In other words, the bounded gradient condition implies convergence, and convergence also implies the condition back, which forms a circular argument. If we can break this circularity of reasoning in a rigorous way (induction or contradiction), both the bounded gradient condition and convergence are proved.

The idea of bounding gradients along the trajectory also allows us to derive an improved convergence result for the Adaptive Moment Estimation (Adam) method proposed in [Kingma and Ba, 2014], which has become one of the most popular optimizers for solving (1.1) when f is the loss for training deep neural networks. Owing to its efficiency and robustness to hyper-parameters, it is widely applied or even sometimes the default choice in many machine learning application domains. It is also well known that Adam significantly outperforms stochastic gradient descent (SGD) for certain models like transformers [Zhang et al., 2020b, Kunstner et al., 2023, Ahn et al., 2023].

Despite its success in practice, theoretical analyses of Adam are still limited. The original proof of convergence in [Kingma and Ba, 2014] was later shown by Reddi et al. [2018] to contain gaps. The authors in [Reddi et al., 2018] also showed that for a range of momentum parameters chosen *independently with the problem instance*, Adam does not necessarily converge even for convex objectives. However, in deep learning practice, the hyper-parameters are in fact *problem-dependent* as they are usually tuned after given the problem and weight initialization. Recently, there have been many works proving the convergence of Adam for non-convex functions with various assumptions and problem-dependent hyper-parameter choices. However, these results leave significant room for improvement. For example, [D'efossez et al., 2020, Guo et al., 2021] prove the convergence to stationary points assuming the gradients are bounded by a constant, either explicitly or implicitly. On the other hand, [Zhang et al., 2022, Wang et al., 2022] consider weak assumptions, but their convergence results are still limited. See Section 1.2 for more detailed discussions of related works.

In this thesis, we develop a new convergence analysis for Adam. Our analysis does not assume the strong assumption of bounded gradients, but prove that gradients are bounded along the trajectory of Adam with high probability. In addition, we consider the generalized ℓ-smoothness with a sub-quadratic ℓ, and obtain the $\mathcal{O}(\epsilon^{-4})$ gradient complexity in the stochastic setting assuming noise is sub-Gaussian, as shown in Table 1.1. Besides, we also propose a varaince-reduced version of Adam and obtain the accelerated $\mathcal{O}(\epsilon^{-3})$ gradient complexity.

However, although our convergence result of Adam improves over existing ones for this method, it is still not better than that of SGD. Specifically, Adam achieves the same $\mathcal{O}(\epsilon^{-4})$ gradient complexity as SGD, but requires a stronger bounded noise condition. In fact, the rate of SGD is already not improvable among first-order methods under the classical smoothness or our ℓ-smoothness condition, as it matches the lower bound in [Arjevani et al., 2023]. Therefore, the above results can not explain why adaptive methods like Adam outperform SGD in some deep learning tasks like training transformers. To bridge this gap, we provide new insights into the benefits of adaptivity to optimization for machine learning applications by delving into the concept of directional smoothness.

For a given vector $u \in \mathbb{R}^d$ and a point $x \in \mathcal{X}$, the directional smoothness at x along u is

Table 1.1: Summary of the convergence results under ℓ-smoothness. ϵ denotes the sub-optimality gap of the function value in convex settings, and the gradient norm in non-convex settings. "$*$" denotes optimal rates.

Method	Convexity	ℓ-smoothness	Gradient complexity
GD	Strongly convex Convex	No requirement	$\mathcal{O}(\log(1/\epsilon))$ (Theorem 3.1.3) $\mathcal{O}(1/\epsilon)$ (Theorem 3.1.2)
	Non-convex	Sub-quadratic ℓ	$\mathcal{O}(1/\epsilon^2)^*$ (Theorem 3.2.2)
		Quadratic ℓ	Ω(exp. in cond #) (Theorem 3.2.4)
NAG	Convex	Sub-quadratic ℓ	$\mathcal{O}(1/\sqrt{\epsilon})^*$ (Theorem 3.1.4)
SGD	Non-convex	Sub-quadratic ℓ	$\mathcal{O}(1/\epsilon^4)^*$ (Theorem 3.2.3)
Adam	Non-convex	Sub-quadratic ℓ	$\tilde{\mathcal{O}}(1/\epsilon^4)$ (Theorem 4.2.1 and 4.2.2)

defined as $\ell_x(u) := u^\top \nabla^2 f(x) u / \|u\|^2$. Recently, [Pan and Li, 2023] empirically computed the directional smoothness $\ell_{x_t}(x_{t+1} - x_t)$ along the trajectories of various optimizers including SGD, Adam, etc. They found that adaptive methods usually have much better (smaller) directional smoothness along the optimization trajectory for training transformers, which they believe may explain why adaptive methods converge faster, as a smaller directional smoothness allows a larger stepsize and potentially faster convergence. In this thesis, we formalize the empirical observations in [Pan and Li, 2023] into a rigorous convergence theory to gain new insights on why and when adaptivity accelerates optimization. In particular, assuming that the directional smoothness around the update direction of memoryless Adam is bounded by some constant L_λ, we can obtain a gradient complexity of $\mathcal{O}(L_\lambda \epsilon^{-2})$ for two special cases of Adam, memoryless Adam and RMSProp, in the deterministic setting. If $L_\lambda \ll L$, which is supported by the empirical observations in [Pan and Li, 2023] and our experiments, our convergence results show acceleration of these adaptive methods compared to the typical $\mathcal{O}(L\epsilon^{-2})$ gradient complexity of gradient descent, providing new insights into the benefits of adaptivity in training transformers.

1.1 Overview of results

Before delving into the detailed problem formulations and convergence analyses, we first provide an overview of our results and summarize the content of each chapter.

- First, we generalize the standard Lipschitz smoothness and also the (L_0, L_1)-smoothness condition to the ℓ-smoothness condition. In Chapter 2, we will present its formal definitions in more detail. Moreover, we will show some useful properties of ℓ-smooth functions and briefly discuss how they can be applied in the convergence analysis. Finally, we provide several interesting examples of ℓ-smooth functions corresponding to different ℓs.

- We develop a new approach for analyzing convergence under this generalized ℓ-smoothness condition by bounding the gradients along the optimization trajectory. In Chapter 3, we apply this approach to prove the convergence of *constant-stepsize* GD, SGD, and NAG in the convex or non-convex settings, and obtain the classical rates for all of them, as summarized in Table 1.1. Our convergence results of *constant-stepsize* methods challenge the folklore belief on the necessity of adaptive stepsize for generalized smooth functions. Notably, we also relax the assumption of bounded noise to the weaker one of bounded variance of noise in the stochastic setting with the simple SGD method.

- In Chapter 4, we further apply the new approach to show that Adam converges to stationary points under relaxed assumptions compared to existing works. In particular, we do not assume bounded gradients or Lipschitzness of the objective function. Furthermore, we also consider the more general ℓ-smoothness condition. Under these more realistic assumptions, we obtain *dimension free* gradient complexities of $\mathcal{O}(\epsilon^{-4})$ if the gradient noise is centered and bounded, or $\mathcal{O}(\epsilon^{-4}\log^{3.25}(1/\epsilon))$ if the gradient noise is centered and has sub-Gaussian norm. Finally, we propose a variance-reduced version of Adam (VRAdam) with provable convergence guarantees. In particular, we obtain the accelerated $\mathcal{O}(\epsilon^{-3})$ gradient complexity.

- In Chapter 5, we propose a directional smoothness condition and analyze the convergence of memoryless Adam and RMSProp under this condition in the deterministic setting. In particular, assuming that the directional smoothness along the update direction of memoryless Adam is bounded by some constant L_λ, we show that memoryless Adam converges with a gradient complexity of $\mathcal{O}(L_\lambda \epsilon^{-2})$. We can also obtain essentially the same convergence rate for RMSProp under stronger assumptions. Compared to the typical $\mathcal{O}(L\epsilon^{-2})$ complexity of gradient descent, our results clearly achieve a faster rate when $L_\lambda \ll L$, which is supported by the the empirical observations in [Pan and Li, 2023] and our experiments. We also provide an example for which $L_\lambda \ll L$ holds and memoryless Adam or RMSProp does converge faster than gradient descent.

1.2 Related work

In this section, we discuss the relevant literature related to various aspects of this thesis.

Gradient-based optimization. The classical gradient-based optimization problems for the standard Lipschitz smooth functions have been well studied for both convex [Nemirovskij and Yudin, 1983, Nesterov, 2003, d'Aspremont et al., 2021] and non-convex functions. In the convex setting, the goal is to reach an ϵ-sub-optimal point x satisfying $f(x) - \inf_x f(x) \leq \epsilon$. It is well known that GD achieves the $\mathcal{O}(1/\epsilon)$ gradient complexity and NAG achieves the accelerated $\mathcal{O}(1/\sqrt{\epsilon})$ complexity which is optimal among all gradient-based methods. For strongly convex functions, GD and NAG achieve the $\mathcal{O}(\kappa \log(1/\epsilon))$ and $\mathcal{O}(\sqrt{\kappa}\log(1/\epsilon))$ complexity respectively, where κ is the condition number and the latter is again optimal. In the non-convex setting, the goal is to find an ϵ-stationary point x satisfying $\|\nabla f(x)\| \leq \epsilon$, since finding a global minimum is NP-hard in general. It is well known that GD achieves the optimal $\mathcal{O}(1/\epsilon^2)$ complexity which matches the lower bound in [Carmon et al., 2017]. In the stochastic setting for unbiased stochastic gradient with bounded variance, SGD achieves the optimal $\mathcal{O}(1/\epsilon^4)$ complexity [Ghadimi and Lan, 2013], matching the lower bound in [Arjevani et al., 2023]. In this thesis, we obtain the classical rates in terms of ϵ for all the above-mentioned methods and settings, under a far more general smoothness condition.

Generalized smoothness. The (L_0, L_1)-smoothness condition proposed by Zhang et al. [2019] was studied by many follow-up works. Under the same condition, [Zhang et al., 2020a] considers momentum in the updates and improves the constant dependency of the convergence rate for SGD with clipping derived in [Zhang et al., 2019]. [Qian et al., 2021] studies gradient clipping in incremental gradient methods, [Zhao et al., 2021] studies stochastic normalized gradient descent, and [Crawshaw et al., 2022] studies a generalized SignSGD method, under the (L_0, L_1)-smoothness condition. [Reisizadeh et al., 2023] studies variance reduction for (L_0, L_1)-smooth functions. [Chen et al., 2023] proposes a new notion of α-symmetric generalized smoothness, which is roughly as general as (L_0, L_1)-smoothness. [Wang et al., 2022] analyzes convergence of Adam and provides a lower bound which shows non-adaptive SGD may diverge. In the stochastic setting, the above-mentioned works either consider the strong assumption of bounded gradient noise or require a very large batch size that depends on ϵ, which essentially reduces the analysis to the deterministic setting. [Faw et al., 2023] proposes an AdaGrad-type algorithm in order to relax the bounded noise assumption. Perhaps due to the lower bounds in [Zhang et al., 2019, Wang et al., 2022], all the above works study methods with an adaptive stepsize. In this thesis, we further generalize the smoothness condition and analyze various methods under this condition through bounding the gradients along the trajectory.

Convergence of Adam. Adam was first proposed by Kingma and Ba [2014] with a theoretical convergence guarantee for convex functions. However, Reddi et al. [2018] found a gap in the proof of this convergence analysis, and also constructed counter-examples for a range of hyper-parameters on which Adam does not converge. That being said, the counter-examples depend on the hyper-parameters of Adam, i.e., they are constructed after picking the hyper-parameters. Therefore, it does not rule out the possibility of obtaining convergence guarantees for problem-dependent hyper-parameters, as also pointed out by [Shi et al., 2021, Zhang et al., 2022].

Many recent works have developed convergence analyses of Adam with various assumptions and hyper-parameter choices. Zhou et al. [2018b] show Adam with certain hyper-parameters can work on the counter-examples of [Reddi et al., 2018]. De et al. [2018] prove convergence for general non-convex functions assuming gradients are bounded and the signs of stochastic gradients are the same along the trajectory. The analysis in [D'efossez et al., 2020] also relies on the bounded gradient assumption. Guo et al. [2021] assume the adaptive stepsize is upper and lower bounded by two constants, which is not necessarily satisfied unless assuming bounded gradients or considering the AdaBound variant [Luo et al., 2019]. [Zhang et al., 2022, Wang et al., 2022] consider very weak assumptions. However, they show either 1) "convergence" only to some neighborhood of stationary points with a constant radius, unless assuming the strong growth condition; or 2) convergence to stationary points but with a slower rate. In a concurrent work [Wang et al., 2023], the authors show the convergence of Adam without assuming strong conditions like bounded gradients. They also consider the bounded varaince assumption on the gradient noise, weaker than our sub-Gaussian noise condition. However, their convergence rate is dimendion-dependent and only considers the standard smoothness condition.

Benefit of adaptivity. Some recent works attempt to provide explanations on why adaptive methods outperform SGD for machine learning tasks like training transformers. For example, [Zhang et al., 2019, Wang et al., 2022] study the (L_0, L_1)-smoothness condition motivated by language model experiments, and claim this condition can theoretically explain the benefit of adaptivity based on their convergence results. However, their claims are undermined by our convergence results of GD/SGD in Chapter 3, which will be discussed in more detail in Section 3.2.3. [Ahn et al., 2024] studies the framework called online learning of updates and connects both SGD and Adam to well-known online learning methods to understand why Adam is better than SGD. [Zhang et al., 2020b] provides both empirical and theoretical evidence that heavy-tailed noise distribution may be one cause of the poor performance of SGD, which they show can be improved with gradient clipping. [Kunstner et al., 2024] empirically shows heavy-tailed class imbalance may lead to difficulty in optimization, which they believe can be counteracted by the normalization used in Adam. [Jiang et al., 2023] shows that adaptive methods bias their trajectories towards regions with a smaller condition number defined in their paper. [Zhang et al., 2024] empirically observe block heterogeneity in the Hessian spectrum of training transformers, and provide evidence showing this may be the reason why SGD performs worse. Finally, [Pan and Li, 2023] empirically shows that adaptive methods like Adam may have a smaller directional smoothness, which may lead to faster convergence. It also motivates our work shown in Chapter 5.

Variants of Adam. After Reddi et al. [2018] pointed out the non-convergence issue with Adam, various variants of Adam that can be proved to converge were proposed [Zou et al., 2018, Gadat and Gavra, 2022, Chen et al., 2018, 2022, Luo et al., 2019, Zhou et al., 2018b]. For example, AMSGrad [Reddi et al., 2018] and AdaFom [Chen et al., 2018] modify the second order momentum so that it is non-decreasing. AdaBound [Luo et al., 2019] explicitly imposes upper and lower bounds on the second order momentum so that the stepsize is also bounded. AdaShift [Zhou et al., 2018b] uses a new estimate of the second order momentum to correct the bias. There are also some works [Zhou et al., 2018a, Gadat and Gavra, 2022, Iiduka, 2023] that provide convergence guarantees of these variants. One closely related work to ours is [Wang and Klabjan, 2022], which considers a variance-reduced version of Adam by combining Adam and SVRG [Johnson and Zhang, 2013]. However, they assume bounded gradients and can only get an asymptotic convergence in the non-convex setting.

Variance reduction methods. The technique of variance reduction was introduced to accelerate convex optimization in the finite-sum setting [Roux et al., 2012, Johnson and Zhang, 2013, Shalev-Shwartz and Zhang, 2013, Mairal, 2013, Defazio et al., 2014]. Later, many works studied variance-reduced methods in the non-convex setting and obtained improved convergence rates for standard smooth functions. For example, SVRG and SCSG improve the $\mathcal{O}(\epsilon^{-4})$ gradient complexity of stochastic gradient descent (SGD) to $\mathcal{O}(\epsilon^{-10/3})$ [Allen-Zhu and Hazan, 2016, Reddi et al., 2016, Lei et al., 2017]. Many new variance reduction methods [Fang et al., 2018, Tran-Dinh et al., 2019, Liu et al., 2020, Li et al., 2021, Cutkosky and Orabona, 2019, Liu et al., 2023] were later proposed to further improve the complexity to $\mathcal{O}(\epsilon^{-3})$, which is optimal and matches the lower bound in [Arjevani et al., 2023]. Recently, [Reisizadeh et al., 2023, Chen et al., 2023] obtained the $\mathcal{O}(\epsilon^{-3})$ complexity for the more general (L_0, L_1)-smooth functions. Our variance-reduced Adam is motivated by the STORM algorithm proposed by [Cutkosky and Orabona, 2019], where an additional term is added in the momentum update to correct the bias and reduce the variance.

1.3 Preliminaries

Notation. For any given vector x, we use $x_{[i]}$ to denote its i-th coordinate and x^2, $\sqrt{x}$, $|x|$ to denote its *coordinate-wise* square, square root, and absolute value respectively. For any two vectors x and y, we use $x \odot y$ and x/y to denote their *coordinate-wise* product and quotient respectively. We also write $x \preceq y$ or $x \succeq y$ to denote the *coordinate-wise* inequality between x and y, which means $x_{[i]} \leq y_{[i]}$ or $x_{[i]} \geq y_{[i]}$ for each coordinate index i respectively. Similarly, for any scalar a and vector x, the coordinate-wise inequality $x \preceq a$ or $x \succeq a$ means $x_{[i]} \leq a$ or $x_{[i]} \geq a$ for each coordinate index i respectively.

For any real-valued function f, we use $\mathrm{dom}(f)$ to denote its domain. We also use $\mathcal{B}(x, R)$ to denote the Euclidean ball with radius $R \geq 0$ centered at point x. Let $\|\cdot\|$ denote the ℓ_2 norm of a vector or spectral norm of a matrix. We also use $\|\cdot\|_1$ and $\|\cdot\|_\infty$ to denote the ℓ_1 and

ℓ_∞ norms of a vector respectively. For any positive semi-definite matrix A and vector x, we denote $\|x\|_A := \sqrt{x^\top Ax}$ as the weighted norm of x with respect to A. For two symmetric real matrices A and B, we say $A \preceq B$ or $A \succeq B$ if $B - A$ or $A - B$ is positive semi-definite (PSD). For any vector x, we use $\operatorname{diag}(x)$ to denote the diagonal matrix whose principle diagonal entries are the coordinates of x. Given two real numbers $a, b \in \mathbb{R}$, we denote $a \wedge b := \min\{a, b\}$ for simplicity. Finally, we use $\mathcal{O}(\cdot)$, $\Theta(\cdot)$, and $\Omega(\cdot)$ for the standard big-O, big-Theta, and big-Omega notation, with $\tilde{\mathcal{O}}(\cdot)$, $\tilde{\Theta}(\cdot)$, and $\tilde{\Omega}(\cdot)$ further hiding logarithmic factors.

1.3.1 Standard assumptions on the objective function

Next, we present the following two standard assumptions in the literature of unconstrained optimization. These will be assumed throughout Chapters 3, 4 and 5.

Assumption 1.1. The objective function f is differentiable and *closed* within its *open* domain $\mathcal{X}$.

Assumption 1.2. The objective function f is bounded from below, i.e., $f^* := \inf_{x\in\mathcal{X}} f(x) > -\infty$.

A function f is said to be closed if its sub-level set $\{x \in \operatorname{dom}(f) \mid f(x) \leq a\}$ is closed for each $a \in \mathbb{R}$. A continuous function f with an open domain is closed if and only $f(x)$ tends to positive infinity when x approaches the boundary of its domain [Boyd and Vandenberghe, 2004]. Assumption 1.1 is necessary for our analysis to ensure that the iterates of a method with a reasonably small stepsize stays within the domain $\mathcal{X}$. Note that for $\mathcal{X} = \mathbb{R}^d$ considered in most unconstrained optimization papers, the assumption is trivially satisfied as all continuous functions over $\mathbb{R}^d$ are closed. We consider a more general domain which may not be the whole space because that is the case for some interesting examples in our generalized function class of interest (see Section 2.3). However, it actually brings us some additional technical difficulties especially in the stochastic setting, as we need to make sure the iterates do not go outside of the domain.

Chapter 2

Generalized smoothness

In this chapter, we formally define the generalized smoothness condition, presenting its properties and examples. In Section 2.1, we introduce two equivalent definitions of the generalized smoothness condition, termed ℓ-smoothness and (r,ℓ)-smoothness, respectively. The (r,ℓ)-smoothness definition represents a local smoothness condition, which we will utilize to derive useful properties for the convergence analysis under our generalized smoothness conditions in Section 2.2. These properties will be extensively applied in our analyses of classical methods and Adam in Chapters 3 and 4. Conversely, the ℓ-smoothness definition is intuitive and mathematically straightforward. We will employ it to explore several interesting examples in Section 2.3. For all lemmas and propositions in this chapter, we present only their statements and defer the proofs to Appendix A.

2.1 Definitions

Definitions 1 and 2 below are two equivalent ways of stating the definition, where we use $\mathcal{B}(x,R)$ to denote the Euclidean ball with radius R centered at x.

Definition 1 (ℓ-smoothness)**.** A real-valued differentiable function $f:\mathcal{X}\to\mathbb{R}$ is ℓ-smooth for some non-decreasing continuous function $\ell:[0,+\infty)\to(0,+\infty)$ if $\|\nabla^2 f(x)\|\le\ell(\|\nabla f(x)\|)$

almost everywhere (with respect to the Lebesgue measure) in $\mathcal{X}$.

Remark 2.1.1. Definition 1 reduces to the classical L-smoothness when $\ell \equiv L$ is a constant function. It reduces to the (L_0, L_1)-smoothness proposed in [Zhang et al., 2019] when $\ell(u) = L_0 + L_1 u$ is an affine function.

Definition 2 ((r,ℓ)-smoothness)**.** A real-valued differentiable function $f : \mathcal{X} \to \mathbb{R}$ is (r,ℓ)-smooth for continuous functions $r, \ell : [0, +\infty) \to (0, +\infty)$ where ℓ is non-decreasing and r is non-increasing, if it satisfies 1) for any $x \in \mathcal{X}$, $\mathcal{B}(x, r(\|\nabla f(x)\|)) \subseteq \mathcal{X}$, and 2) for any $x_1, x_2 \in \mathcal{B}(x, r(\|\nabla f(x)\|))$, $\|\nabla f(x_1) - \nabla f(x_2)\| \leq \ell(\|\nabla f(x)\|) \cdot \|x_1 - x_2\|$.

The requirements that ℓ is non-decreasing and r is non-increasing do not cause much loss in generality. If these conditions are not satisfied, one can replace ℓ and r with the non-increasing function $\tilde{r}(u) := \inf_{0 \leq v \leq u} r(v) \leq r(u)$ and non-decreasing function $\tilde{\ell}(u) := \sup_{0 \leq v \leq u} \ell(v) \geq \ell(u)$ in Definitions 1 and 2. Then the only requirement is $\tilde{r} > 0$ and $\tilde{\ell} < \infty$.

Next, we prove that the above two definitions are equivalent in the following important proposition, whose proof is involved and deferred to Appendix A.2.

Proposition 2.1.2. *An (r,ℓ)-smooth function is ℓ-smooth; and a closed ℓ-smooth function over an open domain is (r,m)-smooth where $m(u) := \ell(u + a)$ and $r(u) := a/m(u)$ for any $a > 0$.*

The condition in Definition 1 is simple and one can easily check whether it is satisfied for a given example function. On the other hand, Definition 2 is a local Lipschitz condition on the gradient that is harder to verify. However, it is useful for deriving several useful properties in the next section.

2.2 Properties

First, we provide the following lemma which is very useful in our analyses of most methods considered in this thesis. Its proof is deferred to Appendix A.3.

Lemma 2.2.1. *If f is (r,ℓ)-smooth, for any $x \in \mathcal{X}$ satisfying $\|\nabla f(x)\| \le G$, we have 1) $\mathcal{B}(x, r(G)) \subseteq \mathcal{X}$, and 2) for any $x_1, x_2 \in \mathcal{B}(x, r(G))$,*

$$\|\nabla f(x_1) - \nabla f(x_2)\| \le L \|x_1 - x_2\|, \quad f(x_1) \le f(x_2) + \left\langle \nabla f(x_2), x_1 - x_2 \right\rangle + \frac{L}{2} \|x_1 - x_2\|^2, \quad (2.1)$$

where $L := \ell(G)$ is the effective smoothness constant.

Remark 2.2.2. Since we have shown the equivalence between ℓ-smoothness and (r,ℓ)-smoothness, Lemma 2.2.1 also applies to ℓ-smooth functions, for which we have $L = \ell(2G)$ and $r(G) = G/L$ if choosing $a = G$ in Proposition 2.1.2.

Lemma 2.2.1 states that, if the gradient at x is bounded by some constant G, then within its neighborhood with a *constant* radius, we can obtain (2.1), the same inequalities that were derived in the textbook analysis [Nesterov, 2003] under the standard Lipschitz smoothness condition. With (2.1), the analysis for generalized smoothness is not much harder than that for standard smoothness. Since we mostly choose $x = x_2 = x_t$ and $x_1 = x_{t+1}$ in the analysis, in order to apply Lemma 2.2.1, we need two conditions: $\|\nabla f(x_t)\| \le G$ and $\|x_{t+1} - x_t\| \le r(G)$ for some constant G. The latter is usually directly implied by the former for most deterministic methods with a small enough stepsize, and the former can be obtained with our new approach that bounds the gradients along the trajectory.

With Lemma 2.2.1, we can derive the following useful lemma which is the reverse direction of a generalized Polyak-Lojasiewicz (PL) inequality, whose proof is deferred to Appendix A.3.

Lemma 2.2.3. *If f is ℓ-smooth satisfying Assumption 1.1, then $\|\nabla f(x)\|^2 \le 2\ell(2\|\nabla f(x)\|) \cdot (f(x) - \inf_x f(x))$ for any $x \in \mathcal{X}$.*

Lemma 2.2.3 provides an inequality involving the gradient norm and the sub-optimality gap. For example, when $\ell(u) = u^\rho$ for some $0 \le \rho < 2$, this lemma suggests $\|\nabla f(x)\| \le \mathcal{O}\left((f(x) - f^*)^{1/(2-\rho)}\right)$, which means the gradient norm is bounded whenever the function value is bounded. The following corollary provides a more formal statement for general sub-quadratic ℓ (i.e., $\lim_{u\to\infty} \ell(u)/u^2 = 0$), and we defer its proof to Appendix A.3.

Corollary 2.2.4. *Suppose f is ℓ-smooth satisfying Assumption 1.1 and 1.2 where ℓ is sub-quadratic. If $f(x) - \inf_x f(x) \leq F$ for some $x \in \mathcal{X}$ and $F \geq 0$, denoting $G := \sup\{u \geq 0 \mid u^2 \leq 2\ell(2u) \cdot F\}$, then they satisfy $G^2 = 2\ell(2G) \cdot F$ and we have $\|\nabla f(x)\| \leq G < \infty$.*

Therefore, in order to bound the gradients along the trajectory as we discussed below Lemma 2.2.1, it suffices to bound the function values, which is usually easier.

2.3 Examples

The most important subset of ℓ-smooth (or (r, ℓ)-smooth) functions are those with a polynomial ℓ, and can be characterized by the (ρ, L_0, L_ρ)-smooth function class defined below.

Definition 3 ((ρ, L_0, L_ρ)-smoothness)**.** A real-valued differentiable function f is (ρ, L_0, L_ρ)-smooth for constants $\rho, L_0, L_\rho \geq 0$ if it is ℓ-smooth with $\ell(u) = L_0 + L_\rho u^\rho$.

Definition 3 reduces to the standard Lipschitz smoothness condition when $\rho = 0$ or $L_\rho = 0$ and to the (L_0, L_1)-smoothness proposed in [Zhang et al., 2019] when $\rho = 1$. We list several univariate examples of (ρ, L_0, L_ρ)-smooth functions for different ρs in Table 2.1 with their rigorous justifications in Appendix A.1. Note that when x goes to infinity, polynomial and exponential functions corresponding to $\rho = 1$ grow much faster than quadratic functions corresponding to $\rho = 0$. Rational and logarithmic functions for $\rho > 1$ grow even faster as they can blow up to infinity near finite points. Note that the domains of such functions are not $\mathbb{R}^d$, which is why we consider a general open domain $\mathcal{X}$ instead of simply assuming $\mathcal{X} = \mathbb{R}^d$.

Aside from logarithmic functions, the $(2, L_0, L_2)$-smooth function class also includes other univariate *self-concordant* functions. This is an important function class in the analysis of Interior Point Methods and coordinate-free analysis of the Newton method [Nesterov, 2003]. More specifically, a convex function $h : \mathbb{R} \to \mathbb{R}$ is self-concordant if $|h'''(x)| \leq 2h''(x)^{3/2}$ for all $x \in \mathbb{R}$. Formally, we have the following proposition whose proof is deferred to Appendix A.1.

Table 2.1: Examples of univariate (ρ, L_0, L_ρ) smooth functions for different ρs. The parameters a, b, p are *real numbers* (not necessarily integers) satisfying $a, b > 1$ and $p < 1$ or $p \geq 2$. We use 1^+ to denote any real number slightly larger than 1.

ρ	0	1	1	1^+	1.5	2	$\frac{p-2}{p-1}$
Examples	Quadratic	Polynomial	a^x	$a^{(b^x)}$	Rational	Logarithmic	x^p

Proposition 2.3.1. *If $h : \mathbb{R} \to \mathbb{R}$ is a self-concordant function satisfying $h''(x) > 0$ over the interval (a, b), then h restricted on (a, b) is $(2, L_0, 2)$-smooth for some $L_0 > 0$.*

Chapter 3

Convergence of classical methods

In this chapter, we analyze the convergence of classical methods in both convex and non-convex settings under the generalized ℓ-smoothness condition introduced in Chapter 2. Section 3.1 details the analysis approach and convergence results of gradient descent (GD) and Nesterov's accelerated gradient method (NAG) in convex and strongly convex settings. Then in Section 3.2, we present the results for gradient descent (GD) and stochastic gradient descent (SGD) in the non-convex setting. All the results recover the classical convergence rates under the standard Lipschitz smoothness condition up to constant factors, where NAG is optimal in the convex setting, and GD or SGD achieves the optimal rate in the deterministic or stochastic non-convex setting. We have no additional requirement on ℓ for GD in the convex setting but require ℓ to be sub-quadratic in all other cases. For GD/SGD in the non-convex setting, we are able to show such a requirement is necessary by providing an exponential lower bound on the iteration complexity when ℓ is not sub-quadratic. However, it is not clear whether it is also necessary for NAG in the convex setting. See Table 1.1 for a summary of their convergence rates.

Notably, we are considering non-adaptive or constant-stepsize GD/SGD unlike in most existing works on generalized smoothness. Our results challenge the folklore belief on the necessity of adaptive stepsize for generalized smooth functions. In addition, we also relax the assumption of bounded noise to the weaker one of bounded variance of noise in the stochastic setting with the simple SGD method. See Section 1.2 for detailed discussions.

3.1 Convex setting

In this section, we present the convergence results of gradient descent (GD) and Nesterov's accelerated gradient method (NAG) in the convex setting. Formally, we define convexity as follows.

Definition 4. A real-valued differentiable function $f : \mathcal{X} \to \mathbb{R}$ is μ-strongly-convex for $\mu \geq 0$ if $\mathcal{X}$ is a convex set and $f(y) - f(x) \geq \left\langle \nabla f(x), y - x \right\rangle + \frac{\mu}{2} \left\| y - x \right\|^2$ for any $x, y \in \mathcal{X}$. A function is convex if it is μ-strongly-convex with $\mu = 0$.

We assume the existence of a global optimal point x^* throughout this section, as in the following assumption. However, we want to note that, for gradient descent, this assumption is just for simplicity rather than necessary.

Assumption 3.1. There exists a point $x^* \in \mathcal{X}$ such that $f(x^*) = f^* = \inf_{x \in \mathcal{X}} f(x)$.

3.1.1 Gradient descent

The gradient descent method with a constant stepsize η is defined via the following update rule

$$x_{t+1} = x_t - \eta \nabla f(x_t). \tag{3.1}$$

As discussed below Lemma 2.2.1, the key in the convergence analysis is to show $\|\nabla f(x_t)\| \leq G$ for all $t \geq 0$ and some constant G. We will prove it by induction relying on the following lemma whose proof is deferred to Appendix B.1.

Lemma 3.1.1. *For any $x \in \mathcal{X}$ satisfying $\|\nabla f(x)\| \leq G$, define $x^+ := x - \eta \nabla f(x)$. If f is convex and (r, ℓ)-smooth, and $\eta \leq \min\left\{\frac{2}{\ell(G)}, \frac{r(G)}{2G}\right\}$, we have $x^+ \in \mathcal{X}$ and $\|\nabla f(x^+)\| \leq \|\nabla f(x)\| \leq G$.*

Lemma 3.1.1 suggests that for gradient descent (3.1) with a small enough stepsize, if the gradient norm at x_t is bounded by G, then we have $\|\nabla f(x_{t+1})\| \leq \|\nabla f(x_t)\| \leq G$, i.e., the gradient norm is also bounded by G at $t+1$. In other words, the gradient norm is indeed a non-increasing potential function for gradient descent in the convex setting. With a standard induction argument, we can show that $\|\nabla f(x_t)\| \leq \|\nabla f(x_0)\|$ for all $t \geq 0$. As discussed below Lemma 2.2.1, then we can basically apply the classical analysis to obtain the convergence guarantee in the convex setting as in the following theorem, whose proof is deferred to Appendix B.1.

Theorem 3.1.2. *Suppose f is convex and (r, ℓ)-smooth satisfying Assumptions 1.1, 1.2, and 3.1. Denote $G := \|\nabla f(x_0)\|$ and $L := \ell(G)$, then the iterates generated by (3.1) with $\eta \leq \min\left\{\frac{1}{L}, \frac{r(G)}{2G}\right\}$ satisfy $\|\nabla f(x_t)\| \leq G$ for all $t \geq 0$ and*

$$f(x_T) - f^* \leq \frac{\|x_0 - x^*\|^2}{2\eta T}.$$

Since η is a constant independent of ϵ or T, Theorem 3.1.2 achieves the classical $\mathcal{O}(1/T)$ rate, or $\mathcal{O}(1/\epsilon)$ gradient complexity to achieve an ϵ-sub-optimal point, under the generalized smoothness condition. Since strongly convex functions are a subset of convex functions, Lemma 3.1.1 still holds for them. Then we immediately obtain the following result in the strongly convex setting, whose proof is deferred to Appendix B.1.

Theorem 3.1.3. *Suppose f is μ-strongly-convex and (r, ℓ)-smooth satisfying Assumptions 1.1, 1.2, and 3.1. Denote $G := \|\nabla f(x_0)\|$ and $L := \ell(G)$, then the iterates generated by (3.1) with $\eta \leq \min\left\{\frac{1}{L}, \frac{r(G)}{2G}\right\}$ satisfy $\|\nabla f(x_t)\| \leq G$ for all $t \geq 0$ and*

$$f(x_T) - f^* \leq \frac{\mu(1-\eta\mu)^T}{2(1-(1-\eta\mu)^T)} \|x_0 - x^*\|^2.$$

Theorem 3.1.3 gives a linear convergence rate and the $\mathcal{O}((\eta\mu)^{-1}\log(1/\epsilon))$ gradient complexity to achieve an ϵ-sub-optimal point. Note that for ℓ-smooth functions, we have $\frac{r(G)}{G} = \frac{1}{L}$ (see Remark 2.2.2), which means we can choose $\eta = \frac{1}{2L}$. Then we obtain the $\mathcal{O}(\kappa\log(1/\epsilon))$ rate, where $\kappa := L/\mu$ is the local condition number around the initial point x_0. For standard Lipschitz smooth functions, it reduces to the classical rate of gradient descent.

3.1.2 Nesterov's accelerated gradient method

Algorithm 1 Nesterov's accelerated gradient method (NAG)

```
1: Input: A convex and $\ell$-smooth function $f$, stepsize $\eta$, initial point $x_0$
2: Initialize $z_0 = x_0$, $B_0 = 0$, and $A_0 = 1/\eta$.
3: for $t = 0, \ldots$ do
4:     $B_{t+1} = B_t + \frac{1}{2}\left(1 + \sqrt{4B_t + 1}\right)$
5:     $A_{t+1} = B_{t+1} + 1/\eta$
6:     $y_t = x_t + (1 - A_t/A_{t+1})(z_t - x_t)$
7:     $x_{t+1} = y_t - \eta\nabla f(y_t)$
8:     $z_{t+1} = z_t - \eta(A_{t+1} - A_t)\nabla f(y_t)$
9: end for
```

In the case of convex and standard Lipschitz smooth functions, it is well known that Nesterov's accelerated gradient method (NAG) achieves the optimal $\mathcal{O}(1/T^2)$ rate. In this section, we show that under the ℓ-smoothness condition with a *sub-quadratic* ℓ, the optimal $\mathcal{O}(1/T^2)$ rate can be achieved by a slightly modified version of NAG shown in Algorithm 1, the only difference between which and the classical NAG is that the latter directly sets $A_{t+1} = B_{t+1}$ in Line 4. Formally, we have the following theorem, whose proof is deferred to Appendix B.2.

Theorem 3.1.4. *Suppose f is convex and ℓ-smooth satisfying Assumptions 1.1, 1.2, and 3.1 where ℓ is sub-quadratic. Then there always exists a constant G satisfying $G \geq \max\left\{8\sqrt{\ell(2G)((f(x_0) - f^*) + \|x_0 - x^*\|^2)}, \|\nabla f(x_0)\|\right\}$. Denote $L := \ell(2G)$ and choose*

$\eta \le \min\left\{\frac{1}{16L^2}, \frac{1}{2L}\right\}$. *The iterates generated by Algorithm 1 satisfy*

$$f(x_T) - f^* \le \frac{4(f(x_0) - f^*) + 4\left\|x_0 - x^*\right\|^2}{\eta T^2 + 4}.$$

It is easy to note that Theorem 3.1.4 achieves the accelerated $\mathcal{O}(1/T^2)$ convergence rate, or equivalently the $\mathcal{O}(1/\sqrt{\epsilon})$ gradient complexity to find an ϵ-sub-optimal point, which is optimal among gradient-based methods [Nesterov, 2003].

In order to prove Theorem 3.1.4, we also use induction to show the gradients along the trajectory of Algorithm 1 are bounded by G. However, unlike gradient descent, the gradient norm is no longer a potential function or monotonically non-increasing, which makes the induction analysis more challenging. Suppose that we have shown $\|\nabla f(y_s)\| \le G$ for $s < t$. To complete the induction, it suffices to prove $\|\nabla f(y_t)\| \le G$. Since $x_t = y_{t-1} - \eta\nabla f(y_{t-1})$ is a gradient descent step by Line 6 of Algorithm 1, Lemma 3.1.1 directly shows $\|\nabla f(x_t)\| \le G$. In order to also bound $\|\nabla f(y_t)\|$, we try to control $\|y_t - x_t\|$, which is the most challenging part of our proof. Since $y_t - x_t$ can be expressed as a linear combination of past gradients $\{\nabla f(y_s)\}_{s<t}$, it might grow linearly with t if we simply apply $\|\nabla f(y_s)\| \le G$ for $s < t$. Fortunately, Lemma 2.2.3 allows us to control the gradient norm with the function value. Thus if the function value is decreasing sufficiently fast, which can be shown by following the standard Lyapunov analysis of NAG, we are able to obtain a good enough bound on $\|\nabla f(y_s)\|$ for $s < t$, which allows us to control $\|y_t - x_t\|$. We defer the detailed proof to Appendix B.2.

Note that Theorem 3.1.4 requires a smaller stepsize $\eta = \mathcal{O}(1/L^2)$, compared to the classical $\mathcal{O}(1/L)$ stepsize for standard Lipschitz smooth functions. The reason is we require a small enough stepsize to get a good enough bound on $\|y_t - x_t\|$. However, if the function is further assumed to be ℓ-smooth with a *sub-linear* ℓ, the requirement of stepsize can be relaxed to $\eta = \mathcal{O}(1/L)$, similar to the classical requirement. See Appendix B.2 for the details.

In the strongly convex setting, we can also prove convergence of NAG with different

$\{A_t\}_{t\geq 0}$ parameters when f is ℓ-smooth with a sub-quadratic ℓ, or (ρ, L_0, L_ρ)-smooth with $\rho < 2$. The rate can be further improved when ρ becomes smaller. However, since the constants G and L are different for GD and NAG, it is not clear whether the rate of NAG is faster than that of GD in the strongly convex setting. We will present the detailed result and analysis in Appendix B.3.

3.2 Non-convex setting

In this section, we present convergence results of gradient descent (GD) and stochastic gradient descent (SGD) in the non-convex setting.

3.2.1 Gradient descent

Similar to the convex setting, we still want to bound the gradients along the trajectory. However, in the non-convex setting, the gradient norm is not necessarily non-increasing. Fortunately, similar to the classical analyses, the function value is still non-increasing and thus a potential function, as formally shown in the following lemma, whose proof is deferred to Appendix B.4.

Lemma 3.2.1. *Suppose f is ℓ-smooth satisfying Assumptions 1.1 and 1.2 where ℓ is sub-quadratic. For any given $F \geq 0$, let $G := \sup\{u \geq 0 \mid u^2 \leq 2\ell(2u) \cdot F\}$ and $L := \ell(2G)$. For any $x \in \mathcal{X}$ satisfying $f(x) - f^* \leq F$, define $x^+ := x - \eta\nabla f(x)$ where $\eta \leq 2/L$, we have $x^+ \in \mathcal{X}$ and $f(x^+) \leq f(x)$.*

Then using a standard induction argument, we can show $f(x_t) \leq f(x_0)$ for all $t \geq 0$. According to Corollary 2.2.4, it implies bounded gradients along the trajectory. Therefore, we can show convergence of gradient descent as in the following theorem, whose proof is deferred to Appendix B.4.

Theorem 3.2.2. *Suppose f is ℓ-smooth satisfying Assumptions 1.1 and 1.2 where ℓ is sub-quadratic. Let $G := \sup\{u \geq 0 \mid u^2 \leq 2\ell(2u) \cdot (f(x_0) - f^*)\}$ and $L := \ell(2G)$. If $\eta \leq 1/L$, the iterates generated by (3.1) satisfy $\|\nabla f(x_t)\| \leq G$ for all $t \geq 0$ and*

$$\frac{1}{T} \sum_{t<T} \|\nabla f(x_t)\|^2 \leq \frac{2(f(x_0) - f^*)}{\eta T}.$$

It is clear that Theorem 3.2.2 gives the classical $\mathcal{O}(1/\epsilon^2)$ gradient complexity to achieve an ϵ-stationary point, which is optimal as it matches the lower bound in [Carmon et al., 2017].

3.2.2 Stochastic gradient descent

In this part, we present the convergence result for stochastic gradient descent defined as follows.

$$x_{t+1} = x_t - \eta \nabla f(x_t, \xi_t), \tag{3.2}$$

where $\nabla f(x_t, \xi_t)$ is an estimate of the gradient $\nabla f(x_t)$ parametrized by the random variable ξ_t. We consider the following standard assumption on the gradient noise $\epsilon_t := \nabla f(x_t, \xi_t) - \nabla f(x_t)$.

Assumption 3.2. $\mathbb{E}_{t-1}[\epsilon_t] = 0$ and $\mathbb{E}_{t-1}\left[\|\epsilon_t\|^2\right] \leq \sigma^2$ for some $\sigma \geq 0$, where $\mathbb{E}_{t-1}$ denotes the expectation conditioned on $\{\xi_s\}_{s<t}$.

Under Assumption 3.2, we can obtain the following theorem.

Theorem 3.2.3. *Suppose f is ℓ-smooth satisfying Assumptions 1.1, 1.2, and 3.2 where ℓ is sub-quadratic. For any $0 < \delta < 1$, we denote $F := 8(f(x_0) - f^* + \sigma)/\delta$ and $G := \sup\{u \geq 0 \mid u^2 \leq 2\ell(2u) \cdot F\} < \infty$. Denote $L := \ell(2G)$ and choose $\eta \leq \min\left\{\frac{1}{2L}, \frac{\epsilon^2}{16G^2F}\right\}$ and $\frac{F}{\eta\epsilon^2} \leq T \leq \frac{1}{16G^2\eta^2}$ for any $\epsilon > 0$. Then with probability at least $1-\delta$, the iterates generated by (3.2) satisfy $\|\nabla f(x_t)\| \leq G$ for all $t < T$ and*

$$\frac{1}{T} \sum_{t<T} \|\nabla f(x_t)\|^2 \leq \epsilon^2.$$

As we choose $\eta = \mathcal{O}(1/\sqrt{T})$, Theorem 3.2.3 gives the classical $\mathcal{O}(1/\epsilon^4)$ gradient complexity, where we ignore non-leading terms. This rate is optimal as it matches the lower bound in [Arjevani et al., 2023]. The key to its proof is again to bound the gradients along the trajectory. However, bounding gradients in the stochastic setting is much more challenging than in the deterministic setting, especially with the heavy-tailed noise in Assumption 3.2. We briefly discuss some of the challenges as well as our approach below and defer the detailed proof of Theorem 3.2.3 to Appendix B.5.

First, due to the existence of heavy-tailed gradient noise as considered in Assumption 3.2, neither the gradient nor the function values is non-increasing. The induction analyses we have used in the deterministic setting hardly work. In addition, to apply Lemma 2.2.1, we need to control the update at each step and make sure $\|x_{t+1} - x_t\| = \eta \|\nabla f(x_t, \xi_t)\| \le G/L$. However, $\nabla f(x_t, \xi_t)$ might be unbounded due to the potentially unbounded gradient noise.

To overcome these challenges, we define the following random variable τ.

$$\begin{aligned} \tau_1 &:= \min\{t \mid f(x_{t+1}) - f^* > F\} \wedge T, \\ \tau_2 &:= \min\left\{t \,\middle|\, \|\epsilon_t\| > \frac{G}{5\eta L}\right\} \wedge T, \\ \tau &:= \min\{\tau_1, \tau_2\}, \end{aligned} \tag{3.3}$$

where we use $a \wedge b$ to denote $\min\{a, b\}$ for any $a, b \in \mathbb{R}$. Then at least before time τ, we know that the function value and gradient noise are bounded, where the former also implies bounded gradients according to Corollary 2.2.4. Therefore, it suffices to show the probability of $\tau < T$ is small, which means with a high probability, $\tau = T$ and thus gradients are always bounded before T.

Since both the gradient and noise are bounded for $t < \tau$, it is straightforward to bound the update $\|x_{t+1} - x_t\|$, which allows us to use Lemma 2.2.1 and other useful properties. However, it is still non-trivial to upper bound $\mathbb{E}[f(x_\tau) - f^*]$ as τ is a random variable instead of a fixed time step. Fortunately, τ is a stopping time with nice properties. That is because both

$f(x_{t+1})$ and $\epsilon_t = \nabla f(x_t, \xi_t) - \nabla f(x_t)$ only depend on $\{\xi_s\}_{s\le t}$, i.e., the stochastic gradients up to t. Therefore, for any fixed t, the events $\{\tau > t\}$ only depend on $\{\xi_s\}_{s\le t}$, which show τ is a stopping time. Then with a careful analysis, we are still able to obtain an upper bound on $\mathbb{E}[f(x_\tau) - f^*] = \mathcal{O}(1)$.

On the other hand, $\tau < T$ means either $\tau = \tau_1 < T$ or $\tau = \tau_2 < T$. If $\tau = \tau_1 < T$, by its definition, we know $f(x_{\tau+1}) - f^* > F$. Roughly speaking, it also suggests $f(x_\tau) - f^* > F/2$. If we choose F such that it is much larger than the upper bound on $\mathbb{E}[f(x_\tau) - f^*]$ we just obtained, by Markov's inequality, we can show the probability of $\tau = \tau_1 < T$ is small. In addition, by union bound and Chebyshev's inequality, the probability of $\tau_2 < T$ can also be bounded by a small constant. Therefore, we have shown $\tau < T$ is unlikely. Then the rest of the analysis is not too hard following the classical analysis.

3.2.3 Reconciliation with existing lower bounds

In this section, we reconcile our convergence results for constant-stepsize GD/SGD in the non-convex setting with existing lower bounds in [Zhang et al., 2019] and [Wang et al., 2022], based on which the authors claim that adaptive methods such as GD/SGD with clipping and Adam are provably faster than non-adaptive GD/SGD. This may seem to contradict our convergence results. In fact, we show that any gain in adaptive methods is at most by constant factors, as GD and SGD already achieve the optimal rates in the non-convex setting.

[Zhang et al., 2019] provides both upper and lower complexity bounds for constant-stepsize GD for (L_0, L_1)-smooth functions, and shows that its complexity is $\mathcal{O}(M\epsilon^{-2})$, where

$$M := \sup\{\|\nabla f(x)\| \mid f(x) \le f(x_0)\}$$

is the supremum gradient norm below the level set of the initial function value. If M is very large, then the $\mathcal{O}(M\epsilon^{-2})$ complexity can be viewed as a negative result, and as evidence that constant-stepsize GD can be slower than GD with gradient clipping, since in the latter case, they obtain the $\mathcal{O}(\epsilon^{-2})$ complexity without M. However, based on our Corollary 2.2.4, their M can be actually bounded by our G, which is a constant. Therefore, the gain in adaptive methods is at most by constant factors.

[Wang et al., 2022] further provides a lower bound which shows non-adaptive GD may diverge for some examples. However, their counter-example does not allow the stepsize to depend on the initial sub-optimality gap. In contrast, our stepsize η depends on the effective smoothness constant L, which depends on the initial sub-optimality gap through G. Therefore, there is no contradiction here either. We should point out that in the practice of training neural networks, the stepsize is usually tuned after fixing the loss function and initialization, so it does depend on the problem instance and initialization.

3.2.4 Lower bound

For (ρ, L_0, L_ρ)-smooth functions with $\rho < 2$, it is easy to verify that the constant G in both Theorem 3.2.2 and Theorem 3.2.3 is a polynomial function of problem-dependent parameters like $L_0, L_\rho, f(x_0) - f^*, \sigma$, etc. In other words, GD and SGD are provably efficient methods in the non-convex setting for $\rho < 2$. In this section, we show that the requirement of $\rho < 2$ is necessary in the non-convex setting with the lower bound for GD in the following Theorem 3.2.4, whose proof is deferred in Appendix B.6. Since SGD reduces to GD when there is no gradient noise, it is also a lower bound for SGD.

Theorem 3.2.4. *Given $L_0, L_2, G_0, \Delta_0 > 0$ satisfying $L_2\Delta_0 \geq 10$, for any $\eta \geq 0$, there exists a $(2, L_0, L_2)$-smooth function f that satisfies Assumptions 1.1 and 1.2, and initial point x_0 that satisfies $\|\nabla f(x_0)\| \leq G_0$ and $f(x_0) - f^* \leq \Delta_0$, such that gradient descent with stepsize η (3.1) either cannot reach a 1-stationary point or takes at least $\exp(L_2\Delta_0/8)/6$ steps to reach a 1-stationary point.*

Chapter 4

Convergence of Adam

In this chapter, we analyze the convergence of Adam under our generalized smoothness condition presented in Chapter 2 in the stochastic non-convex setting.

First, in Section 4.1, we introduce the Adam algorithm and present the key assumptions in our analysis. In particular, we consider the (ρ, L_0, L_ρ)-smoothness condition with $0 \leq \rho < 2$ defined in Definition 3. It is essentially very similar to assuming ℓ-smoothness with a sub-quadratic ℓ, but can better characterize the dependence on ρ in our results. We need to impose a stronger assumption on the gradient noise for Adam compared to that for SGD in Chapter 3 due to additional technical difficulties in the analysis of Adam. In particular, we assume the gradient noise has sub-Gaussian norm, which is stronger than the bounded variance assumption we considered for SGD.

Next, we present the convergence results of Adam in Section 4.2. We can obtain the $\tilde{\mathcal{O}}(\epsilon^{-4})$ gradient complexity as in existing analyses of Adam, under relaxed assumptions. However, since we consider a stronger noise condition than that in the lower bound [Arjevani et al., 2023], it is not clear whether the complexity is optimal or not. Also note that it is not better than that of SGD as the latter is optimal, although Adam outperforms SGD in many deep learning applications like training transformers. We will revisit this gap between theory and practice in Chapter 5.

In Section 4.3, we briefly discuss our analysis approach, bounding gradients along the optimization trajectory. As a warm-up, we show the analysis in detail in the simple deterministic setting to better illustrate our approach. Then we also talk about how to extend the analysis to the stochastic setting.

Finally, in Section 4.4, we proposed a variance-reduced version of Adam which we call VRAdam, inspired by the STORM algorithm in [Cutkosky and Orabona, 2019]. Under stronger assumptions, we are able to obtain an accelerated gradient complexity of $\mathcal{O}(\epsilon^{-3})$. We also present the formal convergence guarantee and discuss how we analyze its convergence in this section.

4.1 Preliminaries

In what follows below, we provide necessary preliminaries for this chapter, including the formal definition of the Adam algorithm and required assumptions for our convergence analysis.

4.1.1 Description of the Adam algorithm

Algorithm 2 Adam

```
1:  **Input:** $\beta, \beta_{\mathrm{sq}}, \eta, \lambda, T, x_{\mathrm{init}}$
2:  **Initialize** $m_0 = v_0 = 0$ and $x_1 = x_{\mathrm{init}}$
3:  **for** $t = 1, \cdots, T$ **do**
4:      Draw a new sample $\xi_t$ and perform the following updates
5:      $m_t = (1-\beta)m_{t-1} + \beta\nabla f(x_t, \xi_t)$
6:      $v_t = (1-\beta_{\mathrm{sq}})v_{t-1} + \beta_{\mathrm{sq}}(\nabla f(x_t, \xi_t))^2$
7:      $\hat{m}_t = \frac{m_t}{1-(1-\beta)^t}$
8:      $\hat{v}_t = \frac{v_t}{1-(1-\beta_{\mathrm{sq}})^t}$
9:      $x_{t+1} = x_t - \frac{\eta}{\sqrt{\hat{v}_t}+\lambda} \odot \hat{m}_t$
10: **end for**
```

The formal definition of Adam proposed in [Kingma and Ba, 2014] is shown in Algorithm 2. Lines 5–9 describe the update rule of iterates $\{x_t\}_{1\le t\le T}$ where all the operators are coordinate-wise. Lines 5–6 are the updates for the first and second order momentum, m_t and v_t, respectively. In Lines 7–8, they are re-scaled to $\hat{m}_t$ and $\hat{v}_t$ in order to correct the initialization bias due to setting $m_0 = v_0 = 0$. Then the iterate is updated by $x_{t+1} = x_t - h_t \odot \hat{m}_t$ where $h_t = \eta/(\sqrt{\hat{v}_t} + \lambda)$ is the adaptive stepsize vector for some parameters η and λ. Note that the algorithm starts at time $t = 1$ instead of $t = 0$ as in the classial methods in Chapter 3 for convenience.

The bias correction steps in Lines 7–8 make Algorithm 2 a bit complicated for analysis. In the following proposition, we provide an equivalent yet simpler update rule of Adam.

Proposition 4.1.1. *Denote* $\alpha_t = \frac{\beta}{1-(1-\beta)^t}$ *and* $\alpha_t^{\mathrm{sq}} = \frac{\beta_{\mathrm{sq}}}{1-(1-\beta_{\mathrm{sq}})^t}$. *Then the update rule in Algorithm 2 is equivalent to*

$$\begin{aligned} \hat{m}_t &= (1-\alpha_t)\hat{m}_{t-1} + \alpha_t \nabla f(x_t, \xi_t), \\ \hat{v}_t &= (1-\alpha_t^{\mathrm{sq}})\hat{v}_{t-1} + \alpha_t^{\mathrm{sq}} (\nabla f(x_t, \xi_t))^2, \\ x_{t+1} &= x_t - \frac{\eta}{\sqrt{\hat{v}_t} + \lambda} \odot \hat{m}_t, \end{aligned} \tag{4.1}$$

where initially we set $\hat{m}_1 = \nabla f(x_1, \xi_1)$ *and* $\hat{v}_1 = (\nabla f(x_1, \xi_1))^2$. *Note that since* $1 - \alpha_1 = 1 - \alpha_1^{\mathrm{sq}} = 0$, *there is no need to define* $\hat{m}_0$ *and* $\hat{v}_0$.

Proof of Proposition 4.1.1. Denote $Z_t = 1-(1-\beta)^t$. Then we know $\alpha_t = \beta/Z_t$ and $m_t = Z_t \hat{m}_t$. By the momentum update rule in Algorithm 2, we have

$$Z_t \hat{m}_t = (1-\beta) Z_{t-1} \hat{m}_{t-1} + \beta \nabla f(x_t, \xi_t).$$

Note that Z_t satisfies the following property

$$(1-\beta) Z_{t-1} = 1 - \beta - (1-\beta)^t = Z_t - \beta.$$

Then we have

$$\begin{aligned}\hat{m}_t =& \frac{Z_t - \beta}{Z_t} \cdot \hat{m}_{t-1} + \frac{\beta}{Z_t} \cdot \nabla f(x_t, \xi_t) \\ =& (1 - \alpha_t)\hat{m}_{t-1} + \alpha_t \nabla f(x_t, \xi_t).\end{aligned}$$

Next, we verify the initial condition. By Algorithm 2, since we set $m_0 = 0$, we have $m_1 = \beta \nabla f(x_1, \xi_1)$. Therefore we have $\hat{m}_1 = m_1/Z_1 = \nabla f(x_1, \xi_1)$ since $Z_1 = \beta$. Then the proof is completed by applying the same analysis on v_t and $\hat{v}_t$. □

4.1.2 Assumptions

Next, we state our main assumptions for the analysis of Adam. First, we still require Assumptions 1.1 and 1.2, the two standard assumptions on the objective function. Besides these, the only additional assumption we make regarding the objective function is the ℓ-smoothness condition with a sub-quadratic ℓ. In particular, we consider the (ρ, L_0, L_ρ)-smoothness condition in Definition 3 with $0 \leq \rho < 2$ to explicitly show the dependence on ρ in the analysis. We state the assumption below for convenience.

Assumption 4.1. The objective function f is (ρ, L_0, L_ρ)-smooth with $0 \leq \rho < 2$.

In addition, we consider one of the following two assumptions on the stochastic gradient $\nabla f(x_t, \xi_t)$ in our analysis of Adam.

Assumption 4.2. The gradient noise is centered and almost surely bounded. In particular, for some $\sigma \geq 0$ and all $t \geq 1$,

$$\mathbb{E}_{t-1}[\nabla f(x_t, \xi_t)] = \nabla f(x_t), \quad \|\nabla f(x_t, \xi_t) - \nabla f(x_t)\| \leq \sigma, \ a.s.,$$

where $\mathbb{E}_{t-1}[\,\cdot\,] := \mathbb{E}[\,\cdot\,|\xi_1, \ldots, \xi_{t-1}]$ is the conditional expectation given $\xi_1, \ldots, \xi_{t-1}$.

Assumption 4.3. The gradient noise is centered with sub-Gaussian norm. In particular, for some $R \geq 0$ and all $t \geq 1$,

$$\mathbb{E}_{t-1}[\nabla f(x_t, \xi_t)] = \nabla f(x_t), \quad \mathbb{P}_{t-1}\left(\|\nabla f(x_t, \xi_t) - \nabla f(x_t)\| \geq s\right) \leq 2e^{-\frac{s^2}{2R^2}}, \; \forall s \in \mathbb{R},$$

where $\mathbb{E}_{t-1}[\,\cdot\,] := \mathbb{E}[\,\cdot\,|\xi_1, \ldots, \xi_{t-1}]$ and $\mathbb{P}_{t-1}[\,\cdot\,] := \mathbb{P}[\,\cdot\,|\xi_1, \ldots, \xi_{t-1}]$ are the conditional expectation and probability given $\xi_1, \ldots, \xi_{t-1}$.

Assumption 4.3 is strictly weaker than Assumption 4.2 since an almost surely bounded random variable clearly has sub-Gaussian norm, but it results in a slightly worse convergece rate up to poly-log factors (see Theorems 4.2.1 and 4.2.2). Both of them are stronger than the most standard bounded variance assumption $\mathbb{E}[\|\nabla f(x_t, \xi_t) - \nabla f(x_t)\|^2] \leq \sigma^2$ for some $\sigma \geq 0$, although Assumption 4.2 is actually a common assumption in existing analyses under the (L_0, L_1)-smoothness condition (see e.g. [Zhang et al., 2019, 2020a]). The extension to the bounded variance assumption is challenging and a very interesting future work as it is also the assumption considered in the lower bound [Arjevani et al., 2023]. We suspect that such an extension would be straightforward if we consider a mini-batch version of Algorithm 2 with a batch size of $S = \Omega(\epsilon^{-2})$, since this results in a very small variance of $\mathcal{O}(\epsilon^2)$ and thus essentially reduces the analysis to the deterministic setting. However, for practical Adam with an $\mathcal{O}(1)$ batch size, the extension is challenging and we leave it as a future work.

4.2 Convergence results

In the section, we provide our convergence results for Adam under Assumptions 1.1, 1.2, 4.1, and 4.2 or 4.3. To keep the statements of the theorems concise, we first define several problem-dependent constants. First, we let $\Delta_1 := f(x_1) - f^* < \infty$ be the initial sub-optimality gap. Next, given a large enough constant $G > 0$, with slight notation abuse, we define

$$L := L_0 + L_\rho (2G)^\rho, \quad r := r(G) = G/L \tag{4.2}$$

where L can be viewed as the effective smoothness constant along the trajectory if one can show $\|\nabla f(x_t)\| \leq G$ and $\|x_{t+1} - x_t\| \leq r$ at each step (see Section 4.3 for more detailed discussions). We will also use c_1, c_2 to denote some small enough numerical constants and C_1, C_2 to denote some large enough ones. The formal convergence results under Assumptions 1.1, 1.2, 4.1, and 4.2 are presented in the following theorem, whose proof is deferred in Appendix C.1.

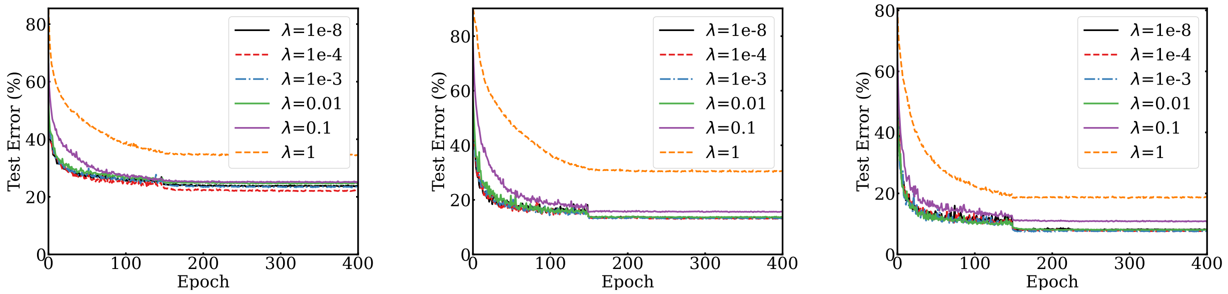
Theorem 4.2.1. *Suppose Assumptions 1.1, 1.2, 4.1, and 4.2 hold. Denote* $\iota := \log(1/\delta)$ *for any* $0 < \delta < 1$*. Let* G *be a constant satisfying* $G \geq \max\left\{2\lambda, 2\sigma, \sqrt{C_1 \Delta_1 L_0}, (C_1 \Delta_1 L_\rho)^{\frac{1}{2-\rho}}\right\}$*. Choose*

$$0 \leq \beta_{\mathrm{sq}} \leq 1, \quad \beta \leq \min\left\{1, \frac{c_1 \lambda \epsilon^2}{\sigma^2 G \sqrt{\iota}}\right\}, \quad \eta \leq c_2 \min\left\{\frac{r\lambda}{G}, \frac{\sigma\lambda\beta}{LG\sqrt{\iota}}, \frac{\lambda^{3/2}\beta}{L\sqrt{G}}\right\}.$$

Let $T = \max\left\{\frac{1}{\beta^2}, \frac{C_2 \Delta_1 G}{\eta \epsilon^2}\right\}$*. Then with probability at least* $1 - \delta$*, we have* $\|\nabla f(x_t)\| \leq G$ *for every* $1 \leq t \leq T$*, and* $\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \leq \epsilon^2$*.*

Note that G, the upper bound of gradients along the trajectory, is a constant that depends on $\lambda, \sigma, L_0, L_\rho$, and the initial sub-optimality gap Δ_1, but not on ϵ. There is no requirement on the second order momentum parameter β_{sq}, although many existing works like [D'efossez et al., 2020, Zhang et al., 2022, Wang et al., 2022] need certain restrictions on it. We choose very small β and η, both of which are $\mathcal{O}(\epsilon^2)$. Therefore, from the choice of T, it is clear that we obtain a gradient complexity of $\mathcal{O}(\epsilon^{-4})$, where we only consider the leading term. We are not clear whether the dependence on ϵ is optimal or not, as the $\Omega(\epsilon^{-4})$ lower bound in [Arjevani et al., 2023] assumes the weaker bounded variance assumption than our Assumpion 4.2. However, it matches the state-of-the-art complexity among existing analyses of Adam.

One limitation of the dependence of our complexity on λ is $\mathcal{O}(\lambda^{-2})$, which might be large since λ is usually small in practice, e.g., the default choice is $\lambda = 10^{-8}$ in the PyTorch implementation. There are some existing analyses on Adam [D'efossez et al., 2020, Zhang et al., 2022, Wang et al., 2022] whose rates do not depend explicitly on λ or only depend on $\log(1/\lambda)$. However, all of them depend on $\mathrm{poly}(d)$, whereas our rate is dimension free. The dimension d is also very large, especially when training transformers, for which Adam is widely used. We believe that independence on d is better than that on λ, because d is fixed given the architecture of the neural network but λ is a hyper-parameter which we have the freedom to tune. In fact, based on our preliminary experimental results on CIFAR-10 shown in Figure 4.1, the performance of Adam is not very sensitive to the choice of λ. Although the default choice of λ is 10^{-8}, increasing it up to 0.01 only makes minor differences.



(a) CNN (b) ResNet-Small (c) ResNet110

Figure 4.1: Test errors of different models trained on CIFAR-10 using the Adam optimizer with $\beta = 0.9, \beta_{\mathrm{sq}} = 0.999, \eta = 0.001$ and different λs. From left to right: (a) a shallow CNN with 6 layers; (b) ResNet-Small with 20 layers; and (c) ResNet110 with 110 layers.

As discussed in Section 4.1.2, we can generalize the bounded gradient noise condition in Assumption 4.2 to the weaker sub-Gaussian noise condition in Assumption 4.3. The following theorem formally shows the convergence result under Assumptions 1.1, 1.2, 4.1, and 4.3, whose proof is deferred in Appendix C.1.5.

Theorem 4.2.2. *Suppose Assumptions 1.1, 1.2, 4.1, and 4.3 hold. Denote* $\iota := \log(2/\delta)$ *and* $\sigma := R\sqrt{2\log(4T/\delta)}$ *for any* $0 < \delta < 1$. *Let* G *be a constant satisfying* $G \geq$

$\max\left\{2\lambda, 2\sigma, \sqrt{C_1\Delta_1 L_0}, (C_1\Delta_1 L_\rho)^{\frac{1}{2-\rho}}\right\}$. *Choose*

$$0 \le \beta_{\mathrm{sq}} \le 1, \quad \beta \le \min\left\{1, \frac{c_1\lambda\epsilon^2}{\sigma^2 G\sqrt{\iota}}\right\}, \quad \eta \le c_2 \min\left\{\frac{r\lambda}{G}, \frac{\sigma\lambda\beta}{LG\sqrt{\iota}}, \quad \frac{\lambda^{3/2}\beta}{L\sqrt{G}}\right\}.$$

Let $T = \max\left\{\frac{1}{\beta^2}, \ \frac{C_2\Delta_1 G}{\eta\epsilon^2}\right\}$. *Then with probability at least* $1-\delta$, *we have* $\|\nabla f(x_t)\| \le G$ *for every* $1 \le t \le T$, *and* $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2 \le \epsilon^2$.

Note that the main difference of Theorem 4.2.2 from Theorem 4.2.1 is that σ is now $\mathcal{O}(\sqrt{\log T})$ instead of a constant. With some standard calculations, one can show that the gradient complexity in Theorem 4.2.2 is bounded by $\mathcal{O}(\epsilon^{-4}\log^p(1/\epsilon))$, where $p = \max\left\{3, \frac{9+2\rho}{4}\right\} < 3.25$.

4.3 Analysis

4.3.1 Bounding the gradients along the optimization trajectory

We want to bound the gradients along the optimization trajectory mainly for two reasons. First, as discussed in Section 1.2, many existing analyses of Adam rely on the assumption of bounded gradients, because unbounded gradient norm leads to unbounded second order momentum $\hat{v}_t$ which implies very small stepsize, and slow convergence. On the other hand, once the gradients are bounded, it is straightforward to control $\hat{v}_t$ as well as the stepsize, and therefore the analysis essentially reduces to the easier one for AdaBound. Second, informally speaking, under Assumption 4.1, bounded gradients also imply bounded Hessians, which essentially reduces the (ρ, L_0, L_ρ)-smoothness to the standard smoothness. See Section 2.2 for more formal discussions.

In this thesis, instead of imposing the strong assumption of globally bounded gradients, we develop a new analysis to show that with high probability, the gradients are always bounded along the trajectory of Adam until convergence. The essential idea can be informally illustrated by the following “circular" reasoning that we will make precise later. On the one hand, if $\|\nabla f(x_t)\| \leq G$ for every $t \geq 1$, it is not hard to show the gradient converges to zero based on our discussions above. On the other hand, we know that a converging sequence must be upper bounded. Therefore there exists some G' such that $\|\nabla f(x_t)\| \leq G'$ for every $t \geq 1$. In other words, the bounded gradient condition implies the convergence result and the convergence result also implies the boundedness condition, forming a circular argument. This circular argument is of course flawed. However, we can break the circularity of reasoning and rigorously prove both the bounded gradient condition and the convergence result using a contradiction argument which we will briefly introduce below.

Define the function $\zeta(u) := \frac{u^2}{2\ell(2u)}$ over $u \geq 0$ where $\ell(u) = L_0 + L_\rho u^\rho$. It is easy to verify that if $\rho < 2$, ζ is increasing and its range is $[0, \infty)$. Therefore, ζ is invertible and ζ^{-1} is also increasing. Then, for any constant $G > 0$, denoting $F = \zeta(G) = \frac{G^2}{2(L_0 + L_\rho (2G)^\rho)}$, Corollary 2.2.4 suggests that if $f(x) - f^* \leq F$, we have $\|\nabla f(x)\| \leq G$. In other words, if $\rho < 2$, the gradient is bounded within any sub-level set, even though the sub-level set could be unbounded. Then, let τ be the first time the sub-optimality gap is strictly greater than F, truncated at $T + 1$, or formally,

$$\tau := \min\{t \mid f(x_t) - f^* > F\} \wedge (T + 1). \tag{4.3}$$

Then at least when $t < \tau$, we have $f(x_t) - f^* \leq F$ and thus $\|\nabla f(x_t)\| \leq G$. Based on our discussions above, it is not hard to analyze the updates before time τ, and one can contruct some Lyapunov function to obtain an upper bound on $f(x_\tau) - f^*$. On the other hand, if $\tau \leq T$, we immediately obtain a lower bound on $f(x_\tau)$, that is $f(x_\tau) - f^* > F$, by the definition of τ in (4.3). If the lower bound is greater than the upper bound, it leads to a contradiction, which shows $\tau = T + 1$, i.e., the sub-optimality gap and the gradient norm are always bounded by F and G respectively before the algorithm terminates.

Before concluding this part, we want to note that our analysis relies on the inequalities in Lemma 2.2.1 for (ρ, L_0, L_ρ)-smooth functions. Since Lemma 2.2.1 is a local condition, we need to make sure the udpate $\|x_{t+1} - x_t\|$ is small enough before applying it. Fortunately, at least before time τ, such a requirement is easy to satisfy for a small enough η, according to the following lemma whose proof is deferred in Appendix C.1.4.

Lemma 4.3.1. *Under Assumption 4.1 and 4.2, if $t < \tau$ and choosing $G \geq \sigma$, we have $\|x_{t+1} - x_t\| \leq \eta D$ where $D := 2G/\lambda$.*

Then as long as $\eta \leq r(G)/D$, we have $\|x_{t+1} - x_t\| \leq r(G)$ which satisfies the requirement in Lemma 2.2.1. Then we can apply the inequalities in it in the same way as the standard smoothness condition. In other words, most classical inequalities derived for standard smooth functions also apply to (ρ, L_0, L_ρ)-smooth functions.

4.3.2 Warm-up: analysis in the deterministic setting

In this section, we consider the simpler deterministic setting where the stochastic gradient $\nabla f(x_t, \xi_t)$ in Algorithm 2 is replaced with the exact gradient $\nabla f(x_t)$. As discussed in Section 4.3.1, the key in our contradiction argument is to obtain both upper and lower bounds on $f(x_\tau) - f^*$. In the following derivations, we focus on illustrating the main idea of our analysis technique and ignore minor proof details. In addition, all of them are under Assumptions 1.1, 1.2, 4.1, and 4.2.

In order to obtain the upper bound, we need the following two lemmas. First, denoting $\epsilon_t := \hat{m}_t - \nabla f(x_t)$, we can obtain the following informal descent lemma for deterministic Adam.

Lemma 4.3.2 (Descent lemma, informal)**.** *For any $t < \tau$, choosing $G \geq \lambda$ and a small enough η,*

$$f(x_{t+1}) - f(x_t) \lessapprox -\frac{\eta}{4G}\left\|\nabla f(x_t)\right\|^2 + \frac{\eta}{2\lambda}\left\|\epsilon_t\right\|^2, \tag{4.4}$$

where "$\lessapprox$" omits less important terms.

Proof Sketch of Lemma 4.3.2. By the definition of τ, for all $t < \tau$, we have $f(x_t) - f^* \leq F$ which implies $\|\nabla f(x_t)\| \leq G$. Then from the update rule (4.1) in Proposition 4.1.1, it is easy to verify $\hat{v}_t \preceq G^2$ since $\hat{v}_t$ is a convex combination of $\{(\nabla f(x_s))^2\}_{s\leq t}$. Let $h_t := \eta/(\sqrt{\hat{v}_t} + \lambda)$ be the stepsize vector and denote $H_t := \operatorname{diag}(h_t)$. We know

$$\frac{\eta}{2G} I \preceq \frac{\eta}{G+\lambda} I \preceq H_t \preceq \frac{\eta}{\lambda} I. \tag{4.5}$$

As discussed in Section 4.3.1, when η is small enough, we can apply Lemma 2.2.1 to obtain

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\lessapprox \left\langle \nabla f(x_t), x_{t+1} - x_t \right\rangle \\ &= - \|\nabla f(x_t)\|_{H_t}^2 - \nabla f(x_t)^\top H_t \epsilon_t \\ &\leq - \frac{1}{2} \|\nabla f(x_t)\|_{H_t}^2 + \frac{1}{2} \|\epsilon_t\|_{H_t}^2 \\ &\leq - \frac{\eta}{4G} \|\nabla f(x_t)\|^2 + \frac{\eta}{2\lambda} \|\epsilon_t\|^2, \end{aligned}$$

where in the first (approximate) inequality we ignore the second order term $\frac{1}{2} L \|x_{t+1} - x_t\|^2 \propto \eta^2$ in Lemma 2.2.1 for small enough η; the equality applies the update rule $x_{t+1} - x_t = -H_t \hat{m}_t = -H_t(\nabla f(x_t) + \epsilon_t)$; in the second inequality we use $2a^\top Ab \leq \|a\|_A^2 + \|b\|_A^2$ for any PSD matrix A and vectors a and b; and the last inequality is due to (4.5). □

Compared with the standard descent lemma for gradient descent, there is an additional term of $\|\epsilon_t\|^2$ in Lemma 4.3.2. In the next lemma, we bound this term recursively.

Lemma 4.3.3 (Informal). *Choosing $\beta = \Theta(\eta G^{\rho+1/2})$, if $t < \tau$, we have*

$$\|\epsilon_{t+1}\|^2 \leq (1 - \beta/4) \|\epsilon_t\|^2 + \frac{\lambda\beta}{16G} \|\nabla f(x_t)\|^2. \tag{4.6}$$

Proof Sketch of Lemma 4.3.3. By the update rule (4.1) in Proposition 4.1.1, we have

$$\epsilon_{t+1} = (1 - \alpha_{t+1}) (\epsilon_t + \nabla f(x_t) - \nabla f(x_{t+1})). \tag{4.7}$$

For small enough η, we can apply Lemma 2.2.1 to get

$$\begin{aligned}\|\nabla f(x_{t+1}) - \nabla f(x_t)\|^2 \leq & L^2 \|x_{t+1} - x_t\|^2 \leq \mathcal{O}(\eta^2 G^{2\rho}) \|\hat{m}_t\|^2 \\ \leq & \mathcal{O}(\eta^2 G^{2\rho})(\|\nabla f(x_t)\|^2 + \|\epsilon_t\|^2), \end{aligned} \tag{4.8}$$

where the second inequality is due to $L = \mathcal{O}(G^\rho)$ and $\|x_{t+1} - x_t\| = \mathcal{O}(\eta)\|\hat{m}_t\|$; and the last inequality uses $\hat{m}_t = \nabla f(x_t) + \epsilon_t$ and Young's inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. Therefore,

$$\begin{aligned}\|\epsilon_{t+1}\|^2 \leq & (1-\alpha_{t+1})(1+\alpha_{t+1}/2)\|\epsilon_t\|^2 + (1+2/\alpha_{t+1})\|\nabla f(x_{t+1}) - \nabla f(x_t)\|^2 \\ \leq & (1-\alpha_{t+1}/2)\|\epsilon_t\|^2 + \mathcal{O}(\eta^2 G^{2\rho}/\alpha_{t+1})\left(\|\nabla f(x_t)\|^2 + \|\epsilon_t\|^2\right) \\ \leq & (1-\beta/4)\|\epsilon_t\|^2 + \frac{\lambda\beta}{16G}\|\nabla f(x_t)\|^2, \end{aligned}$$

where the first inequality uses (4.7) and Young's inequality $\|a+b\|^2 \leq (1+u)\|a\|^2 + (1+1/u)\|b\|^2$ for any $u > 0$; the second inequality uses $(1-\alpha_{t+1})(1+\alpha_{t+1}/2) \leq 1 - \alpha_{t+1}/2$ and (4.8); and in the last inequality we use $\beta \leq \alpha_{t+1}$ and choose $\beta = \Theta(\eta G^{\rho+1/2})$ which implies $\mathcal{O}(\eta^2 G^{2\rho}/\alpha_{t+1}) \leq \frac{\lambda\beta}{16G} \leq \beta/4$. □

Now we can combine the above two lemmas to get the upper bound on $f(x_\tau) - f^*$. Define the function $\Phi_t := f(x_t) - f^* + \frac{2\eta}{\lambda\beta}\|\epsilon_t\|^2$. Note that for any $t < \tau$, (4.4)$+\frac{2\eta}{\lambda\beta}\times$(4.6) gives

$$\Phi_{t+1} - \Phi_t \leq -\frac{\eta}{8G}\|\nabla f(x_t)\|^2. \tag{4.9}$$

The above inequality shows Φ_t is non-increasing and thus a Lyapunov function. Therefore, we have

$$f(x_\tau) - f^* \leq \Phi_\tau \leq \Phi_1 = \Delta_1,$$

where in the last inequality we use $\Phi_1 = f(x_1) - f^* = \Delta_1$ since $\epsilon_1 = \hat{m}_1 - \nabla f(x_1) = 0$ in the deterministic setting.

As discussed in Section 4.3.1, if $\tau \leq T$, we have $F < f(x_\tau) - f^* \leq \Delta_1$. Note that we are able to choose a large enough constant G so that $F = \frac{G^2}{2(L_0 + L_\rho (2G)^\rho)}$ is greater than Δ_1, which leads to a contradiction and shows $\tau = T + 1$. Therefore, (4.9) holds for all $1 \leq t \leq T$. Taking a summation over $1 \leq t \leq T$ and re-arranging terms, we get

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \leq \frac{8G(\Phi_1 - \Phi_{T+1})}{\eta T} \leq \frac{8G\Delta_1}{\eta T} \leq \epsilon^2,$$

if choosing $T \geq \frac{8G\Delta_1}{\eta \epsilon^2}$, i.e., it shows convergence with a gradient complexity of $\mathcal{O}(\epsilon^{-2})$ since both G and η are constants independent of ϵ in the deterministic setting.

4.3.3 Extension to the stochastic setting

In this part, we briefly discuss how to extend the analysis to the more challenging stochastic setting. It becomes harder to obtain an upper bound on $f(x_\tau) - f^*$ because Φ_t is no longer non-increasing due to the existence of noise. In addition, τ defined in (4.3) is now a random variable. Note that all the derivations, such as Lemmas 4.3.2 and 4.3.3, are conditioned on the random event $t < \tau$. Therefore, one can not simply take a total expectation of them to show $\mathbb{E}[\Phi_t]$ is non-increasing.

Fortunately, τ is in fact a stopping time with nice properties. If the noise is almost surely bounded as in Assumption 4.2, by a more careful analysis, we can obtain a high probability upper bound on $f(x_\tau) - f^*$ using concentration inequalities. Then we can still obtain a contradiction and convergence under this high probability event. If the noise has sub-Gaussian norm as in Assumption 4.3, one can change the definition of τ to

$$\tau := \min\{t \mid f(x_t) - f^* > F\} \wedge \min\{t \mid \|\nabla f(x_t) - \nabla f(x_t, \xi_t)\| > \sigma\} \wedge (T + 1)$$

for appropriately chosen F and σ. Then at least when $t < \tau$, the noise is bounded by σ. Hence we can get the same upper bound on $f(x_\tau) - f^*$ as if Assumption 4.2 still holds.

However, when $t \leq T$, the lower bound $f(x_\tau) - f^* > F$ does not necessarily holds, which requires some more careful analyses. The details of the proofs are involved and we defer them in Appendix C.1.

4.4 Varaince-reduced Adam

In this section, we propose a variance-reduced version of Adam (VRAdam). This new algorithm is depicted in Algorithm 3. Its main difference from the original Adam is that in the momentum update rule (Line 6), an additional term of $(1-\beta)\left(\nabla f(x_t, \xi_t) - \nabla f(x_{t-1}, \xi_t)\right)$ is added, inspired by the STORM algorithm [Cutkosky and Orabona, 2019]. This term corrects the bias of m_t so that it is an unbiased estimate of $\nabla f(x_t)$ in the sense of total expectation, i.e., $\mathbb{E}[m_t] = \nabla f(x_t)$. We will also show that it reduces the variance and accelerates the convergence.

Aside from the adaptive stepsize, one major difference between Algorithm 3 and STORM is that our hyper-parameters η and β are fixed constants whereas theirs are decreasing as a function of t. Choosing constant hyper-parameters requires a more accurate estimate at the initialization. That is why we use a mega-batch $\mathcal{S}_1$ to evaluate the gradient at the initial point to initialize m_1 and v_1 (Lines 2–3). In practice, one can also do a full-batch gradient evaluation at initialization. Note that there is no initialization bias for the momentum, so we do not re-scale m_t and only re-scale v_t. We also want to point out that although the initial mega-batch gradient evaluation makes the algorithm a bit harder to implement, constant hyper-parameters are usually easier to tune and more common in training deep neural networks. It should be not hard to extend our analysis to time-decreasing η and β and we leave it as an interesting future work.

In addition to Assumption 1.1 and 1.2, we need to impose the following assumptions which can be viewed as stronger versions of Assumptions 4.1 and 4.2, respectively.

Assumption 4.4. The objective function f and the component function $f(\cdot, \xi)$ for each

Algorithm 3 Variance-Reduced Adam (VRAdam)

```
Input: $\beta, \beta_{\text{sq}}, \eta, \lambda, T, S_1, x_{\text{init}}$
Draw a batch of samples $\mathcal{S}_1$ with size $S_1$ and use them to evaluate the gradient $\nabla f(x_{\text{init}}, \mathcal{S}_1)$.
Initialize $m_1 = \nabla f(x_{\text{init}}, \mathcal{S}_1)$, $v_1 = \beta_{\text{sq}} m_1^2$, and $x_2 = x_{\text{init}} - \frac{\eta m_1}{|m_1|+\lambda}$.
for $t = 2, \cdots, T$ do
    Draw a new sample $\xi_t$ and perform the following updates:
    $m_t = (1-\beta)m_{t-1} + \beta\nabla f(x_t, \xi_t) + (1-\beta)\left(\nabla f(x_t, \xi_t) - \nabla f(x_{t-1}, \xi_t)\right)$
    $v_t = (1-\beta_{\text{sq}})v_{t-1} + \beta_{\text{sq}}(\nabla f(x_t, \xi_t))^2$
    $\hat{v}_t = \frac{v_t}{1-(1-\beta_{\text{sq}})^t}$
    $x_{t+1} = x_t - \frac{\eta}{\sqrt{\hat{v}_t}+\lambda} \odot m_t$
end for
```

fixed ξ are (ρ, L_0, L_ρ)-smooth with $0 \le \rho < 2$.

Assumption 4.5. The random variables $\{\xi_t\}_{1\le t\le T}$ are sampled i.i.d. from some distribution $\mathcal{P}$ such that for any $x \in \mathcal{X}$,

$$\mathbb{E}_{\xi\sim\mathcal{P}}[\nabla f(x,\xi)] = \nabla f(x), \quad \|\nabla f(x,\xi) - \nabla f(x)\| \le \sigma, \ a.s.$$

Remark 4.4.1. Assumption 4.5 is stronger than Assumption 4.2. Assumption 4.2 applies only to the iterates generated by the algorithm, while Assumption 4.5 is a pointwise assumption over all $x \in \mathcal{X}$ and further assumes an i.i.d. nature of the random variables $\{\xi_t\}_{1\le t\le T}$. Also note that, similar to Adam, it is straightforward to generalize the assumption to noise with sub-Gaussian norm as in Assumption 4.3.

4.4.1 Analysis

In this part, we briefly discuss challenges in the analysis of VRAdam. The detailed analysis is deferred in Appendix C.2. Note that Lemma 2.2.1 requires bounded update $\|x_{t+1} - x_t\| \le r(G)$ at each step. For Adam, it is easy to satisfy for a small enough η according to Lemma 4.3.1. However, for VRAdam, obtaining a good enough almost sure bound on the update is challenging even though the gradient noise is bounded. To bypass this difficulty, we directly impose a bound on $\|\nabla f(x_t) - m_t\|$ by changing the definition of the stopping time τ, similar to how we deal with the sub-Gaussian noise condition for Adam. In particular, we define

$$\tau := \min\{t \mid \|\nabla f(x_t)\| > G\} \wedge \min\{t \mid \|\nabla f(x_t) - m_t\| > G\} \wedge (T+1).$$

Then by definition, both $\|\nabla f(x_t)\|$ and $\|\nabla f(x_t) - m_t\|$ are bounded by G before time τ, which directly implies bounded update $\|x_{t+1} - x_t\|$. Of course, the new definition brings new challenges to lower bounding $f(x_\tau) - f^*$, which requires more careful analyses specific to the VRAdam algorithm. Please see Appendix C.2 for the details.

4.4.2 Convergence guarantees for VRAdam

In the section, we provide our main results for convergence of VRAdam under Assumptions 1.1, 1.2, 4.4, and 4.5. We consider the same definitions of problem-dependent constants Δ_1, r, L as those in Section 4.2 to make the statements of theorems concise. Let c be a small enough numerical constant and C be a large enough numerical constant. The formal convergence result is shown in the following theorem.

Theorem 4.4.2. *Suppose Assumptions 1.1, 1.2, 4.4, and 4.5 hold. For any $0 < \delta < 1$, let $G > 0$ be a constant satisfying $G \geq \max\left\{2\lambda, 2\sigma, \sqrt{C\Delta_1 L_0/\delta}, (C\Delta_1 L_\rho/\delta)^{\frac{1}{2-\rho}}\right\}$. Choose $0 \leq \beta_{\mathrm{sq}} \leq 1$ and $\beta = a^2\eta^2$, where $a = 40L\sqrt{G}\lambda^{-3/2}$. Choose*

$$\eta \leq c \cdot \min\left\{\frac{r\lambda}{G}, \quad \frac{\lambda}{L}, \quad \frac{\lambda^2\delta}{\Delta_1 L^2}, \quad \frac{\lambda^2\sqrt{\delta}\epsilon}{\sigma G L}\right\}, \quad T = \frac{64G\Delta_1}{\eta\delta\epsilon^2}, \quad S_1 \geq \frac{1}{2\beta^2 T}.$$

Then with probability at least $1-\delta$, we have $\|\nabla f(x_t)\| \leq G$ for every $1 \leq t \leq T$, and $\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \leq \epsilon^2$.

Note that the choice of G, the upper bound of gradients along the trajectory of VRAdam, is very similar to that in Theorem 4.2.1 for Adam. The only difference is that now it also depends on the failure probability δ. Similar to Theorem 4.2.1, there is no requirement on β_{sq} and we choose a very small $\beta = \mathcal{O}(\epsilon^2)$. However, the variance reduction technique allows us to take a larger stepsize $\eta = \mathcal{O}(\epsilon)$ (compared with $\mathcal{O}(\epsilon^2)$ for Adam) and obtain an accelerated gradient complexity of $\mathcal{O}(\epsilon^{-3})$, where we only consider the leading term. We are not sure whether it is optimal as the $\Omega(\epsilon^{-3})$ lower bound in [Arjevani et al., 2023] assumes the weaker bounded variance condition. However, our result significantly improves upon [Wang and Klabjan, 2022], which considers a variance-reduced version of Adam by combining Adam and SVRG [Johnson and Zhang, 2013] and only obtains asymptotic convergence in the non-convex setting. Similar to Adam, our gradient complexity for VRAdam is dimension free but its dependence on λ is $\mathcal{O}(\lambda^{-2})$. Another limitation is that, the dependence on the failure probability δ is polynomial, worse than the poly-log dependence in Theorem 4.2.1 for Adam.

Chapter 5

Directional smoothness

In this chapter, we propose a directional smoothness condition and analyze the convergence of two special cases of Adam, memoryless Adam (Adam with $\beta = \beta_{\text{sq}} = 1$ in Algorithm 2) and RMSProp (Adam with $\beta = 1$ in Algorithm 2), to better understand why adaptive methods outperform (stochastic) gradient descent for machine learning tasks like training transformers. First, in Section 5.1, we present our main assumption, which essentially assumes that the directional smoothness along the update direction of memoryless Adam is bounded by a constant L_λ, motivated by the empirical observations in [Pan and Li, 2023]. Then we show the convergence of memoryless Adam under this assumption in the deterministic setting in Section 5.2. In particular, we obtain the $\mathcal{O}(L_\lambda \epsilon^{-2})$ gradient complexity for memoryless Adam to converge to ϵ-stationarity points, which is better than the typical gradient complexity of $\mathcal{O}(L\epsilon^{-2})$ of gradient descent when $L_\lambda \ll L$. Under a stronger directional smoothness condition, we are also able to generalize the convergence results to RMSProp and obtain essentially the same gradient complexity. Next, in Section 5.3, we show an example for which $L_\lambda \ll L$ holds and memoryless Adam or RMSProp converges faster than gradient descent if all of them use the stepsizes suggested by theory. Finally, we present some experimental results in Section 5.4 to support our theory.

5.1 Preliminaries

In this section, we provide the formal definition of directional smoothness as well as the assumptions on the objective function f considered in this chapter. First, we consider the following assumption on f.

Assumption 5.1. The objective function $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable and bounded from below, i.e., $f^* := \inf_{x \in \mathbb{R}^d} f(x) > -\infty$.

Note that Assumption 5.1 is a stronger version of Assumptions 1.1 and 1.2 considered in previous chapters, as here we further assume that f is twice differentiable and that its domain is the entire space $\mathbb{R}^d$. Next, we formally define directional smoothness below.

Definition 5 (Directional smoothness)**.** Given a twice differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ and any non-zero vector $u \in \mathbb{R}^d$, the directional smoothness of f along the direction of u at any point $x \in \mathbb{R}^d$ is defined as

$$\ell_x^f(u) := \frac{u^\top \nabla^2 f(x) u}{\|u\|^2},$$

where we usually drop the superscript f when it is clear from the context. We also define $\ell_x^f(0) := \|\nabla^2 f(x)\|$ for convenience.

[Pan and Li, 2023] empirically computed $\ell_{x_t}(x_{t+1} - x_t)$, the directional smoothness along the trajectories of training transformers with various optimizers such as SGD, SignSGD, Adam, etc. They found that adaptive methods tend to have much better (smaller) $\ell_{x_t}(x_{t+1} - x_t)$ compared to SGD, which they believe may intuitively explain why adaptive methods converge faster. In this thesis, we will develop a more rigorous theory based on this observation. In what follows, we first briefly explain why a smaller $\ell_{x_t}(x_{t+1} - x_t)$ is preferred.

For ease of exposition, we denote $\alpha_t = \|x_{t+1} - x_t\|$ and $v_t = \frac{x_{t+1} - x_t}{\|x_{t+1} - x_t\|}$ as the length and direction of the update respectively. Then we can write $x_{t+1} - x_t = \alpha_t \cdot v_t$ and apply Taylor's theorem to get the following informal inequality.

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq \left\langle \nabla f(x_t), x_{t+1} - x_t \right\rangle + \frac{\ell_{x_t}(x_{t+1} - x_t)}{2} \left\| x_{t+1} - x_t \right\|^2 + \mathcal{O}\left(\left\| x_{t+1} - x_t \right\|^3 \right) \quad (5.1) \\ &\approx \alpha_t \left\langle \nabla f(x_t), v_t \right\rangle + \frac{\ell_{x_t}(x_{t+1} - x_t)}{2} \alpha_t^2, \end{aligned}$$

where we have ignored the term $\mathcal{O}\left(\left\| x_{t+1} - x_t \right\|^3 \right)$ whose detailed expression can be found in Lemma D.1.1. Minimizing the RHS over $\alpha_t \geq 0$, assuming $\left\langle \nabla f(x_t), v_t \right\rangle \leq 0$ without loss of generality, we obtain

$$f(x_{t+1}) - f(x_t) \lessapprox -\underbrace{\frac{1}{2\ell_{x_t}(x_{t+1} - x_t)}}_{\text{effective stepsize}} \cdot \underbrace{\left\langle \nabla f(x_t), v_t \right\rangle^2}_{\text{gradient correlation}} . \tag{5.2}$$

The above inequality can be viewed as a descent lemma which bounds the decreased function value at each step. Therefore, to achieve fast convergence, we want to choose a good direction of update v_t so that the RHS is as small as possible. In other words, we want both the terms of effective stepsize and gradient correlation to be large.

If we plug in the expression of $\ell_{x_t}(x_{t+1} - x_t)$ and minimize the RHS over v_t, then we should choose $v_t = (\nabla^2 f(x_t))^{-1} \nabla f(x_t)$, which gives us Newton's method. However, this requires to compute the Hessian $\nabla^2 f(x_t)$, which is usually very expensive for training neural networks. Therefore, we will focus on first-order methods for which v_t only depends on the current and past gradients. For first-order methods, we will see that there is actually a trade-off between the effective stepsize and gradient correlation when choosing v_t.

For example, gradient descent takes the so-called "steepest descent" direction $v_t = -\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}$ which maximizes the gradient correlation term with $\left\langle \nabla f(x_t), v_t \right\rangle^2 = \|\nabla f(x_t)\|^2$. However, it may have a very large $\ell_{x_t}(x_{t+1} - x_t)$ compared to adaptive methods as suggested by the empirical findings in [Pan and Li, 2023]. As a result, it may have a small effective stepsize which leads to slow convergence. On the other hand, the update direction of SignGD is $v_t = -\frac{\operatorname{sign}(\nabla f(x_t))}{\sqrt{d}}$, which may have a smaller $\ell_{x_t}(x_{t+1} - x_t)$ and thus a large effective stepsize.

However, it has a much worse gradient correlation term $\left\langle \nabla f(x_t), v_t \right\rangle^2 = \frac{1}{d} \left\| \nabla f(x_t) \right\|_1^2$. As a result, the convergence rate will depend on the dimension d, which is usually very large for practical transformers. Therefore, it is also hard to obtain a fast convergence rate for SignGD.

To achieve the best of both worlds, we consider an interpolation between the directions of the gradient $\nabla f(x)$ and its coordinate-wise sign $\operatorname{sign}(\nabla f(x))$. Specifically, we define the following soft sign of the gradient

$$u_\lambda(x) := \frac{\nabla f(x)}{|\nabla f(x)| + \lambda}, \tag{5.3}$$

where the operators of absolute value and division are coordinate-wise and $\lambda > 0$ is a small constant. It is an interpolation between the gradient and its sign because one can easily verify that $\lim_{\lambda \to \infty} \frac{u_\lambda(x)}{\|u_\lambda(x)\|} = \frac{\nabla f(x)}{\|\nabla f(x)\|}$ and $\lim_{\lambda \to 0} u_\lambda(x) = \operatorname{sign}(\nabla f(x))$. Therefore, informally speaking, for a very small λ, $u_\lambda(x)$ is close to $\operatorname{sign}(\nabla f(x))$, and thus we call it a soft sign. For this reason, we also define $u_0(x) := \operatorname{sign}(\nabla f(x))$ when $\lambda = 0$ just for convenience. That being said, a non-zero λ is quite essential in our analysis to achieve the best of both worlds and obtain a dimension-free convergence rate. As we will see in Section 5.2, λ could be as small as the level of stationarity ϵ in our analysis.

If viewing $-\eta_t u_\lambda(x_t)$ as the update of an algorithm at time t, where η_t is the stepsize, we obtain

$$x_{t+1} = x_t - \eta_t \cdot \frac{\nabla f(x_t)}{|\nabla f(x_t)| + \lambda},$$

which is essentially the Adam algorithm without memory or momentums, i.e., Algorithm 2 with $\beta = \beta_{\text{sq}} = 1$, in the deterministic setting. Informally speaking, this means memoryless Adam can be viewed as a generalization of SignGD or a good interpolation between GD and SignGD. Note that the update of memoryless Adam is a function of the gradient at the current time and does not depend on the history of the algorithm, which allows us to obtain a condition on the objective function itself, as stated below.

Assumption 5.2. $L_\lambda := \sup_{x\in\mathbb{R}^d} \ell_x(u_\lambda(x)) < \infty$.

We are mostly interested in the scenario where the directional smoothness along $u_\lambda(x)$ is much smaller than that along $\nabla f(x)$ or the standard Lipschitz smoothness constant L, or more formally,

$$L_\lambda \ll L_{\mathrm{g}} \le L \le \infty, \tag{5.4}$$

where we also define

$$L_{\mathrm{g}} := \sup_{x\in\mathbb{R}^d} \ell_x(\nabla f(x)), \quad L := \sup_{x\in\mathbb{R}^d} \left\|\nabla^2 f(x)\right\|.$$

We will present our empirical results in Section 5.4 to show that (5.4) can characterize certain properties of the loss landscape of training transformers. Note that we only assume L_λ is finite as in Assumption 5.2, whereas L_{g} and L could be potentially very large or even infinite.

In addition to Assumption 5.2, we also need the following condition which essentially bounds the third-order derivative of the objective function.

Assumption 5.3. $\|\nabla^2 f(x) - \nabla^2 f(y)\| \le M\,\|x-y\|$ for all $x, y \in \mathbb{R}^d$.

With Assumption 5.3, we are able to derive the formal expression of the Taylor's expansion in (5.1), as will be shown in Lemma D.1.1.

5.2 Convergence results

In this section, we formally show the convergence results of memoryless Adam and RMSProp under the assumptions in Section 5.1. We will compare the obtained rates with that of gradient descent to show the benefit of adaptivity. As the analysis is challenging, we only consider the deterministic setting and leave the extension to the stochastic setting as an interesting future direction. Finally, we also briefly discuss some possible extensions of our results.

5.2.1 Memoryless Adam

We first analyze the convergence of memoryless Adam formally defined by the following update rule. Note that here we assume the iteration starts at $t = 1$ instead of $t = 0$ as in Chapter 3 for convenience.

$$x_{t+1} = x_t - \eta_t u_\lambda(x_t) = x_t - \eta_t \cdot \frac{\nabla f(x_t)}{|\nabla f(x_t)| + \lambda}, \quad \forall t \geq 1. \tag{5.5}$$

In the following theorem, we present the convergence result of the above method with a constant stepsize $\eta_t \equiv \eta$ under our directional smoothness assumptions. Its detailed proof is deferred to Appendix D.2.

Theorem 5.2.1. *Suppose Assumptions 5.1, 5.2, and 5.3 hold. Denote $\Delta_1 := f(x_1) - f^*$. For any $\epsilon > 0$, choose $\eta_t \equiv \eta = \frac{\lambda}{2\max\left\{L_\lambda, M^{2/3}\Delta_1^{1/3}\right\}}$ and*

$$T \geq \frac{2\Delta_1(\epsilon + \lambda)}{\eta\epsilon^2} = 4\max\left\{L_\lambda\Delta_1, M^{2/3}\Delta_1^{4/3}\right\} \cdot \frac{(\epsilon + \lambda)}{\lambda\epsilon^2}.$$

Then the iterates generated by (5.5) satisfy $\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\| \leq \epsilon$.

Theorem 5.2.1 shows that constant-stepsize memoryless Adam always converges to an ϵ-stationary point for any $\epsilon > 0$. If choosing $\lambda \geq \epsilon$, the gradient or iteration complexity is $\mathcal{O}\left(\max\left\{L_\lambda\Delta_1, M^{2/3}\Delta_1^{4/3}\right\}\epsilon^{-2}\right)$, which has the same ϵ dependency as gradient descent. For the dependency on problem-dependent constants, when M or Δ_1 is very large, $M^{2/3}\Delta_1^{4/3}$ will dominate $L_\lambda\Delta_1$, and hence it can not show the benefit of a small L_λ in this case. Fortunately, we will show in the following theorem that the iteration complexity can be further improved if allowing η_t to depend on the gradient $\nabla f(x_t)$. We also defer its proof to Appendix D.2.

Theorem 5.2.2. *Suppose Assumptions 5.1, 5.2, and 5.3 hold. Denote* $\Delta_1 := f(x_1) - f^*$*. For any* $\epsilon > 0$*, choose* $\eta_t = \min\left\{\frac{\lambda}{2L_\lambda}, \frac{\sqrt{\lambda}L_\lambda}{M\|\nabla f(x_t)\|_{H_t}}\right\}$*, where* $H_t = \operatorname{diag}\left(\frac{1}{|\nabla f(x_t)|+\lambda}\right)$*, and*

$$T \geq \frac{4L_\lambda\Delta_1(\epsilon+\lambda)}{\lambda\epsilon^2} + \frac{M^2\Delta_1}{L_\lambda^3}.$$

Then the iterates generated by (5.5) *satisfy* $\min_{t\leq T}\|\nabla f(x_t)\| \leq \epsilon$*.*

For any $\lambda \geq \epsilon$, Theorem 5.2.2 gives an iteration complexity of $\mathcal{O}\left(\frac{L_\lambda\Delta_1}{\epsilon^2} + \frac{M^2\Delta_1}{L_\lambda^3}\right)$. When $\epsilon \leq \frac{L_\lambda^2}{M}$ is small enough, one can ignore the non-leading constant term and the complexity becomes $\mathcal{O}\left(L_\lambda\Delta_1\epsilon^{-2}\right)$. Recall that the typical gradient complexity of gradient descent is $\mathcal{O}\left(L\Delta_1\epsilon^{-2}\right)$ which could be potentially improved to $\mathcal{O}\left(L_{\mathrm{g}}\Delta_1\epsilon^{-2}\right)$. We can see that memoryless Adam has a faster convergence rate if $L_\lambda \ll L_{\mathrm{g}} \leq L$. To support this statement, we will present an example function in Section 5.3, for which $L_\lambda \ll L_{\mathrm{g}} \leq L$ does hold and memoryless Adam can achieve a faster rate.

5.2.2 RMSProp

In this section, we will generalize the convergence result to RMSProp (Adam with $\beta = 1$ as in Algorithm 2). For completeness, we present the definition of RMSProp in the deterministic setting below in Algorithm 4. Note that here we slightly abuse the notation and use β to denote the $1 - \beta_{\mathrm{sq}}$ in Algorithm 2 for simplicity.

Algorithm 4 RMSProp

```
Input: $\beta, \lambda, T, x_{\mathrm{init}}, \{\eta_t\}_{1\leq t\leq T}$
Initialize $v_0 = 0$ and $x_1 = x_{\mathrm{init}}$
for $t = 1, \cdots, T$ do
    $v_t = \beta v_{t-1} + (1-\beta)(\nabla f(x_t))^2$
    $\hat{v}_t = \frac{v_t}{1-\beta^t}$
    $x_{t+1} = x_t - \eta_t \cdot \frac{\nabla f(x_t)}{\sqrt{\hat{v}_t}+\lambda}$
end for
```

The key idea in the extension to RMSProp is to show that $\sqrt{\hat{v}_t}$ is close to $|\nabla f(x_t)|$, which means that the updates of RMSProp and memoryless Adam are close, i.e., $\hat{u}_t \approx u_t$ where we define $\hat{u}_t := \frac{\nabla f(x_t)}{\sqrt{\hat{v}_t}+\lambda}$ and $u_t := \frac{\nabla f(x_t)}{|\nabla f(x_t)|+\lambda}$. However, Assumption 5.2 is a very weak condition, which bounds the directional smoothness only along a single direction at each point. As a result, we are not able to bound the directional smoothness along $\hat{u}_t$ no matter how close it is to u_t unless they exactly align. Therefore we need the following stronger condition which essentially bounds the directional smoothness along all directions near u_t.

Assumption 5.4. There exist numerical constants $r, R > 1$ such that for any vector v satisfying $\frac{1}{r}v \preceq u_\lambda(x) \preceq rv$, we have $\ell_x(v) \leq RL_\lambda$.

Under Assumption 5.4, we can prove the convergence of RMSProp as formally shown in the following theorem, whose proof is deferred to Appendix D.3. Note that here we only consider the convergence result for RMSProp with time-varying stepsizes, although one can also show the convergence of it with a constant stepsize, similar to that of memoryless Adam.

Theorem 5.2.3. *Suppose Assumptions 5.1, 5.3, and 5.4 hold. Denote* $\Delta_1 := f(x_1) - f^*$*. For any* $\epsilon > 0$*, choose* $\beta \leq \frac{1}{4}(1-1/r)^2$*,* $\eta_t = \min\left\{\frac{\lambda}{2RL_\lambda}, \frac{R\sqrt{\lambda}L_\lambda}{\sqrt{r}M\|g_t\|_{H_t}}, \frac{(1-1/r)\lambda^2}{3r^2RL_\lambda\|g_t\|_{H_t}^2}\right\}$*, and*

$$T \geq \frac{4rRL_\lambda\Delta_1(\epsilon+\lambda)}{\lambda\epsilon^2} + \frac{6r^4RL_\lambda\Delta_1}{(r-1)\lambda^2} + \frac{r^2M^2\Delta_1}{R^3L_\lambda^3}.$$

Then the iterates generated by Algorithm 4 satisfy $\min_{t\leq T}\|\nabla f(x_t)\| \leq \epsilon$.

Ignoring the dependency on numerical constants r and R, we can see that Theorem 5.2.3 gives an iteration complexity of $\mathcal{O}\left(\frac{L_\lambda\Delta_1}{\epsilon^2} + \frac{M^2\Delta_1}{L_\lambda^3}\right)$ when $\lambda \geq \epsilon$, which is the same as that of memoryless Adam. Therefore, for a small engouh ϵ, it is also better than the typical $\mathcal{O}(L_{\mathrm{g}}\epsilon^{-2})$ complexity of gradient when $L_\lambda \leq L_{\mathrm{g}}$.

The key idea in our analysis is to bound the following quantity at each time t.

$$E_t := \left\|\frac{\hat{u}_t}{u_t}\right\|_\infty = \left\|\frac{|g_t|+\lambda}{\sqrt{\hat{v}_t}+\lambda}\right\|_\infty.$$

If we can show $1/r \le E_t \le r$ for all $t \ge 1$, then we can apply Assumption 5.4 to bound the directional smoothness along $\hat{u}_t$, and essentially reduce the analysis to that of memoryless Adam. However, bounding E_t is challenging and requires very careful analyses, as we need a very tight and dimension-independent bound on each coordinate of $\hat{v}_t - g_t^2$. Please see Lemma D.3.2 and its proof for the detailed analysis.

5.2.3 Potential extensions

In this part, we briefly discuss some potential extensions of our convergence results. First, we want to note that $u_\lambda(x)$ defined in (5.3) is not the only way of interpolation between $\nabla f(x)$ and $\operatorname{sign}(\nabla f(x))$. For example, if we define the coordinate-wise gradient clipping as

$$v_\lambda(x) := \operatorname{sign}(\nabla f(x)) \min\{\left|\nabla f(x)\right|, \lambda\}$$

where the operators are all coordinate-wise, then it is straightforward to verify that

$$\frac{1}{2} u_\lambda(x) \preceq \frac{1}{\lambda} v_\lambda(x) \preceq 2 u_\lambda(x).$$

Therefore, u_λ and v_λ are essentially equivalent up to a constant factor, meaning that memoryless Adam and gradient descent with coordinate-wise clipping are also equivalent. In fact, [Pan and Li, 2023] uses coordinate-wise gradient clipping as a universal technique to improve the directional smoothness of various algorithms. Then, if assuming $\ell_x(v_\lambda(x))$ is globally bounded by some constant $\bar{L}_\lambda$, following our analysis for memoryless Adam, one can show that gradient descent with coordinate-wise clipping also converges with a gradient complexity of $\mathcal{O}(\bar{L}_\lambda \epsilon^{-2})$, which is also better than that of gradient descent without clipping when $\bar{L}_\lambda \ll L_{\mathrm{g}} \le L$. Actually, if Assumption 5.4 holds with $r \ge 2$, it directly implies $\bar{L}_\lambda \le R L_\lambda$.

Also note that for any $\lambda_1 \geq \lambda_2 > 0$, we have

$$\frac{\lambda_2}{\lambda_1} u_{\lambda_1}(x) \preceq u_{\lambda_2}(x) \preceq \frac{\lambda_1}{\lambda_2} u_{\lambda_1}(x),$$

which means u_{λ_1} and u_{λ_2} are also essentially equivalent. Therefore, the parameter λ in the algorithm could be different than that in our assumption. Specifically, if $\frac{\lambda_1}{\lambda_2} \leq r$, then one can use one of them in Assumption 5.4 and the other in memoryless Adam or RMSProp.

We did not show the convergence of Adam defined in Algorithm 2 in this chapter. It is possible to extend our result to Adam by combining our analyses in both this chapter and Chapter 4. However, the analysis will be very messy and we leave it as an interesting future work. It would also be interesting to extend our analysis to the stochastic setting. The main challenge is to bound $\frac{|\nabla f(x_t)|+\lambda}{|\nabla f(x_t,\xi_t)|+\lambda}$ or something similar in order to apply Assumption 5.4, where $\nabla f(x_t, \xi_t)$ is the stochastic gradient. One possible way to overcome this challenge is to apply the stopping time analysis as in Chapter 3 or 4. We also leave it as an interesting future direction.

5.3 Example

In this section, we provide a simple example objective function satisfying all of Assumptions 5.1, 5.2, 5.3, and 5.4. We will also show that for this example, $L_\lambda \ll L_{\mathrm{g}} \leq L$, and that memoryless Adam or RMSProp converges faster than gradient descent.

Before presenting the example, we first define the auxiliary function $\phi : \mathbb{R} \to \mathbb{R}$ as follows.

$$\phi(z) := \begin{cases} e^z & \text{if } z \leq 0, \\ \frac{1}{2}z^2 + z + 1 & \text{if } z > 0. \end{cases}$$

It is easy to verify that ϕ is twice continuously differentiable. Then the example objective function $f : \mathbb{R}^d \to \mathbb{R}$ with $d \geq 2$ is defined as

$$f(x) := \frac{1}{\alpha}\phi(\alpha \cdot w(x)), \text{ where } \alpha > 0 \text{ and } w(x) := x_{[1]} + \frac{1}{d-1}\left(x_{[2]} + \cdots + x_{[d]}\right). \tag{5.6}$$

First, in the following lemma, we show that the example satisfies all of our assumptions and also bound the constants $L, L_{\mathrm{g}}, M, L_\lambda, r, R$ for it. The proof is deferred to Appendix D.4.

Lemma 5.3.1. *For any $\lambda \leq \frac{1}{d-1}$, the function defined in (5.6) satisfies Assumptions 5.1, 5.2, 5.3, and 5.4 with the following constants.*

1. $L = L_{\mathrm{g}} = \frac{\alpha d}{d-1}$.

2. $M \leq 3\alpha^2$.

3. $L_\lambda \leq 8\alpha \max\left\{\lambda\sqrt{d-1}, \frac{2}{d}\right\}$.

4. *Any $R, r > 1$ satifying $R \geq r^4$.*

Next, we will show in the following theorem that, with appropriate choices of λ, α, d, the example satisfies $L_\lambda \ll L_{\mathrm{g}} \leq L$ and memoryless Adam converges faster than gradient descent.

Theorem 5.3.2. *For any given constants $C \geq c > 0$, there exist some λ, α, d such that the function defined in (5.6) satisfies $L = L_{\mathrm{g}} = C$ and $L_\lambda \leq c$. Let T_{gd} and T_{ma} be the minimum number of iterations required for gradient descent with $\eta_t = \frac{1}{L}$ and memoryless Adam with $\eta_t = \frac{\lambda}{L_\lambda}$ to achieve an ϵ-sub-optimal point respectively. Then for any small enough $\epsilon > 0$ and initial point x_1 satifying $w(x_1) \leq 0$, we always have $T_{\mathrm{gd}}/T_{\mathrm{ma}} = \Omega(C/c)$.*

The proof of the above theorem is deferred to Appendix D.4. Since Theorem 5.3.2 holds for arbitrary C and c, if we choose $c \ll C$, then it shows that memoryless Adam can convergence much faster than gradient descent on this example and that the ratio between their required numbers of iterations is exactly lower bounded by $\Omega(C/c) = \Omega(L/L_\lambda)$.

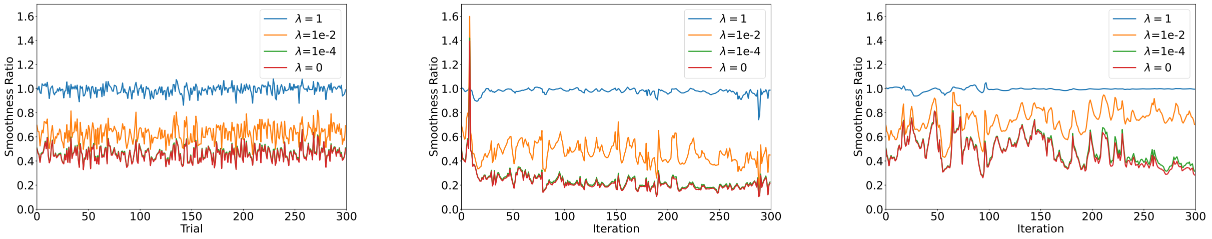
One limitation of the above theorem is that it only works for fixed stepsize choices for both methods. For gradient descent, $\frac{1}{L} = \frac{1}{L_{\mathrm{g}}}$ is the stepsize suggested by classical analyses of gradient descent. For memoryless Adam, $\frac{\lambda}{L_\lambda}$ is the constant term of the stepsize choice in our Theorem 5.2.2. Note that for this simple example, there is no need to use the complex stepsize choice in our theorem. However, one can actually easily show that the stepsize choice in Theorem 5.2.2 results in essentially the same bound on T_{ma} following a similar analysis as in the proof of Theorem 5.3.2 in Appendix D.4. In fact, it is also straightforward to show that RMSProp with the parameter choices in Theorem 5.2.3 also converges faster than gradient descent with $T_{\text{gd}}/T_{\text{rmsprop}} = \Omega(C/c)$.

It is challenging to show a complexity lower bound for gradient descent with arbitrary stepsizes. Therefore, rigorously speaking, Theorem 5.3.2 does not totally rule out the possibility that gradient descent may converge much faster than its typical rate for functions with a small L_λ. We leave it as an interesting future work to derive a more rigorous lower bound for gradient descent.

5.4 Experimental results

In this section, we provide some empirical results from our experiments on simple transformers to support our theory. To show that (5.4) may characterize certain properties of the loss landscape of training transformers, we will empirically compare $\ell_x(u_\lambda(x))$ with $\ell_x(\nabla f(x))$. In particular, as we are not able to enumerate all $x \in \mathbb{R}^d$, we will compute the smoothness ratio $r_\lambda(x) := \ell_x(u_\lambda(x))/\ell_x(\nabla f(x))$ where x is either from the initialization or the optimization trajectories of certain algorithms. Note that the algorithms are just used to generate a sequence of points to evaluate ℓ_λ at. The smoothness ratio function r_λ is algorithm-independent and only depends on the objective function itself.

We mainly consider two optimization problems in this section. The first problem is the training of linear transformers on random instances of linear regression, a recently proposed model for understanding in-context learning. For this problem, we follow the setting and parameter choices in [Ahn et al., 2023]. The second problem we consider is nanoGPT on character-level Shakespeare data[1].

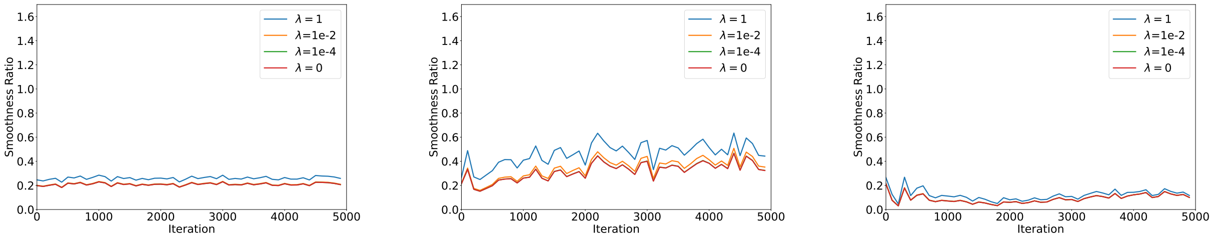


(a) Initialization

(b) SGD trajectory

(c) Adam trajectory

Figure 5.1: Smoothness ratio $r_\lambda(x)$ for the loss of linear transformer on a random instance of linear regression for different values of λ, where x is from the initialization with different seeds or the trajectories generated by SGD or Adam.



(a) Initialization

(b) SGD trajectory

(c) AdamW trajectory

Figure 5.2: Smoothness ratio $r_\lambda(x)$ for the loss of nanoGPT on character-level Shakespeare data for different values of λ, where x is from the initialization with different seeds or the trajectories generated by SGD or AdamW.

Our results on the smoothness ratio r_λ with different values of λ for both problems are shown in Figures 5.1 and 5.2. Recall that we defined $u_0(x) := \text{sign}(\nabla f(x))$. We can see that when λ decreases to zero, the smoothness ratio also decreases, and $\ell_x(u_\lambda(x))$ essentially goes to $\ell_x(\text{sign}(\nabla f(x)))$, consistent with our intuition that $u_\lambda(x)$ is close to $\text{sign}(\nabla f(x))$ for a small λ. When λ is small, $\ell_x((\nabla f(x)))$ is a couple of times larger than $\ell_x(u_\lambda(x))$ for all the random initialized points and most points from the trajectories of both SGD and Adam(W), which supports our conjecture that $L_\lambda \ll L_\text{g}$.

[1] https://github.com/karpathy/nanoGPT

Chapter 6

Conclusion and future work

6.1 Summary

We will conclude this thesis by first summarizing what we have discussed. We investigated the smoothness condition and adaptivity in nonlinear optimization to gain a better understanding of the training behaviors in machine learning applications that can not be well explained by classical optimization theory. In particular, the classical Lipschitz smoothness condition is not only far from being satisfied by the loss function in machine learning applications, but also can not explain why adaptive methods like Adam outperform stochastic gradient descent in tasks like training transformers. To bridge this gap, we proposed a generalized ℓ-smoothness condition and a more fine-grained directional smoothness condition, both motivated by language model experiments, and analyzed the convergence of classical and adaptive methods under such conditions.

First, in Chapter 2, we proposed a generalized ℓ-smoothness condition based on existing works and empirical observations from language model experiments. We provided two equivalent definitions of it for the convenience of both verification and application. To show how general this condition is compared to the standard smoothness condition and the recently proposed (L_0, L_1)-smoothness condition, we provided various examples and theoretical justifications. We also showed some useful properties for ℓ-smooth functions and discussed how they are helpful in the convergence analyses.

Then we developed a new approach for the convergence analysis under our generalized ℓ-smoothness condition. The key idea of our approach is to bound the gradient along the optimization trajectory, which also bounds the Hessian based on the ℓ-smoothness condition and thus essentially reduces the analysis to that under standard smoothness. In Chapter 3, we applied this approach to classical methods including gradient descent, stochastic gradient descent, and Nesterov's accelerated gradient method in convex and/or non-convex settings. For all of them, we achieve the classical convergence rates under the generalized smoothness condition. In Chapter 4, we also applied this approach to Adam and obtained improved results compared to previous works. In particular, we did not assume globally bounded gradients as in some previous works, but use our approach to show gradients are bounded along the trajectory with high probability. In addition, we considered the more general ℓ-smoothness condition for Adam. With this new approach, we also proposed a variance-reduce version of Adam and showed an accelerated convergence rate.

Although the generalized ℓ-smoothness condition is closer to the machine learning practice, we are not able to explain why adaptive methods outperform SGD for certain tasks like training transformers based on this condition. In Chapter 5, to better understand why and when adaptivity accelerates training, we proposed a more fine-grained directional smoothness condition. Instead of assuming the Lipschitzness of the gradient, we only assume directional smoothness around the direction of a soft sign of the gradient, motivated by empirical results from both our experiments and those in previous works. Under this condition, we are able to show the convergence of memoryless Adam and RMSProp in the deterministic setting and obtain convergence rates better than the typical rate of gradient descent. We also provide an example and experimental results to support our theory.

We hope the theoretical and empirical results in this thesis could provide researchers with new ideas and inspire them to gain a better understanding of the training behaviors in real-world machine learning problems and to design more efficient and robust optimizers. However, we know that this work is not perfect, and will discuss some future directions in the next section.

6.2 Future works

In this section, we briefly discuss some possible next steps and directions for future research.

Improving the results for NAG in Chapter 3. First, in Theorem 3.2.4, we provided a lower bound to justify the necessity of requiring a sub-quadratic ℓ for the convergence of constant-stepsize GD on ℓ-smoothness functions. However, it is not clear if such a requirement is also necessary for NAG in the convex setting. It would be interesting if one could either develop a lower bound for NAG or relax this requirement. In addition, we have mentioned below Theorem 3.1.4 that the stepsize we choose for NAG might be too small due to technical difficulties, which may result in a worse dependency on problem-dependent constants. We also leave possible improvement as an interesting future work.

Relaxing the noise condition for Adam in Chapter 4. In Assumption 4.2 or 4.3, we assumed the gradient noise is bounded or sub-Gaussian for our convergence analysis of Adam. However, these conditions are stronger than the bounded variance assumption for SGD in Assumption 3.2, where the latter is also the assumption considered in the lower bound [Arjevani et al., 2023]. As a result, it is not clear whether the $\mathcal{O}(\epsilon^{-4})$ complexity we have obtained is optimal or not. It would be interesting to see if one can relax the noise condition. Note that the concurrent work [Wang et al., 2023] obtained the $\mathcal{O}(\epsilon^{-4})$ complexity under the bounded variance assumption. But their rate is dimension dependent and they consider the standard smoothness condition. That being said, their analysis may be potentially helpful in improving our results.

Potential applications of our technique for bounding gradients along the trajectory. Another interesting future direction is to see if the technique developed in this thesis for bounding gradients in the optimization trajectory can be generalized to improve the convergence results for other optimization problems and algorithms. We believe that with this technique, one can generalize most existing optimization works that assume standard smoothness to those with our generalized ℓ-smoothness functions. In fact, we believe it is also possible to apply this technique to even more general problems to obtain improved convergence results, so long as the function class is well-behaved and the algorithm is efficient enough so that $f(x_\tau) - f^*$ can be well bounded for some appropriately defined stopping time τ.

Limitations of our results in Chapter 5. There are some limitations in our results on directional smoothness in Chapter 5. First, we only analyzed the convergence of memoryless Adam and RMSProp, both of which are simplified versions of Adam in Algorithm 2. It would be very interesting if one can generalize our results to Adam, by e.g. combining our analysis in both Chapters 4 and 5. In addition, we only considered the deterministic setting, as the coordinate-wise condition in Assumption 5.4 is much harder to be satisfied by the algorithm updates when there is noise. One possible way to tackle this challenge is to apply our stopping time analysis developed in Chapters 3 and 4. Finally, we are only able to obtain a lower bound on the ratio between the gradient complexities of gradient descent with the theoretically suggested stepsize and memoryless Adam in Theorem 5.3.2. It would be interesting to see if we can also get a stepsize-independent lower bound to make the comparison more rigorous.

Other explanations on why adaptivity helps. In Chapter 5, we provided an explanation on why adaptive methods outperform SGD for training transformers, motivated by the empirical observations in [Pan and Li, 2023]. However, there are definitely other potential explanations for researchers to explore. For example, we have discussed some of them in Section 1.2, such as heavy-tailed noise distribution or class imbalance, condition number along the trajectory, and heterogeneity of the Hessian spectrum. In fact, the last one may be related to directional smoothness, as the Hessian of the example we considered in Section 5.3 is indeed highly heterogeneous. It would interesting if one could find other interesting empirical observations, or develop more rigorous theoretical understandings based on existing or their own empirical observations.

Appendix A

Proofs for Chapter 2

In this chapter, we provide the proofs of propositions and lemmas related to the generalized ℓ-smoothness condition presented in Chapter 2. First, in Appendix A.1, we justify the examples we presented in Section 2.3. Next, we provide the detailed proof of Proposition 2.1.2 in Appendix A.2. Finally, we provide the proofs of the useful properties of generalized smoothness in Appendix A.3, including Lemma 2.2.1, Lemma 2.2.3, and Corollary 2.2.4 stated in Section 2.2.

A.1 Justification of examples in Section 2.3

In this section, we justify the univariate examples of (ρ, L_0, L_ρ)-smooth functions listed in Table 2.1 and also provide the proof of Propositions 2.3.1.

First, it is well known that all quadratic functions have bounded Hessian and are Lipschitz smooth, corresponding to $\rho = 0$. Next, [Zhang et al., 2019, Lemma 2] shows that any univariate polynomial is (L_0, L_1)-smooth, corresponding to $\rho = 1$. Then, regarding the exponential function $f(x) = a^x$ where $a > 1$, we have $f'(x) = \log(a)a^x$ and $f''(x) = \log(a)^2 a^x = \log(a)f'(x)$, which implies f is $(1, 0, \log(a))$-smooth. Similarly, by standard calculations, it is straight forward to verify that logarithmic functions and x^p, $p \neq 1$ are also (ρ, L_0, L_ρ)-smooth with $\rho = 2$ and $\rho = \frac{p-2}{p-1}$ respectively. So far we have justified all the examples in Table 2.1 except double exponential functions $a^{(b^x)}$ and rational functions, which will be justified rigorously by the two propositions below.

First, for double exponential functions in the form of $f(x) = a^{(b^x)}$ where $a, b > 1$, we have the following proposition, which shows f is (ρ, L_0, L_ρ)-smooth for any $\rho > 1$.

Proposition A.1.1. *For any $\rho > 1$, the double exponential function $f(x) = a^{(b^x)}$, where $a, b > 1$, is (ρ, L_0, L_ρ)-smooth for some $L_0, L_\rho \geq 0$. However, it is not necessarily (L_0, L_1)-smooth for any $L_0, L_1 \geq 0$.*

Proof of Proposition A.1.1. By standard calculations, we can obtain

$$f'(x) = \log(a)\log(b)\, b^x a^{(b^x)}, \quad f''(x) = \log(b)(\log(a)b^x + 1) \cdot f'(x). \tag{A.1}$$

Note that if $\rho > 1$,

$$\lim_{x\to+\infty} \frac{|f'(x)|^\rho}{|f''(x)|} = \lim_{x\to+\infty} \frac{|f'(x)|^{\rho-1}}{\log(b)(\log(a)b^x + 1)} = \lim_{y\to+\infty} \frac{(\log(a)\log(b)y)^{\rho-1}\, a^{(\rho-1)y}}{\log(b)(\log(a)y + 1)} = \infty,$$

where the first equality is a direct calculation based on (A.1); the second equality uses change of variable $y = b^x$; and the last equality is because exponential functions grow faster than affine functions. Therefore, for any $L_\rho > 0$, there exists $x_0 \in \mathbb{R}$ such that $|f''(x)| \leq L_\rho\, |f'(x)|^\rho$ if $x > x_0$. Next, note that $\lim_{x\to-\infty} f''(x) = 0$. Then for any $\lambda_1 > 0$, there exists $x_1 \in \mathbb{R}$ such that $|f''(x)| \leq \lambda_1$ if $x < x_1$. Also, since f'' is continuous, by Weierstrass's Theorem, we have $|f''(x)| \leq \lambda_2$ if $x_1 \leq x \leq x_0$ for some $\lambda_2 > 0$. Then denoting $L_0 = \max\{\lambda_1, \lambda_2\}$, we know f is (ρ, L_0, L_ρ)-smooth.

Next, to show f is not necessarily (L_0, L_1)-smooth, consider the specific double exponential function $f(x) = e^{(e^x)}$. Then we have

$$f'(x) = e^x e^{(e^x)}, \quad f''(x) = (e^x + 1) \cdot f'(x).$$

For any $x \geq \max\{\log(L_0+1), \log(L_1+1)\}$, we can show that

$$|f''(x)| > (L_1+1)f'(x) > L_0 + L_1 |f'(x)|,$$

which shows f is not (L_0, L_1) smooth for any $L_0, L_1 \geq 0$. □

In the next proposition, we show that any univariate rational function $f(x) = P(x)/Q(x)$, where P and Q are two polynomials, is (ρ, L_0, L_ρ)-smooth with $\rho = 1.5$.

Proposition A.1.2. *The rational function $f(x) = P(x)/Q(x)$, where P and Q are two polynomials, is $(1.5, L_0, L_{1.5})$-smooth for some $L_0, L_{1.5} \geq 0$. However, it is not necessarily (ρ, L_0, L_ρ)-smooth for any $\rho < 1.5$ and $L_0, L_\rho \geq 0$.*

Proof of Proposition A.1.2. Let $f(x) = P(x)/Q(x)$ where P and Q are two polynomials. Then the partial fractional decomposition of $f(x)$ is given by

$$f(x) = w(x) + \sum_{i=1}^{m}\sum_{r=1}^{j_i} \frac{A_{ir}}{(x-a_i)^r} + \sum_{i=1}^{n}\sum_{r=1}^{k_i} \frac{B_{ir}x + C_{ir}}{(x^2+b_ix+c_i)^r},$$

where $w(x)$ is a polynomial, $A_{ir}, B_{ir}, C_{ir}, a_i, b_i, c_i$ are all real constants satisfying $b_i^2 - 4c_i < 0$ for each $1 \leq i \leq n$ which implies $x^2 + b_ix + c_i > 0$ for all $x \in \mathbb{R}$. Assume $j_i \geq 1$ and $A_{ij_i} \neq 0$ without loss of generality. Then we know f has only finite singular points $\{a_i\}_{1\leq i\leq m}$ and has continuous first and second order derivatives at all other points. To simplify notation, denote

$$p_{ir}(x) := \frac{A_{ir}}{(x-a_i)^r}, \quad q_{ir}(x) := \frac{B_{ir}x + C_{ir}}{(x^2+b_ix+c_i)^r}.$$

Then we have $f(x) = w(x) + \sum_{i=1}^{m}\sum_{r=1}^{j_i} p_{ir}(x) + \sum_{i=1}^{n}\sum_{r=1}^{k_i} q_{ir}(x)$. We know that $\frac{r+2}{r+1} \leq 1.5$ for any $r \geq 1$. Then we can show that

$$\lim_{x\to a_i} \frac{|f'(x)|^{1.5}}{|f''(x)|} = \lim_{x\to a_i} \frac{\left|p'_{ij_i}(x)\right|^{1.5}}{\left|p''_{ij_i}(x)\right|} \geq \frac{1}{j_i+1}, \tag{A.2}$$

where the first equality is because one can easily verify that the first and second order derivatives of p_{ij_i} dominate those of all other terms when x goes to a_i, and the second equality is by standard calculations noting that $\frac{j_i+2}{j_i+1} \leq 1.5$. Note that (A.2) implies that, for any $L_\rho > j_i + 1$, there exists $\delta_i > 0$ such that

$$|f''(x)| \leq L_\rho \left|f'(x)\right|^{1.5}, \quad \text{if } |x - a_i| < \delta_i. \tag{A.3}$$

Similarly, one can show $\lim_{x\to\infty} \frac{|f'(x)|^{1.5}}{|f''(x)|} = \infty$, which implies there exists $x_0 > 0$ such that

$$|f''(x)| \leq L_\rho \left|f'(x)\right|^{1.5}, \quad \text{if } |x| > x_0. \tag{A.4}$$

Define

$$\mathcal{B} := \left\{x \in \mathbb{R} \mid |x| \leq x_0 \text{ and } |x - a_i| \geq \delta_i, \forall i\right\}.$$

We know $\mathcal{B}$ is a compact set and therefore the continuous function f'' is bounded within $\mathcal{B}$, i.e., there exists some constant $L_0 > 0$ such that

$$|f''(x)| \leq L_0, \quad \text{if } x \in \mathcal{B}. \tag{A.5}$$

Combining (A.3), (A.4), and (A.5), we have shown

$$|f''(x)| \leq L_0 + L_\rho \left|f'(x)\right|^{1.5}, \quad \forall x \in \mathrm{dom}(f),$$

which completes the proof of the first part.

For the second part, consider the ration function $f(x) = 1/x$. Then we know that $f'(x) = -1/x^2$ and $f''(x) = 2/x^3$. Note that for any $\rho < 1.5$ and $0 < x \leq \min\{(L_0 +$

$1)^{-1/3}, (L_\rho+1)^{-1/(3-2\rho)}\}$, we have

$$|f''(x)| = \frac{1}{x^3} + \frac{1}{x^{3-2\rho}} \cdot |f'(x)|^\rho > L_0 + L_\rho \, |f'(x)|^\rho \, ,$$

which shows f is not (ρ, L_0, L_ρ) smooth for any $\rho < 1.5$ and $L_0, L_\rho \geq 0$. □

Finally, we complete this section with the proof of Proposition 2.3.1, which shows self-concordant functions are $(2, L_0, L_2)$-smooth for some $L_0, L_\rho \geq 0$.

Proof of Proposition 2.3.1. Let $h : \mathbb{R} \to \mathbb{R}$ be a self-concordant function. We have $h'''(x) \leq 2h''(x)^{3/2}$. Then, for $x \in (a, b)$, we can obtain

$$\frac{1}{2} h''(x)^{-1/2} h'''(x) \leq h''(x).$$

Integrating both sides from x_0 to y for $x_0, y \in (a, b)$, we have

$$h''(y)^{1/2} - h''(x_0)^{1/2} \leq h'(y) - h'(x_0).$$

Therefore,

$$h''(y) \leq (h''(x_0)^{1/2} - h'(x_0) + h'(y))^2 \leq 2(h''(x_0)^{1/2} - h'(x_0))^2 + 2h'(y)^2.$$

Since $h''(y) > 0$, we have $|h''(y)| = h''(y)$. Therefore, the above inequality shows that h is $(2, L_0, L_2)$-smooth with $L_0 = 2(h''(x_0)^{1/2} - h'(x_0))^2$ and $L_2 = 2$. □

A.2 Proof of Proposition 2.1.2

In order to prove Proposition 2.1.2, we need the following several lemmas. First, the lemma below partially generalizes Grönwall's inequality.

Lemma A.2.1. *Let $\alpha : [a,b] \to [0,\infty)$ and $\beta : [0,\infty) \to (0,\infty)$ be two continuous functions. Suppose $\alpha'(t) \le \beta(\alpha(t))$ almost everywhere over (a,b). Denote function $\phi(u) := \int \frac{1}{\beta(u)} du$. We have for all $t \in [a,b]$,*

$$\phi(\alpha(t)) \le \phi(\alpha(a)) - a + t.$$

Proof of Lemma A.2.1. First, by definition, we know that ϕ is increasing since $\phi' = \frac{1}{\beta} > 0$. Let function $\gamma : [a,b] \to \mathbb{R}$ be the solution of the following differential equation

$$\gamma'(t) = \beta(\gamma(t)) \ \ \forall t \in (a,b), \quad \gamma(a) = \alpha(a). \tag{A.6}$$

Then we have

$$d\phi(\gamma(t)) = \frac{d\gamma(t)}{\beta(\gamma(t))} = dt.$$

Integrating both sides, noting that $\gamma(a) = \alpha(a)$ by (A.6), we obtain

$$\phi(\gamma(t)) - \phi(\alpha(a)) = t - a.$$

Then it suffices to show $\phi(\alpha(t)) \le \phi(\gamma(t))$, $\forall t \in [a,b]$. Note that the following inequality holds almost everywhere.

$$(\phi(\alpha(t)) - \phi(\gamma(t)))' = \phi'(\alpha(t))\alpha'(t) - \phi'(\gamma(t))\gamma'(t) = \frac{\alpha'(t)}{\beta(\alpha(t))} - \frac{\gamma'(t)}{\beta(\gamma(t))} \le 0,$$

where the inequality is because $\alpha'(t) \le \beta(\alpha(t))$ by the assumption of this lemma and $\gamma'(t) = \beta(\gamma(t))$ by (A.6). Since $\phi(\alpha(a)) - \phi(\gamma(a)) = 0$, we know for all $t \in [a,b]$, $\phi(\alpha(t)) \le \phi(\gamma(t))$, which completes the proof. □

With Lemma A.2.1, one can bound the gradient norm within a small enough neighborhood of a given point as in the following lemma.

Lemma A.2.2. *If the objective function f is ℓ-smooth, for any two points $x, y \in \mathbb{R}^d$ such that the closed line segment between x and y is contained in $\mathcal{X}$, if $\|y - x\| \leq \frac{a}{\ell(\|\nabla f(x)\|+a)}$ for any $a > 0$, we have*

$$\|\nabla f(y)\| \leq \|\nabla f(x)\| + a.$$

Proof of Lemma A.2.2. Denote $z(t) := (1-t)x + ty$ for $0 \leq t \leq 1$. Then we know $z(t) \in \mathcal{X}$ for all $0 \leq t \leq 1$ by the assumption made in this lemma. Then we can also define $\alpha(t) := \|\nabla f(z(t))\|$ for $0 \leq t \leq 1$. Note that for any $0 \leq t \leq s \leq 1$, by triangle inequality,

$$\alpha(s) - \alpha(t) \leq \|\nabla f(z(s)) - \nabla f(z(t))\| . \tag{A.7}$$

We know that $\alpha(t) = \|\nabla f(z(t))\|$ is differentiable almost everywhere since f is second order differentiable almost everywhere (Here we assume $\alpha(t) \neq 0$ for $0 < t < 1$ without loss of generality. Otherwise, one can define $t_m = \sup\{0 < t < 1 \mid \alpha(t) = 0\}$ and consider the interval $[t_m, 1]$ instead). Then the following equality holds almost everywhere

$$\begin{aligned}\alpha'(t) =& \lim_{s \downarrow t} \frac{\alpha(s) - \alpha(t)}{s - t} \leq \lim_{s \downarrow t} \frac{\|\nabla f(z(s)) - \nabla f(z(t))\|}{s - t} = \left\| \lim_{s \downarrow t} \frac{\nabla f(z(s)) - \nabla f(z(t))}{s - t} \right\| \\ =& \left\| \nabla^2 f(z(t))(y - x) \right\| \leq \left\| \nabla^2 f(z(t)) \right\| \|y - x\| \leq \ell(\alpha(t)) \|y - x\| ,\end{aligned}$$

where the first inequality is due to (A.7) and the last inequality is by Definition 1. Let $\beta(u) := \ell(u) \cdot \|y - x\|$ and $\phi(u) := \int_0^u \frac{1}{\beta(v)} dv$. By Lemma A.2.1, we know that

$$\phi\left(\|\nabla f(y)\|\right) = \phi(u(1)) \leq \phi(u(0)) + 1 = \phi\left(\|\nabla f(x)\|\right) + 1.$$

Denote $\psi(u) := \int_0^u \frac{1}{\ell(v)} dv = \phi(u) \cdot \|y - x\|$. We have

$$\begin{aligned}
\psi\left(\|\nabla f(y)\|\right) \leq & \psi\left(\|\nabla f(x)\|\right) + \|y - x\| \\
\leq & \psi\left(\|\nabla f(x)\|\right) + \frac{a}{\ell(\|\nabla f(x)\| + a)} \\
\leq & \int_0^{\|\nabla f(x)\|} \frac{1}{\ell(v)} \, dv + \int_{\|\nabla f(x)\|}^{\|\nabla f(x)\|+a} \frac{1}{\ell(v)} \, dv \\
= & \psi(\|\nabla f(x)\| + a).
\end{aligned}$$

Since ψ is increasing, we have $\|\nabla f(y)\| \leq \|\nabla f(x)\| + a$. □

With Lemma A.2.2, we are ready to prove Proposition 2.1.2.

Proof of Proposition 2.1.2. We prove the two directions in this proposition separately.

1. An (r, ℓ)-smooth function is ℓ-smooth.

For each fixed $x \in \mathcal{X}$ where $\nabla^2 f(x)$ exists and any unit-norm vector w, by Definition 2, we know that for any $t \leq r(\|\nabla f(x)\|)$,

$$\|\nabla f(x + tw) - \nabla f(x)\| \leq t \cdot \ell(\|\nabla f(x)\|).$$

Then we know that

$$\begin{aligned}
\left\|\nabla^2 f(x) w\right\| = & \left\|\lim_{t \downarrow 0} \frac{1}{t}(\nabla f(x + tw) - \nabla f(x))\right\| \\
= & \lim_{t \downarrow 0} \frac{1}{t} \|(\nabla f(x + tw) - \nabla f(x))\| \leq \ell(\|\nabla f(x)\|),
\end{aligned}$$

which implies $\|\nabla^2 f(x)\| \leq \ell(\|\nabla f(x)\|)$ for any point x if $\nabla^2 f(x)$ exists.

Then it suffices to show that $\nabla^2 f(x)$ exists almost everywhere. Note that for each $x \in \mathcal{X}$, Definition 2 states that the gradient function is $\ell(\|\nabla f(x)\|)$ Lipschitz within the ball $\mathcal{B}(x, r(\|\nabla f(x)\|))$. Then by Rademacher's Theorem, f is twice differentiable almost everywhere within this ball. Then we can show it is also twice differentiable almost everywhere within the entire domain $\mathcal{X}$ as long as we can cover $\mathcal{X}$ with countably many such balls. Define $\mathcal{S}_n := \{x \in \mathcal{X} \mid n \leq \|\nabla f(x)\| \leq n+1\}$ for integer $n \geq 0$. We have $\mathcal{X} = \cup_{n\geq 0}\mathcal{S}_n$. One can easily find an internal covering of $\mathcal{S}_n$ with balls of size $r(n+1)$[1], i.e., there exist $\{x_{n,i}\}_{i\geq 0}$, where $x_{n,i} \in \mathcal{S}_n$, such that $\mathcal{S}_n \subseteq \cup_{i\geq 0}\mathcal{B}(x_{n,i}, r(n+1)) \subseteq \cup_{i\geq 0}\mathcal{B}(x_{n,i}, r(\|\nabla f(x_{n,i})\|))$. Therefore we have $\mathcal{X} \subseteq \cup_{n,i\geq 0}\mathcal{B}(x_{n,i}, r(\|\nabla f(x_{n,i})\|))$ which completes the proof.

2. An ℓ-smooth function satisfying Assumption 1.1 is (r,m)-smooth where $m(u) := \ell(u+a)$ and $r(u) := a/m(u)$ for any $a > 0$.

For any $y \in \mathbb{R}^d$ satisfying $\|y - x\| \leq r(\|\nabla f(x)\|) = \frac{a}{\ell(\|\nabla f(x)\|+a)}$, denote $z(t) := (1-t)x + ty$ for $0 \leq t \leq 1$. We first show $y \in \mathcal{X}$ by contradiction. Suppose $y \notin \mathcal{X}$, let us define $t_{\mathrm{b}} := \inf\{0 \leq t \leq 1 \mid z(t) \notin \mathcal{X}\}$ and $z_{\mathrm{b}} := z(t_{\mathrm{b}})$. Then we know z_{b} is a boundary point of $\mathcal{X}$. Since f is a closed function with an open domain, we have

$$\lim_{t\uparrow t_{\mathrm{b}}} f(z(t)) = \infty. \tag{A.8}$$

On the other hand, by the definition of t_{b}, we know $z(t) \in \mathcal{X}$ for every $0 \leq t < t_{\mathrm{b}}$. Then by Lemma A.2.2, for all $0 \leq t < t_{\mathrm{b}}$, we have $\|\nabla f(z(t))\| \leq \|\nabla f(x)\| + a$. Therefore for all $0 \leq t < t_{\mathrm{b}}$,

$$\begin{aligned} f(z(t)) \leq & f(x) + \int_0^t \left\langle \nabla f(z(s)), y - x \right\rangle ds \\ \leq & f(x) + (\|\nabla f(x)\| + a) \cdot \|y - x\| \\ < & \infty, \end{aligned}$$

which contradicts (A.8). Therefore we have shown $y \in \mathcal{X}$. Since y is chosen arbitrarily with the ball $\mathcal{B}(x, r(\|\nabla f(x)\|))$, we have $\mathcal{B}(x, r(\|\nabla f(x)\|)) \subseteq \mathcal{X}$. Then for any $x_1, x_2 \in \mathcal{B}(x, r(\|\nabla f(x)\|))$, we denote $w(t) := tx_1 + (1-t)x_2$. Then we know $w(t) \in \mathcal{B}(x, r(\|\nabla f(x)\|))$

[1] We can find an internal covering in the following way. We first cover $\mathcal{S}_n$ with countably many hyper-cubes of length $r(n+1)/\sqrt{d}$, which is obviously doable. Then for each hyper-cube that intersects with $\mathcal{S}_n$, we pick one point from the intersection. Then the ball centered at the picked point with radius $r(n+1)$ covers this hyper-cube. Therefore, the union of all such balls can cover $\mathcal{S}_n$.

for all $0 \leq t \leq 1$ and can obtain

$$\begin{aligned}\|\nabla f(x_1) - \nabla f(x_2)\| =& \left\| \int_0^1 \nabla^2 f(w(t)) \cdot (x_1 - x_2)\, dt \right\| \\ \leq& \|x_1 - x_2\| \cdot \int_0^1 \ell(\|\nabla f(x)\| + a)\, dt \\ =& m(\|\nabla f(x)\|) \cdot \|x_1 - x_2\|,\end{aligned}$$

where the last inequality is due to Lemma A.2.2. □

A.3 Proofs of properties of generalized smoothness

In this part, we provide the proofs of the useful properties stated in Section 2.2, including Lemma 2.2.1, Lemma 2.2.3, and Corollary 2.2.4.

Proof of Lemma 2.2.1. First, note that since ℓ is non-decreasing and r is non-increasing, we have $\ell(\|\nabla f(x)\|) \leq \ell(G) = L$ and $r(G) \leq r(\|\nabla f(x)\|)$. Then by Definition 2, we directly have that $\mathcal{B}(x, r(G)) \subseteq \mathcal{B}(x, r(\|\nabla f(x)\|)) \subseteq \mathcal{X}$, and that for any $x_1, x_2 \in \mathcal{B}(x, r(G))$, we have

$$\|\nabla f(x_1) - \nabla f(x_2)\| \leq \ell(\|\nabla f(x)\|) \|x_1 - x_2\| \leq L \|x_1 - x_2\|.$$

Next, for the second inequality in (2.1), define $z(t) := (1-t)x_2 + tx_1$ for $0 \leq t \leq 1$. We know $z(t) \in \mathcal{B}(x, r(G))$. Note that we have shown

$$\|\nabla f(z(t)) - \nabla f(x_2)\| \leq L \|z(t) - x_2\| = tL \|x_1 - x_2\|. \tag{A.9}$$

Then we have

$$\begin{aligned} f(x_1) - f(x_2) &= \int_0^1 \left\langle \nabla f(z(t), x_1 - x_2 \right\rangle dt \\ &= \int_0^1 \left\langle \nabla f(x_2), x_1 - x_2 \right\rangle + \left\langle \nabla f(z(t)) - \nabla f(x_2), x_1 - x_2 \right\rangle dt \\ &\leq \left\langle \nabla f(x_2), x_1 - x_2 \right\rangle + L \left\| x_1 - x_2 \right\|^2 \int_0^1 t \, dt \\ &= \left\langle \nabla f(x_2), x_1 - x_2 \right\rangle + \frac{L}{2} \left\| x_1 - x_2 \right\|^2, \end{aligned}$$

where the inequality is due to (A.9). □

Proof of Lemma 2.2.3. If f is ℓ-smooth satisfying Assumption 1.1, by Proposition 2.1.2, f is also (r, m)-smooth where $m(u) = \ell(2u)$ and $r(u) = u/\ell(2u)$. Then by Lemma 2.2.1 where we choose $G = \|\nabla f(x)\|$, we have that $\mathcal{B}\left(x, \frac{\|\nabla f(x)\|}{\ell(2\|\nabla f(x)\|)}\right) \subseteq \mathcal{X}$, and that for any $x_1, x_2 \in \mathcal{B}\left(x, \frac{\|\nabla f(x)\|}{\ell(2\|\nabla f(x)\|)}\right)$, we have

$$f(x_1) \leq f(x_2) + \left\langle \nabla f(x_2), x_1 - x_2 \right\rangle + \frac{\ell(2\left\|\nabla f(x)\right\|)}{2} \left\| x_1 - x_2 \right\|.$$

Choosing $x_2 = x$ and $x_1 = x - \frac{\nabla f(x)}{\ell(2\|\nabla f(x)\|)}$, it is easy to verify that $x_1, x_2 \in \mathcal{B}\left(x, \frac{\|\nabla f(x)\|}{\ell(2\|\nabla f(x)\|)}\right)$. Therefore, we have

$$f^* \leq f\left(x - \frac{\nabla f(x)}{\ell(2\left\|\nabla f(x)\right\|)}\right) \leq f(x) - \frac{\left\|\nabla f(x)\right\|^2}{2\ell(2\left\|\nabla f(x)\right\|)},$$

which completes the proof. □

Proof of Corollary 2.2.4. We first show $G < \infty$. Note that since ℓ is sub-quadratic, we know $\lim_{u\to\infty} 2\ell(2u)/u^2 = 0$. Therefore, for any $F > 0$, there exists some $M > 0$ such that $2\ell(2u)/u^2 < 1/F$ for every $u > M$. In other words, for any u satisfying $u^2 \leq 2\ell(2u) \cdot F$, we must have $u \leq M$. Therefore, by definition of G, we have $G \leq M < \infty$ if $F > 0$. If $F = 0$, we trivially get $G = 0 < \infty$. Also, since the set $\{u \geq 0 \mid u^2 \leq 2\ell(2u) \cdot F\}$ is closed and bounded, we know its supremum G is in this set and it is also straightforward to show $G^2 = 2\ell(2G) \cdot F$.

Next, by Lemma 2.2.3, we know

$$\|\nabla f(x)\|^2 \leq 2\ell(2\|\nabla f(x)\|) \cdot (f(x) - f^*) \leq 2\ell(2\|\nabla f(x)\|) \cdot F.$$

Then based on the definition of G, we have $\|\nabla f(x)\| \leq G$. □

Appendix B

Proofs for Chapter 3

B.1 Analysis of GD for convex functions

In this section, we provide the detailed convergence analysis of gradient descent in the convex setting, including the proofs of Lemma 3.1.1 and Theorem 3.1.2, for which the following lemma will be helpful.

Lemma B.1.1 (Co-coercivity)**.** *If f is convex and (r, ℓ)-smooth, for any $x \in \mathcal{X}$ and $y \in \mathcal{B}(x, r(\|\nabla f(x)\|)/2)$, we have $y \in \mathcal{X}$ and*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 ,$$

where $L = \ell(\|\nabla f(x)\|)$.

Proof of Lemma B.1.1. Define the Bregman divergences $\phi_x(w) := f(w) - \langle \nabla f(x), w \rangle$ and $\phi_y(w) := f(w) - \langle \nabla f(y), w \rangle$, which are both convex functions. Since $\nabla \phi_x(w) = \nabla f(w) - \nabla f(x)$, we have $\nabla \phi_x(x) = 0$ which implies $\min_w \phi_x(w) = \phi_x(x)$ as ϕ_x is convex. Similarly we have $\min_w \phi_y(w) = \phi_y(y)$.

Denote $r_x := r(\|\nabla f(x)\|)$. Since f is (r, ℓ)-smooth, we know its gradient ∇f is L-Lipschitz locally in $\mathcal{B}(x, r_x)$. Since $\nabla \phi_x(w) - \nabla f(w) = \nabla f(x)$ is a constant, we know $\nabla \phi_x$ is also

L-Lipschitz locally in $\mathcal{B}(x, r_x)$. Then similar to the proof of Lemma 2.2.1, one can easily show that for any $x_1, x_2 \in \mathcal{B}(x, r_x)$, we have

$$\phi_x(x_1) \leq \phi_x(x_2) + \Big\langle \nabla \phi_x(x_2), x_1 - x_2 \Big\rangle + \frac{L}{2} \|x_1 - x_2\|^2 . \tag{B.1}$$

Note that for any $y \in \mathcal{B}(x, r(\|\nabla f(x)\|)/2)$ as in the lemma statement,

$$\left\| y - \frac{1}{L} \nabla \phi_x(y) - x \right\| \leq \|y - x\| + \frac{1}{L} \|\nabla f(y) - \nabla f(x)\| \leq 2 \|y - x\| \leq r_x,$$

where the first inequality uses triangle inequality and $\nabla \phi_x(y) = \nabla f(y) - \nabla f(x)$; and the second inequality uses Definition 2. It implies that $y - \frac{1}{L} \nabla \phi_x(y) \in \mathcal{B}(x, r_x)$. Then we can obtain

$$\phi_x(x) = \min_w \phi_x(w) \leq \phi_x\Big(y - \frac{1}{L} \nabla \phi_x(y)\Big) \leq \phi_x(y) - \frac{1}{2L} \|\nabla \phi_x(y)\|^2 ,$$

where the last inequality uses (B.1) where we choose $x_1 = y - \frac{1}{L} \nabla \phi_x(y)$ and $x_2 = y$. By the definition of ϕ_x, the above inequality is equivalent to

$$\frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2 \leq f(y) - f(x) - \langle \nabla f(x), x - y \rangle.$$

Similar argument can be made for $\phi_y(\cdot)$ to obtain

$$\frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2 \leq f(x) - f(y) - \langle \nabla f(y), y - x \rangle.$$

Summing up the two inequalities, we can obtain the desired result. □

With Lemma B.1.1, we prove Lemma 3.1.1 as follows.

Proof of Lemma 3.1.1. Let $L = \ell(G)$. We first verify that $x^+ \in \mathcal{B}(x, r(G)/2)$. Note that

$$\left\|x^+ - x\right\| = \|\eta \nabla f(x)\| \leq \eta G \leq r(G)/2,$$

where we choose $\eta \leq r(G)/(2G)$. Thus by Lemma B.1.1, we have

$$\begin{aligned}
\left\|\nabla f(x^+)\right\|^2 &= \|\nabla f(x)\|^2 + 2\langle \nabla f(x^+) - \nabla f(x), \nabla f(x)\rangle + \left\|\nabla f(x^+) - \nabla f(x)\right\|^2 \\
&= \|\nabla f(x)\|^2 - \frac{2}{\eta}\langle \nabla f(x^+) - \nabla f(x), x^+ - x\rangle + \left\|\nabla f(x^+) - \nabla f(x)\right\|^2 \\
&\leq \|\nabla f(x)\|^2 + \left(1 - \frac{2}{\eta L}\right)\left\|\nabla f(x^+) - \nabla f(x)\right\|^2 \\
&\leq \|\nabla f(x)\|^2,
\end{aligned}$$

where the first inequality uses Lemma B.1.1 and the last inequality chooses $\eta \leq 2/L$. □

With Lemma 3.1.1, we are ready to prove both Theorem 3.1.2 and Theorem 3.1.3.

Proof of Theorem 3.1.2. Denote $G := \|\nabla f(x_0)\|$. Then we trivially have $\|\nabla f(x_0)\| \leq G$. Lemma 3.1.1 states that if $\|\nabla f(x_t)\| \leq G$ for any $t \geq 0$, then we also have $\|\nabla f(x_{t+1})\| \leq \|\nabla f(x_t)\| \leq G$. By induction, we can show that $\|\nabla f(x_t)\| \leq G$ for all $t \geq 0$. Then the rest of the proof basically follows the standard textbook analysis. We still provide the detailed proof below for completeness.

Note that $\|x_{t+1} - x_t\| = \eta \|\nabla f(x_t)\| \leq \eta G \leq r(G)$, where we choose $\eta \leq r(G)/(2G)$. Thus we can apply Lemma 2.2.1 to obtain

$$\begin{aligned}
0 &\geq f(x_{t+1}) - f(x_t) - \langle \nabla f(x_t), x_{t+1} - x_t\rangle - \frac{L}{2}\|x_{t+1} - x_t\|^2 \\
&\geq f(x_{t+1}) - f(x_t) - \langle \nabla f(x_t), x_{t+1} - x_t\rangle - \frac{1}{2\eta}\|x_{t+1} - x_t\|^2, \qquad \text{(B.2)}
\end{aligned}$$

where the last inequality chooses $\eta \leq 1/L$. Meanwhile, by convexity between x_t and x^*, we have

$$0 \geq f(x_t) - f^* + \langle \nabla f(x_t), x^* - x_t \rangle. \tag{B.3}$$

Note that $(t+1)\times$(B.2)+(B.3) gives

$$\begin{aligned} 0 \geq\ & f(x_t) - f^* + \langle \nabla f(x_t), x^* - x_t \rangle \\ & + (1+t)\left(f(x_{t+1}) - f(x_t) - \langle \nabla f(x_t), x_{t+1} - x_t \rangle - \frac{1}{2\eta} \left\| x_{t+1} - x_t \right\|^2 \right). \end{aligned}$$

Then reorganizing the terms of the above inequality, noting that

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

we can obtain

$$(t+1)(f(x_{t+1}) - f^*) + \frac{1}{2\eta} \left\| x_{t+1} - x^* \right\|^2 \leq t(f(x_t) - f^*) + \frac{1}{2\eta} \left\| x_t - x^* \right\|^2.$$

The above inequality implies $t(f(x_t) - f^*) + \frac{1}{2\eta} \left\| x_t - x^* \right\|^2$ is a non-increasing potential function, which directly implies the desired result. □

Proof of Theorem 3.1.3. Since strongly convex functions are also convex, by the same argument as in the proof of Theorem 3.1.2, we have $\|\nabla f(x_t)\| \leq G := \|\nabla f(x_0)\|$ for all $t \geq 0$. Moreover, (B.2) still holds. For μ-strongly-convex function, we can obtain a tighter version of (B.3) as follows.

$$0 \geq f(x_t) - f^* + \langle \nabla f(x_t), x^* - x_t \rangle + \frac{\mu}{2} \left\| x^* - x_t \right\|^2. \tag{B.4}$$

Let $A_0 = 0$ and $A_{t+1} = (1 + A_t)/(1 - \eta\mu)$ for all $t \geq 0$. Combining (B.2) and (B.4), we have

$$\begin{aligned} 0 \geq{} & (A_{t+1} - A_t)(f(x_t) - f^* + \langle \nabla f(x_t), x^* - x_t \rangle) \\ & + A_{t+1}\left(f(x_{t+1}) - f(x_t) - \langle \nabla f(x_t), x_{t+1} - x_t \rangle - \frac{1}{2\eta} \left\| x_{t+1} - x_t \right\|^2 \right). \end{aligned}$$

Then reorganizing the terms of the above inequality, noting that

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

we can obtain

$$A_{t+1}(f(x_{t+1}) - f^*) + \frac{1 + \eta\mu A_{t+1}}{2\eta} \left\| x_{t+1} - x^* \right\|^2 \leq A_t(f(x_t) - f^*) + \frac{1 + \eta\mu A_t}{2\eta} \left\| x_t - x^* \right\|^2 .$$

The above inequality means $A_t(f(x_t) - f^*) + \frac{1+\eta\mu A_t}{2\eta} \left\| x_t - x^* \right\|^2$ is a non-increasing potential function. Thus by telescoping we have

$$f(x_T) - f^* \leq \frac{\mu(1 - \eta\mu)^T}{2(1 - (1 - \eta\mu)^T)} \left\| x_0 - x^* \right\|^2 .$$

□

B.2 Analysis of NAG for convex functions

In this section, we provide the detailed analysis of Nesterov's accelerated gradient method in the convex setting. As we discussed in Section 3.1.2, the stepsize size choice in Theorem 3.1.4 is smaller than the classical one. Therefore, we provide a more fine-grained version of the theorem, which allows the stepsize to depend on the degree of ℓ.

Theorem B.2.1. *Suppose f is convex and ℓ-smooth satisfying Assumptions 1.1, 1.2, and 3.1.*

For $\alpha \in (0, 2]$, if $\ell(u) = o(u^\alpha)$, i.e., $\lim_{u\to\infty} \ell(u)/u^\alpha = 0$, then there must exist a constant G such that for $L := \ell(2G)$, we have

$$G \geq \max \left\{ 8 \max\{L^{1/\alpha - 1/2}, 1\} \sqrt{L((f(x_0) - f^*) + \|x_0 - x^*\|^2)}, \|\nabla f(x_0)\| \right\}. \tag{B.5}$$

Choose $\eta \leq \min\left\{ \frac{1}{16L^{3-2/\alpha}}, \frac{1}{2L} \right\}$. Then the iterates of Algorithm 1 satisfy

$$f(x_T) - f^* \leq \frac{4(f(x_0 - f^*) + 4\|x_0 - x^*\|^2}{\eta T^2 + 4}.$$

Note that when $\alpha = 2$, i.e., ℓ is sub-quadratic, Theorem B.2.1 reduces to Theorem 3.1.4 which chooses $\eta \leq \min\{\frac{1}{16L^2}, \frac{1}{2L}\}$. When $\alpha = 1$, i.e., ℓ is sub-linear, the above theorem chooses $\eta \leq \frac{1}{16L}$ as in the classical textbook analysis up to a numerical constant factor.

Throughout this section, we will assume f is convex and ℓ-smooth, and consider the parameter choices in Theorem B.2.1, unless explicitly stated. Note that since f is ℓ-smooth, it is also (r, m)-smooth with $m(u) = \ell(u + G)$ and $r(u) = \frac{G}{\ell(u+G)}$ by Proposition 2.1.2. Note that $m(G) = \ell(2G) = L$ and $r(G) = G/L$. Then the stepsize satisfies $\eta \leq 1/(2L) \leq \min\{\frac{2}{m(G)}, \frac{r(G)}{2G}\}$.

Before proving Theorem B.2.1, we first present several additional useful lemmas. To start with, we provide two lemmas regarding the weights $\{A_t\}_{t\geq 0}$ and $\{B_t\}_{t\geq 0}$ used in Algorithm 1. The lemma below states that $B_t = \Theta(t^2)$.

Lemma B.2.2. *The weights $\{B_t\}_{t\geq 0}$ in Algorithm 1 satisfy $\frac{1}{4}t^2 \leq B_t \leq t^2$ for all $t \geq 0$.*

Proof of Lemma B.2.2. We prove this lemma by induction. First note that the inequality obviously holds for $B_0 = 0$. Suppose its holds up to t. Then we have

$$B_{t+1} = B_t + \frac{1}{2}(1 + \sqrt{4B_t + 1}) \geq \frac{1}{4}t^2 + \frac{1}{2}(1 + \sqrt{t^2 + 1}) \geq \frac{1}{4}(t + 1)^2.$$

Similarly, we have

$$B_{t+1} = B_t + \frac{1}{2}(1 + \sqrt{4B_t + 1}) \leq t^2 + \frac{1}{2}(1 + \sqrt{4t^2 + 1}) \leq (t+1)^2.$$

□

Lemma B.2.2 implies the following useful lemma.

Lemma B.2.3. *The weights $\{A_t\}_{t\geq 0}$ in Algorithm 1 satisfy that*

$$(1 - \frac{A_t}{A_{t+1}})\frac{1}{A_t}\sum_{s=0}^{t-1}\sqrt{A_{s+1}}(A_{s+1} - A_s - 1) \leq 4.$$

Proof of Lemma B.2.3. First, note that it is easy to verify that $A_{s+1} - A_s - 1 = B_{s+1} - B_s - 1 \geq 0$, which implies each term in the LHS of the above inequality is non-negative. Then we have

$$\begin{aligned}
&(1 - \frac{A_t}{A_{t+1}})\frac{1}{A_t}\sum_{s=0}^{t-1}\sqrt{A_{s+1}}(A_{s+1} - A_s - 1) \\
\leq\ &\frac{1}{A_{t+1}\sqrt{A_t}}(A_{t+1} - A_t)\sum_{s=0}^{t-1}(A_{s+1} - A_s - 1) && (A_t \geq A_{s+1}) \\
=\ &\frac{1}{A_{t+1}\sqrt{A_t}}(B_{t+1} - B_t)\sum_{s=0}^{t-1}(B_{s+1} - B_s - 1) && (A_s = B_s + 1/\eta) \\
=\ &\frac{1}{A_{t+1}\sqrt{A_t}} \cdot \frac{1}{2}(1 + \sqrt{4B_t + 1})\sum_{s=0}^{t-1}\Big(-1 + \frac{1}{2}(1 + \sqrt{4B_s + 1})\Big) && (\text{by definition of } B_s) \\
\leq\ &8\frac{1}{(t+1)^2 t} \cdot (t+1)\frac{t^2}{2} && (\text{by } A_t \geq B_t \text{ and Lemma B.2.2}) \\
\leq\ &4.
\end{aligned}$$

□

The following lemma summarizes the results in the classical potential function analysis of NAG in [d'Aspremont et al., 2021]. In order to not deal with the generalized smoothness condition for now, we directly assume the inequality (B.6) holds in the lemma, which will be proved later under the generalized smoothness condition.

Lemma B.2.4. *For any $t \geq 0$, if the following inequality holds,*

$$f(y_t) + \langle \nabla f(y_t), x_{t+1} - y_t \rangle + \frac{1}{2\eta} \left\| x_{t+1} - y_t \right\|^2 \geq f(x_{t+1}), \tag{B.6}$$

then we can obtain

$$A_{t+1}(f(x_{t+1}) - f^*) + \frac{1}{2\eta} \left\| z_{t+1} - x^* \right\|^2 \leq A_t(f(x_t) - f^*) + \frac{1}{2\eta} \left\| z_t - x^* \right\|^2. \tag{B.7}$$

Proof of Lemma B.2.4. These derivations below can be found in [d'Aspremont et al., 2021]. We present them here for completeness.

First, since f is convex, the convexity between x^* and y_t gives

$$f^* \geq f(y_t) + \langle \nabla f(y_t), x^* - y_t \rangle.$$

Similarly the convexity between x_t and y_t gives

$$f(x_t) \geq f(y_t) + \langle \nabla f(y_t), x_t - y_t \rangle.$$

Combining the above two inequalities as well as (B.6) assumed in this lemma, we have

$$\begin{aligned} 0 \geq\ & (A_{t+1} - A_t)(f(y_t) - f^* + \langle \nabla f(y_t), x^* - y_t \rangle) \\ & + A_t(f(y_t) - f(x_t) + \langle \nabla f(y_t), x_t - y_t \rangle) \\ & + A_{t+1}\left(f(x_{t+1}) - f(y_t) - \langle \nabla f(y_t), x_{t+1} - y_t \rangle - \frac{1}{2\eta} \left\| x_{t+1} - y_t \right\|^2 \right). \end{aligned} \tag{B.8}$$

Furthermore, note that

$$\begin{aligned}
&\frac{1}{2\eta}\Big(\|z_{t+1}-x^*\|^2-\|z_t-x^*\|^2\Big)\\
&=\frac{1}{2\eta}\Big(\|z_{t+1}-z_t\|^2+2\langle z_{t+1}-z_t, z_t-x^*\rangle\Big)\\
&=\frac{1}{2\eta}\Big(\eta^2(A_{t+1}-A_t)^2\left\|\nabla f(y_t)\right\|^2-2\eta(A_{t+1}-A_t)\langle\nabla f(y_t), z_t-x^*\rangle\Big)\\
&=\frac{\eta}{2}(A_{t+1}-A_t)^2\left\|\nabla f(y_t)\right\|^2-(A_{t+1}-A_t)\langle\nabla f(y_t), z_t-x^*\rangle. \qquad \text{(B.9)}
\end{aligned}$$

Meanwhile, we have

$$A_{t+1}x_{t+1}=A_{t+1}y_t-\eta A_{t+1}\nabla f(y_t)=A_{t+1}x_t+(A_{t+1}-A_t)(z_t-x_t)-\eta A_{t+1}\nabla f(y_t).$$

Thus we have

$$(A_{t+1}-A_t)z_t=A_{t+1}x_{t+1}-A_tx_t+\eta A_{t+1}\nabla f(y_t).$$

Plugging back in (B.9), we obtain

$$\begin{aligned}
&\frac{1}{2\eta}\Big(\|z_{t+1}-x^*\|^2-\|z_t-x^*\|^2\Big)\\
&=\frac{\eta}{2}(A_{t+1}-A_t)^2\left\|\nabla f(y_t)\right\|^2+(A_{t+1}-A_t)\langle\nabla f(y_t), x^*\rangle\\
&\quad+\langle -A_{t+1}x_{t+1}+A_tx_t-\eta A_{t+1}\nabla f(y_t), \nabla f(y_t)\rangle.
\end{aligned}$$

Thus

$$\begin{aligned}
&(A_{t+1}-A_t)\langle\nabla f(y_t), x^*\rangle+\langle A_tx_t-A_{t+1}x_{t+1}, \nabla f(y_t)\rangle\\
&=\frac{1}{2\eta}\Big(\|z_{t+1}-x^*\|^2-\|z_t-x^*\|^2\Big)+\eta(A_{t+1}-\frac{1}{2}(A_{t+1}-A_t)^2)\left\|\nabla f(y_t)\right\|^2.
\end{aligned}$$

So we can reorganize (B.8) to obtain

$$\begin{aligned}
0 \geq{} & A_{t+1}(f(x_{t+1}) - f^*) - A_t(f(x_t) - f^*) \\
& + (A_{t+1} - A_t)\langle \nabla f(y_t), x^* \rangle + \langle A_t x_t - A_{t+1}x_{t+1}, \nabla f(y_t) \rangle \\
& - \frac{1}{2\eta} A_{t+1} \left\| x_{t+1} - y_t \right\|^2 \\
={} & A_{t+1}(f(x_{t+1}) - f^*) - A_t(f(x_t) - f^*) \\
& + \frac{1}{2\eta}\Big(\left\| z_{t+1} - x^* \right\|^2 - \left\| z_t - x^* \right\|^2 \Big) + \frac{\eta}{2}(A_{t+1} - (A_{t+1} - A_t)^2) \left\| \nabla f(y_t) \right\|^2 .
\end{aligned}$$

Then we complete the proof noting that it is easy to verify

$$A_{t+1} - (A_{t+1} - A_t)^2 = B_{t+1} + \frac{1}{\eta} - (B_{t+1} - B_t)^2 = \frac{1}{\eta} \geq 0.$$

□

In the next lemma, we show that if $\|\nabla f(y_t)\| \leq G$, then the condition (B.6) assumed in Lemma B.2.4 is satisfied at time t.

Lemma B.2.5. *For any $t \geq 0$, if $\|\nabla f(y_t)\| \leq G$, then we have $\|\nabla f(x_{t+1})\| \leq G$, and furthermore,*

$$f(y_t) + \langle \nabla f(y_t), x_{t+1} - y_t \rangle + \frac{1}{2\eta} \left\| x_{t+1} - y_t \right\|^2 \geq f(x_{t+1}).$$

Proof of Lemma B.2.5. As disccued below Theorem B.2.1, the stepsize satisfies $\eta \leq 1/(2L) \leq \min\{\frac{2}{m(G)}, \frac{r(G)}{2G}\}$. Therefore we can apply Lemma 3.1.1 to show $\|\nabla f(x_{t+1})\| \leq \|\nabla f(y_t)\| \leq G$. For the second part, note that $\|x_{t+1} - y_t\| = \eta \|\nabla f(y_t)\| \leq \frac{G}{2L} \leq r(G)$, we can apply Lemma 2.2.1 to show

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

□

With Lemma B.2.4 and Lemma B.2.5, we can show that $\|\nabla f(y_t)\| \leq G$ for all $t \geq 0$, as in the lemma below.

Lemma B.2.6. *For all $t \geq 0$, $\|\nabla f(y_t)\| \leq G$.*

Proof of Lemma B.2.6. We will prove this lemma by induction. First, by Lemma 2.2.3 and the choice of G, it is easy to verify that $\|\nabla f(x_0)\| \leq G$. Then for any fixed $t \geq 0$, suppose that $\|\nabla f(x_s)\| \leq G$ for all $s < t$. Then by Lemma B.2.4 and Lemma B.2.5, we know that $\|\nabla f(x_s)\| \leq G$ for all $0 \leq s \leq t$, and that for all $s < t$,

$$A_{s+1}(f(x_{s+1}) - f^*) + \frac{1}{2\eta} \|z_{s+1} - x^*\|^2 \leq A_s(f(x_s) - f^*) + \frac{1}{2\eta} \|z_s - x^*\|^2 . \tag{B.10}$$

By telescoping (B.10), we have for all $0 \leq s < t$,

$$f(x_{s+1}) - f^* \leq \frac{1}{\eta A_{s+1}}((f(x_0) - f^*) + \|z_0 - x^*\|^2). \tag{B.11}$$

For $0 \leq s \leq t$, since $\|\nabla f(x_s)\| \leq G$, then Lemma 2.2.3 implies

$$\|\nabla f(x_s)\|^2 \leq 2L(f(x_s) - f^*). \tag{B.12}$$

Note that by Algorithm 1, we have

$$z_t - x_t = \frac{A_{t-1}}{A_t}(z_{t-1} - x_{t-1}) - \eta(A_t - A_{t-1})\nabla f(y_{t-1}) + \eta \nabla f(y_{t-1}).$$

Thus we can obtain

$$z_t - x_t = -\frac{1}{A_t} \sum_{s=1}^{t-1} \eta A_{s+1}(A_{s+1} - A_s - 1)\nabla f(y_s).$$

Therefore

$$y_t - x_t = -(1 - \frac{A_t}{A_{t+1}}) \frac{1}{A_t} \sum_{s=1}^{t-1} \eta A_{s+1}(A_{s+1} - A_s - 1) \nabla f(y_s).$$

Thus we have

$$\|y_t - x_t\| \leq (1 - \frac{A_t}{A_{t+1}}) \frac{1}{A_t} \sum_{s=1}^{t-1} \eta A_{s+1}(A_{s+1} - A_s - 1) \|\nabla f(y_s)\| =: \mathcal{I}.$$

Since $\|\nabla f(y_s)\| \leq G$ and $\|x_{s+1} - y_s\| = \|\eta \nabla f(y_s)\| \leq r(G)$ for $s < t$, by Lemma 2.2.1, we have

$$\begin{aligned}
\mathcal{I} &\leq (1 - \frac{A_t}{A_{t+1}}) \frac{1}{A_t} \sum_{s=1}^{t-1} \eta A_{s+1}(A_{s+1} - A_s - 1)(\|\nabla f(x_{s+1})\| + \eta L \|\nabla f(y_s)\|) \\
&\leq \eta L \mathcal{I} + (1 - \frac{A_t}{A_{t+1}}) \frac{1}{A_t} \sum_{s=1}^{t-1} \eta A_{s+1}(A_{s+1} - A_s - 1) \|\nabla f(x_{s+1})\|.
\end{aligned}$$

Thus

$$\begin{aligned}
&\|y_t - x_t\| \\
\leq& \mathcal{I} \leq \frac{1}{1 - \eta L}(1 - \frac{A_t}{A_{t+1}}) \frac{1}{A_t} \sum_{s=1}^{t-1} \eta A_{s+1}(A_{s+1} - A_s - 1) \|\nabla f(x_{s+1})\| \\
\leq& \frac{1}{1 - \eta L}(1 - \frac{A_t}{A_{t+1}}) \frac{1}{A_t} \sum_{s=1}^{t-1} \eta A_{s+1}(A_{s+1} - A_s - 1) \sqrt{2L(f(x_{s+1}) - f^*)} && \text{(by (B.12))} \\
\leq& \frac{1}{1 - \eta L}(1 - \frac{A_t}{A_{t+1}}) \frac{1}{A_t} \sum_{s=1}^{t-1} \eta A_{s+1}(A_{s+1} - A_s - 1) \sqrt{\frac{2L}{A_{s+1}} \cdot \frac{1}{\eta}((f(x_0) - f^*) + \|z_0 - x^*\|^2)} \\
& && \text{(by (B.11))} \\
=& \frac{2\sqrt{\eta L}}{1 - \eta L}(1 - \frac{A_t}{A_{t+1}}) \frac{1}{A_t} \sum_{s=1}^{t-1} \sqrt{A_{s+1}}(A_{s+1} - A_s - 1) \sqrt{(f(x_0) - f^*) + \|z_0 - x^*\|^2} \\
\leq& \frac{8\sqrt{\eta}}{1 - \eta L} \sqrt{L((f(x_0) - f^*) + \|z_0 - x^*\|^2)} && \text{(by Lemma B.2.3)} \\
\leq& \frac{1}{2L^{3/2 - 1/\alpha}} \cdot L^{1/2 - 1/\alpha} G = \frac{G}{2L} \leq r(G). && \text{(by the choices of } \eta \text{ and } G)
\end{aligned}$$

Since $\|\nabla f(x_t)\| \leq G$ and we just showed $\|x_t - y_t\| \leq r(G)$, by Lemma 2.2.1, we have

$$\begin{aligned}
\|\nabla f(y_t)\| &\leq \|\nabla f(x_t)\| + L \|y_t - x_t\| \\
&\leq \sqrt{\frac{2L}{\eta A_t}((f(x_0) - f^*) + \|z_0 - x^*\|^2)} + L \cdot \frac{G}{2L} && \text{(by (B.12) and (B.11))} \\
&\leq G\left(\frac{1}{4} + \frac{1}{2}\right) \leq G. && \text{(by } A_t \geq 1/\eta \text{ and choice of } G\text{)}
\end{aligned}$$

Then we complete the induction as well as the proof.

□

With the three lemmas above, it is straight forward to prove Theorem B.2.1.

Proof of Theorem B.2.1. Combining Lemmas B.2.4, B.2.5, and B.2.6, we know the following inequality holds for all $t \geq 0$.

$$A_{t+1}(f(x_{t+1}) - f^*) + \frac{1}{2\eta} \|z_{t+1} - x^*\|^2 \leq A_t(f(x_t) - f^*) + \frac{1}{2\eta} \|z_t - x^*\|^2 .$$

Then by telescoping, we directly complete the proof. □

B.3 Analysis of NAG for strongly convex functions

In this section, we provide the convergence analysis of the modified version of Nesterov's accelerated gradient method for μ-strongly-convex functions defined in Algorithm 5.

The convergence results is formally presented in the following theorem.

Theorem B.3.1. *Suppose f is μ-strongly-convex and ℓ-smooth satisfying Assumptions 1.1, 1.2,*

and 3.1. For $\alpha \in (0, 2]$, if $\ell(u) = o(u^\alpha)$, i.e., $\lim_{u\to\infty} \ell(u)/u^\alpha = 0$, then there must exist a constant G such that for $L := \ell(2G)$, we have

$$G \geq 8 \max\{L^{1/\alpha - 1/2}, 1\} \sqrt{L((f(x_0) - f^*) + \mu \|z_0 - x^*\|^2)/ \min\{\mu, 1\}}. \tag{B.13}$$

Algorithm 5 NAG for μ-strongly-convex functions

```
**Input** A $\mu$-strongly-convex and $\ell$-smooth function $f$, stepsize $\eta$, initial point $x_0$
**Initialize** $z_0 = x_0$, $B_0 = 0$, and $A_0 = 1/(\eta\mu)$.
**for** $t = 0, \ldots$ **do**
    $B_{t+1} = \frac{2B_t+1+\sqrt{4B_t+4\eta\mu B_t^2+1}}{2(1-\eta\mu)}$
    $A_{t+1} = B_{t+1} + \frac{1}{\eta\mu}$
    $\tau_t = \frac{(A_{t+1}-A_t)(1+\eta\mu A_t)}{A_{t+1}+2\eta\mu A_t A_{t+1}-\eta\mu A_t^2}$ and $\delta_t = \frac{A_{t+1}-A_t}{1+\eta\mu A_{t+1}}$
    $y_t = x_t + \tau_t(z_t - x_t)$
    $x_{t+1} = y_t - \eta\nabla f(y_t)$
    $z_{t+1} = (1-\eta\mu\delta_t)z_t + \eta\mu\delta_t y_t - \eta\delta_t\nabla f(y_t)$
**end for**
```

If we choose

$$\eta \leq \min\left\{\frac{1}{144L^{3-2/\alpha}\log^4\left(e+\frac{144L^{3-2/\alpha}}{\mu}\right)}, \frac{1}{2L}\right\}. \tag{B.14}$$

The iterates generated by Algorithm 5 satisfy

$$f(x_T) - f^* \leq \frac{(1-\sqrt{\eta\mu})^{T-1}(f(x_0 - f^*) + \mu\left\|z_0 - x^*\right\|^2)}{\eta\mu + (1-\sqrt{\eta\mu})^{T-1}}.$$

The above theorem gives a gradient complexity of $\mathcal{O}\left(\frac{1}{\sqrt{\eta\mu}}\log(1/\epsilon)\right)$. Note that Theorem 3.1.2 shows the complexity of GD is $\mathcal{O}\left(\frac{1}{\eta\mu}\log(1/\epsilon)\right)$. It seems NAG gives a better rate at first glance. However, note that the choices of G, L, η in these two theorems are different, it is less clear whether NAG accelerates the optimization in this setting. Below, we informally show that, if $\ell(u) = o(\sqrt{u})$, the rate we obtain for NAG is faster than that for GD.

For simplicity, we informally assume $\ell(u) \asymp x^\rho$ with $\rho \in (0,1)$. Let $G_0 = \|\nabla f(x_0)\|$. Then for GD, by Theorem 3.1.2, we have $\eta_{\text{gd}}\mu \asymp \mu/\ell(G_0) \asymp \mu/G_0^\rho$. For NAG, since ℓ is sub-linear we can choose $\alpha = 1$ in the theorem statement. Since f is μ-strongly-convex, by standard results, we can show that $f(x_0) - f^* \leq \frac{1}{\mu}G_0^2$ and $\|z_0 - x^*\| \leq \frac{1}{\mu}G_0$. Thus the requirement of G in (B.13) can be simplified as $G \gtrsim \ell(G)\cdot G_0/\mu$, which is satisfied if choosing $G \asymp (G_0/\mu)^{1/(1-\rho)}$. Then we also have $\eta_{\text{nag}} \asymp \frac{1}{\ell(G)} \asymp (\mu/G_0)^{\rho/(1-\rho)}$. Thus $\sqrt{\eta_{\text{nag}}\mu} \asymp (\mu/G_0^\rho)^{1/(2-2\rho)}$. This means whenever $1/(2-2\rho) < 1$, i.e., $0 \le \rho < 1/2$, we have $\sqrt{\eta_{\text{nag}}\mu} \gtrsim \eta_{\text{gd}}\mu$, which implies the rate we obtain for NAG is faster than that for GD.

In what follows, we will provide the proof of Theorem B.3.1. We will always use the parameter choices in the theorem throughout this section.

B.3.1 Useful lemmas

In this part, we provide several useful lemmas for proving Theorem B.3.1. To start with, the following two lemmas provide two useful inequalities.

Lemma B.3.2. *For any $0 \le u \le 1$, we have $\log(1+u) \ge \frac{1}{2}u$.*

Lemma B.3.3. *For all $0 < p \le 1$ and $t \ge 0$, we have*

$$t \le \frac{2}{\sqrt{p}} \log(e + \frac{1}{p})(p(1+\sqrt{p})^t + 1).$$

Proof of Lemma B.3.3. Let

$$f(t) = \frac{2}{\sqrt{p}} \log(e + \frac{1}{p})(p(1+\sqrt{p})^t + 1) - t.$$

It is obvious that $f(t) \ge 0$ for $t \le \frac{2}{\sqrt{p}} \log(e + \frac{1}{p})$. For $t > \frac{2}{\sqrt{p}} \log(e + \frac{1}{p})$, we have

$$\begin{aligned}
f'(t) &= 2\sqrt{p} \log(e + \frac{1}{p}) \log(1+\sqrt{p})(1+\sqrt{p})^t - 1 \\
&\ge p(1+\sqrt{p})^t - 1 && \text{(by Lemma B.3.2)} \\
&= p\exp(t\log(1+\sqrt{p})) - 1 \\
&\ge p\exp(t\sqrt{p}/2) - 1 && \text{(by Lemma B.3.2)} \\
&\ge p(e + 1/p) - 1 \ge 0. && (\text{since } t > \tfrac{2}{\sqrt{p}} \log(e + \tfrac{1}{p}))
\end{aligned}$$

Thus f is non-decreasing and

$$f(t) \geq f\left(\frac{2}{\sqrt{p}} \log(e + \frac{1}{p})\right) \geq 0.$$

□

In the next four lemmas, we provide several useful inequalities regarding the weights $\{A_t\}_{t\geq 0}$ and $\{B_t\}_{t\geq 0}$ used in Algorithm 5.

Lemma B.3.4. *For all $s \leq t$, we have*

$$\frac{B_{t+1} - B_t}{B_{t+1}} \cdot \frac{B_{s+1} - B_s}{1 + \eta\mu B_{s+1}} \leq 1,$$

which implies $\tau_t \cdot \delta_s \leq 1$.

Proof of Lemma B.3.4. By Algorithm 5, it is easy to verify

$$(B_{s+1} - B_s)^2 = B_{s+1}(1 + \eta\mu B_{s+1}).$$

This implies

$$B_s = B_{s+1} - \sqrt{B_{s+1}(1 + \eta\mu B_{s+1})}.$$

Thus

$$\frac{B_t}{B_{t+1}} = 1 - \sqrt{\eta\mu + \frac{1}{B_{t+1}}} \geq 1 - \sqrt{\eta\mu + \frac{1}{B_{s+1}}} = \frac{B_s}{B_{s+1}},$$

where in the inequality, we use the fact that B_s is non-decreasing with s. Therefore

$$\frac{B_{t+1} - B_t}{B_{t+1}} \cdot \frac{B_{s+1} - B_s}{1 + \eta\mu B_{s+1}} \leq \frac{B_{s+1} - B_s}{B_{s+1}} \cdot \frac{B_{s+1} - B_s}{1 + \eta\mu B_{s+1}} = 1.$$

Thus we have

$$\begin{aligned}
\tau_t \cdot \delta_s &= \frac{(A_{t+1} - A_t)(1 + \eta\mu A_t)}{A_{t+1} + 2\eta\mu A_t A_{t+1} - \eta\mu A_t^2} \cdot \frac{A_{s+1} - A_s}{1 + \eta\mu A_{s+1}} \\
&\leq \frac{A_{t+1} - A_t}{A_{t+1}} \cdot \frac{A_{s+1} - A_s}{1 + \eta\mu A_{s+1}} && \text{(by } A_{t+1} \geq A_t\text{)} \\
&= \frac{B_{t+1} - B_t}{A_{t+1}} \cdot \frac{B_{s+1} - B_s}{1 + \eta\mu A_{s+1}} && \text{(by } A_{s+1} - A_s = B_{s+1} - B_s\text{)} \\
&\leq \frac{B_{t+1} - B_t}{B_{t+1}} \cdot \frac{B_{s+1} - B_s}{1 + \eta\mu B_{s+1}} \leq 1. && \text{(by } A_{s+1} \geq B_{s+1}\text{)}
\end{aligned}$$

□

Lemma B.3.5. *If $0 < \eta\mu < 1$, then for any $t \geq 1$, we have*

$$\frac{B_t}{1 - \sqrt{\eta\mu}} \leq B_{t+1} \leq \frac{3B_t}{1 - \eta\mu}.$$

Thus

$$B_t \geq \frac{1}{(1 - \sqrt{\eta\mu})^{t-1}} \geq (1 + \sqrt{\eta\mu})^{t-1}.$$

Proof of Lemma B.3.5. For $t \geq 1$, we have $B_t \geq 1$ thus

$$B_{t+1} = \frac{2B_t + 1 + \sqrt{4B_t + 4\eta\mu B_t^2 + 1}}{2(1 - \eta\mu)} \leq \frac{2B_t + 1}{1 - \eta\mu} \leq \frac{3B_t}{1 - \mu\eta}.$$

On the other hand, we have

$$\begin{aligned}
B_{t+1} &= \frac{2B_t + 1 + \sqrt{4B_t + 4\eta\mu B_t^2 + 1}}{2(1 - \eta\mu)} \\
&\geq \frac{2B_t + \sqrt{(2B_t\sqrt{\eta\mu})^2}}{2(1 - \eta\mu)} \\
&= \frac{B_t}{1 - \sqrt{\eta\mu}}.
\end{aligned}$$

Thus

$$B_t \geq \left(\frac{1}{1-\sqrt{\eta\mu}}\right)^{t-1} B_1 \geq \left(\frac{1}{1-\sqrt{\eta\mu}}\right)^{t-1} \geq (1+\sqrt{\eta\mu})^{t-1}.$$

□

Lemma B.3.6. *For $0 < \eta\mu < 1$ and $t \geq 1$, we have*

$$\sum_{s=0}^{t} \sqrt{B_s} \leq (1-\eta\mu) B_{t+1} \leq 3B_t.$$

Proof of Lemma B.3.6.

$$\begin{aligned} B_{t+1} &= \frac{2B_t + 1 + \sqrt{4B_t + 4\eta\mu B_t^2 + 1}}{2(1-\eta\mu)} \\ &\geq B_t + \frac{\sqrt{B_t}}{1-\eta\mu} \\ &\geq \cdots \\ &\geq \sum_{s=0}^{t} \frac{\sqrt{B_s}}{1-\eta\mu}. \end{aligned}$$

Combined with Lemma B.3.5, we have the desired result. □

Lemma B.3.7. *For $t \geq 1$, we have*

$$\sum_{s=0}^{t-1} \frac{\sqrt{A_{s+1}}}{A_t} \leq 3 + 4\log(e + \frac{1}{\eta\mu}).$$

Proof of Lemma B.3.7. By Lemma B.3.5, we have

$$A_t = B_t + \frac{1}{\eta\mu} \geq (1+\sqrt{\eta\mu})^{t-1} + \frac{1}{\eta\mu}. \tag{B.15}$$

Thus, we have

$$
\begin{aligned}
\sum_{s=0}^{t-1} \frac{\sqrt{A_{s+1}}}{A_t} &= \sum_{s=0}^{t-1} \frac{\sqrt{B_{s+1} + 1/(\eta\mu)}}{A_t} \\
&\leq \sum_{s=0}^{t-1} \frac{\sqrt{B_{s+1}}}{A_t} + \frac{t}{\sqrt{\eta\mu} A_t} \\
&\leq 3 + \frac{1}{\sqrt{\eta\mu} A_t} \cdot \frac{2}{\sqrt{\eta\mu}} \log(e + \frac{1}{\eta\mu})(\eta\mu(1+\sqrt{\eta\mu})^t + 1) \\
&\qquad\qquad \text{(by Lemma B.3.6 and Lemma B.3.3)} \\
&\leq 3 + 4\log(e + \frac{1}{\eta\mu}). \qquad\qquad \text{(by Inequality (B.15))}
\end{aligned}
$$

□

B.3.2 Proof of Theorem B.3.1

With all the useful lemmas in the previous section, we proceed to prove Theorem B.3.1, for which we need several additional lemmas. First, similar to Lemma B.2.4, the following lemma summarizes the results in the classical potential function analysis of NAG for strongly convex functions in [d'Aspremont et al., 2021].

Lemma B.3.8. *For any $t \geq 0$, if the following inequality holds*

$$
f(y_t) + \langle \nabla f(y_t), x_{t+1} - y_t \rangle + \frac{1}{2\eta} \|x_{t+1} - y_t\|^2 \geq f(x_{t+1}),
$$

then we can obtain

$$
A_{t+1}(f(x_{t+1}) - f^*) + \frac{1 + \eta\mu A_{t+1}}{2\eta} \|z_{t+1} - x^*\|^2 \leq A_t(f(x_t) - f^*) + \frac{1 + \eta\mu A_t}{2\eta} \|z_t - x^*\|^2 .
$$

Proof of Lemma B.3.8. These derivations can be found in d'Aspremont et al. [2021]. We present it here for completeness.

The strong convexity between x^* and y_t gives

$$f^* \geq f(y_t) + \langle \nabla f(y_t), x^* - y_t \rangle + \frac{\mu}{2} \left\| x^* - y_t \right\|^2 .$$

The convexity between x_t and y_t gives

$$f(x_t) \geq f(y_t) + \langle \nabla f(y_t), x_t - y_t \rangle.$$

Combining the above two inequalities and the one assumed in this lemma, we have

$$\begin{aligned} 0 \geq{}& (A_{t+1} - A_t)(f^* - f(y_t) - \langle \nabla f(y_t), x^* - y_t \rangle - \frac{\mu}{2} \left\| x^* - y_t \right\|^2) \\ &+ A_t(f(y_t) - f(x_t) - \langle \nabla f(y_t), x_t - y_t \rangle) \\ &+ A_{t+1}(f(x_{t+1}) - f(y_t) - \langle \nabla f(y_t), x_{t+1} - y_t \rangle - \frac{1}{2\eta} \left\| x_{t+1} - y_t \right\|^2). \end{aligned}$$

Reorganizing we can obtain

$$\begin{aligned} & A_{t+1}(f(x_{t+1}) - f^*) + \frac{1 + \eta\mu A_{t+1}}{2\eta} \left\| z_{t+1} - x^* \right\|^2 \\ & \leq A_t(f(x_t) - f^*) + \frac{1 + \eta\mu A_t}{2\eta} \left\| z_t - x^* \right\|^2 \\ & \quad + \frac{(A_t - A_{t+1})^2 - A_{t+1} - \eta\mu A_{t+1}^2}{1 + \eta\mu A_{t+1}} \frac{\eta}{2} \left\| \nabla f(y_t) \right\|^2 \\ & \quad - A_t^2 \frac{(A_{t+1} - A_t)(1 + \eta\mu A_t)(1 + \eta\mu A_{t+1})}{(A_{t+1} + 2\eta\mu A_t A_{t+1} - \eta\mu A_t^2)^2} \frac{\mu}{2} \left\| x_t - z_t \right\|^2 . \end{aligned}$$

Then we complete the proof noting that

$$\begin{aligned}
&(A_t - A_{t+1})^2 - A_{t+1} - \eta\mu A_{t+1}^2 \\
&= (B_t - B_{t+1})^2 - B_{t+1} + \frac{1}{\eta\mu} - \eta\mu(B_{t+1} + 1/(\eta\mu))^2 \\
&= \eta\mu B_{t+1}^2 + \frac{1}{\eta\mu} - \eta\mu B_{t+1}^2 - 2B_{t+1} - \frac{1}{\eta\mu} \\
&= -2B_{t+1} \leq 0.
\end{aligned}$$

□

Next, note that Lemma B.2.5 still holds in the strongly convex setting. We repeat it below for completeness.

Lemma B.3.9. *For any $t \geq 0$, if $\|\nabla f(y_t)\| \leq G$, then we have $\|\nabla f(x_{t+1})\| \leq G$, and furthermore,*

$$f(y_t) + \langle \nabla f(y_t), x_{t+1} - y_t \rangle + \frac{1}{2\eta} \|x_{t+1} - y_t\|^2 \geq f(x_{t+1}).$$

With Lemma B.3.8 and Lemma B.3.9, we will show that $\|\nabla f(y_t)\| \leq G$ for all $t \geq 0$ by induction in the following lemma.

Lemma B.3.10. *For all $t \geq 0$, we have $\|\nabla f(y_t)\| \leq G$.*

Proof of Lemma B.3.10. We will prove this lemma by induction. First, by Lemma 2.2.3 and the choice of G, it is easy to verify that $\|\nabla f(x_0)\| \leq G$. Then for any fixed $t \geq 0$, suppose that $\|\nabla f(x_s)\| \leq G$ for all $s < t$. Then by Lemma B.3.8 and Lemma B.3.9, we know that $\|\nabla f(x_s)\| \leq G$ for all $0 \leq s \leq t$, and that for all $s < t$,

$$A_{s+1}(f(x_{s+1}) - f^*) + \frac{1 + \eta\mu A_{s+1}}{2\eta} \|z_{s+1} - x^*\|^2 \leq A_s(f(x_s) - f^*) + \frac{1 + \eta\mu A_s}{2\eta} \|z_s - x^*\|^2 . \tag{B.16}$$

By telescoping (B.16), we have for all $0 \le s < t$,

$$f(x_{s+1}) - f^* \le \frac{1}{A_{s+1}\eta\mu}(f(x_0) - f^* + \mu \|z_0 - x^*\|^2). \tag{B.17}$$

For $0 \le s \le t$, since $\|\nabla f(x_s)\| \le G$, then Lemma 2.2.3 implies

$$\|\nabla f(x_s)\|^2 \le 2L(f(x_s) - f^*). \tag{B.18}$$

Note that by Algorithm 5, we have

$$z_t - x_t = (1 - \eta\mu\delta_{t-1})(1 - \tau_{t-1})(z_{t-1} - x_{t-1}) + \eta(1 - \delta_{t-1})\nabla f(y_{t-1}).$$

Thus

$$z_t - x_t = \eta \sum_{s=0}^{t-1}(1 - \delta_s)\nabla f(y_s) \prod_{i=s+1}^{t-1}(1 - \eta\mu\delta_i)(1 - \tau_i).$$

Therefore

$$y_t - x_t = \eta\tau_t \sum_{s=0}^{t-1}(1 - \delta_s)\nabla f(y_s) \prod_{i=s+1}^{t-1}(1 - \eta\mu\delta_i)(1 - \tau_i).$$

Moreover

$$1 - \eta\mu\delta_i = 1 - \frac{\eta\mu(A_{i+1} - A_i)}{1 + \eta\mu A_{i+1}} = \frac{1 + \eta\mu A_i}{1 + \eta\mu A_{i+1}}$$

and

$$1 - \tau_i = 1 - \frac{(A_{i+1} - A_i)(1 + \eta\mu A_i)}{A_{i+1} + 2\eta\mu A_i A_{i+1} - \eta\mu A_i^2} = \frac{A_i(1 + \eta\mu A_{i+1})}{A_{i+1} + 2\eta\mu A_i A_{i+1} - \eta\mu A_i^2} \le \frac{A_i(1 + \eta\mu A_{i+1})}{A_{i+1}(1 + \eta\mu A_i)}.$$

Thus we have

$$\|y_t - x_t\| \leq \eta\tau_t \sum_{s=0}^{t-1} (\delta_s - 1) \frac{A_{s+1}}{A_t} \|\nabla f(y_s)\| \leq \eta \sum_{s=0}^{t-1} \frac{A_{s+1}}{A_t} \|\nabla f(y_s)\| =: \mathcal{I},$$

where the second inequality follows from Lemma B.3.4. We further control term $\mathcal{I}$ by

$$\begin{aligned}
\mathcal{I} &\leq \eta \sum_{s=0}^{t-1} \frac{A_{s+1}}{A_t} (\|\nabla f(x_{s+1})\| + \eta L \|\nabla f(y_s)\|) \\
&\leq \eta L \mathcal{I} + \eta \sum_{s=0}^{t-1} \frac{A_{s+1}}{A_t} \|\nabla f(x_{s+1})\| .
\end{aligned}$$

Thus we have

$$\begin{aligned}
\|y_t - x_t\| &\leq \frac{\eta}{1-\eta L} \sum_{s=0}^{t-1} \frac{A_{s+1}}{A_t} \|\nabla f(x_{s+1})\| \\
&\leq \frac{\eta}{1-\eta L} \sum_{s=0}^{t-1} \frac{A_{s+1}}{A_t} \sqrt{2L(f(x_{s+1}) - f^*)} && \text{(by (B.18))} \\
&\leq \frac{\eta}{1-\eta L} \sum_{s=0}^{t-1} \frac{A_{s+1}}{A_t} \sqrt{2L \cdot \frac{1}{A_{s+1}\eta\mu} (f(x_0) - f^* + \mu \|z_0 - x^*\|^2)} && \text{(by (B.17))} \\
&= \frac{\sqrt{2\eta L (f(x_0) - f^* + \mu \|z_0 - x^*\|^2)}}{(1-\eta L)\sqrt{\mu}} \sum_{s=0}^{t-1} \frac{\sqrt{A_{s+1}}}{A_t} \\
&\leq \frac{\sqrt{2\eta L (f(x_0) - f^* + \mu \|z_0 - x^*\|^2)}}{(1-\eta L)\sqrt{\mu}} \left(3 + 4\log(e + \frac{1}{\eta\mu})\right). && \text{(by Lemma B.3.7)} \\
&\leq \frac{\sqrt{\eta}}{1-\eta L} \left(3 + 4\log(e + \frac{1}{\eta\mu})\right) \cdot \frac{G \cdot L^{1/2-1/\alpha}}{4} && \text{(by (B.13))} \\
&\leq \frac{3 + 4\log(e + \frac{1}{\eta\mu})}{\log^2\left(e + \frac{144 L^{3-2/\alpha}}{\mu}\right)} \cdot \frac{G}{24L} && \text{(by (B.14))} \\
&\leq \frac{G}{2L} \leq r(G).
\end{aligned}$$

Since $\|\nabla f(x_t)\| \le G$ and we just showed $\|x_t - y_t\| \le r(G)$, by Lemma 2.2.1, we have

$$\begin{aligned}
\|\nabla f(y_t)\| &\le \|\nabla f(x_t)\| + L\,\|y_t - x_t\| \\
&\le \sqrt{\frac{2L}{\eta\mu A_t}((f(x_0) - f^*) + \mu\,\|z_0 - x^*\|^2)} + L \cdot \frac{G}{2L} && \text{(by (B.17))} \\
&\le G\left(\frac{1}{4} + \frac{1}{2}\right) \le G. && \text{(by } A_t \ge 1/(\eta\mu) \text{ and (B.13))}
\end{aligned}$$

Then we complete the induction as well as the proof. □

Proof of Theorem B.3.1. Combining Lemmas B.3.8, B.3.9, and B.3.10, we know the following inequality holds for all $t \ge 0$.

$$A_{t+1}(f(x_{t+1}) - f^*) + \frac{1 + \eta\mu A_{t+1}}{2\eta}\,\|z_{t+1} - x^*\|^2 \le A_t(f(x_t) - f^*) + \frac{1 + \eta\mu A_t}{2\eta}\,\|z_t - x^*\|^2 .$$

Then by telescoping, we get

$$A_t(f(x_t) - f^*) + \frac{1 + \eta\mu A_t}{2\eta}\,\|z_t - x^*\|^2 \le A_0(f(x_0) - f^*) + \frac{1 + \eta\mu A_0}{2\eta}\,\|z_0 - x^*\|^2 .$$

Finally, applying Lemma B.3.5, we have $A_t = B_t + 1/(\eta\mu) \ge 1/(1 - \sqrt{\eta\mu})^{t-1} + 1/(\eta\mu)$. Thus completes the proof. □

B.4 Analysis of GD for non-convex functions

In this section, we provide the proofs related to analysis of gradient descent for non-convex function, including those of Lemma 3.2.1 and Theorem 3.2.2.

Proof of Lemma 3.2.1. First, based on Corollary 2.2.4, we know $\|\nabla f(x)\| \le G < \infty$. Also note that

$$\left\|x^+ - x\right\| = \|\eta\nabla f(x)\| \le \eta G \le G/L.$$

Then by Lemma 2.2.1 and Remark 2.2.2, we have $x^+ \in \mathcal{X}$ and

$$\begin{aligned} f(x^+) \leq & f(x) + \left\langle \nabla f(x), x^+ - x \right\rangle + \frac{L}{2} \left\| x^+ - x \right\|^2 \\ = & f(x) - \eta(1 - \eta L/2) \left\| \nabla f(x) \right\|^2 \\ \leq & f(x). \end{aligned}$$

□

Proof of Theorem 3.2.2. By Lemma 3.2.1, using induction, we directly obtain $f(x_t) \leq f(x_0)$ for all $t \geq 0$. Then by Corollary 2.2.4, we have $\|\nabla f(x_t)\| \leq G$ for all $t \geq 0$. Following the proof of Lemma 3.2.1, we can similarly show

$$f(x_{t+1}) - f(x_t) \leq \eta(1 - \eta L/2) - \frac{\eta}{2} \left\| \nabla f(x_t) \right\|^2 \leq -\frac{\eta}{2} \left\| \nabla f(x_t) \right\|^2.$$

Taking a summation over $t < T$ and rearanging terms, we have

$$\frac{1}{T} \sum_{t<T} \left\| \nabla f(x_t) \right\|^2 \leq \frac{2(f(x_0) - f(x_T))}{\eta T} \leq \frac{2(f(x_0) - f^*)}{\eta T}.$$

□

B.5 Analysis of SGD for non-convex functions

In this section, we provide the detailed convergence analysis of stochastic gradient descent for ℓ-smooth and non-convex functions where ℓ is sub-quadratic. We first state the Optional Stopping Theorem below useful for our analysis.

Lemma B.5.1 (Optional Stopping Theorem). *Let $\{Z_t\}_{t\geq 1}$ be a martingale with respect to a filtration $\{\mathcal{F}_t\}_{t\geq 0}$. Let τ be a bounded stopping time with respect to the same filtration. Then we have $\mathbb{E}[Z_\tau] = \mathbb{E}[Z_0]$.*

Next, we present some useful inequalities related to the parameter choices in Theorem 3.2.3.

Lemma B.5.2. *The parameters choices in Theorem 3.2.3 are valid and the following inequalities hold.*

$$\eta G\sqrt{2T} \leq 1/2, \quad \eta^2\sigma LT \leq 1/2, \quad 100\eta^2 T\sigma^2 L^2 \leq \delta G^2.$$

Proof of Lemma B.5.2. The parameter choices are valide because we have $\frac{F}{\eta\epsilon^2} \leq \frac{1}{16G^2\eta^2}$ by the choice of η. Then, note that by Corollary 2.2.4, we know

$$G^2 = 2LF = 16L(f(x_0) - f^* + \sigma)/\delta \geq 16L\sigma/\delta,$$

i.e., $\sigma L \leq G^2\delta/16$. Note that $\eta \leq \frac{1}{4G\sqrt{T}}$ by the choice of T, we have

$$\begin{aligned} \eta G\sqrt{2T} \leq& \sqrt{2}/4 \leq 1/2, \\ \eta^2\sigma LT \leq& \eta^2 TG^2\delta/16 \leq \delta/256 \leq 1/2, \\ 100\eta^2 T\sigma^2 L^2 \leq& 100\eta^2 TG^4\delta^2/256 \leq \delta G^2. \end{aligned}$$

□

Next, we show the useful lemma which bounds $\mathbb{E}[f(x_\tau) - f^*]$ and $\mathbb{E}\left[\sum_{t<\tau} \|\nabla f(x_t)\|^2\right]$ simultaneously.

Lemma B.5.3. *Under the parameters choices in Theorem 3.2.3, the following inequality holds*

$$\mathbb{E}\left[f(x_\tau) - f^* + \frac{\eta}{2}\sum_{t<\tau} \|\nabla f(x_t)\|^2\right] \leq f(x_0) - f^* + \sigma.$$

Proof of Lemma B.5.3. If $t < \tau$, by the definition of τ, we know $f(x_t) - f^* \leq F$ and $\|\epsilon_t\| \leq \frac{G}{5\eta L}$, and the former also implies $\|\nabla f(x_t)\| \leq G$ by Corollary 2.2.4. Then we can bound

$$\|x_{t+1}-x_t\| = \eta \left\|\nabla f(x_t,\xi_t)\right\| \leq \eta(\|\nabla f(x_t)\| + \|\epsilon_t\|) \leq \eta G + \frac{G}{5L} \leq \frac{G}{L},$$

where we use the choice of $\eta \leq \frac{1}{2L}$. Then based on Lemma 2.2.1 and Remark 2.2.2, for any $t < \tau$, we have

$$\begin{aligned} f(x_{t+1}) - f(x_t) \leq & \Big\langle \nabla f(x_t), x_{t+1}-x_t \Big\rangle + \frac{L}{2}\left\|x_{t+1}-x_t\right\|^2 \\ = & -\eta\Big\langle \nabla f(x_t), \nabla f(x_t,\xi_t)\Big\rangle + \frac{\eta^2 L}{2}\left\|\nabla f(x_t,\xi_t)\right\|^2 \\ \leq & -\eta\left\|\nabla f(x_t)\right\|^2 - \eta\Big\langle \nabla f(x_t),\epsilon_t\Big\rangle + \eta^2 L\left\|\nabla f(x_t)\right\|^2 + \eta^2 L\left\|\epsilon_t\right\|^2 \\ \leq & -\frac{\eta}{2}\left\|\nabla f(x_t)\right\|^2 - \eta\Big\langle \nabla f(x_t),\epsilon_t\Big\rangle + \eta^2 L\left\|\epsilon_t\right\|^2, \end{aligned} \tag{B.19}$$

where the equality is due to (3.2); the second inequality uses $g_t = \epsilon_t + \nabla f(x_t)$ and Young's inequality $\|y+z\|^2 \leq 2\|y\|^2 + 2\|z\|^2$ for any vectors y, z; and the last inequality chooses $\eta \leq 1/(2L)$. Taking a summation over $t < \tau$ and rearanging terms, we have

$$f(x_\tau) - f^* + \frac{\eta}{2}\sum_{t<\tau}\left\|\nabla f(x_t)\right\|^2 \leq f(x_0) - f^* - \eta\sum_{t<\tau}\Big\langle \nabla f(x_t),\epsilon_t\Big\rangle + \eta^2 L\sum_{t<\tau}\left\|\epsilon_t\right\|^2.$$

Now we bound the last two terms on th RHS. First, for the last term, we have

$$\mathbb{E}\left[\sum_{t<\tau}\left\|\epsilon_t\right\|^2\right] \leq \mathbb{E}\left[\sum_{t<T}\left\|\epsilon_t\right\|^2\right] \leq \sigma^2 T,$$

where the first inequality uses $\tau \leq T$ by its defnition; and in the last inequality we use Assumption 3.2.

For the cross term, note that $\mathbb{E}_{t-1}\left[\Big\langle \nabla f(x_t),\epsilon_t\Big\rangle\right] = 0$ by Assumption 3.2. So this term is a sum of a martingale difference sequence. Since τ is a stopping time, we can apply the

Optional Stopping Theorem (Lemma B.5.1) to obtain

$$\mathbb{E}\left[\sum_{t\le\tau}\left\langle\nabla f(x_t),\epsilon_t\right\rangle\right]=0. \tag{B.20}$$

Then we have

$$\begin{aligned}\mathbb{E}\left[-\sum_{t<\tau}\left\langle\nabla f(x_t),\epsilon_t\right\rangle\right]=&\mathbb{E}\left[\left\langle\nabla f(x_\tau),\epsilon_\tau\right\rangle\right]\le G\,\mathbb{E}[\|\epsilon_\tau\|]\le G\sqrt{\mathbb{E}[\|\epsilon_\tau\|^2]}\\ &\le G\sqrt{\mathbb{E}\left[\sum_{t\le T}\|\epsilon_t\|^2\right]}\le\sigma G\sqrt{T+1}\le\sigma G\sqrt{2T},\end{aligned}$$

where the equality is due to (B.20); the first inequality uses $\|\nabla f(x_\tau)\|\le G$ by the definition of τ in (3.3) and Corollary 2.2.4; the fourth inequality uses $\mathbb{E}[X]^2\le\mathbb{E}[X^2]$ for any random variable X; and the last inequality uses Assumption 3.2.

Combining all the bounds above, we get

$$\begin{aligned}\mathbb{E}\left[f(x_\tau)-f^*+\frac{\eta}{2}\sum_{t<\tau}\|\nabla f(x_t)\|^2\right]\le&f(x_0)-f^*+\eta\sigma G\sqrt{2T}+\eta^2\sigma^2LT\\ \le&f(x_0)-f^*+\sigma,\end{aligned}$$

where the last inequality is due to Lemma B.5.2. □

With Lemma B.5.3, we are ready to prove Theorem 3.2.3.

Proof of Theorem 3.2.3. We want to show the probability of $\{\tau<T\}$ is small, as its complement $\{\tau=T\}$ means $f(x_t)-f^*\le F$ for all $t\le T$ which implies $\|\nabla f(x_t)\|\le G$ for all $t\le T$. Note that

$$\{\tau<T\}=\{\tau_2<T\}\cup\{\tau_1<T,\tau_2=T\}.$$

Therefore we only need to bound the probability of each of these two events on the RHS.

We first bound $\mathbb{P}(\tau_2 < T)$. Note that

$$\begin{aligned}\mathbb{P}(\tau_2 < T) =& \mathbb{P}\left(\bigcup_{t<T}\left\{\|\epsilon_t\| > \frac{G}{5\eta L}\right\}\right)\\ \leq& \sum_{t<T}\mathbb{P}\left(\|\epsilon_t\| > \frac{G}{5\eta L}\right)\\ \leq& \frac{25\eta^2 T\sigma^2 L^2}{G^2}\\ \leq& \delta/4,\end{aligned}$$

where the first inequality uses union bound; the second inequality applies Chebyshev's inequality and $\mathbb{E}[\|\epsilon_t\|^2] = \mathbb{E}[\mathbb{E}_{t-1}[\|\epsilon_t\|^2]] \leq \sigma^2$ for each fixed t by Assumption 3.2; the last inequality uses Lemma B.5.2.

Next, we will bound $\mathbb{P}(\tau_1 < T, \tau_2 = T)$. Note that under the event $\{\tau_1 < T, \tau_2 = T\}$, we know that 1) $\tau = \tau_1 < T$ which implies $f(x_{\tau+1}) - f^* > F$; and 2) $\tau < T = \tau_2$ which implies $\|\epsilon_\tau\| \leq \frac{G}{5\eta L}$ by the definition in (3.3). Also note that we always have $f(x_\tau) - f^* \leq F$ which implies $\|\nabla f(x_\tau)\| \leq G$ by Corollary 2.2.4. Then we can show

$$\|x_{\tau+1} - x_\tau\| = \eta\left\|\nabla f(x_\tau, \xi_\tau)\right\| \leq \eta(\|\nabla f(x_\tau)\| + \|\epsilon_\tau\|) \leq \eta G + \frac{G}{5L} \leq \frac{G}{L},$$

where we choose $\eta \leq \frac{1}{2L}$. Then based on Lemma 2.2.1 and Remark 2.2.2, we have

$$\begin{aligned}f(x_{\tau+1}) - f(x_\tau) \leq& -\frac{\eta}{2}\left\|\nabla f(x_\tau)\right\|^2 - \eta\left\langle\nabla f(x_\tau), \epsilon_\tau\right\rangle + \eta^2 L\left\|\epsilon_\tau\right\|^2\\ \leq& \eta\left\|\nabla f(x_\tau)\right\| \cdot \left\|\epsilon_\tau\right\| + \eta^2 L\left\|\epsilon_\tau\right\|^2\\ \leq& \frac{G^2}{4L}\\ =& \frac{F}{2},\end{aligned}$$

where the first inequality is obtained following the same derivation as in (B.19); the last equality is due to Corollary 2.2.4. Therefore we can show that under the event $\{\tau_1 < T, \tau_2 =$

$T\}$,

$$f(x_\tau) - f^* = f(x_\tau) - f(x_{\tau+1}) + f(x_{\tau+1}) - f^* > F/2.$$

Hence,

$$\mathbb{P}(\tau_1 < T, \tau_2 = T) \leq \mathbb{P}\left(f(x_\tau) - f^* > F/2\right) \leq \frac{\mathbb{E}[f(x_\tau) - f^*]}{F/2} \leq \frac{2(f(x_0) - f^* + \sigma)}{F} = \delta/4,$$

where the second inequality uses Markov's inequality; the third inequality uses Lemma B.5.3; and in the last inequality we choose $F = 8(f(x_0) - f^* + \sigma)/\delta$.

Therefore we can show

$$\mathbb{P}(\tau < T) \leq \mathbb{P}(\tau_2 < T) + \mathbb{P}(\tau_1 < T, \tau_2 = T) \leq \delta/2.$$

Then we also know $\mathbb{P}(\tau = T) \geq 1 - \delta/2 \geq 1/2$. Therefore, by Lemma B.5.3,

$$\begin{aligned}\frac{2(f(x_0) - f^* + \sigma)}{\eta} \geq & \mathbb{E}\left[\sum_{t<\tau} \|\nabla f(x_t)\|^2\right] \\ \geq & \mathbb{P}(\tau = T)\mathbb{E}\left[\sum_{t<T} \|\nabla f(x_t)\|^2 \middle| \tau = T\right] \\ \geq & \frac{1}{2}\mathbb{E}\left[\sum_{t<T} \|\nabla f(x_t)\|^2 \middle| \tau = T\right].\end{aligned}$$

Then we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t<T} \|\nabla f(x_t)\|^2 \middle| \tau = T\right] \leq \frac{4(f(x_0) - f^* + \sigma)}{\eta T} = \frac{\delta F}{2\eta T} \leq \frac{\delta}{2} \cdot \epsilon^2,$$

where the last inequality uses the choice of T. Let $\mathcal{E} := \{\frac{1}{T}\sum_{t<T} \|\nabla f(x_t)\|^2 > \epsilon^2\}$ denote the event of not converging to an ϵ-stationary point. By Markov's inequality, we have $\mathbb{P}(\mathcal{E}) \leq \delta/2$. Therefore we have $\mathbb{P}(\{\tau < T\} \cup \mathcal{E}) \leq \delta$, which completes the proof. □

B.6 Lower bound

In this section, we provide the proof of Theorem 3.2.4.

Proof of Theorem 3.2.4. Let $c, \eta_0 > 0$ satisfy $\eta_0 \le c^2/2$. Consider

$$f(x) = \begin{cases} \log(|x| - c), & |x| \ge y \\ 2\log(y - c) - \log(2y - |x| - c), & c/2 \le |x| < y \\ kx^2 + b, & |x| < c/2, \end{cases}$$

where $c > 0$ is a constant and $y = (c + \sqrt{c^2 + 2\eta_0})/2 > 0$ is the fixed point of the iteration

$$x_{t+1} = \left| x_t - \frac{\eta_0}{x_t - c} \right|,$$

and k, b are chosen in such a way that $f(x)$ and $f'(x)$ are continuous. Specifically, choose $k = c^{-1} f'(c/2)$ and $b = f(c/2) - cf'(c/2)/4$. Since $f(-x) = f(x)$, $f(x)$ is symmetric about the line $x = 0$. In a small neighborhood, $f(x)$ is symmetric about $(y, f(y))$, so $f'(x)$ is continuous at y.

Let us first consider the smoothness of f. By symmetry, it suffices to consider $x > 0$. Then,

$$f'(x) = \begin{cases} (x - c)^{-1}, & x \ge y \\ (2y - x - c)^{-1}, & c/2 \le x < y \\ 2kx, & 0 < x < c/2. \end{cases}$$

Its Hessian is given by

$$f''(x) = \begin{cases} -(x-c)^{-2}, & x > y \\ (2y-x-c)^{-2}, & c/2 < x < y \\ 2k, & 0 < x < c/2. \end{cases}$$

Hence, $f(x)$ is $(2, 2k, 1)$-smooth.

Note that $f(x)$ has a stationary point 0. For stepsize η_f satisfying $\eta_0 \leq \eta_f \leq c^2/4$, there exists $z = (c + \sqrt{c^2 + 2\eta_f}) \geq y$ such that $-z = z - \eta_f(y-c)^{-1}$ and by symmetry, once $x_\tau = z$, $x_t = \pm z$ for all $t \geq \tau$, making the GD iterations stuck. Now we choose a proper x_0 such that $f'(x_0)$ and $f(x_0) - f(0)$ are bounded.

We consider arriving at y from above. That is, $x_0 \geq x_1 \geq \ldots x_\tau = z > c > 0$. Since in each update where $x_{t+1} = x_t - \eta_f(x_t - c)^{-1} > c$,

$$x_t - x_{t+1} = x_t - (x_t - \eta_f(x_t-c)^{-1}) = \eta_f(x_t - c)^{-1} \leq \sqrt{\eta_f}.$$

Hence, we can choose τ in such a way that $3c/2 \leq x_0 < 3c/2 + \sqrt{\eta_f}$. Then,

$$\log(c/2) \leq f(x_0) \leq \log(c/2 + \sqrt{\eta_f}), \quad 2/(c + 2\sqrt{\eta_f}) \leq f'(x_0) \leq 2/c.$$

By definition, $y - c = \eta_0(c + \sqrt{c^2 + 2\eta_0})^{-1}$. Hence,

$$\begin{aligned} f(c/2) &= 2\log(y-c) - \log(2y - c/2 - c) \\ &= 2\log(\eta_0) - 2\log(c + \sqrt{c^2+2\eta_0}) - \log(\sqrt{c^2+2\eta_0} - c/2), \\ f'(c/2) &= \frac{1}{\sqrt{c^2+2\eta_0} - c/2} \end{aligned}$$

Then,

$$\begin{aligned} f(x_0) - f(0) &= f(x_0) - f(c/2) + cf'(c/2)/4 \\ &\le \log(c/2 + \sqrt{\eta_f}) + 2\log(\eta_0^{-1}) + 2\log(c + \sqrt{c^2 + 2\eta_0}) \\ &\qquad + \log(\sqrt{c^2 + 2\eta_0} - c/2) + \frac{c}{4}\frac{1}{\sqrt{c^2 + 2\eta_0} - c/2} \\ &\le \log(c) + 2\log(\eta_0^{-1}) + 2\log(2\sqrt{2c^2}) + \log(\sqrt{2c^2}) + \frac{1}{2} \\ &= 4\log(c) + 2\log(\eta_0^{-1}) + \frac{7}{2}\log(2) + \frac{1}{2}. \end{aligned}$$

For stepsize $\eta_f < \eta_0$, reaching below $4c/3$ takes at least

$$(x_0 - 4c/3)/\sqrt{\eta_f} \ge c/(6\sqrt{\eta_f}) > c\eta_0^{-1/2}/6$$

steps to reach $4c/3$, where $f'(4c/3) = \log(c/3)$.

Now we set c and η_0 and scale function $f(x)$ to satisfy the parameter specifications L_0, L_2, G_0, Δ_0. Define $g(x) = L_2^{-1} f(x)$. Then, $g(x)$ is $(2, 2kL_2^{-1}, L_2)$-smooth. Since the gradient of $g(x)$ is L_2^{-1} times $f(x)$, the above analysis for $f(x)$ applies to $g(x)$ by replacing η_0 with $\eta_1 = L_2\eta_0$ and η_f with $\eta = L_2\eta_f$. To ensure that

$$2kL_2^{-1} = 2(cL_2)^{-1}f'(c/2) = \frac{2}{cL_2}\frac{1}{\sqrt{c^2 + 2\eta_1} - c/2} \le \frac{4}{c^2L_2} \le L_0,$$

it suffices to take $c \ge 2/\sqrt{L_0L_2}$. To ensure that

$$g'(x_0) \le \frac{2}{L_2c} \le G_0,$$

it suffices to take $c \ge 2/(L_2G_0)$. To ensure that

$$g(x_0) - g(0) \le (4\log(c) + 2\log(\eta_1^{-1}) + 3.5\log 2 + 0.5)L_2^{-1} \le \Delta_0,$$

it suffices to take

$$\log(\eta_1^{-1}) = \frac{L_2\Delta_0 - 3.5\log 2 - 0.5}{2} - 2\log(c).$$

Since we require $\eta_1 \leq c^2/2$, parameters L_2 and Δ_0 need to satisfy

$$\log 2 - 2\log(c) \leq \frac{L_2\Delta_0 - 3.5\log 2 - 0.5}{2} - 2\log(c),$$

that is, $L_2\Delta_0 \geq 5.5\log 2 + 0.5$, which holds because $L_2\Delta_0 \geq 10$. Take $c = \max\{2/\sqrt{L_0L_2}, 2/(L_2G_0), \sqrt{8/L_0}\}$. Then, as long as $\eta \leq 2/L_0$, the requirement that $\eta \leq c^2/4$ is satisfied. Therefore, on $g(x)$ with initial point x_0, gradient descent with a constant stepsize either gets stuck, or takes at least

$$\begin{aligned} c\eta_1^{-1/2}/6 &= \frac{c}{6}\exp\left(\frac{L_2\Delta_0 - 3.5\log 2 - 0.5}{4} - \log(c)\right) \\ &= \frac{1}{6}\exp(\frac{L_2\Delta_0 - 3.5\log 2 - 0.5}{4}) \\ &\geq \frac{1}{6}\exp(\frac{L_2\Delta_0}{8}) \end{aligned}$$

steps to reach a 1-stationary point.

On the other hand, if $\eta > 2/L_0$, consider the function $f(x) = \frac{L_0}{2}x^2$. For any $x_t \neq 0$, we always have $|x_{t+1}| / |x_t| = |1 - \eta L_0| > 1$, which means the iterates diverge to infinity. □

Appendix C

Proofs for Chapter 4

C.1 Covergence Analysis of Adam

In this section, we provide detailed convergence analysis of Adam. We will focus on proving Theorem 4.2.1 under the bounded noise assumption (Assumption 4.2) in most parts of this section except Appendix C.1.5 where we will show how to generalize the results to noise with sub-Gaussian norm (Assumption 4.3) and provide the proof of Theorem 4.2.2.

For completeness, we repeat some important technical definitions here. First, we define

$$\epsilon_t := \hat{m}_t - \nabla f(x_t) \tag{C.1}$$

as the deviation of the re-scaled momentum from the actual gradient. Given a large enough constant G defined in Theorem 4.2.1, denoting $F = \frac{G^2}{2(L_0+L_\rho(2G)^\rho)}$, we formally define the stopping time τ as

$$\tau := \min\{t \mid f(x_t) - f^* > F\} \wedge (T+1),$$

i.e., τ is the first time when the sub-optimality gap is strictly greater than F, truncated at $T+1$ to make sure it is bounded in order to apply Lemma B.5.1. Based on Corollary 2.2.4

and the discussions in Section 4.3.1, we know that if $t < \tau$, we have both $f(x_t) - f^* \leq F$ and $\|\nabla f(x_t)\| \leq G$. It is clear to see that τ is a stopping time[1] with respect to $\{\xi_t\}_{t\geq 1}$ because the event $\{\tau \geq t\}$ is a function of $\{\xi_s\}_{s<t}$ and independent of $\{\xi_s\}_{s\geq t}$. Next, let

$$h_t := \frac{\eta}{\sqrt{\hat{v}_t} + \lambda}$$

be the stepsize vector and $H_t := \mathrm{diag}(h_t)$ be the diagonal stepsize matrix. Then the update rule can be written as

$$x_{t+1} = x_t - h_t \odot \hat{m}_t = x_t - H_t \hat{m}_t.$$

Finally, as in Lemma 2.2.1 and Lemma 4.3.1, we define the following constants with slight notation abuse.

$$\begin{aligned} &L := L_0 + L_\rho (2G)^\rho, \\ &r := r(G) = G/L, \\ &D := 2G/\lambda. \end{aligned}$$

C.1.1 Useful lemmas for Adam

In this section, we list several useful lemmas for the convergence analysis. Their proofs are all deferred in Appendix C.1.4.

First note that when $t < \tau$, all the quantities in the algorithm are well bounded. In particular, we have the following lemma.

[1] Indeed, $\tau - 1$ is also a stopping time because $\nabla f(x_t)$ only depends on $\{\xi_s\}_{s<t}$, but that is unnecessary for our analysis.

Lemma C.1.1. *If $t < \tau$, we have*

$$\begin{aligned}
&\|\nabla f(x_t)\| \le G, \quad \|\nabla f(x_t, \xi_t)\| \le G + \sigma, \quad \|\hat{m}_t\| \le G + \sigma, \\
&\hat{v}_t \preceq (G+\sigma)^2, \quad \frac{\eta}{G+\sigma+\lambda} \preceq h_t \preceq \frac{\eta}{\lambda}.
\end{aligned}$$

Next, we provide a useful lemma regarding the time-dependent re-scaled momentum parameters in (4.1).

Lemma C.1.2. *Let $\alpha_t = \frac{\beta}{1-(1-\beta)^t}$, then for all $T \ge 2$, we have $\sum_{t=2}^{T} \alpha_t^2 \le 3(1+\beta^2 T)$.*

In the next lemma, we provide an almost sure bound on ϵ_t in order to apply Azuma-Hoeffding inequality (Lemma C.1.10).

Lemma C.1.3. *Denote $\gamma_{t-1} = (1-\alpha_t)(\epsilon_{t-1} + \nabla f(x_{t-1}) - \nabla f(x_t))$. Choosing $\eta \le \min\left\{\frac{r}{D}, \frac{\sigma\beta}{DL}\right\}$, if $t \le \tau$, we have $\|\epsilon_t\| \le 2\sigma$ and $\|\gamma_{t-1}\| \le 2\sigma$.*

Finally, the following lemma hides messy calculations and will be useful in the contradiction argument.

Lemma C.1.4. *Denote*

$$\begin{aligned}
I_1 :=& \frac{8G}{\eta\lambda}\left(\Delta_1\lambda + 8\sigma^2\left(\frac{\eta}{\beta} + \eta\beta T\right) + 20\sigma^2\eta\sqrt{(1/\beta^2 + T)\iota}\right), \\
I_2 :=& \frac{8GF}{\eta} = \frac{4G^3}{\eta L}.
\end{aligned}$$

Under the parameter choices in either Theorem 4.2.1 or Theorem 4.2.2, we have $I_1 \le I_2$ and $I_1/T \le \epsilon^2$.

C.1.2 Proof of Theorem 4.2.1

Before proving the main theorems, several important lemmas are needed. First, we provide a descent lemma for Adam.

Lemma C.1.5. *If $t < \tau$, choosing $G \geq \sigma + \lambda$ and $\eta \leq \min\left\{\frac{r}{D}, \frac{\lambda}{6L}\right\}$, we have*

$$f(x_{t+1}) - f(x_t) \leq -\frac{\eta}{4G}\left\|\nabla f(x_t)\right\|^2 + \frac{\eta}{\lambda}\left\|\epsilon_t\right\|^2.$$

Proof of Lemma C.1.5. By Lemma C.1.1, we have if $t < \tau$,

$$\frac{\eta I}{2G} \leq \frac{\eta I}{G+\sigma+\lambda} \preceq H_t \preceq \frac{\eta I}{\lambda}. \tag{C.2}$$

Since we choose $\eta \leq \frac{r}{D}$, by Lemma 4.3.1, we have $\|x_{t+1} - x_t\| \leq r$ if $t < \tau$. Then we can apply Lemma 2.2.1 to show that for any $t < \tau$,

$$\begin{aligned}
f(x_{t+1}) - f(x_t) &\leq \left\langle \nabla f(x_t), x_{t+1} - x_t \right\rangle + \frac{L}{2}\left\|x_{t+1} - x_t\right\|^2 \\
&= -(\nabla f(x_t))^\top H_t \hat{m}_t + \frac{L}{2}\hat{m}_t^\top H_t^2 \hat{m}_t \\
&\leq -\left\|\nabla f(x_t)\right\|_{H_t}^2 - (\nabla f(x_t))^\top H_t \epsilon_t + \frac{\eta L}{2\lambda}\left\|\hat{m}_t\right\|_{H_t}^2 \\
&\leq -\frac{2}{3}\left\|\nabla f(x_t)\right\|_{H_t}^2 + \frac{3}{4}\left\|\epsilon_t\right\|_{H_t}^2 + \frac{\eta L}{\lambda}\left(\left\|\nabla f(x_t)\right\|_{H_t}^2 + \left\|\epsilon_t\right\|_{H_t}^2\right) \\
&\leq -\frac{1}{2}\left\|\nabla f(x_t)\right\|_{H_t}^2 + \left\|\epsilon_t\right\|_{H_t}^2 \\
&\leq -\frac{\eta}{4G}\left\|\nabla f(x_t)\right\|^2 + \frac{\eta}{\lambda}\left\|\epsilon_t\right\|^2,
\end{aligned}$$

where the second inequality uses (C.1) and (C.2); the third inequality is due to Young's inequality $a^\top A b \leq \frac{1}{3}\|a\|_A^2 + \frac{3}{4}\|b\|_A^2$ and $\|a+b\|_A^2 \leq 2\|a\|_A^2 + 2\|b\|_A$ for any PSD matrix A; the second last inequality uses $\eta \leq \frac{\lambda}{6L}$; and the last inequality is due to (C.2). □

The following lemma bounds the sum of the error term $\|\epsilon_t\|^2$ before the stopping time τ. Since its proof is complicated, we defer it in Appendix C.1.3.

Lemma C.1.6. *If $G \geq 2\sigma$ and $\eta \leq \min\left\{\frac{r}{D}, \frac{\lambda^{3/2}\beta}{6L\sqrt{G}}, \frac{\sigma\beta}{DL}\right\}$, with probability $1-\delta$,*

$$\sum_{t=1}^{\tau-1}\left\|\epsilon_t\right\|^2 - \frac{\lambda}{8G}\left\|\nabla f(x_t)\right\|^2 \leq 8\sigma^2\left(1/\beta + \beta T\right) + 20\sigma^2\sqrt{(1/\beta^2 + T)\log(1/\delta)}.$$

Combining Lemma C.1.5 and Lemma C.1.6, we obtain the following useful lemma, which simultaneously bounds $f(x_t) - f^*$ and $\sum_{t=1}^{\tau-1} \|\nabla f(x_t)\|^2$.

Lemma C.1.7. *If $G \geq 2\max\{\lambda, \sigma\}$ and $\eta \leq \min\left\{\frac{r}{D}, \frac{\lambda^{3/2}\beta}{6L\sqrt{G}}, \frac{\sigma\beta}{DL}\right\}$, then with probability at least $1-\delta$,*

$$\begin{aligned}&\sum_{t=1}^{\tau-1} \|\nabla f(x_t)\|^2 + \frac{8G}{\eta}(f(x_\tau) - f^*)\\ \leq&\frac{8G}{\eta\lambda}\left(\Delta_1\lambda + 8\sigma^2\left(\frac{\eta}{\beta} + \eta\beta T\right) + 20\sigma^2\eta\sqrt{(1/\beta^2 + T)\log(1/\delta)}\right).\end{aligned}$$

Proof of Lemma C.1.7. By telescoping, Lemma C.1.5 implies

$$\sum_{t=1}^{\tau-1} 2\|\nabla f(x_t)\|^2 - \frac{8G}{\lambda}\|\epsilon_t\|^2 \leq \frac{8G}{\eta}(f(x_1) - f(x_\tau)) \leq \frac{8\Delta_1 G}{\eta}. \tag{C.3}$$

Lemma C.1.6 could be written as

$$\sum_{t=1}^{\tau-1} \frac{8G}{\lambda}\|\epsilon_t\|^2 - \|\nabla f(x_t)\|^2 \leq \frac{8G}{\lambda}\left(8\sigma^2(1/\beta + \beta T) + 20\sigma^2\sqrt{(1/\beta^2 + T)\log(1/\delta)}\right). \tag{C.4}$$

(C.3) + (C.4) gives the desired result. □

With Lemma C.1.7, we are ready to complete the contradiction argument and the convergence analysis. Below we provide the proof of Theorem 4.2.1.

Proof of Theorem 4.2.1. According to Lemma C.1.7, there exists some event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}) \geq 1-\delta$, such that conditioned on $\mathcal{E}$, we have

$$\frac{8G}{\eta}(f(x_\tau) - f^*) \leq \frac{8G}{\eta\lambda}\left(\Delta_1\lambda + 8\sigma^2\left(\frac{\eta}{\beta} + \eta\beta T\right) + 20\sigma^2\eta\sqrt{(1/\beta^2 + T)\log(1/\delta)}\right) =: I_1. \tag{C.5}$$

By the definition of τ, if $\tau \leq T$, we have

$$\frac{8G}{\eta}(f(x_\tau) - f^*) > \frac{8GF}{\eta} = \frac{4G^3}{\eta L} =: I_2.$$

Based on Lemma C.1.4, we have $I_1 \leq I_2$, which leads to a contradiction. Therefore, we must have $\tau = T+1$ conditioned on $\mathcal{E}$. Then, Lemma C.1.7 also implies that under $\mathcal{E}$,

$$\frac{1}{T}\sum_{t=1}^{T-1} \|\nabla f(x_t)\|^2 \leq \frac{I_1}{T} \leq \epsilon^2,$$

where the last inequality is due to Lemma C.1.4. □

C.1.3 Proof of Lemma C.1.6

In order to prove Lemma C.1.6, we need the following several lemmas.

Lemma C.1.8. *Denote $\gamma_{t-1} = (1-\alpha_t)(\epsilon_{t-1} + \nabla f(x_{t-1}) - \nabla f(x_t))$. If $G \geq 2\sigma$ and $\eta \leq \min\left\{\frac{r}{D}, \frac{\lambda^{3/2}\beta}{6L\sqrt{G}}\right\}$, we have for every $2 \leq t \leq \tau$,*

$$\|\epsilon_t\|^2 \leq \left(1 - \frac{\alpha_t}{2}\right)\|\epsilon_{t-1}\|^2 + \frac{\lambda\beta}{16G}\|\nabla f(x_{t-1})\|^2 + \alpha_t^2\sigma^2 + 2\alpha_t\left\langle \gamma_{t-1}, \nabla f(x_t, \xi_t) - \nabla f(x_t)\right\rangle.$$

Proof of Lemma C.1.8. According to the update rule (4.1), we have

$$\begin{aligned}\epsilon_t =& (1-\alpha_t)(\epsilon_{t-1} + \nabla f(x_{t-1}) - \nabla f(x_t)) + \alpha_t(\nabla f(x_t, \xi_t) - \nabla f(x_t)) \\ =& \gamma_{t-1} + \alpha_t(\nabla f(x_t, \xi_t) - \nabla f(x_t)). \qquad \text{(C.6)}\end{aligned}$$

Since we choose $\eta \leq \frac{r}{D}$, by Lemma 4.3.1, we have $\|x_t - x_{t-1}\| \leq r$ if $t \leq \tau$. Therefore by Lemma 2.2.1, for any $2 \leq t \leq \tau$,

$$\|\nabla f(x_{t-1}) - \nabla f(x_t)\| \leq L\|x_t - x_{t-1}\| \leq \frac{\eta L}{\lambda}\|\hat{m}_{t-1}\| \leq \frac{\eta L}{\lambda}\left(\|\nabla f(x_{t-1})\| + \|\epsilon_{t-1}\|\right), \qquad \text{(C.7)}$$

Therefore

$$
\begin{aligned}
\left\|\gamma_{t-1}\right\|^2 =& \left\|(1-\alpha_t)\epsilon_{t-1}+(1-\alpha_t)(\nabla f(x_{t-1})-\nabla f(x_t))\right\|^2 \\
\leq& (1-\alpha_t)^2\left(1+\alpha_t\right)\left\|\epsilon_{t-1}\right\|^2+(1-\alpha_t)^2\left(1+\frac{1}{\alpha_t}\right)\left\|\nabla f(x_{t-1})-\nabla f(x_t)\right\|^2 \\
\leq& (1-\alpha_t)\left\|\epsilon_{t-1}\right\|^2+\frac{1}{\alpha_t}\left\|\nabla f(x_{t-1})-\nabla f(x_t)\right\|^2 \\
\leq& \,(1-\alpha_t)\left\|\epsilon_{t-1}\right\|^2+\frac{2\eta^2L^2}{\lambda^2\beta}\left(\left\|\nabla f(x_{t-1})\right\|^2+\left\|\epsilon_{t-1}\right\|^2\right) \\
\leq& \left(1-\frac{\alpha_t}{2}\right)\left\|\epsilon_{t-1}\right\|^2+\frac{\lambda\beta}{16G}\left\|\nabla f(x_{t-1})\right\|^2,
\end{aligned}
$$

where the first inequality uses Young's inequality $\left\|a+b\right\|^2 \leq (1+u)\left\|a\right\|^2+(1+1/u)\left\|b\right\|^2$ for any $u>0$; the second inequality is due to

$$
\begin{aligned}
&(1-\alpha_t)^2\left(1+\alpha_t\right)=(1-\alpha_t)(1-\alpha_t^2)\leq(1-\alpha_t), \\
&(1-\alpha_t)^2\left(1+\frac{1}{\alpha_t}\right)=\frac{1}{\alpha_t}(1-\alpha_t)^2\left(1+\alpha_t\right)\leq\frac{1}{\alpha_t}(1-\alpha_t)\leq\frac{1}{\alpha_t};
\end{aligned}
$$

the third inequality uses (C.7) and Young's inequality; and in the last inequality we choose $\eta \leq \frac{\lambda^{3/2}\beta}{6L\sqrt{G}}$, which implies $\frac{2\eta^2L^2}{\lambda^2\beta} \leq \frac{\lambda\beta}{16G} \leq \frac{\beta}{2} \leq \frac{\alpha_t}{2}$. Then by (C.6), we have

$$
\begin{aligned}
\left\|\epsilon_t\right\|^2 =& \left\|\gamma_{t-1}\right\|^2+2\alpha_t\left\langle\gamma_{t-1},\nabla f(x_t,\xi_t)-\nabla f(x_t)\right\rangle+\alpha_t^2\left\|\nabla f(x_t,\xi_t)-\nabla f(x_t)\right\|^2 \\
\leq& \left(1-\frac{\alpha_t}{2}\right)\left\|\epsilon_{t-1}\right\|^2+\frac{\lambda\beta}{16G}\left\|\nabla f(x_{t-1})\right\|^2+\alpha_t^2\sigma^2+2\alpha_t\left\langle\gamma_{t-1},\nabla f(x_t,\xi_t)-\nabla f(x_t)\right\rangle.
\end{aligned}
$$

□

Lemma C.1.9. *Denote $\gamma_{t-1}=(1-\alpha_t)(\epsilon_{t-1}+\nabla f(x_{t-1})-\nabla f(x_t))$. If $G \geq 2\sigma$ and $\eta \leq \min\left\{\frac{r}{D},\frac{\sigma\beta}{DL}\right\}$, with probability $1-\delta$,*

$$
\sum_{t=2}^{\tau}\alpha_t\left\langle\gamma_{t-1},\nabla f(x_t,\xi_t)-\nabla f(x_t)\right\rangle \leq 5\sigma^2\sqrt{(1+\beta^2T)\log(1/\delta)}.
$$

In order to prove Lemma C.1.9, we need the Azuma-Hoeffding inequality stated below without proofs.

Lemma C.1.10 (Azuma-Hoeffding inequality)**.** *Let $\{Z_t\}_{t\geq 1}$ be a martingale with respect to a filtration $\{\mathcal{F}_t\}_{t\geq 0}$. Assume that $|Z_t - Z_{t-1}| \leq c_t$ almost surely for all $t \geq 0$. Then for any fixed T, with probability at least $1-\delta$,*

$$Z_T - Z_0 \leq \sqrt{2\sum_{t=1}^{T} c_t^2 \log(1/\delta)}.$$

Now we are ready to prove Lemma C.1.9.

Proof of Lemma C.1.9. First note that

$$\sum_{t=2}^{\tau} \alpha_t \Big\langle \gamma_{t-1}, \nabla f(x_t,\xi_t) - \nabla f(x_t) \Big\rangle = \sum_{t=2}^{T} \alpha_t \Big\langle \gamma_{t-1} 1_{\tau\geq t}, \nabla f(x_t,\xi_t) - \nabla f(x_t) \Big\rangle.$$

Since τ is a stopping time, we know that $1_{\tau\geq t}$ is a function of $\{\xi_s\}_{s<t}$. Also, by definition, we know γ_{t-1} is a function of $\{\xi_s\}_{s<t}$. Then, denoting

$$X_t = \alpha_t \Big\langle \gamma_{t-1} 1_{\tau\geq t}, \nabla f(x_t,\xi_t) - \nabla f(x_t) \Big\rangle,$$

we know that $\mathbb{E}_{t-1}[X_t] = 0$, which implies $\{X_t\}_{t\leq T}$ is a martingale difference sequence. Also, by Assumption 4.2 and Lemma C.1.3, we can show that for all $2 \leq t \leq T$,

$$|X_t| \leq \alpha_t \sigma \left\| \gamma_{t-1} 1_{\tau\geq t} \right\| \leq 2\alpha_t \sigma^2.$$

Then by the Azuma-Hoeffding inequality (Lemma C.1.10), we have with probability at least $1-\delta$,

$$\left| \sum_{t=2}^{T} X_t \right| \leq 2\sigma^2 \sqrt{2\sum_{t=2}^{T} \alpha_t^2 \log(1/\delta)} \leq 5\sigma^2 \sqrt{(1+\beta^2 T)\log(1/\delta)},$$

where in the last inequality we use Lemma C.1.2. □

Then we are ready to prove Lemma C.1.6.

Proof of Lemma C.1.6. By Lemma C.1.8, we have for every $2 \le t \le \tau$,

$$\begin{aligned}\frac{\beta}{2}\left\|\epsilon_{t-1}\right\|^2 \le \frac{\alpha_t}{2}\left\|\epsilon_{t-1}\right\|^2 \le \left\|\epsilon_{t-1}\right\|^2 - \left\|\epsilon_t\right\|^2 + \frac{\lambda\beta}{16G}\left\|\nabla f(x_{t-1})\right\|^2 + \alpha_t^2\sigma^2 \\ + 2\alpha_t\left\langle \gamma_{t-1}, \nabla f(x_t,\xi_t) - \nabla f(x_t)\right\rangle.\end{aligned}$$

Taking a summation over t from 2 to τ, we have

$$\begin{aligned}\sum_{t=2}^{\tau}\frac{\beta}{2}\left\|\epsilon_{t-1}\right\|^2 - \frac{\lambda\beta}{16G}\left\|\nabla f(x_{t-1})\right\|^2 \le& \left\|\epsilon_1\right\|^2 - \left\|\epsilon_\tau\right\|^2 + \sigma^2\sum_{t=2}^{\tau}\alpha_t^2 + 10\sigma^2\sqrt{(1+\beta^2T)\log(1/\delta)} \\ \le& 4\sigma^2(1+\beta^2T) + 10\sigma^2\sqrt{(1+\beta^2T)\log(1/\delta)},\end{aligned}$$

where the first inequality uses Lemma C.1.9; and the second inequality uses Lemma C.1.2 and $\left\|\epsilon_1\right\|^2 = \left\|\nabla f(x_1,\xi_1) - \nabla f(x_1)\right\|^2 \le \sigma^2$. Then we complete the proof by multiplying both sides by $2/\beta$. □

C.1.4 Omitted proofs for Adam

In this section, we provide all the omitted proofs for Adam including those of Lemma 4.3.1 and all the lemmas in Appendix C.1.1.

Proof of Lemma 4.3.1. According to Lemma C.1.1, if $t < \tau$,

$$\left\|x_{t+1} - x_t\right\| \le \frac{\eta}{\lambda}\left\|\hat{m}_t\right\| \le \frac{\eta(G+\sigma)}{\lambda} \le \frac{2\eta G}{\lambda}.$$

□

Proof of Lemma C.1.1. By definition of τ, we have $\left\|\nabla f(x_t)\right\| \le G$ if $t < \tau$. Then Assumption 4.2 directly implies $\left\|\nabla f(x_t,\xi_t)\right\| \le G + \sigma$. $\left\|\hat{m}_t\right\|$ can be bounded by a standard induction argument as follows. First note that $\left\|\hat{m}_1\right\| = \left\|\nabla f(x_1,\xi_1)\right\| \le G + \sigma$. Supposing

$\|\hat{m}_{k-1}\| \leq G + \sigma$ for some $k < \tau$, then we have

$$\|\hat{m}_k\| \leq (1 - \alpha_k) \|\hat{m}_{k-1}\| + \alpha_k \|\nabla f(x_k, \xi_k)\| \leq G + \sigma.$$

Then we can show $\hat{v}_t \preceq (G+\sigma)^2$ in a similar way noting that $(\nabla f(x_t, \xi_t))^2 \preceq \|\nabla f(x_t, \xi_t)\|^2 \leq (G+\sigma)^2$. Given the bound on $\hat{v}_t$, it is straight forward to bound the stepsize h_t. □

Proof of Lemma C.1.2. First, when $t \geq 1/\beta$, we have $(1-\beta)^t \leq 1/e$. Therefore,

$$\sum_{1/\beta \leq t \leq T} (1 - (1-\beta)^t)^{-2} \leq (1 - 1/e)^{-2} T \leq 3T.$$

Next, note that when $t < 1/\beta$, we have $(1-\beta)^t \leq 1 - \frac{1}{2}\beta t$. Then we have

$$\sum_{2 \leq t < 1/\beta} (1 - (1-\beta)^t)^{-2} \leq \frac{4}{\beta^2} \sum_{t \geq 2} t^{-m} \leq \frac{3}{\beta^2}.$$

Therefore we have $\sum_{t=2}^{T} \alpha_t^2 \leq 3(1 + \beta^2 T)$. □

Proof of Lemma C.1.3. We prove $\|\epsilon_t\| \leq 2\sigma$ for all $t \leq \tau$ by induction. First, note that for $t = 1$, we have

$$\|\epsilon_1\| = \|\nabla f(x_1, \xi_1) - \nabla f(x_1)\| \leq \sigma \leq 2\sigma.$$

Now suppose $\|\epsilon_{t-1}\| \leq 2\sigma$ for some $2 \leq t \leq \tau$. According to the update rule (4.1), we have

$$\epsilon_t = (1 - \alpha_t)(\epsilon_{t-1} + \nabla f(x_{t-1}) - \nabla f(x_t)) + \alpha_t(\nabla f(x_t, \xi_t) - \nabla f(x_t)),$$

which implies

$$\|\epsilon_t\| \leq (2 - \alpha_t)\sigma + \|\nabla f(x_{t-1}) - \nabla f(x_t)\|.$$

Since we choose $\eta \leq \frac{r}{D}$, by Lemma 4.3.1, we have $\|x_t - x_{t-1}\| \leq \eta D \leq r$ if $t \leq \tau$. Therefore by Lemma 2.2.1, we have for any $2 \le t \le \tau$,

$$\|\nabla f(x_t) - \nabla f(x_{t-1})\| \le L \|x_t - x_{t-1}\| \le \eta DL \le \sigma\alpha_t,$$

where the last inequality uses the choice of η and $\beta \le \alpha_t$. Therefore we have $\|\epsilon_t\| \le 2\sigma$ which completes the induction. Then it is straight forward to show

$$\|\gamma_{t-1}\| \le (1-\alpha_t)(2\sigma + \alpha_t\sigma) \le 2\sigma.$$

□

Proof of Lemma C.1.4. We first list all the related parameter choices below for convenience.

$$G \ge \max\left\{2\lambda, 2\sigma, \sqrt{C_1\Delta_1 L_0}, (C_1\Delta_1 L_\rho)^{\frac{1}{2-\rho}}\right\}, \quad \beta \le \min\left\{1, \frac{c_1\lambda\epsilon^2}{\sigma^2 G\sqrt{\iota}}\right\},$$
$$\eta \le c_2 \min\left\{\frac{r\lambda}{G}, \frac{\sigma\lambda\beta}{LG\sqrt{\iota}}, \frac{\lambda^{3/2}\beta}{L\sqrt{G}}\right\}, \quad T = \max\left\{\frac{1}{\beta^2}, \frac{C_2\Delta_1 G}{\eta\epsilon^2}\right\}.$$

We will show $I_1/I_2 \le 1$ first. Note that if denoting $W = \frac{2L}{\lambda G^2}$, we have

$$I_1/I_2 = W\Delta_1\lambda + 8W\sigma^2\left(\frac{\eta}{\beta} + \eta\beta T\right) + 20W\sigma^2\sqrt{(\eta^2/\beta^2 + \eta^2 T)\iota},$$

Below are some facts that can be easily verified given the parameter choices.

(a) By the choice of G, we have $G^2 \ge 4\Delta_1(L_0 + 4L_\rho G^\rho) \ge 4\Delta_1 L$ for large enough C_1, which implies $W \le \frac{1}{2\Delta_1\lambda}$.

(b) By the choice of T, we have $\eta\beta T \le \frac{\eta}{\beta} + \frac{C_2\Delta_1 G\beta}{\epsilon^2}$.

(c) By the choice of T, we have $\eta^2 T = \max\left\{\left(\frac{\eta}{\beta}\right)^2, \frac{C_2\eta\Delta_1 G}{\epsilon^2}\right\} \le \left(\frac{\eta}{\beta}\right)^2 + \frac{C_2\Delta_1\sigma\beta}{\epsilon^2}\cdot\frac{\eta}{\beta} \le \frac{3}{2}\left(\frac{\eta}{\beta}\right)^2 + \frac{1}{2}\left(\frac{C_2\Delta_1\sigma\beta}{\epsilon^2}\right)^2$.

(d) By the choice of η, we have $\eta/\beta \le \frac{c_2\sigma\lambda}{LG\sqrt{\iota}}$, which implies $W\sigma^2\sqrt{\iota}\cdot\frac{\eta}{\beta} \le \frac{3c_2\sigma^3}{G^3} \le \frac{1}{200}$ for small enough c_2.

(e) By the choice of β and (a), we have $\frac{W\sigma^2\Delta_1 G\sqrt{\iota}\beta}{\epsilon^2} \leq \frac{\sigma^2 G\sqrt{\iota}\beta}{2\lambda\epsilon^2} \leq \frac{1}{100C_2}$ for small enough c_1.

Therefore,

$$\begin{aligned}
I_1/I_2 \leq & \frac{1}{2} + 8W\sigma^2\left(\frac{2\eta}{\beta} + \frac{C_2\Delta_1 G\beta}{\epsilon^2}\right) + 20W\sigma^2\sqrt{\iota}\left(\sqrt{\frac{5\eta^2}{2\beta^2} + \frac{1}{2}\left(\frac{C_2\Delta_1\sigma\beta}{\epsilon^2}\right)^2}\right) \\
\leq & \frac{1}{2} + 48W\sigma^2\sqrt{\iota}\cdot\frac{\eta}{\beta} + \frac{24C_2 W\sigma^2\Delta_1 G\sqrt{\iota}\beta}{\epsilon^2} \\
\leq & 1,
\end{aligned}$$

where the first inequality is due to Facts (a-c); the second inequality uses $\sigma \leq G$, $\iota \geq 1$, and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$; and the last inequality is due to Facts (d-e).

Next, we will show $I_1/T \leq \epsilon^2$. We have

$$\begin{aligned}
I_1/T = & \frac{8G\Delta_1}{\eta T} + \frac{64\sigma^2 G}{\lambda\beta T} + \frac{64\sigma^2 G\beta}{\lambda} + \frac{160\sigma^2 G\sqrt{\iota}}{\lambda}\sqrt{\frac{1}{\beta^2 T^2} + \frac{1}{T}} \\
\leq & \frac{8\epsilon^2}{C_2} + \frac{224\sigma^2 G\sqrt{\iota}}{\lambda\beta T} + \frac{64\sigma^2 G\beta}{\lambda} + \frac{160\sigma^2 G\sqrt{\iota}}{\lambda\sqrt{T}} \\
\leq & \frac{8\epsilon^2}{C_2} + \frac{450\sigma^2 G\sqrt{\iota}\beta}{\lambda} \\
= & \left(\frac{8}{C_2} + 450c_1\right)\epsilon^2 \\
\leq & \epsilon^2,
\end{aligned}$$

where in the first inequality we use $T \geq \frac{C_2\Delta_1 G}{\eta\epsilon^2}$ and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$; the second inequality uses $T \geq \frac{1}{\beta^2}$; the second equality uses the parameter choice of β; and in the last inequality we choose a large enough C_2 and small enough c_1. □

C.1.5 Proof of Theorem 4.2.2

Proof of Theorem 4.2.2. We define stopping time τ as follows

$$\begin{aligned}
\tau_1 &:= \min\{t \mid f(x_t) - f^* > F\} \wedge (T+1),\\
\tau_2 &:= \min\{t \mid \|\nabla f(x_t) - \nabla f(x_t, \xi_t)\| > \sigma\} \wedge (T+1),\\
\tau &:= \min\{\tau_1, \tau_2\}.
\end{aligned}$$

Then it is straightforward to verify that τ_1, τ_2, τ are all stopping times.

Since we want to show $\mathbb{P}(\tau \leq T)$ is small, noting that $\{\tau \leq T\} = \{\tau = \tau_1 \leq T\} \cup \{\tau = \tau_2 \leq T\}$, it suffices to bound both $\mathbb{P}(\tau = \tau_1 \leq T)$ and $\mathbb{P}(\tau = \tau_2 \leq T)$.

First, we know that

$$\begin{aligned}
\mathbb{P}(\tau = \tau_2 \leq T) \leq& \mathbb{P}(\tau_2 \leq T)\\
=& \mathbb{P}\left(\bigcup_{1\leq t\leq T} \|\nabla f(x_t) - \nabla f(x_t, \xi_t)\| > \sigma\right)\\
\leq& \sum_{1\leq t\leq T} \mathbb{P}\left(\|\nabla f(x_t) - \nabla f(x_t, \xi_t)\| > \sigma\right)\\
\leq& \sum_{1\leq t\leq T} \mathbb{E}\left[\mathbb{P}_{t-1}\left(\|\nabla f(x_t) - \nabla f(x_t, \xi_t)\| > \sigma\right)\right]\\
\leq& \sum_{1\leq t\leq T} \mathbb{E}\left[2e^{-\frac{\sigma^2}{2R^2}}\right]\\
=& 2Te^{-\frac{\sigma^2}{2R^2}}\\
\leq& \delta/2,
\end{aligned}$$

where the fourth inequality uses Assumption 4.3; and the last inequality uses $\sigma = R\sqrt{2\log(4T/\delta)}$.

Next, if $\tau = \tau_1 \leq T$, by definition, we have $f(x_\tau) - f^* > F$, or equivalently,

$$\frac{8G}{\eta}(f(x_\tau) - f^*) > \frac{8GF}{\eta} = \frac{4G^3}{\eta L} =: I_2.$$

On the other hand, since for any $t < \tau$, under the new definition of τ, we still have

$$f(x_t) - f^* \leq F, \quad \|f(x_t)\| \leq G, \quad \|\nabla f(x_t) - \nabla f(x_t, \xi_t)\| \leq \sigma.$$

Then we know that Lemma C.1.7 still holds because all of its requirements are still satisfied, i.e., there exists some event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}) \leq \delta/2$, such that under its complement $\mathcal{E}^c$,

$$\sum_{t=1}^{\tau-1} \|\nabla f(x_t)\|^2 + \frac{8G}{\eta}(f(x_\tau) - f^*) \leq \frac{8G}{\eta\lambda}\left(\Delta_1\lambda + 8\sigma^2\left(\frac{\eta}{\beta} + \eta\beta T\right) + 20\sigma^2\eta\sqrt{(1/\beta^2 + T)\iota}\right)$$
$$=: I_1.$$

By Lemma C.1.4, we know $I_1 \leq I_2$, which suggests that $\mathcal{E}^c \cap \{\tau = \tau_1 \leq T\} = \emptyset$, i.e., $\{\tau = \tau_1 \leq T\} \subset \mathcal{E}$. Then we can show

$$\mathbb{P}(\mathcal{E} \cup \{\tau \leq T\}) \leq \mathbb{P}(\mathcal{E}) + \mathbb{P}(\tau = \tau_2 \leq T) \leq \delta.$$

Therefore,

$$\mathbb{P}(\mathcal{E}^c \cap \{\tau = T + 1\}) \geq 1 - \mathbb{P}(\mathcal{E} \cup \{\tau \leq T\}) \geq 1 - \delta,$$

and under the event $\mathcal{E}^c \cap \{\tau = T + 1\}$, we have $\tau = T + 1$ and

$$\frac{1}{T}\sum_{t=1}^{t} \|\nabla f(x_t)\|^2 \leq I_1/T \leq \epsilon^2,$$

where the last inequality is due to Lemma C.1.4. □

C.2 Convergence Anlaysis of VRAdam

In this section, we provide detailed convergence analysis of VRAdam and prove Theorem 4.4.2. To do that, we first provide some technical definitions[2]. Denote

$$\epsilon_t := m_t - \nabla f(x_t)$$

as the deviation of the momentum from the actual gradient. From the update rule in Algorithm 3, we can write

$$\epsilon_t = (1-\beta)\epsilon_{t-1} + W_t, \tag{C.8}$$

where we define

$$W_t := \nabla f(x_t, \xi_t) - \nabla f(x_t) - (1-\beta)\left(\nabla f(x_{t-1}, \xi_t) - \nabla f(x_{t-1})\right).$$

Let G be the constant defined in Theorem 4.4.2 and denote $F := \frac{G^2}{2(L_0 + L_\rho (2G)^\rho)}$. We define the following stopping times as discussed in Section 4.4.1.

$$\begin{aligned}
\tau_1 &:= \min\{t \mid f(x_t) - f^* > F\} \wedge (T+1), \\
\tau_2 &:= \min\{t \mid \|\epsilon_t\| > G\} \wedge (T+1), \\
\tau &:= \min\{\tau_1, \tau_2\}.
\end{aligned} \tag{C.9}$$

It is straight forward to verify that τ_1, τ_2, τ are all stopping times. Then if $t < \tau$, we have

$$f(x_t) - f^* \leq F, \quad \|\nabla f(x_t)\| \leq G, \quad \|\epsilon_t\| \leq G.$$

[2] Note that the same symbol for Adam and VRAdam may have different meanings.

Then we can also bound the update $\|x_{t+1} - x_t\| \leq \eta D$ where $D = 2G/\lambda$ if $t < \tau$ (see Lemma C.2.3 for the details). Finally, we consider the same definition of r and L as those for Adam. Specifically,

$$L := L_0 + L_\rho(2G)^\rho, \quad r := r(G) = G/L. \tag{C.10}$$

C.2.1 Useful lemmas

We first list several useful lemmas in this section without proofs. Their proofs are deferred later in Appendix C.2.3.

To start with, we provide a lemma on the local smoothness of each component function $f(\cdot, \xi)$ when the gradient of the objective function f is bounded.

Lemma C.2.1. *For any constant $G \geq \sigma$ and two points $x \in \mathcal{X}, y \in \mathbb{R}^d$ such that $\|\nabla f(x)\| \leq G$ and $\|y - x\| \leq r/2$, we have $y \in \mathcal{X}$ and*

$$\begin{gathered}
\|\nabla f(y) - \nabla f(x)\| \leq L \|y - x\|, \\
\|\nabla f(y, \xi) - \nabla f(x, \xi)\| \leq 4L \|y - x\|, \quad \forall \xi, \\
f(y) \leq f(x) + \left\langle \nabla f(x), y - x \right\rangle + \frac{1}{2} L \|y - x\|^2,
\end{gathered}$$

where r and L are defined in (C.10).

With the new definition of stopping time τ in (C.9), all the quantities in Algorithm 3 are well bounded before τ. In particular, the following lemma holds.

Lemma C.2.2. *If $t < \tau$, we have*

$$\begin{gathered}
\|\nabla f(x_t)\| \leq G, \quad \|\nabla f(x_t, \xi_t)\| \leq G + \sigma, \quad \|m_t\| \leq 2G, \\
\hat{v}_t \preceq (G + \sigma)^2, \quad \frac{\eta}{G + \sigma + \lambda} \preceq h_t \preceq \frac{\eta}{\lambda}.
\end{gathered}$$

Next, we provide the following lemma which bounds the update at each step before time τ.

Lemma C.2.3. *if $t < \tau$, $\|x_{t+1} - x_t\| \leq \eta D$ where $D = 2G/\lambda$.*

The following lemma bounds $\|W_t\|$ when $t \leq \tau$.

Lemma C.2.4. *If $t \leq \tau$, $G \geq 2\sigma$, and $\eta \leq \frac{r}{2D}$,*

$$\|W_t\| \leq \beta\sigma + \frac{5\eta L}{\lambda} \left(\|\nabla f(x_{t-1})\| + \|\epsilon_{t-1}\|\right).$$

Finally, we present some inequalities regarding the parameter choices, which will simplify the calculations later.

Lemma C.2.5. *Under the parameter choices in Theorem 4.4.2, we have*

$$\frac{2\Delta_1}{F} \leq \frac{\delta}{4}, \quad \frac{\lambda\Delta_1\beta}{\eta G^2} \leq \frac{\delta}{4}, \quad \eta\beta T \leq \frac{\lambda\Delta_1}{8\sigma^2}, \quad \eta \leq \frac{\lambda^{3/2}}{40L}\sqrt{\frac{\beta}{G}}.$$

C.2.2 Proof of Theorem 4.4.2

Before proving the theorem, we will need to present several important lemmas. First, note that the descent lemma still holds for VRAdam.

Lemma C.2.6. *If $t < \tau$, choosing $G \geq \sigma + \lambda$ and $\eta \leq \min\left\{\frac{r}{2D}, \frac{\lambda}{6L}\right\}$, we have*

$$f(x_{t+1}) - f(x_t) \leq -\frac{\eta}{4G}\|\nabla f(x_t)\|^2 + \frac{\eta}{\lambda}\|\epsilon_t\|^2.$$

Proof of Lemma C.2.6. The proof is essentially the same as that of Lemma C.1.5. □

Lemma C.2.7. *Choose $G \geq \max\{2\sigma, 2\lambda\}$, $S_1 \geq \frac{1}{2\beta^2 T}$, and $\eta \leq \min\left\{\frac{r}{2D}, \frac{\lambda^{3/2}}{40L}\sqrt{\frac{\beta}{G}}\right\}$. We have*

$$\mathbb{E}\left[\sum_{t=1}^{\tau-1} \frac{\beta}{2}\|\epsilon_t\|^2 - \frac{\lambda\beta}{16G}\|\nabla f(x_t)\|^2\right] \leq 4\sigma^2\beta^2 T - \mathbb{E}[\|\epsilon_\tau\|^2].$$

Proof of Lemma C.2.7. By Lemma C.2.4, we have

$$\begin{aligned}\|W_t\|^2 \leq& 2\sigma^2\beta^2 + \frac{100\eta^2 L^2}{\lambda^2}\left(\|\nabla f(x_{t-1})\|^2 + \|\epsilon_{t-1}\|^2\right)\\ \leq& 2\sigma^2\beta^2 + \frac{\lambda\beta}{16G}\left(\|\nabla f(x_{t-1})\|^2 + \|\epsilon_{t-1}\|^2\right),\end{aligned}$$

where in the second inequality we choose $\eta \leq \frac{\lambda^{3/2}}{40L}\sqrt{\frac{\beta}{G}}$. Therefore, noting that $\frac{\lambda\beta}{16G} \leq \beta/2$, by (C.8), we have

$$\begin{aligned}\|\epsilon_t\|^2 =& (1-\beta)^2\|\epsilon_{t-1}\|^2 + \|W_t\|^2 + (1-\beta)\Big\langle \epsilon_{t-1}, W_t\Big\rangle\\ \leq& (1-\beta/2)\|\epsilon_{t-1}\|^2 + \frac{\lambda\beta}{16G}\|\nabla f(x_{t-1})\|^2 + 2\sigma^2\beta^2 + (1-\beta)\Big\langle \epsilon_{t-1}, W_t\Big\rangle.\end{aligned}$$

Taking a summation over $2 \leq t \leq \tau$ and re-arranging the terms, we get

$$\sum_{t=1}^{\tau-1}\frac{\beta}{2}\|\epsilon_t\|^2 - \frac{\lambda\beta}{16G}\|\nabla f(x_t)\|^2 \leq \|\epsilon_1\|^2 - \|\epsilon_\tau\|^2 + 2\sigma^2\beta^2(\tau-1) + (1-\beta)\sum_{t=2}^{\tau}\Big\langle \epsilon_{t-1}, W_t\Big\rangle.$$

Taking expectations on both sides, noting that

$$\mathbb{E}\left[\sum_{t=2}^{\tau}\Big\langle \epsilon_{t-1}, W_t\Big\rangle\right] = 0$$

by the Optional Stopping Theorem (Lemma B.5.1), we have

$$\mathbb{E}\left[\sum_{t=1}^{\tau-1}\frac{\beta}{2}\|\epsilon_t\|^2 - \frac{\lambda\beta}{16G}\|\nabla f(x_t)\|^2\right] \leq 2\sigma^2\beta^2 T + \mathbb{E}[\|\epsilon_1\|^2] - \mathbb{E}[\|\epsilon_\tau\|^2] \leq 4\sigma^2\beta^2 T - \mathbb{E}[\|\epsilon_\tau\|^2],$$

where in the second inequality we choose $S_1 \geq \frac{1}{2\beta^2 T}$ which implies $\mathbb{E}[\|\epsilon_1\|^2] \leq \sigma^2/S_1 \leq 2\sigma^2\beta^2 T$. □

Lemma C.2.8. *Under the parameter choices in Theorem 4.4.2, we have*

$$\mathbb{E}\left[\sum_{t=1}^{\tau-1}\|\nabla f(x_t)\|^2\right] \leq \frac{16G\Delta_1}{\eta}, \quad \mathbb{E}[f(x_\tau) - f^*] \leq 2\Delta_1, \quad \mathbb{E}[\|\epsilon_\tau\|^2] \leq \frac{\lambda\Delta_1\beta}{\eta}.$$

Proof of Lemma C.2.8. First note that according to Lemma C.2.5, it is straight forward to verify that the parameter choices in Theorem 4.4.2 satisfy the requirements in Lemma C.2.6 and Lemma C.2.7. Then by Lemma C.2.6, if $t < \tau$,

$$f(x_{t+1}) - f(x_t) \leq -\frac{\eta}{4G} \left\|\nabla f(x_t)\right\|^2 + \frac{\eta}{\lambda} \left\|\epsilon_t\right\|^2 .$$

Taking a summation over $1 \leq t < \tau$, re-arranging terms, multiplying both sides by $\frac{8G}{\eta}$, and taking an expection, we get

$$\mathbb{E}\left[\sum_{t=1}^{\tau-1} 2\left\|\nabla f(x_t)\right\|^2 - \frac{8G}{\lambda}\left\|\epsilon_t\right\|^2\right] \leq \frac{8G}{\eta}\mathbb{E}[f(x_1) - f(x_\tau)] \leq \frac{8G}{\eta}\left(\Delta_1 - \mathbb{E}[f(x_\tau) - f^*]\right). \tag{C.11}$$

By Lemma C.2.7, we have

$$\mathbb{E}\left[\sum_{t=1}^{\tau-1} \frac{8G}{\lambda}\left\|\epsilon_t\right\|^2 - \left\|\nabla f(x_t)\right\|^2\right] \leq \frac{64G\sigma^2\beta T}{\lambda} - \frac{16G}{\lambda\beta}\mathbb{E}[\left\|\epsilon_\tau\right\|^2] \leq \frac{8G\Delta_1}{\eta} - \frac{16G}{\lambda\beta}\mathbb{E}[\left\|\epsilon_\tau\right\|^2], \tag{C.12}$$

where the last inequality is due to Lemma C.2.5. Then (C.11) + (C.12) gives

$$\mathbb{E}\left[\sum_{t=1}^{\tau-1}\left\|\nabla f(x_t)\right\|^2\right] + \frac{8G}{\eta}\mathbb{E}[f(x_\tau) - f^*] + \frac{16G}{\lambda\beta}\mathbb{E}[\left\|\epsilon_\tau\right\|^2] \leq \frac{16G\Delta_1}{\eta},$$

which completes the proof. □

With all the above lemmas, we are ready to prove the theorem.

Proof of Theorem 4.4.2. First note that according to Lemma C.2.5, it is straight forward to verify that the parameter choices in Theorem 4.4.2 satisfy the requirements in all the lemmas for VRAdam.

Then, first note that if $\tau = \tau_1 \leq T$, we know $f(x_\tau) - f^* > F$ by the definition of τ.

Therefore,

$$\mathbb{P}(\tau = \tau_1 \leq T) \leq \mathbb{P}(f(x_\tau) - f^* > F) \leq \frac{\mathbb{E}[f(x_\tau) - f^*]}{F} \leq \frac{2\Delta_1}{F} \leq \frac{\delta}{4},$$

where the second inequality uses Markov's inequality; the third inequality is by Lemma C.2.8; and the last inequality is due to Lemma C.2.5.

Similarly, if $\tau_2 = \tau \leq T$, we know $\|\epsilon_\tau\| > G$. We have

$$\mathbb{P}(\tau_2 = \tau \leq T) \leq \mathbb{P}(\|\epsilon_\tau\| > G) = \mathbb{P}(\|\epsilon_\tau\|^2 > G^2) \leq \frac{\mathbb{E}[\|\epsilon_\tau\|^2]}{G^2} \leq \frac{\lambda\Delta_1\beta}{\eta G^2} \leq \frac{\delta}{4},$$

where the second inequality uses Markov's inequliaty; the third inequality is by Lemma C.2.8; and the last inequality is due to Lemma C.2.5. where the last inequality is due to Lemma C.2.5. Therefore,

$$\mathbb{P}(\tau \leq T) \leq \mathbb{P}(\tau_1 = \tau \leq T) + \mathbb{P}(\tau_2 = \tau \leq T) \leq \frac{\delta}{2}.$$

Also, note that by Lemma C.2.8

$$\begin{aligned}\frac{16G\Delta_1}{\eta} \geq & \mathbb{E}\left[\sum_{t=1}^{\tau-1} \|\nabla f(x_t)\|^2\right] \\ \geq & \mathbb{P}(\tau = T+1)\mathbb{E}\left[\sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \middle| \tau = T+1\right] \\ \geq & \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \middle| \tau = T+1\right],\end{aligned}$$

where the last inequality is due to $\mathbb{P}(\tau = T+1) = 1 - \mathbb{P}(\tau \leq T) \geq 1 - \delta/2 \geq 1/2$. Then we can get

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \middle| \tau = T+1\right] \leq \frac{32G\Delta_1}{\eta T} \leq \frac{\delta\epsilon^2}{2}.$$

Let $\mathcal{F} := \left\{\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\|^2 > \epsilon^2\right\}$ be the event of not converging to stationary points. By

Markov's inequality, we have

$$\mathbb{P}(\mathcal{F}|\tau = T+1) \leq \frac{\delta}{2}.$$

Therefore,

$$\mathbb{P}(\mathcal{F} \cup \{\tau \leq T\}) \leq \mathbb{P}(\tau \leq T) + \mathbb{P}(\mathcal{F}|\tau = T+1) \leq \delta,$$

i.e., with probability at least $1-\delta$, we have both $\tau = T+1$ and $\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \leq \epsilon^2$. □

C.2.3 Proofs of lemmas in Appendix C.2.1

Proof of Lemma C.2.1. This lemma is a direct corollary of Lemma 2.2.1. Note that by Assumption 4.5, we have $\|\nabla f(x,\xi)\| \leq G + \sigma \leq 2G$. Hence, when computing the locality size and smoothness constant for the component function $f(\cdot,\xi)$, we need to replace the constant G in Lemma 2.2.1 with $2G$, that is why we get a smaller locality size of $r/2$ and a larger smoothness constant of $4L$. □

Proof of Lemma C.2.2. The bound on $\|m_t\|$ is by the definition of τ in (C.9). All other quantities for VRAdam are defined in the same way as those in Adam (Algorithm 2), so they have the same upper bounds as in Lemma C.1.1. □

Proof of Lemma C.2.3.

$$\|x_{t+1} - x_t\| \leq \eta \|m_t\| / \lambda \leq 2\eta G/\lambda = \eta D, \tag{C.13}$$

where the first inequality uses the update rule in Algorithm 3 and $h_t \preceq \eta/\lambda$ by Lemma C.2.2; the second inequality is again due to Lemma C.2.2. □

Proof of Lemma C.2.4. By the definition of W_t, it is easy to verify that

$$W_t = \beta(\nabla f(x_t, \xi_t) - \nabla f(x_t)) + (1-\beta)\delta_t,$$

where

$$\delta_t = \nabla f(x_t, \xi_t) - \nabla f(x_{t-1}, \xi_t) - \nabla f(x_t) + \nabla f(x_{t-1}).$$

Then we can bound

$$\begin{aligned}\|\delta_t\| \leq & \|\nabla f(x_t, \xi_t) - \nabla f(x_{t-1}, \xi_t)\| + \|\nabla f(x_t) - \nabla f(x_{t-1})\| \\ \leq & 5L\left\|x_t - x_{t-1}\right\| \\ \leq & \frac{5\eta L}{\lambda}\left(\|\nabla f(x_{t-1})\| + \|\epsilon_{t-1}\|\right),\end{aligned}$$

where the second inequality uses Lemma C.2.1; and the last inequality is due to $\|x_t - x_{t-1}\| \leq \eta\left\|m_{t-1}\right\|/\lambda \leq \eta\left(\|\nabla f(x_{t-1})\| + \|\epsilon_{t-1}\|\right)/\lambda$. Then, we have

$$\|W_t\| \leq \beta\sigma + \frac{5\eta L}{\lambda}\left(\|\nabla f(x_{t-1})\| + \|\epsilon_{t-1}\|\right).$$

□

Proof of Lemma C.2.5. These inequalities can be obtained by direct calculations. □

Appendix D

Proofs for Chapter 5

D.1 Useful lemmas on directional smoothness

In this section, we present two useful lemmas related to directional smoothness along with their proofs. First, the following lemma essentially shows Taylor's expansion in terms of directional smoothness, which is very useful for obtaining descent lemmas in our convergence analyses.

Lemma D.1.1. *Under Assumptions 5.1 and 5.3, the following inequalities hold for all $x, y \in \mathbb{R}^d$.*

$$\left\langle \nabla f(y) - \nabla f(x), y - x \right\rangle \leq \ell_x(y-x) \left\| y - x \right\|^2 + \frac{M}{2} \left\| y - x \right\|^3,$$
$$f(y) - f(x) \leq \left\langle \nabla f(x), y - x \right\rangle + \frac{\ell_x(y-x)}{2} \left\| y - x \right\|^2 + \frac{M}{6} \left\| y - x \right\|^3.$$

Proof of Lemma D.1.1. Denote $z_\alpha := (1-\alpha)x + \alpha y$ for $0 \leq \alpha \leq 1$. Then we have

$$\begin{aligned}
&\Big\langle \nabla f(y) - \nabla f(x), y - x \Big\rangle \\
=&\int_0^1 (y-x)^\top \nabla^2 f(x_\alpha)(y-x)\, d\alpha \\
=&(y-x)^\top \nabla^2 f(x)(y-x) + \int_0^1 (y-x)^\top (\nabla^2 f(x_\alpha) - \nabla^2 f(x))(y-x)\, d\alpha \qquad \text{(D.1)} \\
\leq&\ell_x(y{-}x) \left\| y - x \right\|^2 + \int_0^1 \alpha M \left\| y - x \right\|^3\, d\alpha \\
\leq&\ell_x(y{-}x) \left\| y - x \right\|^2 + \frac{M}{2} \left\| y - x \right\|^3,
\end{aligned}$$

where the first inequality uses the definition of directional smoothness in Definition 5 and Assumption 5.3. Therefore we can also obtain

$$\begin{aligned}
f(y) - f(x) = &\int_0^1 \Big\langle \nabla f(x_\alpha), y - x \Big\rangle\, d\alpha \\
=&\Big\langle \nabla f(x), y - x \Big\rangle + \int_0^1 \frac{1}{\alpha} \Big\langle \nabla f(x_\alpha) - f(x), x_\alpha - x \Big\rangle\, d\alpha \\
\leq&\Big\langle \nabla f(x), y - x \Big\rangle + \int_0^1 \alpha \ell_x(y{-}x) \left\| y - x \right\|^2 + \frac{\alpha^2 M}{2} \left\| y - x \right\|^3\, d\alpha \\
=&\Big\langle \nabla f(x), y - x \Big\rangle + \frac{\ell_x(y{-}x)}{2} \left\| y - x \right\|^2 + \frac{M}{6} \left\| y - x \right\|^3,
\end{aligned}$$

where the inequality uses (D.1). □

The next lemma can be viewed as a generalized reversed Polyak-Lojasiewicz (PL) inequality for functions satisfying our directional smoothness condition.

Lemma D.1.2. *Suppose Assumptions 5.1, 5.2, and 5.3 hold. For any $x \in \mathbb{R}^d$, denoting $H_\lambda(x) := \operatorname{diag}\left(\frac{1}{|\nabla f(x)|+\lambda}\right)$, we have*

$$\left\| \nabla f(x) \right\|_{H_\lambda(x)} \leq \frac{3}{\sqrt{\lambda}} \cdot \max\left\{ \sqrt{L_\lambda (f(x) - f^*)}, M^{1/3} (f(x) - f^*)^{2/3} \right\}.$$

Proof of Lemma D.1.2. Note that by Lemma D.1.1, for any $x \in \mathbb{R}^d$ and $\alpha =$

$\min\left\{\frac{\lambda}{L_\lambda}, \sqrt{\frac{\lambda^{3/2}}{M\|\nabla f(x)\|_{H_\lambda(x)}}}\right\}$, we have

$$\begin{aligned}
f(x-\alpha u_\lambda(x)) - f(x) \leq & -\alpha\Big\langle \nabla f(x), u_\lambda(x)\Big\rangle + \frac{\alpha^2 L_\lambda}{2}\left\|u_\lambda(x)\right\|^2 + \frac{\alpha^3 M}{6}\left\|u_\lambda(x)\right\|^3 \\
\leq & -\alpha\left\|\nabla f(x)\right\|_{H_\lambda(x)}^2 + \frac{\alpha^2 L_\lambda}{2\lambda}\left\|\nabla f(x)\right\|_{H_\lambda(x)}^2 + \frac{\alpha^3 M}{6\lambda^{3/2}}\left\|\nabla f(x)\right\|_{H_\lambda(x)}^3 \\
\leq & -\frac{\alpha}{3}\left\|\nabla f(x)\right\|_{H_\lambda(x)}^2,
\end{aligned}$$

where in the second inequality we use $\Big\langle \nabla f(x), u_\lambda(x)\Big\rangle = \left\|\nabla f(x)\right\|_{H_\lambda(x)}^2$ and $\left\|u_\lambda(x)\right\| \leq \frac{1}{\sqrt{\lambda}}\left\|\nabla f(x)\right\|_{H_\lambda(x)}$ by the definition of $u_\lambda(x)$ and $H_\lambda(x)$; and the last inequality is due to the choice of α. Then noting that $f(x) - f(x-\alpha u_\lambda(x)) \leq f(x) - f^*$, we can show that

$$\left\|\nabla f(x)\right\|_{H_\lambda(x)}^2 \leq \frac{3(f(x)-f^*)}{\alpha} = 3(f(x)-f^*)\max\left\{\frac{L_\lambda}{\lambda}, \sqrt{\frac{M\left\|\nabla f(x)\right\|_{H_\lambda(x)}}{\lambda^{3/2}}}\right\},$$

which implies

$$\left\|\nabla f(x)\right\|_{H_\lambda(x)} \leq \frac{3}{\sqrt{\lambda}} \cdot \max\left\{\sqrt{L_\lambda(f(x)-f^*)}, M^{1/3}(f(x)-f^*)^{2/3}\right\}.$$

□

D.2 Convergence of memoryless Adam

In this section, we provide the proof of Theorem 5.2.1 on the convergence of memoryless Adam under directional smoothness assumptions. To simplify the notation, we denote

$$g_t := \nabla f(x_t), \quad u_t := u_\lambda(x_t) = \frac{g_t}{|g_t|+\lambda}, \quad H_t := H_\lambda(x_t) = \operatorname{diag}\left(\frac{1}{|g_t|+\lambda}\right), \tag{D.2}$$

where H_λ is defined in Lemma D.1.2. In what follows, we first present a descent lemma for memoryless Adam.

Lemma D.2.1. *Suppose Assumptions 5.1, 5.2, and 5.3 hold. Denote* $\Delta_1 := f(x_1) - f^*$*. For* *any $\epsilon > 0$, choose either $\eta_t \equiv \eta = \frac{\lambda}{2\max\left\{L_\lambda, M^{2/3}\Delta_1^{1/3}\right\}}$ or $\eta_t = \min\left\{\frac{\lambda}{2L_\lambda}, \frac{\sqrt{\lambda}L_\lambda}{M\|g_t\|_{H_t}}\right\}$. Then the iterates generated by (5.5) satisfy*

$$f(x_{t+1}) - f(x_t) \le -\frac{\eta_t}{2}\left\|g_t\right\|_{H_t}^2, \quad \forall t \ge 1.$$

Proof of Lemma D.2.1. By Lemma D.1.1, for all $t \ge 1$, we have

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\le -\eta_t\Big\langle g_t, u_t\Big\rangle + \frac{\eta_t^2 L_\lambda}{2}\left\|u_t\right\|^2 + \frac{\eta_t^3 M}{6}\left\|u_t\right\|^3 \\ &\le -\eta_t\left\|g_t\right\|_{H_t}^2 + \frac{\eta_t^2 L_\lambda}{2\lambda}\left\|g_t\right\|_{H_t}^2 + \frac{\eta_t^3 M}{6\lambda^{3/2}}\left\|g_t\right\|_{H_t}^3 \\ &\le -\eta_t\left\|g_t\right\|_{H_t}^2\left(\frac{3}{4} - \frac{\eta_t^2 M}{6\lambda^{3/2}}\left\|g_t\right\|_{H_t}\right), \end{aligned} \tag{D.3}$$

where the first inequality uses Lemma D.1.1; in the second inequality we use $\Big\langle g_t, u_t\Big\rangle = \left\|g_t\right\|_{H_t}^2$ and $\|u_t\| \le \frac{1}{\sqrt{\lambda}}\left\|g_t\right\|_{H_t}$ by the definition of u_t and H_t in (D.2); the third inequality is due to $\eta_t \le \frac{\lambda}{2L_\lambda}$ based on either of our two choices of η_t. Next, we will bound $\frac{\eta_t^2 M}{6\lambda^{3/2}}\left\|g_t\right\|_{H_t}$ for each choice of η_t below.

- If $\eta_t = \min\left\{\frac{\lambda}{2L_\lambda}, \frac{\sqrt{\lambda}L_\lambda}{M\|g_t\|_{H_t}}\right\}$, then we know

  $$\frac{\eta_t^2 M}{6\lambda^{3/2}}\left\|g_t\right\|_{H_t} \le \eta_t \cdot \frac{L_\lambda}{6\lambda} \le \frac{1}{12} \le \frac{1}{4},$$

  which implies

  $$f(x_{t+1}) - f(x_t) \le -\frac{\eta_t}{2}\left\|g_t\right\|_{H_t}^2, \quad \forall t \ge 1.$$

- If $\eta_t = \eta \equiv \frac{\lambda}{2\max\left\{L_\lambda, M^{2/3}\Delta_1^{1/3}\right\}}$, we have

$$\begin{aligned}\frac{\eta_t^2 M}{6\lambda^{3/2}}\left\|g_t\right\|_{H_t} \leq & \frac{M\sqrt{\lambda}\left\|g_t\right\|_{H_t}}{24\max\left\{L_\lambda^2, M^{4/3}\Delta_1^{2/3}\right\}} \\ \leq & \frac{\max\left\{M\sqrt{L_\lambda(f(x_t)-f^*)}, M^{4/3}(f(x_t)-f^*)^{2/3}\right\}}{8\max\left\{L_\lambda^2, M^{4/3}\Delta_1^{2/3}\right\}},\end{aligned} \tag{D.4}$$

where the last inequality uses Lemma D.1.2. Then we will show $f(x_t) \leq f(x_1)$ for all $t \geq 1$ by induction. First note that $f(x_1) \leq f(x_1)$. Suppose $f(x_k) \leq f(x_1)$ for some $k \geq 1$. Also note that

$$M\sqrt{L_\lambda \Delta_1} \leq \begin{cases} L_\lambda^2 & \text{if } L_\lambda \geq M^{2/3}\Delta_1^{1/3}, \\ M^{4/3}\Delta_1^{2/3} & \text{if } L_\lambda \leq M^{2/3}\Delta_1^{1/3}. \end{cases}$$

In other words, $M\sqrt{L_\lambda\Delta_1} \leq \max\left\{L_\lambda^2, M^{4/3}\Delta_1^{2/3}\right\}$, and thus we have

$$\begin{aligned} & \max\left\{M\sqrt{L_\lambda(f(x_k)-f^*)}, M^{4/3}(f(x_k)-f^*)^{2/3}\right\} \\ \leq & \max\left\{M\sqrt{L_\lambda\Delta_1}, M^{4/3}\Delta_1^{2/3}\right\} \\ \leq & \max\{L_\lambda^2, M^{4/3}\Delta_1^{2/3}\}. \end{aligned}$$

Therefore, we have

$$\frac{\eta_t^2 M}{6\lambda^{3/2}}\left\|g_k\right\|_{H_k} \leq \frac{1}{8} \leq \frac{1}{4},$$

which implies

$$f(x_{k+1}) - f(x_k) \leq -\frac{\eta_k}{2}\left\|g_k\right\|_{H_k}^2 \leq 0.$$

Then we know $f(x_{k+1}) \leq f(x_1)$ as well. By induction, the above inequality holds for all $k \geq 1$, which completes the proof.

□

Proof of Theorem 5.2.1. By Lemma D.2.1, we have for all $t \geq 1$,

$$f(x_{t+1}) - f(x_t) \leq -\frac{\eta}{2} \|g_t\|_{H_t}^2 \leq -\frac{\eta}{2} \frac{\|g_t\|^2}{\|g_t\| + \lambda}.$$

Define function $\alpha(z) := \frac{z^2}{z+\lambda}$ for $z \geq 0$. By telescoping, we obtain

$$\frac{1}{T} \sum_{t=1}^{T} \alpha(\|g_t\|) \leq \frac{2(f(x_1) - f(x_{T+1}))}{\eta T} \leq \frac{2\Delta_1}{\eta T} \leq \alpha(\epsilon),$$

where the last inequality uses the choice of T. By standard calculations, we have

$$\alpha'(z) = \frac{z^2 + 2\lambda z}{(z+\lambda)^2} \geq 0, \quad \alpha''(z) = -\frac{2\lambda^2}{(z+\lambda)^3} \geq 0.$$

In other words, α is non-decreasing and convex. Since α is convex, by Jensen's inequality, we have

$$\alpha\left(\frac{1}{T} \sum_{t=1}^{T} \|g_t\|\right) \leq \frac{1}{T} \sum_{t=1}^{T} \alpha(\|g_t\|) \leq \alpha(\epsilon).$$

Since α is increasing, the above inequality implies

$$\frac{1}{T} \sum_{t=1}^{T} \|g_t\| = \frac{1}{T} \sum_{t=1}^{T} \|\nabla f(x_t)\| \leq \epsilon.$$

□

Proof of Theorem 5.2.2. By Lemma D.2.1, we have for all $t \geq 1$,

$$f(x_{t+1}) - f(x_t) \leq -\frac{\eta_t}{2} \|g_t\|_{H_t}^2 .$$

Note that $\eta_t = \min\left\{\frac{\lambda}{2L_\lambda}, \frac{\sqrt{\lambda}L_\lambda}{M\|g_t\|_{H_t}}\right\}$.

- If $\|g_t\|_{H_t} \geq \frac{L_\lambda^2}{M\sqrt{\lambda}}$, then $\eta_t = \frac{\sqrt{\lambda}L_\lambda}{M\|g_t\|_{H_t}}$. Therefore,

$$f(x_{t+1}) - f(x_t) \leq -\frac{\sqrt{\lambda}L_\lambda}{2M}\|g_t\|_{H_t} \leq -\frac{L_\lambda^3}{M^2}.$$

- Otherwise if $\|g_t\|_{H_t} \leq \frac{L_\lambda^2}{M\sqrt{\lambda}}$, then $\eta = \frac{\lambda}{2L_\lambda}$. Then we know that

$$f(x_{t+1}) - f(x_t) \leq -\frac{\lambda}{4L_\lambda}\|g_t\|_{H_t}^2 \leq -\frac{\lambda}{4L_\lambda}\frac{\|g_t\|^2}{\|g_t\| + \lambda} =: -\frac{\lambda}{4L_\lambda}\alpha(\|g_t\|),$$

  where we denote $\alpha(z) := \frac{z^2}{z+\lambda}$ for $z \geq 0$ as in the proof of Theorem 5.2.1.

Therefore we can show that for $1 \leq t \leq T$

$$f(x_{t+1}) - f(x_t) \leq -\min\left\{\frac{L_\lambda^3}{M^2}, \frac{\lambda}{4L_\lambda}\alpha(\|g_t\|)\right\} \leq -\min\left\{\frac{L_\lambda^3}{M^2}, \frac{\lambda}{4L_\lambda}\alpha\left(\min_{t\leq T}\|g_t\|\right)\right\},$$

where the last inequality uses the fact that the function α is increasing. By telescoping, noting that $f(x_1) - f(x_T) \leq \Delta_1$, we have

$$\Delta_1 \geq T\min\left\{\frac{L_\lambda^3}{M^2}, \frac{\lambda}{4L_\lambda}\alpha\left(\min_{t\leq T}\|g_t\|\right)\right\} \geq \min\left\{\Delta_1, \frac{\lambda T}{4L_\lambda}\alpha\left(\min_{t\leq T}\|g_t\|\right)\right\}.$$

Therefore,

$$\alpha\left(\min_{t\leq T}\|g_t\|\right) \leq \frac{4L_\lambda}{\lambda T} \leq \alpha(\epsilon),$$

which implies $\min_{t\leq T}\|g_t\| \leq \epsilon$ since α is increasing. □

D.3 Convergence of RMSProp

In this section, we provide the rigorous proof of Theorem 5.2.3 on the convergence of RMSProp under our directional smoothness assumptions. To simplify the notation, we denote the following quantities in addition to (D.2).

$$\hat{u}_t := \frac{g_t}{\sqrt{\hat{v}_t} + \lambda}, \quad \hat{H}_t := \operatorname{diag}\left(\frac{1}{\sqrt{\hat{v}_t} + \lambda}\right). \tag{D.5}$$

In order to apply Assumption 5.4 on the direction of update $\hat{u}_t$, we need to bound the following quantity.

$$\frac{1}{r} \leq E_t := \left\| \frac{\hat{u}_t}{u_t} \right\|_\infty = \left\| \frac{|g_t| + \lambda}{\sqrt{\hat{v}_t} + \lambda} \right\|_\infty \leq r.$$

Once E_t is bounded as in the above inequality, we are able to obtain a desent lemma similar to Lemma D.2.1 for memoryless Adam. Formally, we can show the following lemma.

Lemma D.3.1. *If $1/r \leq E_t \leq r$ for all $t \leq k$ and choose $\eta_t = \min\left\{\frac{\lambda}{2RL_\lambda}, \frac{R\sqrt{\lambda}L_\lambda}{\sqrt{r}M\|g_t\|_{H_t}}\right\}$, we have*

$$f(x_{t+1}) - f(x_t) \leq -\frac{\eta_t}{2r} \|g_t\|_{H_t}^2, \quad \forall t \leq k.$$

Proof of Lemma D.3.1. For any given $t \leq k$, since $1/r \leq E_t \leq r$, we know that $\frac{1}{r}\hat{u}_t \preceq u_t \preceq r\hat{u}_t$ by definition, which implies $\ell_{x_t}(x_{t+1} - x_t) = \ell_{x_t}(\hat{u}_t) \leq RL_\lambda$ based on Assumption 5.4. Then similar to what we have done in the proof of Theorem 5.2.1, we can show that

$$\begin{aligned} f(x_{t+1}) - f(x_t) \leq& -\eta_t \Big\langle g_t, \hat{u}_t \Big\rangle + \frac{\eta_t^2 R L_\lambda}{2} \left\| \hat{u}_t \right\|^2 + \frac{\eta_t^3 M}{6} \left\| \hat{u}_t \right\|^3 \\ \leq& -\eta_t \left\| g_t \right\|_{\hat{H}_t}^2 + \frac{\eta_t^2 R L_\lambda}{2\lambda} \left\| g_t \right\|_{\hat{H}_t}^2 + \frac{\eta_t^3 M}{6\lambda^{3/2}} \left\| g_t \right\|_{\hat{H}_t}^3 \\ \leq& -\eta_t \left\| g_t \right\|_{\hat{H}_t}^2 \left(\frac{3}{4} - \frac{\eta_t^2 M}{6\lambda^{3/2}} \left\| g_t \right\|_{\hat{H}_t} \right) \\ \leq& -\eta_t \left\| g_t \right\|_{\hat{H}_t}^2 \left(\frac{3}{4} - \frac{\eta_t^2 \sqrt{r} M}{6\lambda^{3/2}} \left\| g_t \right\|_{H_t} \right) \end{aligned} \tag{D.6}$$

where the first inequality uses Lemma D.1.1; in the second inequality we use $\Big\langle g_t, \hat{u}_t \Big\rangle = \left\| g_t \right\|_{\hat{H}_t}^2$ and $\left\| \hat{u}_t \right\| \leq \frac{1}{\sqrt{\lambda}} \left\| g_t \right\|_{\hat{H}_t}$ by the definition of $\hat{u}_t$ and $\hat{H}_t$ in (D.5); the third inequality is due to $\eta \leq \frac{\lambda}{2RL_\lambda}$; the fourth inequality uses $\frac{1}{\sqrt{r}} \left\| g_t \right\|_{H_t} \leq \left\| g_t \right\|_{\hat{H}_t} \leq \sqrt{r} \left\| g_t \right\|_{H_t}$ as $1/r \leq E_t \leq r$.

Then note that by the choice of $\eta_t = \min \left\{ \frac{\lambda}{2RL_\lambda}, \frac{R\sqrt{\lambda} L_\lambda}{\sqrt{r} M \|g_t\|_{H_t}} \right\}$, we have

$$\frac{\eta_t^2 \sqrt{r} M}{6\lambda^{3/2}} \left\| g_t \right\|_{H_t} \leq \eta_t \cdot \frac{RL_\lambda}{6\lambda} \leq \frac{1}{12} \leq \frac{1}{4},$$

which implies

$$f(x_{t+1}) - f(x_t) \leq -\frac{\eta_t}{2} \left\| g_t \right\|_{\hat{H}_t}^2 \leq -\frac{\eta_t}{2r} \left\| g_t \right\|_{H_t}^2,$$

where the last inequality again use $\frac{1}{\sqrt{r}} \left\| g_t \right\|_{H_t} \leq \left\| g_t \right\|_{\hat{H}_t} \leq \sqrt{r} \left\| g_t \right\|_{H_t}$ as $1/r \leq E_t \leq r$.

□

Next, we will bound E_t in the following lemma, which is the most challenging part in our analysis of the convergence RMSProp.

Lemma D.3.2. *Choose $\beta \leq \frac{1}{4}(1 - 1/r)^2$ and $\eta_t = \min \left\{ \frac{R\sqrt{\lambda} L_\lambda}{\sqrt{r} M \|g_t\|_{H_t}}, \frac{(1-1/r)\lambda^2}{3r^2 R L_\lambda \|g_t\|_{H_t}^2} \right\}$, then we have $1/r \leq E_t \leq r$ for all $t \geq 1$.*

Proof of Lemma D.3.2. We will prove this lemma using induction. First note that $\hat{v}_1 = g_1^2$ by the update rule in Algorithm 4, which implies $1/r \leq E_1 = 1 \leq r$. Suppose $1/r \leq E_t \leq r$

for $t < k$. Based on the update rule in Algorithm 4, we also know

$$\hat{v}_k = \frac{1-\beta}{1-\beta^k} \sum_{i=0}^{k-1} \beta^i g_{k-i}^2.$$

Therefore we have

$$\begin{aligned}
\hat{v}_k - g_k^2 =& \frac{1-\beta}{1-\beta^k} \sum_{i=0}^{k-1} \beta^i (g_{k-i}^2 - g_k^2) \\
=& \frac{1-\beta}{1-\beta^k} \sum_{i=0}^{k-1} \beta^i \sum_{j=0}^{i-1} (g_{k-j-1}^2 - g_{k-j}^2) \\
=& \frac{1-\beta}{1-\beta^k} \sum_{j=0}^{k-2} (g_{k-j-1}^2 - g_{k-j}^2) \sum_{i=j+1}^{k-1} \beta^i \\
=& \sum_{j=0}^{k-2} (g_{k-j-1}^2 - g_{k-j}^2) \cdot \beta^{j+1} \frac{1-\beta^{k-j-1}}{1-\beta^k}.
\end{aligned}$$

Then we can show

$$\left|\hat{v}_k - g_k^2\right| \leq \sum_{i=0}^{k-2} \beta^{i+1} \left|g_{k-i-1}^2 - g_{k-i}^2\right|.$$

Therefore, to bound $|\hat{v}_k - g_k^2|$, it suffices to bound $\left|g_t^2 - g_{t+1}^2\right|$ for each $t < k$.

- If $|g_t| \leq |g_{t+1}|$, then obviously $\left|g_t^2 - g_{t+1}^2\right| \leq g_{t+1}^2$.
- Otherwise if $|g_t| > |g_{t+1}|$, then we have

$$\begin{aligned}
\left|g_t^2 - g_{t+1}^2\right| \leq& 2\left|(g_{t+1} - g_t) g_t\right| \\
=& 2\left|(g_{t+1} - g_t)\hat{u}_t\right| \left(\sqrt{\hat{v}_t} + \lambda\right) \\
=& \frac{2}{\eta}\left|(g_{t+1} - g_t)(x_{t+1} - x_t)\right| \left(\sqrt{\hat{v}_t} + \lambda\right) \\
\leq& \frac{2r}{\eta}\left|(g_{t+1} - g_t)(x_{t+1} - x_t)\right| \left(|g_t| + \lambda\right),
\end{aligned}$$

where the last inequality is due to $1/r \leq \mathbb{E}_t \leq r$ for all $t < k$. Note that by Lemma D.1.1, as $\ell(x_t - x_{t+1}) \leq RL_\lambda$, we have

$$\begin{aligned}
\frac{2r}{\eta}\left|(g_{t+1}-g_t)(x_{t+1}-x_t)\right| \leq& \frac{2RrL_\lambda}{\eta}\left\|x_t - x_{t+1}\right\|^2 + \frac{rM}{\eta}\left\|x_t - x_{t+1}\right\|^3 \\
=& 2\eta RrL_\lambda \left\|\hat{u}_t\right\|^2 + \eta^2 rM \left\|\hat{u}_t\right\|^3 \\
\leq& \frac{2\eta rRL_\lambda}{\lambda}\left\|g_t\right\|_{\hat{H}_t}^2 + \frac{\eta^2 rM}{\lambda^{3/2}}\left\|g_t\right\|_{\hat{H}_t}^3 \\
\leq& \frac{2\eta r^2 RL_\lambda}{\lambda}\left\|g_t\right\|_{H_t}^2 + \frac{\eta^2 r^{5/2} M}{\lambda^{3/2}}\left\|g_t\right\|_{H_t}^3 \\
\leq& \frac{3\eta r^2 RL_\lambda}{\lambda}\left\|g_t\right\|_{H_t}^2 \\
\leq& (1-1/r)\lambda.
\end{aligned}$$

where the first inequality uses Lemma D.1.1; in the second inequality is due to $\left\|\hat{u}_t\right\| \leq \frac{1}{\sqrt{\lambda}}\left\|g_t\right\|_{\hat{H}_t}$ by the definition of $\hat{u}_t$ and $\hat{H}_t$ in (D.5); the third inequality uses $\frac{1}{\sqrt{r}}\left\|g_t\right\|_{H_t} \leq \left\|g_t\right\|_{\hat{H}_t} \leq \sqrt{r}\left\|g_t\right\|_{H_t}$ as $1/r \leq E_t \leq r$; the fourth inequality is due to $\eta_t \leq \frac{R\sqrt{\lambda}L_\lambda}{\sqrt{r}M\left\|g_t\right\|_{H_t}}$; and the last inequality is due to $\eta_t \leq \frac{(1-1/r)\lambda^2}{3r^2 RL_\lambda \left\|g_t\right\|_{H_t}^2}$. Therefore, we know that

$$\left|g_t^2 - g_{t+1}^2\right| \leq (1-1/r)(\lambda\left|g_t\right| + \lambda^2) \leq 2\lambda^2 + \frac{1}{4}(1-1/r)^2\left|g_t\right|^2.$$

So far, we have shown that

$$\left|g_t^2 - g_{t+1}^2\right| \leq \max\{g_{t+1}^2, (1-1/r)^2(\lambda^2 + \left|g_t^2\right|)\} \leq g_{t+1}^2 + 2\lambda^2 + \frac{1}{4}(1-1/r)^2\left|g_t\right|^2.$$

Note that

$$\begin{aligned}
&\sum_{i=0}^{k-2}\beta^{i+1}g_{t-(i+1)}^2 \leq \sum_{i=0}^{k-1}\beta^i g_{t-i}^2 = \frac{1-\beta^k}{1-\beta}\hat{v}_k \leq \frac{\hat{v}_k}{1-\beta}, \\
&\sum_{i=0}^{k-2}\beta^{i+1}g_{t-i}^2 \leq \beta\sum_{i=0}^{k-1}\beta^i g_{t-i}^2 \leq \frac{\beta\hat{v}_k}{1-\beta}, \\
&\sum_{i=0}^{k-2}\beta^{i+1} \leq \frac{\beta}{1-\beta}.
\end{aligned}$$

Then we have

$$\left|\hat{v}_k - g_k^2\right| \leq \frac{1}{1-\beta}\left(\beta(2\lambda^2 + \hat{v}_k) + \frac{1}{4}(1-1/r)^2\hat{v}_k\right)$$
$$\leq (1-1/r)^2(\hat{v}_k + \lambda^2).$$

Therefore,

$$|E_k - 1| = \frac{\left|\sqrt{\hat{v}_k} - |g_k|\right|}{\sqrt{\hat{v}_k} + \lambda} \leq \frac{\sqrt{|\hat{v}_k - g_k^2|}}{\sqrt{\hat{v}_k} + \lambda} \leq (1-1/r)\frac{\sqrt{\hat{v}_k + \lambda^2}}{\sqrt{\hat{v}_k} + \lambda} \leq (1-1/r),$$

which implies

$$\frac{1}{r} \leq E_k \leq 2 - \frac{1}{r} \leq r.$$

Then we complete the proof by induction. □

With the above two lemmas, we are ready to prove Theorem 5.2.3.

Proof of Theorem 5.2.3. Combining Lemma D.3.1 and Lemma D.3.2, we know that for all $t \geq 1$,

$$f(x_{t+1}) - f(x_t) \leq -\frac{\eta_t}{2r}\left\|g_t\right\|_{H_t}^2.$$

Note that

$$\eta_t = \min\left\{\frac{\lambda}{2RL_\lambda}, \frac{R\sqrt{\lambda}L_\lambda}{\sqrt{r}M\left\|g_t\right\|_{H_t}}, \frac{(1-1/r)\lambda^2}{3r^2RL_\lambda\left\|g_t\right\|_{H_t}^2}\right\} \geq \min\left\{\frac{\lambda}{2RL_\lambda}, \frac{C}{\left\|g_t\right\|_{H_t}^2}\right\},$$

where we denote

$$C = \min\left\{\frac{2R^3L_\lambda^3}{rM^2}, \frac{(1-1/r)\lambda^2}{3r^2RL_\lambda}\right\}.$$

Then we have

- If $\|g_t\|_{H_t}^2 \geq C \cdot \frac{2RL_\lambda}{\lambda}$, then $\eta_t \geq \frac{C}{\|g_t\|_{H_t}^2}$. Therefore,

$$f(x_{t+1}) - f(x_t) \leq -\frac{C}{2r}.$$

- Otherwise if $\|g_t\|_{H_t}^2 \leq C \cdot \frac{2RL_\lambda}{\lambda}$, then $\eta_t \geq \frac{\lambda}{2RL_\lambda}$. Then we know that

$$f(x_{t+1}) - f(x_t) \leq -\frac{\lambda}{4rRL_\lambda} \|g_t\|_{H_t}^2 \leq -\frac{\lambda}{4rRL_\lambda} \frac{\|g_t\|^2}{\|g_t\| + \lambda} =: -\frac{\lambda}{4rRL_\lambda} \alpha(\|g_t\|),$$

  where we denote $\alpha(z) := \frac{z^2}{z+\lambda}$ for $z \geq 0$ as in the proof of Theorem 5.2.1.

Therefore we can show that for $0 \leq t \leq T$

$$f(x_{t+1}) - f(x_t) \leq -\min\left\{\frac{C}{2r}, \frac{\lambda}{4rRL_\lambda}\alpha(\|g_t\|)\right\} \leq -\min\left\{\frac{C}{2r}, \frac{\lambda}{4rRL_\lambda}\alpha\left(\min_{t\leq T}\|g_t\|\right)\right\},$$

where the last inequality uses the fact that the function α is increasing. By telescoping, noting that $f(x_1) - f(x_T) \leq \Delta_1$, we have

$$\Delta_1 \geq T \min\left\{\frac{C}{2r}, \frac{\lambda}{4rRL_\lambda}\alpha\left(\min_{t\leq T}\|g_t\|\right)\right\} \geq \min\left\{\Delta_1, \frac{\lambda T}{4rRL_\lambda}\alpha\left(\min_{t\leq T}\|g_t\|\right)\right\}.$$

Therefore,

$$\alpha\left(\min_{t\leq T}\|g_t\|\right) \leq \frac{4rRL_\lambda}{\lambda T} \leq \alpha(\epsilon),$$

which implies $\min_{t\leq T}\|g_t\| \leq \epsilon$ since α is increasing.

□

D.4 Proofs related to the example defined in Section 5.3

We first repeat the definition of the example below for completeness. Let

$$\phi(z) := \begin{cases} e^z & \text{if } z \le 0, \\ \frac{1}{2}z^2 + z + 1 & \text{if } z > 0. \end{cases}$$

Then our example objective function $f : \mathbb{R}^d \to \mathbb{R}$ with $d \ge 2$ is defined as

$$f(x) := \frac{1}{\alpha}\phi(\alpha \cdot w(x)), \text{ where } \alpha > 0 \text{ and } w(x) := x_{[1]} + \frac{1}{d-1}\left(x_{[2]} + \cdots + x_{[d]}\right). \tag{D.7}$$

Denote $a := \left(1, \frac{1}{d-1}, \ldots, \frac{1}{d-1}\right)^\top$. By standard calculations, we have

$$\nabla f(x) = \max\{\alpha f(x), \alpha w(x) + 1\} \cdot a = \begin{cases} \alpha f(x) \cdot a & \text{if } w(x) \le 0, \\ (\alpha w(x) + 1) \cdot a & \text{if } w(x) > 0. \end{cases} \tag{D.8}$$

$$\nabla^2 f(x) = \max\{\alpha^2 f(x), \alpha\} \cdot aa^\top = \begin{cases} \alpha^2 f(x) \cdot aa^\top & \text{if } w(x) \le 0, \\ \alpha \cdot aa^\top & \text{if } w(x) > 0. \end{cases} \tag{D.9}$$

Then we are ready to prove Lemma 5.3.1 below.

Proof of Lemma 5.3.1. First note that $\nabla f(x) \propto a$. Then by standard calculations, we can show that for any $x \in \mathbb{R}^d$,

$$\ell_x(\nabla f(x)) = \max\{\alpha^2 f(x), \alpha\} \left\|a\right\|^2 = \left\|\nabla^2 f(x)\right\|.$$

Therefore, by the definition of L and L_{g}, it is easy to verify that

$$L = L_{\mathrm{g}} = \alpha \left\|a\right\|^2 = \frac{\alpha d}{d-1}.$$

Next, it is also easy to see that the following choice of M satisfies Assumption 5.3.

$$M = \alpha^2 \|a\|^3 = \left(\frac{d}{d-1}\right)^{3/2} \alpha^2 \leq 3\alpha^2.$$

Next, we prove the bound on L_λ. Note that based on (D.8), we can denote

$$(p, q, \dots, q)^\top := u_\lambda(x) = \frac{\nabla f(x)}{|\nabla f(x)| + \lambda}.$$

First, for any $x \in \mathbb{R}^d$ such that $w(x) > 0$, we have

$$\nabla f(x) \succeq a \succeq (\lambda, \dots, \lambda)^\top.$$

Then it is easy to see that $\frac{1}{2} \leq p, q \leq 1$. Note that $\nabla^2 f(x) = \alpha \cdot aa^\top$ when $w(x) > 0$. We have

$$\ell_x(u_\lambda(x)) = \frac{\alpha(a^\top u_\lambda(x))^2}{\|u_\lambda(x)\|^2} = \frac{\alpha(p+q)^2}{p^2 + (d-1)q^2} \leq \frac{16\alpha}{d}.$$

Next, we will bound $\ell_x(u_\lambda(x))$ when $w(x) \leq 0$. To simplify the notation, we denote $y := \alpha f(x)/\lambda$. Then according to (D.8), it is easy to verify that

$$p = \frac{y}{y+1}, \quad q = \frac{\frac{y}{d-1}}{\frac{y}{d-1} + 1} = \frac{y}{y + d - 1}.$$

Note that $p \geq q \geq 0$, we can show that

$$\ell_x(u_\lambda(x)) = \frac{\alpha^2 f(x) \cdot (a^\top u_\lambda(x))^2}{\|u_\lambda(x)\|^2} = \frac{\alpha\lambda y(p+q)^2}{p^2 + (d-1)q^2} \leq \frac{4\alpha\lambda y p^2}{p^2 + (d-1)q^2}.$$

Note that $w(x) \leq 0$ corresponds to $0 < y \leq 1/\lambda$. We will bound $\ell_x(u_\lambda(x))$ in the following three cases.

- If $0 < y \leq 1$, we have

$$\ell_x(u_\lambda(x)) \leq 4\alpha\lambda y \leq 4\alpha\lambda.$$

- If $1 \leq y \leq d-1$, we know that $\frac{1}{2} \leq p \leq 1$ and $q \geq \frac{y}{2(d-1)}$. Then we have

$$\ell_x(u_\lambda(x)) \leq \frac{16\alpha\lambda y}{1+\frac{y^2}{d-1}} \leq \sup_{z\geq 0} \frac{16\alpha\lambda z}{1+\frac{z^2}{d-1}} \leq 8\alpha\lambda\sqrt{d-1}.$$

- If $d-1 \leq y \leq 1/\lambda$, we know that $\frac{1}{2} \leq p, q \leq 1$

$$\ell_x(u_\lambda(x)) \leq \frac{4\alpha p^2}{dq^2} \leq \frac{16\alpha}{d}.$$

Therefore, we have

$$L_\lambda = \sup_x \ell_x(u_\lambda(x)) \leq 8\alpha \max\left\{\lambda\sqrt{d-1}, \frac{2}{d}\right\}.$$

Finally, we will show that Assumption 5.4 holds for any $r, R > 1$ satifying $R \geq r^4$. For any $v \in \mathbb{R}^d$ satisfying $\frac{1}{r}v \preceq u_\lambda(x) \leq rv$, we have

$$\begin{aligned}\ell_x(v) =& \max\left\{\alpha^2 f(x), \alpha\right\} \frac{(a^\top v)^2}{\|v\|^2} \\ \leq& r^4 \max\left\{\alpha^2 f(x), \alpha\right\} \frac{(a^\top u_\lambda(x))^2}{\|u_\lambda(x)\|^2} \\ =& r^4 \ell_x(u_\lambda(x)) \leq r^4 L_\lambda \leq R L_\lambda,\end{aligned}$$

where the equality uses the fact that $v \succeq \frac{1}{r}u_\lambda(x) \succeq 0$ and $a \succeq 0$. □

Now we are ready to prove Theorem 5.3.2.

Proof of Theorem 5.3.2. First, note that if we choose

$$d \geq \frac{8C^2}{c^2} + 1, \quad \alpha = \frac{d-1}{d} \cdot C, \quad \lambda = \frac{1}{d-1},$$

then based on Lemma 5.3.1, we have

$$L = L_{\mathrm{g}} = C, \quad L_\lambda \leq c, \quad M \leq 3C^2.$$

Now, we will bound the iteration complexities of gradient descent and memoryless Adam. First, we formally define the minimum required number of iterations for an algorithm $\mathcal{A}$ to achieve an ϵ-sub-optimal point as

$$T_{\mathcal{A}} := \min_{t \geq 1} \left\{ t \mid f(x_{t+1}) - f^* \leq \epsilon \right\},$$

where $\{x_t\}_{t\geq 1}$ are the iterates generated by $\mathcal{A}$. To simplify the notation, we define the following quantities for each time step $t \geq 1$.

$$w_t := w(x_t) = a^\top x_t, \quad f_t := f(x_t) = \frac{1}{\alpha}\phi(\alpha w_t), \quad \delta_t := \alpha \cdot (w_t - w_{t+1}).$$

First, we will show a lower bound on T_{gd} for gradient descent defined by the following update rule when $w_t \leq 0$.

$$x_{t+1} = x_t - \frac{1}{L}\nabla f(x_t) = x_t - \frac{d-1}{d} f_t \cdot a.$$

Therefore, we have

$$\delta_t = \alpha \cdot (w_t - w_{t+1}) = \alpha\, a^\top (x_t - x_{t+1}) = \alpha f_t \geq 1. \tag{D.10}$$

Then by induction, we know that if $w_1 \leq 0$, then $w_t \leq w_{t-1} \leq \cdots \leq w_1 \leq 0$. Therefore, we can also show that $\delta_t = \alpha f_t \le 1$ for all $t \ge 1$, which implies

$$1 \le \frac{f_t}{f_{t+1}} = e^{\delta_t} \le e < 4. \tag{D.11}$$

Since $f_t = \frac{1}{\alpha}\phi(\alpha w_t) = \frac{1}{\alpha}e^{\alpha w_t}$ when $w_t \le 0$, applying Taylor's theorem on the function $\frac{1}{\alpha}e^{\alpha w}$ as a function of w, we have

$$\begin{aligned} f_t - f_{t+1} \le & e^{\alpha w_{t+1}} \cdot (w_t - w_{t+1}) + \frac{\alpha e^{\alpha w_t}}{2}(w_t - w_{t+1})^2 \\ = & f_{t+1}\delta_t + \frac{1}{2} f_t \delta_t^2 \\ \le & \alpha \left(f_t f_{t+1} + \frac{1}{2} f_t^2 \right) \\ \le & 3\alpha f_t f_{t+1}, \end{aligned}$$

where the equality is by the definition of f_t and δ_t; the second inequality uses (D.10) and the fact that $\delta_t \le 1$; and the last inequality is due to (D.11). Dividing both sides by $f_t f_{t+1}$, we have

$$\frac{1}{f_{t+1}} - \frac{1}{f_t} \le 3\alpha,$$

which implies

$$\frac{1}{f_{T_{\mathrm{gd}}+1}} \le \frac{1}{f_1} + 3\alpha T_{\mathrm{gd}}.$$

Therefore, to achieve $f_{T_{\mathrm{gd}}+1} \le \epsilon$, we must have

$$T_{\mathrm{gd}} \ge \frac{1}{3\alpha}\left(\frac{1}{\epsilon} - \frac{1}{f_1} \right).$$

For a small enough $\epsilon \leq f_1/2$, we have

$$T_{\mathrm{gd}} \geq \frac{1}{3\alpha\epsilon}.$$

Next, we will prove an upper bound on T_{ma} for memoryless Adam defined by the following update rule.

$$x_{t+1} = x_t - \eta u_\lambda(x_t),$$

where $\eta = \frac{\lambda}{L_\lambda} = \frac{1}{8\alpha\sqrt{d-1}} \leq \frac{1}{8\alpha}$. With standard calculations, we can obtain

$$\delta_t = \alpha \cdot (w_t - w_{t+1}) = \frac{\eta\alpha^2 f_t}{\alpha f_t + \lambda} + \frac{\eta\alpha^2 f_t}{\alpha f_t + (d-1)\lambda} \geq \frac{\eta\alpha^2 f_t}{\alpha f_t + \lambda}. \tag{D.12}$$

It is easy to see that $0 \leq \delta_t \leq 2\eta\alpha \leq \frac{1}{4}$. Then similar to our analysis for gradient descent, we can easily show that both w_t and f_t are non-increasing with t by induction. Then, by Taylor's theorem on the function $\frac{1}{\alpha}e^{\alpha w}$ as a function of w, we have

$$\begin{aligned} f_t - f_{t+1} \geq & e^{\alpha w_t} \cdot (w_t - w_{t+1}) + \frac{\alpha e^{\alpha w_t}}{2}(w_t - w_{t+1})^2 \\ = & f_t\delta_t - \frac{1}{2}f_t\delta_t^2 \\ \geq & \frac{1}{2}f_t\delta_t \\ \geq & \frac{\eta\alpha^2 f_t^2}{2(\alpha f_t + \lambda)}, \end{aligned}$$

where the equality is by the definition of f_t and δ_t; the second inequality is due to $\delta_t \leq 1/4$; and the last inequality is due to (D.12). Define

$$\tau = \min\{t \mid f_t < \lambda/\alpha\}.$$

Then when $t < \tau$, we have $f_t \geq \lambda/\alpha$ and therefore

$$f_t - f_{t+1} \geq \frac{\eta\alpha}{4} f_t,$$

which implies

$$f_{t+1} \leq \left(1 - \frac{\eta\alpha}{4}\right) f_t \leq e^{-\eta\alpha/4} f_t \leq \cdots \leq e^{-\eta\alpha(t+1)/4} f_1.$$

Also note that $f_{\tau-1} \geq \lambda/\alpha$. We can show that

$$\tau \leq 1 + \frac{4L_\lambda \log(\alpha f_1/\lambda)}{\alpha\lambda}.$$

On that other hand, if $t \geq \tau$, we know that $f_t < \lambda/\alpha$ and thus

$$f_t - f_{t+1} \geq \frac{\eta\alpha^2 f_t^2}{4\lambda} \geq \frac{\eta\alpha^2 f_t f_{t+1}}{4\lambda}.$$

Dividing both sides by $f_t f_{t+1}$, we have

$$\frac{1}{f_{t+1}} - \frac{1}{f_t} \geq \frac{\eta\alpha^2}{4\lambda}.$$

Therefore we have

$$\frac{1}{f_{T_{\mathrm{ma}}}} - \frac{1}{f_\tau} \geq \frac{\eta\alpha^2}{4\lambda}(T_{\mathrm{ma}} - \tau).$$

As $f_{T_{\mathrm{ma}}} \geq \epsilon$, we have

$$\begin{aligned} T_{\mathrm{ma}} \leq & 1 + \tau + \frac{4\lambda}{\eta\alpha^2}\left(\epsilon^{-1} - \lambda^{-1}\right) \\ \leq & 2 + 32\sqrt{d-1}\log(\alpha(d-1)f_1) + \frac{32}{\alpha\epsilon\sqrt{d-1}}. \end{aligned}$$

For a small enough ϵ satifying $\epsilon \leq \frac{1}{\alpha\sqrt{d-1}+\alpha(d-1)\log(\alpha(d-1)f_1)}$, we have

$$T_{\mathrm{ma}} \leq \frac{64}{\alpha\epsilon\sqrt{d-1}} \leq \frac{32}{\alpha\epsilon} \cdot \frac{c}{C}.$$

Therefore we have shown $T_{\mathrm{gd}}/T_{\mathrm{ma}} \geq 96C/c$ which completes the proof.

□

Bibliography

Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). *arXiv preprint arXiv:2310.01082*, 2023.

Kwangjun Ahn, Zhiyu Zhang, Yunbum Kook, and Yan Dai. Understanding adam optimizer via online learning of updates: Adam is ftrl in disguise, 2024.

Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International conference on machine learning*, pages 699–707. PMLR, 2016.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2):165–214, 2023.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, pages 1–50, 2017.

Congliang Chen, Li Shen, Fangyu Zou, and Wei Liu. Towards practical adam: Non-convexity, convergence theory, and mini-batch acceleration. *The Journal of Machine Learning Research*, 23(1):10411–10457, 2022.

Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. *arXiv preprint arXiv:1808.02941*, 2018.

Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. *arXiv preprint arXiv:2303.02854*, 2023.

Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signsgd. *Advances in Neural Information Processing Systems*, 35:9955–9968, 2022.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *ArXiv*, abs/1905.10018, 2019.

Soham De, Anirbit Mukherjee, and Enayat Ullah. Convergence guarantees for rmsprop and adam in non-convex optimization and an empirical comparison to nesterov acceleration. *arXiv: Learning*, 2018.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

Alexandre D'efossez, Léon Bottou, Francis R. Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *arXiv: Machine Learning*, 2020.

Alexandre d'Aspremont, Damien Scieur, and Adrien Taylor. Acceleration methods. *Foundations and Trends® in Optimization*, 5(1-2):1–245, 2021. ISSN 2167-3888. doi:10.1561/2400000036. URL http://dx.doi.org/10.1561/2400000036.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.

Matthew Faw, Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Beyond uniform smoothness: A stopped analysis of adaptive sgd. *ArXiv*, abs/2302.06570, 2023.

Sébastien Gadat and Ioana Gavra. Asymptotic study of stochastic adaptive algorithms in non-convex landscape. *The Journal of Machine Learning Research*, 23(1):10357–10410, 2022.

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

Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. A novel convergence analysis for algorithms of the adam family. *ArXiv*, abs/2112.03459, 2021.

Hideaki Iiduka. Theoretical analysis of adam using hyperparameters close to one without lipschitz smoothness. *Numerical Algorithms*, pages 1–39, 2023.

Kaiqi Jiang, Dhruv Malik, and Yuanzhi Li. How does adaptive optimization impact local neural network geometry? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=gIG8LvTLuc.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, 2013.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between sgd and adam on transformers, but sign descent might be. *arXiv preprint arXiv:2304.13960*, 2023.

Frederik Kunstner, Robin Yadav, Alan Milligan, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why adam outperforms gradient descent on language models, 2024.

Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via scsg methods. *Advances in Neural Information Processing Systems*, 30, 2017.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International conference on machine learning*, pages 6286–6295. PMLR, 2021.

Deyi Liu, Lam M Nguyen, and Quoc Tran-Dinh. An optimal hybrid variance-reduced algorithm for stochastic composite nonconvex optimization. *arXiv preprint arXiv:2008.09055*, 2020.

Zijian Liu, Perry Dong, Srikanth Jagabathula, and Zhengyuan Zhou. Near-optimal high-probability convergence for non-convex stochastic optimization with variance reduction. *arXiv preprint arXiv:2302.06032*, 2023.

Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. Adaptive gradient methods with dynamic bound of learning rate. *ArXiv*, abs/1902.09843, 2019.

Julien Mairal. Optimization with first-order surrogate functions. In *International Conference on Machine Learning*, pages 783–791. PMLR, 2013.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Yan Pan and Yuanzhi Li. Toward understanding why adam converges faster than sgd for transformers. *arXiv preprint arXiv:2306.00204*, 2023.

Jiang Qian, Yuren Wu, Bojin Zhuang, Shaojun Wang, and Jing Xiao. Understanding gradient clipping in incremental gradient methods. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 314–323, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/reddi16.html.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *ArXiv*, abs/1904.09237, 2018.

Amirhossein Reisizadeh, Haochuan Li, Subhro Das, and Ali Jadbabaie. Variance-reduced clipping for non-convex optimization. *arXiv preprint arXiv:2303.00883*, 2023.

Nicolas Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence _rate for finite training sets. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/905056c1ac1dad141560467e0a99e1cf-Paper.pdf.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(1), 2013.

Naichen Shi, Dawei Li, Mingyi Hong, and Ruoyu Sun. Rmsprop converges with proper hyper-parameter. In *International Conference on Learning Representations*, 2021.

Quoc Tran-Dinh, Nhan H Pham, Dzung T Phan, and Lam M Nguyen. Hybrid stochastic gradient descent algorithms for stochastic nonconvex optimization. *arXiv preprint arXiv:1905.05920*, 2019.

Bohan Wang, Yushun Zhang, Huishuai Zhang, Qi Meng, Zhi-Ming Ma, Tie-Yan Liu, and Wei Chen. Provable adaptivity in adam. *arXiv preprint arXiv:2208.09900*, 2022.

Bohan Wang, Jingwen Fu, Huishuai Zhang, Nanning Zheng, and Wei Chen. Closing the gap between the upper bound and lower bound of adam's iteration complexity. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=yDvb3mlogA.

Ruiqi Wang and Diego Klabjan. Divergence results and convergence of a variance reduced version of adam. *ArXiv*, abs/2210.05607, 2022.

Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. *Advances in Neural Information Processing Systems*, 33:15511–15521, 2020a.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020b.

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhimin Luo. Adam can converge without any modification on update rules. *ArXiv*, abs/2208.09632, 2022.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need adam: A hessian perspective, 2024.

Shen-Yi Zhao, Yin-Peng Xie, and Wu-Jun Li. On the convergence and improvement of stochastic normalized gradient descent. *Science China Information Sciences*, 64, 2021.

Dongruo Zhou, Jinghui Chen, Yuan Cao, Yiqi Tang, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018a.

Zhiming Zhou, Qingru Zhang, Guansong Lu, Hongwei Wang, Weinan Zhang, and Yong Yu. Adashift: Decorrelation and convergence of adaptive learning rate methods. *ArXiv*, abs/1810.00143, 2018b.

Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of adam and rmsprop. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11119–11127, 2018.