EPFL

Thèse n° 9077

Universal and adaptive methods for robust stochastic optimization

Présentée le 22 août 2023

Faculté informatique et communications
Laboratoire de systèmes d'information et d'inférence
Programme doctoral en informatique et communications

pour l'obtention du grade de Docteur ès Sciences

par

Ali KAVIS

Acceptée sur proposition du jury

Prof. M. Jaggi, président du jury
Prof. V. Cevher, directeur de thèse
Prof. Y. Nesterov, rapporteur
Prof. S. Wright, rapporteur
Prof. N. Flammarion, rapporteur

École polytechnique fédérale de Lausanne

2023

To my lovely wife Pınar,
and my parents Melike & Şadi . . .

Acknowledgements

During my PhD adventure, I have had the chance to know many amazing people who made this journey enjoyable and memorable.

I would like to begin with a special thanks to my supervisor Volkan Cevher. He has always been a constant source of energy and excitement while providing genuine support during my studies. We had many fruitful research discussions and he always showed his trust in me as an independent researcher. I am truly grateful for the collaborations he initiated with many external researchers through which I have learned a lot and established several invaluable academic partnerships.

It was a privilege to have Yurii Nesterov, Stephen Wright, Martin Jaggi and Nicolas Flammarion in my thesis jury. They made my thesis presentation such an enlightening, stimulating experience and gave constructive feedback to improve my thesis. I also would like to thank Yurii Nesterov for the intriguing research discussions we had.

I consider myself lucky to be able to connect with several external reserachers. Particularly, it was a pleasure to have worked with Kfir Levy and Panayotis Mertikopoulos in several projects during my PhD. Kfir and his papers on adaptive methods have fundamentally inspired my work during my studies. Panayotis has always been available for my technical questions and career advise.

I feel lucky that I was a part of such a diverse, smart and lively group of people at LIONS. Paul Rolland, my first collaborator in academia, is one of those people who is reasonably good at a variety of things; research, climbing, running, cycling, badminton, paragliding and juggling (he literally juggles fire). I must thank him for his optimism and joy in work place, and teaching me various patterns of juggling. Fabian Latorre, the young investor, is one of my first friends in Lausanne. I will always remember his low-pitched “Bye”. I want to thank Ya-Ping Hsieh, and Ahmet, for introducing me to the third-wave coffee brewing. I will miss our morning brew routines, all-you-can-eat feasts and powerlifting trainings. A big thank you goes to Fatih and Baran for our regular döner lunches and afternoon tea sessions (with a secret recipe). My first office mate, Thomas Sanchez, is one of the most proper people I have ever met in many aspects. I have many good memories from the dinner parties we had at his place. Leello Tadesse, the comedian of the lab, has always been a unstoppable source of positivity and laughter. I don’t remember any single moment with him that wasn’t happy. Isn’t it amazing that we also share the same birthday? Speaking of positivity, Thomas Pethick deserves an honorable mention with his ever-present, peaceful smile. I would like to thank him and Elisa

for being incredible road trip partners, and welcoming hosts for parties and movie nights. I owe a huge thank you to Pedro Abranches, the prince (literally), for organizing the Portugal trip, among other things. It was a complete experience with the surfing, incredible sea food of all kinds, city tours (free of charge) and the celebrations in his village. I want to thank Igor for introducing me to dancing and acroyoga and motivating me to continue practicing even after I move to the US. I thank my office mates Grigoris, Stratis and Kimon, the Greek team, Luca and Andrej for the fun we had in the office and the coffee breaks. I hope we could complete our Torino road trip with Luca in the near future for which I am already excited. I have been working with Stratis and Kimon since they joined the lab and I have learned a lot from them. I was fortunate to have them as collaborators and friends. I will miss the The Wire screenings and dinners at Stratis', and the loud laughters and heated research discussions we had with Kimon. I also want to give a big thank you to Fanghui Liu for finding the best Chinese restaurants in town and organizing feasts for the whole lab. I also have a few thanks to some of the alumni of the lab. In my third year, we had a tradition of regular lab outings, which I truly enjoyed, with Yura Malitsky, Maria Vladarean, Panayotis, Nadav Hallak and Kamal Parameswaran and many others. Although they were interrupted by the lockdown at the time, we continued the tradition over video calls. Yura, Nadav and Panayotis are great researchers and wonderful academic mentors, and it was a privilege to have worked with them in several projects. Last but not least, our lab admin Gosia Baltaian deserves a huge thank you for all the things she did for me during my life in Switzerland. She helped me with many things whenever I needed help: PhD extensions, permit renewals, finding a new apartment, filling the tax declaration form, tracking visa applications and many more. I am greatful for all her help even if it wasn't her responsibility in many occasions.

Looking back at my time at EPFL, I was fortunate to have met so many amazing people that made it a truly colorful experience. I would like to thank Ahmet Alacaoglu for being a close friend and helping me adapt to my new life since day one; finding me an apartment, introducing me to powerlifting, coaching me through multiple years and introducing me to many people in Lausanne. We had several road trips and vacations, conference travels, schnitzel and cocktail parties which were always guaranteed to be fun. I must thank Alp Yurtsever for being the host of several cocktail nights and parties, the perfect travel companion and having a reasonable, sound advice for anything from career to life choices. Beril Besbinar is the sole reason we had amazing vacations, planned to the finest details, and house parties. She thought me how to enjoy the outdoors and introduced me to via-ferrata. I am greatful to know the funniest duo, Ceyhun Alp and Okan Altingovde, who make the most extraordinary observations and jokes. I am always impressed by Ceyhun's immense memory of movie lines, memes and comics. Okan is an interesting person who removed the word "worry" from his vocabulary. Without Pinar Akyazi, or Pinob, the 4 pitchers team wouldn't be complete. She has been at the center of many memorable moments with her ability to crack at random moments, uncontrollable laughters and the tendency to put herself in odd, unusual situations.

I am greateful to meet Yigitcan Kaya, who was my classmate and gym partner in the college and the organizer of our road trips in the US. During the lockdowns, our book reading club

with Ahmet, and whisky tasting sessions with Alp and Ahmet, over video calls, made it easier to go through the difficult times. I would like to thank Sigurd Alnes for inviting us to Norway and planning an amazing trip. We learned cross-country skiing, ice-fishing and brewing beer (brewed in a bathroom, fermented in a sauna). It was one of the most memorable vacations that I have ever had.

Finally, and most importantly, I would like to express my gratitude to my wife and family. Without their support, care and efforts, this journey wouldn't be as comfortable and meaningful. My wife and my best friend, Pınar, has always been by my side; whenever I needed somebody to talk to, she would give her full attention until she makes sure that I am feeling better. I am so lucky to have such an incredible person in my life with whom I share a lot in common. I enjoy every second that we spend together; it is nothing but pure joy to explore life with her. Her support and love has been one of the driving forces in my life.

My parents Melike and Şadi have been my biggest supporters my whole life. They always encouraged me to explore and discover what I like to do in life and did everything in their power to make sure I have the means to do so. I wanted to study and live abroad since high school and they have been a constant source of inspiration and support in this journey. This wouldn't be possible without them. I also want to thank my parents-in-law, Ferihan and Mustafa, for welcoming me to their family and providing their true, genuine support during my studies. It is priceless to have such a family by my side.

Lausanne, August 7, 2023 A. K.

Abstract

Within the context of contemporary machine learning problems, efficiency of optimization process depends on the properties of the model and the nature of the data available, which poses a significant problem as the complexity of either increases ad infinitum. An overarching challenge is to design fast, adaptive algorithms which are provably robust to increasing complexities and unknown properties of the optimization landscape, while ensuring scalable implementation at scale.

Having that said, there are two main perspectives to consider: (i) standard proof techniques require precise knowledge of the model and loss function parameters, which are usually prohibitively expensive to estimate; (ii) state-of-the-art methods which show superior performance in practice are mostly heuristics, lacking theoretical basis.

In this dissertation, the reader will be presented with several fundamental problem formulations in machine learning which will be studied from the aforementioned perspectives. Specifically, the focus of this dissertation will be on two fundamental concepts; (i) adaptivity: ability of an algorithm to converge without knowing the problem-dependent parameters and (ii) universality: ability of an algorithm to converge adaptively under multiple problem settings simultaneously without any modifications.

In the light of this terminology, the goal is to unify the discrepancy between the theory of adaptive algorithms and the heuristic approaches employed in practice. To this end, the results are presented in three chapters based on the properties of the optimization problem; convex minimization, non-convex optimization and monotone variational inequalities.

We begin with a universal and adaptive algorithm for compactly constrained convex minimization, which achieves order-optimal convergence rates for smooth/non-smooth problems under deterministic/stochastic oracles, simultaneously. We identify an alternative acceleration scheme together with an appropriate adaptive step-size that enables optimal convergence rates without knowing, a priori, neither the problem parameters nor the problem setting at hand. Then, we propose the first noise-adaptive second-order algorithm which individually adapts to noise in gradient and Hessian estimates.

Moving on non-convex minimization, we have two set of results to present; high probability convergence of adaptive gradient method and adaptive variance reduction methods under two scenarios. For the former, we analyze the AdaGrad algorithm under two noise models; bounded variance and sub-Gaussian noise. We provide order-optimal high probability rates while establishing a set of side results on the boundedness of the iterate sequence. For the latter setting of variance reduction, we study both the more generic setting of streaming

data and the more practical sub-setting of finite-sum minimization. Under both scenarios, we develop the first parameter-free variance reduction methods with optimal rates in their respective problem settings.
Finally, we study the problem of solving monotone variational inequalities under two noise models; standard bounded variance and the relative noise model in which the error in operator computation is proportional to its norm at the evaluation point. With a compatible, mild set of assumptions, we prove that a class of extra-gradient algorithms with a particular adaptive step-size universally adapts to both models of noise without knowing the setting, a priori.

Key words: stochastic optimization, adaptive methods, convex minimization, non-convex minimization, variational inequalities, parameter-free methods, variance reduction, high-probability convergence, first and second-order methods, noise-adaptive algorithms.

Résumé

Dans le cadre des problèmes d'apprentissage automatique contemporains, l'efficacité du processus d'optimisation dépend des propriétés du modèle et de la nature des données disponibles, ce qui pose un problème important lorsque la complexité de l'un ou l'autre augmente drastiquement. Un défi primordial consiste à concevoir des algorithmes rapides et adaptatifs qui sont prouvablement robustes aux complexités croissantes et aux propriétés inconnues du paysage d'optimisation, tout en garantissant une mise en oeuvre efficace à grande échelle.

Ceci étant dit, il y a deux perspectives principales à considérer : (i) les techniques de preuve standard nécessitent une connaissance précise des paramètres du modèle et de la fonction de perte, qui sont généralement très couteux à estimer ; (ii) les méthodes de pointe qui montrent des performances supérieures dans la pratique sont pour la plupart heuristiques, dépourvues de base théorique.

Dans cette thèse, il sera présenté au lecteur plusieurs formulations de problèmes fondamentaux en apprentissage automatique qui seront étudiées à partir des perspectives susmentionnées. Plus précisément, l'accent de cette thèse sera sur deux concepts fondamentaux ; (i) adaptabilité : capacité d'un algorithme à converger sans connaître les paramètres dépendant du problème et (ii) universalité : capacité d'un algorithme à converger de manière adaptative sous plusieurs paramètres de problème simultanément sans aucune modification.

À la lumière de cette terminologie, l'objectif est d'unifier l'écart entre la théorie des algorithmes adaptatifs et les approches heuristiques utilisées dans la pratique. À cet effet, les résultats sont présentés en trois chapitres basés sur les propriétés du problème d'optimisation ; minimisation convexe, optimisation non convexe et inégalités variationnelles monotones.

Nous commençons avec un algorithme universel et adaptatif pour la minimisation convexe contrainte de manière compacte, qui atteint des taux de convergence optimaux pour des problèmes lisses/non lisses sous des oracles déterministes/stochastiques, simultanément. Nous identifions un schéma d'accélération alternatif avec une taille de pas adaptative appropriée qui permet des taux de convergence optimaux sans connaître, a priori, ni les paramètres du problème ni le problème à résoudre. Ensuite, nous proposons le premier algorithme de second ordre adaptatif au bruit qui s'adapte individuellement au bruit dans les estimations de gradient et de Hessienne.

Pour ce qui est de la minimisation non convexe, nous avons deux ensembles de résultats à présenter ; convergence avec probabilité élevée de la méthode du gradient adaptatif et des méthodes de réduction adaptative de variance dans deux scénarios. Pour le premier, nous

analysons l'algorithme AdaGrad sous deux modèles de bruit ; variance bornée et bruit sous-Gaussien. Nous fournissons des taux de convergence optimaux avec grande probabilité tout en établissant un ensemble de résultats secondaires sur la délimitation de la séquence d'itérations. Pour ce dernier cadre de réduction de variance, nous étudions à la fois le cadre plus générique des données en continu et le sous-ensemble plus pratique de la minimisation à somme finie. Dans les deux scénarios, nous développons les premières méthodes de réduction de variance sans paramètre avec des taux optimaux dans leurs paramètres de problème respectifs.
Enfin, nous étudions le problème de résolution des inégalités variationnelles monotones sous deux modèles de bruit ; le modèle standard de variance bornée et le modèle de bruit relatif dans lequel l'erreur de calcul de l'opérateur est proportionnelle à sa norme au point d'évaluation. Avec un ensemble d'hypothèses compatibles et faibles, nous prouvons qu'une classe d'algorithmes extra-gradient avec une taille de pas adaptative particulière s'adapte universellement aux deux modèles de bruit sans connaître le réglage, a priori.

Mots clés : optimisation stochastique, méthodes adaptatives, minimisation convexe, minimisation non convexe, inégalités variationnelles, méthodes sans paramètres, réduction de la variance, convergence à haute probabilité, méthodes du premier et du second ordre, algorithmes adaptatifs au bruit.

Bibliographic Note

This dissertation is based on the following publications:

- A. Kavis et al. “UniXGrad: A Universal, Adaptive Algorithm with Optimal Guarantees for Constrained Optimization”. In: Advances in Neural Information Processing Systems 32. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 6260–6269.
- K. Y. Levy, A. Kavis, and V. Cevher. “STORM+: Fully Adaptive SGD with Recursive Momentum for Nonconvex Optimization”. In: Advances in Neural Information Processing Systems. Ed. by A. Beygelzimer et al. 2021.
- K. Antonakopoulos et al. “Sifting through the noise: Universal first-order methods for stochastic variational inequalities”. In: Advances in Neural Information Processing Systems. Ed. by A. Beygelzimer et al. 2021.
- A. Kavis, K. Y. Levy, and V. Cevher. “High Probability Bounds for a Class of Nonconvex Algorithms with AdaGrad Stepsize”. In: International Conference on Learning Representations. 2022.
- A. Kavis et al. “Adaptive Stochastic Variance Reduction for Non-convex Finite-Sum Minimization”. In: Advances in Neural Information Processing Systems. Ed. by A. H. Oh et al. 2022.
- K. Antonakopoulos, A. Kavis, and V. Cevher. “Extra-Newton: A First Approach to Noise-Adaptive Accelerated Second-Order Methods”. In: Advances in Neural Information Processing Systems. Ed. by A. H. Oh et al. 2022.

The aforementioned papers are individually studied in their own dedicated sections. At the beginning of each section, a detailed bibliographic note has been attached that summarizes the contribution of the author of this dissertation as well as the co-authors of each paper.

In addition to the publications listed above, I am a co-author of the following papers which are not included in this dissertation:

- Y. P. Hsieh, A. Kavis, P. Rolland, and V. Cevher. "Mirrored langevin dynamics." In: Advances in Neural Information Processing Systems. 2018.
- P. Rolland, A. Kavis, A. Immer, A. Singla, and V. Cevher. "Efficient learning of smooth

probability functions from Bernoulli tests with guarantees." In: International Conference on Machine Learning. 2019.

- P. Mertikopoulos, N. Hallak, A. Kavis, and V. Cevher. "On the almost sure convergence of stochastic gradient descent in non-convex problems." In: Advances in Neural Information Processing Systems. 2020.
- P. Rolland, A. Eftekhari, A. Kavis, and V. Cevher. "Double-loop unadjusted langevin algorithm." In International Conference on Machine Learning. 2020.

Contents

1 Introduction

Machine learning has gained overwhelming and rapid popularity thanks to vast availability of processed data, accessibility of tremendous compute power and development of innovative learning models. From simple regression problems to image classification via convolutional neural networks, from generative adversarial networks to large language models, there lies an optimization algorithm at the heart of the training process. Training a machine learning model is essentially equivalent to optimizing an appropriate objective function using an iterative optimization algorithm.

While machine learning research benefits from the advancements in the theory of mathematical optimization, the practical developments and the state-of-the-art approaches predominantly employ heuristic techniques which are not completely verified by the theory. In order to provide meaningful and interpretable conclusions, we sometimes need to make fundamental assumptions on the optimization/learning problem at hand. However, those assumptions might fall short of the reality as they might not be representative of the complexity and generality of what is studied in practice. This creates a gap between the theoretical understanding of the optimization process and heuristics-based state-of-the-art results in practice.

The main objective of my research as presented in this dissertation is taking a step to bridge this gap by designing algorithms that

(i) are capable of solving different optimization problems simultaneously without any modifications,

(ii) could adapt to the local variations on the loss landscape,

(iii) adjust to different levels and types of noise in the computation of the value, gradient and/or Hessian of the objective function necessary for the optimization algorithm.

With increasing complexity and scale of learning models and data dimensions, it becomes more difficult to asses and estimate basic, mathematical properties of the objective function to be optimized. This is crucial as many classical algorithms and the standard analysis approaches require the knowledge of certain problem-dependent parameters to ensure con-

vergence. However more often than not, it is significantly difficult to even estimate such parameters in practice. An important reflection of this in practice is the so-called tuning process. We need to execute several test runs to find a good combination of parameter values for the optimization algorithms, complexity of which grows exponentially with the increasing number of algorithm parameters.

My research provides partial answers to this fundamental problem by designing adaptive and robust algorithms that automatically adjust to the problem formulation at hand, noise in the optimization process and local variations in the loss landscape along the optimization path. The particular design approaches adopted in the development process not only yields more flexible algorithms but also less parameters to tune. In this dissertation, the notions of adaptivity and robustness are investigated under three main problem formulations; constrained convex minimization, smooth non-convex minimization and monotone variational inequalities. Each problem setting will be studied from particular perspectives in their respective chapters.

1.1 Problem formulation

In the whole of this manuscript, we will consider two fundamental optimization problems. First, we will investigate the following generic minimization problem

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

where $f : \mathbb{R}^d \to R$ is a proper, continuous function and $\mathcal{X} \subseteq \mathbb{R}^d$ is a closed, convex subset of $\mathbb{R}^d$. Predominantly, this dissertation focuses on this problem setting, studying it under different structural and regularity assumptions on the objective, and various sets of assumptions on the black-box oracles that provides us with zeroth, first or second-order information with respect to the objective function at hand. We will mostly consider the case when the objective f is first-order L-Lipschitz smooth such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathcal{X}.$$

This is equivalent to saying f has L-Lipschitz continuous gradient. In the sequel, we will also consider the scenarios where objective f or its Hessian is Lipschitz continuous, as well. The second type of problem we will study is the so-called variational inequality problem [FP03],

$$\text{Find } x^* \in \mathbb{R}^d \text{ s.t. } \langle A(x^*), x - x^* \rangle \geq 0 \tag{VI}$$

where the operator A is a *monotone*, *continuous* vector field. The (VI) problem intersects with (Min) when f is a differentiable convex function and $A = \nabla f$ is the vector field induced by the gradient of f. Beyond that, (VI) is a versatile framework that covers many interesting problems in machine learning, e.g., adversarial training, multi-agent games and reinforcement learning [Ant+21].

Now, let us formalize the metrics for evaluating performance of optimization algorithm, running under each problem setup. For the (Min) problem with convex objectives, we are interested in finding an ϵ-optimal point $\hat{x}$ with respect to objective sub-optimality gap,

$$f(\hat{x}) - \min_{x \in \mathcal{X}} f(x) \leq \epsilon. \tag{1.1}$$

In the case of non-convex objectives, our focus would in turn be finding a first-order ϵ-stationary point $\hat{x}$ with respect to the norm of the objective's gradient,

$$\|\nabla f(\hat{x})\| \leq \epsilon. \tag{1.2}$$

Last but not least, for the (VI) problem, our metric of choice is the (restricted) gap [Nes07],

$$\mathrm{Gap}_{\mathcal{X}}(\hat{x}) = \sup_{x \in \mathcal{X}} \langle A(x), \hat{x} - x \rangle \leq \epsilon, \tag{1.3}$$

where $\hat{x}$ is a candidate solution of the problem returned by the algorithm at hand.

1.2 Background and challenges

Our aim in the sequel is to present the reader with a set of representative machine learning and optimization problems under standard sets of assumptions in the literature for (Min) and (VI), respectively identify challenges for the analysis of adaptive and universal methods and exhibit our appropriately-designed algorithms with the key novelties and contributions.

Indeed, (Min) and (VI) formulations embrace a great majority of contemporary and popular learning problems; (non-negative) matrix factorization and completion [GLM16; Hoy04], training convolutional neural networks for image classification [He+16; Den+09a] or attention models for natural language processing [Dev+19], adversarial reinforcement learning [Pin+17] and generative adversarial networks [Goo+14] (GANs) among others. Essentially, with ever-so-increasing model complexity and data sizes, such large-scale problems raise efficiency and scalability concerns. Therefore, we are interested in developing scalable, robust and efficient algorithms to be able to solve such complex and large-scale problems.

Naturally, the *first-order methods*, which basically have access to the gradient of the objective $\nabla f(\cdot)$ for (Min) or operator value $A(\cdot)$ for (VI) at the point of query, have recently been the at the forefront of solving many machine learning problems. Due to their efficient per-iteration complexity, global convergence guarantees and easy extensions to stochastic problems, they have risen to recent fame. Considering the exemplary setting of convex programming, first-order methods have slower, sub-linear convergence rates in comparison to their predecessors interior-point methods [NN94], however, they are also significantly more efficient with respect to per-iteration complexity. Interior-point methods use the Newton's methods as a subsolver, hence they trade-off fast convergence rates with heavier per-iteration cost and non-existance of global gurantees. This poses efficiency issues in the presence of large-scale problems.

In fact, even the simplest first-order, *deterministic* method raises scalability concerns for optimizing models with billions of parameters over datasets with million dimensions because the algorithm must make a complete pass over the whole dataset just to compute a single gradient. This is exactly where the *stochastic* methods come to rescue. By appropriately sampling a smaller batch of the data we could compute much cheaper estimates with lighter memory overhead at the cost of slower $O(1/\sqrt{T})$ convergence rate [NYD83; JNT11].

Now, we will take a look into the thereotical properties of first-order methods and pinpoint some of its shortcomings within the particular context of this dissertation. Let us consider the problem setting (Min) for first-order L-smooth objective for the sake of brevity and take the gradient descent (GD) algorithm as an example;

$$X_{t+1} = X_t - \gamma_t \nabla f(X_t). \tag{1.4}$$

The smoothness condition could easily be verified for many applications such as empirical risk minimization with least-squares and neural networks with sigmoid activations. This regularity condition essentially bounds the variation of the gradient/operator by the distance between the query points, eliminating abrupt changes in the optimization landscape. In turn, standard algorithms must know the smoothness modulus and set their step-sizes respectively small enough to guarantee convergence to a solution. In addition, such algorithms must know *a priori* the nature of the oracle and even the precise variance bounds for solving the problem at the optimal rates [NYD83; Nes18; Lan12].

Specifically, analysis of gradient descent hinges on the (expected) descent of the objective value to verify monotonic convergence to the solution of the problem. Let us provide an abridged proof of GD with a constant step-size $\gamma_t = \gamma$ for the problem at hand when objective f is convex (for a detailed display, see [TM13]). Note that this is one of the many proofs of gradient descent, which is more suitable for the techniques we will present in this manuscript. First, we show that the objective value decreases every iteration using the update rule and smoothness:

$$\begin{aligned}
[f(X_{t+1}) - f(x^*)] - [f(X_t) - f(x^*)] &\leq \langle \nabla f(X_t), X_{t+1} - X_t \rangle + \frac{L}{2}\|X_{t+1} - X_t\|^2 \\
&\leq \left(\frac{\gamma L}{2} - 1\right)\gamma \|\nabla f(X_t)\|^2 \\
&\leq -\frac{1}{2L}\|\nabla f(X_t)\|^2,
\end{aligned}$$

where we set $\gamma = 1/L$ to make sure progress every step. Then, by combining the so-called descent lemma with convexity of the objective and using the quadratic expansion $(a+b)^2 = a^2 + 2ab + b^2$,

$$f(X_{t+1}) - f(x^*) \leq \frac{L}{2}\left(\|X_t - x^*\|^2 - \|X_{t+1} - x^*\|^2\right)$$

Now we sum up for $t = 1, \cdots, T$ and telescope the (RHS) of the expression,

$$\sum_{t=1}^{T} f(X_{t+1}) - f(x^*) \leq \frac{L}{2} \| X_1 - x^* \|^2$$

Then, by using the monotonic decrease of the objective $f(X_t) - f(x^*) \leq f(X_s) - f(x^*)$, for all $s \leq t$, we get the rate

$$f(X_{T+1}) - f(x^*) \leq \frac{L \| X_1 - x^* \|^2}{T}.$$

In order to verify monotonic decrease in sub-optimality gap, and apply the quadratic expansion, we need to set our step-size as $\gamma \leq 1/L$ [Nes03]. In general, the classical algorithms must know not only the smoothness constant, but also whether the function is smooth or not. When the *objective itself* is L-Lipschitz continuous, not the *gradient*, then the step-size should either decrease over time at a rate of $\gamma_t = O(1/\sqrt{t})$ [Nes03], or we compute it with respect to the *fixed* time horizon.

A similar issue arises in the presence of stochasticity. Let us consider the same setting as above under bounded variance assumption [Lan20, Chapter 4.1];

$$\mathbb{E}\left[\| \nabla f(x, \xi) - \nabla f(x) \| | \sigma(x) \right] \leq \sigma^2 \quad \text{and} \quad \mathbb{E}\left[\nabla f(x, \xi) | \sigma(x) \right] = \nabla f(x),$$

where $\nabla f(x, \xi)$ is a stochastic gradient estimate with respect to a random vector $\xi \sim \mathscr{D}$, drawn from some distribution $\mathscr{D}$ and $\sigma(x)$ denotes the sigma algebra generated by the randomness in the computation of vector x. To achieve respective, order-optimal rates, the stochastic gradient descent (SGD) algorithm not only requires the smoothness constant but also needs to set a decreasing step-size as $\gamma_1 \leq \frac{1}{L}$ and $\gamma_t = O\left(\frac{1}{\sqrt{t}}\right)$ [Nem+09, Eq. (2.25)], [Lan12, Theorem 1]. All in all, standard techniques rely on the precise knowledge of the structure and regularity of the *objective* as well as the nature of *oracle* information to ensure convergence of the algorithms at the best possible rates. Similar approaches exist for different variations of both (Min) and (VI) problems and the following question is a natural consequence of this observation:

To what extend we could design adaptive and universal algorithms which are oblivious to problem dependent parameters and also robust to changes along the optimization path and oracle information?

This dissertation aims at answering this question by studying different problem formulations that falls under (Min) and (VI). More specifically, we will identify possible shortcomings of the classical techniques, and propose new algorithmic design approaches with compatible analysis techniques for the particular problem formulation in question. The chapters are separated with respect to the problem formulations they study and we essentially examine different proof techniques that enables us to go beyond the standard analysis under different sets of problem definitions and assumptions.

1.3 Organization

We will study the answer(s) of the foregoing question under three problem settings: convex minimization with compact constraints; smooth, non-convex minimization under different noise models and structure of the objective; monotone variational inequalities with cocoercivity. In each of these chapters, we identify the steps in the analysis that requires the knowledge of problem parameters and develop alternative techniques. Besides, we describe new methodologies in accordance with adaptive algorithm design paradigms in online learning [DHS11] and appropriately modify them for the particular problem formulation(s) at hand. While doing so, we try to establish modular and methodical proof techniques which are extendable to various problems in optimization and machine learning.

Chapter 2: Universal and robust optimization methods for convex minimization. In this first chapter, we focus on the relatively more fundamental setting of convex minimization as described in the following formulation,

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

where $f:\mathbb{R}^d\to\mathbb{R}$ is a proper, closed, convex function and $\mathcal{X}\subseteq\mathbb{R}^d$ is a compact and convex subset of $\mathbb{R}^d$. We will study this problem under two main scenarios; first-order smooth setting in which f has L-Lipschitz continuous gradient and second-order smooth setting in which f has L-Lipschitz continuous Hessian. We study this problem from two aspects; adaptation to levels of noise in first and second-order oracles, and universality across smoothness of the objective in the first-order smooth setting.

For the case of first-order methods, Levy, Yurtsever, and Cevher [LYC18] propose the first adaptive algorithm with accelerated rate of $O\left(\frac{GD+LD^2}{T^2}\right)$ where $D=\max_{x,y\in\mathcal{X}}\|x-y\|$ is the diameter of the constraint set and $G=\max_{x\in\operatorname{dom}\nabla f}\|\nabla f(x)\|$ is a known bound on the gradients. We identify three weak points of their results:

(i) Although they study unconstrained minimization setting, their algorithm requires to know a compact set to which the global minimizer of f belongs.

(ii) They need to know the upper bound for the gradients in order to set the step size.

(iii) Their algorithm does not achieve the optimal rate interpolation $O\left(\frac{LD^2}{T^2}+\frac{D\sigma}{\sqrt{T}}\right)$.

To remedy the aforementioned drawbacks, we study the same problem in the compactly-constrained setting and developed the first noise-adaptive and universal algorithm. Essentially, our method takes the ExtraGradient algorithm [Kor76] as a starting point and combines it with the adaptive step-size strategies in Duchi, Hazan, and Singer [DHS11] and Levy, Yurtsever, and Cevher [LYC18] and Rakhlin and Sridharan [RS13]. While doing so, we developed a new acceleration mechanism that is compatible with the adaptive step-size construction. This mechanism combines weighted averaging with step-size scaling.

The analysis of our algorithm is carefully designed to be modular; we separate the analysis into regret analysis and accelerated conversion scheme. The *offline* regret analysis aptly exploits the smoothness of the objective and that all the losses are computed with respect to the same objective function f. Then, we analyze a new accelerated conversion scheme for the offline regret (concurrrently with Cutkosky [Cut19]) which only uses the convexity of the objective function and quantifies the rate of convergence with respect to the weights in the averaging.

In the second part of this chapter, we further extend our results to second-order methods and achieve faster rates of order $O(1/T^3)$, going beyond the accelerated first-order rate of $O(1/T^2)$. This is mainly due to identifying a one-to-one mapping between the order of smoothness and the degree of weights in the averaging mechanism of the acceleration template. We generalize the conversion scheme to handle different averaging weights and scaling factor for the step-sizes. Our algorithm achieves the first noise-adaptive rate of $O\left(\frac{\sigma_g}{\sqrt{T}} + \frac{\sigma_H}{T^{1.5}} + \frac{LD^3}{T^3}\right)$ by independently adapting to noise in the gradient (σ_G) and the Hessian (σ_H) computations.

With regards to the proof techniques, a key paradigm is to characterize the error due to *not knowing* the Lipschitz constant. As we displayed in the beginning of this section, setting the step-size with respect to the global Lipschitz constant guarantees descent at every iteration. Our strategy for designing adaptive step-sizes are inspired by the AdaGrad algorithm:

$$\gamma_t = \frac{\alpha}{\sqrt{\beta + \sum_{s=1}^{t} \|\nabla f(x_s)\|^2}}.$$

We modify this template construction according to the problem setting, algorithmic framework and the set of assumptions. In simple terms, this monotonically-decreasing step-size gives us an approximation error for the *worst-case optimal* step-size as a function of the observed gradients. In our analysis we prove that the *total approximation error* across the whole of the execution is upper bounded by a constant that depends on α, β as well as L and D. By the foregoing regret-to-rate conversion schemes, the *amortized* error due to the approximation decreases at the optimal convergence rate for the respective problem setting.

Speaking of adaptive, parameter-free step-sizes, we have to briefly talk about *line-search* methods [Arm66; NW06]. As a very well-known and well-studied technique to adaptively set the step-size, it is basically an iterative sub-routine that finds a step-size to guarantee descent with respect to a sufficient decrease condition. The main drawback for such methods is that they require access to objective value oracle and they are usually not designed to handle stochastic feedback. Therefore, we especially deal with data-adaptive step-size in the sense of AdaGrad.

Chapter 3: Adaptive methods and variance reduction for smooth, non-convex optimization. In this chapter, we turn our attention to the more general stochastic, smooth, non-convex

minimization problem formulation,

$$\min_{x\in\mathbb{R}^d} f(x) := \mathbb{E}_{\xi\sim\mathscr{D}}\left[f(x,\xi)\right]$$

where f is a continuous, smooth and possibly non-convex function and the random vector ξ is drawn from the probability distribution $\mathscr{D}$. One can imagine the above formulation as a generalization of the empirical risk minimization problem. We will investigate three settings that falls under this generic formulation.

We begin with the original formulation as above and focus on the high-probability convergence of the AdaGrad. The literature on probabilistic behavior of adaptive methods is significantly sparser than the in-expectation convergence results. The existing results in this context either achieve sub-optimal dependence on probability margin [WWB19] or analyze a modified version of the algorithm [LO20] to avoid fundamental measurability problems. We propose an alternative analysis approach which enables us to prove high-probability convergence of AdaGrad with (almost) optimal convergence rate of $\tilde{O}(1/\sqrt{T})$ with respect to the expectation of *the squared gradient norm*, and best-known dependence of $\log(1/\delta)$ on probability margin. In the meantime, we prove that sub-optimality gap with respect to a global minimizer grows no faster than $O(\log(T))$, proving pseudo-boundedness of iterates/objective values with high probability.

In the second part of the chapter, we delve into two special cases of the above problem setting and describe adaptive *variance reduction* methods that take advantage of the underlying structures to achieve faster rates beyond $O(1/\sqrt{T})$. The first sub-setting is the streaming data regime in which we have access to an oracle that returns a stochastic gradient $\nabla f(x,\xi)$ when queried at the iterate x. The main assumption is that each stochastic feedback $\nabla f(x,\xi)$ is smooth itself: $\|\nabla f(x,\xi) - \nabla f(y,\xi)\| \leq L\|x-y\|$. We make use of the recursive momentum estimator of Cutkosky and Orabona [CO19] which requires only 2 oracle calls per iteration and identify a dynamic, adaptive relationship between the step-size and the momentum parameters. Our parameter-free variance reduction method achieves the optimal noise-adaptive rate of $O\left(\frac{1}{\sqrt{T}} + \frac{\sigma^{1/3}}{T^{1/3}}\right)$ with respect to expectation of *the gradient norm*.

We finally study the more specialized finite-sum minimization setting where $\mathscr{D}$ represents a sampling strategy over the fixed-sized dataset and ξ is a sample data point. Variance reduction has been studied extensively for the case of (strongly) convex minimization [JZ13; Ngu+17; DBL14] and more recently for non-convex problems [AH16b; All17b; Fan+18]. We propose the first parameter-free variance reduction algorithm for smooth, non-convex finite-sum problems. Our algorithm relies on a recursive estimator called SPIDER and combines it with an AdaGrad-type step-size to achieve optimal sample complexity up to logarithmic factors. We identify the correct balance between fast convergence rates and small gradient complexity and prove that the cumulative variance of the whole process grows no faster than $O(\log(T))$.

The proof of the high-probability analysis borrows some techniques from that of Chapter 2, however, we needed to develop new approaches for the variance reduction methods. The

main difference is that the existing work on non-convex variance reduction hinges on showing that the variance at each step monotonically decreases. Then, by validating that this decrease is fast enough, one could show faster rates under these more specialized settings. Note that the monotonic decrease of the variance depends on setting the step-size with respect to the Lipschitz constant [Wan+19] and sometimes a bound on the gradient norms [Cut19]. Our adaptive methods do not guarantee sufficient decrease of the variance at each iteration, but we could show that the cumulative variance could be kept under control. For the finite-sum setting, we show that cumulative variance grows as $O(\log(T))$ while for the streaming data setting, we prove that cumulative variance is comparable to the sum of gradient norms.

Chapter 4: Efficient and robust algorithms for min-max problems and games. In the final chapter, we focus on a different problem than the minimization; solving monotone variational inequalities.

$$\text{Find } x^* \in \mathbb{R}^d \text{ s.t. } \langle A(x^*), x - x^* \rangle \geq 0$$

The variational inequalities have recently been popularized due to an increasing interest in min-max optimization, reinforcement learning, games and GANs. Naturally, the study of adaptive methods are limited [ABM19; Lin+20; ABM21; HAM21; Hsi+22a] and there remians several open problems in different fronts.

In this section we investigate monotone, $(1/L)$-*cocoercive* variational inequality problem such that the operator satisfies

$$\langle A(x) - A(y), x - y \rangle \geq \frac{1}{L} \| A(x) - A(y) \|,$$

which also implies that A is L-Lipschitz continuous. We analyze a generalized extra-gradient algorithm (GEG) with a suitable data-adaptive step-size which recovers extra-gradient, dual averaging and dual extrapolation as its special case. Our goal is to study this template algorithm, without any modifications, under the standard bounded variance regime as well as the *relative noise* setting,

$$\mathbb{E}\left[\| A(x, \xi) - A(x) \|^2 \right] \leq c \| A(x) \|^2,$$

where the noise vanishes as the operator converges to its zero. It is known that under stochastic, monotone setting the best possible convergence rate is $O(1/\sqrt{T})$ [JNT11] and we identify a connection between cocoercivity and relative noise to improve this rate to $O(1/T)$. Our adaptive GEG template achieves $O(1/\sqrt{T})$ and $O(1/T)$ rates for bounded variance and relative noise models, respectively, without any prior knowledge of the setting. We finally show that the last iterate of the GEG template converges to a solution almost surely.

At the heart of the adaptive analysis lies a different technique than previous chapters; we recognize that the rate of convergence is roughly governed by the growth of the inverse step-

size and $\sum_{t=1}^{T} \|A(X_t)\|^2$. Showing that the foregoing quantities are bounded by $O(\sqrt{T})$ and $O(1)$ implies the $O(1/\sqrt{T})$ and $O(1/T)$ rates, respectively.

2 Universal and robust optimization methods for convex minimization

2.1 A Universal Algorithm with Optimal Guarantees for Constrained Minimization

2.1.1 Bibliographic Note

This section (Section 2.1) is based on the published work Kavis et al. [Kav+19], published in the NeurIPS 2019 conference.

Author list of the published work.

- Ali Kavis
- Kfir Y. Levy
- Francis Bach
- Volkan Cevher

Description of contributions. The idea of using optimistic mirror descent algorithm and the adaptive step-size construction by Rakhlin and Sridharan [RS13] comes from Kfir Y. Levy. The candidate designed the final version of Algorithm 2. All the theoretical results and numerical experiments are due to the candidate, while the candidate and Kfir Y. Levy worked jointly for the proof of Theorem 2.1.4.

2.1.2 Introduction

Stochastic constrained optimization with first-order oracles (SCO) is critical in machine learning. Indeed, the scalability of classical machine learning tasks, such as support vector machines (SVMs), linear/logistic regression and Lasso, rely on efficient *stochastic* optimization methods. Importantly, generalization guarantees for such tasks often rely on constraining the

set of possible solutions. The latter induces simple solutions in the form of low norm or low entropy, which in trun enables to establish generalization guarantees.

In the SCO setting, the optimal convergence rates for the cases of non-smooth and smooth objectives are given by $\mathcal{O}(G/\sqrt{T})$ and $\mathcal{O}(L\|X_1 - x^*\|^2/T^2 + \sigma/\sqrt{T})$, respectively [Lan12; Lan20]; where T is the total number of (noisy) gradient queries, L is the smoothness constant of the objective, σ^2 is the variance of the stochastic gradient estimates, and G is a bound on the magnitude of gradient estimates. These rates cannot be improved without additional assumptions.

The optimal rate for the non-smooth case may be obtained by the current optimization algorithms, such as stochastic gradient descent (SGD), (stochastic) mirror descent [NYD83], AdaGrad [DHS11], Adam [KB15], and AmsGrad [RKK18b]. However, in order to obtain the optimal rate for the smooth case, one is required to use more involved *accelerated* methods such as [HPK09; Lan12; Xia10; DO18; CDO18; DCL18]. Unfortunately, all of these accelerated methods require a-priori knowledge of the smoothness parameter L, and in some cases the variance of the gradients σ^2, creating a setup barrier for their use in practice.

This work develops a new *universal* method for *constrained* SCO that obtains the optimal rates in both smooth and non-smooth cases, *without any prior knowledge regarding the smoothness of the problem L, nor the noise magnitude* σ. Such universal methods that implicitly adapt to the properties of the learning objective may be very beneficial in practical large-scale problems where these properties are usually unknown. To our knowledge, this is the first work that achieves this desiderata in the (compactly-)constrained SCO setting.

Our contributions in the context of related work. For the unconstrained setting, Levy, Yurtsever, and Cevher [LYC18] and Cutkosky [Cut19] have recently presented a universal scheme that obtains (almost) optimal rates for both smooth and non-smooth cases.

More specifically, Levy, Yurtsever, and Cevher [LYC18] designs AcceleGrad—a method that obtains respective rates of $\mathcal{O}\left(GD\sqrt{\log T}/\sqrt{T}\right)$ and $\mathcal{O}\left(L\log LD^2/T + \sigma D\sqrt{\log T}/\sqrt{T}\right)$. Note that D denotes the diameter of the constraint set in the rest of this section, however, for the guarantees of AcceleGrad, D denotes an auxiliary, compact set, which is known to contain the global minimizer of the objective.Unfortunately, this result only holds for the unconstrained setting, and the authors leave the truly *constrained* case as an open problem. An important progress towards this open problem is achieved only recently by Cutkosky [Cut19], who proves sub-optimal respective rates of $\mathcal{O}\left(1/\sqrt{T}\right)$ and $\mathcal{O}\left(D^2L/T^{3/2} + \sigma D/\sqrt{T}\right)$ for SCO in the constrained setting.

Our work completely resolves the open problem in Levy, Yurtsever, and Cevher [LYC18] and Cutkosky [Cut19], and proposes the first *universal* method that obtains respective *optimal* rates of $\mathcal{O}\left(GD/\sqrt{T}\right)$ and $\mathcal{O}\left(D^2L/T^2 + \sigma D/\sqrt{T}\right)$ for the constrained setting. When applied to the unconstrained setting, our analysis tightens the rate characterizations by removing the

unnecessary logarithmic factors appearing in [LYC18; Cut19].

Our method is inspired by the Mirror-Prox (MP) method [Nem04; RS13; DO18; BL19], and builds on it using additional techniques from the online learning literature. Among, is an adaptive learning rate rule [DHS11; RS13], that is married with a suitable acceleration mechanism for the MP template. We also adopt an online-to-batch conversion techniques, which was concurrently discovered by [Cut19].

The current part of the chapter is organized as follows. We specify the problem setup, and give the necessary definitions and background information. Then, we motivate our framework and explain the general mechanism. We also introduce the convergence theorems with proof sketches to highlight the technical novelties. We share numerical results in comparison with other adaptive methods and baselines for different machine learning tasks in Section 2.1.6, followed up with conclusions and future proposals.

2.1.3 Setting and preliminaries

Preliminaries. Let $\|\cdot\|$ be a general norm and $\|\cdot\|_*$ be its dual. A function $f : \mathcal{X} \mapsto \mathbb{R}$ is *μ-strongly convex* over a convex set $\mathcal{X}$, if for any $x \in \mathcal{X}$ and a,ny subgradient of f at x $\nabla f(x)$,

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq \frac{\mu}{2}\|x - y\|^2, \quad \forall x, y \in \mathcal{X} \tag{2.1}$$

A function $f : \mathcal{X} \mapsto \mathbb{R}$ is *L-smooth* over $\mathcal{X}$ if it has L-Lipschitz continuous gradient, i.e.,

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|, \quad \forall x, y \in \mathcal{X}. \tag{2.2}$$

Consider a 1-strongly convex differentiable function $h : \mathcal{X} \to \mathbb{R}$. The Bregman divergence with respect to a distance-generating function h is defined as follows $\forall x, y \in \mathcal{X}$,

$$D_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle . \tag{2.3}$$

An important property of Bregman divergence is that $D_h(x, y) \geq \frac{1}{2}\|x - y\|^2$ for all $x, y \in \mathcal{X}$, due to the 1-strong convexity of h. In the sequel we define the diameter of the constraint set $\mathcal{X}$ via the Bregman divergence; $D := \max_{x,y\in\mathcal{X}} \sqrt{D_h(x, y)}$.

Setting. This paper focuses on (approximately) solving the following constrained problem,

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

where $f : \mathcal{X} \mapsto \mathbb{R}$ is a proper, closed, convex function, and $\mathcal{X} \subset \mathbb{R}^d$ is a compact, convex set.

We assume the availability of a first order oracle for $f(\cdot)$, and consider two settings: a deterministic setting where we may access exact gradients, and a stochastic setting where we could

only access unbiased (noisy) gradient estimates. Concretely, we assume that by querying this oracle with a point $x \in \mathscr{X}$, we receive the gradient estimate $\nabla f(x,\xi) \in \mathbb{R}^d$, such that

$$\mathbb{E}\left[\nabla f(x,\xi)\middle|\sigma(x)\right] = \nabla f(x)\,, \tag{2.4}$$

where ξ is the random vector that denotes the noise in the process, and $\sigma(x)$ is the sigma-algebra generated by the (possibly) random vector x. We assume that the random vectors ξ are drawn independently for each call and they are also independent of the iterate sequence generated by the algorithms. We define that a gradient estimate has bounded variance if

$$\mathbb{E}[\|\nabla f(x,\xi) - \nabla f(x)\|_*^2 | \sigma(x)] \le \sigma^2, \quad \forall x \in \mathscr{X}. \tag{2.5}$$

Only in the case of the objective being non-smooth, we define it to be G-Lipschitz continuous. Under the same setting, we assume the sub-gradient estimates have bounded norm,

$$\|\nabla f(x,\xi)\|_* \le G, \qquad \forall x \in \mathscr{X}.$$

We abuse the notation ∇ to denote both gradient and the sub-gradient. We make it clear in the text what ∇ refers to depending on the context.

2.1.4 Method

In this section, we present and analyze our **Uni**versal e**X**tra **Grad**ient (UniXGrad) method. We first discuss the Mirror-Prox (MP) algorithm of [Nem04], and the Optimistic Mirror Descent (OMD) algorithm of [RS13]. Later we present our algorithm which builds on top of both of the aforementioned algorithms with appropriate modifications. Then, we present and analyze the guarantees of our method in non-smooth and smooth settings, respectively.

Our goal is to optimize a convex function f over a compact domain $\mathscr{X}$, and Algorithm 1 offers a well-known framework for solving (Prob). Let us motivate this particular template. Basically, the algorithm takes a step from X_t to $X_{t+\frac{1}{2}}$, using first order information based on X_t. Then, it goes back to X_t and takes another step, but this time, gradient information relies on $X_{t+\frac{1}{2}}$. Each step is a generalized projection with respect to Bregman divergence $D_h(\cdot,\cdot)$.

Algorithm 1: Mirror-Prox Template

Input: Number of iterations T , $X_0 \in \mathscr{X}$, learning rate $\{\gamma_t\}_{t\in[T]}$
for $t = 0$ **to** T **do**
  $X_{t+\frac{1}{2}} = \arg\min_{x\in\mathscr{X}} \langle M_t, x\rangle + \frac{1}{\gamma_t} D_h(x, X_t)$
  $X_{t+1} = \arg\min_{x\in\mathscr{X}} \langle g_t, x\rangle + \frac{1}{\gamma_t} D_h(x, X_t)$
end for

Now, let us explain the salient differences between UniXGrad and MP as well as OMD using the particular choices of M_t, g_t and the distance-generating function $\mathscr{X}$. Optimistic Mirror

Descent takes $g_t = \nabla f(X_{t+\frac{1}{2}})$ and computes $M_t = \nabla f(X_{t-\frac{1}{2}})$, i.e., based on gradient information from previous iterates. This vector is available at the beginning of each iteration and the "optimism" arises in the case where $M_t \approx g_t$. When $M_t = \nabla f(X_t)$ and $g_t = \nabla f(X_{t+\frac{1}{2}})$, the template is known as the famous Mirror-Prox algorithm. One special case of Mirror-Prox is Extra-Gradient scheme [Kor76] where the projections are with respect to Euclidean norm, i.e. $h(x) = 1/2\|x\|_2^2$, instead of general Bregman divergences.

MP has been well-studied, especially in the context of variational inequalities and convex-concave saddle point problems. It achieves fast convergence rate of $\mathcal{O}(1/T)$ for this class of problems, however, in the context of smooth convex optimization, this is the standard slow rate [Nes03]. To date, MP is not known to enjoy the accelerated rate of $\mathcal{O}(1/T^2)$ for smooth convex minimization. We propose three modifications to Algorithm 1, which are the precise choice of g_t and M_t, the adaptive learning rate and the gradient weighting scheme.

The notion of averaging. In different interpretations of acceleration [Nes83a; Nes88; Tse08; AO16], the notion of averaging is always central and we incorporate this notion via gradients taken at weighted average of iterates. Let us define the weight $\alpha_t = t$, $A_t = \sum_{s=1}^{t} \alpha_s$ and the following sequence

$$\bar{X}_{t+\frac{1}{2}} = \frac{\alpha_t X_{t+\frac{1}{2}} + \sum_{s=1}^{t-1} \alpha_s X_{s+\frac{1}{2}}}{A_t}, \qquad \tilde{X}_t = \frac{\alpha_t X_t + \sum_{s=1}^{t-1} \alpha_s X_{s+\frac{1}{2}}}{A_t}. \tag{2.6}$$

Then, UniXGrad algorithm takes $g_t = \nabla f(\bar{X}_{t+\frac{1}{2}})$ and $M_t = \nabla f(\tilde{X}_t)$, which provides a naive interpretation of averaging. Our choice of g_t and M_t coincide with that of the accelerated Extra-Gradient scheme of Diakonikolas and Orecchia [DO18]. While their decision relies on implicit Euler discretization of an accelerated dynamics, we arrive at the same conclusion as a direct consequence of our convergence analysis.

Adaptive learning rate. A key ingredient of our algorithm is the choice of adaptive learning rate γ_t. In light of [RS13], we define our lag-one-behind learning rate as

$$\gamma_t = \frac{2D}{\sqrt{1 + \sum_{s=1}^{t-1} \alpha_s^2 \|g_t - M_t\|_*^2}}, \tag{2.7}$$

where $D^2 = \sup_{x,y \in \mathcal{X}} D_h(x,y)$ is the diameter of the compact set $\mathcal{X}$ with respect to Bregman divergences.

Gradient weighting scheme. We introduce the weights α_t in the sequence updates. One can interpret this as separating step size into learning rate and the scaling factors. It is necessary that $\alpha_t = \Theta(t)$ in order to achieve optimal rates, in fact we precisely choose $\alpha_t = t$. Also notice that they appear in the learning rate, compatible with the update rule. Algorithm 2 summarizes

our framework.

Algorithm 2: UniXGrad

Input: # of iterations T , $X_0 \in K$, diameter D, weight $\alpha_t = t$, learning rate $\{\gamma_t\}_{t\in[T]}$
for $t = 1$ **to** T **do**
$X_{t+\frac{1}{2}} = \arg\min_{x\in\mathcal{X}} \alpha_t\langle M_t, x\rangle + \frac{1}{\gamma_t} D_h(x, X_t)$ $\quad (M_t = \nabla f(\tilde{X}_t) \text{ or } \nabla f(\tilde{X}_t, \xi_t))$
$X_{t+1} = \arg\min_{x\in\mathcal{X}} \alpha_t\langle g_t, x\rangle + \frac{1}{\gamma_t} D_h(x, X_t)$ $\quad (g_t = \nabla f(\bar{X}_{t+\frac{1}{2}}) \text{ or } \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}))$
end for
return $\bar{X}_{T+\frac{1}{2}}$

In the next section, we will present our convergence theorems and provide proof sketches to emphasize the fundamental aspects and novelties. With the purpose of simplifying the analysis, we borrow classical tools in the online learning literature and perform the convergence analysis in the sense of bounding "weighted regret". Then, we use a simple yet essential conversion strategy which enables us to *directly* translate our weighted regret bounds to convergence rates.

2.1.5 Analysis

First off, we will present the conversion scheme from offline, weighted regret to convergence rate, by deferring the proof to the appendix of this chapter. In a concurrent work, [Cut19] proves a similar online-to-offline conversion bound.

Lemma 2.1.1. *Consider weighted average $\bar{X}_{t+\frac{1}{2}}$ as in Eq. (2.6). Let $\text{REG}_T(x^*) = \sum_{t=1}^{T} \alpha_t\langle g_t, X_{t+\frac{1}{2}} - x^*\rangle$ denote the weighted regret after T iterations, $\alpha_t = t$ and $g_t = \nabla f(\bar{X}_{t+\frac{1}{2}})$. Let $x^* = \arg\min_{x\in\mathcal{X}} f(x)$. Then,*

$$f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \le \frac{2\text{REG}_T(x^*)}{T^2}.$$

For the stochastic setting with $g_t = \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}})$, the same bound holds in expectation.

$$\mathbb{E}\left[f(\bar{X}_{T+\frac{1}{2}}) - f(x^*)\right] \le \frac{2\mathbb{E}\left[\text{REG}_T(x^*)\right]}{T^2}.$$

This lemma provides a principled way to convert offline regret to accelerated rates; when the offline, weighted regret is constant, then the respective algorithm achieves the celebrated $O(1/T^2)$ rate. Note that we strictly consider the offline case in which the first-order information is generated with respect to a single, fixed objective function f and we could exploit the smoothness of the objective to obtain a constant regret.

Non-smooth setting

First, we will focus on the convergence analysis in the case of non-smooth objective functions with deterministic/stochastic first-order oracles. We will follow the regret analysis as in [RS13] with essential adjustments that suit our weighted scheme and particular choice of adaptive step-size.

Theorem 2.1.1. *Consider the constrained optimization setting in Problem (Prob), where $f : \mathscr{X} \to \mathbb{R}$ is a proper, convex and G-Lipschitz function defined over compact, convex set $\mathscr{X}$. Let $x^* \in \arg\min_{x\in\mathscr{X}} f(x)$, and define $g_t = \nabla f(\bar{X}_{t+\frac{1}{2}})$, $M_t = \nabla f(\tilde{X}_t)$ where $\nabla f(\cdot)$ represents a sub-gradient of the objective f at the query point. Then, Algorithm 2 guarantees*

$$f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \leq \frac{7D\sqrt{1+\sum_{t=1}^{T}\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2}}{T^2} \leq O\left(\frac{D}{T^2} + \frac{GD}{\sqrt{T}}\right).$$

We establish the basis of our analysis through Lemma 1 and Corollary 2 of [RS13]. Then, we build upon this base by exploiting the structure of the adaptive step-size, the weights α_t and the bound on gradient norms to give adaptive convergence bounds.

Due to the acceleration machinery through the weighted averaging and scaled step-sizes, the offline cumulative regret of UniXGrad in the non-smooth case grows as $O(T^{3/2})$, which is counter intuitive compared to standard results in online convex optimization. However, the offline structure of the problem, together with the acceleration mechanism itself enables a faster conversion scheme, resulting in the standard $O(1/\sqrt{T})$ convergence rate in the average iterate.

Remark 2.1.1. It is important to point out that we do not completely exploit the precise definitions of g_t and M_t in the presence of non-smooth objectives. As far as the regret analysis is concerned, it suffices that these quantities are functions of $\nabla f(\cdot)$ and that, as a corollary, their dual norm is upper bounded. However, in order to bridge the gap between weighted regret and the objective sub-optimality, i.e. $f(\bar{X}_{T+\frac{1}{2}}) - f(x^*)$, we require $g_t = \nabla f(\bar{X}_{t+\frac{1}{2}})$ and $M_t = \nabla f(\tilde{X}_t)$.

Now, we further consider the case of stochastic gradients. We assume that the first-order oracles are unbiased (see Eq. (**??**)). We want to emphasize that our stochastic setting is *not* restricted to the notion of additive noise, i.e. gradients corrupted with zero-mean noise. It essentially includes any estimate that recovers the full gradient in expectation, e.g. estimating gradient using mini batches. Additionally, we propagate the bounded gradient norm assumption to the stochastic oracles, such that $\|\nabla f(x,\xi)\|_* \leq G$, $\forall x \in \mathscr{X}$.

Theorem 2.1.2. *Consider the optimization setting in Problem (Prob), where f is non-smooth, convex and G-Lipschitz. Let $\{X_{t+\frac{1}{2}}\}_{t=1,..,T}$ be a sequence generated by Algorithm 2 such that $g_t = \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}})$, $M_t = \nabla f(\tilde{X}_t, \xi_t)$ and Assumptions in Eq. (2.4) and (2.5) hold. With $\alpha_t = t$*

and step-size as in Eq. (2.7), it holds that

$$\mathbb{E}\left[f(\bar{X}_{T+\frac{1}{2}})\right] - f(x^*) \leq O\left(\frac{D}{T^2} + \frac{GD}{\sqrt{T}}\right).$$

The analysis in the stochastic setting is similar to deterministic setting. The difference is up to replacing g_t and M_t with their stochastic counterparts. With the additional bound on stochastic sub-gradients, the same rate is achieved.

Smooth setting

In terms of theoretical contributions and novelty, the case of L-smooth objective is of greater interest. In the non-smooth analysis, we ignore a particular negative summation term. When coupled with smoothness of the objective and a particular characterization of the growth of step-size γ_t, we will obtain a constant component of the regret. Hence, we achieve the accelerated rates both in deterministic and stochastic settings. We will first start with the deterministic oracle scheme and then introduce the convergence theorem for the noisy setting.

Theorem 2.1.3. *Consider the constrained optimization setting in Problem (Prob), where $f : \mathcal{X} \to \mathbb{R}$ is a proper, convex and L-smooth function defined over compact, convex set $\mathcal{X}$. Let $x^* \in \min_{x\in\mathcal{X}} f(x)$. Then, Algorithm 2 run with $g_t = \nabla f(\bar{X}_{t+\frac{1}{2}})$ and $M_t = \nabla f(\tilde{X}_t)$ ensures the following*

$$f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \leq O\left(\frac{LD^2}{T^2}\right). \tag{2.8}$$

Remark 2.1.2. In the non-smooth setting, we assume that gradients have bounded norms. Our algorithm does **not** need to know this information, but it is used in the analysis in that case. However, when the function is smooth, neither the algorithm nor the analysis requires bounded gradients.

Proof Sketch (Theorem 2.1.3). We follow the proof of Theorem 2.1.1 until the point where we obtain

$$\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle \leq \frac{1}{2}\sum_{t=1}^{T} \gamma_{t+1}\alpha_t^2 \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2 - \frac{1}{2}\sum_{t=1}^{T} \frac{1}{\gamma_{t+1}} \|X_{t+\frac{1}{2}} - X_t\|^2 + D^2\left(\frac{3}{\gamma_{T+1}} + \frac{1}{\gamma_1}\right).$$

By smoothness of the objective function, we have $\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_* \leq L\|\bar{X}_{t+\frac{1}{2}} - \tilde{X}_t\|$,

which implies $-\frac{1}{\gamma_{t+1}}\|X_{t+\frac{1}{2}} - X_t\|^2 \le -\frac{\alpha_t^2}{4L^2\gamma_{t+1}}\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2$. Hence,

$$\le \frac{1}{2}\sum_{t=1}^{T}\left(\gamma_{t+1} - \frac{1}{4L^2\gamma_{t+1}}\right)\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2 + D^2\left(\frac{3}{\gamma_{T+1}} + \frac{1}{\gamma_1}\right).$$

Now we will introduce a time variable to *characterize* the growth of the step-size. Define $\tau^* = \max\left\{t \in \{1,\dots,T\} : \frac{1}{\gamma_{t+1}^2} \le 7L^2\right\}$ such that $\forall t > \tau^*$, $\gamma_{t+1} - \frac{1}{4L^2\gamma_{t+1}} \le -\frac{3}{4}\gamma_{t+1}$. Then,

$$\underbrace{\le D\sum_{t=1}^{\tau^*}\frac{\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2}{\sqrt{1+\sum_{s=1}^{t}\alpha_s^2\|\nabla f(\bar{X}_{s+\frac{1}{2}}) - \nabla f(\tilde{X}_s)\|_*^2}} + \frac{D}{2}}_{\text{(A)}}$$

$$+\underbrace{\frac{3D}{2}\left(\sqrt{1+\sum_{t=1}^{T}\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2} - \sum_{t=\tau^*+1}^{T}\frac{\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2}{\sqrt{1+\sum_{s=1}^{t}\alpha_s^2\|\nabla f(\bar{X}_{s+\frac{1}{2}}) - \nabla f(\tilde{X}_s)\|_*^2}}\right)}_{\text{(B)}},$$

where we wrote γ_{t+1} in open form and used the definition of τ^*. To complete the proof, we will need the following lemma.

Lemma 2.1.2. *Let $\{a_i\}_{i=1,\dots,n}$ be a sequence of non negative numbers. Then, it holds that*

$$\sqrt{\sum_{i=1}^{n} a_i} \le \sum_{i=1}^{n}\frac{a_i}{\sqrt{\sum_{j=1}^{i} a_j}} \le 2\sqrt{\sum_{i=1}^{n} a_i}.$$

Please refer to [MS10; LYC18] for the proof. We jointly use Lemma 2.1.2 and the bound on $\gamma_{\tau*+1}$ to upper bound terms (A) and (B) with $4\sqrt{7}D^2L$ and $6\sqrt{7}D^2L$, respectively. Lemma 2.1.1 immediately establishes the convergence bound. ■

Next, we will present our results for the stochastic extension. The analysis proceeds along similar lines as its deterministic counterpart. However, we execute the analysis using auxiliary terms and attain the optimal accelerated rate without the log factors.

Theorem 2.1.4. *Consider the optimization setting in Problem (Prob), where f is L-smooth and convex. Let $\{X_{t+\frac{1}{2}}\}_{t=1,\dots,T}$ be a sequence generated by Algorithm 2 such that $g_t = \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}})$, $M_t = \nabla f(\tilde{X}_t, \xi_t)$ and the stochastic estimates satisfy the Assumptions in Eq. (2.4) and (2.5). With $\alpha_t = t$, $x^* \in \arg\min_{x\in\mathcal{X}} f(x)$ and step-size as in (2.7), it holds that*

$$\mathbb{E}\left[f(\bar{X}_{T+\frac{1}{2}})\right] - f(x^*) \le O\left(\frac{LD^2}{T^2} + \frac{\sigma D}{\sqrt{T}}\right).$$

Proof Sketch (Theorem 2.1.4). We start in the same spirit as the stochastic, non-smooth set-

ting,

$$\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle = \underbrace{\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle}_{\text{(A)}} + \underbrace{\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle}_{\text{(B)}}.$$

Recall that term (B) is zero in expectation conditioned on $\sigma(X_{t+\frac{1}{2}}) = \sigma(\xi_1, \xi_{1+1/2}, ..., \xi_{t-\frac{1}{2}}, \xi_t)$. Then, we follow the proof steps of Theorem 2.1.1,

$$\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle \leq \frac{7D}{2} \sqrt{1 + \sum_{t=1}^{T} \alpha_t^2 \|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2} - \frac{1}{2} \sum_{t=1}^{T} \frac{1}{\gamma_{t+1}} \|X_{t+\frac{1}{2}} - X_t\|^2. \tag{2.9}$$

Observe that using the smoothness of the objective and the compactness of the constraint set, the first summation on the RHS grows roughly as $O(\sqrt{T})$, which translates to a convergence rate of $O(1/T^{1.5})$. Although this is faster than the gradient method it is still sub-optimal. The negative summation on the RHS is the key to control this term. First, we will obtain $\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2$ from $\|X_{t+\frac{1}{2}} - X_t\|^2$ due to smoothness and the challenge is to relate $\|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2$ and $\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2$ with each other using the variance bound. So let's denote, $B_t^2 := \min\{\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2, \|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2\}$. Using this definition, we could declare an auxiliary step-size which we will *only* use for the analysis,

$$\eta_t = \frac{2D}{\sqrt{1 + \sum_{s=1}^{t-1} \alpha_s^2 B_s^2}}. \tag{2.10}$$

Clearly, for any $t \in [T]$ we have $\gamma_t \leq \eta_t$, which indeed implies $-\frac{1}{\gamma_{t+1}} \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2 \leq -\frac{1}{\eta_{t+1}} B_t^2$. Let us define the shorthand notation for the noise in gradient evaluation,

$$\epsilon_{t+\frac{1}{2}} = \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\bar{X}_{t+\frac{1}{2}}) \quad \text{and} \quad \epsilon_t = \nabla f(\tilde{X}_t, \xi_t) - \nabla f(\tilde{X}_t).$$

Then, we have the following ineqaulities,

$$\|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2 \leq 2\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2 + 2\|\epsilon_{t+\frac{1}{2}} - \epsilon_t\|_*^2, \tag{2.11}$$

and,

$$\|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2 \leq 2B_t^2 + 2\|\epsilon_{t+\frac{1}{2}} - \epsilon_t\|_*^2. \tag{2.12}$$

Therefore, we could rewrite Eq. (2.9) as,

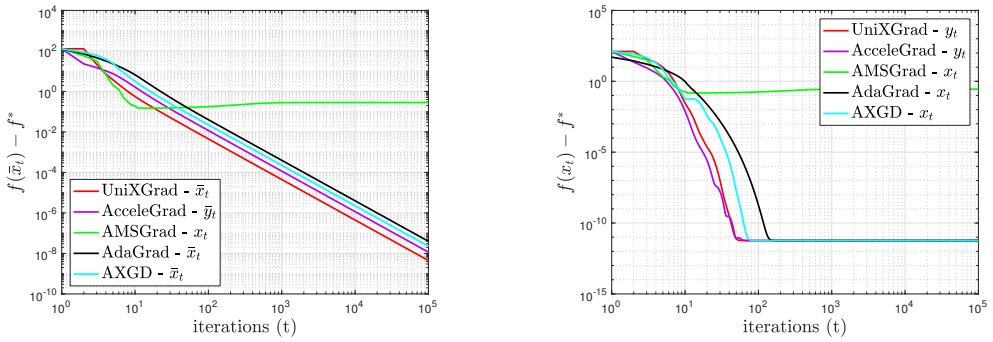
$$\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle \leq \underbrace{\frac{7}{2} \sum_{t=1}^{T} \left(\eta_{t+1} - \frac{1}{28L^2 \eta_{t+1}} \right) \alpha_t^2 B_t^2 + \frac{7D}{2}}_{\text{(A)}} + \underbrace{\frac{7D}{\sqrt{2}} \sqrt{\sum_{t=1}^{T} \alpha_t^2 \|\epsilon_{t+\frac{1}{2}} - \epsilon_t\|_*^2}}_{\text{(B)}}.$$

Using Lemma 2.1.2 and defining a time variable τ_* in the sense of Theorem 2.1.3 (with correct constants), term (A) is upper bounded by $112\sqrt{14}D^2L$. By taking expectation conditioned on $\bar{x}_t$ and using Jensen's inequality, we could upper bound term (B) as $14\sigma D T^{3/2}/\sqrt{2}$, which leads us to the optimal rate of $224\sqrt{14}D^2L/T^2 + 14\sqrt{2}\sigma D/\sqrt{T}$ through Lemma 2.1.1. ■

2.1.6 Experiments

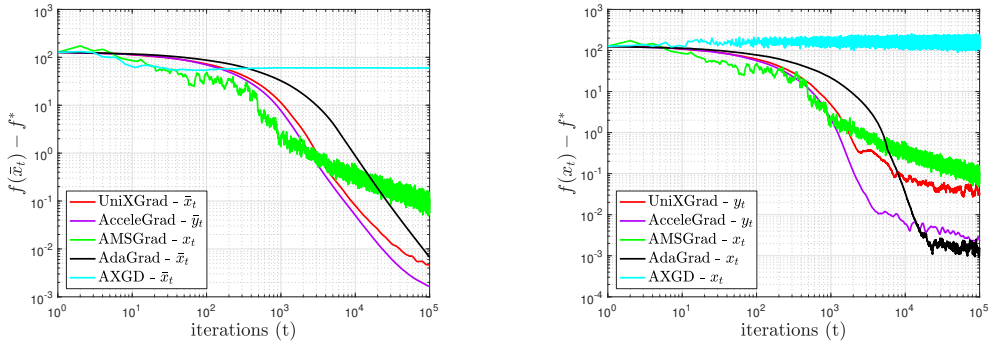
We compare performance of our algorithm for two different tasks against adaptive methods of various characteristics, such as AdaGrad, AMSGrad and AcceleGrad, along with a recent non-adaptive method AXGD. We consider a synthetic setting where we analyze the convergence behavior, as well as a SVM classification task on some LIBSVM dataset. In all the setups, we tuned the hyper-parameters of each algorithm by grid search. In order to compare the adaptive methods on equal grounds, AdaGrad is implemented with a scalar step size based on the template given by Levy [Lev17]. We implement AMSGrad exactly as it is described by Reddi, Kale, and Kumar [RKK18b].

Convergence behavior. We take the least squares problem with L_2-norm ball constraint for this setting, i.e., $\min_{\|x\|_2 < r} \frac{1}{2n}\|Ax - b\|_2^2$, where $A \in \mathbb{R}^{n \times d}$, $A \sim \mathcal{N}(0, \sigma^2 I)$ and $b = Ax^\natural + \epsilon$ such that ϵ is a random vector $\sim \mathcal{N}(0, 10^{-3})$ and $x^\natural$ is the original signal before perturbation. We pick $n = 500$ and $d = 100$. For the rest of this section, we refer to the solution of *constrained* problem as x_*.



(a) Average Iterate **(b)** Last Iterate

Figure 2.1: Convergence rates in the **deterministic** oracle setting when $x^* \in$ Boundary($\mathcal{X}$)

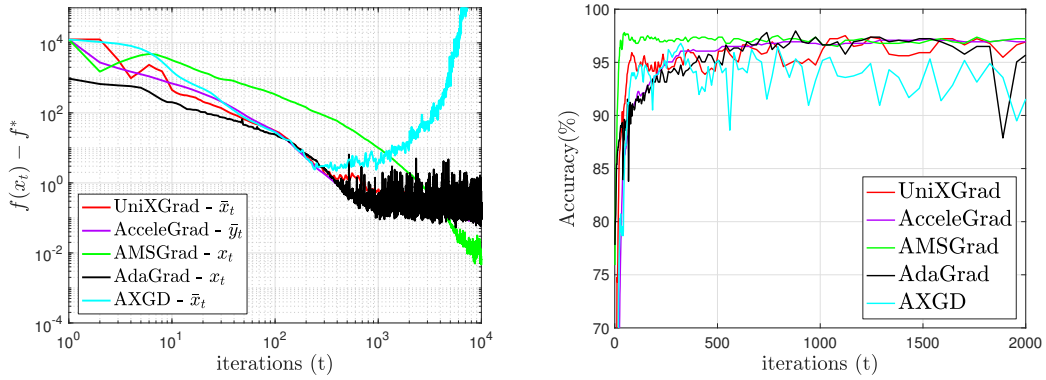


(a) Average Iterate **(b)** Last Iterate

Figure 2.2: Convergence rates in the **stochastic** oracle setting when $x^* \in \text{Boundary}(\mathcal{X})$

In Figure 2.1 and 2.2, we present the convergence rates under deterministic and stochastic oracles, and we pick a problem in which the solution is on the boundary of the constraint set, i.e., $x^* \in \text{Boundary}(\mathcal{X})$. In this setting, our algorithm shows matching performance in comparison with other methods. AXGD has convergence issues in the stochastic setting, as it only handles additive noise and their step size routine does not seem to be compatible with stochastic gradients. Another key observation is that AMSGrad suffers a decrease in its performance when the solution is on the boundary of the set.

SVM classification. In this section, we will tackle SVM classification problem on "breast-cancer" data set taken from LIBSVM. We try to minimize squared Hinge loss with L_2 norm regularization. We split the data set as training and test sets with 80/20 ratio. The models are trained using random mini batches of size 5. Figure 2.3 demonstrates convergence rates and test accuracies of the methods. They represent the average performance of 5 runs, with random initializations. For UniXGrad, AcceleGrad and AXGD, we consider the performance with respect to the average iterate $X_{t+\frac{1}{2}}$ as it shows a more stable behavior, whereas AdaGrad and AMSGrad are evaluated based on their last iterates. AXGD, which has poor convergence behavior in stochastic setting due to its step size rule, shows the worst performance both in terms of convergence and generalization. UniXGrad, AcceleGrad, AdaGrad and AMSGrad achieve comparable generalization performances to each other. AMSGrad achieves a slightly better performance as it has diagonal preconditioner which translates to per-coordinate learning rate. It could possibly adapt to the geometry of the optimization landscape better.



(a) Convergence rates with respect to training data

(b) Test Accuracy

Figure 2.3: Convergence behavior with respect to training data and resulting test accuracies for binary classification task on breast-cancer dataset from LIBSVM [CL11]

2.1.7 Conclusion

In this paper we presented an adaptive and universal framework that achieves the optimal convergence rates in constrained convex optimization setting. To our knowledge, this is the first method that achieves $\mathcal{O}\left(GD/\sqrt{T}\right)$ and $\mathcal{O}\left(LD^2/T^2 + D\sigma/\sqrt{T}\right)$ rates in the constrained setting, without log dependencies. Without any prior information, our algorithm adapts to smoothness of the objective function as well as the variance of the possibly noisy gradients.

One would interpret that our guarantees are extensions of [LYC18] to the constrained setting through a completely different algorithm and a simpler analysis. Our study of their algorithm and proof strategies concludes that:

- It does not seem possible to remove $\log T$ dependency in non-smooth setting for their algorithm, due to their Lemma A.3
- Extending their algorithm to constrained setting (via projecting y sequence) is not trivial, as the analysis requires y sequence to be unbounded (refer to their Appendix A, Eq. (16)).

As a follow up, we would like to investigate three main extensions:

- Proximal version of our algorithm that could handle composite problems in a unified manner. It seems like a rather simple extension as the main difference would be replacing optimality condition for constrained updates with that of proximal operator.
- Extending scalar adaptive learning rate to per-coordinate matrix-like preconditioner. This direction of research would help us create a robust algorithm that is applicable to non-convex problems, such as training deep neural networks.
- Adaptation to strong convexity along with smoothness and noise variance, simultaneously. A first step towards tackling this open problem is proving an improved rate of $O(1/T^2 + \sigma/T)$ for smooth and strongly convex problems, with stochastic gradients.

2.2 A First Approach to Noise-Adaptive Accelerated Second-Order Methods

2.2.1 Bibliographic Note

This section (Section 2.2) is based on the published work Antonakopoulos, Kavis, and Cevher [AKC22], published in the NeurIPS 2022 conference.

Author list of the published work.

- Kimon Antonakopoulos
- Ali Kavis
- Volkan Cevher

Description of contributions. The candidate and Kimon Antonakopoulos equally and jointly contributed to all the theoretical results in this work. The candidate implemented all the numerical experiments.

2.2.2 Introduction

Over the last few decades, first-order (convex) minimization methods have gained popularity for modern machine learning and optimization problems due to their efficient per-iteration cost and *global convergence* properties. The literature on first-order methods is rather dense and extensive with a concrete, thorough understanding of the optimal *global* convergence behavior. Focusing on the settings of smooth, convex minimization, the lower bounds have been well-established; $O(\sigma/\sqrt{T})$ when the gradient feedback is noisy with variance σ^2, and $O(1/T^2)$ under deterministic first-order oracles [NYD83; Nes03]. Under slight variations of the aforementioned problem setting, there exists an extensive amount of work that enjoys the latter, "accelerated" rate [Nes83a; Nes88; Nes05; Tse08; Xia10; Lan12; AO16; LYC18; WA18; DO18; Cut19; Kav+19; Jou+20; Ant+22; Liu+23].

On the contrary to its first-order analogue, the literature on *global convergence* of *second-order*, smooth methods is notably sparse with many open questions standing even in the simplest problem formulations. Following the pioneering works of Bennett [Ben16] and Kantorovich [Kan48], Newton's method and its variations [Lev44; Mar63] are considered as the staple of second-order methods in optimization. Although its powerful local convergence behavior has been repeatedly demonstrated [CGT00; KMR19], studies on its global behavior are relatively limited. Prior attempts at tackling global convergence mostly make additional structural assumptions on the objective function [Pol06; MBR19; KMR19] or assume extra regularity conditions on the Hessian [KSJ18] beyond the simplest smooth and convex setting. Over the last decade, we have witnessed important progress towards a more complete theory of

globally-convergent second-order methods, and yet there remains many important questions unanswered, which we will delve into in this part of the chapter.

To motivate the perspective in our technical endeavour, we take a small detour to introduce the idea of *universality*, which we particularly characterize as *adaptation to the level of noise in oracle feedback*. Our designation of this concept is relevant to that of Nesterov [Nes15] but is fundamentally different in its essence; while Nesterov [Nes15] defines it as continuous adaptation to levels of *Hölder smoothness* [NYD83], we particularly characterize it as *continous adaptation to the variance*.

Enabled by the recent advances in online optimization, universal first-order algorithms essentially attain the $O(\sigma/\sqrt{T}+1/T^2)$ convergence for convex minimization problems, interpolating between stochastic and deterministic rates. There exist a plethora of algorithms that enjoy this rate under different sets of assumptions for both minimization scenarios (for convex and non-convex settings, we refer the reader to [Lan12; Kav+19; ENV21; Jou+20; Ant+22] and [WWB19; LKC21; KLC22; Liu+22b], respectively), and the more general framework of variational inequalities [BL19; Ant+21; VAM21; HAM21; Hsi+22b; ABM21]. However, we observe that such universal results do not exist in second-order literature, hence, it is only natural to ask,

Can we design a simple second-order method that will achieve accelerated universal rates beyond $O(\sigma/\sqrt{T}+1/T^2)$?

More recently, global sub-linear convergence rates for second-order methods have been characterized by [NP06] for second-order smooth setting. Essentially, the so-called Cubic Regularized Newton's Method combines the quadratic Taylor approximation in the typical Newton update with a cubic regularization term. The main challenge of this scheme is efficiently solving the cubic sub-problem, which is possible by explicitly representing it as a one-dimensional convex problem [NP06]. The proposed method achieves $O(1/T^2)$ convergence rate when the objective function is convex. Shortly after, Nesterov [Nes08] proposes an accelerated version of the cubic regularization idea with $O(1/T^3)$ value convergence, pioneering a new direction of research in the study of globally-convergent second-order methods [Mis21]. This idea has been studied further for different settings in convex optimization [JLZ17; JLZ20] with the same accelerated $O(1/T^3)$ rate and extended to non-convex realm [CGT11a; CGT11b], obtaining the analogous rates of $O(1/T^{2/3})$ and $O(1/T^{1/3})$ for finding first-order and second-order stationary points, respectively, leading the way for further investigations [BGM20; DO21; Che+22].

Notice that accelerated cubic regularization is *sub*-optimal such that recent studies prove a respective lower-bound for second-order smooth, convex problems as $O(1/T^{7/2})$ [AH18; ASS19]. The first line of research that shrinks the gap between the upper and lower bounds for achieving an *almost*-optimal (more on this shortly) convergence [Nes18] is the so-called "bisection-type" methods. Pioneered by Monteiro and Svaiter [MS13], these class of algorithms propose a conceptual method where the step-size of the algorithm *implicitly* depends on

the next iterate. To resolve, the authors propose a bisection procedure that simultaneously finds a step-size/next iterate pair that satisfies the conditions of the iterative update, which enables the convergence rate of $O(1/T^{7/2})$, modulo the complexity of bisection procedure. This idea was very recently generalized for higher-order tensor methods [Gas+19]. Not so surprisingly, the same construction finds application in variational inequality (VI) and min-max optimization literature [BL22; JM22]. Very recently and concurrently to our work, [Car+22] propose the first bisection free acceleration for second-order methods, that achieves the optimal $O(1/T^{7/2})$. The authors define an *explicit*, deterministic procedure called MS oracle and compute the step-size using a line-search procedure enabling them to achieve optimal rates while adaptively computing the step-size without needing to know the smoothness constant.

Although there are promising results with an increasing interest into second-order –and also higher-order– methods, we identify three main shortcomings in the literature, which we will systematically address in the sequel. First, bisection-type methods achieve the optimal convergence rate however, the search procedure is computationally very prohibitive [Nes18; LJ22] and the resulting algorithms are complicated with many interconnected components. On the other hand, cubic regularization-based ideas propose a simple construction that achieves acceleration beyond $O(1/T^2)$ however, similar to previous methods, they either require the knowledge of smoothness constant or need to execute a standard line-search procedure to estimate it locally. A common drawback for both approaches is that the algorithmic constructions are designed for handling *only* deterministic oracles and it is an open question whether such frameworks could immediately accommodate stochastic first and second-order information.

Our contributions. To address the aforementioned issues, we developed the first universal and adaptive second-order algorithm, EXTRA-NEWTON, for convex minimization. We summarize our contributions as follows:

1. We prove EXTRA-NEWTON achieves the global convergence rate of $O(\frac{\sigma_g}{\sqrt{T}} + \frac{\sigma_H}{T^{3/2}} + \frac{LD^3}{T^3})$ that adapts simultaneously to the variance in the gradient oracle (σ_g) and Hessian oracle (σ_H) achieving the first universal convergence result in the literature.
2. Our method is completely oblivious to any problem-dependent parameters including smoothness modulus, variance bounds on stochastic oracles, diameter of the constraint set and any possible bounds on the gradient and Hessian.
3. We design the first adaptive step-size, in the sense of [DHS11; RS13], that successfully incorporates second-order information "on-the-fly". While doing so, we bypass any bisection or linesearch procedure, and propose a simple, intuitive algorithmic framework.

From a technical point of view, what will allow us to achieve these results is the combination of three principal ingredients: *(i)* proposing appropriate adjustments to Extra-Gradient [Kor76] that was originally designed for solving variational inequalities and min/max problems; *(ii)* an "optimistic" weighted iterate averaging scheme accompanied by an appropriate gradient rescal-

Table 2.1: A survey on first and second-order algorithms with key properties

	AGD [Nes83a]	**UniXGrad** [Kav+19]	**Reg. Newton** [Mis21]	**Accel. Cubic Reg.** [Nes08]	**ANPE**[1] [MS13]	**OptMS** [Car+22]	**Extra Newton** [AKC22] **[ours]**
Rate	$\frac{1}{T^2}$	$\frac{\sigma_g}{\sqrt{T}} + \frac{1}{T^2}$	$\frac{1}{T^2}$	$\frac{1}{T^3}$	$\frac{1}{T^{7/2}}$	$\frac{1}{T^{7/2}}$	$\frac{\sigma_g}{\sqrt{T}} + \frac{\sigma_H}{T^{3/2}} + \frac{1}{T^3}$
Bisection-free	✓	✓	✓	✓	✗	✓	✓
Adapts to L	✗	✓	✗	Partial	✗	✓	✓
Noise-adaptive	✗	✓	✗	✗	✗	✗	✓

ing strategy in the spirit of [WA18; DO18; Kav+19] which allows us to obtain an accelerated rate of convergence by means of a generalized online-to-batch conversion (Theorem 2.2.3), and *(iii)* the glue that holds these elements together is an adaptive learning rate inspired by [RS13; Kav+19; ABM21] which automatically rescales aggregated gradients and second order information. In what follows, we shall explicate these arguments.

2.2.3 Problem setup

Throughout this section of Chapter 2, we will be focusing on solving the (constrained) convex minimization problems of the general form:

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in \mathcal{X}. \end{aligned} \tag{Opt}$$

Formally, in the above $\mathcal{X}$ is a convex and compact subset of a d- dimensional normed space $\mathcal{V} \cong \mathbb{R}^d$ with diameter $D = \max_{x,y\in\mathcal{X}} \|x-y\|$, and $f : \mathcal{V} \to \mathbb{R} \cup \{+\infty\}$ is a proper, lower semi-continuous, convex function with $\operatorname{dom} f = \{x \in \mathbb{R}^d : f(x) < +\infty\} \subset \mathcal{X}$. To that end, we make a set of blanket assumptions for (Opt). Following the vast literature of constrained convex minimization [NES06; BT09], we consider "simple" constraint sets, i.e.,

Assumption 2.2.1. *The constraint set $\mathcal{X}$ of* (Opt) *possesses favorable geometry which facilitates a tractable projection operator.*

In order to avoid trivialities, we also assume that the said problem admits at least a solution, i.e.

Assumption 2.2.2. *The solution set $\mathcal{X}^* = \arg\min_{x\in\mathcal{X}} f(x)$ of* (Opt) *is non-empty.*

Furthermore, we assume that there exists a Lipschitz continuous selection $x \mapsto \nabla^2 f(x) \in \mathbb{R}^{d\times d}$, i.e.,

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \le L\|x-y\| \quad \forall x, y \in \mathcal{X} \tag{H-smooth}$$

[1] Note that the bisection procedure is computationally prohibitive, we defer the reader to [Nes18], p.304-305.

and in addition it satisfies the second order approximation:

$$f(x) = f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2}\langle \nabla^2 f(y)(x-y), x-y \rangle + O\left(\|x-y\|^3\right) \tag{Taylor}$$

To that end, combining (H-smooth) and (Taylor) we readily get the following inequality:

$$\|\nabla f(x) - \nabla f(y) - \nabla^2 f(y)(x-y)\| \leq \frac{L}{2}\|x-y\|^2 \tag{2.13}$$

The above equivalences are well-established and hence we omit their proofs (we defer for a panoramic view to [Nes19])

Oracle feedback structure. From an algorithmic point of view, we aim to solve (Opt) by using methods that require access to a (stochastic) first and second order-oracle. Before we move forward with the methodology, we shall introduce the definitions and short-hand notation for this oracle model which we will use in algorithm definitions and technical discussions. Let $\nabla f(x,\xi)$ denote the stochastic gradient evaluated at x with randomness defined by ξ and $\nabla^2 f(x,\xi)$ be the stochastic Hessian at x with ξ describing the randomness of the oracle, such that

$$\begin{aligned} \mathbb{E}\left[\nabla f(x,\xi) \mid \sigma(x)\right] &= \nabla f(x), & \mathbb{E}\left[\|\nabla f(x,\xi) - \nabla f(x)\|^2 \mid \sigma(x)\right] &\leq \sigma_g^2 \\ \mathbb{E}\left[\nabla^2 f(x,\xi) \mid \sigma(x)\right] &= \nabla^2 f(x), & \mathbb{E}\left[\|\nabla^2 f(x,\xi) - \nabla^2 f(x)\|^2 \mid \sigma(x)\right] &\leq \sigma_H^2, \end{aligned} \tag{2.14}$$

where $\sigma(x)$ denotes the sigma-algebra generaed by the random variable/iterate x. Due to space constraints, we will also define an operator that accommodates second-order information and its respective stochastic counterpart.

$$\begin{aligned} \mathbf{F}(x;y) &= \nabla f(y) + \frac{1}{2}\nabla^2 f(y)(x-y) \\ \tilde{\mathbf{F}}(x;y,\xi) &= \nabla f(y,\xi) + \frac{1}{2}\nabla^2 f(y,\xi)(x-y) \end{aligned} \tag{2.15}$$

where $\mathbf{F}$ is essentially the gradient (with respect to x) of the second-order Taylor polynomial. By definition, the operator $\mathbf{F}$ satisfies the second-order smoothness property in Eq. (2.13) We present a complete list of definitions and parameter descriptions to make it easier for the reader to follow the technical arguments in the whole of this section.

Table 2.2: A complete list of parameters and expressions, their definitions and descriptions

	Formal Definition	Description
f	$f:\mathbb{R}^d \to \mathbb{R}+\{+\infty\}$	objective function
$\mathscr{X}$	$\mathscr{X} \subset \mathbb{R}^d$	convex and compact constraint set
x^*	$=\arg\min_{x\in\mathscr{X}} f(x)$	solution of the constrained problem (Opt)
D	$=\sup_{x,y\in\mathscr{X}} \|x-y\|$	diameter of the constraint set $\mathscr{X}$
L	$\|\nabla^2 f(x)-\nabla^2 f(y)\| \le L\|x-y\|$	second-order smoothness constant of f
$\nabla f(\cdot,\xi)$	$\mathbb{E}\left[\nabla f(x,\xi) \mid \sigma(x)\right]=\nabla f(x),\ x \perp\!\!\!\perp \xi$	unbiased gradient estimate
$\nabla^2 f(\cdot,\xi)$	$\mathbb{E}\left[\nabla^2 f(x,\xi) \mid \sigma(x)\right]=\nabla^2 f(x),\ x \perp\!\!\!\perp \xi$	unbiased Hessian estimate
$\mathscr{F}_t$	$=\sigma(\xi_1,\xi_{1+\frac{1}{2}},\cdots,\xi_t)$	σ-algebra generated by random variables up to ξ_t
$\mathscr{F}_{t+\frac{1}{2}}$	$=\sigma(\xi_1,\xi_{1+\frac{1}{2}},\cdots,\xi_t,\xi_{t+\frac{1}{2}})$	σ-algebra generated by random variables up to $\xi_{t+\frac{1}{2}}$
σ_g	$\mathbb{E}\left[\|\nabla f(x)-\nabla f(x)\|^2 \mid x\right] \le \sigma_g^2$	variance bound for gradient estimate
σ_H	$\mathbb{E}\left[\|\nabla^2 f(x)-\nabla^2 f(x)\|^2 \mid x\right] \le \sigma_H^2$	variance bound for Hessian estimate
σ	$=\max\{\sigma_g,\sigma_H\}$	maximum variance of oracles
γ_t	Eq. (2.19) and Eq. (2)	adaptive step-size
a_t	$=t^2$	gradient weights
A_t	$=\sum_{s=1}^t a_s$	normalization factor for gradient weights a_t
b_t	$=t^p$, where $p\ge 2$	averaging weights
B_t	$=\sum_{s=1}^t b_s$	normalization factor for averaging weights b_t

2.2.4 Method

In this section, we shall establish our universal second-order framework. Our presentation evolves around three key components: choosing the appropriate algorithmic template with the key motivations behind it, solving implementability issues that commonly arise in higher-order methods and finally designing a universal algorithm that can handle deterministic and noisy oracle feedback simultaneously without having prior knowledge. Our point of departure is the popular Extra-Gradient (EG) template; originally introduced by Korpelevich [Kor76] and further developed in Nemirovski [Nem04],

$$\begin{aligned} X_{t+\frac{1}{2}} &= \Pi_{\mathscr{X}}\left(X_t-\gamma_t\nabla f(X_t)\right)\\ X_{t+1} &= \Pi_{\mathscr{X}}\left(X_t-\gamma_t\nabla f(X_{t+\frac{1}{2}})\right), \end{aligned} \tag{EG}$$

where $\Pi_{\mathscr{X}}(x)=\arg\min_{z\in\mathscr{X}}\|x-z\|^2$ is the standard Euclidean projection onto the set $\mathscr{X}$. In terms of output, the candidate solution returned by (GEG) after T iterations is the so-called

"ergodic average"

$$\bar{X}_{T+\frac{1}{2}} = \frac{\sum_{t=1}^{T} b_t X_{t+\frac{1}{2}}}{\sum_{t=1}^{T} b_t} \tag{2.16}$$

Then, taking $b_t = \gamma_t$ and assuming the method's step-size γ_t is chosen appropriately, $\bar{X}_{T+\frac{1}{2}}$ enjoys the following universal guarantee [JNT11; RS13]:

$$\mathbb{E}[f(\bar{X}_{T+\frac{1}{2}}) - f(x^*)] = \mathcal{O}\left(\frac{1}{T} + \frac{\sigma}{\sqrt{T}}\right) \tag{2.17}$$

where σ signifies the effect of the noisy feedback. However, as it becomes apparent, the vanilla (GEG) template is not capable of matching the iconic $1/T^2$ for the smooth deterministic case. It is well-established in the literature of smooth, convex minimization that iterate averaging (or momentum in the sense of Nesterov [Nes83a]) is essential for matching the $O(1/T^2)$ lower bounds. In fact, plain uniform averaging is not sufficient; one needs to introduce new iterates with *increasing* weights. Precisely, this is equivalent to computing an average by taking $b_t = O(t)$. However, we cannot fully characterize the acceleration machinery without what we like to call "gradient weighting". On top of (weighted) iterate averaging, gradients must be multiplied by the *same order of weights* to achieve acceleration, which is a recurring theme in the literature of accelerated and universal optimization [Tse08; Xia10; Lan12; AO16; LYC18; WA18; Cut19; Kav+19; Jou+20].

Going back to discussion on (GEG), Wang and Abernethy [WA18] and Kavis et al. [Kav+19] provide useful insights into acceleration within the context of (GEG). Wang and Abernethy [WA18] identifies a 2-player game with a particular structure called FENCHELGAME framework, which essentially reduces to minimizing a smooth, convex function when the players cooperate. By introducing an "optimistic" weighted iterate averaging along with a complementary gradient weighting strategy, the framework recovers different acceleration schemes of Nesterov [Nes83a; Nes88; Nes05]. On a related front, Diakonikolas and Orecchia [DO18] proposes the first acceleration of (GEG) by appropriately integrating the optimistic averaging idea [WA18] into the (GEG) template as follows:

$$\tilde{X}_t = \frac{b_t X_t + \sum_{s=1}^{t-1} b_s X_{s+\frac{1}{2}}}{\sum_{s=1}^{t} b_s}, \qquad \bar{X}_{t+\frac{1}{2}} = \frac{\sum_{s=1}^{t} b_s X_{s+\frac{1}{2}}}{\sum_{s=1}^{t} b_s} \tag{2.18}$$

where $b_t = O(t)$ is the "iterate averaging" parameter. Later on, Kavis et al. [Kav+19] designs an adaptive, universal variant of accelerated Mirror-Prox following the same optimistic averaging idea as in Eq. (2.18) (see Eq. (2.6) for the equivalent definition). All in all, it is a recurring theme among accelerated algorithms to adopt weighted iterate averaging ($b_t = O(t)$) with proportionate gradient weighting, and not so surprisingly, prior work establishes clear connections between the degree of weighting and convergence rate. Cutkosky [Cut19] designs a black-box reduction that accelerates a class of online algorithms and proves that the rate of convergence of the reduction is $O(1/\sum_{t=1}^{T} b_t)$ for $b_t \in [1, t]$. In retrospect, we aim at answering the following question;

What algorithmic construction would enable acceleration beyond $O(1/T^2)$?

Implicit algorithm

We give a first affirmative answer to the above question by presenting our implicit accelerated algorithm which is constructed upon (GEG), and establish its convergence properties. Note that the implicitness of the scheme serves as a gentle introduction to the actual explicit second order acceleration, which shall follow. Formally, our scheme is given via the following recursion:

$$\begin{aligned} X_{t+\frac{1}{2}} &= \Pi_{\mathscr{X}}\left(X_t - \gamma_t a_t \mathbf{F}(\bar{X}_{t+\frac{1}{2}}; \tilde{X}_t)\right) \\ &= \arg\min_{x\in\mathscr{X}} a_t\langle \nabla f(\tilde{X}_t) + \frac{1}{2}\nabla^2 f(\tilde{X}_t)(\bar{X}_{t+\frac{1}{2}} - \tilde{X}_t), x - X_t\rangle + \frac{\|x - X_t\|^2}{2\gamma_t} \\ X_{t+1} &= \Pi_{\mathscr{X}}\left(X_t - \gamma_t a_t \nabla f(\bar{X}_{t+\frac{1}{2}})\right) \\ &= \arg\min_{x\in\mathscr{X}} a_t\langle \nabla f(\bar{X}_{t+\frac{1}{2}}), x - X_t\rangle + \frac{\|x - X_t\|^2}{2\gamma_t} \end{aligned} \tag{Implicit}$$

with $\Pi_{\mathscr{X}}(x)$ denoting the Euclidean projection of x onto $\mathscr{X}$, average sequences $\tilde{X}_t$ and $\bar{X}_{t+\frac{1}{2}}$ defined as in (2.18) and the adaptive step-size γ_t defined as (for some $\gamma, \beta_0 > 0$):

$$\gamma_t = \frac{\gamma}{\sqrt{\beta_0 + \sum_{s=1}^{t-1} a_s^2 \|\nabla f(\bar{X}_{s+\frac{1}{2}}) - \mathbf{F}(\bar{X}_{s+\frac{1}{2}}; \tilde{X}_s)\|^2}}. \tag{2.19}$$

The implicit nature of (Implicit) originates from $X_{t+1/2}$ update (which we shall refer to as (corrected) extrapolation step at times) since $\bar{X}_{t+\frac{1}{2}}$ depends upon $X_{t+\frac{1}{2}}$ itself. However, this scheme exhibits several key differences from the vanilla (GEG), which constitute the fundamental parts of our second-order acceleration machinery. In particular, we have:

(i) integration of second-order updates for sharper extrapolation steps - first step of acceleration.

(ii) interplay between averaging (b_t) and gradient weighting (a_t) which allows more aggressive averaging - second step of acceleration.

(iii) adaptive step-size in the sense of Rakhlin and Sridharan [RS13] - key to adaptivity and universality.

Second-order updates. First, we will consider the particular interpretation of (GEG) as an approximation to the Proximal Point method [Roc76] which serves as motivation for the accommodation of second-order information in our scheme.

$$X_{t+1} = X_t - \gamma_t \nabla f(X_{t+1}). \tag{PP}$$

In particular, (GEG) tries to approximate X_{t+1} by generating the extrapolated point $X_{t+\frac{1}{2}}$, and make use of the gradient at $X_{t+\frac{1}{2}}$ to take a step from X_t to X_{t+1}. Therefore, if the algorithm is able to compute a sharper estimate in the extrapolation step, it should be able live up to the fame of (PP) and display faster convergence. To this end, we augment the extrapolation step by introducing second-order term. Essentially, our algorithm makes use of *second-order Taylor approximation*, as opposed to first-order expansion, only for the extrapolation step, trading-off sharper approximation with second-order information.

Iterate averaging and gradient weighting. Now, we turn our attention to the second component in our acceleration machinery; averaging and weighting. Recall that the acceleration framework of Cutkosky [Cut19] guarantees a value convergence rate of $O(1/t^{p+1})$ when weighting factor satisfies $b_t = O(t^p)$ with $p \in [0,1]$. We take this result one step beyond in two fronts; our algorithm exploits higher-order smoothness in order to extend this bound for $p \in [0,2]$, implying the accelerated rate of $O(1/T^3)$. Second, we observe that previous work restricts the choice of gradient weights and averaging weights by taking $a_t \approx b_t$. We decouple those weights by allowing the sequences a_t and b_t to be *different*, which in turn equips us with more aggressive iterate averaging when necessary.

Adaptive step-size. As the final component, we study the adaptive step-size (2.19) from the parameter adaptation perspective (i.e., adaptation to the Lipschitz modulus) and expand on its universal properties in the next section. The vast literature on adaptive methods predominantly rely on constructions of AdaGrad-like decreasing step-size policies by accumulating the observed gradient norms in its denominator. The intuition behind this choice is that whenever the method approaches a solution, the vanishing gradients bring about stabilization, ensuring progress around the solution's neighborhood. However, this idea fails for (compactly) constrained problems; when the solution lies on the boundary. So inspired by [RS13], we design a constraint-aware step-size by accumulating $\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \mathbf{F}(\bar{X}_{t+\frac{1}{2}}; \tilde{X}_t)\|^2$ which converges to 0 as $\bar{X}_{t+\frac{1}{2}} - \tilde{X}_t \to 0$; which in turn implies convergence of the algorithm. To our knowledge, this is the first adaptive step-size that accommodates second order information.

Having established the core components of our design, we are in position to present the first accelerated convergence rate guarantee for (Implicit). Formally, this is given by the following.

Theorem 2.2.1. *Let $\{X_{t+\frac{1}{2}}\}_{t=1}^T$ be generated by* (Implicit) *run with the adaptive step-size policy* (2.19) *where $a_t = t^2$, $b_t = t^p$ with $p \geq 2$. Assume that f satisfies* (H-smooth) *then, it is ensured*

$$f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \leq O\left(\frac{\max\left\{\sqrt{\beta_0}\frac{D^2}{\gamma}, L\frac{D^4+D\gamma^3}{\gamma}\right\}}{T^3}\right)$$

When $\gamma = D$, we obtain the converge rate $O\left(\frac{\max\left\{LD^3, \sqrt{\beta_0}D\right\}}{T^3}\right)$.

Remark 2.2.1. We emphasize that the above rate *does not* require any prior knowledge of

problem paramaters such as L, D, time-horizon T and any bounds on gradient/Hessian norms. In order to have better dependence on D one could set $\gamma = D$, and our rate of $O(1/T^3)$ holds irrespective of γ.

Explicit algorithm

Despite the fact that (Implicit) improves upon the accelerated rate of $O(1/T^2)$, one may easily observe that it exhibits the following drawbacks:

1. (Implicit) is a conceptual algorithm and therefore, *not* implementable in practice.
2. A fortiori, it cannot provide rate interpolation guarantees as it does not have the machinery to simultaneously cope with deterministic and stochastic feedback.

As discussed earlier, a common strategy for overcoming this implicit construction is using a bisection/line-search procedure [JM22; MS13; BL22]. Depending on the context, this procedure serves two *distinct* purposes. Primarily, it tackles the implicit nature of the update rule by simultaneously finding a pair of $(\gamma_t, X_{t+\frac{1}{2}})$ and secondly, it enables adaptation to the second-order smoothness. However, one may identify major setbacks with these approaches; first, it is not clear how to handle stochastic oracles for executing the search procedure, so it is not capable of satisfying any universal guarantees. Moreover, it yields a rather complicated procedure as a byproduct that has many moving parts. To that end, we propose an alternative approach which not only yields a simple scheme, but also provides a universal algorithm that is able to handle noisy feedback on-the-fly. Without further ado, we display our explicit algorithm, EXTRA-NEWTON, with appropriate modifications. Having defined our main scheme, Algorithm 3, we will provide a more detailed description of its components.

Algorithm 3: EXTRA-NEWTON

Input: $X_1 \in \mathcal{X}$, $a_t = t^2$ and $A_t = \sum_{s=1}^t a_s$, $b_t = t^p$ $(p \geq 2)$ and $B_t = \sum_{s=1}^t b_s$, $\gamma > 0$, $\xi_t \sim$ i.i.d.

1: **for** $t = 1$ to T **do**

2: $$\gamma_t = \frac{\gamma}{\sqrt{\beta_0 + \sum_{s=1}^{t-1} a_s^2 \|\nabla f(\bar{X}_{s+\frac{1}{2}}, \xi_{s+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{s+\frac{1}{2}}; \tilde{X}_s, \xi_s)\|^2}}$$

3: $$X_{t+\frac{1}{2}} = \arg\min_{x \in \mathcal{X}} \langle a_t \nabla f(\tilde{X}_t, \xi_t), x \rangle + \frac{a_t b_t}{2B_t} \langle \nabla^2 f(\tilde{X}_t, \xi_t)(x - X_t), x - X_t \rangle + \frac{1}{2\gamma_t} \|x - X_t\|^2$$

4: $$X_{t+1} = \arg\min_{x \in \mathcal{X}} \langle a_t \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), x \rangle + \frac{1}{2\gamma_t} \|x - X_t\|^2$$

5: **end for**

Universal step-size. We modify our step-size (see Line 2, Algorithm 3) in order to operate in the stochastic regime while making it noise-adaptive for rate interpolation. Using the same weighted averaging scheme in Eq. (2.18), we define the universal counterpart of the adaptive step-size. Note that γ_t is independent of any variable/randomness generated at iteration t; it accumulates $a_t^2 \|\nabla f(\bar{X}_{s+\frac{1}{2}}, \xi_{s+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{s+\frac{1}{2}}; \tilde{X}_s, \xi_t)\|^2$ up to $t-1$. Therefore, the step-size is decoupled from the explicit update, *a priori*.

Now, what remains is a new algorithmic design that will retain the accelerated convergence properties demonstrated by (Implicit) while having an explicit construction that is capable of automatically adjusting to noise level in the oracle feedback. Before expanding upon the technical details of our strategy, let us take our time to explain the consequences of our explicit design compared to (Implicit).

From implicit to explicit. To obtain the explicit algorithm, *(i)* we write the projection sub-problem in the argmin form; *(ii)* introduce *stochastic* oracle feedback; *(iii)* for the second-order term, replace $X_{t+\frac{1}{2}}$ in $\bar{X}_{t+\frac{1}{2}}$ with the free variable x; then, *(iv)* simplify as follows:

$$\frac{a_t}{2}\langle \nabla^2 f(\tilde{X}_t,\xi_t)(\bar{X}_{t+\frac{1}{2}}-\tilde{X}_t), x-X_t\rangle$$
$$\Downarrow$$
$$\frac{a_t}{2}\left\langle \nabla^2 f(\tilde{X}_t,\xi_t)\left(\frac{b_t X_{t+\frac{1}{2}}+\sum_{s=1}^{t-1} b_s X_{s+\frac{1}{2}}}{B_t}-\frac{b_t X_t+\sum_{s=1}^{t-1} b_s X_{s+\frac{1}{2}}}{B_t}\right), x-X_t\right\rangle$$
$$\Downarrow$$
$$\frac{a_t}{2}\left\langle \nabla^2 f(\tilde{X}_t,\xi_t)\left(\frac{b_t x+\sum_{s=1}^{t-1} b_s X_{s+\frac{1}{2}}}{B_t}-\frac{b_t X_t+\sum_{s=1}^{t-1} b_s X_{s+\frac{1}{2}}}{B_t}\right), x-X_t\right\rangle$$
$$\Downarrow$$
$$\frac{a_t b_t}{2B_t}\langle \nabla^2 f(\tilde{X}_t,\xi_t)(x-X_t), x-X_t\rangle$$

Given the bisection-type conceptual methods [MS13; JM22; BL22], it is surprising how smoothly we could transition from implicit to explicit *once* we decouple the step-size from the current iteration *apriori*. Moreover, the resulting update rule for the extrapolation step retains the quadratic structure as the X_{t+1} update rule. Having analyzed the components of the explicit scheme, we will first present the universal convergence rates then provide a concise explanation of the proof strategy with particular emphasis on the principal components of the analysis.

Theorem 2.2.2. *Let $\{X_{t+\frac{1}{2}}\}_{t=1}^T$ be a sequence generated by Algorithm 3, run with the adaptive step-size policy (2) and $a_t = t^2, b_t = t^p$ with $p \geq 2$. Assume that f satisfies (H-smooth), and that Assumptions (2.14) hold. Then, the following universal guarantee holds:*

$$\mathbb{E}[f(\bar{X}_{T+\frac{1}{2}})] - f(x^*) = O\left(\frac{\frac{D^2+\gamma^2}{\gamma}\sigma_g}{\sqrt{T}} + \frac{\frac{D^3+D\gamma^2}{\gamma}\sigma_H}{T^{3/2}} + \frac{\max\left\{L\frac{D^4+D\gamma^3}{\gamma}, \sqrt{\beta_0}\frac{D^2+\gamma^2}{\gamma}\right\}}{T^3}\right)$$

When $\gamma = D$, we obtain the target rate $O\left(\frac{D\sigma_g}{\sqrt{T}} + \frac{D^2\sigma_H}{T^{3/2}} + \frac{\max\left\{LD^3, \sqrt{\beta_0}D\right\}}{T^3}\right)$.

Remark 2.2.2. Similar to Theorem 2.2.1, EXTRA-NEWTON achieves the preceding convergence rate independent of the knowledge of problem parameters.

Compatible with the (GEG)-based algorithmic design, our proof has the following main steps

1. We perform an *offline* regret analysis of Algorithm 3 and show adaptive regret bounds - see Proposition 2.2.1.
2. We prove an anytime online-to-batch conversion framework, which generalizes that of Cutkosky [Cut19] and Lemma 2.1.1, through decoupling iterate averaging from gradient weighting - see Theorem 2.2.3.
3. Combining the adaptive regret bound with the conversion theorem immediately implies *universal, accelerated* value convergence of $O(\frac{D\sigma_g}{\sqrt{T}} + \frac{D^2\sigma_H}{T^{3/2}} + \frac{\max\left\{LD^3, \sqrt{\beta_0}D\right\}}{T^3})$ - see Theorem 2.2.2.

Let us begin with clarifying what *offline regret* means for Algorithm 3. We define the (linear) regret considering the convention in both online learning [RS13; Cut19] and first-order acceleration literature [WA18; Kav+19; Jou+20]. We measure the performance of our decisions for the extrapolation sequence such that after playing $X_{t+\frac{1}{2}}$, our algorithm observes and suffers the linear (weighted) loss with respect to $a_t \nabla f(\bar{X}_{t+\frac{1}{2}})$. Hence, we define the regret as

$$\mathrm{REG}_T(x) = \sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x \rangle \tag{Reg}$$

where we run the algorithm for T rounds. Next up, we provide our generalized conversion result.

Theorem 2.2.3. *Let* $\mathrm{REG}_T(x^*)$ *denote the anytime regret for the decision sequence* $\{X_{t+\frac{1}{2}}\}_{t=1}^T$ *as in* (Reg)*, and define two sequences of non-decreasing weights* a_t *and* b_t *such that* $a_t, b_t \geq 1$*. As long as* a_t / b_t *is ensured to be non-increasing,*

$$f(\bar{X}_T) - f(x^*) \leq \frac{\mathrm{REG}_T(x^*)}{a_T \frac{B_T}{b_T}}$$

Remark 2.2.3. This conversion result holds independent of the order of smoothness of the objective as long as f is convex. Moreover, it allows averaging parameter b_t to be asymptotically larger than gradient weights a_t, enabling a more aggressive averaging strategy when necessary.

To complement the lower bound to the regret $\mathrm{REG}_T(x^*)$, we present an upper bound that helps us explain how we exploit second-order smoothness for a more aggressive weighting, hence the rate $O(1/T^3)$.

Proposition 2.2.1. *Let* $\{X_{t+\frac{1}{2}}\}_{t=1}^T$ *be generated by Algorithm 3, run with a non-increasing step-size sequence* γ_t *and non-decreasing sequences of weights* $a_t, b_t \geq 1$ *such that* a_t / b_t *is also non-increasing. Then, the following guarantee holds:*

$$\mathbb{E}\left[\mathrm{REG}_T(x^*)\right] \leq \frac{1}{2}\mathbb{E}\left[\frac{3D^2}{\gamma_{T+1}} + \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}}; \tilde{X}_t, \xi_t)\|^2 - \frac{\|X_{t+\frac{1}{2}} - X_t\|^2}{\gamma_{t+1}}\right]$$

Observe that the inequality in Proposition 2.2.1 is agnostic to the design of our step-size in Algorithm 3 (line 2) as well as the selection of the weights as described in Theorem 2.2.2. It essentially applies to any non-increasing sequence of step-sizes and non-decreasing gradient weight sequence $a_t \geq 1$. To obtain it, we neither used convexity nor the smoothness of the objective. In fact, the structure of the objective function, i.e., its convexity, will not be needed for upper-bounding the regret expression, and required only for the conversion in Theorem 2.2.3.

Now, let us explain how we make use of second-order smoothness for enjoying faster rates, and give a brief discussion of how the regret bound will look in its final form. First, we decompose the stochastic term $\|\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}};\tilde{X}_t,\xi_t)\|^2$ into deterministic feedback and noise. Then, we argue that *the noisy component* grows as $O(\sigma_H T^{3/2} + \sigma_g T^{5/2})$. On the other hand, achieving the accelerated $O(1/T^3)$ component of the universal rate amounts to showing that the regret has a constant, $O(1)$, component. In the worst-case sense, however, *the deterministic component itself* grows as $O(T^{5/2})$. Fortunately, we identify that the negative term is "large enough" in magnitude to control the growth of the deterministic term, permitting a constant component $O(LD^2)$ for the regret. The intuition behind using the negative summation term is of the same spirit as we discussed in Section 2.1.5. However, the main challenge is understanding the correct use of higher-order smoothness in conjunction with acceleration mechanism and adaptive step-size design.

Although the regret bound of $O(LD^3 + D^2\sigma_H T^{3/2} + D\sigma_g T^{5/2})$ seems counter-intuitive from an online-learning perspective, we discuss how second-order smoothness leads to "faster" conversion through more aggressive averaging compared to the result presented in Lemma 2.1.1. As a matter of fact, we will continue our discussion with how second-order smoothness helps us accelerate. It turns out that using (H-smooth), iterate averaging as in Eq.(2.18) and compactness of $\mathscr{X}$, we can bound the negative term as,

$$-\frac{1}{\gamma_{t+1}}\|X_{t+\frac{1}{2}} - X_t\|^2 \leq -\frac{1}{L^2D^2\gamma_{t+1}}t^4\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \mathbf{F}(\bar{X}_{t+\frac{1}{2}};\tilde{X}_t)\|^2$$

Observe that to seamlessly combine the positive and negative terms, our analysis enforces that $a_t = O(t^2)$ and $b_t = \Omega(t^2)$. Then, the conversion implies a convergence rate of $\textsc{Reg}_{\textsc{T}}(x^*)/T^3$, hence the recipe for acceleration. Therefore, the constant component of the regret amounts to $O(1/T^3)$ convergence rate, while the stochastic component of the regret implies $O(\sigma_H/T^{3/2} + \sigma_g/\sqrt{T})$ rate, giving us the first universal acceleration beyond first-order smoothness.

Let us conclude by discussing the intricate relationship between the universal step-size and the regret bounds. Simply put, growth of the summation in the denominator of γ_t is of the same order as the regret bound. Under stochastic gradient and Hessian oracles, the regret bound is of order $O(T^{5/2})$, and we can trivially show using variance bounds that the step-size is lower bounded by $O(T^{-5/2})$. On the other extreme, the regret bound described in Proposition 2.2.1 is bounded by a constant under deterministic oracles, which implies that the summation in the denominator of the step-size is in turn summable, i.e., the step-size has a positive, constant

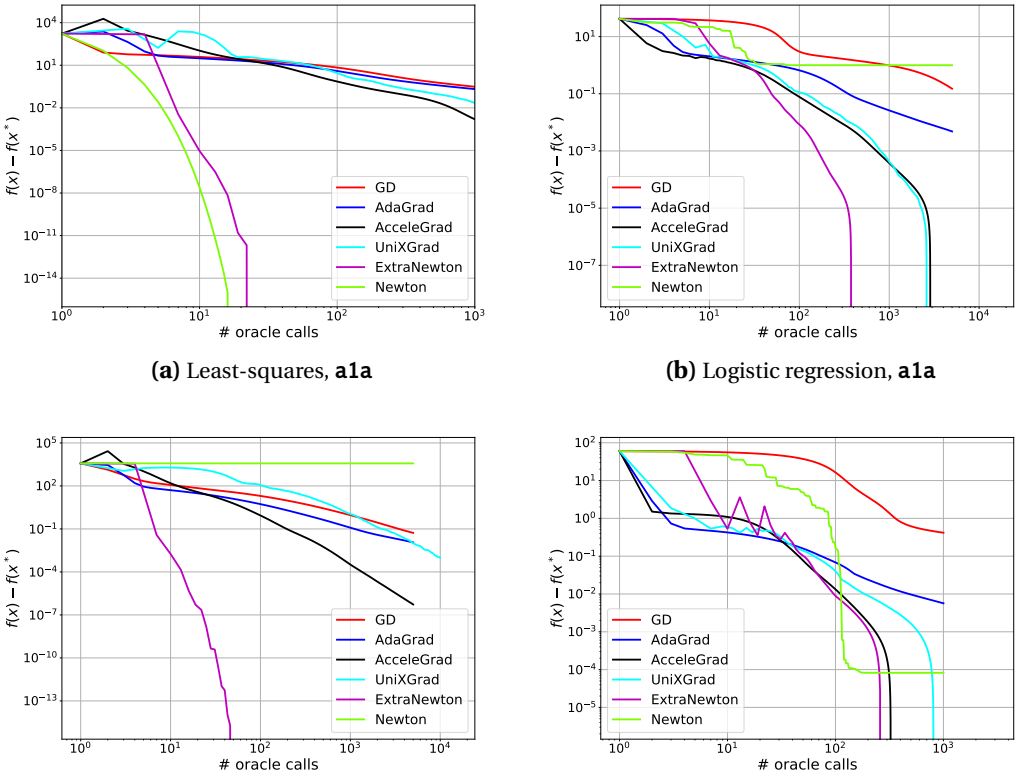
lower bound. This adaptive behavior of our step-size enables automatic adaptation to noise levels and thus the universal rates.

2.2.5 Experiments

In this section, we will present practical performance of EXTRA-NEWTON against a set of first-order algorithms, e.g., GD, SGD, ADAGRAD [DHS11], ACCELEGRAD [LYC18], UNIXGRAD [Kav+19]; and second-order methods, e.g., NEWTON'S, Optimal Monteiro-Svaiter (OPTMS) [Car+22], Cubic Regularization of Newton's method (CRN) [NP06] and Accelerated CRN (ACRN) [Nes08] for least squares and logistic regression problems over a LIBSVM datasets, `a1a`, `a9a` and `w1a`. Our objective is three-folds. First, we compare the performance of our algorithm against first-order methods for least-squares and logistic regression problems (Figure 2.4). Second, we demonstrate the behavior of algorithm against first-order methods under stochastic oracles (Figure 2.5). Finally, we want to understand how our algorithm compares against other second-order methods and we test multiple algorithms for optimizing a regularized logistic regression problem (Figure 2.6). Note that we consider the black-box oracle model in which the algorithms only have access to gradient and Hessians without knowing the actual objective function.

In the whole of Figure 2.4, the statement *# of oracle calls* on the x-axis counts any gradient or Hessian computation as one oracle call. When the problem is suitable, second-order methods show promising performance with truly superior run time. In Figure 2.4a and 2.4c, we display the result for least squares setting under deterministic oracles using `a1a, w1a` datasets from LibSVM. Second-order methods are known to be suitable for quadratic problems, and our method exploits its hybrid construction to converge faster than first-order methods, (almost) matching the behavior of NEWTON'S. For the logistic regression problem in Figure 2.4b and 2.4d, we regularize it with $g(x) = 1/2\|x\|^2$, but use a very small regularization constant to render the problem ill-conditioned, making things slightly more difficult for the algorithms [MBR19; Mis21]. The difference between ours and the first-order methods is less pronounced for this objective, but the faster sublinear rate is still observable especially in Figure 2.4b.

Although we implement NEWTON'S with line-search, we actually observed a sporadic convergence behavior; when the initial point is close to the solution it converges similarly to EXTRA-NEWTON, in fact even faster, however when we initialize further away it doesn't converge. This non-convergent behavior has been known for NEWTON'S, even with line-search present [JT16]. On the contrary, EXTRA-NEWTON consistently converges; even if we perturb the initial step-size and make it adversarially large, it manages to recover due to its adaptive construction. Next, we have the experiments under stochastic oracles in Figure 2.5. We compute mini-batch gradient estimates with a batch-size of 50 samples. We plot the mean of 5 trials for all the methods under mini-batch gradients and also display the variance as the shaded region around the mean curve. The x-axis displays the progress with respect to epochs



(a) Least-squares, **a1a**

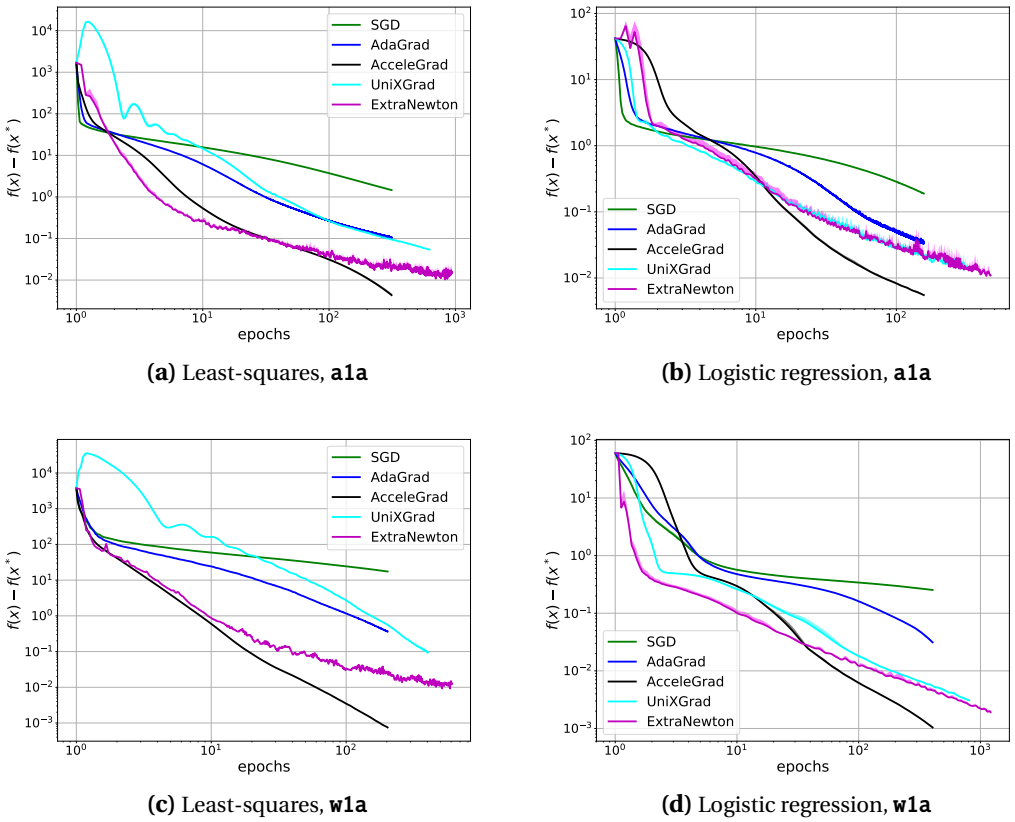
(b) Logistic regression, **a1a**

(c) Least-squares, **w1a**

(d) Logistic regression, **w1a**

Figure 2.4: Comparison of value convergence for regression problems with **deterministic** oracle access

such that one complete pass over the dataset denotes 1 epoch. In terms of performance in the stochastic setting, our method does not offer a significant performance improvement against the others but has faster convergence in the early iterations. ACCELEGRAD seem to have the best performance overall, followed by our EXTRA-NEWTON. We essentially present these results for two main reasons; to show that our method works seamlessly with stochastic gradients without any modifications, and to demonstrate that EXTRA-NEWTON achieves the $O(1/\sqrt{T})$ rate (same as other methods we compare against) when the gradient information is noisy.



(a) Least-squares, **a1a**

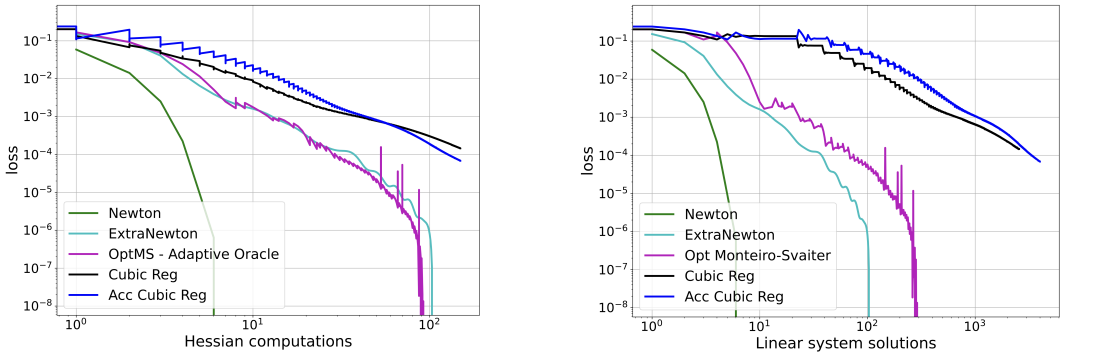
(b) Logistic regression, **a1a**

(c) Least-squares, **w1a**

(d) Logistic regression, **w1a**

Figure 2.5: Comparison of value convergence for regression problems with **stochastic** oracle access

We complement our numerical tests by comparing EXTRA-NEWTON with a set of second-order methods. To that end, we implemented our method within the framework presented in [Car+22]. Using the implementation and the experimental setup provided in their GitHub repository [Hau22], we implemented our method in their code and compared against NEWTON'S, CRN, ACRN and OPTMS algorithms. Figure 2.6 shows that EXTRA-NEWTON has comparable performance to OPTMS, which has the theoretically faster rate $O(1/T^{7/2})$, and marginally outperforms with respect to number of linear system solutions since the linesearch procedure of OPTMS might require multiple system solutions per iteration. While CRN and ACRN has worse convergence than EXTRA-NEWTON, NEWTON'S seems to have the fastest. Note that the initialization favors NEWTON'S as it lies in a close neighborhood of the solution, and NEWTON'S performance sporadically deteriorates when initialized arbitrarily. We observe that the main advantage of our approach, and in general that of second-order methods, becomes apparent when the problem at hand has a compatible structure such as least-squares. Intuitively, second-order methods should benefit when the cost of computing the Hessian is comparable to gradient computation. In fact, quadratic problems like least-squares yield a constant Hessian for any point in the domain, granting a significant advantage to second-order methods. We exemplify this behavior for least-squares problem with deterministic oracles.



(a) Value convergence w.r.t # Hessian oracle calls **(b)** Value convergence w.r.t. # linear system solutions

Figure 2.6: EXTRA-NEWTON vs. second-order methods. Logistic regression with `a9a` dataset

With `w1a` dataset, we couldn't get Newton's method to converge once again. On the contrary, our method shows significant performance upgrade compared to first-order methods while converging consistently in all our trials.

2.2.6 Conclusion

In this work, we present the *first* universal, second-order algorithm, FINEGRAD, which enjoys the value convergence rate of $O(\sigma_g/\sqrt{T}+\sigma_H/T^{3/2}+1/T^3)$. By extending the notion of bounded variance on stochastic gradients to stochastic *Hessian*, we prove adaptation to the noise in first and second-order oracles, simultaneously, while showing accelerated rates matching that of Nesterov [Nes08] under the fully deterministic oracle model. To that end, an important open question is whether we could design a method that achieves an improved rate interpolation guarantee $O(\sigma_g/\sqrt{T}+\sigma_H/T^{3/2}+1/T^{7/2})$ without depending on any line-search/bisection mechanism. We defer this to a future work.

2.3 APPENDIX: Proofs of Chapter 2

2.3.1 Proofs for Section 2.1

First, we discuss a generic scheme that enables us to relate our weighted regret bounds to optimality gap, hence the convergence rate. Once again, note that our analysis borrows tools and techniques from online learning literature and applies them to offline optimization setup. In essence, our conversion scheme applies to a special setting, where the convex loss is fixed across iterations. Let us give the respective Lemma and its proof.

Lemma 2.1.1. *Consider weighted average $X_{t+\frac{1}{2}}$ as in Eq. (2.6). Let $\mathrm{REG}_T(x^*) = \sum_{t=1}^{T} \alpha_t \langle g_t, X_{t+\frac{1}{2}} - x^* \rangle$ denote the weighted regret after T iterations, $\alpha_t = t$ and $g_t = \nabla f(\bar{X}_{t+\frac{1}{2}})$. Then,*

$$f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \leq \frac{2\mathrm{REG}_T(x^*)}{T^2}.$$

Proof. Let's define $A_t = \sum_{s=1}^{t} \alpha_s$. Then, by definition, we could express $X_{t+\frac{1}{2}}$ as

$$X_{t+\frac{1}{2}} = \frac{A_t}{\alpha_t} \bar{X}_{t+\frac{1}{2}} - \frac{A_{t-1}}{\alpha_t} \bar{X}_{t-\frac{1}{2}}. \tag{2.20}$$

Then, use Eq. (2.29) and replace g_t by $\nabla f(\bar{X}_{t+\frac{1}{2}})$ in the weighted regret expression, i.e.

$$\begin{aligned}
\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle &= \sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), \frac{A_t}{\alpha_t} \bar{X}_{t+\frac{1}{2}} - \frac{A_{t-1}}{\alpha_t} \bar{X}_{t-\frac{1}{2}} - x^* \rangle \\
&= \sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), \frac{A_t}{\alpha_t} \left(\bar{X}_{t+\frac{1}{2}} - x^* \right) - \frac{A_{t-1}}{\alpha_t} \left(\bar{X}_{t-\frac{1}{2}} - x^* \right) \rangle \\
&= \sum_{t=1}^{T} A_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), \bar{X}_{t+\frac{1}{2}} - x^* \rangle - A_{t-1} \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), \bar{X}_{t-\frac{1}{2}} - x^* \rangle \\
&= \sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), \bar{X}_{t+\frac{1}{2}} - x^* \rangle + A_{t-1} \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), \bar{X}_{t+\frac{1}{2}} - \bar{X}_{t-\frac{1}{2}} \rangle \\
&\geq \sum_{t=1}^{T} \alpha_t \left(f(\bar{X}_{t+\frac{1}{2}}) - f(x^*) \right) + A_{t-1} \left(f(\bar{X}_{t+\frac{1}{2}}) - f(\bar{X}_{t-\frac{1}{2}}) \right) \\
&= \sum_{t=1}^{T} A_t \left(f(\bar{X}_{t+\frac{1}{2}}) - f(x^*) \right) - A_{t-1} \left(f(\bar{X}_{t-\frac{1}{2}}) - f(x^*) \right) \\
&= A_T f(\bar{X}_{T+\frac{1}{2}}) - f(x^*)
\end{aligned}$$

where we used gradient inequality in the last line. As the final step, we telescope the terms in the summation and take $\alpha_0 = 0$ and $A_0 = 0$. Finally, observe that $A_T \geq \frac{T^2}{2}$,

$$A_T (f(\bar{X}_{T+\frac{1}{2}}) - f(x^*)) \leq \sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle$$

$$f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) = \frac{1}{A_T}\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle$$
$$f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \leq \frac{2\mathrm{REG}_T(x^*)}{T^2}$$

■

As we have mentioned previously, for the weighted regret analysis in the non-smooth case, i.e., f is only G-Lipschitz, please observe that we do not exploit the precise definitions of g_t and M_t. As far as the regret analysis is concerned, their dual norm should be bounded. However, we especially rely on the fact that $g_t = \nabla f(\bar{X}_{t+\frac{1}{2}})$ since it is necessary to obtain converge rates from regret-like bounds using Lemma 2.1.1.

Let us bring up the following relation which we will require for the regret analysis of both smooth and non-smooth objective.

Lemma 2.1.2. *Let $\{a_i\}_{i=1,\dots,n}$ be a sequence of non negative numbers. Then, it holds that*

$$\sqrt{\sum_{i=1}^{n} a_i} \leq \sum_{i=1}^{n} \frac{a_i}{\sqrt{\sum_{j=1}^{i} a_j}} \leq 2\sqrt{\sum_{i=1}^{n} a_i}.$$

Please refer to [LYC18; MS10] for the proof of Lemma 2.1.2, which is due to induction. We will also make use of the following bound (due to Young's Inequality)

$$\alpha_t \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_* \|X_{t+\frac{1}{2}} - X_{t+1}\| = \inf_{\rho>0}\left\{\frac{\rho}{2}\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2 + \frac{\alpha_t^2}{2\rho}\|X_{t+\frac{1}{2}} - X_{t+1}\|^2\right\}. \tag{2.21}$$

Next, we have the proof for the case of non-smooth, deterministic setting.

Theorem 2.1.1. *Consider the constrained optimization setting in Problem (Prob), where $f : \mathcal{X} \to \mathbb{R}$ is a proper, convex and G-Lipschitz function defined over compact, convex set $\mathcal{X}$. Let $x^* \in \arg\min_{x\in\mathcal{X}} f(x)$, and define $g_t = \nabla f(\bar{X}_{t+\frac{1}{2}})$, $M_t = \nabla f(\tilde{X}_t)$ where $\nabla f(\cdot)$ represents a sub-gradient of the objective f at the query point. Then, Algorithm 2 guarantees*

$$f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \leq \frac{7D\sqrt{1+\sum_{t=1}^{T}\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2}}{T^2} \leq \frac{6D}{T^2} + \frac{14GD}{\sqrt{T}} = O\left(\frac{D}{T^2} + \frac{GD}{\sqrt{T}}\right).$$

Proof.

$$\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle$$
$$= \sum_{t=1}^{T} \underbrace{\alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t), X_{t+\frac{1}{2}} - X_{t+1} \rangle}_{\text{(A)}} + \underbrace{\alpha_t \langle \nabla f(\tilde{X}_t), X_{t+\frac{1}{2}} - X_{t+1} \rangle}_{\text{(B)}} + \underbrace{\alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+1} - x^* \rangle}_{\text{(C)}}.$$

Bounding (A)

$$\begin{aligned}
&\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t), X_{t+\frac{1}{2}} - X_{t+1} \rangle \\
&\leq \sum_{t=1}^{T} \alpha_t \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_* \|X_{t+\frac{1}{2}} - X_{t+1}\| \quad \text{(Hölder's Inequality)} \\
&\leq \sum_{t=1}^{T} \frac{\rho}{2} \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2 + \frac{\alpha_t^2}{2\rho} \|X_{t+\frac{1}{2}} - X_{t+1}\|^2 \quad \text{(Equation (2.21))}.
\end{aligned}$$

By setting $\rho = \alpha_t^2 \gamma_{t+1}$, we get the following upper bound for term (A),

$$\begin{aligned}
&\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t), X_{t+\frac{1}{2}} - X_{t+1} \rangle \\
&\quad \leq \sum_{t=1}^{T} \frac{\alpha_t^2 \gamma_{t+1}}{2} \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2 + \frac{1}{2\gamma_{t+1}} \|X_{t+\frac{1}{2}} - X_{t+1}\|^2
\end{aligned}$$

Bounding (B)

$$\begin{aligned}
\sum_{t=1}^{T} \alpha_t \langle \nabla f(\tilde{X}_t), X_{t+\frac{1}{2}} - X_{t+1} \rangle &\leq \sum_{t=1}^{T} \frac{1}{\gamma_t} \nabla_x D_h(X_{t+\frac{1}{2}}, X_t)^T (X_{t+1} - X_{t+\frac{1}{2}}) \quad \text{(Optimality for } X_{t+\frac{1}{2}}) \\
&= \sum_{t=1}^{T} \frac{1}{\gamma_t} \Big(D_h(X_{t+1}, X_t) - D_h(X_{t+\frac{1}{2}}, X_t) - D_h(X_{t+1}, X_{t+\frac{1}{2}}) \Big).
\end{aligned}$$

Bounding (C)

$$\begin{aligned}
\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+1} - x^* \rangle &\leq \sum_{t=1}^{T} \frac{1}{\gamma_t} \nabla_x D_h(X_{t+1}, X_t)^T (x^* - X_{t+1}) \quad \text{(Optimality for } X_{t+1}) \\
&= \sum_{t=1}^{T} \frac{1}{\gamma_t} \big(D_h(x^*, X_t) - D_h(X_{t+1}, X_t) - D_h(x^*, X_{t+1}) \big).
\end{aligned}$$

Final Bound

$$\begin{aligned}
&\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle \\
&\leq \sum_{t=1}^{T} \frac{\alpha_t^2 \gamma_{t+1}}{2} \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2 + \frac{1}{2\gamma_{t+1}} \|X_{t+\frac{1}{2}} - X_{t+1}\|^2 \\
&\quad + \frac{1}{\gamma_t} \Big(D_h(x^*, X_t) - D_h(x^*, X_{t+1}) - D_h(X_{t+\frac{1}{2}}, X_t) - D_h(X_{t+1}, X_{t+\frac{1}{2}}) \Big) \\
&\leq \sum_{t=1}^{T} \frac{\alpha_t^2 \gamma_{t+1}}{2} \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2 + \frac{1}{2\gamma_{t+1}} \|X_{t+\frac{1}{2}} - X_{t+1}\|^2
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{\gamma_t}\left(D_h(x^*,X_t)-D_h(x^*,X_{t+1})\right)-\frac{1}{2\gamma_t}\left(\|X_{t+\frac{1}{2}}-X_{t+1}\|^2+\|X_{t+\frac{1}{2}}-X_t\|^2\right)\\
&\le \sum_{t=1}^{T}\frac{\alpha_t^2\gamma_{t+1}}{2}\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2+\sum_{t=1}^{T-1}\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right)D_h(x^*,X_{t+1})\\
&+\frac{1}{2}\sum_{t=1}^{T}\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right)\|X_{t+\frac{1}{2}}-X_{t+1}\|^2+\frac{1}{\gamma_1}D^2\\
&\le \sum_{t=1}^{T}\frac{\alpha_t^2\gamma_{t+1}}{2}\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2+\frac{1}{2}\sum_{t=1}^{T}\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right)\|X_{t+\frac{1}{2}}-X_{t+1}\|^2+\frac{D^2}{\gamma_T}+\frac{D}{2}\\
&\le \sum_{t=1}^{T}\frac{\alpha_t^2\gamma_{t+1}}{2}\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2+D^2\left(\frac{2}{\gamma_{T+1}}+\frac{1}{\gamma_T}\right)+\frac{D}{2}\\
&\le D\sum_{t=1}^{T}\frac{\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2}{\sqrt{1+\sum_{s=1}^{t}\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2}}+\frac{3}{2}D\sqrt{1+\sum_{t=1}^{T}\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2}+\frac{D}{2}\\
&\le \frac{7}{2}D\sqrt{1+\sum_{t=1}^{T}\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2}-\frac{D}{2}\\
&\le 3D+7GD\sqrt{\sum_{t=1}^{T}\alpha_t^2}\\
&\le 3D+7GDT^{3/2}.
\end{aligned}
$$

We obtain the rate by applying Lemma 2.1.1 to the weighted regret bound above. ■

We now complement the previous result with the analysis in the non-smooth, stochastic setting.

Theorem 2.1.2. *Consider the optimization setting in Problem (Prob), where f is non-smooth, convex and G-Lipschitz. Let $\{X_{t+\frac{1}{2}}\}_{t=1,\dots,T}$ be a sequence generated by Algorithm 2 such that $g_t=\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}})$, $M_t=\nabla f(\tilde{X}_t,\xi_t)$ and Assumptions in Eq. (2.4) and (2.5) hold. With $\alpha_t=t$ and step-size as in Eq. (2.7), it holds that*

$$
\mathbb{E}\left[f(\bar{X}_{T+\frac{1}{2}})\right]-f(x^*)\le\frac{6D}{T^2}+\frac{14GD}{\sqrt{T}}=O\left(\frac{D}{T^2}+\frac{GD}{\sqrt{T}}\right).
$$

Proof. Note that $x^*\in\min_{x\in\mathcal{X}}f(x)$. We start with weighted regret bound,

$$
\mathrm{REG}_T(x^*)=\sum_{t=1}^{T}\alpha_t\langle\nabla f(\bar{X}_{t+\frac{1}{2}}),X_{t+\frac{1}{2}}-x^*\rangle.
$$

We separate $\nabla f(\bar{X}_{t+\frac{1}{2}})=\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}})+(\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}}))$ and re-write the above term as

$$
\sum_{t=1}^{T}\alpha_t\langle\nabla f(\bar{X}_{t+\frac{1}{2}}),X_{t+\frac{1}{2}}-x^*\rangle
$$

$$= \underbrace{\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle}_{\text{(A)}} + \underbrace{\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle}_{\text{(B)}}.$$

Due to unbiasedness of the gradient estimates, expected value of $\alpha_t \langle X_{t+\frac{1}{2}} - x^*, \nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) \rangle$, conditioned on the sigma-algebra $\sigma(\bar{X}_{t+\frac{1}{2}}) = \sigma(\xi_1, \xi_{1+1/2}, ..., \xi_{t-\frac{1}{2}}, \xi_t)$ evaluates to 0. We will only need to bound the first summation whose analysis is identical to its deterministic counterpart up to replacing $\nabla f(\bar{X}_{t+\frac{1}{2}})$ with $\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}})$, and $\nabla f(\tilde{X}_t)$ with $\nabla f(\tilde{X}_t, \xi_t)$. Hence, term (A) is upper bounded by $6D + 14GDT^{3/2}$.

In addition to the setup in the deterministic setting, we put forth the assumption that stochastic gradients have bounded norms, which is natural in the constrained optimization framework. Using Lemma 2.1.1, we translate the regret bound into the convergence rate, i.e,

$$\mathbb{E}\left[f(\bar{X}_{T+\frac{1}{2}}) \right] - \min_{x \in \mathcal{X}} f(x) \leq \frac{6D}{T^2} + \frac{14GD}{\sqrt{T}}.$$

■

We will now introduce the L-smooth setting (see Eq. (2.2)). In the sequel, we provide the weighted regret analysis for smooth functions in the presence of deterministic and stochastic oracles and convert these bound into sub-optimality gap via our regret-to-rate scheme.

Theorem 2.1.3. *Consider the constrained optimization setting in Problem (Prob), where $f : \mathcal{X} \to \mathbb{R}$ is a proper, convex and L-smooth function defined over compact, convex set $\mathcal{X}$. Let $x^* \in \min_{x \in \mathcal{X}} f(x)$. Then, Algorithm 2 run with $g_t = \nabla f(\bar{X}_{t+\frac{1}{2}})$ and $M_t = \nabla f(\tilde{X}_t)$ ensures the following*

$$f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \leq \frac{20\sqrt{7}D^2L}{T^2} = O\left(\frac{LD^2}{T^2}\right). \tag{2.22}$$

Proof. Recall the regret analysis for the non-smooth, convex objective

$$\begin{aligned}
&\text{REG}_T(x^*) \\
&\leq \frac{1}{2}\sum_{t=1}^{T} \gamma_{t+1}\alpha_t^2 \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2 + \frac{1}{\gamma_{t+1}}\|X_{t+\frac{1}{2}} - X_{t+1}\|^2 \\
&\quad + \sum_{t=1}^{T} \frac{1}{\gamma_t}\left(D_h(x^*, X_t) - D_h(x^*, X_{t+1})\right) - \frac{1}{2\gamma_t}\left(\|X_{t+\frac{1}{2}} - X_{t+1}\|^2 + \|X_{t+\frac{1}{2}} - X_t\|^2\right) \\
&\leq \frac{1}{2}\sum_{t=1}^{T} \gamma_{t+1}\alpha_t^2 \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2 - \frac{1}{\gamma_t}\|X_{t+\frac{1}{2}} - X_t\|^2 \\
&\quad + \frac{1}{2}\sum_{t=1}^{T}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right)\|X_{t+\frac{1}{2}} - X_{t+1}\|^2 + \sum_{t=1}^{T-1}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right)D_h(x^*, X_{t+1}) + \frac{D^2}{\gamma_1} \\
&= \frac{1}{2}\sum_{t=1}^{T} \gamma_{t+1}\alpha_t^2 \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2 - \frac{1}{\gamma_{t+1}}\|X_{t+\frac{1}{2}} - X_t\|^2 + \frac{1}{2}\sum_{t=1}^{T}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right)\|X_{t+\frac{1}{2}} - X_t\|^2
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2}\sum_{t=1}^{T}\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right)\|X_{t+\frac{1}{2}}-X_{t+1}\|^2+\sum_{t=1}^{T-1}\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right)D_h(x^*,X_{t+1})+\frac{D^2}{\gamma_1}\\
&\leq\frac{1}{2}\sum_{t=1}^{T}\gamma_{t+1}\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2-\frac{1}{2}\sum_{t=1}^{T}\frac{1}{\gamma_{t+1}}\|X_{t+\frac{1}{2}}-X_t\|^2+D^2\left(\frac{2}{\gamma_{T+1}}+\frac{1}{\gamma_T}+\frac{1}{\gamma_1}\right).
\end{aligned}
$$

The key challenge in this analysis is to exploit the negative term, i.e., $-\frac{1}{2}\sum_{t=1}^{T}\frac{1}{\gamma_{t+1}}\|X_{t+\frac{1}{2}}-X_t\|^2$, such that we could tighten the regret bound from non-smooth analysis. Using the smoothness of f and that $\alpha_t=t$, $A_t=\sum_{s=1}^{t}\alpha_s$, $g_t=\nabla f\bar{X}_{t+\frac{1}{2}}$ and $M_t=\nabla f\tilde{X}_t$

$$
\begin{aligned}
\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2&\leq\frac{L^2\alpha_t^2}{A_t^2}\|X_{t+\frac{1}{2}}-X_t\|^2\\
&=\frac{4L^2t^2}{t^2(t+1)^2}\|X_{t+\frac{1}{2}}-X_t\|^2\\
&=\frac{4L^2}{\alpha_{t+1}^2}\|X_{t+\frac{1}{2}}-X_t\|^2\\
&\leq\frac{4L^2}{\alpha_t^2}\|X_{t+\frac{1}{2}}-X_t\|^2.
\end{aligned}
$$

Hence,

$$
-\frac{1}{\gamma_{t+1}}\|X_{t+\frac{1}{2}}-X_t\|^2\leq-\frac{\alpha_t^2}{4L^2\gamma_{t+1}}\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2.
$$

After applying this upper bound and regrouping the terms we have

$$
\mathrm{REG}_T(x^*)\leq\frac{1}{2}\sum_{t=1}^{T}\left(\gamma_{t+1}-\frac{1}{4L^2\gamma_{t+1}}\right)\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2+D^2\left(\frac{2}{\gamma_{T+1}}+\frac{1}{\gamma_T}+\frac{1}{\gamma_1}\right).
$$

Define that $\tau^*=\max\left\{t\in\{1,...,T\}:\frac{1}{\gamma_{t+1}^2}\leq 7L^2\right\}$ such that $\forall t>\tau^*$, $\gamma_{t+1}-\frac{1}{4L^2\gamma_{t+1}}\leq-\frac{3}{4}\gamma_{t+1}$. We can rewrite the above term as

$$
\begin{aligned}
&\mathrm{REG}_T(x^*)\\
&\leq\frac{1}{2}\sum_{t=1}^{\tau^*}\left(\gamma_{t+1}-\frac{1}{4L^2\gamma_{t+1}}\right)\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2+\frac{D^2}{\gamma_1}\\
&\quad+\frac{1}{2}\sum_{t=\tau^*+1}^{T}\left(\gamma_{t+1}-\frac{1}{4L^2\gamma_{t+1}}\right)\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2+\frac{3D^2}{\gamma_{T+1}}\\
&\leq\underbrace{\frac{1}{2}\sum_{t=1}^{\tau^*}\gamma_{t+1}\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2+\frac{D}{2}}_{\text{(A)}}+\underbrace{\frac{3D^2}{\gamma_{T+1}}-\frac{3}{4}\sum_{t=\tau^*+1}^{T}\gamma_{t+1}\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2}_{\text{(B)}}.
\end{aligned}
$$

Bounding (A). We will simply need to use the definition of τ^* and Lemma 2.1.2

$$\begin{aligned}\frac{1}{2}\sum_{t=1}^{\tau^*}\gamma_{t+1}\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2+\frac{D}{2} &= D\sum_{t=1}^{\tau^*}\frac{\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2}{\sqrt{1+\sum_{s=1}^t\alpha_s^2\|\nabla f(\bar{X}_{s+\frac{1}{2}})-\nabla f(\tilde{X}_s)\|_*^2}}+\frac{D}{2}\\ &\le 2D\sqrt{1+\sum_{t=1}^{\tau^*}\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2}\\ &= \frac{4D^2}{\gamma_{\tau^*+1}}\\ &\le 4\sqrt{7}D^2L.\end{aligned}$$

Bounding (B).

$$\begin{aligned}\text{(B)} &\le \frac{3D}{2}\left(\sqrt{1+\sum_{t=1}^T\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2}-\sum_{t=\tau^*+1}^T\frac{\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2}{\sqrt{1+\sum_{s=1}^t\alpha_s^2\|\nabla f(\bar{X}_{s+\frac{1}{2}})-\nabla f(\tilde{X}_s)\|_*^2}}\right)\\ &\le \frac{3D}{2}+\frac{3D}{2}\left(\sum_{t=1}^T\frac{\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2}{\sqrt{1+\sum_{s=1}^t\alpha_s^2\|\nabla f(\bar{X}_{s+\frac{1}{2}})-\nabla f(\tilde{X}_s)\|_*^2}}-\sum_{t=\tau^*+1}^T\frac{\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2}{\sqrt{1+\sum_{s=1}^t\alpha_s^2\|\nabla f(\bar{X}_{s+\frac{1}{2}})-\nabla f(\tilde{X}_s)\|_*^2}}\right)\\ &\le \frac{3D}{2}+\frac{3D}{2}\sum_{t=1}^{\tau^*}\frac{\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2}{\sqrt{1+\sum_{s=1}^t\alpha_s^2\|\nabla f(\bar{X}_{s+\frac{1}{2}})-\nabla f(\tilde{X}_s)\|_*^2}}\\ &\le 3D\sqrt{1+\sum_{t=1}^{\tau^*}\alpha_s^2\|\nabla f(\bar{X}_{s+\frac{1}{2}})-\nabla f(\tilde{X}_s)\|_*^2}\\ &= \frac{6D^2}{\gamma_{\tau^*+1}}\\ &\le 6\sqrt{7}D^2L.\end{aligned}$$

Final Bound. What remains is to simply bring the term (A) and (B) together.

$$\begin{aligned}\text{Reg}_T(x^*) &\le \frac{1}{2}\sum_{t=1}^{\tau^*}\gamma_{t+1}\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2+\frac{D}{2}+\frac{3D^2}{\gamma_{T+1}}-\frac{3}{4}\sum_{t=\tau^*+1}^T\gamma_{t+1}\alpha_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\nabla f(\tilde{X}_t)\|_*^2\\ &\le 10\sqrt{7}D^2L.\end{aligned}$$

We conclude the proof by applying Lemma 2.1.1 and get $f(\bar{X}_{T+\frac{1}{2}})-f(x^*)\le\frac{20\sqrt{7}D^2L}{T^2}$. ■

Finally, we are at a position to present convergence of Algorithm 2 in the smooth convex minimization setting under stochastic oracle information.

Theorem 2.1.4. *Consider the optimization setting in Problem (Prob), where f is L-smooth and convex. Let $\{X_{t+\frac{1}{2}}\}_{t=1,..,T}$ be a sequence generated by Algorithm 2 such that $g_t=\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}})$*

and $M_t = \nabla f(\tilde{X}_t, \xi_t)$. *With* $\alpha_t = t$, $x^* = \arg\min_{x \in \mathscr{X}} f(x)$ *and learning rate as in (2.7), it holds that*

$$\mathbb{E}\left[f(\bar{X}_{T+\frac{1}{2}})\right] - f(x^*) \leq \frac{224\sqrt{14}D^2L}{T^2} + \frac{14\sqrt{2}\sigma D}{\sqrt{T}} = O\left(\frac{LD^2}{T^2} + \frac{\sigma D}{\sqrt{T}}\right).$$

Proof. We start out with weighted regret, the same way as in Theorem 2.1.2

$$\begin{aligned}
&\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle \\
&\quad \leq \underbrace{\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle}_{\text{(A)}} + \underbrace{\sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle}_{\text{(B)}}.
\end{aligned}$$

We already know that term (B) is zero in expectation conditioned on $\sigma(X_{t+\frac{1}{2}}) = \sigma(\xi_1, \xi_{1+1/2}, ..., \xi_{t-\frac{1}{2}}, \xi_t)$. Following the proof steps of Theorem 2.1.2, we could upper bound term (A) as

$$\begin{aligned}
&\leq \frac{1}{2}\sum_{t=1}^{T} \gamma_{t+1}\alpha_t^2 \|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2 - \frac{1}{2}\sum_{t=1}^{T} \frac{1}{\gamma_{t+1}} \|X_{t+\frac{1}{2}} - X_t\|^2 + D^2\left(\frac{3}{\gamma_{T+1}} + \frac{1}{\gamma_1}\right) \\
&= \frac{D}{2} + D\sum_{t=1}^{T} \frac{\alpha_t^2 \|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2}{\sqrt{1 + \sum_{s=1}^{t} \alpha_s^2 \|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2}} \\
&\quad + \frac{3D}{2}\sqrt{1 + \sum_{t=1}^{T} \alpha_t^2 \|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2} - \sum_{t=1}^{T} \frac{\|X_{t+\frac{1}{2}} - X_t\|^2}{2\gamma_{t+1}} \\
&\leq \frac{7D}{2}\sqrt{1 + \sum_{t=1}^{T} \alpha_t^2 \|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2} - \frac{1}{2}\sum_{t=1}^{T} \frac{1}{\gamma_{t+1}} \|X_{t+\frac{1}{2}} - X_t\|^2.
\end{aligned}$$

Now lets denote,

$$B_t^2 := \min\{\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2, \|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2\},$$

as well as an auxiliary learning rate which we will only use for the analysis

$$\eta_t = \frac{2D}{\sqrt{1 + \sum_{s=1}^{t-1} \alpha_s^2 B_s^2}}. \tag{2.23}$$

Clearly, for any $t \in [T]$ we have $1/\eta_t \leq 1/\gamma_t$, and therefore,

$$-\frac{1}{\gamma_{t+1}} \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2 \leq -\frac{1}{\eta_{t+1}} B_t^2. \tag{2.24}$$

Recall the bounded variance assumption in Eq. (2.5) and let us define

$$\epsilon_{t+\frac{1}{2}} = \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\bar{X}_{t+\frac{1}{2}}) \text{ and } \epsilon_t = \nabla f(\tilde{X}_t, \xi_t) - \nabla f(\tilde{X}_t).$$

Hence, we use the shorthand notation,

$$\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t) = \nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t) + \epsilon_{t+\frac{1}{2}} - \epsilon_t.$$

Now, let us quantify the variance bound with respect to $\epsilon_{t+\frac{1}{2}}$ and ϵ_t.

$$\mathbb{E}\left[\|\epsilon_{t+\frac{1}{2}} - \epsilon_t\|_*^2\right] \le \mathbb{E}\left[2\|\epsilon_{t+\frac{1}{2}}\|^2 + 2\|\epsilon_t\|^2\right] \tag{2.25}$$

$$= \mathbb{E}\left[2\mathbb{E}\left[\|\epsilon_{t+\frac{1}{2}}\|^2 \mid \sigma(X_{t+\frac{1}{2}})\right] + 2\mathbb{E}\left[\|\epsilon_t\|^2 \mid \sigma(X_t)\right]\right] \le 4\sigma^2. \tag{2.26}$$

Using the above definitions we can write,

$$\|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2 \le 2\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2 + 2\|\epsilon_{t+\frac{1}{2}} - \epsilon_t\|_*^2. \tag{2.27}$$

Thus,

$$\begin{aligned}
&\|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2 \\
&= B_t^2 + \left(\|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2 - \min\{\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2, \|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2\}\right) \\
&= B_t^2 + \max\{0, \|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2 - \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2\} \\
&\le 2B_t^2 + 2\|\epsilon_{t+\frac{1}{2}} - \epsilon_t\|_*^2.
\end{aligned}$$

To explain how we obtain the last inequality, consider the following,

$$\|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2 \ge \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2 \implies B_t^2 = \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2.$$

Also, Eq. (2.27) implies that,

$$\|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2 - \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2 \le B_t^2 + 2\|\epsilon_{t+\frac{1}{2}} - \epsilon_t\|_*^2,$$

and we obtain the last line. In the other case with $\|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2 \le \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)\|_*^2$, we have $B_t = \|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)\|_*^2$ and the last line hold with strict inequality. We will take conditional expectation after we simplify the expression. Now, we plug Eq. (2.24) and (2.27) into above bound,

$$\begin{aligned}
&\le \frac{7D}{2}\sqrt{1 + 2\sum_{t=1}^{T}\alpha_t^2 B_t^2 + \alpha_t^2\|\epsilon_{t+\frac{1}{2}} - \epsilon_t\|_*^2} - \frac{1}{2}\sum_{t=1}^{T}\frac{1}{4L^2\eta_{t+1}}\alpha_t^2 B_t^2 \\
&\le \frac{7D}{\sqrt{2}}\sqrt{\sum_{t=1}^{T}\alpha_t^2\|\epsilon_{t+\frac{1}{2}} - \epsilon_t\|_*^2} + \frac{7D}{2}\sqrt{1 + 2\sum_{t=1}^{T}\alpha_t^2 B_t^2} - \frac{1}{2}\sum_{t=1}^{T}\frac{1}{4L^2\eta_{t+1}}\alpha_t^2 B_t^2 \\
&\le \frac{7D}{2} + \frac{7D}{\sqrt{2}}\sqrt{\sum_{t=1}^{T}\alpha_t^2\|\epsilon_{t+\frac{1}{2}} - \epsilon_t\|_*^2} + 7D\sum_{t=1}^{T}\frac{\alpha_t^2 B_t^2}{\sqrt{1 + 2\sum_{s=1}^{t}\alpha_s^2 B_s^2}} - \frac{1}{2}\sum_{t=1}^{T}\frac{1}{4L^2\eta_{t+1}}\alpha_t^2 B_t^2 \quad \text{(Lem 2.1.2)}
\end{aligned}$$

$$\leq \frac{7D}{2} + \frac{7D}{\sqrt{2}}\sqrt{\sum_{t=1}^{T}\alpha_t^2\|\epsilon_{t+\frac{1}{2}}-\epsilon_t\|_*^2} + 7D\sum_{t=1}^{T}\frac{\alpha_t^2 B_t^2}{\sqrt{1+\sum_{s=1}^{t}\alpha_s^2 B_s^2}} - \frac{1}{2}\sum_{t=1}^{T}\frac{1}{4L^2\eta_{t+1}}\alpha_t^2 B_t^2$$

$$\leq \underbrace{\frac{7}{2}\sum_{t=1}^{T}\left(\eta_{t+1}-\frac{1}{28L^2\eta_{t+1}}\right)\alpha_t^2 B_t^2 + \frac{7D}{2}}_{\text{(A)}} + \underbrace{\frac{7D}{\sqrt{2}}\sqrt{\sum_{t=1}^{T}\alpha_t^2\|\epsilon_{t+\frac{1}{2}}-\epsilon_t\|_*^2}}_{\text{(B)}}.$$

Bounding (A). We will make use of the exact same approach as we did in Theorem 2.1.3, where we defined an auxiliary time variable τ^* to characterize the behavior of the learning rate.

Now, let us denote $\tau^* = \max\left\{t \in \{1,...,T\} : \frac{1}{\eta_{t+1}^2} \leq 56L^2\right\}$. It implies that

$$\eta_{t+1} - \frac{1}{28L^2\eta_{t+1}} \leq -\eta_{t+1}, \quad \forall t > \tau^*. \tag{2.28}$$

Then, we could proceed as

$$\begin{aligned}
\text{(A)} &= \frac{7}{2}\sum_{t=1}^{\tau^*}\left(\eta_{t+1}-\frac{1}{28L^2\eta_{t+1}}\right)\alpha_t^2 B_t^2 + \frac{7}{2}\sum_{t=\tau^*+1}^{T}\left(\eta_{t+1}-\frac{1}{28L^2\eta_{t+1}}\right)\alpha_t^2 B_t^2 + \frac{7D}{2} \\
&\leq \frac{7}{2}\sum_{t=1}^{\tau^*}\eta_{t+1}\alpha_t^2 B_t^2 - \frac{7}{2}\sum_{t=\tau^*+1}^{T}\eta_{t+1}\alpha_t^2 B_t^2 + \frac{7D}{2} \\
&\leq \frac{7}{2}\sum_{t=1}^{\tau^*}\eta_{t+1}\alpha_t^2 B_t^2 + \frac{7D}{2} \\
&= 7D\sum_{t=1}^{\tau^*}\frac{\alpha_t^2 B_t^2}{\sqrt{1+\sum_{s=1}^{t}\alpha_s^2 B_s^2}} + \frac{7D}{2} \\
&\leq 14D\sqrt{1+\sum_{t=1}^{\tau^*}\alpha_t^2 B_t^2} \\
&\leq \frac{28D^2}{\widetilde{\eta}_{\tau^*+1}} \\
&\leq 112\sqrt{14}D^2L.
\end{aligned}$$

Bounding (B). Following bounded variance definition in Eq. (2.5), we can write $\mathbb{E}[\|\epsilon_{t+\frac{1}{2}} - \epsilon_t\|_*^2] \leq 4\sigma^2$. After taking expected value of the whole expression,

$$\begin{aligned}
\mathbb{E}\left[\frac{7D}{\sqrt{2}}\sqrt{\sum_{t=1}^{T}\alpha_t^2\|\epsilon_{t+\frac{1}{2}}-\epsilon_t\|_*^2}\right] &\leq \frac{7D}{\sqrt{2}}\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t^2\|\epsilon_{t+\frac{1}{2}}-\epsilon_t\|_*^2\right]} \qquad \text{(Jensen's ineq.)} \\
&= \frac{7D}{\sqrt{2}}\sqrt{\sum_{t=1}^{T}\alpha_t^2\mathbb{E}\left[\|\epsilon_{t+\frac{1}{2}}-\epsilon_t\|_*^2\right]}
\end{aligned}$$

$$\leq \frac{7D}{\sqrt{2}}\sqrt{\sum_{t=1}^{T} 4\alpha_t^2\sigma^2} \qquad \text{(Eq. (2.25))}$$
$$\leq \frac{14D\sigma}{\sqrt{2}}\sqrt{T^3}$$
$$= \frac{14\sigma D T^{3/2}}{\sqrt{2}}.$$

Finally, we combine all these bounds together and feed them through Lemma 2.1.1 to obtain the final rate.

$$\mathbb{E}\left[f(\bar{X}_{T+\frac{1}{2}})\right] - f(x^*) \leq \frac{224\sqrt{14}D^2L}{T^2} + \frac{14\sqrt{2}\sigma D}{\sqrt{T}}.$$

■

2.3.2 Proofs for Section 2.2

This section is dedicated for the proofs of Lemma and Theorems in Section 2.2. We being with the generalized online-to-batch conversion scheme which connects the optimality gap $f(\bar{X}_{T+\frac{1}{2}}) - f(x^*)$ with the "weighted" regret $\mathrm{REG}_{\mathrm{T}}(x^*) = \sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle$. This is a generalization of Lemma 2.1.1 and the proof follows the same arguments.

Theorem 2.2.3. *Let* $\mathrm{REG}_{\mathrm{T}}(x^*)$ *denote the anytime regret for the decision sequence* $\{X_{t+\frac{1}{2}}\}_{t=1}^{T}$ *as in* (Reg)*, and define two sequences of non-decreasing weights* a_t *and* b_t *such that* $a_t, b_t \geq 1$*. As long as* a_t / b_t *is ensured to be non-increasing,*

$$f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \leq \frac{\mathrm{REG}_{\mathrm{T}}(x^*)}{a_T \frac{B_T}{b_T}}$$

Proof. First, recall the definition of the offline regret:

$$\mathrm{REG}_{\mathrm{T}}(x^*) = \sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^* \rangle$$

Devising our analysis in the spirit of [Cut19; Kav+19], we need to relate $X_{t+\frac{1}{2}}$ to the average iterate $\bar{X}_{t+\frac{1}{2}}$ in order to exploit the convexity of the objective function. Notice that we could write the iterate $X_{t+\frac{1}{2}}$ as the difference of consecutive *average* iterates,

$$a_t X_{t+\frac{1}{2}} = a_t \frac{B_t}{b_t} \bar{X}_{t+\frac{1}{2}} - a_t \frac{B_{t-1}}{b_t} \bar{X}_{t-\frac{1}{2}}. \tag{2.29}$$

Also, we could subsequently express $a_t x^* = a_t \frac{B_t}{b_t} x^* - a_t \frac{B_{t-1}}{b_t} x^*$. Combining them together,

$$\begin{aligned}
\mathrm{REG}_{\mathrm{T}}(x^*) &= \sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^* \rangle \\
&= \sum_{t=1}^{T} a_t \frac{B_t}{b_t} \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), \bar{X}_{t+\frac{1}{2}} - x^* \rangle - a_t \frac{B_{t-1}}{b_t} \langle \nabla f(\bar{X}_{t+\frac{1}{2}}, \bar{X}_{t-\frac{1}{2}} - x^* \rangle \\
&= \sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), \bar{X}_{t+\frac{1}{2}} - x^* \rangle + a_t \frac{B_{t-1}}{b_t} \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), \bar{X}_{t+\frac{1}{2}} - \bar{X}_{t-\frac{1}{2}} \rangle
\end{aligned}$$

where we added and subtracted $a_t \frac{B_{t-1}}{b_t} \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), \bar{X}_{t+\frac{1}{2}} \rangle$ to obtain the second equality. Having expressed both inner products in the form we want, we could apply convexity and telescope.

$$\begin{aligned}
&\sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^* \rangle \\
&\quad \geq \sum_{t=1}^{T} a_t \left(f(\bar{X}_{t+\frac{1}{2}}) - f(x^*) \right) + a_t \frac{B_{t-1}}{b_t} \left(f(\bar{X}_{t+\frac{1}{2}}) - f(\bar{X}_{t-\frac{1}{2}}) \right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{t=1}^{T} a_t \left(f(\bar{X}_{t+\frac{1}{2}}) - f(x^*) \right) + a_t \frac{B_{t-1}}{b_t} \left(f(\bar{X}_{t+\frac{1}{2}}) - f(x^*) \right) - a_t \frac{B_{t-1}}{b_t} \left(f(\bar{X}_{t-\frac{1}{2}}) - f(x^*) \right) \\
&= \sum_{t=1}^{T} a_t \frac{B_t}{b_t} \left(f(\bar{X}_{t+\frac{1}{2}}) - f(x^*) \right) - a_t \frac{B_{t-1}}{b_t} \left(f(\bar{X}_{t-\frac{1}{2}}) - f(x^*) \right) \\
&= a_T \frac{B_T}{b_T} \left(f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \right) - a_1 \frac{B_0}{b_1} \left(f(\bar{X}_{-1/2}) - f(x^*) \right) + \sum_{t=1}^{T-1} B_t \left(\frac{a_t}{b_t} - \frac{a_{t+1}}{b_{t+1}} \right) \left(f(\bar{X}_{t+\frac{1}{2}}) - f(x^*) \right)
\end{aligned}
$$

Setting $B_0 = 0$ eliminates the second term. To conclude the proof, we need to show that the summation term in the above expression is always non-negative. This is ensured when the sequence $\frac{a_t}{b_t}$ is monotonically non-increasing, which is specified in the theorem statement (and subsequently satisfied by the algorithms). Hence,

$$
\begin{aligned}
&\sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^* \rangle \\
&\qquad = a_T \frac{B_T}{b_T} \left(f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \right) + \sum_{t=1}^{T-1} B_t \left(\frac{a_t}{b_t} - \frac{a_{t+1}}{b_{t+1}} \right) \left(f(\bar{X}_{t+\frac{1}{2}}) - f(x^*) \right) \\
&\qquad \geq a_T \frac{B_T}{b_T} \left(f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \right).
\end{aligned}
$$

Rearranging the terms gives us the final result

$$
f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \leq \frac{\sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle}{a_T \frac{B_T}{b_T}} = \frac{\mathrm{REG}_T(x^*)}{a_T \frac{B_T}{b_T}}.
$$

■

As the next step, we will prove the template inequality in Proposition 2.2.1 in the case of stochastic oracles. This inequality will give us the main departure point for both Theorem 2.2.1 and Theorem 2.2.2. We will prove a corollary of the following result later on, specifically for the deterministic setup, which will follow the same steps as Proposition 2.2.1.

For ease of navigation, we present EXTRA-NEWTON once more.

EXTRA-NEWTON

Input: $X_1 \in \mathscr{X}$, $a_t = t^2$ and $A_t = \sum_{s=1}^{t} a_s$, $b_t = t^p$ $(p \geq 2)$ and $B_t = \sum_{s=1}^{t} b_s$, $\gamma > 0$, $\xi_t \sim$ i.i.d.

1: **for** $t = 1$ to T **do**

2: $\gamma_t = \dfrac{\gamma}{\sqrt{\beta_0 + \sum_{s=1}^{t-1} a_s^2 \|\nabla f(\bar{X}_{s+\frac{1}{2}}, \xi_{s+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{s+\frac{1}{2}}; \tilde{X}_s, \xi_s)\|^2}}$

3: $X_{t+\frac{1}{2}} = \arg\min_{x \in \mathscr{X}} \langle a_t \nabla f(\tilde{X}_t, \xi_t), x \rangle + \frac{a_t b_t}{2B_t} \langle \nabla^2 f(\tilde{X}_t, \xi_t)(x - X_t), x - X_t \rangle + \frac{1}{2\gamma_t} \|x - X_t\|^2$

4: $X_{t+1} = \arg\min_{x \in \mathscr{X}} \langle a_t \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), x \rangle + \frac{1}{2\gamma_t} \|x - X_t\|^2$

5: **end for**

Proposition 2.2.1. *Let $\{X_{t+\frac{1}{2}}\}_{t=1}^{T}$ be generated by Algorithm 3, run with a non-increasing step-size sequence γ_t and non-decreasing sequences of weights $a_t, b_t \geq 1$ such that a_t/b_t is also non-increasing. Then, the following guarantee holds:*

$$\mathbb{E}\mathrm{REG}_{\mathrm{T}}(x^*) \leq \frac{1}{2}\mathbb{E}\left[\frac{3D^2}{\gamma_{T+1}} + \sum_{t=1}^{T}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}};\tilde{X}_t,\xi_t)\|^2 - \frac{\|X_{t+\frac{1}{2}}-X_t\|^2}{\gamma_{t+1}}\right]$$

Proof. We take off from the optimality conditions associated with each update sequence for our explicit algorithm EXTRA-NEWTON (Algorithm 3). Optimality condition for $X_{t+\frac{1}{2}}$ implies for any $z_0 \in \mathscr{X}$,

$$\begin{aligned}
&\langle a_t\nabla f(\tilde{X}_t,\xi_t) + a_t\frac{b_t}{B_t}\nabla^2 f(\tilde{X}_t,\xi_t)(X_{t+\frac{1}{2}}-X_t), X_{t+\frac{1}{2}}-z_0\rangle \\
&= \langle a_t\nabla f(\tilde{X}_t,\xi_t) + a_t\nabla^2 f(\tilde{X}_t,\xi_t)(\bar{X}_{t+\frac{1}{2}}-\tilde{X}_t), X_{t+\frac{1}{2}}-z_0\rangle \\
&= \langle a_t\tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}};\tilde{X}_t,\xi_t), X_{t+\frac{1}{2}}-z_0\rangle \\
&\leq \frac{1}{\gamma_t}\langle X_{t+\frac{1}{2}}-X_t, z_0-X_{t+\frac{1}{2}}\rangle \\
&= \frac{1}{2\gamma_t}\left(\|X_t-z_0\|^2 - \|X_{t+\frac{1}{2}}-z_0\|^2 - \|X_{t+\frac{1}{2}}-X_t\|^2\right)
\end{aligned} \tag{2.30}$$

Similarly, optimality of X_{t+1} update yields for any $z_1 \in \mathscr{X}$,

$$\begin{aligned}
\langle a_t\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}}), X_{t+1}-z_1\rangle &\leq \frac{1}{2\gamma_t}\langle X_{t+1}-X_t, z_1-X_{t+1}\rangle \\
&= \frac{1}{2\gamma_t}\left(\|X_t-z_1\|^2 - \|X_{t+1}-z_1\|^2 - \|X_{t+1}-X_t\|^2\right)
\end{aligned} \tag{2.31}$$

First, we will set $z_1 = x^*$ to establish the telescoping summation over $\|X_t-x^*\|^2 - \|X_{t+1}-x^*\|^2$. Then, we will simply align the above expression with the regret as follows,

$$\begin{aligned}
&\langle a_t\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}}-x^\star\rangle \\
&\qquad = \langle a_t\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}}-X_{t+1}\rangle + \langle a_t\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}}), X_{t+1}-x^\star\rangle \\
&\qquad \leq \langle a_t\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}}-X_{t+1}\rangle \\
&\qquad\quad + \frac{1}{2\gamma_t}\left(\|X_t-x^\star\|^2 - \|X_{t+1}-x^\star\|^2 - \|X_{t+1}-X_t\|^2\right)
\end{aligned} \tag{2.32}$$

Now, observe that setting $z_0 = X_{t+1}$ in Eq. (2.30) and rearranging we have

$$\begin{aligned}
&-\frac{1}{2\gamma_t}\|X_{t+1}-X_t\|^2 \\
&\leq -\langle a_t\tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}};\tilde{X}_t,\xi_t), X_{t+\frac{1}{2}}-X_{t+1}\rangle - \frac{1}{2\gamma_t}\left(\|X_{t+\frac{1}{2}}-X_{t+1}\|^2 + \|X_{t+\frac{1}{2}}-X_t\|^2\right)
\end{aligned}$$

Plugging the above expression into Eq. (2.32) and summing over $t = 1, ..., T$, we will obtain,

$$\begin{aligned}
&\sum_{t=1}^{T} \langle a_t \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^\star \rangle \\
&\quad \leq \sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}}; \tilde{X}_t, \xi_t), X_{t+\frac{1}{2}} - X_{t+1} \rangle \\
&\qquad + \sum_{t=1}^{T} \frac{1}{2\gamma_t} \left(\|X_t - x^\star\|^2 - \|X_{t+1} - x^\star\|^2 - \|X_{t+\frac{1}{2}} - X_{t+1}\|^2 - \|X_{t+\frac{1}{2}} - X_t\|^2 \right)
\end{aligned}$$

First off, we bound the inner product term using Cauchy-Schwarz and a slight generalization of Young's inequality [RS13]

$$\begin{aligned}
&\sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}}; \tilde{X}_t, \xi_t), X_{t+\frac{1}{2}} - X_{t+1} \rangle \\
&\quad \leq \sum_{t=1}^{T} a_t \|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}}; \tilde{X}_t, \xi_t)\| \|X_{t+\frac{1}{2}} - X_{t+1}\| \\
&\quad \leq \frac{1}{2} \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}}; \tilde{X}_t, \xi_t)\|^2 + \frac{1}{\gamma_{t+1}} \|X_{t+\frac{1}{2}} - X_{t+1}\|^2.
\end{aligned}$$

We merge the expressions together,

$$\begin{aligned}
&\sum_{t=1}^{T} \langle a_t \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^\star \rangle \\
&\quad \leq \frac{1}{2} \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}}; \tilde{X}_t, \xi_t)\|^2 + \frac{1}{\gamma_{t+1}} \|X_{t+\frac{1}{2}} - X_{t+1}\|^2 \\
&\qquad + \sum_{t=1}^{T} \frac{1}{2\gamma_t} \left(\|X_t - x^\star\|^2 - \|X_{t+1} - x^\star\|^2 - \|X_{t+\frac{1}{2}} - X_{t+1}\|^2 - \|X_{t+\frac{1}{2}} - X_t\|^2 \right)
\end{aligned}$$

It is important that we invoke generalized Young's inequality with step-size at time $t+1$. Since the step-size lags one iteration behind, γ_t does not include $\|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}}; \tilde{X}_t, \xi_t)\|^2$ and this would pose some problems in the later stages of the proof. Hence, we add/subtract $\frac{1}{\gamma_{t+1}} \|X_{t+\frac{1}{2}} - X_t\|^2$ and regroup the terms,

$$\begin{aligned}
&\sum_{t=1}^{T} \langle a_t \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^\star \rangle \\
&\quad \leq \frac{1}{2} \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \|\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}}; \tilde{X}_t, \xi_t)\|^2 - \frac{1}{\gamma_{t+1}} \|X_{t+\frac{1}{2}} - X_t\|^2 \\
&\qquad + \frac{1}{2} \sum_{t=1}^{T} \left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t} \right) \left(\|X_{t+\frac{1}{2}} - X_{t+1}\|^2 + \|X_{t+\frac{1}{2}} - X_t\|^2 \right) \\
&\qquad + \frac{1}{2} \sum_{t=1}^{T} \frac{1}{\gamma_t} \left(\|X_t - x^\star\|^2 - \|X_{t+1} - x^\star\|^2 \right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{2}\sum_{t=1}^{T}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}})-\tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}};\tilde{X}_t,\xi_t)\|^2-\frac{1}{\gamma_{t+1}}\|X_{t+\frac{1}{2}}-X_t\|^2\\
&\quad+\frac{\|X_1-x^*\|^2}{2\gamma_1}+\frac{1}{2}\sum_{t=1}^{T-1}\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right)\|X_{t+1}-x^\star\|^2+D^2\sum_{t=1}^{T}\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right)\\
&\leq \frac{3D^2}{2\gamma_{T+1}}+\frac{1}{2}\sum_{t=1}^{T}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}})-\tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}};\tilde{X}_t,\xi_t)\|^2-\frac{1}{2}\sum_{t=1}^{T}\frac{1}{\gamma_{t+1}}\|X_{t+\frac{1}{2}}-X_t\|^2
\end{aligned}
$$

where we have rewritten the telescoping summation for $\|X_t-x^\star\|^2-\|X_{t+1}-x^\star\|^2$ and used that $D^2=\sup_{x,y\in\mathscr{X}}\|x-y\|^2$ (diameter of the constraint set) to obtain the second inequality. The final line follows from telescoping the summations, plugging in the diameter D and rearranging the resulting terms.

Now, what remains is to obtain the (expected) regret from $\sum_{t=1}^{T}\langle a_t\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}}),X_{t+\frac{1}{2}}-x^\star\rangle$. Recall the definitions of $\mathscr{F}_t=\sigma(\xi_1,\xi_{1+\frac{1}{2}},\cdots,\xi_t)$ and $\mathscr{F}_{t+\frac{1}{2}}=\sigma(\xi_1,\xi_{1+\frac{1}{2}},\cdots,\xi_t,\xi_{t+\frac{1}{2}})$ from Table 2.2. Taking expectation over all randomness,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\langle a_t\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}}),X_{t+\frac{1}{2}}-x^\star\rangle\right]\\
&=\mathbb{E}\left[\sum_{t=1}^{T}a_t\langle\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}})-\nabla f(\bar{X}_{t+\frac{1}{2}}),X_{t+\frac{1}{2}}-x^\star\rangle+a_t\langle\nabla f(\bar{X}_{t+\frac{1}{2}}),X_{t+\frac{1}{2}}-x^\star\rangle\right]\\
&=\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}\left[a_t\langle\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}})-\nabla f(\bar{X}_{t+\frac{1}{2}}),X_{t+\frac{1}{2}}-x^\star\rangle\mid\mathscr{F}_t\right]\right]\\
&\quad+\mathbb{E}\left[\sum_{t=1}^{T}a_t\langle\nabla f(\bar{X}_{t+\frac{1}{2}}),X_{t+\frac{1}{2}}-x^\star\rangle\right]\\
&=\mathbb{E}\left[\sum_{t=1}^{T}a_t\langle\mathbb{E}\left[\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}})\mid\mathscr{F}_t\right]-\nabla f(\bar{X}_{t+\frac{1}{2}}),X_{t+\frac{1}{2}}-x^\star\rangle\right]\\
&\quad+\mathbb{E}\left[\sum_{t=1}^{T}a_t\langle\nabla f(\bar{X}_{t+\frac{1}{2}}),X_{t+\frac{1}{2}}-x^\star\rangle\right]\\
&=\mathbb{E}\left[\sum_{t=1}^{T}a_t\langle\nabla f(\bar{X}_{t+\frac{1}{2}}),X_{t+\frac{1}{2}}-x^\star\rangle\right]
\end{aligned}
$$

We used towering property of expectation (equivalently total law of expectation) to have the second inequality, and the last line from the unbiasedness assumption of gradient oracles in Eq. (2.14) such that $\mathbb{E}\left[\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}})\mid\mathscr{F}_t\right]=\nabla f(\bar{X}_{t+\frac{1}{2}})$. Hence, we obtain that

$$
\begin{aligned}
\mathbb{E}\left[\mathrm{REG}_T(x^*)\right]&=\mathbb{E}\left[\sum_{t=1}^{T}a_t\langle\nabla f(\bar{X}_{t+\frac{1}{2}}),X_{t+\frac{1}{2}}-x^\star\rangle\right]\\
&=\mathbb{E}\left[\sum_{t=1}^{T}\langle a_t\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}}),X_{t+\frac{1}{2}}-x^\star\rangle\right],
\end{aligned}
$$

which concludes the target result,

$$\mathbb{E}\left[\,\mathrm{REG}_{\mathrm{T}}(x^*)\,\right] \leq \frac{1}{2}\mathbb{E}\left[\frac{3D^2}{\gamma_{T+1}} + \sum_{t=1}^{T}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}};\bar{X}_t,\xi_t)\|^2 - \frac{\|X_{t+\frac{1}{2}}-X_t\|^2}{\gamma_{t+1}}\right]$$

■

Theorem 2.2.2. *Let $\{X_{t+\frac{1}{2}}\}_{t=1}^{T}$ be a sequence generated by Algorithm 3, run with the adaptive step-size policy (2) and $a_t = t^2, b_t = t^p$ for $p \geq 2$. Assume that f satisfies (H-smooth), and that Assumptions (2.14) hold . Then, the following universal guarantee holds:*

$$\mathbb{E}\left[\,f(\bar{X}_{T+\frac{1}{2}})\,\right] - f(x^*) \leq O\left(\frac{\frac{D^2+\gamma^2}{\gamma}\sigma_g}{\sqrt{T}} + \frac{\frac{D^3+D\gamma^2}{\gamma}\sigma_H}{T^{3/2}} + \frac{\max\left\{L\frac{D^4+D\gamma^3}{\gamma}, \sqrt{\beta_0}\frac{D^2+\gamma^2}{\gamma}\right\}}{T^3}\right)$$

When $\gamma = D$, we obtain the target rate $O\left(\frac{D\sigma_g}{\sqrt{T}} + \frac{D^2\sigma_H}{T^{3/2}} + \frac{\max\left\{LD^3, \sqrt{\beta_0}D\right\}}{T^3}\right)$.

Proof. We take Proposition 2.2.1 as our departure point for the analysis. After proving an offline regret bound, we will use Theorem 2.2.3 to obtain the optimality gap from the regret bound. Recall the template regret bound,

$$\mathbb{E}\mathrm{REG}_{\mathrm{T}}(x^*) \leq \frac{1}{2}\mathbb{E}\left[\frac{3D^2}{\gamma_{T+1}} + \sum_{t=1}^{T}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}};\bar{X}_t,\xi_t)\|^2 - \frac{\|X_{t+\frac{1}{2}}-X_t\|^2}{\gamma_{t+1}}\right]$$

Now, we want to unify the first two terms through numerical inequalities. We will write the *second term in terms of the first term.* Due to Lemma 2.1.2, we can upper the bound second term as,

$$\begin{aligned}
&\frac{1}{2}\sum_{t=1}^{T}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}};\bar{X}_t,\xi_t)\|^2 \\
&\qquad = \frac{\gamma}{2}\sum_{t=1}^{T}\frac{a_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}};\bar{X}_t,\xi_t)\|^2}{\sqrt{\beta_0 + \sum_{s=1}^{t}a_s^2\|\nabla f(\bar{X}_{s+\frac{1}{2}},\xi_{s+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{s+\frac{1}{2}};\bar{X}_s,\xi_s)\|^2}} \\
&\qquad \leq \gamma\sqrt{\beta_0 + \sum_{t=1}^{T}a_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}};\bar{X}_t,\xi_t)\|^2} - \frac{\gamma}{2\sqrt{\beta_0}}
\end{aligned}$$

Plugging this back into the original expression gives us

$$\mathbb{E}\left[\,\mathrm{REG}_{\mathrm{T}}(x^*)\,\right] \leq \left(\frac{3D^2}{2\gamma} + \gamma\right)\sqrt{\beta_0 + \sum_{t=1}^{T}a_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_s) - \tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}};\bar{X}_t,\xi_t)\|^2} - \sum_{t=1}^{T}\frac{\|X_{t+\frac{1}{2}}-X_t\|^2}{2\gamma_{t+1}}$$

Next up, we will handle the negative term in the above expression. As we have discussed in the main text, the key for faster rates beyond $O(1/T^2)$ is understanding how to manipulate

the negative term in the above expression. A crucial part of our analysis is understanding the implications of second-order smoothness and how to unlock its potential. This next derivation will demonstrate how (H-smooth) allows for a more aggressive gradient weighting and in turn faster convergence rate implied by our generalized conversion technique. Next, we will relate the negative term to the positive terms using smoothness and primal averaging, similar to the approaches in [WA18; Kav+19].

$$\begin{aligned}
-\frac{\|X_{t+\frac{1}{2}} - X_t\|^2}{\gamma_{t+1}} &= -\frac{D^2}{D^2\gamma_{t+1}}\|X_{t+\frac{1}{2}} - X_t\|^2 \\
&\leq -\frac{1}{D^2\gamma_{t+1}}\|X_{t+\frac{1}{2}} - X_t\|^4 \\
&= -\frac{1}{D^2\gamma_{t+1}}\frac{B_t^4}{b_t^4}\|\frac{b_t}{B_t}X_{t+\frac{1}{2}} - \frac{b_t}{B_t}X_t\|^4 \\
&= -\frac{1}{D^2\gamma_{t+1}}\frac{B_t^4}{b_t^4}\|\frac{b_t X_{t+\frac{1}{2}} + \sum_{s=1}^{t-1} b_s X_{s+\frac{1}{2}}}{B_t} - \frac{b_t X_t + \sum_{s=1}^{t-1} b_s X_{s+\frac{1}{2}}}{B_t}\|^4 \\
&= -\frac{1}{D^2\gamma_{t+1}}c^4 t^4 \|\bar{X}_{t+\frac{1}{2}} - \tilde{X}_t\|^4 \\
&\leq -\frac{4c^4 t^4}{L^2 D^2 \gamma_{t+1}}\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \mathbf{F}(\bar{X}_{t+\frac{1}{2}}; \tilde{X}_t)\|^2
\end{aligned}$$

First, notice that for any sequence $b_t = O(t^p)$ with $p \geq 0$, we have $B_t = \sum_{s=1}^t b_s = O(t^{p+1})$, which implies $\frac{B_t}{b_t} \leq ct$, where $c > 0$ is an absolute constant depending on how b_t is defined. Then, we use the definitions of average sequences $\bar{X}_{t+\frac{1}{2}}$ and $\tilde{X}_t$ to go from $\|X_{t+\frac{1}{2}} - X_t\|^4$ to $\|\bar{X}_{t+\frac{1}{2}} - \tilde{X}_t\|^4$ to obtain equalities 3-5, and apply smoothness to obtain the last line. On a related note, we want to highlight the importance of optimistic weighted averaging that is central for obtaining the above expression. Since the averaged pairs $\bar{X}_{t+\frac{1}{2}}$ and $\tilde{X}_t$ differ by only the last element, we can seamlessly relate $\|X_{t+\frac{1}{2}} - X_t\|$ to $\|\bar{X}_{t+\frac{1}{2}} - \tilde{X}_t\|$.

Now, we are at a position to explain how we will go beyond $O(1/T^2)$ convergence rate, which fundamentally depends on the gradient weights a_t and jointly relies on our generalized online-to-batch conversion in Theorem 2.2.3. The negative term above is monotonically decreasing (increases in magnitude) which is essential to (partially) control the growth of remaining positive term. More specifically, one can notice that in order to align the summands of the positive and negative term, the algebra dictates that we need to select $a_t = O(t^2)$, which implies $b_t = \Omega(t^2)$. Notice that our averaging and weighting parameters grow at least $O(t)$ faster than the existing accelerated schemes for first-order smoothness, which grants the improved $O(1/T^3)$ rate. On the contrary, first-order smoothness would only allow t^2 factor in front of the norm, leading to the slower rate. This is what we have observed precisely in Section 2.1 with the conversion scheme in Lemma 2.1.1 and the choice of weighting factors in the Theorem 2.1.3 and 2.1.4.

Due to (margin-wise) space constraints, we will use a slightly more compact notation for certain expressions. Let us first define a shorthand notation for noise in gradient and Hessian

evaluations, respectively.

$$\begin{aligned}\epsilon_t &= [\nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t, \xi_t)] - [\nabla f(\bar{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)] \\ \delta_t &= \nabla^2 f(\tilde{X}_t, \xi_t) - \nabla^2 f(\tilde{X}_t)\end{aligned} \tag{2.33}$$

Then, we define following deterministic/stochastic placeholders:

$$\begin{aligned}\nabla_t &= \nabla f(\bar{X}_{t+\frac{1}{2}}) - \mathbf{F}(\bar{X}_{t+\frac{1}{2}}, \tilde{X}_t) \\ \widetilde{\nabla}_t &= \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{\mathbf{F}}(\bar{X}_{t+\frac{1}{2}}, \tilde{X}_t, \xi_t) = \nabla_t + \epsilon_t - \delta_t(\bar{X}_{t+\frac{1}{2}} - \tilde{X}_t)\end{aligned} \tag{2.34}$$

Setting $a_t = t^2$, combining all the terms and introducing the compact notation,

$$\begin{aligned}&\sum_{t=1}^{T} \langle a_t \nabla f(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^\star \rangle \\ &\leq \left(\frac{3D^2}{2\gamma} + \gamma\right) \sqrt{\beta_0 + \sum_{t=1}^{T} a_t^2 \|\widetilde{\nabla}_t\|^2 - \sum_{t=1}^{T} \frac{2c^4}{L^2 D^2 \gamma_{t+1}} a_t^2 \|\nabla_t\|^2}\end{aligned}$$

Now, we describe how to relate $\|\nabla_t\|^2$ and $\|\widetilde{\nabla}_t\|^2$ while treating the step-size γ_{t+1} accordingly. From the perspective of step-size, we need to find a relevant, if not matching, lower bound for $\|\nabla_t\|^2$ and $\|\widetilde{\nabla}_t\|^2$. Indeed, we follow the ideas presented in the proof of Theorem 2.1.4, and begin by (trivially) lower bounding both terms with the same expression,

$$\begin{aligned}\|\widetilde{\nabla}_t\|^2 &\geq \min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\} \\ \|\nabla_t\|^2 &\geq \min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}\end{aligned} \tag{2.35}$$

Now, we will decompose $\|\widetilde{\nabla}_t\|^2$ into $\|\nabla_t\|^2$ and the noise terms. Using the definitions in Eq. (2.33) and (2.34) and applying triangular inequality with quadratic expansion,

$$\|\widetilde{\nabla}_t\|^2 \leq 2\|\nabla_t\|^2 + 4\|\delta_t(\bar{X}_{t+\frac{1}{2}} - \tilde{X}_t)\|^2 + 4\|\epsilon_t\|^2 \tag{2.36}$$

We can also have the following trivial upper bound,

$$\begin{aligned}\|\widetilde{\nabla}_t\|^2 &\leq 2\|\widetilde{\nabla}_t\|^2 \\ &\leq 2\|\widetilde{\nabla}_t\|^2 + 4\|\delta_t(\bar{X}_{t+\frac{1}{2}} - \tilde{X}_t)\|^2 + 4\|\epsilon_t\|^2\end{aligned} \tag{2.37}$$

We simplify Eq. (2.36) and Eq. (2.37) using the same arguments in Theorem 2.1.4; if $\|\nabla_t\|^2 \leq \|\widetilde{\nabla}_t\|^2$, then Eq. (2.36) is tighter, otherwise Eq. (2.37) is tighter. Hence, we could select the minimum of $\|\nabla_t\|^2$ and $\|\widetilde{\nabla}_t\|$:

$$\|\widetilde{\nabla}_t\|^2 \leq 2\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\} + 4\|\delta_t(\bar{X}_{t+\frac{1}{2}} - \tilde{X}_t)\|^2 + 4\|\epsilon_t\|^2 \tag{2.38}$$

Using this intuition, we can construct a variable η_t that always upper bounds the step-size.

$$\eta_t = \frac{\gamma}{\sqrt{\beta_0 + \sum_{s=1}^{t-1} a_s^2 \min\left\{\|\widetilde{\nabla}_s\|^2, \|\nabla_s\|^2\right\}}} \tag{2.39}$$

It is immediate that $\gamma_t \leq \eta_t$. Essentially, we will replace the terms $\|\nabla_t\|^2$ and $\|\widetilde{\nabla}_t\|^2$ with $\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}$, $\|\delta_t(\bar{X}_{t+\frac{1}{2}} - \tilde{X}_t)\|^2$ and $\|\epsilon_t\|^2$.

$$\begin{aligned}
&\mathbb{E}\left[\,\mathrm{REG}_T(x^*)\,\right] \\
&\leq \mathbb{E}\left[\frac{3D^2+2\gamma^2}{2\gamma}\sqrt{\beta_0 + \sum_{t=1}^{T} a_t^2\|\widetilde{\nabla}_t\|^2} - \sum_{t=1}^{T}\frac{2c^4}{L^2D^2\gamma_{t+1}}a_t^2\|\nabla_t\|^2\right] \\
&\leq \mathbb{E}\left[\frac{3D^2+2\gamma^2}{2\gamma}\sqrt{\beta_0 + \sum_{t=1}^{T} 2a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\} + 4a_t^2\|\delta_t(\bar{X}_{t+\frac{1}{2}} - \tilde{X}_t)\|^2 + 4a_t^2\|\epsilon_t\|^2}\right. \\
&\quad \left. - \sum_{t=1}^{T}\frac{2c^4}{L^2D^2\eta_{t+1}}a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}\right] \\
&\leq \mathbb{E}\left[\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma}\sqrt{\beta_0 + \sum_{t=1}^{T} a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}} - \sum_{t=1}^{T}\frac{2c^4a_t^2}{L^2D^2\eta_{t+1}}\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}\right. \\
&\quad \left. + 2\left(\frac{3D^2}{2\gamma}+\gamma\right)\sqrt{\sum_{t=1}^{T} a_t^2\|\delta_t(\bar{X}_{t+\frac{1}{2}} - \tilde{X}_t)\|^2} + 2\left(\frac{3D^2}{2\gamma}+\gamma\right)\sqrt{\sum_{t=1}^{T} a_t^2\|\epsilon_t\|^2}\;\right] \\
&\leq \mathbb{E}\left[\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma^2}\left(\gamma\sqrt{\beta_0} + \sum_{t=1}^{T}\eta_{t+1}a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}\right) - \sum_{t=1}^{T}\frac{2c^4a_t^2}{L^2D^2\eta_{t+1}}\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}\right. \\
&\quad \left. + 2\left(\frac{3D^2}{2\gamma}+\gamma\right)\sqrt{\sum_{t=1}^{T} a_t^2\|\delta_t(\bar{X}_{t+\frac{1}{2}} - \tilde{X}_t)\|^2} + 2\left(\frac{3D^2}{2\gamma}+\gamma\right)\sqrt{\sum_{t=1}^{T} a_t^2\|\epsilon_t\|^2}\;\right] \\
&\leq \frac{3D^2+2\gamma^2}{\sqrt{2}\gamma}\sqrt{\beta_0} + \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma^2} - \frac{2c^4}{L^2D^2\eta_{t+1}^2}\right)\eta_{t+1}a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}\right. \\
&\quad \left. + 2\left(\frac{3D^2}{2\gamma}+\gamma\right)\sqrt{\sum_{t=1}^{T} a_t^2\|\delta_t(\bar{X}_{t+\frac{1}{2}} - \tilde{X}_t)\|^2} + 2\left(\frac{3D^2}{2\gamma}+\gamma\right)\sqrt{\sum_{t=1}^{T} a_t^2\|\epsilon_t\|^2}\;\right]
\end{aligned}$$

Next, we will simplify the first summation and eventually show that it has a finite, constant upper bound. First off, notice that $\left(\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma^2} - \frac{2c^4}{L^2D^2\eta_{t+1}^2}\right)$ is a decreasing quantity and we are interested in the time point at which it changes signs. Let us define,

$$T_0 = \max\left\{t \in \mathbb{Z} \mid \left(\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma^2} - \frac{2c^4}{L^2D^2\eta_{t+1}^2}\right) \geq 0\right\}.$$

This immediately implies that for any $t \leq T_0$,

$$\frac{1}{\eta_{t+1}} \leq \frac{LD\sqrt{3D^2+2\gamma^2}}{2^{3/4}\gamma c^2}. \tag{2.40}$$

There is a critical cut-off point for the possible values of T_0 depending on the value of β_0. When the initial step-size is small enough, i.e., β_0 is too large, then $T_0 < 0$. This occurs when $\beta_0 \geq \frac{L^2D^2(3D^2+2\eta^2)}{2^{3/2}\eta^2c^4}$, which implies,

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma^2} - \frac{2c^4}{L^2D^2\eta_{t+1}^2}\right)\eta_{t+1}a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}\right] \leq 0$$

We get the same bound when $T_0 = 0$. For any other value of T_0, i.e., $T_0 > 0$, observe that the way we define T_0 enables us to *upper bound* the summation up to T, with the summation up to T_0. Hence,

$$\begin{aligned}
&\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma}\sqrt{\beta_0} + \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma^2} - \frac{2c^4}{L^2D^2\eta_{t+1}^2}\right)\eta_{t+1}a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}\right] \\
&\quad\leq \frac{3D^2+2\gamma^2}{\sqrt{2}\gamma}\sqrt{\beta_0} + \mathbb{E}\left[\sum_{t=1}^{T_0}\left(\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma^2} - \frac{2c^4}{L^2D^2\eta_{t+1}^2}\right)\eta_{t+1}a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}\right] \\
&\quad\leq \frac{3D^2+2\gamma^2}{\sqrt{2}\gamma}\sqrt{\beta_0} + \frac{3D^2+2\gamma^2}{\sqrt{2}\gamma}\sum_{t=1}^{T_0}\frac{a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}}{\sqrt{\beta_0 + \sum_{s=1}^{t-1}a_s^2\min\left\{\|\widetilde{\nabla}_s\|^2, \|\nabla_s\|^2\right\}}} \\
&\quad\leq \frac{3\sqrt{2}D^2+2\sqrt{2}\gamma^2}{\gamma}\sqrt{\beta_0 + \sum_{t=1}^{T_0}a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}} \\
&\quad= \left(3\sqrt{2}D^2+2\sqrt{2}\gamma^2\right)\frac{1}{\lambda_{T_0+1}} \\
&\quad\leq \frac{LD\left(3D^2+2\gamma^2\right)^{3/2}}{2^{1/4}\gamma c^2}
\end{aligned}$$

To make sure we incorporate the effect of the initial step-size, we combine the bounds to get

$$\begin{aligned}
&\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma}\sqrt{\beta_0} + \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma^2} - \frac{2c^4}{L^2D^2\eta_{t+1}^2}\right)\eta_{t+1}a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}\right] \\
&\qquad\leq \frac{3D^2+2\eta^2}{2^{1/4}\eta}\max\left\{\frac{\sqrt{\beta_0}}{2^{1/4}}, \frac{LD\sqrt{3D^2+2\eta^2}}{c^2}\right\}
\end{aligned}$$

This gives us the constant part of the regret, which will lead to the $O(1/T^3)$ part of the convergence rate. Now, what remains is to handle the "stochasticity". We will bound the remaining stochastic terms with respect to the stochastic gradient and the stochastic Hessian. Plugging the expected regret in to the bound and combining all the expressions together,

$\mathbb{E}\left[\,\mathrm{REG}_T(x^*)\,\right]$

$$
\begin{aligned}
&\leq \frac{3D^2+2\gamma^2}{\gamma}\mathbb{E}\left[\sqrt{\sum_{t=1}^{T} a_t^2\|\delta_t(\bar{X}_{t+\frac{1}{2}}-\tilde{X}_t)\|^2}+\sqrt{\sum_{t=1}^{T} a_t^2\|\epsilon_t\|^2}\right]+\frac{3D^2+2\gamma^2}{2^{1/4}\gamma}\max\left\{\frac{\sqrt{\beta_0}}{2^{1/4}},\frac{LD\sqrt{3D^2+2\gamma^2}}{c^2}\right\}\\
&\leq \frac{3D^2+2\gamma^2}{2^{1/4}\gamma}\max\left\{\frac{\sqrt{\beta_0}}{2^{1/4}},\frac{LD\sqrt{3D^2+2\gamma^2}}{c^2}\right\}+\frac{3D^2+2\gamma^2}{\gamma}\left(\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[a_t^2\|\delta_t\|^2\|(\bar{X}_{t+\frac{1}{2}}-\tilde{X}_t)\|^2\right]}\right)\\
&\quad+\frac{3D^2+2\gamma^2}{\gamma}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[a_t^2[\|\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}})-\nabla f(\bar{X}_{t+\frac{1}{2}})\|^2+\|\nabla f(\tilde{X}_t,\xi_t)-\nabla f(\tilde{X}_t)\|^2]\right]}\\
&= \frac{3D^2+2\gamma^2}{2^{1/4}\gamma}\max\left\{\frac{\sqrt{\beta_0}}{2^{1/4}},\frac{LD\sqrt{3D^2+2\gamma^2}}{c^2}\right\}+\frac{3D^2+2\gamma^2}{\gamma}\sqrt{D^2\sum_{t=1}^{T}\mathbb{E}\left[a_t^2\frac{b_t^2}{B_t^2}\mathbb{E}\left[\|\delta_t\|^2\mid\mathscr{F}_t\right]\right]}\\
&\quad+\frac{3D^2+2\gamma^2}{\gamma}\sqrt{\sum_{t=1}^{T}a_t^2\mathbb{E}\left[\mathbb{E}\left[\|\nabla f(\bar{X}_{t+\frac{1}{2}},\xi_{t+\frac{1}{2}})-\nabla f(\bar{X}_{t+\frac{1}{2}})\|^2\mid\mathscr{F}_t\right]+\mathbb{E}\left[\|\nabla f(\tilde{X}_t,\xi_t)-\nabla f(\tilde{X}_t)\|^2\mid\mathscr{F}_{t-\frac{1}{2}}\right]\right]}\\
&\leq \frac{3D^2+2\gamma^2}{\gamma}\left(\sqrt{D^2\sigma_H^2\sum_{t=1}^{T}a_t^2\frac{b_t^2}{B_t^2}}+\sqrt{4\sigma_g^2\sum_{t=1}^{T}a_t^2}\right)+\frac{3D^2+2\gamma^2}{2^{1/4}\gamma}\max\left\{\frac{\sqrt{\beta_0}}{2^{1/4}},\frac{LD\sqrt{3D^2+2\gamma^2}}{c^2}\right\}\\
&\leq \frac{3D^2+2\gamma^2}{\gamma}\left(\sqrt{\frac{D^2\sigma_H^2}{c^2}\sum_{t=1}^{T}a_t}+2\sigma_g T^{5/2}\right)+\frac{3D^2+2\gamma^2}{2^{1/4}\gamma}\max\left\{\frac{\sqrt{\beta_0}}{2^{1/4}},\frac{LD\sqrt{3D^2+2\gamma^2}}{c^2}\right\}\\
&\leq \frac{3D^2+2\gamma^2}{2^{1/4}\gamma}\max\left\{\frac{\sqrt{\beta_0}}{2^{1/4}},\frac{LD\sqrt{3D^2+2\gamma^2}}{c^2}\right\}+\frac{3D^3+2D\gamma^2}{c\gamma}\sigma_H T^{3/2}+\frac{6D^2+4\gamma^2}{\gamma}\sigma_g T^{5/2}
\end{aligned}
$$

Before concluding the convergence proof, we would like to have a quick detour on the value of c. The value of c is roughly between $[1/p,1]$, where p is the exponent of the averaging weight, $b_t=t^p$. For instance, when we pick $b_t=t^2$, we have $t^3/3\leq B_t\leq t^3$; and when $b_t=t^3$, $t^4/4\leq B_t\leq t^4$. Hence, we can avoid its effect in the final bound. Running the above expression through Theorem 2.2.3 and taking the full expectation we obtain,

$$
\mathbb{E}\left[f(\bar{X}_{T+\frac{1}{2}})\right]-f(x^*)\leq O\left(\frac{\frac{D^2+\gamma^2}{\gamma}\sigma_g}{\sqrt{T}}+\frac{\frac{D^3+D\gamma^2}{\gamma}\sigma_H}{T^{3/2}}+\frac{\max\left\{L\frac{D^4+D\gamma^3}{\gamma},\sqrt{\beta_0}\frac{D^2+\gamma^2}{\gamma}\right\}}{T^3}\right)
$$

■

Having established the main components of our analysis in the more general case of stochastic minimization, we will provide the analysis of the implicit algorithm (Implicit) under deterministic oracles. To do so, we will first start with a corollary result based on Proposition 2.2.1 that essentially proves the same template inequality under deterministic oracle model. In fact, one could easily show that Proposition 2.2.1 holds exactly up to replacing stochastic evaluations $\nabla f(\cdot)$ and $\tilde{\mathbf{F}}(\cdot;\cdot)$ with $\nabla f(\cdot)$ and $\mathbf{F}(\cdot;\cdot)$. For completeness, we will formalize the aforementioned result in Proposition 2.3.1 which follows the same steps as the proof of Proposition 2.2.1.

Proposition 2.3.1. *Let $\{X_{t+\frac{1}{2}}\}_{t=1}^{T}$ be generated by (Implicit), run with a non-increasing step-*

size sequence γ_t and non-decreasing sequences of weights $a_t, b_t \geq 1$ such that a_t/b_t is also non-increasing. Then, the following guarantee holds:

$$\text{REG}_T(x^*) \leq \frac{1}{2}\left(\frac{3D^2}{\gamma_{T+1}} + \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \mathbf{F}(\bar{X}_{t+\frac{1}{2}}; \tilde{X}_t)\|^2 - \frac{\|X_{t+\frac{1}{2}} - X_t\|^2}{\gamma_{t+1}}\right).$$

Proof. The proof of this theorem is analogous to that of Proposition 2.2.1 in Section **??**, up to replacing the stochastic feedback with the deterministic oracle calls. ■

Theorem 2.2.1. *Let $\{X_{t+\frac{1}{2}}\}_{t=1}^T$ be a sequence generated by (Implicit), run with the adaptive step-size policy (2.19) where $a_t = t^2$, $b_t = t^3$. Assume that f satisfies (H-smooth) and denote the diameter of the set as D. Then, the following guarantee holds:*

$$f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \leq O\left(\frac{\max\left\{\sqrt{\beta_0}\frac{D^2}{\gamma}, L\frac{D^4 + D\gamma^3}{\gamma}\right\}}{T^3}\right)$$

When $\gamma = D$, we obtain the converge rate $O\left(\frac{\max\left\{LD^3, \sqrt{\beta_0}D\right\}}{T^3}\right)$.

Proof. We will initiate our proof at template regret inequality as we proved in Proposition 2.3.1. Our overall strategy is straightforward; we first prove a constant upper bound for the offline weighted regret, then make use of the conversion result in Theorem 2.2.3 to obtain a convergence rate of order $O(1/T^3)$.

Due to Proposition 2.3.1 we have,

$$\text{REG}_T(x^*) \leq \frac{1}{2}\left(\frac{3D^2}{\gamma_{T+1}} + \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \mathbf{F}(\bar{X}_{t+\frac{1}{2}}; \tilde{X}_t)\|^2 - \frac{\|X_{t+\frac{1}{2}} - X_t\|^2}{\gamma_{t+1}}\right)$$

We will merge the first two terms and express the first term in the form of the second one using Lemma 2.1.2. Observe that for the proof of Theorem 2.2.2, we did the opposite and converted the summation into the form of the first term, $\frac{3D^2}{2\gamma}$.

$$\begin{aligned}
&\text{REG}_T(x^*) \\
&\leq \frac{3D^2}{2\gamma}\sqrt{\beta_0 + \sum_{t=1}^{T} a_t^2 \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \mathbf{F}(\bar{X}_{t+\frac{1}{2}}; \tilde{X}_t)\|^2} \\
&\quad + \frac{1}{2}\sum_{t=1}^{T} \gamma_{t+1} a_t^2 \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \mathbf{F}(\bar{X}_{t+\frac{1}{2}}; \tilde{X}_t)\|^2 - \frac{\|X_{t+\frac{1}{2}} - X_t\|^2}{\gamma_{t+1}} \\
&\leq \frac{3D^2\sqrt{\beta_0}}{2\gamma} + \frac{3D^2}{2\gamma}\sum_{t=1}^{T} \frac{a_t^2 \|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \mathbf{F}(\bar{X}_{t+\frac{1}{2}}; \tilde{X}_t)\|^2}{\sqrt{\beta_0 + \sum_{s=1}^{t} a_s^2 \|\nabla f(\bar{X}_{s+\frac{1}{2}}) - \mathbf{F}(\bar{X}_{s+\frac{1}{2}}; \tilde{X}_s)\|^2}}
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2}\sum_{t=1}^{T}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\mathbf{F}(\bar{X}_{t+\frac{1}{2}};\tilde{X}_t)\|^2-\frac{\|X_{t+\frac{1}{2}}-X_t\|^2}{\gamma_{t+1}}\\
&=\frac{3D^2\sqrt{\beta_0}}{2\gamma}+\frac{3D^2}{2\gamma^2}\sum_{t=1}^{T}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\mathbf{F}(\bar{X}_{t+\frac{1}{2}};\tilde{X}_t)\|^2\\
&+\frac{1}{2}\sum_{t=1}^{T}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\mathbf{F}(\bar{X}_{t+\frac{1}{2}};\tilde{X}_t)\|^2-\frac{\|X_{t+\frac{1}{2}}-X_t\|^2}{\gamma_{t+1}}\\
&=\frac{3D^2\sqrt{\beta_0}}{2\gamma}+\frac{1}{2}\sum_{t=1}^{T}\frac{3D^2+\gamma^2}{\gamma^2}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\mathbf{F}(\bar{X}_{t+\frac{1}{2}};\tilde{X}_t)\|^2-\frac{\|X_{t+\frac{1}{2}}-X_t\|^2}{\gamma_{t+1}},
\end{aligned}
$$

where we obtain the second inequality due to Lemma 2.1.2 and the last two lines follow from the definition of the step-size in Eq. (2.19) and appropriate regrouping. Similar to the proof in the explicit algorithm, we upper bound the negative term using appropriate averaging constants and smoothness.

$$
-\frac{\|X_{t+\frac{1}{2}}-X_t\|^2}{\gamma_{t+1}}\leq-\frac{4c^4}{L^2D^2\gamma_{t+1}}t^4\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\mathbf{F}(\bar{X}_{t+\frac{1}{2}};\tilde{X}_t)\|^2
$$

Setting $a_t=t^2$, plugging the bound on the negative term into the original expression we have,

$$
\leq\frac{3D^2\sqrt{\beta_0}}{2\gamma}+\frac{1}{2}\sum_{t=1}^{T}\left(\frac{3D^2+\gamma^2}{\gamma^2}-\frac{4c^4}{L^2D^2\gamma_{t+1}^2}\right)\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\mathbf{F}(\bar{X}_{t+\frac{1}{2}};\tilde{X}_t)\|^2 \tag{2.41}
$$

Our main objective is to show that the above summation is summable so we could show the constant upper bound for the offline regret, hence the acceleration. First off, notice that $\left(\frac{3D^2+\gamma^2}{\gamma^2}-\frac{4c^4}{L^2D^2\gamma_{t+1}^2}\right)$ is a non-increasing quantity and we are interested in the time point at which this quantity becomes negative. For that reason, we define the following time point,

$$
T_0=\max\left\{t\in\mathbb{Z}\mid\left(\frac{3D^2+\gamma^2}{\gamma^2}-\frac{4c^4}{L^2D^2\gamma_{t+1}^2}\right)\geq 0\right\}.
$$

This immediately implies that for any $t\leq T_0$,

$$
\frac{1}{\gamma_{t+1}}\leq\frac{LD\sqrt{3D^2+\gamma^2}}{2\gamma c^2}. \tag{2.42}
$$

To paint a complete picture, we would like to have a brief discussion on the possible values for T_0.

1. $T_0\leq 0$ implies that the step-size is small enough from the very beginning and that the summation term in Eq. (2.41) is always bounded by a constant, which immediately implies constant regret and $O(1/T^3)$ rate.
2. $T_0=\infty$ implies that the step-size is always lower bounded by the *inverse of the constant on the right-hand side* of Eq.(2.42). This is equivalent to saying $\sum_{t=1}^{\infty}a_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}})-\mathbf{F}(\bar{X}_{t+\frac{1}{2}};\tilde{X}_t)\|^2\leq$

C for some constant C, which in turn ensures that the summation in Eq. (2.41) is summable. Once again, we will have the constant regret and $O(1/T^3)$ rate.

3. When T_0 is a finite positive integer, we can upper bound the summation in Eq. (2.41) with the same summation up to iteration T_0. Note that it is not important whether T is larger or smaller than T_0, as the summands change sign and become negative after T_0.

Same as in the proof of EXTRA-NEWTON, we need to understand the effect of the initial step-size choice due to β_0. Imagine the case $\sqrt{\beta_0} \geq \frac{LD\sqrt{3D^2+\gamma^2}}{2\gamma c^2}$. This implies that $T_0 < 0$ and that the step-size is already small enough to make the summation negative from the first step onwards. In that scenario, the condition in Eq. (2.42) doesn't hold so we should consider the effect of this initial setup for the final bound. For the case when $T_0 > 0$, we can safely unify all the 3 cases above and simply upper bound the expression in Eq. (2.41) by rewriting the summation up to T_0. Therefore,

$$
\begin{aligned}
&\leq \frac{3D^2\sqrt{\beta_0}}{2\gamma} + \frac{1}{2}\sum_{t=1}^{T}\left(\frac{3D^2+\gamma^2}{\gamma^2} - \frac{4c^4}{L^2D^2\gamma_{t+1}^2}\right)\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \mathbf{F}(\bar{X}_{t+\frac{1}{2}};\tilde{X}_t)\|^2 \\
&\leq \frac{3D^2+\gamma^2}{2\gamma}\sqrt{\beta_0} + \frac{3D^2+\gamma^2}{2\gamma^2}\sum_{t=1}^{T_0}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \mathbf{F}(\bar{X}_{t+\frac{1}{2}};\tilde{X}_t)\|^2 \\
&= \frac{3D^2+\gamma^2}{2\gamma}\sqrt{\beta_0} + \frac{3D^2+\gamma^2}{2\gamma}\sum_{t=1}^{T_0}\frac{a_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \mathbf{F}(\bar{X}_{t+\frac{1}{2}};\tilde{X}_t)\|^2}{\sqrt{\beta_0 + \sum_{s=1}^{t}a_s^2\|\nabla f(\bar{X}_{s+\frac{1}{2}}) - \mathbf{F}(\bar{X}_{s+\frac{1}{2}};\tilde{X}_s)\|^2}} \\
&\leq \frac{3D^2+\gamma^2}{\gamma}\sqrt{\beta_0 + \sum_{t=1}^{T_0}a_t^2\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - \mathbf{F}(\bar{X}_{t+\frac{1}{2}};\tilde{X}_t)\|^2} \\
&= \left(3D^2+\gamma^2\right)\frac{1}{\gamma_{T_0+1}} \\
&\leq \frac{LD\left(3D^2+\gamma^2\right)^{3/2}}{2\gamma c^2}
\end{aligned}
$$

We combine the case for $T_0 < 0$ with the one above to established the constant regret bound

$$\mathrm{REG}_T(x^*) \leq O\left(\max\left\{\sqrt{\beta_0}\frac{D^2}{\gamma}, L\frac{D^4+D\gamma^3}{\gamma}\right\}\right)$$

Plugging this result in its place we obtain the convergence rate,

$$f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \leq O\left(\frac{\max\left\{\sqrt{\beta_0}\frac{D^2}{\gamma}, L\frac{D^4+D\gamma^3}{\gamma}\right\}}{T^3}\right)$$

■

3 Adaptive methods and variance reduction for smooth, non-convex optimization

3.1 High Probability Bounds for a Class of AdaGrad-type Non-convex Algorithms

3.1.1 Bibliographic Note

This section (Section 3.1) is based on the published work Kavis, Levy, and Cevher [KLC22], published in the ICLR 2022 conference.

Author list of the published work.

- Ali Kavis
- Kfir Y. Levy
- Volkan Cevher

Description of contributions. The candidate and Kfir Y. Levy jointly proved an earlier version of Theorem 3.1.2 for which Kfir Y. Levy proposed to use the alternative approach in Eq. (3.9) by assuming bounded function values. The candidate removed the restrictive bounded objective value assumption (due to Proposition 3.1.2) and obtained the final version of the results in Theorem 3.1.2. The candidate further extended the proof technique for Algorithm 5 (Theorem 3.1.4) and proved the noise-adaptive rates under sub-Gaussian noise model (Proposition 3.1.3 and Theorem 3.1.3). Numerical experiments are due to the candidate.

3.1.2 Introduction

In this section, we will focus on solving the following, simple minimization problem, which slightly generalizes the setting considered in the whole of Chapter 2,

$$\min_{x\in\mathbb{R}^d} f(x) := \mathbb{E}_{\xi\sim\mathscr{D}}\left[f(x,\xi)\right], \tag{P}$$

where the objective $f(x)$ is continuous and possibly non-convex, and $\mathscr{D}$ is a probability distribution from which the random vector ξ is drawn. Problem (P) captures, for instance, empirical risk minimization or finite-sum minimization [SB14] problems, where ξ represents the mini-batches and $\mathscr{D}$ corresponds to the distribution governing the data generation process or the sampling strategy.

Within the context of large-scale problems, including streaming data, computing full gradients is extremely costly, if not impossible. Hence, stochastic iterative methods are the main optimizer choice in these scenarios. The so-called adaptive methods such as AdaGrad [DHS11], Adam [KB15] and AmsGrad [RKK18a] have witnessed a surge of interest both theoretically and practically due to their off-the-shelf performance. For instance, adaptive optimization methods are known to show superior performance in various learning tasks such as machine translation [Zha+20; Vas+17].

From a theoretical point of view, existing literature provides a quite comprehensive understanding regarding the *expected* behaviour of existing stochastic optimization algorithm, including adaptive methods. Nevertheless, these results inherently cannot capture the behavior of the algorithms for a single run, which is related to the *probabilistic* nature of the optimization process. While there exists *high probability* analysis of vanilla SGD for non-convex problems [GL13], adaptive methods have received limited attention in this context.

Our main goal in the first part of this section is to understand the probabilistic convergence properties of adaptive algorithms, specifically AdaGrad, while focusing on their *problem parameter* adaptation capabilities in the non-convex setting. This result essentially complements the well-documented, expected convergence behavior of various adaptive algorithms. Compatible with Chapter 2, adaptivity refers to the ability of an algorithm to ensure convergence without requiring the knowledge of quantities such as smoothness modulus or variance of noise. Studies along this direction largely exist for the convex objectives [LYC18; Kav+19; Jou+20; AKC22]; for instance, Levy, Yurtsever, and Cevher [LYC18] shows that AdaGrad can (implicitly) exploit smoothness and adapt to the magnitude of noise in the gradients when $f(x)$ is convex in (P).

This particular perspective to adaptivity is crucial because most existing analysis, both for classical and adaptive methods, assume to have access to smoothness constant, bound on gradients [RKK18a] and even noise variance [GL13]. In practice, it is difficult, if not impossible, to compute or even estimate such quantities. For this purpose, in the setting of (P) we study a class of adaptive gradient methods that enable us to handle noisy gradient feedback without

requiring the knowledge of the objective's smoothness modulus, noise variance or a bound on gradient norms.

We summarize the contributions of this section as follows:

1. We provide a modular, simple high probability analysis for AdaGrad-type adaptive methods.
2. We present the first *optimal high probability convergence result* of the *original AdaGrad* algorithm for non-convex smooth problems. Concretely,
   (a) we analyze a fully adaptive step-size, oblivious to Lipschitz constant and noise variance,
   (b) we obtain the best known dependence of $\log(1/\delta)$ on the probability margin δ.
   (c) we show that under sub-Gaussian noise model, AdaGrad adapts to noise level with high probability, i.e, as variance $\sigma \to 0$, convergence rate improves, $1/\sqrt{T} \to 1/T$.
3. We present a new extension of AdaGrad that include averaging and momentum primitives, and prove similar high probability bounds for this framework, as well. Concretely, we study a general adaptive template which individually recovers AdaGrad and (adaptive) RSAG [GL16] for different parameter choices.

In the next section, we will provide a broad overview of related work with an emphasis on the recent developments. Section 3.1.4 formalizes the problem setting and states our blanket assumptions. Section 3.1.5 introduces the building blocks of our proposed proof technique while proving convergence results for AdaGrad. We generalize the convergence results of AdaGrad for a class of nonconvex, adaptive algorithms in Section 3.1.6.

3.1.3 Related Work

Adaptive methods for stochastic optimization and online learning As an extended version of the online (projected) GD [Zin03], AdaGrad [DHS11] is the pioneering work behind most of the contemporary adaptive optimization algorithms Adam, AmsGrad and RmsProp [TH12] to name a few. Simply put, adaptive methods compute step-sizes on-the-fly by accumulating gradient information and achieve data-dependent regret bounds as a function of gradient history [TP19; Ala+20; LXL19; HWD19]. Through standard online-to-offline conversions [CL06; Sha12], the resulting regret bounds imply order-optimal convergence rates of $O(1/\sqrt{T})$ when the environment is stochastic and the online losses are generated randomly, i.e., the environment is not adversarial.

Universality, adaptive methods and acceleration. We call an algorithm *universal* if it achieves optimal rates under different settings, simultaneously, without any modifications. For *convex* minimization problems, Levy, Yurtsever, and Cevher [LYC18] showed that AdaGrad attains a rate of $O(1/T + \sigma/\sqrt{T})$ by implicitly adapting to smoothness and noise levels; here T is the number of oracle queries and σ is the noise variance. They also proposed an acceler-

ated AdaGrad variant with scalar step-size. The latter result was extended for compactly constrained problems via accelerated Mirror-Prox algorithm [Kav+19], and for composite objectives [Jou+20]. Recently, Ene, Nguyen, and Vladu [ENV21] have further generalized the latter results by designing a novel adaptive, accelerated algorithm with per-coordinate step-sizes. Convergence properties of such algorithms under smooth, non-convex losses are unknown as the acceleration mechanism is not necessarily compatible with non-convex problems.

Adaptive methods for nonconvex optimization. Following the popularity of neural networks, adaptive methods have attracted massive attention due to their favorable performance in training and their ease of tuning. The literature is quite vast, which is impossible to cover exhaustively here [Che+19; Zah+18; LO19; Zou+19; Def+20; AMC21; Che+21; LKC21]. The majority of the existing results on adaptive methods for nonconvex problems focus on *in expectation* performance.

High probability results. The literature on high probability behavior of stochastic algorithms is relatively thin for first-order methods. For non-smooth optimization, Harvey et al. [Har+19] verifies that SGD, converges with the rates $O(\log(T)\log(1/\delta)/\sqrt{T})$ and $O(\log(T)\log(1/\delta)/T)$ for convex and strongly-convex problems, respectively, while Rakhlin, Shamir, and Sridharan [RSS12] proves a rate of $O(\log(\log(T)/\delta)/T)$ for strongly-convex objectives. Under smooth objectives satisfying Polyak-Lojasiewicz condition, Madden, Dall'Anese, and Becker [MDB21] proves a complementary $O(\sigma^2\log(1/\delta)/T)$ rate.

For the more relevant non-convex realm, Ghadimi and Lan [GL13] are the first to analyze probabilistic convergence of SGD and provide tight bounds. Nevertheless, their method requires prior knowledge of the smoothness modulus and noise variance. In the context of adaptive methods, Li and Orabona [LO20] considers delayed AdaGrad (with lag-one-behind step-size) for smooth, non-convex losses under sub-Gaussian noise and proved $O(\sigma\sqrt{\log(T/\delta)}/\sqrt{T})$ rate. Under similar conditions, Zhou et al. [Zho+18] proves convergence of order $O((\sigma^2\log(1/\delta))/T + 1/\sqrt{T})$ for AdaGrad. However, both works require the knowledge of smoothness to set the step-size. Moreover, Ward, Wu, and Bottou [WWB19] guarantees that AdaGrad with scalar step-size convergences at $O((1/\delta)\log(T)/\sqrt{T})$ rate with high probability. Although their framework is oblivious to smoothness constant, their dependence of probability margin is $1/\delta$.

More recently, under heavy-tailed noise having bounded p^{th} moment for $p \in (1,2)$, Cutkosky and Mehta [CM21] proves a rate of $O(\log(T/\delta)/T^{(p-1)/(3p-2)})$ for clipped normalized SGD with momentum; nevertheless their method requires the knowledge of (a bound on) the behavior of the heavy tails.

3.1.4 Setup and preliminaries

As we stated in the introduction, we consider the unconstrained minimization setting

$$\min_{x\in\mathbb{R}^d} f(x) := \mathbb{E}_{\xi\sim\mathscr{D}}\left[f(x,\xi)\right],$$

where the differentiable function $f:\mathbb{R}^d\to\mathbb{R}$ is a smooth and (possibly) non-convex function.

We are interested in finding a first-order ϵ-stationary output point satisfying $\|\nabla f(\hat{X})\|^2 \le \epsilon$, where $\|\cdot\|$ denotes the Euclidean norm for the sake of simplicity and $\hat{X}\in\mathbb{R}^d$ is a candidate solution. As the standard measure of convergence in the literature, we will quantify the performance of algorithms with respect to *average* gradient norm, $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|^2$. It immediately implies convergence in the *minimum* gradient norm across the whole of the execution, $\min_{t\in[T]}\|\nabla f(X_t)\|^2$. Moreover, the algorithm could compute a particular output $\hat{X}$ which is selected uniformly at random from the set of iterates generated by the algorithm $\{X_1,\dots,X_T\}$. Then, any bound on the average gradient norm, $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|^2$, ensures a convergence guarantee of the form $\mathbb{E}\left[\,\|\nabla f(\hat{X})\|^2\,\right]$, where the expectation is computed with respect to the randomness in oracle information and the *uniform* selection of the output point [GL13]. This is a notation we will use to simplify the presentation.

A function is called G-Lipschitz continuous if it satisfies

$$\left|f(x)-f(y)\right| \le G\|x-y\|, \quad \forall x,y\in\mathrm{dom}(f), \tag{3.1}$$

which immediately implies that

$$\|\nabla f(x)\| \le G, \quad \forall x\in\mathrm{dom}(f). \tag{3.2}$$

A differentiable function is called L-smooth if it has L-Lipschitz gradient

$$\|\nabla f(x)-\nabla f(y)\| \le L\|x-y\|, \quad \forall x,y\in\mathrm{dom}(\nabla f). \tag{3.3}$$

An equivalent characterization is also referred to as the "descent lemma" [WWB19; Bec17],

$$\left|f(x)-f(y)-\langle\nabla f(y),x-y\rangle\right| \le (L/2)\|x-y\|^2. \tag{3.4}$$

Assumptions on oracle model. We denote stochastic gradients with $\nabla f(x,\xi)$, for some random vector drawn from distribution $\xi\sim\mathscr{D}$. Since our template embraces single-call algorithms for this section of the manuscript, we might use the shorthand notation $\widetilde{\nabla} f(x) = \nabla f(x,\xi)$ for simplicity, at times. An oracle is called unbiased if

$$\mathbb{E}\left[\,\nabla f(x,\xi)|\sigma(x)\,\right] = \nabla f(x), \quad \forall x\in\mathrm{dom}(\nabla f). \tag{3.5}$$

Gradient estimates generated by a first-order oracle have bounded variance if they satisfy

$$\mathbb{E}\left[\|\nabla f(x,\xi)-\nabla f(x)\|^2|\sigma(x)\right]\leq\sigma^2,\quad\forall x\in\operatorname{dom}(\nabla f). \tag{3.6}$$

Finally, we assume that the stochastic gradients are bounded almost surely, i.e.,

$$\|\nabla f(x,\xi)\|\leq\tilde{G},\quad\forall x\in\operatorname{dom}(\nabla f). \tag{3.7}$$

Remark 3.1.1. Bounded variance assumption (3.16) is standard in the analysis of stochastic methods [Lan20] for non-convex problems with the general form as in (P). Similarly, for the analysis of adaptive methods in the nonconvex realm, it is common to assume bounded stochastic gradients (see [Zah+18; Zho+18; Che+19; LO20] and references therein). It is of independent interest to investigate to what extent these assumptions could be relaxed.

3.1.5 Method and Analysis

We will now introduce our proposed proof technique as well as our main theoretical results for AdaGrad with proof sketches and discussions on the key elements of our theoretical findings. We will present a high-level overview of our simplified, modular proof strategy while proving a complementary convergence result for AdaGrad under deterministic oracles. In the sequel, we refer to the name AdaGrad as the scalar step-size version (also known as AdaGrad-Norm) as presented in Algorithm 4.

Algorithm 4: AdaGrad

Input: time horizon T , $X_1\in\mathbb{R}^d$, step-size $\{\gamma_t\}_{t\in[T]}$, $G_0>0$

1: **for** $t=1$ to T **do**

2: Generate $\nabla f(X_t,\xi_t)$

3: $\gamma_t=\dfrac{1}{\sqrt{G_0^2+\sum_{s=1}^t\|\nabla f(X_s,\xi_s)\|^2}}$

4: $X_{t+1}=X_t-\gamma_t\nabla f(X_t,\xi_t)$

5: **end for**

Before moving forward with the analysis, let us first establish the notation we will use to simplify the presentation. In the sequel, we use $[T]$ as a shorthand expression for the set $\{1,2,...,T\}$. We will use $\Delta_t=f(X_t)-\min_{x\in\mathbb{R}^d}f(x)$ as a concise notation for objective sub-optimality and $\Delta_{\max}=\max_{t\in[T+1]}\Delta_t$ will denote the maximum over Δ_t.

Notice that AdaGrad (Alg. 4) does not require any prior knowledge regarding the smoothness modulus nor the noise variance. The main results in this section is Theorem 3.1.2, where we show that with high probability AdaGrad obtains the convergence rate $\tilde{O}(\log(1/\delta)/\sqrt{T})$ for finding an approximate stationary point. Moreover, Theorem 3.1.1 shows that in the determin-

istic case AdaGrad achieves the optimal rate of $O(1/T)$, thus establishing its universality.

Technical Lemmas

We make use of a few technical lemmas while proving our main results, which we refer to in our proof sketches. We present them all at once before the main theorems for completeness. We have previously introduced Lemma 2.1.2 in Section 2.1 which was crucial to bound the terms of the form $\sum_{t=1}^{T} \gamma_t \|\nabla f(X_t)\| \leq \frac{1}{\gamma_T}$ when the adaptive step-size is constructed in accordance with the algorithm at hand. Besides, we introduce a new numerical inequality3.1.1 which is another well-known result from online learning, essential for handling adaptive stepsizes. We restate Lemma 2.1.2 for ease of navigation.

Lemma 2.1.2. *Let $a_1, ..., a_n$ be a sequence of non-negative real numbers. Then, it holds that*

$$\sqrt{\sum_{i=1}^{n} a_i} \leq \sum_{i=1}^{n} \frac{a_i}{\sqrt{\sum_{k=1}^{i} a_k}} \leq 2\sqrt{\sum_{i=1}^{n} a_i}$$

Lemma 3.1.1. *Let $a_1, ..., a_n$ be a sequence of non-negative real numbers. Then, it holds that*

$$\sum_{i=1}^{n} \frac{a_i}{\sum_{k=1}^{i} a_i} \leq 1 + \log\left(1 + \sum_{i=1}^{n} a_i\right)$$

In order to achieve high probability bounds, we need to quantify the probabilistic behavior of the cumulative noise. The next lemma is the key for achieving this relationship; roughly speaking, it enables us to upper bound the cumulative noise via the square root of cumulative variance.

Lemma 3.1.2 (Lemma 3 in [KT08])**.** *Let X_t be a martingale difference sequence such that $|X_t| \leq b$. Let us also define*

$$\mathbf{Var}_{t-1}(X_t) = \mathbf{Var}\left(X_t \mid \sigma(X_1, ..., X_{t-1})\right) = \mathbb{E}\left[X_t^2 \mid \sigma(X_1, ..., X_{t-1})\right],$$

and $V_T = \sum_{t=1}^{T} \mathbf{Var}_{t-1}(X_t)$ as the sum of variances. For $\delta < 1/e$ and $T \geq 3$, it holds that

$$\mathbb{P}\left(\sum_{t=1}^{T} X_t > \max\left\{2\sqrt{V_T}, 3b\sqrt{\log(1/\delta)}\right\}\sqrt{\log(1/\delta)}\right) \leq 4\log(T)\delta \tag{3.8}$$

Overview of proposed analysis

We will start by presenting the individual steps of our proof and provide insight into its advantages. In the rest of this section, we solely focus on AdaGrad, however, the same intuition applies to the more general Algorithm 5 as we will make clear in the sequel. The classical

analysis begins with,

$$f(X_{t+1}) - f(X_t) \leq -\gamma_t \|\nabla f(X_t)\|^2 - \gamma_t \langle \nabla f(X_t), \zeta_t \rangle + \frac{L\gamma_t^2}{2} \|\nabla f(X_t, \xi_t)\|^2.$$

which is due to the smoothness property in Eq. (3.4), and we define $\zeta_t = \nabla f(X_t, \xi_t) - \nabla f(X_t)$ as the noise vector. Re-arranging and summing over $t \in [T]$ yields,

$$\sum_{t=1}^{T} \gamma_t \|\nabla f(X_t)\|^2 \leq f(X_1) - f(x^*) + \sum_{t=1}^{T} -\gamma_t \langle \nabla f(X_t), \zeta_t \rangle + \frac{L}{2} \sum_{t=1}^{T} \gamma_t^2 \|\nabla f(X_t, \xi_t)\|^2.$$

The main issue in this expression is the $\gamma_t \langle \nabla f(X_t), \zeta_t \rangle$ term, which creates measurability problems due to the fact hat γ_t and ζ_t are dependent random variables. On the left hand side, the mismatch between γ_t and $\|\nabla f(X_t)\|^2$ prohibits the use of technicals lemmas as we accumulate *stochastic* gradients for γ_t. Moreover, we cannot make use of Holder-type inequalities as we deal with high probability results. Instead, we divide both sides by γ_t, then sum over t and re-arrange to obtain a bound of the form,

$$\begin{aligned} \sum_{t=1}^{T} \|\nabla f(X_t)\|^2 &\leq \frac{\Delta_1}{\gamma_1} + \sum_{t=2}^{T} \left(\frac{1}{\gamma_t} - \frac{1}{\gamma_{t-1}} \right) \Delta_t + \sum_{t=1}^{T} -\langle \nabla f(X_t), \zeta_t \rangle + \frac{L}{2} \sum_{t=1}^{T} \gamma_t \|\nabla f(X_t, \xi_t)\|^2 \\ &\leq \frac{\Delta_{\max}}{\gamma_T} + \sum_{t=1}^{T} -\langle \nabla f(X_t), \zeta_t \rangle + \frac{L}{2} \sum_{t=1}^{T} \gamma_t \|\nabla f(X_t, \xi_t)\|^2 \end{aligned} \tag{3.9}$$

This modification solves the two aforementioned problems, but we now need to ensure boundedness of function values, specifically the maximum distance to the optimum, $\Delta_{\max}$. In fact, neural networks with bounded activations (e.g. sigmoid function) in the last layer and some objective functions in robust non-convex optimization (e.g. Welsch loss [Bar19]) satisfy bounded function values. However, this is a restrictive assumption to make for general smooth problems and we will *prove* that it is bounded or at least it grows no faster than $O(\log(T))$. As a key element of our approach, we show that it is the case for Algorithms 4 & 5. Now, we are at a position to state an overview of our proof:

1. Show that $\Delta_{\max} \leq O(\log(T))$ with high probability or $\Delta_{\max} \leq O(1)$ (deterministic).

2. Prove that $\sum_{t=1}^{T} -\langle \nabla f(X_t), \zeta_t \rangle \leq \tilde{O}(\sqrt{T})$ with high probability using Lemma 3.1.2.

3. Show that $\frac{L}{2} \sum_{t=1}^{T} \gamma_t \|\nabla f(X_t, \xi_t)\|^2 \leq O(\sqrt{T})$ by using Lemma 2.1.2.

For completeness, we will propose a simple proof for AdaGrad in the deterministic setting. This will showcase advantages of our approach, while providing some insight into the theoretical behavior of the algorithm. We provide a sketch of the proof, whose full version will be accessible in the appendix at the end of the chapter.

Theorem 3.1.1. *Let $\{X_t\}$ be a sequence generated by Algorithm 4 with $G_0 = 0$ for simplicity.*

Then, it holds that

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|^2 \leq O\left(\frac{(\Delta_1+L)^2}{T}\right).$$

Proof Sketch (Theorem 3.1.1). In the presence of only deterministic oracle, we have $\nabla f(X_t,\xi_t) = \nabla f(X_t)$ in Eq. (3.9). By replacing the stochastic gradient with the true gradient we obtain,

$$\sum_{t=1}^{T}\|\nabla f(X_t)\|^2 \;\leq\; \frac{\Delta_{\max}}{\gamma_T} + \frac{L}{2}\sum_{t=1}^{T}\gamma_t\|\nabla f(X_t)\|^2 \;\leq\; (\Delta_{\max}+L)\sqrt{\sum_{t=1}^{T}\|\nabla f(X_t)\|^2},$$

where we obtain the final inequality using Lemma 2.1.2. Now, we show that Δ_{T+1} is bounded for any T. Using the descent lemma and summing over $t \in [T]$,

$$f(X_{T+1}) - f(x^*) \leq f(X_1) - f(x^*) + \sum_{t=1}^{T}\left(\frac{L\gamma_t}{2}-1\right)\gamma_t\|\nabla f(X_t)\|^2.$$

Notice that the step-size is monotonically-decreasing, and we want to identify the time point at which we will have $\frac{L\gamma_t}{2}-1 \leq 0$ the terms in the summation on the RHS becomes negative. Now, define $t_0 = \max\left\{t \in [T] \mid \gamma_t > \frac{2}{L}\right\}$, such that $\left(\frac{L\gamma_t}{2}-1\right) \leq 0$ for any $t > t_0$. Then,

$$\begin{aligned} f(X_{T+1}) - f(x^*) &\leq \Delta_1 + \sum_{t=1}^{t_0}\left(\frac{L\gamma_t}{2}-1\right)\gamma_t\|\nabla f(X_t)\|^2 + \sum_{t=t_0+1}^{T}\left(\frac{L\gamma_t}{2}-1\right)\gamma_t\|\nabla f(X_t)\|^2 \\ &\leq \Delta_1 + \frac{L}{2}\sum_{t=1}^{t_0}\gamma_t^2\|\nabla f(X_t)\|^2 \leq \Delta_1 + \frac{L}{2}\left(1+\log\left(1+L^2/4\right)\right), \end{aligned}$$

where we use the definition of t_0 and Lemma 3.1.1 for the last inequality. Since this is true for any T, the bound holds for $\Delta_{\max}$, as well. Then, we will represent the resulting inequality as a quadratic inequality and optimize respectively. Defining $X = \sqrt{\sum_{t=1}^{T}\|\nabla f(X_t)\|^2}$, the original expression reduces to $X^2 \leq (\Delta_{\max}+L)\,X$. Solving for X, plugging in the bound for $\Delta_{\max}$ and dividing by T results in

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|^2 \leq \frac{\left(\Delta_1 + \frac{L}{2}\left(3+\log\left(L^2/4\right)\right)\right)^2}{T}.$$

■

Remark 3.1.2. To our knowledge, the most relevant analysis was provided by Ward, Wu, and Bottou [WWB19], which achieves $\tilde{\mathcal{O}}(\log(T)/T)$ convergence rate. Our new approach enables us to remove $\log(T)$ factor.

High probability convergence under stochastic oracle

Having introduced the building blocks, we will now present the high probability convergence bound for AdaGrad (Algorithm 4). Let us begin by the departure point of our proof, which is Eq. (3.9)

$$\sum_{t=1}^{T}\|\nabla f(X_t)\|^2 \leq \underbrace{\frac{\Delta_{\max}}{\gamma_T}}_{(*)} + \underbrace{\sum_{t=1}^{T}-\langle\nabla f(X_t),\zeta_t\rangle}_{(**)} + \underbrace{\frac{L}{2}\sum_{t=1}^{T}\gamma_t\|\nabla f(X_t,\xi_t)\|^2}_{(***)}. \tag{3.10}$$

We can readily bound expression $(***)$ using Lemma 2.1.2. Hence, what remains is to argue about high probability bounds for expressions $(*)$ and $(**)$, which we do in the following propositions.

Proposition 3.1.1. *Using Lemma 3.1.2, with probability* $1-4\log(T)\delta$ *and* $\delta < 1/e$*, we have*

$$\sum_{t=1}^{T}-\langle\nabla f(X_t),\zeta_t\rangle \leq 2\sigma\sqrt{\log(1/\delta)}\sqrt{\sum_{t=1}^{T}\|\nabla f(X_t)\|^2} + 3(G^2+G\tilde{G})\log(1/\delta).$$

The last ingredient of the analysis is the bound on Δ_t. The following proposition ensures a high probability bound of order $O(\log(t))$ on Δ_t under Algorithm 4.

Proposition 3.1.2. *Let* $\{X_t\}$ *be generated by AdaGrad for* $G_0 > 0$*. With probability at least* $1-4\log(t)\delta$*,*

$$\Delta_{t+1} \leq \Delta_1 + 2L\left(1+\log\left(\max\{1,G_0^2\}+\tilde{G}^2 t\right)\right) + G_0^{-1}(M_1+\sigma^2)\log(1/\delta) + M_2,$$

where $M_1 = 3(G^2+G\tilde{G})$ *and* $M_2 = G_0^{-1}(2G^2+G\tilde{G})$*.*

As an immediate corollary, since the statement of Proposition 3.1.2 holds for any t, it holds for $\Delta_{\max}$ by definition. Hence, we have that $\max_{t\in[T]}\Delta_t = \Delta_{\max} \leq O\left(\Delta_1 + L\log(T) + \sigma^2\log(1/\delta)\right)$ with high probability for any time horizon T. In the light of the above results, we are now able to present our high probability bound for AdaGrad.

Theorem 3.1.2. *Let* $\{X_t\}$ *be the sequence of iterates generated by AdaGrad. Under Assumptions 3.2, 3.6, 3.7, for* $\Delta_{max} \leq O\left(\Delta_1 + L\log(T) + \sigma^2\log(1/\delta)\right)$*, with probability at least* $1-8\log(T)\delta$*,*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|^2 \leq \frac{(\Delta_{max}+L)\,G_0 + 3(G^2+G\tilde{G})\log(1/\delta)}{T} + \frac{(\Delta_{max}+L)\,\tilde{G} + 2G\sigma\sqrt{\log(1/\delta)}}{\sqrt{T}}.$$

Proof Sketch (Theorem 3.1.2). By Eq. (3.10),

$$\sum_{t=1}^{T}\|\nabla f(X_t)\|^2 \leq \frac{\Delta_{\max}}{\gamma_T} + \sum_{t=1}^{T}-\langle\nabla f(X_t),\zeta_t\rangle + \frac{L}{2}\sum_{t=1}^{T}\gamma_t\|\nabla f(X_t,\xi_t)\|^2.$$

Invoking Lemma 2.1.2 on the last sum and using Proposition 3.1.1 for the second expression, we have with probability at least $1-4\log(T)\delta$

$$\sum_{t=1}^{T}\|\nabla f(X_t)\|^2 \leq (\Delta_{\max}+L)\sqrt{G_0^2+\sum_{t=1}^{T}\|\nabla f(X_t,\xi_t)\|^2}+2\sigma\sqrt{\log(1/\delta)}\sqrt{\sum_{t=1}^{T}\|\nabla f(X_t)\|^2}+3(G^2+G\tilde{G})\log(1/\delta)$$

Finally, we use the bounds on the gradient norms $\|\nabla f(X_t)\| \leq G$ and $\|\nabla f(X_t,\xi_t)\| \leq \tilde{G}$, rearrange the terms and divide both sides by T. Due to Proposition 3.1.2, with probability at least $1-8\log(T)\delta$,

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|^2 \leq \frac{(\Delta_{\max}+L)\,G_0+3(G^2+G\tilde{G})\log(1/\delta)}{T}+\frac{(\Delta_{\max}+L)\,\tilde{G}+2G\sigma\sqrt{\log(1/\delta)}}{\sqrt{T}},$$

where $\Delta_{\max} \leq O\big(\Delta_1+L\log(T)+\sigma^2\log(\frac{1}{\delta})\big)$. We keep $\Delta_{\max}$ in the bound due to lack of space.

■

Noise adaptation under sub-Gaussian noise model

To our knowledge, under the standard setting we consider (unbiased stochastic gradients with bounded variance), noise adaptation is not achieved for high probability convergence to first-order stationary points, specifically for AdaGrad-type adaptive methods. We call an algorithm *noise adaptive* if the convergence rate improves $1/\sqrt{T} \to 1/T$ as variance $\sigma \to 0$. Following the technical results and approach proposed by Li and Orabona [LO20], we will prove that high probability convergence of AdaGrad (Algorithm 4) exhibits adaptation to noise under sub-Gaussian noise model. First, we will introduce the additional assumption on the noise. We assume that the tails of the noise behaves as sub-Gaussian if,

$$\mathbb{E}\left[\exp(\|\nabla f(x,\xi)-\nabla f(x)\|^2) \mid \sigma(x)\right] \leq \exp(\sigma^2). \tag{3.11}$$

This last assumption on the noise is more restrictive than standard assumption of bounded variance. Indeed, Eq. (3.11) implies bounded variance (Eq. (3.16)), but the converse is not true. Finally, we conclude with the main theorems. We first present the compatible concentration inequality that we will use under sub-Gaussian noise (Lemma 3.1.3), which is due to Li and Orabona [LO20, Lemma 1]. Then, we establish a new high probability bound on $\Delta_{\max}$ before presenting the noise-adaptive rates for AdaGrad.

Lemma 3.1.3 (Lemma 1 from Li and Orabona [LO20])**.** *Let $Z_1,\cdots,Z_T$ be a martingale difference sequence (MDS) with respect to random vectors $\xi_1,\cdots,\xi_T$ and Y_t be a sequence of random variables which is $\sigma(\xi_1,\cdots,\xi_{t-1})$-measurable. Given that $\mathbb{E}\left[\exp(Z_t^2/Y_t^2) \mid \xi_1,\cdots\xi_{t-1}\right] \leq \exp(1)$, for any $\lambda>0$ and $\delta\in(0,1)$ with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} Z_t \leq \frac{3}{4}\lambda\sum_{t=1}^{T} Y_t^2+\frac{1}{\lambda}\log(1/\delta)$$

Let us explain the difference between the above concentration inequality and the one we used for the bounded variance setting in Lemma 3.1.2. The MDS Z_t defined in the above statement corresponds to the noise term (equivalent to X_t in Lemma 3.1.2), and observe that we do not have any almost-sure bound on Z_t sequence. This helps us remove the dependence on a bound on stochastic gradients in the final convergence rate in Theorem 3.1.3. Moreover, we are also able to achieve the noise adaptation by using the above Lemma thorugh sub-Gaussian tail assumption. Next, we prove the high probability boundedness of sub-optimality gap.

Proposition 3.1.3. *Let $\{X_t\}$ be generated by AdaGrad and define $\Delta_t = f(X_t) - \min_{x\in\mathbb{R}^d} f(x)$. Under sub-Gaussian noise assumption as in Eq. (3.11), with probability at least $1-3\delta$,*

$$\Delta_{t+1} \leq \Delta_1 + 3G_0^{-1}G^2 + 2G_0^{-1}\sigma^2 \log\left(\frac{et}{\delta}\right) + \frac{3}{4G_0}\sigma^2\log(1/\delta) + \frac{L}{2}\left(1+\log\left(\max\{1,G_0^2\}+2G^2t+2\sigma^2 t\log\left(\frac{et}{\delta}\right)\right)\right).$$

A fundamental difference between Proposition 3.1.2 and 3.1.3 is that the latter does not depend on a bound on the stochastic gradients, which is a direct consequence of the fact that concentration inequality in Lemma 3.1.3 does *not* assume boundedness of the MDS Z_t. This implies that we will *not* need an almost-sure bound on the noise vector. Combining the above results yields the noise-adaptive high probability convergence of AdaGrad.

Theorem 3.1.3. *Let $\{X_t\}$ be generated by AdaGrad and define $\Delta_t = f(X_t) - \min_{x\in\mathbb{R}^d} f(x)$. Under sub-Gaussian noise assumption as in Eq. (3.11) and considering high probability boundedness of Δ_{max} due to Proposition 3.1.3, with probability at least $1-5\delta$,*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|^2 \leq \frac{32\,(\Delta_{max}+L)^2 + 8\,(\Delta_{max}+L)\,(G_0+\sigma\sqrt{2\log(1/\delta)}) + 8\sigma^2\log(1/\delta)}{T} + \frac{8\sqrt{2}\,(\Delta_{max}+L)\,\sigma}{\sqrt{T}}.$$

Remark 3.1.3. By introducing the sub-Gaussian noise model, we manage to achieve a high probability convergence bound that is *adaptive to noise*, while *removing the dependence on a bound on stochastic gradients* in the final result. As explained previously, this noise assumption helps us use a more general concentration inequality (without boundedness assumption).

3.1.6 Generalized Method and Analysis

Having proven the high probability convergence for AdaGrad, we will now present an extension of our analysis to the more general accelerated gradient (AGD) template, which corresponds to a specific reformulation of Nesterov's acceleration [GL16].

Algorithm 5: Generic AGD Template

Input: Horizon T , $\tilde{X}_1 = X_1 \in \mathbb{R}^d$, $\alpha_t \in (0,1]$, step-sizes $\{\eta_t\}_{t\in[T]}$, $\{\gamma_t\}_{t\in[T]}$

1: **for** $t = 1$ to T **do**
2: $\quad \bar{X}_t = \alpha_t X_t + (1-\alpha_t)\tilde{X}_t$
3: $\quad$ Set $g_t = \nabla f(\bar{X}_t, \xi_t)$ (or $g_t = \nabla f(\bar{X}_t)$)
4: $\quad X_{t+1} = X_t - \eta_t g_t$
5: $\quad \tilde{X}_t = \bar{X}_t - \lambda_t g_t$
6: **end for**

Algorithm 5 is a small modification of Nesterov's optimal scheme for smooth convex minimization [Nes05]. A reformulation of the aforementioned optimal scheme was recently referred to as linear coupling [AO16], which is an intricate combination of mirror descent (MD), gradient descent (GD) and averaging. In the sequel, we focus on two aspects of the algorithm; averaging parameter α_t and selection of (adaptive) step-sizes η_t and λ_t. We could recover some well-known algorithms from this generic scheme depending on parameter choices, which we display in Table 3.1.

Our reason behind choosing this generic algorithm is two-fold. First, it helps us demonstrate flexibility of our simple, modular proof technique by extending it to a generalized algorithmic template. Second, as an integral element of this scheme, we want to investigate the notion of averaging (equivalently momentum [Def21]), which is an important primitive for machine learning and optimization problems. For instance, it is necessary for achieving *accelerated* convergence in convex optimization [Nes83a; Kav+19], while it helps improve performance in neural network training and stabilizes the effect of noise [Def21; Sut+13; LGY20]. We will analyze in what scenarios it plays in our favor and what are the possible limitations in terms of theoretical behavior of the algorithms.

Let us briefly introduce some instances of Algorithm 5, their properties and the corresponding parameter choices; specifically the averaging parameter α_t and step sizes η_t and λ_t. The averaging parameter α_t has two main forms: $\alpha_t = 2/(t+1)$ for weighted averaging and $\alpha_t = 1/t$ for uniform averaging. We take $\alpha_t = 2/(t+1)$ by default in our analysis, as it is a key element in achieving acceleration in the convex setting. Our convergence results could immediately be extended to uniform averaging, too, at the expense of an additional $\log(T)$ factor. Let us define the AdaGrad step size once more, which we use to define η_t and λ_t,

$$\gamma_t = \frac{1}{\sqrt{G_0^2 + \sum_{s=1}^t \|g_s\|^2}}, \qquad G_0 > 0. \tag{3.12}$$

The first instance of Algorithm 5 is the AdaGrad itself, i.e. $X_{t+1} = X_t - \gamma_t g_t$. Since $\tilde{X}_1 = X_1$ by initialization, we have $\bar{X}_1 = \tilde{X}_1 = X_1$. The fact that $\eta_t = \lambda_t = \gamma_t$ implies the equivalence $\tilde{X}_t = X_t$ for any $t \in [T]$, which ignores the averaging step. The second instance we obtain is AdaGrad

with averaging,

$$\begin{aligned} \bar{X}_t &= \alpha_t X_t + (1-\alpha_t)\bar{X}_{t-1} \\ X_{t+1} &= X_t - \alpha_t \gamma_t g_t, \end{aligned} \tag{3.13}$$

where we set $\eta_t = \gamma_t$ and $\lambda_t = 0$ in Algorithm 5. For the initialization $X_1 = \tilde{X}_1$, we can obtain by induction that $\tilde{X}_t = \bar{X}_{t-1}$, hence the scheme above. For the reasons we will make clear in the sequel, our analysis will not apply to plain AdaGrad with averaging. Introducing the gradients evaluated at averaged iterates into the algorithm requires more intricate design and analyzing convergence of the algorithm in Eq. (3.13) proves to be a challenge that we could not manage to do.

The final scheme we will analyze is (adaptive) RSAG algorithm proposed by Ghadimi and Lan [GL16], which keeps track of multiple sequences to handle gradients at averaged iterates. It selects a step size pair that satisfies $\lambda_t \approx (1+\alpha_t)\gamma_t$ and $\eta_t = \gamma_t$, generating a 3-sequence algorithm as in the original form of Algorithm 5.

Table 3.1: Example methods covered by the generic AGD template. We analyze the algorithms in **boldface**, and *italized* algorithms are not analyzed with our technique.

ALGORITHM	**WEIGHTS** (α_t)	**STEP-SIZE** (η_t,λ_t)
AdaGrad	N/A	$\eta_t = \gamma_t$, $\lambda_t = \gamma_t$
Adaptive RSAG[GL16]	$\alpha_t = \frac{2}{t+1}$	$\eta_t = \gamma_t$, $\lambda_t = (1+\alpha_t)\gamma_t$
AdaGrad w/ Averaging	$\alpha_t = \frac{2}{t+1}$ or $\frac{1}{t}$	$\eta_t = \alpha_t\gamma_t$, $\lambda_t = 0$
AcceleGrad[LYC18]	$\alpha_t = \frac{2}{t+1}$	$\eta_t \approx \frac{1}{\alpha_t}\gamma_t$, $\lambda_t \approx \gamma_t$

Before moving on to convergence results, we have an observation concerning time-scale difference between step sizes for the aforementioned algorithms. Precisely, λ_t is always *only* a constant factor away from η_t for AdaGrad and RSAG; as $t \to \infty$, $\eta_t \to \lambda_t$ and $\lim_{t\to\infty} \eta_t/\lambda_t = 1$. On the contrary, AdaGrad with averaging and AcceleGrad exhibits a different behavior; $\lim_{t\to\infty} \eta_t/\lambda_t \neq 1$. This phenomenon has an immediate connection to acceleration in the convex realm, and we will independently expand upon it at the end of this section.

Having defined instances of Algorithm 5, we will present high probability convergence rates for them. Similar to Eq. (3.10) for AdaGrad, we first define a departure point of similar structure in Proposition 3.1.4, then apply Proposition 3.1.1 and 3.1.2 in the same spirit as before to finalize the bound.

Following the notation in [GL16], let us define the following geometric sequence,

$$\Gamma_t = (1-\alpha_t)\Gamma_{t-1} \quad \text{where} \quad \Gamma_1 = 1.$$

Now, we can begin with the departure point of the proof, which is due to Ghadimi and Lan [GL16].

Proposition 3.1.4. *Let $\{X_t\}$ be generated by Algorithm 5. Then, it holds that*

$$\sum_{t=1}^{T} \|\nabla f(\bar{X}_t)\|^2 \leq \frac{\Delta_{max} + 2L}{\eta_T} + \frac{L}{2\eta_T} \sum_{t=1}^{T} \underbrace{\left[\sum_{k=t}^{T} (1-\alpha_k)\Gamma_k\right] \frac{\alpha_t}{\Gamma_t}}_{(*)} \frac{(\eta_t - \lambda_t)^2}{\alpha_t^2} \|\nabla f(\bar{X}_t, \xi_t)\|^2 + \underbrace{\sum_{t=1}^{T} -\langle \nabla f(\bar{X}_t), \zeta_t \rangle}_{(**)}.$$

Next, we will deliver the complementary bound on term $(*)$ in Proposition 3.1.4.

Proposition 3.1.5. *Using the recursive definition of Γ, we have*

$$\left[\sum_{k=t}^{T} (1-\alpha_k)\Gamma_k\right] \frac{\alpha_t}{\Gamma_t} \leq \begin{cases} 2 & \text{if } \alpha_t = \frac{2}{t+1}; \\ \log(T+1) & \text{if } \alpha_t = \frac{1}{t}. \end{cases}$$

Finally, we present the high probability convergence rates for Algorithm 5, specifically adaptive RSAG.

Theorem 3.1.4. *Let $\{X_t\}$ be the sequence generated by adaptive RSAG. Under Assumptions 3.2, 3.6, 3.7, with probability $1 - 8\log(T)\delta$,*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(\bar{X}_t)\|^2 \leq \frac{G_0(\Delta_{max} + 3L + L\log(\max\{1, G_0^{-2}\} + \tilde{G}^2 T)) + 3(G^2 + G\tilde{G})\log(1/\delta))}{T} + \frac{\tilde{G}(\Delta_{max} + 3L + L\log(\max\{1, G_0^{-2}\} + \tilde{G}^2 T)) + 2G\sigma\sqrt{\log(1/\delta)}}{\sqrt{T}},$$

where $\Delta_{max} \leq O\big(\Delta_1 + L\log(T) + \sigma^2\log(1/\delta)\big)$.

this results is a mere generalization of the main result in the previous section (Theorem 3.1.2). The core challenge is to handle the additional error introduced due to the gradient computation at the averaged iterates. There are multiple mechanisms in place to make up for this extra error term, one of which is the correct choice of time-scale difference between the step-sizes.

Let us focus on the first summation in Proposition 3.1.4, and let us take $\alpha_t = 2/(t+1)$ so that the summation becomes $\frac{L}{\eta_T}\sum_{t=1}^{T} \frac{(\eta_t - \lambda_t)^2}{\alpha_t^2} \|\nabla f(\bar{X}_t, \xi_t)\|^2$. Recall that we recover AdaGrad itself when $\eta_t = \lambda_t$, which immediately eliminates this error term. In the other case of $\eta_t \neq \lambda_t$, we are required to eliminate $\frac{1}{\alpha_t^2} = O(t^2)$. RSAG eliminates it by selecting $\eta_t = \gamma_t$ and $\lambda_t = (1+\alpha_t)\gamma_t$. Finally, applying Lemma 3.1.1, the auxiliary error incurred by the averaged gradients amounts to $O(\log(T))$.

On the contrary, AdaGrad with averaging (Eq. (3.13)) selects $\eta_t = \alpha_t \gamma_t$ and $\lambda_t = 0$ and achieve $O(\log(T))$ error for the same summation term. However, the first term on the RHS of the inequality in Proposition 3.1.4 now has the form $\frac{\Delta_{\max}+2L}{\alpha_T \gamma_T}$, which now suffers the additional $\frac{1}{\alpha_T} = O(T)$ factor. This is the very reason that we cannot handle AdaGrad with averaging using our analysis. In fact, it is of independent interest to analyze the convergence of the scheme in Eq. (3.13) for general smooth problems.

A discussion on acceleration and nonconvex analysis. AGD and its variants are able to converge at the fast rate of $\mathcal{O}(1/T^2)$ [Nes03] for smooth, convex objectives. It has been established that the accelerated gradient template in Algorithm 5 can achieve the $O(1/T)$ rate for smooth, non-convex objectives under deterministic oracles, however, the set of parameters under which the Algorithm 5 is run for the non-convex setting is fundamentally different than that of the convex case [GL16]. In fact, the mechanism that allows the accelerated algorithms to converge faster could even prhibit convergence when convexity assumption is lifted. We will conclude this section with a brief discussion on this phenomenon.

As we mentioned previously, step-sizes for AdaGrad and RSAG have the same time-scale up to a constant factor. However, AcceleGrad, an accelerated algorithm, has a time-scale difference of $\mathcal{O}(t)$ between $\eta_t = \alpha_t^{-1}\gamma_t$ and $\lambda_t = \gamma_t$, such that it runs with a modified step-size

$$\gamma_t = \frac{1}{\sqrt{1 + \sum_{s=1}^{t} \alpha_s^{-2} \|g_s\|^2}}.$$

This scale difference is not possible to handle with standard approaches or our proposed analysis, to the best of our knowledge. Specifically, if we look at the second term in Proposition 3.1.4, it roughly evaluates to

$$\frac{L}{2\eta_T} \sum_{t=1}^{T} \frac{\eta_t^2}{\alpha_t^2} \|g_t\|^2 \leq \frac{L\alpha_T}{2\gamma_T} \sum_{t=1}^{T} \frac{\gamma_t^2}{\alpha_t^2} \alpha_t^{-2} \|g_t\|^2,$$

where each summand is $O(t^4)$ orders of magnitude larger compared to non-accelerated methods. A factor of $\alpha_t^{-2} = O(t^2)$ is absorbed by the modified step-size, but this term still grows faster than we can manage. Moreover, the scaling factor $\frac{1}{\alpha_T}$ in the step-size $\eta_T = \frac{1}{\alpha_T}\gamma_T$ amounts to an additional $O(T)$ error multiplying the summation. We aim to understand it further in our future work.

3.1.7 Conclusion

We propose a simple and modular high probability analysis for a class of AdaGrad-type algorithms. Bringing AdaGrad into the focus, we show that our new analysis techniques goes beyond and generalizes to the accelerated gradient template (Algorithm 5) which individually recovers AdaGrad and the adaptive version of RSAG [GL16]. By proposing a modification

over standard analysis and relying on concentration bounds for martingale difference sequences, we achieve high probability convergence bounds for the aforementioned algorithms *without* requiring the knowledge of smoothness L and variance σ while having best-known dependence of $\log(1/\delta)$ on δ. Futhermore, through a more refined notion of variance, i.e., assuming that noise has sub-Gaussian tail, we prove noise-adaptive high probability rates that interpolate between $1/T$ and $1/\sqrt{T}$.

3.2 Fully-adaptive SGD with Recursive Momentum for Non-convex Minimization

3.2.1 Bibliographic Note

This section (Section 3.2) is based on the published work Levy, Kavis, and Cevher [LKC21], published in the NeurIPS 2021 conference.

Author list of the published work .

- Kfir Y. Levy
- Ali Kavis
- Volkan Cevher

Description of contributions. The candidate worked on the earlier version of Algorithm 6 in which the step-size was defined as $\gamma_t = \frac{1}{(\sum_{s=1}^{t} \|d_s\|^2)^{1/3}}$ and proved noise-adaptive rates *without* parameter-free properties. Kfir Y. Levy identified the adaptive connection between step-size and the momentum and achieved the final parameter-free, noise-adaptive rates in Theorem 3.2.1. The candidate has complementary contributions via the auxiliary, intermediate results used in the proof of the main theorem. Numerical experiments are due to the candidate.

3.2.2 Introduction

Over the past decade non-convex models have attracted significant attention in machine learning (ML), and in data-science. This predominantly includes deep models, as well as phase retrieval [CLS15], non-negative matrix factorization [Hoy04], and matrix completion problems [GLM16] among others.

The main workhorse for training such machine learning models is SGD and its numerous variants. One parameter that has fundamental effect on the performance is the step-size, which often requires a careful and costly hyper-parameter tuning due to the complext problem lanscape. Adaptive approaches to setting the step-size proposed in AdaGrad [DHS11] and Adam [KB15], as well as non-adaptive heuristics [LH17; He+19] are very popular in modern ML applications, yet these methods also require some tuning of hyper-parameters like momentum and the scale of the step-size (initial step-size) schedule.

A popular SGD heuristic that has proven to be crucial in many applications is the use of *momentum*, i.e., the use of a weighted average of past gradients instead of only the current gradient [Sut+13; KB15]. Although adaptive approaches to setting the momentum have been investigated in the past [SSB18; HKS16], principled and theoretically-grounded approaches to doing so are less investigated. Another aspect that has not been extensively studied, which we

take into account in this section, is the interplay between step-size and momentum. To this end, we will explore momentum-based adaptive and parameter-free methods for stochastic non-convex optimization problems. Concretely, we focus on the setting where the objective is an *expectation over smooth losses* (see Eq. (3.17)), and the goal is to find an approximate stationary point.

In the general case of smooth non-convex objectives it is known that one can approach a stationary point at a rate of $O(1/T^{1/4})$, where T is the total number of samples [GL13]. While this rate is optimal in the general case, it is known that one can obtain an improved rate of $O(1/T^{1/3})$ if the objective is an *expectation over smooth losses* [Fan+18; ZXG18; CO19; Tra+19]. One could interpret this structure as a generalization of the finite-sum structure to the scenarios in which the data arrives at a streaming fashion. Besides, this rate was recently shown to be tight [Arj+19].

Nevertheless, most of the methods developed for this setting rely on variance reduction techniques [JZ13; ZMJ13; MZJ13; Wan+13], which require careful maintenance of anchor points in conjunction with appropriately selected large batch-sizes. This leads to a challenging hyper-parameter tuning problem, weakening their practicality. One exception is the recent STORM algorithm of [CO19].

STORM does not require large batches nor anchor points; instead, it uses a corrected momentum-based gradient update that leads to implicit variance reduction, which in turn facilitates fast convergence. Unfortunately, none of the aforementioned methods (including STORM) is parameter-free. Indeed, the knowledge of smoothness parameter together with either the noise variance or a bound on the norm of the gradients are crucial to establish their guarantees.

In this work, we essentially develop a parameter-free variant of STORM algorithm. We summarize our contributions as follows,

- We present STORM+, a *parameter-free* momentum-based method that ensures the optimal $O(1/T^{1/3})$ rate for the *expectation over smooth losses* setting. Similarly to STORM, our method requires neither large-batches nor anchor points.
- STORM+ implicitly adapts to the variance of the gradients. Concretely, it obtains the convergence rate of $O(1/\sqrt{T}+\sigma^{1/3}/T^{1/3})$, which recovers the optimal $O(1/\sqrt{T})$ rate in the noiseless case. We also improve over STORM by shaving off a $(\log T)^{3/4}$ factor from the $1/\sqrt{T}$ term at the expense of an assumption on the range of sub-optimality gap.
- We demonstrate a novel way to set the step-size by introducing an adaptive interplay between step-size and momentum parameters.

3.2.3 Related Work

In the context of stochastic non-convex optimization with general smooth losses, it was shown in Ghadimi and Lan [GL13] that SGD with an appropriately selected step-size can

obtain a rate of $O(1/T^{1/4})$ for finding an approximate stationary point, which is known to match the respective lower bound [Arj+19]. While the method of Ghadimi and Lan [GL13] requires knowledge of the smoothness and variance parameters, recent works have shown that adaptive methods like AdaGrad are able to obtain this bound in a parameter free manner, while adapting to the noise levels of the first-order oracles [LO19; WWB19; RKK18b]. These results, in a sense, explain the success of adaptive methods like AdaGrad [DHS11], Adam [KB15], and RMSProp [TH12] in handling non-convex problems.

The idea of using variance reduction techniques for non-convex problems was first suggested in the context of finite-sum problems [AH16b; Red+16], showing a rate of $O(1/T^{1/4})$. This was later improved [Lei+17] to a rate of $O(1/T^{3/10})$. The first works that have obtained the optimal $O(1/T^{1/3})$ for this setting were Fang et al. [Fan+18] and Zhou, Xu, and Gu [ZXG18]. Additionally, Fang et al. [Fan+18] shows that the same convergence behavior applies to the more general *expectation over smooth losses* setting (see Eq. (3.17)) – a setting that captures finite-sum problems as a private case.

The STORM algorithm suggested in Cutkosky and Orabona [CO19] is the first algorithm to obtain the optimal $O(1/T^{1/3})$ for this setting without the need to maintain anchor points and large batches. Instead, it relies on a clever correction of the momentum by making only one extra call to the oracle, which leads to an implicit variance reduction effect. Moreover, STORM adapts to the variance of the problem by obtaining a rate of $O((\log T)^{3/4}/\sqrt{T} + \sigma^{1/3}/T^{1/3})$ without any prior knowledge of variance parameter. However, it needs to know the smoothness and a bound on the gradient norms to set the step size and momentum parameters. Simultaneously to the work of [CO19], another paper [Tra+19] have obtained the same optimal bound by proposing a similar update rule. Note that [Tra+19] does calculate a single anchor point, and it still requires the knowledge of the smoothness and variance parameters.

3.2.4 Preliminaries

We discuss stochastic non-convex optimization problems where the objective $f : \mathbb{R}^d \mapsto \mathbb{R}$ is of the following form,

$$f(x) := \mathbb{E}_{\xi\sim\mathscr{D}}[f(x;\xi)]\,,$$

and $\mathscr{D}$ is an unknown distribution from which we may draw i.i.d. samples. Our goal is to find an approximate stationary point of f, i.e. after T draws from $\mathscr{D}$ we should output a point $\hat{x} \in \mathbb{R}^d$ such that $\mathbb{E}\|\nabla f(x)\| \le \text{Poly}(1/T)$.

We focus on first order methods, i.e., methods that may access the gradients of $f(\cdot,\xi)$, and make the following assumptions regarding the noisy gradients and function values.
Bounded values: There exists an absolute constant $B > 0$ such that,

$$\max_{x,y\in\mathbb{R}^d} |f(x) - f(y)| \le B. \tag{3.14}$$

Bounded gradients: There exists an absolute constant $G > 0$ such that,

$$\|\nabla f(x;\xi)\|^2 \leq G^2 \; ; \quad \forall x \in \mathbb{R}^d, \; \xi \in \text{supp}\{\mathscr{D}\}. \tag{3.15}$$

Bounded variance: There exists $\sigma > 0$ such that,

$$\mathbb{E}\left[\|\nabla f(x;\xi) - \nabla f(x)\|^2 \mid \sigma(\xi) \right] \leq \sigma^2 \; ; \quad \forall x \in \mathbb{R}^d. \tag{3.16}$$

Expectation over smooth losses: There exists $L > 0$ such that,

$$\|\nabla f(x;\xi) - \nabla f(y;\xi)\| \leq L\|x-y\| \; ; \quad \forall x, y \in \mathbb{R}^d, \; \xi \in \text{supp}\{\mathscr{D}\} \; . \tag{3.17}$$

The last assumption also implies that the expected loss $f(\cdot)$ is L-smooth. A property of smooth functions that we will exploit throughout the paper is the following,

$$f(y) \leq f(x) + \langle \nabla f(x), y-x \rangle + (L/2)\|y-x\|^2 \; ; \quad \forall x, y \in \mathbb{R}^d \tag{3.18}$$

In the rest of this manuscript, $\nabla f(x;\xi)$ relates to gradients with respect to x, i.e., $\nabla := \nabla_x$. We use $\|\cdot\|$ to denote the Euclidean norm, and x^* denotes a global minima of $f(\cdot)$, i.e., $x^* = \min_{x \in \mathbb{R}^d} f(x)$.

3.2.5 Method

In this section we present STORM+ (STochastic Recursive Momentum +); a parameter-free stochastic optimization method that finds approximate stationary points at an optimal rate. We describe our method in Algorithm 6 and Eq. (3.22), and state its guarantees in Theorem 3.2.1. We will build towards the ultimate result by first analyzing a simplified version which will help us explain the individual roles of adaptive step-size and the adaptive momentum term. We also aim to describe the dependence between the step-size and momentum, which helps use prove the noise-adaptive convergence rates without needing to know the smoothness parameter.

The original STORM algorithm. The original STORM template of [CO19] relies on an SGD-style update with a corrected momentum. Concretely, the idea is to maintain a gradient estimate d_t which is a *corrected* weighted average of past stochastic gradients, and then update the iterates similarly to SGD,

$$X_{t+1} = X_t - \gamma_t d_t \; . \tag{3.19}$$

Standard momentum is a weighted average of past gradients,

$$d_t = a_t \nabla f(X_t, \xi_t) + (1-a_t) d_{t-1} \; ; \text{ where } a_t \in [0,1] \; .$$

Under this construction, d_t is generally a biased estimate of $\nabla f(X_t)$. In STORM it is suggested to add a correction term ,$(1-a_t)(\nabla f(X_t,\xi_t)-\nabla f(X_{t-1},\xi_t))$, which leads to the following update rule (again, $a_t \in [0,1]$),

$$d_t = \nabla f(X_t,\xi_t) + (1-a_t)(d_{t-1} - \nabla f(X_{t-1},\xi_t)) \,, \tag{3.20}$$

The correction term plays a crucial role here: it exploits the smoothness of $f(\cdot,\xi)$ in a way that leads to a variance reduction effect. To see this effect one can inspect the error of the momentum d_t compared to the exact gradient at X_t,

$$\epsilon_t := d_t - \nabla f(X_t) \,.$$

The STORM update rule induces the following error dynamics,

$$\epsilon_t = (1-a_t)\epsilon_{t-1} + a_t(\nabla f(X_t,\xi_t) - \nabla f(X_t)) + (1-a_t)Z_t$$

where $Z_t := (\nabla f(X_t,\xi_t) - \nabla f(X_{t-1},\xi_t)) - (\nabla f(X_t) - \nabla f(X_{t-1}))$. Roughly speaking, we want to argue that the error has some decreasing behavior; whether in sequence or on average. To do so, we want to argue that the second term and the third term either show some decreasing behavior as we iterate the algorithm and hopefully converge to a stationary point. First, the second term in the above dynamics, $a_t(\nabla f(x_t,\xi_t) - \nabla f(x_t))$, which is the variance term, can be controlled by correctly choosing the momentum term a_t. As a_t decreases, the contribution of the variance term shrinks. For the last term, Lipschitz continuity of stochastic gradients with respect to the same sample (Eq. (3.17)) immediately implies $\|Z_t\| \leq O(\|X_t - X_{t-1}\|) = O(\gamma_{t-1}\|d_{t-1}\|)$. Intuitively, as we approach a stationary point (and use a small enough step-size) then $\gamma_{t-1}\|d_{t-1}\|$ should decrease, which in turn reduces the magnitude of Z_t. Thus, carefully controlling the step-size and momentum parameters leads to a variance reduction effect which facilitates fast convergence. The novelty of our approach is identifying the relationship between the step-size and momentum while ensuring parameter-free algorithm design.

The original STORM paper [CO19] makes the following choices,

$$\gamma_t = \theta / \left(w + \sum_{s=1}^{t} \|g_s\|^2 \right)^{1/3} \qquad \& \qquad a_t = cL^2\gamma_{t-1}^2 \,, \tag{3.21}$$

where we denote $g_t := \nabla f(x_t,\xi_t)$. The above choice of step-size is inspired by AdaGrad [DHS11], which also sets the step-size inversely proportional to the cumulative square norms of past gradients. Note that θ and w are constants that depend on the smoothness of the objective L, as well as on the bound on the gradients G, and c is an absolute constant independent of the problem's characteristics. These choices of the constants and especially the choice of $a_t \propto L^2\gamma_{t-1}^2$ is crucial for the analysis of the original STORM. In fact, the convergence proof for STORM breaks down unless we encode this prior knowledge into γ_t and a_t. Next, we describe our parameter-free version.

Algorithm 6: STORM+

Input: # of iterations T , $X_1 \in \mathbb{R}^d$

1: Sample ξ_1 and set $d_1 = g_1 = \nabla f(X_1, \xi_1)$
2: **for** $t = 1, ..., T$ **do**
3: $\quad a_{t+1} = \dfrac{1}{\left(1 + \sum_{s=1}^{t} \|g_s\|^2\right)^{2/3}} \quad \& \quad \gamma_t = \dfrac{1}{\left(\sum_{s=1}^{t} \|d_s\|^2 / a_{s+1}\right)^{1/3}}$
4: $\quad X_{t+1} = X_t - \gamma_t d_t$
5: $\quad$ Sample ξ_{t+1} and set $g_{t+1} := \nabla f(X_{t+1}; \xi_{t+1})$, and $\tilde{g}_t := \nabla f(X_t; \xi_{t+1})$
6: $\quad d_{t+1} = g_{t+1} + (1 - a_{t+1})(d_t - \tilde{g}_t)$
7: **end for**
8: **return** $\hat{X}_T$ by choosing uniformly at random from $\{X_1, \ldots, X_T\}$

STORM+ relies on the original STORM template described in Equations (3.19) and (3.20), with the following parameter-free choices of step-size and momentum parameter,

$$\gamma_t = 1 / \left(\sum_{s=1}^{t} \|d_s\|^2 / a_{s+1} \right)^{1/3} \qquad \& \qquad a_t = 1 / \left(1 + \sum_{s=1}^{t-1} \|g_s\|^2 \right)^{2/3} , \tag{3.22}$$

where, again, we denote $g_t := \nabla f(X_t, \xi_t)$. Note that in contrast to the original STORM, our adaptive step-size builds on history of estimates $\{d_1, \ldots, d_t\}$ as well as on the momentum parameters $\{a_1, \ldots, a_{t+1}\}$. Our momentum term is similar to the adaptive choice of STORM, yet it does not require a bound on the gradients nor on the smoothness parameter, which was crucial for the original analysis. Finally, note that the above choice ensures $a_t \in [0, 1]$. For completeness we present our method in Algorithm 6, where it can be seen that STORM+ is a combination of the original STORM template (Equations (3.19) and (3.20)) together with the specific choices of γ_t and a_t appearing in Eq. (3.22). Similar to the discussion in Section 3.1.4, the solution that STORM+ outputs is a point chosen uniformly at random among all iterates, which is quite standard in (stochastic) non-convex optimization.

A note on notation: In Algorithm 6 and throughout the rest of the paper we will employ the following notation due to margin-wise (horizontal) space constraints

$$g_t := \nabla f(X_t, \xi_t) \; ; \;\; \tilde{g}_t := \nabla f(X_t, \xi_{t+1}) \; ; \;\; \bar{g}_t := \nabla f(X_t) \; .$$

One could interpret g_t as the standard stochastic oracle feedback, the "tilde" on $\tilde{g}_t$ denotes the correction oracle call and the "bar" on $\bar{g}_t$ signifies the exact oracle information. Now, we are at a position to present our main theorem regarding STORM+ (Algorithm 6):

Theorem 3.2.1. *Under the assumption in Eq.* (3.14), (3.15), (3.16) *and* (3.17) STORM+ *ensures,*

$$\mathbb{E}\left[\|\nabla f(\hat{X}_T)\| \right] \leq O\left(\frac{M}{\sqrt{T}} + \frac{\kappa \sigma^{1/3}}{T^{1/3}} \right) ,$$

where $\kappa = O(1 + B^{3/4} + L^{3/2})$; $M = O(1 + L^{9/4} + B^{9/8} + G^5 + (LG^4)^{3/2})$, *and the expectation is with*

respect to the randomization of the samples as well as the selection of the output point.

Theorem 3.2.1 demonstrates that in the stochastic case STORM+ achieves the optimal $O(1/T^{1/3})$ rate for our setting. Moreover, it can be seen that STORM+ implicitly adapts to the variance of the noise; in the noiseless case where $\sigma = 0$, STORM+ recovers the optimal $O(1/\sqrt{T})$ rate. We note that scaling the step-size by some (absolute) constant factor may enable us to obtain better dependence on L and B.

3.2.6 Analysis

In this section we provide the convergence analysis of the STORM+ algorithm. We begin with the analysis in the offline case where $\sigma = 0$, and establish a convergence rate of $O(1/\sqrt{T})$ for completeness. Then, we introduce a simplified version of STORM+, with a non-adaptive momentum parameter of the form $a_{t+1} := 1/t^{2/3}$. Due to simplicity and space limitations, this simplified version enables us to illustrate the main steps of the original proof. We show that this version achieves a convergence rate of $O(1/T^{1/3})$ in the stochastic case, though it does not adapt to the variance. Finally, we provide a proof sketch for STORM+ in Algorithm 6 that establishes the result in Theorem 3.2.1.

Offline Case

Here we analyze STORM+ in the case where $\sigma = 0$, and demonstrate a rate of $O(1/\sqrt{T})$ for finding an approximate stationary point.

Theorem 3.2.2. *Let f satisfy Eq. (3.14), (3.17) and $\hat{X}_T$ be generated after running Algorithm 6 for T iterations under deterministic oracle. Then it holds that,*

$$\mathbb{E}\left[\,\|\nabla f(\hat{X}_T)\|\,\right] \leq O(\sqrt{1+L^3+B^{9/4}}/\sqrt{T})\,.$$

where we take expectation due to the selection of $\hat{X}_T$, uniformly at random among the iterate sequence $\{X_t\}$ generated by the algorithm (see line 8 in Algorithm 6).

Proof. In the case where $\sigma = 0$ one can directly show by induction that $d_t = \bar{g}_t = \nabla f(X_t)$. So the update rule becomes $X_{t+1} = X_t - \gamma_t \nabla f(X_t)$. Now, using the smoothness of the objective implies,

$$\Delta_{t+1} - \Delta_t = f(X_{t+1}) - f(X_t) \leq -\gamma_t \|\nabla f(X_t)\|^2 + L\gamma_t^2 \|\nabla f(X_t)\|^2/2\,,$$

here we denoted $\Delta_t := f(X_t) - f(x^*)$, where $x^* \in \arg\min_{x\in\mathbb{R}^d} f(x)$. Dividing by γ_t, re-arranging and summing gives,

$$\sum_{t=1}^{T} \|\nabla f(X_t)\|^2 \leq \frac{\Delta_1}{\gamma_1} - \frac{\Delta_{T+1}}{\gamma_T} + \sum_{t=2}^{T}\left(\frac{1}{\gamma_t} - \frac{1}{\gamma_{t-1}}\right)\Delta_t + \frac{L}{2}\sum_{t=1}^{T}\gamma_t\|\nabla f(X_t)\|^2$$

$$\leq \frac{B}{\gamma_1} + B\sum_{t=2}^{T}\left(\frac{1}{\gamma_t} - \frac{1}{\gamma_{t-1}}\right) + \frac{L}{2}\sum_{t=1}^{T}\frac{\|\nabla f(X_t)\|^2}{\left(\sum_{s=1}^{t}\|\bar{g}_s\|^2\right)^{1/3}}$$
$$\leq \frac{B}{\gamma_T} + L\left(\sum_{t=1}^{T}\|\bar{g}_t\|^2\right)^{2/3} \leq B\left(\sum_{t=1}^{T}\|\nabla f(X_t)\|^2/a_{t+1}\right)^{1/3} + L\left(\sum_{t=1}^{T}\|\nabla f(X_t)\|^2\right)^{2/3}$$
$$\leq B\left(1+\sum_{t=1}^{T}\|\nabla f(X_t)\|^2\right)^{2/9}\left(\sum_{t=1}^{T}\|\nabla f(X_t)\|^2\right)^{1/3} + L\left(\sum_{t=1}^{T}\|\nabla f(X_t)\|^2\right)^{2/3} \tag{3.23}$$

where the second inequality uses $\gamma_t = \left(\sum_{s=1}^{t}\|\bar{g}_s\|^2/a_{s+1}\right)^{-1/3} \leq \left(\sum_{s=1}^{t}\|\bar{g}_s\|^2\right)^{-1/3}$ which holds since $d_t = \nabla f(X_t)$ and $a_t \leq 1$. We also use that $\Delta_t \in [0,B]$ together with $\gamma_t^{-1} - \gamma_{t-1}^{-1} \geq 0$. The third inequality uses Lemma 3.4.3 below; and the last inequality uses $1/a_{t+1} \leq (1/a_{T+1}) = \left(1+\sum_{t=1}^{T}\|\nabla f(X_t)\|^2\right)^{2/3}$, which holds since a_t is monotonically non-increasing.

To solve the above system, we want to represent it as a root finding for polynomia inequalities. First, add 1 to both sides and define $x^9 = 1+\sum_{t=1}^{T}\|\nabla f(X_t)\|^2$. Treating the inequality in Eq. (3.23) as a polynomial of x results in the equivalent expression

$$x^9 - Lx^6 - Bx^5 - 1 \leq 0.$$

Using the same arguments as we did before, one coudl verify that setting $x = O((1+L^3+B^{9/4})^{1/9})$ satisfies the polynomial inequality. This implies that

$$\sum_{t=1}^{T}\|\nabla f(X_t)\|^2 \leq 1+\sum_{t=1}^{T}\|\nabla f(X_t)\|^2 \leq O(1+L^3+B^{9/4})\,.$$

Using the definition of $\hat{X}_T$ as well as Jensen's inequality gives us the final result,

$$\mathbb{E}\left[\,\|\nabla f(\hat{X}_T)\|\,\right] = \mathbb{E}\left[\,\|\bar{g}(\hat{X}_T)\|\,\right] \leq \sqrt{\mathbb{E}\left[\,\|\bar{g}(\hat{X}_T)\|^2\,\right]} = \sqrt{\sum_{t=1}^{T}\|\nabla f(X_t)\|^2/T} \leq O(\sqrt{1+L^3+B^{9/4}}/\sqrt{T})\,,$$

which establishes the bound. In the proof we have used the technical lemma below, which is a generalization of Lemma 2.1.2 for any exponent $0 < p < 1$.

Lemma 3.2.1. *Let $b_1 > 0$, $b_2, \ldots, b_n \geq 0$ be a sequence of real numbers, $p \in (0,1)$ be a real number.*

$$\sum_{s=1}^{n}\frac{b_i}{\left(\sum_{j=1}^{i}b_j\right)^p} \leq \frac{1}{1-p}\left(\sum_{s=1}^{n}b_i\right)^{1-p}$$

■

Stochastic Case for Simplified STORM+

Here we analyze a simplified version of STORM+ in the stochatic setting. While this version does not adapt to the noise variance, it exhibits the optimal rate of $O(1/T^{1/3})$ in the stochastic

case, and its analysis illustrates some of the main ideas that we employ in the proof of the fully adaptive STORM+ (which is more involved).

The version that we analyze here differs from STORM+ in the choice of the momentum parameters. Here we choose $a_1 = 1$ and $a_{t+1} = 1/t^{2/3}$; $\forall t \geq 1$, in contrast to the adaptive choice that we make in Algorithm 6. Note that we keep the same expression for the step size, $\gamma_t = 1/\left(\sum_{s=1}^{t} \|d_s\|^2 / a_{s+1}\right)^{1/3}$.

Theorem 3.2.3. *Under Assumptions in Eq.* (3.14), (3.15), (3.16) *and* (3.17), simplified STORM+ *ensures,*

$$\mathbb{E}\left[\, \|\nabla f(\hat{X}_T)\| \,\right] = O(\sqrt{L^3 + \sigma^2 + B^{3/2}}/T^{1/3}) \,,$$

Note that the estimator d_t is not (conditionally) unbiased. This eliminates a standard SGD-type analysis. Similar to the proof technique in Cutkosky and Orabona [CO19], our approach hinges on bounding the cumulative error $\mathbb{E}\left[\sum_{t=1}^{T} \|\epsilon_t\|^2\right]$, where ϵ_t is the difference between the corrected momentum d_t and the exact gradient $\bar{g}_t$, i.e. $\epsilon_t = d_t - \bar{g}_t$. On top of that, we go a step further and understand the behavior of the cumulative error with respect to the sum of *exact gradients*, $\mathbb{E}\left[\sum_{t=1}^{T} \|\bar{g}_t\|^2\right]$.

Before we proceed with the proof sketch of the simplified version, let us present an intermediate result, which provides us with the main departure point for the main results (Theorem 3.2.1) and the analysis of the simplified algorithm (Theorem 3.2.3).

Lemma 3.2.2. *Let* $\{X_t\}$ *be generated by* STORM+ *(Algorithm 6) under the assumptions in Theorem 3.2.1. Then, it holds that*

$$\sum_{t=1}^{T} \|\bar{g}_t\|^2 \leq \sum_{t=1}^{T} \|\epsilon_t\|^2 + 2B a_{T+1}^{-1/3} (\sum_{t=1}^{T} \|d_t\|^2)^{1/3} + \frac{3}{2} L (\sum_{t=1}^{T} \|d_t\|^2)^{2/3}$$

Proof Sketch (Theorem 3.2.3). We divide the analysis into two cases:

- First, we take into account the large error regime where the cumulative error is greater than the sum of exact gradients.
- Second, we analyze the more involved small error regime where we make additional use of the Lipschitzness of individual stochastic gradients.

Bounding $\mathbb{E}\left[\sum_{t=1}^{T} \|\epsilon_t\|^2\right]$.
Let us mention the notation once more for completeness:

$$g_t := \nabla f(X_t, \xi_t) \,; \;\; \tilde{g}_t := \nabla f(X_t, \xi_{t+1}) \,; \;\; \bar{g}_t := \nabla f(X_t) \,; \;\; \epsilon_t = d_t - \bar{g}_t \,.$$

Also, recall that the update rule for d_t induces the following error dynamics,

$$\epsilon_t = (1 - a_t)\epsilon_{t-1} + a_t(g_t - \bar{g}_t) + (1 - a_t) Z_t \tag{3.24}$$

where $Z_t := (g_t - \tilde{g}_{t-1}) - (\bar{g}_t - \bar{g}_{t-1})$. Letting $\mathscr{F}_t$ be the sigma-algebra defined by the randomness up to and including time t, i.e., $\mathscr{F}_t := \sigma(x_1, \xi_1, \xi_2, \ldots, \xi_t)$ and recalling that both a_t and X_t depend on history up to $t-1$, i.e., $\mathscr{F}_{t-1}$, immediately implies that

$$\mathbb{E}\left[\, a_t(g_t - \bar{g}_t) | \mathscr{F}_{t-1} \,\right] = 0 \quad \& \quad \mathbb{E}\left[\, (1-a_t) Z_t | \mathscr{F}_{t-1} \,\right] = 0.$$

Similarly, we have the following conditionally independent expression, $\mathbb{E}\left[\, (1-a_t)\epsilon_{t-1} | \mathscr{F}_{t-1} \,\right] = (1-a_t)\epsilon_{t-1}$. Thus, taking the square of the above equation and then computing the expected value gives,

$$\begin{aligned}
\mathbb{E}\left[\, \|\epsilon_t\|^2 \,\right] &\le (1-a_t)^2 \mathbb{E}\left[\, \|\epsilon_{t-1}\|^2 \,\right] + \mathbb{E}\left[\, \|(1-a_t) Z_t + a_t(g_t - \bar{g}_t)\|^2 \,\right] \\
&\le (1-a_t)^2 \mathbb{E}\left[\, \|\epsilon_{t-1}\|^2 \,\right] + 2(1-a_t)^2 \mathbb{E}\left[\, \|Z_t\|^2 \,\right] + 2a_t^2 \mathbb{E}\left[\, \|g_t - \bar{g}_t\|^2 \,\right] \\
&\le (1-a_t) \mathbb{E}\left[\, \|\epsilon_{t-1}\|^2 \,\right] + 8L^2 \mathbb{E}\left[\, \gamma_{t-1}^2 \|d_{t-1}\|^2 \,\right] + 2a_t^2 \sigma^2 \,, \qquad (3.25)
\end{aligned}$$

where the second line uses $\|b+c\|^2 \le 2\|b\|^2 + 2\|c\|^2$, and the last line uses $\mathbb{E}\left[\, \|g_t - \bar{g}_t\|^2 \,\right] \le \sigma^2$ and $(1-a_t) \in [0,1]$, as well as the smoothness assumption that implies

$$\|Z_t\| \le \|g_t - \tilde{g}_{t-1}\| + \|\bar{g}_t - \bar{g}_{t-1}\| \le 2L\|X_t - X_{t-1}\| = 2L\gamma_{t-1}\|d_{t-1}\|.$$

Dividing Eq. (3.25) by a_t and re-arranging implies,

$$\mathbb{E}\left[\, \|\epsilon_{t-1}\|^2 \,\right] \le \frac{1}{a_t}(\mathbb{E}\left[\, \|\epsilon_{t-1}\|^2 \,\right] - \mathbb{E}\left[\, \|\epsilon_t\|^2 \,\right]) + 8L^2 \mathbb{E}\left[\, \gamma_{t-1}^2 \|d_{t-1}\|^2 / a_t \,\right] + 2a_t\sigma^2 \,.$$

Summing the above, and using $\epsilon_0 := 0$ gives,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \|\epsilon_{t-1}\|^2 \right] \le \underbrace{-\frac{\mathbb{E}\left[\, \|\epsilon_T\|^2 \,\right]}{a_T}}_{\text{(A)}} + \underbrace{\sum_{t=1}^{T-1} \left(\frac{1}{a_{t+1}} - \frac{1}{a_t}\right) \mathbb{E}\left[\, \|\epsilon_t\|^2 \,\right]}_{\text{(B)}} \\
+ 8L^2 \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \gamma_{t-1}^2 \|d_{t-1}\|^2 / a_t \right]}_{\text{(C)}} + 2\sigma^2 \underbrace{\sum_{t=1}^{T} a_t}_{\text{(D)}} \qquad (3.26)
\end{aligned}$$

Next, we bound all the term on the RHS of the above equation.

Bounding (A). Since $a_T \le 1$ we can bound $-\mathbb{E}\left[\, \|\epsilon_T\|^2 \,\right] / a_T \le -\mathbb{E}\left[\, \|\epsilon_T\|^2 \,\right]$

Bounding (B). Note that $G(z) = z^{2/3}$ is a concave function in $\mathbb{R}_+$. Thus, the gradient inequality for concave functions implies that $\forall z_1, z_2 \ge 0$, we have $(z_1 + z_2)^{2/3} - z_1^{2/3} \le \frac{2}{3} z_1^{-1/3} z_2$. Hence, for all $t \ge 2$,

$$1/a_{t+1} - 1/a_t = t^{2/3} - (t-1)^{2/3} \le 2(t-1)^{-1/3}/3 \le 2/3 \,.$$

Moreover, $1/a_2 - 1/a_1 = 0$. Combining the above expressions together, we have (B) $\le (2/3)\mathbb{E}\left[\, \sum_{t=1}^{T} \|\epsilon_t\|^2 \,\right]$.

Bounding (C). By the definition of γ_t we have,

$$\begin{aligned}(\mathrm{C}) &= \mathbb{E}\left[\sum_{t=1}^{T}\frac{\|d_{t-1}\|^2/a_t}{\left(\sum_{s=1}^{t-1}\|d_s\|^2/a_{s+1}\right)^{2/3}}\right]\\ &\leq 3\mathbb{E}\left[\left(\sum_{t=1}^{T-1}\|d_t\|^2/a_{t+1}\right)^{1/3}\right]\\ &\leq 3T^{2/9}\left(\mathbb{E}\left[\sum_{t=1}^{T}\|d_t\|^2\right]\right)^{1/3}.\end{aligned}$$

where the first inequality uses Lemma 3.4.3, and the second inequality uses $1/a_t \leq 1/a_{T+1} \leq T^{2/3}$ as well as Jensen's inequality with respect to the concave function $U(z) = z^{1/3}$ defined over $\mathbb{R}_+$.

Bounding (D). Lemma 3.4.3 immediately implies that $(\mathrm{D}) = 1 + \sum_{t=1}^{T-1} 1/t^{2/3} \leq 1 + 3T^{1/3} \leq 4T^{1/3}$.

Plugging these bounds into Eq. (3.26) and re-arranging yields,

$$\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right] \leq 72L^2T^{2/9}\left(6\mathbb{E}\left[\sum_{t=1}^{T}\|d_t\|^2\right]\right)^{1/3} + 24\sigma^2T^{1/3}. \tag{3.27}$$

Now, we use the bound of Eq. (3.27) in order to bound the sum of square gradients. Let us divide the analysis into two cases.

Case 1 (Large error regime): $\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right] \geq \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right]$.
Combining the condition of Case 1 with $\|d_t\|^2 \leq 2\|\bar{g}_t\|^2 + 2\|\epsilon_t\|^2$, implies that $\mathbb{E}\left[\sum_{t=1}^{T}\|d_t\|^2\right] \leq 6\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right]$. Plugging this inside Eq. (3.27) yields,

$$\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right] \leq 72\cdot 6^{1/3}L^2T^{2/9}\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right]^{1/3} + 24\sigma^2T^{1/3}.$$

Let us denote $x = \mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right]^{1/3}$, then the above expression could be represented as a cubic polynomial of the form (ignoring the absolute constants for simplicity),

$$x^3 - (LT^{1/9})^2x - \sigma^2T^{3/9} \leq 0.$$

One could easily see that taking $x = O(LT^{1/9})$ satisfies the inequality. However, this is not satisfactory as we would lose the dependence on σ. By a bit more involved analysis, we end up taking $x = O((L^3+\sigma^2)^{1/3}T^{1/9})$, which immediately implies that $\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right] \leq O((L^3+\sigma^2)T^{1/3})$. Finally, due to the condition of Case 1 we therefore have,

$$\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right] \leq O((L^3+\sigma^2)T^{1/3}),$$

concluding the first part of the proof under Case 1.

Case 2 (Small error regime): $\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right] \leq \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right]$.
Once again, $\|d_t\|^2 \leq 2\|\bar{g}_t\|^2 + 2\|\epsilon_t\|^2$ and the condition of Case 2 together imply that $\mathbb{E}\left[\sum_{t=1}^{T}\|d_t\|^2\right] \leq 3\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right]$.

Now using the update rule $X_{t+1} = X_t - \gamma_t d_t$ together with smoothness of $f(\cdot)$, one can show in a similar manner to the derivation of Eq. (3.23) the following bound,

$$\sum_{t=1}^{T}\|\bar{g}_t\|^2 \leq \sum_{t=1}^{T}\|\epsilon_t\|^2 + 2BT^{2/9}\left(\sum_{t=1}^{T}\|d_t\|^2\right)^{1/3} + \frac{3}{2}L\left(\sum_{t=1}^{T}\|d_t\|^2\right)^{2/3} \tag{3.28}$$

which is due to Lemma 3.2.2 by replacing the adaptive definition of momentum parameter with $a_{t+1} = t^{-}2/3$. Taking the expectation of the above equation and plugging in $\mathbb{E}\sum_{t=1}^{T}\|d_t\|^2 \leq 3\mathbb{E}\sum_{t=1}^{T}\|\bar{g}_t\|^2$ as well as $\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right] \leq \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right]$ gives,

$$\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right] \leq \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right] + 2\cdot 3^{1/3}BT^{2/9}\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right]^{1/3} + \frac{3^{4/3}}{2}L\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right]^{2/3} \tag{3.29}$$

where we also used Jensen's inequality with respect to the concave functions $z^{1/3}$ and $z^{2/3}$ defined over $\mathbb{R}_+$. Using the same arguments regarding cubic polynomials in the Case 1, one could realize that the above immediately implies, $\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right] \leq O((L^3 + B^{3/2})T^{1/3})$. This concludes Case 2.

Final bound: Combining case 1 and 2.
By summing up the two possible bounds, we have shown that $\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right] \leq O((L^3 + \sigma^2 + B^{3/2})T^{1/3})$, combining this with the definition of $\hat{X}_T$ and using Jensen's inequality similarly to what we did in the offline analysis provides,

$$\mathbb{E}\left[\|\nabla f(\hat{X}_T)\|\right] = O(\sqrt{L^3 + \sigma^2 + B^{3/2}}/T^{1/3}),$$

which concludes the proof. ■

So far in this sectiion, the techniques which enabled us to adapt to the smoothness and variance bounds is fundamentally different than what we have seen in Chapter 2 and Section 3.1 of this chapter. The main goal of the previous sections was to primarily show that the error with respect to *not knowing the smoothness or the variance* could be compensated on average. By using the data-adaptive, monotonically decreasing step-sizes in the sense of AdaGrad, the step-size approximates the (inverse) smoothness constant well enough, and we can argue that the cumulative error of such an approximation is bounded by a constant. This, in turn, implies that the additional error term subsequently decreases in the same order as the optimal convergence rate.

In this section, we go down a different route, and try to understand how the cumulative

variance behaves. Essentially, knowing the smoothness and gradient bounds enables to have an expected decrease in the variance. Our solution is to understand the growth of the cumulative variance with respect to the sum of exact gradients, which helps us relate the cumulative error and noise in the estimator directly to the convergence metric. This approach is fundamentally different than what we have presented so far within the context of analyzing parameter-free, data-adaptive algorithms. The next section will make use of similar arguments to prove convergence rates in the more specific (non-convex) finite-sum minimization.

Stochastic Case for the Original STORM+ (Algorithm 6)

Finally, we provide a sketch of the proof for the main theorem with respect to our original algorithm. At a high level, the analysis follows similar lines to that of the simplified STORM+ in the previous section. There are two extra challenges we need to handle:

1. Now a_t is a random variable that depends on the noisy samples. This introduces additional measurability problems.
2. The differences $1/a_{t+1} - 1/a_t$ are not necessarily smaller than 1.

Among the tools that we use to address the first challenge is a version of Young's inequality. It is essential to identify the correct exponents and constants for different terms that appear on the right-hand side. We will defer the details to the appendix and focus on the handling of the resulting terms.

Recall that in the simplified version of our algorithm, we have shown that $1/a_{t+1} - 1/a_t \leq 2/3$, which was crucial in bounding the term (B). To cope with the second challenge, we bound the expectation of $\sum_{t=1}^{T} \|\epsilon_t\|^2$ by splitting the summation into two regimes,

$$\sum_{t=1}^{T} \|\epsilon_t\|^2 = \sum_{t=1}^{\tau^*} \|\epsilon_t\|^2 + \sum_{t=\tau^*+1}^{T} \|\epsilon_t\|^2$$

where τ^* is a time-step after which we can ensure that $1/a_{t+1} - 1/a_t \leq 2/3$. This technique is similar to what we have done in Chapter 2; we tried to identify a point in time beyond which the step-size is going to be smaller than C/L, where C is some positive constant that depends on the algorithm and the problem setting at hand. In the proof sketch, we delve into the details of how we handle the two aforementioned challenges.

Before we proceed with the proof sketch of the main theorem, we would like to present a complementary numerical inequality, which considers the case of Lemma 3.4.3 when the exponent is $p = 4/3$, simultaneously going beyond the context of Lemma 3.1.1.

Lemma 3.2.3. *For any non-negative real numbers $a_1, \ldots, a_n \in [0, a_{\max}]$,*

$$\sum_{i=1}^{n} \frac{a_i}{(1 + \sum_{j=1}^{i-1} a_j)^{4/3}} \leq 12 + 2a_{\max} .$$

Proof Sketch of Theorem 3.2.1. The proof is composed of three parts:

1. In the first part we bound the cumulative expectation of errors $\mathbb{E}\left[\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2\right]$, where $\epsilon_t := d_t - \bar{g}_t$, and τ^* is a stopping time after which we can ensure that $1/a_{t+1} - 1/a_t \le 2/3$. This solves the first challenge.

2. In the second part we use our bound on $\mathbb{E}\left[\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2\right]$ in order to bound the *total sum of square errors*, $\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right]$.

3. Then, in the last part, we divide into two sub-cases as we did in the simpflified proof sketch such that we first analyze the setting if $\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right] \le (1/2)\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right]$ and then its complement. We also use the smoothness of the objective together with the update rule, similarly to what we do in Eq. (3.23).

Part (1): Bounding $\mathbb{E}\left[\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2\right]$.
Recall the error dynamics of STORM+ in Eq. (3.24). Taking the square and summing up to some $\tau^* \in [T]$ enables us to bound,

$$\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2 \le \sum_{t=1}^{\tau^*}(1-a_t)\|\epsilon_{t-1}\|^2 + 2\sum_{t=1}^{\tau^*}\|Z_t\|^2 + 2\sum_{t=1}^{\tau^*}a_t^2\|g_t-\bar{g}_t\|^2 + \sum_{t=1}^{\tau^*}M_t\,,$$

where $M_t = 2\langle(1-a_t)\epsilon_{t-1}, a_t(g_t-\bar{g}_t)+(1-a_t)Z_t\rangle$ is a martingale difference sequence such that $\mathbb{E}[M_t|\mathscr{F}_{t-1}] = 0$, where $\mathscr{F}_t$ is the history upto and including iteration t, i.e., $\mathscr{F}_t := \{x_1, \xi_1, \xi_2, \xi_3 \ldots, \xi_t\}$. Also, recall that we have defined $Z_t := (g_t - \tilde{g}_{t-1}) - (\bar{g}_t - \bar{g}_{t-1})$.

Now let us define $\beta := \min\{1, 1/G^4\}$, and $\tau^* = \max\{t \in [T] : a_t \ge \beta\}$. Recalling that a_{t+1} is measurable with respect to $\mathscr{F}_t$ implies that $\tau^* \in [T]$ is a stopping time adapted to the same sigma-algebra sequence, $\{\mathscr{F}_t\}$. Re-arranging the above and using the definition of τ^* implies,

$$\beta\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2 \le \|\epsilon_{\tau^*}\|^2 + \sum_{t=1}^{\tau^*-1}a_{t+1}\|\epsilon_t\|^2 \le \underbrace{2\sum_{t=1}^{T}\|Z_t\|^2}_{\text{(i)}} + \underbrace{2\sum_{t=1}^{T}a_t^2\|g_t-\bar{g}_t\|^2}_{\text{(ii)}} + \underbrace{\sum_{t=1}^{\tau^*}M_t}_{\text{(iii)}}$$

where we used $\tau^* \le T$, as well as $\beta \le 1$. Note that we haven't used the particular definition of β yet. Next we bound the expected value of the above terms.

Bounding (i). As in the previous section, the smoothness property implies that $\|Z_t\|^2 \le 4L^2\gamma_{t-1}^2\|d_{t-1}\|^2$. Using the expression for γ_{t-1} together with Lemma 3.4.3 enables to show,

$$\text{(i)} \le 4L^2\sum_{t=1}^{T}\frac{\|d_{t-1}\|^2}{(\sum_{s=1}^{t-1}\|d_s\|^2)^{2/3}} \le 12L^2(\sum_{t=1}^{T}\|d_t\|^2)^{1/3}\,.$$

Bounding (ii). Observe that a_t and g_t are conditionally independent given $\mathscr{F}_t$. Therefore, one can directly show that $\mathbb{E}[a_t^2\|g_t-\bar{g}_t\|^2] \le \mathbb{E}[a_t^2\|g_t\|^2]$ by quadratic expansion. Together with the

expression for a_t, it is possible to show that,

$$\mathbb{E}\,[\,(\text{ii})\,] \leq \mathbb{E}\sum_{t=1}^{T} \frac{\|g_t\|^2}{(1+\sum_{s=1}^{t-1}\|g_s\|^2)^{4/3}} \leq C_1\ .$$

where C_1 is a constant, and the second inequality is due to Lemma 3.2.3, proof of which we describe in the appendix.

Bounding (iii). Since $\tau^* \in [T]$ is a bounded stopping time, and M_t is a martingale difference sequence, then Doob's optional stopping theorem [LP17] implies $\mathbb{E}\,[\,(\text{iii})\,] = \mathbb{E}\left[\sum_{t=1}^{\tau^*} M_t\right] = 0$.

Final bound. Combining all three bounds above, and applyting Jensen's inequality for $U(z) = z^{1/3}$ the concave function defined over $\mathbb{R}_+$ gives us the first part of the proof,

$$\mathbb{E}\left[\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2\right] \leq 2C_1/\beta + 24(L^2/\beta)\mathbb{E}\left[\sum_{t=1}^{T}\|d_t\|^2\right]^{1/3}. \tag{3.30}$$

Part (2): Bounding $\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right]$.
Recall the error dynamics of STORM+ in Eq. (3.24). Dividing by $\sqrt{a_t}$, taking the square and summing up to some T enables to bound,

$$\frac{1}{a_t}\|\epsilon_t\|^2 \leq (\frac{1}{a_t}-1)\|\epsilon_{t-1}\|^2 + 2\frac{\|Z_t\|^2}{a_t} + 2a_t\|g_t-\bar{g}_t\|^2 + Y_t$$

where $Y_t = 2\langle\frac{1-a_t}{\sqrt{a_t}}\epsilon_{t-1}, \sqrt{a_t}(g_t-\bar{g}_t) + \frac{1-a_t}{\sqrt{a_t}}Z_t\rangle$ is a martingale difference sequence such that $\mathbb{E}\,[\,Y_t|\mathscr{F}_{t-1}\,] = 0$. Re-arranging the above and summing over t yields,

$$\sum_{t=1}^{T}\|\epsilon_{t-1}\|^2 \leq \underbrace{-\frac{1}{a_T}\|\epsilon_T\|^2}_{\text{(A)}} + \underbrace{\sum_{t=1}^{T}(\frac{1}{a_{t+1}}-\frac{1}{a_t})\|\epsilon_t\|^2}_{\text{(B)}} + \underbrace{2\sum_{t=1}^{T}\frac{\|Z_t\|^2}{a_t}}_{\text{(C)}} + \underbrace{2\sum_{t=1}^{T}a_t\|g_t-\bar{g}_t\|^2}_{\text{(D)}} + \underbrace{\sum_{t=1}^{T}Y_t}_{\text{(E)}}$$

First off, due to the martingale property for Y_t, we have $\mathbb{E}\,[\,(\text{E})\,] = 0$. Terms (A), (C), (D) will be bounded using the same arguments and techniques as we have done in the proof of the simplified algorithm. The challenge is bounding (B) in the case where a_t is data-adaptive and hence a random variable.

Bounding (B). Using the definition of τ^* one can show that $1/a_{t+1} \leq 1/\tilde{\beta}\ ; \forall\, t \leq \tau^*$, where $1/\tilde{\beta} := (1/\beta^{3/2} + G^2)^{2/3}$ due to boundedness of stochastic gradients. Moreover, by using the gradient inequality for the function $U(z) = z^{2/3}$ where $z = 1+\sum_{s=1}^{t}\|iter[g]\|^2$ and the definition of τ^* we can show,

$$1/a_{t+1} - 1/a_t \leq 2/3\ ; \quad \forall\, t \geq \tau^*+1$$

This enables to decompose and bound (B) according to τ^*,

$$\sum_{t=1}^{T}(\frac{1}{a_{t+1}}-\frac{1}{a_t})\|\epsilon_t\|^2=\sum_{t=1}^{\tau^*}(\frac{1}{a_{t+1}}-\frac{1}{a_t})\|\epsilon_t\|^2+\sum_{t=\tau^*+1}^{T}(\frac{1}{a_{t+1}}-\frac{1}{a_t})\|\epsilon_t\|^2$$
$$\leq\frac{1}{\tilde{\beta}}\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2+\frac{2}{3}\sum_{t=\tau^*+1}^{T}\|\epsilon_t\|^2\leq\frac{1}{\tilde{\beta}}\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2+\frac{2}{3}\sum_{t=1}^{T}\|\epsilon_t\|^2\,. \tag{3.31}$$

Then, by plugging the expression in Eq. (3.30) into the above expression, the expected value of term (B) is bounded.

From here the analysis of the other terms and bounding $\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_{t-1}\|^2\right]$ is done similarly to our analysis of simplified STORM+.

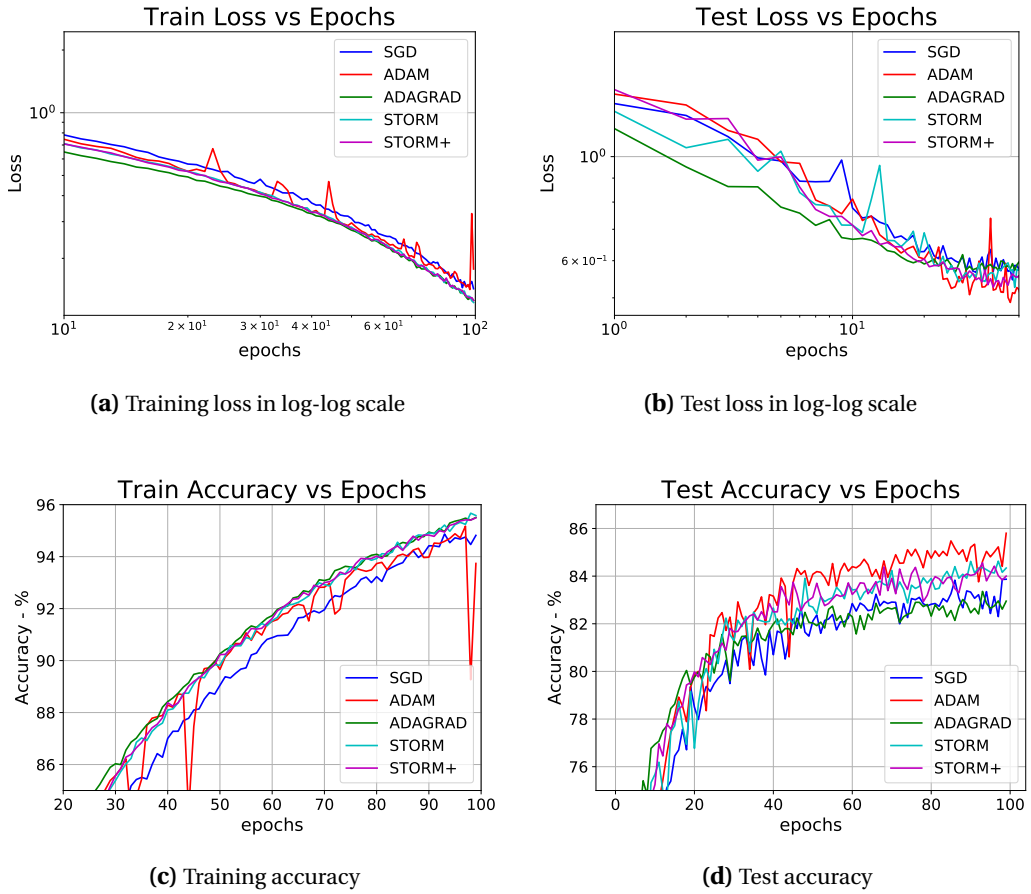
Part (3): Bounding $\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right]$. Finally, we analyze the growth of cumulative error $\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right]$ with respect to $(1/2)\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right]$ in two cases; $\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right]\leq(1/2)\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right]$ and its complement expression. The rest is very similar to our analysis in the simplified setting, details of which is presented in the appendix. ■

3.2.7 Experiments

In this section we provide numerical performance of STORM+ for a multi-class classification task. Specifically, we train ResNet34 architecture on CIFAR10 dataset using SGD with momentum, STORM and STORM+, as well as AdaGrad and Adam. We implemented the whole setup in *pytorch* [Pas+19] retrieving the model and the dataset from *torchvision* package. We executed the experiments on NVIDIA DGX infrastructure. Specifically, our code ran on NVIDIA A100-SXM4-40GB graphics card. We use mini-batches of 100 samples both for training and testing, while using the default train/test data split provided in the package.

To be fair to all methods, we fixed all the parameters to their default value except for the step-size. Then, we executed an initial step-size sweep over the same logarithmic range for all the algorithms. All methods use a constant step-size schedule without any heuristic strategies. All methods are run with the best performing initial step-size after tuning and the results for a single run are presented in Figure 3.1. In the plots, epoch refers to the number of passes over dataset, *not* number of gradient calls. Per iteration cost of STORM and STORM+ are twice that of other methods with respect to forward/backward passes.

The results do not exhibit a noticeable practical advantage for STORM+, however, they verify that it achieves comparable performance with respect to other adaptive methods. The performance of STORM and STORM+ are quite close to each other under all 4 metrics. In the training phase, STORM and STORM+ seem to outperform other methods by a small margin, both in training accuracy and training loss. Adam and SGD seem to achieve a relatively small training accuracy and relatively large training loss compared to other methods. In the test phase, we observe a different picture where Adam generalizes slightly better than other methods, followed by STORM and STORM+ as we could see in Figure 3.1d.



(a) Training loss in log-log scale

(b) Test loss in log-log scale

(c) Training accuracy

(d) Test accuracy

Figure 3.1: Comparison of SGD and adaptive methods, Resnet34 on CIFAR10

In terms of ease of tuning, provably, STORM+ does not require the knowledge of *any* problem parameters to operate and only initial step-size tuning suffices, while STORM additionally needs to tune the initial momentum parameter as, in theory, it requires the knowledge of smoothness and bound on the gradients. Adam would need tuning for its moving average parameters β_1 and β_2, while SGD has a momentum parameter which is subject to a search over admissible values. Similar to STORM+, AdaGrad does not require tuning beyond initial step-size.

3.2.8 Conclusion

We have presented a novel parameter-free and adaptive algorithm for non-convex optimization that obtains the optimal rate in the setting of expectation over smooth losses while adapting to variance in gradient estimates. Our approach suggests a new way to set the step-size and momentum jointly and adaptively throughout the learning process. We also present a new analysis approach for the study of parameter-free methods, specifically applicable for the case of *expectation over smooth losses*, and variance-reduced estimator d_t with corrected

momentum as in Cutkosky and Orabona [CO19]. Different than the techniques used in Chapter 2, we especially focus on the growth of cumulative variance/error, and quantify its growth with respect to the sum of exact gradients. Instead of guaranteeing a monotonic decrease in variance $\|\epsilon_t\|^2$ (in expectation), we control its evolution for the whole of the execution *on average*. We believe that our alternative approaches will open up new avenues to both practical and theoretical developments in the study of non-convex machine learning problems and variance reduction techniques.

3.3 Adaptive Stochastic Variance Reduction for Non-convex Finite-Sum Minimization

3.3.1 Bibliographic Note

This section (Section 3.3) is based on the published work Kavis et al. [Kav+22], published in the NeurIPS 2022 conference.

Author list of the published work.

- Ali Kavis
- Stratis Skoulakis
- Kimon Antonakopoulos
- Leello Tadesse Dadi
- Volkan Cevher

Description of contributions. The candidate worked on the earlier version of Algorithm 7 and proved parameter-free rates with $\tilde{O}\left(1/\sqrt{T}\right)$ *iteration complexity*. However, the dependence on number of components n in the sample complexity was sub-optimal. Stratis Skoulakis identified a different n-dependence in the adaptive step-size (as in line 9 in Algorithm 7) and improved the sample complexity to its final form in Theorem 3.3.1. The candidate and Stratis Skoulakis jointly contributed to all the theoretical results in this work. Leello Tadesse Dadi implemented the neural network experiments (see Figure 3.3 and Table 3.3) while the rest of the experimental results under "Convex loss with non-convex regularizer" are due to the candidate (Figure 3.2).

3.3.2 Introduction

In this last section, we will study smooth, non-convex minimization problems with the following finite-sum structure:

$$\min_{x\in\mathbb{R}^d} f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{Prob}$$

where each component function $f_i : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and is possibly non-convex, and we further assume f is also non-convex. We seek to find an ϵ-approximate first-order stationary point $\hat{x}$ of f, such that $\|\nabla f(\hat{x})\| \le \epsilon$, where $\epsilon > 0$ is the accuracy of the desired solution.

This structure captures many interesting learning problems from empirical risk minimization to training of neural networks. *First-order methods* have been the standard choice for solving (Prob), due to their efficiency and favorable practical behavior. In that regard, while

gradient descent (GD) requires $O(n/\epsilon^2)$ gradient computations, stochastic gradient descent (SGD) requires $O(1/\epsilon^4)$ overall gradient computations. In many interesting machine learning applications n tends to be large, e.g., training a neural network for image classification with very big image datasets [Den+09b], hence SGD typically leads to better practical performance.

To leverage the best of both regimes, GD and SGD, the so-called variance reduction (VR) framework combines the *faster convergence rate* of GD with the *low per-iteration complexity* of SGD. Originally proposed for solving strongly-convex problems [JZ13; DBL14; Ngu+17], variance reduction frameworks essentially generate low-variance gradient estimates by maintaining a balance between periodic full gradient computations and stochastic (mini-batch) gradients. VR methods and their theoretical behavior for *convex problems* have been well-studied under various problem setups and assumptions, including μ-strongly convex functions with $O(n+(L/\mu)\log(1/\epsilon))$ complexity [JZ13; Ngu+17; DBL14]; μ-strongly convex functions with accelerated $O(n+\sqrt{L/\mu}\log(1/\epsilon))$ complexity [All17a; LLZ19; SJM20] and smooth, convex functions with $\tilde{O}(n+1/\epsilon)$ complexity [AY16; SJM20; Dub+22].

For non-convex minimization, earlier attempts extended the existing VR frameworks, achieving the first rates of order $O(n+n^{2/3}/\epsilon^2)$ with sub-optimal dependence on n [Red+16; ZXG18; All17b; LL18]. The most recent non-convex VR methods [Fan+18; Wan+19; Li+21; Pha+20; LHR21] close this gap and achieve the optimal gradient oracle complexity of $O(n+\sqrt{n}/\epsilon^2)$ [Fan+18].

Adaptivity and First-order Optimization

The selection of the step-size is of great importance in both the theoretical and practical performance of first-order methods, including the aforementioned VR methods. In the case of L-smooth minimization, first-order methods need the knowledge of L so as to adequately select their step-size [Nes03], otherwise the method is not guaranteed to be convergent and might even diverge [Dub+22; Liu+22a]. To elucidate, classical analysis relies on the (expected) descent property and guarantees that the algorithm monotonically makes progress every iteration. To enforce this property everywhere on the optimization landscape, one needs to pick the step-size as $\gamma_t \leq O(1/L)$, which restricts the step length of the algorithm with respect to the worst-case constant L. On the other hand, estimating the smoothness constant for an objective of interest, such as neural networks, is a very hard task [GRC20]. At the same time, using crude bounds on the smoothness constant leads to very small step-sizes and consequently to poorer convergence. In practice the step-size is tuned through an empirical search over a range of hand-picked values that adds a considerable computational overhead and burden. In order to alleviate the burden of tuning process, we need step-sizes that adjust in accordance with the optimization path.

A popular line of research studies first-order methods that *adaptively* select their step-size by taking advantage of the previously produced point. In many settings of interest, these *adaptive methods* are able to guarantee optimal convergence rates without requiring the knowledge of the smoothness constant L while they often admit superior empirical performance due to their ability to decrease the step-size according to the local geometry of the objective function.

Inspired by AdaGrad introduced in the concurrent seminal works of [DHS11; MS10], a recent line of works [LYC18; Kav+19; Jou+20; ENV21] propose adaptive gradient methods that given access to *noiseless gradient-estimates* achieve accelerated rates in the case of L-smooth convex minimization without requiring the knowledge of smoothness constant L. Similarly, Ene, Nguyen, and Vladu [ENV21] and Antonakopoulos, Belmega, and Mertikopoulos [ABM21] propose adaptive methods with optimal convergence rates for monotone variational inequalities while Antonakopoulos et al. [Ant+21] provide adaptive methods for monotone variational inequalities assuming access to *relative noise gradient-estimates*. Hsieh, Antonakopoulos, and Mertikopoulos [HAM21] and Vu, Antonakopoulos, and Mertikopoulos [VAM21] study the convergence properties of adaptive first-order methods for routing and generic games.

Adaptive non-convex methods for general noise. Related to our work is a recent line of papers studying adaptive first-order methods under the *general noise model*. In this setting, a method is assumed to access unbiased stochastic estimate of the gradient with bounded variance. This is a more general setting than finite-sum optimization that comes with worse lower bounds, i.e. $\Omega(1/\epsilon^4)$ gradient-estimates are needed so as to compute an ϵ-stationary point. We remark that in the case of finite-sum minimization there exist variance reduction methods with $O(n+\sqrt{n}/\epsilon^2)$ gradient complexity [Fan+18; Wan+19]. A recent line of works study adaptive first-order methods that are able to achieve near-optimal oracle-complexity while being oblivious to the smoothness constant L and the variance of the estimator [WWB19; Faw+22; LO19; KLC22]. For example, Ward, Wu, and Bottou [WWB19] established that the adaptive method called AdaGrad-Norm is able to achieve $\tilde{O}(1/\epsilon^4)$ gradient-complexity in the general noise model. In their recent work, Faw et al. [Faw+22] significantly extended the results of Ward et al. [WWB19] by showing that AdaGrad-Norm achieves the same rates even in case the gradient admits unbounded norm (a restrictive assumption in [WWB19]) while their result persists even if the variance increases with the gradient norm. In a slightly more restrictive setting in which the objective function admits the form $f(x) := \mathbb{E}_{\xi\sim\mathscr{D}} f(x,\xi)$ and individual components $f(x,\xi)$ are L-smooth with respect to x for all ξ, Levy, Kavis, and Cevher [LKC21] proposed an adaptive method called STORM+ that achieves $O(1/\epsilon^3)$ gradient-complexity. Their result simultanesouly removes the requirement of the knowledge on problem parameters (e.g., smoothness constant, absolute bounds on gradient norms) that the original STORM method [CO19] requires. The latter gradient-complexity matches the $\Omega(1/\epsilon^3)$ lower bound of Arjevani et al. [Arj+19].

Adaptivity and finite-sum minimization. In parallel with what we discussed earlier, existing variance-reduction methods (VR) crucially need to know the smoothness constant L to select their step-size appropriately to guarantee their convergence. To this end, the following natural question arises

*Can we design **adaptive** VR methods that achieve the optimal gradient computation complexity?*

Table 3.2: In the following table we present the gradient computation complexity of the existing non-adaptive and adaptive variance reduction methods for both convex and non-convex finite-sum minimization. Since for there are multiple non-adaptive VR methods, we present the earliest-proposed method matching up to logarithmic factors the respective lower bounds.

$f(x)$	**Non-Adaptive VR**	**Adaptive VR**	**Lower Bound**
convex	$\tilde{O}\left(n+\sqrt{\frac{n}{\epsilon}}\right)$	$\tilde{O}\left(n+\sqrt{\frac{n}{\epsilon}}\right)$	$\Omega(n+\sqrt{\frac{n}{\epsilon}})$
(ϵ-*optimal solution*)	[LLZ19]	[Liu+22a]	[WS16]
convex	$\tilde{O}\left(n+\frac{1}{\epsilon}\right)$	$\tilde{O}\left(n+\frac{1}{\epsilon}\right)$	$\Omega(n+\sqrt{\frac{n}{\epsilon}})$
(ϵ-*optimal solution*)	[ZY16]	[Dub+22]	[WS16]
non-convex	$O\left(n+\frac{\sqrt{n}}{\epsilon^2}\right)$	$\tilde{O}\left(n+\frac{\sqrt{n}}{\epsilon^2}\right)$	$\Omega\left(n+\frac{\sqrt{n}}{\epsilon^2}\right)$
(ϵ-*stationary point*)	[Fan+18]	[**This work**]	[Fan+18]

Li, Wang, and Giannakis [LWG20] and Tan et al. [Tan+16] were the first to propose adaptive variance-reduction methods by using the Barzilai-Borwein step-size [BB88]. Despite their promising empirical performance, these methods do not admit formal convergence guarantees. When the objective function f in (Prob) is convex, [Dub+22] recently proposed an adaptive VR method requiring $O(n+1/\epsilon)$ gradient computation while, shortly after, [Liu+22a] proposed an *accelerated* adaptive VR method requiring $O(n+\sqrt{n}/\sqrt{\epsilon})$ gradient computations.

To the best of our knowledge, there is no adaptive VR method in the case where f is *non-convex*. We remark that f being non-convex captures the most interesting settings such as minimizing the empirical loss of deep neural network where each f_i stands for the loss with respect to i-th data point and thus is a non-convex function in the parameters of the neural architecture. Through this particular example, we could motivate adaptive VR methods in two fronts: first, even estimating the smoothness constant L of a deep neural network is prohibitive [GRC20], and at the same time, the parameter n in (Prob) equals the number of data samples, which can be very large in practice and is prohibitive for the use of deterministic methods.

Contribution and Techniques. In this work we present an adaptive VR method, called AdaSpider, that converges to an ϵ-*stationary point* for (Prob) by using $\tilde{O}(n+\sqrt{n}L^2/\epsilon^2)$ gradient computations. Our gradient complexity bound matches the existing lower bounds up to logarithmic factors [Fan+18]. AdaSpider combines an adaptive step-size schedule in the lines proposed by AdaGrad [DHS11] with the variance-reduction mechanism based on the *stochastic path integrated differential estimator* of the Spider algorithm [Fan+18]. More precisely, AdaSpider selects the step-size by aggregating the norm of its recursive estimator, while following a single-loop structure as in Fang et al. [Fan+18].

Our contributions and techniques can be summarized as follows:

- To our knowledge, AdaSpider is *the first* parameter-free VR method for smooth, non-

convex problems with finite-sum structure in the sense that it is both *accuracy-independent* and is oblivious to the knowledge of any problem parameters including L. Moreover, ϵ-independence enables us to provide *any-iterate* guarantees. While SPIDER needs both ϵ and L to set its step-size as $\min(\frac{\epsilon}{L\sqrt{n}\|\nabla_t\|}, \frac{1}{2\sqrt{n}L})$ to achieve optimal gradient complexity [Fan+18], all other existing non-convex methods must know at least the value of L in order to guarantee convergence [AH16a; Wan+19].

- We introduce a novel step-size schedule $\gamma_t := n^{-1/4}\left(\sqrt{n} + \sum_{s=0}^{t}\|\nabla_s\|^2\right)^{-1/2}$ where ∇_s is the recursive variance-reduced estimator at round s. By identifying a unique additive/multiplicative form for integrating n, we manage to achieve optimal dependence on the number of components. We note that Adaspider can be viewed as SPIDER with the step-size of AdaGrad-Norm [Faw+22; WWB19; SM10; OP15] where the parameters are respectively selected as $\eta := n^{1/4}$ and $b_0^2 := \sqrt{n}$ [Faw+22].
- We show how to combine the above adaptive step-size schedule with the recursive SPIDER estimator in order to ensure that the *average variance* $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_t - \nabla f(X_t)\|\right]$ decays at a rate $\tilde{O}\left(n^{1/4}/\sqrt{T}\right)$. This might be of independent interest for other variance reduction techniques.

We follow a novel technical path that uses the adaptivity of the step-size to bound the overall variance of the process. This fact differentiates our approach from the previous adaptive and non-adaptive VR approaches and provides us with a surprisingly concise analysis.

Remark 3.3.1. Our convergence results do not require bounded gradients that is typically a restrictive assumption that the analysis of the adaptive methods for stochastic optimization requires. We overcome this obstacle by using the fact $\|X_t - X_{t-1}\| \leq 1$ (due to the step-size selection) and thus $\|\nabla f(X_t) - \nabla f(X_{t-1})\| \leq L\|X_t - X_{t-1}\| \leq L$. The latter leads to the following upper bound on the gradient norm, $\|\nabla f(X_t)\| \leq LT + \|\nabla f(X_0)\|$ that leads to only a logarithmic overhead in the final bound (see Lemma 3.3.2). A similar idea is used by Faw et al. [Faw+22] (Lemma 2) in order to remove the bounded gradient assumption on the convergence rates of AdaGrad-Norm under general noise.

3.3.3 Setup and Preliminaries

During the whole of this manuscript, we consider that the non-convex objective function $f : \mathbb{R}^d \mapsto \mathbb{R}$ possesses a finite-sum structure

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

where each component function f_i is L-*smooth* (or alternatively has L-*Lipschitz gradient*) and (possibly) non-convex. To quantify the performance of our algorithm within the context of non-convex minimization, we want to find an ϵ-*first order stationary point* $\hat{x} \in \mathbb{R}^d$ such that

$$\|\nabla f(\hat{x})\| \leq \epsilon.$$

For notational simplicity we define $\|\cdot\|$ as the Euclidean norm. Then, we say that a continuously differentiable function f is *L-smooth* if

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \tag{3.32}$$

which admits the following equivalent form,

$$f(x) \leq f(y) + \nabla f(y)^\top (x - y) + \frac{L}{2}\|x - y\|^2 \quad \text{for all } x, y \in \mathbb{R}^d. \tag{3.33}$$

Observe that smoothness of each component immediately suggests that objective f is L-smooth itself. Since we are studying randomized algorithms for finite-sum minimization problems, we do not consider any variance bounds on the gradients of components. We only assume that we have access to an oracle which returns the gradient of individual components when queried.

3.3.4 Method

In this section, we present our adaptive variance reduction method, called AdaSpider (Algorithm 7) which exploits the variance reduction properties of the *stochastic path integrated differential estimator* proposed in [Fan+18] while combining it with an AdaGrad-type step-size construction [DHS11]. Unlike the original Spider method [Fan+18], our algorithm admits anytime guarantees, i.e., we don't need to specify the accuracy ϵ a priori. Additionally, our algorithm does not need to know the smoothness parameter L and guarantees convergence without any tuning procedure.

Algorithm 7: Adaptive SPIDER (AdaSpider)

Input: $x_0 \in \mathbb{R}^d, \beta_0 > 0, G_0 > 0$

$G = 0$
for $t = 0, \ldots, T-1$ **do**
  if $t \bmod n = 0$ **then**
    $\nabla_t = \nabla f(X_t)$
  else
    pick $i_t \in \{1, \ldots, n\}$ uniformly at random
    $\nabla_t = \nabla f_{i_t}(X_t) - \nabla f_{i_t}(X_{t-1}) + \nabla_{t-1}$
  end if
  $\gamma_t = 1/\left(n^{1/4}\beta_0\sqrt{n^{1/2}G_0^2 + \sum_{s=0}^{t}\|\nabla_s\|^2}\right)$
  $X_{t+1} = X_t - \gamma_t \cdot \nabla_t$
end for
return uniformly at random $\{X_0, \ldots, X_{T-1}\}$.

As Algorithm 7 indicates, AdaSpider performs a full-gradient computation every n iterations

and in the remaining steps it updates the variance-reduced gradient estimator in a recursive manner, $\nabla_t \leftarrow \nabla f_{i_t}(X_t) - \nabla f_{i_t}(X_{t-1}) + \nabla_{t-1}$. The adaptive nature of AdaSpider comes from the selection of the step-size at Algorithm 7, line 9 that only depends on the norms of estimators produced by the algorithm in the previous steps.

Before presenting the formal convergence guarantees of AdaSpider (stated in Theorem 3.3.1), we present the cornerstone idea behind its design and motivate the analysis for controlling the overall variance of the process through the *adaptivity of the step-size*. This conceptual novelty differentiates our work form the previous adaptive VR methods [Dub+22; Liu+22a] for which the adaptive step-size only helps with adapting to the smoothness constant L, and their constructions come with additional challenges in bounding the (cumulative) variance in the whole of the execution. As a result, the following challenge is the first to be tackled by the design of a VR method.

Challenge 1. *Does the average variance of the estimator, $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_t - \nabla f(X_t)\|\right]$, diminishes at a sufficiently fast rate?*

Up next we explain why combining the variance-reduction estimator of Step 7 with the adaptive step-size of Step 9 provides a surprisingly concise answer to Challenge 1. We remark that Spider is able to control the variance at any iterations by choosing $\gamma_t := \min(\frac{\epsilon}{L\sqrt{n}\|\nabla_t\|}, \frac{1}{2\sqrt{n}L})$ as step-size. The latter enforces the method to make tiny steps, $\|X_t - X_{t-1}\| \leq \epsilon/L\sqrt{n}$ which results in ϵ-bounded variance at any iteration. The latter proposed SpiderBoost [Wan+19] provides the same gradient-complexity bounds with Spider but through the *accuracy-independent* step-size $\gamma = 1/L$. SpiderBoost handles Challenge 1 by using a dense gradient-computations schedule[1] combined with amortization arguments based on the descent inequality (this is why the knowledge of L is necessary in its analysis). We remark that AdaSpider, despite being oblivious to L and accuracy ϵ, admits a significantly simpler analysis by exploiting the adaptability of its step-size.

In the rest of the section we present our approach to Challenge 1 and we conclude the section with Theorem 3.3.1 stating the formal convergence guarantees of AdaSpider.

Handling the variance with adaptive step-size We start with the following variance aggregation lemma that is folklore in (VR) literature (e.g. [AH16a]).

Lemma 3.3.1. *Define the gradient estimator at current point x^+ as $\nabla_{x^+} := \nabla f_i(x^+) - \nabla f_i(x) + \nabla_x$ where x denotes the previous step of the execution and i is sampled uniformly at random from $\{1,\dots,n\}$. Then,*

$$\mathbb{E}\left[\|\nabla_{x^+} - \nabla f(x^+)\|^2\right] \leq L^2\|x^+ - x\|^2 + \mathbb{E}\left[\|\nabla_x - \nabla f(x)\|^2\right]$$

Now, let us apply Lemma 3.3.1 on Spider estimator, $\nabla_t := \nabla f_{i_t}(X_t) - \nabla f_{i_t}(X_{t-1}) + \nabla_{t-1}$ to

[1] SpiderBoost computes a full-gradient every $\sqrt{n}$ steps and at the intermediate steps uses batches of size $\sqrt{n}$.

measure its variance at step X_t.

$$\begin{aligned}\mathbb{E}\left[\|\nabla_t - \nabla f(X_t)\|^2\right] &\leq L^2\mathbb{E}\left[\|X_t - X_{t-1}\|^2\right] + \mathbb{E}\left[\|\nabla_{t-1} - \nabla f(X_{t-1})\|^2\right] \\ &\leq L^2\mathbb{E}\left[\gamma_{t-1}^2\|\nabla_{t-1}\|^2\right] + \mathbb{E}\left[\|\nabla_{t-1} - \nabla f(X_{t-1})\|^2\right] \\ &\leq L^2\mathbb{E}\left[\gamma_{t-1}^2\|\nabla_{t-1}\|^2\right] + \ldots + \mathbb{E}\left[\|\nabla_{t-(t \bmod n)} - \nabla f(X_{t-(t \bmod n)})\|^2\right] \\ &= \sum_{\tau=t-(t \bmod n)+1}^{t-1} L^2\mathbb{E}\left[\gamma_\tau^2 \cdot \|\nabla_\tau\|^2\right]\end{aligned}$$

where the last equality follows by the fact $\mathbb{E}\left[\|\nabla_{t-(t \bmod n)} - \nabla f(X_{t-(t \bmod n)})\|^2\right] = 0$ since Algorithm 7 performs a full-gradient computations for every t with $t \bmod n = 0$ (Line 3, Algorithm 7). By telescoping the summation we get,

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_t - \nabla f(X_t)\|^2\right] \leq \sum_{t=0}^{T-1}\sum_{\tau=t-(t \bmod n)+1}^{t-1} L^2\mathbb{E}\left[\gamma_\tau^2 \cdot \|\nabla_\tau\|^2\right] \leq L^2 n \cdot \sum_{t=0}^{T-1}\mathbb{E}\left[\gamma_t^2 \cdot \|\nabla_t\|^2\right]$$

where the n factor on the right-hand side is due to the fact that each term $\mathbb{E}\left[\gamma_t^2\|\nabla_t\|^2\right]$ appears at most n times in the total summation. To this end, using the structure of the *stochastic path integrated differential estimator* we have been able to bound the overall variance of the process as follows,

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_t - \nabla f(X_t)\|^2\right] \leq L^2 n \cdot \sum_{t=0}^{T-1}\mathbb{E}\left[\gamma_t^2 \cdot \|\nabla_t\|^2\right] \tag{3.34}$$

However, it is not clear at all why the above bound is helpful. At this point the adaptive selection of the step-size (Step 9 in Algorithm 7) comes into play by providing the following surprisingly simple answer,

$$\begin{aligned}&\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_t - \nabla f(X_t)\|^2\right] \leq L^2 n\,\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t^2 \cdot \|\nabla_t\|^2\right] \\ = \quad &\frac{L^2\sqrt{n}}{\beta_0^2}\,\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{\|\nabla_t\|^2/G_0^2}{\sqrt{n+\sum_{s=0}^{t}\|\nabla_s\|^2/G_0^2}}\right] \leq \frac{L^2\sqrt{n}}{\beta_0^2}\log\left(1+\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_t\|^2/G_0^2\right]\right)\end{aligned}$$

where the last inequality comes from Lemma 3.1.1. To finalize the bound, we require the following expression that follows by the fact that $\gamma_t \leq 1/\|\nabla_t\|$ and thus $\|X_t - X_{t-1}\| \leq 1$.

Lemma 3.3.2. *Let $\{X_t\}$ be the points produced by Algorithm 7. Then,*

$$\sum_{t=0}^{T-1}\|\nabla_t\|^2 \leq \mathcal{O}\left(n^2 T^3 \cdot \left(\frac{L^2}{\beta_0^2} + \|\nabla f(X_0)\|^2\right)\right)$$

In simple terms, Lemma 3.3.2 helps us avoid the *bounded gradient norm* assumption that is common among in the literature of adaptive methods for smooth non-convex optimization

(both stochastic and randomized). We trade-off the removal of bounded gradient assumption with the $O(\log(T))$ dependence as we will see in Eq. (3.35). As a result, AdaSpider admits the following cumulative variance bound,

$$\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla_t - \nabla f(X_t)\|^2\right] \leq O\left(\frac{L^2\sqrt{n}}{\beta_0^2} \log\left(1 + nT \cdot \left(\frac{L}{\beta_0 G_0} + \frac{\|\nabla f(X_0)\|}{G_0}\right)\right)\right). \tag{3.35}$$

Remark 3.3.2. To this end one might notice that using a more aggressive n dependence on γ_t leads to smaller variance of the estimator which is obviously favorable (see Eq. (3.34) and the effect of step-size to the variance bound). However more aggressive dependence on n leads to smaller step-sizes and thus to sub-optimal overall gradient complexity with respect to the dependence on n. In the analysis section, we explain why the optimal way to inject the n dependence into the step-size is through the simultaneous multiplicative/additive way as described in Step 9 of AdaSpider. Even though it may seem counter-intuitive at the first sight, we claim it is necessary for the correct balance between gradient complexity and n-dependence for the final rate.

We will conclude this discussion with a complementary remark on the interplay between our adaptive step-size and the convergence rate. As we demonstrated in Eq. (3.35), using a data-adaptive step-size leads to a decreasing variance bound *in an amortized sense* as opposed to *any iterate* variance bound of Spider. The trade-off in our favor is the parameter-free step-size that is independent of ϵ and L. For a fair exposition of our results, notice that the aforementioned advantages of an adaptive step-size comes at an additional $\log(T)$ term in our final bound due to Eq. (3.35). This has a negligible effect on the convergence as even in the large iteration regime when T is in the order billions, it amounts to a small constant factor.

We conclude the section with Theorem 3.3.1 that formally establishes the convergence rate of AdaSpider. The proof of Theorem 3.3.1 is deferred to the next section.

Theorem 3.3.1. *Let $\{X_t\}$ be the sequence of points produced by Algorithm 7 in case $f(\cdot)$ is L-smooth. Let us also define $\Delta_0 := f(X_0) - f^*$. Then,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(X_t)\|\right] \leq O\left(n^{1/4} \cdot \frac{\Theta}{\sqrt{T}} \cdot \log\left(1 + nT \cdot \left(\frac{L}{\beta_0 G_0} + \frac{\|\nabla f(X_0)\|}{G_0}\right)\right)\right)$$

where $\Theta = \Delta_0 \cdot \beta_0 + G_0 + L/\beta_0 + L^2/(\beta_0^2 G_0)$. Overall, Algorithm 7 with $\beta_0 := 1$ and $G_0 := 1$ needs at most $\tilde{O}\left(n + \sqrt{n} \cdot \frac{\Delta_0^2 + L^4}{\epsilon^2}\right)$ oracle calls to reach an ϵ-stationary point.

3.3.5 Analysis

In this section we present the key steps for proving Theorem 3.3.1. We first use the triangle inequality to derive,

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(X_t)\|\right] \leq \sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_t - \nabla f(X_t)\|\right] + \sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_t\|\right] \tag{3.36}$$

We have previously discussed how to bound the first term, cumulative variance previous section. More precisely, by the Jensen's inequality and the arguments presented in the previous part, we obtain the following variance bound.

Lemma 3.3.3. *Let $\{X_t\}$ be a sequence of points produced by Algorithm 7. Then,*

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_t - \nabla f(X_t)\|^2\right] \leq O\left(\frac{Ln^{1/4}}{\beta_0}\sqrt{\log\left(1+nT\left(\frac{L}{\beta_0 G_0}+\frac{\|\nabla f(X_0)\|}{G_0}\right)\right)}\right).$$

We continue with presenting how to treat the term $\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_t\|\right]$. By the smoothness of the function and through a telescopic summation one can easily establish the following bound,

$$\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t\|\nabla_t\|^2\right] \leq 2(f(X_0)-f^*) + L\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t^2\|\nabla_t\|^2\right] + \mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t\|\nabla f(X_t)-\nabla_t\|^2\right]$$

As we already explained, the term $\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t^2\|\nabla_t\|^2\right]$ can be upper bounded by the data-adaptive construction of the step-size γ_t through Lemma 3.1.1. There remains two main technical challenges to establish the bound in Lemma 3.3.3;

1. Showing that $\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_t\|^2\right]$ could be upper bounded by $O\left(n^{1/4}\sqrt{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t\|\nabla_t\|^2\right]\right)$ so that we could upper bound the $\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_t\|\right]$ in Eq. (3.36). Note that the challenge arises due to the fact that γ_t and $\|\nabla_t\|$ are dependent random variables.
2. Due to the same measurability problem between γ_T and ∇_t, the *scaled* cumulative variance term $\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t\|\nabla f(X_t)-\nabla_t\|^2\right]$ in the above inequality should be treated separately; Lemma 3.3.1.

Handling the first challenge requires the use of numerical inequality in Lemma 2.1.2 and the derivation exploits the particular dependence on n in our step-size γ_t to establish the necessary bound. Specifically, we formalize our solution to the first challenge in Lemma 3.3.4, which provides a bound on $\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_t\|\right]$ which we will eventually use to bound $\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(X_t)\|\right]$ in Eq. (3.36).

Lemma 3.3.4. *Let $\{X_t\}$ be the sequence of points produced by Algorithm 7 and $\Delta_0 := f(X_0)-f^*$. Then,*

$$\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_t\|\right] \leq \tilde{\mathcal{O}}\left(\Delta_0\beta_0 + G_0 + \frac{L}{\beta_0} + \mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t\|\nabla f(X_t)-\nabla_t\|^2\right]\right)n^{1/4}\sqrt{T}.$$

We simply use the data-adaptive structure of the step-size together with Lemma 3.1.1. To cope with the second challenge, we will prove a complementary result for the *scaled cumulative variance* in the presence of adaptive step-sizes. As γ_T and ∇_t are dependent random objects, the weighted variance term $\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t\|\nabla f(X_t)-\nabla_t\|^2\right]$ cannot be handled by Lemma 3.3.1. To overcome this challenge, we use the monotonic behavior of the step-size γ_t to establish the following refinement.

Lemma 3.3.5. *Let $\{X_t\}$ be the sequence of points produced by Algorithm 7. Then,*

$$\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t\|\nabla_t-\nabla f(X_t)\|^2\right]\leq L^2 n\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t^3\|\nabla_t\|^2\right]$$

Having established the main ingredients, we are now ready to summarize the importance of simultaneous additive/multiplicative n dependence of γ_t. This selection permits us to do achieve two *orthogonal goals* at the same time;

- Bounding the variance of the process, $\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_t-\nabla f(x_t)\|\right]\leq\tilde{O}\left(n^{1/4}\sqrt{T}\right)$ (see Lemma 3.3.3).
- Bounding the sum, $\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_t\|\right]\leq\tilde{O}\left(n^{5/4}\sqrt{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t^3\|\nabla_t\|^2\right]\right)$ (see Lemma 3.3.4 and Lemma 3.3.5).

Another important thing that the selection of γ_t does is that it enables us to upper bound the term $\tilde{O}\left(n^{5/4}\,\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t^3\|\nabla_t\|^2\right]\right)$ by $\tilde{O}\left(n^{1/4}\right)$, the derivation of which can be found in the proof of Theorem 3.3.1.

3.3.6 Experiments

We complement our theoretical findings with an evaluation of the numerical performance of the algorithm under different experimental setups. We aim to highlight the sample complexity improvements over simple stochastic methods, while displaying the advantages of adaptive step-size strategies. For that purpose we design two setups; first, we consider the minimization of a convex loss with a non-convex regularizer in the sense of Wang et al. [Wan+19] and in a second part we consider an image classification task with neural networks.

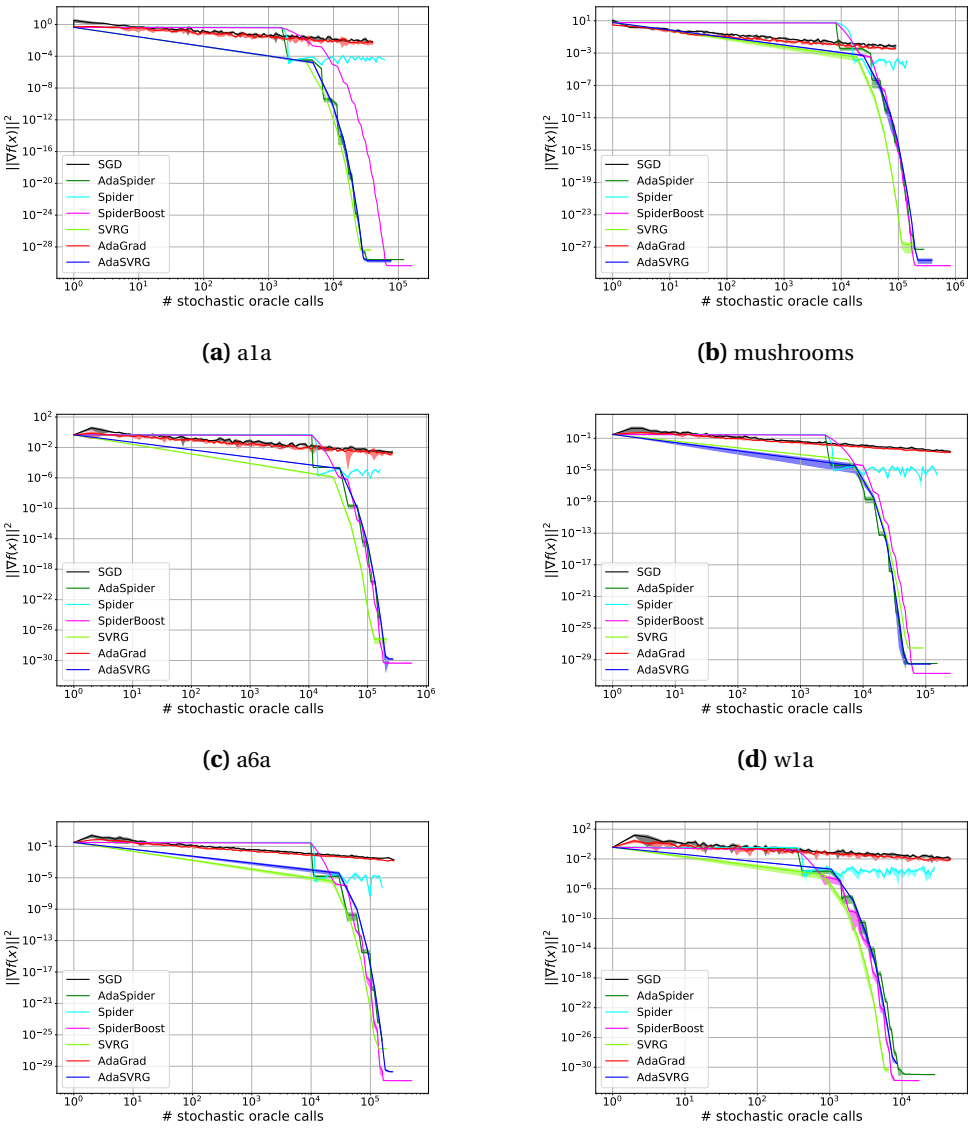
Convex loss with a non-convex regularizer

We consider the following problem:

$$\min_{x\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^{n}\ell(x,(a_i,b_i))+\lambda g(x)$$

where $\ell(x,(a_i,b_i))$ is the loss with respect to the decision variable/weights x with (a_i,b_i) denoting the (feature vector, label) pair. We select $g(x)=\sum_{i=1}^{d}\frac{x_i^2}{1+x_i^2}$, similar to Wang et al. [Wan+19], where the subscript denotes the corresponding dimension of x. We compare

ADASPIDER against the original SPIDER, SPIDERBOOST, SVRG, ADASVRG and two non-VR methods, SGD and ADAGRAD. We picked two datasets from LibSVM, namely `a1a, mushrooms`. We initialize each algorithm from the same point and repeat the experiments 5 times, then report the mean convergence with standard deviation as the shaded region around the mean curves. We tune the algorithms by executing a parameter sweep for their initial step-size over an interval of values which are exponentially scaled as $\{10^{-3}, 10^{-2}, ..., 10^{2}, 10^{3}\}$. After tuning the algorithms on one dataset, we run them with the same parameters for the others.



(a) a1a

(b) mushrooms

(c) a6a

(d) w1a

(e) w5a

(f) ionosphere-scale

Figure 3.2: Logistic regression with non-convex regularizer on LibSVM datasets

First, we clearly observe the difference between SGD & ADAGRAD, and the rest of the pack,

which demonstrates the superior sample complexity of VR methods in general. Among VR algorithms, there does not seem to be any concrete differences with similar convergence, except for SPIDER. The performance of ADASPIDER is on par with other VR methods, and superior to SPIDER. The unexpected behavior of SPIDER algorithm has previously been documented in Wang et al. [Wan+19]. From a technical point of view, this behavior is predominantly due to the *accuracy dependence* in the step-size, making the step-size unusually small. We had to run SPIDER beyond its prescribed setting and tune the step-size with a large initial value to make sure the algorithm makes observable progress.

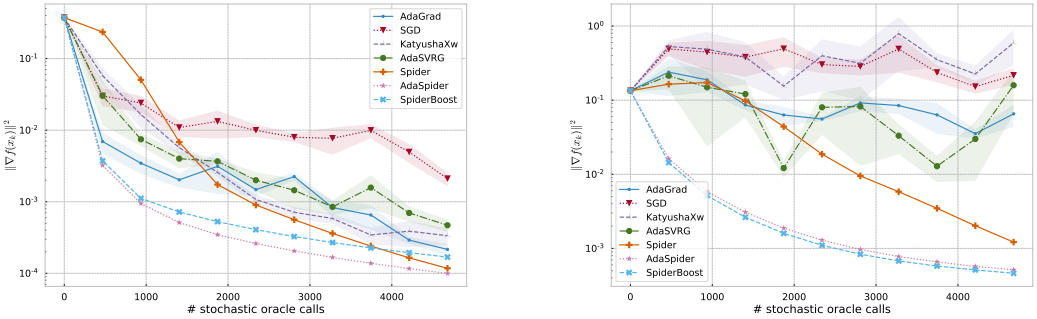
Experiments with neural networks

In our second setup, we train neural networks with our variance reduction scheme. Our focus is on standard image classification tasks trained with the cross entropy loss [Bri89; Bri90]. Denoting by C the number of classes, the considered datasets in this section consist of n pairs (a_i, b_i) where a_i is a vectorized image and $b_i \in \mathbb{R}^C$ is a one-hot encoded class label. A neural network is parameterized with weights $x \in \mathbb{R}^d$ and its output on is denoted net(x, a) $\in \mathbb{R}^C$, where a is the input image. The training of the network consists of solving the following optimization problem: $\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}\left(-b_i^\top \text{net}(x, a_i) + \texttt{logsumexp}(\text{net}(x, a_i))\right)$. This is the default setup for doing image classification and we test our algorithm on two benchmark datasets : `MNIST`[Lec+98] and `FashionMNIST`[XRV17]. We choose 3-layer fully connected network with dimensions $[28 * 28, 512, 512, 10]$. The activation function is the `ELU` [CUH16].

Initialization. The initialization of the network is a crucial component to guarantee good performance. We find that a slight modification of the `Kaiming Uniform` initialization [He+15] improves the stability of the tested variance reduction schemes. For each layer in the network with d_{in} inputs, the original method initializes the weights with independent uniform random variables with variance $\frac{1}{d_{in}}$. Our modification initializes with a smaller variance of $\frac{c_{\text{init}}}{d_{in}}$ with c_{init} in the order of 0.01. With this choice, we observed that fewer variance reductions schemes diverged, and standard algorithms like `SGD` and `AdaGrad`(for which the original method was tuned), were not penalized and performed well. This often overlooked initialization heuristic is the only "tuning" needed for `AdaSpider`.

	`MNIST` Batch Size = 32, $c_{\text{init}} = 0.03$		`FashionMNIST` Batch Size = 128, $c_{\text{init}} = 0.01$	
Algorithm	**Parameters**	**Test Accuracy**	**Parameters**	**Test Accuracy**
`AdaGrad`[DHS11]	$\eta = 0.01$, $\epsilon = 10^{-4}$	97.86	$\eta = 0.01$, $\epsilon = 10^{-4}$	86.19
`SGD`[RM51]	$\eta = 0.01$	98.11	$\eta = 0.01$	85.83
`KatyushaXw`[All18]	$\eta = 0.005$	97.93	$\eta = 0.01$	86.27
`AdaSVRG`[Dub+22]	$\eta = 0.1$	98.03	$\eta = 0.1$	86.82
`Spider`[Fan+18]	$\epsilon = 0.01$, $L = 100.0$, $n_0 = 1$	97.53	$\epsilon = 0.01$, $L = 50.0$, $n_0 = 1$	82.22
`SpiderBoost`[Wan+19; Ngu+17]	$L = 200$	97.01	$L = 120$	84.42
`AdaSpider`	$n = 60000$	97.49	$n = 60000$	84.09

Table 3.3: Algorithm parameters and test accuracies (average of 5 runs, in %)



(a) `MNIST` **(b)** `FashionMNIST`

Figure 3.3: Gradient norms throughout the epochs for image classification with neural networks (curves are averaged over 5 independent runs and the shaded region are the standard error).

Observations. We observe (Figure 3.3) that `AdaSpider` performs as well as other variance reduction methods in terms of minimizing the gradient norm. The key message here is that it does so without the need for extensive tuning. This diminished need for tuning is a welcome feature for deep learning optimization, but, often the true metric of interest is not the gradient norm, but the accuracy on unseen data, and on this metric variance reduction schemes are not yet competitive with simpler methods like `SGD`. With `AdaSpider`, the focus can go to finding the right initialization scheme and architecture to ensure good generalization without being distracted by other parameters like the step-size choice.

3.4 APPENDIX: Proofs of Chapter 3

3.4.1 Proofs of Section 3.1

Lemma 3.1.1. *Let $a_1, ..., a_n$ be a sequence of non-negative real numbers. Then, it holds that*

$$\sum_{s=1}^{n} \frac{a_i}{\sum_{s=1}^{i} a_i} \leq 1 + \log\left(1 + \sum_{s=1}^{n} a_i\right)$$

Proof. We will follow the proof steps of Levy, Yurtsever, and Cevher [LYC18] with a slight modification. The proof is due to induction.

For the base case of $n = 1$:

$$\frac{a_1}{a_1} = 1 \leq 1 + \log(1 + a_1)$$

Assume that the statement holds up to and including $n - 1 > 1$. Then, for n:

$$\sum_{s=1}^{n} \frac{a_i}{\sum_{j=1}^{i} a_j} \leq 1 + \log\left(1 + \sum_{s=1}^{n-1} a_i\right) + \frac{a_n}{\sum_{s=1}^{n} a_i} \overset{?}{\leq} 1 + \log\left(1 + \sum_{s=1}^{n} a_i\right)$$

We want to show that for any a_n, the second inequality with the question mark (?) holds. Let us define $x = \frac{a_n}{\sum_{s=1}^{n-1} a_i}$. Focusing on the second inequality and re arranging the terms we get,

$$\begin{aligned}
\frac{a_n}{\sum_{s=1}^{n} a_i} &\leq \log\left(\frac{1 + \sum_{s=1}^{n} a_i}{1 + \sum_{s=1}^{n-1} a_i}\right) \\
&= \log\left(1 + \frac{a_n}{1 + \sum_{s=1}^{n-1} a_i}\right) \\
&\leq \log\left(1 + \frac{a_n}{\sum_{s=1}^{n-1} a_i}\right)
\end{aligned}$$

Notice that

$$\begin{aligned}
\frac{a_n}{\sum_{s=1}^{n} a_i} &= \frac{a_n}{\sum_{s=1}^{n-1} a_i} \cdot \frac{\sum_{s=1}^{n-1} a_i}{\sum_{s=1}^{n} a_i} = \frac{a_n}{\sum_{s=1}^{n-1} a_i} \cdot \frac{1}{\frac{\sum_{s=1}^{n} a_i}{\sum_{s=1}^{n-1} a_i}} = \frac{a_n}{\sum_{s=1}^{n-1} a_i} \cdot \frac{1}{\left(1 + \frac{a_n}{\sum_{s=1}^{n-1} a_i}\right)} \\
&= x \frac{1}{1 + x}
\end{aligned}$$

Combining both expressions,

$$\frac{x}{1 + x} \leq \log(1 + x)$$

which *always* holds whenever $x \geq 0$. ■

Lemma 3.1.2 (Lemma 3 in [KT08]). *Let X_t be a martingale difference sequence such that $|X_t| \leq b$. Let us also define*

$$\mathbf{Var}_{t-1}(X_t) = \mathbf{Var}\left(X_t \mid \sigma(X_1, ..., X_{t-1})\right) = \mathbb{E}\left[X_t^2 \mid \sigma(X_1, ..., X_{t-1})\right],$$

and define $V_T = \sum_{t=1}^{T} \mathbf{Var}_{t-1}(X_t)$ as the sum of variances. For $\delta < 1/e$ and $T \geq 3$, it holds that

$$\mathbb{P}\left(\sum_{t=1}^{T} X_t > \max\left\{2\sqrt{V_T}, 3b\sqrt{\log(1/\delta)}\right\}\sqrt{\log(1/\delta)}\right) \leq 4\log(T)\delta \tag{3.37}$$

Proof. The proof of this lemma could be found at the beginning of the Appendix section of Kakade and Tewari [KT08], which is their Lemma 3 in the main text.

■

Theorem 3.1.1. *Let $\{X_t\}$ be a sequence generated by Algorithm 4 with $G_0 = 0$ for simplicity. Then, it holds that*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(X_t)\|^2 \leq O\left(\frac{(\Delta_1 + L)^2}{T}\right).$$

Proof (Theorem 3.1.1). In the presence of only deterministic oracle, we have $\nabla f(X_t, \xi_t) = \nabla f(X_t)$ in Eq. (3.9). By replacing the stochastic gradient with the true gradient we obtain,

$$\sum_{t=1}^{T} \|\nabla f(X_t)\|^2 \;\leq\; \frac{\Delta_{\max}}{\gamma_T} + \frac{L}{2}\sum_{t=1}^{T} \gamma_t \|\nabla f(X_t)\|^2 \;\leq\; (\Delta_{\max} + L)\sqrt{\sum_{t=1}^{T} \|\nabla f(X_t)\|^2},$$

where we obtain the final inequality using Lemma 2.1.2. Now, we show that Δ_{T+1} is bounded for any T. Using descent lemma and the update rule for X_t,

$$\begin{aligned} f(X_{t+1}) - f(X_t) &\leq \langle \nabla f(X_t), X_{t+1} - X_t \rangle + \frac{L}{2}\|X_{t+1} - X_t\|^2 \\ &\leq -\gamma_t \|\nabla f(X_t)\|^2 + \frac{L\gamma_t^2}{2}\|\nabla f(X_t)\|^2 \end{aligned}$$

Summing over $t \in [T]$, telescoping function values and re-arranging right-hand side,

$$\begin{aligned} f(X_{T+1}) - f(X_t) &\leq \sum_{t=1}^{T}\left(\frac{L\gamma_t}{2} - 1\right)\gamma_t \|\nabla f(X_t)\|^2 \\ f(X_{T+1}) - f(x^*) &\leq f(X_1) - f(x^*) + \sum_{t=1}^{T}\left(\frac{L\gamma_t}{2} - 1\right)\gamma_t \|\nabla f(X_t)\|^2 \end{aligned}$$

where $x^* = \arg\min_{x \in \mathbb{R}^d} f(x)$. Now, define $t_0 = \max\left\{t \in [T] \mid \gamma_t > \frac{2}{L}\right\}$, such that $\left(\frac{L\gamma_t}{2} - 1\right) \leq 0$ for

any $t > t_0$. Then,

$$\begin{aligned}
f(X_{T+1}) - f(x^*) &\leq \Delta_1 + \sum_{t=1}^{t_0} \left(\frac{L\gamma_t}{2} - 1\right)\gamma_t \|\nabla f(X_t)\|^2 + \sum_{t=t_0+1}^{T} \left(\frac{L\gamma_t}{2} - 1\right)\gamma_t \|\nabla f(X_t)\|^2 \\
&\leq \Delta_1 + \frac{L}{2}\sum_{t=1}^{t_0} \gamma_t^2 \|\nabla f(X_t)\|^2 && \text{(Lemma 3.1.1)} \\
&\leq \Delta_1 + \frac{L}{2}\left(1 + \log\left(1 + \sum_{t=1}^{t_0} \|\nabla f(X_t)\|^2\right)\right) && \text{(Definition of } \gamma_t\text{)} \\
&\leq \Delta_1 + \frac{L}{2}\left(1 + \log\left(1 + \frac{1}{\gamma_{t_0}^2}\right)\right) && \text{(Definition of } t_0\text{)} \\
&\leq \Delta_1 + \frac{L}{2}\left(1 + \log\left(1 + \frac{L^2}{4}\right)\right),
\end{aligned}$$

where we use the definition of t_0 and Lemma 3.1.1 for the last inequality. Since this is true for any T, the bound holds for $\Delta_{\max}$ such that $\Delta_{\max} \leq \Delta_1 + \frac{L}{2}\left(1 + \log\left(L^2/4\right)\right)$. Now, define $X = \sqrt{\sum_{t=1}^{T} \|\nabla f(X_t)\|^2}$, then the original expression reduces to $X^2 \leq (\Delta_{\max} + L)\, X$. Solving for X trivially yields

$$X \leq (\Delta_{\max} + L) \implies X^2 = \sum_{t=1}^{T} \|\nabla f(X_t)\|^2 \leq (\Delta_{\max} + L)^2 .$$

Plugging in the bound for $\Delta_{\max}$ and dividing by T gives,

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(X_t)\|^2 \leq \frac{\left(\Delta_1 + \frac{L}{2}\left(3 + \log\left(1 + \frac{L^2}{4}\right)\right)\right)^2}{T}$$

■

Proposition 3.1.1. *Using Lemma 3.1.2, with probability* $1 - 4\log(T)\delta$ *with* $\delta < 1/e$*, we have*

$$\sum_{t=1}^{T} -\langle \nabla f(X_t), \zeta_t \rangle \leq 2\sigma\sqrt{\log(1/\delta)}\sqrt{\sum_{t=1}^{T} \|\nabla f(X_t)\|^2} + 3(G^2 + G\tilde{G})\log(1/\delta).$$

Proof. We have to show that the random variable $-\langle \nabla f(X_t), \zeta_t \rangle$ is a martingale difference sequence and satisfies the conditions in Lemma 3.1.2. Recall that $\nabla f(X_t, \xi_t)$ is the stochastic gradient evaluated at X_t where ξ_t represents the randomness in oracle feedback, and we use the shorthand notation $\zeta_t = \nabla f(X_t, \xi_t) - \nabla f(X_t)$. Let us define $\mathscr{F}_t = \sigma(\xi_t, ..., \xi_1)$ as the σ-algebra generated by randomness up to, and including ξ_t. Notice that $\mathscr{F}_t$ is the natural filtration of $-\langle \nabla f(X_t), \zeta_t \rangle$. Then, we need to show that

1. $-\langle \nabla f(X_t), \zeta_t \rangle$ is integrable,
2. *martingale (difference) property* holds, $\mathbb{E}\left[-\langle \nabla f(X_t), \zeta_t \rangle | \mathscr{F}_{t-1}\right] = 0$.

First off, we show that $-\langle \nabla f(X_t), \zeta_t \rangle$ is integrable:

$$\begin{aligned}
\mathbb{E}\left[\left|\langle \nabla f(X_t), \zeta_t \rangle\right|\right] &\leq \mathbb{E}\left[\|\nabla f(X_t)\| \|\zeta_t\|\right] \\
&= \mathbb{E}\left[\|\nabla f(X_t)\|^2 + \|\zeta_t\|^2\right] \\
&\leq G^2 + \mathbb{E}\left[\mathbb{E}\left[\|\zeta_t\| | \mathscr{F}_{t-1}\right]\right] \\
&\leq G^2 + \sigma^2 < +\infty,
\end{aligned}$$

where the second inequality is due to the towering property of expectation. Then, the martingale property:

$$\begin{aligned}
\mathbb{E}\left[-\langle \nabla f(X_t), \zeta_t \rangle | \mathscr{F}_{t-1}\right] &= -\langle \nabla f(X_t), \mathbb{E}\left[\zeta_t | \mathscr{F}_{t-1}\right] \rangle \\
&= -\langle \nabla f(X_t), 0 \rangle = 0
\end{aligned}$$

Before applying Lemma 3.1.2, we need to verify that $\left|-\langle \nabla f(X_t), \zeta_t \rangle\right|$ is bounded:

$$\begin{aligned}
\left|-\langle \nabla f(X_t), \zeta_t \rangle\right| &= \left|-\langle \nabla f(X_t), \nabla f(X_t, \xi_t) - \nabla f(X_t) \rangle\right| \\
&= \left|\|\nabla f(X_t)\|^2 - \langle \nabla f(X_t), \nabla f(X_t, \xi_t) \rangle\right| \\
&\leq \|\nabla f(X_t)\|^2 + \left|-\langle \nabla f(X_t), \nabla f(X_t, \xi_t) \rangle\right| \\
&\leq \|\nabla f(X_t)\|^2 + \|\nabla f(X_t)\| \|\nabla f(X_t, \xi_t)\| \\
&\leq G^2 + G\tilde{G},
\end{aligned}$$

where we used G-Lipschitzness of f and almost sure boundedness of stochastic gradients $\nabla f(X_t, \xi_t)$. Now, we are able make the high probability statement. By Lemma 3.1.2, with probability $1 - 4\log(T)\delta$ for $\delta < 1/e$, we have

$$\begin{aligned}
\sum_{t=1}^{T} -\langle \nabla f(X_t), \zeta_t \rangle &\leq \max\left\{2\sqrt{\sum_{t=1}^{T} \mathbb{E}\left[\langle \nabla f(X_t), \zeta_t \rangle^2 | \mathscr{F}_{t-1}\right]}, 3(G^2 + G\tilde{G})\sqrt{\log(1/\delta)}\right\}\sqrt{\log(1/\delta)} \\
&\overset{(1)}{\leq} \sqrt{\log(1/\delta)}\left(2\sqrt{\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla f(X_t)\|^2 \|\zeta_t\|^2 | \mathscr{F}_{t-1}\right]} + 3(G^2 + G\tilde{G})\sqrt{\log(1/\delta)}\right) \\
&\overset{(2)}{\leq} \sqrt{\log(1/\delta)}\left(2\sqrt{\sigma^2 \sum_{t=1}^{T} \|\nabla f(X_t)\|^2} + 3(G^2 + G\tilde{G})\sqrt{\log(1/\delta)}\right) \\
&\leq 2\sigma\sqrt{\log(1/\delta)}\sqrt{\sum_{t=1}^{T} \|\nabla f(X_t)\|^2} + 3(G^2 + G\tilde{G})\log(1/\delta)
\end{aligned}$$

where we used Cauchy-Schwarz inequality for the inner product to obtain inequality (1) and bounded variance assumption to obtain (2). ■

Proposition 3.1.2. *Let $\{X_t\}$ be generated by AdaGrad for $G_0 > 0$. With probability at least*

$1-4\log(t)\delta$,

$$\Delta_{t+1} \leq \Delta_1 + 2L\left(1+\log\left(\max\{1, G_0^2\}+\tilde{G}^2 t\right)\right) + G_0^{-1}(M_1+\sigma^2)\log(1/\delta) + M_2,$$

where $M_1 = 3(G^2+G\tilde{G})$ *and* $M_2 = G_0^{-1}(2G^2+G\tilde{G})$.

Proof. We will handle this bound in two cases. First, we show the bound for AdaGrad, and then for RSAG. Indeed, the bounds for the two cases differ by a factor of constants, hence we will use the larger bound for all algorithms.

Case 1 (AdaGrad)
First off by smoothness,

$$\begin{aligned} f(X_{t+1}) - f(X_t) &\leq -\gamma_t \langle \nabla f(X_t), \nabla f(X_t, \xi_t)\rangle + \frac{L\gamma_t^2}{2}\|\nabla f(X_t,\xi_t)\|^2 \\ &= -\gamma_t \|\nabla f(X_t)\|^2 - \gamma_t\langle \nabla f(X_t), \zeta_t\rangle + \frac{L\gamma_t^2}{2}\|\nabla f(X_t,\xi_t)\|^2 \end{aligned}$$

Defining $x^* = \min_{x\in\mathbb{R}^d} f(x)$ as the global minimizer of f and summing over $t \in [T]$,

$$f(X_{T+1}) - f(x^*) \leq f(X_1) - f(x^*) + \underbrace{\sum_{t=1}^{T} -\gamma_t\|\nabla f(X_t)\|^2}_{\text{(A)}} + \underbrace{\frac{L}{2}\sum_{t=1}^{T}\gamma_t^2\|\nabla f(X_t,\xi_t)\|^2}_{\text{(B)}} + \underbrace{\sum_{t=1}^{T} -\gamma_t\langle\nabla f(X_t),\zeta_t\rangle}_{\text{(C)}} \tag{3.38}$$

Term (A) At this point, we will keep this term as it will be coupled with the sum-of-conditional-variances term which will be obtained through martingale concentration.

Term (B)

$$\begin{aligned} \frac{L}{2}\sum_{t=1}^{T}\gamma_t^2\|\nabla f(X_t,\xi_t)\|^2 &= \frac{L}{2}\sum_{t=1}^{T}\frac{\|\nabla f(X_t,\xi_t)\|^2}{G_0^2+\sum_{s=1}^{t}\|g_i\|^2} && \text{(Lemma 3.1.1)} \\ &\leq \frac{L}{2}\left(1+\log\left(\max\{1,G_0^2\}+\sum_{t=1}^{T}\|\nabla f(X_t,\xi_t)\|^2\right)\right) && \text{(Bounded gradients)} \\ &\leq \frac{L}{2}\left(1+\log\left(\max\{1,G_0^2\}+\tilde{G}^2 T\right)\right) \end{aligned}$$

Bounding term (C)

$$\sum_{t=1}^{T} -\gamma_t \langle \nabla f(X_t), \zeta_t \rangle \leq \underbrace{\sum_{t=1}^{T} -\gamma_{t-1} \langle \nabla f(X_t), \zeta_t \rangle}_{\text{(C.1)}} + \underbrace{\sum_{t=1}^{T} (\gamma_{t-1} - \gamma_t) \langle \nabla f(X_t), \zeta_t \rangle}_{\text{(C.2)}}$$

We will make use of Lemma 3.1.2 to achieve high probability bounds on term (C.1). To do so, we need to prove that $X_t = -\gamma_{t-1} \langle \nabla f(X_t), \zeta_t \rangle$ is a martingale difference sequence and validate some of its properties:

1. $-\gamma_{t-1} \langle \nabla f(X_t), \zeta_t \rangle$ is absolutely integrable:

$$\begin{aligned} \mathbb{E}[\left| -\gamma_{t-1} \langle \nabla f(X_t), \zeta_t \rangle \right|] &\leq G_0^{-1} \mathbb{E}[\left| \langle \nabla f(X_t), \zeta_t \rangle \right|] \\ &\leq G_0^{-1} \mathbb{E}[\|\nabla f(X_t)\|^2 + \|\zeta_t\|^2] \\ &\leq G_0^{-1} (G^2 + \sigma^2) < +\infty \end{aligned}$$

2. $-\gamma_{t-1} \langle \nabla f(X_t), \zeta_t \rangle$ is adapted to its natural filtration $\mathscr{F}_t = \sigma(\xi_1, ..., \xi_t)$

3. It satisfies the martingale (difference) property:

$$\mathbb{E}\left[-\gamma_{t-1} \langle \nabla f(X_t), \xi_t \rangle \mid \mathscr{F}_{t-1} \right] = -\gamma_{t-1} \langle \nabla f(X_t), \mathbb{E}\left[\zeta_t \mid \mathscr{F}_{t-1} \right] \rangle = 0$$

4. $X_t = -\gamma_{t-1} \langle \nabla f(X_t), \zeta_t \rangle$ is bounded:

$$-\gamma_{t-1} \langle \nabla f(X_t), \zeta_t \rangle \leq G_0^{-1} \left| \langle \nabla f(X_t), \zeta_t \rangle \right| \leq G_0^{-1} (\|\nabla f(X_t)\|^2 + \|\nabla f(X_t)\| \|\nabla f(X_t, \xi_t)\|) \leq G_0^{-1} (G^2 + G\tilde{G})$$

5. Conditional variance of $X_t = -\gamma_{t-1} \langle \nabla f(X_t), \zeta_t \rangle$:

$$\begin{aligned} \mathbf{Var}_{t-1}(X_t) &= \mathbb{E}\left[(\gamma_{t-1} \langle \nabla f(X_t), \zeta_t \rangle)^2 \mid \mathscr{F}_{t-1} \right] \\ &\leq G_0^{-2} \mathbb{E}\left[(\langle \nabla f(X_t), \zeta_t \rangle)^2 \mid \mathscr{F}_{t-1} \right] \\ &\leq G_0^{-2} \|\nabla f(X_t)\|^2 \mathbb{E}\left[\|\zeta_t\|^2 \mid \mathscr{F}_{t-1} \right] \\ &\leq G_0^{-2} \sigma^2 \|\nabla f(X_t)\|^2 \end{aligned}$$

Term (C.1) Now, we are at a position to apply Lemma 3.1.2 on term (C.1). With probability $1 - 4\log(T)\delta$,

$$\begin{aligned} \sum_{t=1}^{T} -\gamma_{t-1} \langle \nabla f(X_t), \zeta_t \rangle &\leq \max \left\{ 2 \sqrt{\sum_{t=1}^{T} \mathbb{E}\left[(\gamma_{t-1} \langle \tilde{g}_t, \xi_t \rangle)^2 \mid \mathscr{F}_{t-1} \right]}, 3 G_0^{-1} (G^2 + G\tilde{G}) \sqrt{\log(1/\delta)} \right\} \sqrt{\log(1/\delta)} \\ &\leq \max \left\{ 2 \sqrt{\sum_{t=1}^{T} \sigma^2 \gamma_{t-1}^2 \|\nabla f(X_t)\|^2}, 3 G_0^{-1} (G^2 + G\tilde{G}) \sqrt{\log(1/\delta)} \right\} \sqrt{\log(1/\delta)} \end{aligned}$$

$$\leq \underbrace{2\sigma\sqrt{\log(1/\delta)}\sqrt{\sum_{t=1}^{T}\gamma_{t-1}^2\|\nabla f(X_t)\|^2}}_{\text{(D)}} + 3G_0^{-1}(G^2+G\tilde{G})\log(1/\delta)$$

Term (C.2):

$$\begin{aligned}\sum_{t=1}^{T}(\gamma_{t-1}-\gamma_t)\langle\nabla f(X_t),\zeta_t\rangle &\leq \sum_{t=1}^{T}(\gamma_{t-1}-\gamma_t)\left|\langle\nabla f(X_t),\zeta_t\rangle\right| \\ &\leq (G^2+G\tilde{G})\sum_{t=1}^{T}(\gamma_{t-1}-\gamma_t) \\ &\leq (G^2+G\tilde{G})\gamma_0\end{aligned}$$

Terms (A) + (D). All the underbraced term but expression (D) either grows as $O\big(\log(T)\big)$, or is upper bounded by a constant. The worst-case growth of term (D) is $O(\sqrt{T})$, which we will keep under control via term (A).

$$\begin{aligned}\text{(A) + (D)} &\leq 2\sigma\sqrt{\log(1/\delta)}\sqrt{\sum_{t=1}^{T}\gamma_{t-1}^2\|\nabla f(X_t)\|^2} - \sum_{t=1}^{T}\gamma_t\|\nabla f(X_t)\|^2 \\ &\leq 2\sigma\sqrt{\log(1/\delta)}\sqrt{\sum_{t=1}^{T}\gamma_{t-1}^2\|\nabla f(X_t)\|^2} - G_0\sum_{t=1}^{T}\gamma_t^2\|\nabla f(X_t)\|^2 \\ &\leq 2\sigma\sqrt{\log(1/\delta)}\sqrt{\sum_{t=1}^{T}\gamma_{t-1}^2\|\nabla f(X_t)\|^2} - G_0\sum_{t=1}^{T}\gamma_{t-1}^2\|\nabla f(X_t)\|^2 + G_0\sum_{t=1}^{T}(\gamma_{t-1}^2-\gamma_t^2)\|\nabla f(X_t)\|^2 \\ &\leq 2\sigma\sqrt{\log(1/\delta)}\sqrt{\sum_{t=1}^{T}\gamma_{t-1}^2\|\nabla f(X_t)\|^2} - G_0\sum_{t=1}^{T}\gamma_{t-1}^2\|\nabla f(X_t)\|^2 + G_0G^2\gamma_0^2\end{aligned}$$

In order to characterize the growth of this expression, let us define $f(x) = 2\sigma\sqrt{\log(1/\delta)}\sqrt{x} - G_0x$, which is a concave function as its second derivative is non-positive. Now, looking at derivative of f,

$$\frac{d}{dx}f(x) = \frac{\sigma\sqrt{\log(1/\delta)}}{\sqrt{x}} - G_0,$$

which is 0 at $x = G_0^{-2}\sigma^2\log(1/\delta)$. This is indeed the point at which the function attains its maximum. For the final step of the proof, we define $Z_T = \sum_{t=1}^{T}\gamma_{t-1}^2\|\nabla f(X_t)\|^2$. Then,

$$\text{(A) + (D)} \leq f(Z_T) + G_0G^2\gamma_0^2$$

$$\leq f(G_0^{-2}\sigma^2\log(1/\delta)) + G_0G^2\gamma_0^2$$
$$= G_0^{-1}\sigma^2\log(1/\delta) + G_0G^2\gamma_0^2$$

Final bound Plugging all the expression together and setting $\gamma_0 = \gamma_1$, with probability at least $1 - 4\log(T)\delta$,

$$\begin{aligned} f(X_{T+1}) - f(x^*) \leq f(X_1) - f(x^*) + \frac{L}{2}\left(1 + \log\left(\max\{1, G_0^2\} + \tilde{G}^2 T\right)\right) \\ + G_0^{-1}(3(G^2 + G\tilde{G}) + \sigma^2)\log(1/\delta) \\ + G_0^{-1}(2G^2 + G\tilde{G}) \end{aligned}$$

Since this result holds for any T, to make it consistent with the statement of the proposition, we re-state the bound with t,

$$\begin{aligned} f(X_{t+1}) - f(x^*) \leq f(X_1) - f(x^*) + \frac{L}{2}\left(1 + \log\left(\max\{1, G_0^2\} + \tilde{G}^2 t\right)\right) \\ + G_0^{-1}(3(G^2 + G\tilde{G}) + \sigma^2)\log(1/\delta) \\ + G_0^{-1}(2G^2 + G\tilde{G}) \end{aligned}$$

Case 2 (Adaptive RSAG) For consistency, let us re-state the notation and definitions for Algorithm 5. Let $\zeta_t = \nabla f(\bar{X}_t, \xi_t) - \nabla f(\bar{X}_t)$. Again, by smoothness and the update rule for X_t sequence,

$$\begin{aligned} & f(X_{t+1}) - f(X_t) \\ & \leq \langle \nabla f(X_t), X_{t+1} - X_t \rangle + \frac{L}{2}\|X_{t+1} - X_t\|^2 \\ & = -\eta_t \langle \nabla f(\bar{X}_t), \nabla f(\bar{X}_t, \xi_t) \rangle - \eta_t \langle \nabla f(X_t) - \nabla f(\bar{X}_t), \nabla f(\bar{X}_t, \xi_t) \rangle + \frac{L\eta_t^2}{2}\|\nabla f(\bar{X}_t, \xi_t)\|^2 \\ & \qquad \text{(Cauchy-Schwarz)} \\ & = -\eta_t \|\nabla f(X_t) - \eta_t \langle \nabla f(\bar{X}_t), \zeta_t \rangle + \nabla f(\bar{X}_t)\| \|\nabla f(\bar{X}_t, \xi_t)\| + \frac{L\eta_t^2}{2}\|\nabla f(\bar{X}_t, \xi_t)\|^2 \quad \text{(Smoothness)} \\ & = -\eta_t \|\nabla f(\bar{X}_t)\|^2 - \eta_t \langle \nabla f(\bar{X}_t), \zeta_t \rangle + L\eta_t \|\bar{X}_t - X_t\| \|\nabla f(\bar{X}_t, \xi_t)\| + \frac{L\eta_t^2}{2}\|\nabla f(\bar{X}_t, \xi_t)\|^2 \\ & \qquad \text{(Young's ineq.)} \\ & = -\eta_t \|\nabla f(\bar{X}_t)\|^2 - \eta_t \langle \nabla f(\bar{X}_t), \zeta_t \rangle + \frac{L}{2}\|\bar{X}_t - X_t\|^2 + L\eta_t^2 \|\nabla f(\bar{X}_t, \xi_t)\|^2 \end{aligned}$$

Using recursive expansion of $\|\bar{X}_t - X_t\|^2$ and summing over $t \in [T]$,

$$\begin{aligned}\Delta_{T+1} &\leq \Delta_1 + \frac{L}{2}\sum_{t=1}^{T}\left[(1-\alpha_t)\Gamma_t \sum_{s=1}^{t}\frac{\alpha_s}{\Gamma_s}\frac{(\eta_s-\lambda_s)^2}{\alpha_s^2}\|\nabla f(\bar{X}_t,\xi_t)\|^2\right] \\ &\quad + \sum_{t=1}^{T} L\eta_t^2\|\nabla f(\bar{X}_t,\xi_t)\|^2 - \eta_t\|\nabla f(\bar{X}_t)\|^2 - \eta_t\langle\nabla f(\bar{X}_t),\zeta_t\rangle \\ &\leq \Delta_1 + \frac{L}{2}\sum_{t=1}^{T}\left[\sum_{s=t}^{T}(1-\alpha_s)\Gamma_s\right]\frac{\alpha_t}{\Gamma_t}\frac{(\eta_t-\lambda_t)^2}{\alpha_t^2}\|\nabla f(\bar{X}_t,\xi_t)\|^2 \\ &\quad + \sum_{t=1}^{T} L\eta_t^2\|\nabla f(\bar{X}_t,\xi_t)\|^2 - \eta_t\|\nabla f(\bar{X}_t)\|^2 - \eta_t\langle\nabla f(\bar{X}_t),\zeta_t\rangle\end{aligned}$$

First, we plug in $\alpha_t = 2/(t+1)$ and invoke Proposition 3.1.5 to obtain $\left[\sum_{s=t}^{T}(1-\alpha_s)\Gamma_s\right]\frac{\alpha_t}{\Gamma_t} \leq 2$. Recognizing that $\left|\lambda_t - \eta_t\right| = \alpha_t\gamma_t$ and $\eta_t = \gamma_t$, where we accumulate $\nabla f(\bar{X}_t,\xi_t)$ in the step-size γ_t for RSAG,

$$\Delta_{T+1} \leq \Delta_1 + \underbrace{\sum_{t=1}^{T} -\gamma_t\|\nabla f(\bar{X}_t)\|^2}_{\text{(A)}} + \underbrace{2L\sum_{t=1}^{T}\gamma_t^2\|\nabla f(\bar{X}_t,\xi_t)\|^2}_{\text{(B)}} + \underbrace{\sum_{t=1}^{T} -\gamma_t\langle\nabla f(\bar{X}_t),\zeta_t\rangle}_{\text{(C)}}$$

Observe that this expression is the same as Eq. (3.38) up to replacing $\frac{L}{2}$ in term (B) of AdaGrad with $2L$. Hence, the same bounds hold up to incorporating the aforementioned change. With probability $1 - 4\log(T)\delta$,

$$\begin{aligned}f(X_{T+1}) - f(x^*) \leq f(X_1) - f(x^*) &+ 2L\left(1 + \log\left(\max\{1, G_0^2\} + \tilde{G}^2 T\right)\right) \\ &+ G_0^{-1}(3(G^2 + G\tilde{G}) + \sigma^2)\log(1/\delta) \\ &+ G_0^{-1}(2G^2 + G\tilde{G})\end{aligned}$$

Similarly, since this holds for any T, we re-state the results with t for consistency,

$$\begin{aligned}f(X_{t+1}) - f(x^*) \leq f(X_1) - f(x^*) &+ 2L\left(1 + \log\left(\max\{1, G_0^2\} + \tilde{G}^2 t\right)\right) \\ &+ G_0^{-1}(3(G^2 + G\tilde{G}) + \sigma^2)\log(1/\delta) \\ &+ G_0^{-1}(2G^2 + G\tilde{G})\end{aligned}$$

■

Now, we are at a position to present the main hig probability convergence result for AdaGrad.

Theorem 3.1.2. *Let $\{X_t\}$ be the sequence of iterates generated by AdaGrad. Under Assump-*

tions 3.2, 3.6, 3.7, with probability at least $1-8\log(T)\delta$,

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|^2 \leq \frac{(\Delta_{max}+L)\,G_0+3(G^2+G\tilde{G})\log(1/\delta)}{T}+\frac{(\Delta_{max}+L)\,\tilde{G}+2G\sigma\sqrt{\log(1/\delta)}}{\sqrt{T}}.$$

where $\Delta_{max}\leq O\left(\Delta_1+L\log(T)+\sigma^2\log(1/\delta)\right)$.

Proof (Theorem 3.1.2). Let $\zeta_t=\nabla f(X_t,\xi_t)-\nabla f(X_t)$. By Eq. (3.10),

$$\sum_{t=1}^{T}\|\nabla f(X_t)\|^2 \leq \frac{\Delta_{\max}}{\gamma_T}+\sum_{t=1}^{T}-\langle\nabla f(X_t),\zeta_t\rangle+\frac{L}{2}\sum_{t=1}^{T}\gamma_t\|\nabla f(X_t,\xi_t)\|^2$$

Invoking Lemma 2.1.2 and plugging the bound for the term $(**)$ from Proposition 3.1.1 we achieve with probability $1-4\log(T)\delta$,

$$\begin{aligned}\sum_{t=1}^{T}\|\nabla f(X_t)\|^2 &\leq (\Delta_{\max}+L)\sqrt{G_0^2+\sum_{t=1}^{T}\|\nabla f(X_t,\xi_t)\|^2}\\ &\quad+2\sigma\sqrt{\log(1/\delta)}\sqrt{\sum_{t=1}^{T}\|\nabla f(X_t)\|^2+3(G^2+G\tilde{G})\log(1/\delta)}\\ &\leq (\Delta_{\max}+L)\sqrt{G_0^2+\tilde{G}^2T}+2G\sigma\sqrt{\log(1/\delta)}\sqrt{T}+3(G^2+G\tilde{G})\log(1/\delta)\\ &\leq (\Delta_{\max}+L)\,G_0+3(G^2+G\tilde{G})\log(1/\delta)+\left[(\Delta_{\max}+2L)\,\tilde{G}+2G\sigma\sqrt{\log(1/\delta)}\right]\sqrt{T}\end{aligned}$$

Dividing both sides by T, we achieve the bound,

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|^2 \leq \frac{(\Delta_{\max}+L)\,G_0+3(G^2+G\tilde{G})\log(1/\delta)}{T}+\frac{(\Delta_{\max}+L)\,\tilde{G}+2G\sigma\sqrt{\log(1/\delta)}}{\sqrt{T}}$$

Now, we will incorporate the high probability bound for $\Delta_{\max}$ to complete the convergence proof. Essentially, we are interested in scenarios in which both the statement of Proposition 3.1.1 and the statement of Proposition 3.1.2 holds, simultaneously, with high probability. Formally, let the statement of Proposition 3.1.1 be denoted as event A and the statement of Proposition 3.1.2 as event B. We have already proven that

$$\mathbb{P}(A)\geq 1-4\log(T)\delta \qquad \& \qquad \mathbb{P}(B)\geq 1-4\log(T)\delta$$

What we want to obtain is a lower bound to $\mathbb{P}(A\cap B)$, which is

$$\begin{aligned}\mathbb{P}(A\cap B)&=\mathbb{P}(A)+\mathbb{P}(B)-\mathbb{P}(A\cup B)\\ &\geq 1-4\log(T)\delta+1-4\log(T)\delta-\mathbb{P}(A\cup B)\\ &\geq 2-8\log(T)\delta-1=1-8\log(T)\delta,\end{aligned}$$

which is the best we could do due to the unknown extent of dependence between events A and

B. Hence, integrating the results of Proposition 3.1.2, with probability at least $1-8\log(T)\delta$,

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|^2 \leq \frac{(\Delta_{\max}+L)\,G_0+3(G^2+G\tilde{G})\log(1/\delta)}{T}+\frac{(\Delta_{\max}+L)\,\tilde{G}+2G\sigma\sqrt{\log(1/\delta)}}{\sqrt{T}}$$

where

$$\Delta_{\max}\leq \Delta_1+2L\left(1+\log\left(\max\{1,G_0^2\}+\tilde{G}^2T\right)\right)+G_0^{-1}(3(G^2+G\tilde{G})+\sigma^2)\log(1/\delta)+G_0^{-1}(2G^2+G\tilde{G})$$

■

Proposition 3.1.4. *Let $\{X_t\}$ be generated by Algorithm 5. Then, it holds that*

$$\sum_{t=1}^{T}\|\nabla f(\bar{X}_t)\|^2 \leq \frac{\Delta_{max}+2L}{\eta_T}+\frac{L}{2\eta_T}\underbrace{\sum_{t=1}^{T}\left[\sum_{s=t}^{T}(1-\alpha_s)\Gamma_s\right]\frac{\alpha_t}{\Gamma_t}\frac{(\eta_t-\lambda_t)^2}{\alpha_t^2}\|\nabla f(\bar{X}_t;\xi_t)\|^2}_{(*)}+\underbrace{\sum_{t=1}^{T}-\langle\nabla f(\bar{X}_t),\zeta_t\rangle}_{(**)}.$$

Proof. This result is due to Ghadimi and Lan [GL16] and Lan [Lan20] up to introducing adaptive step-sizes. We follow their derivations in the deterministic setting and incorporate it with our high probability analysis. Then,

$$\begin{aligned}
f(X_{t+1})-f(X_t) &\leq \langle\nabla f(X_t),X_{t+1}-X_t\rangle+\frac{L}{2}\|X_{t+1}-X_t\|^2\\
&\leq -\eta_t\langle\nabla f(X_t),\nabla f(\bar{X}_t)+\zeta_t\rangle+\frac{L\eta_t^2}{2}\|\nabla f(\bar{X}_t,\xi_t)\|^2\\
&= -\eta_t\|\nabla f(\bar{X}_t)\|^2-\eta_t\langle\nabla f(\bar{x}_t),\zeta_t\rangle-\eta_t\langle\nabla f(X_t)-\nabla f(\bar{X}_t),\nabla f(\bar{X}_t,\xi_t)\rangle+\frac{L\eta_t^2}{2}\|\nabla f(\bar{X}_t,\xi_t)\|^2\\
&\leq -\eta_t\|\nabla f(\bar{X}_t)\|^2-\eta_t\langle\nabla f(\bar{X}_t),\zeta_t\rangle+\frac{L}{2}\|\bar{X}_t-X_t\|^2+L\eta_t^2\|\nabla f(\bar{X}_t,\xi_t)\|^2
\end{aligned}$$

where we used descent lemma (Eq. (3.4)) in the first inequality, and update rule for X_{t+1} in Algorithm 5, line 4 in the second inequality. For the last line, we use Cauchy-Schwarz, apply smoothnness definition in Eq. (3.3) and finally use Young's inequality. Let us define $\Delta_t=f(X_t)-\min_{x\in\mathbb{R}^d}f(x)$ and $\Delta_{\max}=\max_{t\in[T]}\Delta_t$. Dividing both sides by η_t, rearranging, and summing over $t=1,\ldots,T$ we obtain,

$$\sum_{t=1}^{T}\|\nabla f(\bar{X}_t)\|^2\leq\sum_{t=1}^{T}\frac{1}{\eta_t}(\Delta_t-\Delta_{t+1})+\frac{L}{2}\sum_{t=1}^{T}\frac{1}{\eta_t}\|\bar{X}_t-X_t\|^2+L\sum_{t=1}^{T}\eta_t\|\nabla f(\bar{X}_t,\xi_t)\|^2+\sum_{t=1}^{T}-\langle\nabla f(\bar{X}_t),\zeta_t\rangle$$

Now, we express the term $\bar{X}_t-X_t$ recursively, as a function of gradient norms.

$$\bar{X}_t-X_t=(1-\alpha_t)\left[\tilde{X}_t-X_t\right]$$

$$
\begin{aligned}
&= (1-\alpha_t)\left[\bar{X}_{t-1} - X_{t-1} + (\eta_{t-1}-\lambda_{t-1})\nabla f(\bar{X}_{t-1},\xi_{t-1})\right] \\
&= (1-\alpha_t)\left[(1-\alpha_{t-1})(\tilde{X}_{t-1} - X_{t-1}) + (\eta_{t-1}-\lambda_{t-1})\nabla f(\bar{X}_{t-1},\xi_{t-1})\right] \\
&= (1-\alpha_t)\sum_{s=1}^{t-1}\left(\prod_{j=s+1}^{t-1}(1-\alpha_j)\right)(\eta_s-\lambda_s)\nabla f(\bar{X}_s,\xi_s) \\
&= (1-\alpha_t)\sum_{s=1}^{t-1}\frac{\Gamma_{t-1}}{\Gamma_s}(\eta_s-\lambda_s)\nabla f(\bar{X}_s,\xi_s) \\
&= (1-\alpha_t)\Gamma_{t-1}\sum_{s=1}^{t-1}\frac{\alpha_s}{\Gamma_s}\frac{(\eta_s-\lambda_s)}{\alpha_s}\nabla f(\bar{X}_s,\xi_s),
\end{aligned}
$$

Hence, by convexity of squared norm and (absolute) homogeneity of vector norms,

$$
\begin{aligned}
\|\bar{X}_t - X_t\|^2 &= \|(1-\alpha_t)\Gamma_{t-1}\sum_{s=1}^{t-1}\frac{\alpha_s}{\Gamma_s}\frac{(\eta_s-\lambda_s)}{\alpha_s}\nabla f(\bar{X}_s,\xi_s)\|^2 \\
&\le (1-\alpha_t)^2\Gamma_{t-1}\sum_{s=1}^{t-1}\frac{\alpha_s}{\Gamma_s}\frac{(\eta_s-\lambda_s)^2}{\alpha_s^2}\|\nabla f(\bar{X}_s,\xi_s)\|^2 \\
&\le (1-\alpha_t)\Gamma_t\sum_{s=1}^{t}\frac{\alpha_s}{\Gamma_s}\frac{(\eta_s-\lambda_s)^2}{\alpha_s^2}\|\nabla f(\bar{X}_s,\xi_s)\|^2
\end{aligned}
$$

Finally, we plug this in the original expression,

$$
\begin{aligned}
\sum_{t=1}^{T}\|\nabla f(\bar{X}_t)\|^2 &\le \sum_{t=1}^{T}\frac{1}{\eta_t}(\Delta_t - \Delta_{t+1}) + \frac{L}{2}\sum_{t=1}^{T}\left[(1-\alpha_t)\frac{\Gamma_t}{\eta_t}\sum_{s=1}^{t}\frac{\alpha_s}{\Gamma_s}\frac{(\eta_s-\lambda_s)^2}{\alpha_s^2}\|\nabla f(\bar{X}_s,\xi_s)\|^2\right] \\
&\quad + L\sum_{t=1}^{T}\eta_t\|\nabla f(\bar{X}_t,\xi_t)\|^2 + \sum_{t=1}^{T}-\langle\nabla f(\bar{X}_t),\zeta_t\rangle \\
&\le \frac{\Delta_1}{\eta_1} + \sum_{t=1}^{T-1}(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t})\Delta_{t+1} + \frac{L}{2}\sum_{t=1}^{T}\left[\sum_{s=t}^{T}(1-\alpha_s)\frac{\Gamma_s}{\eta_s}\right]\frac{\alpha_t}{\Gamma_t}\frac{(\eta_t-\lambda_t)^2}{\alpha_t^2}\|\nabla f(\bar{X}_t,\xi_t)\|^2 \\
&\quad + L\sum_{t=1}^{T}\frac{\|\nabla f(\bar{X}_t,\xi_t)\|^2}{\sqrt{G_0^2 + \sum_{s=1}^{t}\|\nabla f(\bar{X}_s,\xi_s)\|^2}} + \sum_{t=1}^{T}-\langle\nabla f(\bar{X}_t),\zeta_t\rangle \\
&\le \frac{\Delta_{\max}}{\eta_1} + \Delta_{\max}\sum_{t=1}^{T-1}(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}) + \frac{L}{2}\sum_{t=1}^{T}\left[\sum_{s=t}^{T}(1-\alpha_s)\frac{\Gamma_s}{\eta_s}\right]\frac{\alpha_t}{\Gamma_t}\frac{(\eta_t-\lambda_t)^2}{\alpha_t^2}\|\nabla f(\bar{X}_t,\xi_t)\|^2 \\
&\quad + 2L\sqrt{G_0^2 + \sum_{t=1}^{T}\|\nabla f(\bar{X}_t,\xi_t)\|^2} + \sum_{t=1}^{T}-\langle\nabla f(\bar{X}_t),\zeta_t\rangle \\
&\le \frac{\Delta_{\max}+2L}{\eta_T} + \frac{L}{2\eta_T}\sum_{t=1}^{T}\underbrace{\left[\sum_{s=t}^{T}(1-\alpha_s)\Gamma_s\right]\frac{\alpha_t}{\Gamma_t}}_{(*)}\frac{(\eta_t-\lambda_t)^2}{\alpha_t^2}\|\nabla f(\bar{X}_t,\xi_t)\|^2 + \underbrace{\sum_{t=1}^{T}-\langle\nabla f(\bar{X}_t),\zeta_t\rangle}_{(**)}.
\end{aligned}
$$

We rearranged the summations to obtain the second inequality, used the assumption that $\Delta_t \le \Delta_{\max}$ for any t together with Lemma 2.1.2, and we telescope the first summation on the right hand side to obtain the result.

■

Next, we provide the proof for term $(*)$ in Proposition 3.1.4.

Proposition 3.1.5. *We have*

$$\left[\sum_{s=t}^{T}(1-\alpha_s)\Gamma_s\right]\frac{\alpha_t}{\Gamma_t} \leq \begin{cases} 2 & \text{if } \alpha_t = \frac{2}{t+1}; \\ \log(T+1) & \text{if } \alpha_t = \frac{1}{t}. \end{cases}$$

Proof. First, we begin with the weighted averaging setting, i.e., $\alpha_t = 2/(t+1)$. Using the recursive definition of Γ, one could easily show that for any $\alpha_t \in (0,1)$,

$$\sum_{s=1}^{t}\frac{\alpha_s}{\Gamma_s} = \frac{1}{\Gamma_t} \quad \Longrightarrow \quad \Gamma_t\sum_{s=1}^{t}\frac{\alpha_s}{\Gamma_s} = 1.$$

Defining $A_t = \sum_{s=1}^{t} s = \frac{t(t+1)}{2}$ and $A_0 = 1$, we have that $\alpha_t = \frac{t}{A_t}$ and

$$\Gamma_t = \prod_{s=1}^{t}(1-\alpha_s) = \prod_{s=1}^{t}(1-\frac{s}{A_s}) = \prod_{s=1}^{t}\frac{A_{s-1}}{A_s} = \frac{1}{A_t}$$

Hence, we can express term $(*)$ as

$$\begin{aligned}\left[\sum_{s=t}^{T}(1-\alpha_s)\Gamma_s\right]\frac{\alpha_t}{\Gamma_t} &\leq \left[\sum_{s=t}^{T}\frac{1}{A_s}\right]t \\ &= \left[2\sum_{s=t}^{T}\frac{1}{s(s+1)}\right]t \\ &= \left[2\sum_{s=t}^{T}\frac{1}{s}-\frac{1}{s+1}\right]t \\ &= 2\left(\frac{1}{t}-\frac{1}{T+1}\right)t \\ &\leq 2\end{aligned}$$

For the uniform averaging setting with $\alpha_t = \frac{1}{t}$, for $t > 1$,

$$\Gamma_t = \prod_{s=1}^{t}(1-\alpha_s) = \prod_{i=2}^{t}\frac{k-1}{k} = \frac{1}{k}$$

Hence, again for $t > 1$,

$$\left[\sum_{s=t}^{T}(1-\alpha_s)\Gamma_s\right]\frac{\alpha_t}{\Gamma_t} = \sum_{s=t}^{T}\frac{1}{k} \leq \sum_{k=2}^{T}\frac{1}{k} \leq \log(T+1),$$

where the last inequality is due to that fact that integral of $f(x) = 1/x$ over the range $[1, k]$ upper bounds the summation above.

■

Finally, we conclude with the high probability convergence theorem for RSAG.

Theorem 3.1.4. *Let $\{X_t\}$ be the sequence generated by adaptive RSAG. Under Assumptions 3.2, 3.6, 3.7, with probability $1 - 8\log(T)\delta$,*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\bar{X}_t)\|^2 \leq \frac{G_0(\Delta_{max} + 3L + L\log(\max\{1, G_0^{-2}\} + \tilde{G}^2T)) + 3(G^2 + G\tilde{G})\log(1/\delta))}{T} + \frac{\tilde{G}(\Delta_{max} + 3L + L\log(\max\{1, G_0^{-2}\} + \tilde{G}^2T)) + 2G\sigma\sqrt{\log(1/\delta)}}{\sqrt{T}},$$

where $\Delta_{max} \leq O\left(\Delta_1 + L\log(T) + \sigma^2\log(1/\delta)\right)$.

Proof. Again by Proposition 3.1.4,

$$\sum_{t=1}^{T}\|\nabla f(\bar{X}_t)\|^2 \leq \frac{\Delta_{\max} + 2L}{\eta_T} + \frac{L}{2\eta_T}\sum_{t=1}^{T}\underbrace{\left[\sum_{s=t}^{T}(1-\alpha_s)\Gamma_s\right]\frac{\alpha_t}{\Gamma_t}}_{(*)}\frac{(\eta_t - \lambda_t)^2}{\alpha_t^2}\|\nabla f(\bar{X}_t, \xi_t)\|^2 + \underbrace{\sum_{t=1}^{T} -\langle\nabla f(\bar{X}_t), \zeta_t\rangle}_{(**)}.$$

Recall that we use weighted averaging and the particular step-size choices; $\eta_t = \gamma_t$ and $\lambda_t = (1+\alpha_t)\gamma_t$ where γ_t is defined as in Eq. (3.12). Combining with the previous expression we get

$$\sum_{t=1}^{T}\|\nabla f(\bar{X}_t)\|^2 \leq \frac{\Delta_{\max} + 2L}{\gamma_T} + \frac{L}{2\gamma_T}\sum_{t=1}^{T}\underbrace{\left[\sum_{s=t}^{T}(1-\alpha_s)\Gamma_s\right]\frac{\alpha_t}{\Gamma_t}}_{(*)}\gamma_t^2\|\nabla f(\bar{X}_t, \xi_t)\|^2 + \underbrace{\sum_{t=1}^{T} -\langle\nabla f(\bar{X}_t), \zeta_t\rangle}_{(**)}.$$

We introduce the bounds in Proposition 3.1.5 and Proposition 3.1.1 for the respective marked term,

$$\begin{aligned}\sum_{t=1}^{T}\|\nabla f(\bar{X}_t)\|^2 &\leq \frac{\Delta_{\max} + 2L + L\sum_{t=1}^{T}\gamma_t^2\|\nabla f(\bar{X}_t, \xi_t)\|^2}{\gamma_T} \\ &\quad + 2\sigma\sqrt{\log(1/\delta)}\sqrt{\sum_{t=1}^{T}\|\nabla f(\bar{X}_t)\|^2} + 3(G^2 + G\tilde{G})\log(1/\delta) \\ &\leq \frac{\Delta_{\max} + 3L + L\log\left(\max\{1, G_0^{-2}\} + \sum_{t=1}^{T}\|\nabla f(\bar{X}_t, \xi_t)\|^2\right)}{\gamma_T}\end{aligned}$$

$$
\begin{aligned}
&+2G\sigma\sqrt{\log(1/\delta)}\sqrt{T}+3(G^2+G\tilde{G})\log(1/\delta)\\
&\leq\left(\Delta_{\max}+3L+L\log\left(\max\{1,G_0^{-2}\}+\sum_{t=1}^{T}\|\nabla f(\bar{X}_t,\xi_t)\|^2\right)\right)\sqrt{G_0^2+\sum_{t=1}^{T}\|\nabla f(\bar{X}_t,\xi_t)\|^2}\\
&+2G\sigma\sqrt{\log(1/\delta)}\sqrt{T}+3(G^2+G\tilde{G})\log(1/\delta)\\
&\leq\left(\tilde{G}\left(\Delta_{\max}+3L+L\log\left(\max\{1,G_0^{-2}\}+\tilde{G}^2T\right)\right)+2G\sigma\sqrt{\log(1/\delta)}\right)\sqrt{T}\\
&+G_0\left(\Delta_{\max}+3L+L\log\left(\max\{1,G_0^{-2}\}+\tilde{G}^2T\right)\right)+3(G^2+G\tilde{G})\log(1/\delta)
\end{aligned}
$$

where we used Lemma 3.1.1 in the second inequality, while boundedness of $\nabla f(\bar{X}_t)$ and almost sure boundedness of $\nabla f(\bar{X}_t,\xi_t)$ in the last line. Dividing both sides by T, and using the same argument as in the proof of Theorem 3.1.2, with probability at least $1-8\log(T)\delta$,

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\bar{X}_t)\|^2\leq&\frac{G_0\left(\Delta_{\max}+3L+L\log\left(\max\{1,G_0^{-2}\}+\tilde{G}^2T\right)\right)+3(G^2+G\tilde{G})\log(1/\delta))}{T}\\
&+\frac{\tilde{G}\left(\Delta_{\max}+3L+L\log\left(\max\{1,G_0^{-2}\}+\tilde{G}^2T\right)\right)+2G\sigma\sqrt{\log(1/\delta)}}{\sqrt{T}}
\end{aligned}
$$

where

$$
\Delta_{\max}\leq\Delta_1+2L\left(1+\log\left(\max\{1,G_0^2\}+\tilde{G}^2T\right)\right)+G_0^{-1}(3(G^2+G\tilde{G})+\sigma^2)\log(1/\delta)+G_0^{-1}(2G^2+G\tilde{G})
$$

■

In this part of the appendix, we focus on the sub-Gaussian noise model and proofs of the relevant results. Specifically, we present the proof of Proposition 3.1.3 and Theorem 3.1.3 along with the Lemmas that we will require in the proofs. We first prove a bound on $\Delta_{\max}$ and then show noise-adaptive rates for AdaGrad; we will argue about high probability bounds on objective sub-optimality under sub-Gaussian assumption. First, we will present Lemma 1 from Li and Orabona [LO20], as well as our modified version of it that we use in our derivations.

Lemma 3.1.3. *Let $Z_1,\cdots,Z_T$ be a martingale difference sequence (MDS) with respect to random vectors $\xi_1,\cdots,\xi_T$ and Y_t be a sequence of random variables which is $\sigma(\xi_1,\cdots,\xi_{t-1})$-measurable. Given that $\mathbb{E}\left[\exp(Z_t^2/Y_t^2)\mid\sigma(\xi_1,\cdots\xi_{t-1})\right]\leq\exp(1)$, for any $\lambda>0$ and $\delta\in(0,1)$ with probability at least $1-\delta$,*

$$
\sum_{t=1}^{T}Z_t\leq\frac{3}{4}\lambda\sum_{t=1}^{T}Y_t^2+\frac{1}{\lambda}\log(1/\delta)
$$

Next, we present a slightly modified version of the above lemma. Its proof follows the same lines with Lemma 3.1.3 up to replacing Y_t with a deterministic quantity, selecting a particular choice of λ and dealing with the MDS Z_t itself rather than its square, Z_t^2.

Lemma 3.4.1. *Let $Z_1, \cdots, Z_T$ be a martingale difference sequence (MDS) with respect to random vectors $\xi_1, \cdots, \xi_T$ and $\sigma^2 \in \mathbb{R}$ such that $\mathbb{E}\left[\exp(Z_t/\sigma^2) \mid \sigma(\xi_1, \cdots \xi_{t-1})\right] \leq 1$. Then, with probability as least $1-\delta$,*

$$\sum_{t=1}^{T} Z_t \leq \sigma^2 \log(1/\delta)$$

We will also make use of another relevant result (Lemma 5 in Li and Orabona [LO20]) regarding the probabilistic behavior of maximum over norms of noise vectors.

Lemma 3.4.2 (Lemma 5 in Li and Orabona [LO20])**.** *Under assumptions as in Eq. (3.5) and (3.11), let $\zeta_t = \nabla f(X_t, \xi_t) - \nabla f(X_t)$. For $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\max_{1 \leq t \leq T} \|\zeta_t\|^2 \leq \sigma^2 \log\left(\frac{eT}{\delta}\right)$$

We begin by the high probability bound on the function sub-optimality.

Proposition 3.1.3. *Let $\{X_t\}$ be generated by AdaGrad and define $\Delta_t = f(X_t) - \min_{x \in \mathbb{R}^d} f(x)$. Under sub-Gaussian noise assumption as in Eq. (3.11), with probability at least $1-3\delta$,*

$$\begin{aligned}\Delta_{t+1} \leq \Delta_1 + 3G_0^{-1}G^2 + 2G_0^{-1}\sigma^2 \log\left(\frac{et}{\delta}\right) + \frac{3}{4G_0}\sigma^2 \log(1/\delta) \\ + \frac{L}{2}\left(1 + \log\left(\max\{1, G_0^2\} + 2G^2 t + 2\sigma^2 t \log\left(\frac{et}{\delta}\right)\right)\right).\end{aligned}$$

Proof. Using the initial steps of the proof in the original derivation,

$$\Delta_{T+1} \leq \Delta_1 + \underbrace{\sum_{t=1}^{T} -\gamma_t \|\nabla f(X_t)\|^2}_{\text{(A)}} + \underbrace{\sum_{t=1}^{T} -\gamma_t \langle \nabla f(X_t), \zeta_t \rangle}_{\text{(B)}} + \underbrace{\frac{L}{2}\sum_{t=1}^{T} \gamma_t^2 \|\nabla f(X_t, \xi_t)\|^2}_{\text{(C)}}$$

Term (A) + (B). In order to deal with measurability issues, we will divide term (B) into two parts:

$$\begin{aligned}\sum_{t=1}^{T} -\gamma_t \langle \nabla f(X_t), \xi_t \rangle &= \sum_{t=1}^{T} -\gamma_{t-1} \langle \nabla f(X_t), \zeta_t \rangle + \sum_{t=1}^{T} (\gamma_{t-1} - \gamma_t) \langle \nabla f(X_t), \zeta_t \rangle \\ &\overset{(1)}{\leq} \sum_{t=1}^{T} -\gamma_{t-1} \langle \nabla f(X_t), \zeta_t \rangle + 2(G^2 + \max_{1 \leq t \leq T} \|\zeta_t\|^2) \sum_{t=1}^{T} (\gamma_{t-1} - \gamma_t) \\ &\leq \sum_{t=1}^{T} -\gamma_{t-1} \langle \nabla f(X_t), \zeta_t \rangle + 2(G^2 + \max_{1 \leq t \leq T} \|\zeta_t\|^2)\gamma_0\end{aligned}$$

where we used Cauchy-Schwarz together with Young's inequality to obtain inequality (1) and telescoped in the last line. Moreover, we pick $\eta_0 \geq \eta_1$ to make sure monotonicity. Without loss of generality, a natural choice would be $\gamma_0 = G_0^{-1}$, which aligns with the definition in Algorithm 4. By Lemma 3.4.2, with probability at least $1-\delta$,

$$\sum_{t=1}^{T} -\gamma_t \langle \nabla f(X_t), \zeta_t \rangle \leq \sum_{t=1}^{T} -\gamma_{t-1} \langle \nabla f(X_t), \zeta_t \rangle + 2G_0^{-1} \left(G^2 + \sigma^2 \log\left(\frac{eT}{\delta}\right)\right)$$

Now, we will invoke Lemma 3.1.3 on the term $\sum_{t=1}^{T} -\gamma_{t-1} \langle \nabla f(X_t), \zeta_t \rangle$ by setting $Z_t = -\gamma_{t-1} \langle \nabla f(X_t), \zeta_t \rangle$, $Y_t^2 = \gamma_{t-1}^2 \|\nabla f(X_t)\|^2 \sigma^2$, with probability at least $1-\delta$,

$$\begin{aligned}\sum_{t=1}^{T} -\gamma_{t-1} \langle \nabla f(X_t), \zeta_t \rangle &\leq \frac{3}{4}\lambda \sum_{t=1}^{T} Y_t^2 + \frac{1}{\lambda}\log(1/\delta) \\ &= \frac{3}{4}\lambda\sigma^2 \sum_{t=1}^{T} \gamma_{t-1}^2 \|\nabla f(X_t)\|^2 + \frac{1}{\lambda}\log(1/\delta)\end{aligned}$$

Now, summing up the expression above with term (A) and leaving $\frac{1}{\lambda}\log(1/\delta)$ aside for now,

$$\begin{aligned}\frac{3}{4}\lambda\sigma^2 \sum_{t=1}^{T} \gamma_{t-1}^2 \|\nabla f(X_t)\|^2 - \sum_{t=1}^{T} \gamma_t \|\nabla f(X_t)\|^2 & \\ \leq \frac{3}{4}\lambda\sigma^2 \sum_{t=1}^{T} \gamma_{t-1}^2 \|\nabla f(X_t)\|^2 - G_0 \sum_{t=1}^{T} \gamma_t^2 \|\nabla f(X_t)\|^2 & \\ \leq \frac{3}{4}\lambda\sigma^2 \sum_{t=1}^{T} \eta_{t-1}^2 \|\nabla f(X_t)\|^2 - G_0 \sum_{t=1}^{T} \gamma_{t-1}^2 \|\nabla f(X_t)\|^2 + G_0 \sum_{t=1}^{T} \left(\gamma_{t-1}^2 - \gamma_t^2\right) \|\nabla f(X_t)\|^2 & \\ \leq \left(\frac{3}{4}\lambda\sigma^2 - G_0\right) \sum_{t=1}^{T} \gamma_{t-1}^2 \|\nabla f(X_t)\|^2 + G_0 G^2 \gamma_0^2 &\end{aligned}$$

where we used $G_0\gamma_t^2 \leq \gamma_t$ in the first inequality and added/subtracted $\sum_{t=1}^{T} \gamma_{t-1}^2 \|\nabla f(X_t)\|^2$ in the second inequality. Since we have a free variable to choose, λ, we could set it to $\lambda = \frac{4G_0}{3\sigma^2}$ to obtain,

$$\frac{3}{4}\lambda\sigma^2 \sum_{t=1}^{T} \gamma_{t-1}^2 \|\nabla f(X_t)\|^2 - \sum_{t=1}^{T} \gamma_t \|\nabla f(X_t)\|^2 \leq G_0^{-1} G^2$$

Hence, summing up all the expressions together, with probability at least $1-2\delta$,

$$(A) + (B) \leq 3G_0^{-1}G^2 + 2G_0^{-1}\sigma^2 \log\left(\frac{eT}{\delta}\right) + \frac{3}{4G_0}\sigma^2 \log(1/\delta)$$

Term (C). This term is easy to prove using online learning lemmas as we did previously, but we introduce a slight change in order to avoid bounded stochastic gradient assumption.

$$\begin{aligned}\frac{L}{2}\sum_{t=1}^{T}\gamma_t^2\|\nabla f(X_t,\xi_t)\|^2 &\leq \frac{L}{2}\left(1+\log\left(\max\{1,G_0^2\}+\sum_{t=1}^{T}\|\nabla f(X_t,\xi_t)\|^2\right)\right)\\ &\leq \frac{L}{2}\left(1+\log\left(\max\{1,G_0^2\}+2\sum_{t=1}^{T}\|\nabla f(X_t)\|^2+2\sum_{t=1}^{T}\|\zeta_t\|^2\right)\right)\\ &\leq \frac{L}{2}\left(1+\log\left(\max\{1,G_0^2\}+2G^2T+2\left(\max_{1\leq t\leq T}\|\zeta_t\|^2\right)T\right)\right)\end{aligned}$$

We invoked Lemma 3.1.1 to obtain the first inequality. Once again via Lemma 3.4.2, with probability at least $1-\delta$,

$$\frac{L}{2}\sum_{t=1}^{T}\gamma_t^2\|\nabla f(X_t,\xi_t)\|^2 \leq \frac{L}{2}\left(1+\log\left(\max\{1,G_0^2\}+2G^2T+2\sigma^2T\log\left(\frac{eT}{\delta}\right)\right)\right)$$

Finally, merging all the expression, with probability at least $1-3\delta$,

$$\begin{aligned}\Delta_{T+1} &\leq \Delta_1+3G_0^{-1}G^2+2G_0^{-1}\sigma^2\log\left(\frac{eT}{\delta}\right)+\frac{3}{4G_0}\sigma^2\log(1/\delta)\\ &\quad+\frac{L}{2}\left(1+\log\left(\max\{1,G_0^2\}+2G^2T+2\sigma^2T\log\left(\frac{eT}{\delta}\right)\right)\right)\\ &= O\left(\Delta_1+\sigma^2\log\left(\frac{eT}{\delta}\right)+\sigma^2\log(1/\delta)+L\log\left(T+\sigma^2T\log\left(\frac{eT}{\delta}\right)\right)\right)\end{aligned}$$

■

Now, we are at a position to prove noise-adaptive bounds.

Theorem 3.1.3. *Let $\{X_t\}$ be generated by AdaGrad and define $\Delta_t = f(X_t)-\min_{x\in\mathbb{R}^d}f(x)$. Under sub-Gaussian noise assumption as in Eq. (3.11) and considering high probability boundedness of Δ_{max} due to Proposition 3.1.3, with probability at least $1-5\delta$,*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|^2 \leq \frac{32\left(\Delta_{max}+L\right)^2+8\left(\Delta_{max}+L\right)\left(G_0+\sigma\sqrt{2\log(1/\delta)}\right)+8\sigma^2\log(1/\delta)}{T}+\frac{8\sqrt{2}\left(\Delta_{max}+L\right)\sigma}{\sqrt{T}}.$$

Proof. We take off from the same step of the original analysis by defining $\zeta_t=\nabla f(X_t,\xi_t)-\nabla f(X_t)$,

$$\begin{aligned}\sum_{t=1}^{T}\|\nabla f(X_t)\|^2 &\leq \frac{\Delta_{\max}}{\gamma_T}+\sum_{t=1}^{T}-\langle\nabla f(X_t),\zeta_t\rangle+\frac{L}{2}\sum_{t=1}^{T}\gamma_t\|\nabla f(X_t,\xi_t)\|^2\\ &\leq (\Delta_{\max}+L)\sqrt{G_0^2+\sum_{t=1}^{T}\|\nabla f(X_t,\xi_t)\|^2}+\sum_{t=1}^{T}-\langle\nabla f(X_t),\zeta_t\rangle\end{aligned}$$

$$\leq (\Delta_{\max}+L)\sqrt{G_0^2+2\sum_{t=1}^{T}\left(\|\nabla f(X_t)\|^2+\|\zeta_t\|^2+\sigma^2-\sigma^2\right)}+\sum_{t=1}^{T}-\langle\nabla f(X_t),\zeta_t\rangle$$

$$\leq (\Delta_{\max}+L)\left(G_0+\sqrt{2\sum_{t=1}^{T}\|\nabla f(X_t)\|^2}+\left(\max\Big\{0,\underbrace{2\sum_{t=1}^{T}(\|\zeta_t\|^2-\sigma^2)}_{(*)}\Big\}\right)^{1/2}+\sigma\sqrt{2T}\right)$$

$$+\underbrace{\sum_{t=1}^{T}-\langle\nabla f(X_t),\zeta_t\rangle}_{(**)}$$

We already showed that $-\langle\nabla f(X_t),\zeta_t\rangle$ is a MDS. Similarly, we could show that martingale property holds for $\|\zeta_t\|^2-\sigma^2$,

$$\mathbb{E}\left[\,\|\zeta_t\|^2-\sigma^2\mid\sigma(\xi_1,\cdots,\xi_{t-1})\,\right]=\mathbb{E}\left[\,\|\zeta_t\|^2\mid\sigma(\xi_1,\cdots,\xi_{t-1})\,\right]-\sigma^2\leq\sigma^2-\sigma^2=0. \tag{3.39}$$

Lemma 3.4.1 immediately implies for term $(*)$ that with probability at least $1-\delta$,

$$\sum_{t=1}^{T}(\|\zeta_t\|^2-\sigma^2)\leq\sigma^2\log(1/\delta)$$

For term $(**)$, we apply Lemma 3.1.3 with $Y_t^2=\sigma^2\|\nabla f(X_t)\|^2$ and $\lambda=1/\sigma^2$ to obtain with probability at least $1-\delta$,

$$\sum_{t=1}^{T}-\langle\nabla f(X_t),\zeta_t\rangle\leq\frac{3}{4}\sum_{t=1}^{T}\|\nabla f(X_t)\|^2+\sigma^2\log(1/\delta)$$

Plugging these values in and re-arranging,

$$\frac{1}{4}\sum_{t=1}^{T}\|\nabla f(X_t)\|^2\leq\sqrt{2}\,(\Delta_{\max}+L)\sqrt{\sum_{t=1}^{T}\|\nabla f(X_t)\|^2}+(\Delta_{\max}+L)\left(G_0+\sigma\sqrt{2\log(1/\delta)}+\sigma\sqrt{2T}\right)+\sigma^2\log(1/\delta)$$

We will conclude our proof by treating the above inequality as a quadratic inequality with respect to $x=\sqrt{\sum_{t=1}^{T}\|\nabla f(X_t)\|^2}$. Defining $c=(\Delta_{\max}+L)\left(G_0+\sigma\sqrt{2\log(1/\delta)}+\sigma\sqrt{2T}\right)+\sigma^2\log(1/\delta)$,

$$x^2-4\sqrt{2}\,(\Delta_{\max}+L)\,x-4c\leq 0,$$

where the roots of the inequality are

$$x=\frac{4\sqrt{2}\,(\Delta_{\max}+L)\pm\sqrt{32\,(\Delta_{\max}+L)^2+16\,(\Delta_{\max}+L)\left(G_0+\sigma\sqrt{2\log(1/\delta)}+\sigma\sqrt{2T}\right)+16\sigma^2\log(1/\delta)}}{2}$$

Since $x > 0$ by default, we will take into account the positive root above, which yields,

$$\sum_{t=1}^{T} \|\nabla f(X_t)\|^2 \leq 4\left(\sqrt{2}\left(\Delta_{\max}+L\right)+\sqrt{\left(\Delta_{\max}+L\right)\left(G_0+2\left(\Delta_{\max}+L\right)+\sigma\sqrt{2\log(1/\delta)}+\sigma\sqrt{2T}\right)+\sigma^2\log(1/\delta)}\right)^2$$

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(X_t)\|^2 \leq \frac{32\left(\Delta_{\max}+L\right)^2+8\left(\Delta_{\max}+L\right)\left(G_0+\sigma\sqrt{2\log(1/\delta)}\right)+8\sigma^2\log(1/\delta)}{T}+\frac{8\sqrt{2}\left(\Delta_{\max}+L\right)\sigma}{\sqrt{T}}$$

■

3.4.2 Proofs of Section 3.2

We first present proof of the following lemma, which provides us with a departure point for our STORM+ and its simplified version.

Lemma 3.2.2. *Let $\{X_t\}$ be generated by* STORM+ *(Algorithm 6) under the assumptions in Theorem 3.2.1. Then, it holds that*

$$\sum_{t=1}^{T}\|\bar{g}_t\|^2 \le \sum_{t=1}^{T}\|\epsilon_t\|^2 + 2Ba_{T+1}^{-1/3}(\sum_{t=1}^{T}\|d_t\|^2)^{1/3} + \frac{3}{2}L(\sum_{t=1}^{T}\|d_t\|^2)^{2/3}$$

Proof. Using smoothness together with the update rule implies,

$$\begin{aligned}\Delta_{t+1}-\Delta_t = f(X_{t+1})-f(X_t) &\le -\gamma_t\bar{g}_t^\top d_t + \frac{L\gamma_t^2}{2}\|d_t\|^2\\ &= -\gamma_t\|\bar{g}_t\|^2 - \gamma_t\bar{g}_t^\top\epsilon_t + \frac{L\gamma_t^2}{2}\|d_t\|^2\\ &\le -\gamma_t\|\bar{g}_t\|^2 + \frac{\gamma_t}{2}\|\bar{g}_t\|^2 + \frac{\gamma_t}{2}\|\epsilon_t\|^2 + \frac{L\gamma_t^2}{2}\|d_t\|^2\,,\end{aligned}$$

where we defined $\Delta_t := f(x_t)-f(x^*)$. The second line above uses $d_t = \bar{g}_t+\epsilon_t$, and the third line uses $z^\top y \le \frac{1}{2}(\|z\|^2+\|y\|^2)$.

Re-arranging the above we get,

$$\|\bar{g}_t\|^2 \le \|\epsilon_t\|^2 + \frac{2}{\gamma_t}(\Delta_t-\Delta_{t+1}) + L\gamma_t\|d_t\|^2$$

Summing over t gives,

$$\begin{aligned}\sum_{t=1}^{T}\|\bar{g}_t\|^2 &\le \sum_{t=1}^{T}\|\epsilon_t\|^2 - \frac{2}{\gamma_T}\Delta_{T+1} + 2\sum_{t=1}^{T}(\frac{1}{\gamma_t}-\frac{1}{\eta_{t-1}})\Delta_t + L\sum_{t=1}^{T}\gamma_t\|d_t\|^2\\ &\le \sum_{t=1}^{T}\|\epsilon_t\|^2 + 2B\sum_{t=1}^{T}(\frac{1}{\gamma_t}-\frac{1}{\eta_{t-1}}) + L\sum_{t=1}^{T}\frac{\|d_t\|^2}{(\sum_{i=1}^{t}\|d_t\|^2)^{1/3}}\\ &\le \sum_{t=1}^{T}\|\epsilon_t\|^2 + 2B\frac{1}{\gamma_T} + \frac{3}{2}L(\sum_{t=1}^{T}\|d_t\|^2)^{2/3}\\ &\le \sum_{t=1}^{T}\|\epsilon_t\|^2 + 2B(\sum_{t=1}^{T}\|d_t\|^2/a_{t+1})^{1/3} + \frac{3}{2}L(\sum_{t=1}^{T}\|d_t\|^2)^{2/3}\\ &\le \sum_{t=1}^{T}\|\epsilon_t\|^2 + 2B(1/a_{T+1})^{1/3}(\sum_{t=1}^{T}\|d_t\|^2)^{1/3} + \frac{3}{2}L(\sum_{t=1}^{T}\|d_t\|^2)^{2/3} \qquad (3.40)\end{aligned}$$

The second line uses $\Delta_t \in [0,B]$, the third line uses Lemma 3.4.3, and the last line uses the fact that a_t is monotonically decreasing. ■

Note that the proof of Equation (3.28) directly follows from Lemma 3.2.2 by taking $a_{T+1} = 1/T^{2/3}$.

Before we begin with the proof of the main result in Theorem 3.2.1, we present a set of numerical results, which are critical in handling the adaptive step-size and momentum parameters under different formulations; different range of exponents of the summations in the denominator, variation in the limits of the summation etc.

Lemma 3.4.3. *Let $b_1 > 0$, $b_2, ..., b_n \geq 0$ be a sequence of real numbers, $p \in (0,1)$ be a real number.*

$$\sum_{s=1}^{n} \frac{b_i}{\left(\sum_{j=1}^{i} b_j\right)^p} \leq \frac{1}{1-p}\left(\sum_{s=1}^{n} b_i\right)^{1-p}$$

Proof. We will prove the lemma by induction on n. The proof relies on the arguments in [MS10] and generalizes it for any $p \in (0,1)$. For the base case of $n = 1$, we can easily show that the hypothesis holds.

$$\frac{b_1}{b_1^p} = b_1^{1-p} \leq \frac{1}{1-p} b_1^{1-p}$$

Now, assuming that the hypothesis holds for some arbitrary number $n-1 > 1$, we want to show that it holds for n, too. Let us define $Z = \sum_{t=1}^{n} b_t$ and $x = b_n$. Then, using the inductive hypothesis for $n-1$,

$$\begin{aligned}\sum_{t=1}^{n} \frac{b_n}{(\sum_{i=1}^{t} b_i)^p} &\leq \frac{1}{1-p}\left(\sum_{t=1}^{n-1} b_t\right)^{1-p} + \frac{b_n}{(\sum_{t=1}^{n} b_t)^p} \\ &= \frac{1}{1-p}(Z-x)^{1-p} + \frac{x}{Z^p}\end{aligned}$$

Let us denote $h(x) = \frac{1}{1-p}(Z-x)^{1-p} + \frac{x}{Z^p}$ is concave in x. What we need to show is that, for any choice of allowable x, $h(x) \leq \frac{1}{1-p} Z^{1-p}$. Specifically, we want to prove that

$$\max_{0 \leq x < Z} h(x) \leq \frac{1}{1-p} Z^{1-p}$$

First, observe that $h(x)$ is a concave function, hence at the maximum the derivative evaluates to zero. Our aim is to find such x. Taking derivative wrt x,

$$\frac{dh(x)}{dx} = \frac{1}{Z^p} - \frac{1}{(Z-x)^p},$$

which evaluates to zero when $x = 0$. Hence,

$$\max_{0 \leq x < Z} h(x) = h(0) = \frac{1}{1-p} Z^{1-p} = \frac{1}{1-p}\left(\sum_{t=1}^{n} b_t\right)^{1-p}$$

which implies that the hypothesis is true:

$$\sum_{t=1}^{n} \frac{b_t}{\left(\sum_{i=1}^{t} b_i\right)^p} \leq \frac{1}{1-p}\left(\sum_{t=1}^{n} b_t\right)^{1-p}.$$

■

Next lemma forms basis of the proof of Lemma 3.2.3 in the main text, which has a sum that lag one iteration behind. We first prove a version of it iin the next lemma without any time delay in the summation in the denominator. Then, we use the result of Lemma 3.4.4 in the proof of Lemma 3.2.3.

Lemma 3.4.4. *For any non-negative real numbers* $a_1, \ldots, a_n \in [0, a_{\max}]$,

$$\sum_{i=1}^{n} \frac{a_i}{(1+\sum_{j=1}^{i} a_j)^{4/3}} \leq 12\,.$$

Proof. Define,

$$N_0 = \max\left\{ i \in [n] : \sum_{j=1}^{i} a_j \leq 2 \right\}.$$

as well as for any $k \geq 1$

$$N_k = \max\left\{ i \in [n] : 2^k < \sum_{j=1}^{i} a_j \leq 2^{k+1} \right\}.$$

Now lets split the sum according to the N_k's

$$\begin{aligned}
\sum_{i=1}^{n} \frac{a_i}{(1+\sum_{j=1}^{i} a_j)^{4/3}} &= \sum_{i=1}^{N_0} \frac{a_i}{(1+\sum_{j=1}^{i} a_j)^{4/3}} + \sum_{k=1}^{\infty} \sum_{i=N_{k-1}+1}^{N_k} \frac{a_i}{(1+\sum_{j=1}^{i} a_j)^{4/3}} \\
&\leq \sum_{i=1}^{N_0} a_i + \sum_{k=1}^{\infty} \frac{1}{(2^k)^{4/3}} \sum_{i=1}^{N_k} a_i \\
&\leq 2 + \sum_{k=1}^{\infty} \frac{2^{k+1}}{(2^k)^{4/3}} \\
&= 2 + \sum_{k=1}^{\infty} \frac{2^{k+1}}{(2^k)^{4/3}} \\
&= 2 + 2\sum_{k=1}^{\infty} \left(\frac{1}{2^{1/3}}\right)^k \\
&\leq 2 + 2 \cdot \frac{1}{1-2^{-1/3}} \\
&\leq 12\,.
\end{aligned}$$

■

Lemma 3.2.3. *For any non-negative real numbers $a_1, \ldots, a_n \in [0, a_{\max}]$,*

$$\sum_{i=1}^{n} \frac{a_i}{(1+\sum_{j=1}^{i-1} a_j)^{4/3}} \leq 12 + 2a_{\max} .$$

Proof. Lets define,

$$N_0 = \min \left\{ i \in [n] : \sum_{j=1}^{i-1} a_j \geq a_{\max} \right\} .$$

Thus, we can decompose the sum as follows,

$$\begin{aligned}
\sum_{i=1}^{n} \frac{a_i}{(1+\sum_{j=1}^{i-1} a_j)^{4/3}} &= \sum_{i=1}^{N_0-1} \frac{a_i}{(1+\sum_{j=1}^{i-1} a_j)^{4/3}} + \sum_{i=N_0}^{n} \frac{a_i}{(1+\sum_{j=1}^{i-1} a_j)^{4/3}} \\
&\leq \sum_{i=1}^{N_0-1} a_i + \sum_{i=N_0}^{n} \frac{a_i}{(1+\sum_{j=1}^{N_0-1} a_j + \sum_{j=N_0}^{i-1} a_i)^{4/3}} \\
&\leq 2a_{\max} + \sum_{i=N_0}^{n} \frac{a_i}{(1+a_{\max}+\sum_{j=N_0}^{i-1} a_i)^{4/3}} \\
&\leq 2a_{\max} + \sum_{i=N_0}^{n} \frac{a_i}{(1+a_i+\sum_{j=N_0}^{i-1} a_i)^{4/3}} \\
&\leq 2a_{\max} + 12
\end{aligned}$$

where the second and third lines use the definition of N_0 and definition of $a_{\max}$, the fourth line uses $a_i \leq a_{\max}$, and the last line uses the helper Lemma 3.4.4. ■

This lemma is the time-shifted version of Lemma 3.4.3, such that the summation lags one iteration behind.

Lemma 3.4.5. *Let $b_1, ..., b_n \in (0, b]$ be a sequence of non-negative real numbers for some positive real number b, $b_0 > 0$ and $p \in (0,1)$ a rational number. Then,*

$$\sum_{i=1}^{n} \frac{b_i}{\left(b_0 + \sum_{j=1}^{i-1} b_j\right)^p} \leq \frac{b}{(b_0)^p} + \frac{2}{1-p} \left(b_0 + \sum_{i=1}^{n} b_i\right)^{1-p}$$

Proof. The proof of this lemma relies on the arguments of Lemma A.1 from [BL19] and makes use of Lemma 3.4.3 we proved earlier. We consider two cases for the proof depending on whether $b_0 \leq b$ or $b_0 \geq b$.

Case 1 : $b_0 \geq b$.

$$\sum_{i=1}^{n} \frac{b_i}{\left(b_0 + \sum_{j=1}^{i-1} b_j\right)^p} \leq \sum_{i=1}^{n} \frac{b_i}{\left(b + \sum_{j=1}^{i-1} b_j\right)^p}$$

$$\leq \sum_{i=1}^{n} \frac{b_i}{\left(\sum_{j=1}^{i} b_j\right)^p}$$
$$\leq \frac{1}{1-p}\left(\sum_{i=1}^{n} b_i\right)^{1-p}$$
$$\leq \frac{b}{(b_0)^p} + \frac{2}{1-p}\left(b_0 + \sum_{i=1}^{n} b_i\right)^{1-p}$$

Case 2 : $b_0 \leq b$.
Let us denote a time variable

$$T_0 = \min\left\{ i \in [n] : \sum_{j=1}^{i-1} b_j \geq b \right\}$$

Then, we could separate the summation as

$$\begin{aligned}
\sum_{i=1}^{n} \frac{b_n}{(b_0 + \sum_{j=1}^{i-1} b_j)^p} &= \sum_{i=1}^{T_0-1} \frac{b_n}{(b_0 + \sum_{j=1}^{i-1} b_j)^p} + \sum_{i=T_0}^{n} \frac{b_n}{(b_0 + \sum_{j=1}^{i-1} b_j)^p} \\
&\leq \frac{1}{(b_0)^p} \sum_{i=1}^{T_0-1} b_n + \sum_{i=T_0}^{n} \frac{b_n}{(\frac{1}{2}\sum_{j=1}^{i-1} b_j + \frac{1}{2}\sum_{j=1}^{i-1} b_j)^p} \\
&\leq \frac{b}{(b_0)^p} + \sum_{i=T_0}^{n} \frac{b_n}{(\frac{1}{2}b + \frac{1}{2}\sum_{j=1}^{i-1} b_j)^p} && \text{(Use definition of } T_0\text{)} \\
&\leq \frac{b}{(b_0)^p} + 2\sum_{i=T_0}^{n} \frac{b_n}{(\sum_{j=1}^{i} b_j)^p} && \text{(Use } b_i \leq b\text{)} \\
&\leq \frac{b}{(b_0)^p} + \frac{2}{1-p}\left(\sum_{i=T_0}^{n} b_i\right)^{1-p} && \text{(Use Lemma 3.4.3)} \\
&\leq \frac{b}{(b_0)^p} + \frac{2}{1-p}\left(b_0 + \sum_{i=1}^{n} b_i\right)^{1-p}
\end{aligned}$$

■

The final numerical inequality we will have helps us quantify the behavior of the difference $1/a_{t+1} - 1/a_t$ for the original STORM+.

Lemma 3.4.6. *Let the momentum parameter sequence $\{a_t\}$ be generated by Algorithm 6, and bounded (stochastic) gradient assumption in Eq. (3.15) hold. Also define the stopping time $\tau^* = \max\left(t \in [T] : a_t \geq \beta\right)$ where $\beta = \min\left(1, G^{-4}\right)$. Then, it holds that*

$$1/a_{t+1} - 1/a_t \leq 2/3\,; \quad \forall t \geq \tau^* + 1.$$

Moreover,

$$1/a_{t+1} \le 1/\tilde{\beta}\,; \forall\, t \le \tau^*$$

where $1/\tilde{\beta} := (1/\beta^{3/2} + G^2)^{2/3}$.

Proof. The lemma has two parts. We prove them separately.

Proof of the first part.
First note that the function $H(y) := y^{2/3}$ is concave over $\mathbb{R}_+$. Applying the gradient inequality for concave functions imply that,

$$\forall\, y_1, y_2 \ge 0\,; H(y_2) - H(y_1) \le \nabla H(y_1)^\top (y_2 - y_1) = \frac{2}{3}\frac{1}{y_1^{1/3}} \cdot (y_2 - y_1)\,.$$

Therefore, for any $t \ge \tau^* + 1$

$$\begin{aligned}
\frac{1}{a_{t+1}} - \frac{1}{a_t} &= (1 + \sum_{s=1}^{t-1} \|g_s\|^2 + \|g_t\|^2)^{2/3} - (1 + \sum_{s=1}^{t-1} \|g_s\|^2)^{2/3} \\
&\le \frac{2}{3}\frac{\|g_t\|^2}{(1 + \sum_{s=1}^{t-1} \|g_s\|^2)^{1/3}} \\
&= \frac{2}{3}\sqrt{a_t}\|g_t\|^2 \\
&\le \frac{2}{3}\sqrt{\beta}G^2 \\
&\le \frac{2}{3}\,.
\end{aligned}$$

where the fourth line uses the definition of τ^*, and the last line uses the definition of β.

Proof of the second part.
Recalling that $\tau^* = \max\{t \in [T] : a_t \ge \beta\}$ for $\beta = \min\{1, 1/G^4\}$ implies that $1/a_t \le 1/\beta\,; \forall\, t \le \tau^*$. Moreover, using the definition of a_t and boundedness of gradients we obtain,

$$(1/a_{\tau^*+1})^{3/2} = (1/a_{\tau^*})^{3/2} + \|g_{\tau^*}\|^2 \le \frac{1}{\beta^{3/2}} + G^2$$

Defining $\frac{1}{\tilde{\beta}} := \left(\frac{1}{\beta^{3/2}} + G^2\right)^{2/3}$ implies that,

$$1/a_t \le 1/\tilde{\beta}\,; \qquad \forall\, t \le \tau^* + 1\,.$$

■

Now, we are at a position to present the proof of the main result of our STORM+. Note that the proof of the simplified version checks out using the same derivation up to replacing the non-adaptive version of the momentum parameter.

Theorem 3.2.1. *Under the assumption in Eq.* (3.14), (3.15), (3.16) *and* (3.17) STORM+ *ensures,*

$$\mathbb{E}\left[\,\|\nabla f(\hat{X}_T)\|\,\right] \leq O\left(\frac{M}{\sqrt{T}} + \frac{\kappa\sigma^{1/3}}{T^{1/3}}\right),$$

where $\kappa = O(B^{3/4} + L^{3/2})$; $M = O(1 + L^{9/4} + B^{9/8} + G^5 + (LG^4)^{3/2})$, *and the expectation is taken over the randomization of the (gradient) samples as well as the selection of the output point.*

Proof Sketch of Theorem 3.2.1. The proof is composed of three parts:

1. In the first part we bound the cumulative expectation of errors $\mathbb{E}\left[\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2\right]$, where $\epsilon_t := d_t - \bar{g}_t$, and τ^* is a stopping time after which we can ensure that $1/a_{t+1} - 1/a_t \leq 2/3$. This solves the first challenge.

2. In the second part we use our bound on $\mathbb{E}\left[\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2\right]$ in order to bound the *total sum of square errors*, $\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right]$.

3. Then, in the last part, we divide into two sub-cases as we did in the simpflified proof sketch such that we first analyze the setting if $\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right] \leq (1/2)\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right]$ and then its complement. We also use the smoothness of the objective together with the update rule, similarly to what we do in Eq. (3.23).

Part (1): Bounding $\mathbb{E}\left[\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2\right]$.
Recall the error dynamics of STORM+ in Eq. (3.24). Taking the square and summing up to some $\tau^* \in [T]$ enables us to bound,

$$\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2 \leq \sum_{t=1}^{\tau^*}(1-a_t)\|\epsilon_{t-1}\|^2 + 2\sum_{t=1}^{\tau^*}\|Z_t\|^2 + 2\sum_{t=1}^{\tau^*}a_t^2\|g_t - \bar{g}_t\|^2 + \sum_{t=1}^{\tau^*}M_t\,,$$

where $M_t = 2\langle(1-a_t)\epsilon_{t-1}, a_t(g_t - \bar{g}_t) + (1-a_t)Z_t\rangle$ is a martingale difference sequence such that $\mathbb{E}[M_t|\mathscr{F}_{t-1}] = 0$, where $\mathscr{F}_t$ is the history upto and including iteration t, i.e., $\mathscr{F}_t := \{x_1, \xi_1, \xi_2, \xi_3 \ldots, \xi_t\}$. Also, recall that we have defined $Z_t := (g_t - \tilde{g}_{t-1}) - (\bar{g}_t - \bar{g}_{t-1})$.

Now let us define $\beta := \min\{1, 1/G^4\}$, and $\tau^* = \max\{t \in [T] : a_t \geq \beta\}$. Recalling that a_{t+1} is measurable with respect to $\mathscr{F}_t$ implies that $\tau^* \in [T]$ is a stopping time adapted to the same sigma-algebra sequence, $\{\mathscr{F}_t\}$. Re-arranging the above and using the definition of τ^* implies,

$$\beta\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2 \leq \|\epsilon_{\tau^*}\|^2 + \sum_{t=1}^{\tau^*-1}a_{t+1}\|\epsilon_t\|^2 \leq 2\underbrace{\sum_{t=1}^{T}\|Z_t\|^2}_{\text{(i)}} + 2\underbrace{\sum_{t=1}^{T}a_t^2\|g_t - \bar{g}_t\|^2}_{\text{(ii)}} + \underbrace{\sum_{t=1}^{\tau^*}M_t}_{\text{(iii)}} \tag{3.41}$$

where we used $\tau^* \leq T$, as well as $\beta \leq 1$. Note that we haven't used the particular definition of β yet. Next we bound the expected value of the above terms.

Bounding (i). Using smoothness property implies that $\|Z_t\| \leq 2L\|X_t - X_{t-1}\| = 2L\gamma_{t-1}\|d_{t-1}\|$.

Using the expression for γ_{t-1} together with Lemma 3.4.3 enables to show,

$$\text{(i)} \leq 4L^2 \sum_{t=1}^{T} \frac{\|d_{t-1}\|^2}{(\sum_{s=1}^{t-1}\|d_s\|^2)^{2/3}} \leq 12L^2 (\sum_{t=1}^{T}\|d_t\|^2)^{1/3} .$$

where the first inequality uses $\gamma_t = 1/\left(\sum_{s=1}^{t}\|d_s\|^2/a_{i+1}\right)^{1/3} \leq 1/\left(\sum_{s=1}^{t}\|d_s\|^2\right)^{1/3}$.

Bounding (ii). Since $\mathbb{E}[g_t|H_{t-1}] = \bar{g}_t$ and a_t is measurable with respect to $\mathscr{F}_{t-1}$, it follows that

$$\mathbb{E}[a_t^2\|g_t - \bar{g}_t\|^2] \leq \mathbb{E}[a_t^2(\|g_t\|^2 - \|\bar{g}_t\|^2)] \leq \mathbb{E}[a_t^2\|g_t\|^2]$$

Using this together with the expression for a_t, it is possible to show that,

$$\mathbb{E}(\text{ii}) \leq \mathbb{E}\left[\sum_{t=1}^{T} \frac{\|g_t\|^2}{(1+\sum_{s=1}^{t-1}\|g_s\|^2)^{4/3}}\right] \leq C_1 .$$

where $C_1 := 12 + 2G^2$ (recall G is a bound on the gradient norms), and the last inequality is due to Lemma 3.2.3.

Bounding (iii). Since $\tau^* \in [T]$ is a bounded stopping time, and M_t is a martingale difference sequence, then Doob's optional stopping theorem [LP17] implies $\mathbb{E}(\text{iii}) = \mathbb{E}\left[\sum_{t=1}^{\tau^*} M_t\right] = 0$.

Final bound. Combining the above bounds inside Eq. (3.41) together with Jensen's inequality for $U(z) = z^{1/3}$ defined over $\mathbb{R}_+$, yields,

$$\mathbb{E}\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2 \leq 2C_1/\beta + 24(L^2/\beta)(\mathbb{E}\sum_{t=1}^{T}\|d_t\|^2)^{1/3} . \tag{3.42}$$

Part (2): Bounding $\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right]$.
Recall the error dynamics of STORM+ in Eq. (3.24). Dividing by $\sqrt{a_t}$, and taking the square gives,

$$\begin{aligned}\frac{1}{a_t}\|\epsilon_t\|^2 &= (\frac{1}{a_t} - 2 + a_t)\|\epsilon_{t-1}\|^2 + \|(1-a_t)\frac{Z_t}{\sqrt{a_t}} + \sqrt{a_t}(g_t - \bar{g}_t)\|^2 + Y_t \\ &\leq (\frac{1}{a_t} - 1)\|\epsilon_{t-1}\|^2 + 2\frac{\|Z_t\|^2}{a_t} + 2a_t\|g_t - \bar{g}_t\|^2 + Y_t ,\end{aligned}$$

where we used $a_t \in [0,1]$, and $(1-a_t) \in [0,1]$, as well as $\|b+c\|^2 \leq 2\|b\|^2 + 2\|c\|^2$. We also defined $Y_t = 2\langle\frac{1-a_t}{\sqrt{a_t}}\epsilon_{t-1}, \sqrt{a_t}(g_t - \bar{g}_t) + \frac{1-a_t}{\sqrt{a_t}}Z_t\rangle$. Note that $\mathbb{E}[Y_t|\mathscr{F}_{t-1}] = 0$; therefore Y_t is a martingale difference sequence. Re-arranging the above and summing gives,

$$\sum_{t=1}^{T}\|\epsilon_{t-1}\|^2 \leq \underbrace{-\frac{1}{a_T}\|\epsilon_T\|^2}_{\text{(A)}} + \underbrace{\sum_{t=1}^{T}(\frac{1}{a_{t+1}} - \frac{1}{a_t})\|\epsilon_t\|^2}_{\text{(B)}} + 2\underbrace{\sum_{t=1}^{T}\frac{\|Z_t\|^2}{a_t}}_{\text{(C)}} + 2\underbrace{\sum_{t=1}^{T}a_t\|g_t - \bar{g}_t\|^2}_{\text{(D)}} + \underbrace{\sum_{t=1}^{T}Y_t}_{\text{(E)}} . \tag{3.43}$$

Next, we bound *the expected value* of each term separately.

Bounding (A): Since $a_T \leq 1$ we can bound $-\mathbb{E}\left[\,\|\epsilon_T\|^2\,\right]/a_T \leq -\mathbb{E}\left[\,\|\epsilon_T\|^2\,\right]$

Bounding (B). Lemma 3.4.6 enables to decompose the summation at the stopping time τ^* and bound (B) as follows,

$$\begin{aligned}\sum_{t=1}^{T}\left(\frac{1}{a_{t+1}}-\frac{1}{a_t}\right)\|\epsilon_t\|^2 &= \sum_{t=1}^{\tau^*}\left(\frac{1}{a_{t+1}}-\frac{1}{a_t}\right)\|\epsilon_t\|^2 + \sum_{t=\tau^*+1}^{T}\left(\frac{1}{a_{t+1}}-\frac{1}{a_t}\right)\|\epsilon_t\|^2 \\ &\leq \frac{1}{\tilde{\beta}}\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2 + \frac{2}{3}\sum_{t=\tau^*+1}^{T}\|\epsilon_t\|^2 \leq \frac{1}{\tilde{\beta}}\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2 + \frac{2}{3}\sum_{t=1}^{T}\|\epsilon_t\|^2 . \end{aligned} \tag{3.44}$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\,(\mathrm{B})\,\right] &\leq \frac{1}{\tilde{\beta}}\mathbb{E}\left[\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2\right] + \frac{2}{3}\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right] \\ &\leq \frac{2C_1}{\beta\tilde{\beta}} + 24\frac{L^2}{\beta\tilde{\beta}}\mathbb{E}\left[\sum_{t=1}^{T}\|d_t\|^2\right]^{1/3} + \frac{2}{3}\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right] \end{aligned} \tag{3.45}$$

where we have used Eq. (3.42) to obtain the last inequality.

Bounding (C). Recall that due to smoothness property $\|Z_t\| \leq 2L\|X_t - X_{t-1}\| = 2L\gamma_{t-1}\|d_{t-1}\|$, and using the expression for γ_{t-1} together with Lemma 3.4.3 enables to show,

$$\begin{aligned}\sum_{t=1}^{T}\frac{\|Z_t\|^2}{a_t} &\leq 4L^2\sum_{t=1}^{T}\gamma_{t-1}^2\|d_{t-1}\|^2/a_t \\ &= 4L^2\sum_{t=1}^{T}\frac{\|d_{t-1}\|^2/a_t}{\left(\sum_{s=1}^{t-1}\|d_s\|^2/a_{i+1}\right)^{2/3}} && \text{(Lemma 3.4.3)} \\ &\leq 12L^2\left(\sum_{t=1}^{T-1}\|d_t\|^2/a_{t+1}\right)^{1/3} \\ &\leq 12L^2\frac{1}{a_T^{1/3}}\left(\sum_{t=1}^{T-1}\|d_t\|^2\right)^{1/3} \\ &\leq 12L^2\left(1+\sum_{t=1}^{T}\|g_t\|^2\right)^{2/9}\left(\sum_{t=1}^{T}\|d_t\|^2\right)^{1/3}, \end{aligned} \tag{3.46}$$

where we used the fact that a_t is non-increasing.

Now, let us recall Young's inequality which states that for any $a, b > 0$, and $p, q > 1$ that satisfies $\frac{1}{p}+\frac{1}{q}=1$, we have $ab \leq a^p/p + b^q/q$. This implies that for any $a, b, \rho > 0$ and $p = \frac{3}{2}, q = 3$, we have,

$$a^{2/9}b^{1/3} = (a\rho^{9/2})^{2/9}(b/\rho^3)^{1/3} \leq \frac{(a\rho^{9/2})^{2p/9}}{p} + \frac{(b/\rho^3)^{q/3}}{q} = \frac{2}{3}a^{1/3}\rho^{3/2} + \frac{b}{3\rho^3} \tag{3.47}$$

Thus, taking $\rho = (512L^2)^{1/3}$, $a = 1 + \sum_{t=1}^{T} \|g_t\|^2$, $b = \sum_{t=1}^{T} \|d_t\|^2$, and using Young's inequality inside Eq. (3.46) implies,

$$\sum_{t=1}^{T} \frac{\|Z_t\|^2}{a_t} \leq 512L^3 \left(1 + \sum_{t=1}^{T} \|g_t\|^2\right)^{1/3} + \frac{1}{128} \sum_{t=1}^{T} \|d_t\|^2 \tag{3.48}$$

Bounding (D). Note that a_t is measurable with respect to $\mathscr{F}_{t-1}$, and $\mathbb{E}[g_t|\mathscr{F}_{t-1}] = \bar{g}_t$, therefore using smoothing gives,

$$\mathbb{E}[a_t \|g_t - \bar{g}_t\|^2] = \mathbb{E}[a_t(\|g_t\|^2 - \|\bar{g}_t\|^2)] \leq \mathbb{E}[a_t \|g_t\|^2]$$

Thus,

$$\begin{aligned}
\mathbb{E}[(\mathrm{D})] &:= \mathbb{E}\left[\sum_{t=1}^{T} a_t \|g_t - \bar{g}_t\|^2\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} a_t \|g_t\|^2\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \frac{\|g_t\|^2}{(1 + \sum_{s=1}^{t-1} \|g_s\|^2)^{2/3}}\right] \\
&\leq G^2 + 6\mathbb{E}\left[1 + \sum_{t=1}^{T} \|g_t\|^2\right]^{1/3},
\end{aligned}$$

where the last line is due to Lemma 3.4.5, which is a modified and time-shifted version of Lemma 3.4.3.

Bounding (E). Since $\{Y_t\}_{t\in[T]}$ is a martingale difference sequence, as we have argued previously, we have by definition,

$$\mathbb{E}[(\mathrm{E})] = \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] = 0\,.$$

Final bound: Combining (A)-(E). Combining the above bounds inside Eq. (3.43) we conclude that,

$$\begin{aligned}
\frac{1}{3}\mathbb{E}\sum_{t=1}^{T} \|\epsilon_t\|^2 &\leq \frac{24L^2}{\beta\tilde{\beta}}\mathbb{E}(\sum_{t=1}^{T} \|d_t\|^2)^{1/3} + \frac{2C_1}{\beta\tilde{\beta}} + 2G^2 \\
&\quad + (1024L^3 + 12)\mathbb{E}\left(1 + \sum_{t=1}^{T} \|g_t\|^2\right)^{1/3} + \frac{1}{64}\mathbb{E}\sum_{t=1}^{T} \|d_t\|^2 \\
&\leq \frac{24L^2}{\beta\tilde{\beta}}(\mathbb{E}\sum_{t=1}^{T} \|d_t\|^2)^{1/3} + \frac{2C_1}{\beta\tilde{\beta}} + 2G^2 \\
&\quad + (1024L^3 + 12)\left(1 + \mathbb{E}\sum_{t=1}^{T} \|g_t\|^2\right)^{1/3} + \frac{1}{64}\mathbb{E}\sum_{t=1}^{T} \|d_t\|^2
\end{aligned} \tag{3.49}$$

where we have used Jensen's inequality for the concave function $G(z) = z^{1/3}$, $z \geq 0$.

Part (3): Bounding $\mathbb{E}\left[\sum_{t=1}^{T} \|\bar{g}_t\|^2\right]$.
We divide the final part of the proof into two cases with respect to the growth of the cumulative noise:

Case 1 (Large error regime): $\mathbb{E}\left[\sum_{t=1}^{T} \|\epsilon_t\|^2\right] \geq (1/2)\mathbb{E}\left[\sum_{t=1}^{T} \|\bar{g}_t\|^2\right]$.
Using the statement of Case 1 implies

$$\mathbb{E}\left[\sum_{t} \|d_t\|^2\right] \leq 2\mathbb{E}\left[\sum_{t=1}^{T} \|\bar{g}_t\|^2\right] + 2\mathbb{E}\left[\sum_{t=1}^{T} \|\epsilon_t\|^2\right] \leq 6\mathbb{E}\left[\sum_{t=1}^{T} \|\epsilon_t\|^2\right].$$

Plugging this into Eq. (3.49) gives,

$$\begin{aligned}\frac{1}{3}\mathbb{E}\left[\sum_{t=1}^{T} \|\epsilon_t\|^2\right] \leq \frac{24L^2}{\beta\tilde{\beta}}\left(6\mathbb{E}\left[\sum_{t=1}^{T} \|\epsilon_t\|^2\right]\right)^{1/3} + \frac{2C_1}{\beta\tilde{\beta}} + 2G^2 \\ + (1024L^3+12)\left(1+\sigma^2 T + \mathbb{E}\left[\sum_{t=1}^{T} \|\bar{g}_t\|^2\right]\right)^{1/3} + \frac{6}{64}\mathbb{E}\left[\sum_{t=1}^{T} \|\epsilon_t\|^2\right]\end{aligned}$$

where the first line uses $\mathbb{E}\|g_t\|^2 = \mathbb{E}\|\bar{g}_t\|^2 + \mathbb{E}\|g_t - \bar{g}_t\|^2 \leq \mathbb{E}\|\bar{g}_t\|^2 + \sigma^2$.

Re-arranging and using $\mathbb{E}\sum_{t=1}^{T} \|\bar{g}_t\|^2 \leq 2\mathbb{E}\sum_{t=1}^{T} \|\epsilon_t\|^2$ gives,

$$\begin{aligned}\frac{1}{5}\mathbb{E}\left[\sum_{t=1}^{T} \|\epsilon_t\|^2\right] \leq \frac{24L^2}{\beta\tilde{\beta}}\left(6\mathbb{E}\left[\sum_{t=1}^{T} \|\epsilon_t\|^2\right]\right)^{1/3} + \frac{2C_1}{\beta\tilde{\beta}} + 2G^2 \\ + (1024L^3+12)\left(1+\sigma^2 T + 2\mathbb{E}\left[\sum_{t=1}^{T} \|\epsilon_t\|^2\right]\right)^{1/3}\end{aligned}$$

We now want to simplify the expression by freeing the RHS from the term $O\left(\mathbb{E}\left[\sum_{t=1}^{T} \|\epsilon_t\|^2\right]^{1/3}\right)$. Once again, we express thwhole expression as a polynomial inequality. First, we upper bound the last term as

$$\begin{aligned}(1024L^3+12)\left(1+\sigma^2 T + 2\mathbb{E}\left[\sum_{t=1}^{T} \|\epsilon_t\|^2\right]\right)^{1/3} \leq 2^{1/3}(1024L^3+12)\mathbb{E}\left[\sum_{t=1}^{T} \|\epsilon_t\|^2\right]^{1/3} \\ + (1024L^3+12) + (1024L^3+12)\sigma^{2/3}T^{1/3}.\end{aligned}$$

Then, defining $x = \mathbb{E}\left[\sum_{t=1}^{T} \|\epsilon_t\|^2\right]^{1/3}$ and simplifying absolute constants to 1 for ease of navigation gives us an expression of the form

$$x^3 - \left(1 + L^3 + \frac{L^2}{\beta\tilde{\beta}}\right)x - br1 + L^3 + G^2 + \frac{C_1}{\beta\tilde{\beta}} + (L^3+1)\sigma^{2/3}T^{1/3} \leq 0$$

Using the same approach as we have shown in the proof of the offline (deterministic) setting and that of the simplified STORM+, one can easily validate that the following choice of x

satisfies the inequality

$$x = \mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right]^{1/3} = O\left(\left[1+1+L^3+G^2+\frac{C_1}{\beta\tilde{\beta}}+(L^3+1)\sigma^{2/3}T^{1/3}+\left(\frac{L^2}{\beta\tilde{\beta}}\right)^{3/2}+L^{9/2}\right]^{1/3}\right)$$

The above selection, together with the statement of Case 1 implies

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right] &\leq 2\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right] \\ &\leq O\left(1+1+L^3+G^2+\frac{C_1}{\beta\tilde{\beta}}+(L^3+1)\sigma^{2/3}T^{1/3}+\left(\frac{L^2}{\beta\tilde{\beta}}\right)^{3/2}+L^{9/2}\right) \qquad (3.50)\end{aligned}$$

Case 2 (Small error regime): $\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right] \leq (1/2)\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right]$. Using Lemma 3.2.2 we obtain the departure point of the analysis of this case,

$$\begin{aligned}\sum_{t=1}^{T}\|\bar{g}_t\|^2 &\leq \sum_{t=1}^{T}\|\epsilon_t\|^2 + 2B(1+\sum_{t=1}^{T}\|g_t\|^2)^{2/9}(\sum_{t=1}^{T}\|d_t\|^2)^{1/3}+\frac{3}{2}L(\sum_{t=1}^{T}\|d_t\|^2)^{2/3} \\ &\leq \sum_{t=1}^{T}\|\epsilon_t\|^2+\frac{3}{2}L(\sum_{t=1}^{T}\|d_t\|^2)^{2/3}+20B^{3/2}(1+\sum_{t=1}^{T}\|g_t\|^2)^{1/3}+\frac{1}{64}\sum_{t=1}^{T}\|d_t\|^2 \qquad (3.51)\end{aligned}$$

where the second line uses a version of Young's inequality in Eq. (3.47) by selecting a different set of variables; $\rho := (128B/3)^{1/3}$, $a := 1+\sum_{t=1}^{T}\|g_t\|^2$ and $b := \sum_{t=1}^{T}\|d_t\|^2$. The condition of this case implies

$$\mathbb{E}\left[\sum_{t}\|d_t\|^2\right] \leq 2\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right]+2\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right] \leq 3\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right]$$

Taking expectation of Eq. (3.51), plugging in the upper bound for $\mathbb{E}\left[\sum_t\|d_t\|^2\right]$ as described above and upper bounding the cumulative noise term $\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right]$ using the condition of Case 2 gives us,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|^2\right]+\frac{3}{2}L(\mathbb{E}\left[\sum_{t=1}^{T}\|d_t\|^2\right])^{2/3}+20B^{3/2}(1+\mathbb{E}\left[\sum_{t=1}^{T}\|g_t\|^2\right])^{1/3}+\frac{1}{64}\mathbb{E}\left[\sum_{t=1}^{T}\|d_t\|^2\right] \\ &\leq \left(\frac{1}{2}+\frac{3}{64}\right)\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right]+\frac{3}{2}L(3\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right])^{2/3}+20B^{3/2}(1+\sigma^2T+\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right])^{1/3}\end{aligned}$$

where we have used Jensen's inequality for the function $z^{1/3}$ and $z^{2/3}$ defined over $\mathbb{R}_+$. We also use the fact that $\mathbb{E}\left[\|g_t\|^2\right] = \mathbb{E}\left[\|\bar{g}_t\|^2\right]+\mathbb{E}\left[\|g_t-\bar{g}_t\|^2\right] \leq \mathbb{E}\left[\|\bar{g}_t\|^2\right]+\sigma^2$. By re-arranging and simplifying the above we obtain,

$$\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right] \leq 18L\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right]^{2/3}+80B^{3/2}(1+\sigma^2T+\mathbb{E}\left[\sum_{t=1}^{T}\|\bar{g}_t\|^2\right])^{1/3}$$

With the same technique as in Case 1, let us denote $x = \mathbb{E}\left[\sum_{t=1}^{T} \|\bar{g}_t\|^2\right]^{1/3}$, represents absolute constants as 1 to obtain the polynomial inequality

$$x^3 - Lx^2 - B^{3/2}x - B^{3/2}\left(1 + \sigma^2/3T^{1/3}\right) \leq 0$$

We recognize that setting

$$x = \mathbb{E}\left[\sum_{t=1}^{T} \|\bar{g}_t\|^2\right]^{1/3} = O\left(\left[B^{3/2} + B^{9/4} + B^{3/2}\sigma^{2/3}T1/3 + L^3\right]^{1/3}\right)$$

satisfies the inequality, which directly translates to

$$\mathbb{E}\left[\sum_{t=1}^{T} \|\bar{g}_t\|^2\right] \leq \left(B^{3/2} + B^{9/4} + B^{3/2}\sigma^{2/3}T^{1/3} + L^3\right) \tag{3.52}$$

Final bound: Combining Case 1 and 2.
Combining Eq. (3.50) and (3.53) together with the definitions of C_1, β and $\tilde{\beta}$,

$$\mathbb{E}\left[\sum_{t=1}^{T} \|\bar{g}_t\|^2\right] \leq O(M^2 + \kappa^2\sigma^{2/3}T^{1/3}) \tag{3.53}$$

where $\kappa^2 := 1 + B^{3/2} + L^3$, and $M^2 := 1 + L^{9/2} + B^{9/4} + G^{10} + (L^2G^8)^{3/2}$.

Using the definition of $\hat{X}_T$ together with Jensen's inequality gives,

$$\mathbb{E}\left[\|\nabla f(\hat{X}_T)\|\right] = O\left(\frac{M}{\sqrt{T}} + \frac{\kappa\sigma^{1/3}}{T^{1/3}}\right).$$

which concludes the proof. ■

3.4.3 Proofs of Section 3.3

Lemma 3.3.1. *Define the gradient estimator at current point x^+ as $\nabla_{x^+} := \nabla f_i(x^+) - \nabla f_i(x) + \nabla_x$ where x denotes the previous step of the execution and i is sampled uniformly at random from $\{1,\dots,n\}$. Then,*

$$\mathbb{E}\left[\|\nabla_{x^+} - \nabla f(x^+)\|^2\right] \leq L^2\|x^+ - x\|^2 + \mathbb{E}\left[\|\nabla_x - \nabla f(x)\|^2\right]$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\|\nabla_{x^+} - \nabla f(x^+)\|^2\right] &= \mathbb{E}\left[\|\nabla f_i(x^+) - \nabla f_i(x) + \nabla_x - \nabla f(x^+)\|^2\right] \\
&= \mathbb{E}\left[\|\nabla f_i(x^+) - \nabla f_i(x) + \nabla_x - \nabla f(x^+) + \nabla f(x) - \nabla f(x)\|^2\right] \\
&= \mathbb{E}\left[\|\nabla f_i(x^+) - \nabla f_i(x) - \nabla f(x^+) + \nabla f(x)\|^2\right] \\
&+ 2\mathbb{E}\left[\left\langle \nabla f_i(x^+) - \nabla f_i(x) - \nabla f(x^+) + \nabla f(x), \nabla_x - \nabla f(x)\right\rangle\right] \\
&+ \mathbb{E}\left[\|\nabla_x - \nabla f(x)\|^2\right]
\end{aligned}$$

Notice that $\mathbb{E}\left[\nabla f_i(x^+) - \nabla f_i(x) - \nabla f(x^+) + \nabla f(x)\right] = 0$ due to the fact that i is selected uniformly at random in $\{1,\dots,n\}$ and thus $\mathbb{E}\left[\nabla f_i(x^+) - \nabla f_i(x)\right] = \nabla f(x^+) - \nabla f(x)$. The latter implies that,

$$\begin{aligned}
\mathbb{E}\left[\|\nabla_{x^+} - \nabla f(x^+)\|^2\right] &= \mathbb{E}\left[\|\nabla f_i(x^+) - \nabla f_i(x) - \nabla f(x^+) + \nabla f(x)\|^2\right] + \mathbb{E}\left[\|\nabla_x - \nabla f(x)\|^2\right] \\
&\leq \mathbb{E}\left[\|\nabla f_i(x^+) - \nabla f_i(x)\|^2\right] + \mathbb{E}\left[\|\nabla_x - \nabla f(x)\|^2\right] \\
&\leq L^2\mathbb{E}\left[\|x^+ - x\|^2\right] + \mathbb{E}\left[\|\nabla_x - \nabla f(x)\|^2\right]
\end{aligned}$$

where the first inequality follows by the identity $\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] = \mathbb{E}[\|X\|^2] - \|\mathbb{E}[X]\|^2$ and the second inequality by the smoothness of the function $f_i(\cdot)$. ■

Lemma 3.3.2. *Let $\{X_t\}$ be the points produced by Algorithm 7. Then,*

$$\sum_{t=0}^{T-1}\|\nabla_t\|^2 \leq \mathcal{O}\left(n^2T^3\left(\frac{L^2}{\beta_0^2} + \|\nabla f(X_0)\|^2\right)\right)$$

Proof. The selection of the step-size in Step 9 of Algorithm 7 implies that $\|X_{t+1} - X_t\|^2 = \|\gamma_t\nabla_t\|^2 \leq 1/\beta_0^2$. Due to the fact that every n iterations a full-gradient computation is performed, the estimator $\nabla_t := \nabla f_{i_t}(X_t) - \nabla f_{i_t}(X_{t-1}) + \nabla_{t-1}$ can be equivalently written as

$$\nabla_t = \sum_{s=t-(t \bmod n)+1}^{t}\left(\nabla f_{i_s}(X_s) - \nabla f_{i_s}(X_{s-1})\right) + \nabla f(X_{t-(t \bmod n)})$$

As a result,

$$\|\nabla_t\|^2 = \|\sum_{s=t-(t \bmod n)+1}^{t}\left(\nabla f_{i_s}(X_s) - \nabla f_{i_s}(X_{s-1})\right) + \nabla f(X_{t-(t \bmod n)})\|^2$$

$$
\begin{aligned}
&\leq 2\| \sum_{s=t-(t \bmod n)+1}^{t} \nabla f_{i_s}(X_s) - \nabla f_{i_s}(X_{s-1})\|^2 + 2\|\nabla f(X_{t-(t \bmod n)})\|^2 \\
&\leq 2n \sum_{s=t-(t \bmod n)+1}^{t} \|\nabla f_{i_s}(X_s) - \nabla f_{i_s}(X_{s-1})\|^2 + 2\|\nabla f(X_{t-(t \bmod n)})\|^2 \\
&\leq 2nL^2 \sum_{s=t-(t \bmod n)+1}^{t} \|X_s - X_{s-1}\|^2 + 2\|\nabla f(X_{t-(t \bmod n)})\|^2 \\
&\leq \frac{2L^2n^2}{\beta_0^2} + 2\|\nabla f(X_{t-(t \bmod n)})\|^2
\end{aligned}
$$

Now, we want to upper bound $\|\nabla f(X_t)\|$ for any $t \leq T$ with respect to the initial gradient norm. Using again the step-size selection γ_t we get,

$$
\begin{aligned}
\|\nabla f(X_t)\| &= \|\nabla f(X_t) - \nabla f(X_0) + \nabla f(X_0)\| \\
&\leq \|\nabla f(X_t) - \nabla f(X_0)\| + \|\nabla f(X_0)\| && \text{(Triangular inequality)} \\
&\leq L\|X_t - X_0\| + \|\nabla f(X_0)\| && \text{(Smoothness)} \\
&\leq L\|X_t - X_{t-1}\| + L\|X_{t-1} - X_0\| + \|\nabla f(X_0)\| && \text{(Triangular inequality)} \\
&\leq L\sum_{s=1}^{t} \|X_s - X_{s-1}\| + \|\nabla f(X_0)\| \\
&\leq \frac{Lt}{\beta_0} + \|\nabla f(X_0)\|
\end{aligned}
$$

As a result,

$$
\begin{aligned}
\sum_{t=0}^{T-1} \|\nabla_t\|^2 &\leq \sum_{t=0}^{T-1} \left(\frac{2L^2n^2}{\beta_0^2} + 2\|\nabla f(X_{t-(t \bmod n)})\|^2 \right) \\
&\leq \frac{2L^2n^2}{\beta_0^2} T + 2\sum_{t=0}^{T-1} \|\nabla f(X_t)\|^2 \\
&\leq \frac{2L^2n^2}{\beta_0^2} T + 2\sum_{t=0}^{T-1} (\frac{Lt}{\beta_0} + \|\nabla f(X_0)\|)^2 \\
&= \frac{2L^2n^2}{\beta_0^2} T + 2\sum_{t=0}^{T-1} \left(\frac{L^2t^2}{\beta_0^2} + 2\frac{Lt}{\beta_0}\|\nabla f(X_0)\| + \|\nabla f(X_0)\|^2 \right) \\
&\leq \frac{2L^2n^2}{\beta_0^2} T + \frac{2L^2T^3}{\beta_0^2} + \frac{4LT^2\|\nabla f(X_0)\|}{\beta_0} + 2T\|\nabla f(X_0)\|^2
\end{aligned}
$$

■

Lemma 3.3.3. *Let $\{X_t\}$ be a sequence of points produced by Algorithm 7. Then,*

$$
\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla_t - \nabla f(X_t)\|^2\right] \leq O\left(\frac{Ln^{1/4}}{\beta_0} \sqrt{\log\left(1 + nT\left(\frac{L}{\beta_0 G_0} + \frac{\|\nabla f(X_0)\|}{G_0}\right)\right)} \right).
$$

Proof.

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_t-\nabla f(X_t)\|\right] &= \mathbb{E}\left[\sqrt{\left(\sum_{t=0}^{T-1}\|\nabla_t-\nabla f(X_t)\|\right)^2}\right]\\ &\leq \sqrt{\mathbb{E}\left[\left(\sum_{t=0}^{T-1}\|\nabla_t-\nabla f(X_t)\|\right)^2\right]} \qquad \text{(Jensen's ineq.)}\\ &\leq \sqrt{T}\sqrt{\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_t-\nabla f(X_t)\|^2\right]}\end{aligned}$$

where the last inequality follows by the fact that $\|\sum_{t=0}^{T-1} y_t\|^2 \leq T\sum_{t=0}^{T-1}\|y_t\|^2$. By applying Lemma 3.3.1 to the estimator $\nabla_t := \nabla f_{i_t}(X_t) - \nabla f_{i_t}(X_{t-1}) + \nabla_{t-1}$ we get,

$$\begin{aligned}\mathbb{E}\left[\|\nabla_t-\nabla f(X_t)\|^2\right] &\leq L^2\mathbb{E}\left[\|X_t-X_{t-1}\|^2\right]+\mathbb{E}\left[\|\nabla_{t-1}-\nabla f(X_{t-1})\|^2\right]\\ &\leq L^2\mathbb{E}\left[\gamma_{t-1}^2\|\nabla_{t-1}\|^2\right]+\mathbb{E}\left[\|\nabla_{t-1}-\nabla f(X_{t-1})\|^2\right]\\ &\leq L^2\mathbb{E}\left[\gamma_{t-1}^2\|\nabla_{t-1}\|^2\right]+\ldots+\mathbb{E}\left[\|\nabla_{t-(t \bmod n)}-\nabla f(X_{t-(t \bmod n)})\|^2\right]\\ &= \sum_{\tau=t-(t \bmod n)+1}^{t-1} L^2\mathbb{E}\left[\gamma_\tau^2\|\nabla_\tau\|^2\right]\end{aligned}$$

where the last equality follows by the fact that $\mathbb{E}\left[\|\nabla_{t-(t \bmod n)}-\nabla f(X_{t-(t \bmod n)})\|^2\right]=0$ (see Step 3 of Algorithm 7). As explained in Section 3.3.5, by a telescoping summation over t we get that

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_t-\nabla f(X_t)\|^2\right] \leq L^2 n\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t^2\|\nabla_t\|^2\right].$$

Now, as discussed in Section 3.3.5, using the step-size selection γ_t of Algorithm 7 we can provide a bound on the total variance $\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_t-\nabla f(X_t)\|^2\right]$.

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_t-\nabla f(X_t)\|^2\right] &\leq L^2 n\,\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t^2\|\nabla_t\|^2\right] &(3.54)\\ &= \frac{L^2\sqrt{n}}{\beta_0^2}\,\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{\|\nabla_t\|^2}{\sqrt{n}G_0^2+\sum_{s=0}^{t}\|\nabla_s\|^2}\right] &(3.55)\\ &\leq \frac{L^2\sqrt{n}}{\beta_0^2}\log\left(1+\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_t\|^2/G_0^2\right]\right) &(3.56)\\ &\leq \frac{L^2\sqrt{n}}{\beta_0^2}\mathcal{O}\left(\log\left(1+nT\left(\frac{L}{\beta_0 G_0}+\frac{\|\nabla f(X_0)\|}{G_0}\right)\right)\right) &(3.57)\end{aligned}$$

where the second inequality follows by Lemma 3.1.1 and the third inequality by Lemma 3.3.2.

Putting everything together we get

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_t - \nabla f(X_t)\|\right] \leq \frac{Ln^{1/4}}{\beta_0\sqrt{T}}\mathcal{O}\left(\sqrt{\log\left(1+nT\left(\frac{L}{\beta_0 G_0}+\frac{\|\nabla f(X_0)\|}{G_0}\right)\right)}\right)$$

■

Lemma 3.3.4. *Let $\{X_t\}$ be the sequence of points produced by Algorithm 7 and $\Delta_0 := f(X_0) - f^*$. Then,*

$$\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_t\|\right] \leq \tilde{\mathcal{O}}\left(\Delta_0\beta_0 + G_0 + \frac{L}{\beta_0} + \mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t\|\nabla f(X_t)-\nabla_t\|^2\right]\right)n^{1/4}\sqrt{T}.$$

Proof. Let $\mathscr{F}_t$ denote the filtration at round t, i.e., all the random choices $\{i_0,\ldots,i_t\}$ and the selection of the initial point X_0. By the smoothness of f we get that,

$$\begin{aligned}
\mathbb{E}\left[f(X_{t+1}) \mid \mathscr{F}_t\right] &\leq \mathbb{E}\left[f(X_t) + \langle\nabla f(X_t), X_{t+1}-X_t\rangle + \frac{L}{2}\|X_t - X_{t+1}\|^2 \mid \mathscr{F}_t\right] \\
&= \mathbb{E}\left[f(X_t) - \gamma_t\langle\nabla f(X_t),\nabla_t\rangle + \frac{L}{2}\gamma_t^2\|\nabla_t\|^2 \mid \mathscr{F}_t\right] \\
&\leq \mathbb{E}\left[f(X_t) + \frac{\gamma_t}{2}\|\nabla_t - \nabla f(X_t)\|^2 - \frac{\gamma_t}{2}(1-L\gamma_t)\|\nabla_t\|^2 \mid \mathscr{F}_t\right]
\end{aligned}$$

Thus,

$$\mathbb{E}\left[\gamma_t\|\nabla_t\|^2\right] \leq 2\mathbb{E}\left[f(X_t) - f(X_{t+1})\right] + \mathbb{E}\left[L\gamma_t^2\|\nabla_t\|^2\right] + \beta_0\mathbb{E}\left[\gamma_t\|\nabla f(X_t)-\nabla_t\|^2\right].$$

and by summing from $t=0$ to $T-1$ and telescoping the function values we get,

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\gamma_t\|\nabla_t\|^2\right] \leq 2\Delta_0 + \mathbb{E}\left[\sum_{t=0}^{T-1}L\gamma_t^2\|\nabla_t\|^2\right] + \mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t\|\nabla f(X_t)-\nabla_t\|^2\right]$$

Using the fact that $\gamma_t := n^{-1/4}\beta_0^{-1}\left(n^{1/2}G_0^2 + \sum_{s=0}^t\|\nabla_s\|^2\right)^{-1/2}$ on the second summation term,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t\|\nabla_t\|^2\right] &\leq 2\Delta_0 + \mathbb{E}\left[\sum_{t=0}^{T-1}L\gamma_t^2\|\nabla_t\|^2\right] + \mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t\|\nabla f(X_t)-\nabla_t\|^2\right] \\
&\leq 2\Delta_0 + \frac{L}{\beta_0^2}\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{\|\nabla_t\|^2}{\sqrt{n}G_0^2 + \sum_{s=0}^t\|\nabla_t\|^2}\right] \\
&+ \mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t\|\nabla f(X_t)-\nabla_t\|^2\right] \\
&\leq 2\Delta_0 + \frac{L}{\beta_0^2}\mathcal{O}\left(\log\left(1+nT\left(\frac{L}{\beta_0 G_0}+\frac{\|\nabla f(X_0)\|}{G_0}\right)\right)\right) + \mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t\|\nabla f(X_t)-\nabla_t\|^2\right]
\end{aligned}$$

where the last line is due to Eq. (3.54) in the proof of Lemma 3.3.3. Using again the definition of the step-size $\gamma_t := n^{-1/4}\beta_0^{-1}\left(n^{1/2}G_0^2 + \sum_{s=0}^t\|\nabla_s\|^2\right)^{-1/2}$ we *lower bound the right-hand side*

as follows,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t\|\nabla_t\|^2\right] &\geq \mathbb{E}\left[\frac{\sum_{t=0}^{T-1}\|\nabla_t\|^2}{n^{1/4}\beta_0\sqrt{n^{1/2}G_0^2+\sum_{t=0}^{T-1}\|\nabla_t\|^2}}\right] \\
&\geq \frac{G_0}{\beta_0}\mathbb{E}\left[\frac{\sum_{t=0}^{T-1}\|\nabla_t\|^2/\sqrt{n}G_0^2}{\sqrt{1+\sum_{t=0}^{T-1}\|\nabla_t\|^2/\sqrt{n}G_0^2}}\right] \\
&\geq \frac{G_0}{\beta_0}\left(\mathbb{E}\left[\sqrt{\sum_{t=0}^{T-1}\|\nabla_t\|^2/\sqrt{n}G_0^2}\right]-1\right) \\
&\geq \frac{1}{\beta_0 n^{1/4}\sqrt{T}}\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_t\|\right]-\frac{G_0}{\beta_0}
\end{aligned}$$

By putting everything together we get,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_t\|\right] &\leq O\left(\Delta_0\beta_0+G_0+\frac{L}{\beta_0}\log\left(1+nT\left(\frac{L}{\beta_0 G_0}+\frac{\|\nabla f(X_0)\|}{G_0}\right)\right)\right) \\
&+ \mathcal{O}\left(\beta_0\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t\|\nabla f(X_t)-\nabla_t\|^2\right]\right)n^{1/4}\sqrt{T}.
\end{aligned}$$

■

Lemma 3.3.5. *Let $\{X_t\}$ be the sequence of points produced by Algorithm 7. Then,*

$$\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t\|\nabla_t-\nabla f(X_t)\|^2\right]\leq L^2 n\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t^3\|\nabla_t\|^2\right]$$

Proof. Let $\mathscr{F}_t$ denotes the filtration at round t, i.e., all the random choices $\{i_0,\ldots,i_t\}$ and the selection of theinitial point X_0. At first notice that by the definition of γ_t in Line 9 of Algorithm 7, $\gamma_t\leq\gamma_{t-1}$, which we have to do to circumvent non-measurability issues, and thus

$$\mathbb{E}\left[\gamma_t\|\nabla_t-\nabla f(X_t)\|^2\mid\mathscr{F}_{t-1}\right]\leq\mathbb{E}\left[\gamma_{t-1}\|\nabla_t-\nabla f(X_t)\|^2\mid\mathscr{F}_{t-1}\right]$$

Up next we derive a bound on $\mathbb{E}\left[\gamma_{t-1}\|\nabla_t-\nabla f(X_t)\|^2\mid\mathscr{F}_{t-1}\right]$ using similar arguments with the ones used in Lemma 3.3.3. Notice that γ_{t-1} is $\mathscr{F}_{t-1}$-measurable, hence we can treat in independent of the conditional expectation.

$$\begin{aligned}
&\mathbb{E}\left[\gamma_{t-1}\|\nabla_t-\nabla f(X_t)\|^2\mid\mathscr{F}_{t-1}\right] \\
=\;&\gamma_{t-1}\mathbb{E}\left[\|\nabla f_{i_t}(X_t)-\nabla f_{i_t}(X_{t-1})-\nabla f(X_t)+\nabla f(X_{t-1})+(\nabla_{t-1}-\nabla f(X_{t-1}))\|^2\mid\mathscr{F}_{t-1}\right] \\
=\;&\gamma_{t-1}\mathbb{E}\left[\|\nabla f_{i_t}(X_t)-\nabla f_{i_t}(X_{t-1})-\nabla f(X_t)+\nabla f(X_{t-1})\|^2\mid\mathscr{F}_{t-1}\right]
\end{aligned}$$

$$
\begin{aligned}
&+ \quad \gamma_{t-1}\underbrace{\mathbb{E}\left[(\nabla f_{i_t}(X_t)-\nabla f_{i_t}(X_{t-1})-\nabla f(X_t)+\nabla f(X_{t-1}))^\top(\nabla_{t-1}-\nabla f(X_{t-1}))\mid \mathscr{F}_{t-1}\right]}_{0}\\
&+ \quad \gamma_{t-1}\mathbb{E}\left[\|\nabla_{t-1}-\nabla f(X_{t-1})\|^2\mid \mathscr{F}_{t-1}\right]\\
&= \quad \gamma_{t-1}\mathbb{E}\left[\|\nabla f_{i_t}(X_t)-\nabla f_{i_t}(X_{t-1})\|^2\mid \mathscr{F}_{t-1}\right]+\gamma_{t-1}\mathbb{E}\left[\|\nabla_{t-1}-\nabla f(X_{t-1})\|^2\mid \mathscr{F}_{t-1}\right]\\
&\leq \quad L^2\gamma_{t-1}\mathbb{E}\left[\|X_t-X_{t-1}\|^2\mid \mathscr{F}_{t-1}\right]+\gamma_{t-1}\mathbb{E}\left[\|\nabla_{t-1}-\nabla f(X_{t-1})\|^2\mid \mathscr{F}_{t-1}\right]\\
&= \quad L^2\gamma_{t-1}^3\mathbb{E}\left[\|\nabla_{t-1}\|^2\mid \mathscr{F}_{t-1}\right]+\gamma_{t-1}\mathbb{E}\left[\|\nabla_{t-1}-\nabla f(X_{t-1})\|^2\mid \mathscr{F}_{t-1}\right]
\end{aligned}
$$

Taking full expectation over all randomness and by the law of total expctation, we get that,

$$\mathbb{E}\left[\gamma_t\|\nabla_t-\nabla f(X_t)\|^2\right]\leq L^2\mathbb{E}\left[\gamma_{t-1}^3\|\nabla_{t-1}\|^2\right]+\mathbb{E}\left[\gamma_{t-1}\|\nabla_{t-1}-\nabla f(X_{t-1})\|^2\right]$$

Due to the fact that $\mathbb{E}\left[\|\nabla_t-\nabla f(X_t)\|\right]=0$ for $t \bmod n == 0$ we get that

$$\mathbb{E}\left[\gamma_t\|\nabla_t-\nabla f(X_t)\|^2\right]\leq L^2\mathbb{E}\left[\sum_{s=t-t \bmod n}^{t-1}\gamma_s^3\|\nabla_s\|^2\right]$$

and thus

$$\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t\|\nabla_t-\nabla f(X_t)\|^2\right]\leq L^2 n\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t^3\|\nabla_t\|^2\right]$$

■

We are not at a position to prove the main result of Section 3.3.

Theorem 3.3.1. *Let $\{X_t\}$ be the sequence of points produced by Algorithm 7 in case $f(\cdot)$ is L-smooth. Let us also define $\Delta_0 := f(X_0)-f^*$. Then,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(X_t)\|\right]\leq O\left(n^{1/4}\frac{\Theta}{\sqrt{T}}\log\left(1+nT\left(\frac{L}{\beta_0 G_0}+\frac{\|\nabla f(X_0)\|}{G_0}\right)\right)\right)$$

where $\Theta=\Delta_0\beta_0+G_0+L/\beta_0+L^2/(\beta_0^2 G_0)$. Overall, Algorithm 7 with $\beta_0:=1$ and $G_0:=1$ needs at most $\tilde{O}\left(n+\sqrt{n}\frac{\Delta_0^2+L^4}{\epsilon^2}\right)$ oracle calls to reach an ϵ-stationary point.

Proof of Theorem 3.3.1. By the triangle inequality we get that

$$\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla f(X_t)\|\right]\leq\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla_t\|\right]+\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla f(X_t)-\nabla_t\|\right]$$

Using the bounds obtained in Lemma 3.3.3 and Lemma 3.3.4 we get that,

$$\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla f(X_t)\|\right] \quad\leq\quad \tilde{O}\left(\Delta_0\beta_0+G_0+\frac{L}{\beta_0}\right)n^{1/4}\sqrt{T}$$

$$+ \quad \beta_0 \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t \|\nabla f(X_t) - \nabla_t\|^2\right] n^{1/4}\sqrt{T}$$

Then by Lemma 3.3.5 we get that,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{T-1} \|\nabla f(X_t)\|\right] &\leq \tilde{O}\left(\Delta_0\beta_0 + G_0 + \frac{L}{\beta_0}\right) n^{1/4}\sqrt{T} \\ &+ \underbrace{\beta_0 n^{5/4}\sqrt{T}L^2\mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t^3 \|\nabla_t\|^2\right]}_{\text{(A)}}\end{aligned}$$

Substituing the selection of γ_t in term (A) we get,

$$\begin{aligned}\beta_0\sqrt{T}L^2\mathbb{E}\left[\sum_{t=0}^{T-1} n^{5/4}\gamma_t^3\|\nabla_t\|^2\right] &= \frac{\sqrt{T}L^2}{\beta_0^2}\mathbb{E}\left[\sum_{t=0}^{T-1} \frac{n^{5/4}}{n^{3/4}\sqrt{n^{1/2}G_0^2 + \sum_{s=0}^t \|\nabla_t\|^2}} \frac{\|\nabla_t\|^2}{n^{1/2}G_0^2 + \sum_{s=0}^t \|\nabla_t\|^2}\right] \\ &\leq \frac{\sqrt{T}L^2}{\beta_0^2 G_0}\mathbb{E}\left[\sum_{t=0}^{T-1} \frac{n^{5/4}}{n^{3/4}\sqrt{n^{1/2}}} \frac{\|\nabla_t\|^2/G_0^2}{n^{1/2} + \sum_{s=0}^t \|\nabla_t\|^2/G_0^2}\right] \\ &\leq \frac{\sqrt{T}L^2}{\beta_0^2 G_0} n^{1/4}\mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\|\nabla_t\|^2/G_0^2}{1 + \sum_{s=0}^t \|\nabla_t\|^2/G_0^2}\right] \\ &\leq \frac{\sqrt{T}L^2}{\beta_0^2 G_0} n^{1/4}\mathcal{O}\left(\log\left(1 + nT\left(\frac{L}{\beta_0 G_0} + \frac{\|\nabla f(X_0)\|}{G_0}\right)\right)\right)\end{aligned}$$

where the forth inequality follows by Lemma 3.1.1 and the last by Lemma 3.3.2. Theorem 3.3.1 then follows by dividing both sides with T. ■

4 Efficient and robust algorithms for min-max problems and games

4.1 Universal First-Order Methods for Stochastic Variational Inequalities

4.1.1 Bibliographic Note

This section (Section 4.1) is based on the published work Antonakopoulos et al. [Ant+21], published in the NeurIPS 2021 conference.

Author list of the published work .

- Kimon Antonakopoulos
- Thomas Pethick
- Ali Kavis
- Panayotis Mertikopoulos
- Volkan Cevher

Description of contributions. The convergence proof for the non-adaptive setting in Section 4.1.5 (Theorems 4.1.1 and 4.1.2) are due to the candidate. Kimon Antonakopoulos identified the relationship between cocoercivity and relative noise assumption for achieving fast rates of order $O(1/T)$ for monotone variational inequalities. The analysis for the adaptive algorithm under Section 4.1.5 (Theorems 4.1.3 and 4.1.4) and the almost-sure convergence proofs in the whole of the manuscript (Propositions 4.1.5, 4.1.7 and 4.2.4) are due to Kimon Antonakopoulos. Thomas Pethick implemented all the numerical experiments of the paper, many of which are not included in this manuscript due to space constraints.

4.1.2 Introduction

In this section, we focus on solving the variational inequality (VI) problem of the form

$$\text{Find } x^* \in \mathcal{V} \text{ such that } \langle A(x^*), x - x^* \rangle \geq 0 \text{ for all } x \in \mathcal{V}, \tag{VI}$$

where $A\colon \mathbb{R}^d \to \mathbb{R}^d$ is a *monotone cocoercive* operator, i.e.,

$$\langle A(y) - A(x), y - x \rangle \geq L\|A(y) - A(x)\|^2 \quad \text{for some } L > 0 \text{ and all } x, y \in \mathcal{V}. \tag{CC}$$

The study of (VI) is a classical topic in optimization that provides a powerful and elegant unifying framework for a broad spectrum of "convex-structured" problems – including convex minimization, saddle-point problems, and games [FP03; BC17]. In particular, such problems have recently attracted considerable attention in the fields of machine learning (ML) and data science because of their potential applications to generative adversarial networks [Goo+14], multi-agent and robust reinforcement learning [Pin+17], auction theory [Syr+15], and many other areas of interest where the minimization of a single empirical loss function does not suffice.

The golden standard for solving (VI) is provided by first-order methods: these methods can be run with computationally cheap updates that only require (noisy) access to A, so they are ideal for problems with very high dimensionality and moderate-to-low precision needs (as is typically the case in ML). More precisely, when A is monotone cocoercive as above, the min-max optimal convergence rate for solving (VI) is $\mathcal{O}(1/T)$ after T oracle calls, and it is achieved by the extra-gradient / mirror-prox algorithm [Kor76; Nem04] with Polyak-Rupert averaging [PJ92]. However, this method requires access to a *perfect* oracle; if the method is run with an imperfect, *stochastic* first-order oracle, its convergence rate drops to $\mathcal{O}(1/\sqrt{T})$ [JNT11], and this rate cannot be improved without additional assumptions [Nes03].

One case where the $\mathcal{O}(1/\sqrt{T})$ convergence rate *can* be improved is when the underlying operator is *strongly monotone* – i.e., the RHS of (CC) is replaced by $\kappa\|y - x\|^2$ for some $\kappa > 0$. In this case, we can obtain a fast $\mathcal{O}(1/T)$ rate with a rapidly decreasing step-size [Hsi+19]; however, this acceleration requires knowledge of the strong monotonicity modulus, and there is no known way to adapt to it. In particular, if a stochastic method that has been fine-tuned for strongly monotone operators is run on a merely monotone problem, its rate of convergence suffers a catastrophic drop to $\mathcal{O}(1/\log T)$.

These considerations naturally lead to two key research questions:

1. *Are there any conditions for the method's oracle that would close the stochastic-deterministic convergence gap outlined above?*

2. *Is it possible to design a class of methods that are capable of adapting to the quality of the oracle, and that achieve order-optimal rates without prior knowledge of the problem's parameters?*

Our contributions in the context of related work. Our goal in this work is to provide a range of positive answers to the above questions, both in terms of the required oracle conditions, as well as methods that are able to gracefully interpolate between an $\mathcal{O}(1/T)$ and an $\mathcal{O}(1/\sqrt{T})$ rate depending on the setting at hand. While doing so, we aim to adopt a parameter-free approach such that the proposed algorithm runs with a data-adaptive step-size which doesn't need to know any problem-dependent parameters, e.g., Lipschitz constant, type of the noise and the respective variance parameter.

With regard to the first question, our point of departure is the "relative noise" framework of Polyak [Pol87], in which the variance of the oracle is upper bounded by the square norm of the operator at the queried point. This noise model is particularly relevant in coordinate descent methods for unconstrained problems as well as applications to control theory and signal processing where the operator is calculated based on actual, physical measurements that are only accurate up to a percentage of their true value. In recent applications to ML, this noise model has also been studied in the context of overparametrization [OS19] and representation learning [Zha+17]. Moreover, this oracle model has also been studied under the umbrella of multiplicative noise [Ius+19] or growth conditions [VBS19; CV19; SR13; XWW20], and it is known to improve the convergence rate of stochastic gradient algorithms with non-adaptive step-sizes, even in non-smooth problems [FFF21].

Finally, in the online learning literature (multi-agent learning) the same noise model that we consider has been studied [Lin+20]. This particular noisy feedback model, as it is called in the community, allows them to get finite-time last-iterate convergence also in the unconstrained setting under cocoercivity with unknown constant but for a standard gradient update. However, crucially, they require the relative noise factor to vanish. We get rid of this requirement by employing an extragradient scheme with a different adaptivity, obtaining a $\mathcal{O}(1/T)$-rate for the ergodic average iterate.

With regard to the second question, we introduce a flexible first-order algorithmic template that includes as special cases the dual averaging [Nes09], dual extrapolation [Nes07] and optimistic gradient methods [Pop80; RS13], and which accounts for both adaptive and non-adaptive variants thereof.

To address the last remark on adaptivity, we make use of the tools and design paradigms in the literature and adapt them to the problem setting at hand. To summarize the relevant, prior work on adaptive methods, such schemes that achieve optimal rates even without knowing the noise constant have been considered before in the min-max optimization setting [BL19]. However, [BL19] focuses on the general noise model where only $\mathcal{O}(1/\sqrt{T})$ is possible. It is also worth pointing out that their step-size relies on the gradient mapping since they consider constrained min-max problems, while ours is based on the operator difference since we consider unconstrained VI problems.

In this sense, [BL19] is closer to the scheme in [ABM21], where they focus on adapting to non-smooth/smooth problems with unbounded domains in the *deterministic* setting. For

		Lipschitz		Cocoercive + rel. noise	
	V_t	Ergodic	Last Iterate	Ergodic	Last Iterate
Adapt. dual averaging	0	$1/\sqrt{T}$ [DHS11]	Unknown	$1/T$	Asym.
Adapt. dual extrapolation	AX_t + rel.noise	$1/\sqrt{T}$ [RS13]	Unknown	$1/T$	Asym.
Adapt. optimistic gradient	$AX_{t-1/2}$ + rel.noise	$1/\sqrt{T}$ [EN20]	Unknown	$1/T$	Asym.

Table 4.1: The best known convergence rates in stochastic monotone VIs with our contributions highlighted in gray. *Adaptive* refers to our particular adaptive step-size choice in (Adapt). We obtain various schemes with particular choices of V_t. For the nomenclature, please refer to Section 4.1.4.

the stochastic setting, there exists results for a single-call method using the same adaptive step-size as ours [EN20]. This work allows us to recover the $\mathcal{O}(T^{-1/2})$ convergence in the general case of Lipschitz operators for the particular instantiation of our algorithmic template.

In the light of related work, let us summarize our contributions as follows:

1. For oracles with bounded variance, we show that the proposed methods achieve an $\mathcal{O}(1/\sqrt{T})$ rate of convergence if run with a *non-adaptive*, decreasing step-size.
2. In the relative noise model, this rate improves to $\mathcal{O}(1/T)$, and it is achieved with a *constant* step-size that *does need to be tuned* as a function of T.
3. Finally, we provide an *adaptive* step-size rule that allows the method to achieve a fast, $\mathcal{O}(1/T)$ rate under relative noise, and an order-optimal $\mathcal{O}(1/\sqrt{T})$ rate in the absolute noise case.

Importantly, our work shows that an extra-gradient mechanism is *not* required to obtain a fast $\mathcal{O}(1/T)$ rate, as this can be achieved by vanilla dual-averaging methods with a *constant* step-size. This is an elegant consequence of the interplay between cocoercivity and the relative noise model; to the best of our knowledge, the only other work considering these models in tandem is the recent paper [Lin+20]. Our work closes several open threads in [Lin+20], which requires a *vanishing* relative noise level to obtain faster convergence in models with relative noise. A summary of our results in the context of related work can be found in Table 4.1.

4.1.3 Preliminaries

Examples and motivation. Throughout the sequel, we will focus on solving the variational inequality problem (VI). For completeness (and a certain degree of posterity), we briefly mention some examples below, and we defer to [FP03; Scu+10] for a panoramic survey of the field.

Example 1 (Convex Minimization)**.** If $A = \nabla f$ for some convex function f, the solutions of (VI) are precisely the minimizers of f.

Example 2 (Min-Max Problems)**.** If $A = (\nabla_{x_1} L, -\nabla_{x_2} L)$ for some convex-concave function $L(x_1, x_2)$, then the solutions of (VI) coincide with the (global) saddle points of L. More precisely,

$x^* = (x_1^*, x_2^*)$ is a solution of (VI) if and only if it holds that

$$L(x_1^*, x_2) \leq L(x_1^*, x_2^*) \leq L(x_1, x_2^*) \text{ for all } x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2. \tag{SP}$$

In this case, (VI) is sometimes referred to as the "vector field formulation" of (SP).

Example 3 (Monotone games)**.** Going beyond the min-max setting, a *continuous game in normal form* is defined as follows: First, consider a finite set of players $\mathcal{N} = \{1, \dots, N\}$, each with their own action space $\mathcal{X}_i = \mathbb{R}^{d_i}$. During play, each player selects an action x_i from $\mathcal{X}_i$ with the aim of minimizing a loss $\ell_i(x_i; x_{-i})$ determined by the ensemble $x := (x_i; x_{-i}) := (x_1, \dots, x_N)$ of all players' actions. In this context, a Nash equilibrium is any action profile $x^* \in \mathcal{X}$ that is *unilaterally stable*, i.e.,

$$\ell_i(x_i^*; x_{-i}^*) \leq \ell_i(x_i; x_{-i}^*) \quad \text{for all } x_i \in \mathcal{X}_i \text{ and all } i \in \mathcal{N}. \tag{NE}$$

The corresponding operator associated to the game is $A(x) = (\nabla_{x_i} \ell_i(x_i; x_{-i}))_{i \in \mathcal{N}}$. If A is monotone, then the game is itself called *monotone*, and its Nash equilibria coincide with the solutions of (VI), cf. [MS17; BLM18; MZ19; LS19; MS18; Mer+19; FP03; Scu+10] and references therein.

Regularity conditions. As we discussed in the introduction, our blanket regularity assumption for (VI) is that the defining operator A is β*-cocoercive* in the sense of (CC); for a panoramic overview of cocoercive operators we refer the reader to [BC17].

Some further comments for the cocoercivity condition are in order. First, one may easily observe that if A is β-cocoercive, it is also $1/\beta$-Lipschitz. The converse does not hold for the general setting of operators; however, when A is the gradient of a smooth convex function, this is indeed the case [BH77]. Moreover, even though cocoercivity implies that A is monotone, it does not imply that it is *strictly* monotone – a condition which is usually invoked to ensure the existence and uniqueness of solutions to (VI). Therefore, to avoid pathologies, we make the following assumption for our setting:

Assumption 4.1.1. *The set* $\mathcal{X}^* = \{x^* \in \mathbb{R}^d : x^* \text{ is a solution of (VI)}\}$ *is non-empty.*

Together with cocoercivity, the existence of a solution will be our only blanket assumption in the sequel.

The gap function. With the above setup in hand, a widely used performance measure in order to evaluate a candidate solution of (VI) is the so-called *restricted gap function*:

$$\mathrm{Gap}_{\mathcal{X}}(\hat{x}) = \sup_{x \in \mathcal{X}} \langle A(x), \hat{x} - x \rangle, \tag{Gap}$$

where the "test domain" $\mathcal{X}$ is a non-empty compact subset of $\mathbb{R}^d$. The motivation for this choice of merit function is that it characterizes the solutions of the (VI) via its zeros. Formally,

we have the following:

Proposition 4.1.1. *Let $\mathscr{X}$ be a non-empty convex subset of $\mathbb{R}^d$. Then, the following holds*

1. $\mathrm{Gap}_{\mathscr{X}}(\hat{x}) \geq 0$, *whenever* $\hat{x} \in \mathscr{X}$
2. *If* $\mathrm{Gap}_{\mathscr{X}}(\hat{x}) = 0$ *and* $\mathscr{X}$ *contains a neighbourhood of* $\hat{x}$, *then* $\hat{x}$ *is a solution of* (VI)

Proposition 4.1.1 is a generalization of an earlier characterization by Nesterov [Nes07]; see also [ABM19; Nes09] and references therein. Moreover, it provides a formal justification for the use of $\mathrm{Gap}_{\mathscr{X}}(\hat{x})$ as a merit function for (VI). To streamline our presentation we defer the proof of the above proposition to the appendix of this chapter.

Oracle structure and profiles of randomness

From an algorithmic point of view, in order to solve (VI) we will use iterative methods that require access to a stochastic first-order oracle [Nes03]. Formally, this is a black-box feedback mechanism which, when called at x, returns a random dual vector $g(x;\xi)$ with ξ drawn from some (complete) probability space $(\Omega, \mathscr{F}, \mathbb{P})$. In practice, the oracle will be called repeatedly at a (possibly random) sequence of points generated by the algorithm at play. Therefore, once the iterate of the method is generated at each round, the oracle draws an i.i.d. sample $\xi \in \Omega$ and returns a dual vector:

$$g(x;\xi) = A(x) + U(x;\xi) \tag{4.1}$$

with $U(x;\xi)$ denoting the "measurement error".

In this general setting, we make the following statistical assumptions for the oracle:

Assumption 4.1.2 (Absolute noise)**.** *The oracle $g(x;\xi)$ enjoys the following properties:*

1. Almost sure boundedness: *There exists some strictly positive numbers $M > 0$ such that:*

$$\|g(x;\xi)\|_* \leq M \ \textit{almost surely} \tag{4.2}$$

2. Unbiasedness: $\mathbb{E}\left[g(x;\xi)\right] = A(x)$
3. Bounded absolute variance: $\mathbb{E}\left[\|U(x;\xi)\|_*^2 | \sigma(x)\right] \leq \sigma^2$

Such type of conditions for the oracle are standard, especially in the context of adaptive methods cf. [Kav+19; LYC18; BL19]. Also, because the variance of the noise is independent of the value of the operator at the queried point, this type of randomness in the oracle will be called *absolute*.

By contrast, following Polyak [Pol87], the *relative* noise model is defined as follows:

Assumption 4.1.3 (Relative noise)**.** *The oracle $g(x;\xi)$ enjoys the following properties:*

1. Almost sure boundedness: *There exists some strictly positive numbers* $M > 0$ *such that:* $\|g(x;\xi)\|_* \leq M$ *almost surely*
2. Unbiasedness: $\mathbb{E}\left[g(x;\xi)\right] = A(x)$
3. Bounded relative variance: *There exists some positive* $c > 0$ *such that:*

$$\mathbb{E}\left[\|U(x;\xi)\|_*^2\right] \leq c\|A(x)\|_*^2 \tag{4.3}$$

Assumption 4.1.2 is standard for obtaining the typical $\mathcal{O}(1/\sqrt{T})$ convergence rate for stochastic optimization scenarios (see for example [Nem+09; JNT11] and references therein). That said, Assumption 4.1.3 will prove itself as the crucial statistical condition that will allow us to recover the well known order-optimal bound $\mathcal{O}(1/T)$ for deterministic settings. For concreteness, we provide an example below:

Example 4 (Random coordinate descent)**.** Consider a smooth convex function f over $\mathbb{R}^d$, as per Example 1. Then the randomized coordinate descent (RCD) algorithm draws one coordinate i_t at random at each stage, and calculates the partial derivative $v_{i,t} = \partial f/\partial x_{i_t}$. Subsequently, the i-th derivative is updated as $X_{i,t+1} = X_{i,t} - d\gamma_t v_{i,t}$.

This update rule can be written in abstract recursive form as $x^+ = x - g(x;\xi)$ where $g_i(x;\xi) = d\cdot \partial f/\partial x_i \cdot \xi$, d is the dimension of the domain of the objective and ξ is drawn uniformly at random from the set of basis vectors $\{e_1,\ldots,e_d\}$ of $\mathbb{R}^d$. Clearly, $\mathbb{E}[g(x;\xi)] = \nabla f(x)$ by construction; moreover, since $\partial f/\partial x_i = 0$ at the minimum points of f, we also have $g(x^*;\xi) = 0$ whenever x^* is a minimizer of f – i.e., the variance of the estimator $g(x;\xi)$ vanishes at the minimum points of f. It is then straightforward to verify that $\mathbb{E}[\|g(x;\xi) - \nabla f(x)\|^2] = \mathcal{O}(\|\nabla f(x)\|^2)$, which is precisely the relative noise condition for $A = \nabla f$.

4.1.4 Method

We now present the generalized extra-gradient (GEG) family of algorithms. More precisely, given two sequences of dual vectors V_t and $V_{t+1/2}$, (GEG) is given by the following recursive formula:

$$\begin{aligned} X_{t+1/2} &= X_t - \gamma_t V_t \\ Y_{t+1} &= Y_t - V_{t+1/2} \\ X_{t+1} &= \gamma_{t+1} Y_{t+1} \end{aligned} \tag{GEG}$$

Heuristically, the machinery behind (GEG) suggests to first generate a leading state $X_{t+1/2}$ by taking a step along V_t, then aggregate the vector $V_{t+1/2}$ observed at the leading state by incorporating the second dual sequence $V_{t+1/2}$ and finally update the method by applying a dual averaging step [Nes09; Xia10]. This idea is well-known in the literature of extra-gradient methods [Kor76; Nem04; Nes07]. However, up to this point, we have not assumed anything particular for the sequences of V_t and $V_{t+1/2}$, except that they are dual vectors (but not necessarily queries of a stochastic oracle). This generic choice is the building block that will allow us to include various popular algorithmic schemes and provide a unified framework of

their analysis.

For simplicity of notation, we denote $g(X_t, \xi_t) = A(X_t) + U(X_t, \xi_t)$ equivalently as $g_t = A(X_t) + U_t$. The same shorthand notation applies to g_{t-1}, $g_{t+\frac{1}{2}}$ and $g_{t-\frac{1}{2}}$, as well. We use the same notation for the sigma-algebra generated by the random sequence $\{\xi\}$ as in Table 2.2, such that $\mathscr{F}_t = \sigma(\xi_1, \xi_{1/2}, \cdots, \xi_{t-\frac{1}{2}}, \xi_t)$ and $\mathscr{F}_{t+1/2} = \sigma(\xi_1, \xi_{1/2}, \cdots, \xi_{t-\frac{1}{2}}, \xi_t, \xi_{t+\frac{1}{2}})$.

To begin with, we provide the following examples that illustrate the fact that Dual Averaging, Dual Extrapolation and Optimistic Dual Averaging can all be written in the form of (GEG) under different choices of V_t and $V_{t+1/2}$.

Example 5. Stochastic Dual Averaging [Nes09]: Consider the case $V_t \equiv 0$ and $V_{t+1/2} \equiv g_{t+1/2} = A(X_{t+1/2}) + U_{t+1/2}$. Then, this yields that $X_{t+1/2} = X_t$ and hence $g_{t+1/2} = g_t = V_{t+1/2}$. Therefore, (GEG) reduces to the dual averaging scheme:

$$\begin{aligned} Y_{t+1} &= Y_t - g_t \\ X_{t+1} &= \gamma_{t+1} Y_{t+1} \end{aligned} \tag{DA}$$

Example 6. Stochastic Dual Extrapolation [Nes07]: Consider the case now where $V_t \equiv g_t = A(X_t) + U_t$ and $V_{t+1/2} \equiv g_{t+1/2} = A(X_{t+1/2}) + U_{t+1/2}$ are noisy oracle queries at X_t and $X_{t+1/2}$ respectively. Then (GEG) readily yields Nesterov's dual extrapolation method [Nes07]:

$$\begin{aligned} X_{t+1/2} &= X_t - \gamma_t g_t \\ Y_{t+1} &= Y_t - g_{t+1/2} \\ X_{t+1} &= \gamma_{t+1} Y_{t+1} \end{aligned} \tag{DE}$$

Example 7. Stochastic Optimistic Dual Averaging [Pop80; RS13; HAM21; Hsi+22a]: Consider the case $V_t \equiv g_{t-1/2} = A(X_{t-1/2}) + U_{t-1/2}$ and $V_{t+1/2} \equiv g_{t+1/2} = A(X_{t+1/2}) + U_{t+1/2}$ are the noisy oracle feedback at $X_{t-1/2}$ and $X_{t+1/2}$ respectively. We then get the optimistic dual averaging method:

$$\begin{aligned} X_{t+1/2} &= X_t - \gamma_t g_{t-1/2} \\ Y_{t+1} &= Y_t - g_{t+1/2} \\ X_{t+1} &= \gamma_{t+1} Y_{t+1} \end{aligned} \tag{OptDA}$$

The next crucial step is to provide the key ingredient that will allow us to unify the approach for all algorithms belonging to the family (GEG). This is done by a shared "energy" inequality satisfied by all (GEG)-type schemes. Formally, this is described by the following proposition:

Proposition 4.1.2. *Assume that $X_t, X_{t+1/2}$ are the iterates of (GEG) run with a non-negative, non-increasing step-size γ_t. Then, for all $x \in \mathbb{R}^d$ the following inequality holds:*

$$\sum_{t=1}^{T} \langle V_{t+1/2}, X_{t+1/2} - x \rangle \leq \frac{\|x\|^2}{2\gamma_{T+1}} + \frac{1}{2} \sum_{t=1}^{T} \gamma_t \|V_{t+1/2} - V_t\|_*^2 - \frac{1}{2} \sum_{t=1}^{T} \frac{1}{\gamma_t} \|X_{t+1/2} - X_t\|^2 \tag{4.4}$$

Proving Proposition 4.1.2 requires tiresome computations, so we defer it to the end of the

section in order not to disrupt the narrative. The proof strategy follows the same steps as the proof of Proposition 2.2.1, but takes a slightly generalized approach to accommodate all the algorithmic instances under (GEG) template. Essentially, this template inequality is the basis of establishing the "regret" of the algorithms in question which immediately translates to the Gap function via averaging.

Moving forward, we conclude this section by illustrating the various method-specific template inequalities:

1. (Stochastic Dual Averaging): For $V_{t+1/2} = g_{t+1/2}$ and $V_t = 0$, then (4.4) becomes:

$$\sum_{t=1}^{T} \langle g_t, X_t - x \rangle \leq \frac{\|x\|^2}{2\gamma_{T+1}} + \frac{1}{2}\sum_{t=1}^{T} \gamma_t \|V_t\|_*^2 \tag{4.5}$$

2. (Stochastic Dual Extrapolation): For $V_t = g_t$ for all $t = 1, 2, \ldots$ then (4.4) becomes:

$$\sum_{t=1}^{T} \langle g_{t+1/2}, X_{t+1/2} - x \rangle \leq \frac{\|x\|^2}{2\gamma_{T+1}} + \frac{1}{2}\sum_{t=1}^{T} \gamma_t \left\|g_{t+1/2} - g_t\right\|_*^2 - \frac{1}{2}\sum_{t=1}^{T} \frac{1}{\gamma_t} \|X_{t+1/2} - X_t\|^2 \tag{4.6}$$

3. (Stochastic Optimistic Dual Averaging): For $V_t = g_{t-1/2}$ and $V_{t+1/2} = g_{t+1/2}$ then (4.4) becomes:

$$\sum_{t=1}^{T} \langle g_{t+1/2}, X_{t+1/2} - x \rangle \leq \frac{\|x\|^2}{2\gamma_{T+1}} + \frac{1}{2}\sum_{t=1}^{T} \gamma_t \left\|g_{t+1/2} - g_{t-1/2}\right\|_*^2 - \frac{1}{2}\sum_{t=1}^{T} \frac{1}{\gamma_t} \|X_{t+1/2} - X_t\|^2 \tag{4.7}$$

4.1.5 Analysis

Non-adaptive algorithm

In this section, we derive a series of tight convergence rates for (GEG) under both oracle/noise profiles but with a *non-adaptive* step-size sequences. We defer the full analysis to the appendix; however, we provide here a proof sketch of our main results via an appropriate "energy inequality" in Proposition 4.1.2.

Absolute random noise. In the context of monotone VIs, assumptions induced by the random oracle model are common and well-understood. Indeed, for the general case of bounded variance, i.e., $\mathbb{E}[U_{t+1/2}|\mathscr{F}_t] \leq \sigma$, extra-gradient/mirror-prox is known to converge at a rate $\mathcal{O}(1/\sqrt{T})$ [JNT11], with a decreasing step-size of order $\mathcal{O}(1/\sqrt{t})$. The decreasing step-size is essential in order to drive the effect of the variance to zero at a fast enough rate while maintaining the correct balance between the bias term and the (bounded) variance.

For completeness, we analyze (GEG) under a random oracle profile, i.e., for $V_{t+1/2} = g_{t+1/2} \equiv g(X_{t+1/2}; \xi_{t+1/2})$ satisfying Assumption 4.1.2 and V_t being an almost surely bounded sequence of dual vectors. To that end, we employ a decreasing step-size choice, which is summarized in

the next theorem.

Theorem 4.1.1. *Let $X_t, X_{t+1/2}$ be generated by (GEG) with a decreasing step-size $\gamma_t = \mathcal{O}(1/\sqrt{t})$. Then, for every compact neighborhood $\mathcal{X} \subset \mathbb{R}_d$ of x^*, with $\bar{X}_T = \frac{1}{T}\sum_{t=1}^{T} X_{t+1/2}$, it holds that:*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{X}}\left(\bar{X}_T\right)\right] = \mathcal{O}(1/\sqrt{T}).$$

The arguments for the proof of Theorem 4.1.1 are standard and we defer them to the appendix due to space constraints. Thanks to this result, we can now derive the respective method specific rates as special instances. More precisely, we have the following proposition:

Proposition 4.1.3. *Under Assumption 4.1.2 the iterates of (DA), (DE), (OptDA) enjoy the following rate:*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{X}}\left(\bar{X}_T\right)\right] = \mathcal{O}(1/\sqrt{T}) \tag{4.8}$$

Relative random noise We now turn our attention to the relative random oracle framework, i.e. $V_{t+1/2} = g_{t+1/2}$ satisfying Assumption 4.1.2 along with:

$$\mathbb{E}\left[\|V_t\|_*^2 | \mathcal{F}_{t-1/2}\right] \leq c\|A(X_t)\|_*^2 \text{ for all } t = 1, 1/2, \ldots \tag{4.9}$$

In particular, with a carefully chosen constant step-size, under the assumption of relative variance, it is possible to achieve an accelerated rate of $\mathcal{O}(1/T)$. One needs to depart from the standard approach to fully exploit the problem setting, i.e., making the correct use of cocoercivity and understanding the advantages of relative variance. Essentially, it amounts to ensuring that $\sum_{t=1}^{T}\|A_t\|_*^2$ and $\sum_{t=1}^{T}\|A_{t+1/2}\|_*^2$ are summable. Let us briefly motivate the idea behind handling the vanishing noise in this setting.

In the standard bounded variance regime, the error due to the noise in the orale information has, at best, a constant upper bound in expectation. Therefore, the algorithm suffers the same degree of measurement error no matter how close the iterates get to a solution. In the relative noise scheme, the error at iteration t is upper bounded by $O(A(X_t))$. Since we would like to find a solution which is also a zero of the (VI) problem, in the presence of Lipschitz continuity, our expectation is to observe a decrease in the noise magnitude as we approach to a solution. Then, what remains is to show that the error vanishes at a fast enough rate that the cumulative variance is summable, hence the faster rate. In this setting, what also enables us to use the constant step-size is the fact that cumulative variance is summable, and the algorithm could take more agressive steps without needing to drive the variance to zero in the limit.

Now, we present our result under the respective setting with a proof sketch that highlights its main ingredients.

Theorem 4.1.2. *Let $X_t, X_{t+1/2}$ be generated by (GEG) with a constant step-size that satisfies*

$$\min\left\{(2L)^{-1}, (4L^2\gamma)^{-1}\right\} - 2\gamma c > 0 \text{ with } L = 1/\beta. \tag{4.10}$$

Then, for every compact neighbourhood $\mathscr{X} \subset \mathbb{R}_d$ of x^*, with $\bar{X}_T = \frac{1}{T}\sum_{t=1}^{T} X_{t+1/2}$, we have:

$$\mathbb{E}\left[\mathrm{Gap}_{\mathscr{X}}\left(\bar{X}_T\right)\right] = \mathbb{E}\left[\sup_{X\in\mathscr{X}} \left\langle A(X), \bar{X}_T - X\right\rangle\right] = \mathscr{O}(1/T)$$

Proof. With a constant step-size, Proposition 4.1.2 implies

$$\sum_{t=1}^{T} \langle V_{t+1/2}, X_{t+1/2} - X\rangle = \frac{\|X\|^2}{2\gamma} + \frac{\gamma}{2}\sum_{t=1}^{T}\|V_{t+1/2} - V_t\|^2 - \frac{1}{2\gamma}\sum_{t=1}^{T}\|X_t - X_{t+1/2}\|^2$$

We show that using smoothness and cocoercivity of the operator, along with the relative noise condition,

$$\left(\min\left\{(2L)^{-1}, (4L^2\gamma)^{-1}\right\} - 2\gamma c\right)\sum_{t=1}^{T}\left(\mathbb{E}\left[\|A(X_t)\|^2\right] + \mathbb{E}\left[\|A(X_{t+1/2})\|^2\right]\right) \le \frac{\mathbb{E}\left[\|X\|^2\right]}{\gamma}$$

If constant step-size γ satisfies Eq. (4.42), then there exists some strictly positive real number β, such that $\mathbb{E}\left[\sum_{t=1}^{T}\left(\|A(X_t)\|^2 + \|A(X_{t+1/2})\|^2\right)\right] \le \mathbb{E}\left[\|X\|^2/\beta\gamma\right] < +\infty$, which concludes that both $\sum_{t=1}^{T}\|A_t\|_*^2$ and $\sum_{t=1}^{T}\|A_{t+1/2}\|_*^2$ are summable. Using the same arguments as in the proof of Theorem 4.1.1 and taking $\bar{[X_T]} = (1/T)\sum_{t=1}^{T} X_{t+\frac{1}{2}}$, we obtain an upper bound for the gap,

$$\mathbb{E}\left[\mathrm{Gap}_{\mathscr{X}}(\bar{[X_T]})\right] \le \frac{\frac{D^2}{2\gamma} + 2\gamma c\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_{t+1/2})\|^2 + \|A(X_t)\|^2\right] + \sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\|V_{t+1/2}\|_*^2\right]}}{T}.$$

By relative variance and summability of operators,

$$\mathbb{E}\left[\mathrm{Gap}_{\mathscr{X}}(\bar{[X_T]})\right] = \mathscr{O}(1/T)$$

■

Observe that it is important to design the *non-adaptive* step-size based specifically on the noise model; without the correct knowledge of the nature and properties of the noise, the non-adaptive algorithm wouldn't be able to achieve order-optimal convergence rates. For instance, using a decreasing step-size of the form $O(1/\sqrt{t})$ for the relative noise model, yields the same slow rate of $O(1/\sqrt{T})$ as in the standard bounded variance model. Exploiting the vanishing structure of the (stochastic) measurement error requires a more aggressive, constant step-size that incorporates the relative noise parameter c.

Similar to the setting of absolutely random noise, Theorem 4.1.2 implies algorithm-specific convergence bounds, which are presented below:

Proposition 4.1.4. *Under Assumption 4.1.3 the iterates of* (DA), (DE), (OptDA) *enjoy the following rate:*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathscr{X}}\left(\bar{X}_T\right)\right] = \mathscr{O}(1/T) \tag{4.11}$$

An extra appealing feature of the above is that we are able to derive an asymptotic last iterate trajectory result, i.e., the asymptotic convergence of the iterates themselves before any averaging occurs, almost surely. More precisely, we have the following proposition:

Proposition 4.1.5. *Under Assumption 4.1.3 the iterates of* (DA), (DE), (OptDA) *converge to a* (VI) *solution* x^*.

The proof Proposition 4.1.5 relies on the fact that the distance of the iterates towards any solution of (VI) is decreasing almost surely along with the fact that the summability of $\|A(X_t)\|_*^2$ guarantees that every limit point of the iterate is also a solution of (VI). To streamline our presentation, we defer the detailed proof to the appendix.

Adaptive algorithm

By the results of the previous section on non-adaptive methods, one may easily observe the interplay between the $\mathcal{O}(1/\sqrt{T})$ to $\mathcal{O}(1/T)$ convergence rates under different noise profiles and step-sizes policies. Therefore a natural question that arises from this context is the following:

Can we derive a universal step-size policy that is able to optimally adjust the performance of (GEG) *without any prior knowledge of the oracle's noise profile?*

In what follows, this desired property is achieved by running (GEG) with the following adaptive step-size:

$$\gamma_t = \frac{1}{\sqrt{1 + \sum_{j=1}^{t-1} \|V_j - V_{j+1/2}\|_*^2}} \tag{Adapt}$$

The step-size (Adapt) is inspired by [RS13]; however, in our analysis, we provide a generalized point of view *which does not assume* that V_t necessarily is the oracle query at the respective points as in [RS13]. This allows us to include in the (Adapt) formulation all the adaptive step-sizes typically used for the archetypical schemes introduced in Section 4.1.4. More precisely, we have:

1. *Adaptive Stochastic Dual Averaging:* For $V_t \equiv 0$ (Adapt) becomes the standard AdaNorm stepsize, studied in various works [DHS11; MS10]:

$$\gamma_t = \frac{1}{\sqrt{1 + \sum_{j=1}^{t-1} \|g_j\|_*^2}} \tag{4.12}$$

2. *Adaptive Stochastic Dual Extrapolation:* For $V_t = g_{t+1/2}$ (Adapt) becomes

$$\gamma_t = \frac{1}{\sqrt{1 + \sum_{j=1}^{t-1} \|g_j - g_{j+1/2}\|_*^2}} \tag{4.13}$$

   as used in, e.g., [RS13; Syr+15; ABM21].

3. *Adaptive Stochastic Optimistic Dual Averaging:*. For $V_t = g_{t-1/2}$ (Adapt) becomes the step-size used in [Hsi+22a; HAM21]:

$$\gamma_t = \frac{1}{\sqrt{1 + \sum_{j=1}^{t-1} \|g_{j+1/2} - g_{j-1/2}\|_*^2}} \tag{4.14}$$

Our results in the non-adaptive setting heuristically suggests that the success of γ_t should hinge on a simultaneous performance; decreasing roughly at a rate of $1/\sqrt{t}$ for the absolute random oracle feedback and behaving as a constant step-size whenever the relative random feedback kicks in. This important interpolation feature is what will show in the sequel.

Absolute random noise. We will first treat oracles subject to absolute random noise. As we have mentioned above, the main goal of the analysis for this noise regime is to prove that the step-size decreases at a particular rate, similar to that of the non-adaptive counterpart. In this case, we have the following result.

Theorem 4.1.3. *Assume that $X_t, X_{t+1/2}$ are the iterates of* (GEG) *run with the step-size* (Adapt). *Then, for every compact neighborhood $\mathcal{X} \subset \mathbb{R}^d$ of a solution x^* of* (VI), *we have:*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{X}}(\overline{X}_T)\right] = \mathcal{O}(1/\sqrt{T}) \tag{4.15}$$

with $\overline{X}_T = (1/T)\sum_{t=1}^{T} X_{t+1/2}$

Now let us provide an insight into the elements of the proof. As we have stated earlier, the proof hinges on quantifying the behavior of the adaptive step-size. Our claim is that in the presence of absolute noise, the adaptive step-size behaves similary to its non-adaptive counterpart. This is justified by the fact that the almost sure boundedness conditions for the sequences,

$$\|V_t\|_* \leq M \text{ almost surely for all } t = 1, 1/2, \ldots \tag{4.16}$$

implies that $\gamma_t = \Omega(1/\sqrt{t})$, which is in line with the respective result for the non-adaptive algorithm. To handle the adaptive step-sizes, we once again resort to numerical inequalities presented in Lemma 2.1.2 and 3.1.1. Through our generalized analysis and representations, the above result implies the same order of convergence for each particular algorithmic instance.

Proposition 4.1.6. *Under Assumption 4.1.2 the iterates of* (DA), (DE), (OptDA) *enjoy the following:*

$$\mathbb{E}\left[Gap_{\mathcal{X}}(\overline{X}_T)\right] = \mathcal{O}(1/\sqrt{T}) \tag{4.17}$$

Relative random noise. Under the relative random noise condition, we can obtain the improved rate $\mathcal{O}(1/T)$, going beyond the $\mathcal{O}(1/\sqrt{T})$ rate above. Recall that the faster rate under the relative noise setting was possible due to the use of a carefully tuned, *constant* step-size which depends on the knowledge of smoothness and relative noise parameters of the problem.

The literature on parameter-free methods and the results we have seen in the preivous chapters suggest that data-adaptive step-sizes of the form (Adapt) are capable of adapting to smoothness constant. Our goal is to argue that they simultaneously adapt to the relative noise parameter c as well as *the type of the noise model* without prior knowledge of the setting. We will elaborate on the main techniques through a concise sketch of the analysis in the sequel. Formally, we have the following result which is achieved without making any changes on the algorithmic framework.

Theorem 4.1.4. *Assume that $X_t, X_{t+1/2}$ are the iterates of (GEG) run with the step-size (Adapt). Then, for every compact neighborhood $\mathcal{X} \subset \mathbb{R}^d$ of a solution x^* of (VI), we have:*

$$\mathbb{E}\left[Gap_{\mathcal{X}}(\overline{X}_T)\right] = \mathcal{O}(1/T) \tag{4.18}$$

with $\overline{X}_T = 1/T \sum_{t=1}^{T} X_{t+1/2}$

Note that the convergence results under both the absolute noise (Assumption 4.1.2) and relative noise (Assumption 4.1.3) holds for exactly the same adaptive step-size defined in (Adapt) without any modification. We identify a delicate relationship between the growth of the adaptive step-size and the vanishing nature of the noise. The crucial ingredient for the proof of Theorem 4.1.4 consists of showing that the adaptive step size stabilizes to a positive constant $\gamma_\infty > 0$ when the noise gradually vanishes towards the solution of the VI problem according to the dynamics defined in Eq. (4.3). In order to obtain this, the first step is to show that the template inequality of Proposition 4.1.2 yields

$$\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right] \leq \left(8c\max\{L, 2L^2\}\left(\frac{\|x^* - x_1\|^2}{2} + 2G^2 + 1\right) + 1\right)\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] \tag{4.19}$$

Moreover, due to the definition of (Adapt) and Jensen's inequality we have:

$$\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] = \mathbb{E}\left[\sqrt{1 + \sum_{t=1}^{T}\|V_t - V_{t+1/2}\|^2}\right] \leq \sqrt{\mathbb{E}\left[1 + \sum_{t=1}^{T}\|V_t - V_{t+1/2}\|^2\right]} = \sqrt{\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right]} \tag{4.20}$$

Therefore, after combining (4.19) and (4.20) we get that $\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right] < +\infty$. This directly implies (by the monotone convergence theorem) that:

$$\frac{1}{\gamma_{T+1}^2} = 1 + \sum_{t=1}^{T}\|V_t - V_{t+1/2}\|_*^2 < +\infty \text{ almost surely} \tag{4.21}$$

which in turn yields that $\sum_{t=1}^{T}\|V_t - V_{t+1/2}\|_*^2$ is summable almost surely. Therefore due to the definition of γ_t we have almost surely the following:

$$\gamma_{T+1} = \frac{1}{\sqrt{1 + \sum_{t=1}^{T}\|V_t - V_{t+1/2}\|_*^2}} \to \frac{1}{\sqrt{1 + \sum_{t=1}^{+\infty}\|V_t - V_{t+1/2}\|_*^2}} = \gamma_\infty > 0 \tag{4.22}$$

Finally, we conclude by providing the respective method-specific result.

Proposition 4.1.7. *Under Assumption 4.1.3 the iterates of* (DA), (DE), (OptDA) *enjoy the following:*

1. *The convergence rate in terms of the restricted gap function for the time-average:*

$$\mathbb{E}\left[Gap_{\mathscr{X}}(\overline{X}_T)\right] = \mathcal{O}(1/T) \tag{4.23}$$

2. *Their last iterate trajectory converges to a* (VI) *solution* x^* *almost surely.*

The last iterate convergence result of Proposition 4.1.7 refers to the asymptotic convergence of the actual sequences of the methods-before any averaging takes place- and it hinges on the fact that the (random) sequences $\|V_t - V_{t+1/2}\|_*^2$ and $\|A(X_t)\|_*^2$ are summable with probability 1. Having established this, we show that X_t is a (stochastic) quasi-Fejér sequence [CP15b] (with respect to the solution set $\mathscr{X}^*$) along with the fact that every limit point of X_t belongs to $\mathscr{X}^*$. These two building blocks are sufficient in order to derive the almost sure convergence of the iterate's trajectory.

4.1.6 Experiments

In this section we validate and explore the consequences of the theoretical results. We adopt the experimental setting considered in [GP19] which is a particular instance of the Kelly auction with $N = 4$. In its generality in a single resource Kelly auction, there are N players sharing a total amount of $Q \in \mathbb{R}_{>0}$ resources. At every round, each bidder, p, submits a bid $x^p \in \mathbb{R}_{\geq 0}$ and receives proportional resources, $\rho^p = \frac{Qx^p}{Z+\sum_p x^p}$, where Z is the auction entry price. The payoff for player p is then given as $u^p(x^p; x^{-p}) = G^p \rho^p - x^p$, where G^p is the marginal gain in utility for player p. One can easily verify that the vector field associated with the payoff functions is cocoercive. In addition, the assumption of relative noise can be justified since each player can be seen as performing a measurement when querying the payoff. In such settings, it is common to assume that the error is proportional to the measured quantity and this uncertainty propagates to the gradient information in the form of relative noise. Since players act without communication in this example, it is particularly important that our results extends to single-call extragradient variants (see for instance [RS13] for elaboration). However, note that our proposed adaptive step-size (Adapt) still relies on global information of all players so our non-adaptive results for known problem constants is also important for this example.

In order to simulate the presence of relative noise we add a term proportional to the norm of the operator. In our notation we can thus capture both relative noise and absolute noise through the error term U_t in the following way,

$$U_t = \epsilon_{\text{rel}} \|A(X_t)\| + \epsilon_{\text{abs}}, \tag{4.24}$$

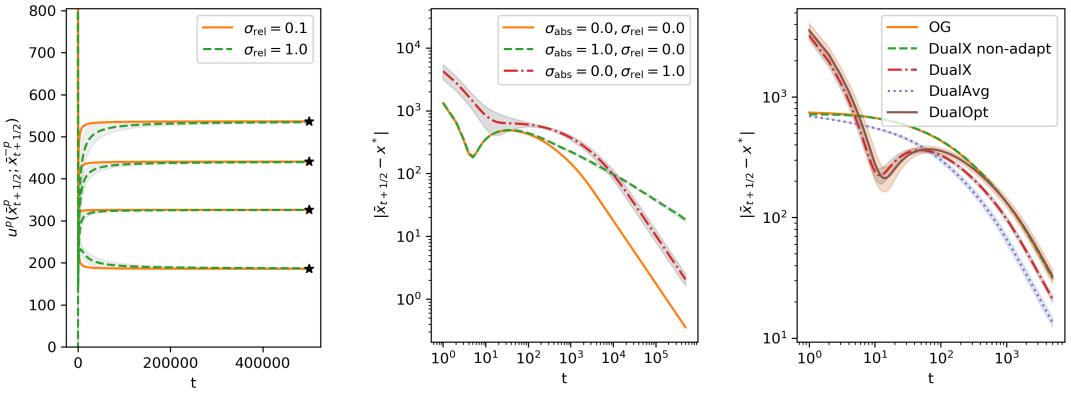


Figure 4.1: (left) Player utility using adaptive DualX for various relative noise levels. Even at relative high levels of noise do we converge to the optimal depicted with ($\star$). (center) Average iterate for deterministic, absolute noise and relative noise using adaptive DualX. We observe the $\mathcal{O}(1/\sqrt{T})$ rate under absolute noise while $\mathcal{O}(1/T)$ is achieved both in the noiseless setting and under relative noise. In addition, as expected, the last iterate only converges under the deterministic and relative noise oracle. (right) Average iterate comparing various methods for $\sigma_{\text{rel}} = 0.1$. All methods shares convergence rate with adaptive methods being slightly faster possibly because of difficulty of step-size tuning for non-adaptive methods. Error bars indicate one standard deviation computed using 10 independent executions.

where $\epsilon_{\text{rel}} \sim \mathcal{N}(0, \sigma^2_{\text{rel}})$ and $\epsilon_{\text{abs}} \sim \mathcal{N}(0, \sigma^2_{\text{abs}})$. To validate the convergence rate we compute the optimal strategy in the deterministic setting (i.e. $\sigma_{\text{rel}} = \sigma_{\text{abs}} = 0$) using Mathematica.

In Fig. 4.1 we illustrate the behavior of the different instantiations of our algorithmic template under different choices of σ_{rel} and σ_{abs}. To denote (DA), (DE) and (OptDA) we use DualAvg, DualX and DualOpt respectively. In addition we include optimistic gradient (OG) from [Das+18] for comparison. For higher dimensional experiments see the appendix, where we additionally apply our adaptive method to the non-convex problem of learning a covariance matrix [Das+18; Hsi+20].

4.1.7 Conclusion

In this paper we provide rate interpolation guarantees for different noise profiles; namely that of absolute and relative random noise. That being said our analysis crucially depends on the cocoercivity of the associated operator that defines the respective (VI). It thus remains open whether it is possible to achieve the same $\mathcal{O}(1/T)$ rate for monotone (VI) by only assuming Lipschitz continuity of the said operator and relative noise. Moreover, an additional interesting direction for future research is investigate the impact of relative noise for adaptive accelerated methods and whether it is possible to recover the iconic $\mathcal{O}(1/T^2)$ rate. We postpone these questions to the future.

4.2 APPENDIX: Proofs of Chapter 4

Let us begin with the basic properties of the restricted merit function $\text{Gap}_{\mathscr{X}}$ introduced in (Gap). For completeness, we provide the proof of Proposition 4.1.1, which itself is an extension of a similar result by [Nes07]:

Proposition 4.1.1. *Let $\mathscr{X}$ be a non-empty convex subset of $\mathbb{R}^d$. Then, the following holds*

1. $\text{Gap}_{\mathscr{X}}(\hat{x}) \geq 0$, *whenever* $\hat{x} \in \mathscr{X}$
2. *If* $\text{Gap}_{\mathscr{X}}(\hat{x}) = 0$ *and* $\mathscr{X}$ *contains a neighbourhood of* $\hat{x}$, *then* $\hat{x}$ *is a solution of* (VI)

Proof. Let $x^* \in \mathscr{X}$ be a solution of (VI) so $\langle A(x^*), x - x^* \rangle \geq 0$ for all $x \in \mathscr{X}$. Then, by monotonicity, we get:

$$\begin{aligned} \langle A(x), x^* - x \rangle &\leq \langle A(x) - A(x^*), x^* - x \rangle + \langle A(x^*), x^* - x \rangle \\ &= -\langle A(x^*) - A(x), x^* - x \rangle - \langle A(x^*), x - x^* \rangle \leq 0, \end{aligned} \tag{4.25}$$

so $\text{Gap}_{\mathscr{X}}(x^*) \leq 0$. On the other hand, if $x^* \in \mathscr{X}$, we also get $\text{Gap}(x^*) \geq \langle A(x^*), x^* - x^* \rangle = 0$, so we conclude that $\text{Gap}_{\mathscr{X}}(x^*) = 0$.

For the converse statement, assume that $\text{Gap}_{\mathscr{X}}(\hat{x}) = 0$ for some $\hat{x} \in \mathscr{X}$ and suppose that $\mathscr{X}$ contains a neighborhood of $\hat{x}$ in $\mathscr{X}$. First, we claim that the following inequality holds:

$$\langle A(x), x - \hat{x} \rangle \geq 0 \quad \text{for all } x \in \mathscr{X}. \tag{4.26}$$

Indeed, assume to the contrary that there exists some $x_1 \in \mathscr{X}$ such that

$$\langle A(x_1), x_1 - \hat{x} \rangle < 0. \tag{4.27}$$

This would then give

$$0 = \text{Gap}_{\mathscr{X}}(\hat{x}) \geq \langle A(x_1), \hat{x} - x_1 \rangle > 0, \tag{4.28}$$

which is a contradiction. Now, we further claim that $\hat{x}$ is a solution of (VI),i.e.,:

$$\langle A(\hat{x}), x - \hat{x} \rangle \geq 0 \text{ for all } x \in \mathscr{X}. \tag{4.29}$$

If we suppose that there exists some $z_1 \in \mathscr{X}$ such that $\langle A(\hat{x}), z_1 - \hat{x} \rangle < 0$, then, by the continuity of A, there exists a neighborhood $\mathscr{U}'$ of $\hat{x}$ in $\mathscr{X}$ such that

$$\langle A(x), z_1 - x \rangle < 0 \quad \text{for all } x \in \mathscr{U}'. \tag{4.30}$$

Hence, assuming without loss of generality that $\mathscr{U}' \subset \mathscr{U} \subset \mathscr{X}$ (the latter assumption due to the assumption that $\mathscr{X}$ contains a neighborhood of $\hat{x}$), and taking $\lambda > 0$ sufficiently small so that $x = \hat{x} + \lambda(z_1 - \hat{x}) \in \mathscr{U}'$, we get that $\langle A(x), x - \hat{x} \rangle = \lambda \langle A(x), z_1 - \hat{x} \rangle < 0$, in contradiction to (4.26). We conclude that $\hat{x}$ is a solution of (VI), as claimed. ■

Next, we shall provide the proof of the template inequality of Proposition 4.1.2. As we already argued in the main, this energy inequality will serve as a template for deriving the method specific convergence rates in the sequel, in a similar sense to template inequality (Proposition 2.2.1) in Section 2.2.4. Formally, we have the following:

Proposition 4.1.2. *Assume that $X_t, X_{t+1/2}$ are the iterates of* (GEG) *run with a non-negative, non-increasing step-size γ_t. Then, for all $x \in \mathbb{R}^d$ the following inequality holds:*

$$\sum_{t=1}^{T} \langle V_{t+1/2}, X_{t+1/2} - x\rangle \leq \frac{\|x\|^2}{2\gamma_{T+1}} + \frac{1}{2}\sum_{t=1}^{T} \gamma_t \|V_{t+1/2} - V_t\|_*^2 - \frac{1}{2}\sum_{t=1}^{T} \frac{1}{\gamma_t}\|X_{t+1/2} - X_t\|^2 \tag{4.31}$$

Proof. By the update rule for X_{t+1} in (GEG) we get the following:

$$\begin{aligned}
\langle V_{t+1/2}, X_{t+1} - x\rangle &= \left\langle \frac{1}{\gamma_t}\gamma_t Y_t - \frac{1}{\gamma_{t+1}}\gamma_{t+1}Y_{t+1}, X_{t+1} - x\right\rangle \\
&= \left\langle \frac{1}{\gamma_t}\gamma_t Y_t - \frac{1}{\gamma_t}\gamma_{t+1}Y_{t+1}, X_{t+1} - x\right\rangle + \left\langle \frac{1}{\gamma_t}\gamma_{t+1}Y_{t+1} - \frac{1}{\gamma_{t+1}}\gamma_{t+1}Y_{t+1}, X_{t+1} - x\right\rangle \\
&= \frac{1}{\gamma_t}\left\langle \gamma_t Y_t - \gamma_{t+1}Y_{t+1}, X_{t+1} - x\right\rangle + \left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right)\left\langle 0 - \gamma_{t+1}Y_{t+1}, X_{t+1} - x\right\rangle. \\
&= \frac{1}{\gamma_t}\left\langle X_t - X_{t+1}, X_{t+1} - x\right\rangle + \left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right)\left\langle 0 - X_{t+1}, X_{t+1} - x\right\rangle
\end{aligned}$$

Therefore, by quadratic expansion of the scalar products $\langle X_t - X_{t+1}, X_{t+1} - x\rangle$and $\langle 0 - X_{t+1}, X_{t+1} - x\rangle$ we get:

$$\begin{aligned}
\langle V_{t+1/2}, X_{t+1} - x\rangle = \frac{1}{\gamma_t}&\left[\frac{1}{2}\|X_{t+1} - x + X_t - X_{t+1}\|^2 - \frac{1}{2}\|X_t - X_{t+1}\|^2 - \frac{1}{2}\|X_{t+1} - x\|^2\right] \\
&+ \left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right)\left[\frac{1}{2}\|X_{t+1} - x - X_{t+1}\|^2 - \frac{1}{2}\|X_{t+1}\|^2 - \frac{1}{2}\|X_{t+1} - x\|^2\right]
\end{aligned} \tag{4.32}$$

which in turn yields:

$$\begin{aligned}
\langle V_{t+1/2}, X_{t+1} - x\rangle \leq \frac{1}{2\gamma_t}\|X_t - x\|^2 - \frac{1}{2\gamma_t}\|X_t - X_{t+1}\|^2 - \frac{1}{2\gamma_t}\|X_{t+1} - x\|^2 + \frac{1}{2}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right)\|x\|^2 \\
- \frac{1}{2\gamma_{t+1}}\|X_{t+1} - x\|^2 + \frac{1}{2\gamma_t}\|X_{t+1} - x\|^2
\end{aligned} \tag{4.33}$$

Therefore, after rearranging we have first part of the proof,

$$\begin{aligned}
\frac{1}{2\gamma_{t+1}}\|X_{t+1} - x\|^2 &\leq \frac{1}{2\gamma_t}\|X_t - x\|^2 + \frac{1}{2}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right)\|x\|^2 - \langle V_{t+1/2}, X_{t+1} - x\rangle - \frac{1}{2\gamma_t}\|X_t - X_{t+1}\|^2 \\
&= \frac{1}{2\gamma_t}\|X_t - x\|^2 + \frac{1}{2}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right)\|x\|^2 - \langle V_{t+1/2}, X_{t+1/2} - x\rangle \\
&\quad + \langle V_{t+1/2}, X_{t+1/2} - X_{t+1}\rangle - \frac{1}{2\gamma_t}\|X_t - X_{t+1}\|^2.
\end{aligned}$$

On the other hand, by invoking the update rule of $X_{t+1/2}$ in (GEG) we have:

$$\begin{aligned}\gamma_t \langle V_t, X_{t+1/2} - x\rangle &= \langle X_t - X_{t+1/2}, X_{t+1/2} - x\rangle \\ &= \frac{1}{2}\|X_{t+1/2} - x + X_t - X_{t+1/2}\|^2 - \frac{1}{2}\|X_t - X_{t+1/2}\|^2 - \frac{1}{2}\|X_{t+1/2} - x\|^2 \\ &= \frac{1}{2}\|X_t - x\|^2 - \frac{1}{2}\|X_t - X_{t+1/2}\|^2 - \frac{1}{2}\|X_{t+1/2} - x\|^2,\end{aligned} \tag{4.34}$$

and after dividing with γ_t and rearranging and setting $x = X_{t+1}$

$$\frac{1}{2\gamma_t}\|X_t - X_{t+1/2}\|^2 + \frac{1}{2\gamma_t}\|X_{t+1/2} - X_{t+1}\|^2 + \langle V_t, X_{t+1/2} - X_{t+1}\rangle = \frac{1}{2\gamma_t}\|X_t - X_{t+1}\|^2. \tag{4.35}$$

So, combining the above, we get

$$\begin{aligned}\frac{1}{2\gamma_{t+1}}\|X_{t+1} - x\|^2 &\le \frac{1}{2\gamma_t}\|X_t - x\|^2 + \frac{1}{2}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right)\|x\|^2 - \langle V_{t+1/2}, X_{t+1/2} - x\rangle \\ &\quad + \langle V_{t+1/2}, X_{t+1/2} - X_{t+1}\rangle - \langle V_t, X_{t+1/2} - X_{t+1}\rangle \\ &\quad - \frac{1}{2\gamma_t}\|X_t - X_{t+1/2}\|^2 - \frac{1}{2\gamma_t}\|X_{t+1/2} - X_{t+1}\|^2 \\ &\le \frac{1}{2\gamma_t}\|X_t - x\|^2 - \langle V_{t+1/2}, X_{t+1/2} - x\rangle + \frac{1}{2}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right)\|x\|^2 \\ &\quad \underbrace{+ \langle V_{t+1/2} - V_t, X_{t+1/2} - X_{t+1}\rangle - \frac{1}{2\gamma_t}\|X_{t+1/2} - X_{t+1}\|^2}_{\text{(A)}} - \frac{1}{2\gamma_t}\|X_t - X_{t+1/2}\|^2.\end{aligned} \tag{4.36}$$

Now, we handle term (A):

$$\begin{aligned}&\langle V_{t+1/2} - V_t, X_{t+1/2} - X_{t+1}\rangle - \frac{1}{2\gamma_t}\|X_{t+1/2} - X_{t+1}\|^2 \\ &\le \frac{1}{2}\gamma_t\|V_{t+1/2} - V_t\|_*^2 + \frac{1}{2\gamma_t}\|X_{t+1/2} - X_{t+1}\|^2 - \frac{1}{2\gamma_t}\|X_{t+1/2} - X_{t+1}\|^2 \\ &\le \frac{1}{2}\gamma_t\|V_{t+1/2} - V_t\|_*^2.\end{aligned} \tag{4.37}$$

By integrating the bound on (A) back into the original expresion yields,

$$\begin{aligned}\frac{1}{2\gamma_{t+1}}\|X_{t+1} - x\|^2 \le \frac{1}{2\gamma_t}\|X_t - x\|^2 - \langle V_{t+1/2}, X_{t+1/2} - x\rangle + \frac{1}{2}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right)\|x\|^2 \\ + \frac{1}{2}\gamma_t\|V_t - V_{t+1/2}\|_*^2 - \frac{1}{2\gamma_t}\|X_t - X_{t+1/2}\|^2.\end{aligned} \tag{4.38}$$

So, after rearranging and telescoping over $t = 1, \dots, T$ we get:

$$\begin{aligned}\sum_{t=1}^{T} \langle V_{t+1/2}, X_{t+1/2} - x\rangle \leq \frac{\|X_1 - x\|^2}{2\gamma_1} + \frac{\|x\|^2}{2\gamma_{T+1}} - \frac{\|x\|^2}{2\gamma_1} \\ + \frac{1}{2}\sum_{t=1}^{T} \gamma_t \|V_t - V_{t+1/2}\|_*^2 - \frac{1}{2}\sum_{t=1}^{T} \frac{\|X_t - X_{t+1/2}\|^2}{\gamma_t}.\end{aligned} \tag{4.39}$$

The result follows by setting $X_1 = 0$ and simplifying the expression. ■

Before we begin with the proof of the main results, we need to present a lemma which is crucial in handling the fixed point which essentially depends on the randomness in the whole of the process due to the definition of the Gap function. We have the following result that will help us to deal with the martingale difference component of the "noise", which is due to Bach and Levy [BL19].

Lemma 4.2.1. *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set and $h : \mathcal{X} \to \mathbb{R}$ be a 1-strongly-convex with respect to a $\|\cdot\|$ over $\mathcal{X}$. Also, assume that $\forall x \in \mathcal{X}$, $h(x) - \min_{x\in\mathcal{X}} h(x) \leq \frac{D^2}{2}$. Then, for any martingale difference $(Z_t)_{t=1}^T \in \mathbb{R}^d$, and any random vector $x \in \mathcal{X}$, we have:*

$$\mathbb{E}\left[\left\langle \sum_{t=1}^{T} Z_t, x \right\rangle\right] \leq \frac{D}{2}\sqrt{\sum_{t=1}^{T} \mathbb{E}\left[\|Z_t\|_*^2\right]} \tag{4.40}$$

The proof of the above lemma could be found in [BL19], where they present the same result under the label Proposition B.1.

We are now at a position to present the main results, starting with the proofs for the non-adaptive schemes studied in this chapter.

Theorem 4.1.1. *Let $X_t, X_{t+1/2}$ be generated by (GEG) with a decreasing step-size $\gamma_t = \mathcal{O}(1/\sqrt{t})$. Then, for every compact neighborhood $\mathcal{X} \subset \mathbb{R}_d$ of x^*, with $\bar{X}_T = \frac{1}{T}\sum_{t=1}^{T} X_{t+1/2}$, it holds that:*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{X}}\left(\bar{X}_T\right)\right] = \mathcal{O}(1/\sqrt{T}).$$

Proof. Since we adopt a decreasing step-size schedule of $O(1/\sqrt{t})$, Proposition 4.1.2 immediately applies to this setting. Combining this with almost sure boundedness of stochastic operators,

$$\begin{aligned}\sum_{t=1}^{T} \langle V_{t+1/2}, X_{t+1/2} - x\rangle &\leq \frac{\|X\|^2}{2\gamma_{T+1}} + \frac{1}{2}\sum_{t=1}^{T} \gamma_t \|V_{t+1/2} - V_t\|_*^2 - \frac{1}{2}\sum_{t=1}^{T} \frac{1}{\gamma_t}\|X_{t+1/2} - X_t\|^2 \\ &\leq \frac{\|x\|^2}{2}\sqrt{T+1} + \sum_{t=1}^{T} \gamma_t \|V_{t+1/2}\|_*^2 + \gamma_t \|V_t\|_*^2 \\ &\leq \frac{\|x\|^2}{2}\sqrt{T+1} + 2M^2\sqrt{T}.\end{aligned}$$

By monotonicity, and the definition that $V_{t+1/2} = A(X_{t+1/2}) + U_{t+1/2}$,

$$\begin{aligned}\sum_{t=1}^{T} \langle V_{t+1/2}, X_{t+1/2} - x\rangle &= \sum_{t=1}^{T} \langle A(X_{t+1/2}), X_{t+1/2} - x\rangle + \langle U_{t+1/2}, x - X_{t+1/2}\rangle \\ &\geq \sum_{t=1}^{T} \langle A(x), X_{t+1/2} - x\rangle + \langle U_{t+1/2}, X_{t+1/2} - x\rangle \\ &= T\left\langle A(x), \bar{X}_T - x\right\rangle + \sum_{t=1}^{T} \langle U_{t+1/2}, X_{t+1/2} - x\rangle\end{aligned}$$

Plugging this lower bound into the first expression,

$$\left\langle A(x), \bar{X}_T - x\right\rangle \leq \frac{\left(\frac{\|x\|^2}{2} + 2M^2\right)\sqrt{T+1} + \sum_{t=1}^{T} \langle U_{t+1/2}, x - X_{t+1/2}\rangle}{T}$$

Taking supremum over $x \in \mathscr{X}$ and finally computing expectation with respect to all randomness we obtain

$$\mathbb{E}\left[\mathrm{Gap}_{\mathscr{X}}(\bar{X}_T)\right] \leq \frac{\mathbb{E}\left[\sup_{x\in\mathscr{X}}\left\{\left(\frac{\|x\|^2}{2} + 2M^2\right)\sqrt{T+1} + \underbrace{\sum_{t=1}^{T} \langle U_{t+1/2}, x\rangle}_{\text{(A)}} - \underbrace{\sum_{t=1}^{T} \langle U_{t+1/2}, X_{t+1/2}\rangle}_{\text{(B)}}\right\}\right]}{T}.$$

For term (A),

$$\begin{aligned}\mathbb{E}\left[\sup_{x\in\mathscr{X}} \sum_{t=1}^{T} \langle U_{t+1/2}, x\rangle\right] &\leq \mathbb{E}\left[\max_{x\in\mathscr{X}} \left\langle \sum_{t=1}^{T} U_{t+1/2}, x\right\rangle\right] \\ &= \mathbb{E}\left[\left\langle \sum_{t=1}^{T} U_{t+1/2}, \tilde{x}\right\rangle\right] && \text{(for some } \tilde{x} \in \mathscr{X} \text{ which attains the maximum)} \\ &= \frac{D}{2}\sqrt{\sum_{t=1}^{T} \mathbb{E}\left[\|U_{t+1/2}\|_*^2\right]} && \text{((by Lemma 4.2.1))} \\ &= \frac{D}{2}\sqrt{\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\|U_{t+1/2}\|_*^2 | \mathscr{F}_t\right]\right]} \\ &= \frac{D}{2}\sigma\sqrt{T} && \text{(Bounded variance)}\end{aligned}$$

Also, for term (B),

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle U_{t+1/2}, X_{t+1/2}\rangle\right] = \sum_{t=1}^{T} \mathbb{E}\left[\langle U_{t+1/2}, X_{t+1/2}\rangle\right]$$

$$
\begin{aligned}
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\langle U_{t+1/2}, X_{t+1/2}\rangle \mid \mathscr{F}_t\right]\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\langle \mathbb{E}\left[U_{t+1/2} \mid \mathscr{F}_t\right], X_{t+1/2}\rangle\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\langle 0, X_{t+1/2}\rangle\right] && \text{(unbiasedness of } V_{t+1/2}\text{)} \\
&= 0.
\end{aligned}
$$

Finally recognizing $\sup_{x\in\mathscr{X}} \|x\| < D$ and combining the expressions for term (A) and (B),

$$
\mathbb{E}\left[\mathrm{Gap}_{\mathscr{X}}(\bar{X}_T)\right] \leq \frac{\mathbb{E}\left[\sup_{x\in\mathscr{X}} \left\{\left(\frac{D^2}{2} + 2M^2\right)\sqrt{T+1} + \frac{\tilde{D}}{2}\sigma\sqrt{T}\right\}\right]}{T},
$$

which concludes our derivation

$$
\mathbb{E}\left[\mathrm{Gap}_{\mathscr{X}}(\bar{X}_T)\right] = \mathscr{O}(1/\sqrt{T})
$$

■

Proposition 4.1.3. *Under Assumption 4.1.2 the iterates of (DA), (DE), (OptDA) enjoy the following rate:*

$$
\mathbb{E}\left[\mathrm{Gap}_{\mathscr{X}}\left(\bar{X}_T\right)\right] = \mathscr{O}(1/\sqrt{T}) \tag{4.41}
$$

Proof. Directly obtained by Theorem 4.1.1 by setting $V_t = 0$ for (DA), $V_t = g_{t+1/2}$ for (DE) and $V_t = g_{t-1/2}$ for (OptDA). ■

Theorem 4.1.2. *Let $X_t, X_{t+1/2}$ be generated by (GEG) with a constant step-size that satisfies*

$$
\min\left\{(2L)^{-1}, (4L^2\gamma)^{-1}\right\} - 2\gamma c > 0 \text{ with } L = 1/\beta. \tag{4.42}
$$

Then, for every compact neighbourhood $\mathscr{X} \subset \mathbb{R}_d$ of x^*, with $\bar{X}_T = \frac{1}{T}\sum_{t=1}^{T} X_{t+1/2}$, we have:

$$
\mathbb{E}\left[\mathrm{Gap}_{\mathscr{X}}\left(\bar{X}_T\right)\right] = \mathbb{E}\left[\sup_{X\in\mathscr{X}} \left\langle A(X), \bar{X}_T - X\right\rangle\right] = \mathscr{O}(1/T)
$$

Proof.

$$
\begin{aligned}
&\sum_{t=1}^{T} \langle V_{t+1/2}, X_{t+1/2} - x\rangle \\
&\quad = \sum_{t=1}^{T} \langle V_{t+1/2} - V_t, X_{t+1/2} - X_{t+1}\rangle + \langle V_t, X_{t+1/2} - X_{t+1}\rangle + \langle V_{t+1/2}, X_{t+1} - x\rangle \\
&\quad \leq \sum_{t=1}^{T} \|V_{t+1/2} - V_t\|\|X_{t+1/2} - X_{t+1}\| + \frac{1}{\gamma}\langle X_t - X_{t+1/2}, X_{t+1/2} - X_{t+1}\rangle + \frac{1}{\gamma}\left\langle \gamma Y_t - X_{t+1}, X_{t+1} - x\right\rangle
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{t=1}^{T} \frac{\gamma}{2} \|V_{t+1/2} - V_t\|^2 + \frac{1}{2\gamma} \|X_{t+1/2} - X_{t+1}\|^2 \\
&\quad + \frac{1}{2\gamma} \left(\|X_t - X\|^2 - \|X_{t+1} - X\|^2 - \|X_t - X_{t+1/2}\|^2 - \|X_{t+1/2} - X_{t+1}\|^2 \right) \\
&= \frac{\|X_1 - x\|^2}{2\gamma} + \frac{\gamma}{2} \sum_{t=1}^{T} \|V_{t+1/2} - V_t\|^2 - \frac{1}{2\gamma} \sum_{t=1}^{T} \|X_t - X_{t+1/2}\|^2 \\
&= \frac{\|x\|^2}{2\gamma} + \frac{\gamma}{2} \sum_{t=1}^{T} \|V_{t+1/2} - V_t\|^2 - \frac{1}{2\gamma} \sum_{t=1}^{T} \|X_t - X_{t+1/2}\|^2,
\end{aligned}
$$

where we set $X_1 = 0$. At this point the question is how to introduce the relative noise into the analysis such that we show that the stochastic/deterministic operator norms are summable. This would enable us to achieve the anticipated 1/T rate. In other words, we want to show that

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \|A(X_{t+1/2})\|^2 \right] &< +\infty \\
\mathbb{E}\left[\sum_{t=1}^{T} \|A(X_t)\|^2 \right] &< +\infty
\end{aligned}
$$

We take expectation with respect to all randomness and lower bound the left hand side with the norm of the operator using cocoercivity. Setting $x = x^*$, where x^* is a solution of (VI),

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \left\langle V_{t+1/2}, X_{t+1/2} - x^* \right\rangle \right] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[\left\langle V_{t+1/2}, X_{t+1/2} - x^* \right\rangle | \mathscr{F}_t \right] \right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \mathbb{E}\left[V_{t+1/2} | \mathscr{F}_t \right], X_{t+1/2} - x^* \right\rangle \right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \left\langle A(X_{t+1/2}) - A(x^*), X_{t+1/2} - x^* \right\rangle \right] \qquad \text{(Cocoercivity)} \\
&\geq \frac{1}{L} \mathbb{E}\left[\sum_{t=1}^{T} \|A(X_{t+1/2})\|^2 \right]
\end{aligned}
$$

Plugging this into the original expression yields

$$
\frac{1}{L} \mathbb{E}\left[\sum_{t=1}^{T} \|A(X_{t+1/2})\|^2 \right] \leq \frac{\|x^*\|^2}{2\gamma} + \frac{\gamma}{2} \sum_{t=1}^{T} \|V_{t+1/2} - V_t\|^2 - \frac{1}{2\gamma} \sum_{t=1}^{T} \|X_t - X_{t+1/2}\|^2
$$

With a similar approach,

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{L} \sum_{t=1}^{T} \|A(X_{t+1/2})\|^2 + \frac{1}{2\gamma} \sum_{t=1}^{T} \|X_t - X_{t+1/2}\|^2 \right] \\
&\qquad \geq \mathbb{E}\left[\frac{1}{L} \sum_{t=1}^{T} \|A(X_{t+1/2})\|^2 + \frac{1}{2L^2\gamma} \sum_{t=1}^{T} \|A(X_t) - A(X_{t+1/2})\|^2 \right]
\end{aligned}
$$

$$\geq \mathbb{E}\left[\min\left\{\frac{1}{2L}, \frac{1}{4L^2\gamma}\right\} \sum_{t=1}^{T} 2\|A(X_{t+1/2})\|^2 + 2\|A(X_t) - A(X_{t+1/2})\|^2\right]$$
$$\geq \mathbb{E}\left[\min\left\{\frac{1}{2L}, \frac{1}{4L^2\gamma}\right\} \sum_{t=1}^{T} \|A(X_t)\|^2\right]$$

Hence,

$$\mathbb{E}\left[\sum_{t=1}^{T} \min\left\{\frac{1}{2L}, \frac{1}{4L^2\gamma}\right\} \|A(X_t)\|^2 + \frac{1}{L}\|A(X_{t+1/2})\|^2\right] \leq \mathbb{E}\left[\frac{\|x^*\|^2}{\gamma} + \gamma \sum_{t=1}^{T} \|V_{t+1/2} - V_t\|^2\right]$$

We now use the relative variance in the expression on the right hand side. Relying on the towering property of expectation,

$$\mathbb{E}\left[\frac{\|x^*\|^2}{\gamma} + \gamma \sum_{t=1}^{T} \|V_{t+1/2} - V_t\|^2\right] \leq \mathbb{E}\left[\frac{\|x^*\|^2}{\gamma} + 2\gamma \sum_{t=1}^{T} \mathbb{E}\left[\|V_{t+1/2}\|^2 | \mathscr{F}_t\right] + \mathbb{E}\left[\|V_t\|^2 | \mathscr{F}_{t-1/2}\right]\right]$$
$$\leq \mathbb{E}\left[\frac{\|x^*\|^2}{\gamma} + 2\gamma c \sum_{t=1}^{T} \|A(X_{t+1/2})\|^2 + \|A(X_t)\|^2\right]$$

Combining last two expressions together yields

$$\mathbb{E}\left[\sum_{t=1}^{T} \min\left\{\frac{1}{2L}, \frac{1}{4L^2\gamma}\right\} \left(\|A(X_t)\|^2 + \|A(X_{t+1/2})\|^2\right)\right] \leq \mathbb{E}\left[\frac{\|x^*\|^2}{\gamma} + 2\gamma c \sum_{t=1}^{T} \|A(X_{t+1/2})\|^2 + \|A(X_t)\|^2\right]$$

Grouping the same terms on the same side of the inequality,

$$\mathbb{E}\left[\sum_{t=1}^{T} \left(\min\left\{\frac{1}{2L}, \frac{1}{4L^2\gamma}\right\} - 2\gamma c\right) \left(\|A(X_t)\|^2 + \|A(X_{t+1/2})\|^2\right)\right] \leq \mathbb{E}\left[\frac{\|x^*\|^2}{\gamma}\right]$$

As long as $\min\left\{\frac{1}{2L}, \frac{1}{4L^2\gamma}\right\} - 2\gamma c > 0$, we show that sum of operator norms with respect to both sequences are summable.

To obtain the gap, we will decompose $V_{t+1/2}$ into the full operator plus the noise,

$$\sum_{t=1}^{T} \langle V_{t+1/2}, X_{t+1/2} - x\rangle = \sum_{t=1}^{T} \langle A(X_{t+1/2}), X_{t+1/2} - x\rangle + \sum_{t=1}^{T} \langle U_{t+1/2}, X_{t+1/2} - x\rangle$$
$$\geq \sum_{t=1}^{T} \langle A(x), X_{t+1/2} - x\rangle + \sum_{t=1}^{T} \langle U_{t+1/2}, X_{t+1/2} - x\rangle \qquad \text{(Monotonicity)}$$
$$= T\left\langle A(x), \bar{X}_{t+1/2} - x\right\rangle + \sum_{t=1}^{T} \langle U_{t+1/2}, X_{t+1/2} - x\rangle$$

Rearranging and incorporating into the original bound,

$$\begin{aligned}&\langle A(x),\ \bar{X}_T - x\rangle \\ &\qquad \leq \frac{1}{T}\left(\frac{\|x\|^2}{2\gamma} + \sum_{t=1}^{T}\frac{\gamma}{2}\|V_{t+1/2}-V_t\|^2 - \frac{1}{2\gamma}\|X_t - X_{t+1/2}\|^2 + \langle U_{t+1/2}, x - X_{t+1/2}\rangle\right),\end{aligned}$$

We take supremum over x to retrieve the gap function and taking expectation,

$$\begin{aligned}&\mathbb{E}\left[\mathrm{Gap}_{\mathscr{X}}(\bar{X}_T)\right] \\ &\leq \mathbb{E}\left[\sup_{x\in\mathscr{X}}\left\{\frac{1}{T}\left(\frac{\|x\|^2}{2\gamma} + \sum_{t=1}^{T}\frac{\gamma}{2}\|V_{t+1/2}-V_t\|^2 - \frac{1}{2\gamma}\|X_t - X_{t+1/2}\|^2 + \langle U_{t+1/2}, x - X_{t+1/2}\rangle\right)\right\}\right] \\ &\leq \frac{1}{T}\left(\frac{D^2}{2\gamma} + \sum_{t=1}^{T}\mathbb{E}\left[\gamma\|V_{t+1/2}\|^2 + \gamma\|V_t\|^2\right] + \mathbb{E}\left[\sup_{x\in\mathscr{X}}\{\langle U_{t+1/2}, x\rangle\}\right] - \mathbb{E}\left[\langle U_{t+1/2}, X_{t+1/2}\rangle\right]\right) \\ &\leq \frac{1}{T}\left(\frac{D^2}{2\gamma} + \gamma c\underbrace{\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_{t+1/2})\|^2 + \|A(X_t)\|^2\right]}_{\text{(i)}} + \underbrace{\sum_{t=1}^{T}\mathbb{E}\left[\sup_{x\in\mathscr{X}}\{\langle U_{t+1/2}, x\rangle\}\right]}_{\text{(ii)}} - \underbrace{\sum_{t=1}^{T}\mathbb{E}\left[\langle U_{t+1/2}, X_{t+1/2}\rangle\right]}_{\text{(iii)}}\right),\end{aligned}$$

where we define that $\sup_{x\in\mathscr{X}}\|x\| \leq D$ and use relative variance in the last inequality.

For term (i), we have already proven that this particular summation is finite.

For term (ii),

$$\begin{aligned}\mathbb{E}\left[\sup_{x\in\mathscr{X}}\sum_{t=1}^{T}\langle U_{t+1/2}, x\rangle\right] &\leq \mathbb{E}\left[\max_{x\in\mathscr{X}}\left\langle\sum_{t=1}^{T}U_{t+1/2}, x\right\rangle\right] \\ &= \mathbb{E}\left[\left\langle\sum_{t=1}^{T}U_{t+1/2}, \tilde{x}\right\rangle\right] && \text{(for some } \tilde{x}\in\mathscr{X} \text{ which attains the maximum)} \\ &= \frac{D}{2}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\|U_{t+1/2}\|_*^2\right]} && \text{((by Lemma 4.2.1))} \\ &= \frac{D}{2}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\|V_{t+1/2} - A(X_{t+1/2})\|_*^2\right]} && \text{(unbiasedness of } V_{t+1/2}) \\ &= \frac{D}{2}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\|V_{t+1/2}\|_*^2\right]} && \text{(Towering property)} \\ &= \frac{D}{2}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[c\,\|A(X_{t+1/2})\|_*^2\right]} < +\infty && \text{(Relative variance)}\end{aligned}$$

Finally for term (iii),

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\langle U_{t+1/2}, X_{t+1/2}\rangle\right] &= \sum_{t=1}^{T}\mathbb{E}\left[\langle U_{t+1/2}, X_{t+1/2}\rangle\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\langle U_{t+1/2}, X_{t+1/2}\rangle \mid \mathscr{F}_t\right]\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\langle \mathbb{E}\left[U_{t+1/2} \mid \mathscr{F}_t\right], X_{t+1/2}\rangle\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\langle 0, X_{t+1/2}\rangle\right] && \text{(unbiasedness of } V_{t+1/2}\text{)} \\
&= 0.
\end{aligned}$$

Since we have shown that either the terms are finite or 0, it immediately implies that

$$\mathbb{E}\left[\mathrm{Gap}_{\mathscr{X}}(\bar{X}_T)\right] = \mathscr{O}(1/T)$$

■

Proposition 4.1.4. *Under Assumption 4.1.3 the iterates of (DA), (DE), (OptDA) enjoy the following rate:*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathscr{X}}\left(\bar{X}_T\right)\right] = \mathscr{O}(1/T) \tag{4.43}$$

Proof. Directly obtained by Theorem 4.1.2 by setting $V_t = 0$ for (DA), $V_t = g_{t+1/2}$ for (DE) and $V_t = g_{t-1/2}$ for (OptDA). ■

Having concluded the proof for the non-adaptive step-size schedules, we shall now provide the proof for (GEG) run with adaptive step-sizes for the various noise profiles. We will start presenting our analysis with the absolute random noise setting.

Theorem 4.1.3. *Assume that $X_t, X_{t+1/2}$ are the iterates of (GEG) run with the step-size (Adapt). Then, for every compact neighborhood $\mathscr{X} \subset \mathbb{R}^d$ of a solution x^* of (VI), we have:*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathscr{X}}(\overline{X}_T)\right] = \mathscr{O}(1/\sqrt{T}) \tag{4.44}$$

with $\overline{X}_T = (1/T)\sum_{t=1}^{T} X_{t+1/2}$

Proof. Invoking Proposition 4.1.2 and removing the negative term will give us the following inequality:

$$\sum_{t=1}^{T}\langle V_{t+1/2}, X_{t+1/2} - x\rangle \leq \frac{\|x\|^2}{2\gamma_{T+1}} + \frac{1}{2}\sum_{t=1}^{T}\gamma_t \|V_{t+1/2} - V_t\|_*^2 \tag{4.45}$$

Then, we replace the definition $V_{t+1/2} = A(X_{t+1/2}) + U_{t+1/2}$ and rewrite the above expression:

$$\sum_{t=1}^{T} \langle A(X_{t+1/2}), X_{t+1/2} - x \rangle \leq \frac{\|x\|^2}{2\gamma_{T+1}} + \frac{1}{2}\sum_{t=1}^{T} \gamma_t \|V_{t+1/2} - V_t\|_*^2 + \sum_{t=1}^{T} \langle U_{t+1/2}, x - X_{t+1/2} \rangle \tag{4.46}$$

Now, by applying the monotonicity of A we can bound the (LHS) from below and obtain:

$$\sum_{t=1}^{T} \langle A(x), X_{t+1/2} - x \rangle \leq \frac{\|x\|^2}{2\gamma_{T+1}} + \frac{1}{2}\sum_{t=1}^{T} \gamma_t \|V_{t+1/2} - V_t\|_*^2 + \sum_{t=1}^{T} \langle U_{t+1/2}, x - X_{t+1/2} \rangle \tag{4.47}$$

By taking suprema on both sides over a compact neighbourhood of a solution x^* and taking expectations:

$$\begin{aligned} T\mathbb{E}\left[\sup_{x\in\mathscr{X}} \left\langle A(x), \overline{X}_T - x \right\rangle\right] \leq D^2/2\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] + \frac{1}{2}\sum_{t=1}^{T} \mathbb{E}\left[\gamma_t \|V_{t+1/2} - V_t\|_*^2\right] \\ + \sum_{t=1}^{T} \mathbb{E}\left[\sup_{x\in\mathscr{X}} \langle U_{t+1/2}, x - X_{t+1/2} \rangle\right] \end{aligned} \tag{4.48}$$

where we plugged in the definition of $\overline{X_T}$, and used the fact that $\|x\| \leq D$ for any $x \in \mathscr{X}$. We rewrite the (LHS) with respect to the definition of the Gap function to obtain,

$$\begin{aligned} T\mathbb{E}\left[\mathrm{Gap}_{\mathscr{X}}(\overline{X}_T)\right] \leq D^2/2\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] + \frac{1}{2}\sum_{t=1}^{T} \mathbb{E}\left[\gamma_t \|V_{t+1/2} - V_t\|_*^2\right] \\ + \sum_{t=1}^{T} \mathbb{E}\left[\sup_{x\in\mathscr{X}} \langle U_{t+1/2}, x - X_{t+1/2} \rangle\right] \end{aligned} \tag{4.49}$$

Therefore, we are left to bound from above the (RHS). We will bound each term individually. Let us begin with the first term $D^2/2\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right]$.

$$D^2/2\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] = D^2/2\mathbb{E}\left[\sqrt{1 + \sum_{t=1}^{T} \|V_t - V_{t+1/2}\|_*^2}\right] \leq D^2/2\sqrt{1 + 4M^2T} \tag{4.50}$$

where we used the definition of the step-size and almost-sure boundedness of the sequence V_t to obtain the last inequality. Secondly, for the term $\frac{1}{2}\sum_{t=1}^{T} \mathbb{E}\left[\gamma_t \|V_{t+1/2} - V_t\|_*^2\right]$ we have:

$$\begin{aligned} \frac{1}{2}\sum_{t=1}^{T} \mathbb{E}\left[\gamma_t \|V_{t+1/2} - V_t\|_*^2\right] &= \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T} \left(\gamma_t - \gamma_{t+1}\right) \|V_{t+1/2} - V_t\|_*^2 + \sum_{t=1}^{T} \gamma_{t+1} \|V_{t+1/2} - V_t\|_*^2\right] \\ &\leq \frac{1}{2}\left[4M^2\mathbb{E}\left[\cdot \sum_{t=1}^{T} \left(\gamma_t - \gamma_{t+1}\right)\right] + \mathbb{E}\left[\sum_{t=1}^{T} \gamma_{t+1} \|V_{t+1/2} - V_t\|_*^2\right]\right] \\ &\leq \frac{1}{2}\left[4M^2\gamma_1 + \mathbb{E}\left[\sum_{t=1}^{T} \gamma_{t+1} \|V_{t+1/2} - V_t\|_*^2\right]\right] \end{aligned} \tag{4.51}$$

Notice that $\gamma_1 = 1$ due to the lag-one-behind step-size. To handle the summation on the (RHS), we will make use of the same approach as in the previous chapters; use adaptive structure of the step-size. Now by applying Lemma 2.1.2 we have:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t+1}\left\|V_{t+1/2}-V_t\right\|_*^2\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\frac{\left\|V_{t+1/2}-V_t\right\|_*^2}{\sqrt{1+\sum_{j=1}^{t}\left\|V_{j+1/2}-V_j\right\|_*^2}}\right]\\ &\leq 2\mathbb{E}\left[\sqrt{1+\sum_{t=1}^{T}\left\|V_{t+1/2}-V_t\right\|_*^2}\right]\\ &\leq 2\sqrt{1+4M^2T}.\end{aligned}$$

The last inequality is once again due to the boundedness of V_t. What remains is to bound the noise term, which is essentially the same as the proof in the non=adaptive setting as the expression itself is free of the step-size. Let us divide the term (B) into two and proceed.

$$\mathbb{E}\left[\sup_{x\in\mathscr{X}}\sum_{t=1}^{T}\langle U_{t+1/2},x-X_{t+1/2}\rangle\right] = \underbrace{\mathbb{E}\left[\sup_{x\in\mathscr{X}}\sum_{t=1}^{T}\langle U_{t+1/2},x\rangle\right]}_{\text{(B1)}} - \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\langle U_{t+1/2},X_{t+1}\rangle\right]}_{\text{(B2)}} \tag{4.52}$$

We will handle the term (B2) by the unbiasedness property of the noisy evaluations.

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\langle U_{t+1/2},X_{t+1/2}\rangle\right] &= \sum_{t=1}^{T}\mathbb{E}\left[\langle U_{t+1/2},X_{t+1/2}\rangle\right]\\ &= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\langle U_{t+1/2},X_{t+1/2}\rangle \mid \mathscr{F}_t\right]\right]\\ &= \sum_{t=1}^{T}\mathbb{E}\left[\langle \mathbb{E}\left[U_{t+1/2}\mid\mathscr{F}_t\right],X_{t+1/2}\rangle\right]\\ &= \sum_{t=1}^{T}\mathbb{E}\left[\langle 0,X_{t+1/2}\rangle\right] && \text{((unbiasedness of } V_{t+1/2}\text{))}\\ &= 0.\end{aligned}$$

For the term (B1) we will use Lemma 4.2.1 and we get:

$$\begin{aligned}\mathbb{E}\left[\sup_{x\in\mathscr{X}}\sum_{t=1}^{T}\langle U_{t+1/2},x\rangle\right] &\leq \mathbb{E}\left[\max_{x\in\mathscr{X}}\left\langle\sum_{t=1}^{T}U_{t+1/2},x\right\rangle\right]\\ &= \mathbb{E}\left[\left\langle\sum_{t=1}^{T}U_{t+1/2},\tilde{x}\right\rangle\right] && \text{(for some } \tilde{x}\in\mathscr{X} \text{ attaining the maximum)}\\ &\leq \frac{D}{2}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\left\|U_{t+1/2}\right\|_*^2\right]} && \text{(by Lemma 4.2.1)}\\ &\leq \frac{D\sigma}{2}\sqrt{T}\end{aligned}$$

Observe that all the terms on the (RHS) grow as $O(\sqrt{T})$. By dividing both sides by T and combining the individual bounds together gives us the following, which concludes the proof.

$$\mathbb{E}\left[\mathrm{Gap}_{\mathscr{X}}(\overline{X}_T)\right] = \frac{4M^2 + \left(\frac{D^2}{2}+2\right)\sqrt{1+4M^2}T + \frac{D\sigma}{2}\sqrt{T}}{T} = O\left(\frac{1}{\sqrt{T}}\right)$$

■

Next, we have the proof of Proposition 4.1.6. Similar to the non-adaptive counterpart, the proof immediately follows by Theorem 4.1.3 upto replacing $V_{t+1/2}$ and V_t with the respective definitions.

Proposition 4.1.6. *Under Assumption 4.1.2 the iterates of* (DA), (DE), (OptDA) *enjoy the following:*

$$\mathbb{E}\left[Gap_{\mathscr{X}}(\overline{X}_T)\right] = \mathscr{O}(1/\sqrt{T}) \tag{4.53}$$

Proof. Directly obtained by Theorem 4.1.3 by setting $V_t = 0$ for (DA), $V_t = g_{t+1/2}$ for (DE) and $V_t = g_{t-1/2}$ for (OptDA). ■

Now, we turn our attention towards the relative random noise in the adaptive step-size setting, which requires a bit more involved analysis than its non-adaptive counterpart. In particular, in order to show our main results for this context we will need the following proposition as a stepping stone. As a prelude, we point out that the following result will also play a crucial role for establishing the last iterate convergence results at the end of the appendix.

Proposition 4.2.1. *Assume that $X_t, X_{t+1/2}$ are the iterates of* (GEG) *run with* (Adapt)*. Then, we have:*

$$\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right] = \mathbb{E}\left[1 + \sum_{t=1}^{T} \|V_t - V_{t+1/2}\|_*^2\right] < +\infty \tag{4.54}$$

and

$$\mathbb{E}\left[\sum_{t=1}^{T} \|A(X_{t+1/2})\|_*^2\right] < +\infty \tag{4.55}$$

and

$$\mathbb{E}\left[\sum_{t=1}^{T} \|A(X_t)\|_*^2\right] < +\infty \tag{4.56}$$

and

$$\mathbb{E}\left[\sum_{t=1}^{T} \|X_{t+1/2} - X_t\|^2\right] < +\infty \tag{4.57}$$

Proof. Applying Proposition 4.1.2 and for $x = x^*$ with x^* being a solution of (VI), we have

$$\sum_{t=1}^{T} \langle V_{t+1/2}, X_{t+1/2} - x^* \rangle \leq \frac{\|x^*\|^2}{2\gamma_{T+1}} + \underbrace{\frac{1}{2}\sum_{t=1}^{T} \gamma_k \|V_{t+1/2} - V_t\|_*^2}_{\text{(A)}} - \frac{1}{2}\sum_{t=1}^{T} \frac{1}{\gamma_t} \|X_{t+1/2} - X_t\|^2 \tag{4.58}$$

First, we shall bound the term (A).

$$\begin{aligned}
\frac{1}{2}\sum_{t=1}^{T} \gamma_t \|V_{t+1/2} - V_t\|_*^2 &= \frac{1}{2}\left[\sum_{t=1}^{T} (\gamma_t - \gamma_{t+1}) \|V_{t+1/2} - V_t\|_*^2 + \sum_{t=1}^{T} \gamma_{t+1} \|V_{t+1/2} - V_t\|_*^2\right] \\
&\leq \frac{1}{2}\left[4G^2 \cdot \sum_{t=1}^{T} (\gamma_t - \gamma_{t+1}) + \sum_{t=1}^{T} \gamma_{t+1} \|V_{t+1/2} - V_t\|_*^2\right] \\
&\leq 2G^2 + \frac{1}{2}\sum_{t=1}^{T} \gamma_{t+1} \|V_{t+1/2} - V_t\|_*^2 \\
&\leq 2G^2 \sqrt{1 + \sum_{t=1}^{T} \|V_{t+1/2} - V_t\|_*^2} + \frac{1}{2}\sum_{t=1}^{T} \frac{\|V_{t+1/2} - V_t\|_*^2}{\sqrt{1 + \sum_{j=1}^{t} \|V_{j+1/2} - V_j\|_*^2}} \\
&\leq 2G^2 \sqrt{1 + \sum_{t=1}^{T} \|V_{t+1/2} - V_t\|_*^2} + 2 \cdot \frac{1}{2}\sqrt{1 + \sum_{t=1}^{T} \|V_{t+1/2} - V_t\|_*^2} \\
&= (2G^2 + 1)\sqrt{1 + \sum_{t=1}^{T} \|V_{t+1/2} - V_t\|_*^2} \\
&= (2G^2 + 1)\frac{1}{\gamma_{T+1}}
\end{aligned} \tag{4.59}$$

We plug the above expression into the original inequality, take expectation on both sides to obtain,

$$\begin{aligned}
\text{(B)} = \mathbb{E}\left[\sum_{t=1}^{T} \langle V_{t+1/2}, X_{t+1/2} - x^* \rangle\right] &\leq \frac{\|x^*\|^2}{2} \cdot \mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] + (2G^2 + 1) \cdot \mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] \\
&\quad - \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T} \frac{1}{\gamma_t} \|X_{t+1/2} - X_t\|^2\right] \\
&= \left[\frac{\|x^*\|^2}{2} + 2G^2 + 1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] - \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T} \frac{1}{\gamma_t} \|X_{t+1/2} - X_t\|^2\right]
\end{aligned} \tag{4.60}$$

Now, we will focus on the (LHS) of this expression. We make use of the (conditional) unbiased-

ness of the oracle information and subsequently apply cocoercivity.

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\left\langle V_{t+1/2}, X_{t+1/2}-x^*\right\rangle\right] &= \sum_{t=1}^{T}\mathbb{E}\left[\left\langle V_{t+1/2}, X_{t+1/2}-x^*\right\rangle\right] \\ &= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\left\langle V_{t+1/2}, X_{t+1/2}-x^*\right\rangle \mid \mathscr{F}_{t+1/2}\right]\right] \\ &= \sum_{t=1}^{T}\mathbb{E}\left[\left\langle \mathbb{E}\left[V_{t+1/2}\mid \xi_{t+1/2}\right], X_{t+1/2}-x^*\right\rangle\right] \\ &= \sum_{t=1}^{T}\mathbb{E}\left[\left\langle A(X_{t+1/2}), X_{t+1/2}-x^*\right\rangle\right]\end{aligned} \tag{4.61}$$

Recall that the continuous operator A is $1/L$-cocoercive by definition, which enables us to obtain,

$$\mathbb{E}\left[\sum_{t=1}^{T}\left\langle V_{t+1/2}, X_{t+1/2}-x^*\right\rangle\right] \geq \sum_{t=1}^{T}\frac{1}{L}\mathbb{E}\left[\|A(X_{t+1/2})\|_*^2\right] \tag{4.62}$$

By combining (4.60) and (4.62) we have

$$\frac{1}{L}\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_{t+1/2})\|_*^2\right] \leq \left[\frac{\|x^*\|^2}{2}+2G^2+1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right]-\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{\gamma_t}\|X_{t+1/2}-X_t\|^2\right] \tag{4.63}$$

Eq. (4.63) is the first cornerstone inequality, and we will show that the (RHS) is bounded by a constant, verifying the summability of $\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_{t+1/2})\|_*^2\right]$. Next, we will use the above inequality to obtain a bound on $\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_t)\|_*^2\right]$. Rearranging the terms an using the fact that $1/\gamma_t \geq 1$,

$$(\mathrm{C}) = \frac{1}{L}\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_{t+1/2})\|_*^2\right]+\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\left[\|X_{t+1/2}-X_t\|^2\right] \leq \left[\frac{\|x^*\|^2}{2}+2G^2+1\right]E\left[\frac{1}{\gamma_{T+1}}\right]. \tag{4.64}$$

Then, we bound the (LHS) from below using the L-Lipschitz continuity of the operator (implied by $1/L$-cocoercivity) and using quadratic inequality $(a+b)^2 \leq 2a^2+2b^2$.

$$\begin{aligned}&\frac{1}{L}\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_{t+1/2})\|_*^2\right]+\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\left[\|X_{t+1/2}-X_t\|^2\right] \\ &\qquad\geq \frac{1}{L}\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_{t+1/2})\|_*^2\right]+\frac{1}{2L^2}\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_{t+1/2})-A(X_t)\|_*^2\right] \\ &\qquad\geq \min\left\{\frac{1}{L},\frac{1}{2L^2}\right\}\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_{t+1/2})\|_*^2+\|A(X_{t+1/2})-A(X_t)\|_*^2\right] \\ &\qquad= \min\left\{\frac{1}{2L},\frac{1}{4L^2}\right\}\sum_{t=1}^{T}\mathbb{E}\left[2\|A(X_{t+1/2})\|_*^2+2\|A(X_{t+1/2})-A(X_t)\|_*^2\right] \\ &\qquad\geq \min\left\{\frac{1}{2L},\frac{1}{4L^2}\right\}\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_t)\|_*^2\right]\end{aligned} \tag{4.65}$$

To summarize, we get the following inequalities:

$$\begin{aligned} \frac{1}{L}\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_{t+1/2})\|_*^2\right] &\leq \left[\frac{\|x^*\|^2}{2}+2G^2+1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] \\ \min\left\{\frac{1}{2L},\frac{1}{4L^2}\right\}\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_t)\|_*^2\right] &\leq \left[\frac{\|x^*\|^2}{2}+2G^2+1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right]. \end{aligned} \tag{Ineq}$$

Then, we sum up the two inequalities and proceed with the next step of our analysis.

$$\frac{1}{L}\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_{t+1/2})\|_*^2\right]+\min\left\{\frac{1}{2L},\frac{1}{4L^2}\right\}\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_t)\|_*^2\right]\leq 2\left[\frac{\|x^*\|^2}{2}+2G^2+1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] \tag{4.66}$$

Now, our goal is to lower bound the expression above via the stochastic operators. This is where we will make use of the relative noise assumption. The strategy is to lower bound the (LHS) by $O(1/\gamma_T^2)$ and ultimately show that $\lim_{t\to\infty}\gamma_t=\gamma_\infty>0$. Then,

$$\begin{aligned} &\frac{1}{L}\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_{t+1/2})\|_*^2\right]+\min\left\{\frac{1}{2L},\frac{1}{4L^2}\right\}\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_t)\|_*^2\right] \\ &\qquad\geq \min\left\{\frac{1}{2L},\frac{1}{4L^2}\right\}\left[\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_{t+1/2})\|_*^2\right]+\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_t)\|_*^2\right]\right] \\ &\qquad\geq \min\left\{\frac{1}{2L},\frac{1}{4L^2}\right\}\left[\sum_{t=1}^{T}\frac{1}{c}\mathbb{E}\left[\|V_{t+1/2}\|_*^2\right]+\sum_{t=1}^{T}\frac{1}{c}\mathbb{E}\left[\|V_t\|_*^2\right]\right] \qquad \text{(Assumption 4.1.3)} \\ &\qquad\geq \frac{1}{c\max\{4L,8L^2\}}\left[\sum_{t=1}^{T}\mathbb{E}\left[2\|V_{t+1/2}\|_*^2+2\|V_t\|_*^2\right]\right] \\ &\qquad\geq \frac{1}{c\max\{4L,8L^2\}}\sum_{t=1}^{T}\mathbb{E}\left[\|V_{t+1/2}-V_t\|_*^2\right] \end{aligned}$$

By combining the last inequality with the previous expression Eq. (4.66) gives us,

$$\mathbb{E}\left[\sum_{t=1}^{T}\|V_{t+1/2}-V_t\|_*^2\right]\leq 8c\max\{L,2L^2\}\cdot\left[\frac{\|x^*\|^2}{2}+2G^2+1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] \tag{4.67}$$

Using the definition of the adaptive step-size,

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^{T}\|V_{t+1/2}-V_t\|_*^2\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\|V_{t+1/2}-V_t\|_*^2+1-1\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T}\|V_{t+1/2}-V_t\|_*^2+1\right]-1 \\ &= \mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right]-1 \end{aligned} \tag{4.68}$$

After combining everything we have so far, we will obtain (through Jensen's inequality) a

quadratic inequality with respect to the variable $x = \mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right]$.

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right] &\leq 8c\max\{L, 2L^2\}\left[\frac{\|x^*\|^2}{2} + 2G^2 + 1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] + 1 \\
&\leq 8c\max\{L, 2L^2\}\left[\frac{\|x^*\|^2}{2} + 2G^2 + 1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] + \mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] \\
&= \left[8c\max\{L, 2L^2\}\left[\frac{\|x^*\|^2}{2} + 2G^2 + 1\right] + 1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] \\
&= \left[8c\max\{L, 2L^2\}\left[\frac{\|x^*\|^2}{2} + 2G^2 + 1\right] + 1\right]\mathbb{E}\left[\sqrt{1 + \sum_{t=1}^{T}\|V_{t+1/2} - V_t\|_*^2}\right] \\
&= \left[8c\max\{L, 2L^2\}\left[\frac{\|x^*\|^2}{2} + 2G^2 + 1\right] + 1\right]\sqrt{\mathbb{E}\left[1 + \sum_{t=1}^{T}\|V_{t+1/2} - V_t\|_*^2\right]} \\
&= \left[8c\max\{L, 2L^2\}\left[\frac{\|x^*\|^2}{2} + 2G^2 + 1\right] + 1\right]\sqrt{\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right]}
\end{aligned} \tag{4.69}$$

By simplifying the above expression gives us,

$$\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right] \leq \left(8c\max\{L, 2L^2\}\left[\frac{\|x^*\|^2}{2} + 2G^2 + 1\right] + 1\right)^2 \tag{4.70}$$

which proves the first inequality in the proposition. The second and third claim is derived directly by combining the first claim with (Ineq). To verify the final statement of the proposition, we go back to (4.63). We rearrange the terms to get,

$$\begin{aligned}
\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{\gamma_t}\|X_{t+1/2} - X_t\|^2\right] &\leq \left[\frac{\|x^*\|^2}{2} + 2G^2 + 1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] \\
\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\left[\|X_{t+1/2} - X_t\|^2\right] &\leq \left[\frac{\|x^*\|^2}{2} + 2G^2 + 1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] \qquad (\tfrac{1}{\gamma_t} \geq 1)
\end{aligned}$$

The summability of the last expression follows by Eq. (4.70). ■

Finally, we present the proof of the main result under relative random noise

Theorem 4.1.4. *Assume that $X_t, X_{t+1/2}$ are the iterates of* (GEG) *run with the step-size* (Adapt). *Then, for every compact neighborhood $\mathcal{X} \subset \mathbb{R}^d$ of a solution x^* of* (VI), *we have:*

$$\mathbb{E}\left[Gap_{\mathcal{X}}(\overline{X}_T)\right] = \mathcal{O}(1/T) \tag{4.71}$$

with $\overline{X}_T = 1/T\sum_{t=1}^{T} X_{t+1/2}$

Proof. Once again, we begin with the template inequality in Proposition 4.1.2.

$$\sum_{t=1}^{T} \langle V_{t+1/2}, X_{t+1/2} - x \rangle \leq \frac{\|x\|^2}{2\gamma_{T+1}} + \sum_{t=1}^{T} \gamma_t \|V_{t+1/2} - V_t\|_*^2 \tag{4.72}$$

which holds for all $x \in \mathbb{R}^d$. By the definition of $V_{t+1/2} = A(X_{t+1/2}) + U_{t+1/2}$ the above becomes

$$\begin{aligned} \sum_{t=1}^{T} \langle A(X_{t+1/2}), X_{t+1/2} - x \rangle &\leq \sum_{t=1}^{T} \langle U_{t+1/2}, x - X_{t+1/2} \rangle + \frac{\|x\|^2}{2\gamma_{T+1}} + \sum_{t=1}^{1} \gamma_1 \|V_{t+1} - V_t\|_*^2 \\ \sum_{t=1}^{T} \langle A(x), X_{t+1/2} - x \rangle &\leq \sum_{t=1}^{T} \langle U_{t+1/2}, x - X_{t+1/2} \rangle + \frac{\|x\|^2}{2\gamma_{T+1}} + \sum_{t=1}^{1} \gamma_1 \|V_{t+1} - V_t\|_*^2 \end{aligned} \tag{4.73}$$

Taking the summation inside the inner produc in the (LHS) and dividing both sides by T,

$$\left\langle A(x), \bar{X}_T - x \right\rangle \leq \frac{1}{T} \left[\sum_{t=1}^{T} \langle U_{t+1/2}, x - X_{t+1/2} \rangle + \frac{\|x\|^2}{2\gamma_{T+1}} + \sum_{t=1}^{T} \gamma_t \|V_{t+1/2} - V_t\|_*^2 \right] \tag{4.74}$$

Now to obtain the Gap from the above expression, we take suprema on both sides over $\mathscr{X}$ (defining $D^2 = \sup_{x \in \mathscr{X}} \|x - x_1\|^2$), and then take the expectation to have,

$$\mathbb{E}\left[\mathrm{Gap}_{\mathscr{X}}\left(\bar{X}_T\right)\right] \leq \frac{1}{T} \Bigg[\underbrace{\mathbb{E}\left[\sup_{x \in \mathscr{X}} \sum_{t=1}^{T} \langle U_{t+1/2}, x - X_{t+1/2} \rangle \right]}_{\text{(B)}} + \underbrace{\frac{D^2}{2} \mathbb{E}\left[\frac{1}{\gamma_{T+1}} \right]}_{\text{(C)}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \gamma_t \|V_{t+1/2} - V_1\|_*^2 \right]}_{\text{(D)}} \Bigg] \tag{4.75}$$

Now, we bound each term individually. We leave the proof of term (B) to the end as it is relatively more involved than the rest.

Bounding (C).

$$\begin{aligned} \frac{D^2}{2} \mathbb{E}\left[\frac{1}{\gamma_{T+1}} \right] &= \frac{D^2}{2} \mathbb{E}\left[\sqrt{1 + \sum_{t=1}^{T} \|V_{t+1/2} - V_t\|_*^2} \right] \\ &\leq \frac{D^2}{2} \sqrt{\mathbb{E}\left[1 + \sum_{t=1}^{T} \|V_{t+1/2} - V_t\|_*^2 \right.} \\ &= \frac{D^2}{2} \sqrt{\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2} \right]} \\ &< +\infty.. \qquad \text{(by Proposition 4.2.1)} \end{aligned}$$

Bounding (D).

$$\mathbb{E}\left[\sum_{t=1}^{T} \gamma_t \|V_{t+1/2} - V_t\|_*^2 \right] = \mathbb{E}\left[\sum_{t=1}^{T} (\gamma_t - \gamma_{t+1}) \|V_{t+1/2} - V_t\|_*^2 \right] + \mathbb{E}\left[\sum_{t=1}^{T} \gamma_{t+1} \|V_{t+1/2} - V_t\|_*^2 \right]$$

$$\leq 2G^2 + 2\mathbb{E}\left[\sqrt{1+\sum_{t=1}^{T}\|V_{t+1/2}-V_t\|_*^2}\right]$$
$$\leq 2G^2 + 2\sqrt{\mathbb{E}\left[1+\sum_{t=1}^{T}\|V_{t+1/2}-V_t\|_*^2\right]}$$
$$\leq 2G^2 + 2\sqrt{\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right]}$$
$$< +\infty. \qquad \text{(by Proposition 4.2.1)}$$

Bounding (B).

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}}\sum_{t=1}^{T}\langle U_{t+1/2}, x - X_{t+1/2}\rangle\right] = \underbrace{\mathbb{E}\left[\sup_{x\in\mathcal{X}}\sum_{t=1}^{T}\langle U_{t+1/2}, x\rangle\right]}_{\text{(B1)}} - \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\langle U_{t+1/2}, X_{t+1/2}\rangle\right]}_{\text{(B2)}} \tag{4.76}$$

By working in the same spirit as Theorem 4.1.3, the term (B2) evaluates as

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle U_{t+1/2}, X_{t+1/2}\rangle\right] = 0. \tag{4.77}$$

We will handle term (B1) in a different way than before:

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}}\sum_{t=1}^{T}\langle U_{t+1/2}, x\rangle\right] \leq \frac{D}{2}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\|U_{t+1/2}\|_*^2\right]} \tag{4.78}$$

Due to the definition of $V_{t+1/2} = A(X_{t+1/2}) + U_{t+1/2}$ we have $U_{t+1/2} = A(X_{t+1/2}) - V_{t+1/2}$. So,

$$\mathbb{E}\left[\sup_{x\in e}\sum_{t=1}^{T}\langle U_{t+1/2}, x\rangle\right] \leq \frac{D}{2}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_{t+1}) - V_{t+1/2}\|_*^2\right]}$$
$$\leq \frac{D}{2}\sqrt{2\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_{t+1/2})\|_*^2\right] + 2\sum_{t=1}^{T}\mathbb{E}\left[\|V_{t+1/2}\|_*^2\right]}$$
$$< +\infty. \qquad \text{(by Proposition 4.2.1)}$$

Since all the terms (B), (C) and (D) are bounded, we immediately have the fast rate of order $O(1/T)$ with respect to the restricted Gap. ■

Similar to Proposition 4.1.6, Theorem 4.1.4 allows us to obtain the following result.

Proposition 4.1.7. *Under Assumption 4.1.3 the iterates of* (DA), (DE), (OptDA) *enjoy the following:*

$$\mathbb{E}\left[Gap_{\mathcal{X}}(\overline{X}_T)\right] = \mathcal{O}(1/T) \tag{4.79}$$

Proof. Directly obtained by Theorem 4.1.4 by setting $V_t = 0$ for (DA), $V_t = g_{t+1/2}$ for (DE) and $V_t = g_{t-1/2}$ for (OptDA). ■

We conclude by showing that the iterates $X_{t+1/2}, X_t$ of (GEG) run with the adaptive step-size policy (Adapt) converge towards some (VI) solution x^* almost surely. In doing so, we will need the following proposition:

Proposition 4.2.2. *Let there be a non-empty closed set F and let a sequence $(x_t)_t \in \mathbb{R}^d$. Suppose that for all $z \in F$ there exists $(\beta_t)_t$ sequence of random variables satisfying the following almost surely:*

$$\mathbb{E}\left[\|x_{t+1} - z\|^2 \mid \mathscr{F}_t\right] \leq \|x_t - z\|^2 + \beta_t \tag{4.80}$$

with $\sum_{t=1}^{\infty} \beta_t < +\infty$ almost surely. Then, the following hold:

1. *$\|x_t - z\|^2$ converges almost surely.*
2. *If the set of almost sure limit points, i.e.*

$$\hat{\mathscr{X}} = \{\hat{x} \in \mathbb{R}^d : \text{there exists a subsequence } x_{t_n} \to \hat{x} \text{ almost surely}\} \tag{4.81}$$

is non-empty and $\hat{\mathscr{X}} \subset F$, then x_t converges almost surely to some random variable $\hat{x} \in F$.

Proof. This proposition is a special case of Combettes and Pesquet [CP15a, Proposition 2.3], and we refer the reader to the respective manuscript for the proof of this result. ■

Moreover, we will also use the following classical convergence theorem.

Proposition 4.2.3 (Monotone Convergence Theorem)**.** *Let (Ω, Σ, μ) be a measure space and $\mathscr{X} \in \Sigma$. Consider a pointwise non-decreasing sequence of $(\Sigma, \mathscr{B}_{\mathbb{R}_{>0}})$-measurable, non-negative functions: $f_t : \mathscr{X} \to [0, +\infty]$. Set the pointwise limit of the sequence (f_n) as,*

$$\lim_t f_t(x) = f(x) \tag{4.82}$$

Then, f is $(\Sigma, \mathscr{B}_{\mathbb{R}_{>0}})$ -measurable and

$$\lim_{t \to +\infty} \int_{\mathscr{X}} f_t \, d\mu = \int_{\mathscr{X}} \lim_{t \to +\infty} f_t \, d\mu = \int_{\mathscr{X}} f \, d\mu. \tag{4.83}$$

Essentially, this is necessary to interchange limit with the expectation (integration) and argue that the expression at hand satisfies summability of β_t sequence in Proposition 4.2.2. Having all these at hand, we are now in the position to illustrate the last iterate convergence result for the iterates of (DA)/(DE)/(OptDA). For the ease of presentation we shall provide the convenience of the general choice for the $V_{t+1/2}$.

Proposition 4.2.4. *The iterates of* (DA)/(DE)/(OptDA) *converge towards a* (VI) *solution x^*.*

Proof. We are left to show that the iterates $X_{t+1/2}$ satisfies the requirements of Proposition 4.2.2. In particular, invoking Proposition 4.1.2 we have:

$$\frac{1}{2\gamma_{t+1}}\left\|X_{t+1}-x^*\right\|^2 \leq \frac{1}{2\gamma_t}\left\|X_t-x^*\right\|^2 - \langle V_{t+1/2}, X_{t+1/2}-x^*\rangle + \frac{D^2}{2}\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right)+\gamma_t\|V_t-V_{t+1/2}\|_*^2 \tag{4.84}$$

with $D^2 = \sup_{x^*\in\mathcal{X}}\|x^*\|^2$. Now, by multiplying both sides with $2\gamma_t$ and using the fact that γ_t is non-decreasing and $\gamma_t \leq 1$ we get:

$$\left\|X_{t+1}-x^*\right\|^2 \leq \|X_t-x^*\| - \gamma_t\langle V_{t+1/2}, X_{t+1/2}-x^*\rangle + \frac{D^2}{2}\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right)+\|V_t-V_{t+1/2}\|_*^2 \tag{4.85}$$

Now, by taking conditional expectations we obtain:

$$\begin{aligned}\mathbb{E}\left[\left\|X_{t+1}-x^*\right\|^2 \mid \mathcal{F}_t\right] \leq \left\|X_t-x^*\right\|^2 &- \gamma_t\mathbb{E}\left[\langle V_{t+1/2}, X_{t+1/2}-x^*\rangle \mid \mathcal{F}_t\right] \\ &+ \frac{D^2}{2}\mathbb{E}\left[\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right)\mid \mathcal{F}_t\right] + \gamma_t\mathbb{E}\left[\|V_t-V_{t+1}\|_*^2 \mid \mathcal{F}_t\right]\end{aligned} \tag{4.86}$$

since γ_t is $\mathcal{F}_{t-1/2}$-measurable and $\mathcal{F}_{t-1/2}\subset\mathcal{F}_t$, γ_t is $\mathcal{F}_t$-measurable. Therefore, $\mathbb{E}\left[\gamma_t \mid \mathcal{F}_t\right] = \gamma_t$ almost surely. Also note that $X_{t+\frac{1}{2}}$ is $\mathcal{F}_t$-measurable. Then, we have

$$\begin{aligned}\gamma_t\mathbb{E}\left[\langle V_{t+1/2}, X_{t+1/2}-x^*\rangle \mid \mathcal{F}_t\right] &= \gamma_t\langle \mathbb{E}[V_{t+1/2}|\mathcal{F}_t], X_{t+1/2}-x^*\rangle \\ &= \langle A(X_{t+\frac{1}{2}}), X_{t+1/2}-x^*\rangle \\ &\leq 0.\end{aligned}$$

where we used the fact that $V_{t+1/2}$ is an unbiased estimator of $A(X_{t+1/2})$, conditioned on $\mathcal{F}_t$. Last line follows from that x^* is a solution of (VI). Combining all we obtain

$$\mathbb{E}\left[\left\|X_{t+1}-x^*\right\|^2 \mid \mathcal{F}_t\right] \leq \left\|X_t-x^*\right\|^2 + \frac{D^2}{2}\mathbb{E}\left[\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right)\mid \mathcal{F}_t\right] + \mathbb{E}\left[\|V_t-V_{t+1}\|_*^2 \mid \mathcal{F}_t\right] \tag{4.87}$$

Now let us define

$$\beta_t = \frac{D^2}{2}\mathbb{E}\left[\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right)\mid \mathcal{F}_t\right] + \mathbb{E}\left[\|V_t-V_{t+1}\|_*^2 \mid \mathcal{F}_t\right] \tag{4.88}$$

The first step is to show that β_t sequence satisfies the summability statement in Proposition 4.2.2. We will show that $\mathbb{E}\left[\sum_{t=1}^T \beta_t\right] < +\infty$ and use the Monotone Convergence Theorem to argue that this implies $\sum_{t=1}^T \beta_t < +\infty$. Indeed, we have that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T \beta_t\right] &= \frac{D^2}{2}\mathbb{E}\left[\sum_{t=1}^T\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right)\right] + \mathbb{E}\left[\sum_{t=1}^T\|V_t-V_{t+1}\|_*^2\right] \\ &\leq \frac{D^2}{2}\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] + \mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right] \qquad \text{(Definition of } \gamma_t\text{)}\end{aligned}$$

$$\leq \left(\frac{D^2}{2}+1\right)\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right] \qquad \left(\tfrac{1}{\gamma_t}\leq\tfrac{1}{\gamma_t^2}\right)$$
$$< +\infty,$$

and the last line is due to Proposition 4.2.1. On the other hand, $\sum_{t=1}^{T}\beta_t$ is a non-decreasing (random) sequence; therefore is converges almost surely to some random value $\beta_\infty = \sum_{t=1}^{+\infty}\beta_t \in (0,\infty]$. Assume that $\beta_\infty = +\infty$. Then, by applying Proposition 4.2.3 we get:

$$+\infty = \mathbb{E}\left[\sum_{t=1}^{+\infty}\beta_t\right] = \lim_T \mathbb{E}\left[\sum_{t=1}^{T}\beta_t\right] < +\infty \tag{4.89}$$

which is a contradiction. Therefore $\sum_{t=1}^{+\infty}\beta_t < +\infty$ almost surely. Therefore, we are left to show that every almost sure limit point of X_t is a (VI) solution. Let $\hat{x}\in\mathbb{R}^d$ be a limit point of X_t. Then, there exists a subsequence X_{t_n} which converges almost surely towards $\hat{x}$. Then, by invoking Proposition 4.2.1 (Eq. (4.56)), we have that:

$$\mathbb{E}\left[\sum_{t=1}^{T}\|A(X_t)\|_*^2\right] < +\infty \tag{4.90}$$

Therefore by the same reasoning as above, Proposition 4.2.3 ensures that:

$$\sum_{t=1}^{T}\|A(X_t)\|_*^2 < +\infty \text{ almost surely} \tag{4.91}$$

which yields a fortiori that $\|A(X_t)\|_*^2 \to 0$ almost surely. On the other hand, we have that: $\|A(X_{t_n})\|_* \to \|A(\hat{x})\|_*$. Thus, by limit uniqueness we get that that $\|A(\hat{x})\|_* = 0$, so $\hat{x}$ is a (VI) solution, hence the result follows by Proposition 4.2.2. Finally, in order to show that $X_{t+1/2}$ converges also towards a solution, we shall invoke Proposition 4.2.1 (Eq. (4.57)) that:

$$\mathbb{E}\left[\sum_{t=1}^{T}\|X_t - X_{t+1/2}\|^2\right] < +\infty \tag{4.92}$$

Hence, by the same reasoning we obtain that:

$$\|X_t - X_{t+1/2}\|^2 \to 0 \text{ almost surely} \tag{4.93}$$

and so our proof is completed. ■

5 Conclusion and future extensions

5.1 Summary of the thesis

In this dissertation, we have studied adaptive and universal algorithms for three main optimization problems; constrained convex minimization, smooth non-convex minimization and monotone variational inequalities. Particularly, we have developed parameter-free, simple algorithms and suitable, novel analysis techniques that are oblivious to problem-dependent parameters, e.g., Lipschitz constant, noise levels in the oracle feedback. When possible, we demonstrate universal properties of our proposed frameworks in the sense that the algorithms could achieve (optimal) convergence under different problem settings simultaneously without knowing the problem at hand a priori. Let us summarize the contributions with respect to the chapters.

In Chapter 2, we have designed adaptive first and second-order methods for compactly constrained convex minimization setting. We have answered an open problem in the field by developing the first adaptive and universal algorithm UNIXGRAD, which achieves optimal convergence rates for smooth/non-smooth problems under deternistic/stochastic oracles, simultaneously. The proposed algorithm accomplishes these results without knowing neither the smoothness of the problems nor the nature of the oracle ahead of time. This result is possible due to an alternative accelerated scheme based on the extra-gradient template and we complete it with a modular proof which consists of an offline regret analysis and (accelerated) regret-to-rate conversion.

We further extend these results for second-order methods and propose the first noise-adaptive, accelerated second-order method EXTRA-NEWTON. Under the bounded variance condition for stochastic gradient and the Hessian oracles, we generalize the noise adaptation capabilities of universal first-order methods to the second-order realm, which is a direction that has not been studied to the best of our knowledge. We generalize the conversion scheme and the regret analysis for our first-order scheme and identified a delicate connection between the order of smoothness, adaptive step-size design and averaging parameters that enables faster sublinear rates beyond $O(1/T^2)$. Moreover, the techniques we use to handle approximation

error due to not knowing L for setting the step-size are easily extensible for other problem formulations, including monotone VIs.

In Chapter 3, we turned our attention to smooth, non-convex minimization. We initially focused on the high-probability convergence analysis for the AdaGrad, and proved optimal convergence up to logarithmic factors with best known dependence on the probability margin. Under the standard bounded variance setting, we show that the scalar step-size version of the original AdaGrad achieves $O\left(\frac{L\log(T)+\sigma\sqrt{log(1/\delta)}}{\sqrt{T}}\right)$ with probability at least $O(1-\delta)$. Under the finer model of sub-Gaussian noise, we obtain the celebrated noise-adaptive rate of $O\left(\frac{L^2+\sigma^2\log(1/\delta)}{T}+\frac{L\sigma}{\sqrt{T}}\right)$. We propose an alternative proof strategy by showing the sub-optimality gap grows no faster than $\log(T)$ with high probability, which verifies *pseudo*-boundedness of the iterates. We combine the best of existing results [WWB19; LO20] by keeping the original construction of the adaptive step-size while achieving best known dependence on probability margin.

The second focus of this chapter is the variance reduction algorithms under two main problem structures. For the more general case of *expectation over smooth losses*, i.e., $\min_x \mathbb{E}_{\xi\sim\mathscr{D}}\left[f(x,\xi)\right]$, we designed the first parameter-free, noise-adaptive variance reduction algorithm that obtains the convergence rate of $O\left(\frac{1}{\sqrt{T}}+\frac{\sigma^{1/3}}{T^{1/3}}\right)$, with optimal dependence on time horizon T, but sub-optimal dependence on L, G and initial sub-optimality gap. Our construction relies on the recursive momentum estimator of Cutkosky and Orabona [CO19] but we identify a concrete and recursive relationship between the step-size and momentum parameters. For the more specific case of finite-sum minimization, we design the first parameter-free variance reduction algorithm with sample complexity $\tilde{O}\left(n+\frac{L^4\sqrt{n}}{\epsilon^2}\right)$, which is optimal in dependence on time horizon T (up to logarithmic factors) and number of components n, but sub-optimal in L dependence. Our main finding that made these results possible is the correct quantification of the cumulative variance across the whole execution. While standard approaches guarantee that the variance decreases at a particular rate, we show that cumulative behavior of the variance under adaptive step-size grows roughly in the same rate as in the non-adaptive case. We deem the additional log factors and sub-optimal parameter dependence is indeed a direct consequence of the error due to non-monotonic behavior of the variance.

In the last part, Chapter 4, we study the intricate relationship between cocoercivity and different noise models. We consider a generalized algorithm that recover 3 staple algorithms in the study of VI problem, and propose a simple adaptive step-size scheme, which is capable of simultaneously adapting to cocoercivity constant and the type of the noise without knowing the precise setting a priori. Unlike the previous sections, we have recognized a one-to-one connection between the cumulative growth of the norm of the operator evaluated at the decision sequence, and the rate of convergence. Under standard bounded variance setting, the cumulative quantity grows as $O(\sqrt{T})$, leading to the $O(1/\sqrt{T})$, while the vanishing, relative noise model implies summability of the operator norms, yielding the fast rate of $O(1/T)$. While doing so, we provide a generalized analysis technique for different algorithms, and propose an alternative look into the convergence analysis by identifying a connection between the growth

of the adaptive step-size and the rate of convergence.

5.2 Future directions

As a follow up to the presented work in the main body of this dissertation, we have several directions for some of which we have preliminary results or existing attempts.

Per-coordinate step-sizes. The proposed algorithms that we studied in this dissertation all have a scalar step-size, and in practice it is popular and very common to use *per-coordinate* step-sizes. For most of our algorithms, this appears as an immediate and relatively easy extension. For instance, our preliminary attempts suggest that we could prove a convergence rate of the same order for UNIXGRAD and EXTRA-NEWTON when the step-size is generalized to the per-coordinate version.

Indeed, some of the techniques we used in our analysis is not compatible with vector-valued step-sizes. For instance, in the analysis of high-probability AdaGrad and STORM+ in Sections 3.1 and 3.2, we divide both sides by the scalar step-size. In the presence of vector-valued step-sizes, we cannot apply the same technique and need to come up with an appropriate modification.

Adaptation to approximation and sampling errors: bias-variance trade-offs. A fundamental concept we have presented in this manuscript, both for the function minimization and variational inequality setting is adaptation to noise levels and types of oracle errors. To complement our existing results, we would like to investigate the types of *biased* estimates and the underlying procedure that generates them. Concrete examples towards this direction come from offline reinforcement learning [Nac+19], online learning [Str+10] and distributed stochastic optimization [Bez+22]. As a preliminary result, we have pinpointed a type of bias called *relative bias*, which is defined as,

$$\|\mathbb{E}\left[\,\tilde{\nabla} f(x)\,\right] - \nabla f(x)\| \leq \beta \|\nabla f(x)\|,$$

where $\tilde{\nabla} f(x)$ is a biased and possibly stochastic gradient estimate for the true value $\nabla f(x)$. The motivation comes from the use of importance sampling in reinforcement learning [Met+18], which yields biased estimates of the above form. Our results show that the AdaGrad algorithm implicitly adapts to the bias levels, but requires (almost surely) bounded stochastic gradients. Note that this is barely a first step as there are many open problems to be answered for displaying a concrete understanding of the bias adaptation.

First, we need to quantify respective lower bounds for convex/non-convex minimization under the relative bias assumption. It is important to verify the limits with respect to the iteration count T, as well as the correct dependence on the bias parameter β. Second, we will focus on the accelerated algorithms in the smooth, convex setting and investigate uni-

versal convergence properties with respect to smoothness, noise levels and bias parameter, simultaneously. The fact that the effect of bias is not eliminated in expectation, our analysis needs an additional mechanism to handle the systematic error due to the bias. Besides, this is only one type of bias that appears in application, and we aim to study different bias types and quantifications to expand the literature towards this direction, taking another step towards closing the boundary between theoretical works and practical approaches.

Simultaneous adaptation to different problem formulations. When run with the same decreasing step-size strategy, SGD could achieve the same, and order-optimal, convergence rate of $O(1/\sqrt{T})$ for minimizing smooth convex and non-convex problems. The same applies to gradient descent algorithm with a particular range of fixed step-size. However, in the presence of accelerated algorithms, this no longer holds. Ghadimi and Lan [GL16] shows that an accelerated gradient algorithm that achieves the optimal $O(1/T^2)$ rate for smooth, convex problems, achieves the optimal rate of $O(1/T)$ for smooth, *non-convex* minimization as long as the algorithm parameters are changed according to the non-convex setting. With the "accelerated" set of parameters, to the best of our knowledge, showing the optimal non-asymptotic rates is not trivial for non-convex problems.

Motivated by this idea, we have investigated the *min-min to min-max* adaptation for variational inequalities. Let us take the UNIXGRAD algorithm we studied as an example, which achieves the noise adaptive rate of $O\left(\frac{LD^2}{T^2}+\frac{\sigma}{\sqrt{T}}\right)$. The rate of convergence is deemed by the averaging scheme and the appropriate scaling of the step-size; computing the averages as $\bar{X}_{t+\frac{1}{2}}=\frac{\alpha_t}{A_t}X_{t+\frac{1}{2}}+\frac{A_{t-1}}{A_t}\bar{X}_{t-\frac{1}{2}}$ and the *effective* step-size as $\alpha_t\gamma_t$ where $\alpha_t=t$ and $A_t=\sum_{s=1}^{t}\alpha_s$ yield the desired rate.

If we run this algorithm exactly for solving monotone VIs, then the algorithm diverges. This is a phenomenon that we observed in practice for simple bilinear games, too. In fact, choosing α_t as any increasing function of t such as $\alpha_t=\log(t)$ or $\alpha_t=t^{1/p}$ for $p>1$, yields the same divergent behavior. An important qualitative assesment within this context is that the vector field induced by a monotone operator might have cycles, whereas this is not the case for the dynamics defined by accelerated algorithms. Our intuitive understanding is that weighted averaging might be causing unexpected behavior around cycles and lead to divergence away from the solution set. Given that the gradient descent-ascent diverges for the simple bilinear setting indicates how important it is to understand the structural difference between the problem formulations.

Our proposition is to make the averaging scheme itself adaptive. Roughly speaking, α_t should remain constant if the vector field behaves in accordance with a monotone operator, and should grow linearly when the structure is locally convex. A proposition we have studied is to form a recursive relationship between the averaging parameter α_t and the step-size γ_t such that $\alpha_t=\gamma_{t-1}^2$. However, we couldn't make the conversion scheme work for this construction. Moreover, when α_t is a random variable, analysis in the stochastic setting runs into measurability problems.

Although this is an unorthodox approach in optimization research, we see it as the first step to understand the transition between convex and non-convex minimization in terms of universality.

Another direction we want to investigate is the continuous adaptation to degrees of convexity. The literature on convex minimization validates that accelerated gradient method [Nes83b] cannot adapt to strong convexity, while gradient descent does and achieves a linear rate of convergence with sub-optimal constants. A known approach is to use restarts, and setting the restart period adaptively (such that it doesn't depend on problem parameters like L and μ) is one of the goals we have for future publications. Similarly, we want to investigate the effect of averaging and adaptive step-size selection in achieving optimal linear rates for strongly convex case, and $O(1/T^2)$ rate for the convex case, simultaneously.

Technically speaking, the biggest challenge for the latter direction is the *additive error accumulation* for the analysis of adaptive methods. It is not enough to show the summability of error due to not knowing the Lipschitz constant and the strong convexity parameter to achieve the linear rate. We also need to prove that such an error decreases exponentially.

Higher-order methods. There has been a recent interest in the study of higher-order methods for minimization and variational inequalities. While most of such work focus on deterministic and non-adaptive algorithm design, there is some decent progress towards parameter-free algorithm development [DMN22]. We have two main goals towards this direction.

First, we want to extend our results for second-order convex minimization to higher-order smooth settings. Under the assumption that the higher-order sub-problem in the extrapolation step (as in Algorithm 3) is efficiently computable, we believe our analysis technique can accommodate higher-order of smoothness to achieve faster rates. However, we must note that these faster rates will not match the lower bounds exactly. Hence, we will investigate the optimal scheme of Carmon et al. [Car+22] and adaptive scheme of [DMN22] in conjunction with our EXTRA-NEWTON to develop order-optimal strategies for higher-order methods in the presence of *stochastic* oracles.

For variational inequalities, acceleration mechanism that enables faster rates for higher-order formulations is fundamentally different than that of in convex minimization. For the smooth, convex setting, computing gradient at the weighted averages sits at the heart of the analysis. On the contrary, such an approach will not work for monotone variational inequalities. If we take the extra-gradient scheme as a representative example, acceleration of higher-order VIs is related to the regularization degree in the extrapolation step and the growth of $\sum_{t=1}^{T} \frac{1}{\|X_{t+\frac{1}{2}} - X_t\|^p}$, where p is the degree of smoothness of the operator [BL22; Adi+22]. Under stochastic oracles, we will need to deal with additional measurability problems due to the aforementioned terms. Therefore, developing an adaptive and universal scheme for higher-order variational inequalities is an interesting but challenging problem.

A second direction we want to investigate is the faster convergence of second-order methods under third-order smoothness. Inspired by Nesterov [Nes21], we would like to develop a second-order VI method which approximates the second-order Jacobian (the so-called third-order term) with lower-order terms. Then, we need to develop a fast sub-solver in the spirit of Bauschke, Bolte, and Teboulle [BBT17] and Lu, Freund, and Nesterov [LFN18] for solving the auxiliary VI problem for the extrapolation step. We have made some progress along this direction up to verifying a fast sub-solver.

Bibliography

[Adi+22] D. Adil et al. *Optimal Methods for Higher-Order Smooth Monotone Variational Inequalities*. 2022. DOI: 10.48550/ARXIV.2205.06167.

[AH18] N. Agarwal and E. Hazan. "Lower Bounds for Higher-Order Convex Optimization". In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 774–792.

[AMC21] A. Alacaoglu, Y. Malitsky, and V. Cevher. "Convergence of adaptive algorithms for constrained weakly convex optimization". In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*. Ed. by M. Ranzato et al. 2021, pp. 14214–14225.

[Ala+20] A. Alacaoglu et al. "A new regret analysis for Adam-type algorithms". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 202–210.

[All17a] Z. Allen-Zhu. "Katyusha: The First Direct Acceleration of Stochastic Gradient Methods". In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. STOC 2017. Association for Computing Machinery, 2017, pp. 1200–1205.

[All17b] Z. Allen-Zhu. "Natasha: Faster Non-Convex Stochastic Optimization via Strongly Non-Convex Parameter". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 89–97.

[All18] Z. Allen-Zhu. "Katyusha X: Practical Momentum Method for Stochastic Sum-of-Nonconvex Optimization". In: *Proceedings of the 35th International Conference on Machine Learning*. ICML '18. Full version available at http://arxiv.org/abs/1802.03866. 2018.

[AH16a] Z. Allen-Zhu and E. Hazan. "Variance Reduction for Faster Non-Convex Optimization". In: *Proceedings of the 33rd International Conference on International*

Conference on Machine Learning - Volume 48. ICML'16. New York, NY, USA: JMLR.org, 2016.

[AH16b] Z. Allen-Zhu and E. Hazan. "Variance reduction for faster non-convex optimization". In: *International conference on machine learning*. PMLR. 2016, pp. 699–707.

[AO16] Z. Allen-Zhu and L. Orecchia. *Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent*. 2016.

[AY16] Z. Allen-Zhu and Y. Yuan. "Improved SVRG for Non-Strongly-Convex or Sum-of-Non-Convex Objectives". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by M. F. Balcan and K. Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1080–1089.

[ABM19] K. Antonakopoulos, E. V. Belmega, and P. Mertikopoulos. "An adaptive mirror-prox algorithm for variational inequalities with singular operators". In: *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 2019.

[ABM21] K. Antonakopoulos, V. Belmega, and P. Mertikopoulos. "Adaptive Extra-Gradient Methods for Min-Max Optimization and Games". In: *International Conference on Learning Representations*. 2021.

[AKC22] K. Antonakopoulos, A. Kavis, and V. Cevher. "Extra-Newton: A First Approach to Noise-Adaptive Accelerated Second-Order Methods". In: *Advances in Neural Information Processing Systems*. Ed. by A. H. Oh et al. 2022.

[Ant+21] K. Antonakopoulos et al. "Sifting through the noise: Universal first-order methods for stochastic variational inequalities". In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. 2021.

[Ant+22] K. Antonakopoulos et al. "UnderGrad: A universal black-box optimization method with almost dimension-free convergence rate guarantees". In: *ICML '22: Proceedings of the 39th International Conference on Machine Learning*. 2022.

[ASS19] Y. Arjevani, O. Shamir, and R. Shiff. "Oracle complexity of second-order methods for smooth convex optimization". In: *Mathematical Programming* (2019), pp. 1–34.

[Arj+19] Y. Arjevani et al. "Lower bounds for non-convex stochastic optimization". In: *Mathematical Programming* 199 (2019), pp. 165–214.

[Arm66] L. Armijo. "Minimization of functions having Lipschitz continuous first partial derivatives." In: *Pacific Journal of Mathematics* 16 (1966), pp. 1–3.

[BL19] F. R. Bach and K. Y. Levy. "A Universal Algorithm for Variational Inequalities Adaptive to Smoothness and Noise". In: *Annual Conference Computational Learning Theory*. 2019.

[BH77] J.-B. Baillon and G. Haddad. “Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones”. In: *Israel Journal of Mathematics* 26 (1977), pp. 137–150.

[Bar19] J. T. Barron. “A General and Adaptive Robust Loss Function”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4326–4334. DOI: 10.1109/CVPR.2019.00446.

[BB88] J. Barzilai and J. M. Borwein. “Two-Point Step Size Gradient Methods”. In: *IMA Journal of Numerical Analysis* 8.1 (Jan. 1988), pp. 141–148. DOI: 10.1093/imanum/8.1.141.

[BBT17] H. H. Bauschke, J. Bolte, and M. Teboulle. “A Descent Lemma Beyond Lipschitz Gradient Continuity: First-Order Methods Revisited and Applications”. In: *Math. Oper. Res.* 42.2 (May 2017), pp. 330–348. DOI: 10.1287/moor.2016.0817.

[BC17] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. 2nd ed. New York, NY, USA: Springer, 2017.

[Bec17] A. Beck. *First-Order Methods in Optimization*. Philadelphia, PA, USA: SIAM-Society for Industrial and Applied Mathematics, 2017.

[BT09] A. Beck and M. Teboulle. “A fast iterative shrinkage-thresholding algorithm for linear inverse problems”. In: *SIAM journal on imaging sciences* 2.1 (2009), pp. 183–202.

[BGM20] S. Bellavia, G. Gurioli, and B. Morini. “Adaptive cubic regularization methods with dynamic inexact Hessian information and applications to finite-sum minimization”. In: *IMA Journal of Numerical Analysis* 41.1 (Apr. 2020), pp. 764–799. DOI: 10.1093/imanum/drz076.

[Ben16] A. A. Bennett. “Newton’s Method in General Analysis”. In: *Proceedings of the National Academy of Sciences* 2.10 (1916), pp. 592–598. DOI: 10.1073/pnas.2.10.592.

[Bez+22] A. Beznosikov et al. *On Biased Compression for Distributed Learning*. 2022.

[BLM18] M. Bravo, D. S. Leslie, and P. Mertikopoulos. “Bandit learning in concave N-person games”. In: *NeurIPS ’18: Proceedings of the 32nd International Conference of Neural Information Processing Systems*. 2018.

[Bri89] J. Bridle. “Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters”. In: *Advances in neural information processing systems* 2 (1989).

[Bri90] J. S. Bridle. “Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition”. In: *Neurocomputing*. Springer, 1990, pp. 227–236.

[BL22] B. Bullins and K. A. Lai. “Higher-Order Methods for Convex-Concave Min-Max Optimization and Monotone Variational Inequalities”. In: *SIAM Journal on Optimization* 32.3 (2022), pp. 2208–2229. DOI: 10.1137/21M1395764.

[CLS15] E. J. Candes, X. Li, and M. Soltanolkotabi. “Phase retrieval via Wirtinger flow: Theory and algorithms”. In: *IEEE Transactions on Information Theory* 61.4 (2015), pp. 1985–2007.

[Car+22] Y. Carmon et al. “Optimal and Adaptive Monteiro-Svaiter Acceleration”. In: *Advances in Neural Information Processing Systems*. Ed. by A. H. Oh et al. 2022.

[CGT11a] C. Cartis, N. I. Gould, and P. L. Toint. “Adaptive Cubic Regularisation Methods for Unconstrained Optimization. Part I: Motivation, Convergence and Numerical Results”. In: *Math. Program.* 127.2 (Apr. 2011), pp. 245–295.

[CGT11b] C. Cartis, N. I. M. Gould, and P. L. Toint. “Adaptive Cubic Regularisation Methods for Unconstrained Optimization. Part II: Worst-Case Function- and Derivative-Evaluation Complexity”. In: *Math. Program.* 130.2 (Dec. 2011), pp. 295–319. DOI: 10.1007/s10107-009-0337-y.

[CL06] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006, pp. I–XII, 1–394.

[CV19] V. Cevher and B. C. Vũ. “On the linear convergence of the stochastic gradient method with constant step-size”. In: *Optimization Letters* 13.5 (July 2019), pp. 1177–1187. DOI: 10.1007/s11590-018-1331-1.

[CL11] C.-C. Chang and C.-J. Lin. “LIBSVM: A library for support vector machines”. In: *ACM Transactions on Intelligent Systems and Technology* 2 (3 2011), 27:1–27:27.

[Che+21] J. Chen et al. “Closing the Generalization Gap of Adaptive Gradient Methods in Training Deep Neural Networks”. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. IJCAI’20. Yokohama, Yokohama, Japan, 2021.

[Che+22] X. Chen et al. “Accelerating Adaptive Cubic Regularization of Newton’s Method via Random Sampling”. In: *Journal of Machine Learning Research* 23.90 (2022), pp. 1–38.

[Che+19] X. Chen et al. “On the Convergence of A Class of Adam-Type Algorithms for Non-Convex Optimization”. In: *International Conference on Learning Representations*. 2019.

[CUH16] D. Clevert, T. Unterthiner, and S. Hochreiter. “Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)”. In: *4th International Conference on Learning Representations*. Ed. by Y. Bengio and Y. LeCun. 2016.

[CDO18] M. Cohen, J. Diakonikolas, and L. Orecchia. “On Acceleration with Noise-Corrupted Gradients”. In: *International Conference on Machine Learning*. 2018.

[CP15a] P. L. Combettes and J.-C. Pesquet. “Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping”. In: *SIAM Journal on Optimization* 25.2 (2015), pp. 1221–1248.

[CP15b] P. L. Combettes and J.-C. Pesquet. “Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping”. In: *SIAM Journal on Optimization* 25.2 (2015), pp. 1221–1248.

[CGT00] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust Region Methods*. Society for Industrial and Applied Mathematics, 2000. DOI: 10.1137/1.9780898719857.

[Cut19] A. Cutkosky. “Anytime Online-to-Batch Conversions, Optimism, and Acceleration”. In: *the International Conference on Machine Learning (ICML)* (June 2019).

[CM21] A. Cutkosky and H. Mehta. “High-probability Bounds for Non-Convex Stochastic Optimization with Heavy Tails”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. 2021.

[CO19] A. Cutkosky and F. Orabona. “Momentum-Based Variance Reduction in Non-Convex SGD”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019.

[Das+18] C. Daskalakis et al. “Training GANs with optimism”. In: *ICLR ’18: Proceedings of the 2018 International Conference on Learning Representations*. 2018.

[Def21] A. Defazio. *Momentum via Primal Averaging: Theoretical Insights and Learning Rate Schedules for Non-Convex Optimization*. 2021.

[DBL14] A. Defazio, F. Bach, and S. Lacoste-Julien. “SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives”. In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani et al. Vol. 27. Curran Associates, Inc., 2014.

[Def+20] A. Defossez et al. *On the Convergence of Adam and Adagrad*. Mar. 2020.

[Den+09a] J. Deng et al. “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[Den+09b] J. Deng et al. “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[DCL18] Q. Deng, Y. Cheng, and G. Lan. “Optimal Adaptive and Accelerated Stochastic Gradient Descent”. In: *arXiv preprint arXiv:1810.00553* (2018).

[Dev+19] J. Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran, and T. Solorio. Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/n19-1423.

[DO18] J. Diakonikolas and L. Orecchia. “Accelerated Extra-Gradient Descent: A Novel Accelerated First-Order Method”. In: *ITCS*. 2018.

[DMN22] N. Doikov, K. Mishchenko, and Y. Nesterov. *Super-Universal Regularized Newton Method*. 2022. DOI: 10.48550/ARXIV.2208.05888.

[Dub+22] B. Dubois-Taine et al. "SVRG Meets AdaGrad: Painless Variance Reduction". In: *Mach. Learn.* 111.12 (Dec. 2022), pp. 4359–4409. DOI: 10.1007/s10994-022-06265-x.

[DHS11] J. Duchi, E. Hazan, and Y. Singer. "Adaptive subgradient methods for online learning and stochastic optimization". In: *Journal of Machine Learning Research* 12.Jul (2011), pp. 2121–2159.

[DO21] J.-P. Dussault and D. Orban. "Scalable Adaptive Cubic Regularization Methods". In: (2021). DOI: 10.13140/RG.2.2.18142.15680.

[EN20] A. Ene and H. L. Nguyen. "Adaptive and Universal Algorithms for Variational Inequalities with Optimal Convergence". In: *AAAI Conference on Artificial Intelligence*. 2020.

[ENV21] A. Ene, H. L. Nguyen, and A. Vladu. "Adaptive Gradient Methods for Constrained Convex Optimization and Variational Inequalities". In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021, pp. 7314–7321.

[FP03] F. Facchinei and J.-S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Series in Operations Research. Springer, 2003.

[Fan+18] C. Fang et al. "Near-optimal non-convex optimization via stochastic path integrated differential estimator". In: *Advances in Neural Information Processing Systems* 31 (2018), p. 689.

[FFF21] H. Fang, Z. Fan, and M. P. Friedlander. "FAST CONVERGENCE OF STOCHASTIC SUBGRADIENT METHOD UNDER INTERPOLATION". In: *ICLR '21: Proceedings of the 2021 International Conference on Learning Representations*. 2021.

[Faw+22] M. Faw et al. "The Power of Adaptivity in SGD: Self-Tuning Step Sizes with Unbounded Gradients and Affine Variance". In: *Conference on Learning Theory, 2-5 July 2022, London, UK*. Ed. by P. Loh and M. Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, 2022, pp. 313–355.

[GP19] B. Gao and L. Pavel. "Discounted Mirror Descent Dynamics in Concave Games". In: *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE. 2019, pp. 5942–5947.

[Gas+19] A. Gasnikov et al. "Optimal Tensor Methods in Smooth Convex and Uniformly ConvexOptimization". In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by A. Beygelzimer and D. Hsu. Vol. 99. Proceedings of Machine Learning Research. PMLR, 25–28 Jun 2019, pp. 1374–1391.

[GLM16] R. Ge, J. D. Lee, and T. Ma. "Matrix completion has no spurious local minimum". In: *Advances in Neural Information Processing Systems* (2016), pp. 2981–2989.

[GL13] S. Ghadimi and G. Lan. "Stochastic first-and zeroth-order methods for nonconvex stochastic programming". In: *SIAM Journal on Optimization* 23.4 (2013), pp. 2341–2368.

[GL16] S. Ghadimi and G. Lan. "Accelerated Gradient Methods for Nonconvex Nonlinear and Stochastic Programming". In: *Math. Program.* 156.1–2 (Mar. 2016), pp. 59–99. DOI: 10.1007/s10107-015-0871-8.

[GRC20] F. L. Gómez, P. Rolland, and V. Cevher. "Lipschitz constant estimation of Neural Networks via sparse polynomial optimization". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[Goo+14] I. J. Goodfellow et al. "Generative adversarial nets". In: *NIPS '14: Proceedings of the 28th International Conference on Neural Information Processing Systems*. 2014.

[HKS16] A. A. Hameed, B. Karlik, and M. S. Salman. "Back-propagation algorithm with variable adaptive momentum". In: *Knowledge-Based Systems* 114 (2016), pp. 79–87.

[Har+19] N. J. A. Harvey et al. "Tight analyses for non-smooth stochastic gradient descent". In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by A. Beygelzimer and D. Hsu. Vol. 99. Proceedings of Machine Learning Research. PMLR, 25–28 Jun 2019, pp. 1579–1613.

[Hau22] D. Hausler. *Optimal and Adaptive Monteiro-Svaiter Acceleration*. https://github.com/danielle-hausler/ms-optimal. 2022.

[He+15] K. He et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[He+16] K. He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[He+19] T. He et al. "Bag of tricks for image classification with convolutional neural networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 558–567.

[Hoy04] P. O. Hoyer. "Non-negative matrix factorization with sparseness constraints." In: *Journal of machine learning research* 5.9 (2004).

[Hsi+19] Y.-G. Hsieh et al. "On the convergence of single-call stochastic extra-gradient methods". In: *Neural Information Processing Systems*. 2019.

[Hsi+20] Y.-G. Hsieh et al. “Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling”. In: *NeurIPS ’20: Proceedings of the 34th International Conference on Neural Information Processing Systems*. 2020.

[Hsi+22a] Y.-G. Hsieh et al. “Multi-Agent Online Optimization with Delays: Asynchronicity, Adaptivity, and Optimism”. In: *Journal of Machine Learning Research* 23.78 (2022), pp. 1–49.

[Hsi+22b] Y.-G. Hsieh et al. “No-regret learning in games with noisy feedback: Faster rates and adaptivity via learning rate separation”. In: *Advances in Neural Information Processing Systems*. Ed. by A. H. Oh et al. 2022.

[HAM21] Y. Hsieh, K. Antonakopoulos, and P. Mertikopoulos. “Adaptive Learning in Continuous Games: Optimal Regret Bounds and Convergence to Nash Equilibrium”. In: *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*. Ed. by M. Belkin and S. Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, 2021, pp. 2388–2422.

[HPK09] C. Hu, W. Pan, and J. T. Kwok. “Accelerated gradient methods for stochastic optimization and online learning”. In: *Advances in Neural Information Processing Systems*. 2009, pp. 781–789.

[HWD19] H. Huang, C. Wang, and B. Dong. “Nostalgic Adam: Weighting More of the Past Gradients When Designing the Adaptive Learning Rate”. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, July 2019, pp. 2556–2562. DOI: 10.24963/ijcai.2019/355.

[Ius+19] A. N. Iusem et al. “Variance-based extragradient methods with line search for stochastic variational inequalities”. In: *SIAM Journal on Optimization* 29.1 (2019), pp. 175–206.

[JT16] F. Jarre and P. L. Toint. “Simple Examples for the Failure of Newton’s Method with Line Search for Strictly Convex Minimization”. In: *Math. Program.* 158.1–2 (July 2016), pp. 23–34. DOI: 10.1007/s10107-015-0913-2.

[JLZ17] B. Jiang, T. Lin, and S. Zhang. *A Unified Scheme to Accelerate Adaptive Cubic Regularization and Gradient Methods for Convex Optimization*. 2017. DOI: 10.48550/ARXIV.1710.04788.

[JLZ20] B. Jiang, T. Lin, and S. Zhang. “A Unified Adaptive Tensor Approximation Scheme to Accelerate Composite Convex Optimization”. In: *SIAM Journal on Optimization* 30.4 (2020), pp. 2897–2926. DOI: 10.1137/19M1286025.

[JM22] R. Jiang and A. Mokhtari. *Generalized Optimistic Methods for Convex-Concave Saddle Point Problems*. 2022. DOI: 10.48550/ARXIV.2202.09674.

[JZ13] R. Johnson and T. Zhang. “Accelerating Stochastic Gradient Descent Using Predictive Variance Reduction”. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*. NIPS’13. Curran Associates Inc., 2013.

[Jou+20] P. Joulani et al. “A simpler approach to accelerated optimization: iterative averaging meets optimism”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 4984–4993.

[JNT11] A. Juditsky, A. Nemirovski, and C. Tauvel. “Solving variational inequalities with stochastic mirror-prox algorithm”. In: *Stochastic Systems* 1.1 (2011). DOI: 10.1214/10-SSY011.

[KT08] S. M. Kakade and A. Tewari. “On the Generalization Ability of Online Strongly Convex Programming Algorithms”. In: *Proceedings of the 21st International Conference on Neural Information Processing Systems*. NIPS’08. Vancouver, British Columbia, Canada: Curran Associates Inc., 2008, pp. 801–808.

[Kan48] L. V. Kantorovich. “Functional analysis and applied mathematics”. In: *Uspekhi Mat. Nauk* 3 (6(28) 1948), pp. 89–185.

[KSJ18] S. P. Karimireddy, S. U. Stich, and M. Jaggi. *Global linear convergence of Newton’s method without strong-convexity or Lipschitz gradients*. 2018. DOI: 10.48550/ARXIV.1806.00413.

[KLC22] A. Kavis, K. Y. Levy, and V. Cevher. “High Probability Bounds for a Class of Nonconvex Algorithms with AdaGrad Stepsize”. In: *International Conference on Learning Representations*. 2022.

[Kav+19] A. Kavis et al. “UniXGrad: A Universal, Adaptive Algorithm with Optimal Guarantees for Constrained Optimization”. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 6260–6269.

[Kav+22] A. Kavis et al. “Adaptive Stochastic Variance Reduction for Non-convex Finite-Sum Minimization”. In: *Advances in Neural Information Processing Systems*. Ed. by A. H. Oh et al. 2022.

[KB15] D. P. Kingma and J. Ba. “Adam: A Method for Stochastic Optimization”. In: *3rd International Conference on Learning Representations*. Ed. by Y. Bengio and Y. LeCun. 2015.

[Kor76] G. M. Korpelevich. “The extragradient method for finding saddle points and other problems”. In: *Matecon* 12 (1976), pp. 747–756.

[KMR19] D. Kovalev, K. Mishchenko, and P. Richtárik. *Stochastic Newton and Cubic Newton Methods with Simple Local Linear-Quadratic Rates*. 2019. DOI: 10.48550/ARXIV.1912.01597.

[Lan12] G. Lan. “An optimal method for stochastic composite optimization”. In: *Mathematical Programming* 133.1-2 (2012), pp. 365–397.

[Lan20] G. Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer, 2020.

[LLZ19] G. Lan, Z. Li, and Y. Zhou. *A unified variance-reduced accelerated gradient method for convex optimization*. 2019.

[Lec+98] Y. Lecun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[Lei+17] L. Lei et al. “Non-convex finite-sum optimization via SCSG methods”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pp. 2345–2355.

[Lev44] K. Levenberg. “A METHOD FOR THE SOLUTION OF CERTAIN NON – LINEAR PROBLEMS IN LEAST SQUARES”. In: *Quarterly of Applied Mathematics* 2 (1944), pp. 164–168.

[LP17] D. A. Levin and Y. Peres. *Markov chains and mixing times*. Vol. 107. American Mathematical Soc., 2017.

[Lev17] K. Levy. “Online to Offline Conversions, Universality and Adaptive Minibatch Sizes”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 1612–1621.

[LYC18] K. Y. Levy, A. Yurtsever, and V. Cevher. “Online Adaptive Methods, Universality and Acceleration”. In: *Neural and Information Processing Systems (NeurIPS)*. Dec. 2018.

[LKC21] K. Y. Levy, A. Kavis, and V. Cevher. “STORM+: Fully Adaptive SGD with Recursive Momentum for Nonconvex Optimization”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. 2021.

[LWG20] B. Li, L. Wang, and G. B. Giannakis. “Almost Tune-Free Variance Reduction”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 5969–5978.

[LO19] X. Li and F. Orabona. “On the Convergence of Stochastic Gradient Descent with Adaptive Stepsizes”. In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by K. Chaudhuri and M. Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 16–18 Apr 2019, pp. 983–992.

[LO20] X. Li and F. Orabona. *A High Probability Analysis of Adaptive SGD with Momentum*. 2020.

[LHR21] Z. Li, S. Hanzely, and P. Richtárik. *ZeroSARAH: Efficient Nonconvex Finite-Sum Optimization with Zero Full Gradient Computation*. 2021. DOI: 10.48550/ARXIV.2103.01447.

[LL18] Z. Li and J. Li. “A Simple Proximal Stochastic Gradient Method for Nonsmooth Nonconvex Optimization”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS’18. Montréal, Canada: Curran Associates Inc., 2018, pp. 5569–5579.

[Li+21] Z. Li et al. “PAGE: A Simple and Optimal Probabilistic Gradient Estimator for Nonconvex Optimization”. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 6286–6295.

[LS19] T. Liang and J. Stokes. “Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks”. In: *AISTATS ’19: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*. 2019.

[LJ22] T. Lin and M. I. Jordan. *Perseus: A Simple High-Order Regularization Method for Variational Inequalities*. 2022. DOI: 10.48550/ARXIV.2205.03202.

[Lin+20] T. Lin et al. “Finite-time last-iterate convergence for multi-agent learning in games”. In: *ICML ’20: Proceedings of the 37th International Conference on Machine Learning*. 2020.

[LGY20] Y. Liu, Y. Gao, and W. Yin. “An Improved Analysis of Stochastic Gradient Descent with Momentum”. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS’20. Vancouver, BC, Canada: Curran Associates Inc., 2020.

[Liu+22a] Z. Liu et al. “Adaptive Accelerated (Extra-)Gradient Methods with Variance Reduction”. In: *CoRR* abs/2201.12302 (2022).

[Liu+22b] Z. Liu et al. *META-STORM: Generalized Fully-Adaptive Variance Reduced SGD for Unbounded Functions*. 2022.

[Liu+23] Z. Liu et al. “On the Convergence of AdaGrad(Norm) on $\mathbb{R}^d$: Beyond Convexity, Non-Asymptotic Rate and Acceleration”. In: *The Eleventh International Conference on Learning Representations*. 2023.

[LH17] I. Loshchilov and F. Hutter. “SGDR: Stochastic Gradient Descent with Warm Restarts”. In: *International Conference on Learning Representations*. 2017.

[LFN18] H. Lu, R. M. Freund, and Y. Nesterov. “Relatively Smooth Convex Optimization by First-Order Methods, and Applications”. In: *SIAM Journal on Optimization* 28.1 (2018), pp. 333–354. DOI: 10.1137/16M1099546.

[LXL19] L. Luo, Y. Xiong, and Y. Liu. “Adaptive Gradient Methods with Dynamic Bound of Learning Rate”. In: *International Conference on Learning Representations*. 2019.

[MDB21] L. Madden, E. Dall’Anese, and S. Becker. *High probability convergence bounds for stochastic gradient descent assuming the Polyak-Lojasiewicz inequality*. 2021.

[MZJ13] M. Mahdavi, L. Zhang, and R. Jin. “Mixed optimization for smooth functions”. In: *Advances in neural information processing systems* 26 (2013), pp. 674–682.

[Mar63] D. W. Marquardt. "An Algorithm for Least-Squares Estimation of Nonlinear Parameters". In: *Journal of the Society for Industrial and Applied Mathematics* 11.2 (1963), pp. 431–441. DOI: 10.1137/0111030.

[MBR19] U. Marteau-Ferey, F. R. Bach, and A. Rudi. "Globally Convergent Newton Methods for Ill-conditioned Generalized Self-concordant Losses". In: *NeurIPS*. 2019.

[MS10] H. B. McMahan and M. Streeter. "Adaptive Bound Optimization for Online Convex Optimization". In: *COLT 2010* (2010), p. 244.

[MS17] P. Mertikopoulos and M. Staudigl. "Convergence to Nash equilibrium in continuous games with noisy first-order feedback". In: *CDC '17: Proceedings of the 56th IEEE Annual Conference on Decision and Control*. 2017.

[MS18] P. Mertikopoulos and M. Staudigl. "Stochastic mirror descent dynamics and their convergence in monotone variational inequalities". In: *Journal of Optimization Theory and Applications* 179.3 (Dec. 2018), pp. 838–867.

[MZ19] P. Mertikopoulos and Z. Zhou. "Learning in games with continuous action sets and unknown payoff functions". In: *Mathematical Programming* 173.1-2 (Jan. 2019), pp. 465–507.

[Mer+19] P. Mertikopoulos et al. "Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile". In: *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations*. 2019.

[Met+18] A. M. Metelli et al. "Policy optimization via importance sampling". In: *Advances in Neural Information Processing Systems* 31 (2018).

[Mis21] K. Mishchenko. *Regularized Newton Method with Global $O(1/k^2)$ Convergence*. 2021. DOI: 10.48550/ARXIV.2112.02089.

[MS13] R. D. C. Monteiro and B. F. Svaiter. "An Accelerated Hybrid Proximal Extragradient Method for Convex Optimization and Its Implications to Second-Order Methods". In: *SIAM Journal on Optimization* 23.2 (2013), pp. 1092–1125. DOI: 10.1137/110833786.

[Nac+19] O. Nachum et al. "DualDICE: Behavior-Agnostic Estimation of Discounted Stationary Distribution Corrections". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[Nem04] A. Nemirovski. "Prox-Method with Rate of Convergence O(1/t) for Variational Inequalities with Lipschitz Continuous Monotone Operators and Smooth Convex-Concave Saddle Point Problems". In: *SIAM Journal on Optimization* 15.1 (2004), pp. 229–251.

[Nem+09] A. S. Nemirovski et al. "Robust Stochastic Approximation Approach to Stochastic Programming". In: *SIAM Journal on Optimization* 19.4 (2009), pp. 1574–1609.

[NYD83] A. Nemirovskii, D. B. Yudin, and E. Dawson. "Problem complexity and method efficiency in optimization". In: (1983).

[Nes05] Y. Nesterov. "Smooth Minimization of Non-Smooth Functions". In: *Math. Program.* 103.1 (May 2005), pp. 127–152. DOI: 10.1007/s10107-004-0552-5.

[NES06] Y. NESTEROV. *Cubic regularization of Newton's method for convex problems with constraints*. LIDAM Discussion Papers CORE 2006039. Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), Apr. 2006.

[Nes08] Y. Nesterov. "Accelerating the cubic regularization of Newton's method on convex problems". In: *Mathematical Programming* (2008).

[Nes83a] Y. Nesterov. "A method for solving the convex programming problem with convergence rate $O(1/k^2)$". In: *Dokl. Akad. Nauk SSSR* 269 (1983), pp. 543–547.

[Nes83b] Y. Nesterov. "A method of solving a convex programming problem with convergence rate O (1/k2)". In: *Soviet Mathematics Doklady*. Vol. 27. 2. 1983, pp. 372–376.

[Nes88] Y. Nesterov. "On an approach to the construction of optimal methods of minimization of smooth convex functions". In: *Ekonomika i Mateaticheskie Metody* 24.3 (1988), pp. 509–517.

[Nes03] Y. Nesterov. *Introductory Lectures on Convex Optimization. 2004*. 2003.

[Nes07] Y. Nesterov. "Dual Extrapolation and Its Applications to Solving Variational Inequalities and Related Problems". In: *Math. Program.* 109.2–3 (Mar. 2007), pp. 319–344.

[Nes09] Y. Nesterov. "Primal-Dual Subgradient Methods for Convex Problems". In: *Mathematical Programming* 120.1 (2009), pp. 221–259.

[Nes15] Y. Nesterov. "Universal gradient methods for convex optimization problems". In: *Mathematical Programming* 152.1-2 (2015), pp. 381–404.

[Nes19] Y. Nesterov. "Implementable tensor methods in unconstrained convex optimization". In: *Mathematical Programming* 186 (2019), pp. 157–183.

[Nes21] Y. Nesterov. "Superfast second-order methods for Unconstrained Convex Optimization". In: *Journal of Optimization Theory and Applications* (Aug. 2021). DOI: 10.1007/s10957-021-01930-y.

[NN94] Y. Nesterov and A. Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics, 1994. DOI: 10.1137/1.9781611970791.

[NP06] Y. Nesterov and B. Polyak. "Cubic regularization of Newton method and its global performance". In: *Math. Program.* 108 (Aug. 2006), pp. 177–205. DOI: 10.1007/s10107-006-0706-8.

[Nes18] Y. E. Nesterov. *Lectures on Convex Optimization*. Vol. 137. Springer Optimization and Its Applications. Springer, 2018.

[Ngu+17] L. M. Nguyen et al. “SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient”. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 2613–2621.

[NW06] J. Nocedal and S. J. Wright. *Numerical Optimization*. 2e. New York, NY, USA: Springer, 2006.

[OP15] F. Orabona and D. Pál. “Scale-free algorithms for online linear optimization”. In: *International Conference on Algorithmic Learning Theory*. Springer. 2015, pp. 287–301.

[OS19] S. Oymak and M. Soltanolkotabi. “Toward Moderate Overparameterization: Global Convergence Guarantees for Training Shallow Neural Networks”. In: *IEEE Journal on Selected Areas in Information Theory* 1 (2019), pp. 84–105.

[Pas+19] A. Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035.

[Pha+20] N. H. Pham et al. “ProxSARAH: An Efficient Algorithmic Framework for Stochastic Composite Nonconvex Optimization”. In: *Journal of Machine Learning Research* 21.110 (2020), pp. 1–48.

[Pin+17] L. Pinto et al. “Robust adversarial reinforcement learning”. In: *ICML ’17: Proceedings of the 34th International Conference on Machine Learning*. 2017.

[PJ92] B. T. Polyak and A. B. Juditsky. “Acceleration of Stochastic Approximation by Averaging”. In: *SIAM Journal on Control and Optimization* 30.4 (1992), pp. 838–855. DOI: 10.1137/0330046.

[Pol06] B. Polyak. “Newton-Kantorovich Method and Its Global Convergence”. In: *Journal of Mathematical Sciences* 133 (2006), pp. 1513–1523.

[Pol87] B. T. Polyak. *Introduction to Optimization*. New York, NY, USA: Optimization Software, 1987.

[Pop80] L. D. Popov. “A modification of the Arrow-Hurwicz method for search of saddle points”. In: *Mathematical notes of the Academy of Sciences of the USSR* 28 (1980), pp. 845–848.

[RSS12] A. Rakhlin, O. Shamir, and K. Sridharan. “Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization”. In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*. ICML’12. Edinburgh, Scotland: Omnipress, 2012, pp. 1571–1578.

[RS13] S. Rakhlin and K. Sridharan. “Optimization, learning, and games with predictable sequences”. In: *Advances in Neural Information Processing Systems*. 2013, pp. 3066–3074.

[RKK18a] S. Reddi, S. Kale, and S. Kumar. “On the convergence of Adam and Beyond”. In: *International Conference on Learning Representations*. 2018.

[Red+16] S. J. Reddi et al. “Stochastic variance reduction for nonconvex optimization”. In: *International conference on machine learning*. PMLR. 2016, pp. 314–323.

[RKK18b] S. J. Reddi, S. Kale, and S. Kumar. “On the Convergence of Adam and Beyond”. In: *International Conference on Learning Representations*. 2018.

[RM51] H. Robbins and S. Monro. “A Stochastic Approximation Method”. In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407. DOI: 10.1214/aoms/1177729586.

[Roc76] R. T. Rockafellar. “Monotone Operators and the Proximal Point Algorithm”. In: *SIAM Journal on Control and Optimization* 14.5 (1976), pp. 877–898. DOI: 10.1137/0314056.

[SR13] M. Schmidt and N. L. Roux. “Fast convergence of stochastic gradient descent under a strong growth condition”. In: *arXiv preprint arXiv:1308.6370* (2013).

[Scu+10] G. Scutari et al. “Convex optimization, game theory, and variational inequality theory in multiuser communication systems”. In: *IEEE Signal Processing Magazine* 27.3 (May 2010), pp. 35–49.

[Sha12] S. Shalev-Shwartz. “Online Learning and Online Convex Optimization”. In: *Found. Trends Mach. Learn.* 4.2 (Feb. 2012), pp. 107–194. DOI: 10.1561/2200000018.

[SB14] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. USA: Cambridge University Press, 2014.

[SJM20] C. Song, Y. Jiang, and Y. Ma. “Variance Reduction via Accelerated Dual Averaging for Finite-Sum Optimization”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 833–844.

[SSB18] V. Srinivasan, A. R. Sankar, and V. N. Balasubramanian. “ADINE: An adaptive momentum method for stochastic gradient descent”. In: *Proceedings of the ACM india joint international conference on data science and management of data*. 2018, pp. 249–256.

[SM10] M. J. Streeter and H. B. McMahan. “Less Regret via Online Conditioning”. In: *CoRR* abs/1002.4862 (2010).

[Str+10] A. Strehl et al. “Learning from logged implicit exploration data”. In: *Advances in neural information processing systems* 23 (2010).

[Sut+13] I. Sutskever et al. “On the importance of initialization and momentum in deep learning”. In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by S. Dasgupta and D. McAllester. Vol. 28. Proceedings of Machine Learning Research. Atlanta, Georgia, USA: PMLR, 17–19 Jun 2013, pp. 1139–1147.

[Syr+15] V. Syrgkanis et al. “Fast convergence of regularized learning in games”. In: *NIPS’15: Proceedings of the 29th International Conference on Neural Information Processing Systems*. 2015, pp. 2989–2997.

[Tan+16] C. Tan et al. “Barzilai-Borwein Step Size for Stochastic Gradient Descent”. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS’16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 685–693.

[TM13] R. Tibshirani and M. Marchetti-Bowick. *Lecture 6*. Feb. 2013.

[TH12] T. Tieleman and G. Hinton. “Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude”. In: *COURSERA: Neural networks for machine learning* 4.2 (2012).

[TP19] P. T. Tran and L. T. Phong. “On the Convergence Proof of AMSGrad and a New Version”. In: *IEEE Access* 7 (2019), pp. 61706–61716.

[Tra+19] Q. Tran-Dinh et al. “Hybrid stochastic gradient descent algorithms for stochastic nonconvex optimization”. In: *arXiv preprint arXiv:1905.05920* (2019).

[Tse08] P. Tseng. “On accelerated proximal gradient methods for convex-concave optimization”. In: *submitted to SIAM Journal on Optimization* (Jan. 2008).

[Vas+17] A. Vaswani et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017.

[VBS19] S. Vaswani, F. R. Bach, and M. Schmidt. “Fast and Faster Convergence of SGD for Over-Parameterized Models and an Accelerated Perceptron”. In: *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*. Ed. by K. Chaudhuri and M. Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 1195–1204.

[VAM21] D. Q. Vu, K. Antonakopoulos, and P. Mertikopoulos. “Fast Routing under Uncertainty: Adaptive Learning in Congestion Games via Exponential Weights”. In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*. Ed. by M. Ranzato et al. 2021, pp. 14708–14720.

[Wan+13] C. Wang et al. “Variance reduction for stochastic gradient optimization”. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 1*. 2013, pp. 181–189.

[WA18] J.-K. Wang and J. D. Abernethy. “Acceleration through Optimistic No-Regret Dynamics”. In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio et al. Curran Associates, Inc., 2018, pp. 3824–3834.

[Wan+19] Z. Wang et al. “SpiderBoost and Momentum: Faster Variance Reduction Algorithms”. In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by H. M. Wallach et al. 2019, pp. 2403–2413.

[WWB19] R. Ward, X. Wu, and L. Bottou. "AdaGrad Stepsizes: Sharp Convergence Over Nonconvex Landscapes". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 6677–6686.

[WS16] B. Woodworth and N. Srebro. "Tight Complexity Bounds for Optimizing Composite Objectives". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 3646–3654.

[XRV17] H. Xiao, K. Rasul, and R. Vollgraf. "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms". In: *arXiv preprint arXiv:1708.07747* (2017).

[Xia10] L. Xiao. "Dual averaging methods for regularized stochastic learning and online optimization". In: *Journal of Machine Learning Research* 11.Oct (2010), pp. 2543–2596.

[XWW20] Y. Xie, X. Wu, and R. Ward. "Linear convergence of adaptive stochastic gradient descent". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 1475–1485.

[Zah+18] M. Zaheer et al. "Adaptive Methods for Nonconvex Optimization". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018.

[Zha+17] C. Zhang et al. "Understanding deep learning requires rethinking generalization". In: *International Conference on Learning Representations*. 2017.

[Zha+20] J. Zhang et al. "Why Are Adaptive Methods Good for Attention Models?" In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Vancouver, BC, Canada: Curran Associates Inc., 2020.

[ZMJ13] L. Zhang, M. Mahdavi, and R. Jin. "Linear convergence with condition number independent access of full gradients". In: *Advances in Neural Information Processing Systems* 26 (2013), pp. 980–988.

[ZXG18] D. Zhou, P. Xu, and Q. Gu. "Stochastic Nested Variance Reduction for Nonconvex Optimization". In: *Advances in Neural Information Processing Systems* 31 (2018), pp. 3921–3932.

[Zho+18] D. Zhou et al. "On the Convergence of Adaptive Gradient Methods for Nonconvex Optimization". In: *ArXiv* abs/1808.05671 (2018).

[ZY16] Z. A. Zhu and Y. Yuan. "Improved SVRG for Non-Strongly-Convex or Sum-of-Non-Convex Objectives". In: *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*. Ed. by M. Balcan and K. Q. Weinberger. Vol. 48. JMLR Workshop and Conference Proceedings. JMLR.org, 2016, pp. 1080–1089.

[Zin03] M. Zinkevich. “Online Convex Programming and Generalized Infinitesimal Gradient Ascent”. In: *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*. ICML’03. Washington, DC, USA: AAAI Press, 2003, pp. 928–935.

[Zou+19] F. Zou et al. “A Sufficient Condition for Convergences of Adam and RMSProp”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

Ali Kavis

École Polytechnique Fédérale de Lausanne
School of Computer and Communication Sciences
Address: EPFL STI IEL LIONS
ELD 243, 1015 Lausanne, CH

E-mail: alikavis@gmail.com
Website: alikavis.github.io
Google Scholar: sPrPq6oAAAAJ

Research Interests

convex/nonconvex optimization, stochastic and adaptive methods, variational inequalities, online learning

Education

École Polytechnique Fédérale de Lausanne, 9/2017-8/2023
PhD in School of Computer and Communication Sciences

Bilkent University, Ankara, Turkey, 8/2012 - 6/2017
B.S. in Computer Engineering (Salutatorian with GPA: 3.99/4.00)

Profesional Experience

Research Assistant, 9/2017 - 8/2023
École Polytechnique Fédérale de Lausanne (EPFL)
School of Computer and Communication Sciences

Tutorial at EUSIPCO 2020, 1/2021
Adaptive Optimization Methods for Machine Learning and Signal Processing
Lecturers: Kfir Levy, **Ali Kavis**, Ahmet Alacaoglu, Volkan Cevher

Minisymposium at SIAM OP23, 5/2023
Adaptivity and Universality: First-Order Methods and Beyond
Organizers: Volkan Cevher, **Ali Kavis**, Kimon Antonakopoulos

Honors and Awards

Swiss National Science Foundation Postdoc.Mobility Grant (CHF 120K), 2023-2025

EPFL IC Doctoral Studies Fellowship, 2017-2018

Travel Award, NeurIPS, 2019

Spotlight Paper, NeurIPS, 2018 & 2019

Bilkent University Comprehensive Scholarship (full tuition waiver with monthly stipend), 2012-2017

Publications

A. Kavis*, S. Skoulakis*, K. Antonakopoulos, L. T. Dadi, V. Cevher. *Adaptive Stochastic Variance Reduction for Non-convex Finite-Sum Minimization*, Conference on Neural Information Processing Systems (NeurIPS), 2022

K. Antonakopoulos*, **A. Kavis***, V. Cevher. *Extra-Newton: A First Approach to Noise-Adaptive Accelerated Second-Order Methods*, Conference on Neural Information Processing Systems (NeurIPS), 2022

A. Kavis, K. Y. Levy, V. Cevher. *High Probability Bounds for a Class of Nonconvex Algorithms with AdaGrad Stepsize*, International Conference on Learning Representations (ICLR), 2022

K. Y. Levy, **A. Kavis**, V. Cevher. *STORM+: Fully Adaptive SGD with Recursive Momentum for Nonconvex Optimization*, Conference on Neural Information Processing Systems (NeurIPS), 2021

K. Antonakopoulos, T. Pethick, **A. Kavis**, P. Mertikopoulos, V. Cevher. *Sifting through the noise: Universal first-order methods for stochastic variational inequalities*, Conference on Neural Information Processing Systems (NeurIPS), 2021

P. Mertikopoulos, N. Hallak, **A. Kavis**, V. Cevher. *On the almost sure convergence of stochastic gradient descent in non-convex problems*, Conference on Neural Information Processing Systems (NeurIPS), 2020

P. T. Y. Rolland, A. Eftekhari, **A. Kavis**, V. Cevher. *Double-Loop Unadjusted Langevin Algorithm*, International Conference on Machine Learning (ICML), 2020

A. Kavis*, K. Y. Levy*, F. Bach, V. Cevher. *UniXGrad: A Universal, Adaptive Algorithm with Optimal Guarantees for Constrained Optimization*, Conference on Neural Information Processing Systems (NeurIPS), 2019

P. T. Y. Rolland, **A. Kavis**, A. Immer, A. Singla, V. Cevher. *Efficient learning of smooth probability functions from Bernoulli tests with guarantees*, International Conference on Machine Learning (ICML), 2019

Y. P. Hsieh, **A. Kavis**, P. T. Y. Rolland, V. Cevher. *Mirrored Langevin Dynamics*, Conference on Neural Information Processing Systems (NeurIPS), 2018

Talks

SIAM Conference on Mathematics of Data Science, 2022.
High Prob. Bounds for a Class of Nonconvex Algorithms with AdaGrad Stepsize.

SIAM Conference on Optimization, 2023.
A First Approach to Noise-Adaptive Accelerated Second-Order Methods

Teaching Experience

Teaching Assistant
Mathematics of data: from theory to computation (2018-2022, Head TA in 2021)
Theory and Methods for Reinforcement Learning (2022)
Information, Computation, Communication (2020)
Practice of object-oriented programming (2018)

Professional Service

Reviewer
NeurIPS, COLT, ICLR, Springer Machine Learning

Coding Skills

Python, Pytorch, C/C++, MATLAB