École doctorale de sciences mathématiques de Paris centre
Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Sorbonne Université

THÈSE DE DOCTORAT

Discipline : Mathématiques
Specialité : Statistiques

présentée par

Nicklas Werge

Learning from time-dependent streaming data with online stochastic algorithms

---

dirigée par: Antoine GODICHON-BAGGIONI and Olivier WINTENBERGER

Soutenue le 29 Septembre 2022

Composition du Jury:

Alain DURMUS	ENS Paris-Saclay	Examinateur
Gersende FORT	Institut de Mathématiques de Toulouse	Examinatrice
Sébastien GADAT	Université Toulouse I Capitole	Rapporteur
Antoine GODICHON-BAGGIONI	Sorbonne Université	Co-directeur de thèse
Sylvain LE CORFF	Sorbonne Université	Président du jury
Olivier WINTENBERGER	Sorbonne Université	Directeur de thèse

Laboratoire de Probabilités,
Statistiques et Modélisation.
UMR 8001
Boîte courrier 158
4 place Jussieu
75252 Paris Cedex 05

Sorbonne Université
École doctorale de sciences
mathématiques de Paris centre.
Boîte courrier 290
4 place Jussieu
75252 Paris Cedex 05

This work was supported by the Paris Ile-de-France Region Ph.D. program.

"Sometimes I'll start a sentence and
I don't even know where it's going.
I just hope I find it along the way."

Michael Scott, The Office (US).

Abstract

Learning from time-dependent streaming data with online stochastic algorithms

In recent decades, intelligent systems, such as machine learning and artificial intelligence, have become mainstream in many parts of society. However, many of these methods often work in a batch or offline learning setting, where the model is re-trained from scratch when new data arrives. Such learning methods suffer some critical drawbacks, such as expensive re-training costs when dealing with new data and thus poor scalability for large-scale and real-world applications. At the same time, these intelligent systems generate a practically infinite amount of large datasets, many of which come as a continuous stream of data, so-called streaming data. Therefore, first-order methods with low per-iteration computational costs have become predominant in the literature in recent years, in particular the Stochastic Gradient (SG) descent [124]. These SG methods have proven scalable and robust in many areas ranging from smooth and strongly convex problems to complex non-convex ones, which makes them applicable in many learning tasks for real-world applications where data are large in size (and dimension) and arrive at a high velocity. Such first-order methods have been intensively studied in theory and practice in recent years [21]. Nevertheless, there is still a lack of theoretical understanding of how dependence and biases affect these learning algorithms.

A central theme in this thesis is to learn from time-dependent streaming data and examine how changing data streams affect learning. To achieve this, we first construct the Stochastic Streaming Gradient (SSG) algorithm, which can handle streaming data; this includes several SG-based methods, such as the well-known SG descent and mini-batch methods, along with their Polyak-Ruppert average estimates [118, 129]. The SSG combines SG-based methods' applicability, computational benefits, variance-reducing properties through mini-batching, and the accelerated convergence from Polyak-Ruppert averaging. Our analysis links the dependency and convexity level, enabling us to improve convergence. Roughly speaking, SSG methods can converge using non-decreasing streaming batches, which break long-term and short-term dependence, even using biased gradient estimates. More surprisingly, these results form a heuristic that can help increase the stability of SSG methods in practice. In particular, our analysis reveals how noise reduction and accelerated convergence can be achieved by processing the dataset in a specific pattern, which is beneficial for large-scale learning problems.

At last, we propose an online adaptive recursive estimation routine for Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) models called AdaVol. The AdaVol procedure relies on stochastic algorithms combined with Variance Targeting Estimation (VTE); AdaVol has computationally efficient properties, while VTE overcomes some convergence difficulties due to the lack of convexity of the Quasi-Maximum Likelihood (QML) procedure. Empirical demonstrations show favorable trade-offs between AdaVol's stability and its ability to adapt to time-varying estimates.

Keywords: *stochastic optimization, machine learning, stochastic algorithms, online learning, streaming, time-dependent data*

Résumé

Apprentissage à partir de données en continu dépendant du temps avec des algorithmes stochastiques en ligne

Au cours des dernières décennies, les systèmes intelligents, tels que l'apprentissage automatique et l'intelligence artificielle, se sont imposés dans de nombreux secteurs de la société. Cependant, bon nombre de ces méthodes fonctionnent souvent dans un cadre d'apprentissage batch ou hors ligne, où le modèle est réentraîné à partir de zéro lorsque de nouvelles données arrivent. Ces méthodes d'apprentissage présentent des inconvénients majeurs, tels que des coûts de réentraînement élevés en cas de nouvelles données, et donc une faible adaptabilité aux données massives et en pratique. Dans le même temps, ces systèmes intelligents génèrent une quantité pratiquement infinie de grands jeux de données, dont beaucoup se présentent sous la forme d'un flux quasi-continu de données, appelé streaming. C'est pourquoi les méthodes du premier ordre à faible coût de calcul par itération sont devenues prédominantes dans la littérature ces dernières années, en particulier la descente de gradient stochastique (SG) [124]. Ces méthodes SG sont adaptées et robustes dans de nombreux domaines allant de problèmes lisses et fortement convexes aux problèmes complexes non convexes, ce qui les rend applicables à de nombreuses tâches d'apprentissage pour des applications réelles où les données sont de grande taille (et de grande dimension) et arrivent à une vitesse élevée. Ces méthodes du premier ordre ont été intensivement étudiées en théorie et en pratique au cours des dernières années [21]. Néanmoins, il y a encore un manque de compréhension théorique sur la façon dont la dépendance et le biais affectent ces algorithmes d'apprentissage.

Un thème central de cette thèse est d'apprendre à partir de données en streaming dépendantes du temps et d'examiner comment les flux de données changeants affectent l'apprentissage. Pour y parvenir, nous construisons d'abord l'algorithme de gradient stochastique en streaming (SSG), qui peut gérer des données quasi-continues ; il comprend diverses méthodes SG, telles que la descente SG (c'est-à-dire l'algorithme de Robbins-Monro), les méthodes SG à mini-batch, ainsi que leurs estimations moyennes Polyak-Ruppert [118, 129]. La descente SSG combine l'applicabilité des méthodes fondées sur les SG, les avantages en termes de calcul, les propriétés de réduction de la variance grâce au mini-batching, et la convergence accélérée grâce à la moyénisation de Polyak-Ruppert. Notre analyse repose sur le niveau de dépendance et de convexité du problème, et nous permet d'améliorer la convergence. En résumé, les méthodes SSG peuvent converger en utilisant des mini-batchs de tailles croissante en streaming, qui rompent la dépendance à long terme et à court terme, et ce, même en utilisant des estimations de gradient biaisées. De manière plus surprenante, ces résultats forment une heuristique qui peut aider à augmenter la stabilité des méthodes SSG en pratique. En particulier, notre analyse révèle comment une réduction du bruit et une convergence accélérée peuvent être obtenues en traitant l'ensemble de données selon une procédure spécifique, ce qui est bénéfique pour les problèmes d'apprentissage à grande échelle.

Enfin, nous proposons une méthode d'estimation récursive adaptative en ligne pour les modèles GARCH appelée AdaVol. La procédure AdaVol repose sur des algorithmes stochastiques combinés à la méthode de ciblage de la variance (VTE) ; AdaVol présente des propriétés efficaces sur le plan du calcul grâce à la VTE qui permet de surmonter certaines difficultés de convergence dues au manque de convexité de la procédure de vraisemblance quasi-maximale (QML). Des démonstrations empiriques montrent des compromis favorables entre la stabilité d'AdaVol et sa capacité à s'adapter à des estimations variant dans le temps.

Mots clés: *optimisation stochastique, apprentissage automatique, algorithmes stochastiques, apprentissage en ligne, streaming, données dépendantes*

Acknowledgements

It is now three years since I moved to Paris for this thesis project. A lot has happened since then. I left with high expectations for my new Parisian life. I had to learn French, eat buttery food every night, taste a lot of local grapes, and truly enjoy the Parisian culture. I *almost* made it, not because of my own efforts but because of the people around me. I owe immense gratitude to everyone who made this journey enjoyable. Therefore, I will now start to name-drop all those I want to thank, but don't feel neglected if I have forgotten your name.

First of all, I would like to thank my supervisors, Antoine and Olivier. This project would not have been possible without your enormous commitment and support, and the many hours you have invested in me. Olivier, it has been over seven years since we got to know each other, and I am so grateful for everything you have taught me. I have truly lived by the saying that *an expert is someone who has made all the mistakes that can be made in a narrow field.* Without your extraordinary patience and support, it would not have been possible for me to make all these mistakes; without saying I have made enough. Antoine, I really appreciate all the hours you spent evaluating my work. Without your enthusiasm and knowledge, I would never have made it through. I am sorry that my lack of language skills has excluded me from all your jokes; as far as I could tell, they were hilarious. I have nothing more to say than that it has been a great pleasure working with you.

Thanks to DIM Math Innov for the opportunity they gave me through the Paris Ile-de-France Region program, and in particular to Dominique Wetzel, who constantly took care of us candidates.

I would also like to express my gratitude to Sebastien Gadat and Genero Sucarrat for their careful reading of my thesis and for many insightful comments and suggestions. Thanks to Alain Durmus, Gersende Fort, and Sylvain Le Corff for being my jury members.

Thanks to my office family. You all made this thesis so much more fun! I especially want to thank Joseph, Grâce, Cyril, Camila, Ludovic, Antonio, Miguel, Ariane, Pierre, Iqraa, Franceso, Alice, Alexis, Patrick, and those who went before me; Adeline, Aude, Nicolas, Riccardo, and Vincent. Thanks to the LPSM organization for their work and help; Hugues, Louise, Valérie, and Nathalie. Last but not least, I want to thank Gloria and Thibault, it would never have been the same without you. I am already sad not to see you every day, but I wish you all the best. I hope we see each other again soon.

Thanks to my friends and family for supporting and listening to me during this thesis. It has been the most stressful and challenging period of my life, but you have helped me make it the happiest. Finally, I would like to thank the most important person, Amalie. All these years, you stayed by my side, supporting and reassuring me, pushing and cheering me on, and sometimes putting up with me. Our relationship is my greatest achievement.

Table of Contents

Chapter 1: Introduction

Contents

1.1 Learning from Streaming Data

Machine learning and intelligent systems have become an integral part of modern society, e.g., through online learning, deep learning, reinforcement learning and supervised learning [59, 70, 71, 140]. This interest is particularly driven by readily available datasets of enormous size and increasingly powerful and cheaper computer systems. Many of these systems follow the traditional learning scheme where we observe an entire dataset, build our model and then predict/label new observations, e.g., see the left-hand side of Figure 1.1. At the same time, the use of these intelligent systems generates a practically infinite amount of large-scale datasets, many of which come as a continuous data stream, so-called *streaming data*, such as internet traffic (e.g., tweets, search engines, advertising), self-driving cars, financial investments, weather data, or other sensor data [1, 87, 90]. These data streams should be processed sequentially with the property that the data stream may change over time. To be able to distinguish between these, we introduce *streaming learning*. In streaming learning, we assume that data arrives sequentially over time, in which we update our models during this continuous influx of data, e.g., see right-hand side of Figure 1.1.

Streaming data arrives as an endless sequence of samples (data points), which means that at any given time, the model must be able to adapt to the samples observed (so far) to predict/label new samples accurately. Such streaming models can never be seen as complete but must be updated continuously as newer samples arrive. Methods that recalculate the model from scratch on the arrival of new samples are impractical due to their high computational cost. Therefore we need procedures that effectively update the model as more samples arrive. This computational efficiency should not be at the expense of accuracy; the model's accuracy should be close to that achieved if

Figure 1.1: Learning schemes: large- and small-scale learning vs. learning from streaming data.

we built a model from scratch using all the samples [20].

Machine learning is rooted in statistics and is highly dependent on the efficiency of numerical algorithms. One of the main ingredients in machine learning is to choose the optimization method. The optimization model must numerically estimate the parameters of a given model so that the model can make accurate decisions based on future data. These model parameters must be selected optimally for a given learning problem based on currently available data. Traditional gradient-based batch approaches can effectively solve small learning problems but are unfeasible for streaming (and large-scale) learning problems [19]. Therefore, there is a need for effective optimization methods that process data samples at low computational costs while having sufficient theoretical guarantees. This setting goes beyond the traditional optimization methods [37, 78], which gives first-order gradient methods an important role.

A hallmark of learning from streaming data (or large-scale learning) is the uncertainty from limited (or no) access to accurate information about incoming data. This can be implemented by assuming that the objectives of the optimization are stochastic, leading to Stochastic Optimization (SO) [102]. Solving the SO problem in a streaming framework means we approach the objective using the gradually arriving samples drawn according to an unknown process. Stochastic algorithms, such as the Stochastic Gradient (SG) descent [124], have been one of the core methods of efficiently dealing with SO problems. Since then, much work has been done to analyze, improve and develop methods dealing with stochasticity [21, 54, 89, 101]. An essential extension is Polyak-Ruppert averaging [118, 129]; this technique sequentially aggregates the estimates, which leads to a smoother curves (i.e., variance reduction in the estimation trajectories), and accelerates the convergence.

The classical analyses of SO problems typically require unbiased gradients drawn independently and identically distributed (i.i.d.) from some underlying (and unknown) data generation process [34]. However, in practice, learning often involves a data-generating process that produces highly dependent data samples, which are known to heavily bias the SO problem and slow down the convergence of learning; these time-dependent streaming data could, e.g., be meteorological or financial time series. Nevertheless, stochastic algorithms for dependent data are not as well understood as for i.i.d. data. Stochastic algorithms can converge even when they only have access to biased gradients,

but most analysis has been developed with specific applications in mind [4, 15, 37, 40, 130]. Yet, some researchers have examined the convergence of stochastic algorithms these difficult settings, e.g., see Agarwal and Duchi [3], Karimi et al. [79].

While the above works utilized concepts of data dependence to characterize different stochastic algorithms over dependent data, there is still a lack of theoretical understanding on how different levels of data dependence affects these algorithms. In particular, the learning scheme of stochastic algorithms critically affects the bias and variance of the learning process. In fact, under i.i.d. data and convexity, we have shown that SG-based methods achieves only slightly improved convergence bounds by using constant mini-batch vs. single batch [57]. However, these learning schemes may lead to substantially different convergence behaviors over highly dependent data, as the gradients are no longer unbiased estimates. Therefore, it is vital to understand the interplay between data dependence and stochastic algorithms. In this thesis, we go beyond these standard assumptions by allowing dependent and biased gradients. Specifically, we study convergence rates of stochastic algorithms over a broad spectrum of data dependence levels under various streaming schemes, including mini-batch and averaged mini-batches.

Organization. In this introduction, we summarize this thesis's main ideas and challenges. Section 1.1.1 explains some of the benefits of SG-based methods and emphasizes where our contributions should be placed in the literature. Next, this thesis's contributions (and a brief description of them) are given in Section 1.1.2. Section 1.2 presents the stochastic streaming algorithms that solve the SO problem, namely the Stochastic Streaming Gradient (SSG) and Averaged SSG (ASSG). In addition, we highlight some of its extensions used to design optimization methods, e.g., noise/variance reduction methods that use the power of mini-batch methods to reduce noise during optimization and iterative averaging that improves convergence rates (Section 1.2.3). Finally, we provide a summary of the main results of this thesis (Section 1.3); here we show how our SSG and ASSG methods overcome these challenges and achieve convergence in difficult settings with long- and short-range dependencies, biased estimates, and changing data streams.

1.1.1 Examples of Applications

In statistics and machine learning, one often encounters the following optimization problem [70]: let $l_1, \dots, l_n$ be a sequence of random differentiable functions from $\mathbb{R}^d$ to $\mathbb{R}$. Our goal is to find an approximate solution $\theta \in \mathbb{R}^d$ of the following optimization problem,

$$L_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} l_t(\theta). \tag{1.1.1}$$

We say that $L_n : \mathbb{R}^d \to \mathbb{R}$ yields the empirical risk, i.e., empirical loss. Many problems, from classification, and regression to ranking, can be written on this form (1.1.1), e.g., see Teo et al. [141] for examples of scalar and vectorial loss functions and their derivatives.

For example, consider the simple case where we have some samples (X_t, Y_t), $t = 1, \dots, n$ from a couple of random variables (X, Y) in $\mathcal{X} \times \mathcal{Y}$. Our interest is to find predictor $h_\theta : \mathcal{X} \to \mathbb{R}$ over

some parameterization $\{h_\theta\}_{\theta\in\mathbb{R}^d}$, by minimizing (1.1.1) with $l_t(\theta) = l(h_\theta(X_t), Y_t) + \lambda\Omega(\theta)$, where l is some loss function, $\lambda > 0$ a regulizer parameter, and $\Omega : \mathbb{R}^d \to \mathbb{R}$ some regularizer, e.g., the l_1 or l_2 regularization. The loss l could be the quadratic loss, logistic loss, (squared) hinge loss, or Huber's (robust) loss, but it depends on the experiments that one wants to perform [21, 26, 31, 104, 110].

More specifically, in classification one has $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{-1, 1\}$, thus, taking the hinge loss $\max\{0, 1 - Y_t h_\theta(X_t)\}$, $h_\theta(X_t) = \theta^T X_t$ and $\Omega(\theta) = \|\theta\|_2^2$ one obtains the SVM problem. On the other hand, taking the logistic loss $\log(1 + \exp(-Y_t h_\theta(X_t)))$ and again $h_\theta(X_t) = \theta^T X_t$ and $\Omega(\theta) = \|\theta\|_2^2$ one obtains the (regularized) logistic regression problem. In regression one has $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$, thus, taking the squared loss $(h_\theta(X_t) - Y_t)^2$, $h_\theta(X_t) = \theta^T X_t$ and $\Omega(\theta) = 0$ one obtains the vanilla least-squares problem. This problem can be rewritten in vector notation as $\min_{\theta\in\mathbb{R}^d} \|X\theta - Y\|^2$, where $X \in \mathbb{R}^{n\times d}$ is the matrix with X_t the t'th row and $Y = (Y_1, \ldots, Y_n)^T$. Hence, with $\Omega(\theta) = \|\theta\|_2^2$ one obtains the ridge regression problem, while with $\Omega(\theta) = \|\theta\|_1$ this is the LASSO problem [142]. The regularizer $\Omega(\theta)$ can be seen as a simple convex function that, when added to a non-convex loss function l, may convexifies it, thereby helping gradient-based optimization techniques avoid poor solutions at its saddle points or flat areas.

In addition to these simple methods, there are many other more complex methods, such as those for linear/non-linear time series. These methods have been successfully used in a wide range of applications due to their ability to describe or predict time-varying (dependent) processes, e.g., the AutoRegressive (AR), Moving-Average (MA), and AutoRegressive Moving-Average (ARMA) models are the most well-known models for time series [25, 30, 66]. Standard time series analysis often relies on independence and constant noise, but it can be relaxed by, e.g., the AutoRegressive Conditional Heteroskedasticity (ARCH) model [45].

Let us give some examples of (1.1.1) with our notation: let (Z_t) denote some real-valued time series. For an AR(d) model our interest is in explaining $Y_t = Z_t$ using the explanatory values $X_t = (Z_{t-1}, \ldots, Z_{t-d})$. In other words, an AR model explains the variable of interest by a linear combination of past values of the variable, whereas a multiple regression model explains the variable of interest by a linear combination of predictors. A reasonable measure could be to compare to the best possible AR model, i.e., at time t, we make a prediction $h_\theta(X_t) = \sum_{i=1}^d \theta_i Z_{t-i}$, after which Y_t is revealed, and we suffer a loss $l_t(\theta) = (Y_t - h_\theta(X_t)^2$, where $\theta = (\theta_1, \ldots, \theta_d)$ are the coefficients.

Another example could be an ARCH(d) model, e.g, see Francq and Zakoïan [48]: in this case, we are interested in predicting the volatility of Z_t (i.e., $Y_t = Z_t^2$) using the explanatory values $X_t = (Z_{t-1}^2, \ldots, Z_{t-d}^2)$. Thus, an ARCH($d$) process with parameters $\theta = (\omega, \alpha_1, \ldots, \alpha_d)$ has $h_\theta(X_t) = \omega + \sum_{i=1}^d \alpha_i Z_{t-i}^2$. The natural estimator for θ is the Quasi-Maximum Likelihood Estimator (QMLE) due to its theoretically appealing properties and robustness to extreme values, e.g., see Patton [113]. This means we are considering Quasi-Likelihood (QL) losses of form $l_t(\theta) = \log h_\theta(X_t) + Y_t^2/h_\theta(X_t)$. However, the concavity of the QL loss raises some issues, which we will return to in Chapter 4.

More sophisticated models such as ARMA, ARIMA, and Generalized ARCH (GARCH) are studied in Anava et al. [7], Liu et al. [92], Werge and Wintenberger [150]. In Chapter 4, we propose an online adaptive estimation routine for GARCH models called AdaVol. AdaVol uses

online stochastic algorithms to estimate the parameters of a GARCH model using the QMLE. More generally, online learning algorithms of (both stationary and non-stationary) dependent time series have been studied in Agarwal and Duchi [3], Wintenberger [152].

To sum up, the optimization problem in (1.1.1) contains numerous models, whether we want to incorporate dependency or not. Even though (1.1.1) may look remarkably simple, it ranges widely. We will see several different examples of (l_t) over the following chapters; in particular, we will consider each l_t as a varying block of streaming data, i.e., a *streaming batch* of varying size.

Computational Trade-offs by Stochastic Gradient Methods

Let us now illustrate the computational advantages of stochastic gradient-based methods. For this purpose, we introduce some fundamental gradient-based optimization algorithms to minimize the empirical risk L_n in (1.1.1). We are currently introducing them in the context of minimizing the empirical risk L_n, but our later analysis will focus on algorithms for minimizing the expected risk L in (1.2.4). Optimization methods can be divided into two broad categories: stochastic and batch. Gradient-based batch optimization methods can, in their simplest form with $k \in \mathbb{N}$, be defined as

$$\theta_k = \theta_{k-1} - \gamma_k g_k(\theta_{k-1}), \text{ with } g_k(\theta_{k-1}) = \begin{cases} \frac{1}{|C_k|} \sum_{i \in C_k} \nabla_\theta l_i(\theta_{k-1}), & \text{(mini-batch)} \\ \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta l_i(\theta_{k-1}), & \text{(batch gradient)} \end{cases} \tag{1.1.2}$$

with $\theta_0 \in \mathbb{R}^d$, $C_k \subseteq \{1, \ldots, n\}$ and (γ_k) is the learning rate. Solving (1.1.1) by traditional iterative gradient-based methods (1.1.2) can be effective in solving small-scale learning problems where n and d are small. A batch approach would have a computational cost of $\mathcal{O}(dn)$ per-iteration, i.e., $\mathcal{O}(kdn)$ computations after k iterations. In practice, one would instead use an iterative mini-batch method in which a subset of samples C_k is chosen randomly in each iteration. A mini-batch have a computational costs of $\mathcal{O}(|C_k|d)$ per-iteration, i.e., $\mathcal{O}(k|C_k|d)$ computations after k iterations. These (mini-)batch approaches will converge quickly, but these approaches are too computationally expensive and will thus be prohibitive for streaming data (or large datasets), where n and/or d are large, since we would have computational costs of $\mathcal{O}(kdn)$ (or $\mathcal{O}(k|C_k|d)$) every time new data arrives. Instead, online algorithms have been the core method of interest [71, 132], particularly the SG descent [124], which, in the context of minimizing L_n in (1.1.1), is defined for $t = 1, \ldots, n$ by

$$\theta_t = \theta_{t-1} - \gamma_t \nabla_\theta l_t(\theta_{t-1}). \tag{1.1.3}$$

The SG descent in (1.1.3) has a computational cost of only $\mathcal{O}(d)$ for each new data point, corresponding to a mini-batch with $|C_k| = 1$ and $k = 1$, which is very cheap. These SG methods and batch approaches have different trade-offs in computational costs and expected convergence rates. One may ask, why have SG methods become so famous for large-scale problems? This question requires careful consideration of the computational trade-offs between stochastic and batch methods and a thorough investigation of their convergence capabilities [19–21].

1.1.2 Contributions and Outline of Thesis

Main goals. The central theme of this thesis is to learn from time-dependent streaming data, where traditional optimization techniques are unsustainable due to their high computational cost. We want to explore SG-based methods robustness and convergence guarantees under various settings. In short, the main objectives of this thesis are

1. To allow learning algorithms to handle streaming data,
2. To improve learning by adapting streaming learning to the hardness of the problem; the level of dependence, noisiness, and convexity.

Chapter 2 introduces the Stochastic Optimization (SO) problem in a streaming framework. In this streaming setting, we propose techniques for minimizing convex objectives through unbiased estimates of their gradients. Our analysis extends the work of Moulines and Bach [96] to a streaming framework. A fundamental aspect of this chapter is to explore how changing data streams affect these techniques; this include everything from vanilla SG descent and Averaged SG (ASG)[1], mini-batch SG and ASG, to more exotic learning designs. Our main theoretical contribution is the non-asymptotic analysis of the SSG and Averaged SSG (ASSG) method in this streaming framework. Our results show a noticeable improvement in convergence rates by having learning rates that adapt to the expected data streams. In particular, we show how to obtain improved convergence rates robust to any data streaming rate. In addition, noise reduction can be achieved by processing the data in a specific pattern, which is advantageous for large-scale machine learning problems. These theoretical results are illustrated for various data streams, showing the effectiveness of the proposed streaming algorithms.

- Godichon-Baggioni, A., Werge, N., Wintenberger, O. (2021). Non-asymptotic analysis of stochastic approximation algorithms for streaming data. *arXiv preprint arXiv:2109.07117*.

Next, in Chapter 3, we investigate the SO problem in a streaming framework [57], where the data comes from a dependent stochastic process. We provide non-asymptotic analysis and quantify the magnitude of achievable convergence rates under various dependency structures (sometimes leading to divergence). Our framework covers many applications with dependence and biased gradients. Our results build a connection between the level of dependency and convexity, enabling us to improve convergence. Roughly speaking, SSG methods can break short-term and even long-term dependence by using increasing batch sizes, which counteracts the dependency structures. We show that biased SSG methods converge, and that they can converge with the same accuracy as unbiased SSG methods if the bias is not too large. More surprisingly, our results give an explicit heuristic that can be used in practice to help increase the stability of SSG methods. In particular, we show that mini-batch is essential to break dependence and ensure convexity. In addition, we can accelerate convergence by simultaneously averaging.

[1]The ASG estimate is referring to the Polyak-Ruppert averaging estimate [118, 129].

- Godichon-Baggioni, A., Werge, N., and Wintenberger, O. (2022). Learning from time-dependent streaming data with online stochastic algorithms. *arXiv preprint arXiv:2205.12549*.

In Chapter 4, we propose an online adaptive estimation routine for GARCH models called AdaVol. The AdaVol procedure relies on stochastic algorithms combined with the technique of Variance Targeting Estimation (VTE) [49]. This AdaVol method has computationally efficient properties, while VTE alleviates some convergence difficulties encountered by the usual Quasi-Maximum Likelihood (QML) estimation due to a lack of convexity. Empirical demonstrations show favorable trade-offs between AdaVol's stability and its ability to adapt to time-varying estimates. The adaptation to time-varying parameters was a surprising advantage that appeared when we applied our method to real-life observations.[2]

- Werge, N., Wintenberger, O. (2022). Adavol: An adaptive recursive volatility prediction method. *Econometrics and Statistics*, 23:19–35.

During my Ph.D. first year, I worked part-time at Advestis as an AI researcher. Advestis is an award-winning french fintech company making AI-based quantitative trading and research. My mission at Advestis was to conduct independent research with the aim of constructing a regime-shifting model that can distinguish between market regimes in a wide range of financial markets. This work resulted in the paper below, where I propose an asset-independent regime-switching model for risk-adjusted return forecasts based on hidden Markov models; a full version paper is included in Chapter A.

- Werge, N. (2021). Predicting risk-adjusted returns using an asset independent regime-switching model. *Expert Systems with Applications*, 184:115576.

1.2 Stochastic Optimization for Streaming Data

The computational complexity of an algorithm is a limited element when handling streaming data (or large-scale datasets). The main focus of this thesis is to study stochastic algorithms for solving stochastic optimization problems in a steaming framework. We focus on gradient-based optimization algorithms with convex objectives that we will analyze in a non-asymptotic way. Remark that gradient estimates can be constructed from function values if gradients are unavailable, e.g., see Nesterov and Spokoiny [106]. In Section 1.2.1, we formalize stochastic optimization problems in a steaming framework. Next, in Section 1.2.2, we define the Stochastic Streaming Gradient (SSG) methods [57, 58] and highlight some of the main benefits of SSG methods. This section is concluded with a preview of some advanced optimization techniques discussed in the literature (Section 1.2.3).

[2] AdaVol was recently ranked third among the best probability forecasters in the M6 financial forecast competition, e.g., see https://m6competition.com. The M6 competition will be live, lasting for twelve months, starting in February 2022 and ending a year later in 2023. Quarterly prizes will be awarded for each of the four quarters of the competition, of which we ranked third in the first quarter.

1.2.1 Problem Formulation

In statistics and machine learning, we often want to describe the behavior of a real system of interest, usually in the form of a parameterized mathematical model [21, 70]. Therefore, we set up a mathematical function representing how well the model describes the system of interest with the model parameters as arguments. Throughout the thesis, we refer to this as the objective function. We can now describe our streaming setting formally: at each time $t \in \mathbb{N}$, a *block* consisting of $n_t \in \mathbb{N}$ random functions $l_t = (l_{t,1}, \dots, l_{t,n_t})$ arrives. The objective of the Stochastic Optimization (SO) problem is to minimize functions of the form

$$L(\theta) = \mathbb{E}[l_t(\theta)], \tag{1.2.4}$$

with respect to $\theta \in \Theta$, where Θ is a closed convex set in $\mathbb{R}^d$ and $l_t : \Theta \to \mathbb{R}$ is some random differentiable functions (possibly non-convex), e.g, see Boyd et al. [26], Nesterov et al. [104]. The function L is called the objective function (and sometimes also the risk). We assume that L is μ-quasi-strongly convex and Lipschitz continuous, e.g., see definitions in Section 1.3.1. Our goal is to find this unique global minimizer $\theta^* \in \Theta$ of L. Minimization of the objective function L is achieved without evaluating it directly but by with use of random functions $\nabla_\theta l_{t,i}$ as estimates of the gradient of L. These random functions $(l_{t,i})$ can be seen as observations (or random loss functions) depending on L and some underlying noise sequence.

An example of such a SO problem (1.2.4) can be given as follows [87]: there is an unknown one-to-one mapping $L : \Theta \to \mathbb{R}$ embedded into the system by nature, which we are interested in. Thus, in order to approximate L (and recover θ from it), we use the gradient estimates $(\nabla_\theta l_t)$, where l_t is e.g., the loss between the predicted $h_\theta(X_t)$ and true Y_t outputs, respectively; here we assume that the prediction function h_θ has a fixed form and is parameterized by a real vector $\theta \in \Theta$ over which the optimization is to be performed. Hence, the aim is to find θ such that the prediction function h_θ minimizes the risk L.

In order to compare our streaming methods fairly, we should always compare in terms of the number of observations used, namely using $N_t = \sum_{i=1}^{t} n_i$, which is the (accumulated) sum of observations at time t. For example, for empirical risk minimization (1.1.1), one chooses the predictor by minimizing the empirical risk over a parameterized set of predictors potentially with regularization [89, 141, 145], e.g, for a parameterization $\{h_\theta\}_{\theta\in\Theta}$, a regulizer parameter λ, and a regularizer $\Omega : \Theta \to \mathbb{R}$, this requires to minimize $L_{N_t}(\theta) = \frac{1}{N_t}\sum_{i=1}^{t} l_i(\theta)$ with $l_i(\theta) = \sum_{j=1}^{n_i} l(h_\theta(X_{i,j}), Y_{i,j}) + \lambda\Omega(\theta)$, where $X_i = (X_{i,1}, \dots, X_{i,n_i})$ and $Y_i = (Y_{i,1}, \dots, Y_{i,n_i})$ are the blocks of n_i observations that arrive at each i (a.k.a. streaming batches). Consequently, if $n_t = 1$, we have the classic setting of Moulines and Bach [96], which we described in Section 1.1.1. Whereas, if n_t is constant, we have a (constant) mini-batch. Having n_t varying means that we have varying streaming batches depending on the time t, which is what we are particularly interested in.

1.2.2 Stochastic Streaming Gradients

The prototypical method for solving the SO problem is SG-based methods [24, 68, 133, 144, 153, 156]. But to solve the SO problem (1.2.4) in a streaming framework, we use the Stochastic Streaming Gradient (SSG) method proposed by Godichon-Baggioni et al. [57], given as

$$\theta_t = \theta_{t-1} - \frac{\gamma_t}{n_t} \sum_{i=1}^{n_t} \nabla_\theta l_{t,i} \left(\theta_{t-1}\right), \quad \theta_0 \in \Theta, \tag{1.2.5}$$

where γ_t is the learning rate satisfying the conditions $\sum_{i=1}^{\infty} \gamma_i = \infty$ and $\sum_{i=1}^{\infty} \gamma_i^2 < \infty$ [124]. Note that if $\forall t$, $n_t = 1$, SSG becomes the well-known SG descent. In many models, there may be constraints on the parameter space, which would require a projection of the parameters; therefore, we also introduce the Projected Stochastic Streaming Gradient (PSSG) estimate, defined by

$$\theta_t = \mathcal{P}_\Theta \left(\theta_{t-1} - \frac{\gamma_t}{n_t} \sum_{i=1}^{n_t} \nabla_\theta l_{t,i} \left(\theta_{t-1}\right) \right), \quad \theta_0 \in \Theta, \tag{1.2.6}$$

where $\mathcal{P}_\Theta$ denotes the the Euclidean projection onto Θ, i.e., $\mathcal{P}_\Theta(\theta) = \arg\min_{\theta' \in \Theta} \|\theta - \theta'\|_2$. To shorten notation, we let $\nabla_\theta l_t(\theta) = n_t^{-1} \sum_{i=1}^{n_t} \nabla_\theta l_{t,i}(\theta)$.

In this streaming setting, we are also interested in acceleration approaches to the existing algorithms. An essential extension is the Polyak-Ruppert averaging [118, 129], which guarantees optimal statistical efficiency without jeopardizing the computational cost; the Averaged Stochastic Streaming Gradient (ASSG) is given by

$$\bar{\theta}_t = \frac{1}{N_t} \sum_{i=0}^{t-1} n_{i+1} \theta_i, \quad \bar{\theta}_0 = 0, \tag{1.2.7}$$

where $N_t = \sum_{i=1}^{t} n_i$ is the sum of observations at time t. Likewise, PASSG denotes the averaged estimate of PSSG (1.2.6). In addition, (1.2.7) can be modified to a weighted average version, giving greater weight to the latest estimates and thereby improving the convergence while limiting the effect of poor initializations; examples of these can be found in Boyer and Godichon-Baggioni [27], Mokkadem and Pelletier [95].

These averaging methods sequentially aggregates the estimates, which leads to a smoother curves (i.e., variance reduction in the estimation trajectories), and accelerates the convergence. Practically, as we handle data sequentially, we will make use of the rewritten formula: $\bar{\theta}_t = (N_{t-1}/N_t)\bar{\theta}_{t-1} + (n_t/N_t)\theta_{t-1}$ with $\bar{\theta}_0 = 0$. Pseudo-code of these streaming estimates are presented in Algorithm 1.1. Each update of these methods is very cheap, involving only the computation of n_t gradients $\nabla_\theta l_t(\theta_{t-1}) = n_t^{-1} \sum_{i=1}^{n_t} \nabla_\theta l_{t,i}(\theta_{t-1})$, i.e., a computational costs of $\mathcal{O}(dn_t)$. Thus, we have the same computationally efficiency as the SG descent in (1.1.3), e.g., see Section 1.1.1. These methods are notable as (θ_t) is a stochastic process whose behavior is determined by the random sequence (l_t) and the learning rate (γ_t). Still, as we shall see in our analysis in Section 1.3, the

direction of $-\gamma_t \nabla_\theta l_t(\theta_t)$ might not point at θ_t. However, if it does in expectation, we follow the gradient of L, and thereby the sequence (θ_t) can be guided toward the minimizer of L.

Algorithm 1.1: Stochastic streaming gradient estimates (SSG/PSSG/ASSG/PASSG)

Inputs : $\theta_0 \in \Theta$, project: **True** or **False**, average: **True** or **False**
Outputs: θ_t, $\bar{\theta}_t$ (resulting estimates)
$\bar{\theta}_0 = 0$
for *each* $t \geq 1$, *a block of* n_t *data arrives* **do**
  $\theta_t \leftarrow \theta_{t-1} - \frac{\gamma_t}{n_t} \sum_{i=1}^{n_t} \nabla_\theta l_{t,i}(\theta_{t-1})$
  if *project is* **True then**
    $\theta_t \leftarrow \mathcal{P}_\Theta(\theta_t)$ `/* project estimate */`
  if *average is* **True then**
    $\bar{\theta}_t \leftarrow (N_{t-1}/N_t)\bar{\theta}_{t-1} + (n_t/N_t)\theta_{t-1}$ `/* average estimate */`

In the same way as the streaming methods in (1.2.5) to (1.2.7), one can define the (full) batch gradient method in our streaming setting, given as

$$\theta_{k+1} = \theta_k - \frac{\gamma_k}{N_t} \sum_{i=1}^{t} \sum_{j=1}^{n_i} \nabla_\theta l_{i,j}(\theta_k). \tag{1.2.8}$$

The computational cost of each step (1.2.8) is prohibitive for N_t very large (e.g., 10^6 or 10^9), although one would expect a better step estimate when all N_t samples are considered at each iteration. Per-iteration would have a computational cost of $\mathcal{O}(dN_t)$, i.e., $\mathcal{O}(kdN_t)$ computations after k iterations (similarly to what we saw in Section 1.1.1). In addition, we must also take into account that in streaming settings, one has limited response time between new observations; thus, computational costs become even more crucial.

1.2.3 Beyond Stochastic Streaming Gradients

Due to the massive popularity of SG methods, it is obvious to ask how we can make SG even more efficient, robust, and user-friendly for several different optimization methods. This question has led to very many variants, of which we will outline some of the most common, but we will omit algorithms that are impractical for large-scale datasets, e.g., see Boyd et al. [26], Nesterov et al. [104] for more details on second-order methods (such as Newton's method), or other extensions.

The choice of learning rate (γ_t) has a significant impact on the convergence of SG methods; if it is too small, it will slow down the convergence, while too high a learning rate may prevent convergence or even divergence, as the loss function will fluctuate around the minimum. Thus, an adaptive learning rate would be much more effortless to adjust and more user-friendly, as it requires less fine-tuning. In addition, it would be preferable to have a learning rate per dimension, which thereby adjust learning individually as convergence evolves. Some of the most common adaptive learning algorithms for SG optimization is Momentum [119], Nesterov accelerated gradient [103], Adagrad [42], Adadelta [155], RMSprop [74], and Adam [83]. Ruder [127] gives an overview of

various SG methods for (convex and non-convex) optimization, including how to parallelize and distribute SG updates.

If one were to look at the trajectory of the noisy gradients SG uses as estimates, one would be surprised. This lack of robustness (or high noise level) can prevent SG methods from converging or lead to slow convergence. There are several techniques to improve the robustness of SG methods, of which some of the most known are mini-batch SG, gradient aggregation methods, and iterate averaging methods. Such methods have proven effective in practice and possess attractive theoretical properties because they reduce the noise in the gradient estimates [38, 76, 77, 108, 126]. The mini-batch SG uses a small subset of gradient estimates in each iteration, which intuitively reduces variance, makes it easier to tune the learning rate (γ_t), and improves the quality of each iteration. Gradient aggregation methods enhance the quality of the gradient estimates more adaptively; these methods smooth the iterations using past gradient estimates, e.g., using a weighted average of these past estimates [39]. On the other hand, iterate averaging methods do not accomplish noise reduction by averaging gradient estimates but instead by averaging the iterates computed during the optimization [118, 129]. These methods also have some appealing convergence acceleration properties, which we will come back to later.

1.3 Non-asymptotic Analysis of Stochastic Streaming Gradient Estimates

This section presents a summary of this thesis's main results [57, 58]. These results are shown in a simple form to be able to highlight the main conclusions, but an extended (and fully non-asymptotic) version can be found in the papers themselves, e.g., see Chapters 2 and 3. Before examining the stochastic streaming estimates in more detail, we briefly present the mathematical framework in Section 1.3.1. Next, Section 1.3.2 provides the analysis in the i.i.d. streaming setting. In Section 1.3.3, we expand these assumptions and notions to include time-dependency. Through this section, we will show examples of our findings.

Throughout this introduction, we consider the stochastic algorithms in (1.2.5) to (1.2.7) with learning rates on the form $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ with hyper-parameters $C_\gamma > 0$, $\beta \in [0, 1]$, and $\alpha > 0$ chosen accordingly to the expected streaming batches denoted by n_t. The streaming batches n_t are on the form $C_\rho t^\rho$ with $C_\rho \geq 1$ and $\rho \in (-1, 1)$; here $C_\rho = 1$ and $\rho = 0$ corresponds to the classical SG descent, where we process observations one-by-one [96]. If $C_\rho \in \mathbb{N}$ and $\rho = 0$ we consider mini-batch procedures of size C_ρ, and likewise, if $C_\rho \in \mathbb{N}$ and $\rho \in (-1, 1)$ we have varying streaming batches with initial batch size of C_ρ. We will refer to ρ as the *streaming rate*. Our aim is to bound $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ and $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ non-asymptotically such that these bounds depend solely on the problem's parameters. These non-asymptotic bounds are derived by explicitly bounding the t-th estimate of (1.2.5) and (1.2.6) using classical techniques from stochastic approximations [13, 87]. Remark that almost sure convergence of SO algorithms were shown in Pelletier [115].

1.3.1 Mathematical Framework

The analysis of SO algorithms requires assumptions on the objective function L: the SO problem (1.2.4) is specified over a convex domain Θ, which in this thesis we always take to be a compact subset of $\mathbb{R}^d$, $d \geq 1$, and an objective function $L : \Theta \to \mathbb{R}$ which is convex with respect to its argument $\theta \in \Theta$. This problem is a closely related branch of optimization tools for (online) convex optimization [26, 71, 104].

Quasi-strong Convex Objectives

Following Gower et al. [60], Moulines and Bach [96], we assume that L has a unique global minimizer $\theta^* \in \Theta$ such that $\nabla_\theta L(\theta^*) = 0$, and it is μ-quasi-strongly convex [80, 99], i.e, there exists $\mu > 0$ such that $\forall \theta \in \Theta$,

$$L(\theta^*) \geq L(\theta) + \langle \nabla_\theta L(\theta), \theta^* - \theta \rangle + \frac{\mu}{2} \|\theta^* - \theta\|^2. \tag{1.3.9}$$

Teo et al. [141] provides a comprehensive record of various convex functions L used in machine learning applications. Milder degrees of convexity have been studied by, e.g., Karimi et al. [80], which studied stochastic gradient methods under the Polyak-Łojasiewicz condition [93, 117], or Gadat and Panloup [50], which studied the Ruppert-Polyak averaging estimate under some Kurdyka-Łojasiewicz-type condition [86, 93]. Relaxations of convexity is crucial in practice to ensure robustness and adaptiveness of the algorithms, e.g., for non-strongly convex SO, see Bach and Moulines [9], Necoara et al. [99], Nemirovski et al. [101].

Smoothness of the Objectives

Some additional assumptions are needed for bounding the averaging estimate $(\bar{\theta}_t)$ in (1.2.7): let the function L have C_∇-Lipschitz continuous gradients, i.e., there exists a constant $C_\nabla > 0$, $\forall \theta, \theta' \in \Theta \subseteq \mathbb{R}^d$,

$$\|\nabla_\theta L(\theta) - \nabla_\theta L(\theta')\| \leq C_\nabla \|\theta - \theta'\|. \tag{1.3.10}$$

Remark that one has $\mu \mathbb{I}_d \preceq \nabla^2_\theta L(\theta) \preceq C_\nabla \mathbb{I}_d$ in the case L is μ-quasi-strongly convex and twice differentiable, e.g., see Nesterov et al. [104]. As discussed in Bottou et al. [21], this assumption ensures that $\nabla_\theta L$ does not vary arbitrarily, making the gradient $\nabla_\theta L$ a useful indicator on how to decrease L. Moreover, note that when L is μ-quasi-convex and C_∇-smooth, the convergence of gradient methods will depend on the number $C_\nabla/\mu \geq 1$. If C_∇/μ is small, we have fast convergence, and conversely, if it is large, we get oscillations [21]. Next, assume that the Hessian of L is C'_∇-Lipschitz-continuous, that is, there exists $C'_\nabla \geq 0$ such that $\forall \theta, \theta' \in \Theta \subseteq \mathbb{R}^d$,

$$\|\nabla^2_\theta L(\theta) - \nabla^2_\theta L(\theta')\| \leq C'_\nabla \|\theta - \theta'\|. \tag{1.3.11}$$

Note that (1.3.10) and (1.3.11) only needs to hold true for $\theta' = \theta^*$.

1.3.2 Learning from Streaming Data

A fundamental aspect of Godichon-Baggioni et al. [57] is to explore how changing data streams affect these SO methods. These data streams includes everything from vanilla SG and ASG descent, mini-batch SG and ASG, to more exotic learning designs. Our analysis extends the work of Moulines and Bach [96] to a streaming framework. Our main theoretical contribution is the non-asymptotic analysis of the SSG methods in this streaming framework. Our results show a noticeable improvement in convergence rates by having learning rates that adapt to the expected data streams. In particular, we show how to obtain improve convergence, while being robust to any data streaming rate.

Remember the description of our streaming framework in which we solve our SO problem (1.2.4): at each time $t \in \mathbb{N}$, a *block* consisting of $n_t \in \mathbb{N}$ random functions $l_t = (l_{t,1}, \ldots, l_{t,n_t})$ arrive. Let (l_t) constitute a sequence of independent differentiable random functions (possibly non-convex) and their gradients unbiased estimates of $\nabla_\theta L$, e.g., see Nesterov et al. [104] for definitions and properties of such functions. We assume the following about the $l_{t,i}$ functions at each $t \in \mathbb{N}$ with $i = 1, \ldots, n_t$:

Assumption 1.3.1 (unbiased gradients)**.** *The random variable $\nabla_\theta l_{t,i}(\theta)$ is square-integrable and $\forall \theta \in \Theta$, $\mathbb{E}[\nabla_\theta l_{t,i}(\theta)] = \nabla_\theta L(\theta)$.*

In the classical convergence analysis of SG methods, one assumes that the SGs are uniformly bounded [72, 101, 121, 133]. However, this assumption is too restrictive as it only may hold for some losses, e.g., see Bottou et al. [21], Nguyen et al. [107]. Instead, we follow the same ideas as in Gower et al. [60], Moulines and Bach [96], to make the following assumption about the expected smoothness of the stochastic gradients $(\nabla_\theta l_{t,i})$.

Assumption 1.3.2-p (C_l-expected smoothness)**.** *For $p \geq 1$, there exists $C_l > 0$ such that $\forall \theta, \theta' \in \Theta$, $\mathbb{E}[\|\nabla_\theta l_{t,i}(\theta) - \nabla_\theta l_{t,i}(\theta')\|^p] \leq C_l^p \mathbb{E}[\|\theta - \theta'\|^p]$.*

Assumption 1.3.2-p can be seen as an assumption about the smoothness properties of $(l_{t,i})$, and it only needs to hold for $\theta' = \theta^*$. Moreover, under Assumption 1.3.1, Assumption 1.3.2-p with $p = 1$ implies the condition in (1.3.10) by Jensen's inequality. The last fundamental assumption (Assumption 1.3.3-p) is about the finitude of $(\nabla_\theta l_{t,i}(\theta^*))$:

Assumption 1.3.3-p (σ-gradient noise)**.** *For $p \geq 1$, there exists $\sigma > 0$ such that $\mathbb{E}[\|\nabla_\theta l_{t,i}(\theta^*)\|^p] \leq \sigma^p$.*

These assumptions are modified versions of the standard assumptions for stochastic approximations as they hold for any $i = 1, \ldots, n_t$, e.g., see [13, 87, 96]. By the smoothness assumption (Assumption 1.3.2-p), we avoid the unfavorable uniformly bounded gradients assumption, which is too restrictive and only holds for a few losses. Assumption 1.3.3-p enables to give an upper bound of the Frobenius norm of the variance of the gradient for $p = 2$ and is very usual (see [96] for

instance). For SSG and PSSG, we only need Assumptions 1.3.2-p and 1.3.3-p to hold for $p = 2$, whereas, for ASSG and PASSG, we need $p = 4$ in order to bound the fourth-order moment. Our framework include classic examples: stochastic approximation (Robbins-Monro setting [124]) and learning from i.i.d. data, such as linear regression, logistic regression, general ridge regressions and quantile regression, p-means, and softmax regression, under regularity conditions [33, 136]; here it is important to remark that most of these examples lead to the minimization of only locally strongly convex objectives, and the assumptions are only verified for the projected estimates, i.e., PSSG and PASSG [141].

Stochastic Streaming Estimates

Recall that we consider learning rates on the form $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ with hyper-parameters $C_\gamma > 0$, $\beta \in [0, 1]$, and α chosen accordingly to the expected streaming batches denoted by $n_t = C_\rho t^\rho$. But before we present our results, we want to recall the result for the classical SG descent (i.e., $n_t = 1$ as $C_\rho = 1$ and $\rho = 0$) shown by Moulines and Bach [96]:

Theorem 1.3.1. *Denote $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ for some $\delta_0 \geq 0$, where (θ_t) follows (1.2.5) or (1.2.6). Assume that Assumption 1.3.1, Assumptions 1.3.2-p and 1.3.3-p for $p = 2$ hold true. Then there exists $C_\delta > 0$ such that for $\alpha \in (1/2, 1)$, we have*

$$\delta_t \leq \frac{2^{1+\alpha}\sigma^2 C_\gamma}{\mu N_t^\alpha} + \mathcal{O}(\exp(-C_\delta N_t^{1-\alpha})). \tag{1.3.12}$$

The non-asymptotic bound in (1.3.12) depends explicitly upon the problem's parameters. Such bounds were the first of their kind, whereas previous results focused mainly on almost sure convergence., e.g., see [115].

Decay of the initial conditions. The condition of having $\alpha \in (1/2, 1)$ is a natural restriction from Robbins and Monro [124], ensuring $\sum_{i=1}^\infty \gamma_i = \infty$ and $\sum_{i=1}^\infty \gamma_i^2 < \infty$. The primary conclusion of Theorem 1.3.1 is that (1.3.12) can be divided into a noise term $\frac{2^{1+\alpha}\sigma^2 C_\gamma}{\mu N_t^\alpha}$ and a sub-exponential term $\mathcal{O}(\exp(-C_\delta N_t^{1-\alpha}))$ (an explicit version of this term can be found in Moulines and Bach [96]). Thus, we should focus on reducing the noise term without harming the natural decay of the sub-exponential term.

Now, let us present our first result [57]: we start by considering constant streaming batches (i.e., mini-batch SSG) where n_t follows the constant streaming batch size $C_\rho \in \mathbb{N}$:

Theorem 1.3.2. *Denote $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ for some $\delta_0 \geq 0$, where (θ_t) follows (1.2.5) or (1.2.6). Assume that Assumption 1.3.1, Assumptions 1.3.2-p and 1.3.3-p for $p = 2$ hold true. Then there exists $C'_\delta > 0$ such that for $\alpha \in (1/2, 1)$, we have*

$$\delta_t \leq \frac{2^{1+\alpha}\sigma^2 C_\gamma}{\mu C_\rho^{1-\alpha-\beta} N_t^\alpha} + \mathcal{O}(\exp(-C'_\delta N_t^{1-\alpha})). \tag{1.3.13}$$

Variance reduction. Not surprisingly, the bound in (1.3.13) has the same structure as (1.3.12),

whereby we can make equivalent conclusions. However, the noise term in (1.3.13) is divided by $C_\rho^{1-\alpha-\beta}$, implying we could achieve variance reduction by taking $\alpha+\beta \leq 1$. Thus taking a large streaming batch size C_ρ will give us some variance reduction, but it will not increase the convergence rate, which is still determined by $\alpha \in (1/2, 1)$. Remark that too large streaming batch sizes C_ρ would be unsuitable in practice as it would mean we would only take a few steps before convergence is achieved.

These fixed-sized streaming batches are not the most realistic streaming setting. It is far more likely to vary in size depending on the data streams. Thus, let us now consider varying streaming batches where n_t are on the form $C_\rho t^\rho$ with $C_\rho \geq 1$ and $\rho \in (-1, 1)$ such that $n_t \geq 1$ for all t. We will refer to ρ as the *streaming rate*. For the convenience of notation, let $\tilde{\rho} = \rho \mathbb{1}_{\{\rho \geq 0\}}$.

Theorem 1.3.3. *Denote $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ for some $\delta_0 \geq 0$, where (θ_t) follows (1.2.5) or (1.2.6). Assume that Assumption 1.3.1, Assumptions 1.3.2-p and 1.3.3-p for $p = 2$ hold true. Then there exists $C''_\delta > 0$ such that for $\alpha - \beta\tilde{\rho} \in (1/2, 1)$, we have*

$$\delta_t \leq \frac{2^{1+(2+\rho)\phi}\sigma^2 C_\gamma}{\mu C_\rho^{(1-\beta)\mathbb{1}_{\{\rho\geq 0\}}-\phi} N_t^\phi} + \mathcal{O}(\exp(-C''_\delta N_t^{1-\phi})), \tag{1.3.14}$$

with $\phi = ((1-\beta)\tilde{\rho} + \alpha)/(1+\tilde{\rho})$.

When $\rho = 0$, Theorem 1.3.3 yields the same as Theorem 1.3.2. Moreover, when $C_\rho = 1$ and $\rho = 0$, we obtain the usual SG descent studied in Moulines and Bach [96], e.g., see Theorem 1.3.1.

Accelerated decay and variance reduction. The condition of having $\alpha - \beta\tilde{\rho} \in (1/2, 1)$ relaxes the usual condition of having $\alpha \in (1/2, 1)$ for ρ non-negative. In particular, accelerated convergence could be achieved by, e.g., setting $\alpha = 2/3$ and $\beta = 0$ for streaming rates $\rho > 0$, giving us $\delta_t = \mathcal{O}(N_t^{-(2/3+\rho)/(1+\rho)})$, meaning increasing streaming batches ($\rho > 0$) can accelerate convergence. Moreover, the noise term is scaled by $C_\rho^{1-\beta-\phi}$ for $\rho \geq 0$, implying we should take $\alpha + \beta \leq 1$ to obtain variance reduction (as we saw for Theorem 1.3.2).

Acceleration by Averaging

In what follows, we consider the averaging estimate $(\bar{\theta}_n)$ given in (1.2.7) derived with use of (θ_t) from (1.2.5) or (1.2.6). Instead of first considering the vanilla case $\{C_\rho = 1, \rho = 0\}$, then the mini-batch case $\{C_\rho \in \mathbb{N}, \rho = 0\}$, and finally the streaming case $\{C_\rho \in \mathbb{N}, \rho \in (-1, 1)\}$, we only consider the streaming case from which the other cases will follow.

Besides having Assumptions 1.3.2-p and 1.3.3-p to hold for $p = 4$, an additional assumption is needed for bounding the *rest* term of the averaging estimate.

Assumption 1.3.4. *There exists a non-negative self-adjoint operator Σ such that $\mathbb{E}[\nabla_\theta l_{t,i}(\theta^*)\nabla_\theta l_{t,i}(\theta^*)^\top] \preceq \Sigma$.*

Note that the operator Σ always exists when σ is finite for order $p = 4$ in Assumption 1.3.3-p. Moreover, to avoid calculating the six-order moment when considering projected average estimate

PASSG, we make the unnecessary assumption that $\|\nabla_\theta l_{t,i}(\theta)\|$ is uniformly bounded for any $\theta \in \Theta$; the derivation of the six-order moment can be found in Godichon-Baggioni [55].

Assumption 1.3.5. *Let* $d_{\min} = \inf_{\theta\in\partial\Theta}\|\theta-\theta^*\| > 0$ *with* $\partial\Theta$ *denoting the frontier of* Θ. *Moreover, there exists* $G_\Theta > 0$ *such that* $\forall t \geq 1$, $\sup_{\theta\in\Theta}\|\nabla_\theta l_{t,i}(\theta)\|^2 \leq G_\Theta^2$ *a.s., with* $i = 1, \ldots, n_t$.

Theorem 1.3.4. *Denote* $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ *with* $(\bar{\theta}_t)$ *given by* (1.2.7), *where* (θ_t) *follows* (1.2.5) *or* (1.2.6). *Assume that Assumption 1.3.1, Assumptions 1.3.2-p and 1.3.3-p for* $p = 4$, *and Assumption 1.3.4 hold true. Moreover, let* (1.3.10) *and* (1.3.11) *hold true. In addition, Assumption 1.3.5 must hold true only if* (θ_t) *follows* (1.2.6). *For* $\alpha - \beta\tilde{\rho} \in (1/2, 1)$, *we have*

$$\bar{\delta}_t^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + \mathcal{O}(\max\{N_t^{-1+\phi/2}, N_t^{-\phi}\}), \tag{1.3.15}$$

where $\Lambda = \mathrm{Tr}(\nabla_\theta^2 L(\theta^*)^{-1}\Sigma\nabla_\theta^2 L(\theta^*)^{-1})$ *and* $\phi = ((1-\beta)\tilde{\rho} + \alpha)/(1+\tilde{\rho})$.

Accelerated decay. As noticed in Polyak and Juditsky [118], the leading term Λ/N_t achieves the desirable Cramer-Rao bound, namely, the leading term Λ/N_t could obtain the optimal and incorrigible rate of $\mathcal{O}(N_t^{-1})$ [50, 97]. Moreover, this bound is achieved without inverting the Hessian, and it is invariant of the learning rate (γ_t). Thus, by averaging, we have increased the rate of convergence from $\mathcal{O}(N_t^{-\phi})$ (in Theorem 1.3.3) to the optimal rate $\mathcal{O}(N_t^{-1})$. As discussed in Gadat and Panloup [50], the bound of $\bar{\delta}_t$ can be seen as a bias-variance decomposition between the first and second term in (1.3.15).

Next, it is worth noting that there are no sub-exponential decaying terms for the initial conditions in (1.3.15), which is a common problem for averaging. This means we should be more careful when picking our hyper-parameters, e.g., taking C_γ too large. Nevertheless, these hyper-parameters decay at a rate of at least $\mathcal{O}(N_t^{-2})$.

Robustness towards streaming rates ρ. The main remainder term $\mathcal{O}(\max\{N_t^{-1+\phi/2}, N_t^{-\phi}\})$ reveal that $\phi = 2/3 \Leftrightarrow \alpha - \beta\tilde{\rho} = (2-\tilde{\rho})/3$, e.g., by setting $\beta = 0$, we should pick $\alpha = (2-\tilde{\rho})/3$. Likewise, if $\rho = 0$, we yield the same conclusion as in Moulines and Bach [96], namely $\alpha = 2/3$. However, these hyper-parameter choices are not resilient against any arrival schedule ρ. Nonetheless, we can robustly achieve $\phi = 2/3$ for any $\rho \in (-1, 1)$ by setting $\alpha = 2/3$ and $\beta = 1/3$. In other words, we can achieve optimal convergence for any data stream by having $\alpha = 2/3$ and $\beta = 1/3$. It is important to remark that these choices of hyper-parameters are not derived from exact bounds. Gadat and Panloup [50] establishes even tighter bounds (which also are optimal relative to Cramer-Rao's lower bound) under the Kurdyka-Łojasiewicz-type condition [86, 93], where they show that $\alpha = 3/4$, leading to the main remainder term $\mathcal{O}(N_t^{-5/4})$.

Example 1.3.1 (Geometric median)**.** The geometric median is a generalization of the real median introduced by Haldane [64]. Robust estimators such as the geometric median may be preferred over the mean when the data is noisy. Moreover, in our streaming framework, stochastic algorithms are preferred as they efficiently handle large samples of high-dimensional data [33, 55]. The geometric

median of $X \in \mathbb{R}^d$ is defined by $\theta^* \in \mathbb{R}^d$ which minimizes the convex function $L(\theta) = \mathbb{E}[\|X - \theta\| - \|X\|]$, e.g., see Gervini [53], Kemperman [81] for properties such as existence, uniqueness, and robustness (breakdown point). Thus, the gradient $\nabla_\theta L(\theta) = \mathbb{E}[\nabla_\theta l_t(\theta)]$ with $\nabla_\theta l_t(\theta) = -(X_t - \theta)/\|X_t - \theta\|$ is bounded as $\|\nabla_\theta l_t(\theta)\| \leq 1$. We omit to project our estimates as this would hide the errors we want to explore (which we will see more clearly in Example 1.3.2, where we consider real-life time-dependent streaming data). Instead of projecting the estimates, one could adapt the proof of Gadat and Panloup [50] to a streaming setting. Otherwise, if X_t is bounded, one can adapt Cardot et al. [32] to the streaming setting showing that the streaming estimates are bounded.

To measure the performance, we use the mean quadratic error of the parameter estimates over one-hundred replications, given by $(\mathbb{E}[\|\theta_{N_t} - \theta^*\|^2])_{t\geq 1}$. Note that averaging over several iterations gives a reduction in variability, which mainly benefits the SSG. Suppose (X_t) is standard Gaussian centered at $(\theta_i)_{1\leq i\leq d}$ with θ_i taken randomly in the range $[-d, d]$. Moreover, following the reasoning of Cardot et al. [33], we set $C_\gamma = \sqrt{d}$, and let $\alpha = 2/3$. For this example we take $d = 10$ (Figure 1.2).

In Figure 1.2a, we consider constant data streams to illustrate the results in Theorems 1.3.1 and 1.3.2; this figure shows the variance reduction effect for different constant streaming batches $C_\rho \in \{1, 8, 64, 128\}$ with $\beta = 0$ (as shown in Theorem 1.3.2). However, the robustness of the geometric median leaves only a small positive impact for further variance reduction. Thus, too large (constant) streaming batch sizes C_ρ hinders the convergence as we make too few iterations. In addition, we see an acceleration in decay by averaging, as explained in Theorem 1.3.4.

These findings can be extended to Figures 1.2b to 1.2e, where we vary the streaming rate ρ for streaming batch sizes $C_\rho = 1$, 8, 64, and 128, respectively, with $\beta = 0$. These figures shows an increase in decay of the SSG when the streaming rate ρ increase as mentioned after Theorem 1.3.3. But the lack of convergence improvements in Figures 1.2d and 1.2e comes from $\beta = 0$, which means we do not exploit the potential of using more observations to accelerate convergence.

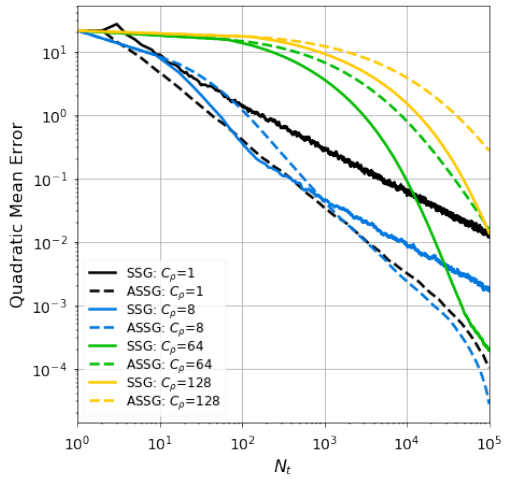
As discussed after Theorem 1.3.4, one example of this could be achieved by setting $\alpha = 2/3$ and $\beta = 1/3$ such that $\phi = 2/3$ for any ρ. As shown in Figure 1.2f, we can achieve this acceleration by simply taking $\beta = 1/3$. In addition, $\beta = 1/3$ provides optimal convergence robust to any streaming rate ρ. Choosing a proper $\beta > 0$ is particularly important when C_ρ is large, as robustness is an integral part of the geometric median method.

1.3.3 Learning from Time-dependent Streaming Data

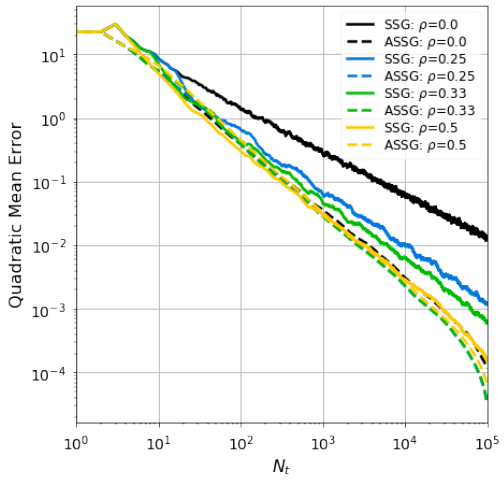
In this section, we go beyond the classical assumptions that require unbiased gradients (e.g., see Bottou et al. [21], Lacoste-Julien et al. [88]) by allowing the gradients to be dependent and biased estimators [58]. Convergence rates of SG descent with biased estimators has previously been studied in, e.g., Ajalloeian and Stich [4], Bertsekas [15], but not in a streaming setting. In this section, we start by replacing Assumptions 1.3.1 to 1.3.4 with the new assumptions; Assumptions 1.3.6-p to 1.3.9. These new assumptions are milder than the standard assumptions for stochastic approximations, e.g., see [13, 57, 87, 96]. We show some examples of how these assumptions could be verified using mixing conditions. Next, our convergence results are presented, with and without

Figure 1.2: Geometric median for various data streams $n_t = C_\rho t^\rho$. See Example 1.3.1 for details.

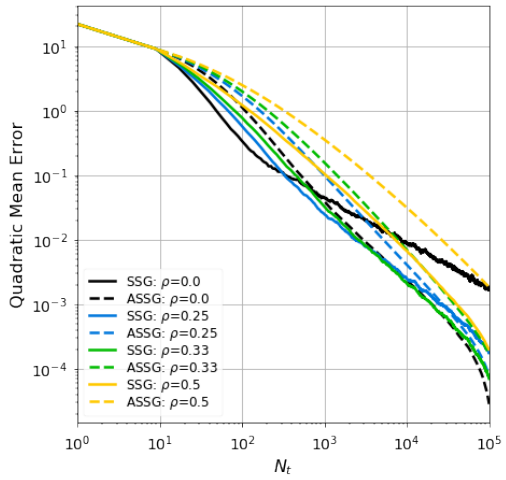
(a) Constant streaming batches, $\rho = 0$, $\beta = 0$



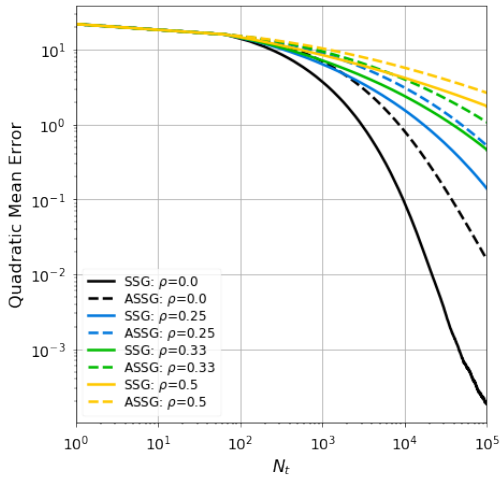
(b) Varying streaming batches, $C_\rho = 1$, $\beta = 0$



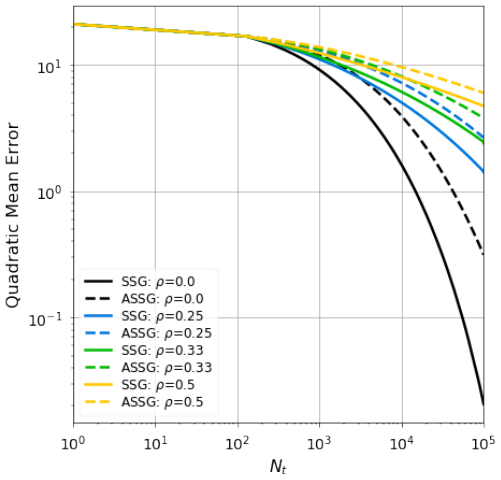
(c) Varying streaming batches, $C_\rho = 8$, $\beta = 0$



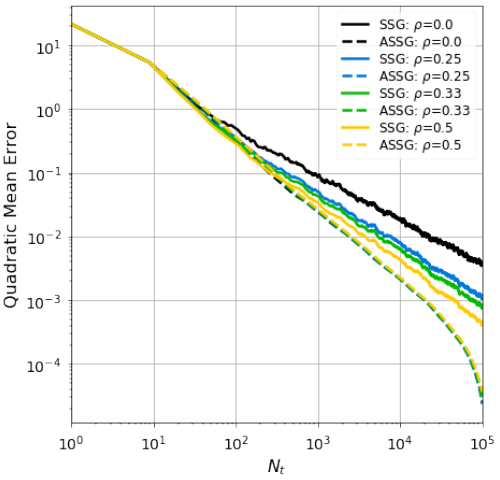
(d) Varying streaming batches, $C_\rho = 64$, $\beta = 0$



(e) Varying streaming batches, $C_\rho = 128$, $\beta = 0$



(f) Varying streaming batches, $C_\rho = 8$, $\beta = 1/3$



averaging. At last, experiments of our findings are illustrated. Let $\mathcal{F}_t = \sigma(l_i : i \leq t)$ denote the natural filtration of the SO problem (1.2.4).

Assumption 1.3.6-p ($D_\nu\nu_t$-dependence and $B_\nu\nu_t$-bias)**.** *Let θ_0 be $\mathcal{F}_0$-measurable. For each $t \geq 1$, the random function $\nabla_\theta l_t(\theta)$ is square-integrable, $\mathcal{F}_t$-measurable, and there exists a positive integer p such that for all $\mathcal{F}_{t-1}$-measurable $\theta \in \Theta$,*

$$\mathbb{E}[\|\mathbb{E}[\nabla_\theta l_t(\theta)|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta)\|^p] \leq \nu_t^p(D_\nu^p\mathbb{E}[\|\theta - \theta^*\|^p] + B_\nu^p), \tag{1.3.16}$$

for some positive sequence $(\nu_t)_{t\geq 1}$ with $D_\nu, B_\nu \geq 0$.

Assumption 1.3.7-p (κ_t-expected smoothness)**.** *There exists a positive integer p such that $\forall \theta, \theta' \in \Theta$, $\mathbb{E}[\|\nabla_\theta l_t(\theta) - \nabla_\theta l_t(\theta')\|^p] \leq \kappa_t^p\mathbb{E}[\|\theta - \theta'\|^p]$ for some positive sequence $(\kappa_t)_{t\geq 1}$.*

Assumption 1.3.8-p (σ_t-gradient noise)**.** *There exists a positive integer p such that $\mathbb{E}[\|\nabla_\theta l_t(\theta^*)\|^p] \leq \sigma_t^p$ for some positive sequence $(\sigma_t)_{t\geq 1}$.*

Assumption 1.3.6-p is on the form of mixing conditions for weakly dependence sequences, implying that dependence dilutes with the rate of ν_t. It is possible to verify Assumption 1.3.6-p by using moment inequalities for partial sums of strongly mixing sequences [123]; we will refer to this as short-range dependence. Note that for any positive integer p, Assumption 1.3.6-p can be upper bounded by

$$\mathbb{E}[\|\mathbb{E}[\nabla_\theta l_t(\theta)|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta)\|^p] \leq \mathbb{E}[\|\nabla_\theta l_t(\theta) - \nabla_\theta L(\theta)\|^p] = n_t^{-p}\mathbb{E}[\|S_t\|^p], \tag{1.3.17}$$

using Jensen's inequality, where $S_t = \sum_{i=1}^{n_t}(\nabla_\theta l_{t,i}(\theta) - \nabla_\theta L(\theta))$ is a d-dimensional vector. Let $(\nabla_\theta l_{t,i})$ be a strictly stationary sequence and assume that there exists some $r > p$ such that $\sup_{x>0}(x^r Q(x))^{1/r} < \infty$, where $Q(x)$ denotes the quantile function of $\|\nabla_\theta l_{t,i}\|$. Suppose that $(\nabla_\theta l_{t,i})$ is strongly α-mixing in the sense of Rosenblatt [125], with strong mixing coefficients $(\alpha_t)_{t\geq 1}$ satisfying $\alpha_t = \mathcal{O}(t^{-pr/(2r-2p)})$. Then by Rio [123, Corollary 6.1], we have that $\mathbb{E}[\|S_t\|^p] = \mathcal{O}(n_t^{p/2})$, meaning, (1.3.17) is at most $\mathcal{O}(n_t^{-p/2})$; this includes several linear, non-linear, and Markovian time series, e.g., see Bradley [29], Doukhan [41] for more examples, other mixing coefficients of weak dependence and the relations between them. In relation to the form of Assumption 1.3.6-p, this means that $B_\nu \neq 0$ in this case. However, having $B_\nu = 0$ is possible in well-specified examples. Note that Assumptions 1.3.7-p and 1.3.8-p can be verified using α-mixing conditions by analogues arguments as for Assumption 1.3.6-p such that κ_t^p and σ_t^p is $\mathcal{O}(n_t^{-p/2})$.

The (ν_t), (κ_t), and (σ_t) sequences may be considered as uncertain terms depending on the streaming-batch n_t. Thus, let $\nu_t = n_t^{-\nu}$, $\kappa_t = C_\kappa n_t^{-\kappa}$, and $\sigma_t = C_\sigma n_t^{-\sigma}$ with $\nu \in (0, \infty)$, $\kappa, \sigma \in [0, 1/2]$, and $C_\kappa, C_\sigma > 0$. Having, $\sigma, \kappa \in [0, 1/2]$ follows directly from Godichon-Baggioni et al. [57], since $\sigma = \kappa = 1/2$ corresponds to the i.i.d. case, whereas $\sigma, \kappa < 1/2$ allows noisier outputs. Similarly, $v_t = 0$ corresponds to the classical unbiased i.i.d. setting. Having $\nu_t = n_t^{-\nu}$ means Assumption 1.3.6-p, allow so-called long-range dependence (also known as long memory or long-range persistence) when $\nu \in (0, 1/2)$ and short-range dependence when $\nu \in [1/2, \infty)$. Thus, the

i.i.d. case is when $\nu \to \infty$. In this section, we continue to consider streaming-batches (n_t) on the form $C_\rho t^\rho$ but with $\rho \in [0, 1)$ (compared to Section 1.3.2 where $\rho \in (-1, 1)$).

Stochastic Streaming Estimates

Theorem 1.3.5. *Denote $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ for some $\delta_0 \geq 0$, where (θ_t) follows the recursion in (1.2.5) or (1.2.6). Assume that Assumptions 1.3.6-p to 1.3.8-p hold true for $p = 2$. Suppose $n_t = C_\rho t^\rho$ with $\rho \in [0, 1)$ and $C_\rho \in \mathbb{N}$, such that $\mu_\nu = \mu - \mathbb{1}_{\{\rho=0\}} 2D_\nu C_\rho^{-\nu} > 0$. There exists $C_\delta''' > 0$ such that for $\alpha - \rho\beta \in (1/2, 1)$, we have*

$$\delta_t \leq \frac{2^{\frac{7+6\rho\sigma}{1+\rho}} C_\sigma^2 C_\gamma}{\mu_\nu C_\rho^{\frac{2\sigma-\beta-\alpha}{1+\rho}} N_t^{\frac{\rho(2\sigma-\beta)+\alpha}{1+\rho}}} + \frac{2^{\frac{2+6\rho\nu}{1+\rho}} B_\nu^2}{\mu\mu_\nu C_\rho^{\frac{2\nu}{1+\rho}} N_t^{\frac{2\rho\nu}{1+\rho}}} + \mathcal{O}(\exp(-C_\delta''' N_t^{(1+\rho\beta-\alpha)/(1+\rho)})). \tag{1.3.18}$$

Theorem 1.3.5 replicate the results of the unbiased i.i.d. case (with $B_\nu = 0$ and $\kappa = \sigma = 1/2$) considered in Section 1.3.2 [57]. Our findings also reproduce the results of Moulines and Bach [96], where they considered the unbiased i.i.d. case (under slightly different assumptions) using the vanilla SG descent, namely, when $C_\rho = 1$ and $\rho = 0$. Moreover, if the function L has C_∇-Lipschitz continuous gradients, then Theorem 1.3.5 implies the bound on the objective function values of L, $\mathbb{E}[L(\theta_t) - L(\theta^*)] \leq C_\nabla \delta_t/2$ by Cauchy–Schwarz's inequality.

Decay of the initial conditions. Note that the positivity of the dependence penalised convexity constant $\mu_\nu = \mu - \mathbb{1}_{\{\rho=0\}} 2D_\nu C_\rho^{-\nu}$ is essential in all terms of (1.3.18). Having $\mu_\nu > 0$ depends solely on the level of dependence D_ν but it is scaled by $C_\rho^{-\nu}$, meaning if D_ν is so large that μ_ν is no longer positive, then we should take C_ρ large enough such that μ_ν becomes positive again; this is illustrated in Chapter 3 for ARCH models [58]. The streaming constant C_ρ contributes positively to all terms in (1.3.18), either directly or though μ_ν.

The last term of (1.3.18) can be seen as the noise term decaying with $\mathcal{O}(N_t^{-(\rho(2\sigma-\beta)+\alpha)/(1+\rho)})$ for $\alpha - \rho\beta \in (1/2, 1)$, e.g., for any $\rho \in [0, 1)$, $\delta_t = \mathcal{O}(N_t^{-2/3})$ when $\alpha = 2/3$, $\beta = 1/3$, and $\sigma = 1/2$. In addition, the noise term is positively affected by large streaming constants C_ρ when $\alpha + \beta < 2\sigma$, which will be expressed as a variance reduction, e.g., see Example 1.3.2 below. In unbiased cases ($B_\nu = 0$) the noise term would also be the asymptotic term.

Behavior for B_ν. The second term of (1.3.18) can be seen as an dependency term as it is determined solely by the level of dependence ν, the bias error B_ν, and the convexity constant μ_ν. It is remarkable that the dependence term is unconnected from the choice of the learning rate (γ_t) but instead by the streaming rate through C_ρ and ρ. The dependence term decay with $\mathcal{O}(N_t^{-2\rho\nu/(1+\rho)})$, which requires ρ positive to decay since $\nu \in (0, \infty)$, e.g., if $\nu = 1/2$, we would need $\rho = 1$ to obtain $\mathcal{O}(N_t^{-1/2})$. It is surprising that Theorem 1.3.5 allows both long-range and short-range dependence. Indeed, long-range dependence leads to slow convergence (slower than $\mathcal{O}(N_t^{-1/2})$) but it will still converge. Obviously, this only matters if $B_\nu \neq 0$. To conclude, by taking $\rho > 0$ and C_ρ large enough to ensure that μ_ν stays positive, then we will converge with $\delta_t = \mathcal{O}(\max\{\mathbb{1}_{\{B_\nu \neq 0\}} N_t^{-2\rho\nu/(1+\rho)}, N_t^{-(\rho(2\sigma-\beta)+\alpha)/(1+\rho)}\})$.

Acceleration by Averaging

Similarly to Section 1.3.2, besides having Assumptions 1.3.6-p to 1.3.8-p to hold for $p = 4$, an additional assumption is needed to control the rest term. Thus, in continuation of Assumption 1.3.8-p with $\sigma_t = C_\sigma n_t^{-\sigma}$ for $\sigma \in [0, 1/2]$, we make the following assumption:

Assumption 1.3.9. *There exists a non-negative self-adjoint operator Σ such that $\forall t \geq 1$, we have $n_t^{2\sigma}\mathbb{E}[\nabla_\theta l_t(\theta^*)\nabla_\theta l_t(\theta^*)^\top] \preceq \Sigma + \Sigma_t$, where Σ_t is a positive symmetric matrix with $Tr(\Sigma_t) = C'_\sigma n_t^{-2\sigma'}$, $C'_\sigma \geq 0$, and $\sigma' \in (0, 1/2]$.*

In the unbiased and independent case, Assumption 1.3.9 is verified with $\sigma = 1/2$ and $C'_\sigma = 0$ [57]. The short-range dependence cases is when $\sigma = 1/2$, whereas, the long-range dependence case is for $\sigma < 1/2$. Moreover, Assumption 1.3.9 allows us to obtain leading term Λ/N_t with $\Lambda = \mathrm{Tr}(\nabla^2_\theta L(\theta^*)^{-1}\Sigma\nabla^2_\theta L(\theta^*)^{-1})$ as in Theorem 1.3.4. To consider the projected average estimate $\bar{\theta}_n$ given in (1.2.7), an additional assumption is needed to avoid calculating the six-order moment. Thus, we make the unnecessary assumption that $(\nabla_\theta l_t)$ is uniformly bounded.[3]

Assumption 1.3.10. *Let $D_\Theta = \inf_{\theta\in\partial\Theta}\|\theta - \theta^*\| > 0$ with $\partial\Theta$ denoting the frontier of Θ. Moreover, there exists $G_\Theta > 0$ such that $\forall t \geq 1$, $\sup_{\theta\in\Theta}\|\nabla_\theta l_t(\theta)\|^2 \leq G^2_\Theta$ a.s.*

Theorem 1.3.6. *Denote $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ with $\bar{\theta}_n$ given by (1.2.7), where (θ_t) follows the recursion in (1.2.5) or (1.2.6). Assume that Assumptions 1.3.6-p to 1.3.8-p for $p = 4$ and Assumption 1.3.9 hold true. Moreover, let (1.3.10) and (1.3.11) hold true. In addition, Assumption 1.3.10 must hold true if (θ_t) follows the recursion in (1.2.6). Suppose $n_t = C_\rho t^\rho$ with $\rho \in [0, 1)$ and $C_\rho \in \mathbb{N}$, such that $\mu_\nu = \mu - \mathbb{1}_{\{\rho=0\}}2D_\nu C_\rho^{-\nu} > 0$. For $\alpha - \rho\beta \in (1/2, 1)$, we have*

$$\bar{\delta}_t^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}}\mathbb{1}_{\{\sigma=1/2\}} + \frac{2^{1/2}\Lambda^{1/2}C_\rho^{\frac{1-2\sigma}{2(1+\rho)}}}{N_t^{\frac{1+2\rho\sigma}{2(1+\rho)}}}\mathbb{1}_{\{\sigma<1/2\}} + \frac{2^{1/2}C_\sigma'^{1/2}C_\rho^{\frac{1-2(\sigma+\sigma')}{2(1+\rho)}}}{\mu N_t^{\frac{1+2\rho(\sigma+\sigma')}{2(1+\rho)}}} \tag{1.3.19}$$

$$+\mathcal{O}\left(\max\left\{N_t^{-\frac{2+\rho(2\sigma+\beta)-\alpha}{2(1+\rho)}}, N_t^{-\frac{\rho(2\sigma-\beta)+\alpha}{1+\rho}}\right\}\right) + \tilde{\mathcal{O}}\left(N_t^{-\frac{\delta+\rho\nu}{2(1+\rho)}}\right) + \mathbb{1}_{\{B_\nu\neq 0\}}\Psi_t, \tag{1.3.20}$$

with $\delta = \mathbb{1}_{\{B_\nu=0\}}(\rho(2\sigma-\beta)+\alpha) + \mathbb{1}_{\{B_\nu\neq 0\}}\min\{\rho(2\sigma-\beta)+\alpha, 2\rho\nu\}$ and Ψ_t given such that

$$\Psi_t = \tilde{\mathcal{O}}\left(\max\left\{N_t^{-\frac{\rho(\sigma+\nu)}{2(1+\rho)}}, N_t^{-\frac{1+\rho(\beta+\nu)-\alpha}{1+\rho}}, N_t^{-\frac{1+2\rho\nu}{2(1+\rho)}}, N_t^{-\frac{\delta/2+\rho\nu}{2(1+\rho)}}, N_t^{-\frac{2\rho\nu}{1+\rho}}\right\}\right).$$

Accelerated decay. Theorem 1.3.6 replicate the results of Theorem 1.3.4 with Λ/N_t as leading term in the unbiased i.i.d. case. Thus, by averaging it is possible to achieve the incorrigible rate of $\mathcal{O}(N_t^{-1})$, e.g., this is always achieved in the unbiased case with $\sigma = 1/2$, even under short-range dependence (i.e., when $\nu \geq 1/2$). Remark that each term in (1.3.19) is a direct consequence of Assumption 1.3.9. Furthermore, all terms of (1.3.19) are independent of the learning rate (γ_t) but the two last terms are dependent on streaming batches through C_ρ and ρ. As in Theorem 1.3.5,

[3] The derivation of the six-order moment can be found in Godichon-Baggioni [55].

the positivity of μ_ν is essential for all terms in (1.3.20) even if it does not appear directly; a long version can be found in Chapter 3. For objectives that lack convexity μ or have high levels of dependence D_ν, we can only ensure convergence by increasing C_ρ, i.e., ensuring positivity of μ_ν; this is illustrated in Chapter 3 for ARCH models [58].

The first term of (1.3.20) decay at the rate $\mathcal{O}(\max\{N_t^{-(2+\rho(\beta+2\sigma)-\alpha)/(1+\rho)}, N_t^{-2(\rho(2\sigma-\beta)+\alpha)/(1+\rho)}\})$, which suggests choosing α, β such that $\alpha + \rho(2\sigma/3 - \beta) = 2/3$, e.g., $\alpha = 2/3$, $\beta = 1/3$ and $\sigma = 1/2$ yields a decay of $\mathcal{O}(N_t^{-4/3})$ for any ρ. Thus, we can robustly achieve $\mathcal{O}(N_t^{-4/3})$ for any streaming rate ρ by setting $\alpha = 2/3$ and $\beta = 1/3$ if $\sigma = 1/2$. In general, the convergence is resilient to any streaming rate ρ by having $\alpha = 2/3$ and $\beta = 2\sigma/3$. But taking $\beta > 0$ would damage the variance reduction effect from having C_ρ large (e.g., see discussion after Theorem 3.3.1). Thus, there is a trade-off between accelerating the convergence by taking $\beta = 2\sigma/3 > 0$ or taking $\beta = 0$ to favor from variance reduction. In practice, an immediate choice would be to take $\beta = 0$, but if the data or model contains a low amount of noise, it can be advantageous to raise β to improve convergence [57]; this is illustrated in Examples 1.3.1 and 1.3.2.

Next, the decay of the second term in (1.3.20) is tricky to interpret in a simple manner as it is a mixture of the learning rate α and β, streaming rate ρ, dependence rate ν, and bias B_ν. Nevertheless, some observations can be made: first, having $\beta = 0$ is beneficial for the decay rate δ in all cases. Second, increasing streaming rate ρ would also increase the decay.

Behavior for B_ν. The influence of B_ν is exclusively contained in Ψ_t, with the exception of the second term of (1.3.20). Also, increasing ρ will always diminish the bad influence of this bias term. Surprisingly, $\Psi_t \to 0$ as $t \to \infty$ for any ν, but long-range dependence is excluded if we wish to obtain the desired rate of $\bar{\delta}_t = \mathcal{O}(N^{-1})$. However, it does not seem to have any major influence in our experiments, e.g., see Example 1.3.2. To conclude, by taking ρ positive and C_ρ large enough to ensure that μ_ν stays positive, then we will converge under long- or short-range dependence with biased gradient estimates.

Example 1.3.2 (Geometric median, continuation of Example 1.3.1)**.** To illustrate our methodology on real-life time-dependent streaming data, we consider some historical hourly weather data[4]. The dataset contains around five years (roughly 45000 data points) of high temporal resolution hourly measurements over various weather attributes, such as temperature, humidity, and air pressure. These measurements are available for thirty US and Canadian cities and six Israeli cities, meaning the dimension $d = 36$. In our study, we consider the hourly temperature measurements, which we filter for monthly and annual seasonality by subtracting the monthly and annual averages.

Figure 1.3 shows the results of the geometric median estimated in the same way as described in Example 1.3.1; but here we compare our estimates to the geometric median estimate calculated by the Weiszfeld's algorithm [148]. Although the (geometric) median is a robust metric (as seen in Example 1.3.1), we see a considerable amount of fluctuations in Figure 1.3, which comes from the time-dependency and the noise in the weather measurements. Figure 1.3a shows that it is essential

[4]The historical hourly weather dataset can be found on https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data.

to use a mini-batch of a certain size to stabilize the optimization. But to achieve reasonable convergence, we need to have increasing streaming batches (i.e., positive streaming rates $\rho > 0$), which we can see in Figures 1.3b to 1.3e. Most surprising is that we can achieve excellent convergence (with a final error of only 10^{-5}) by combining increasing streaming batches with averaging, e.g., see Figure 1.3f with $C_\rho = 64$, $\rho > 0$ and $\beta = 1/3$.

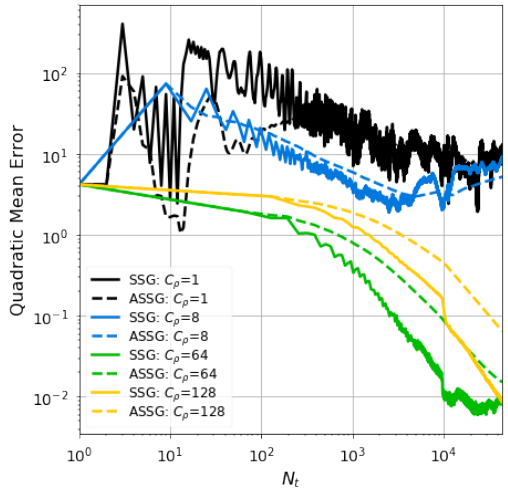
An extended study of stochastic algorithms for time-dependent data can be found in Chapter 4, where we propose AdaVol as an online adaptive recursive estimation routine of GARCH parameters. Here, we make a natural adaptation of the Quasi-Maximum Likelihood (QML) procedure to a streaming setting. This method is based on stochastic algorithms combined with the Variance Targeting Estimation (VTE) technique [49]. However, AdaVol was made before the theory of stochastic algorithms of time-dependent streaming data (with lack of convexity) was shown. This can also be reflected in Chapter 4, as we use VTE to include all models parameters in the projection and thereby achieved convergence. With this new theory in place, a modified AdaVol algorithm could thus be made without VTE by increasing streaming batches to remedy the convexity problems and break the dependency. Thus, some theoretical guarantees could be proved for this modified version. Nevertheless, in Chapter 4, we will see that AdaVol goes beyond this stationary setting, e.g., AdaVol's ability to adapt to time-varying parameters was beneficial in the M6 financial forecast competition, where it recently ranked third among the best probability forecasters.

Summary

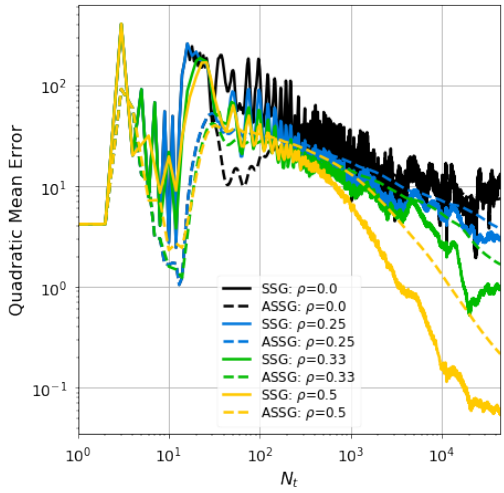
We examined the SO problem in a streaming framework using time-dependent and biased (gradient) estimates. In particular, we explored convergence rates of the SSG and ASSG algorithms in a non-asymptotic manner. The theoretical results formed heuristics that links the level of dependency and convexity to the rest of the model parameters. These heuristics provided new insights into determining learning rates, which can help increase the stability of SG-based methods. Our experiments verified these findings, suggesting using increasing streaming batches for highly dependent data sources. Moreover, in large-scale learning problems with dependence, noisy variables, and lack of convexity, we know now how to accelerate convergence and reduce noise through the learning rate and the treatment pattern of the datasets.

Figure 1.3: Geometric median for various data streams $n_t = C_\rho t^\rho$. See Example 1.3.2 for details.

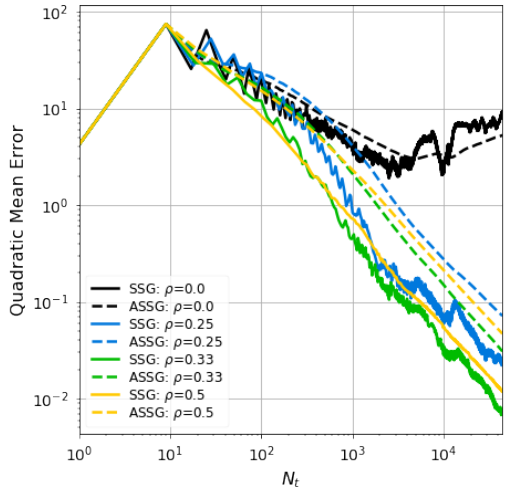
(a) Constant streaming batches, $\rho = 0$, $\beta = 0$



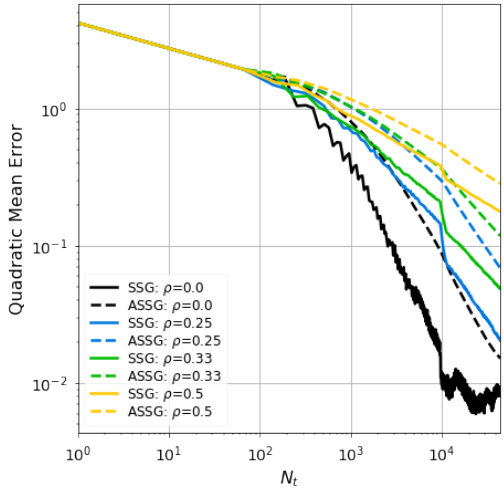
(b) Varying streaming batches, $C_\rho = 1$, $\beta = 0$



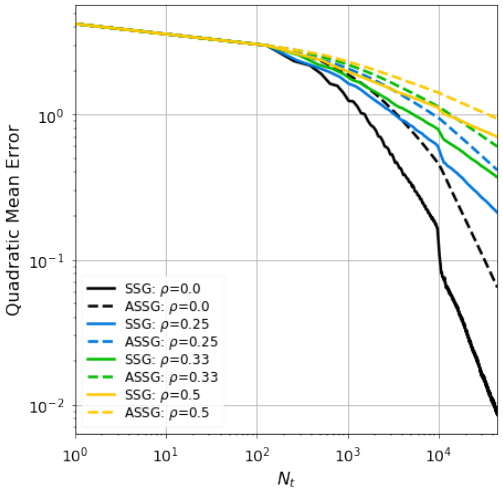
(c) Varying streaming batches, $C_\rho = 8$, $\beta = 0$



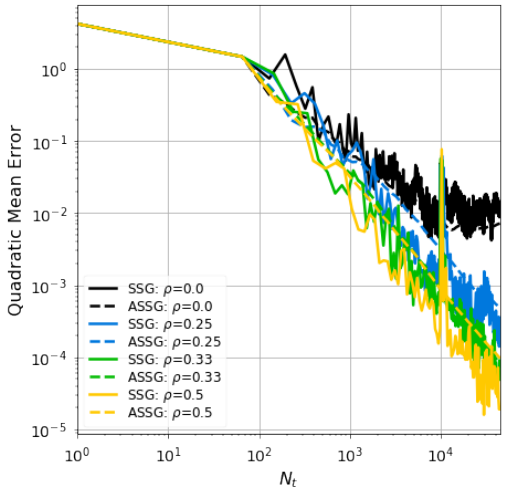
(d) Varying streaming batches, $C_\rho = 64$, $\beta = 0$



(e) Varying streaming batches, $C_\rho = 128$, $\beta = 0$



(f) Varying streaming batches, $C_\rho = 64$, $\beta = 1/3$



Chapter 2: Non-asymptotic Analysis of Stochastic Algorithms for Streaming Data

Abstract

Motivated by the high-frequency data streams continuously generated, real-time learning is becoming increasingly important. These data streams should be processed sequentially with the property that the data stream may change over time. In this streaming setting, we propose techniques for minimizing convex objectives through unbiased estimates of their gradients, commonly referred to as stochastic approximation problems. Our methods rely on stochastic approximation algorithms because of their applicability and computational advantages. The reasoning includes iterate averaging that guarantees optimal statistical efficiency under classical conditions. Our non-asymptotic analysis shows accelerated convergence by selecting the learning rate according to the expected data streams. We show that the average estimate converges optimally and robustly for any data stream rate. In addition, noise reduction can be achieved by processing the data in a specific pattern, which is advantageous for large-scale machine learning problems. These theoretical results are illustrated for various data streams, showing the effectiveness of the proposed algorithms.

keywords: *machine learning, large-scale, stochastic approximation, stochastic optimization, streaming data.*

Contents

2.1 Introduction

Machine learning and artificial intelligence have become an integral part of modern society. This massive utilization of intelligent systems generates an endless sequence of data, many of which come as *streaming data* such as internet traffic data, financial investments, self-driving cars, or sensor data. It requires robust and time-efficient algorithms to analyze and process such data without compromising accuracy. This problem has attracted a lot of attention in the machine learning community [19, 20, 133, 153, 156].

Even after 70 years, Stochastic Approximation (SA) algorithms are still widely used for handling large amounts of data [124]; the most well-known is presumably the Stochastic Gradient (SG) method, which has led to numerous extensions [42, 83, 122, 128, 143, 155]. An essential extension is the Polyak-Ruppert averaging (ASG) proposed by Polyak and Juditsky [118] and Ruppert [129], which guarantees optimal statistical efficiency without jeopardizing the computational cost. Bottou et al. [21] reviews these stochastic algorithms for large-scale machine learning, including noise reduction and second-order methods, among others.

Contributions. A fundamental aspect of this paper is to explore how changing data streams affect these stochastic optimization methods. Our analysis extends the work of Moulines and Bach [96] to a streaming framework. We examine two different kinds of data streams: constant and varying streaming-batches. These data streams includes everything from vanilla SG and ASG, mini-batch SG and ASG, to more exotic learning designs. Our main theoretical contribution is the non-asymptotic analysis of the SG and ASG method in this streaming framework. Our results show a noticeable improvement in convergence rates by having learning rates that adapt to the expected data streams. In particular, we show how to obtain *optimal* convergence rates *robust* to any data streaming rate.

Organization. Section 2.2 presents the streaming framework on which the non-asymptotic

analysis relies. Our convergence results are presented in Sections 2.3 and 2.4, with and without averaging. Both sections includes analysis of unbounded and uniform bounded gradients. These theoretical results are illustrated in Section 2.5 for a variety of data streams. At last, some final remarks are done in Section 2.6.

2.2 Problem Formulation

The objective of the stochastic optimization problem is to minimize functions of the form $L(\theta) = \mathbb{E}[l_t(\theta)]$ with respect to $\theta \in \Theta$, where Θ is a closed convex set in $\mathbb{R}^d$. The minimization of L is achieved without evaluating it directly but by unbiased functions $l_t : \mathbb{R}^d \to \mathbb{R}$. Observe that the principles for biased functions (l_t) are rather different [37, 130]. Let (l_t) constitute a sequence of independent differentiable random functions (possibly non-convex) and their gradients unbiased estimates of $\nabla_\theta L$, e.g., see Nesterov et al. [104] for definitions and properties of such functions. Let us now describe our streaming framework in which we will solve our SA problem: at each time $t \in \mathbb{N}$, a *block* consisting of $n_t \in \mathbb{N}$ random functions $l_t = (l_{t,1}, \dots, l_{t,n_t})$ arrive. These random functions $(l_{t,i})$ can be seen as observations (or random loss functions) depending on the true minimizer θ^* and some underlying noise sequence. To solve this, we introduce the Stochastic Streaming Gradient (SSG) defined as

$$\theta_t = \theta_{t-1} - \frac{\gamma_t}{n_t} \sum_{i=1}^{n_t} \nabla_\theta l_{t,i}(\theta_{t-1}), \quad \theta_0 \in \Theta, \tag{2.2.1}$$

where (γ_t) is a decreasing sequence of positive numbers also referred to as the *learning rate* satisfying $\sum_{i=1}^t \gamma_i = \infty$ and $\sum_{i=1}^t \gamma_i^2 < \infty$ for $t \to \infty$ [124]. In the same way, we introduce the Projected Stochastic Streaming Gradient (PSSG), defined by

$$\theta_t = \mathcal{P}_\Theta \left(\theta_{t-1} - \frac{\gamma_t}{n_t} \sum_{i=1}^{n_t} \nabla_\theta l_{t,i}\left(\theta_{t-1}\right) \right), \quad \theta_0 \in \Theta, \tag{2.2.2}$$

where $\mathcal{P}_\Theta$ denotes the projection onto Θ. The PSSG estimate in (2.2.2) is very convenient for models with conditions on the parameters space, and thereby, requires a projection of the parameters. Next, to guarantee optimal convergence properties [118, 129], we introduce the Averaged Stochastic Streaming Gradient (ASSG), given as

$$\bar{\theta}_t = \frac{1}{N_t} \sum_{i=0}^{t-1} n_{i+1} \theta_i, \quad \bar{\theta}_0 = 0, \tag{2.2.3}$$

where (θ_t) follow (2.2.1) and N_t denotes the accumulated sum of observations, $\sum_{i=1}^t n_i$. Similarly, we define the PASSG estimate as when $(\bar{\theta}_t)$ (in (2.2.3)) is derived using (2.2.2). Practically, as we handle data sequentially, we will make use of the rewritten formula: $\bar{\theta}_t = (N_{t-1}/N_t)\bar{\theta}_{t-1} + (n_t/N_t)\theta_{t-1}$ with $\bar{\theta}_0 = 0$.

2.2.1 Quasi-strong Convex and Lipschitz Smooth Objectives

Following Moulines and Bach [96], Sridharan et al. [135], we make the following assumptions about the objective function L: assume that $\theta^* \in \Theta$ is the unique global minimizer of L with $\nabla_\theta L(\theta^*) = 0$. Also, let L be μ-quasi-strong convex [80, 99], that is, the exists $\mu > 0$ such that $\forall \theta \in \Theta$ the following inequality holds,

$$L(\theta^*) \geq L(\theta) + \langle \nabla_\theta L(\theta), \theta^* - \theta \rangle + \frac{\mu}{2}\|\theta^* - \theta\|^2. \tag{2.2.4}$$

Teo et al. [141] provides a comprehensive record of various convex functions L used in machine learning applications. Milder degrees of convexity have been studied by, e.g., Karimi et al. [80], which studied stochastic gradient methods under the Polyak-Łojasiewicz condition [93, 117], or Gadat and Panloup [50], which studied the Ruppert-Polyak averaging estimate under some Kurdyka-Łojasiewicz-type condition [86, 93]. Next, let the function $\nabla_\theta L$ be C_∇-Lipschitz continuous, i.e., there exists $C_\nabla > 0$ such that $\forall \theta, \theta' \in \Theta$,

$$\|\nabla_\theta L(\theta) - \nabla_\theta L(\theta')\| \leq C_\nabla \|\theta - \theta'\|. \tag{2.2.5}$$

Furthermore, for the averaging estimate in (2.2.3), we need the function L to be twice differentiable with C_δ-Lipschitz continuous Hessian operator $\nabla^2_\theta L$, meaning, there exists $C_\delta > 0$ such that $\forall \theta, \theta' \in \Theta$,

$$\|\nabla^2_\theta L(\theta) - \nabla^2_\theta L(\theta')\| \leq C_\delta \|\theta - \theta'\|. \tag{2.2.6}$$

Note that (2.2.5) and (2.2.6) only needs to hold for $\theta' = \theta^*$.

2.3 Stochastic Streaming Gradients

This section considers the SSG and PSSG methods with streaming batches arriving in constant and varying streams. Our aim is to provide bounds on the quadratic mean $\mathbb{E}[\|\theta_t - \theta^*\|^2]$, which depends explicitly upon the problem's parameters. In order to do this, we assume the following about the function $l_{t,i}$ for each $t \in \mathbb{N}$ with $i = 1, \ldots, n_t$:

Assumption 2.3.1. *The random variable $\nabla_\theta l_{t,i}(\theta)$ is square-integrable and $\forall \theta \in \Theta$, $\mathbb{E}[\nabla_\theta l_{t,i}(\theta)] = \nabla_\theta L(\theta)$.*

Assumption 2.3.2-p (C_l-expected smoothness)**.** *For $p \geq 1$, there exists $C_l > 0$ such that $\forall \theta, \theta' \in \Theta$, $\mathbb{E}[\|\nabla_\theta l_{t,i}(\theta) - \nabla_\theta l_{t,i}(\theta')\|^p] \leq C_l^p \mathbb{E}[\|\theta - \theta'\|^p]$.*

Assumption 2.3.3-p (σ-gradient noise)**.** *For $p \geq 1$, there exists $\sigma > 0$ such that $\mathbb{E}[\|\nabla_\theta l_{t,i}(\theta^*)\|^p] \leq \sigma^p$.*

These assumptions are modified versions of the standard assumptions for stochastic approximations [13, 87] as they hold for any $i = 1, \ldots, n_t$. Note that Assumption 2.3.2-p only needs to

hold for $\theta' = \theta^*$. By the smoothness assumption (Assumption 2.3.2-p), we avoid the unfavorable uniformly bounded gradients assumption, which is too restrictive and only holds for a few losses. Assumption 2.3.3-p is a weak assumption that should be seen as an assumption on Θ rather than on $(l_{t,i})$. For SSG and PSSG, we only need Assumptions 2.3.2-p and 2.3.3-p to hold for $p = 2$, whereas, for ASSG and PASSG, we need $p = 4$ in order to bound the fourth-order moment. Our framework include classic examples: stochastic approximation (Robbins-Monro setting [124]) and learning from i.i.d. data, such as linear regression, logistic regression, general ridge regressions and quantile regression, p-means, and softmax regression, under regularity conditions [33, 136]. In the following theorem, we derive an explicit upper bound on the t-th estimate of (2.2.1) and (2.2.2) for any learning rate (γ_t) using classical techniques from stochastic approximations [13, 87].

Theorem 2.3.1 (SSG/PSSG). *Denote $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ for some $\delta_0 \geq 0$, where (θ_t) follows (2.2.1) or (2.2.2). Under Assumption 2.3.1, Assumptions 2.3.2-p and 2.3.3-p with $p = 2$, we have for any learning rate (γ_t) that*

$$\delta_t \leq \exp\left(-\mu \sum_{i=t/2}^{t} \gamma_i\right) \pi_t^\delta + \frac{2\sigma^2}{\mu} \max_{t/2 \leq i \leq t} \frac{\gamma_i}{n_i}, \tag{2.3.7}$$

with $\pi_t^\delta = \exp(4C_l^2 \sum_{i=1}^t \gamma_i^2/n_i) \exp(2C_\nabla^2 \sum_{i=1}^t \mathbb{1}_{\{n_i > 1\}} \gamma_i^2)(\delta_0 + 2\sigma^2/C_l^2)$.

Sketch of proof. Under Assumption 2.3.1, Assumptions 2.3.2-p and 2.3.3-p with $p = 2$, we derive from (2.2.1) that (δ_t) satisfies the recursive relation

$$\delta_t \leq [1 - 2\mu\gamma_t + (2C_l^2 + (n_t - 1)C_\nabla^2) n_t^{-1} \gamma_t^2] \delta_{t-1} + 2\sigma^2 n_t^{-1} \gamma_t^2, \tag{2.3.8}$$

for any (n_t) and (γ_t) fulfilling the conditions imposed on the learning rate [124]. This recursive relation is then bounded in a non-asymptotic manner using Proposition B.1.5 in Chapter B. Bounding the projected estimate in (2.2.2) follows directly from the fact that $\mathbb{E}[\|\mathcal{P}_\Theta(\theta) - \theta^*\|^2] \leq \mathbb{E}[\|\theta - \theta^*\|^2]$, $\forall \theta \in \Theta$ [157].

Related work. When $n_t = 1$ in (2.3.7), we obtain the usual SG method studied in Moulines and Bach [96]. Similarly, Theorem 2.3.1 provides an upper bound on the function values, $\mathbb{E}[L(\theta_t) - L(\theta^*)] \leq C_l \delta_t / 2$; this follows by Cauchy-Schwarz inequality and Assumption 2.3.2-p.

Natural decay imposed by Robbins and Monro [124]. The learning rate (γ_t) should satisfy the following requirements: $\sum_{i=1}^t \gamma_i = \infty$ and $\sum_{i=1}^t \gamma_i^2 / n_i \leq \sum_{i=1}^t \gamma_i^2 < \infty$ for $t \to \infty$. These conditions directly imply that $\pi_t^\delta < \infty$ as $t \to \infty$. Thus, our attention is on reducing the noise term $\max_{t/2 \leq i \leq t} \gamma_i / n_i$ without damaging the natural decay of the sub-exponential term $\exp(-\mu \sum_{i=t/2}^t \gamma_i)$. In particular, this non-asymptotic bound shows convergence in quadratic mean for any learning rate, fulfilling these conditions. In addition, the scaling with (n_t) in the noise term shows an apparent variance reduction when we increase the streaming batches (n_t).

Throughout this paper, we will consider learning rates on the form $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ with hyper-parameters $C_\gamma > 0$, $\beta \in [0, 1]$, and α chosen accordingly to the expected streaming batches denoted

by n_t. We start by considering constant streaming batches (i.e., mini-batch SG) where n_t follows the constant streaming batch size $C_\rho \in \mathbb{N}$:

Corollary 2.3.1 (SSG/PSSG, constant streaming batches)**.** *Denote $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ for some $\delta_0 \geq 0$, where (θ_t) follows (2.2.1) or (2.2.2). Suppose $\gamma_t = C_\gamma C_\rho^\beta t^{-\alpha}$ such that $\alpha \in (1/2, 1)$. Under Assumption 2.3.1, Assumptions 2.3.2-p and 2.3.3-p with $p = 2$, we have*

$$\delta_t \leq \exp\left(-\frac{\mu C_\gamma N_t^{1-\alpha}}{2^{1-\alpha} C_\rho^{1-\alpha-\beta}}\right) \pi_\infty^c + \frac{2^{1+\alpha}\sigma^2 C_\gamma}{\mu C_\rho^{1-\alpha-\beta} N_t^\alpha}, \tag{2.3.9}$$

where $\pi_\infty^c = \exp(4\alpha C_\gamma^2(2C_l^2 + C_\rho \mathbb{1}_{\{C_\rho > 1\}} C_\nabla^2)/(2\alpha - 1)C_\rho^{1-2\beta})(\delta_0 + 2\sigma^2/C_l^2)$ is a finite constant.

Decay of the initial conditions. The bound in Corollary 2.3.1 depends on the initial condition $\delta_0 = \|\theta_0 - \theta^*\|^2$ and the variance σ^2 in the noise term. The initial condition δ_0 vanish sub-exponentially fast for $\alpha \in (1/2, 1)$. Thus, the asymptotic term is $2^{1+\alpha}\sigma^2 C_\gamma / \mu C_\rho^{1-\alpha-\beta} N_t^\alpha$, i.e., $\delta_t = \mathcal{O}(N_t^{-\alpha})$. Moreover, the bound in (2.3.9) is optimal (up to some constants) for quadratic functions $(l_{t,i})$, since the deterministic recursion in (2.3.8) would be with equality. It is worth noting that if $C_\gamma C_l$ or $C_\gamma C_\nabla$ is chosen too large, they may produce a large π_∞^c constant. In addition, π_∞^c is positively affected by C_ρ when $\beta < 1/2$. Obviously, the hyper-parameter β only comes into play if the streaming batch size is larger than one, i.e., $C_\rho > 1$. Nonetheless, the effect of π_∞^c will decrease exponentially fast due to the sub-exponentially decaying factor in front.

Variance reduction. The asymptotic term is divided by $C_\rho^{1-\alpha-\beta}$, implying we could achieve variance reduction by taking $\alpha + \beta \leq 1$ when C_ρ is large. Taking a large streaming batch size, e.g., $C_\rho = t$, one accelerates the vanilla SG convergence rate to $\mathcal{O}(N_t^{1-\beta})$. However, this large streaming batch size would be unsuitable in practice, and it would mean that we would take few steps until convergence is achieved.

The *safe* choice of having $\beta = 0$ functions well for the SSG method for any streaming batch size C_ρ, but fixed-sized streaming batches are not the most realistic streaming setting. These streaming batches are far more likely to vary in size depending on the data streams. For the sake of simplicity, we consider varying streaming batches where n_t are on the form $C_\rho t^\rho$ with $C_\rho \geq 1$ and $\rho \in (-1, 1)$ such that $n_t \geq 1$ for all t. We will refer to ρ as the *streaming rate*. For the convenience of notation, let $\tilde{\rho} = \rho \mathbb{1}_{\{\rho \geq 0\}}$.

Corollary 2.3.2 (SSG/PSSG, varying streaming batches)**.** *Denote $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ for some $\delta_0 \geq 0$, where (θ_t) follows (2.2.1) or (2.2.2). Suppose $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ where $n_t = C_\rho t^\rho$ with $C_\rho \geq 1$ and $\rho \in (-1, 1)$, such that $\alpha - \beta\tilde{\rho} \in (1/2, 1)$. Under Assumption 2.3.1, Assumptions 2.3.2-p and 2.3.3-p with $p = 2$, we have*

$$\delta_t \leq \exp\left(-\frac{\mu C_\gamma N_t^{1-\phi}}{2^{(2+\rho)(1-\phi)} C_\rho^{1-\beta-\phi}}\right) \pi_\infty^v + \frac{2^{1+(2+\rho)\phi}\sigma^2 C_\gamma}{\mu C_\rho^{(1-\beta)\mathbb{1}_{\{\rho \geq 0\}} - \phi} N_t^\phi}, \tag{2.3.10}$$

where $\phi = ((1-\beta)\tilde{\rho} + \alpha)/(1 + \tilde{\rho})$ and $\pi_\infty^v = \exp(4(\alpha - \beta\tilde{\rho})C_\gamma^2 C_\rho^{2\beta}(2C_l^2 + C_\nabla^2)/(2(\alpha - \beta\tilde{\rho}) - 1))(\delta_0 +$

$2\sigma^2/C_l^2)$ *is a finite constant.*

Decay of the initial conditions. As mentioned for Corollary 2.3.1, the condition of having $\alpha - \beta\tilde{\rho} \in (1/2, 1)$ is a natural restriction from Robbins and Monro [124], which relaxes the usual condition of having $\alpha \in (1/2, 1)$ for ρ non-negative. For $\rho \in (-1, 1/2)$, setting $\alpha = 2/3$ and $\beta = 1/3$ would give same decay rate, $\delta_t = \mathcal{O}(N_t^{-2/3})$ as we saw for Corollary 2.3.1 when $\alpha = 2/3$. However, accelerated convergence could be achieved by, e.g., setting $\alpha = 1$ and $\beta = 1/2$ for streaming rate $\rho \in (0, 1)$, giving us $\delta_t = \mathcal{O}(N_t^{-(1+\rho/2)/(1+\rho)})$.

Variance reduction. Similarly to Corollary 2.3.1, the sub-exponential and asymptotic term is scaled by $C_\rho^{1-\beta-\phi}$ for $\rho \geq 0$, implying we should take $\alpha + \beta \leq 1$ to obtain variance reduction. These conclusions will change when we consider the averaging estimate in Section 2.4.

The reasoning in Corollary 2.3.2 could be expanded to include *random* streaming batches where n_t is given such that $C_L t^{\rho_L} \leq n_t \leq C_H t^{\rho_H}$ with $\rho_L, \rho_H \in (-1, 1)$ and $C_L, C_H \geq 1$. This yields the modified rate $\phi' = ((1-\beta)\rho_L + \alpha)/(1 + \rho_H)$; nevertheless, we will leave the proof to the reader.

2.4 Averaged Stochastic Streaming Gradients

In what follows, we consider the averaging estimate $(\bar{\theta}_n)$ given in (2.2.3) derived with use of (θ_t) from (2.2.1) (Section 2.4.1) or (2.2.2) (Section 2.4.2). Besides having Assumptions 2.3.2-p and 2.3.3-p to hold for $p = 4$, an additional assumption is needed for bounding the *rest* term of the averaging estimate.

Assumption 2.4.1. *There exists a non-negative self-adjoint operator* Σ *such that* $\mathbb{E}[\nabla_\theta l_{t,i}(\theta^*)\nabla_\theta l_{t,i}(\theta^*)^\top] \preceq \Sigma$.

Note that the operator Σ always exists when σ is finite for order $p = 4$ in Assumption 2.3.3-p.

2.4.1 Unbounded Gradients

As in Section 2.3, we conduct a general study for any learning rate (γ_t) when applying the Polyak-Ruppert averaging estimate from (2.2.3):

Theorem 2.4.1 (ASSG). *Denote* $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ *with* $(\bar{\theta}_t)$ *given by* (2.2.3) *using* (θ_t) *from* (2.2.1). *Under Assumption 2.3.1, Assumptions 2.3.2-p and 2.3.3-p with* $p = 4$, *and Assumption 2.4.1, we have for any learning rate* (γ_t) *that*

$$\begin{aligned}\bar{\delta}_t^{1/2} \leq & \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{1}{\mu N_t}\sum_{i=1}^{t-1}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right|\delta_i^{1/2} + \frac{n_t}{\mu\gamma_t N_t}\delta_t^{1/2} + \frac{n_1}{\mu N_t}\left(\frac{1}{\gamma_1} + C_l\right)\delta_0^{1/2} \\ & + \frac{C_l}{\mu N_t}\left(\sum_{i=1}^{t-1} n_{i+1}\delta_i\right)^{1/2} + \frac{C_\delta}{\mu N_t}\sum_{i=0}^{t-1} n_{i+1}\Delta_i^{1/2}, \end{aligned} \tag{2.4.11}$$

where $\Lambda = \mathrm{Tr}(\nabla_\theta^2 L(\theta^*)^{-1}\Sigma\nabla_\theta^2 L(\theta^*)^{-1})$ *and* $\Delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^4]$ *for some* $\Delta_0 \geq 0$.

As noticed in Polyak and Juditsky [118], the leading term Λ/N_t achieves the Cramer-Rao bound [50, 97]. Note that the leading term Λ/N_t is invariant of the learning rate (γ_t). Moreover, this bound of $\mathcal{O}(N_t^{-1})$ is achieved without inverting the Hessian. Next, the processes (δ_t) and (Δ_t) can be bounded by the recursive relations in (2.3.7) and (2.7.21). There are no sub-exponential decaying terms for the initial conditions in Theorem 2.4.1, which is a common problem for averaging. However, as mentioned previously, we are more interested in advancing the decay of the asymptotic terms. To ease notation, we make use of the functions $\psi_x^y(t) : \mathbb{R} \to \mathbb{R}$, given as

$$\psi_x^y(t) = \begin{cases} t^{(1-x)/(1+y)}/(1-x) & \text{if } x < 1, \\ (1+y)\log(t) & \text{if } x = 1, \\ x/(x-1) & \text{if } x > 1, \end{cases}$$

with $y \in \mathbb{R}_+$, such that $\sum_{i=1}^t i^{-x} \leq \psi_x^0(t)$ for any $x \in \mathbb{R}_+$. Note that $\psi_x^y(t)/t = \mathcal{O}(t^{-(x+y)/(1+y)})$ if $x < 1$, $\psi_x^y(t)/t = \mathcal{O}(\log(t)t^{-1})$ if $x = 1$, and $\psi_x^y(t)/t = \mathcal{O}(t^{-1})$ if $x > 1$. Hence, for any $x, y \in \mathbb{R}_+$, $\psi_x^y(t)/t = \tilde{\mathcal{O}}(t^{-(x+y)/(1+y)})$, where the $\tilde{\mathcal{O}}(\cdot)$ notation hides logarithmic factors.

Corollary 2.4.1 (ASSG, constant streaming batches)**.** *Denote $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ with $(\bar{\theta}_t)$ given by (2.2.3) using (θ_t) from (2.2.1). Suppose $\gamma_t = C_\gamma C_\rho^\beta t^{-\alpha}$ such that $\alpha \in (1/2, 1)$. Under Assumption 2.3.1, Assumptions 2.3.2-p and 2.3.3-p with $p = 4$, and Assumption 2.4.1, we have*

$$\begin{aligned} \bar{\delta}_t^{1/2} \leq & \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{6\sigma C_\rho^{(1-\alpha-\beta)/2}}{\mu^{3/2} C_\gamma^{1/2} N_t^{1-\alpha/2}} + \frac{2^\alpha 6 C_\delta \sigma^2 C_\gamma}{\mu^2 C_\rho^{1-\alpha-\beta} N_t^\alpha} + \frac{2 C_l \sigma C_\gamma^{1/2}}{\mu^{3/2} C_\rho^{(1-\alpha-\beta)/2} N_t^{(1+\alpha)/2}} + \frac{C_\rho \Gamma_c}{\mu N_t} \\ & + \frac{C_\rho^{2-\alpha-\beta} \sqrt{\pi_\infty^c} A_\infty^c}{\mu C_\gamma N_t^{2-\alpha}} + \frac{(6 + 7\mathbb{1}_{\{C_\rho > 1\}}) 2^{3\alpha/2} C_\delta \sigma^2 C_\gamma^{3/2} C_\rho^{3\beta/2} \psi_{3\alpha/2}^0(N_t/C_\rho)}{\mu^{3/2} N_t}, \end{aligned}$$

with Γ_c given by $(1/C_\gamma C_\rho^\beta + C_l)\delta_0^{1/2} + C_l\sqrt{\pi_\infty^c A_\infty^c / C_\rho} + \sqrt{\pi_\infty^c} A_\infty^c / C_\gamma C_\rho^\beta + C_\delta \sqrt{\Pi_\infty^c} A_\infty^c$, consisting of the finite constants π_∞^c, Π_∞^c and A_∞^c, that only depends on μ, δ_0, Δ_0, C_l, σ, C_∇, C_δ, C_γ, C_ρ, β and α.

Accelerated decay the initial conditions. By averaging, we have increased the rate of convergence from $\mathcal{O}(N_t^{-\alpha})$ to the optimal rate $\mathcal{O}(N_t^{-1})$. The two subsequent terms are the main remaining terms decaying at the rate $\mathcal{O}(N_t^{\alpha-2})$ and $\mathcal{O}(N_t^{-2\alpha})$, which suggests setting $\alpha = 2/3$ would be *optimal*. The remaining terms are negligible. Next, it is worth noting that having $\alpha + \beta = 1$ in Corollary 2.4.1, we would give no impact in the main remaining terms from the streaming batch size C_ρ. Moreover, taking $\alpha = 2/3$ and $\beta \leq 1/3$ would be an *optimal* choice of hyper-parameters such that the streaming batch size C_ρ have a positive or no impact. At last, as we do not rely on sub-exponentially decaying terms, we need to be more careful when picking our hyper-parameters, e.g., taking $C_\gamma C_l$ too large may cause Γ_c to be significant. Nevertheless, the term consisting of Γ_c decay at a rate of at least $\mathcal{O}(N_t^{-2})$.

Corollary 2.4.2 (ASSG, varying streaming batches)**.** *Denote $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ with $(\bar{\theta}_t)$ given by (2.2.3) using (θ_t) from (2.2.1). Suppose $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ where $n_t = C_\rho t^\rho$ with $C_\rho \geq 1$ and $\rho \in (-1, 1)$,*

such that $\alpha - \beta\tilde{\rho} \in (1/2, 1)$. *Under Assumption 2.3.1, Assumptions 2.3.2-p and 2.3.3-p with* $p = 4$, *and Assumption 2.4.1, we have*

$$\bar{\delta}_t^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{2^{3+\phi(1+\tilde{\rho})}\sigma C_\rho^{(1-\phi-\beta)/2\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}C_\gamma^{1/2}N_t^{1-\phi/2}} + \frac{2^{(1+\phi)(1+\tilde{\rho})-2}C_\delta\sigma^2 C_\gamma}{\mu^2 C_\rho^{1-\phi-\beta}N_t^{\phi}} + \frac{2^{\phi(1+\tilde{\rho})/2}C_l\sigma C_\gamma^{1/2}}{\mu^{3/2}C_\rho^{(1-\phi-\beta)/2\mathbb{1}_{\{\rho\geq 0\}}}N_t^{(1+\phi)/2}}$$
$$+ \frac{C_\rho\Gamma_v}{\mu N_t} + \frac{C_\rho^{2-\phi-\beta}\sqrt{\pi_\infty^v}A_\infty^v}{\mu C_\gamma N_t^{2-\phi}} + \frac{2^{3(1+\phi)(1+\tilde{\rho})/2}C_\delta\sigma^2 C_\gamma^{3/2}C_\rho^{1+3\beta/2}\psi_{3(\alpha-\beta\tilde{\rho})/2}^{\tilde{\rho}}(N_t/C_\rho)}{\mu^{3/2}C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}}N_t},$$

with Γ_v *given by* $(1/C_\gamma C_\rho^\beta + C_l)\delta_0^{1/2} + 2^{\tilde{\rho}}C_l\sqrt{\pi_\infty^v A_\infty^v/C_\rho} + 2\sqrt{\pi_\infty^v}A_\infty^v/C_\gamma C_\rho^\beta + 2^{\tilde{\rho}}C_\delta\sqrt{\Pi_\infty^v}A_\infty^v$, *consisting of the finite constants* π_∞^v, Π_∞^v *and* A_∞^v, *that only depends on* μ, δ_0, Δ_0, C_l, σ, C_∇, C_δ, C_γ, C_ρ, β *and* α.

Robustness towards streaming rates ρ**:** Following the arguments above, the two main remainder terms reveal that $\phi = 2/3 \Leftrightarrow \alpha - \beta\tilde{\rho} = (2-\tilde{\rho})/3$, e.g., by setting $\beta = 0$, we should pick $\alpha = (2-\tilde{\rho})/3$. Likewise, if $\rho = 0$, we yield the same conclusion as in Corollary 2.4.1, namely $\alpha = 2/3$. However, these hyper-parameter choices are not resilient against any arrival schedule ρ. Nonetheless, we can *robustly* achieve $\phi = 2/3$ for any $\rho \in (-1, 1)$ by setting $\alpha = 2/3$ and $\beta = 1/3$. In other words, we can achieve *optimal* convergence for any data stream by having $\alpha = 2/3$ and $\beta = 1/3$.

2.4.2 Bounded Gradients

In what follows, we consider the averaging estimate $\bar{\theta}_n$ given in (2.2.3) but with the use of the projected estimate PSSG from (2.2.2). To avoid calculating the six-order moment, we make the unnecessary assumption that $\|\nabla_\theta l_{t,i}(\theta)\|$ is uniformly bounded for any $\theta \in \Theta$; the derivation of the six-order moment can be found in Godichon-Baggioni [55].

Assumption 2.4.2. *Let* $d_{\min} = \inf_{\theta\in\partial\Theta}\|\theta - \theta^*\| > 0$ *with* $\partial\Theta$ *denoting the frontier of* Θ. *Moreover, there exists* $G_\Theta > 0$ *such that* $\forall t \geq 1$, $\sup_{\theta\in\Theta}\|\nabla_\theta l_{t,i}(\theta)\|^2 \leq G_\Theta^2$ *a.s., with* $i = 1, \ldots, n_t$.

Corollary 2.4.3 (PASSG, constant streaming batches)**.** *Denote* $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ *with* $(\bar{\theta}_t)$ *given by (2.2.3) using* (θ_t) *from (2.2.2). Suppose* $\gamma_t = C_\gamma C_\rho^\beta t^{-\alpha}$ *such that* $\alpha \in (1/2, 1)$. *Under Assumption 2.3.1, Assumptions 2.3.2-p and 2.3.3-p with* $p = 4$, *Assumptions 2.4.1 and 2.4.2, we have*

$$\bar{\delta}_t^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{6\sigma C_\rho^{(1-\alpha-\beta)/2}}{\mu^{3/2}C_\gamma^{1/2}N_t^{1-\alpha/2}} + \frac{2^\alpha 6 C_\delta\sigma^2 C_\gamma}{\mu^2 C_\rho^{1-\alpha-\beta}N_t^{\alpha}} + \frac{2C_l\sigma C_\gamma^{1/2}}{\mu^{3/2}C_\rho^{(1-\alpha-\beta)/2}N_t^{(1+\alpha)/2}} + \frac{C_\rho\Gamma_c}{\mu N_t}$$
$$+ \frac{C_\rho^{2-\alpha-\beta}\sqrt{\pi_\infty^c}A_\infty^c}{\mu C_\gamma N_t^{2-\alpha}} + \frac{(6 + 7\mathbb{1}_{\{C_\rho>1\}})2^{3\alpha/2}C_\delta'\sigma^2 C_\gamma^{3/2}C_\rho^{3\beta/2}\psi_{3\alpha/2}^0(N_t/C_\rho)}{\mu^{3/2}N_t},$$

with $C_\delta' = C_\delta + 2^2 G_\Theta/d_{\min}^2$ *and* Γ_c *given by* $(1/C_\gamma C_\rho^\beta + C_l)\delta_0^{1/2} + C_l\sqrt{\pi_\infty^c A_\infty^c/C_\rho} + \sqrt{\pi_\infty^c}A_\infty^c/C_\gamma C_\rho^\beta + C_\delta\sqrt{\Pi_\infty^c}A_\infty^c$, *consisting of the finite constants* π_∞^c, Π_∞^c *and* A_∞^c, *that only depends on* μ, δ_0, Δ_0, C_l, σ, C_∇, C_δ, C_γ, C_ρ, β *and* α.

Corollary 2.4.4 (PASSG, varying streaming batches)**.** *Denote* $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ *with* $(\bar{\theta}_t)$ *given by* (2.2.3) *using* (θ_t) *from* (2.2.2)*. Suppose* $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ *where* $n_t = C_\rho t^\rho$ *with* $C_\rho \geq 1$ *and* $\rho \in (-1, 1)$*, such that* $\alpha - \beta\tilde{\rho} \in (1/2, 1)$*. Under Assumption 2.3.1, Assumptions 2.3.2-p and 2.3.3-p with* $p = 4$*, Assumptions 2.4.1 and 2.4.2, we have*

$$\bar{\delta}_t^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{2^{3+\phi(1+\tilde{\rho})}\sigma C_\rho^{(1-\phi-\beta)/2\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}C_\gamma^{1/2}N_t^{1-\phi/2}} + \frac{2^{(1+\phi)(1+\tilde{\rho})-2}C_\delta\sigma^2 C_\gamma}{\mu^2 C_\rho^{1-\phi-\beta}N_t^\phi} + \frac{2^{\phi(1+\tilde{\rho})/2}C_l\sigma C_\gamma^{1/2}}{\mu^{3/2}C_\rho^{(1-\phi-\beta)/2\mathbb{1}_{\{\rho\geq 0\}}}N_t^{(1+\phi)/2}}$$
$$+\frac{C_\rho\Gamma_v}{\mu N_t} + \frac{C_\rho^{2-\phi-\beta}\sqrt{\pi_\infty^v}A_\infty^v}{\mu C_\gamma N_t^{2-\phi}} + \frac{2^{3(1+\phi)(1+\tilde{\rho})/2}C'_\delta\sigma^2 C_\gamma^{3/2}C_\rho^{1+3\beta/2}\psi^{\tilde{\rho}}_{3(\alpha-\beta\tilde{\rho})/2}(N_t/C_\rho)}{\mu^{3/2}C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}}N_t},$$

with $C'_\delta = C_\delta + 2^2 G_\Theta/d_{\min}^2$ *and* Γ_v *given by* $(1/C_\gamma C_\rho^\beta + C_l)\delta_0^{1/2} + 2\tilde{\rho}C_l\sqrt{\pi_\infty^v A_\infty^v/C_\rho} + 2\sqrt{\pi_\infty^v}A_\infty^v/C_\gamma C_\rho^\beta + 2\tilde{\rho}C_\delta\sqrt{\Pi_\infty^v}A_\infty^v$*, consisting of the finite constants* π_∞^v*,* Π_∞^v *and* A_∞^v*, that only depends on* μ*,* δ_0*,* Δ_0*,* C_l*,* σ*,* C_∇*,* C_δ*,* C_γ*,* C_ρ*,* β *and* α*.*

2.5 Experiments

In this section, we demonstrate the theoretical results presented in Sections 2.3 and 2.4 for various data streams. In Section 2.5.1, we illustrate the unbounded gradient case (Sections 2.3 and 2.4.1) using linear regression. Where in Section 2.5.2, we present the bounded gradient case (Sections 2.3 and 2.4.2) by considering the geometric median. To measure the performance, we use the quadratic mean error of the parameter estimates over 100 replications, given by $(\mathbb{E}[\|\theta_{N_t} - \theta^*\|^2])_{t\geq 1}$. Note that averaging over several iterations gives a reduction in variability, which mainly benefits the SSG and PSSG.

2.5.1 Linear Regression

Consider the linear regression defined by $y_t = X_t^T\theta + \epsilon_t$, where $X_t \in \mathbb{R}^d$ is a random features vector, $\theta \in \mathbb{R}^d$ is the parameters vector, and ϵ_t is a random variable with zero mean, independent from X_t. Moreover, $(X_t, \epsilon_t)_{t\geq 1}$ are independent and identically distributed. Thus, θ^* is the minimizer of $L(\theta) = \mathbb{E}[(y_t - X_t^T\theta)^2]$. In this example, we fix $d = 10$, set $\theta = (-4, -3, 2, 1, 0, 1, 2, 3, 4, 5)^T \in \mathbb{R}^{10}$, and let (X_t) and (ϵ_t) be standard Gaussian. It is well-known that C_γ can substantially impact convergence; when C_γ is too large, instability can occur, leading to an explosion during the first iterations. If C_γ is too small, the convergence can become very slow and destroy the desired rate α. To focus on the various data streams, we set $C_\gamma = 1/2$ and $\alpha = 2/3$.

In Figure 2.1a, we consider constant data streams to illustrate the results in Corollaries 2.3.1 and 2.4.1. The figures show a solid decay rate proportional to $\alpha = 2/3$ for any streaming batch size $C_\rho \in \{1, 8, 64, 128\}$ with $\beta = 0$, as shown in Corollary 2.3.1. In addition, we see an acceleration in decay by averaging, as explained in Corollary 2.4.1. Both methods show a noticeable reduction in variance when C_ρ increases which are particularly beneficial in the beginning. Moreover, as mentioned in Remark 2.7.1, the *stationary* phase may also commence earlier when we raise the

streaming batch size C_ρ. Next, in Figures 2.1b to 2.1e, we vary the streaming rate ρ for streaming batch sizes $C_\rho = 1$, 8, 64, and 128, respectively, with $\beta = 0$. These figures shows an increase in decay of the SSG when the streaming rate ρ increase. Despite this, we still achieve better convergence for the ASSG method, which seems more immune to the different choices of streaming rate ρ, e.g., see the discussion after Corollary 2.4.2. We know this from Corollary 2.3.2, as $\phi = (\tilde{\rho} + \alpha)/(1 + \tilde{\rho}) \geq \alpha$ for $\beta = 0$. In addition, we see that C_ρ has a positive effect on the noise (i.e, variance reduction), but if C_ρ becomes too large, it may slow down convergence (as seen in Figure 2.1e). Alternatively, we could think around the problem in another way: how can we choose α and β such that we have *optimal* decay of $\phi = 2/3$ for any ρ. In other words, for any arrival schedule that may occur, how should we choose our hyper-parameters such that we achieve optimal decay of $\phi = 2/3$. As discussed after Corollary 2.4.2, one example of this could be achieved by setting $\alpha = 2/3$ and $\beta = 1/3$ such that $\phi = 2/3$ for any ρ. Figure 2.1f shows an example of this where we (indeed) achieve the same decay rate for any streaming rate ρ.

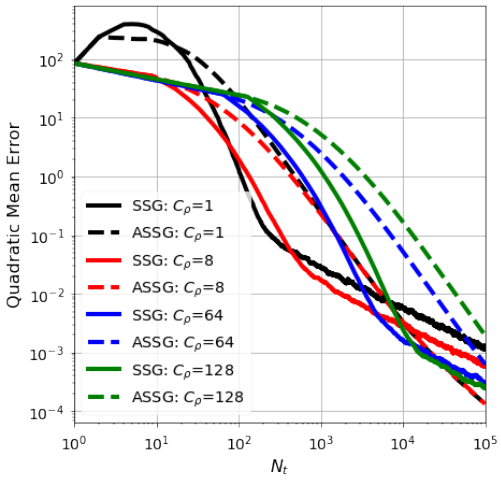
2.5.2 Geometric Median

The geometric median is a generalization of the real median introduced by Haldane [64]. Robust estimators such as the geometric median may be preferred over the mean when the data is noisy. Moreover, in our streaming framework, stochastic algorithms are preferred as they efficiently handle large samples of high-dimensional data [33, 55]. The geometric median of $X \in \mathbb{R}^d$ is defined by $\theta^* \in \mathbb{R}^d$ which minimizes the convex function $L(\theta) = \mathbb{E}[\|X - \theta\| - \|X\|]$, e.g., see Gervini [53], Kemperman [81] for properties such as existence, uniqueness, and robustness (breakdown point). Thus, the gradient $\nabla_\theta L(\theta) = \mathbb{E}[\nabla_\theta l_t(\theta)]$ with $\nabla_\theta l_t(\theta) = -(X_t - \theta)/\|X_t - \theta\|$ is bounded as $\|\nabla_\theta l_t(\theta)\| \leq 1$. We omit to project our estimates as this would hide the errors we want to explore (which we will see more clearly in Example 1.3.2, where we consider real-life time-dependent streaming data). Instead of projecting the estimates, one could adapt the proof of Gadat and Panloup [50] to a streaming setting. Otherwise, if X_t is bounded, one can adapt Cardot et al. [32] to the streaming setting showing that the streaming estimates are bounded. Similarly to above, we fix $d = 10$ and let (X_t) be standard Gaussian centered at $\theta = (-4, -3, 2, 1, 0, 1, 2, 3, 4, 5)^T \in \mathbb{R}^{10}$. Moreover, following the reasoning of Cardot et al. [33], we set $C_\gamma = \sqrt{d} = \sqrt{10}$, and let $\alpha = 2/3$.

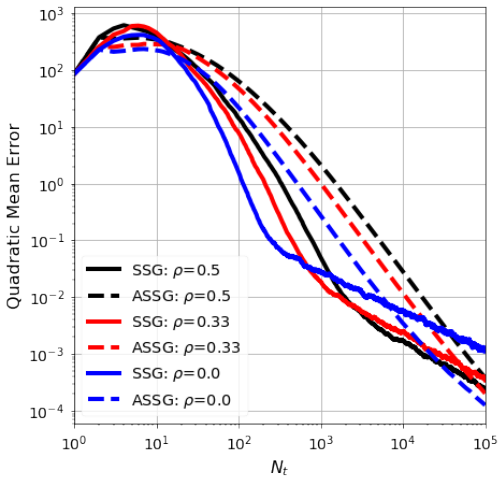
Figure 2.2a shows the variance reduction effect for different constant streaming batches C_ρ with $\beta = 0$. However, the robustness of the geometric median leaves only a small positive impact for further variance reduction. Thus, too large (constant) streaming batch sizes C_ρ hinders the convergence as we make too few iterations. These findings can be extended to Figures 2.2b to 2.2e, where we vary the streaming rate ρ for streaming batch sizes $C_\rho = 1$, 8, 64, and 128, respectively, with $\beta = 0$. The lack of convergence improvements comes from $\beta = 0$, which means we do not exploit the potential of using more observations to accelerate convergence. As shown in Figure 2.2f, we can achieve this acceleration by simply taking $\beta = 1/3$. In addition, $\beta = 1/3$ provides optimal convergence robust to any streaming rate ρ. Choosing a proper $\beta > 0$ is particularly important when C_ρ is large, as robustness is an integral part of the geometric median method.

Figure 2.1: Linear regression for various data streams $n_t = C_\rho t^\rho$. See Section 2.5.1 for details.

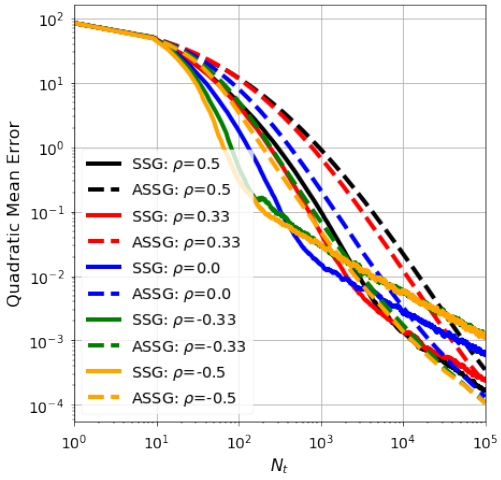
(a) Constant streaming batches, $\rho = 0$, $\beta = 0$



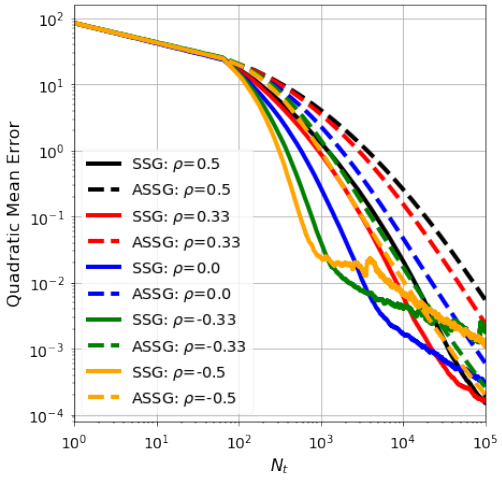
(b) Varying streaming batches, $C_\rho = 1$, $\beta = 0$



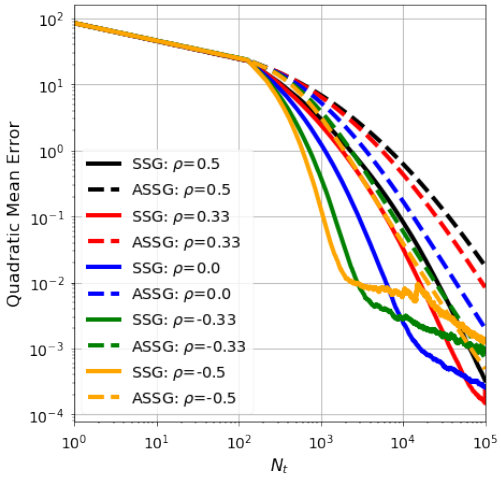
(c) Varying streaming batches, $C_\rho = 8$, $\beta = 0$



(d) Varying streaming batches, $C_\rho = 64$, $\beta = 0$



(e) Varying streaming batches, $C_\rho = 128$, $\beta = 0$



(f) Varying streaming batches, $C_\rho = 8$, $\beta = 1/3$

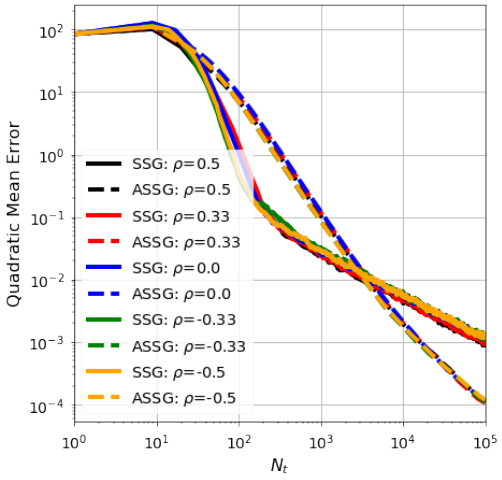
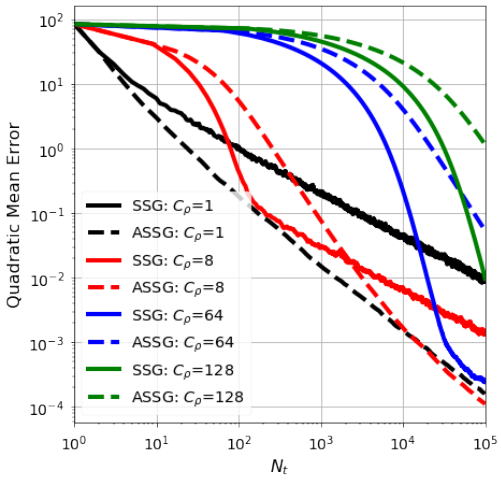
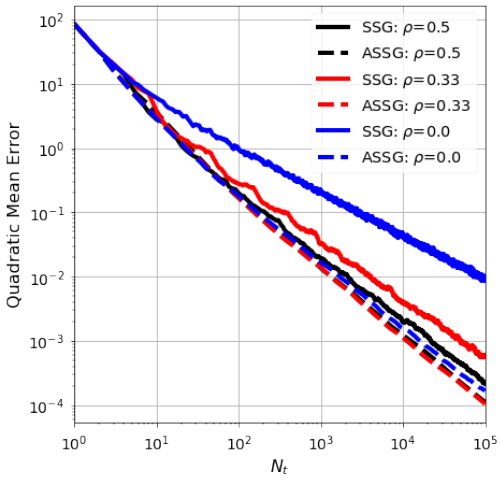


Figure 2.2: Geometric median for various data streams $n_t = C_\rho t^\rho$. See Section 2.5.2 for details.

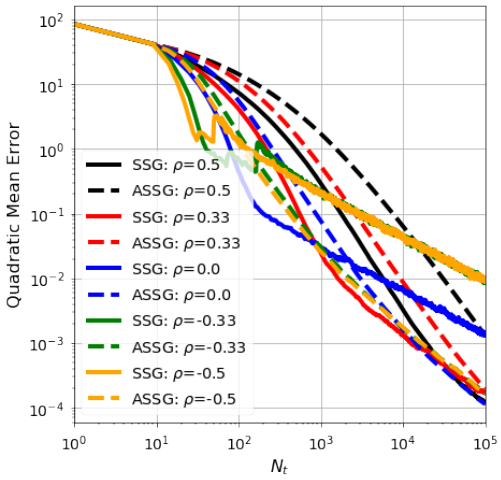
(a) Constant streaming batches, $\rho = 0$, $\beta = 0$



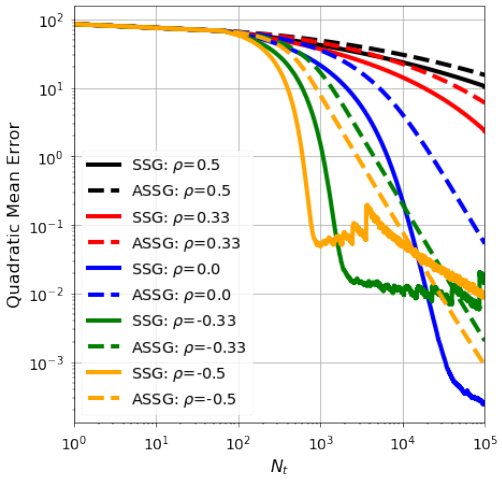
(b) Varying streaming batches, $C_\rho = 1$, $\beta = 0$



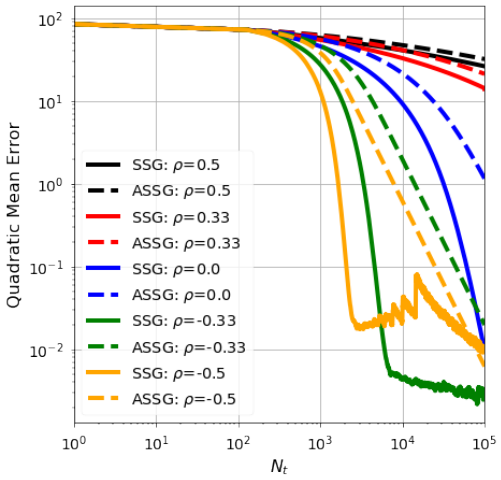
(c) Varying streaming batches, $C_\rho = 8$, $\beta = 0$



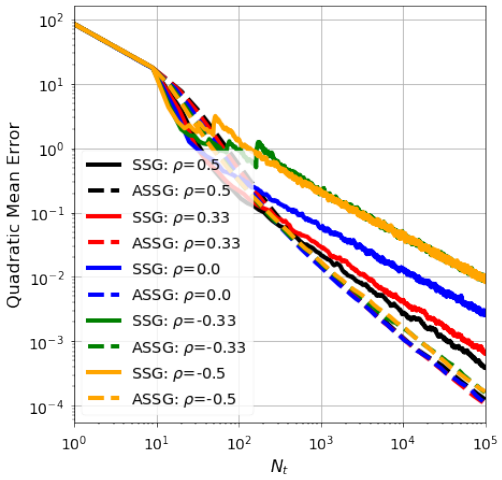
(d) Varying streaming batches, $C_\rho = 64$, $\beta = 0$



(e) Varying streaming batches, $C_\rho = 128$, $\beta = 0$



(f) Varying streaming batches, $C_\rho = 8$, $\beta = 1/3$



2.6 Conclusions

We considered the SO problem in a streaming framework where we had to minimize objectives using only unbiased estimates of its gradients. We introduced and studied the convergence rates of the stochastic streaming algorithms in a non-asymptotic manner. This investigation was derived using learning rates of the form $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ under varying data streams of n_t. The theoretical results and our experiments showed a noticeable improvement in the convergence rate by choosing the learning rate (hyper-parameters) according to the expected data streams. For ASSG and PASSG, we showed that this choice of learning rate led to optimal convergence rates and was robust to any data stream rate we may encounter. Moreover, in large-scale learning problems, we know how to accelerate convergence and reduce noise through the learning rate and the treatment pattern of the data.

There are several ways to expand our work but let us give some examples: first, we can extend our analysis to include streaming batches of any size in the spirit of the discussion after Corollary 2.3.2. Second, many machine learning problems encounter correlated variables and high-dimensional data, making an extension to non-strongly convex objectives advantageous, e.g, in Werge and Wintenberger [150], they use SG-based optimization methods for volatility prediction through GARCH modeling. Third, Assumption 2.3.1 requires independent random functions, thus, an obvious extension could incorporate a more realistic dependency assumption, thereby increasing the applicability for more models. Moreover, studying dependence may give insight into how to process dependent information *optimally*. Next, a natural extension would be to modify our averaging estimate from (2.2.3) to a weighted averaged version (WASSG) proposed by Mokkadem and Pelletier [95] and Boyer and Godichon-Baggioni [27], given as

$$\bar{\theta}_{t,\lambda} = \frac{1}{\sum_{i=1}^{t} n_i \log(1+i)^\lambda} \sum_{i=1}^{t} n_i \log(1+i)^\lambda \theta_{i-1}, \tag{2.6.12}$$

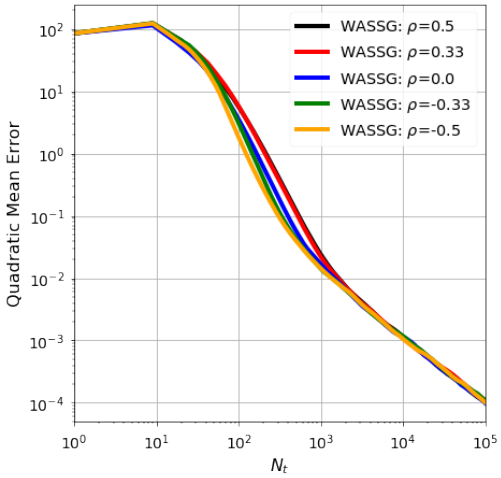
where $\bar{\theta}_{0,\lambda} = 0$, $\lambda > 0$ and (θ_t) is coming from (2.2.1). By giving more importance to the latest estimates, we should improve convergence and limit the effect of bad initializations. Following the demonstrations in Section 2.5, an example of this WASSG estimate $(\bar{\theta}_{t,\lambda})$ can be found in Figure 2.3 with use of $\lambda = 2$. Here we see that although the WASSG estimate in (2.6.12) may not achieve a better final error (compared to the ASSG and PASSG estimates in Figures 2.1f and 2.2f), it still achieves a better decay along the way, often referred to as *parameter tracking*.

2.7 Proofs

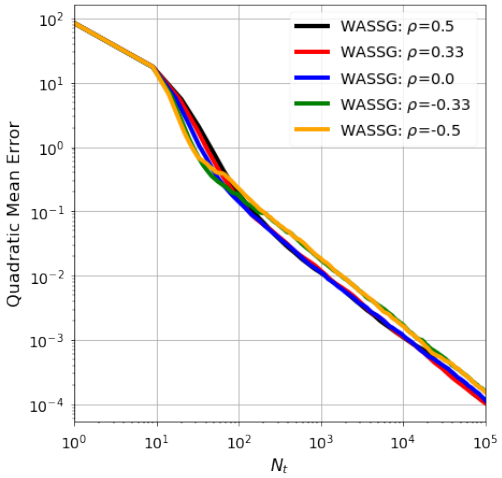
In this section, we provide detailed proofs of the results presented in the manuscript. Purely technical results used in the proofs can be found in Chapter B. Let $(\mathcal{F}_t)_{t\geq 1}$ be an increasing family of σ-fields, namely $\mathcal{F}_t = \sigma(l_1, \ldots, l_t)$ with $l_t = (l_{t,1}, \ldots, l_{t,n_t})$. Furthermore, we expand the notation with $\mathcal{F}_{t-1,i} = \sigma(l_{1,1}, \ldots, l_{t-1,n_{t-1}}, l_{t,1}, \ldots, l_{t,i})$ such that $\mathcal{F}_{t-1,0} = \mathcal{F}_{t-1}$. Meaning, $\forall 0 \leq i < j$, we

Figure 2.3: WASSG for various data streams $n_t = C_\rho t^\rho$. See Section 2.6 for details.

(a) Linear regression, varying streaming batches, $C_\rho = 8$, $\beta = 1/3$



(b) Geometric median, varying streaming batches, $C_\rho = 8$, $\beta = 1/3$



have $\mathcal{F}_{t-1} \subseteq \mathcal{F}_{t-1,i} \subset \mathcal{F}_{t-1,j}$. Thus, by the independence of the random (differentiable) functions $(l_{t,i})$, Assumption 2.3.1 yields that $\forall t \geq 1$, $\mathbb{E}[\nabla_\theta l_{t,i}(\theta_{t-1})|\mathcal{F}_{t-1,i-1}] = \nabla_\theta L(\theta_{t-1})$ with $i = 1, \ldots, n_t$.

2.7.1 Proofs for Section 2.3

The section is structured such that we start by analyzing the recursive relations and bounding them for every choice of learning rate. Next, we look at specific choices of learning rates.

Proof of Theorem 2.3.1. Taking the quadratic norm on both sides of (2.2.1), expanding it, and take the conditional expectation, yields

$$\begin{aligned}\mathbb{E}[\|\theta_t - \theta^*\|^2|\mathcal{F}_{t-1}] =& \|\theta_{t-1} - \theta^*\|^2 + \frac{\gamma_t^2}{n_t^2}\mathbb{E}\left[\left\|\sum_{i=1}^{n_t} \nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2 \middle| \mathcal{F}_{t-1}\right] \\ &- \frac{2\gamma_t}{n_t}\sum_{i=1}^{n_t}\mathbb{E}[\langle \nabla_\theta l_{t,i}(\theta_{t-1}), \theta_{t-1} - \theta^*\rangle|\mathcal{F}_{t-1}]. \end{aligned} \tag{2.7.13}$$

To bound the second term (on the right-hand side) of (2.7.13), we first expand it as follows,

$$\sum_{i=1}^{n_t}\mathbb{E}[\|\nabla_\theta l_{t,i}(\theta_{t-1})\|^2|\mathcal{F}_{t-1}] + \sum_{i\neq j}^{n_t}\mathbb{E}[\langle \nabla_\theta l_{t,i}(\theta_{t-1}), \nabla_\theta l_{t,j}(\theta_{t-1})\rangle|\mathcal{F}_{t-1}]. \tag{2.7.14}$$

For first term of (2.7.14), we utilize the Lipschitz continuity of $\nabla_\theta l_{t,i}$, together with Assump-

tions 2.3.1 to 2.3.3-p, to obtain

$$\begin{aligned}\mathbb{E}[\|\nabla_\theta l_{t,i}(\theta_{t-1})\|^2|\mathcal{F}_{t-1}] \leq& 2\mathbb{E}[\|\nabla_\theta l_{t,i}(\theta_{t-1}) - \nabla_\theta l_{t,i}(\theta^*)\|^2|\mathcal{F}_{t-1}] + 2\mathbb{E}[\|\nabla_\theta l_{t,i}(\theta^*)\|^2|\mathcal{F}_{t-1}] \\ \leq& 2C_l^2\|\theta_{t-1} - \theta^*\|^2 + 2\sigma^2, \end{aligned} \tag{2.7.15}$$

using $\|x+y\|^2 \leq 2(\|x\|^2 + \|y\|^2)$. Next, for the second term in (2.7.14): as $\mathcal{F}_{t-1} \subseteq \mathcal{F}_{t-1,i} \subset \mathcal{F}_{t-1,j}$ for all $0 \leq i < j$, we have

$$\mathbb{E}[\langle \nabla_\theta l_{t,i}(\theta_{t-1}), \nabla_\theta l_{t,j}(\theta_{t-1})\rangle|\mathcal{F}_{t-1}] = \mathbb{E}[\mathbb{E}[\langle \nabla_\theta l_{t,i}(\theta_{t-1}), \nabla_\theta L(\theta_{t-1})\rangle|\mathcal{F}_{t-1,i-1}]|\mathcal{F}_{t-1}],$$

since θ_{t-1} and $l_{t,i}$ are $\mathcal{F}_{t-1,j-1}$-measurable for all $0 \leq i < j$, and similarly, as θ_{t-1} is $\mathcal{F}_{t-1}$-measurable and $\mathcal{F}_{t-1,i-1}$-measurable for all $i \geq 0$, we also have

$$\begin{aligned}\mathbb{E}[\mathbb{E}[\langle \nabla_\theta l_{t,i}(\theta_{t-1}), \nabla_\theta L(\theta_{t-1})\rangle|\mathcal{F}_{t-1,i-1}]|\mathcal{F}_{t-1}] =& \mathbb{E}[\langle \mathbb{E}[\nabla_\theta l_{t,i}(\theta_{t-1})|\mathcal{F}_{t-1,i-1}], \nabla_\theta L(\theta_{t-1})\rangle|\mathcal{F}_{t-1}] \\ =& \|\nabla_\theta L(\theta_{t-1})\|^2, \end{aligned}$$

where $\|\nabla_\theta L(\theta_{t-1})\|^2 \leq C_\nabla^2\|\theta_{t-1} - \theta^*\|^2$ as $\nabla_\theta L$ is C_∇-Lipschitz continuous and $\nabla_\theta L(\theta^*) = 0$. Thus, we obtained a bound for the second term (on the right-hand side) of (2.7.13) using the bounds of the two terms in (2.7.14):

$$\sum_{i=1}^{n_t}(2C_l^2\|\theta_{t-1} - \theta^*\|^2 + 2\sigma^2) + \sum_{i\neq j}^{n_t} C_\nabla^2\|\theta_{t-1} - \theta^*\|^2 = (2C_l^2 n_t + C_\nabla^2(n_t - 1)n_t)\|\theta_{t-1} - \theta^*\|^2 + 2\sigma^2 n_t. \tag{2.7.16}$$

For the third term (on the right-hand side) of (2.7.13) we use that L is μ-quasi-strong convex and θ_{t-1} is $\mathcal{F}_{t-1}$-measurable,

$$\begin{aligned}\mathbb{E}[\langle \nabla_\theta l_{t,i}(\theta_{t-1}), \theta_{t-1} - \theta^*\rangle|\mathcal{F}_{t-1}] =& \langle \mathbb{E}[\nabla_\theta l_{t,i}(\theta_{t-1})|\mathcal{F}_{t-1}], \theta_{t-1} - \theta^*\rangle \\ =& \langle \nabla_\theta L(\theta_{t-1}), \theta_{t-1} - \theta^*\rangle \\ \geq& \mu\|\theta_{t-1} - \theta^*\|^2, \end{aligned} \tag{2.7.17}$$

by Assumption 2.3.1. Combining inequalities from (2.7.16) and (2.7.17) into (2.7.13) and taking the expectation on both sides of the inequality, yields the recursive relation (2.3.8):

$$\delta_t \leq [1 - 2\mu\gamma_t + (2C_l^2 + (n_t - 1)C_\nabla^2)n_t^{-1}\gamma_t^2]\delta_{t-1} + 2\sigma^2 n_t^{-1}\gamma_t^2,$$

with $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ with some $\delta_0 \geq 0$. At last, by Proposition B.1.5, we obtain the desired

inequality in (2.3.7), namely

$$\delta_t \leq \exp\left(-\mu \sum_{i=t/2}^{t} \gamma_i\right) \exp\left(4C_l^2 \sum_{i=1}^{t} \frac{\gamma_i^2}{n_i}\right) \exp\left(2C_\nabla^2 \sum_{i=1}^{t} \mathbb{1}_{\{n_i>1\}} \gamma_i^2\right) \left(\delta_0 + \frac{2\sigma^2}{C_l^2}\right) + \frac{2\sigma^2}{\mu} \max_{t/2\leq i\leq t} \frac{\gamma_i}{n_i}.$$

using that $(n_t - 1)n_t^{-1} \leq \mathbb{1}_{\{n_t>1\}}$, $n_t \geq 1$, and that

$$\max_{1\leq i\leq t} \frac{2\sigma^2}{2C_l^2 + (n_i - 1)C_\nabla^2} \leq \max_{1\leq i\leq t} \frac{2\sigma^2}{2C_l^2} = \frac{\sigma^2}{C_l^2}.$$

□

Remark 2.7.1. The decrease of $(2C_l^2 + (n_t - 1)C_\nabla^2)n_t^{-1}\gamma_t$ determines when the *stationary* phase occurs. This is more clearly seen in Proposition B.1.4, where the inner terms directly depend on the inception of the stationary phase. Thus, by increasing n_t, we decrease $(2C_l^2 + (n_t - 1)C_\nabla^2)n_t^{-1}\gamma_t$, and especially it dominates the constant C_l.

Proof of Corollary 2.3.1. By Theorem 2.3.1, we have the upper bound giving as

$$\delta_t \leq \exp\left(-\mu \sum_{i=t/2}^{t} \gamma_i\right) \pi_t^c + \frac{2\sigma^2}{\mu C_\rho} \max_{t/2\leq i\leq t} \gamma_i. \tag{2.7.18}$$

as $n_t = C_\rho$, with $\pi_t^c = \exp((4C_l^2/C_\rho)\sum_{i=1}^t \gamma_i^2) \exp(2C_\nabla^2 \mathbb{1}_{\{C_\rho>1\}} \sum_{i=1}^t \gamma_i^2)(\delta_0 + \sigma^2/C_l^2)$. The sum term $\sum_{i=1}^t \gamma_i^2 = C_\gamma^2 C_\rho^{2\beta} \sum_{i=1}^t i^{-2\alpha}$ in π_t^c can be bounded with the help of integral tests for convergence, $\sum_{i=1}^t i^{-2\alpha} = 1 + \sum_{i=2}^t i^{-2\alpha} \leq 1 + \int_1^t x^{-2\alpha}\, dx \leq 1 + 1/(2\alpha - 1) = 2\alpha/(2\alpha - 1)$, as $\alpha \in (1/2, 1)$. Likewise, plugging $\gamma_t = C_\gamma C_\rho^\beta t^{-\alpha}$ into the first term of (2.7.18), gives

$$\exp\left(-\mu \sum_{i=t/2}^{t} \gamma_i\right) = \exp\left(-\mu C_\gamma C_\rho^\beta \sum_{i=t/2}^{t} i^{-\alpha}\right) \leq \exp\left(-\mu C_\gamma C_\rho^\beta \int_{t/2}^{t} x^{-\alpha}\, dx\right) \leq \exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1-\alpha}}{2^{1-\alpha}}\right),$$

using the integral test for convergence. Next, as $(\gamma_t)_{t\geq 1}$ is decreasing, then $\max_{t/2\leq i\leq t} \gamma_t = \gamma_{t/2}$. Combining all these findings into (2.7.18), gives us

$$\delta_t \leq \exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1-\alpha}}{2^{1-\alpha}}\right) \pi_\infty^c + \frac{2^{1+\alpha}\sigma^2 C_\gamma}{\mu C_\rho^{1-\beta} t^\alpha}, \tag{2.7.19}$$

with $\pi_\infty^c = \exp(4\alpha C_\gamma^2 (2C_l^2 + C_\rho \mathbb{1}_{\{C_\rho>1\}} C_\nabla^2)/(2\alpha - 1)C_\rho^{1-2\beta})(\delta_0 + 2\sigma^2/C_l^2)$. At last, converting (2.7.19) into terms of N_t using $N_t = C_\rho t$, yields the desired. □

Proof of Corollary 2.3.2. For convenience, we divided the proof into two cases to comprehend that $n_t \geq 1$ for all t: first, we bound each term of (2.3.7) (from Theorem 2.3.1) after inserting, $\gamma_t = C_\gamma n_t^\beta t^{-\alpha} = C_\gamma C_\rho^\beta t^{\beta\rho-\alpha}$ if $\rho \geq 0$, or $\gamma_t \geq C_\gamma t^{-\alpha}$ if $\rho < 0$ (using that $\beta \geq 0$) into the inequality. If

$\rho \geq 0$, the first term of (2.3.7) can be bounded, as follows:

$$\exp\left(-\mu \sum_{i=t/2}^{t} \gamma_i\right) = \exp\left(-\mu C_\gamma C_\rho^\beta \sum_{i=t/2}^{t} i^{\beta\rho-\alpha}\right) \leq \exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1+\beta\rho-\alpha}}{2^{1+\beta\rho-\alpha}}\right),$$

using that $\alpha - \beta\rho \in (1/2, 1)$ and the integral test for convergence. In a same way, if $\rho < 0$, one has

$$\exp\left(-\mu \sum_{i=t/2}^{t} \gamma_i\right) \leq \exp\left(-\mu C_\gamma \sum_{i=t/2}^{t} i^{-\alpha}\right) \leq \exp\left(-\frac{\mu C_\gamma t^{1-\alpha}}{2^{1-\alpha}}\right).$$

Likewise, with the help of integral tests for convergence, we have for $\rho \geq 0$, that $\sum_{i=1}^t \gamma_i^2/n_i \leq \sum_{i=1}^t \gamma_i^2 \leq 2(\alpha-\beta\rho)C_\gamma^2 C_\rho^{2\beta}/(2(\alpha-\beta\rho)-1)$, as $n_t \geq 1$ and $\alpha - \rho\beta > 1/2$. If $\rho < 0$, one has $\sum_{i=1}^t \gamma_i^2/n_i \leq \sum_{i=1}^t \gamma_i^2 \leq 2\alpha C_\gamma^2 C_\rho^{2\beta}/(2\alpha-1)$ since $C_\rho \geq n_t \geq 1$. Next, as $(1-\beta)\rho + \alpha > 0$ for $\rho \geq 0$, then we can bound the last term of (2.3.7) by

$$\frac{2\sigma^2}{\mu} \max_{t/2 \leq i \leq t} \frac{\gamma_i}{n_i} = \frac{2\sigma^2 C_\gamma}{\mu C_\rho^{1-\beta}} \max_{t/2 \leq i \leq t} \frac{1}{i^{(1-\beta)\rho+\alpha}} \leq \frac{2^{1+(1-\beta)\rho+\alpha}\sigma^2 C_\gamma}{\mu C_\rho^{1-\beta} t^{(1-\beta)\rho+\alpha}}.$$

Likewise, if $\rho < 0$, we have

$$\frac{2\sigma^2}{\mu} \max_{t/2 \leq i \leq t} \frac{\gamma_i}{n_i} = \frac{2\sigma^2 C_\gamma}{\mu} \max_{t/2 \leq i \leq t} \frac{1}{n_i^{1-\beta} i^\alpha} \leq \frac{2^{1+\alpha}\sigma^2 C_\gamma}{\mu t^\alpha},$$

since $n_t \geq 1$ and $\beta \leq 1$. Combining all these findings gives

$$\delta_t \leq \exp\left(-\frac{\mu C_\gamma C_\rho^{\beta \mathbb{1}_{\{\rho \geq 0\}}} t^{(1-\phi)(1+\tilde{\rho})}}{2^{(1-\phi)(1+\tilde{\rho})}}\right) \pi_\infty^v + \frac{2^{1+\phi(1+\tilde{\rho})}\sigma^2 C_\gamma}{\mu C_\rho^{(1-\beta)\mathbb{1}_{\{\rho\geq 0\}}} t^{\phi(1+\tilde{\rho})}}, \tag{2.7.20}$$

where $\pi_\infty^v = \exp(4(\alpha-\beta\tilde{\rho})C_\gamma^2 C_\rho^{2\beta}(2C_l^2 + C_\nabla^2)/2(\alpha-\beta\tilde{\rho}) - 1)$ with $\tilde{\rho} = \rho \mathbb{1}_{\{\rho \geq 0\}}$ and $\phi = ((1-\beta)\tilde{\rho} + \alpha)/(1+\tilde{\rho})$. To write this in terms of N_t, we use that $N_t = \sum_{i=1}^t n_i = C_\rho \sum_{i=1}^t i^\rho = C_\rho(t^\rho + \sum_{i=1}^{t-1} i^\rho) \leq C_\rho(t^\rho + \int_1^t x^\rho \, dx) \leq C_\rho(t^\rho + t^\rho \int_1^t dx) = C_\rho(t^\rho + t^{1+\rho}) \leq 2C_\rho t^{1+\rho}$, for $\rho \geq 0$, thus, $t \geq (N_t/2C_\rho)^{1/(1+\rho)}$. For $\rho < 0$, we have $N_t \leq C_\rho t$, i.e, $t \geq N_t/C_\rho$. □

2.7.2 Proofs for Section 2.4

Lemma 2.7.1 (ASSG/PASSG). *Denote $\Delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^4]$ for some $\Delta_0 \geq 0$, where (θ_t) follows (2.2.1) or (2.2.2). Under Assumption 2.3.1, Assumptions 2.3.2-p and 2.3.3-p with $p = 4$ and Assumption 2.4.1, we have for any learning rate (γ_t) that*

$$\Delta_t \leq \exp\left(-\mu \sum_{i=t/2}^{t} \gamma_i\right) \Pi_t^\Delta + \frac{32\sigma^4}{\mu^2} \max_{t/2 \leq i \leq t} \frac{\gamma_i^2}{n_i^2} + \frac{48\sigma^4}{\mu} \max_{t/2 \leq i \leq t} \frac{\gamma_i^3}{n_i^3} + \frac{114\sigma^4}{\mu} \max_{t/2 \leq i \leq t} \frac{\gamma_i^3 \mathbb{1}_{\{n_i > 1\}}}{n_i^2}, \tag{2.7.21}$$

with Π_t^{Δ} *given in* (2.7.29).

Proof of Lemma 2.7.1. We will now derive the recursive step sequence for the fourth-order moment using the same arguments as in proof for Theorem 2.3.1. Thus, one can show that

$$\begin{aligned}\mathbb{E}[\|\theta_t-\theta^*\|^4|\mathcal{F}_{t-1}] \leq & \|\theta_{t-1}-\theta^*\|^4+\frac{\gamma_t^4}{n_t^4}\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^4\middle|\mathcal{F}_{t-1}\right] \\ & +\frac{4\gamma_t^2}{n_t^2}\mathbb{E}\left[\left\langle\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right),\theta_{t-1}-\theta^*\right\rangle^2\middle|\mathcal{F}_{t-1}\right] \\ & +\frac{2\gamma_t^2}{n_t^2}\|\theta_{t-1}-\theta^*\|^2\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2\middle|\mathcal{F}_{t-1}\right] \\ & -\frac{4\gamma_t}{n_t}\|\theta_{t-1}-\theta^*\|^2\sum_{i=1}^{n_t}\langle\mathbb{E}[\nabla_\theta l_{t,i}(\theta_{t-1})|\mathcal{F}_{t-1}],\theta_{t-1}-\theta^*\rangle \\ & +\frac{4\gamma_t^3}{n_t^3}\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2\left\langle\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right),\theta_{t-1}-\theta^*\right\rangle\middle|\mathcal{F}_{t-1}\right],\end{aligned}$$

using θ_{t-1} is $\mathcal{F}_{t-1}-$measurable. Note, by Assumption 2.3.1, we have

$$\langle\mathbb{E}[\nabla_\theta l_{t,i}(\theta_{t-1})|\mathcal{F}_{t-1}],\theta_{t-1}-\theta^*\rangle=\langle\nabla_\theta L(\theta_{t-1}),\theta_{t-1}-\theta^*\rangle\geq\mu\|\theta_{t-1}-\theta^*\|^2,$$

as L is μ-quasi-strong convex. Combining this with the Cauchy-Schwarz inequality $\langle x,y\rangle\leq\|x\|\|y\|$, we obtain the simplified expression:

$$\begin{aligned}\mathbb{E}[\|\theta_t-\theta^*\|^4|\mathcal{F}_{t-1}] \leq & \|\theta_{t-1}-\theta^*\|^4+\frac{\gamma_t^4}{n_t^4}\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^4\middle|\mathcal{F}_{t-1}\right] \\ & +\frac{6\gamma_t^2}{n_t^2}\|\theta_{t-1}-\theta^*\|^2\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2\middle|\mathcal{F}_{t-1}\right] \\ & -4\mu\gamma_t\|\theta_{t-1}-\theta^*\|^4+\frac{4\gamma_t^3}{n_t^3}\|\theta_{t-1}-\theta^*\|\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^3\middle|\mathcal{F}_{t-1}\right].\end{aligned}$$

Next, recall Young's Inequality, i.e., for any $a_t,b_t,c_t>0$ we have $a_tb_t\leq a_t^2c_t^2/2+b_t^2/2c_t^2$,

$$\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^3\leq\frac{\gamma_t}{2n_t\left\|\theta_{t-1}-\theta^*\right\|}\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^4+\frac{2n_t\left\|\theta_{t-1}-\theta^*\right\|}{\gamma_t}\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2,$$

giving us

$$\mathbb{E}[\|\theta_t - \theta^*\|^4|\mathcal{F}_{t-1}] \leq (1-4\mu\gamma_t)\|\theta_{t-1}-\theta^*\|^4 + \frac{3\gamma_t^4}{n_t^4}\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^4\middle|\mathcal{F}_{t-1}\right] + \frac{8\gamma_t^2}{n_t^2}\|\theta_{t-1}-\theta^*\|^2\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2\middle|\mathcal{F}_{t-1}\right]. \tag{2.7.22}$$

To bound the second and fourth-order terms in (2.7.22), we would need to study the recursive sequences: firstly, utilizing the Lipschitz continuity of $\nabla_\theta l_{t,i}$, together with Assumptions 2.3.2-p and 2.3.3-p, and that θ_{t-1} is $\mathcal{F}_{t-1}$-measurable (Assumption 2.3.1), we obtain

$$\begin{aligned}\mathbb{E}[\|\nabla_\theta l_{t,i}(\theta_{t-1})\|^p|\mathcal{F}_{t-1}] \leq& 2^{p-1}[\mathbb{E}[\|\nabla_\theta l_{t,i}(\theta_{t-1}) - \nabla_\theta l_{t,i}(\theta^*)\|^p|\mathcal{F}_{t-1}] + \mathbb{E}[\|\nabla_\theta l_{t,i}(\theta^*)\|^p|\mathcal{F}_{t-1}]] \\ \leq& 2^{p-1}[C_l^p\|\theta_{t-1}-\theta^*\|^p + \sigma^p],\end{aligned} \tag{2.7.23}$$

for any $p \in [1,4]$ using the bound $\|x+y\|^p \leq 2^{p-1}(\|x\|^p + \|y\|^p)$. Thus, we can bound the second-order term in (2.7.22) by

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{i=1}^{t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2\middle|\mathcal{F}_{t-1}\right] \leq& [2C_l^2 n_t + C_\nabla^2(n_t-1)n_t]\|\theta_{t-1}-\theta^*\|^2 + 2\sigma^2 n_t \\ \leq& [2C_l^2 n_t + C_\nabla^2 n_t^2 \mathbb{1}_{\{n_t>1\}}]\|\theta_{t-1}-\theta^*\|^2 + 2\sigma^2 n_t,\end{aligned} \tag{2.7.24}$$

following the same steps in the proof of Theorem 2.3.1, but with use of (2.7.23). Bounding the fourth-order term is a bit heavier computationally, but let us recall that $\|\sum_i x_i\|^2 = \sum_i\|x_i\|^2 + \sum_{i\neq j}\langle x_i, x_j\rangle = \sum_i\|x_i\|^2 + 2\sum_{i\leq j}\langle x_i, x_j\rangle$. Then, we have that

$$\begin{aligned}\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}(\theta_{t-1})\right\|^4 =& \left(\sum_{i=1}^{n_t}\|\nabla_\theta l_{t,i}(\theta_{t-1})\|^2 + \sum_{i\neq j}^{n_t}\langle\nabla_\theta l_{t,i}(\theta_{t-1}), \nabla_\theta l_{t,j}(\theta_{t-1})\rangle\right)^2 \\ \leq& 2\left(\sum_{i=1}^{n_t}\|\nabla_\theta l_{t,i}(\theta_{t-1})\|^2\right)^2 + 4\left(\sum_{i<j}^{n_t}\langle\nabla_\theta l_{t,i}(\theta_{t-1}), \nabla_\theta l_{t,j}(\theta_{t-1})\rangle\right)^2,\end{aligned} \tag{2.7.25}$$

as $(x+y)^2 \leq 2x^2 + 2y^2$. For the first term of (2.7.25), we have

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{n_1}\|\nabla_\theta l_{t,i}(\theta_{t-1})\|^2\right)^2\middle|\mathcal{F}_{t-1}\right] =& \sum_{i=1}^{n_t}\mathbb{E}[\|\nabla_\theta l_{t,i}(\theta_{t-1})\|^4|\mathcal{F}_{t-1}] + \sum_{i\neq j}^{n_t}\mathbb{E}[\|\nabla_\theta l_{t,i}(\theta_{t-1})\|^2\|\nabla_\theta l_{t,j}(\theta_{t-1})\|^2|\mathcal{F}_{t-1}] \\ \leq& 8n_t[C_l^4\|\theta_{t-1}-\theta^*\|^4 + \sigma^4] + 4n_t^2\mathbb{1}_{\{n_t>1\}}[C_l^2\|\theta_{t-1}-\theta^*\|^2 + \sigma^2]^2,\end{aligned}$$

using the bound from (2.7.23), $n_t(n_t-1) \leq n_t^2\mathbb{1}_{\{n_t>1\}}$, and that $\mathcal{F}_{t-1} \subseteq \mathcal{F}_{t-1,i} \subset \mathcal{F}_{t-1,j}$ for all $0 \leq i < j$. To bound the second term of (2.7.25), we ease notation by denoting $\nabla_\theta l_{t,i}(\theta_{t-1})$ by υ_i,

giving us

$$\left(\sum_{i<j}^{n_t}\langle v_i, v_j\rangle\right)^2 = \sum_{i<j}^{n_t}\langle v_i, v_j\rangle^2 + \sum_{\substack{i<j,k<l\\(i,j)\neq(k,l)}}^{n_t}\langle v_i, v_j\rangle\langle v_k, v_l\rangle$$
$$= \underbrace{\sum_{i<j}^{n_t}\langle v_i, v_j\rangle^2}_{A} + \underbrace{\sum_{\substack{i<j,k<l\\(i,j)\neq(k,l),j=l}}^{n_t}\langle v_i, v_j\rangle\langle v_k, v_l\rangle}_{B} + \underbrace{\sum_{\substack{i<j,k<l\\(i,j)\neq(k,l),j\neq l}}^{n_t}\langle v_i, v_j\rangle\langle v_k, v_l\rangle}_{C}.$$

By Cauchy-Schwarz inequality, we can bound the first term A, by

$$\begin{aligned}\mathbb{E}[A|\mathcal{F}_{t-1}] \leq& \sum_{i<j}^{n_t}\mathbb{E}[\|v_i\|^2\|v_j\|^2|\mathcal{F}_{t-1}]\\ \leq& 2n_t(n_t-1)[C_l^2\|\theta_{t-1}-\theta^*\|^2+\sigma^2]^2\\ \leq& 2n_t^2\mathbb{1}_{\{n_t>1\}}[C_l^2\|\theta_{t-1}-\theta^*\|^2+\sigma^2]^2,\end{aligned}$$

using that $\mathcal{F}_{t-1}\subseteq\mathcal{F}_{t-1,i}\subset\mathcal{F}_{t-1,j}$ for all $0\leq i<j$. Next, since $l=j$ implies $i\neq k$, we have

$$\begin{aligned}\mathbb{E}[B|\mathcal{F}_{t-1}] =& \sum_{i<j,k<l,i\neq k,j=l}^{n_t}\mathbb{E}[\langle v_i, v_j\rangle\langle v_k, v_l\rangle|\mathcal{F}_{t-1}]\\ =& \sum_{i<j,k<l,i\neq k,j=l}^{n_t}\mathbb{E}[\mathbb{E}[\langle\mathbb{E}[v_i|\mathcal{F}_{t-1,i-1}], v_j\rangle\langle\mathbb{E}[v_k|\mathcal{F}_{t-1,k-1}], v_l\rangle|\mathcal{F}_{t-1,l-1}]|\mathcal{F}_{t-1}]\\ =& \sum_{i<j,k<l,i\neq k,j=l}^{n_t}\mathbb{E}[\mathbb{E}[\langle\nabla_\theta L(\theta_{t-1}), v_l\rangle^2|\mathcal{F}_{t-1,l-1}]|\mathcal{F}_{t-1}]\\ \leq& \sum_{i<j,k<l,i\neq k,j=l}^{n_t}\mathbb{E}[\|\nabla_\theta L(\theta_{t-1})\|^2\mathbb{E}[\|v_l\|^2|\mathcal{F}_{t-1,l-1}]|\mathcal{F}_{t-1}]\\ \leq& \sum_{i<j,k<l,i\neq k,j=l}^{n_t}2C_\nabla^2\|\theta_{t-1}-\theta^*\|^2[C_l^2\|\theta_{t-1}-\theta^*\|^2+\sigma^2]\\ =& n_t(n_t-1)(n_t-2)C_\nabla^2\|\theta_{t-1}-\theta^*\|^2[C_l^2\|\theta_{t-1}-\theta^*\|^2+\sigma^2]\\ \leq& n_t^3\mathbb{1}_{\{n_t>1\}}C_\nabla^2\|\theta_{t-1}-\theta^*\|^2[C_l^2\|\theta_{t-1}-\theta^*\|^2+\sigma^2],\end{aligned}$$

using the Cauchy-Schwarz inequality and the bound in (2.7.23). In the same way, as $j\neq l$ includes $(i,j)\neq(k,l)$, we can rewrite C as

$$C = \sum_{i<j,k<l,j\neq l}^{n_t}\langle v_i, v_j\rangle\langle v_k, v_l\rangle = \underbrace{\sum_{i<j,k<l,i=k,j\neq l}^{n_t}\langle v_i, v_j\rangle\langle v_k, v_l\rangle}_{C_1} + \underbrace{\sum_{i<j,k<l,i\neq k,j\neq l}^{n_t}\langle v_i, v_j\rangle\langle v_k, v_l\rangle}_{C_2},$$

where $\mathbb{E}[C_1|\mathcal{F}_{t-1}] = \mathbb{E}[B|\mathcal{F}_{t-1}]$. Finally, we can rewrite C_2 as

$$C_2 = \underbrace{\sum_{i<j,k<l,i\neq k,j\neq l,i=l,j\neq k}^{n_t} \langle \upsilon_i \upsilon_j \rangle \langle \upsilon_k \upsilon_l \rangle}_{C_{2,1}} + \underbrace{\sum_{i<j,k<l,i\neq k,j\neq l,i\neq l,j=k}^{n_t} \langle \upsilon_i \upsilon_j \rangle \langle \upsilon_k \upsilon_l \rangle}_{C_{2,2}} + \underbrace{\sum_{i<j,k<l,i\neq j\neq k\neq l}^{n_t} \langle \upsilon_i \upsilon_j \rangle \langle \upsilon_k \upsilon_l \rangle}_{C_{2,3}},$$

where $\mathbb{E}[C_{2,1}|\mathcal{F}_{t-1}] = \mathbb{E}[C_{2,2}|\mathcal{F}_{t-1}] = \mathbb{E}[B|\mathcal{F}_{t-1}]$, and

$$\begin{aligned}
\mathbb{E}[C_{2,3}|\mathcal{F}_{t-1}] = & \sum_{i<j,k<l,i\neq j\neq k\neq l}^{n_t} \mathbb{E}[\|\nabla_\theta L(\theta_{t-1})\|^4|\mathcal{F}_{t-1}] \\
\leq & n_t(n_t-1)(n_t-2)(n_t-3)C_\nabla^4 \|\theta_{t-1}-\theta^*\|^4 \\
\leq & n_t^4 \mathbb{1}_{\{n_t>1\}} C_\nabla^4 \|\theta_{t-1}-\theta^*\|^4.
\end{aligned}$$

Thus, the fourth-order term of (2.7.22), is bounded by

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{i=1}^{n_t} \nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^4 \middle| \mathcal{F}_{t-1}\right] \leq 16n_t[C_l^4\|\theta_{t-1}-\theta^*\|^4+\sigma^4] + 16n_t^2\mathbb{1}_{\{n_t>1\}}[C_l^2\|\theta_{t-1}-\theta^*\|^2+\sigma^2]^2 \\
&\quad + 12n_t^3\mathbb{1}_{\{n_t>1\}}C_\nabla^2\|\theta_{t-1}-\theta^*\|^2[C_l^2\|\theta_{t-1}-\theta^*\|^2+\sigma^2] + 4n_t^4\mathbb{1}_{\{n_t>1\}}C_\nabla^4\|\theta_{t-1}-\theta^*\|^4 \\
&\leq [16C_l^4 n_t + 16C_l^4 n_t^2\mathbb{1}_{\{n_t>1\}} + 12C_\nabla^2 C_l^2 n_t^3\mathbb{1}_{\{n_t>1\}} + 4C_\nabla^4 n_t^4\mathbb{1}_{\{n_t>1\}}]\|\theta_{t-1}-\theta^*\|^4 \\
&\quad + [32C_l^2\sigma^2 n_t^2\mathbb{1}_{\{n_t>1\}} + 12C_\nabla^2\sigma^2 n_t^3\mathbb{1}_{\{n_t>1\}}]\|\theta_{t-1}-\theta^*\|^2 + 16\sigma^4 n_t + 16\sigma^4 n_t^2\mathbb{1}_{\{n_t>1\}}. \qquad (2.7.26)
\end{aligned}$$

Combining the bound from (2.7.24) and (2.7.26) into (2.7.22), we obtain the recursive relation for the fourth-order moment:

$$\begin{aligned}
\mathbb{E}[\|\theta_t-\theta^*\|^4|\mathcal{F}_{t-1}] \leq & [1-4\mu\gamma_t + 8C_\nabla^2\mathbb{1}_{\{n_t>1\}}\gamma_t^2 + 16C_l^2 n_t^{-1}\gamma_t^2 + 48C_l^4 n_t^{-3}\gamma_t^4 + 48C_l^4 n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4 \\
& + 36C_\nabla^2 C_l^2 n_t^{-1}\mathbb{1}_{\{n_t>1\}}\gamma_t^4 + 12C_\nabla^4\mathbb{1}_{\{n_t>1\}}\gamma_t^4]\|\theta_{t-1}-\theta^*\|^4 \\
& + [16\sigma^2 n_t^{-1}\gamma_t^2 + 96C_l^2\sigma^2 n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4 + 36C_\nabla^2\sigma^2 n_t^{-1}\mathbb{1}_{\{n_t>1\}}\gamma_t^4]\|\theta_{t-1}-\theta^*\|^2 \\
& + 48\sigma^4 n_t^{-3}\gamma_t^4 + 48\sigma^4 n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4.
\end{aligned}$$

Using the Young's inequalities, $2C_\nabla^2 C_l^2 \leq n_t C_\nabla^4 + n_t^{-1}C_l^4$, $16\sigma^2 n_t^{-1}\gamma_t^2\|\theta_{t-1}-\theta^*\|^2 \leq 2\mu\gamma_t\|\theta_t-\theta^*\|^4 + 32\sigma^4\mu^{-1}n_t^{-2}\gamma_t^3$, $2C_l^2\sigma^2 n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4\|\theta_{t-1}-\theta^*\|^2 \leq C_l^4 n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4\|\theta_t-\theta^*\|^4 + \sigma^4 n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4$, and $2C_\nabla^2\sigma^2 n_t^{-1}\mathbb{1}_{\{n_t>1\}}\gamma_t^4\|\theta_{t-1}-\theta^*\|^2 \leq C_\nabla^4\mathbb{1}_{\{n_t>1\}}\gamma_t^4\|\theta_t-\theta^*\|^4 + \sigma^4 n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4$, yields,

$$\begin{aligned}
\mathbb{E}[\|\theta_t-\theta^*\|^4|\mathcal{F}_{t-1}] \leq & [1-2\mu\gamma_t + 8C_\nabla^2\mathbb{1}_{\{n_t>1\}}\gamma_t^2 + 16C_l^2 n_t^{-1}\gamma_t^2 + 48C_l^4 n_t^{-3}\gamma_t^4 + 114C_l^4 n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4 \\
& + 48C_\nabla^4\mathbb{1}_{\{n_t>1\}}\gamma_t^4]\|\theta_{t-1}-\theta^*\|^4 + 32\mu^{-1}\sigma^4 n_t^{-2}\gamma_t^3 + 48\sigma^4 n_t^{-3}\gamma_t^4 + 114\sigma^4 n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4. \qquad (2.7.27)
\end{aligned}$$

Taking, the expectation on both sides of the inequality in (2.7.27) yields the recursive relation for

the fourth-order moment:

$$\begin{aligned}\Delta_t \leq &[1 - 2\mu\gamma_t + 8C_\nabla^2 \mathbb{1}_{\{n_t>1\}}\gamma_t^2 + 16C_l^2 n_t^{-1}\gamma_t^2 + 48C_l^4 n_t^{-3}\gamma_t^4 + 114C_l^4 n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4 \\ &+ 48C_\nabla^4 \mathbb{1}_{\{n_t>1\}}\gamma_t^4]\Delta_{t-1} + 32\mu^{-1}\sigma^4 n_t^{-2}\gamma_t^3 + 48\sigma^4 n_t^{-3}\gamma_t^4 + 114\sigma^4 n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4. \end{aligned} \tag{2.7.28}$$

with $\Delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^4]$ for some $\Delta_0 \geq 0$. By Proposition B.1.5, we achieve the (upper) bound of Δ_t in (2.7.28), given as

$$\Delta_t \leq \exp\left(-\mu \sum_{i=t/2}^{t} \gamma_i\right) \Pi_t^\Delta + \frac{32\sigma^4}{\mu^2} \max_{t/2\leq i\leq t} \frac{\gamma_i^2}{n_i^2} + \frac{48\sigma^4}{\mu} \max_{t/2\leq i\leq t} \frac{\gamma_i^3}{n_i^3} + \frac{114\sigma^4}{\mu} \max_{t/2\leq i\leq t} \frac{\gamma_i^3 \mathbb{1}_{\{n_i>1\}}}{n_i^2}.$$

where Π_t^Δ is given by

$$\begin{aligned}&\exp\left(32C_l^2 \sum_{i=1}^{t} \frac{\gamma_i^2}{n_i}\right) \exp\left(96C_l^4 \sum_{i=1}^{t} \frac{\gamma_i^4}{n_i^3}\right) \exp\left(228C_l^4 \sum_{i=1}^{t} \frac{\mathbb{1}_{\{n_i>1\}}\gamma_i^4}{n_i^2}\right) \\ &\exp\left(16C_\nabla^2 \sum_{i=1}^{t} \mathbb{1}_{\{n_i>1\}}\gamma_i^2\right) \exp\left(96C_\nabla^4 \sum_{i=1}^{t} \mathbb{1}_{\{n_i>1\}}\gamma_i^4\right) \left(\Delta_0 + \frac{2\sigma^4}{C_l^4} + \frac{4\sigma^4\gamma_1}{\mu C_l^2 n_1}\right), \end{aligned} \tag{2.7.29}$$

with use of

$$\max_{1\leq i\leq t} \frac{32\mu^{-1}\sigma^4 n_i^{-2}\gamma_i + 48\sigma^4 n_i^{-3}\gamma_i^2 + 114\sigma^4 n_i^{-2}\mathbb{1}_{\{n_i>1\}}\gamma_i^2}{8C_\nabla^2 \mathbb{1}_{\{n_i>1\}} + 16C_l^2 n_i^{-1} + 48C_l^4 n_i^{-3}\gamma_i^2 + 114C_l^4 n_i^{-2}\mathbb{1}_{\{n_i>1\}}\gamma_i^2 + 48C_\nabla^4 \mathbb{1}_{\{n_i>1\}}\gamma_i^2} \leq \frac{\sigma^4}{C_l^4} + \frac{2\sigma^4\gamma_1}{\mu C_l^2 n_1}.$$

At last, bounding the projected estimate (2.2.2) follows from that $\mathbb{E}[\|\mathcal{P}_\Theta(\theta) - \theta^*\|^2] \leq \mathbb{E}[\|\theta - \theta^*\|^2]$, $\forall \theta \in \Theta$. □

Proofs for Section 2.4.1

Proof of Theorem 2.4.1. Following Polyak and Juditsky [118], we rewrite (2.2.1) to

$$\theta_t = \theta_{t-1} - \frac{\gamma_t}{n_t} \sum_{i=1}^{n_t} \nabla_\theta l_{t,i}(\theta_{t-1}) \iff \frac{1}{\gamma_t}(\theta_{t-1} - \theta_t) = \nabla_\theta l_t(\theta_{t-1}), \tag{2.7.30}$$

where $\nabla_\theta l_t(\theta_{t-1})$ denotes $n_t^{-1}\sum_{i=1}^{n_t} \nabla_\theta l_{t,i}(\theta_{t-1})$. Note $\nabla_\theta l_t(\theta_{t-1}) \approx \nabla_\theta l_t(\theta^*) + \nabla_\theta^2 l_t(\theta^*)(\theta_{t-1} - \theta^*)$, and that $\nabla_\theta l_t(\theta^*)$ and $\nabla_\theta l_t(\theta) - \nabla_\theta L(\theta)$ behaves almost like an i.i.d. sequences with zero mean. Thus, $\bar{\theta}_t - \theta^*$ behaves like $-\nabla_\theta L(\theta^*)^{-1} N_t^{-1} \sum_{i=1}^t n_i \nabla_\theta l_i(\theta^*)$ leading to a bound in $\mathcal{O}(\sqrt{N_t})$. Observe that

$$\begin{aligned}\nabla_\theta^2 L(\theta^*)(\theta_{t-1} - \theta^*) =& \nabla_\theta l_t(\theta_{t-1}) - \nabla_\theta l_t(\theta^*) \\ &- \underbrace{[\nabla_\theta l_t(\theta_{t-1}) - \nabla_\theta l_t(\theta^*) - \nabla_\theta L(\theta_{t-1})]}_{\text{martingale term}} - \underbrace{[\nabla_\theta L(\theta_{t-1}) - \nabla_\theta^2 L(\theta^*)(\theta_{t-1} - \theta^*)]}_{\text{rest term}}, \end{aligned}$$

where $\nabla^2_\theta L(\theta^*)$ is invertible with lowest eigenvalue greater than μ, i.e., $\nabla^2_\theta L(\theta^*) \geq \mu$. Thus, summing the parts and using the Minkowski's inequality, we obtain the inequality:

$$\begin{aligned}\left(\mathbb{E}\left[\left\|\bar{\theta}_t - \theta^*\right\|^2\right]\right)^{\frac{1}{2}} \leq & \left(\mathbb{E}\left[\left\|\nabla^2_\theta L\left(\theta^*\right)^{-1} \frac{1}{N_t}\sum_{i=1}^{t} n_i \nabla_\theta l_i\left(\theta^*\right)\right\|^2\right]\right)^{\frac{1}{2}} \\ & + \left(\mathbb{E}\left[\left\|\nabla^2_\theta L\left(\theta^*\right)^{-1} \frac{1}{N_t}\sum_{i=1}^{t} n_i \nabla_\theta l_i\left(\theta_{i-1}\right)\right\|^2\right]\right)^{\frac{1}{2}} \\ & + \left(\mathbb{E}\left[\left\|\nabla^2_\theta L\left(\theta^*\right)^{-1} \frac{1}{N_t}\sum_{i=1}^{t} n_i \left[\nabla_\theta l_i\left(\theta_{i-1}\right) - \nabla_\theta l_i\left(\theta^*\right) - \nabla_\theta L\left(\theta_{i-1}\right)\right]\right\|^2\right]\right)^{\frac{1}{2}} \\ & + \left(\mathbb{E}\left[\left\|\nabla^2_\theta L\left(\theta^*\right)^{-1} \frac{1}{N_t}\sum_{i=1}^{t} n_i \left[\nabla_\theta L\left(\theta_{i-1}\right) - \nabla^2_\theta L\left(\theta^*\right)\left(\theta_{i-1} - \theta^*\right)\right]\right\|^2\right]\right)^{\frac{1}{2}}.\end{aligned}$$

As $(\nabla_\theta l_{t,i}(\theta^*))$ is a square-integrable martingale increment sequences on $\mathbb{R}^d$ (Assumption 2.3.1), we have

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla^2_\theta L\left(\theta^*\right)^{-1} \frac{1}{N_t}\sum_{i=1}^{t} n_i \nabla_\theta l_i\left(\theta^*\right)\right\|^2\right] = & \mathbb{E}\left[\left\|\nabla^2_\theta L\left(\theta^*\right)^{-1} \frac{1}{N_t}\sum_{i=1}^{t}\sum_{j=1}^{n_i} \nabla_\theta l_{i,j}\left(\theta^*\right)\right\|^2\right] \\ \leq & \frac{1}{N_t^2}\sum_{i=1}^{t}\sum_{j=1}^{n_i} \mathbb{E}\left[\left\|\nabla^2_\theta L\left(\theta^*\right)^{-1} \nabla_\theta l_{i,j}\left(\theta^*\right)\right\|^2\right] \\ \leq & \frac{\operatorname{Tr}\left[\nabla^2_\theta L(\theta^*)^{-1}\Sigma\nabla^2_\theta L(\theta^*)^{-1}\right]}{N_t}, \qquad (2.7.31)\end{aligned}$$

using Assumption 2.4.1. To ease notation, we denote $\operatorname{Tr}[\nabla^2_\theta L(\theta^*)^{-1}\Sigma\nabla^2_\theta L(\theta^*)^{-1}]$ by Λ. Next, note that for all $t \geq 1$, we have the relation in (2.7.30), giving us

$$\begin{aligned}\frac{1}{N_t}\sum_{i=1}^{t} n_i \nabla_\theta l_i\left(\theta_{i-1}\right) = & \frac{1}{N_t}\sum_{i=1}^{t} \frac{n_i}{\gamma_i}\left(\theta_{i-1} - \theta_i\right) \\ = & \frac{1}{N_t}\sum_{i=1}^{t-1}\left(\theta_i - \theta^*\right)\left(\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right) - \frac{1}{N_t}\left(\theta_t - \theta^*\right)\frac{n_t}{\gamma_t} + \frac{1}{N_t}\left(\theta_0 - \theta^*\right)\frac{n_1}{\gamma_1},\end{aligned}$$

leading to

$$\begin{aligned}\left\|\nabla^2_\theta L\left(\theta^*\right)^{-1} \frac{1}{N_t}\sum_{i=1}^{t} n_i \nabla_\theta l_i\left(\theta_{i-1}\right)\right\| \leq & \frac{1}{N_t\mu}\sum_{i=1}^{t-1}\left\|\theta_i - \theta^*\right\|\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| \\ & + \frac{1}{N_t\mu}\left\|\theta_t - \theta^*\right\|\frac{n_t}{\gamma_t} + \frac{1}{N_t\mu}\left\|\theta_0 - \theta^*\right\|\frac{n_1}{\gamma_1}.\end{aligned}$$

Hence, with the notion of $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ this expression can be simplified to

$$\left(\mathbb{E}\left[\left\|\nabla_\theta^2 L(\theta^*)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i \nabla_\theta l_i(\theta_{i-1})\right\|^2\right]\right)^{\frac{1}{2}} \leq \frac{1}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{\frac{1}{2}}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| + \frac{n_t}{N_t\gamma_t\mu}\delta_t^{\frac{1}{2}} + \frac{n_1}{N_t\gamma_1\mu}\delta_0^{\frac{1}{2}}. \tag{2.7.32}$$

For the martingale term, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla_\theta^2 L(\theta^*)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i\left[\nabla_\theta l_i(\theta_{i-1}) - \nabla_\theta l_i(\theta^*) - \nabla_\theta L(\theta_{i-1})\right]\right\|^2\right] \\
&\leq \frac{1}{N_t^2\mu^2}\sum_{i=1}^{t} n_i^2\mathbb{E}\left[\|\nabla_\theta l_i(\theta_{i-1}) - \nabla_\theta l_i(\theta^*)\|^2\right] = \frac{1}{N_t^2\mu^2}\sum_{i=1}^{t}\mathbb{E}\left[\left\|\sum_{j=1}^{n_i}\nabla_\theta l_{i,j}(\theta_{i-1}) - \nabla_\theta l_{i,j}(\theta^*)\right\|^2\right] \\
&\leq \frac{1}{N_t^2\mu^2}\sum_{i=1}^{t}\sum_{j=1}^{n_i}\left(\mathbb{E}\left[\|\nabla_\theta l_{i,j}(\theta_{i-1}) - \nabla_\theta l_{i,j}(\theta^*)\|^2\right]\right)^{\frac{1}{2}} \leq \frac{C_l^2}{N_t^2\mu^2}\sum_{i=1}^{t} n_i\delta_{i-1},
\end{aligned} \tag{2.7.33}$$

by the Cauchy-Schwarz inequality and Assumption 2.3.2-p. For all $t \geq 1$, the rest term is directly bounded by (2.2.6):

$$\left(\mathbb{E}\left[\left\|\nabla_\theta^2 L(\theta^*)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i\left[\nabla_\theta L(\theta_{i-1}) - \nabla_\theta^2 L(\theta^*)(\theta_{i-1} - \theta^*)\right]\right\|^2\right]\right)^{\frac{1}{2}} \leq \frac{C_\delta}{N_t\mu}\sum_{i=1}^{t} n_i\Delta_{i-1}^{\frac{1}{2}}, \tag{2.7.34}$$

with the notion $\Delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^4]$. Finally, combining the terms from (2.7.31) to (2.7.34), gives us

$$\begin{aligned}
\bar{\delta}_t^{1/2} \leq &\frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{1}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{1/2}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| + \frac{n_t}{N_t\gamma_t\mu}\delta_t^{1/2} + \frac{n_1}{N_t\gamma_1\mu}\delta_0^{1/2} \\
&+ \frac{C_l}{N_t\mu}\left(\sum_{i=1}^{t} n_i\delta_{i-1}\right)^{1/2} + \frac{C_\delta}{N_t\mu}\sum_{i=1}^{t} n_i\Delta_{i-1}^{1/2},
\end{aligned} \tag{2.7.35}$$

where $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$, which can be simplified into (2.4.11) by shifting the indices and collecting the δ_0 terms. □

Proof of Corollary 2.4.1. As $n_t = C_\rho$ for all $t \geq 1$, we simplify the bound for $\bar{\delta}_t$ in (2.4.11) to

$$\begin{aligned}
\bar{\delta}_t^{1/2} \leq &\frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{C_\rho}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{1/2}\left|\frac{1}{\gamma_{i+1}} - \frac{1}{\gamma_i}\right| + \frac{C_\rho}{N_t\gamma_t\mu}\delta_t^{1/2} + \frac{C_\rho}{N_t\mu}\left(\frac{1}{\gamma_1} + C_l\right)\delta_0^{1/2} \\
&+ \frac{C_l C_\rho^{\frac{1}{2}}}{N_t\mu}\left(\sum_{i=1}^{t-1}\delta_i\right)^{1/2} + \frac{C_\delta C_\rho}{N_t\mu}\sum_{i=0}^{t-1}\Delta_i^{1/2}.
\end{aligned} \tag{2.7.36}$$

The second-order moment δ_t is bounded by Corollary 2.3.1 but with use of (2.7.19) as we work in terms of t. The fourth-order moment Δ_t from Lemma 2.7.1 can be simplified to:

$$\begin{aligned}\Delta_t \leq & \exp\left(-\mu\sum_{i=t/2}^{t}\gamma_i\right)\Pi_\infty^c + \frac{1}{\mu}\left(\frac{32\sigma^4}{\mu C_\rho^2}\max_{t/2\leq i\leq t}\gamma_i^2 + \frac{48\sigma^4}{C_\rho^3}\max_{t/2\leq i\leq t}\gamma_i^3 + \frac{114\sigma^4\mathbb{1}_{\{C_\rho>1\}}}{C_\rho^2}\max_{t/2\leq i\leq t}\gamma_i^3\right)\\ \leq & \exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1-\alpha}}{2^{1-\alpha}}\right)\Pi_\infty^c + \frac{1}{\mu}\left(\frac{2^{2\alpha}32\sigma^4C_\gamma^2C_\rho^{2\beta}}{\mu C_\rho^2 t^{2\alpha}} + \frac{2^{3\alpha}48\sigma^4C_\gamma^3C_\rho^{3\beta}}{C_\rho^3t^{3\alpha}} + \frac{2^{3\alpha}114\sigma^4C_\gamma^3C_\rho^{3\beta}\mathbb{1}_{\{C_\rho>1\}}}{C_\rho^2t^{3\alpha}}\right),\end{aligned}$$

using that $\gamma_t = C_\gamma C_\rho^\beta t^{-\alpha}$ is decreasing as $\alpha \in (1/2, 1)$. Regarding Π_t^Δ defined in (2.7.29), we can bound it by

$$\begin{aligned}\Pi_\infty^c = & \exp\left(\frac{64\alpha C_l^2C_\gamma^2C_\rho^{2\beta}}{(2\alpha-1)\,C_\rho}\right)\exp\left(\frac{(192+456C_\rho\mathbb{1}_{\{C_\rho>1\}})C_l^4C_\gamma^4C_\rho^{4\beta}}{C_\rho^3}\right)\exp\left(\frac{32\alpha C_\nabla^2C_\gamma^2C_\rho^{2\beta}\mathbb{1}_{\{C_\rho>1\}}}{2\alpha-1}\right)\\ & \exp\left(192C_\nabla^4C_\gamma^4C_\rho^{4\beta}\mathbb{1}_{\{C_\rho>1\}}\right)\left(\Delta_0 + \frac{2\sigma^4}{C_l^4} + \frac{4\sigma^4C_\gamma}{\mu C_l^2C_\rho^{1-\beta}}\right),\end{aligned}$$

using $\sum_{i=1}^t i^{-2\alpha} \leq 2\alpha/(2\alpha-1)$ and $\sum_{i=1}^t i^{-4\alpha} \leq 2$. Note that Π_∞^c is a finite constant, independent of t. To bound the first term of (2.7.36), namely $\frac{C_\rho}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{1/2}|\gamma_{i+1}^{-1} - \gamma_i^{-1}|$, we remark that $|\gamma_{t+1}^{-1} - \gamma_t^{-1}| \leq C_\gamma^{-1}C_\rho^{-\beta}\alpha t^{\alpha-1}$, one has (since $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$),

$$\frac{C_\rho}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{\frac{1}{2}}\left|\frac{1}{\gamma_{i+1}} - \frac{1}{\gamma_i}\right| \leq \frac{C_\rho^{1-\beta}\alpha}{C_\gamma\mu N_t}\sum_{i=1}^{t} i^{\alpha-1}\left(\exp\left(-\frac{\mu C_\gamma C_\rho^\beta i^{1-\alpha}}{2^{2-\alpha}}\right)\sqrt{\pi_\infty^c} + \frac{2^{\frac{1+\alpha}{2}}\sigma\sqrt{C_\gamma}}{\sqrt{\mu}C_\rho^{\frac{1-\beta}{2}}i^{\alpha/2}}\right). \tag{2.7.37}$$

For simplicity, let us denote

$$A_\infty^c = \sum_{i=0}^{\infty}\exp\left(-\frac{\mu C_\gamma C_\rho^\beta i^{1-\alpha}}{2^{2-\alpha}}\right) \geq \sum_{i=0}^{\infty} i^{\alpha-1}\exp\left(-\frac{\mu C_\gamma C_\rho^\beta i^{1-\alpha}}{2^{2-\alpha}}\right),$$

as $\alpha < 1$. Thus, the first part of (2.7.37) is bounded as follows:

$$\frac{C_\rho^{1-\beta}\alpha\sqrt{\pi_\infty^c}}{C_\gamma\mu N_t}\sum_{i=1}^{t} i^{\alpha-1}\exp\left(-\frac{\mu C_\gamma C_\rho^\beta i^{1-\alpha}}{2^{2-\alpha}}\right) \leq \frac{C_\rho^{1-\beta}\alpha\sqrt{\pi_\infty^c}A_\infty^c}{C_\gamma\mu N_t}.$$

Furthermore, with the help of an integral test for convergence, one has $\sum_{i=1}^t i^{\alpha/2-1} \leq 1 + \int_1^t s^{\alpha/2-1}\,ds = 1 + (2/\alpha)t^{\alpha/2} - (2/\alpha) \leq (2/\alpha)t^{\alpha/2}$, such that the second part of (2.7.37) can be bounded by

$$\frac{2^{\frac{1+\alpha}{2}}\sigma C_\rho^{\frac{1-\beta}{2}}\alpha}{C_\gamma^{1/2}\mu^{3/2}N_t}\sum_{i=1}^{t} i^{\alpha/2-1} \leq \frac{2^{\frac{3+\alpha}{2}}\sigma C_\rho^{\frac{1-\beta}{2}}t^{\alpha/2}}{C_\gamma^{1/2}\mu^{3/2}N_t} = \frac{2^{\frac{3+\alpha}{2}}\sigma C_\rho^{\frac{1-\alpha-\beta}{2}}}{C_\gamma^{1/2}\mu^{3/2}N_t^{1-\alpha/2}}.$$

By combining this, we get

$$\frac{C_\rho}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{\frac{1}{2}}\left|\frac{1}{\gamma_{i+1}}-\frac{1}{\gamma_i}\right|\leq\frac{C_\rho^{1-\beta}\alpha\sqrt{\pi_\infty^c}A_\infty^c}{C_\gamma\mu N_t}+\frac{2^{\frac{3+\alpha}{2}}\sigma C_\rho^{\frac{1-\alpha-\beta}{2}}}{\sqrt{C_\gamma}\mu^{3/2}N_t^{1-\alpha/2}}. \tag{2.7.38}$$

Similarly, second term of (2.7.36), can be bounded by

$$\begin{aligned}\frac{C_\rho}{N_t\gamma_t\mu}\delta_t^{\frac{1}{2}}\leq&\frac{C_\rho^{1-\alpha-\beta}}{C_\gamma\mu N_t^{1-\alpha}}\left(\exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1-\alpha}}{2^{2-\alpha}}\right)\sqrt{\pi_\infty^c}+\frac{2^{\frac{1+\alpha}{2}}\sigma\sqrt{C_\gamma}}{\sqrt{\mu}C_\rho^{\frac{1-\beta}{2}}t^{\alpha/2}}\right)\\\leq&\frac{C_\rho^{2-\alpha-\beta}\sqrt{\pi_\infty^c}A_\infty^c}{C_\gamma\mu N_t^{2-\alpha}}+\frac{2^{\frac{1+\alpha}{2}}C_\rho^{\frac{1-\alpha-\beta}{2}}\sigma}{\sqrt{C_\gamma}\mu^{3/2}N_t^{1-\alpha/2}},\end{aligned}$$

using $\exp(-\mu C_\gamma C_\rho^\beta t^{1-\alpha}/2^{2-\alpha})=A_t^c\leq t^{-1}\sum_{i=1}^t A_i^c\leq t^{-1}A_\infty^c$ as A_t^c is decreasing. In a same way, one has

$$\frac{C_lC_\rho^{\frac{1}{2}}}{N_t\mu}\left(\sum_{i=1}^{t-1}\delta_i\right)^{\frac{1}{2}}\leq\frac{C_lC_\rho^{\frac{1}{2}}}{N_t\mu}\left(A_\infty^c\pi_\infty^c+\frac{2^{1+\alpha}\sigma^2C_\gamma t^{1-\alpha}}{(1-\alpha)\,\mu C_\rho^{1-\beta}}\right)^{1/2}\leq\frac{C_lC_\rho^{\frac{1}{2}}\sqrt{\pi_\infty^c}\sqrt{A_\infty^c}}{N_t\mu}+\frac{2^{\frac{1+\alpha}{2}}C_l\sigma\sqrt{C_\gamma}}{C_\rho^{\frac{1-\alpha-\beta}{2}}\mu^{3/2}N_t^{\frac{1+\alpha}{2}}}.$$

Bound the last term of (2.7.36), is done as follows,

$$\begin{aligned}\frac{C_\delta C_\rho}{N_t\mu}\sum_{i=0}^{t-1}\Delta_i^{\frac{1}{2}}\leq&\frac{C_\delta C_\rho}{N_t\mu}\sum_{i=0}^{t-1}\exp\left(-\frac{\mu C_\gamma C_\rho^\beta i^{1-\alpha}}{2^{2-\alpha}}\right)\sqrt{\Pi_\infty^c}+\frac{2^\alpha 6C_\delta\sigma^2C_\gamma C_\rho^\beta}{N_t\mu^2}\sum_{i=1}^{t-1}i^{-\alpha}\\&+\frac{(6+7\mathbb{1}_{\{C_\rho>1\}})2^{3\alpha/2}C_\delta\sigma^2C_\gamma^{3/2}C_\rho^{3\beta/2}}{N_t\mu^{3/2}}\sum_{i=1}^{t-1}i^{-3\alpha/2}\\\leq&\frac{C_\delta C_\rho\sqrt{\Pi_\infty^c}A_\infty^c}{N_t\mu}+\frac{2^\alpha 6C_\delta\sigma^2C_\gamma}{C_\rho^{1-\alpha-\beta}\mu^2N_t^\alpha}+\frac{(6+7\mathbb{1}_{\{C_\rho>1\}})2^{3\alpha/2}C_\delta\sigma^2C_\gamma^{3/2}C_\rho^{3\beta/2}\psi^0_{3\alpha/2}(N_t/C_\rho)}{\mu^{3/2}N_t}.\end{aligned}$$

Thus, by collecting the terms above, we obtain:

$$\begin{aligned}\bar{\delta}_t^{1/2}\leq&\frac{\Lambda^{1/2}}{N_t^{1/2}}+\frac{6\sigma C_\rho^{\frac{1-\alpha-\beta}{2}}}{\sqrt{C_\gamma}\mu^{3/2}N_t^{1-\alpha/2}}+\frac{2^\alpha 6C_\delta\sigma^2C_\gamma}{C_\rho^{1-\alpha-\beta}\mu^2N_t^\alpha}+\frac{C_\rho^{2-\alpha-\beta}\sqrt{\pi_\infty^c}A_\infty^c}{C_\gamma\mu N_t^{2-\alpha}}+\frac{2^{\frac{1+\alpha}{2}}C_l\sigma\sqrt{C_\gamma}}{C_\rho^{\frac{1-\alpha-\beta}{2}}\mu^{3/2}N_t^{\frac{1+\alpha}{2}}}\\&+\frac{C_\rho\Gamma_c}{\mu N_t}+\frac{(6+7\mathbb{1}_{\{C_\rho>1\}})2^{3\alpha/2}C_\delta\sigma^2C_\gamma^{3/2}C_\rho^{3\beta/2}}{\mu^{3/2}\psi^0_{3\alpha/2}(N_t/C_\rho)^{-1}N_t},\end{aligned}$$

where $\Gamma_c=(1/C_\gamma C_\rho^\beta+C_l)\delta_0^{1/2}+C_l\sqrt{\pi_\infty^cA_\infty^c}/C_\rho^{1/2}+\sqrt{\pi_\infty^c}A_\infty^c/C_\gamma C_\rho^\beta+C_\delta\sqrt{\Pi_\infty^c}A_\infty^c$. □

Proof of Corollary 2.4.2. The steps of the proof follows the ones of Corollary 2.4.1 with the smart

notation of ϕ and $\tilde{\rho}$: The bound for $\bar{\delta}_t$ in (2.4.11) is given by

$$\begin{aligned}\bar{\delta}_t^{1/2} \leq & \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{1}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{1/2}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| + \frac{n_t}{N_t\gamma_t\mu}\delta_t^{1/2} + \frac{n_1}{N_t\mu}\left(\frac{1}{\gamma_1} + C_l\right)\delta_0^{1/2} \\ & + \frac{C_l}{N_t\mu}\left(\sum_{i=1}^{t-1} n_{i+1}\delta_i\right)^{1/2} + \frac{C_\delta}{N_t\mu}\sum_{i=0}^{t-1} n_{i+1}\Delta_i^{1/2}, \end{aligned} \tag{2.7.39}$$

where the learning rate is on the form $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ with $n_t = C_\rho t^\rho$. The second-order moment δ_t is upper bounded by (2.7.20) from Corollary 2.3.2. The fourth-order moment Δ_t from Lemma 2.7.1 can be simplified as follows,

$$\Delta_t \leq \exp\left(-\mu\sum_{i=t/2}^{t}\gamma_i\right)\Pi_\infty^v + \frac{32\sigma^4}{\mu^2}\max_{t/2\leq i\leq t}\frac{\gamma_i^2}{n_i^2} + \frac{162\sigma^4}{\mu}\max_{t/2\leq i\leq t}\frac{\gamma_i^3}{n_i^2},$$

as $n_t \geq 1$ for any $t \geq 1$ and $\beta \leq 1$, and

$$\Pi_\infty^v = \exp\left(\frac{32(\alpha-\beta\tilde{\rho})C_\gamma^2 C_\rho^{2\beta}(2C_l^2 + C_\nabla^2)}{2(\alpha-\beta\tilde{\rho})-1}\right)\exp\left(192C_\gamma^4 C_\rho^{4\beta}(4C_l^4 + C_\nabla^4)\right)\left(\Delta_0 + \frac{2\sigma^4}{C_l^4} + \frac{4\sigma^4 C_\gamma}{\mu C_l^2 C_\rho^{1-\beta}},\right)$$

using that $\sum_{i=1}^t i^{-a} \leq 2$ for $a \geq 2$. Next, for $\rho \geq 0$, we have

$$\Delta_t \leq \exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1+\beta\rho-\alpha}}{2^{1+\beta\rho-\alpha}}\right)\Pi_\infty^v + \frac{2^{2\alpha-2\beta\rho+2\rho}32\sigma^4 C_\gamma^2 C_\rho^{2\beta}}{\mu^2 C_\rho^2 t^{2\alpha-2\beta\rho+2\rho}} + \frac{2^{3\alpha-3\beta\rho+2\rho}162\sigma^4 C_\gamma^3 C_\rho^{3\beta}}{\mu C_\rho^2 t^{3\alpha-3\beta\rho+2\rho}},$$

using that $\alpha - \beta\rho \in (1/2, 1)$. If $\rho < 0$, one directly have

$$\Delta_t \leq \exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1-\alpha}}{2^{1-\alpha}}\right)\Pi_\infty^v + \frac{2^{2\alpha}32\sigma^4 C_\gamma^2 C_\rho^{2\beta}}{\mu^2 t^{2\alpha}} + \frac{2^{3\alpha}162\sigma^4 C_\gamma^3 C_\rho^{3\beta}}{\mu t^{3\alpha}}.$$

With the notion of ϕ and $\tilde{\rho}$, we can combine the two ρ-cases as follows:

$$\Delta_t \leq \exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} t^{(1-\phi)(1+\tilde{\rho})}}{2^{(1-\phi)(1+\tilde{\rho})}}\right)\Pi_\infty^v + \frac{2^{2\phi(1+\tilde{\rho})}32\sigma^4 C_\gamma^2 C_\rho^{2\beta}}{\mu^2 C_\rho^{2\mathbb{1}_{\{\rho\geq 0\}}} t^{2\phi(1+\tilde{\rho})}} + \frac{2^{3\phi(1+\tilde{\rho})-\tilde{\rho}}162\sigma^4 C_\gamma^3 C_\rho^{3\beta}}{\mu C_\rho^{2\mathbb{1}_{\{\rho\geq 0\}}} t^{3\phi(1+\tilde{\rho})-\tilde{\rho}}}.$$

We will in the following bound the terms for t but afterwards we will translate it to terms in N_t. If $\rho \geq 0$, the first relation is $t \geq (N_t/2C_\rho)^{1/(1+\rho)}$ since $N_t = C_\rho(t^\rho + \sum_{i=1}^{t-1} i^\rho) \leq C_\rho(t^\rho + \int_1^t x^\rho\,dx) \leq C_\rho(t^\rho + t^\rho\int_1^t dx) = C_\rho(t^\rho + t^{1+\rho}) \leq 2C_\rho t^{1+\rho}$ by use the integral test for convergence. Similarly, $N_t = C_\rho\sum_{i=1}^t i^\rho \geq C_\rho\int_0^t x^\rho\,dx = C_\rho t^{\rho+1}$, thus, $t \leq (N_t/C_\rho)^{1/(1+\rho)}$. If $\rho < 0$, one has $t \leq N_t$ and $N_t \leq C_\rho t$, i.e., $t \geq N_t/C_\rho$.

Bounding $\frac{1}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{1/2}|n_{i+1}/\gamma_{i+1} - n_i/\gamma_i|$, we first note $n_t/\gamma_t = C_\gamma^{-1}C_\rho^{1-\beta}t^{(1-\beta)\rho+\alpha}$ for $\rho \geq 0$.

Thus, by the mean value theorem, we obtain:

$$\left|\frac{n_{i+1}}{\gamma_{i+1}}-\frac{n_i}{\gamma_i}\right| \leq ((1-\beta)\rho+\alpha)\,\frac{C_\rho^{1-\beta}}{C_\gamma}\sup_{\nu\in(i,i+1)}\left|\nu^{(1-\beta)\rho+\alpha-1}\right| \leq \frac{((1-\beta)\rho+\alpha)\,C_\rho^{1-\beta}}{C_\gamma i^{1-(1-\beta)\rho-\alpha}}, \tag{2.7.40}$$

as $\alpha+(1-\beta)\rho \leq 1-\rho$ since $\alpha-\beta\rho \in (1/2,1)$. For $\rho<0$, the mean value theorem gives us

$$\begin{aligned}\left|\frac{n_{i+1}}{\gamma_{i+1}}-\frac{n_i}{\gamma_i}\right| =& \frac{1}{C_\gamma}\left|n_{i+1}^{1-\beta}(i+1)^\alpha - n_i^{1-\beta}i^\alpha\right| \leq \frac{C_\rho^{1-\beta}}{C_\gamma}\left|(i+1)^\alpha - i^\alpha\right| \\ \leq& \frac{\alpha C_\rho^{1-\beta}}{C_\gamma}\sup_{\nu\in(i,i+1)}\left|\nu^{\alpha-1}\right| \leq \frac{\alpha C_\rho^{1-\beta}}{C_\gamma i^{1-\alpha}},\end{aligned}$$

as $(n_t)_{t\geq 1}$ is a decreasing sequence and $\beta \leq 1$. Thus, for any $\rho \in (-1,1)$, we have

$$\left|\frac{n_{i+1}}{\gamma_{i+1}}-\frac{n_i}{\gamma_i}\right| \leq \frac{\phi(1+\tilde{\rho})C_\rho^{1-\beta}}{C_\gamma i^{1-\phi(1+\tilde{\rho})}}.$$

By using this, we obtain:

$$\begin{aligned}\frac{1}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{\frac{1}{2}}\left|\frac{n_{i+1}}{\gamma_{i+1}}-\frac{n_i}{\gamma_i}\right| \leq& \frac{\phi(1+\tilde{\rho})C_\rho^{1-\beta}}{N_t\mu C_\gamma}\sum_{i=1}^{t} i^{\phi(1+\tilde{\rho})-1} \\ & \left(\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} i^{(1-\phi)(1+\tilde{\rho})}}{2^{1+(1-\phi)(1+\tilde{\rho})}}\right)\sqrt{\pi_\infty^v} + \frac{2^{\frac{1+\phi(1+\tilde{\rho})}{2}}\sigma\sqrt{C_\gamma}}{\sqrt{\mu}C_\rho^{\frac{(1-\beta)}{2}\mathbb{1}_{\{\rho\geq 0\}}} i^{\frac{\phi(1+\tilde{\rho})}{2}}}\right).\end{aligned}$$

Next, let us denote

$$A_\infty^v = \sum_{i=0}^{\infty} i^{\tilde{\rho}}\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} i^{(1-\phi)(1+\tilde{\rho})}}{2^{1+(1-\phi)(1+\tilde{\rho})}}\right) \geq \sum_{i=0}^{\infty} i^{\phi(1+\tilde{\rho})-1}\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} i^{(1-\phi)(1+\tilde{\rho})}}{2^{1+(1-\phi)(1+\tilde{\rho})}}\right),$$

since $\phi(1+\tilde{\rho})-1 = \alpha+(1-\beta)\tilde{\rho}-1 \leq \tilde{\rho}$. Thus,

$$\frac{\phi(1+\tilde{\rho})C_\rho^{1-\beta}\sqrt{\pi_\infty^v}}{N_t\mu C_\gamma}\sum_{i=1}^{t} i^{\phi(1+\tilde{\rho})-1}\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} i^{(1-\phi)(1+\tilde{\rho})}}{2^{1+(1-\phi)(1+\tilde{\rho})}}\right) \leq \frac{\phi(1+\tilde{\rho})C_\rho^{1-\beta}\sqrt{\pi_\infty^v}A_\infty^v}{N_t\mu C_\gamma}.$$

Furthermore, with the help of an integral test for convergence, we have

$$\begin{aligned}\frac{\phi(1+\tilde{\rho})2^{\frac{1+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma}N_t}\sum_{i=1}^{t} i^{\frac{\phi(1+\tilde{\rho})}{2}-1} \leq& \frac{2^{\frac{3+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}} t^{\frac{\phi(1+\tilde{\rho})}{2}}}{\mu^{3/2}\sqrt{C_\gamma}N_t} \\ \leq& \frac{2^{\frac{3+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma}N_t^{1-\phi/2}}.\end{aligned}$$

Summarising, with use of $\phi(1+\tilde{\rho}) < 2$, we obtain

$$\begin{aligned}\frac{1}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{\frac{1}{2}}\left|\frac{n_{i+1}}{\gamma_{i+1}}-\frac{n_i}{\gamma_i}\right| \leq & \frac{\phi(1+\tilde{\rho})C_\rho^{1-\beta}\sqrt{\pi_\infty^v}A_\infty^v}{N_t\mu C_\gamma}+\frac{2^{\frac{3+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma}N_t^{1-\phi/2}} \\ \leq & \frac{2C_\rho^{1-\beta}\sqrt{\pi_\infty^v}A_\infty^v}{\mu C_\gamma N_t}+\frac{2^{\frac{3+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma}N_t^{1-\phi/2}}.\end{aligned}$$

Similarly, for $\frac{n_t}{N_t\gamma_t\mu}\delta_t^{1/2}$, one have

$$\begin{aligned}\frac{n_t}{N_t\gamma_t\mu}\delta_t^{\frac{1}{2}} \leq & \frac{C_\rho^{1-\beta}\sqrt{\pi_\infty^v}t^{\phi(1+\tilde{\rho})}}{N_tC_\gamma\mu}\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}}t^{(1-\phi)(1+\tilde{\rho})}}{2^{1+(1-\phi)(1+\tilde{\rho})}}\right)+\frac{2^{\frac{1+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}t^{\frac{\phi(1+\tilde{\rho})}{2}}}{\mu^{3/2}\sqrt{C_\gamma}N_t} \\ \leq & \frac{C_\rho^{2-\phi-\beta}\sqrt{\pi_\infty^v}A_\infty^v}{\mu C_\gamma N_t^{2-\phi}}+\frac{2^{\frac{1+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma}N_t^{1-\phi/2}}.\end{aligned}$$

For $\frac{n_1}{N_t\mu}(\gamma_1^{-1}+C_l)\delta_0^{1/2}$, we insert the definition of our learning functions, giving us

$$\frac{n_1}{N_t\mu}\left(\frac{1}{\gamma_1}+C_l\right)\delta_0^{1/2}=\frac{C_\rho}{N_t\mu}\left(\frac{1}{C_\gamma C_\rho^\beta}+C_l\right)\delta_0^{1/2}.$$

Bounding $\frac{C_l}{N_t\mu}(\sum_{i=1}^{t-1}n_{i+1}\delta_i)^{1/2}$, follows the ideas from above, using that $n_{t+1}\leq 2^{\tilde{\rho}}n_t$, to obtain

$$\begin{aligned}&\frac{C_l}{N_t\mu}\left(\sum_{i=1}^{t-1}n_{i+1}\delta_i\right)^{\frac{1}{2}} \leq \frac{2^{\tilde{\rho}/2}C_l}{N_t\mu}\left(\sum_{i=1}^{t}n_i\delta_i\right)^{\frac{1}{2}} \\ \leq & \frac{2^{\tilde{\rho}/2}C_l}{N_t\mu}\left(C_\rho\sum_{i=1}^{t}i^{\tilde{\rho}}\left(\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}}i^{(1-\phi)(1+\tilde{\rho})}}{2^{(1-\phi)(1+\tilde{\rho})}}\right)\pi_\infty^v+\frac{2^{1+\phi(1+\tilde{\rho})}\sigma^2C_\gamma}{\mu C_\rho^{(1-\beta)\mathbb{1}_{\{\rho\geq 0\}}}i^{\phi(1+\tilde{\rho})}}\right)\right)^{\frac{1}{2}} \\ = & \frac{2^{\tilde{\rho}/2}C_l}{N_t\mu}\left(C_\rho\pi_\infty^v\sum_{i=1}^{t}i^{\tilde{\rho}}\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}}i^{(1-\phi)(1+\tilde{\rho})}}{2^{(1-\phi)(1+\tilde{\rho})}}\right)+\frac{2^{1+\phi(1+\tilde{\rho})}\sigma^2C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}}}{\mu}\sum_{i=1}^{t}i^{\beta\tilde{\rho}-\alpha}\right)^{\frac{1}{2}} \\ \leq & \frac{2^{\tilde{\rho}/2}C_l}{N_t\mu}\left(C_\rho\pi_\infty^vA_\infty^v+\frac{2^{\phi(1+\tilde{\rho})}\sigma^2C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}}t^{(1-\phi)(1+\tilde{\rho})}}{\mu}\right)^{\frac{1}{2}} \\ \leq & \frac{2^{\tilde{\rho}/2}C_l\sqrt{C_\rho}\sqrt{\pi_\infty^v}\sqrt{A_\infty^v}}{\mu N_t}+\frac{2^{\frac{\phi(1+\tilde{\rho})}{2}}C_l\sigma\sqrt{C_\gamma}C_\rho^{\beta/2\mathbb{1}_{\{\rho\geq 0\}}}t^{\frac{(1-\phi)(1+\tilde{\rho})}{2}}}{\mu^{3/2}N_t} \\ \leq & \frac{2^{\tilde{\rho}/2}C_l\sqrt{C_\rho}\sqrt{\pi_\infty^v}\sqrt{A_\infty^v}}{\mu N_t}+\frac{2^{\frac{\phi(1+\tilde{\rho})}{2}}C_l\sigma\sqrt{C_\gamma}}{\mu^{3/2}C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}N_t^{\frac{1+\phi}{2}}}.\end{aligned}$$

Likewise, for $\frac{C_\delta}{N_t\mu}\sum_{i=0}^{t-1} n_{i+1}\Delta_i^{1/2}$, we get

$$\begin{aligned}
&\frac{C_\delta}{N_t\mu}\sum_{i=0}^{t-1} n_{i+1}\Delta_i^{\frac{1}{2}} \\
&\leq \frac{2^{\tilde\rho}C_\delta C_\rho}{N_t\mu}\sum_{i=1}^{t-1} i^{\tilde\rho}\left(\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} i^{(1-\phi)(1+\tilde\rho)}}{2^{(1-\phi)(1+\tilde\rho)}}\right)\Pi_\infty^v + \frac{2^{2\phi(1+\tilde\rho)}32\sigma^4 C_\gamma^2 C_\rho^{2\beta}}{\mu^2 C_\rho^{2\mathbb{1}_{\{\rho\geq 0\}}} i^{2\phi(1+\tilde\rho)}} + \frac{2^{3\phi(1+\tilde\rho)-\tilde\rho}162\sigma^4 C_\gamma^3 C_\rho^{3\beta}}{\mu C_\rho^{2\mathbb{1}_{\{\rho\geq 0\}}} i^{3\phi(1+\tilde\rho)-\tilde\rho}}\right)^{\frac{1}{2}} \\
&\leq \frac{2^{\tilde\rho}C_\delta C_\rho}{N_t\mu}\sum_{i=1}^{t-1} i^{\tilde\rho}\left(\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} i^{(1-\phi)(1+\tilde\rho)}}{2^{1+(1-\phi)(1+\tilde\rho)}}\right)\sqrt{\Pi_\infty^v} + \frac{2^{\phi(1+\tilde\rho)}6\sigma^2 C_\gamma C_\rho^{\beta}}{\mu C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} i^{\phi(1+\tilde\rho)}} + \frac{2^{3\phi(1+\tilde\rho)/2-\tilde\rho/2}13\sigma^2 C_\gamma^{3/2} C_\rho^{3\beta/2}}{\mu^{1/2} C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} i^{3\phi(1+\tilde\rho)/2}}\right) \\
&\leq \frac{2^{\tilde\rho}C_\delta C_\rho\sqrt{\Pi_\infty^v}A_\infty^v}{\mu N_t} + \frac{2^{\phi(1+\tilde\rho)+\tilde\rho}C_\delta\sigma^2 C_\gamma C_\rho^{1+\beta}}{\mu^2 C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} N_t}\sum_{i=1}^{t-1} i^{\beta\tilde\rho-\alpha} + \frac{2^{3\phi(1+\tilde\rho)/2+\tilde\rho/2}C_\delta\sigma^2 C_\gamma^{3/2} C_\rho^{1+3\beta/2}}{\mu^{3/2} C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} N_t}\sum_{i=1}^{t-1} i^{3(\beta\tilde\rho-\alpha)/2},
\end{aligned}$$

where the second term can be bounded as

$$\begin{aligned}
\frac{2^{(1+\phi)(1+\tilde\rho)-1}C_\delta\sigma^2 C_\gamma C_\rho^{1+\beta}}{\mu^2 C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} N_t}\sum_{i=1}^{t-1} i^{\beta\tilde\rho-\alpha} \leq& \frac{2^{(1+\phi)(1+\tilde\rho)-1}C_\delta\sigma^2 C_\gamma C_\rho^{1+\beta} t^{1+\beta\tilde\rho-\alpha}}{(1+\beta\tilde\rho-\alpha)\mu^2 C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} N_t} \\
\leq& \frac{2^{(1+\phi)(1+\tilde\rho)-2}C_\delta\sigma^2 C_\gamma}{\mu^2 C_\rho^{1-\phi-\beta} N_t^\phi},
\end{aligned}$$

and the third term by

$$\frac{2^{3(1+\phi)(1+\tilde\rho)/2}C_\delta\sigma^2 C_\gamma^{3/2} C_\rho^{1+3\beta/2}}{\mu^{3/2} C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} N_t}\sum_{i=1}^{t-1} i^{3(\beta\tilde\rho-\alpha)/2} \leq \frac{2^{3(1+\phi)(1+\tilde\rho)/2}C_\delta\sigma^2 C_\gamma^{3/2} C_\rho^{1+3\beta/2}\psi^{\tilde\rho}_{3(\alpha-\beta\tilde\rho)/2}(N_t/C_\rho)}{\mu^{3/2} C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} N_t}.$$

By collecting these bounds, we get

$$\begin{aligned}
\frac{C_\delta}{N_t\mu}\sum_{i=0}^{t-1} n_{i+1}\Delta_i^{\frac{1}{2}} \leq& \frac{2^{\tilde\rho}C_\delta C_\rho\sqrt{\Pi_\infty^v}A_\infty^v}{\mu N_t} + \frac{2^{(1+\phi)(1+\tilde\rho)-2}C_\delta\sigma^2 C_\gamma}{\mu^2 C_\rho^{1-\phi-\beta} N_t^\phi} \\
&+ \frac{2^{3(1+\phi)(1+\tilde\rho)/2}C_\delta\sigma^2 C_\gamma^{3/2} C_\rho^{1+3\beta/2}\psi^{\tilde\rho}_{3(\alpha-\beta\tilde\rho)/2}(N_t/C_\rho)}{\mu^{3/2} C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} N_t}.
\end{aligned}$$

Combining our findings from above, we have

$$\begin{aligned}
\bar\delta_t^{1/2} \leq& \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{2C_\rho^{1-\beta}\sqrt{\pi_\infty^v}A_\infty^v}{\mu C_\gamma N_t} + \frac{2^{\frac{3+\phi(1+\tilde\rho)}{2}}\sigma C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma}N_t^{1-\phi/2}} + \frac{C_\rho^{2-\phi-\beta}\sqrt{\pi_\infty^v}A_\infty^v}{\mu C_\gamma N_t^{2-\phi}} + \frac{2^{\frac{1+\phi(1+\tilde\rho)}{2}}\sigma C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma}N_t^{1-\phi/2}} \\
&+ \frac{C_\rho}{N_t\mu}\left(\frac{1}{C_\gamma C_\rho^\beta} + C_l\right)\delta_0^{\frac{1}{2}} + \frac{2^{\tilde\rho/2}C_l\sqrt{C_\rho}\sqrt{\pi_\infty^v}\sqrt{A_\infty^v}}{\mu N_t} + \frac{2^{\frac{\phi(1+\tilde\rho)}{2}}C_l\sigma\sqrt{C_\gamma}}{\mu^{3/2}C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}N_t^{\frac{1+\phi}{2}}} + \frac{2^{\tilde\rho}C_\delta C_\rho\sqrt{\Pi_\infty^v}A_\infty^v}{\mu N_t} \\
&+ \frac{2^{(1+\phi)(1+\tilde\rho)-2}C_\delta\sigma^2 C_\gamma}{\mu^2 C_\rho^{1-\phi-\beta} N_t^\phi} + \frac{2^{3(1+\phi)(1+\tilde\rho)/2}C_\delta\sigma^2 C_\gamma^{3/2} C_\rho^{1+3\beta/2}\psi^{\tilde\rho}_{3(\alpha-\beta\tilde\rho)/2}(N_t/C_\rho)}{\mu^{3/2} C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} N_t}.
\end{aligned}$$

This can be simplified to the desired using Γ_v given by $(1/C_\gamma C_\rho^\beta + C_l)\delta_0^{1/2} + 2\tilde{\rho}C_l\sqrt{\pi_\infty^v A_\infty^v}/C_\rho^{1/2} + 2\sqrt{\pi_\infty^v}A_\infty^v/C_\gamma C_\rho^\beta + 2\tilde{\rho}C_\delta\sqrt{\Pi_\infty^v}A_\infty^v$, consisting of the finite constants π_∞^v, Π_∞^v and A_∞^v. □

Proofs for Section 2.4.2

Theorem 2.7.1 (PASSG). *Denote $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ with $(\bar{\theta}_t)$ given by (2.2.3) using (θ_t) from (2.2.2). Under Assumption 2.3.1, Assumptions 2.3.2-p and 2.3.3-p with $p = 4$, Assumptions 2.4.1 and 2.4.2, we have for any learning rate (γ_t) that*

$$\begin{aligned}\bar{\delta}_t^{1/2} \leq & \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{1}{N_t\mu}\sum_{i=1}^{t-1}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right|\delta_i^{1/2} + \frac{n_t}{N_t\gamma_t\mu}\delta_t^{1/2} + \frac{n_1}{N_t\mu}\left(\frac{1}{\gamma_1} + C_l\right)\delta_0^{1/2} \\ & + \frac{C_l}{N_t\mu}\left(\sum_{i=1}^{t-1} n_{i+1}\delta_i\right)^{1/2} + \frac{C'_\delta}{N_t\mu}\sum_{i=0}^{t} n_{i+1}\Delta_i^{1/2}\end{aligned}$$

where $\Lambda = \mathrm{Tr}(\nabla_\theta^2 L(\theta^*)^{-1}\Sigma\nabla_\theta^2 L(\theta^*)^{-1})$ and $C'_\delta = C_\delta + 2^2 G_\Theta/d_{\min}^2$.

Proof of Theorem 2.7.1. Denote $\mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ by $\bar{\delta}_t$ with $(\bar{\theta}_t)$ given by (2.2.3) using (θ_t) from (2.2.2). As in the proof Theorem 2.4.1, we follow the steps of Polyak and Juditsky [118], in which, we can rewrite (2.2.2) to

$$\frac{1}{\gamma_t}\left(\theta_{t-1} - \theta_t\right) = \nabla_\theta l_t\left(\theta_{t-1}\right) - \frac{1}{\gamma_t}\Omega_t,$$

where $\nabla_\theta l_t(\theta_{t-1}) = n_t^{-1}\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}(\theta_{t-1})$ and $\Omega_t = \mathcal{P}_\Theta(\theta_{t-1} - \gamma_t\nabla_\theta l_t(\theta_{t-1})) - (\theta_{t-1} - \gamma_t\nabla_\theta l_t(\theta_{t-1}))$. Thus, summing the parts, using the Minkowski's inequality, and bounding each term gives us the same bound as in Theorem 2.4.1, but with an additional term regarding Ω_t, namely

$$\begin{aligned}\left(\mathbb{E}\left[\left\|\nabla_\theta^2 L\left(\theta^*\right)^{-1}\frac{1}{N_t}\sum_{i=1}^{t}\frac{n_i}{\gamma_i}\Omega_i\right\|^2\right]\right)^{\frac{1}{2}} \leq & \frac{1}{\mu N_t}\sum_{i=1}^{t}\frac{n_i}{\gamma_i}\sqrt{\mathbb{E}\left[\|\Omega_i\|^2\right]} \\ = & \frac{1}{\mu N_t}\sum_{i=1}^{t}\frac{n_i}{\gamma_i}\sqrt{\mathbb{E}\left[\|\Omega_i\|^2\mathbb{1}_{\{\theta_{i-1}-\gamma_i\nabla_\theta l_i(\theta_{i-1})\notin\Theta\}}\right]}, \qquad (2.7.41)\end{aligned}$$

using Godichon-Baggioni [55, Lemma 4.3]. Next, we note that $\mathbb{E}[\|\Omega_t\|^2\mathbb{1}_{\{\theta_{t-1}-\gamma_t\nabla_\theta l_t(\theta_{t-1})\notin\Theta\}}] = 4\gamma_t^2 G_\Theta^2\mathbb{P}[\theta_{t-1} - \gamma_t\nabla_\theta l_t(\theta_{t-1}) \notin \Theta]$, since

$$\begin{aligned}\|\Omega_t\|^2 = & \left\|\mathcal{P}_\Theta\left(\theta_{t-1} - \gamma_t\nabla_\theta l_t\left(\theta_{t-1}\right)\right) - \theta_{t-1} + \gamma_t\nabla_\theta l_t\left(\theta_{t-1}\right)\right\|^2 \\ \leq & 2\left\|\mathcal{P}_\Theta\left(\theta_{t-1} - \gamma_t\nabla_\theta l_t\left(\theta_{t-1}\right)\right) - \theta_{t-1}\right\|^2 + 2\gamma_t^2\left\|\nabla_\theta l_t\left(\theta_{t-1}\right)\right\|^2 \\ = & 2\left\|\mathcal{P}_\Theta\left(\theta_{t-1} - \gamma_t\nabla_\theta l_t\left(\theta_{t-1}\right)\right) - \mathcal{P}_\Theta\left(\theta_{t-1}\right)\right\|^2 + 2\gamma_t^2\left\|\nabla_\theta l_t\left(\theta_{t-1}\right)\right\|^2 \\ \leq & 2\left\|\theta_{t-1} - \gamma_t\nabla_\theta l_t\left(\theta_{t-1}\right) - \theta_{t-1}\right\|^2 + 2\gamma_t^2\left\|\nabla_\theta l_t\left(\theta_{t-1}\right)\right\|^2 \\ = & 4\gamma_t^2\left\|\nabla_\theta l_t\left(\theta_{t-1}\right)\right\|^2 \leq 4\gamma_t^2 G_\Theta^2,\end{aligned}$$

as $\mathcal{P}_\Theta$ is Lipschitz and $\|\nabla_\theta l_{t,i}(\theta)\|^2 \leq G_\Theta^2$ for any $\theta \in \Theta$. Moreover, as in Godichon-Baggioni and Portier [56, Theorem 4.2], we know that $\mathbb{P}[\theta_{t-1} - \gamma_t \nabla_\theta l_t(\theta_{t-1}) \notin \Theta] \leq \Delta_t / d_{\min}^4$, where $d_{\min} = \inf_{\theta \in \partial\Theta} \|\theta - \theta^*\|$ with $\partial\Theta$ denoting the frontier of Θ. Thus, (2.7.41) can then be bounded by

$$\frac{1}{\mu N_t} \sum_{i=1}^{t} \frac{n_i}{\gamma_i} \sqrt{\mathbb{E}\left[\|\Omega_i\|^2 \mathbb{1}_{\{\theta_{i-1} - \gamma_i \nabla_\theta l_i(\theta_{i-1}) \notin \Theta\}}\right]} \leq \frac{2G_\Theta}{\mu d_{\min}^2 N_t} \sum_{i=1}^{t} n_i \Delta_i^{1/2} \leq \frac{2^2 G_\Theta}{\mu d_{\min}^2 N_t} \sum_{i=1}^{t} n_{i+1} \Delta_i^{1/2},$$

using that the sequence (n_t) is either constant or varying, meaning $n_{t+1}/n_t \leq 2$. □

Proof of Corollary 2.4.3. The proof follows directly from Corollary 2.4.1 but with use of Theorem 2.7.1. □

Proof of Corollary 2.4.4. The proof follows directly from Corollary 2.4.2 but with use of Theorem 2.7.1. □

Chapter 3: Learning from Time-dependent Streaming Data with Online Stochastic Algorithms

Abstract

We study stochastic algorithms in a streaming framework, trained on samples coming from a dependent data source. In this streaming framework, we analyze the convergence of Stochastic Gradient (SG) methods in a non-asymptotic manner; this includes various SG methods such as the well-known stochastic gradient descent (i.e., Robbins-Monro algorithm), mini-batch SG methods, together with their averaged estimates (i.e., Polyak-Ruppert averaged). Our results form a heuristic by linking the level of dependency and convexity to the rest of the model parameters. This heuristic provides new insights into choosing the optimal learning rate, which can help increase the stability of SG-based methods; these investigations suggest large streaming batches with slow decaying learning rates for highly dependent data sources.

keywords: *stochastic optimization, machine learning, stochastic algorithms, online learning, streaming, time-dependent data*

Contents

3.1 Introduction

Over the past decade, machine learning and artificial intelligence have become mainstream in many parts of society; substantial improvements in the performance and cost of mass storage devices and network systems have contributed to this. Traditional machine learning methods often work in a batch or offline learning setting, where the model is re-trained from scratch when new data arrive. Such learning methods suffer some critical drawbacks, such as expensive re-training costs when dealing with new data and thus poor scalability for large-scale and real-world applications. At the same time, these intelligent systems generate a practically infinite amount of large-scale data sets, many of which come as a continuous data stream, so-called streaming data.

Streaming data arrives as an endless sequence of samples (data points), which means that at any given time, the model must be able to adapt to the samples observed (so far) to predict/label new samples accurately. Such (streaming) models can never be seen as complete but must be updated continuously as newer samples arrive. Methods that recalculate the model from scratch on the arrival of new samples are impractical due to their high computational cost. Therefore we need procedures that effectively update the model as more samples arrive. This computational efficiency should not be at the expense of accuracy; the model's accuracy should be close to that achieved if we built a model from scratch using all the samples [20].

Stochastic algorithms have proven effective in overcoming the drawbacks of traditional (batch/offline) machine learning methods as they only use samples one by one without knowing their number in advance, especially the Stochastic Gradient (SG) method [124]. These SG methods have proven scalable and robust in many areas ranging from smooth and strongly convex problems to complex

non-convex ones, which makes them applicable in many large-scale machine learning tasks for real-world applications where data are large in size (and dimension) and arrive at a high velocity. Such first-order methods have been intensively studied in theory and practice in recent years [21].

The classical analyses for SG methods typically require unbiased gradients drawn independently and identically distributed (i.i.d.) from some underlying (and unknown) data generation process [34]. However, in practice, learning often happens with non-i.i.d. (and biased) data, e.g., network traffic, meteorological, financial time series, or other sensor data. We go beyond these standard assumptions by allowing dependent and biased gradients. SG methods can converge even when they only have access to biased gradients, but most analysis has been developed with specific applications in mind [4, 15, 37, 40, 130]. Stochastic learning algorithms for non-i.i.d. data are not as well understood as for i.i.d. data; however, some researchers have examined the convergence of statistical learning algorithm in non-i.i.d. settings [3, 94, 154].

Solving the problem of stochastic approximations using streaming SGs methods means we must approach the objective using the gradually arriving samples drawn according to some unknown dependent process. This leads to some new challenges, e.g., this endless stream of samples (may) changes at each step (and arrives sequentially), meaning that streaming SGs must be able to adapt to varying arrival speeds without compromising accuracy. We present and analyze streaming SGs that overcome these challenges and achieve convergence in various settings with long- and short-range dependence, biased gradient estimates, and changing data streams.

Contributions. In this paper, we investigate SG methods in a streaming framework [57], where the data comes from a dependent stochastic process. We provide non-asymptotic analysis and quantify the magnitude of achievable convergence rates under various dependency structures (sometimes leading to divergence). Our framework covers many applications with dependence and biased gradients under weak gradient assumptions. Our results builds a connection between dependency, the level of convexity, and the achievable learning rate to obtain optimal convergence. Roughly speaking, SG methods can achieve convergence using increasing batch sizes, which counteract the long-range (and short-range) dependence. We show that biased SG methods converge, and that they can converge with the same accuracy as unbiased SG methods if the bias is not too large. More surprisingly, our results show a precise heuristic that can be used in practice to help increase the stability of SG methods.

Organization. Section 3.2 presents the streaming framework on which the non-asymptotic analysis relies; we introduce some key concepts, definitions, and assumptions. In particular, Section 3.2.2 contains the assumptions about dependency structures and gradients, with some examples of how these could be verified using mixing conditions. Our convergence results are presented in Section 3.3, with and without averaging (Sections 3.3.1 and 3.3.2). Each result is followed by a thorough discussion that relates to other work. All our convergence analysis depends on the assumptions in Section 3.2, and some additional conditions for the averaged case (Section 3.3.2). At last, experimentations of our findings are illustrated in Section 3.4, with some final remarks in Section 3.5.

3.2 Problem Formulation

We consider the Stochastic Optimization (SO) problem $\min_{\theta\in\Theta} L(\theta) = \mathbb{E}_t[l_t(\theta)]$, where Θ is a closed convex set in $\mathbb{R}^d$ and $l_t : \Theta \to \mathbb{R}$ is some differentiable random functions (possibly non-convex), e.g, see Nesterov et al. [104]. We solve the SO problem in a streaming framework, where a *block* $l_t = (l_{t,1}, \ldots, l_{t,n_t})$ of $n_t \in \mathbb{N}$ random functions arrives at any given time $t \in \mathbb{N}$. In solving the SO problem, we use the Stochastic Streaming Gradient (SSG) estimate proposed by Godichon-Baggioni et al. [57], given as

$$\theta_t = \theta_{t-1} - \frac{\gamma_t}{n_t}\sum_{i=1}^{n_t} \nabla_\theta l_{t,i}\left(\theta_{t-1}\right), \quad \theta_0 \in \Theta, \tag{3.2.1}$$

where γ_t is the learning rate satisfying the conditions $\sum_{i=1}^{\infty}\gamma_i = \infty$ and $\sum_{i=1}^{\infty}\gamma_i^2 < \infty$ [124]. Note that if $\forall t$, $n_t = 1$, SSG becomes the well-known SG method, which has attracted a lot of attention [24, 68, 133, 153, 156]. Almost sure convergence of SO algorithms were shown in Pelletier [115]. In many models, there may be constraints on the parameter space, which would require a projection of the parameters; therefore, we also introduce the Projected Stochastic Streaming Gradient (PSSG) estimate, defined by

$$\theta_t = \mathcal{P}_\Theta\left(\theta_{t-1} - \frac{\gamma_t}{n_t}\sum_{i=1}^{n_t} \nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right), \quad \theta_0 \in \Theta, \tag{3.2.2}$$

where $\mathcal{P}_\Theta$ denotes the Euclidean projection onto Θ, i.e., $\mathcal{P}_\Theta(\theta) = \arg\min_{\theta'\in\Theta}\|\theta - \theta'\|_2$. To shorten notation, we let $\nabla_\theta l_t(\theta) = n_t^{-1}\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}(\theta)$. An essential extension is the Polyak-Ruppert averaging [118, 129], which guarantees optimal statistical efficiency without jeopardizing the computational cost; the Averaged Stochastic Streaming Gradient (ASSG) is given by

$$\bar{\theta}_t = \frac{1}{N_t}\sum_{i=0}^{t-1} n_{i+1}\theta_i, \quad \bar{\theta}_0 = 0, \tag{3.2.3}$$

where $N_t = \sum_{i=1}^{t} n_i$ is the accumulated sum of observations. Likewise, let PASSG denote the (Polyak-Ruppert) averaged estimate of PSSG (3.2.2).

3.2.1 Quasi-strong Convex Objectives

Following Gower et al. [60], Moulines and Bach [96], we assume that L has a unique global minimizer $\theta^* \in \Theta$ such that $\nabla_\theta L(\theta^*) = 0$, and it is μ-quasi-strongly convex [80, 99], i.e, there exists $\mu > 0$ such that $\forall\theta \in \Theta$,

$$L(\theta^*) \geq L(\theta) + \langle \nabla_\theta L(\theta), \theta^* - \theta\rangle + \frac{\mu}{2}\|\theta^* - \theta\|^2. \tag{3.2.4}$$

The μ-quasi-strongly convexity assumption is a non-strongly convex relaxation of the SO problem, which is more conservative than μ-strongly convexity. Relaxations of convexity is crucial in practice to ensure robustness and adaptiveness of the algorithms, e.g., for non-strongly convex SO, see Bach and Moulines [9], Necoara et al. [99], Nemirovski et al. [101].

3.2.2 Stochastic Streaming Gradient Assumptions: Dependence, Biased Gradients, Expected Smoothness, and Gradient Noise

We go beyond the classical assumptions that require unbiased (uniformly bounded) gradients by allowing the gradients to be dependent and biased estimates. Our aim is to non-asymptotically bound the SSG estimates (3.2.1) to (3.2.3) explicitly using the SO problem parameters. In order to do this, we let the natural filtration of the SO problem $\mathcal{F}_t = \sigma(l_i : i \leq t)$, and assume the following about the gradients $(\nabla_\theta l_t)$:

Assumption 3.2.1-p ($D_\nu\nu_t$-dependence and $B_\nu\nu_t$-bias)**.** *Let θ_0 be $\mathcal{F}_0$-measurable. For each $t \geq 1$, the random function $\nabla_\theta l_t(\theta)$ is square-integrable, $\mathcal{F}_t$-measurable, and there exists a positive integer p such that for all $\mathcal{F}_{t-1}$-measurable $\theta \in \Theta$,*

$$\mathbb{E}[\|\mathbb{E}[\nabla_\theta l_t(\theta)|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta)\|^p] \leq \nu_t^p (D_\nu^p \mathbb{E}[\|\theta - \theta^*\|^p] + B_\nu^p), \tag{3.2.5}$$

for some positive sequence $(\nu_t)_{t\geq 1}$ with $D_\nu, B_\nu \geq 0$.

In the classical convergence analysis of SG methods, one assumes that the SGs are uniformly bounded. However, this assumption is too restrictive as it only may hold for some losses [21, 107]. Instead, we follow the same ideas as in Gower et al. [60], Moulines and Bach [96], to make the following assumption about the expected smoothness of the stochastic gradients $(\nabla_\theta l_t)$.

Assumption 3.2.2-p (κ_t-expected smoothness)**.** *There exists a positive integer p such that $\forall \theta, \theta' \in \Theta$, $\mathbb{E}[\|\nabla_\theta l_t(\theta) - \nabla_\theta l_t(\theta')\|^p] \leq \kappa_t^p \mathbb{E}[\|\theta - \theta'\|^p]$ for some positive sequence $(\kappa_t)_{t\geq 1}$.*

Assumption 3.2.2-p can be seen as an assumption about the smoothness properties of (l_t). The last fundamental assumption (Assumption 3.2.3-p) is a very weak assumption, and should be seen as an assumption on Θ rather than on (l_t):

Assumption 3.2.3-p (σ_t-gradient noise)**.** *There exists a positive integer p such that $\mathbb{E}[\|\nabla_\theta l_t(\theta^*)\|^p] \leq \sigma_t^p$ for some positive sequence $(\sigma_t)_{t\geq 1}$.*

These assumptions (Assumptions 3.2.1-p to 3.2.3-p) are milder than the standard assumptions for stochastic approximations, e.g., see [13, 57, 87, 96]. They include classic examples such as stochastic approximation and learning from dependent data, which we will demonstrate later in Section 3.4. Assumption 3.2.1-p is on the form of mixing conditions for weakly dependence sequences, implying that dependence dilutes with the rate of ν_t. It is possible to verify Assumption 3.2.1-p by using moment inequalities for partial sums of strongly mixing sequences [123]; we will refer to this

as short-range dependence. Note that for any positive integer p, Assumption 3.2.1-p can be upper bounded by

$$\mathbb{E}[\|\mathbb{E}[\nabla_\theta l_t(\theta)|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta)\|^p] \leq \mathbb{E}[\|\nabla_\theta l_t(\theta) - \nabla_\theta L(\theta)\|^p] = n_t^{-p}\mathbb{E}[\|S_t\|^p], \tag{3.2.6}$$

using Jensen's inequality, where $S_t = \sum_{i=1}^{n_t}(\nabla_\theta l_{t,i}(\theta) - \nabla_\theta L(\theta))$ is a d-dimensional vector. Let $(\nabla_\theta l_{t,i})$ be a strictly stationary sequence and assume that there exists some $r > p$ such that $\sup_{x>0}(x^r Q(x))^{1/r} < \infty$, where $Q(x)$ denotes the quantile function of $\|\nabla_\theta l_{t,i}\|$. Suppose that $(\nabla_\theta l_{t,i})$ is strongly α-mixing in the sense of Rosenblatt [125], with strong mixing coefficients $(\alpha_t)_{t\geq 1}$ satisfying $\alpha_t = \mathcal{O}(t^{-pr/(2r-2p)})$. Then by Rio [123, Corollary 6.1], we have that $\mathbb{E}[\|S_t\|^p] = \mathcal{O}(n_t^{p/2})$, meaning, (3.2.6) is at most $\mathcal{O}(n_t^{-p/2})$; this includes several linear, non-linear, and Markovian time series, e.g., see Bradley [29], Doukhan [41] for more examples, other mixing coefficients of weak dependence and the relations between them. In relation to the form of Assumption 3.2.1-p, this means that $B_\nu \neq 0$ in this case. However, having $B_\nu = 0$ is possible in well-specified examples, which we will see later in Section 3.4. Note that Assumptions 3.2.2-p and 3.2.3-p can be verified using α-mixing conditions by analogues arguments as for Assumption 3.2.1-p such that κ_t^p and σ_t^p is $\mathcal{O}(n_t^{-p/2})$.

3.3 Convergence Analysis

In this section, we consider the stochastic streaming estimates in (3.2.1) to (3.2.3) with streaming-batches (n_t) arriving in non-decreasing streams. We aim to non-asymptotically bound $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ and $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$, such that they only depend on the parameters of the problem.

Learning rate and function forms. Throughout this paper, we consider learning rates on the form $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ with $C_\gamma > 0$, $\beta \in [0, 1]$, and α chosen accordingly to the expected streaming-batches n_t. Obviously, (ν_t), (κ_t), and (σ_t) may be considered as uncertain terms depending on the streaming-batch n_t. Thus, let $\nu_t = n_t^{-\nu}$, $\kappa_t = C_\kappa n_t^{-\kappa}$, and $\sigma_t = C_\sigma n_t^{-\sigma}$ with $\nu \in (0, \infty)$, $\kappa, \sigma \in [0, 1/2]$, and $C_\kappa, C_\sigma > 0$. Having, $\sigma, \kappa \in [0, 1/2]$ follows directly from Godichon-Baggioni et al. [57], since $\sigma = \kappa = 1/2$ corresponds to the i.i.d. case[1], whereas $\sigma, \kappa < 1/2$ allows noisier outputs. Similarly, $v_t = 0$ corresponds to the classical i.i.d. setting. Having $\nu_t = n_t^{-\nu}$ means Assumption 3.2.1-p, allow so-called long-range dependence (also known as long memory or long-range persistence) when $\nu \in (0, 1/2)$ and short-range dependence when $\nu \in [1/2, \infty)$. Thus, the i.i.d. case is when $\nu \to \infty$.

For the sake of simplicity, we consider streaming-batches (n_t) on the form $C_\rho t^\rho$ with $C_\rho \in \mathbb{N}$ and $\rho \in [0, 1)$ such that $n_t \in \mathbb{N}$. This form of streaming-batches means that we are considering everything from vanilla SG and mini-batch SG methods, to more exotic learning designs, e.g., $C_\rho > 1$ and $\rho = 0$ correspond to mini-batch SG of size C_ρ. We will refer to C_ρ as the *streaming constant* size and ρ as the *streaming rate*.

[1] You can't beat the system.

3.3.1 Stochastic Streaming Gradients

Theorem 3.3.1. *Denote* $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ *for some* $\delta_0 \geq 0$, *where* (θ_t) *follows the recursion in* (3.2.1) *or* (3.2.2). *Assume that Assumptions* 3.2.1-p *to* 3.2.3-p *hold true for* $p = 2$. *Suppose* $n_t = C_\rho t^\rho$ *with* $\rho \in [0, 1)$ *and* $C_\rho \in \mathbb{N}$, *such that* $\mu_\nu = \mu - \mathbb{1}_{\{\rho=0\}} 2D_\nu C_\rho^{-\nu} > 0$. *For* $\alpha - \rho\beta \in (1/2, 1)$, *we have*

$$\delta_t \leq \pi_t + \frac{2^{\frac{2+6\rho\nu}{1+\rho}} B_\nu^2}{\mu\mu_\nu C_\rho^{\frac{2\nu}{1+\rho}} N_t^{\frac{2\rho\nu}{1+\rho}}} + \frac{2^{\frac{7+6\rho\sigma}{1+\rho}} C_\sigma^2 C_\gamma}{\mu_\nu C_\rho^{\frac{2\sigma-\beta-\alpha}{1+\rho}} N_t^{\frac{\rho(2\sigma-\beta)+\alpha}{1+\rho}}}, \tag{3.3.7}$$

with π_t *given in* (3.6.22) *such that* $\pi_t = \mathcal{O}(\exp(-N_t^{(1+\rho\beta-\alpha)/(1+\rho)}))$.

Sketch of proof. Under Assumptions 3.2.1-p to 3.2.3-p with $p = 2$, it can be shown that (δ_t) satisfies the recursive relation (3.6.20),

$$\delta_t \leq [1 - (\mu - 2D_\nu\nu_t)\gamma_t + 2\kappa_t^2\gamma_t^2]\delta_{t-1} + \frac{B_\nu^2}{\mu}\nu_t^2\gamma_t + 2\sigma_t^2\gamma_t^2,$$

for any γ_t, ν_t, κ_t, σ_t, and n_t. This recursive relation can be explicitly upper bounded in a non-asymptotic way (by Proposition 3.6.1) using classical techniques from stochastic approximations [13, 87]. As mentioned in Zinkevich [157], bounding the projected estimate in (3.2.2) follows directly from that $\mathbb{E}[\|\mathcal{P}_\Theta(\theta) - \theta^*\|^2] \leq \mathbb{E}[\|\theta - \theta^*\|^2]$, $\forall\theta \in \mathbb{R}^d$, $\forall\theta^* \in \Theta$, as Θ is a closed convex set.

Related work. Theorem 3.3.1 replicate the results of the unbiased i.i.d. case (with $B_\nu = 0$ and $\kappa = \sigma = 1/2$) considered in Godichon-Baggioni et al. [57]. Our findings also reproduce the results of Moulines and Bach [96], where they considered the unbiased i.i.d. case (under slightly different assumptions) using the vanilla SG method, namely, when $C_\rho = 1$ and $\rho = 0$. Moreover, if the function L has C_∇-Lipschitz continuous gradients[2], then (3.3.7) implies the bound on the objective function values of L, $\mathbb{E}[L(\theta_t) - L(\theta^*)] \leq C_\nabla\delta_t/2$ by Cauchy–Schwarz's inequality.

Decay of the initial conditions. The initial conditions that π_t contains will be forgotten sub-exponentially fast, since $\pi_t = \mathcal{O}(\exp(-N_t^{(1+\rho\beta-\alpha)/(1+\rho)}))$ as long as $\mu_\nu = \mu - \mathbb{1}_{\{\rho=0\}} 2D_\nu C_\rho^{-\nu} > 0$. Note that the positivity of the dependence penalised convexity constant μ_ν is essential in all terms of (3.3.7). Having $\mu_\nu > 0$ depends solely on the level of dependence D_ν but it is scaled by $C_\rho^{-\nu}$, meaning if D_ν is so large that μ_ν is no longer positive, then we should take C_ρ large enough such that μ_ν becomes positive again; this is illustrated in Sections 3.4.2 and 3.4.3. The streaming constant C_ρ contributes positively to all terms in (3.3.7), either directly or though μ_ν.

The last term of (3.3.7) can be seen as the noise term decaying with $\mathcal{O}(N_t^{-(\rho(2\sigma-\beta)+\alpha)/(1+\rho)})$ for $\alpha - \rho\beta \in (1/2, 1)$, e.g., for any $\rho \in [0, 1)$, $\delta_t = \mathcal{O}(N_t^{-2/3})$ when $\alpha = 2/3$, $\beta = 1/3$, and $\sigma = 1/2$. In addition, the noise term is positively affected by large streaming constants C_ρ when $\alpha + \beta < 2\sigma$, which will be expressed as a variance reduction, e.g., see Section 3.4. In unbiased cases ($B_\nu = 0$) the noise term would also be the asymptotic term.

[2]Later, in Section 3.3.2 for the averaged estimate (3.2.3), we assume in (3.3.8) that the function L has C_∇-Lipschitz continuous gradients.

Behavior for B_ν. The second term of (3.3.7) can be seen as an dependency term as it is determined solely by the level of dependence ν, the bias error B_ν, and the convexity constant μ_ν; It is remarkable that the dependence term is unconnected from the choice of the learning rate (γ_t) but instead by the streaming rate through C_ρ and ρ. The dependence term decay with $\mathcal{O}(N_t^{-2\rho\nu/(1+\rho)})$ which requires ρ positive to decay since $\nu \in (0, \infty)$, e.g., to obtain $\mathcal{O}(N_t^{-1/2})$ we would need $\rho = 1$ and $\nu = 1/2$. It is surprising that Theorem 3.3.1 allows both long-range and short-range dependence. Indeed, long-range dependence leads to slow convergence (slower than $\mathcal{O}(N_t^{-1/2})$) but it will still converge. Obviously, this only matters if $B_\nu \neq 0$. Overall, $\delta_t = \mathcal{O}(\max\{\mathbb{1}_{\{B_\nu \neq 0\}} N_t^{-2\rho\nu/(1+\rho)}, N_t^{-(\rho(2\sigma-\beta)+\alpha)/(1+\rho)}\})$.

3.3.2 Averaged Stochastic Streaming Gradients

In what follows, we consider the averaging estimate $\bar{\theta}_n$ given in (3.2.3) with (θ_t) following the SSG estimate in (3.2.1) or the PSSG estimate in (3.2.2). Some additional assumptions is needed for bounding the *rest* terms of the averaging estimate: let the function L have C_∇-Lipschitz continuous gradients, i.e., there exists a constant $C_\nabla > 0$, $\forall \theta, \theta' \in \Theta \subseteq \mathbb{R}^d$,

$$\|\nabla_\theta L(\theta) - \nabla_\theta L(\theta')\| \leq C_\nabla \|\theta - \theta'\|. \tag{3.3.8}$$

As discussed in Bottou et al. [21], this assumption ensures that $\nabla_\theta L$ does not vary arbitrarily, making the gradient $\nabla_\theta L$ a useful indicator on how to decrease L. Next, assume that the Hessian of L is C'_∇-Lipschitz-continuous, that is, there exists $C'_\nabla > 0$ such that $\forall \theta, \theta' \in \Theta \subseteq \mathbb{R}^d$,

$$\|\nabla^2_\theta L(\theta) - \nabla^2_\theta L(\theta')\| \leq C'_\nabla \|\theta - \theta'\|. \tag{3.3.9}$$

Note that (3.3.8) and (3.3.9) only needs to hold true for $\theta' = \theta^*$. Moreover, in continuation of Assumption 3.2.3-p with $\sigma_t = C_\sigma n_t^{-\sigma}$ for $\sigma \in [0, 1/2]$, we make the following assumption:

Assumption 3.3.1. *There exists a non-negative self-adjoint operator Σ such that $\forall t \geq 1$, we have $n_t^{2\sigma} \mathbb{E}[\nabla_\theta l_t(\theta^*) \nabla_\theta l_t(\theta^*)^\top] \preceq \Sigma + \Sigma_t$, where Σ_t is a positive symmetric matrix with $Tr(\Sigma_t) = C'_\sigma n_t^{-2\sigma'}$, $C'_\sigma \geq 0$, and $\sigma' \in (0, 1/2]$.*

Remark that in the unbiased case, such as in Section 3.4.1, Assumption 3.3.1 is verified with $\sigma = 1/2$ and $C'_\sigma = 0$ [57]. The short-range dependence cases is when $\sigma = 1/2$, as in Section 3.4.1, whereas, the long-range dependence case is for $\sigma < 1/2$. Moreover, Assumption 3.3.1 allows us to obtain leading term Λ/N_t with $\Lambda = \mathrm{Tr}(\nabla^2_\theta L(\theta^*)^{-1} \Sigma \nabla^2_\theta L(\theta^*)^{-1})$, which attains the Cramer-Rao bound; we will see this in Theorem 3.3.2.

To consider the averaging estimate $\bar{\theta}_n$ given in (3.2.3), an additional assumption is needed in order to avoid calculating the six-order moment: we make the unnecessary assumption that $(\nabla_\theta l_t)$ is uniformly bounded; the derivation of the six-order moment can be found in Godichon-Baggioni [55].

Assumption 3.3.2. *Let $D_\Theta = \inf_{\theta\in\partial\Theta}\|\theta-\theta^*\| > 0$ with $\partial\Theta$ denoting the frontier of Θ. Moreover, there exists $G_\Theta > 0$ such that $\forall t \geq 1$, $\sup_{\theta\in\Theta}\|\nabla_\theta l_t(\theta)\|^2 \leq G_\Theta^2$ a.s.*

Theorem 3.3.2. *Denote $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t-\theta^*\|^2]$ with $\bar{\theta}_n$ given by (3.2.3), where (θ_t) follows the recursion in (3.2.1) or (3.2.2). Assume that Assumptions 3.2.1-p to 3.2.3-p for $p=4$ and Assumption 3.3.1 hold true. In addition, Assumption 3.3.2 must hold true if (θ_t) follows the recursion in (3.2.2). Suppose $n_t = C_\rho t^\rho$ with $\rho \in [0,1)$ and $C_\rho \in \mathbb{N}$, such that $\mu_\nu = \mu - \mathbb{1}_{\{\rho=0\}}2D_\nu C_\rho^{-\nu} > 0$. For $\alpha-\rho\beta \in (1/2,1)$, we have*

$$\bar{\delta}_t^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}}\mathbb{1}_{\{\sigma=1/2\}} + \frac{2^{1/2}\Lambda^{1/2}C_\rho^{\frac{1-2\sigma}{2(1+\rho)}}}{N_t^{\frac{1+2\rho\sigma}{2(1+\rho)}}}\mathbb{1}_{\{\sigma<1/2\}} + \frac{2^{1/2}C_\sigma'^{1/2}C_\rho^{\frac{1-2(\sigma+\sigma')}{2(1+\rho)}}}{\mu N_t^{\frac{1+2\rho(\sigma+\sigma')}{2(1+\rho)}}} \tag{3.3.10}$$

$$+\mathcal{O}\left(\max\left\{N_t^{-\frac{2+\rho(2\sigma+\beta)-\alpha}{2(1+\rho)}}, N_t^{-\frac{\rho(2\sigma-\beta)+\alpha}{1+\rho}}\right\}\right) + \tilde{\mathcal{O}}\left(N_t^{-\frac{\delta+\rho\nu}{2(1+\rho)}}\right) + \mathbb{1}_{\{B_\nu\neq 0\}}\Psi_t, \tag{3.3.11}$$

with $\delta = \mathbb{1}_{\{B_\nu=0\}}(\rho(2\sigma-\beta)+\alpha) + \mathbb{1}_{\{B_\nu\neq 0\}}\min\{\rho(2\sigma-\beta)+\alpha, 2\rho\nu\}$ and Ψ_t given in (3.6.36), such that

$$\Psi_t = \tilde{\mathcal{O}}\left(\max\left\{N_t^{-\frac{\rho(\sigma+\nu)}{2(1+\rho)}}, N_t^{-\frac{1+\rho(\beta+\nu)-\alpha}{1+\rho}}, N_t^{-\frac{1+2\rho\nu}{2(1+\rho)}}, N_t^{-\frac{\delta/2+\rho\nu}{2(1+\rho)}}, N_t^{-\frac{2\rho\nu}{1+\rho}}\right\}\right).$$

An explicit version of the bound is given in (3.6.37).

Sketch of proof. In Lemma 3.6.3, we conduct a general study of the Polyak-Ruppert averaging estimate $(\bar{\theta}_t)$ defined in (3.2.3) for (γ_t), (ν_t), (κ_t), (σ_t) and (n_t) on any form. Thus, Theorem 3.3.2 follows by Lemma 3.6.3 using the (specific) bounds of $\delta_t = \mathbb{E}[\|\theta_t-\theta^*\|^2]$ and $\Delta_t = \mathbb{E}[\|\theta_t-\theta^*\|^4]$ in Theorem 3.3.1 (eq. (3.6.21)) and Lemma 3.6.2.

Related work. Theorem 3.3.2 replicate the results of Godichon-Baggioni et al. [57] with Λ/N_t as leading term in the unbiased i.i.d. case. Thus, by averaging it is possible to achieve the incorrigible rate of $\mathcal{O}(N_t^{-1})$, e.g., this is always achieved in the unbiased case with $\sigma = 1/2$, even under short-range dependence (i.e., when $\nu \geq 1/2$).

Accelerated decay. Remark that each term in (3.3.10) is a direct consequence of Assumption 3.3.1. Furthermore, all terms of (3.3.10) are independent of the learning rate (γ_t) but the two last terms are dependent on streaming batches through C_ρ and ρ. As in Theorem 3.3.1, the positivity of μ_ν is essential for all terms in (3.3.11) even if it does not appear directly. For objectives that lack convexity μ or have high levels of dependence D_ν, we can only ensure convergence by increasing C_ρ, i.e., ensuring positivity of μ_ν; this is illustrated in Sections 3.4.2 and 3.4.3 for ARCH models.

The first term of (3.3.11) decay at the rate $\mathcal{O}(\max\{N_t^{-(2+\rho(\beta+2\sigma)-\alpha)/(1+\rho)}, N_t^{-2(\rho(2\sigma-\beta)+\alpha)/(1+\rho)}\})$, which suggests choosing α, β such that $\alpha+\rho(2\sigma/3-\beta) = 2/3$, e.g., $\alpha = 2/3$, $\beta = 1/3$ and $\sigma = 1/2$ yields a decay of $\mathcal{O}(N_t^{-4/3})$ for any ρ. Thus, we can robustly achieve $\mathcal{O}(N_t^{-4/3})$ for any streaming rate ρ by setting $\alpha = 2/3$ and $\beta = 1/3$ if $\sigma = 1/2$. In general, the convergence is resilient to any streaming rate ρ by having $\alpha = 2/3$ and $\beta = 2\sigma/3$. But taking $\beta > 0$ would damage the variance

reduction effect from having C_ρ large (e.g., see discussion after Theorem 3.3.1). Thus, there is a trade-off between accelerating the convergence by taking $\beta = 2\sigma/3 > 0$ or taking $\beta = 0$ to favor from variance reduction. In practice, an immediate choice would be to take $\beta = 0$, but if the data or model contains a low amount of noise, it can be advantageous to raise β to improve convergence [57].

Next, the decay of the second term in (3.3.11) is tricky to interpret in a simple manner as it is a mixture of the learning rate, streaming rate, dependence, and bias. Nevertheless, some observations can be made: first, having $\beta = 0$ is beneficial for the decay rate δ in all cases. Second, increasing streaming rate ρ would also increase the decay.

Behavior for B_ν. The influence of B_ν is exclusively contained in Ψ_t, with the exception of the second term of (3.3.11). Also, increasing ρ will always diminish the bad influence of this bias term. Surprisingly, $\Psi_t \to 0$ as $t \to \infty$ for any ν, but long-range dependence is excluded if we wish to obtain the desired rate of $\bar{\delta}_t = \mathcal{O}(N^{-1})$. However, it does not seem to have any major influence in our experiments, e.g., see Section 3.4. To conclude, by taking ρ positive and C_ρ large enough to ensure that μ_ν stays positive, then we will converge under long- or short-range dependence with biased gradient estimates.

3.4 Experiments

A way to illustrate our findings is by use of classical time-series methods that aims to construct models for time-series analysis, modeling, and prediction of the underlying sequences of real-valued signals (X_t). These methods have been successfully used in a wide range of applications such as statistics, econometrics, and signal processing because of their ability to describe or predict time-varying (dependent) processes, e.g., the AutoRegressive (AR), Moving-Average (MA), and AutoRegressive Moving-Average (ARMA) models are the most well-known models for time-series [25, 30, 66]. The standard time-series analysis often relies on independence and constant noise, but it can be relaxed by, e.g., the AutoRegressive Conditional Heteroskedasticity (ARCH) model [45]. Online learning algorithms of (both stationary and non-stationary) dependent time-series have been studied in Agarwal and Duchi [3], Anava et al. [7], Wintenberger [152].

Our experiment measures the performance by the quadratic mean error $\mathbb{E}[\|\theta_{N_t} - \theta^*\|^2]$ over one thousand replications with θ_0 and θ^* drawn randomly according to the models' specifications. It should be noted that averaging over several replications gives a reduction in variability, that mainly benefits the SSG. The experiments will demonstrate how the choice of C_ρ and ρ affects the dependence D_ν, bias B_ν, and the (dependence) penalised convexity constant μ_ν. To compare different data streams $n_t = C_\rho t^\rho$ through the selection of C_ρ and ρ, we fix the following parameters: $C_\gamma = 1$, $\alpha = 2/3$, and $\beta = 0$.

3.4.1 AutoRegressive (AR) Model

A process (X_t) is called a (zero-mean) AR(1) process, if there exists real-valued parameter θ such that $X_t = \theta X_{t-1} + \epsilon_t$, where (ϵ_t) is some noise process with zero mean and noise σ_ϵ. To illustrate the versatility of our results, we construct some noisy (heavy-tailed) data with long-range dependence: the noisiness is integrated using a Student's t-distribution with degrees of freedom above four, denoted by (z_t). The long-range dependence is incorporated by multiplying (z_t) with the fractional Gaussian noise $G_t(H) = B_{t+1}(H) - B_t(H)$, where $(B_t(H))$ is a fractional Brownian motion with Hurst index $H \in (0, 1)$. $(B_t(H))$ can also be seen as a (zero-mean) Gaussian process with stationary and self-similar increments. Thus, let the AR(1) process X_t be constructed using the noise process $\epsilon_t = \sqrt{G_t(3/4)}z_t$, where a Hurst index of $H = 3/4$ corresponds to $\nu_t^2, \kappa_t^2, \sigma_t^2$ is $\mathcal{O}(n_t^{-1/2})$ and $\nu_t^4, \kappa_t^4, \sigma_t^4$ is $\mathcal{O}(n_t^{-3/4})$ in Assumptions 3.2.1-p to 3.2.3-p and 3.3.1.

AutoRegressive (AR) Model

Consider the example, in which, we estimate an AR(1) model $X_t = \theta X_{t-1} + \epsilon_t$ from the underlying stationary AR(1) process $X_t = \theta^* X_{t-1} + \epsilon_t$ with $|\theta^*| < 1$. We omit to project our estimates as this will hide the dependence coming from D_ν, which is what we wish to explore. For constant streaming-batch sizes of one, the squared loss is $l_t(\theta) = (X_t - \theta X_{t-1})^2$ with $\nabla_\theta l_t(\theta) = -2X_{t-1}(X_t - \theta X_{t-1})$. This gives a mean squared loss

$$L(\theta) = \mathbb{E}_t[l_t(\theta)] = \mathbb{E}[(X_t - \theta X_{t-1})^2] = \mathbb{E}[(\theta^* X_{t-1} + \epsilon_t - \theta X_{t-1})^2] = (\theta^* - \theta)^2 \mathbb{E}[X_{t-1}^2] + \sigma_\epsilon^2,$$

with $\nabla_\theta L(\theta) = -2(\theta^* - \theta)\mathbb{E}[X_{t-1}^2]$. Thus, Assumption 3.2.1-p (for $p = 2$ with $\sigma(X_{t-1}) \subseteq \mathcal{F}_{t-1}$) yields

$$\begin{aligned}\mathbb{E}[\|\mathbb{E}[\nabla_\theta l_t(\theta)|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta)\|^2] =& \mathbb{E}[(\mathbb{E}[2X_{t-1}(\theta X_{t-1} - X_t)|\mathcal{F}_{t-1}] - 2(\theta - \theta^*)\mathbb{E}[X_{t-1}^2])^2] \\ =& 4(\theta - \theta^*)^2 \mathbb{E}[(X_{t-1}^2 - \mathbb{E}[X_{t-1}^2])^2],\end{aligned}$$

meaning that Assumption 3.2.1-p is fulfilled if X_t has bounded moments of order p. Moreover, from this we can directly deduce that $B_\nu = 0$. Likewise, the remaining assumptions can be verified, in particular Assumption 3.3.1 is satisfied with $\Sigma_t = 0$.

AutoRegressive (AR) Model: misspecified case

Next, assume that the underlying data generating process follows the MA(1)-process, $X_t = \phi\epsilon_{t-1} + \epsilon_t$, with $\phi \in \mathbb{R}$. The misspecification error of fitting an AR(1) model to a MA(1) process

can be found by minimizing $L(\theta)$,

$$\begin{aligned}\theta^* &= \arg\min_\theta \mathbb{E}[(X_t - \theta X_{t-1})^2] = \arg\min_\theta \mathbb{E}[(\epsilon_t + \phi\epsilon_{t-1} - \theta(\epsilon_{t-1} + \phi\epsilon_{t-2}))^2] \\ &= \arg\min_\theta \mathbb{E}[(\epsilon_t + (\phi-\theta)\epsilon_{t-1} - \theta\phi\epsilon_{t-2})^2] = \arg\min_\theta \sigma_\epsilon^2 + (\phi-\theta)^2\sigma_\epsilon^2 + \theta^2\phi^2\sigma_\epsilon^2 \\ &\equiv \arg\min_\theta (\phi-\theta)^2 + \theta^2\phi^2 = \arg\min_\theta L(\theta),\end{aligned}$$

where $L(\theta) = (\phi-\theta)^2+\theta^2\phi^2$ is a strictly convex function in θ. Thus, $\nabla_\theta L(\theta) = 0 \Leftrightarrow 2(\theta-\phi)+2\theta\phi^2 = 0 \Leftrightarrow 2\theta(1+\phi^2) = 2\phi \Leftrightarrow \theta = \phi/(1+\phi^2)$, meaning for $\phi \in \mathbb{R}$ we have $\theta \in (-1/2, 1/2)$. With this in mind, we can conduct our study of fitting an AR(1) model to the MA(1) process with ϕ drawn randomly from $\mathbb{R}$.

3.4.2 AutoRegressive Conditional Heteroskedasticity (ARCH) Model

A key element of time series analysis is modeling heteroscedasticity of the conditional variance, e.g., volatility clustering in financial time series. AutoRegressive Conditional Heteroscedasticity (ARCH) models are some of the most well-known models that incorporate this feature. A process (ϵ_t) is called an ARCH(1) process with parameters α_0 and α_1 if it satisfies

$$\begin{cases}\epsilon_t = \sigma_t z_t, \\ \sigma_t^2 = \alpha_0 + \alpha_1\epsilon_{t-1}^2,\end{cases} \tag{3.4.12}$$

where $\alpha_0 > 0$ and $\alpha_1 \geq 0$ ensures the non-negativity of the conditional variance process (σ_t^2), and the innovations (z_t) is white noise. The ARCH process parameters are known to be challenging to estimate in empirical applications as the optimization algorithms can quickly fail or converge to irregular solutions. Therefore, projecting the estimates is vital for the optimization procedure. A well-discussed problem for the ARCH models is that small values of α_0 are tricky to estimate. Stabilizing the estimation of α_0 would not only improve the α_0 estimate but also have a positive impact on the other model parameters. One way to deal with small values of α_0 is by the using the models homogeneity, i.e., scaling an ARCH process (X_t) with parameters (α_0, α_1) gives us an ARCH process $(\sqrt{\lambda}X_t)$ with parameters $(\lambda\alpha_0, \alpha_1)$ with same innovations. To simplify our analysis we consider a stationary ARCH(1) model, where we fix α_0 at 1 and initialize it at $1/2$. We employ the quasi-maximum likelihood procedure for the statistical inference as outlined in Werge and Wintenberger [150]; the quasi likelihood losses is given by $l_t(\theta) = 2^{-1}(X_t^2/\sigma_t^2(\theta) + \log(\sigma_t^2(\theta))$ with first-order derivative

$$\nabla_\theta l_t(\theta) = \nabla_\theta\sigma_t^2(\theta)\left(\frac{\sigma_t^2(\theta) - X_t^2}{2\sigma_t^4(\theta)}\right)$$

where $\nabla_\theta\sigma_t^2(\theta) = (1, X_{t-1}^2)^T$. Observe that the loss function (l_t) itself is not strongly convex but only the objective function L may be strongly convex; convexity conditions of ARCH was investigated

in Wintenberger [152]. There are different ways to overcome lack of convexity: first, projecting the estimates such that the (conditional) variance process (σ_t^2) stays away from zero (and close to the unconditional variance). Second, in the specific example of ARCH model, one could also recover convexity by implementing variance targeting techniques; an example using Generalized ARCH (GARCH) models can be found in Werge and Wintenberger [150].

3.4.3 AutoRegressive (AR)-AutoRegressive Conditional Heteroskedasticity (ARCH) Model

We complete our experiments by considering an AR models with ARCH noise: the process (X_t) is called an AR(1)-ARCH(1) process with parameters θ, α_0 and α_1 if it satisfies

$$\begin{cases} X_t = \theta X_{t-1} + \epsilon_t, \\ \epsilon_t = \sigma_t z_t, \\ \sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2. \end{cases} \tag{3.4.13}$$

where the innovations (z_t) is white noise. The statistical inference of this model is done using the squared loss for the AR-part and the QMLE for the ARCH part, e.g., see Sections 3.4.1 and 3.4.2.

3.4.4 Discussion of Experiments

The experiments described earlier in Sections 3.4.1 to 3.4.3 can be found in Figure 3.1; here $\{C_\rho = 1, \rho = 0\}$ corresponds to the classical SG method and its (Polyak-Ruppert) average estimate, $\{C_\rho = 64, \rho = 0\}$ is a mini-batch SSG/ASSG, and $\{C_\rho = 64, \rho = 1/2\}$ is an increasing SSG/ASSG with initial batch size of $C_\rho = 64$.

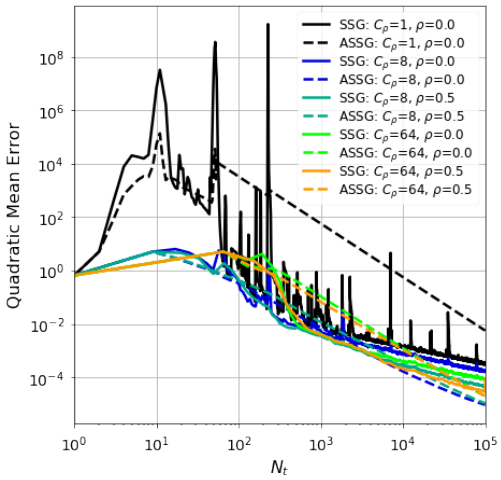
First consider the AR illustration in Figures 3.1a and 3.1b: each pair of data streams converges, but it is clear that the traditional SG method experiences a large amount of noise initially, particularly affecting the average estimate period but not its decay rate.[3] Both methods show a noticeable reduction in variance when C_ρ increases, which is particularly beneficial in the beginning. Nevertheless, too large streaming batch sizes C_ρ may hinder the convergence as this leads to too few iterations. Moreover, Figures 3.1a and 3.1b indicates improving decay for SSG when increasing the streaming rate ρ. Conversely, ASSG does not see improvements in the same way, as we do not exploit the potential of using multiple observations through the β parameter, which could accelerate convergence, e.g., see Godichon-Baggioni et al. [57] for a discussion in the (unbiased) i.i.d. case. It is surprising that we do not see any effect from Σ_t in Assumption 3.3.1, but this seems to be an artifact effect in the proof as we need fourth-order moments.

In Figures 3.1c and 3.1d, we have the experiments for the stationary ARCH(1) models, with and without the AR-part, respectively, as outlined in Sections 3.4.2 and 3.4.3. These figures illustrate the lack of convexity when using small streaming batch sizes, e.g., the traditional SG method and its average estimate, $\{C_\rho = 1, \rho = 0\}$ diverges. Remark that the lack of convexity is expressed through

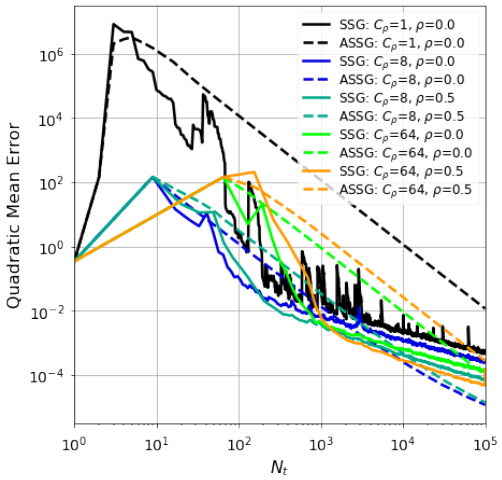
[3] A modification of our average estimate to a weighted average version could improve convergence as it could limit the effect of poor initializations [27, 95]. But despite this, we still achieve better convergence for the ASSG method.

Figure 3.1: Simulation of various data streams $n_t = C_\rho t^\rho$. See Section 3.4 for details.

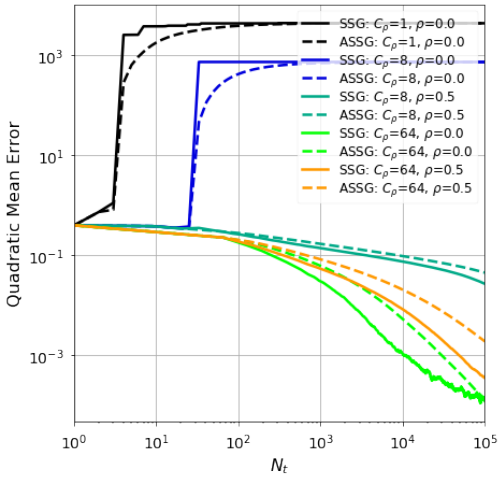
(a) AR(1): well-specified case. See Section 3.4.1 for details.



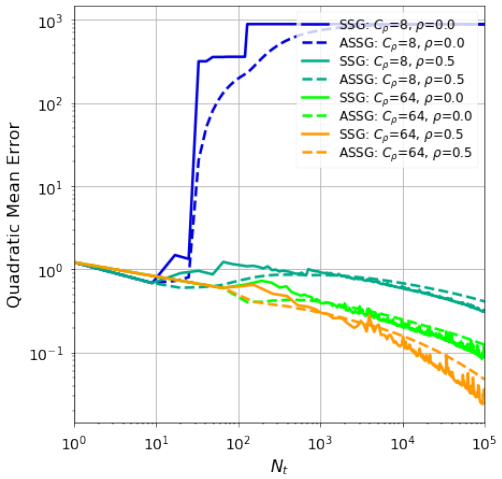
(b) AR(1): misspecified case. See Section 3.4.1 for details.



(c) ARCH(1). See Section 3.4.2 for details.



(d) AR(1)-ARCH(1). See Section 3.4.3 for details.



the lack positively of μ_ν, which only larger streaming batch sizes C_ρ can counteract. Moreover, the traditional SG method, $\{C_\rho = 1, \rho = 0\}$ is omitted in Figure 3.1d due to lack of convexity. Figure 3.1d shows that large ($C_\rho = 64$) and non-decreasing ($\rho \geq 0$) streaming batches can converge under difficult settings. To conclude, by taking ρ positive and C_ρ large enough to ensure that μ_ν stays positive, then we will converge under long- or short-range dependence with biased gradient estimates.

3.5 Conclusion

We studied the SO problem in a streaming framework using dependent and biased (gradient) estimates. In particular, we explored convergence rates of the SSG and ASSG algorithms in a non-asymptotic manner. The theoretical results formed heuristics that links the level of dependency and convexity to the rest of the model parameters. These heuristics provided new insights into determining optimal learning rates, which can help increase the stability of SG-based methods. In short, by taking positive increasing streaming batches, we will converge under long- or short-range dependence with biased gradient estimates. Our experimentation verified these investigations suggesting large streaming batches for highly dependent data sources. Moreover, in large-scale learning problems with dependence, noisy variables, and lack of convexity, we know how to accelerate convergence and reduce noise through the learning rate and the treatment pattern of the data.

There are several ways to expand our work: first, we can extend our analysis to include streaming batches of any size (not in terms of streaming batch size and streaming rates). Second, an extension to non-strongly convex goals could be beneficial as it will provide more insight into how we can choose robust learning rates [9, 99, 101]. At the same time, this learning rate could be made adaptive such that it is robust to poor initialization and requires less fine-tuning. This last objective is the most important for practitioners as it builds a universality across applications.

3.6 Proofs

Let us start by giving a short sketch of how our proofs section is structured: we start by deriving recursive relations to the desired quantities. Next, we derive a general bounds to the recursive relationship for any (γ_t), (ν_t), (κ_t), (σ_t), and (n_t). Finally, we insert the specific functions forms of (γ_t), (ν_t), (κ_t), (σ_t), and (n_t), which yield the results seen in Theorems 3.3.1 and 3.3.2. Before doing the proofs, we recall a repeating argument used to non-asymptotically bound recursive relations of form (3.6.14):

Proposition 3.6.1 (Godichon-Baggioni et al. [57])**.** *Suppose (ω_t), (α_t), (η_t), and (β_t) to be some non-negative sequences satisfying the recursive relation,*

$$\omega_t \leq [1 - 2\lambda\alpha_t + \eta_t\alpha_t]\omega_{t-1} + \beta_t\alpha_t, \tag{3.6.14}$$

with $\omega_0 \geq 0$ and $\lambda > 0$. Let $C_\omega \geq 1$ be such that $\lambda\alpha_t \leq 1$ for all $t \geq t_\omega$ with $t_\omega = \inf\{t \geq 1 : C_\omega\eta_t \leq$

$\lambda\}$. *Then, for* (α_t) *and* (η_t) *decreasing, we have the upper bound on* (ω_t) *given by*

$$\omega_t \leq \tau_t + \frac{1}{\lambda} \max_{t/2 \leq i \leq t} \beta_i, \tag{3.6.15}$$

with

$$\tau_t = \exp\left(-\lambda \sum_{i=t/2}^{t} \alpha_i\right) \left[\exp\left(C_\omega \sum_{i=1}^{t} \eta_i \alpha_i\right) \left(\omega_0 + \frac{1}{\lambda} \max_{1 \leq i \leq t} \beta_i\right) + \sum_{i=1}^{t/2-1} \beta_i \alpha_i\right].$$

Proposition 3.6.1 shows a simple way to bound (ω_t) in (3.6.14); the bound in (3.6.15) consists of a sub-exponential term τ_t and a noise term $\lambda^{-1} \max_{t/2 \leq i \leq t} \beta_i$. Thus, our attention is on reducing the noise term without damaging the natural decay of the sub-exponential term where $\tau_t \to 0$ exponentially fast as $t \to \infty$.

Later in the proofs, we will insert some specific types of the sequences above, resulting in different generalized harmonic numbers, which can be bounded with the integral test for convergence. Moreover, to present our results in terms of $N_t = \sum_{i=1}^{t} n_i$, we will use that $(N_t/2C_\rho)^{1/(1+\rho)} \leq t \leq (2N_t/C_\rho)^{1/(1+\rho)}$. To ease notation, we will make use of the functions $\psi_x(t), \psi_x^y(t) : \mathbb{R}_+ \setminus \{0\} \to \mathbb{R}$, given as

$$\psi_x(t) = \begin{cases} t^{1-x}/(1-x) & \text{if } x < 1, \\ 1 + \log(t) & \text{if } x = 1, \\ x/(x-1) & \text{if } x > 1, \end{cases} \quad \text{and} \quad \psi_x^y(t) = \begin{cases} t^{(1-x)/(1+y)}/(1-x) & \text{if } x < 1, \\ 1 + \log(t^{1/(1+y)}) & \text{if } x = 1, \\ x/(x-1) & \text{if } x > 1, \end{cases} \tag{3.6.16}$$

with $y \in \mathbb{R}_+$ such that $\psi_x^y(t) = \psi_x(t^{1/(1+y)})$. Thus, $\sum_{i=1}^{t} i^{-x} \leq \psi_x(t)$ for any $x \geq 0$. Furthermore, with this notation, we have that $\psi_x^y(t)/t = \mathcal{O}(t^{-(x+y)/(1+y)})$ if $x < 1$, $\psi_x^y(t)/t = \mathcal{O}(\log(t)t^{-1})$ if $x = 1$, and $\psi_x^y(t)/t = \mathcal{O}(t^{-1})$ if $x > 1$. Hence, for any $x_0, x_1, x_2, y \geq 0$, $\psi_{x_0}^y(t)/t = \tilde{\mathcal{O}}(t^{-(x_0+y)/(1+y)})$ and $\psi_{x_1}^y(t)\psi_{x_2}^y(t)/t = \tilde{\mathcal{O}}(t^{-(x_1+x_2+y-1)/(1+y)})$, where the $\tilde{\mathcal{O}}(\cdot)$ notation suppress logarithmic factors.

3.6.1 Proofs for Section 3.3.1

In the following lemma, we derive an explicit recursive relation of $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ to non-asymptotically bound the t-th estimate of (3.2.1) for any (γ_t), (ν_t), (κ_t), (σ_t), and (n_t) using classical techniques from stochastic approximations [13, 87]. As mentioned in Zinkevich [157], bounding the projected estimate in (3.2.2) follows directly from that $\mathbb{E}[\|\mathcal{P}_\Theta(\theta) - \theta^*\|^2] \leq \mathbb{E}[\|\theta - \theta^*\|^2]$, $\forall \theta \in \mathbb{R}^d$, $\forall \theta^* \in \Theta$, as Θ is a closed convex set.

Lemma 3.6.1 (Second-order moment)**.** *Assume that Assumptions 3.2.1-p to 3.2.3-p hold true for* $p = 2$. *Suppose that* $\mu_\nu = \mu - \mathbb{1}_{\{\nu_t = \mathcal{C}\}} 2D_\nu \nu_t > 0$. *Let* $\mathbb{1}_{\{\nu_t = \mathcal{C}\}}$ *and* $\mathbb{1}_{\{\nu_t \neg \mathcal{C}\}}$ *indicate whether* (ν_t) *is constant or not. Denote* $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ *for some* $\delta_0 \geq 0$, *where* (θ_t) *follows the recursion in*

(3.2.1) *or* (3.2.2). *For any learning rate* (γ_t), *we have*

$$\delta_t \leq \pi_t + \frac{2B_\nu^2}{\mu\mu_\nu} \max_{t/2\leq i\leq t} \nu_i^2 + \frac{4}{\mu_\nu} \max_{t/2\leq i\leq t} \sigma_i^2\gamma_i,$$

with

$$\pi_t = \exp\left(-\frac{\mu_\nu}{2}\sum_{i=t/2}^{t}\gamma_i\right)\left[\exp\left(\mathbb{1}_{\{\nu_t\neg\subset\}}2C_\delta D_\nu\sum_{i=1}^{t}\nu_i\gamma_i\right)\exp\left(2C_\delta\sum_{i=1}^{t}\kappa_i^2\gamma_i^2\right)\right.$$
$$\left.\left(\delta_0 + \frac{2B_\nu^2}{\mu\mu_\nu}\max_{1\leq i\leq t}\nu_i^2 + \frac{4}{\mu_\nu}\max_{1\leq i\leq t}\sigma_i^2\gamma_i\right) + \frac{B_\nu^2}{\mu}\sum_{i=1}^{t/2-1}\nu_i^2\gamma_i + 2\sum_{i=1}^{t/2-1}\sigma_i^2\gamma_i^2\right].$$

Proof of Lemma 3.6.1. By taking the quadratic norm on (3.2.1), expanding the norm, and taking the expectation, we can derive the equation,

$$\delta_t = \delta_{t-1} + \gamma_t^2\mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1})\|^2] - 2\gamma_t\mathbb{E}[\langle\nabla_\theta l_t(\theta_{t-1}), \theta_{t-1}-\theta^*\rangle], \tag{3.6.17}$$

where $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ with $\delta_0 \geq 0$. To bound the second term in (3.6.17), we use Assumptions 3.2.2-p and 3.2.3-p for $p=2$, to obtain that

$$\begin{aligned}\mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1})\|^2] =& \mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1}) - \nabla_\theta l_t(\theta^*) + \nabla_\theta l_t(\theta^*)\|^2]\\ \leq& 2\mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1}) - \nabla_\theta l_t(\theta^*)\|^2] + 2\mathbb{E}[\|\nabla_\theta l_t(\theta^*)\|^2]\\ \leq& 2\kappa_t^2\delta_{t-1} + 2\sigma_t^2,\end{aligned} \tag{3.6.18}$$

as $\|x+y\|^p \leq 2^{p-1}(\|x\|^p + \|y\|^p)$. As noted in Bottou et al. [21], Nesterov et al. [104], (3.2.4) implies that $\langle\nabla_\theta L(\theta), \theta-\theta^*\rangle \geq \mu\|\theta-\theta^*\|^2$ for all $\theta\in\Theta\subseteq\mathbb{R}^d$. Next, since L is $\mu-$strongly convex (3.2.4) and θ_{t-1} is $\mathcal{F}_{t-1}$-measurable (Assumption 3.2.1-p), we can bound the third term on the right-hand side of (3.6.17) by

$$\begin{aligned}\mathbb{E}[\langle\nabla_\theta l_t(\theta_{t-1}), \theta_{t-1}-\theta^*\rangle] =& \mathbb{E}[\langle\nabla_\theta L(\theta_{t-1}), \theta_{t-1}-\theta^*\rangle]\\ &+ \mathbb{E}[\langle\mathbb{E}[\nabla_\theta l_t(\theta_{t-1})|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta_{t-1}), \theta_{t-1}-\theta^*\rangle]\\ \geq& \mu\delta_{t-1} - D_\nu\nu_t\delta_{t-1} - B_\nu\nu_t\delta_{t-1}^{1/2},\end{aligned} \tag{3.6.19}$$

since

$$\begin{aligned}&\mathbb{E}[\langle\mathbb{E}[\nabla_\theta l_t(\theta_{t-1})|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta_{t-1}), \theta_{t-1}-\theta^*\rangle] \geq -\mathbb{E}[\|\mathbb{E}[\nabla_\theta l_t(\theta_{t-1})|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta_{t-1})\|\|\theta_{t-1}-\theta^*\|]\\ &\geq -\sqrt{\mathbb{E}[\|\mathbb{E}[\nabla_\theta l_t(\theta_{t-1})|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta_{t-1})\|^2]}\sqrt{\mathbb{E}[\|\theta_{t-1}-\theta^*\|^2]}\\ &\geq -\sqrt{\nu_t^2(D_\nu^2\delta_{t-1} + B_\nu^2)}\sqrt{\delta_{t-1}} \geq -D_\nu\nu_t\delta_{t-1} - B_\nu\nu_t\sqrt{\delta_{t-1}},\end{aligned}$$

by Jensen's inequality, Cauchy–Schwarz inequality, Hölder's inequality, and Assumption 3.2.1-p

with $p = 2$. Hence, applying the inequalities (3.6.18) and (3.6.19) to (3.6.17), yields

$$\begin{aligned}\delta_t \leq & [1 - 2\mu\gamma_t + 2D_\nu\nu_t\gamma_t + 2\kappa_t^2\gamma_t^2]\delta_{t-1} + 2B_\nu\nu_t\gamma_t\delta_{t-1}^{1/2} + 2\sigma_t^2\gamma_t^2 \\ \leq & [1 - (\mu - 2D_\nu\nu_t)\gamma_t + 2\kappa_t^2\gamma_t^2]\delta_{t-1} + \frac{B_\nu^2}{\mu}\nu_t^2\gamma_t + 2\sigma_t^2\gamma_t^2,\end{aligned}$$

using Young's inequality[4]; $2B_\nu\nu_t\gamma_t\delta_{t-1}^{1/2} \leq \mu\gamma_t\delta_{t-1} + B_\nu^2\nu_t^2\gamma_t/\mu$. Next, we introduce the indicator function for whether (ν_t) is constant $(= \mathcal{C})$ or not $(\neg\mathcal{C})$, such that

$$\delta_t \leq [1 - (\mu_\nu - \mathbb{1}_{\{\nu_t \neg\mathcal{C}\}}2D_\nu\nu_t)\gamma_t + 2\kappa_t^2\gamma_t^2]\delta_{t-1} + \frac{B_\nu^2}{\mu}\nu_t^2\gamma_t + 2\sigma_t^2\gamma_t^2, \tag{3.6.20}$$

with $\mu_\nu = \mu - \mathbb{1}_{\{\nu_t = \mathcal{C}\}}2D_\nu\nu_t > 0$. Let C_δ be the constant fulfilling the conditions of Proposition 3.6.1 such that C_δ is chosen larger than 1 verifying $C_\delta(\mathbb{1}_{\{\nu_t\neg\mathcal{C}\}}2D_\nu\nu_t + 2\kappa_t^2\gamma_t) \leq \mu_\nu/2$ such that it's imply $\mu_\nu\gamma_t/2 \leq 1$, which is possible as the sequence (ν_t) is non-increasing, and (κ_t) and (γ_t) is decreasing. At last, bounding (3.6.20) by Proposition 3.6.1 concludes the proof. □

Proof of Theorem 3.3.1. Inserting the functions $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$, $\nu_t = n_t^{-\nu}$, $\kappa_t = C_\kappa n_t^{-\kappa}$, $\sigma_t = C_\sigma n_t^{-\sigma}$, and $n_t = C_\rho t^\rho$ into the bound of Lemma 3.6.1 yields

$$\begin{aligned}\delta_t \leq & \pi_t + \frac{2^{1+2\rho\nu}B_\nu^2}{\mu\mu_\nu C_\rho^{2\nu}t^{2\rho\nu}} + \frac{2^{2+\rho(2\sigma-\beta)+\alpha}C_\sigma^2 C_\gamma C_\rho^\beta}{\mu_\nu C_\rho^{2\sigma}t^{\rho(2\sigma-\beta)+\alpha}} \\ \leq & \pi_t + \frac{2^{(2+6\rho\nu)/(1+\rho)}B_\nu^2}{\mu\mu_\nu C_\rho^{2\nu/(1+\rho)}N_t^{2\rho\nu/(1+\rho)}} + \frac{2^{(7+6\rho\sigma)/(1+\rho)}C_\sigma^2 C_\gamma}{\mu_\nu C_\rho^{(2\sigma-\beta-\alpha)/(1+\rho)}N_t^{(\rho(2\sigma-\beta)+\alpha)/(1+\rho)}},\end{aligned} \tag{3.6.21}$$

[4] If $a, b, c > 0$, $p, q > 1$ such that $1/p + 1/q = 1$, then $ab \leq a^p c^p/p + b^q/qc^q$.

with $\mu_\nu = \mu - \mathbb{1}_{\{\rho=0\}} 2D_\nu C_\rho^{-\nu} > 0$, and π_t can be bounded by

$$
\begin{aligned}
&\exp\left(-\frac{\mu_\nu C_\gamma C_\rho^\beta}{2}\sum_{i=t/2}^{t} i^{\rho\beta-\alpha}\right)\left[\exp\left(\frac{\mathbb{1}_{\{\rho\neq 0\}} 2C_\delta D_\nu C_\gamma C_\rho^\beta}{C_\rho^\nu}\sum_{i=1}^{t} i^{\rho(\beta-\nu)-\alpha}\right)\right.\\
&\exp\left(\frac{2C_\delta C_\kappa^2 C_\gamma^2 C_\rho^{2\beta}}{C_\rho^{2\kappa}}\sum_{i=1}^{t} i^{2\rho(\beta-\kappa)-2\alpha}\right)\left(\delta_0 + \frac{2B_\nu^2}{\mu\mu_\nu C_\rho^{2\nu}} + \frac{4C_\sigma^2 C_\gamma C_\rho^\beta}{\mu_\nu C_\rho^{2\sigma}}\right)\\
&\left.+\frac{B_\nu^2 C_\gamma C_\rho^\beta}{\mu C_\rho^{2\nu}}\sum_{i=1}^{t/2-1} i^{\rho(\beta-2\nu)-\alpha} + \frac{2C_\sigma^2 C_\gamma^2 C_\rho^{2\beta}}{C_\rho^{2\sigma}}\sum_{i=1}^{t/2-1} i^{2\rho(\beta-\sigma)-2\alpha}\right]\\
&\le \exp\left(-\frac{\mu_\nu C_\gamma C_\rho^\beta t^{1+\rho\beta-\alpha}}{2^2}\right)\left[\exp\left(\frac{\mathbb{1}_{\{\rho\neq 0\}} 2C_\delta D_\nu C_\gamma C_\rho^\beta \psi_{\alpha-\rho(\beta-\nu)}(t)}{C_\rho^\nu}\right)\right.\\
&\exp\left(\frac{4(\alpha-\rho(\beta-\kappa))C_\delta C_\kappa^2 C_\gamma^2 C_\rho^{2\beta}}{(2\alpha-2\rho(\beta-\kappa)-1)C_\rho^{2\kappa}}\right)\left(\delta_0 + \frac{2B_\nu^2}{\mu\mu_\nu C_\rho^{2\nu}} + \frac{4C_\sigma^2 C_\gamma C_\rho^\beta}{\mu_\nu C_\rho^{2\sigma}}\right)\\
&\left.+\frac{B_\nu^2 C_\gamma C_\rho^\beta \psi_{\alpha-\rho(\beta-2\nu)}(t/2)}{\mu C_\rho^{2\nu}} + \frac{4(\alpha-\rho(\beta-\sigma))C_\sigma^2 C_\gamma^2 C_\rho^{2\beta}}{(2\alpha-2\rho(\beta-\sigma)-1)C_\rho^{2\sigma}}\right]\\
&\le \exp\left(-\frac{\mu C_\gamma N_t^{(1+\rho\beta-\alpha)/(1+\rho)}}{2^{(3+\rho(2+\beta)-\alpha)/(1+\rho)} C_\rho^{(1-\beta-\alpha)/(1+\rho)}}\right)\left[\exp\left(\frac{\mathbb{1}_{\{\rho\neq 0\}} 2C_\delta D_\nu C_\gamma C_\rho^\beta \psi^\rho_{\alpha-\rho(\beta-\nu)}(2N_t/C_\rho)}{C_\rho^\nu}\right)\right.\\
&\exp\left(\frac{4(\alpha-\rho(\beta-\kappa))C_\delta C_\kappa^2 C_\gamma^2 C_\rho^{2\beta}}{(2\alpha-2\rho(\beta-\kappa)-1)C_\rho^{2\kappa}}\right)\left(\delta_0 + \frac{2B_\nu^2}{\mu\mu_\nu C_\rho^{2\nu}} + \frac{4C_\sigma^2 C_\gamma C_\rho^\beta}{\mu_\nu C_\rho^{2\sigma}}\right)\\
&\left.+\frac{B_\nu^2 C_\gamma C_\rho^\beta \psi^\rho_{\alpha-\rho(\beta-2\nu)}(N_t/C_\rho)}{\mu C_\rho^{2\nu}} + \frac{4(\alpha-\rho(\beta-\sigma))C_\sigma^2 C_\gamma^2 C_\rho^{2\beta}}{(2\alpha-2\rho(\beta-\sigma)-1)C_\rho^{2\sigma}}\right],
\end{aligned}
\tag{3.6.22}
$$

with help of an integral test for convergence[5], $\psi_x(t)$ and $\psi_x^y(t)$ from (3.6.16), and by use of $(N_t/2C_\rho)^{1/(1+\rho)} \le t \le (2N_t/C_\rho)^{1/(1+\rho)}$. □

3.6.2 Proofs for Section 3.3.2

As in Section 3.6.1, we begin the following sections by conducting a general study for any (γ_t), (ν_t), (κ_t), (σ_t), and (n_t), when applying the Polyak-Ruppert averaging estimate $(\bar{\theta}_t)$ from (3.2.3). Moreover, we need to study fourth-order rate $\Delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^4]$ of the recursive estimates (3.2.1) and (3.2.2).

Lemma 3.6.2 (Fourth-order moment)**.** *Assume that Assumptions 3.2.1-p to 3.2.3-p hold true for $p = 4$. Suppose that $\mu'_\nu = \mu - \mathbb{1}_{\{\nu_t=\mathcal{C}\}} 2D_\nu^4 \nu_t^4/\mu^3 > 0$. Let $\mathbb{1}_{\{\nu_t=\mathcal{C}\}}$ and $\mathbb{1}_{\{\nu_t\neg\mathcal{C}\}}$ indicate whether (ν_t) is constant or not. Denote $\Delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^4]$ for some $\Delta_0 \ge 0$, where (θ_t) follows the recursion in*

[5]Note that $\sum_{i=1}^t i^{2\rho(\beta-\kappa)-2\alpha} \le (2\alpha - 2\rho(\beta-\kappa))/(2\alpha-2\rho(\beta-\kappa)-1)$ and $\sum_{i=1}^t i^{2\rho(\beta-\sigma)-2\alpha} \le (2\alpha-2\rho(\beta-\sigma))/(2\alpha-2\rho(\beta-\sigma)-1)$ as $\nu > 0$, $\sigma, \kappa \in [0, 1/2]$, $\rho \in [0, 1)$, $\beta \in [0, 1]$, and $\alpha - \rho\beta \in (1/2, 1)$.

(3.2.1) *or* (3.2.2). *For any learning rate* (γ_t), *we have*

$$\Delta_t \leq \Pi_t + \frac{4B_\nu^4}{\mu^3\mu'_\nu}\max_{t/2\leq i\leq t}\nu_i^4 + \frac{1024}{\mu\mu'_\nu}\max_{t/2\leq i\leq t}\sigma_i^4\gamma_i^2 + \frac{96}{\mu'_\nu}\max_{t/2\leq i\leq t}\sigma_i^4\gamma_i^3,$$

with

$$\begin{aligned}\Pi_t = &\exp\left(-\frac{\mu'_\nu}{4}\sum_{i=t/2}^{t}\gamma_i\right)\left[\exp\left(\frac{\mathbb{1}_{\{\nu_t\neg\mathcal{C}\}}C_\Delta D_\nu^4}{\mu^3}\sum_{i=1}^{t}\nu_i^4\gamma_i\right)\exp\left(\frac{256C_\Delta}{\mu}\sum_{i=1}^{t}\kappa_i^4\gamma_i^3\right)\right.\\
&\exp\left(24C_\Delta\sum_{i=1}^{t}\kappa_i^4\gamma_i^4\right)\left(\Delta_0 + \frac{4B_\nu^4}{\mu^3\mu'_\nu}\max_{1\leq i\leq t}\nu_i^4 + \frac{1024}{\mu\mu'_\nu}\max_{1\leq i\leq t}\sigma_i^4\gamma_i^2 + \frac{96}{\mu'_\nu}\max_{1\leq i\leq t}\sigma_i^4\gamma_i^3\right)\\
&\left.+\frac{B_\nu^4}{\mu^3}\sum_{i=1}^{t/2-1}\nu_i^4\gamma_i + \frac{256}{\mu}\sum_{i=1}^{t/2-1}\sigma_i^4\gamma_i^3 + 24\sum_{i=1}^{t/2-1}\sigma_i^4\gamma_i^4\right].\end{aligned}$$

Proof of Lemma 3.6.2. The derivation of the recursive step sequence for the fourth-order moment Δ_t of (3.2.1) follows the same methodology as for the second-order moment in Lemma 3.6.1. In the same way we deduced (3.6.17), we can take the quadratic norm on (3.2.1), expand the norm, and take the square on both sides, to derive the equation

$$\begin{aligned}\|\theta_t-\theta^*\|^4 =&(\|\theta_{t-1}-\theta^*\|^2+\gamma_t^2\|\nabla_\theta l_t(\theta_{t-1})\|^2-2\gamma_t\langle\nabla_\theta l_t(\theta_{t-1}),\theta_{t-1}-\theta^*\rangle)^2\\
=&\|\theta_{t-1}-\theta^*\|^4+\gamma_t^4\|\nabla_\theta l_t(\theta_{t-1})\|^4+4\gamma_t^2\langle\nabla_\theta l_t(\theta_{t-1}),\theta_{t-1}-\theta^*\rangle^2+2\gamma_t^2\|\theta_{t-1}-\theta^*\|^2\|\nabla_\theta l_t(\theta_{t-1})\|^2\\
&-4\gamma_t\|\theta_{t-1}-\theta^*\|^2\langle\nabla_\theta l_t(\theta_{t-1}),\theta_{t-1}-\theta^*\rangle-4\gamma_t^3\|\nabla_\theta l_t(\theta_{t-1})\|^2\langle\nabla_\theta l_t(\theta_{t-1}),\theta_{t-1}-\theta^*\rangle.\end{aligned}$$

Taking the expectation on both sides of the equality above gives us

$$\begin{aligned}\Delta_t =&\Delta_{t-1}+\gamma_t^4\mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1})\|^4]+4\gamma_t^2\mathbb{E}[\langle\nabla_\theta l_t(\theta_{t-1}),\theta_{t-1}-\theta^*\rangle^2]+2\gamma_t^2\mathbb{E}[\|\theta_{t-1}-\theta^*\|^2\|\nabla_\theta l_t(\theta_{t-1})\|^2]\\
&-4\gamma_t\mathbb{E}[\|\theta_{t-1}-\theta^*\|^2\langle\nabla_\theta l_t(\theta_{t-1}),\theta_{t-1}-\theta^*\rangle]-4\gamma_t^3\mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1})\|^2\langle\nabla_\theta l_t(\theta_{t-1}),\theta_{t-1}-\theta^*\rangle]\\
\leq&\Delta_{t-1}+\gamma_t^4\mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1})\|^4]+6\gamma_t^2\mathbb{E}[\|\theta_{t-1}-\theta^*\|^2\|\nabla_\theta l_t(\theta_{t-1})\|^2]\\
&-4\gamma_t\mathbb{E}[\|\theta_{t-1}-\theta^*\|^2\langle\nabla_\theta l_t(\theta_{t-1}),\theta_{t-1}-\theta^*\rangle]+4\gamma_t^3\mathbb{E}[\|\theta_{t-1}-\theta^*\|\|\nabla_\theta l_t(\theta_{t-1})\|^3],\end{aligned}$$

by use of Cauchy-Schwarz inequality. Next, Young's inequality yields $4\gamma_t^3\|\theta_{t-1}-\theta^*\|\|\nabla_\theta l_t(\theta_{t-1})\|^3 \leq 2\gamma_t^4\|\nabla_\theta l_t(\theta_{t-1})\|^4+2\gamma_t^2\|\theta_{t-1}-\theta^*\|^2\|\nabla_\theta l_t(\theta_{t-1})\|^2$ and $8\gamma_t^2\|\theta_{t-1}-\theta^*\|^2\|\nabla_\theta l_t(\theta_{t-1})\|^2 \leq (\mu\gamma_t/2)\|\theta_{t-1}-\theta^*\|^4+32\mu^{-1}\gamma_t^3\|\nabla_\theta l_t(\theta_{t-1})\|^4$, which helps us to obtain the simplified expression,

$$\begin{aligned}\Delta_t \leq&[1+\mu\gamma_t/2]\Delta_{t-1}+3\gamma_t^4\mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1})\|^4]+32\mu^{-1}\gamma_t^3\mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1})\|^4]\\
&-4\gamma_t\mathbb{E}[\|\theta_{t-1}-\theta^*\|^2\langle\nabla_\theta l_t(\theta_{t-1}),\theta_{t-1}-\theta^*\rangle].\end{aligned}$$

To bound the fourth-order term $\mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1})\|^4]$, we make use of the Lipschitz continuity of $\nabla_\theta l_t$ (Assumption 3.2.2-p), Assumption 3.2.3-p, and that θ_{t-1} is $\mathcal{F}_{t-1}$-measurable (Assumption 3.2.1-p),

to have that

$$\mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1})\|^4] \leq 8\kappa_t^4 \Delta_{t-1} + 8\sigma_t^4, \tag{3.6.23}$$

using that $\|x+y\|^p \leq 2^{p-1}(\|x\|^p + \|y\|^p)$ for any $p \in \mathbb{N}$. Thus,

$$\begin{aligned}\Delta_t \leq & [1 + \mu\gamma_t/2 + 256\mu^{-1}\kappa_t^4\gamma_t^3 + 24\kappa_t^4\gamma_t^4]\Delta_{t-1} + 256\mu^{-1}\sigma_t^4\gamma_t^3 + 24\sigma_t^4\gamma_t^4 \\ & - 4\gamma_t \mathbb{E}[\|\theta_{t-1} - \theta^*\|^2 \langle \nabla_\theta l_t(\theta_{t-1}), \theta_{t-1} - \theta^* \rangle].\end{aligned} \tag{3.6.24}$$

Next, using the same arguments as in the proof of Lemma 3.6.1, Young's inequality, and Assumption 3.2.1-p with $p = 4$, we have

$$\begin{aligned}& 4\gamma_t \mathbb{E}[\|\theta_{t-1} - \theta^*\|^2 \langle \mathbb{E}[\nabla_\theta l_t(\theta_{t-1})|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta_{t-1}), \theta_{t-1} - \theta^* \rangle] \\ & \geq -4\gamma_t \mathbb{E}[\|\theta_{t-1} - \theta^*\|^3 \|\mathbb{E}[\nabla_\theta l_t(\theta_{t-1})|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta_{t-1})\|] \\ & \geq -3\mu\gamma_t\Delta_{t-1} - \mu^{-3}\gamma_t \mathbb{E}[\|\mathbb{E}[\nabla_\theta l_t(\theta_{t-1})|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta_{t-1})\|^4] \\ & \geq -3\mu\gamma_t\Delta_{t-1} - \mu^{-3}\gamma_t D_\nu^4 \nu_t^4 \Delta_{t-1} - \mu^{-3}\gamma_t B_\nu^4 \nu_t^4,\end{aligned}$$

such that the last term of (3.6.24) can be bounded as follows,

$$\begin{aligned}4\gamma_t \mathbb{E}[\|\theta_{t-1} - \theta^*\|^2 \langle \nabla_\theta l_t(\theta_{t-1}), \theta_{t-1} - \theta^* \rangle] =& 4\gamma_t \mathbb{E}[\|\theta_{t-1} - \theta^*\|^2 \langle \mathbb{E}[\nabla_\theta l_t(\theta_{t-1})|\mathcal{F}_{t-1}], \theta_{t-1} - \theta^* \rangle] \\ =& 4\gamma_t \mathbb{E}[\|\theta_{t-1} - \theta^*\|^2 \langle \nabla_\theta L(\theta_{t-1}), \theta_{t-1} - \theta^* \rangle] \\ & + 4\gamma_t \mathbb{E}[\|\theta_{t-1} - \theta^*\|^2 \langle \mathbb{E}[\nabla_\theta l_t(\theta_{t-1})|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta_{t-1}), \theta_{t-1} - \theta^* \rangle] \\ \geq & \mu\gamma_t\Delta_{t-1} - \mu^{-3}\gamma_t D_\nu^4 \nu_t^4 \Delta_{t-1} - \mu^{-3}\gamma_t B_\nu^4 \nu_t^4.\end{aligned}$$

Indeed, inserting this into (3.6.24) gives us

$$\Delta_t \leq \left[1 - \left(\frac{\mu}{2} - \frac{D_\nu^4 \nu_t^4}{\mu^3}\right)\gamma_t + \frac{256\kappa_t^4\gamma_t^3}{\mu} + 24\kappa_t^4\gamma_t^4\right]\Delta_{t-1} + \frac{B_\nu^4 \nu_t^4 \gamma_t}{\mu^3} + \frac{256\sigma_t^4\gamma_t^3}{\mu} + 24\sigma_t^4\gamma_t^4,$$

which can be modified with use the indicator function that determines whether (ν_t) is constant $(= \mathcal{C})$ or not $(\neg\mathcal{C})$, such that

$$\Delta_t \leq \left[1 - \left(\frac{\mu_\nu}{2} - \frac{\mathbb{1}_{\{\nu_t \neg\mathcal{C}\}} D_\nu^4 \nu_t^4}{\mu^3}\right)\gamma_t + \frac{256\kappa_t^4\gamma_t^3}{\mu} + 24\kappa_t^4\gamma_t^4\right]\Delta_{t-1} + \frac{B_\nu^4 \nu_t^4 \gamma_t}{\mu^3} + \frac{256\sigma_t^4\gamma_t^3}{\mu} + 24\sigma_t^4\gamma_t^4, \tag{3.6.25}$$

with $\mu_\nu' = \mu - \mathbb{1}_{\{\nu_t = \mathcal{C}\}} 2D_\nu^4 \nu_t^4/\mu^3 > 0$. Note that μ_ν from Lemma 3.6.1 is lower bounded by μ_ν', and strictly lower bounded for (ν_t) constant, i.e., $\mu_\nu > \mu_\nu' > 0$. Let $C_\Delta \geq 1$ fulfill the conditions of Proposition 3.6.1; the C_Δ constant is chosen such that $C_\Delta(\mathbb{1}_{\{\nu_t \neg\mathcal{C}\}} D_\nu^4 \nu_t^4/\mu^3 + 256\kappa_t^4\gamma_t^2/\mu + 24\kappa_t^4\gamma_t^3) \leq \mu_\nu'/2$ implying $\mu_\nu'\gamma_t/2 \leq 1$, which is possible as the sequence (ν_t) is non-increasing, and (κ_t) and (γ_t) decrease. Hence, by applying Proposition 3.6.1 on (3.6.25), we obtain the desired bound for

Δ_t. □

Corollary 3.6.1 (Fourth-order moment). *Assume that Assumptions 3.2.1-p to 3.2.3-p hold true for $p = 4$. Let $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$, $\nu_t = n_t^{-\nu}$, $\kappa_t = C_\kappa n_t^{-\kappa}$, and $\sigma_t = C_\sigma n_t^{-\sigma}$ with $\nu \in (0, \infty)$, $\beta \in [0, 1]$, $\kappa, \sigma \in [0, 1/2]$, and $C_\gamma, C_\kappa, C_\sigma > 0$. Suppose $n_t = C_\rho t^\rho$ with $\rho \in [0, 1)$ and $C_\rho \in \mathbb{N}$, such that $\mu'_\nu = \mu - \mathbb{1}_{\{\rho=0\}} 2D_\nu^4/\mu^3 C_\rho^{4\nu} > 0$. Denote $\Delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^4]$ for some $\Delta_0 \geq 0$, where (θ_t) follows the recursion in (3.2.1) or (3.2.2). For $\alpha - \rho\beta \in (1/2, 1)$, we have*

$$\Delta_t \leq \Pi_t + \frac{2^{2+4\rho\nu} B_\nu^4}{\mu^3 \mu'_\nu C_\rho^{4\nu} t^{4\rho\nu}} + \frac{2^{2\rho(2\sigma-\beta)+2\alpha}(2^{10}\mu^{-1} + 2^7 C_\gamma C_\rho^\beta) C_\sigma^4 C_\gamma^2 C_\rho^{2\beta}}{\mu'_\nu C_\rho^{4\sigma} t^{2\rho(2\sigma-\beta)+2\alpha}}, \tag{3.6.26}$$

with Π_t given in (3.6.27) such that $\Pi_t = \mathcal{O}(\exp(-N_t^{(1+\rho\beta-\alpha)/(1+\rho)}))$.

Proof of Corollary 3.6.1. Inserting the functions $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$, $\nu_t = n_t^{-\nu}$, $\kappa_t = C_\kappa n_t^{-\kappa}$, $\sigma_t = C_\sigma n_t^{-\sigma}$, and $n_t = C_\rho t^\rho$ into the bound of Lemma 3.6.2 and using $\gamma_t^3 \leq C_\gamma C_\rho^\beta \gamma_t^2$ as $\alpha - \rho\beta \in (1/2, 1)$, yields (3.6.26) with $\mu'_\nu = \mu - \mathbb{1}_{\{\rho=0\}} 2D_\nu^4/\mu^3 C_\rho^{4\nu} > 0$, where Π_t can be bounded as follows,

$$\begin{aligned}
&\exp\left(-\frac{\mu'_\nu C_\gamma C_\rho^\beta}{4} \sum_{i=t/2}^{t} i^{\rho\beta-\alpha}\right)\left[\exp\left(\frac{\mathbb{1}_{\{\rho\neq 0\}} C_\Delta D_\nu^4 C_\gamma C_\rho^\beta}{\mu^3 C_\rho^{4\nu}} \sum_{i=1}^{t} i^{\rho(\beta-4\nu)-\alpha}\right)\right. \\
&\exp\left(\frac{2^8 C_\Delta C_\kappa^4 C_\gamma^3 C_\rho^{3\beta}}{\mu C_\rho^{4\kappa}} \sum_{i=1}^{t} i^{\rho(3\beta-4\kappa)-3\alpha}\right) \exp\left(\frac{24 C_\Delta C_\kappa^4 C_\gamma^4 C_\rho^{4\beta}}{C_\rho^{4\kappa}} \sum_{i=1}^{t} i^{4\rho(\beta-\kappa)-4\alpha}\right) \\
&\left(\Delta_0 + \frac{4B_\nu^4}{\mu^3 \mu'_\nu C_\rho^{4\nu}} + \frac{1024 C_\sigma^4 C_\gamma^2 C_\rho^{2\beta}}{\mu \mu'_\nu C_\rho^{4\sigma}} + \frac{96 C_\sigma^4 C_\gamma^3 C_\rho^{3\beta}}{\mu'_\nu C_\rho^{4\sigma}}\right) \\
&\left. + \frac{B_\nu^4 C_\gamma C_\rho^\beta}{\mu^3 C_\rho^{4\nu}} \sum_{i=1}^{t/2-1} i^{\rho(\beta-4\nu)-\alpha} + \frac{256 C_\sigma^4 C_\gamma^3 C_\rho^{3\beta}}{\mu C_\rho^{4\sigma}} \sum_{i=1}^{t/2-1} i^{\rho(3\beta-4\sigma)-3\alpha} + \frac{24 C_\sigma^4 C_\gamma^4 C_\rho^{4\beta}}{C_\rho^{4\sigma}} \sum_{i=1}^{t/2-1} i^{4\rho(\beta-\sigma)-4\alpha}\right] \\
&\leq \exp\left(-\frac{\mu'_\nu C_\gamma C_\rho^\beta t^{1+\rho\beta-\alpha}}{2^3}\right)\left[\exp\left(\frac{\mathbb{1}_{\{\rho\neq 0\}} C_\Delta D_\nu^4 C_\gamma C_\rho^\beta \psi^0_{\alpha-\rho(\beta-4\nu)}(t)}{\mu^3 C_\rho^{4\nu}}\right) \exp\left(\frac{2^{10} C_\Delta C_\kappa^4 C_\gamma^3 C_\rho^{3\beta}}{\mu C_\rho^{4\kappa}}\right)\right. \\
&\exp\left(\frac{2^6 C_\Delta C_\kappa^4 C_\gamma^4 C_\rho^{4\beta}}{C_\rho^{4\kappa}}\right)\left(\Delta_0 + \frac{2^2 B_\nu^4}{\mu^3 \mu'_\nu C_\rho^{4\nu}} + \frac{2^{10} C_\sigma^4 C_\gamma^2 C_\rho^{2\beta}}{\mu \mu'_\nu C_\rho^{4\sigma}} + \frac{2^7 C_\sigma^4 C_\gamma^3 C_\rho^{3\beta}}{\mu'_\nu C_\rho^{4\sigma}}\right) \\
&\left. + \frac{B_\nu^4 C_\gamma C_\rho^\beta \psi^0_{\alpha-\rho(\beta-4\nu)}(t/2)}{\mu^3 C_\rho^{4\nu}} + \frac{2^{10} C_\sigma^4 C_\gamma^3 C_\rho^{3\beta}}{\mu C_\rho^{4\sigma}} + \frac{2^6 C_\sigma^4 C_\gamma^4 C_\rho^{4\beta}}{C_\rho^{4\sigma}}\right],
\end{aligned} \tag{3.6.27}$$

with help of the integral test for convergence; $\sum_{i=1}^t i^{\rho(3\beta-4x)-3\alpha} \leq 3 < 2^2$ and $\sum_{i=1}^t i^{4\rho(\beta-x)-4\alpha} \leq 2$ for any $x \geq 0$ as $\alpha - \rho\beta \in (1/2, 1)$. □

Lemma 3.6.3. *Assume that Assumptions 3.2.1-p to 3.2.3-p for $p = 4$ and Assumption 3.3.1 hold true. Denote $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ with $\bar{\theta}_n$ given by (3.2.3), where (θ_t) follows the recursion in (3.2.1) or (3.2.2). In addition, Assumption 3.3.2 must hold true if (θ_t) follows the recursion in (3.2.2),*

which is indicated by $\mathbb{1}_{\{D_\Theta<\infty\}}$. *For any learning rate* (γ_t), *we have*

$$\begin{aligned}\bar{\delta}_t^{1/2} \leq & \frac{\Lambda^{1/2}}{N_t}\left(\sum_{i=1}^{t} n_i^{2(1-\sigma)}\right)^{1/2} + \frac{C_\sigma'^{1/2}}{\mu N_t}\left(\sum_{i=1}^{t} n_i^{2(1-\sigma-\sigma')}\right)^{1/2} + \frac{2^{1/2}B_\nu^{1/2}}{\mu N_t}\left(\sum_{j=2}^{t}\left(n_j\nu_j\sum_{i=1}^{j-1} n_i\sigma_i\right)\right)^{1/2} \\ & + \frac{1}{\mu N_t}\sum_{i=1}^{t-1}\delta_i^{1/2}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| + \frac{n_t}{\mu\gamma_t N_t}\delta_t^{1/2} + \frac{n_1}{\mu N_t}\left(\frac{1}{\gamma_1} + 2^{1/2}(C_\nabla + \kappa_1)\right)\delta_0^{1/2} \\ & + \frac{2^{1/2}}{\mu N_t}\left(\sum_{i=1}^{t-1} n_{i+1}^2(C_\nabla^2 + \kappa_{i+1}^2)\delta_i\right)^{1/2} + \frac{C_\nabla''}{\mu N_t}\sum_{i=0}^{t-1} n_{i+1}\Delta_i^{1/2} \\ & + \frac{2^{3/4}}{\mu N_t}\left(\sum_{j=1}^{t-1}\left((D_\nu\delta_j^{1/2} + 2^{1/2}B_\nu)n_{j+1}\nu_{j+1}\sum_{i=0}^{j-1}(C_\nabla + \kappa_{i+1})n_{i+1}\delta_i^{1/2}\right)\right)^{1/2},\end{aligned}$$

with $\Lambda = \mathrm{Tr}(\nabla_\theta^2 L(\theta^*)^{-1}\Sigma\nabla_\theta^2 L(\theta^*)^{-1})$ *and* $C_\nabla'' = C_\nabla'/2 + \mathbb{1}_{\{D_\Theta<\infty\}}2G_\Theta/D_\Theta^2$.

Proof of Lemma 3.6.3. The proof is divided into two parts; in the first part, (θ_t) follows (3.2.1), and the second part considers (3.2.2). Assume that (θ_t) is derived from the recursion in (3.2.1): following Polyak and Juditsky [118], we rewrite (3.2.1) to

$$\frac{1}{\gamma_t}(\theta_{t-1} - \theta_t) = \nabla_\theta l_t(\theta_{t-1}), \tag{3.6.28}$$

where $\nabla_\theta l_t(\theta_{t-1}) = n_t^{-1}\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}(\theta_{t-1})$. Observe that

$$\begin{aligned}\nabla_\theta^2 L(\theta^*)(\theta_{t-1} - \theta^*) = & -\nabla_\theta l_t(\theta^*) + \nabla_\theta l_t(\theta_{t-1}) - [\nabla_\theta l_t(\theta_{t-1}) - \nabla_\theta l_t(\theta^*) - \nabla_\theta L(\theta_{t-1})] \\ & - [\nabla_\theta L(\theta_{t-1}) - \nabla_\theta^2 L(\theta^*)(\theta_{t-1} - \theta^*)],\end{aligned}$$

where $\nabla_\theta^2 L(\theta^*)$ is invertible with lowest eigenvalue greater than μ, i.e., $\nabla_\theta^2 L(\theta^*) \geq \mu\mathbb{I}_d$. Thus, summing the parts, taking the quadratic norm and expectation, and using Minkowski's inequality,

gives us the inequality,

$$\begin{aligned}\left(\mathbb{E}\left[\left\|\bar{\theta}_t-\theta^*\right\|^2\right]\right)^{1/2} \leq & \left(\mathbb{E}\left[\left\|\nabla_\theta^2 L\left(\theta^*\right)^{-1} \frac{1}{N_t} \sum_{i=1}^t n_i \nabla_\theta l_i\left(\theta^*\right)\right\|^2\right]\right)^{1/2} \\ & +\left(\mathbb{E}\left[\left\|\nabla_\theta^2 L\left(\theta^*\right)^{-1} \frac{1}{N_t} \sum_{i=1}^t n_i \nabla_\theta l_i\left(\theta_{i-1}\right)\right\|^2\right]\right)^{1/2} \\ & +\left(\mathbb{E}\left[\left\|\nabla_\theta^2 L\left(\theta^*\right)^{-1} \frac{1}{N_t} \sum_{i=1}^t n_i\left[\nabla_\theta l_i\left(\theta_{i-1}\right)-\nabla_\theta l_i\left(\theta^*\right)-\nabla_\theta L\left(\theta_{i-1}\right)\right]\right\|^2\right]\right)^{1/2} \\ & +\left(\mathbb{E}\left[\left\|\nabla_\theta^2 L\left(\theta^*\right)^{-1} \frac{1}{N_t} \sum_{i=1}^t n_i\left[\nabla_\theta L\left(\theta_{i-1}\right)-\nabla_\theta^2 L\left(\theta^*\right)\left(\theta_{i-1}-\theta^*\right)\right]\right\|^2\right]\right)^{1/2} .\end{aligned} \tag{3.6.29}$$

As $(\nabla_\theta l_t(\theta^*))$ is a square-integrable sequences on $\mathbb{R}^d$ (Assumption 3.2.1-p), we have

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla_\theta^2 L\left(\theta^*\right)^{-1} \frac{1}{N_t} \sum_{i=1}^t n_i \nabla_\theta l_i\left(\theta^*\right)\right\|^2\right] = & \frac{1}{N_t^2} \sum_{i=1}^t n_i^2 \mathbb{E}\left[\left\|\nabla_\theta^2 L\left(\theta^*\right)^{-1} \nabla_\theta l_i\left(\theta^*\right)\right\|^2\right] \\ & +\frac{2}{N_t^2} \sum_{1 \leq i<j \leq t} n_i n_j \mathbb{E}\left[\left\langle\nabla_\theta^2 L\left(\theta^*\right)^{-1} \nabla_\theta l_i\left(\theta^*\right), \nabla_\theta^2 L\left(\theta^*\right)^{-1} \nabla_\theta l_j\left(\theta^*\right)\right\rangle\right],\end{aligned}$$

where the first term can be bounded by Assumption 3.3.1,

$$\begin{aligned}\frac{1}{N_t^2} \sum_{i=1}^t n_i^2 \mathbb{E}\left[\left\|\nabla_\theta^2 L\left(\theta^*\right)^{-1} \nabla_\theta l_i\left(\theta^*\right)\right\|^2\right] \leq & \frac{1}{N_t^2} \sum_{i=1}^t n_i^{2(1-\sigma)}\left(\operatorname{Tr}\left[\nabla_\theta^2 L(\theta^*)^{-1} \Sigma \nabla_\theta^2 L(\theta^*)^{-1}\right]+\frac{C_\sigma'}{\mu^2 n_i^{2 \sigma'}}\right) \\ = & \frac{\Lambda}{N_t^2} \sum_{i=1}^t n_i^{2(1-\sigma)}+\frac{C_\sigma'}{\mu^2 N_t^2} \sum_{i=1}^t n_i^{2(1-\sigma-\sigma')},\end{aligned}$$

where Λ denotes $\mathrm{Tr}[\nabla_\theta^2 L(\theta^*)^{-1}\Sigma\nabla_\theta^2 L(\theta^*)^{-1}]$. For the next term,

$$\begin{aligned}
&\frac{2}{N_t^2}\sum_{1\le i<j\le t} n_i n_j \mathbb{E}\left[\left\langle \nabla_\theta^2 L\left(\theta^*\right)^{-1}\nabla_\theta l_i\left(\theta^*\right), \nabla_\theta^2 L\left(\theta^*\right)^{-1}\nabla_\theta l_j\left(\theta^*\right)\right\rangle\right] \\
&\le \frac{2}{\mu^2 N_t^2}\sum_{1\le i<j\le t} n_i n_j \mathbb{E}\left[\left\langle \nabla_\theta l_i\left(\theta^*\right), \nabla_\theta l_j\left(\theta^*\right)\right\rangle\right] \\
&= \frac{2}{\mu^2 N_t^2}\sum_{1\le i<j\le t} n_i n_j \mathbb{E}\left[\left\langle \nabla_\theta l_i\left(\theta^*\right), \mathbb{E}[\nabla_\theta l_j(\theta^*)|\mathcal{F}_{j-1}] - \nabla_\theta L(\theta^*)\right\rangle\right] \\
&\le \frac{2}{\mu^2 N_t^2}\sum_{1\le i<j\le t} n_i n_j \mathbb{E}\left[\left\|\nabla_\theta l_i\left(\theta^*\right)\right\| \left\|\left[\mathbb{E}[\nabla_\theta l_j(\theta^*)|\mathcal{F}_{j-1}] - \nabla_\theta L(\theta^*)\right]\right\|\right] \\
&\le \frac{2}{\mu^2 N_t^2}\sum_{1\le i<j\le t} n_i n_j \sqrt{\mathbb{E}\left[\left\|\nabla_\theta l_i\left(\theta^*\right)\right\|^2\right]}\sqrt{\mathbb{E}\left[\left\|\left[\mathbb{E}[\nabla_\theta l_j(\theta^*)|\mathcal{F}_{j-1}] - \nabla_\theta L(\theta^*)\right]\right\|^2\right]} \\
&\le \frac{2B_\nu}{\mu^2 N_t^2}\sum_{1\le i<j\le t} n_i n_j \sigma_i \nu_j = \frac{2B_\nu}{\mu^2 N_t^2}\sum_{j=2}^{t}\left(n_j\nu_j\sum_{i=1}^{j-1} n_i\sigma_i\right),
\end{aligned}$$

by Cauchy-Schwarz inequality, Hölder's inequality, and Assumptions 3.2.1-p and 3.2.3-p. Thus,

$$\begin{aligned}
\left(\mathbb{E}\left[\left\|\nabla_\theta^2 L\left(\theta^*\right)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i\nabla_\theta l_i\left(\theta^*\right)\right\|^2\right]\right)^{1/2} \le& \frac{\Lambda^{1/2}}{N_t}\left(\sum_{i=1}^{t} n_i^{2(1-\sigma)}\right)^{1/2} + \frac{C_\sigma'^{1/2}}{\mu N_t^{1/2}}\left(\sum_{i=1}^{t} n_i^{2(1-\sigma-\sigma')}\right)^{1/2} \\
&+\frac{2^{1/2}B_\nu^{1/2}}{\mu N_t}\left(\sum_{j=2}^{t}\left(n_j\nu_j\sum_{i=1}^{j-1} n_i\sigma_i\right)\right)^{1/2}. \qquad (3.6.30)
\end{aligned}$$

Next, by the relation in (3.6.28), we have

$$\begin{aligned}
\frac{1}{N_t}\sum_{i=1}^{t} n_i\nabla_\theta l_i\left(\theta_{i-1}\right) =& \frac{1}{N_t}\sum_{i=1}^{t}\frac{n_i}{\gamma_i}\left(\theta_{i-1}-\theta_i\right) \\
=& \frac{1}{N_t}\sum_{i=1}^{t-1}\left(\theta_i-\theta^*\right)\left(\frac{n_{i+1}}{\gamma_{i+1}}-\frac{n_i}{\gamma_i}\right) - \frac{1}{N_t}\left(\theta_t-\theta^*\right)\frac{n_t}{\gamma_t} + \frac{1}{N_t}\left(\theta_0-\theta^*\right)\frac{n_1}{\gamma_1},
\end{aligned}$$

leading to

$$\begin{aligned}
\left\|\nabla_\theta^2 L\left(\theta^*\right)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i\nabla_\theta l_i\left(\theta_{i-1}\right)\right\| \le& \frac{1}{\mu N_t}\sum_{i=1}^{t-1}\left\|\theta_i-\theta^*\right\|\left|\frac{n_{i+1}}{\gamma_{i+1}}-\frac{n_i}{\gamma_i}\right| + \frac{1}{\mu N_t}\left\|\theta_t-\theta^*\right\|\frac{n_t}{\gamma_t} \\
&+\frac{1}{\mu N_t}\left\|\theta_0-\theta^*\right\|\frac{n_1}{\gamma_1}.
\end{aligned}$$

Hence, with the notation of $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$, the second term can be bounded by

$$\left(\mathbb{E}\left[\left\|\nabla_\theta^2 L\left(\theta^*\right)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i \nabla_\theta l_i\left(\theta_{i-1}\right)\right\|^2\right]\right)^{1/2} \leq \frac{1}{\mu N_t}\sum_{i=1}^{t-1}\delta_i^{\frac{1}{2}}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| + \frac{n_t}{\mu\gamma_t N_t}\delta_t^{\frac{1}{2}} + \frac{n_1}{\mu\gamma_1 N_t}\delta_0^{\frac{1}{2}}. \tag{3.6.31}$$

For the third term, we can derive it as

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla_\theta^2 L(\theta^*)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i\left[\nabla_\theta l_i(\theta_{i-1}) - \nabla_\theta l_i(\theta^*) - \nabla_\theta L(\theta_{i-1})\right]\right\|^2\right] \\
&= \frac{1}{\mu^2 N_t^2}\sum_{i=1}^{t} n_i^2 \mathbb{E}\left[\left\|\nabla_\theta l_i(\theta_{i-1}) - \nabla_\theta l_i(\theta^*) - \nabla_\theta L(\theta_{i-1})\right\|^2\right] \\
&+ \frac{2}{\mu^2 N_t^2}\sum_{i<j}^{t} n_i n_j \mathbb{E}\left[\langle \nabla_\theta l_i(\theta_{i-1}) - \nabla_\theta l_i(\theta^*) - \nabla_\theta L(\theta_{i-1}), \nabla_\theta l_j(\theta_{j-1}) - \nabla_\theta l_j(\theta^*) - \nabla_\theta L(\theta_{j-1})\rangle\right],
\end{aligned}$$

where

$$\begin{aligned}
\sum_{i=1}^{t} n_i^2 \mathbb{E}\left[\left\|\nabla_\theta l_i\left(\theta_{i-1}\right) - \nabla_\theta l_i\left(\theta^*\right) - \nabla_\theta L\left(\theta_{i-1}\right)\right\|^2\right] \leq & 2\sum_{i=1}^{t} n_i^2 \mathbb{E}\left[\left\|\nabla_\theta l_i\left(\theta_{i-1}\right) - \nabla_\theta l_i\left(\theta^*\right)\right\|^2\right] \\
& + 2\sum_{i=1}^{t} n_i^2 \mathbb{E}\left[\left\|\nabla_\theta L\left(\theta_{i-1}\right)\right\|^2\right] \\
\leq & 2\sum_{i=1}^{t} n_i^2 \kappa_i^2 \delta_{i-1} + 2C_\nabla^2 \sum_{i=1}^{t} n_i^2 \delta_{i-1},
\end{aligned}$$

by the Cauchy-Schwarz inequality, Assumption 3.2.2-p and (3.3.8). For the other term, we note that

$$\begin{aligned}
&\mathbb{E}[\langle \nabla_\theta l_i(\theta_{i-1}) - \nabla_\theta l_i(\theta^*) - \nabla_\theta L(\theta_{i-1}), \nabla_\theta l_j(\theta_{j-1}) - \nabla_\theta l_j(\theta^*) - \nabla_\theta L(\theta_{j-1})\rangle] \\
=&\mathbb{E}[\langle \nabla_\theta l_i(\theta_{i-1}) - \nabla_\theta l_i(\theta^*) - [\nabla_\theta L(\theta_{i-1}) - \nabla_\theta L(\theta^*)], \\
&\mathbb{E}[\nabla_\theta l_j(\theta_{j-1})|\mathcal{F}_{j-1}] - \nabla_\theta L(\theta_{j-1}) - [\mathbb{E}[\nabla_\theta l_j(\theta^*)|\mathcal{F}_{j-1}] - \nabla_\theta L(\theta^*)]\rangle] \\
\leq&\sqrt{\mathbb{E}[\|\nabla_\theta l_i(\theta_{i-1}) - \nabla_\theta l_i(\theta^*) - [\nabla_\theta L(\theta_{i-1}) - \nabla_\theta L(\theta^*)]\|^2]} \\
&\sqrt{\mathbb{E}[\|\mathbb{E}[\nabla_\theta l_j(\theta_{j-1})|\mathcal{F}_{j-1}] - \nabla_\theta L(\theta_{j-1}) - [\mathbb{E}[\nabla_\theta l_j(\theta^*)|\mathcal{F}_{j-1}] - \nabla_\theta L(\theta^*)]\|^2]} \\
\leq&\sqrt{2\mathbb{E}[\|\nabla_\theta l_i(\theta_{i-1}) - \nabla_\theta l_i(\theta^*)\|^2] + 2\mathbb{E}[\|\nabla_\theta L(\theta_{i-1}) - \nabla_\theta L(\theta^*)\|^2]} \\
&\sqrt{2\mathbb{E}[\|\mathbb{E}[\nabla_\theta l_j(\theta_{j-1})|\mathcal{F}_{j-1}] - \nabla_\theta L(\theta_{j-1})\|^2] + 2\mathbb{E}[\|\mathbb{E}[\nabla_\theta l_j(\theta^*)|\mathcal{F}_{j-1}] - \nabla_\theta L(\theta^*)\|^2]} \\
\leq&\sqrt{2\kappa_i^2\delta_{i-1} + 2C_\nabla^2\delta_{i-1}}\sqrt{2D_\nu^2\nu_j^2\delta_{j-1} + 4B_\nu^2\nu_j^2} \leq 2^{1/2}(\kappa_i\delta_{i-1}^{1/2} + C_\nabla\delta_{i-1}^{1/2})(D_\nu\nu_j\delta_{j-1}^{1/2} + 2^{1/2}B_\nu\nu_j),
\end{aligned}$$

using $\mathcal{F}_{i-1} \subset \mathcal{F}_{j-1}$ since $i < j$, Cauchy–Schwarz inequality, Hölder's inequality, $\|a+b\|^p \leq$

$2^{p-1}(\|a\|^p + \|b\|^p)$ with $p \in \mathbb{N}$, Assumptions 3.2.1-p and 3.2.2-p, and (3.3.8). Thus,

$$\begin{aligned}
&\left(\mathbb{E}\left[\left\|\nabla_\theta^2 L(\theta^*)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i\left[\nabla_\theta l_i(\theta_{i-1}) - \nabla_\theta l_i(\theta^*) - \nabla_\theta L(\theta_{i-1})\right]\right\|^2\right]\right)^{1/2} \\
&\leq \frac{2^{1/2}}{\mu N_t}\left(\sum_{i=1}^{t} n_i^2\kappa_i^2\delta_{i-1}\right)^{1/2} + \frac{2^{1/2}C_\nabla}{\mu N_t}\left(\sum_{i=1}^{t} n_i^2\delta_{i-1}\right)^{1/2} \\
&+ \frac{2^{3/4}}{\mu N_t}\left(\sum_{j=2}^{t}\left((D_\nu\delta_{j-1}^{1/2} + 2^{1/2}B_\nu)n_j\nu_j\sum_{i=1}^{j-1}(C_\nabla + \kappa_i)n_i\delta_{i-1}^{1/2}\right)\right)^{1/2}.
\end{aligned} \tag{3.6.32}$$

The last term is directly bounded by (3.3.9), using that (3.3.9) implies $\forall\theta$, $\|\nabla_\theta L(\theta) - \nabla_\theta^2 L(\theta^*)(\theta - \theta^*)\| \leq C'_\nabla\|\theta - \theta^*\|^2/2$ [105], giving us

$$\left(\mathbb{E}\left[\left\|\nabla_\theta^2 L\left(\theta^*\right)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i\left[\nabla_\theta L\left(\theta_{i-1}\right) - \nabla_\theta^2 L\left(\theta^*\right)\left(\theta_{i-1} - \theta^*\right)\right]\right\|^2\right]\right)^{\frac{1}{2}} \leq \frac{C'_\nabla}{2\mu N_t}\sum_{i=1}^{t} n_i\Delta_{i-1}^{1/2}, \tag{3.6.33}$$

with the notion $\Delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^4]$. Combining the terms (3.6.30) to (3.6.33) into (3.6.29), gives us

$$\begin{aligned}
\bar{\delta}_t^{1/2} \leq &\frac{\Lambda^{1/2}}{N_t}\left(\sum_{i=1}^{t} n_i^{2(1-\sigma)}\right)^{1/2} + \frac{C_\sigma'^{1/2}}{\mu N_t}\left(\sum_{i=1}^{t} n_i^{2(1-\sigma-\sigma')}\right)^{1/2} + \frac{2^{1/2}B_\nu^{1/2}}{\mu N_t}\left(\sum_{j=2}^{t}\left(n_j\nu_j\sum_{i=1}^{j-1} n_i\sigma_i\right)\right)^{1/2} \\
&+ \frac{1}{\mu N_t}\sum_{i=1}^{t-1}\delta_i^{1/2}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| + \frac{n_t}{\mu\gamma_t N_t}\delta_t^{1/2} + \frac{n_1}{\mu\gamma_1 N_t}\delta_0^{1/2} + \frac{2^{1/2}}{\mu N_t}\left(\sum_{i=1}^{t} n_i^2\kappa_i^2\delta_{i-1}\right)^{1/2} \\
&+ \frac{2^{1/2}C_\nabla}{\mu N_t}\left(\sum_{i=1}^{t} n_i^2\delta_{i-1}\right)^{1/2} + \frac{C'_\nabla}{2\mu N_t}\sum_{i=1}^{t} n_i\Delta_{i-1}^{1/2} \\
&+ \frac{2^{3/4}}{\mu N_t}\left(\sum_{j=2}^{t}\left((D_\nu\delta_{j-1}^{1/2} + 2^{1/2}B_\nu)n_j\nu_j\sum_{i=1}^{j-1}(C_\nabla + \kappa_i)n_i\delta_{i-1}^{1/2}\right)\right)^{1/2},
\end{aligned}$$

which gives the desired by shifting the indices and collecting the δ_0 terms,

$$\begin{aligned}\bar{\delta}_t^{1/2} \leq & \frac{\Lambda^{1/2}}{N_t}\left(\sum_{i=1}^{t} n_i^{2(1-\sigma)}\right)^{1/2} + \frac{C_\sigma'^{1/2}}{\mu N_t}\left(\sum_{i=1}^{t} n_i^{2(1-\sigma-\sigma')}\right)^{1/2} + \frac{2^{1/2}B_\nu^{1/2}}{\mu N_t}\left(\sum_{j=2}^{t}\left(n_j\nu_j\sum_{i=1}^{j-1} n_i\sigma_i\right)\right)^{1/2} \\ & + \frac{1}{\mu N_t}\sum_{i=1}^{t-1}\delta_i^{1/2}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| + \frac{n_t}{\mu\gamma_t N_t}\delta_t^{1/2} + \frac{n_1}{\mu N_t}\left(\frac{1}{\gamma_1} + 2^{1/2}(C_\nabla + \kappa_1)\right)\delta_0^{1/2} \\ & + \frac{2^{1/2}}{\mu N_t}\left(\sum_{i=1}^{t-1} n_{i+1}^2(C_\nabla^2 + \kappa_{i+1}^2)\delta_i\right)^{1/2} + \frac{C_\nabla'}{2\mu N_t}\sum_{i=0}^{t-1} n_{i+1}\Delta_i^{1/2} \\ & + \frac{2^{3/4}}{\mu N_t}\left(\sum_{j=1}^{t-1}\left((D_\nu\delta_j^{1/2} + 2^{1/2}B_\nu)n_{j+1}\nu_{j+1}\sum_{i=0}^{j-1}(C_\nabla + \kappa_{i+1})n_{i+1}\delta_i^{1/2}\right)\right)^{1/2}. \end{aligned} \tag{3.6.34}$$

Now, assume that (θ_t) is derived from the recursion in (3.2.2): as above, we follow the steps of Polyak and Juditsky [118], in which, we can rewrite (3.2.2) to

$$\frac{1}{\gamma_t}(\theta_{t-1} - \theta_t) = \nabla_\theta l_t(\theta_{t-1}) - \frac{1}{\gamma_t}\Omega_t,$$

where $\Omega_t = \mathcal{P}_\Theta(\theta_{t-1} - \gamma_t\nabla_\theta l_t(\theta_{t-1})) - (\theta_{t-1} - \gamma_t\nabla_\theta l_t(\theta_{t-1}))$. Thus, summing the parts, taking the norm and expectation, and using the Minkowski's inequality, yields the same terms as in (3.6.29), but with an additional term regarding Ω_t, namely

$$\begin{aligned}\left(\mathbb{E}\left[\left\|\nabla_\theta^2 L\left(\theta^*\right)^{-1}\frac{1}{N_t}\sum_{i=1}^{t}\frac{n_i}{\gamma_i}\Omega_i\right\|^2\right]\right)^{1/2} \leq & \frac{1}{\mu N_t}\sum_{i=1}^{t}\frac{n_i}{\gamma_i}\sqrt{\mathbb{E}\left[\|\Omega_i\|^2\right]} \\ = & \frac{1}{\mu N_t}\sum_{i=1}^{t}\frac{n_i}{\gamma_i}\sqrt{\mathbb{E}\left[\|\Omega_i\|^2\mathbb{1}_{\{\theta_{i-1}-\gamma_i\nabla_\theta l_i(\theta_{i-1})\notin\Theta\}}\right]}, \end{aligned} \tag{3.6.35}$$

using Godichon-Baggioni [55, Lemma 4.3]. Next, we note that $\mathbb{E}[\|\Omega_t\|^2\mathbb{1}_{\{\theta_{t-1}-\gamma_t\nabla_\theta l_t(\theta_{t-1})\notin\Theta\}}] = 4\gamma_t^2 G_\Theta^2\mathbb{P}[\theta_{t-1} - \gamma_t\nabla_\theta l_t(\theta_{t-1}) \notin \Theta]$, since

$$\begin{aligned}\|\Omega_t\|^2 = & \left\|\mathcal{P}_\Theta\left(\theta_{t-1} - \gamma_t\nabla_\theta l_t\left(\theta_{t-1}\right)\right) - \theta_{t-1} + \gamma_t\nabla_\theta l_t\left(\theta_{t-1}\right)\right\|^2 \\ \leq & 2\left\|\mathcal{P}_\Theta\left(\theta_{t-1} - \gamma_t\nabla_\theta l_t\left(\theta_{t-1}\right)\right) - \theta_{t-1}\right\|^2 + 2\gamma_t^2\left\|\nabla_\theta l_t\left(\theta_{t-1}\right)\right\|^2 \\ = & 2\left\|\mathcal{P}_\Theta\left(\theta_{t-1} - \gamma_t\nabla_\theta l_t\left(\theta_{t-1}\right)\right) - \mathcal{P}_\Theta\left(\theta_{t-1}\right)\right\|^2 + 2\gamma_t^2\left\|\nabla_\theta l_t\left(\theta_{t-1}\right)\right\|^2 \\ \leq & 2\left\|\theta_{t-1} - \gamma_t\nabla_\theta l_t\left(\theta_{t-1}\right) - \theta_{t-1}\right\|^2 + 2\gamma_t^2\left\|\nabla_\theta l_t\left(\theta_{t-1}\right)\right\|^2 = 4\gamma_t^2\left\|\nabla_\theta l_t\left(\theta_{t-1}\right)\right\|^2 \leq 4\gamma_t^2 G_\Theta^2, \end{aligned}$$

as $\mathcal{P}_\Theta$ is Lipschitz and $\|\nabla_\theta l_t(\theta)\|^2 \leq G_\Theta^2$ for any $\theta \in \Theta$. Moreover, as in Godichon-Baggioni and Portier [56, Theorem 4.2] with use of Lemma 3.6.2, we know that $\mathbb{P}[\theta_{t-1} - \gamma_t\nabla_\theta l_t(\theta_{t-1}) \notin \Theta] \leq \Delta_t/D_\Theta^4$, where $D_\Theta = \inf_{\theta\in\partial\Theta}\|\theta - \theta^*\|$ with $\partial\Theta$ denoting the frontier of Θ. Thus, (3.6.35) can then

be bounded by

$$\frac{1}{\mu N_t}\sum_{i=1}^{t}\frac{n_i}{\gamma_i}\sqrt{\mathbb{E}\left[\|\Omega_i\|^2\,\mathbb{1}_{\{\theta_{i-1}-\gamma_i\nabla_\theta l_i(\theta_{i-1})\notin\Theta\}}\right]}\leq\frac{2G_\Theta}{\mu D_\Theta^2 N_t}\sum_{i=1}^{t}n_i\Delta_i^{1/2}\leq\frac{2G_\Theta}{\mu D_\Theta^2 N_t}\sum_{i=1}^{t}n_{i+1}\Delta_i^{1/2},$$

since the sequence (n_t) is either constant or increasing, meaning $\forall t, n_t/n_{t+1}\leq 1$. At last, this term can be combined into (3.6.34) with use of $C''_\nabla = C'_\nabla/2 + \mathbb{1}_{\{D_\Theta<\infty\}}2G_\Theta/D_\Theta^2$, which indicates whether (θ_t) follows (3.2.2) or not. □

Proof of Theorem 3.3.2. The result follows by simplifying and bounding each term of Lemma 3.6.3, with use of Theorem 3.3.1 and Lemma 3.6.2. Thus, by inserting $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$, $\nu_t = n_t^{-\nu}$, $\kappa_t = C_\kappa n_t^{-\kappa}$, $\sigma_t = C_\sigma n_t^{-\sigma}$, and $n_t = C_\rho t^\rho$ into the bound of Lemma 3.6.3, we obtain

$$\begin{aligned}
\bar{\delta}_t^{1/2}\leq&\frac{\Lambda^{1/2}}{N_t^{1/2}}\mathbb{1}_{\{\sigma=1/2\}}+\frac{\Lambda^{1/2}C_\rho^{1-\sigma}}{N_t}\left(\sum_{i=1}^{t}i^{2\rho(1-\sigma)}\right)^{1/2}\mathbb{1}_{\{\sigma\neq1/2\}}+\frac{C_\sigma'^{1/2}C_\rho^{1-\sigma-\sigma'}}{\mu N_t}\left(\sum_{i=1}^{t}i^{2\rho(1-\sigma-\sigma')}\right)^{1/2}\\
&+\frac{2^{1/2}B_\nu^{1/2}C_\sigma^{1/2}C_\rho}{\mu C_\rho^{(\sigma+\nu)/2}N_t}\left(\sum_{j=2}^{t}\left(j^{\rho(1-\nu)}\sum_{i=1}^{j-1}i^{\rho(1-\sigma)}\right)\right)^{1/2}+\frac{(\rho(1-\beta)+\alpha)C_\rho}{\mu C_\gamma C_\rho^\beta N_t}\sum_{i=1}^{t-1}i^{\rho(1-\beta)+\alpha-1}\delta_i^{1/2}\\
&+\frac{C_\rho t^{\rho(1-\beta)+\alpha}}{\mu C_\gamma C_\rho^\beta N_t}\delta_t^{1/2}+\frac{C_\rho}{\mu N_t}\left(\frac{1}{C_\gamma C_\rho^\beta}+2^{1/2}\left(\frac{C_\kappa}{C_\rho^\kappa}+C_\nabla\right)\right)\delta_0^{1/2}+\frac{2^{1/2+\rho(1-\kappa)}C_\kappa C_\rho}{\mu C_\rho^\kappa N_t}\left(\sum_{i=1}^{t-1}i^{2\rho(1-\kappa)}\delta_i\right)^{1/2}\\
&+\frac{2^{1/2+\rho}C_\nabla C_\rho}{\mu N_t}\left(\sum_{i=1}^{t-1}i^{2\rho}\delta_i\right)^{1/2}+\frac{2^\rho C''_\nabla C_\rho}{\mu N_t}\sum_{i=0}^{t-1}i^\rho\Delta_i^{1/2}\\
&+\frac{2^{3/4+\rho(2-\nu)/2}C_\rho}{\mu C_\rho^{\nu/2}N_t}\left(\sum_{j=1}^{t-1}\left((D_\nu\delta_j^{1/2}+2^{1/2}B_\nu)j^{\rho(1-\nu)}\sum_{i=1}^{j-1}\left(C_\nabla+\frac{2^{\rho\kappa}C_\kappa}{C_\rho^\kappa i^{\rho\kappa}}\right)i^\rho\delta_i^{1/2}\right)\right)^{1/2},
\end{aligned}$$

using $n_{i+1}/n_i\leq 2^\rho$ and that $|n_{i+1}/\gamma_{i+1}-n_i/\gamma_i|\leq(\rho(1-\beta)+\alpha)C_\rho^{1-\beta}/C_\gamma i^{1-\rho(1-\beta)-\alpha}$ as $\rho(1-\beta)+\alpha\leq 1-\rho$ with $\rho\in[0,1)$. Next, as $\sigma\in[0,1/2]$ and $\sigma'\in(0,1/2]$, we have $\sum_{i=1}^t i^{2\rho(1-\sigma-\sigma')}\leq t^{1+2\rho(1-\sigma-\sigma')}/(1+2\rho(1-\sigma-\sigma'))$, where $t\leq(2N_t/C_\rho)^{1/(1+\rho)}$. Similarly, as $\nu\in(0,\infty)$, we have that

$$\begin{aligned}
\sum_{j=2}^{t-1}\left(j^{\rho(1-\nu)}\sum_{i=1}^{j-1}i^{\rho(1-\sigma)}\right)\leq&\sum_{j=1}^{t-1}j^{\rho(1-\nu)}\sum_{i=1}^{t-1}i^{\rho(1-\sigma)}\\
\leq&\psi_{\rho(\nu-1)}(t)\psi_{\rho(\sigma-1)}(t)\\
\leq&\psi^\rho_{\rho(\nu-1)}(2N_t/C_\rho)\psi^\rho_{\rho(\sigma-1)}(2N_t/C_\rho),
\end{aligned}$$

using the ψ-function defined in (3.6.16), such that

$$\sqrt{\psi^\rho_{\rho(\sigma-1)}(2N_t/C_\rho)\psi^\rho_{\rho(\nu-1)}(2N_t/C_\rho)}/N_t\leq\tilde{\mathcal{O}}(N_t^{-\rho(\sigma+\nu)/2(1+\rho)}).$$

Let D_{∇}^{κ} denote $C_{\nabla} + 2^{\rho\kappa} C_{\kappa}/C_{\rho}^{\kappa}$ with $\kappa \in [0, 1/2]$, such that

$$\frac{2^{1/2+\rho(1-\kappa)} C_{\kappa} C_{\rho}}{\mu C_{\rho}^{\kappa} N_t} \left(\sum_{i=1}^{t-1} i^{2\rho(1-\kappa)} \delta_i \right)^{1/2} + \frac{2^{1/2+\rho} C_{\nabla} C_{\rho}}{\mu N_t} \left(\sum_{i=1}^{t-1} i^{2\rho} \delta_i \right)^{1/2} \leq \frac{2^{1/2+\rho} D_{\nabla}^{\kappa} C_{\rho}}{\mu N_t} \left(\sum_{i=1}^{t-1} i^{2\rho} \delta_i \right)^{1/2},$$

and, likewise, we have that

$$\sum_{j=1}^{t-1} \left((D_{\nu} \delta_j^{1/2} + 2^{1/2} B_{\nu}) j^{\rho(1-\nu)} \sum_{i=1}^{j-1} \left(C_{\nabla} + \frac{2^{\rho\kappa} C_{\kappa}}{C_{\rho}^{\kappa} i^{\rho\kappa}} \right) i^{\rho} \delta_i^{1/2} \right)$$
$$\leq D_{\nabla}^{\kappa} \sum_{j=1}^{t-1} \left((D_{\nu} \delta_j^{1/2} + 2^{1/2} B_{\nu}) j^{\rho(1-\nu)} \sum_{i=1}^{j-1} i^{\rho} \delta_i^{1/2} \right).$$

From (3.6.21) we know that $\delta_t \leq D_{\delta}/t^{\delta}$ with

$$D_{\delta} = \sup_{t \in \mathbb{N}} \pi_t t^{\delta} + \frac{2^{1+2\rho\nu} B_{\nu}^2}{\mu \mu_{\nu} C_{\rho}^{2\nu}} + \frac{2^{2+\rho(2\sigma-\beta)+\alpha} C_{\sigma}^2 C_{\gamma} C_{\rho}^{\beta}}{\mu_{\nu} C_{\rho}^{2\sigma}},$$

and $\delta = \mathbb{1}_{\{B_{\nu}=0\}}(\rho(2\sigma-\beta)+\alpha) + \mathbb{1}_{\{B_{\nu}\neq 0\}} \min\{\rho(2\sigma-\beta)+\alpha, 2\rho\nu\}$, yielding

$$\begin{aligned}
&\sum_{j=1}^{t-1} \left((D_{\nu} \delta_j^{1/2} + 2^{1/2} B_{\nu}) j^{\rho(1-\nu)} \sum_{i=1}^{j-1} i^{\rho} \delta_i^{1/2} \right) \\
&\leq D_{\delta}^{1/2} \sum_{j=1}^{t-1} \left((D_{\nu} D_{\delta}^{1/2} j^{-\delta/2} + 2^{1/2} B_{\nu}) j^{\rho(1-\nu)} \sum_{i=1}^{j-1} i^{\rho-\delta/2} \right) \\
&\leq D_{\delta}^{1/2} \sum_{j=1}^{t-1} \left((D_{\nu} D_{\delta}^{1/2} j^{-\delta/2} + 2^{1/2} B_{\nu}) j^{\rho(1-\nu)} \psi_{\delta/2-\rho}(t) \right) \\
&\leq D_{\nu} D_{\delta} \psi_{\delta/2-\rho}(t) \psi_{\delta/2+\rho(\nu-1)}(t) + 2^{1/2} B_{\nu} D_{\delta}^{1/2} \psi_{\delta/2-\rho}(t) \psi_{\rho(\nu-1)}(t) \\
&\leq D_{\nu} D_{\delta} \psi_{\delta/2-\rho}^{\rho}(2N_t/C_{\rho}) \psi_{\delta/2+\rho(\nu-1)}^{\rho}(2N_t/C_{\rho}) + 2^{1/2} B_{\nu} D_{\delta}^{1/2} \psi_{\delta/2-\rho}^{\rho}(2N_t/C_{\rho}) \psi_{\rho(\nu-1)}^{\rho}(2N_t/C_{\rho}),
\end{aligned}$$

if $\delta/2 - \rho \geq 0$. Hence, $\sqrt{\psi_{\delta/2-\rho}^{\rho}(2N_t/C_{\rho}) \psi_{\delta/2+\rho(\nu-1)}^{\rho}(2N_t/C_{\rho})}/N_t = \tilde{\mathcal{O}}(N_t^{-(\delta+\rho\nu)/2(1+\rho)})$, and $\sqrt{\psi_{\delta/2-\rho}^{\rho}(2N_t/C_{\rho}) \psi_{\rho(\nu-1)}^{\rho}(2N_t/C_{\rho})}/N_t = \tilde{\mathcal{O}}(N_t^{-(\delta/2+\rho\nu)/2(1+\rho)})$. Next, we define $\bar{\pi}_t = \sum_{i=1}^{t} i^2 \pi_i \geq \sum_{i=1}^{t} \pi_i$ such that $\pi_t \leq t^{-1} \sum_{i=1}^{t} \pi_i \leq t^{-1} \bar{\pi}_t \leq t^{-1} \bar{\pi}_{\infty}$ since π_t is decreasing. Similarly, let $\bar{\Pi}_t = \sum_{i=1}^{t} i^{\rho} \Pi_i$. Both $\bar{\pi}_t$ and $\bar{\Pi}_t$ convergences to some finite constant depending on the model's

parameters. With use of these notions, one can show that

$$
\begin{aligned}
\bar{\delta}_t^{1/2} \leq\ & \frac{\Lambda^{1/2}}{N_t^{1/2}} \mathbb{1}_{\{\sigma=1/2\}} + \frac{2^{1+2\rho(1-\sigma)/2(1+\rho)} \Lambda^{1/2} C_\rho^{(1-2\sigma)/2(1+\rho)}}{\sqrt{1+2\rho(1-\sigma)} N_t^{(1+2\rho\sigma)/2(1+\rho)}} \mathbb{1}_{\{\sigma\neq 1/2\}} \\
& + \frac{2^{1+2\rho(1-\sigma-\sigma')/2(1+\rho)} C_\sigma'^{1/2} C_\rho^{(1-2\sigma-2\sigma')/2(1+\rho)}}{\sqrt{1+2\rho(1-\sigma-\sigma')} \mu N_t^{(1+2\rho(\sigma+\sigma'))/2(1+\rho)}} + \frac{2^{1/2} B_\nu^{1/2} C_\sigma^{1/2} C_\rho \sqrt{\psi^\rho_{\rho(\sigma-1)}(2N_t/C_\rho) \psi^\rho_{\rho(\nu-1)}(2N_t/C_\rho)}}{\mu C_\rho^{(\sigma+\nu)/2} N_t} \\
& + \frac{(\rho(1-\beta)+\alpha) C_\rho \bar{\pi}_\infty}{\mu C_\gamma C_\rho^\beta N_t} + \frac{(\rho(1-\beta)+\alpha) 2^{1/2+\rho\nu} B_\nu C_\rho \psi^\rho_{1+\rho(\beta+\nu-1)-\alpha}(2N_t/C_\rho)}{\mu^{3/2} \mu_\nu^{1/2} C_\gamma C_\rho^{\beta+\nu} N_t} \\
& + \frac{(\rho(1-\beta)+\alpha) 2^{(4+\rho(2+2\sigma-3\beta)+3\alpha)/2(1+\rho)} C_\sigma C_\rho^{(2-2\sigma-\beta-\alpha)/2(1+\rho)}}{(\rho(1-\sigma)+(\alpha-\rho\beta)/2) \mu \mu_\nu^{1/2} C_\gamma^{1/2} N_t^{(2+\rho(\beta+2\sigma)-\alpha)/2(1+\rho)}} \\
& + \frac{2^{(1+\rho(1-\beta)+\alpha)/(1+\rho)} C_\rho^{(2+\beta-\alpha)/(1+\rho)} \bar{\pi}_\infty}{\mu C_\gamma N_t^{(2+\rho\beta-\alpha)/(1+\rho)}} + \frac{2^{(1+\rho(1+3\nu-\beta)+\alpha))/(1+\rho)} B_\nu C_\rho^{(1-\beta-\nu-\alpha)/(1+\rho)}}{\mu^{3/2} \mu_\nu^{1/2} C_\gamma N_t^{(1+\rho(\beta+\nu)-\alpha)/(1+\rho)}} \\
& + \frac{2^{(2+\rho(1-2\beta+\sigma)+2\alpha)/(1+\rho)} C_\sigma C_\rho^{(2-2\sigma-\beta-\alpha)/2(1+\rho)}}{\mu \mu_\nu^{1/2} C_\gamma^{1/2} N_t^{(2+\rho(\beta+2\sigma)-\alpha)/2(1+\rho)}} + \frac{2^{1/2+\rho} D_\nabla^\kappa C_\rho \bar{\pi}_\infty^{1/2}}{\mu N_t} \\
& + \frac{2^{3/2+\rho(1+\nu)} B_\nu D_\nabla^\kappa C_\rho \sqrt{\psi^\rho_{2\rho(\nu-1)}(2N_t/C_\rho)}}{\mu^{3/2} \mu_\nu^{1/2} C_\rho^\nu N_t} + \frac{2^{(3+\rho(5-2\sigma+\beta)-\alpha)/2(1+\rho)} D_\nabla^\kappa C_\sigma C_\gamma^{1/2} C_\rho^{(1+\beta-2\sigma+\alpha)/2(1+\rho)}}{\mu \mu_\nu^{1/2} N_t^{(1+\rho(2\sigma-\beta)+\alpha)/(2(1+\rho))}} \\
& + \frac{2^{3/4+\rho(2-\nu)/2} \sqrt{D_\nabla^\kappa} D_\nu^{1/2} D_\delta^{1/2} C_\rho \sqrt{\psi^\rho_{\delta/2-\rho}(2N_t/C_\rho) \psi^\rho_{\delta/2+\rho(\nu-1)}(2N_t/C_\rho)}}{\mu C_\rho^{\nu/2} N_t} \\
& + \frac{2^{1+\rho(2-\nu)/2} B_\nu^{1/2} \sqrt{D_\nabla^\kappa} D_\delta^{1/4} C_\rho \sqrt{\psi^\rho_{\delta/2-\rho}(2N_t/C_\rho) \psi^\rho_{\rho(\nu-1)}(2N_t/C_\rho)}}{\mu C_\rho^{\nu/2} N_t} \\
& + \frac{C_\rho}{\mu N_t} \left(\frac{1}{C_\gamma C_\rho^\beta} + 2^{1/2} D_\nabla^\kappa \right) \delta_0^{1/2} + \frac{2^\rho C_\nabla'' C_\rho \bar{\Pi}_\infty}{\mu N_t} + \frac{2^{1+\rho(1+2\nu)} B_\nu^2 C_\nabla'' C_\rho \psi^\rho_{\rho(2\nu-1)}(2N_t/C_\rho)}{\mu^{5/2} \sqrt{\mu_\nu'} C_\rho^{2\nu} N_t} \\
& + \frac{2^{(1+\rho(1+2\sigma-\beta)+\alpha)/(1+\rho)} (2^5 \mu^{-1/2} + 2^4 C_\gamma^{1/2} C_\rho^{\beta/2}) C_\nabla'' C_\sigma^2 C_\gamma}{\mu \sqrt{\mu_\nu'} C_\rho^{(1-2\rho\sigma-\alpha)/(1+\rho)} N_t^{(\rho(2\sigma-\beta)+\alpha)/(1+\rho)}},
\end{aligned}
$$

where $\mu_\nu' = \mu - \mathbb{1}_{\{\rho=0\}} 2D_\nu^4/\mu^3 C_\rho^{4\nu}$, $D_\nabla^\kappa = C_\nabla + 2^\kappa C_\kappa / C_\rho^\kappa$ and $C_\nabla'' = C_\nabla' + \mathbb{1}_{\{D_\Theta < \infty\}} 2G_\Theta / D_\Theta^2$, which

can be simplified into

$$\begin{aligned}\bar{\delta}_t^{1/2} \leq & \frac{\Lambda^{1/2}}{N_t^{1/2}}\mathbb{1}_{\{\sigma=1/2\}} + \frac{2^{1/2}\Lambda^{1/2}C_\rho^{(1-2\sigma)/2(1+\rho)}}{N_t^{(1+2\rho\sigma)/2(1+\rho)}}\mathbb{1}_{\{\sigma\neq 1/2\}} + \frac{2^{1/2}C_\sigma'^{1/2}C_\rho^{(1-2(\sigma+\sigma'))/2(1+\rho)}}{\mu N_t^{(1+2\rho(\sigma+\sigma'))/2(1+\rho)}} \\ & + \frac{2^{2+(7+2\rho(1+\sigma))/2(1+\rho)}C_\sigma C_\rho^{(2-2\sigma-\beta-\alpha)/2(1+\rho)}}{\mu\mu_\nu^{1/2}C_\gamma^{1/2}N_t^{(2+\rho(\beta+2\sigma)-\alpha)/2(1+\rho)}} + \frac{2^{(1+\rho(1+2\sigma-\beta)+\alpha)/(1+\rho)}(2^5\mu^{-1/2}+2^4C_\gamma^{1/2}C_\rho^{\beta/2})C_\nabla''C_\sigma^2C_\gamma}{\mu\sqrt{\mu_\nu'}C_\rho^{(1-2\rho\sigma-\alpha)/(1+\rho)}N_t^{(\rho(2\sigma-\beta)+\alpha)/(1+\rho)}} \\ & + \frac{2^{(5/2+\rho(5-2\sigma))/2(1+\rho)}D_\nabla^\kappa C_\sigma C_\gamma^{1/2}C_\rho^{(1+\beta-2\sigma+\alpha)/2(1+\rho)}}{\mu\mu_\nu^{1/2}N_t^{(1+\rho(2\sigma-\beta)+\alpha)/(2(1+\rho))}} + \frac{\Gamma C_\rho}{\mu N_t} + \frac{2^{(2+\rho)/(1+\rho)}C_\rho^{(2+\beta-\alpha)/(1+\rho)}\bar{\pi}_\infty}{\mu C_\gamma N_t^{(2+\rho\beta-\alpha)/(1+\rho)}} \\ & + \frac{2^{3/4+\rho(2-\nu)/2}\sqrt{D_\nabla^\kappa}D_\nu^{1/2}D_\delta^{1/2}C_\rho\sqrt{\psi^\rho_{\delta/2-\rho}(2N_t/C_\rho)\psi^\rho_{\delta/2+\rho(\nu-1)}(2N_t/C_\rho)}}{\mu C_\rho^{\nu/2}N_t} + \mathbb{1}_{\{B_\nu\neq 0\}}\Psi_t,\end{aligned}$$

as $\alpha-\rho\beta \in (1/2,1)$ with use of $\Gamma = 2\bar{\pi}_\infty/C_\gamma C_\rho^\beta + (1/C_\gamma C_\rho^\beta + 2^{1/2}D_\nabla^\kappa)\delta_0^{1/2} + 2^{1/2+\rho}D_\nabla^\kappa\bar{\pi}_\infty^{1/2} + 2^\rho C_\nabla''\bar{\Pi}_\infty$, $D_\nabla^\kappa = C_\nabla + 2^\kappa C_\kappa/C_\rho^\kappa$, $\delta = \mathbb{1}_{\{B_\nu=0\}}(\rho(2\sigma-\beta)+\alpha) + \mathbb{1}_{\{B_\nu\neq 0\}}\min\{\rho(2\sigma-\beta)+\alpha, 2\rho\nu\}$, and Ψ_t given as

$$\begin{aligned}& \frac{2^{1/2}B_\nu^{1/2}C_\sigma^{1/2}C_\rho\sqrt{\psi^\rho_{\rho(\sigma-1)}(2N_t/C_\rho)\psi^\rho_{\rho(\nu-1)}(2N_t/C_\rho)}}{\mu C_\rho^{(\sigma+\nu)/2}N_t} + \frac{2^{3/2+\rho\nu}B_\nu C_\rho\psi^\rho_{1+\rho(\beta+\nu-1)-\alpha}(2N_t/C_\rho)}{\mu^{3/2}\mu_\nu^{1/2}C_\gamma C_\rho^{\beta+\nu}N_t} \\ & + \frac{2^{3/2+\rho(1+\nu)}B_\nu D_\nabla^\kappa C_\rho\sqrt{\psi^\rho_{2\rho(\nu-1)}(2N_t/C_\rho)}}{\mu^{3/2}\mu_\nu^{1/2}C_\rho^\nu N_t} + \frac{2^{3(1+\rho\nu)}B_\nu C_\rho^{(1-\beta-\nu-\alpha)/(1+\rho)}}{\mu^{3/2}\mu_\nu^{1/2}C_\gamma N_t^{(1+\rho(\beta+\nu)-\alpha)/(1+\rho)}} \\ & + \frac{2^{1+\rho(2-\nu)/2}B_\nu^{1/2}\sqrt{D_\nabla^\kappa}D_\delta^{1/4}C_\rho\sqrt{\psi^\rho_{\delta/2-\rho}(2N_t/C_\rho)\psi'^\rho_{\rho(\nu-1)}(2N_t/C_\rho)}}{\mu C_\rho^{\nu/2}N_t} + \frac{2^{2(1+\rho\nu)}B_\nu^2C_\nabla''C_\rho\psi^\rho_{\rho(2\nu-1)}(2N_t/C_\rho)}{\mu^{5/2}\sqrt{\mu_\nu'}C_\rho^{2\nu}N_t} \\ & = \tilde{\mathcal{O}}(N_t^{-\rho(\sigma+\nu)/2(1+\rho)}) + \tilde{\mathcal{O}}(N_t^{-(1+\rho(\beta+\nu)-\alpha)/(1+\rho)}) + \tilde{\mathcal{O}}(N_t^{-(1+2\rho\nu)/2(1+\rho)}) \\ & + \mathcal{O}(N_t^{-(1+\rho(\beta+\nu)-\alpha)/(1+\rho)}) + \tilde{\mathcal{O}}(N_t^{-(\delta/2+\rho\nu)/2(1+\rho)}) + \tilde{\mathcal{O}}(N_t^{-2\rho\nu/(1+\rho)}),\end{aligned} \tag{3.6.36}$$

Furthermore, with $\tilde{\mathcal{O}}$-notation one can yield,

$$\begin{aligned}\bar{\delta}_t^{1/2} \leq & \frac{\Lambda^{1/2}}{N_t^{1/2}}\mathbb{1}_{\{\sigma=1/2\}} + \frac{2^{1/2}\Lambda^{1/2}C_\rho^{(1-2\sigma)/2(1+\rho)}}{N_t^{(1+2\rho\sigma)/2(1+\rho)}}\mathbb{1}_{\{\sigma\neq 1/2\}} + \frac{2^{1/2}C_\sigma'^{1/2}C_\rho^{(1-2(\sigma+\sigma'))/2(1+\rho)}}{\mu N_t^{(1+2\rho(\sigma+\sigma'))/2(1+\rho)}} \\ & + \frac{2^6C_\sigma C_\rho^{(2-2\sigma-\beta-\alpha)/2(1+\rho)}}{\mu\mu_\nu^{1/2}C_\gamma^{1/2}N_t^{(2+\rho(\beta+2\sigma)-\alpha)/2(1+\rho)}} + \mathbb{1}_{\{B_\nu\neq 0\}}\Psi_t + \frac{2^7(\mu^{-1/2}+C_\gamma^{1/2}C_\rho^{\beta/2})C_\nabla''C_\sigma^2C_\gamma}{\mu\sqrt{\mu_\nu'}C_\rho^{(1-2\rho\sigma-\alpha)/(1+\rho)}N_t^{(\rho(2\sigma-\beta)+\alpha)/(1+\rho)}} \\ & + \frac{2^2D_\nabla^\kappa C_\sigma C_\gamma^{1/2}C_\rho^{(1+\beta-2\sigma+\alpha)/2(1+\rho)}}{\mu\mu_\nu^{1/2}N_t^{(1+\rho(2\sigma-\beta)+\alpha)/(2(1+\rho))}} + \frac{\Gamma C_\rho}{\mu N_t} + \frac{2^2C_\rho^{(2+\beta-\alpha)/(1+\rho)}\bar{\pi}_\infty}{\mu C_\gamma N_t^{(2+\rho\beta-\alpha)/(1+\rho)}} + \tilde{\mathcal{O}}(N_t^{-(\delta+\rho\nu)/2(1+\rho)}),\end{aligned} \tag{3.6.37}$$

where $\Psi_t = \tilde{\mathcal{O}}(N_t^{-\rho(\sigma+\nu)/2(1+\rho)}) + \tilde{\mathcal{O}}(N_t^{-(1+\rho(\beta+\nu)-\alpha)/(1+\rho)}) + \tilde{\mathcal{O}}(N_t^{-(1+2\rho\nu)/2(1+\rho)}) + \tilde{\mathcal{O}}(N_t^{-(\delta/2+\rho\nu)/2(1+\rho)}) + \tilde{\mathcal{O}}(N_t^{-2\rho\nu/(1+\rho)})$, implying that $\nu > 1/2$ to obtain the desired rate $\bar{\delta}_t = \mathcal{O}(N^{-1})$ if $B_\nu = 0$. □

Chapter 4: AdaVol: An Adaptive Recursive Volatility Prediction Method

Abstract

Quasi-Maximum Likelihood (QML) procedures are theoretically appealing and widely used for statistical inference. While there are extensive references on QML estimation in batch settings, it has attracted little attention in streaming settings until recently. An investigation of the convergence properties of the QML procedure in a general conditionally heteroscedastic time series model is conducted, and the classical batch optimization routines extended to the framework of streaming and large-scale problems. An adaptive recursive estimation routine for GARCH models named AdaVol is presented. The AdaVol procedure relies on stochastic approximations combined with the technique of Variance Targeting Estimation (VTE). This recursive method has computationally efficient properties, while VTE alleviates some convergence difficulties encountered by the usual QML estimation due to a lack of convexity. Empirical results demonstrate a favorable trade-off between AdaVol's stability and the ability to adapt to time-varying estimates for real-life data.

keywords : *volatility models, quasi-likelihood, recursive algorithm, GARCH, prediction method, stock index*

Contents

4.1 Introduction

Time series analysis has attracted much attention in the last three decades. A central aspect of time series analysis is modeling heteroscedasticity of the conditional variance, e.g., volatility clustering in financial time series. Some well-known models incorporating this feature are the AutoRegressive Conditional Heteroscedasticity (ARCH) model and the Generalized ARCH (GARCH) model introduced by [45] and [18], respectively. Many reasons can explain these models' success; they constitute a stationary time series model with a time-varying conditional variance, and secondly, they may model time series with heavier tails than the Gaussian ones, which often occurs in financial time series.

Quasi-Maximum Likelihood (QML) estimation is widely used for statistical inference in GARCH models due to their appealing theoretical nature and tolerance to overdispersion, which is often observed in empirical data. This paper studies the Quasi-Maximum Likelihood Estimator (QMLE) for the broader class of conditionally heteroscedastic time series models of multiplicative form given by

$$X_t = h_t(\theta_0)Z_t, \quad t \in \mathbb{Z}, \tag{4.1.1}$$

where θ_0 is the true underlying parameter vector, (Z_t) is a sequence of i.i.d. random variables with $\mathbb{E}[Z_0] = 0$ and $\mathbb{E}[Z_0^2] = 1$, and the (non-negative) volatility process $(h_t)_{t\in\mathbb{Z}}$ is defined as

$$h_t(\theta) = g_\theta\big(X_{t-1}, \ldots, X_{t-p}, h_{t-1}(\theta), \ldots, h_{t-q}(\theta)\big), \quad p, q \geq 0. \tag{4.1.2}$$

Suppose that the parameter set $\Theta \subset \mathbb{R}^d$ and $\{g_\theta | \theta \in \Theta\}$ denotes the (finite) parametric family of non-negative functions on $\mathbb{R}^p \times [0, \infty)^q$ satisfying certain regularity conditions. We also require that h_t is $\mathcal{F}_{t-1}$-measurable for all $t \in \mathbb{Z}$, where $\mathcal{F}_t = \sigma(Z_k : k \leq t)$ denotes the σ-field generated by the

random variables $\{Z_k : k \leq t\}$.

The stability of model (4.1.1)-(4.1.2) is accomplished under the assumption that g_θ is a contraction. This condition is a random Lipschitz coefficient condition, where the Lipschitz coefficient has a negative logarithmic moment. The notion of contractivity is clarified in [138] where they study QML inference of general conditionally heteroscedastic models with emphasis on the approximation $(\widehat{h}_t)$ of the stochastic volatility (h_t).

QML estimation of the parameters in the class of conditionally heteroscedastic time series models has been studied frequently in recent years, see e.g., [14], [48], [138], and [151]. However, all these references consider iterative estimation, where one assembles a batch of data and afterward performs the statistical inference. Thus, one evaluates an objective function consisting of a sum of n loss terms. Each iteration would then have a cost of $\mathcal{O}(nd)$, making the recursion cost $\mathcal{O}(mnd)$, where m is the number of iterations. As the amount of data grows, these optimizers become prohibitively expensive and increasingly computationally inefficient. Moreover, iterative optimizers become unsuitable for streaming settings where we are modeling and predicting data as they arrive.

Many financial practices, such as banks, asset managers, and financial services institutes, find themselves estimating thousands of volatility models every day for risk and pricing purposes. In addition, the sampling of financial time series is increasingly at high frequency. Therefore, recursive procedures must undoubtedly be advantageous since one only processes observations once. In recursive QML estimation, we update the previous QML estimate with the new observations at time t in order to produce the QML estimate of the parameters at time t.

Thus, in modern statistical analysis, it is becoming increasingly common to work with streaming data where one observes only a group of observations at a time. Naturally, this has led to an expanded interest in time-scalable recursive estimation procedures with a cost of only $\mathcal{O}(d)$ computations per recursion, e.g., see [19]. However, there has only been given a little amount of attention to recursive estimation in conditionally heteroscedastic time series models.

[36] presented a recursive method for estimating the parameters of an ARCH process. Under sufficient assumptions on the underlying process, [6] showed consistency of their recursive least squares method for GARCH processes, and [82] also developed a recursive estimation method for GARCH processes supported by empirical evidence. Convergence analysis of the recursive QML estimator for GARCH processes based on stochastic approximations with Markovian dynamics using a resetting mechanism has been previously presented ([52]). A self-weighted recursive estimation algorithm for GARCH models was proposed in [35] with a robustification in [73]. However, none of the above references mention problems with convexity or address the obstacles that may occur when the true parameter θ_0 is close to the boundary of the parameter space.

The difficulty of estimating time-varying parameters of statistical models increases in the setting of streaming data. To sustain computational efficiency and be adaptive to changes in the estimates, one may decrease the number of observations in each iteration in the optimization procedure, which may decrease the stability of the statistical inference. We propose a natural adaptation of the QML method, relying on stochastic approximations combined with the Variance Targeting Estimation

(VTE) technique, which we call AdaVol. This recursive method is time-scalable and memory-efficient, as it only requires the previous estimate to process new observations, and it only needs to treat the observations once. We present empirical evidence that AdaVol achieves a favorable trade-off between adaptability and stability.

The rest of the paper is organized as follows: Section 4.2 introduces the QML procedure for the general class of conditionally heteroscedastic time series models of multiplicative form and investigates the asymptotic properties of the Quasi-Likelihood (QL) function (Section 4.2.1). Next, in Section 4.2.2, we present the QML estimation of the GARCH parameters. In Section 4.3, we present our adaptive approach for recursively estimating GARCH parameters named AdaVol. We examine the AdaVol procedure on simulated and real-life observations in Section 4.4, and some concluding remarks are made in Section 4.5.

4.2 QML Estimation in Conditionally Heteroscedastic Time Series Models

The approximate QMLE $\widehat{\theta}_n^*$ is defined as

$$\widehat{\theta}_n^* \in \underset{\theta \in \mathcal{K}}{\arg\min}\, \widehat{L}_n(\theta), \tag{4.2.3}$$

where the parameter set $\mathcal{K}$ is a suitable compact subset of the parameter space Θ. The QL function $L_n(\theta)$ and approximate QL function $\widehat{L}_n(\theta)$ are given by

$$L_n(\theta) = \sum_{t=1}^{n} l_t(\theta) \text{ and } \widehat{L}_n(\theta) = \sum_{t=1}^{n} \widehat{l}_t(\theta), \tag{4.2.4}$$

with QL losses, denoted $l_t(\theta)$ and $\widehat{l}_t(\theta)$, given as

$$l_t(\theta) = \frac{1}{2}\left(\frac{X_t^2}{h_t(\theta)} + \log h_t(\theta)\right) \text{ and } \widehat{l}_t(\theta) = \frac{1}{2}\left(\frac{X_t^2}{\widehat{h}_t(\theta)} + \log \widehat{h}_t(\theta)\right), \tag{4.2.5}$$

where $(\widehat{h}_t)$ is an approximation of (h_t) defined recursively for $t \geq 1$ as in (4.1.2) with initialization $\widehat{h}_{-q+1} = \cdots = \widehat{h}_0 = 0$ or any deterministic constant. From [137, Proposition 5.2.12], we know the initialization error between $(\widehat{h}_t)$ and the true (h_t) will vanish exponentially fast almost surely. Assuming Z_0 is standard normal distributed, we may note X_t is also Gaussian with variance h_t conditioned on $\mathcal{F}_{t-1}$. The QL function $L_n(\cdot)$ in (4.2.4) is derived under this Gaussian assumption.

The consistency and asymptotic properties of the QMLE $\widehat{\theta}_n^*$ combined with the robustness of the QL function for overdispersion make the method highly used in practice (e.g., see [113]). Under the assumptions in [138, N.1, N.2, N.3 and N.4], the QMLE $\widehat{\theta}_n^*$ is strongly consistent and asymptotically normal, that is

$$\widehat{\theta}_n^* \overset{\text{a.s.}}{\to} \theta_0 \text{ and } \sqrt{n}\left(\widehat{\theta}_n^* - \theta_0\right) \to \mathcal{N}(0, V_0) \text{ as } n \to \infty, \tag{4.2.6}$$

with θ_0 as the true parameter vector and V_0 the asymptotic covariance matrix.

Unfortunately, these asymptotic properties in (4.2.6) come with a drawback on the QL loss; the consistency is achieved through careful domination of logarithmic moments. The concavity of logarithms makes the criterion insensitive to extreme values, but it also implies that the criterion itself behaves as a concave function. As most optimization algorithms are based on convex assumptions, this is striking.

In the next section, we show that the approximate Hessian $\widehat{H}_n(\theta) = n^{-1}\nabla^2_\theta \widehat{L}_n(\theta)$ admits strictly positive eigenvalues for n sufficiently large dependent on the model specifications and the underlying data process. This means that for sufficiently large batch sizes of observations, the QMLE $\widehat{\theta}^*_n$ can be seen as the unique solution of a locally strongly convex optimization problem; the existence and uniqueness of $\widehat{\theta}^*_n$ ensure that usual iterative optimization routines can efficiently approximate it for n large enough.

4.2.1 Asymptotic Properties of the QL Function

To establish the asymptotic local convexity of the QL function of the model described by (4.1.1)-(4.1.2), we need the following assumptions: Assumption 4.2.1, 4.2.2, and 4.2.3, which naturally emerges from the arguments and properties [138] made to ensure stability of the QL function and QMLE procedure. We will use two different matrix norms, namely, let $\|A\|_{op}$ denote the matrix operator norm of the matrix $A \in \mathbb{R}^{d\times d}$ with respect to the Euclidean norm, i.e., $\|A\|_{op} = \sup_{v\neq 0} |Av|/|v|$, and denote $\|A\|_{\mathcal{K}}$ the norm of the continuous matrix-valued function A on $\mathcal{K}$, i.e., $\|A\|_{\mathcal{K}} = \sup_{x\in\mathcal{K}}\|A(x)\|_{op}$, where $\mathcal{K}$ is a compact set of $\mathbb{R}^d$.

Assumption 4.2.1. *The model (4.1.1)-(4.1.2) with $\theta = \theta_0$ admits a unique stationary ergodic solution.*

Assumption 4.2.2. *Let $\mathcal{K} \subset \Theta$ be a compact set with true parameter vector $\theta_0 \in \mathcal{K}$ in the interior. The random functions fulfill certain conditions, such that $\mathbb{E}[\|l_0\|_{\mathcal{K}}] < \infty$, $\mathbb{E}[\|\nabla^2_\theta l_0\|_{\mathcal{K}}] < \infty$, and furthermore have the following uniform convergences $\|n^{-1}\widehat{L}_n - L_n\|_{\mathcal{K}} \xrightarrow{a.s.} 0$ and $n^{-1}\|\nabla^2_\theta \widehat{L}_n - \nabla^2_\theta L_n\|_{\mathcal{K}} \xrightarrow{a.s.} 0$ as $n \to \infty$.*

Assumption 4.2.3. *The components of the vector $\nabla_\theta g_\theta(X_0, h_0)$ from (4.1.2) with $\theta = \theta_0$ are linearly independent random variables.*

The following Theorem 4.2.1 is an extension of [75], which established similar results for the likelihood function of GARCH models under the assumption that (X_t) is strictly stationary and strongly mixing with geometric rate, and (Z_t) is Gaussian. Solving the QML estimation problem in (4.2.3) for $\widehat{\theta}^*_n$ is known to be computationally heavy as one has to find the solution of a non-linear equation, namely (4.2.4). Nonetheless, Theorem 4.2.1 ensures the existence of an N such that we have a unique global QMLE $\widehat{\theta}^*_n$ for all $n \geq N$.

Theorem 4.2.1. *Under Assumption 4.2.1, 4.2.2, and 4.2.3, there exist positive constants $C, \delta > 0$, and a random positive integer $N \in \mathbb{N}$ such that*

$$g^T \widehat{H}_n(\theta) g > C g^T g, \ \forall n \geq N, \quad a.s., \tag{4.2.7}$$

for all $\theta \in B(\theta_0, \delta)$ and $g \in \mathbb{R}^d \setminus \{0\}$.

The result above shows local strong convexity of the QL function $\widehat{L}_n$. The following corollary arises from the proof of Theorem 4.2.1:

Corollary 4.2.1. *Under Assumption 4.2.1, 4.2.2, and 4.2.3, the QMLE $\widehat{\theta}_n^*$ exists and is unique, that is*

$$\widehat{\theta}_n^* = \arg\min_{\theta \in \mathcal{K}} \widehat{L}_n(\theta).$$

Local strong convexity is crucial for guaranteeing the convergence of an optimization algorithm, although some methods go beyond this point ([147]). Thus, Theorem 4.2.1 is an essential result for computing the QMLE $\widehat{\theta}_n^*$ parameters of the model in (4.1.1)-(4.1.2). Nevertheless, to guarantee the property in (4.2.7), we need a sufficiently large (and maybe unbounded) random N, which depends on the true parameter vector θ_0, the parameter estimates $(\widehat{\theta}_t^*)$, and the observations (X_t). One often has a fixed size of observations in practice, so the iterative algorithm may not converge. To our experience, this phenomenon may occur when the true parameter vector θ_0 is close to the boundary of $\mathcal{K}$, or if the initial values $\widehat{\theta}_0^*$ are far away from the true parameters θ_0.

4.2.2 QML Estimation of GARCH(p,q) Parameters

The general class of conditionally heteroscedastic time series models includes the very popular ARCH and GARCH models. For more than three decades, these models have attracted considerable amounts of attention in the literature since their introduction. A process (X_t) is called a GARCH(p, q) process with parameter vector $\theta = (\omega, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)^T$, if it satisfies

$$\begin{cases} X_t = \sigma_t Z_t, \\ \sigma_t^2 = \omega + \sum_{i=1}^p \alpha_i X_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2, \end{cases} \tag{4.2.8}$$

where ω, α_i, and β_j for $1 \leq i \leq p$ and $1 \leq j \leq q$ are non-negative parameters ensuring the non-negativity of the conditional variance process (σ_t^2). The innovations (Z_t) is a sequence of i.i.d. random variables with $\mathbb{E}[Z_0] = 0$ and $\mathbb{E}[Z_0^2] = 1$. Likewise, one can define an ARCH(p) process by setting $\beta_j = 0$ for $1 \leq j \leq q$ in (4.2.8). The GARCH(p, q) process (X_t) given in (4.2.8) has QL losses given by $\widehat{l}_t(\theta) = 2^{-1}(X_t^2/\widehat{\sigma}_t^2(\theta) + \log \widehat{\sigma}_t^2(\theta))$ with first-order derivative

$$\nabla_\theta \widehat{l}_t(\theta) = \nabla_\theta \widehat{\sigma}_t^2(\theta) \left(\frac{\widehat{\sigma}_t^2(\theta) - X_t^2}{2\widehat{\sigma}_t^4(\theta)} \right) \tag{4.2.9}$$

and second-order derivative

$$\nabla_\theta^2 \widehat{l}_t(\theta) = \nabla_\theta \widehat{\sigma}_t^2(\theta)^T \nabla_\theta \widehat{\sigma}_t^2(\theta) \left(\frac{2X_t^2 - \widehat{\sigma}_t^2(\theta)}{2\widehat{\sigma}_t^6(\theta)} \right) + \nabla_\theta^2 \widehat{\sigma}_t^2(\theta) \left(\frac{\widehat{\sigma}_t^2(\theta) - X_t^2}{2\widehat{\sigma}_t^4(\theta)} \right), \tag{4.2.10}$$

where $\nabla_\theta \widehat{\sigma}_t^2(\theta) = \vartheta_t(\theta) + \sum_{j=1}^q \beta_j \nabla_\theta \widehat{\sigma}_{t-j}^2(\theta)$ with $\vartheta_t(\theta) = (1, X_{t-1}^2, \ldots, X_{t-p}^2, \widehat{\sigma}_{t-1}^2(\theta), \ldots, \widehat{\sigma}_{t-q}^2(\theta))^T \in \mathbb{R}^{p+q+1}$ and Hessian $\widehat{H}_n(\theta) = n^{-1} \sum_{t=1}^n \nabla_\theta^2 \widehat{l}_t(\theta)$.

The equations (4.2.8) creates a complicated probabilistic structure that is not easily understood, although it looks relatively simple. The conditions ensuring the existence and uniqueness of a stationary solution to the equations (4.2.8) for GARCH$(1,1)$ was provided by [100]. [23] later showed it for the GARCH(p,q) model using that GARCH(p,q) can be embedded in a Iterated Random Lipschitz Map (IRLM). See [22] for a formal definition of IRLMs.

We can illustrate the IRLM method on the GARCH$(1,1)$ model with parameter vector $\theta = (\omega, \alpha_1, \beta_1)^T$. The IRLM for σ_t^2 is then given by $\sigma_t^2 = A_t \sigma_{t-1}^2 + B_t$ with $t \in \mathbb{Z}$, where $A_t = \alpha_1 Z_{t-1}^2 + \beta_1$ and $B_t = \omega$. Note $((A_t, B_t))$ constitutes an i.i.d. sequence. From the literature on IRLMs it is well known that the conditions $\mathbb{E}[\log |A_0|] < 0$ and $\mathbb{E}[\log^+ |B_0|] < \infty$ guarantee the existence and uniqueness of a strictly stationary solution of the IRLM $Y_t = A_t Y_{t-1} + B_t$ for $t \in \mathbb{Z}$ provided $((A_t, B_t))$ is a stationary ergodic sequence. Applying this to the GARCH$(1,1)$ model, we get the known sufficient condition for the existence of a stationary solution, namely $\mathbb{E}[\log(\alpha_1 Z_0^2 + \beta_1)] < 0$. This also implies $\beta_1 < 1$ since $\log(\beta_1) \leq \mathbb{E}[\log(\alpha_1 Z_0^2 + \beta_1)] < 0$. Likewise, the ARCH$(1)$ process $(\beta_1 = 0)$ then requires $\mathbb{E}[\log(\alpha_1 Z_0^2)] < 0$, which is the same as $\alpha < 2e^\epsilon \approx 3.56$ with Z_0 being Gaussian. Thus, the stationary condition is much weaker than the second-order stationary condition in which we require $\alpha_1 + \beta_1 < 1$.

The statistical inference leads to further nontrivial problems since the exact distribution of (Z_t) remains unspecified, and so one usually determines the likelihoods under the hypothesis of standard Gaussian innovations. Moreover, the volatility (σ_t) is an unobserved quantity approximated by mimicking the recursion (4.2.8) with an initialization, for instance $X_{-p+1} = \cdots = X_0 = 0$ and $\sigma_{-q+1}^2 = \cdots = \sigma_0^2 = 0$. [14] showed under minimal assumptions that the QMLE is strongly consistent and asymptotically normal.

Furthermore, under Assumption 4.2.1-4.2.3, we have asymptotic local strong convexity of the QL function in GARCH(p,q) models by Theorem 4.2.1. However, the number of observations needed to guarantee local strong convexity vary. This can easily be seen by looking at the simplest case, namely when (X_t) follows an ARCH(1) process with parameter vector $\theta = (\omega, \alpha_1)^T$. The volatility process $\sigma_t^2(\theta)$ is given as $\omega + \alpha_1 X_{t-1}^2$. The eigenvalues of $\nabla_\theta^2 l_t(\theta)$ are given by $\lambda_t = (\lambda_{t,1}, \lambda_{t,2}) = (0, \lambda_{t,2})$ with $\lambda_{t,2} = (1 + X_{t-1}^4)(2X_t^2 - \sigma_t^2(\theta))2^{-1}\sigma_t^{-6}(\theta)$. Thus, the non-negativity of $\lambda_{t,2}$ would ensure convexity at time t in our QML procedure. However, the probability of having convexity at each t is unlikely as $\mathbb{P}(\cap_{t=1}^n \nabla_\theta^2 l_t(\theta) \geq 0) = \mathbb{P}(\cap_{t=1}^n Z_t^2 \geq 1/2) = \mathbb{P}(Z_0^2 \geq 1/2)^n$ is approximately 0.52^n with i.i.d. Gaussian innovations (Z_t), i.e., (Z_t^2) is χ^2-distributed with 1 degree of freedom. On the other hand, increasing the number of observations used at each iteration would increase the probability of having local strong convexity.

4.3 Adaptive Recursive QML Estimation

Our recursive QML method relies on stochastic approximations introduced by [124], which only requires the previous parameter estimate to update the parameter estimate using the new observation. We perform the first-order stochastic gradient method defined as

$$\widehat{\theta}_t = \widehat{\theta}_{t-1} - \eta_{t-1}\nabla_\theta \widehat{l}_t(\widehat{\theta}_{t-1}), \tag{4.3.11}$$

where $\eta_{t-1} > 0$ is the step-size at the $t-1$ step, and $\nabla_\theta \widehat{l}_t(\widehat{\theta}_{t-1})$ is the gradient using the X_t observation and the QMLE estimate $\widehat{\theta}_{t-1}$. This method is computationally efficient as it only requires a cost of $\mathcal{O}(d)$ per recursion. Depending on the number of observations, we have a trade-off between the accuracy of the recursive QML estimates and the time it takes to perform a parameter update ([19]).

According to [124], we must schedule the step-size such that $\sum_{t=1}^{\infty} \eta_t = \infty$ and $\sum_{t=1}^{\infty} \eta_t^2 < \infty$, but these bounds do not make the choice of an appropriate step-size η_t easier in practice. A more suitable approach is an adaptive learning rate, which updates the step-size in (4.3.11) on the fly pursuant to the gradient $\nabla_\theta \widehat{l}_t(\cdot)$. Thus, our choice of step-size η_t have less impact on performance, making convergence more robust and lower the demand for manually fine-tuning. Such an approach is often used in settings of streaming data as generic methods are preferred. Adaptive and separate learning rates for each parameter was proposed by [42] in their AdaGrad procedure. A different learning rate speeds up convergence in situations where the appropriate learning rates vary across parameters. Other well-known examples of adaptive learning rates could be AdaDelta by [155], RMSProp by [143] and ADAM by [83]. As we may expect a lack of convexity, we select the AdaGrad algorithm since it has shown promising results in non-convex optimization ([147]). The AdaGrad procedure is given by the updates

$$\widehat{\theta}_t = \widehat{\theta}_{t-1} - \frac{\eta}{\sqrt{\sum_{i=1}^{t} \nabla_\theta \widehat{l}_i(\widehat{\theta}_{i-1})^2 + \epsilon}} \nabla_\theta \widehat{l}_t(\widehat{\theta}_{t-1}), \tag{4.3.12}$$

where $\eta > 0$ is a constant learning rate and $\epsilon > 0$ a small number ensuring positivity. Good default values are $\eta = 0.1$ and $\epsilon = 10^{-8}$, e.g., see AdaVol in Algorithm 4.1. Note $\nabla_\theta \widehat{l}_i(\widehat{\theta}_{i-1})^2$ denotes the element-wise square $\nabla_\theta \widehat{l}_i(\widehat{\theta}_{i-1}) \odot \nabla_\theta \widehat{l}_i(\widehat{\theta}_{i-1})$.

As the QL loss is defined only for $\widehat{\theta}_n \in \mathcal{K}$, we will require that the recursive algorithm always takes values in $\mathcal{K}$. [157] suggests we project our approximation $\widehat{\theta}_n$ onto $\mathcal{K}$, preventing large jumps and enforcing the convergence of our stochastic gradient method. By implementing this projection on (4.3.12), we have our method for updating estimates, namely

$$\widehat{\theta}_t = \mathrm{P}_{\mathcal{K}}\left[\widehat{\theta}_{t-1} - \frac{\eta}{\sqrt{\sum_{i=1}^{t} \nabla_\theta \widehat{l}_i(\widehat{\theta}_{i-1})^2 + \epsilon}} \nabla_\theta \widehat{l}_t(\widehat{\theta}_{t-1})\right]. \tag{4.3.13}$$

4.3.1 Adaptive Recursive QML Estimation for GARCH Models

The GARCH process (X_t) parameters can be numerically challenging to estimate in empirical applications. The numerical optimization algorithms can quickly fail or converge to irregular solutions ([159]). Therefore, examining the approximative QMLE $\widehat{\theta}_n^*$ must be made with a healthy amount of skepticism. A well-discussed problem for the GARCH(p,q) models is that the QMLE performs poorly for numerically small (but still positive) values of ω. The parameter ω is vital and often tricky to estimate. Stabilizing the estimation of ω would not only improve the ω estimate but also have a positive impact on the other model parameters.

On way to overcome small values of ω for the GARCH(p,q) model is by scaling (X_t) with some factor $\lambda > 0$ as we have homogeneity; let (X_t) follow a GARCH(p,q) process with parameter vector $\theta = (\omega, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)^T$ and innovations (Z_t). Then for any $\lambda > 0$, the process $(\sqrt{\lambda}X_t)$ is a GARCH(p,q) process with parameter vector $\theta = (\lambda\omega, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)^T$ and identical innovations (Z_t).

However, we wish to avoid this form of inference in our recursive algorithm as one then needs to come up with a scaling parameter that has to be estimated beforehand. Instead, we circumvent this issue by introducing a concept called Variance Targeting Estimation (VTE) ([49]). We apply VTE for estimating ω by use of γ^2, which is the unconditional variance estimated by the sample variance (as seen in (4.3.14)). Thus we have a two-step estimator where we estimate the sample variance γ^2 recursively, and the remaining parameters $\theta = (\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)^T$ are estimated by the QML method. Pseudo-code of the AdaVol algorithm is presented in Algorithm 4.1. The reparametrization is obtained by defining

$$\omega = \gamma^2 \left(1 - \sum_{i=1}^{p} \alpha_i - \sum_{j=1}^{q} \beta_j\right). \tag{4.3.14}$$

The volatility process in the GARCH(p,q) process can then be rewritten as

$$(\sigma_t^2 - \gamma^2) = \sum_{i=1}^{p} \alpha_i (X_{t-i}^2 - \gamma^2) + \sum_{j=1}^{q} \beta_j (\sigma_{t-j}^2 - \gamma^2). \tag{4.3.15}$$

Similarly, one can define an ARCH(p) process by setting $\beta_j = 0$ for $1 \leq j \leq q$. The GARCH(p,q) process (X_t) in (4.3.15) has similar QL losses as before except $\nabla_\theta \widehat{\sigma}_t^2(\theta)$ in (4.2.9) and (4.2.10), where $\vartheta_t(\theta)$ is given as $(X_{t-1}^2 - \gamma^2, \ldots, X_{t-p}^2 - \gamma^2, \widehat{\sigma}_{t-1}^2(\theta) - \gamma^2, \ldots, \widehat{\sigma}_{t-q}^2(\theta) - \gamma^2)^T \in \mathbb{R}^{p+q}$ and the parameter space is defined by $\mathcal{K} = \left\{(\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q) \in \mathbb{R}_+^{p+q} \middle| \sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1\right\}$.

The VTE is not a requirement for the recursive method, but it provides additional speed and numerical stability. Namely, the VTE ensures a consistent estimate of the long-run variance, even if the model is misspecified. Additionally, presuming γ is well estimated, we reduce the parameter space dimension and increase the speed of convergence of the recursive optimization routines. Moreover, the geometry of the new set of optimization $\mathcal{K}$ allows the projection step in (4.3.13) to

Algorithm 4.1: AdaVol: Adaptive recursive QML estimation for GARCH(p, q) models using the technique of VTE.

Data: $(X_t)_{t\geq 1}$ (observations)
Inputs : $\widehat{\theta}_0$ (initial parameter vector), $\eta = 0.1$, $\epsilon = 10^{-8}$
Outputs: $\widehat{\theta}_t$ (resulting estimates), $\widehat{\sigma}^2_{t+1}$ (predicted volatility)
initialize: $\widehat{\sigma}^2_1 = X^2_1$, $\widehat{\mu}_0 = 0$, $\widehat{\gamma}^2_0 = 0$, $\widehat{G}_0 = \epsilon$ and $t = 0$
while $\widehat{\theta}_t$ **not converged do**
  $t = t + 1$
  $\widehat{\mu}_t = t(t+1)^{-1}\widehat{\mu}_{t-1} + (t+1)^{-1}X_t$
  $\widehat{\gamma}^2_t = (t-1)t^{-1}\widehat{\gamma}^2_{t-1} + t^{-1}\left(X_t - \widehat{\mu}_t\right)^2$
  $\widehat{g}_t = \nabla_\theta \widehat{l}_t(\widehat{\theta}_{t-1})$
  $\widehat{G}_t = \widehat{G}_{t-1} + \widehat{g}^2_t$
  $\widehat{\theta}_t = \mathrm{P}_{\mathcal{K}}\left[\widehat{\theta}_{t-1} - \eta \widehat{G}^{-1/2}_t \widehat{g}_t\right]$
  $\widehat{\sigma}^2_{t+1} = \widehat{\gamma}^2_t + \sum_{i=1}^p \widehat{\alpha}^{(t)}_i (X^2_{t-i} - \widehat{\gamma}^2_t) + \sum_{j=1}^q \widehat{\beta}^{(t)}_j (\widehat{\sigma}^2_{t-j} - \widehat{\gamma}^2_t)$

be efficiently implemented following [43].

One should be aware that the VTE requires stronger assumptions for the existence of the variance and is likely to suffer from efficiency loss. [49] also showed that the VTE would never be asymptotically more accurate than the QMLE. Another drawback of using the VTE is the need for a finite fourth moment of the process (X_t). Meaning, one would need $\alpha_1 < 0.57$ for an ARCH(1) model using standard Gaussian noise as $EX^4_t < \infty$ if and only if $\alpha^2_1 + (EZ^4_0 - 1)\alpha^2_1 < 1$. For a GARCH$(1, 1)$ model, we should have $(\alpha_1 + \beta_1)^2 + (EZ^4_0 - 1)\alpha^2_1 < 1$. These parameter bounds restrict the usefulness and range of applications for the VTE techniques. Fortunately, these constraints solely concern the batch setting.

4.4 Applications

In this section, we examine the AdaVol algorithm on simulated and real-life observations. Our implementation of AdaVol is provided in a repository at [149], and a relative speed comparison can be found in 4.7. We compare our approach to the Iterative QMLE (IQMLE) approximation $\widetilde{\theta}_n$, which is estimated at every two thousand increments using all observations up to this point, i.e., $(\widetilde{\theta}_t)_{(k-2000)+1\leq t\leq k}$ is estimated using $(X_t)_{1\leq t\leq k}$ for $k = 2000, 4000, \ldots, n$. In this way, we illuminate the large-scale learning trade-off of applying our recursive method instead of the iterative method, which is forward-looking with up to two thousand observations ([19]). As suggested by [75], we use the (bounded) *L-BFGS* algorithm to solve the nonlinear optimization problem in (4.2.3) for $\widetilde{\theta}_n$ with initial guess $\widetilde{\theta}_0 \in \mathcal{K}$. Our recursive QMLE approximation $\widehat{\theta}_n$ is produced by the AdaVol algorithm (described in Algorithm 4.1). It takes our initial value $\widehat{\theta}_0 \in \mathcal{K}$, learning rate $\eta = 0.1$ and $\epsilon = 10^{-8}$ as input. At last, for a fair comparison, we always use the same initial guess for both methods, namely $\widehat{\theta}_0 = \widetilde{\theta}_0 \in \mathcal{K}$.

It is possible to customize AdaVol by tuning the learning parameter η, e.g., by choosing the best

performing learning rate evaluated on the first part of the observations. We use a fixed learning rate $\eta = 0.1$ across all applications (simulated and real-life observations) to avoid the learning rate's potential influence in our experiments. However, one should be aware of the versatility achieved with different learning rate choices. The choice of learning rates is cumbersome, as an excessive learning rate can cause the algorithm to deviate from the true parameter estimate. In contrast, a learning rate that is too small can lead to slow convergence. Nevertheless, a small learning rate may be preferred if one only wants to keep track of minor parameter estimation changes.

4.4.1 Simulations

All simulations are performed by the use of twenty thousand observations ($n = 20000$), and the simulated data (X_t) is always generated using Gaussian innovations with zero mean and unit variance. To avoid possible bias due to the choice of the true parameter vector θ_0 and initial values $\widehat{\theta}_0, \widetilde{\theta}_0$, we conduct our experiments using random parameter vectors $\theta_0 \in \mathcal{K}$ and random initial guesses $\widehat{\theta}_0, \widetilde{\theta}_0 \in \mathcal{K}$. These parameter vectors are drawn randomly from our parameter space $\mathcal{K}$. The ω parameter is generated by taking a positive number from a uniform distribution, and then we multiply it with $10^{-\tau}$, where τ is some random positive integer up to eight. In this way, we cover a broad parameter domain while having parameter values close to the boundary. Similarly, the $(\alpha_i)_{1\leq i\leq p}$ and $(\beta_j)_{1\leq j\leq q}$ parameters is generated from a uniform distribution with the condition of having $\sum_{i=1}^{p}\alpha_i + \sum_{j=1}^{q}\beta_j < 1$. Note that the initial guesses $\widehat{\theta}_0$ and $\widetilde{\theta}_0$ are generated the same way. Thus, when we mention random parameters for the rest of the paper, we refer to this generation procedure.

ARCH Models

As discussed earlier, the iterative QMLE approximation $\widetilde{\theta}_n$ performs poorly for numerically small $\omega > 0$ values, which are often encountered in financial time series. Before moving on to the case of small ω parameter values, we have in Figure 4.1a the trajectories of both QMLE approximations using an ARCH(1) process with true parameter vector and initial values given by

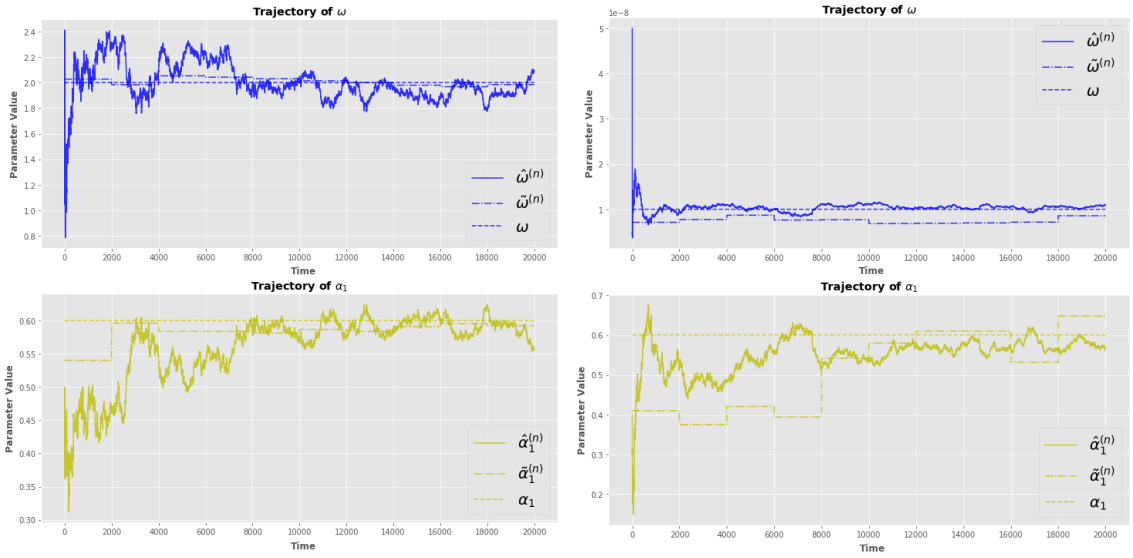
$$\theta_0 = \begin{pmatrix} \omega \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} 2.0 \\ 0.6 \end{pmatrix} \text{ and } \widehat{\theta}_0 = \widetilde{\theta}_0 = \begin{pmatrix} 1.5 \\ 0.4 \end{pmatrix}. \tag{4.4.16}$$

Figure 4.1a shows a very reasonable convergence of both estimators, $\widehat{\theta}_n = (\widehat{\omega}^{(n)}, \widehat{\alpha}_1^{(n)})^T$ and $\widetilde{\theta}_n = (\widetilde{\omega}^{(n)}, \widetilde{\alpha}_1^{(n)})^T$, when the true parameter $\omega = 2.0$. Not surprisingly, our method experiences some fluctuations initially, but as the learning rate decreases, the fluctuation likewise evaporates, and within the first few thousand observations, we hit the true parameter values.

Likewise, in Figure 4.1b, we have the QMLE approximations' trajectories for an ARCH(1)

Figure 4.1: Trajectory of $\widehat{\theta}_n$ (solid line) and $\widetilde{\theta}_n$ (semi-dotted line) for an ARCH(1) process with true parameter vector (dotted line) and initial guess given in sub-caption.

(a) (4.4.16). (b) (4.4.17).



process, but now with true parameter vector and initial guess given as

$$\theta_0 = \begin{pmatrix} 1 \cdot 10^{-8} \\ 0.6 \end{pmatrix} \text{ and } \widehat{\theta}_0 = \widetilde{\theta}_0 = \begin{pmatrix} 5 \cdot 10^{-8} \\ 0.4 \end{pmatrix}. \tag{4.4.17}$$

Figure 4.1b indicates a modest convergence of $\widehat{\theta}_n$ but shows slow convergence of $\widetilde{\alpha}_n$ towards the true α_1 parameter. In addition, $\widetilde{\alpha}_n$ seems biased concerning the initial value $\widetilde{\alpha}_0 = 0.4$ as it processes almost half of the observations before moving closer to the true $\alpha_1 = 0.6$.

A way of demonstrating the variation of $\widehat{\theta}_n$ and $\widetilde{\theta}_n$ performance for small ω values is presented in Figure 4.2a and Figure 4.2b, where we have the average trajectory of one hundred trajectories with their corresponding boxplots showing the distribution of these one hundred trajectories.

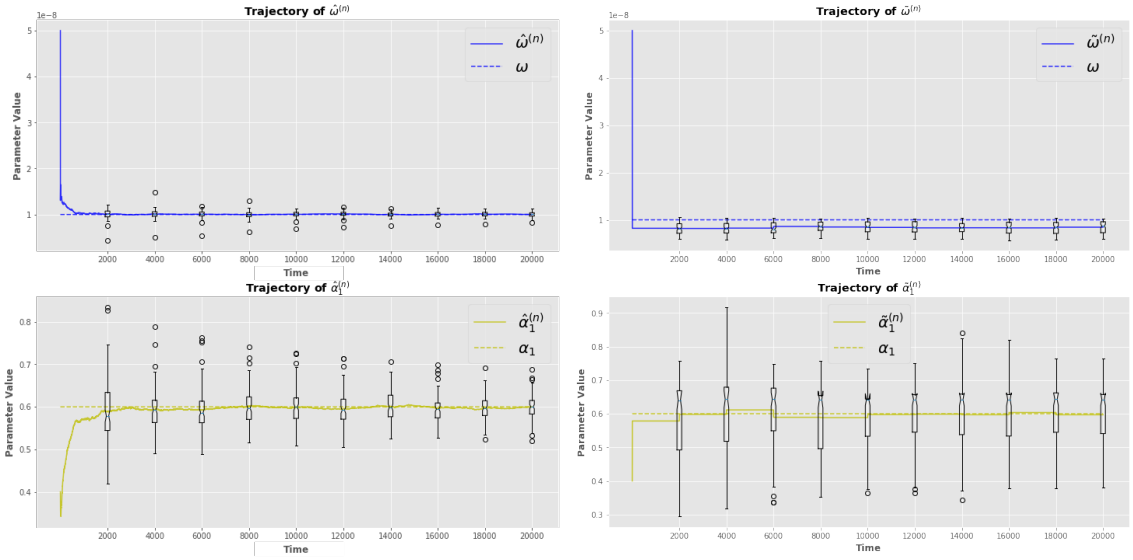
Here, in Figure 4.2a, we can see that AdaVol converges to the true parameter values with low sensitivity to the choice of initial values. Moreover, this convergence occurs within the first few thousand observations. However, in Figure 4.2b, we see the opposite in which $\widetilde{\theta}_n$ has convergence issues; it is consistently underestimating the ω parameter. Furthermore, the α_1 parameter range does not appear to be decreasing over time, and the range seems larger than AdaVol's.

As we observe the true volatility process (σ_t) in this section, we can evaluate the predicted volatility processes' accuracy. We do this using the Mean Percentage Errors (MPE) given as

$$\widehat{\sigma}_{\text{MPE}} = \frac{1}{n}\sum_{t=1}^{n} \frac{\sigma_t - \widehat{\sigma}_t}{\sigma_t} \text{ and } \widetilde{\sigma}_{\text{MPE}} = \frac{1}{n}\sum_{t=1}^{n} \frac{\sigma_t - \widetilde{\sigma}_t}{\sigma_t}, \tag{4.4.18}$$

Figure 4.2: Average trajectory (solid line) of one hundred $\widehat{\theta}_n, \widetilde{\theta}_n$'s for an ARCH(1) process with true parameter vector (dotted line) and initial guess from (4.4.17). The boxplots shows the distribution of the one hundred trajectories.

(a) $(\widehat{\theta}_n)$ (b) $(\widetilde{\theta}_n)$



and the Mean Absolute Percentage Errors (MAPE) given by

$$\widehat{\sigma}_{\text{MAPE}} = \frac{1}{n}\sum_{t=1}^{n}\frac{|\sigma_t - \widehat{\sigma}_t|}{\sigma_t} \text{ and } \widetilde{\sigma}_{\text{MAPE}} = \frac{1}{n}\sum_{t=1}^{n}\frac{|\sigma_t - \widetilde{\sigma}_t|}{\sigma_t}, \tag{4.4.19}$$

where $(\widehat{\sigma}_t)$ is coming from AdaVol and $(\widetilde{\sigma}_t)$ from the IQMLE approximation. Note that $\widetilde{\sigma}_t$'s estimation is the same as for the IQMLE approximation $\widetilde{\theta}_t$, i.e., $(\widetilde{\sigma}_t)_{(k-2000)+1\leq t\leq k}$ is estimated using $(X_t)_{1\leq t\leq k}$ for $k = 2000, 4000, \ldots, n$.

In the rest of this section, we will use random parameters to generalize our studies, limiting the potential bias from having fixed parameters (See Section 4.4.1). Our routine is as follows: We draw a random true parameter vector $\theta_0 \in \mathcal{K}$ from which we generate our observations (X_t). Based on these observations (X_t), we calculate our estimates using (a random) $\widehat{\theta}_0 = \widetilde{\theta}_0 \in \mathcal{K}$. Then, we evaluate our estimates using an accuracy score, e.g., MPE and MAPE. Finally, we repeat all these steps the desired number of times. Boxplots of one hundred accuracy scores, MPE in (4.4.18) and MAPE in (4.4.19), can be found in Figure 4.3. In the top graph of Figure 4.3, one can observe the MPE (for both methods) is symmetric around zero, but $\widetilde{\sigma}_{\text{MPE}}$ has a negative tail, meaning the iterative method may overestimate the volatility in some cases. Also, the spread of $\widetilde{\sigma}_{\text{MPE}}$ is higher than the $\widehat{\sigma}_{\text{MPE}}$, which is clearly seen by looking at $\widetilde{\sigma}_{\text{MAPE}}$ in the bottom graph of Figure 4.3.

Another way of measuring the accuracy can be made by studying the conditional quantiles using the recursive $(\widehat{\sigma}_t)$ and iterative $(\widetilde{\sigma}_t)$ predicted volatility processes ([16]). Under the assumption of standard Gaussian innovations, X_t is Gaussian with zero mean and variance σ_t^2. Thus, for any

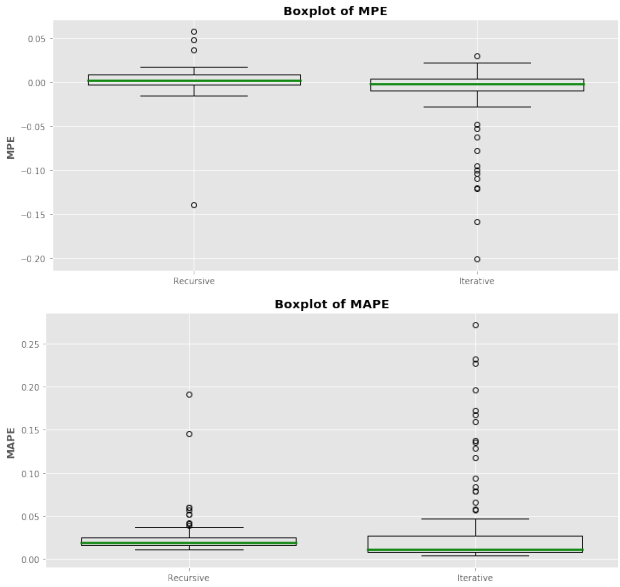


Figure 4.3: Boxplots of one hundred accuracy scores MPE (4.4.18) and MAPE (4.4.19) using an ARCH(1) process with random true parameter vector and initial guess in $\mathcal{K}$.

$\alpha \in (0,1)$, the α-quantile of a Gaussian distribution $\mathcal{N}(0,\sigma_t^2)$ is $\sigma_t\Phi^{-1}(\alpha)$, where $\Phi^{-1}(\alpha)$ is the α-quantile of the standard Gaussian distribution. We use the so-called α-quantile loss function proposed by [84]: The α-quantile loss function ρ_α using the volatility process σ_t is defined as

$$\rho_\alpha(X_t,\sigma_t) = \begin{cases} \alpha\left(X_t - \Phi^{-1}(\alpha)\sigma_t\right), & \text{for } X_t > \Phi^{-1}(\alpha)\sigma_t, \\ (1-\alpha)\left(\Phi^{-1}(\alpha)\sigma_t - X_t\right), & \text{for } X_t \leq \Phi^{-1}(\alpha)\sigma_t, \end{cases} \tag{4.4.20}$$

with tilting parameter $\alpha \in (0,1)$. The idea behind the α-quantile loss function is to penalize quantiles of low probability more for overestimation than for underestimation (and reversely for high probability quantiles). We evaluate across the α-quantile scores ρ_α of (σ_t) by the (normalized) cumulative α-quantile scoring function QS_α:

$$QS_\alpha(X_n,\sigma_n) = \frac{1}{n}\sum_{t=1}^{n}\sum_{m=1}^{M}\rho_{\alpha_m}(X_t,\sigma_t), \tag{4.4.21}$$

with M as the number of quantiles $\alpha = \{\alpha_1,\ldots,\alpha_M\}$. The lowest QS_α score indicates the best ability of volatility forecast. The findings of one hundred $QS_\alpha(X_n,\widehat{\sigma}_n)$ and $QS_\alpha(X_n,\widetilde{\sigma}_n)$ scores is presented in Figure 4.4, where we have used $\alpha = \{0.01, 0.02, \ldots, 0.99\}$, a random true parameter vector and random initialization in $\mathcal{K}$. The QS_α scores in Figure 4.4 are indistinguishable. This indicates no loss of generality in using our recursive method even though our estimates are calculated only once, making them more adaptable over time. Surprisingly, the iterative method is not superior, even when forward-looking (with up to two thousand observations).

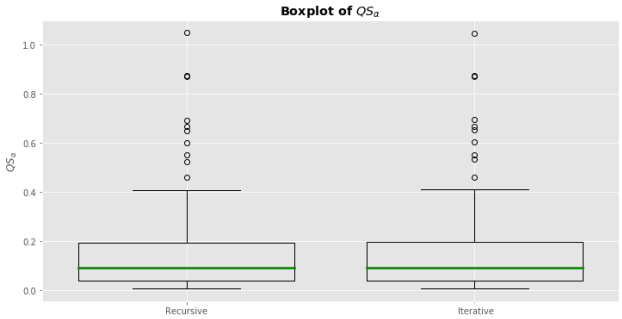


Figure 4.4: Boxplots of one hundred QS_α scores with $\alpha = \{0.01, 0.02, \ldots, 0.99\}$ using an ARCH(1) model with random true parameter vector and initial value in $\mathcal{K}$.

GARCH Models

Figure 4.5a and 4.5b shows the trajectories of the parameter estimates $\widehat{\theta}_n = (\widehat{\omega}^{(n)}, \widehat{\alpha}_1^{(n)}, \widehat{\beta}_1^{(n)})^T$ and $\widetilde{\theta}_n = (\widetilde{\omega}^{(n)}, \widetilde{\alpha}_1^{(n)}, \widetilde{\beta}_1^{(n)})^T$ for a GARCH$(1,1)$ model with the true parameter vector and initial guess given by

$$\theta_0 = \begin{pmatrix} \omega \\ \alpha_1 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 1 \cdot 10^{-8} \\ 0.2 \\ 0.7 \end{pmatrix} \text{ and } \widehat{\theta}_0 = \widetilde{\theta}_0 = \begin{pmatrix} 5 \cdot 10^{-8} \\ 0.1 \\ 0.8 \end{pmatrix}. \tag{4.4.22}$$

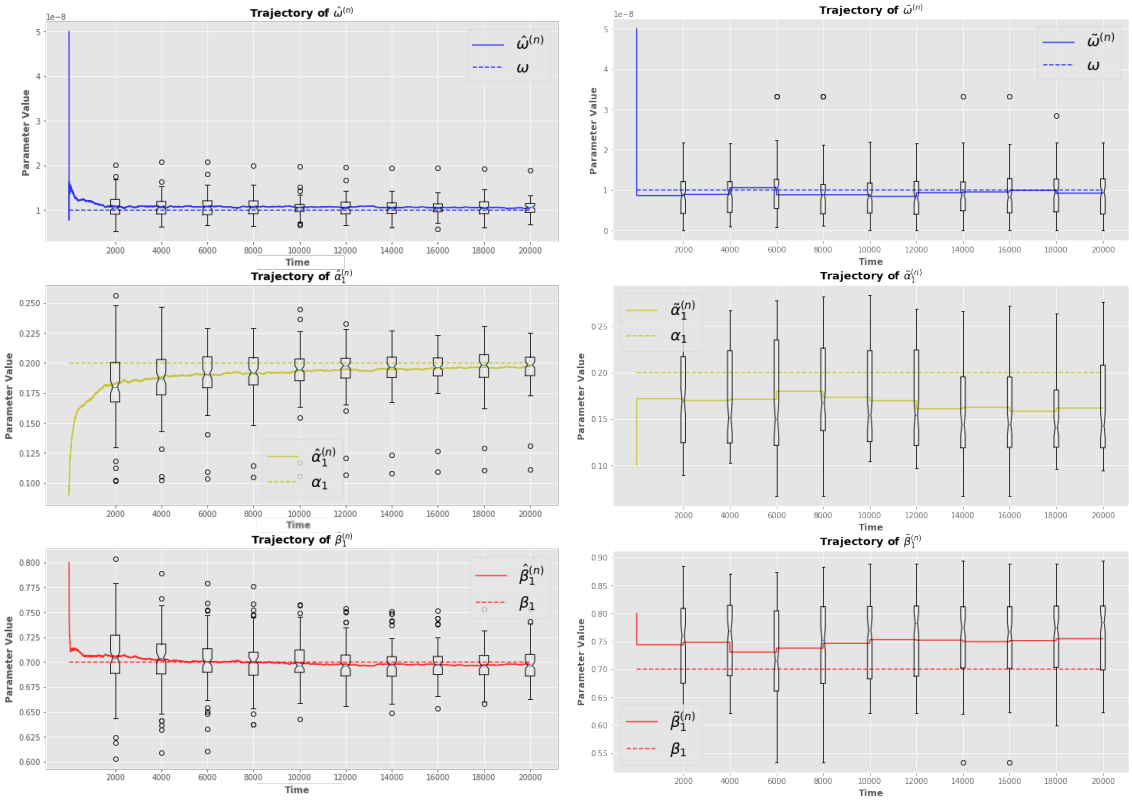
As for the ARCH(1) model, we observe a lower spread in the parameter trajectories coming from AdaVol $\widehat{\theta}_n$ than from the IQMLE approximation $\widetilde{\theta}_n$. Moreover, the iterative $\widetilde{\theta}_n$ is consistently overestimating the β_1 parameter (and underestimating the α_1 parameter), indicating a bias relative to the initial value. It is worth mentioning that even if all initial values are in the stationary region, i.e., $\widehat{\theta}_0 = \widetilde{\theta}_0 = \theta_0 \in \mathcal{K}$, we still have a proper amount of fluctuation in the parameter trajectories. As discussed before, this may partially be due to the volatility introduced by the gradient method and the flatness of the QL loss ([159]). Nevertheless, our recursive method possesses a remarkable convergence already after the first few thousand observations.

The accuracy scores, namely MPE from (4.4.18) and MAPE from (4.4.19), can be found in Figure 4.6 for the GARCH$(1,1)$ model using random true parameter vector and random initial values in $\mathcal{K}$. By comparing our methods using random initializations, we circumvent the possible bias from the initial guess, which we observed in Figure 4.5b for the iterative method. As in the ARCH(1) case, we obtain a lower spread for $\widehat{\sigma}_{\text{MPE}}$ than $\widetilde{\sigma}_{\text{MPE}}$. Nevertheless, one should still expect some probability of ending up with an irregular solution where the AdaVol algorithm fails to converge.

Figure 4.7 presents the results of one hundred QS_α scores with random true parameter vector and initial value in $\mathcal{K}$. Again, the QS_α scores are indistinguishable (even when the iterative method is forward-looking).

Figure 4.5: Average trajectory (solid line) of one hundred $\widehat{\theta}_n, \widetilde{\theta}_n$'s for a GARCH$(1,1)$ process with true parameter vector (dotted line) and initial guess given in (4.4.22). The boxplots shows the distribution of the one hundred trajectories.

(a) $(\widehat{\theta}_n)$ (b) $(\widetilde{\theta}_n)$



4.4.2 Real-life Observations

We will now demonstrate AdaVol's abilities on real-life observations showing how our technique works in practice. Table 4.1 shows an overview of the used stock market indices. All empirical studies use the GARCH$(1,1)$ model, but higher-order parameters may yield a better fit for some stock market indices. As the observation period spans over a long time, it is unlikely that the log-return series is stationary. To exhibit AdaVol's ability to adapt to time-varying estimates, we begin by considering the S&P500 Index in Section 4.4.2. Afterward, in Section 4.4.2, we investigate the remaining six stock market indices presented in Table 4.1, namely the CAC, DAX, DJIA, NDAQ, NKY, and RUT index.

Application to the S&P500 Index

We apply our method on the S&P500 Index from January 1950 to September 2020, consisting of $n = 17672$ observations to test real-life data performance. We employ the GARCH$(1,1)$ model

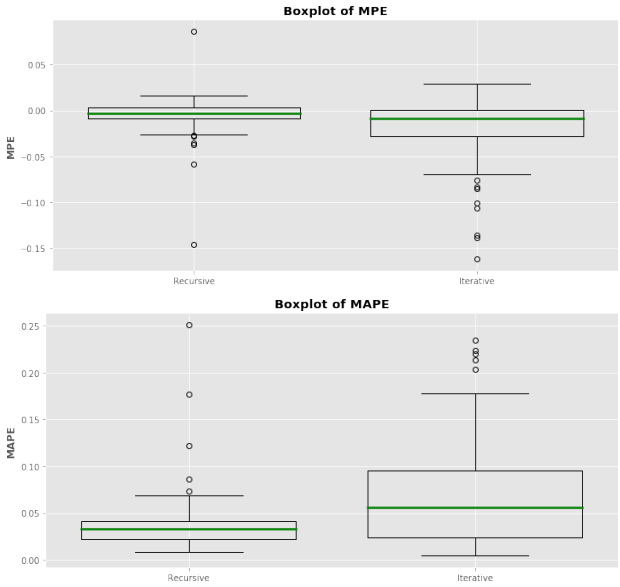


Figure 4.6: Boxplots of one hundred accuracy scores MPE (4.4.18) and MAPE (4.4.19) using a GARCH$(1,1)$ process with true parameter vector and random initial guess in $\mathcal{K}$.

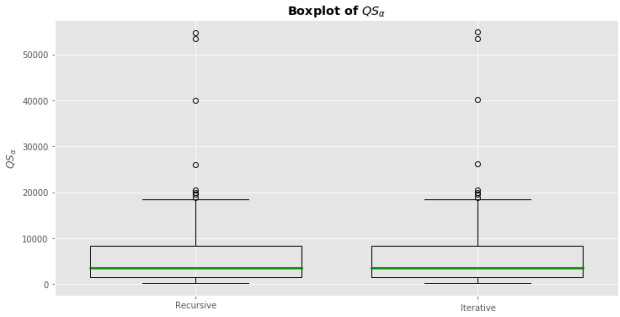


Figure 4.7: Boxplots of one hundred QS_α scores with $\alpha = \{0.01, 0.02, \ldots, 0.99\}$ using the GARCH$(1,1)$ model with random true parameter vector and initial value in $\mathcal{K}$.

Stock Market Index		Period
CAC 40	(CAC)	March 1990 - Sep. 2020
DAX 30	(DAX)	Jan. 1988 - Sep. 2020
Dow Jones	(DJIA)	Feb. 1985 - Sep. 2020
NASDAQ Composite	(NDAQ)	Feb. 1971 - Sep. 2020
Nikkei 225	(NKY)	Jan. 1965 - Sep. 2020
Russell 2000	(RUT)	Nov. 1987 - Sep. 2020
Standard & Poor's 500	(S&P500)	Jan. 1950 - Sep. 2020

Table 4.1: Overview of considered stock market indices including their observation periods. The observations consist of daily log-returns which are defined as log differences of the closing prices of the index between two consecutive days.

with initial values:

$$\widehat{\theta}_0 = \widetilde{\theta}_0 = \begin{pmatrix} 5 \cdot 10^{-5} \\ 0.05 \\ 0.9 \end{pmatrix}. \tag{4.4.23}$$

The QML trajectories can be seen in Figure 4.8. The produced AdaVol estimates $\widehat{\theta}_n = (\widehat{\omega}^{(n)}, \widehat{\alpha}_1^{(n)}, \widehat{\beta}_1^{(n)})^T$ experience some fluctuations initially, but as it vaporizes, it is clear that our estimates change over time. Most remarkable are the shifts our estimates make around some historical market crashes, e.g., Black Monday, the financial crisis, and COVID-19. The instant shift in our estimates is an appealing property for detecting structural breaks. It is noteworthy that the estimates of the IQMLE approximation $\widetilde{\theta}_n = (\widetilde{\omega}^{(n)}, \widetilde{\alpha}_1^{(n)}, \widetilde{\beta}_1^{(n)})^T$ are predominantly constant over time with minor changes except for some years between 1990 and 2000, where we detect a shift to lower $\widetilde{\beta}_1^{(n)}$ values and higher $\widetilde{\omega}^{(n)}$ values.

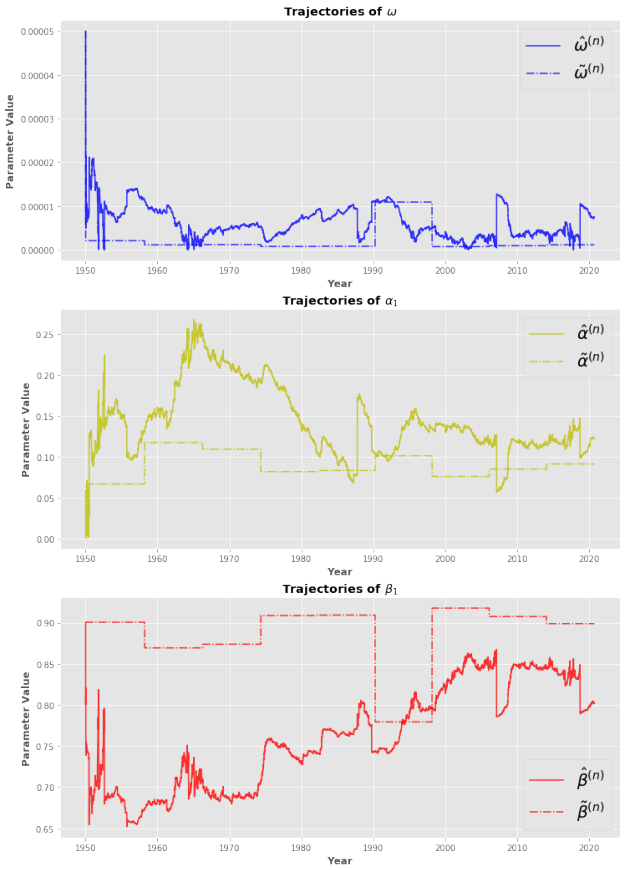


Figure 4.8: Trajectory of the recursive $\widehat{\theta}_n$ (solid line) and iterative $\widetilde{\theta}_n$ (semi-dotted line) QML estimate using a GARCH$(1,1)$ model on S&P500 Index log-returns from year 1950 to 2020. Both methods use initial value given in (4.4.23).

In Figure 4.9, we have the log-returns r_t of the S&P500 Index, and the confidence intervals $\bar{r} \pm 1.96\widehat{\sigma}_t$ and $\bar{r} \pm 1.96\widetilde{\sigma}_t$ using the recursive $\widehat{\sigma}_t$ and iterative $\widetilde{\sigma}_t$ predicted volatilities, where $\bar{r}$ is the mean of the log-returns r_t. It seems that the recursive method $\widehat{\sigma}_t$ adapts more rapidly than the iterative one $\widetilde{\sigma}_t$ to changes in the S&P500 Index observations r_t. Especially in Figure 4.9, under the COVID-19 crisis, we encountered a period with a substantial volatility increase. Here, we observe $\widehat{\sigma}_t$'s ability to track changing volatilities better than $\widetilde{\sigma}_t$.

In the absence of the true (unobserved) variance process (σ_t^2), the efficiency of our recursive $(\widehat{\sigma}_t)$

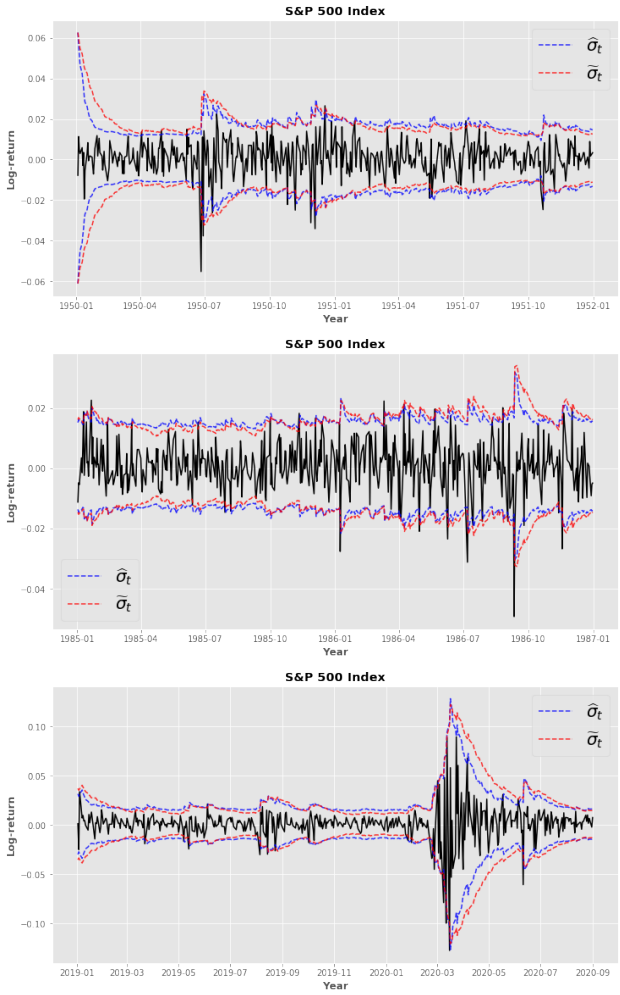


Figure 4.9: Log-returns r_t of S&P500 Index (solid lines) and confidence intervals $\bar{r} \pm 1.96\widehat{\sigma}_t$ and $\bar{r} \pm 1.96\widetilde{\sigma}_t$ (dotted lines) using the recursive $\widehat{\sigma}_t$ (blue) and iterative $\widetilde{\sigma}_t$ (red) predicted volatilities, where $\bar{r}$ is the mean of the log-returns r_t. From top to bottom, we have Jan. 1950 to Jan. 1952, Jan. 1985 to Jan. 1987, and Jan. 2019 to Sep. 2020.

and the iterative ($\widetilde{\sigma}_t$) volatility can be appraised with the use of the squared log-returns (r_t^2). We use the Mean Absolute Errors (MAE) defined by

$$\widehat{\sigma}^2_{\text{MAE}} = \frac{1}{n}\sum_{t=1}^{n} |r_t^2 - \widehat{\sigma}_t^2| \text{ and } \widetilde{\sigma}^2_{\text{MAE}} = \frac{1}{n}\sum_{t=1}^{n} |r_t^2 - \widetilde{\sigma}_t^2|. \tag{4.4.24}$$

In Table 4.2, we consider the MAEs for the same periods used in Figure 4.9, including for the full dataset. The results in Table 4.2 confirm our conclusions about Figure 4.9; the AdaVol method tracks the volatility better than the iterative method.

Figure 4.10 contains the results of one hundred QS_α scores using the recursive ($\widehat{\sigma}_t$) and iterative ($\widetilde{\sigma}_t$) volatility process, respectively, with random initial values in $\mathcal{K}$. Remarkably, AdaVol outperforms the iterative method, although the latter uses future information, i.e., $(\widetilde{\sigma}_t)_{(k-2000)+1 \le t \le k}$

Period	$\widehat{\sigma}^2_{\text{MAE}}$	$\widetilde{\sigma}^2_{\text{MAE}}$
Jan. 1950 - Jan. 1952	8.2388	8.9049
Jan. 1985 - Jan. 1987	7.1214	7.4723
Jan. 2018 - Sep. 2020	26.9205	30.4775
Jan. 1950 - Sep. 2020	10.1861	10.6731

Table 4.2: MAEs (4.4.24) using log-returns r_t of S&P500 Index with the recursive $\widehat{\sigma}_t$ and iterative $\widetilde{\sigma}_t$ predicted volatilities. Both methods has initial value given in (4.4.23). The $\widehat{\sigma}^2_{\text{MAE}}$ and $\widetilde{\sigma}^2_{\text{MAE}}$ numbers are scaled by 10^{-5}.

is estimated using $(r_t)_{1\leq t\leq k}$ for $k = 2000, 4000, \ldots, 16000, 17505$. This indicates that one could achieve better performance using the recursive method, even if it only predicts volatility using previous information.

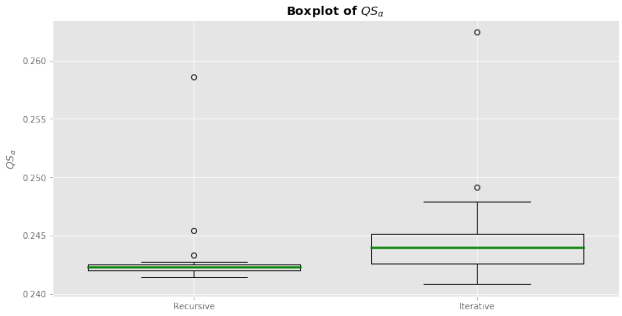


Figure 4.10: Boxplots of one hundred QS_α scores with use of the recursive $\widehat{\sigma}_t$ and iterative $\widetilde{\sigma}_t$ volatility process, respectively, for $\alpha = \{0.01, 0.02, \ldots, 0.99\}$, using the GARCH$(1,1)$ model on the log-returns r_t of S&P500 Index with random initial value in $\mathcal{K}$.

Other Stock Market Indices

We now extend our analysis to the remaining stock market indices from Table 4.1, namely the CAC, DAX, DJIA, NDAQ, NKY, and RUT index. In Figure 4.11, we can observe AdaVol's ability to adapt to time-varying parameters seems to hold for several stock market indices. These figures show a clear benefit in recursive estimation as it increases adaptivity that may be advantageous under a financial crisis such as the COVID-19.

These conclusions are confirmed in Figure 4.12, where we have one hundred QS_α scores using the recursive $(\widehat{\sigma}_t)$ and iterative $(\widetilde{\sigma}_t)$ volatility process with random initial values in $\mathcal{K}$. As for the S&P500 Index (in Figure 4.10), our findings indicate that the recursive approach estimates the QS_α quantiles better than the iterative method, both on average and with a lower spread.

The assumption of having an underlying data generation process with constant "true" parameters may not hold in real-life examples. Thus, AdaVol seems to have an advantage compared to the iterative method, as it estimates the parameters step-by-step. In contrast, the iterative method always has to estimate the parameters using all observations over an extensive period of time.

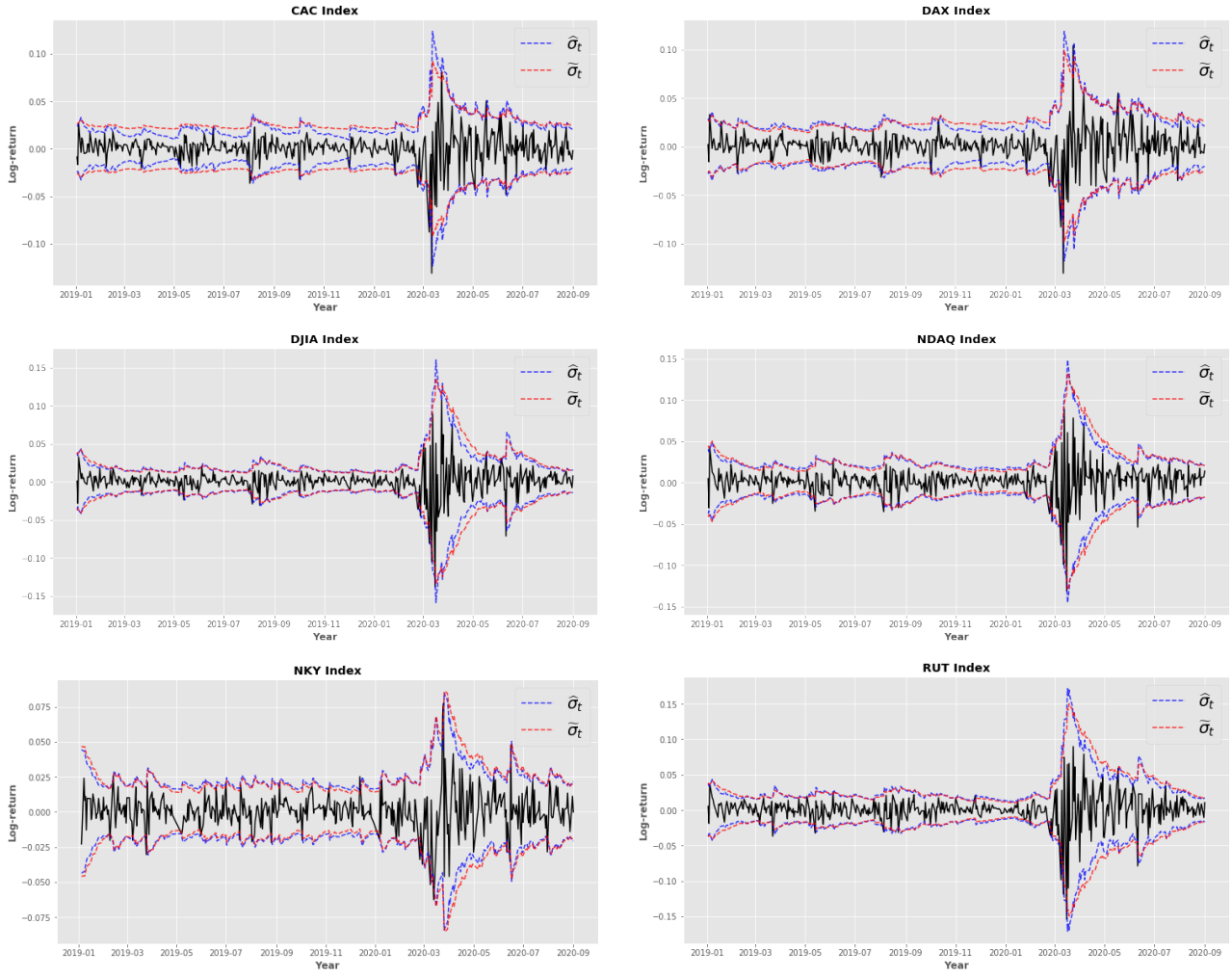


Figure 4.11: Log-returns r_t of the CAC (top-left), DAX (top-right), DJIA (mid-left), NDAQ (mid-right), NKY (bottom-left) and RUT (bottom-right) index (solid lines) and confidence intervals $\bar{r} \pm 1.96\widehat{\sigma}_t$ and $\bar{r} \pm 1.96\widetilde{\sigma}_t$ (dotted lines) using the recursive $\widehat{\sigma}_t$ (blue) and iterative $\widetilde{\sigma}_t$ (red) predicted volatilities, where $\bar{r}$ is the mean of the log-returns r_t. The period is Jan. 2019 to Sep. 2020.

4.5 Conclusion

We proposed an adaptive approach to recursively estimate GARCH model parameters in a streaming setting using the VTE technique (AdaVol). AdaVol's design showed to produce resilient and adaptive estimates in our empirical investigations. The adaptation to time-varying parameters was a surprising advantage that appeared when we applied our method to real-life observations. As the assumption of having constant estimates seems not to be the case for the stock indices we analyze, then it is beneficial to have the ability to adapt. One could facilitate this ability more by incorporating a rolling volatility estimation of γ instead of using the sample volatility. Combining this with a different learning rate than AdaGrad, which enables continuous learning (e.g., ADAM by [83]), could encourage adaptability.

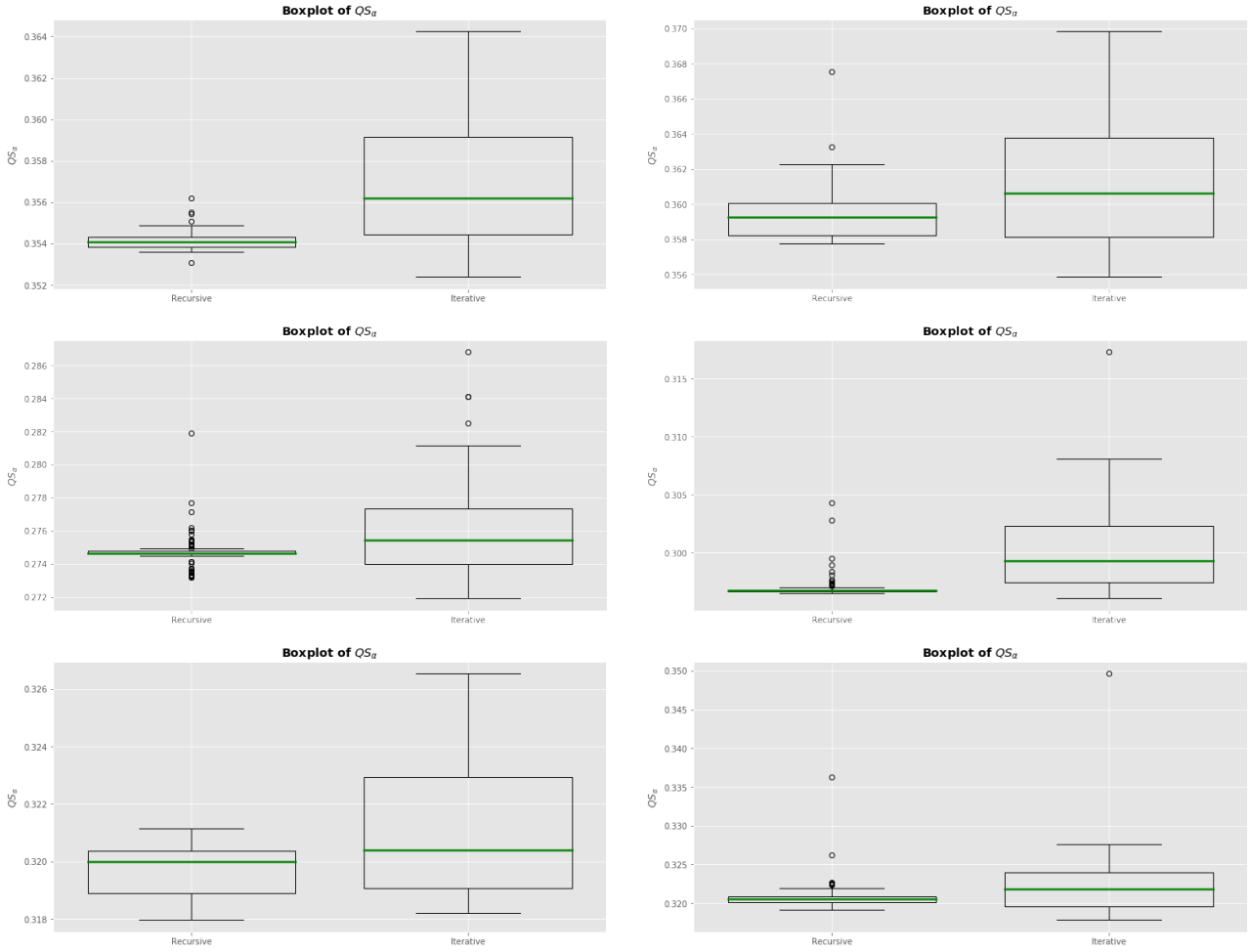


Figure 4.12: Boxplots of one hundred QS_α scores with the use of the recursive $\widehat{\sigma}_t$ and iterative $\widetilde{\sigma}_t$ volatility process, respectively, for $\alpha = \{0.01, 0.02, \ldots, 0.99\}$, using the GARCH$(1,1)$ model on the log-returns r_t of the CAC (top-left), DAX (top-right), DJIA (mid-left), NDAQ (mid-right), NKY (bottom-left) and RUT (bottom-right) index with random initial values in $\mathcal{K}$.

4.5.1 Future Perspectives

We proved asymptotic local convexity of the QL function in general conditionally heteroscedastic time series models of multiplicative form. An interesting question arises: can one prove Theorem 4.2.1 for a bounded set of N observations? Expressed differently, can one find a N bounded, such that we have convergence/convexity of recursive algorithms, e.g., for the GARCH, EGARCH, and AGARCH models. To our knowledge, this has not been proved yet.

The stability of using our recursive approach to solve the QML problem could be improved by using a mini-batch approach. A mini-batch approach will lower each incremental volatility as one uses more observations per recursion to update the QML estimate. Applying a mini-batch method does not require much more computational power than the stochastic gradient descent, only $\mathcal{O}(bd)$, where b is the number of observations used in each (mini-batch) recursion. Using more observations,

we could achieve more consistency and smoothness in the estimation procedure's convergence while keeping favorable computational costs.

Furthermore, an accelerated convergence of our estimates could be obtained by recursion averaging, also called Polyak-Ruppert averaging, which is guaranteed under fairly relaxed conditions ([118, 129]). This Polyak-Ruppert average estimate could be utilized solely or employed as a benchmark to detect structural breaks.

Finally, it could be interesting to extend all of these concepts around our work with Adavol to the multivariate case [91, 114].

4.6 Proofs

Proof of Theorem 4.2.1. To prove local strong convexity for the approximate QL function $\widehat{L}_n$ using the approximate QMLE $\widehat{\theta}_n^*$, we first list some bounds for the Hessians: under the regularity conditions on the derivatives of h_t, then using (4.2.5), we can write

$$\nabla_\theta l_t(\theta) = \frac{1}{2}\frac{\nabla_\theta h_t(\theta)}{h_t(\theta)}\left(1 - \frac{X_t^2}{h_t(\theta)}\right)$$

and

$$\nabla_\theta^2 l_t(\theta) = \frac{1}{2h_t^2(\theta)}\left(\nabla_\theta h_t(\theta)^T \nabla_\theta h_t(\theta)\left(\frac{2X_t^2}{h_t(\theta)} - 1\right) + \nabla_\theta^2 h_t(\theta)\left(h_t(\theta) - X_t^2\right)\right),$$

where the Hessian $H_n(\theta)$ is defined as $n^{-1}\nabla_\theta^2 L_n(\theta) = n^{-1}\sum_{t=1}^n \nabla_\theta^2 l_t(\theta)$. Similarly, for $\nabla_\theta \widehat{l}_t(\theta)$, $\nabla_\theta^2 \widehat{l}_t(\theta)$, and $\widehat{H}_n(\theta)$, we replace $h_t(\theta), \nabla_\theta h_t(\theta)$ and $\nabla_\theta^2 h_t(\theta)$ by $\widehat{h}_t(\theta), \nabla_\theta \widehat{h}_t(\theta)$ and $\nabla_\theta^2 \widehat{h}_t(\theta)$, respectively. From Assumption 4.2.2, we know $n^{-1}\|\nabla_\theta^2 \widehat{L}_n - \nabla_\theta^2 L_n\|_{\mathcal{K}} \xrightarrow{\text{a.s.}} 0$ for $n \to \infty$. Hence, for some random N_1 large enough, there exists $\epsilon > 0$ such that $n^{-1}\|\nabla_\theta^2 \widehat{L}_n - \nabla_\theta^2 L_n\|_{\mathcal{K}} < \epsilon$ for all $n \geq N_1$ a.s. As a consequence, we get

$$\|\widehat{H}_n - H_n\|_{\mathcal{K}} < \epsilon, \quad \text{a.s.}, \tag{4.6.25}$$

for all $n \geq N_1$. Similarly, applying the ergodic theorem on the integrable sequence (uniformly over $\mathcal{K}$) $(\nabla_\theta^2 l_t)$ of continuous functions over the compact set $\mathcal{K}$, we obtain $\|n^{-1}\sum_{t=1}^n \nabla_\theta^2 l_t - \mathbb{E}[\nabla_\theta^2 l_0]\|_{\mathcal{K}} \xrightarrow{\text{a.s.}} 0$ for $n \to \infty$. Then there exists N_2 such that

$$\|H_n - H_0\|_{\mathcal{K}} < \epsilon, \quad \text{a.s.}, \tag{4.6.26}$$

for all $n \geq N_2$. Thus, by equation (4.6.25) and (4.6.26), we know there exists $N = \max(N_1, N_2)$ such that for all $n \geq N$, we have

$$\|\widehat{H}_n - H_0\|_{\mathcal{K}} \leq \|\widehat{H}_n - H_n\|_{\mathcal{K}} + \|H_n - H_0\|_{\mathcal{K}} < 2\epsilon, \quad \text{a.s.}$$

Especially, as $\|\widehat{H}_n - H_0\|_{\mathcal{K}}$ is defined as $\sup_{\theta\in\mathcal{K}}\|\widehat{H}_n(\theta) - H_0(\theta)\|_{op}$, then

$$\|\widehat{H}_n(\theta) - H_0(\theta)\|_{op} < 2\epsilon, \tag{4.6.27}$$

for all $\theta \in \mathcal{K}$.

From [138, Lemma 7.2], the asymptotic Hessian $H_0(\theta_0) = \mathbb{E}[\nabla^2_\theta l_0(\theta_0)]$ is a symmetric positive definite matrix a.s. under Assumption 4.2.3. As $H_0(\theta)$ is the limit of the continuous matrix-valued function $H_n(\theta)$, it is itself a continuous matrix-valued function. Thus, the eigenvalue function $\lambda^i_0(\theta)$ for $1 \le i \le d$ of $H_0(\theta)$ is also continuous. The eigenvalues $\lambda^i_0(\theta_0)$ are positive real numbers with the smallest one $\lambda^{\min}_0(\theta_0)$ denoted by

$$\lambda^{\min}_0(\theta_0) = \min_{1\le i\le d} \lambda^i_0(\theta_0) > 0,$$

satisfying $g^T H_0(\theta_0) g \ge \lambda^{\min}_0(\theta_0) g^T g$ for all $g \in \mathbb{R}^d \setminus \{0\}$.

To shorten the notation, we write with no ambiguity $H_0(\theta_0) \succeq \lambda^{\min}_0(\theta_0) I_d$ where I_d denotes the d-dimensional identity matrix. By continuity, $\lambda^{\min}_0(\theta)$ is positive on a neighborhood $B(\theta_0, \delta)$ such there exist $\epsilon > 0$ satisfying $\lambda^{\min}_0(\theta_0) - \epsilon > 0$, meaning

$$H_0(\theta) \succeq (\lambda^{\min}_0(\theta_0) - \epsilon) I_d,$$

for $\theta \in B(\theta_0, \delta)$. Hence, for $\theta \in B(\theta_0, \delta)$ and $g \in \mathbb{R}^d \setminus \{0\}$, we have

$$\begin{aligned}
\frac{g^T \widehat{H}_n(\theta)}{g^T g} &= \frac{g^T H_0(\theta) g}{g^T g} + \frac{g^T \left(\widehat{H}_n(\theta) - H_0(\theta)\right) g}{g^T g} \\
&\ge \lambda_{\min} - \epsilon - \frac{g^T \|\widehat{H}_n(\theta) - H_0(\theta)\|_{op} g}{g^T g} \\
&> \lambda_{\min} - 3\epsilon \\
&> C, \quad \text{a.s.},
\end{aligned}$$

using (4.6.27) for all $n \ge N$ by taking $0 < \epsilon < 6^{-1}\lambda_{\min}$ and letting $C = 2^{-1}\lambda_{\min}$. Then we have the desired inequality (4.2.7). □

Proof of Corollary 4.2.1. The uniqueness of the QMLE $\widehat{\theta}^*_n$ follows from a Pfanzagl argument ([116]). By Theorem 4.2.1, we know there exists N such that

$$\inf_{\theta\in B(\theta_0,\delta_0)} g^T \widehat{H}_n(\theta) g > C g^T g, \quad \text{a.s.},$$

for all $n \ge N$ where $B(\theta_0, \delta_0)$ denotes the open ball around θ_0 with radius $\delta_0 > 0$. For each element $\theta_i \in \mathcal{K}$, we make an open ball $B(\theta_i, \delta_i)$ for $\delta_i > 0$ such that the union of $B(\theta_i, \delta_i)$ for all i only contains θ_0 once, i.e., $\theta_0 \notin B(\theta_i, \delta_i)$ for $i \ne 0$. As $\mathcal{K}$ is compact and contained in the union of all $B(\theta_i, \delta_i)$, then there is a finite covering of $\mathcal{K}$, i.e., $\mathcal{K} \subseteq \bigcup_{i=0}^k B(\theta_i, \delta_i)$. Let $\mathcal{K}' = \mathcal{K} \setminus B(\theta_0, \delta_0)$. As $\mathcal{K}'$

is compact, the minimum of the continuous QL function $\mathbb{E}[l_0]$ exists. Moreover, as $\mathbb{E}[l_0]$ is a unique minimum at θ_0 under Assumption 4.2.1, we get

$$\inf_{\theta \in \mathcal{K}'} \mathbb{E}[l_0(\theta)] > \mathbb{E}[l_0(\theta_0)], \quad \text{a.s.}$$

From Assumption 4.2.2, we know that $\|n^{-1}\widehat{L}_n - L_0\|_{\mathcal{K}'} \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$. Hence, we have

$$\inf_{\theta \in \mathcal{K}'} n^{-1}\widehat{L}_n(\theta) \xrightarrow{\text{a.s.}} \inf_{\theta \in \mathcal{K}'} L_0(\theta),$$

where $\inf_{\theta \in \mathcal{K}'} L_0(\theta) > \mathbb{E}[l_0(\theta_0)]$. Thus, the $B(\theta_0, \delta_0)$ gives us a unique global minimum of the QL function $\widehat{L}_n$, i.e.,

$$\inf_{\theta \in \mathcal{K}} n^{-1}\widehat{L}_n(\theta) \geq \mathbb{E}[l_0(\theta_0)], \quad \text{a.s.},$$

where equality only is attained when $\theta = \theta_0$. □

4.7 Relative Speed Comparison

It is argued that the recursive procedure AdaVol is computationally advantageous as it only processes observations once. In order to illustrate this advantage, a relative computational speed comparison as in [139] is presented. The code is not optimized; it is solely for illustration purposes. In the streaming data framework, the parameters are estimated recursively as described in Section 4.4. Meaning, for each t, the iterative estimate $\widetilde{\theta}_t$ is estimated using the observations $(X_i)_{1 \leq i \leq t}$ and the previous iterative estimate $\widetilde{\theta}_{t-1}$ as initialization.

An ARCH(1), GARCH(1, 1), and GARCH(2, 2) model is considered for the computational speed analysis. Table 4.3 shows the relative speed comparison for these models with sample sizes $n = 1000$ and $n = 2000$. The overall conclusion is that the AdaVol procedure is faster than the iterative one, e.g., the iterative estimation of a GARCH(1, 1) model is about 205 times slower. Another important observation is the relative speed for different sample sizes n, namely, the larger the sample size n, the greater the relative speed gain is for the AdaVol procedure.

Model	n	AdaVol	`arch`
ARCH(1)	1000	1.00	163.64
	2000	1.00	190.12
GARCH(1, 1)	1000	1.00	204.89
	2000	1.00	233.86
GARCH(2, 2)	1000	1.00	322.33
	2000	1.00	328.50

Table 4.3: Relative speed comparison between AdaVol ([149]) and `arch` version 4.15 ([134]). A value of 1.00 means the method is the fastest. A value of 163.64 means the estimation time of the method is 163.64 times larger than the fastest.

Conclusion and Future Perspectives

The central theme of this thesis was to learn from time-dependent streaming data. We examined the robustness and convergence guarantees of SG-based methods under different settings, covering many applications with dependence and biased gradients. Our analysis explored convergence rates of the stochastic streaming algorithms in a non-asymptotic manner. The theoretical results formed heuristics that links the level of dependency and convexity to the rest of the model parameters. These heuristics provided new insights into determining optimal learning rates, which can help increase the stability of SG-based methods. Roughly speaking, SG-based methods brooked short-term and even long-term dependence by using non-decreasing batch sizes, which counteracted the dependency structures. In particular, we showed that mini-batch is essential to break dependence and ensure convexity. In addition, we can accelerate convergence by simultaneously averaging. Our experimentation verified these investigations suggesting large streaming batches with slow decaying learning rates for highly dependent data sources. Moreover, in large-scale learning problems with dependence, noisy variables, and lack of convexity, we know how to achieve (and accelerate) convergence and reduce noise through the learning rate and the treatment pattern of the data.

Future perspectives. There are several ways to expand our work about stochastic algorithms: (a) we can extend our analysis to include streaming batches of any size (and not as a function of streaming batch size C_ρ and streaming rates ρ). (b) an extension to non-strongly convex objectives could be advantageous as it will provide more insight into how we should choose our learning rates [9, 50, 99, 101]. (c) learning rates should be made adaptive so they are robust to poor initialization and require less tuning; an adaptive learning rate is essential for practitioners as it builds a form of universality across applications, e.g., see [42, 83]. (d) non-parametric analysis could improve our theoretical results for large values of d. (e) we have focused on results in quadratic mean but another way to strengthen our non-asymptotic guarantees could be high probability bounds [44]; for any $\delta \in (0, 1)$, we could obtain bounds on the sequence $\{\|\theta_t - \theta^*\| : t \in \mathbb{N}\}$ that holds with probability at least $1 - \delta$. (f) concerning AdaVol, we need to prove that Adavol works in theory, both under stationarity and non-stationarity, which should also include the proof of Polyak-Ruppert averaging (under stationarity). (g) at last, it could be interesting to generalize our work to other time series models but especially to the multivariate case [91, 114].

Bibliography

[1] Abu-Mostafa, Y. S., Magdon-Ismail, M., and Lin, H.-T. (2012). *Learning from data*, volume 4. AMLBook New York.

[2] Adams, S., Beling, P., and Cogill, R. (2016). Feature selection for hidden markov models and hidden semi-markov models. *IEEE Access*, 4:1–1.

[3] Agarwal, A. and Duchi, J. C. (2012). The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587.

[4] Ajalloeian, A. and Stich, S. U. (2020). On the convergence of sgd with biased gradients. *arXiv preprint arXiv:2008.00051*.

[5] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

[6] Aknouche, A. and Guerbyenne, H. (2006). Recursive estimation of garch models. *Communications in Statistics - Simulation and Computation*, 35(4):925–938.

[7] Anava, O., Hazan, E., Mannor, S., and Shamir, O. (2013). Online learning for time series prediction. In *Conference on learning theory*, pages 172–184. PMLR.

[8] Ang, A. and Timmermann, A. (2012). Regime changes and financial markets. *Annual Review of Financial Economics*, 4:313–337.

[9] Bach, F. and Moulines, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). *Advances in neural information processing systems*, 26.

[10] Baum, L. E. and Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bull. Amer. Math. Soc.*, 73(3):360–363.

[11] Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Statist.*, 41(1):164–171.

[12] Baum, L. E. and Sell, G. R. (1968). Growth transformations for functions on manifolds. *Pacific J. Math.*, 27(2):211–227.

[13] Benveniste, A., Métivier, M., and Priouret, P. (2012). *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media.

[14] Berkes, I., Horváth, L., and Kokoszka, P. (2003). GARCH processes: structure and estimation. *Bernoulli*, 9(2):201–227.

[15] Bertsekas, D. (2016). *Nonlinear Programming*, volume 4. Athena Scientific.

[16] Biau, G. and Patra, B. (2011). Sequential quantile prediction of time series. *Information Theory, IEEE Transactions on*, 57:1664 – 1674.

[17] Bilmes, J. A. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models.

[18] Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, 31(3):307–327.

[19] Bottou, L. and Bousquet, O. (2007). The tradeoffs of large scale learning. *Advances in neural information processing systems*, 20.

[20] Bottou, L. and Cun, Y. L. (2003). Large scale online learning. *Advances in neural information processing systems*, 16.

[21] Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311.

[22] Bougerol, P. (1993). Kalman filtering with random coefficients and contractions. *SIAM Journal on Control and Optimization*, 31(4):942–959.

[23] Bougerol, P. and Picard, N. (1992). Stationarity of GARCH processes and of some nonnegative time series. *Journal of Econometrics*, 52(1-2):115–127.

[24] Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526.

[25] Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

[26] Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[27] Boyer, C. and Godichon-Baggioni, A. (2020). On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. *arXiv preprint arXiv:2011.09706*.

[28] Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52:345–370.

[29] Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probability surveys*, 2:107–144.

[30] Brockwell, P. J. and Davis, R. A. (2009). *Time series: theory and methods*. Springer Science & Business Media.

[31] Bubeck, S. et al. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357.

[32] Cardot, H., Cénac, P., and Monnez, J.-M. (2012). A fast and recursive algorithm for clustering large datasets with k-medians. *Computational Statistics & Data Analysis*, 56(6):1434–1449.

[33] Cardot, H., Cénac, P., and Zitt, P.-A. (2013). Efficient and fast estimation of the geometric median in hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19(1):18–43.

[34] Cesa-Bianchi, N., Conconi, A., and Gentile, C. (2004). On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057.

[35] Cipra, T. and Hendrych, R. (2018). Robust recursive estimation of garch models. *Kybernetika -Praha*, 54:1138–1155.

[36] Dahlhaus, R. and Subba Rao, S. (2007). A recursive online algorithm for the estimation of time-varying arch parameters. *Bernoulli*, 13(2):389–422.

[37] d'Aspremont, A. (2008). Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183.

[38] Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27.

[39] Défossez, A. and Bach, F. (2017). Adabatch: Efficient gradient aggregation rules for sequential and parallel stochastic gradient methods. *arXiv preprint arXiv:1711.01761*.

[40] Devolder, O. et al. (2011). Stochastic first order methods in smooth convex optimization. Technical report, CORE.

[41] Doukhan, P. (2012). *Mixing: properties and examples*, volume 85. Springer Science & Business Media.

[42] Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

[43] Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (2008). Efficient projections onto the l1-ball for learning in high dimensions. *Proceedings of the 25th International Conference on Machine Learning*, pages 272–279.

[44] Durmus, A., Moulines, E., Naumov, A., Samsonov, S., Scaman, K., and Wai, H.-T. (2021). Tight high probability bounds for linear stochastic approximation with fixed stepsize. *Advances in Neural Information Processing Systems*, 34.

[45] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the econometric society*, pages 987–1007.

[46] Fons, E., Dawson, P., Yau, J., jun Zeng, X., and Keane, J. (2021). A novel dynamic asset allocation system using feature saliency hidden markov models for smart beta investing. *Expert Systems with Applications*, 163:113720.

[47] Ford, J. and Moore, J. (1998). Adaptive estimation of hmm transition probabilities. *IEEE Transactions on Signal Processing*, 46:1374–1385.

[48] Francq, C. and Zakoïan, J.-M. (2004). Maximum likelihood estimation of pure garch and arma-garch processes. *Bernoulli*, 10(4):605–637.

[49] Francq, C., Zakoïan, J.-M., and Horvath, L. (2011). Merits and drawbacks of variance targeting in garch models. *Journal of Financial Econometrics*, 9:619–656.

[50] Gadat, S. and Panloup, F. (2017). Optimal non-asymptotic bound of the ruppert-polyak averaging without strong convexity. *arXiv preprint arXiv:1709.03342*.

[51] Gauvain, J.-L. and Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298.

[52] Gerencsér, L., Orlovits, Z., and Torma, B. (2010). Recursive estimation of garch processes. In *The 19th International Symposium on Mathematical Theory of Networks and Systems, (MTNS 2010), Budapest, Hungary, forthcoming*, pages 2415–2422.

[53] Gervini, D. (2008). Robust functional estimation using the median and spherical principal components. *Biometrika*, 95(3):587–600.

[54] Ghadimi, S. and Lan, G. (2012). Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492.

[55] Godichon-Baggioni, A. (2016). Estimating the geometric median in hilbert spaces with stochastic gradient algorithms: Lp and almost sure rates of convergence. *Journal of Multivariate Analysis*, 146:209–222.

[56] Godichon-Baggioni, A. and Portier, B. (2017). An averaged projected robbins-monro algorithm for estimating the parameters of a truncated spherical distribution. *Electronic Journal of Statistics*, 11(1):1890–1927.

[57] Godichon-Baggioni, A., Werge, N., and Wintenberger, O. (2021). Non-asymptotic analysis of stochastic approximation algorithms for streaming data. *arXiv preprint arXiv:2109.07117*.

[58] Godichon-Baggioni, A., Werge, N., and Wintenberger, O. (2022). Learning from time-dependent streaming data with online stochastic algorithms. *arXiv preprint arXiv:2205.12549*.

[59] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

[60] Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). Sgd: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR.

[61] Guidolin, M. and Timmermann, A. (2007a). Asset allocation under multivariate regime switching. *Journal of Economic Dynamics and Control*, 31(11):3503–3544.

[62] Guidolin, M. and Timmermann, A. (2007b). Size and value anomalies under regime shifts. *Journal of Financial Econometrics*, 6:1–48.

[63] Gupta, A. and Dhingra, B. (2012). Stock market prediction using hidden markov models. *2012 Students Conference on Engineering and Systems, SCES 2012*, pages 1–4.

[64] Haldane, J. (1948). Note on the median of a multivariate distribution. *Biometrika*, 35(3-4):414–417.

[65] Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57:357–384.

[66] Hamilton, J. D. (2020). *Time series analysis*. Princeton university press.

[67] Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):190–195.

[68] Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR.

[69] Hassan, M. and Nath, B. (2005). Stock market forecasting using hidden markov model: A new approach. *Proceedings - 5th International Conference on Intelligent Systems Design and Applications 2005, ISDA '05*, 2005:192–196.

[70] Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

[71] Hazan, E. et al. (2016). Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325.

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

[73] Hendrych, R. and Cipra, T. (2018). Self-weighted recursive estimation of garch models. *Communications in Statistics - Simulation and Computation*, 47(2):315–328.

[74] Hinton, G., Srivastava, N., and Swersky, K. (2012). Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2.

[75] Ip, W.-C., Wong, H., Pan, J., and Li, D. (2006). The asymptotic convexity of the negative likelihood function of garch models. *Computational Statistics & Data Analysis*, 50:311–331.

[76] Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26.

[77] Jothimurugesan, E., Tahmasbi, A., Gibbons, P., and Tirthapura, S. (2018). Variance-reduced stochastic gradient descent on streaming data. *Advances in neural information processing systems*, 31.

[78] Juditsky, A., Nemirovski, A., et al. (2011). First order methods for nonsmooth convex large-scale optimization, i: general purpose methods. *Optimization for Machine Learning*, 30(9):121–148.

[79] Karimi, B., Miasojedow, B., Moulines, E., and Wai, H.-T. (2019). Non-asymptotic analysis of biased stochastic approximation scheme. In *Conference on Learning Theory*, pages 1944–1974. PMLR.

[80] Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer.

[81] Kemperman, J. (1987). The median of a finite measure on a banach space. *Statistical data analysis based on the L1-norm and related methods (Neuchâtel, 1987)*, pages 217–230.

[82] Kierkegaard, J., Jensen, L., and Madsen, H. (2000). Estimating garch models using recursive methods.

[83] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[84] Koenker, R. W. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

[85] Kritzman, M., Page, S., and Turkington, D. (2012). Regime shifts: Implications for dynamic strategies. *Financial Analysts Journal*, 68.

[86] Kurdyka, K. (1998). On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, 48(3):769–783.

[87] Kushner, H. J. and Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. Springer-Verlag.

[88] Lacoste-Julien, S., Schmidt, M., and Bach, F. (2012). A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*.

[89] Lan, G. (2012). An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397.

[90] Lan, G. (2020). *First-order and stochastic optimization methods for machine learning*. Springer.

[91] Laurent, S., Rombouts, J. V., and Violante, F. (2012). On the forecasting accuracy of multivariate garch models. *Journal of Applied Econometrics*, 27(6):934–955.

[92] Liu, C., Hoi, S. C., Zhao, P., and Sun, J. (2016). Online arima algorithms for time series prediction. In *Thirtieth AAAI conference on artificial intelligence*.

[93] Lojasiewicz, S. (1963). A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2.

[94] Mohri, M. and Rostamizadeh, A. (2010). Stability bounds for stationary φ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11(2).

[95] Mokkadem, A. and Pelletier, M. (2011). A generalization of the averaging procedure: The use of two-time-scale algorithms. *SIAM Journal on Control and Optimization*, 49(4):1523–1543.

[96] Moulines, E. and Bach, F. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24.

[97] Murata, N. and Amari, S.-i. (1999). Statistical analysis of learning dynamics. *Signal Processing*, 74(1):3–28.

[98] Murphy, K. P. (2013). *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.].

[99] Necoara, I., Nesterov, Y., and Glineur, F. (2019). Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1):69–107.

[100] Nelson, D. (1990). Stationarity and persistence in the garch(1,1) model. *Econometric Theory*, 6:318–334.

[101] Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609.

[102] Nemirovskij, A. S. and Yudin, D. B. (1983). *Problem complexity and method efficiency in optimization*. Wiley-Interscience.

[103] Nesterov, Y. (1983). A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2). In *Doklady an ussr*, volume 269, pages 543–547.

[104] Nesterov, Y. et al. (2018). *Lectures on convex optimization*, volume 137. Springer.

[105] Nesterov, Y. and Polyak, B. T. (2006). Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205.

[106] Nesterov, Y. and Spokoiny, V. (2017). Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566.

[107] Nguyen, L., Nguyen, P. H., Dijk, M., Richtárik, P., Scheinberg, K., and Takác, M. (2018). Sgd and hogwild! convergence without the bounded gradients assumption. In *International Conference on Machine Learning*, pages 3750–3758. PMLR.

[108] Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. (2017). Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR.

[109] Nguyen, N. (2018). Hidden markov model for stock trading. *International Journal of Financial Studies*, 6(2):1–17.

[110] Nocedal, J. and Wright, S. J. (1999). *Numerical optimization*. Springer.

[111] Nystrup, P., Madsen, H., and Lindström, E. (2015). Stylised facts of financial time series and hidden markov models in continuous time. *Quantitative Finance*, 15(9):1531–1541.

[112] Nystrup, P., Madsen, H., and Lindström, E. (2017). Long memory of financial time series and hidden markov models with time-varying parameters. *Journal of Forecasting*, 36(8):989–1002.

[113] Patton, A. (2006). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160:246–256.

[114] Pedersen, R. S. and Rahbek, A. (2014). Multivariate variance targeting in the bekk–garch model. *The Econometrics Journal*, 17(1):24–55.

[115] Pelletier, M. (1998). On the almost sure asymptotic behaviour of stochastic algorithms. *Stochastic processes and their applications*, 78(2):217–244.

[116] Pfanzagl, J. (1969). On the measurability and consistency of minimum contrast estimates. *Metrika*, 14:249–272.

[117] Polyak, B. T. (1963). Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653.

[118] Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855.

[119] Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151.

[120] Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286.

[121] Rakhlin, A., Shamir, O., and Sridharan, K. (2011). Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*.

[122] Reddi, S. J., Kale, S., and Kumar, S. (2019). On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*.

[123] Rio, E. (2017). *Asymptotic theory of weakly dependent random processes*, volume 80. Springer.

[124] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

[125] Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America*, 42(1):43.

[126] Roux, N., Schmidt, M., and Bach, F. (2012). A stochastic gradient method with an exponential convergence _rate for finite training sets. *Advances in neural information processing systems*, 25.

[127] Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

[128] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

[129] Ruppert, D. (1988). Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering.

[130] Schmidt, M., Roux, N., and Bach, F. (2011). Convergence rates of inexact proximal-gradient methods for convex optimization. *Advances in neural information processing systems*, 24.

[131] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

[132] Shalev-Shwartz, S. et al. (2012). Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194.

[133] Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A. (2011). Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30.

[134] Sheppard, K. (2020). bashtage/arch:. Release 4.15 (version 4.15), Zenodo.

[135] Sridharan, K., Shalev-Shwartz, S., and Srebro, N. (2008). Fast rates for regularized objectives. *Advances in neural information processing systems*, 21.

[136] Steinwart, I. and Christmann, A. (2011). Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225.

[137] Straumann, D. (2005). Maximum likelihood estimation in conditionally heteroscedastic time series models. *Estimation in Conditionally Heteroscedastic Time Series Models*, pages 141–168.

[138] Straumann, D. and Mikosch, T. (2006). Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. *Annals of Statistics*, 34(5):2449–2495.

[139] Sucarrat, G. (2020). garchx: Flexible and Robust GARCH-X Modelling. MPRA Paper 100301, University Library of Munich, Germany.

[140] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

[141] Teo, C. H., Smola, A., Vishwanathan, S., and Le, Q. V. (2007). A scalable modular convex solver for regularized risk minimization. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 727–736.

[142] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

[143] Tieleman, T., Hinton, G., et al. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31.

[144] Toulis, P., Tran, D., and Airoldi, E. (2016). Towards stability and optimality in stochastic gradient descent. In *Artificial Intelligence and Statistics*, pages 1290–1298. PMLR.

[145] Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.

[146] Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.

[147] Ward, R., Wu, X., and Bottou, L. (2018). Adagrad stepsizes: Sharp convergence over non-convex landscapes, from any initialization.

[148] Weiszfeld, E. and Plastria, F. (2009). On the point for which the sum of the distances to n given points is minimum. *Annals of Operations Research*, 167(1):7–41.

[149] Werge, N. (2019). Adavol. *GitHub repository*.

[150] Werge, N. and Wintenberger, O. (2022). Adavol: An adaptive recursive volatility prediction method. *Econometrics and Statistics*, 23:19–35.

[151] Wintenberger, O. (2013). Continuous invertibility and stable qml estimation of the egarch(1,1) model. *Scandinavian Journal of Statistics*, 40(4):846–867.

[152] Wintenberger, O. (2021). Stochastic online convex optimization; application to probabilistic time series forecasting. *arXiv preprint arXiv:2102.00729*.

[153] Xiao, L. (2009). Dual averaging method for regularized stochastic learning and online optimization. *Advances in Neural Information Processing Systems*, 22.

[154] Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116.

[155] Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

[156] Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116.

[157] Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936.

[158] Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. New York: Chapman and Hall/CRC.

[159] Zumbach, G. (2000). The pitfalls in fitting garch (1, 1) processes. In *Advances in Quantitative Asset Management*, pages 179–200. Springer.

Appendix A: Predicting Risk-adjusted Returns using an Asset Independent Regime-switching Model

Abstract

Financial markets tend to switch between various market regimes over time, making stationarity-based models unsustainable. We construct a regime-switching model independent of asset classes for risk-adjusted return predictions based on hidden Markov models. This framework can distinguish between market regimes in a wide range of financial markets such as the commodity, currency, stock, and fixed income market. The proposed method employs sticky features that directly affect the regime stickiness and thereby changing turnover levels. An investigation of our metric for risk-adjusted return predictions is conducted by analyzing daily financial market changes for almost twenty years. Empirical demonstrations of out-of-sample observations obtain an accurate detection of bull, bear, and high volatility periods, improving risk-adjusted returns while keeping a preferable turnover level.

keywords: *hidden Markov model, financial time series, non-stationary, regime-switching, prediction markets, trading strategies*

A.1 Introduction

Financial markets are known to shift between economic cycles; some of the most well-known regimes are the bull, bear, and high-volatility markets. Each of these market regimes may have financial characteristics unique to this particular regime. One of the most common methods of financial market analysis is time series analysis. Time series models are used to predict future prices, price changes, and volatilities in a wide range of financial markets. Some of the most famous models are the AutoRegressive Integrated Moving Average (ARIMA) models. However, analyzing financial time series through these traditional time series methods may result in misleading resolutions as they cannot embrace the nonlinear characteristics of financial time series, e.g., the stationarity assumption often seems dubious in practice. Therefore, non-stationary-based time series models are more suitable for financial time series. One could comprehend this by modifying these time series models by incorporating a time-dependent variable to adjust for the non-stationarity, e.g., the threshold autoregressive time series model.

Another way to capture financial markets' tendency to switch between regimes is the Hidden Markov Model (HMM), as it "only" assumes local or state-conditioned stationarity. Modeling times series data using HMMs became mainstream after [11] and [120] applied it across many areas (e.g., speech recognition, medical applications, and text classification). The idea of making a Markov-switching approach to analyze financial time series became popular after [65] applied this approach to identify economic cycles of GNP levels. More recently, the HMM has been used to predict market regimes in the financial markets due to their ability to capture multiple characteristics from financial return series such as time-varying correlations, fat tails, volatility clustering, skewness, and kurtosis, while also providing reasonable approximations even for processes in which the underlying model is unknown ([8, 111, 112]). Besides, HMMs are advantageous as they allow ample interpretability of the results; thinking in market regimes is a natural approach for financial practitioners. Nevertheless, the lack of data availability makes the linking between investment purposes and business cycles a complex and challenging task. As the market regimes are not observable, one has to extract them from the time series. However, this extraction is not unambiguous, as some specific regimes may be up for discussion in the financial practitioner's community, e.g., high and low volatility regimes depend on the given risk-aversion. Consequently, we demand a model to apprehend the various economic sentiments of the financial markets.

Many researchers have applied HMMs to analyze and predict economic (non-linear) trends and future financial asset prices. [85] studied an HMM with two states to predict regimes in market turbulence, inflation, and economic growth index. [69] and [109] used the HMM to forecast prices in the stock market. A combination of open, close, low, and high prices was used in [63] for stock price prediction. All of the above references use four hidden states in their study on the stock market. [61] and [46] used a four-state and two-state HMM, respectively, in their studies of asset allocation decisions using various time series. As suggested by [62], a range between two and four hidden states in the HMM is often encountered in financial studies. However, studies of applying HMMs to predict trends across a broad range of assets are sparse.

In this study, we focus on predicting risk-adjusted returns using a single regime-switching model. Using only one HMM to analyze a wide range of assets, we enforce generalizations in the model. This framework is made with so-called "sticky" features that naturally enhance regime stickiness by an adjustable hyperparameter. Finally, we demonstrate our methodology on a broad range of asset classes by analyzing daily financial market changes for almost twenty years. The investigation illustrates our metric ability to predict risk-adjusted returns for different regime stickiness choices. Our experiments are conducted using out-of-sample observations, showing an accurate detection of bull, bear, and high volatility periods, improving risk-adjusted returns while keeping a preferable turnover level.

A.2 Hidden Markov Models (HMMs)

There is much literature about HMMs, but to have the necessary notions, we briefly sketch the elements of the HMM, how to estimate the parameters, and select the number of hidden states in the HMMs. For a comprehensive introduction of the inference of HMMs, we refer to [158] and [98].

A.2.1 Elements of HMM

The HMM is a probabilistic model in which a sequence of observations $x = (x_1, \dots, x_n)$ with $x_t \in \mathbb{R}^d$ for $t = 1, \dots, n$ is generated by a latent finite-state Markov chain $z = (z_1, \dots, z_n)$. Denote by d the dimension of the observations. We call z the sequence of hidden states where $z_t \in \{1, \dots, S\}$ for $t = 1, \dots, n$ with S the number of hidden states. The HMM can be specified by the initial probability vector $\boldsymbol{\pi} = \{\pi_i\}_{i=1,\dots S} \in \mathbb{R}^S$, a transition probability matrix $\mathbf{A} = \{A_{ij}\}_{i,j=1,\dots,S} \in \mathbb{R}^{S\times S}$ and the emission probabilities $\mathbf{B}$ which can be any distribution conditioned on the current hidden state. The parameters of the HMM are given by $\Lambda = \{\boldsymbol{\pi}, \mathbf{A}, \mathbf{B}\}$ and have to be estimated from the observed sequence x. Note that $\pi_i = \mathbb{P}(z_1 = i)$ is the probability for being in hidden state i at time $t = 1$ where $\sum_{i=1}^S \pi_i = 1$, $A_{ij} = \mathbb{P}(z_t = j | z_{t-1} = i)$ is the transition probability of moving from hidden state i at time $t-1$ to hidden state j at time t with $\sum_{j=1}^S A_{ij} = 1$, and $\mathbf{B}$ is the parameters of the conditional densities $p(x_t|z_t = j)$.

When working with financial time series, a typical choice of emission probabilities is the Gaussian Mixture Model (GMM). However, other density functions could likewise be considered. A gentle introduction of HMMs with GMM emissions is made in [17]. The authors of [8] and [111] show evidence on the HMMs ability to comprehend several stylized facts, such as leptokurtosis, heteroskedasticity, skewness, and time-varying correlations, by use of the GMM as emission probability. For simplicity, we assume the distribution of emission probabilities $\mathbf{B}$ to be Gaussian; $\mathbf{B} = p(x_t|z_t = j, \Lambda) = \mathcal{N}(x_t|\mu_j, \Sigma_j)$ where $\boldsymbol{\mu} = \{\mu_j\}_{j=1,\dots,S}$ is the mean vectors and $\boldsymbol{\Sigma} = \{\Sigma_j\}_{j=1,\dots,S}$ the co-variance matrices with $\mu_j \in \mathbb{R}^d$ and $\Sigma_j \in \mathbb{R}^{d\times d}$ for $j = 1, \dots, S$. Thus, the model parameters of our HMM is given as $\Lambda = \{\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$.

A.2.2 Parameter Estimation

There are three fundamental problems in estimating the HMM:

- Given the observations sequence $x = (x_1, \ldots, x_n)$ and HMM parameters $\Lambda = \{\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, how can we estimate $\mathbb{P}(x|\Lambda)$ the likelihood of the given observation sequence.
- Given the observations sequence $x = (x_1, \ldots, x_n)$ and model parameters $\Lambda = \{\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, how can we choose a sequence of hidden states $z = (z_1, \ldots, z_n)$, which is optimal.
- How do we adjust the HMM parameters $\Lambda = \{\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ to maximize $\mathbb{P}(x|\Lambda)$.

There are several approaches to solve these problems since there are several possible optimal criteria. We choose to solve the first and the second problem by the dynamic programming algorithms known as the forward-backward algorithm proposed by [10] and [12], and the Viterbi algorithm ([146]). The third problem is solved by the iterative Baum-Welch (BW) algorithm, a type of the Expectation-Maximization (EM) algorithm ([120]).

The BW algorithm alternates between an expectation step and a maximization step until convergence is reached, often abbreviated as the E-step and M-step. In the E-step, we calculate the expected log-likelihood of the hidden state given the observation sequence x and model parameters Λ. Next, in the M-step we maximize the expected log-likelihood from the E-step to update our model parameters Λ. We denote by $Q(\Lambda, \bar{\Lambda})$ the function of the expectation of the complete log-likelihood given as

$$Q(\Lambda, \bar{\Lambda}) = \mathbb{E}[\log \mathbb{P}(x, z|\Lambda)|x, \bar{\Lambda}], \tag{A.2.1}$$

where the current model is Λ and the previous model as $\bar{\Lambda}$.

It can be proven that $\mathbb{P}(x|\Lambda) \geq \mathbb{P}(x|\bar{\Lambda})$, but it is essential to remember that the BW algorithm does not guarantee a global solution. As suggested in [2] and [46], we modify the $Q(\Lambda, \bar{\Lambda})$ function with the priors of the model parameters $G(\Lambda)$, namely

$$Q(\Lambda, \bar{\Lambda}) + \log(G(\Lambda)), \tag{A.2.2}$$

which is called Maximum a Posteriori (MAP) estimation ([51]). Thus, in the E-step, we calculate the $Q(\Lambda, \bar{\Lambda})$ function from (A.2.1), and for the M-step, we maximize (A.2.2).

A.2.3 Prediction

The prediction of the hidden states sequence $(z_1, \ldots, z_n)$ is estimated using the observation sequence $(x_1, \ldots, x_n)$ as described in Section A.2.2. We denote by $\alpha_{n|n}$ the vector of state probabilities at time n (given the sequence of observations $x = (x_1, \ldots, x_n)$) with the jth entry $(\alpha_{n|n})_j = \mathbb{P}(z_n = j|x)$ for $j = 1, \ldots, S$. Thus, one can forecast the state probability $h \geq 0$ steps ahead by

$$\alpha_{n+h|n} = \alpha_{n|n}\mathbf{A}^h, \tag{A.2.3}$$

as the model parameters $\mathbf{A}$ are assumed to be constant over time.

A.2.4 Model Selection

A drawback of using the HMM is the necessity of knowing the number of hidden states in advance (such as the hyper-parameter k in the k-nearest neighbor algorithm and k-means clustering). There are several criteria used for this model selection: the lazy approach is to use statistical criteria such as the Akaike's Information Criterion (AIC) by [5], Bayesian Information Criterion (BIC) by [131], Hannan-Quinn Information Criterion (HQIC) by [67], and Bozdogan Consistent Akaike Information Criterion (BCAIC) by [28]. These criteria are defined as follows:

$$\begin{aligned}
\text{AIC} &= -2\log(L) + 2p, \\
\text{BIC} &= -2\log(L) + p\log(n), \\
\text{HQIC} &= -2\log(L) + p\log(\log(n)), \\
\text{BCAIC} &= -2\log(L) + p(\log(n) + 1),
\end{aligned}$$

where $\log(L)$ is the log-likelihood of the model, n indicates the number of observations in the time series, and p denotes the number of independent parameters of the model. In the case of an HMM with GMM emissions, we have $p = S(S+cm)$, where S is the number of hidden states in the Markov chain of the model, m is the number of Gaussian mixtures, and c is the number of parameters of the underlying distribution of the observation process. Note that a d-dimensional multivariate Gaussian with full covariance matrix process has $c = d + d(d+1)/2$ parameters to estimate. Thus, an HMM with three hidden states ($S = 3$), a single 2-dimensional Gaussian process in each hidden state, has a total of 24 parameters.

Suppose one were to see the number of hidden states as the number of strategies we have to make to produce proper predictions. Then the number should be neither too small nor too large. If the number of hidden states is too small, then the risk of misclassification will increase. Too many hidden states will make the distinction between each hidden state vague and, therefore, increase the risk for overfitting and increase the computational cost. A similar observation can be made regarding the number of Gaussian mixture components.

However, if one wishes to maintain a high degree of interpretability of the hidden states in the model, we should keep the number of hidden states low. Another approach is the greedy approach, where we decide the number of hidden states in the HMM by constructing different portfolios based on HMMs with different numbers of hidden states and then select the number of hidden states associated with the portfolios of the best performance, e.g., evaluated by the Sharpe Ratio (SR). One should be aware that we may find different optimal numbers of states for each asset using these criteria.

A.3 Data

Our objective is to identify market regimes on various asset classes, namely commodity (CO), currency (FX), equity (EQ), and fixed income (FI). We consider $d = 15$ instruments defined as $I = (I_1, \ldots, I_{15})^T$, consisting of four different instruments per asset type, except for commodities where we have only three instruments. All instruments I are future contracts generated automatically by selecting the nearest contract. The data analyzed are closing returns of daily frequency from January 2000 to October 2019, consisting of $n = 4972$ observations (per instrument).

Table A.1 presents an overview of the performance of each asset. This confirms a high degree of variation of the considered asset; commodities and equities are the most volatile asset classes, whereas fixed income volatility is several times lower. Currencies appear to be in the middle of the levels we observe for equities and fixed income. Fixed income seems to be the most coherent asset class, whereas we find some large variations in returns, volatility, and maximum drawdown in commodities.

#	**Instr.**	**Ret.**	**Vol.**	**SR**	**DD**
1	CO1	6.36%	17.23%	0.45	16.81%
2	CO2	−20.19%	50.4%	−0.2	37.71%
3	CO3	5.46%	36.12%	0.33	24.13%
4	FX1	−0.83%	9.17%	−0.04	10.83%
5	FX2	−0.02%	9.58%	0.05	5.83%
6	FX3	−1.82%	9.79%	−0.14	8.29%
7	FX4	−0.98%	7.96%	−0.08	4.98%
8	EQ1	1.1%	23.34%	0.16	18.39%
9	EQ2	2.64%	18.1%	0.24	15.79%
10	EQ3	1.08%	24.61%	0.17	27.98%
11	EQ4	4.14%	18.67%	0.31	21.02%
12	FI1	3.73%	5.91%	0.66	5.02%
13	FI2	4.22%	5.24%	0.82	3.49%
14	FI3	3.65%	5.88%	0.65	4.19%
15	FI4	2.35%	3.05%	0.79	2.87%

Table A.1: Performance overview (annualized return, annualized volatility, Sharpe ratio, and maximum drawdown) of our instruments I evaluated from January 2000 to October 2019. Note $I \in \mathbb{R}^{n \times d}$ with $n = 4972$ and $d = 15$.

To further emphasize our instruments' diversity, we show the range (minimum; maximum) of the one-year rolling mean, standard deviation, skewness, and (excess) kurtosis in Table A.2. The instruments I show a considerable amount of variability, both within and across instrument types, with commodities showing the most variation and fixed income showing the least fluctuation. In particular, it is not abnormal that skewness exceeds one (in absolute terms), nor kurtosis is negative (platykurtic) or very positive (leptokurtic), e.g., CO2 have a kurtosis above thirty-five.

#	Instr.	Mean	Std.	Skew.	Kurt.
1	CO1	(-3.26;3.93)	(5.09;23.69)	(-3.63;2.20)	(-0.42;19.33)
2	CO2	(-8.94;12.6)	(14.01;58.96)	(-1.01;4.15)	(-0.64;35.64)
3	CO3	(-13.48;5.33)	(8.61;56.73)	(-2.54;1.79)	(-0.69;11.50)
4	FX1	(-3.12;1.6)	(2.7;13.02)	(-4.18;0.97)	(-0.66;25.85)
5	FX2	(-2.37;1.57)	(2.35;12.07)	(-1.36;1.10)	(-0.61;6.59)
6	FX3	(-1.89;2.12)	(2.67;11.58)	(-2.29;2.65)	(-0.36;13.96)
7	FX4	(-1.35;1.92)	(2.07;9.87)	(-0.88;1.30)	(-0.70;4.62)
8	EQ1	(-5.55;2.61)	(5.84;35.71)	(-2.01;1.90)	(-0.42;13.29)
9	EQ2	(-5.03;2.41)	(4.66;32.84)	(-1.41;1.21)	(-0.56;5.26)
10	EQ3	(-8.62;4.55)	(6.47;49.24)	(-2.61;1.25)	(-0.80;13.03)
11	EQ4	(-5.77;2.16)	(3.64;39.59)	(-4.05;1.49)	(-0.43;22.06)
12	FI1	(-0.86;1.8)	(1.66;8.11)	(-1.40;1.39)	(-0.75;6.52)
13	FI2	(-0.62;1.34)	(1.97;6.25)	(-1.32;0.70)	(-0.77;3.83)
14	FI3	(-0.81;1.4)	(1.94;7.3)	(-0.90;1.60)	(-0.70;6.92)
15	FI4	(-0.74;0.58)	(0.48;4.49)	(-4.34;2.24)	(-0.49;26.33)

Table A.2: Range (min; max) of one-year rolling mean, standard deviation, skewness and (excess) kurtosis of instruments I. Rolling mean and standard deviation are scaled by 10^3.

A.4 Feature Engineering

A.4.1 Exponential Weighted Moving Moments

When the underlying parameters are believed to follow a random walk, it is natural to use exponential forgetting. One of the most popular methods for calculating moments is the Exponential Weighted Moving Moment (EWMM) method, which is applied extensively in many different fields due to its computational efficiency. This EWMM method is often used to reduce noisy time-series data, also called "smoothing" the data. We can define the EWMM_t^i of order $i \in \mathbb{N}$ at time t by

$$\text{EWMM}_t^i = \lambda \text{M}_t^i + (1-\lambda)\text{EWMM}_{t-1}^i,$$

where $\lambda = \frac{2}{s+1}$ with $s \in \mathbb{N}$ defined as the span. For daily data, letting our span $s = 5$ would correspond to a half-life of 5 days. The choice of s can be seen as a smoothing factor where high (low) values of s would mean a high (low) degree of smoothing our time series. Using this method to calculate the well-known exponential weighted moving average of observations $(x_1, \ldots, x_n)$ is done by letting $\text{M}_t^1 = x_t$ for $t = 1, \ldots, n$. Furthermore, setting $s = 2t - 1$ would give us the usual average estimate. Hence, there is a trade-off between the sensitivity to noise and its ability to adapt to parameter changes.

A.4.2 Feature Extraction

Our interest is to predict risk-adjusted returns, where we incorporate an adjustable hyperparameter that changes the stickiness of the regimes. We extract the features of our instruments I

according to the description of EWMMs in Section A.4.1. Denote our features for the first and second moment by $(f_s^i)_{i=1,2} = (\text{EWMM}_t^i)_{i=1,2}$, where s denotes the feature span. All features $(f_s^i)_{i=1,2}$ are normalized to zero mean and unit variance using a z-score normalization fitted on the training data. After normalization, we concatenate our features depending on the moment's order into one feature before passing it onto our HMM. Thus, our complete features space is $f_s = (f_s^1, f_s^2)$.

The span s in our features f_s will work as a smoothing factor and determine the frequency of regime shifts, namely the regime stickiness. The larger we make our smoothing factor s, the slower our features f_s would change, making our hidden states more sticky, i.e., large diagonal values in the transition matrix $\mathbf{A}$ (See Section A.2.1). Thus, portfolio turnover will decrease.

There are different approaches in the literature on how to deal with this increased noise of hidden state prediction; the authors of [63] use the notion of latency days, in which they forecast the hidden states at time $n+1$ using only the ten previous days of observations. Others detect a regime change by considering the number of consecutive days in the same new hidden state, given a rolling window of days (which one has to estimate/select). Intuitively, smaller window sizes will lead to a larger number of regime changes, whereas large window sizes will increase regimes' length. Putting into an economic scenario, one would like to find a window size according to the preferences for turnover adjusted for transaction costs.

A.4.3 Prediction of Expected SR

The unsupervised classification computed by the HMM using our features $f_s = (f_s^1, f_s^2)$ results in some mean and variance estimates of every feature in each hidden state S. We aim to combine these resulting mean and variance estimates into a self-explanatory financial metric that reflects the underlying risk-adjusted returns.

Before defining the risk-adjusted return metric we need to introduce the following notions: let $\boldsymbol{\mu} = \{\mu_j\}_{j=1,\dots,S}$ denote the mean vectors and $\boldsymbol{\Sigma} = \{\Sigma_j\}_{j=1,\dots,S}$ the co-variance matrices with $\mu_j = (\mu_j(f_s^1), \mu_j(f_s^2))^T \in \mathbb{R}^2$ and $\Sigma_j = \Sigma_j(f_s^1, f_s^2) \in \mathbb{R}^{2\times 2}$ for $j = 1, \dots, S$. Thus, by dividing our mean estimate of our first moment by the mean estimate of the second moment at each hidden state, we have an Expected SR (ESR) in each hidden state called ESR_s^j. Meaning, for each hidden state $j \in \{1, \dots, S\}$, then $\text{ESR}_s^j = \mu_j(f_s^1)/\mu_j(f_s^2) \in \mathbb{R}$. We denote by $\mathbf{ESR}_s^S$ the vector $(\text{ESR}_s^1, \dots, \text{ESR}_s^S)^T \in \mathbb{R}^S$, where S is the number of hidden states in the HMM and s the span used to calculate our features.

We can use our $\mathbf{ESR}_s^S$ metric to predict an expected SR $h \geq 0$ steps ahead by combining this with the estimated vector of state probabilities α and the transition matrix $\mathbf{A}$. Recall from (A.2.3) that $\alpha_{n+h|n} = \alpha_{n|n}\mathbf{A}^h$, where $\alpha_{n|n}$ is the vector of state probabilities at time n and $\mathbf{A}$ the transition matrix (given the sequence of observations $(x_1, \dots, x_n)$) with the jth entry $(\alpha_{n|n})_j = \mathbb{P}(z_n = j|x)$ for $j = 1, \dots, S$. Hence, we can define the predicted ESR (PESR) metric by the product of

$$\text{PESR}_s^S(h) = (\mathbf{ESR}_s^S)^T \alpha_{n+h|n}, \tag{A.4.4}$$

where $h \geq 0$ and $\text{PESR}_s^S(h) \in \mathbb{R}$. This $\text{PESR}_s^S(h)$ number tells us what SR to expect $h \geq 0$ times ahead.

Summarizing, $\mathbf{ESR}_s^S$ is a vector containing an expected SR of each hidden state of our HMM. Thus, by incorporation the transition estimates, we obtain $\text{PESR}_s^S(h)$ as a metric for predicting expected risk-adjusted returns $h \geq 0$ steps ahead given the HMM with S hidden states. Both metrics are fitted on the features using span s, extracted from the past observations $(x_1, \ldots, x_n)$. One may note that more elaborating functions could be made by including higher order of moments, incorporating the downside risk of returns. Extracting features using closing and opening prices, high and low prices, and volume may also be of interest, as long as the features are not linearly correlated.

A.5 Experiments

In our experiments, we divide the data set into three parts: training (up to the year 2012 $\approx$ twelve years), validation (the year 2012 to 2016 $\approx$ four years), and test set (from the year 2016 $\approx$ four years).

We train our HMM using the features $f_s = (f_s^1, f_s^2)$ extracted from our training data. Then we validate the (out-of-sample) performance by evaluating our model on the validation data. Selecting training data with suitable variability will help us improve the models' ability to generalize. Thus, we identify the desired pattern(s) in our training data, which explains our validation data's behavior the best. To avoid getting stuck in a local maximum, we select the HMM with the highest score over many trained models, where each model is randomly initialized.

Our goal is to enhance the risk-adjusted returns with the use of our proposed PESR metric $\text{PESR}_s^S(h)$ from (A.4.4). We choose the number of hidden states relatively low to have high interpretability of each hidden state in our HMM. Thus, our choice is an HMM with three hidden states ($S = 3$), where the hidden states can be labeled as a bull, bear, and high volatility regime. Our labeling comes from the fact that our estimated ESR metric outputs a positive, negative, and (close to) zero value, which can be labeled into a bull, bear, and high volatility regime. Our high volatility regimes have an estimated ESR metric close to zero as the estimated volatility dominates, i.e., $\mu_j(f_s^2)$ is sufficiently larger than $\mu_j(f_s^1)$ and $\mu_j(f_s^1)$ is close to zero.

We model the outcomes/predictions of the PESR metric $\text{PESR}_s^S(h)$ into the two different holding strategies; a long-only strategy and long/short strategy. We will not restrict the turnover level, but we incorporate a transaction cost of 5bps for buying and selling. Lastly, as we are disallowing gearing, we cap our holdings onto the range $[0, 1]$ for the long-only strategy and $[-1, 1]$ for the long/short strategy. If we were to increase the number of hidden states (and/or adding other features) in our HMM, then the PESR metric's outcomes may be transformed into a more advanced holding strategy.

From our training and validation data, we observe that spans $s \in \{15, 30, 60\}$ seems preferable to have some different levels of transitions within the four years of testing. Thus, we will in the next section consider span $s \in \{15, 30, 60\}$. This range of spans s would also illustrate how the choice of

span affects our method's regime stickiness. Recall that the choice of span s will directly affect the turnover, meaning a lower span s may increase (absolute) performance and lower regime stickiness, i.e., increase the level of turnover.

All results in the following section are made using the (out-of-sample) test period from January 2016 to October 2019. Before we move to the results of our experiments, then we may need an overview of the instrument's performance metrics to compare with the outcome of our strategies. In Table A.3, we have the annualized returns, annualized volatilities, Sharpe ratios, and maximum drawdowns of each instrument in $I = (I_1, \ldots, I_{15})^T$. As we earlier discussed in Section A.3, each instrument's performance metrics vary a lot, but also within each asset class, we have large variations. However, most annualized returns are positive (with only a few exceptions) but achieved under different volatility levels.

#	**Instr.**	**Ret.**	**Vol.**	**SR**	**DD**
1	CO1	7.54%	12.02%	0.67	7.66%
2	CO2	−12.94%	39.25%	−0.16	20.0%
3	CO3	10.22%	33.09%	0.46	17.44%
4	FX1	−6.01%	9.52%	−0.6	9.93%
5	FX2	−1.68%	6.84%	−0.21	3.78%
6	FX3	1.65%	8.57%	0.23	5.56%
7	FX4	0.29%	5.9%	0.08	3.39%
8	EQ1	6.44%	15.35%	0.49	11.8%
9	EQ2	7.51%	12.85%	0.64	6.15%
10	EQ3	5.43%	20.56%	0.36	14.16%
11	EQ4	11.23%	11.35%	1.0	7.82%
12	FI1	0.92%	3.78%	0.27	2.31%
13	FI2	4.23%	4.19%	1.02	2.1%
14	FI3	4.91%	5.35%	0.94	3.23%
15	FI4	1.59%	1.71%	0.96	1.39%

Table A.3: Realized performance metrics; annualized returns, annualized volatility, Sharpe ratios, and maximum drawdowns of instruments I in the test period from January 2016 to October 2019.

A.5.1 Results

The results of our long-only strategy based on the outcomes of $\text{PESR}_s^3(1)_{s\in\{15,30,60\}}$ are presented in Table A.4. Table A.4 confirms our claim that lower (higher) levels of span s delivers a higher (lower) level of turnover. However, different choices of span s affect the performance metrics individually due to both the "true" length of market regimes and the transaction costs. If we consider span $s = 30$, then what first comes to mind is that all (annualized) returns are positive with slightly lower (annualized) volatility leading to an improved SR, now above one for all assets (except from FX1, which have a SR of 0.86). Furthermore, CO1, EQ4, and FI1, now have a SR above two. The daily turnover range from 1.64% to 4.08%, giving an investment horizon of approximately 25 to over 60 days. Thus, one would have a monthly re-balancing scheme for this long-only strategy.

The overall results presented in Table A.4 show a convincing improvement of SR with a feasible turnover rate (which can be changed after preferences through the selection of span s). Nevertheless, we cannot guarantee that the cumulative return will be improved using our PESR metric, as the aim is to improve risk-adjusted returns. FI4 is an example of this as we see an improved SR but not a cumulative return. In such cases, additional span sizes should be included to embrace these assets. Several factors affect the investment strategy, but the choice of span s has a significant influence since it operates as a smoothing factor and determines the regime shifts' frequency (i.e., the regime stickiness). Thus, assets with low volatility may not require much smoothing, suggesting that we should use higher levels of span s. In addition, transaction costs play a significant role as the absolute returns are small.

Long-only		**PESR$^3_{15}(1)$**					**PESR$^3_{30}(1)$**					**PESR$^3_{60}(1)$**				
#	**Instr.**	**Ret.**	**Vol.**	**SR**	**DD**	**Turn.**	**Ret.**	**Vol.**	**SR**	**DD**	**Turn.**	**Ret.**	**Vol.**	**SR**	**DD**	**Turn.**
1	CO1	11.65%	7.8%	2.4	6.29%	4.87%	13.21%	8.27%	2.44	6.48%	2.67%	9.61%	8.39%	1.74	6.67%	1.96%
2	CO2	29.74%	24.13%	2.08	17.7%	3.87%	15.56%	25.29%	1.2	17.7%	2.49%	10.57%	26.39%	0.73	17.7%	2.46%
3	CO3	31.57%	20.62%	1.95	10.94%	4.62%	22.08%	18.35%	1.59	12.81%	3.19%	15.55%	17.88%	1.25	12.81%	2.54%
4	FX1	3.48%	5.71%	0.84	4.09%	4.56%	2.98%	4.89%	0.86	3.67%	2.52%	1.65%	4.51%	0.55	3.29%	2.01%
5	FX2	4.63%	4.26%	1.93	3.78%	3.69%	3.75%	3.89%	1.77	2.57%	2.36%	2.95%	3.89%	1.41	2.57%	1.47%
6	FX3	4.33%	5.31%	1.19	3.59%	4.02%	4.66%	5.19%	1.27	4.72%	2.7%	6.73%	5.75%	1.54	4.72%	2.05%
7	FX4	3.37%	3.16%	1.94	2.13%	3.82%	2.89%	3.7%	1.29	2.71%	2.46%	3.44%	4.13%	1.08	2.42%	2.06%
8	EQ1	15.64%	8.56%	2.48	11.32%	4.49%	11.51%	8.92%	1.73	11.32%	3.74%	8.04%	9.84%	1.04	11.32%	2.66%
9	EQ2	13.44%	6.97%	2.55	4.63%	4.91%	8.02%	7.68%	1.36	4.91%	3.77%	9.81%	8.18%	1.54	5.67%	2.24%
10	EQ3	23.22%	11.22%	2.46	7.58%	5.4%	16.85%	11.47%	1.8	7.58%	3.71%	15.42%	12.05%	1.58	9.58%	2.06%
11	EQ4	14.4%	6.35%	3.0	5.17%	4.79%	12.59%	7.17%	2.01	7.72%	3.43%	10.93%	8.28%	1.44	7.72%	2.07%
12	FI1	2.01%	1.91%	2.49	1.65%	2.44%	2.26%	1.97%	2.58	1.69%	1.64%	1.83%	1.99%	1.92	1.82%	1.44%
13	FI2	5.21%	2.74%	2.62	1.53%	4.91%	4.09%	2.85%	1.93	2.07%	4.08%	3.72%	3.01%	1.6	2.07%	2.97%
14	FI3	6.0%	3.97%	1.8	2.18%	4.64%	4.17%	3.71%	1.35	2.22%	3.87%	3.86%	3.73%	1.31	3.1%	3.09%
15	FI4	0.5%	1.04%	1.25	1.39%	1.63%	0.74%	1.01%	1.62	1.12%	1.68%	1.36%	1.21%	1.97	1.39%	1.27%

Table A.4: Realized performance metrics; annualized returns, annualized volatilities, Sharpe ratios, maximum drawdowns, and daily turnovers of long-only strategies $\text{PESR}^3_s(1)_{s\in\{15,30,60\}}$ in the test period from January 2016 to October 2019.

Next, in Table A.5, we have the results of our long/short strategy; this strategy seems to provide larger (absolute) returns but with increased volatility, leading to a lower SR than for the long-only strategy. This means the short leg of our strategies adds some more volatility to the strategy. Naturally, as we can be short now, this leads to increasing daily turnover, e.g., for span $s = 30$, the turnover now ranges from 3.37% to 7.34% giving an investment horizon of approximately 15 to 30 days. As the turnover increase, the same do transaction costs, which for some strategies/assets may represent a significant part of the overall performance. Particularly, FI4 has a negative SR (and cumulative return), however, with lower volatility than the asset itself.

Time-series plots of cumulative returns of each instrument I for both HMM strategies (long-only and long/short) can be found in A.7, including their corresponding holdings. These figures show that we mostly shift between the bull and bear regime, and only in the high volatility state for some short periods. Overall, as we seek to increase our risk-adjusted returns, then the long-only strategy would be preferred. However, if we relaxed our risk-aversion, we could maximize total return using the long/short strategy.

Long/short		$\text{PESR}^3_{15}(1)$					$\text{PESR}^3_{30}(1)$					$\text{PESR}^3_{60}(1)$				
#	Instr.	Ret.	Vol.	SR	DD	Turn.	Ret.	Vol.	SR	DD	Turn.	Ret.	Vol.	SR	DD	Turn.
1	CO1	15.69%	11.53%	1.31	7.76%	9.68%	18.45%	11.62%	1.52	6.66%	5.22%	9.94%	9.27%	1.07	6.67%	3.13%
2	CO2	51.61%	33.75%	1.4	17.7%	7.61%	33.52%	34.48%	1.01	17.7%	4.72%	26.5%	36.1%	0.83	17.7%	4.81%
3	CO3	53.3%	27.25%	1.7	11.4%	9.03%	26.26%	26.62%	1.01	13.88%	6.63%	16.58%	26.12%	0.72	13.11%	5.27%
4	FX1	10.74%	7.25%	1.43	4.09%	8.22%	9.38%	6.96%	1.32	3.67%	5.11%	6.77%	7.13%	0.95	3.93%	3.9%
5	FX2	10.64%	6.64%	1.55	3.78%	7.56%	9.62%	6.52%	1.44	3.31%	4.74%	8.03%	6.54%	1.22	3.31%	2.98%
6	FX3	7.42%	6.89%	1.06	3.98%	7.38%	7.68%	6.64%	1.15	5.02%	4.92%	9.51%	6.9%	1.35	5.02%	4.1%
7	FX4	6.05%	5.68%	1.06	3.74%	7.72%	5.34%	5.7%	0.94	3.32%	5.0%	5.91%	5.71%	1.04	3.32%	4.16%
8	EQ1	27.3%	13.5%	1.87	11.68%	8.66%	13.81%	13.16%	1.06	11.32%	7.18%	10.74%	13.88%	0.81	12.07%	5.02%
9	EQ2	23.33%	11.27%	1.93	6.72%	9.85%	13.14%	11.36%	1.15	6.6%	7.34%	11.61%	11.99%	0.98	6.72%	4.58%
10	EQ3	43.07%	16.19%	2.3	8.15%	10.64%	29.11%	16.81%	1.62	12.54%	6.98%	22.73%	16.86%	1.31	12.15%	4.43%
11	EQ4	18.51%	10.55%	1.66	6.65%	9.16%	15.01%	10.66%	1.36	9.17%	6.51%	10.67%	11.06%	0.97	7.86%	3.96%
12	FI1	3.28%	3.68%	0.9	1.81%	4.74%	3.58%	3.67%	0.99	1.87%	3.37%	2.32%	2.66%	0.87	1.82%	2.46%
13	FI2	6.28%	4.0%	1.52	1.98%	9.92%	4.93%	4.0%	1.21	2.07%	7.48%	3.35%	3.59%	0.93	2.07%	5.07%
14	FI3	6.98%	4.66%	1.47	2.33%	9.13%	4.24%	4.41%	0.97	2.36%	6.92%	3.49%	4.45%	0.8	3.15%	5.31%
15	FI4	−0.57%	1.66%	−0.34	1.39%	3.29%	−0.2%	1.57%	−0.12	1.12%	3.45%	1.22%	1.35%	0.91	1.39%	2.03%

Table A.5: Realized performance metrics; annualized returns, annualized volatilities, Sharpe ratios, maximum drawdowns, and daily turnovers of long/short strategies $\text{PESR}^3_s(1)_{s\in\{15,30,60\}}$ in the test period from January 2016 to October 2019.

A.6 Discussion

HMMs have previously been applied to finance time series with great success but never on a broad class of assets, at least not to our knowledge. We proposed an asset independent three-state HMM for predicting risk-adjusted returns trained using only the first two moments as features. The model outcome was combined into a metric for predicting expected SRs. Our investigation showed a proper ability to predict bull, bear, and high-volatility regimes, which lead to enhanced risk-adjusted returns (compared to buying the underlying asset) while keeping a preferable turnover level. However, this could be improved by fine-tuning the choice of span s as transaction costs could otherwise dominate.

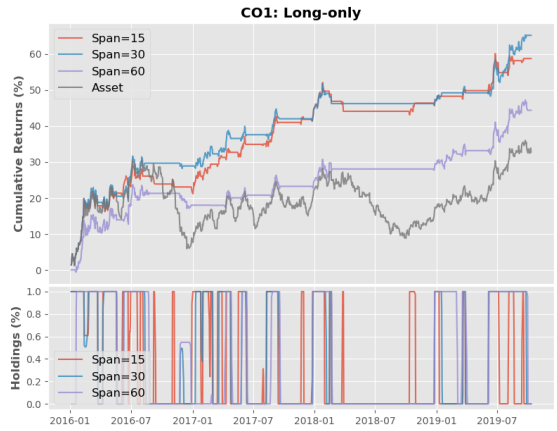
As our findings were made using the entire test dataset to predict the hidden state sequence, our next focus will then be an extension to a setting in which we make incremental predictions of tomorrow's expected SR using only past information. As this may increase noise, we could increase our model's predictability by introducing time-varying parameters, i.e., an adaptive model where the model parameters are updated as new observations arrive (e.g., see [47] and [112]). Expanding this analysis with a larger group of features, e.g., volume, higher-order moments, short-term oscillators, and associated gradients, could be appealing. All this could be combined with the feature saliency HMM proposed by [2], which comprises the treatment of "irrelevant" features.

A.7 Cumulative Returns of HMM Strategies

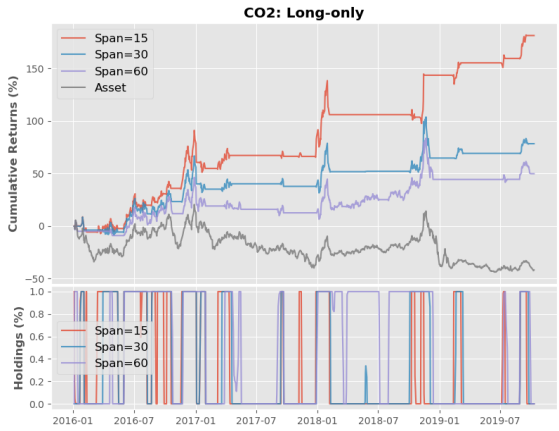
Figure A.1a-A.3c and Figure A.3d-A.5f shows the HMM strategies long-only and long/short, respectively, based on the outcomes of $\text{PESR}^3_s(1)_{s\in\{15,30,60\}}$ for the instruments I.

Figure A.1: Cumulative returns HMM strategies in the test period from January 2016 to October 2019.

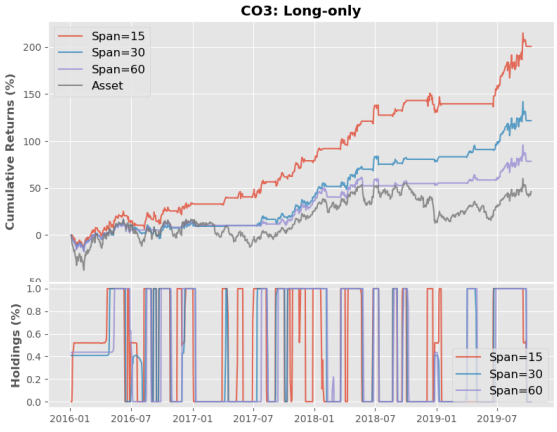
(a) I_1 (CO1)



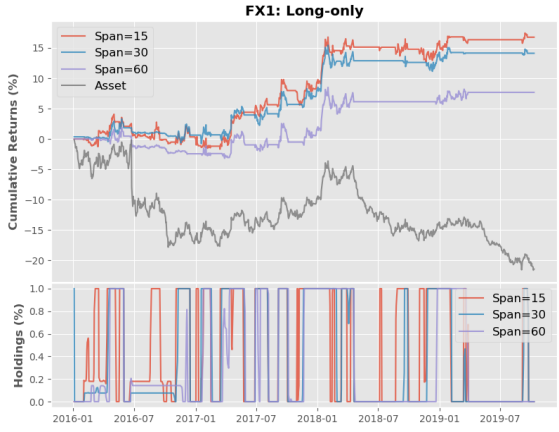
(b) I_2 (CO2)



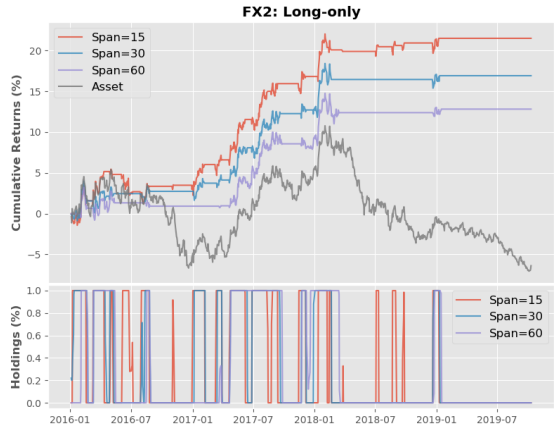
(c) I_3 (CO3)



(d) I_4 (FX1)



(e) I_5 (FX2)



(f) I_6 (FX3)

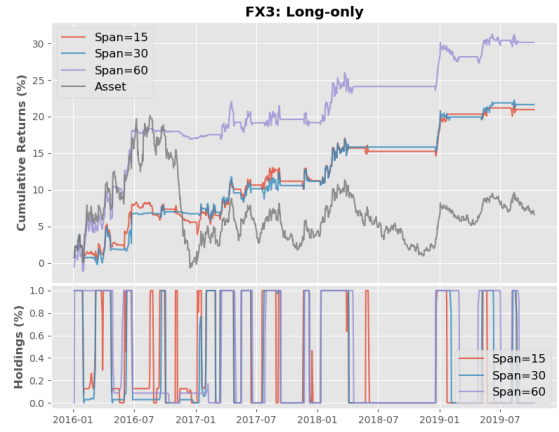
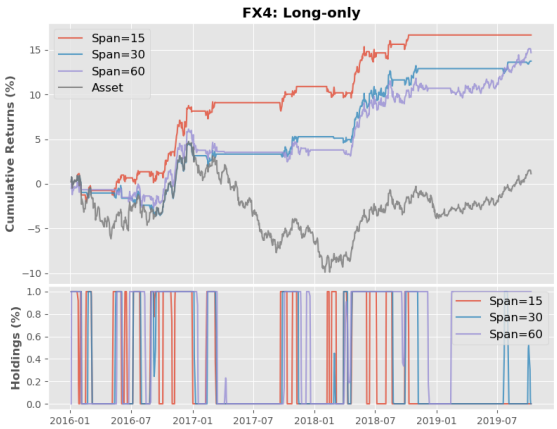
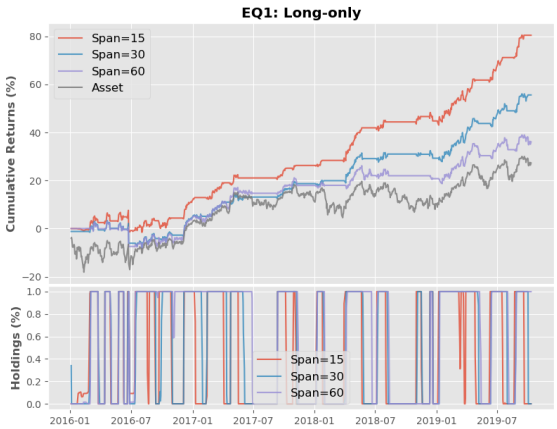


Figure A.2: Cumulative returns HMM strategies in the test period from January 2016 to October 2019.

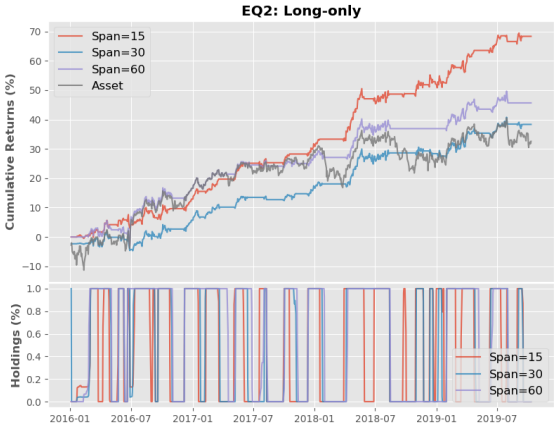
(a) I_7 (FX4)



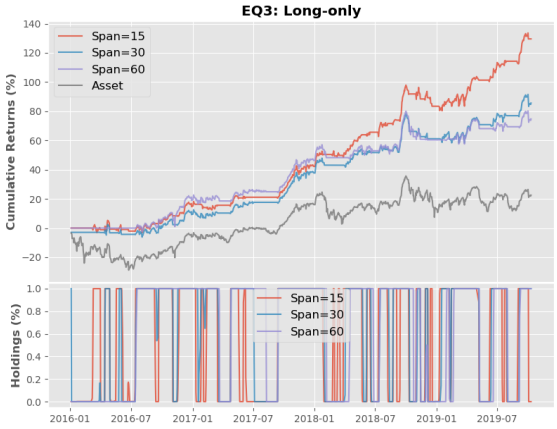
(b) I_8 (EQ1)



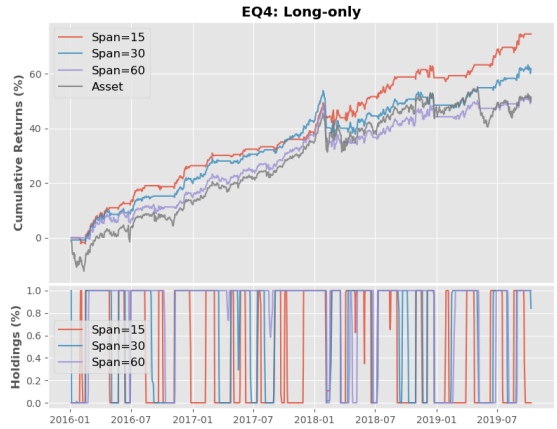
(c) I_9 (EQ2)



(d) I_{10} (EQ3)



(e) I_{11} (EQ4)



(f) I_{12} (FI1)

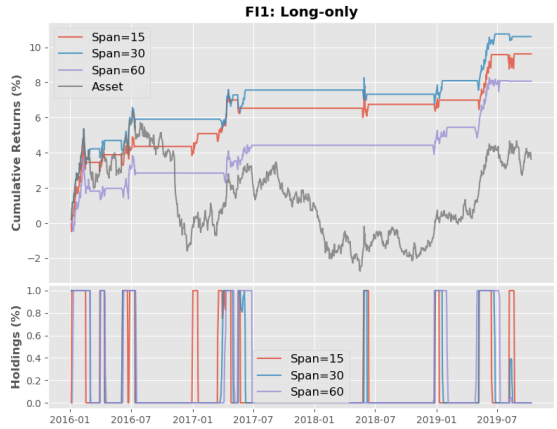
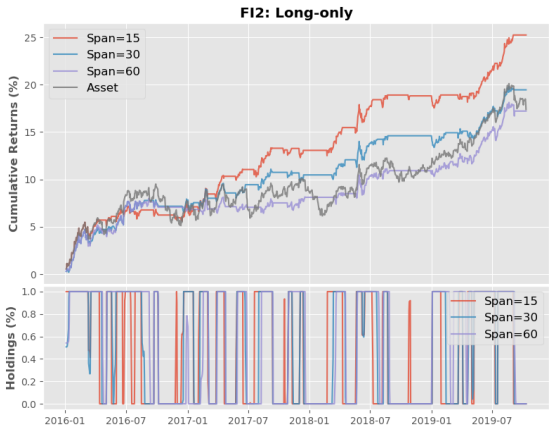
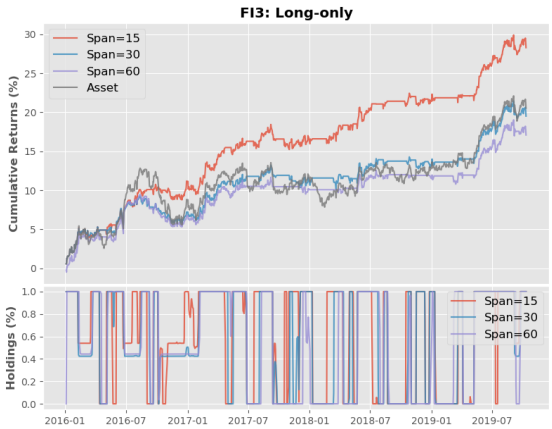


Figure A.3: Cumulative returns HMM strategies in the test period from January 2016 to October 2019.

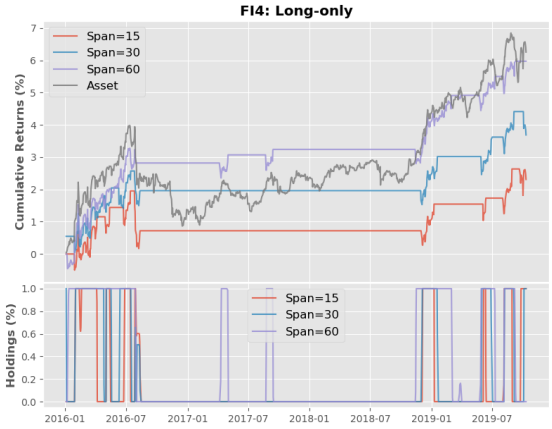
(a) I_{13} (FI2)



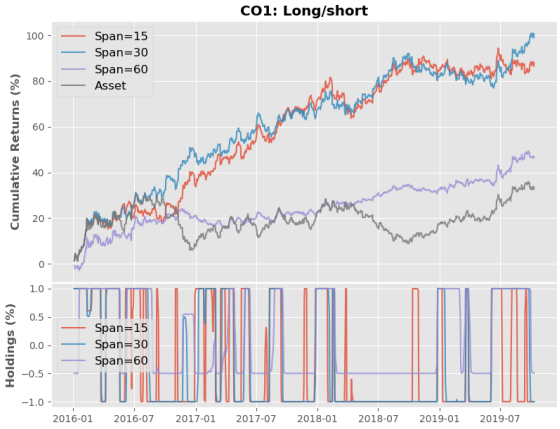
(b) I_{14} (FI3)



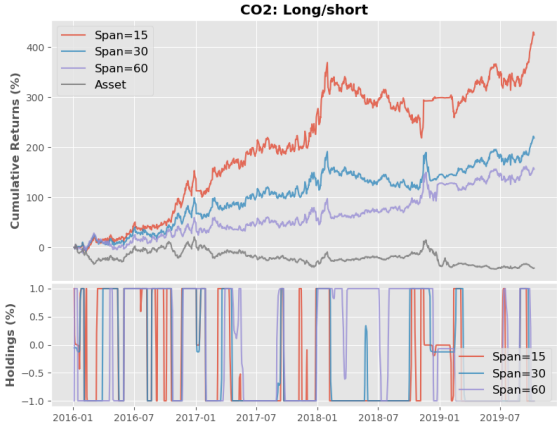
(c) I_{15} (FI4)



(d) I_1 (CO1)



(e) I_2 (CO2)



(f) I_3 (CO3)

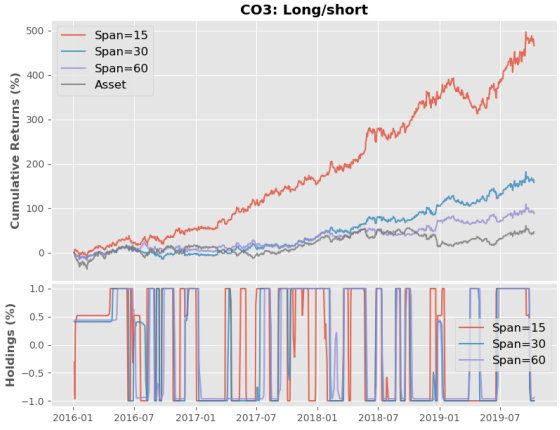
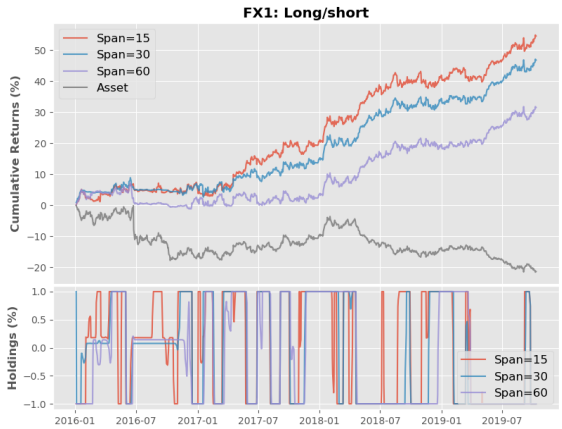
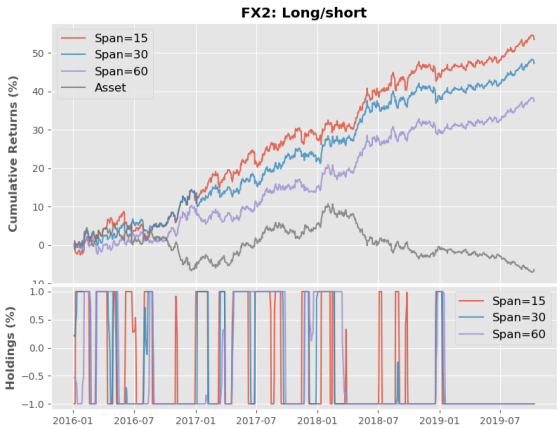


Figure A.4: Cumulative returns HMM strategies in the test period from January 2016 to October 2019.

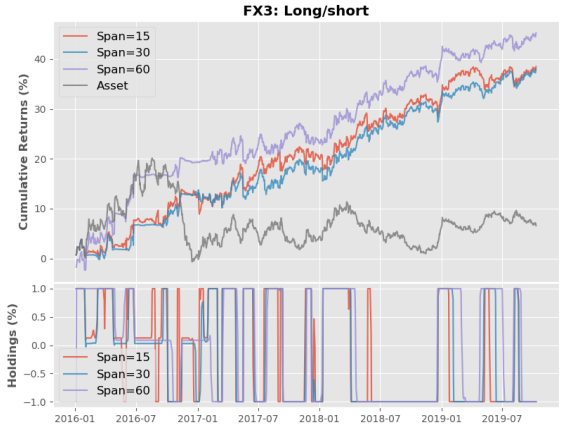
(a) I_4 (FX1)



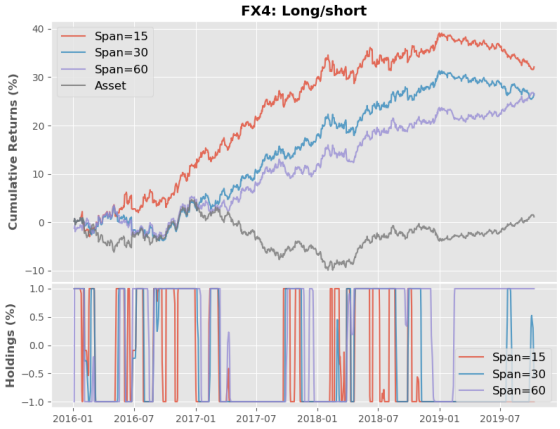
(b) I_5 (FX2)



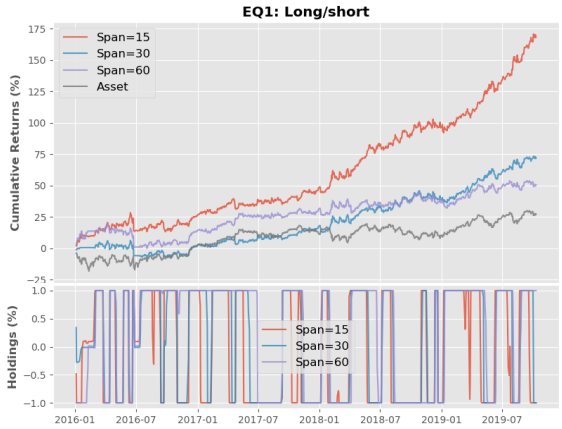
(c) I_6 (FX3)



(d) I_7 (FX4)



(e) I_8 (EQ1)



(f) I_9 (EQ2)

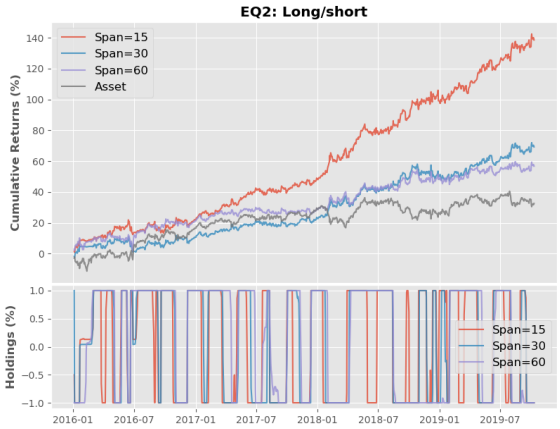
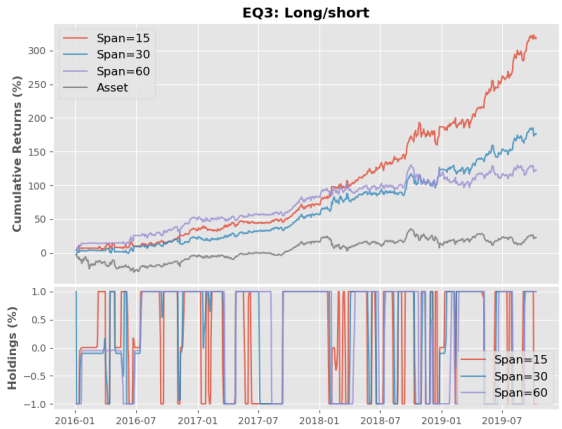
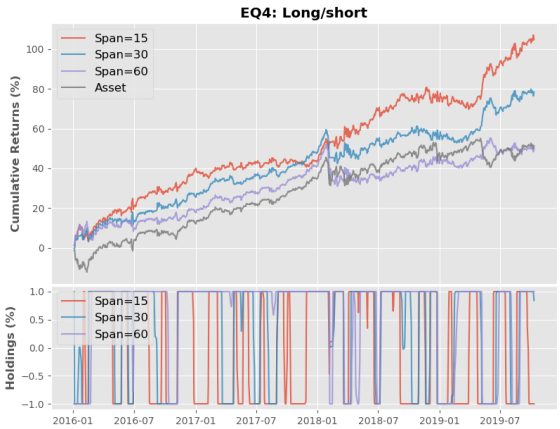


Figure A.5: Cumulative returns HMM strategies in the test period from January 2016 to October 2019.

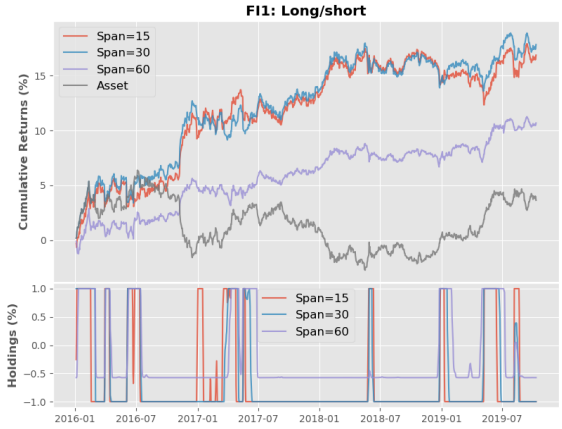
(a) I_{10} (EQ3)



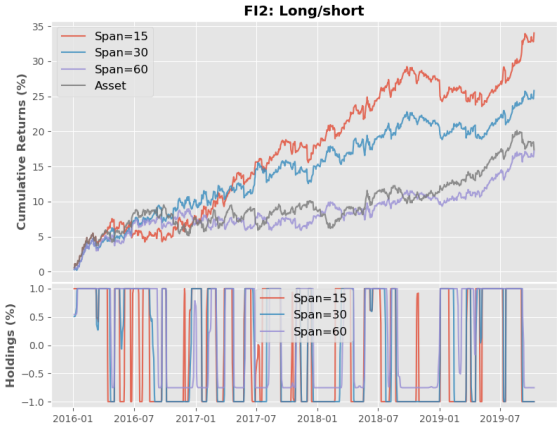
(b) I_{11} (EQ4)



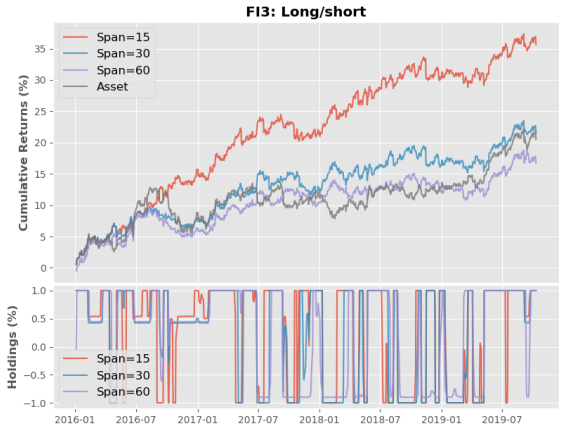
(c) I_{12} (FI1)



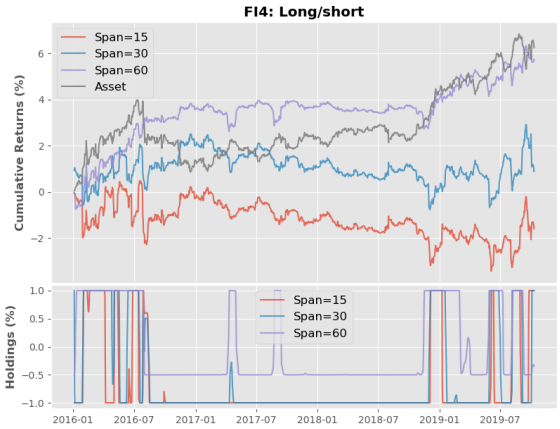
(d) I_{13} (FI2)



(e) I_{14} (FI3)



(f) I_{15} (FI4)



Appendix B: Technical Results and Their Proofs

B.1 Outline

This chapter contains purely technical results used in the proofs presented in Chapters 2 and 3 [57, 58]. In what follows, we use the conventions $\inf \emptyset = 0$, $\sum_{t=1}^{0} = 0$ and $\prod_{t=1}^{0} = 1$.

Proposition B.1.1. *Let $(\gamma_t)_{t\geq 1}$ be a positive sequence. For any $k \leq t$, and $\omega > 0$, we have*

$$\sum_{i=k}^{t} \prod_{j=i+1}^{t} [1+\omega\gamma_j]\,\gamma_i \leq \frac{1}{\omega} \prod_{j=k}^{t} [1+\omega\gamma_j] \leq \frac{1}{\omega} \exp\left(\omega \sum_{j=k}^{t} \gamma_j\right). \tag{B.1.1}$$

Proof of Proposition B.1.1. We begin with considering the first inequality in (B.1.1), which follows by expanding the sum of product:

$$\begin{aligned}
\sum_{i=k}^{t} \prod_{j=i+1}^{t} [1+\omega\gamma_j]\,\gamma_i =& \frac{1}{\omega} \sum_{i=k}^{t} \prod_{j=i+1}^{t} [1+\omega\gamma_j]\,\omega\gamma_i = \frac{1}{\omega} \sum_{i=k}^{t} \prod_{j=i+1}^{t} [1+\omega\gamma_j]\,[1+\omega\gamma_i - 1] \\
=& \frac{1}{\omega} \sum_{i=k}^{t} \left[\prod_{j=i+1}^{t} [1+\omega\gamma_j]\,[1+\omega\gamma_i] - \prod_{j=i+1}^{t} [1+\omega\gamma_j] \right] \\
=& \frac{1}{\omega} \sum_{i=k}^{t} \left[\prod_{j=i}^{t} [1+\omega\gamma_j] - \prod_{j=i+1}^{t} [1+\omega\gamma_j] \right].
\end{aligned}$$

As the (positive) terms cancel out, we end up with the first inequality in (B.1.1):

$$\begin{aligned}
\frac{1}{\omega} \sum_{i=k}^{t} \left[\prod_{j=i}^{t} [1+\omega\gamma_j] - \prod_{j=i+1}^{t} [1+\omega\gamma_j] \right] =& \frac{1}{\omega} \left[\prod_{j=k}^{t} [1+\omega\gamma_j] - \prod_{j=k+1}^{t} [1+\omega\gamma_j] + \cdots - \prod_{j=t+1}^{t} [1+\omega\gamma_j] \right] \\
=& \frac{1}{\omega} \left[\prod_{j=k}^{t} [1+\omega\gamma_j] - \prod_{j=t+1}^{t} [1+\omega\gamma_j] \right] \\
=& \frac{1}{\omega} \left[\prod_{j=k}^{t} [1+\omega\gamma_j] - 1 \right] \leq \frac{1}{\omega} \prod_{j=k}^{t} [1+\omega\gamma_j],
\end{aligned}$$

as $\prod_{t+1}^{t} = 1$ for all $t \in \mathbb{N}$. Using the (simple) bound of $1 + t \leq \exp(t)$ for all $t \in \mathbb{R}$, we obtain the

second inequality of (B.1.1):

$$\frac{1}{\omega}\prod_{j=k}^{t}\left[1+\omega\gamma_j\right] \leq \frac{1}{\omega}\prod_{j=k}^{t}\exp\left(\omega\gamma_j\right) = \frac{1}{\omega}\exp\left(\omega\sum_{j=k}^{t}\gamma_j\right).$$

□

Proposition B.1.2. *Let $(\gamma_t)_{t\geq 1}$ be a positive sequence. Let $\omega > 0$ and $k \leq t$ such that for all $i \geq k$, $\omega\gamma_i \leq 1$, then*

$$\sum_{i=k}^{t}\prod_{j=i+1}^{t}\left[1-\omega\gamma_j\right]\gamma_i \leq \frac{1}{\omega}. \tag{B.1.2}$$

Proof of Proposition B.1.2. We start with expanding the sums of products term in (B.1.2), given us

$$\begin{aligned}
\sum_{i=k}^{t}\prod_{j=i+1}^{t}\left[1-\omega\gamma_j\right]\gamma_i =& \frac{1}{\omega}\sum_{i=k}^{t}\prod_{j=i+1}^{t}\left[1-\omega\gamma_j\right]\omega\gamma_i = -\frac{1}{\omega}\sum_{i=k}^{t}\prod_{j=i+1}^{t}\left[1-\omega\gamma_j\right]\left[-\omega\gamma_i\right] \\
=& -\frac{1}{\omega}\sum_{i=k}^{t}\prod_{j=i+1}^{t}\left[1-\omega\gamma_j\right]\left[1-\omega\gamma_i-1\right] \\
=& -\frac{1}{\omega}\sum_{i=k}^{t}\left[\prod_{j=i+1}^{t}\left[1-\omega\gamma_j\right]\left[1-\omega\gamma_i\right] - \prod_{j=i+1}^{t}\left[1-\omega\gamma_j\right]\right] \\
=& -\frac{1}{\omega}\sum_{i=k}^{t}\left[\prod_{j=i}^{t}\left[1-\omega\gamma_j\right] - \prod_{j=i+1}^{t}\left[1-\omega\gamma_j\right]\right] \\
=& \frac{1}{\omega}\sum_{i=k}^{t}\left[\prod_{j=i+1}^{t}\left[1-\omega\gamma_j\right] - \prod_{j=i}^{t}\left[1-\omega\gamma_j\right]\right].
\end{aligned}$$

As we only have positive terms, we can upper bound the term:

$$\frac{1}{\omega}\sum_{i=k}^{t}\left[\prod_{j=i+1}^{t}\left[1-\omega\gamma_j\right] - \prod_{j=i}^{t}\left[1-\omega\gamma_j\right]\right] \leq \frac{1}{\omega}\left[1-\prod_{j=k}^{t}\left[1-\omega\gamma_j\right]\right] \leq \frac{1}{\omega},$$

using $\prod_{j=k}^{t}[1-\omega\gamma_j] \geq 0$, showing the inequality in (B.1.2). □

Proposition B.1.3. *Let $(\gamma_t)_{t\geq 1}$ and $(\eta_t)_{t\geq 1}$ be positive sequences. For any $k \leq t$, we can obtain the (upper) bounds:*

$$\sum_{i=k}^{t}\prod_{j=i+1}^{t}\left[1+\omega\gamma_j\right]\eta_i\gamma_i \leq \frac{1}{\omega}\max_{k\leq i\leq t}\eta_i\exp\left(\omega\sum_{j=k}^{t}\gamma_j\right), \tag{B.1.3}$$

with $\omega > 0$. Furthermore, suppose that for all $i \geq k$, $\omega\gamma_i \leq 1$, then

$$\sum_{i=k}^{t} \prod_{j=i+1}^{t} [1 - \omega\gamma_j] \eta_i \leq \frac{1}{\omega} \max_{k \leq i \leq t} \eta_i. \tag{B.1.4}$$

Proof of Proposition B.1.3. We obtain the inequality in (B.1.3) directly by Proposition B.1.1:

$$\begin{aligned}
\sum_{i=k}^{t} \prod_{j=i+1}^{t} [1 + \omega\gamma_j] \eta_i\gamma_i \leq & \max_{k \leq i \leq t} \eta_i \sum_{i=k}^{t} \prod_{j=i+1}^{t} [1 + \omega\gamma_j] \gamma_i \\
\leq & \frac{1}{\omega} \max_{k \leq i \leq t} \eta_i \prod_{j=k}^{t} [1 + \omega\gamma_j] \\
\leq & \frac{1}{\omega} \max_{k \leq i \leq t} \eta_i \exp\left(\omega \sum_{j=k}^{t} \gamma_j\right).
\end{aligned}$$

Similarly, for the inequality in (B.1.4), we have

$$\sum_{i=k}^{t} \prod_{j=i+1}^{t} [1 - \omega\gamma_j] \eta_i\gamma_i \leq \max_{k \leq i \leq t} \eta_i \sum_{i=k}^{t} \prod_{j=i+1}^{t} [1 - \omega\gamma_j] \gamma_i \leq \frac{1}{\omega} \max_{k \leq i \leq t} \eta_i,$$

by Proposition B.1.2. □

Proposition B.1.4. *Let $(\delta_t)_{t\geq 0}$, $(\gamma_t)_{t\geq 1}$, $(\eta_t)_{t\geq 1}$, and $(\nu_t)_{t\geq 1}$ be some positive sequences satisfying the recursive relation:*

$$\delta_t \leq (1 - 2\omega\gamma_t + \eta_t\gamma_t) \delta_{t-1} + \nu_t\gamma_t, \tag{B.1.5}$$

with $\delta_0 \geq 0$ and $\omega > 0$. Denote $t_0 = \inf\{t \geq 1 : \eta_t \leq \omega\}$, and suppose that for all $t \geq t_0 + 1$, one has $\omega\gamma_t \leq 1$. Then, for γ_t and η_t decreasing, we have the upper bound on (δ_t):

$$\delta_t \leq \exp\left(-\omega \sum_{i=t/2}^{t} \gamma_i\right) \left[\exp\left(\sum_{i=1}^{t_0} \eta_i\gamma_i\right) \left(\delta_0 + \max_{1 \leq i \leq t_0} \frac{\nu_i}{\eta_i}\right) + \sum_{i=t_0+1}^{t/2-1} \nu_i\gamma_i\right] + \frac{1}{\omega} \max_{t/2 \leq i \leq t} \nu_i, \tag{B.1.6}$$

for all $t \in \mathbb{N}$ with the convention that $\sum_{t_0}^{t/2} = 0$ if $t/2 < t_0$.

Proof of Proposition B.1.4. Applying the recursive relation from (B.1.5) t times, we derive:

$$\delta_t \leq \underbrace{\prod_{i=1}^{t} [1 - 2\omega\gamma_i + \eta_i\gamma_i]}_{B_t} \delta_0 + \underbrace{\sum_{i=1}^{t} \prod_{j=i+1}^{t} [1 - 2\omega\gamma_j + \eta_j\gamma_j] \nu_i\gamma_i}_{A_t},$$

where B_t can be seen as a transient term only depending on the initialisation δ_0, and a stationary

term A_t. The transient term B_t can be divided into two products, before and after t_0,

$$B_t = \prod_{i=1}^{t} [1 - 2\omega\gamma_i + \eta_i\gamma_i] = \left(\prod_{i=1}^{t_0} [1 - 2\omega\gamma_i + \eta_i\gamma_i]\right) \left(\prod_{i=t_0+1}^{t} [1 - 2\omega\gamma_i + \eta_i\gamma_i]\right).$$

Using that $t_0 = \inf\{t \geq 1 : \eta_t \leq \omega\}$, and since for all $t \geq t_0 + 1$, we have $2\omega\gamma_t - \eta_t\gamma_t \geq \omega\gamma_t$, it comes

$$\begin{aligned} B_t &\leq \left(\prod_{i=1}^{t_0} [1 - 2\omega\gamma_i + \eta_i\gamma_i]\right) \left(\prod_{i=t_0+1}^{t} [1 - \omega\gamma_i]\right) \leq \left(\prod_{i=1}^{t_0} \exp\left(-2\omega\gamma_i + \eta_i\gamma_i\right)\right) \left(\prod_{i=t_0+1}^{t} \exp\left(-\omega\gamma_i\right)\right) \\ &= \exp\left(-2\omega \sum_{i=1}^{t_0} \gamma_i\right) \exp\left(\sum_{i=1}^{t_0} \eta_i\gamma_i\right) \exp\left(-\omega \sum_{i=t_0+1}^{t} \gamma_i\right) \leq \exp\left(-\omega \sum_{i=1}^{t} \gamma_i\right) \exp\left(\sum_{i=1}^{t_0} \eta_i\gamma_i\right) \end{aligned}$$

by applying the (simple) bound $1 + t \leq \exp(t)$ for all $t \in \mathbb{R}$. We derive that

$$B_t \leq \exp\left(-\omega \sum_{i=t/2}^{t} \gamma_i\right) \exp\left(\sum_{i=1}^{t_0} \eta_i\gamma_i\right). \tag{B.1.7}$$

Next, the stationary term A_t can (similarly) be divided into two sums (after and before t_0):

$$A_t = \underbrace{\sum_{i=t_0+1}^{t} \prod_{j=i+1}^{t} [1 - 2\omega\gamma_j + \eta_j\gamma_j] \nu_i\gamma_i}_{A_{t,1}} + \underbrace{\sum_{i=1}^{t_0} \prod_{j=i+1}^{t} [1 - 2\omega\gamma_j + \eta_j\gamma_j] \nu_i\gamma_i}_{A_{t,2}}.$$

The first stationary term $A_{t,1}$ (with $t > t_0$) can be bounded as follows: if $t/2 \leq t_0 + 1$, we have

$$A_{t,1} \leq \max_{t_0+1 \leq i \leq t} \nu_i \sum_{i=t_0+1}^{t} \prod_{j=i+1}^{t} [1 - \omega\gamma_j] \gamma_i = \frac{1}{\omega} \max_{t_0+1 \leq i \leq t} \nu_i \leq \frac{1}{\omega} \max_{t/2 \leq i \leq t} \nu_i,$$

by Proposition B.1.3. Furthermore, if $t/2 > t_0 + 1$, we get

$$\begin{aligned} A_{t,1} &\leq \sum_{i=t_0+1}^{t} \prod_{j=i+1}^{t} [1 - \omega\gamma_j] \nu_i\gamma_i = \sum_{i=t_0+1}^{t/2-1} \prod_{j=i+1}^{t} [1 - \omega\gamma_j] \nu_i\gamma_i + \sum_{i=t/2}^{t} \prod_{j=i+1}^{t} [1 - \omega\gamma_j] \nu_i\gamma_i \\ &\leq \sum_{i=t_0+1}^{t/2-1} \prod_{j=t/2}^{t} [1 - \omega\gamma_j] \nu_i\gamma_i + \max_{t/2 \leq i \leq t} \nu_i \sum_{i=t/2}^{t} \prod_{j=i+1}^{t} [1 - \omega\gamma_j] \gamma_i \\ &= \prod_{j=t/2}^{t} [1 - \omega\gamma_j] \sum_{i=t_0+1}^{t/2-1} \nu_i\gamma_i + \frac{1}{\omega} \max_{t/2 \leq i \leq t} \nu_i, \end{aligned}$$

where $\prod_{j=t/2}^{t}[1-\omega\gamma_j] \le \exp(-\omega\sum_{j=t/2}^{t}\gamma_j)$ as $1+t \le \exp(t)$ for all $t \in \mathbb{R}$. Thus, for all $t \in \mathbb{R}$,

$$A_{t,1} \le \exp\left(-\omega \sum_{j=t/2}^{t} \gamma_j\right) \sum_{i=t_0+1}^{t/2-1} \nu_i\gamma_i + \frac{1}{\omega} \max_{t/2\le i\le t} \nu_i, \tag{B.1.8}$$

where $\sum_{t_0}^{t/2} = 0$ if $t/2 < t_0$. The second stationary term $A_{t,2}$ can be bounded, thanks to Proposition B.1.1, as follows:

$$\begin{aligned}
A_{t,2} &= \sum_{i=1}^{t_0} \prod_{j=i+1}^{t} [1-2\omega\gamma_j+\eta_j\gamma_j]\,\nu_i\gamma_i = \left(\prod_{j=t_0+1}^{t} [1-2\omega\gamma_j+\eta_j\gamma_j]\right) \sum_{i=1}^{t_0} \prod_{j=i+1}^{t_0} [1-2\omega\gamma_j+\eta_j\gamma_j]\,\nu_i\gamma_i \\
&\le \left(\prod_{j=t_0+1}^{t} [1-\omega\gamma_j]\right) \sum_{i=1}^{t_0} \prod_{j=i+1}^{t_0} [1+\eta_j\gamma_j]\,\nu_i\gamma_i \le \exp\left(-\omega \sum_{j=t_0+1}^{t} \gamma_j\right) \max_{1\le i\le t_0} \frac{\nu_i}{\eta_i} \sum_{i=1}^{t_0} \prod_{j=i+1}^{t_0} [1+\eta_j\gamma_j]\,\eta_i\gamma_i \\
&\le \exp\left(-\omega \sum_{j=t_0+1}^{t} \gamma_j\right) \max_{1\le i\le t_0} \frac{\nu_i}{\eta_i} \exp\left(\sum_{i=1}^{t_0} \eta_i\gamma_i\right) \le \exp\left(-\omega \sum_{j=1}^{t} \gamma_j\right) \max_{1\le i\le t_0} \frac{\nu_i}{\eta_i} \exp\left(2\sum_{i=1}^{t_0} \eta_i\gamma_i\right),
\end{aligned}$$

by the definition of t_0, thus

$$A_{t,2} \le \exp\left(-\omega \sum_{j=1}^{t} \gamma_j\right) \max_{1\le i\le t_0} \frac{\nu_i}{\eta_i} \exp\left(2\sum_{i=1}^{t_0} \eta_i\gamma_i\right) \le \exp\left(-\omega \sum_{j=t/2}^{t} \gamma_j\right) \max_{1\le i\le t_0} \frac{\nu_i}{\eta_i} \exp\left(2\sum_{i=1}^{t_0} \eta_j\gamma_j\right). \tag{B.1.9}$$

Then, using the bound for $A_{t,1}$ in (B.1.8) and $A_{t,2}$ in (B.1.9), we can bound A_t by

$$A_t \le \exp\left(-\omega \sum_{j=t/2}^{t} \gamma_j\right) \left[\exp\left(2\sum_{i=1}^{t_0} \eta_j\gamma_j\right) \max_{1\le i\le t_0} \frac{\nu_i}{\eta_i} + \sum_{i=t_0+1}^{t/2-1} \nu_i\gamma_i\right] + \frac{1}{\omega} \max_{t/2\le i\le t} \nu_i. \tag{B.1.10}$$

Finally, combining the bound for B_t in (B.1.7) and A_t in (B.1.10), we achieve the bound for $\delta_t \le B_t\delta_0 + A_t$, namely the upper bound in (B.1.6). □

The following proposition is a more simplistic but rougher version of the bound in Proposition B.1.4.

Proposition B.1.5. *Let $(\delta_t)_{t\ge0}$, $(\gamma_t)_{t\ge1}$, $(\eta_t)_{t\ge1}$, and $(\nu_t)_{t\ge1}$ be some positive sequences satisfying the recursive relation in (B.1.5). Denote $t_0 = \inf\{t \ge 1 : \eta_t \le \omega\}$, and suppose that for all $t \ge t_0+1$, one has $\omega\gamma_t \le 1$. Then, for γ_t and η_t decreasing, we have for all $t \in \mathbb{N}$,*

$$\delta_t \le \exp\left(-\omega \sum_{i=t/2}^{t} \gamma_i\right) \exp\left(2\sum_{i=1}^{t} \eta_i\gamma_i\right) \left(\delta_0 + 2\max_{1\le i\le t} \frac{\nu_i}{\eta_i}\right) + \frac{1}{\omega} \max_{t/2\le i\le t} \nu_i. \tag{B.1.11}$$

Proof of Proposition B.1.5. The resulting (upper) bound in (B.1.11) follows directly from (B.1.6)

by noting that $t_0 \leq t$, giving us

$$\sum_{i=t_0+1}^{t/2-1} \nu_i \gamma_i \leq \sum_{i=1}^{t} \nu_i \gamma_i \leq \max_{1 \leq i \leq t} (\nu_i / \eta_i) \sum_{i=1}^{t} \eta_i \gamma_i \leq \max_{1 \leq i \leq t} (\nu_i / \eta_i) \exp(2 \sum_{i=1}^{t} \eta_i \gamma_i),$$

as (ν_t) and (γ_t) are positive sequences. $\square$

List of Figures

List of Tables