On Adaptive Stochastic Optimization for Streaming Data: A Newton's Method with O(dN) Operations

Antoine Godichon-Baggioni, Nicklas Werge

On Adaptive Stochastic Optimization for Streaming Data: A Newton's Method with $\mathcal{O}(dN)$ Operations

Antoine Godichon-Baggioni[1] and Nicklas Werge[2]

[1] Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, Paris, France, antoine.godichon_baggioni@sorbonne-universite.fr

[2] Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark, werge@sdu.dk

Abstract

Stochastic optimization methods encounter new challenges in the realm of streaming, characterized by a continuous flow of large, high-dimensional data. While first-order methods, like stochastic gradient descent, are the natural choice, they often struggle with ill-conditioned problems. In contrast, second-order methods, such as Newton's methods, offer a potential solution, but their computational demands render them impractical. This paper introduces adaptive stochastic optimization methods that bridge the gap between addressing ill-conditioned problems while functioning in a streaming context. Notably, we present an adaptive inversion-free Newton's method with a computational complexity matching that of first-order methods, $\mathcal{O}(dN)$, where d represents the number of dimensions/features, and N the number of data. Theoretical analysis confirms their asymptotic efficiency, and empirical evidence demonstrates their effectiveness, especially in scenarios involving complex covariance structures and challenging initializations. In particular, our adaptive Newton's methods outperform existing methods, while maintaining favorable computational efficiency.

Keywords: stochastic optimization, adaptive methods, Newton's method, online learning, large-scale, streaming

1 Introduction

The focus of this paper is on the stochastic optimization problem, where the objective is to minimize a convex function $F : \mathbb{R}^d \to \mathbb{R}$ with $d \in \mathbb{N}$. The problem is formulated as follows:

$$\min_{\theta \in \mathbb{R}^d} \{F(\theta) := \mathbb{E}_{\xi \sim \Xi}[f(\theta; \xi)]\}, \tag{1}$$

where f is a loss function, ξ is a random variable following an unknown distribution Ξ, and θ is the parameter of interest. The challenge in (1) is widespread in machine learning applications (Kushner and Yin, 2003; Shapiro et al., 2021; Bottou et al., 2018; Sutton and Barto, 2018). For instance, in the context of an input-output pair $\xi = (x, y)$, the function f typically takes the form $f(\theta; \xi) = f(\theta; x, y) = l(h_\theta(x); y)$, where l is a loss function onto $\mathbb{R}$ and h_θ is a prediction model parameterized by θ.

We address the stochastic optimization problem (1) within a streaming context, where data are both large in size and dimensionality. Similar to prior work by Godichon-Baggioni et al. (2023b,a), streaming data continuously arrives in blocks, resembling time-varying mini-batches, as independent and identically distributed (i.i.d.) samples of the random variable ξ. More formally, we consider an endless sequence of i.i.d. copies: $\{\xi_{1,1}, \ldots, \xi_{1,n_1}\}, \ldots, \{\xi_{t,1}, \ldots, \xi_{t,n_t}\}, \ldots$, where $\{\xi_{t,1}, \ldots, \xi_{t,n_t}\}$ represents a block of n_t data points arriving at time t. This setup mirrors the incremental and block-based nature of real-world streaming data.

Our adaptive stochastic optimization methods go beyond the conventional stochastic gradient-based methods by incorporating a Hessian matrix approximation A_t at each step t to refine the descent direction. In a general form, these adaptive methods can be expressed recursively as:

$$\theta_{t+1} = \theta_t - \gamma_{t+1} A_t \nabla_\theta f(\theta_t; \xi_{t+1}), \quad \theta_0 \in \mathbb{R}^d, \tag{2}$$

where $\nabla_\theta f(\theta_t; \xi_{t+1}) = n_{t+1}^{-1} \sum_{i=1}^{n_{t+1}} \nabla_\theta f(\theta_t; \xi_{t+1,i})$, $(\nabla_\theta f(\theta_t; \xi_{t+1,i}))$ is unbiased gradients in $\mathbb{R}^d$, (A_t) is a sequence of random matrices in $\mathbb{R}^{d\times d}$, and (γ_t) is the learning rate.

Specially, if we set $A_t = \mathbb{I}_d$ and $n_t = 1$, the update in (2) reduces to the classical Robbins-Monro method (Robbins and Monro, 1951), commonly known as Stochastic Gradient Descent (SGD). When $n_t \in \mathbb{N}$ (with $A_t = \mathbb{I}_d$), we obtain a streaming version of SGD, akin to time-varying mini-batch SGD, as considered in Godichon-Baggioni et al. (2023b,a). For Adagrad (Duchi et al., 2011), A_t serves as an estimate of the inverse square root of the diagonal of the variance of the gradients. Furthermore, the update in (2) transforms into Newton's method, when A_t serves as an approximation of the inverse Hessian matrix $\nabla_\theta^2 F(\theta_t)$.

The central question in this paper is twofold: *Can we construct a sequence of Hessian approximations (A_t) in a manner that is both computationally efficient and ensures the robustness of our adaptive methods to ill-conditioned problems?*

Contributions. In this work, we present adaptive stochastic optimization methods capable of robustly handling ill-conditioned problems while ensuring computational efficiency in streaming contexts. These adaptive methods dynamically adjust learning per-dimension, leveraging historical gradient and Hessian information. Additionally, we propose iterative weighted average versions of our adaptive methods. These acceleration techniques both provide variance-reduction during learning and accelerated convergence. Theoretical analysis establishes their asymptotic efficiencies, encompassing strong consistency, rate of convergence, and asymptotic normality. Empirical evidence further validates their effectiveness, particularly in scenarios with complex covariance structures and challenging initializations.

A noteworthy contribution of our work is the introduction of inversion-free adaptive Newton's methods, designed to match the computational complexity of first-order methods—$\mathcal{O}(dN_t)$, where $N_t = \sum_{i=1}^t n_i$ is the total quantity of data up to time t. These adaptive Newton's methods not only achieve the computational efficiency of first-order methods but also incorporate acceleration techniques for enhanced convergence, while harnessing the power of second-order information.

Related work. Stochastic optimization and adaptive methods have been extensively researched, as evident in works such as Bottou et al. (2018); Chau et al. (2022). Theoretical investigations into SGD span topics from in-depth non-asymptotic analysis to its asymptotic efficiency (Moulines and Bach, 2011; Kushner and Yin, 2003; Toulis and Airoldi, 2017; Pelletier, 1998; Fabian, 1968; Pelletier, 2000; Gadat and Gavra, 2022; Nemirovski et al., 2009; Lacoste-Julien et al., 2012). A noteworthy extension of SGD is the concept of averaging, known for its role in accelerating convergence. This averaging scheme, referred to as Polyak-Ruppert averaging or averaged SGD (ASGD), was introduced by Ruppert (1988); Polyak and Juditsky (1992). They demonstrated that using a learning rate with slower decays, combined with uniform averaging, robustly leads to information-theoretically optimal asymptotic variance. While these estimates are known to be asymptotically efficient (Pelletier, 2000), their non-asymptotic properties have been thoroughly investigated (Moulines and Bach, 2011; Needell et al., 2014; Gadat and Panloup, 2023). However, it's important to note that this averaging concept can be sensitive to ill-conditioned problems among others, leading to sub-optimal performance in practice (Leluc and Portier, 2023; Boyer and Godichon-Baggioni, 2023).

To address this practical challenge, recent strategies have emerged to enhance the performance of stochastic optimization methods, focusing on adaptive approaches. These methods involve tuning the learning rate, also known as the step-size sequence, through strategies that adapt to the gradient. One of the most well-known adaptive techniques is Adagrad (Duchi et al., 2011), which incorporates an estimation of the square root of the inverse of the gradient's covariance into the step-size. Subsequently, this method has undergone various modifications and improvements. Notable among these adaptations are RMSProp (Tieleman and Hinton, 2012), ADAM (Kingma and Ba, 2015), AdaDelta (Zeiler, 2012), NADAM (Dozat, 2016), and AMSGrad (Reddi et al., 2018). Nevertheless, these adaptive methods do not fully tackle the challenge of poor conditioning. Another limitation of these methods is their reliance on information solely from the diagonal of the gradient covariance estimator. Consequently, in scenarios with strong correlations, this restricted information may result in sub-optimal practical outcomes.

To address this, an alternative approach involves considering inversion-free stochastic Newton's methods (Bercu et al., 2020; Boyer and Godichon-Baggioni, 2023; Leluc and Portier, 2023), where an estimate of the inverse of the Hessian is integrated into the step-size. Alternatively, stochastic Gauss-Newton methods (Cénac

et al., 2020; Bercu et al., 2023) can be employed. These stochastic Newton's methods, relying on the Sherman-Morrison formula (Sherman and Morrison, 1950),[1] require a specific form of the Hessian. Nevertheless, they find applications in various scenarios, including linear, logistic, softmax, and ridge regressions (Bercu et al., 2020; Boyer and Godichon-Baggioni, 2023; Godichon-Baggioni et al., 2024), as well as tasks such as estimation of the geometric median (Godichon-Baggioni and Lu, 2023), non-linear regression (Cénac et al., 2020), and optimal transport (Bercu et al., 2023).

Our adaptive stochastic optimization methods aim to integrate the strengths of acceleration techniques (weighted Polyak-Ruppert averaging), adaptive methods, and stochastic Newton's methods. We believe that this integration provides an effective solution to solving the challenges posed by ill-conditioned problems in a streaming context.

Organization. This paper is organized as follows: Section 2 presents the underlying theoretical framework within we analyse our adaptive stochastic optimization methods. In Section 3, we analyse our adaptive stochastic optimization methods and its weighted averaged version in Section 4. In Section 5, we apply our adaptive methods to Adagrad and Newton's method. In particular, Section 5 details the construction of our adaptive Newton's methods with $\mathcal{O}(dN_t)$ operations. In Section 6, we present our experimental results, demonstrating the efficiency of our proposed methods.

Notations. We represent the Euclidean norm as $\|\cdot\|$ and the operator norm as $\|\cdot\|_{\text{op}}$. The notation $M \succ 0$ indicates that M is positive definite, while $M \succeq 0$ indicates that it is positive semi-definite. The minimum and maximum eigenvalues of matrix M are denoted by $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$, respectively.

2 Underlying Theoretical Framework

In this section, we provide the theoretical framework that underpins our analysis. Our objective is to solve the stochastic optimization problem in (1), while operating within a streaming contexts. As a reminder, we consider stochastic optimization problems of the form:

$$\min_{\theta \in \mathbb{R}^d} \{F(\theta) := \mathbb{E}_{\xi \sim \Xi}[f(\theta; \xi)]\},$$

where ξ is a random variable sampled following an unknown distribution Ξ.

To lay the foundation for our analyses, we introduce three key assumptions. These assumptions, contingent upon the differentiability of the function F, serve as the bedrock for our theoretical framework. These assumptions are standard in the realms of stochastic optimization, stochastic approximation, and adaptive methods (Bottou et al., 2018; Leluc and Portier, 2023; Boyer and Godichon-Baggioni, 2023; Kushner and Yin, 2003; Godichon-Baggioni, 2019b,a; Benveniste et al., 1990; Duflo, 2013; Godichon-Baggioni and Tarrago, 2023).

Assumption 1 *For almost any ξ, the function $f(\cdot; \xi)$ is differentiable and there exists non-negative constants C and C' for all $\theta \in \mathbb{R}^d$, such that*

$$\mathbb{E}[\|\nabla_\theta f(\theta; \xi)\|^2] \leq C + C'(F(\theta) - F(\theta^*)). \tag{3}$$

In addition, there exists $\theta^* \in \mathbb{R}^d$ such that $\nabla_\theta F(\theta^*) = 0$, and the functional $\Sigma : \theta \to \mathbb{E}[\nabla_\theta f(\theta; \xi) \nabla_\theta f(\theta; \xi)^\top]$ is continuous at θ^*.

In Assumption 1, we do not confine $\mathbb{E}[\|\nabla_\theta f(\theta; \xi)\|^2]$ by a constant or the squared errors $\|\theta - \theta\|^2$. Instead, we utilize the functional error $F(\theta) - F(\theta^*)$, a condition known as expected smoothness (Gower et al., 2019; Gazagnadou et al., 2019; Gower et al., 2021). Moreover, when $C = 0$, (3) is known as the weak growth condition (Vaswani et al., 2019; Nguyen et al., 2018). Notably, it is worth mentioning that, in the context of μ-strong convexity of the functional F, the squared errors condition implies the functional error condition, as $\|\theta - \theta^*\|^2 \leq 2/\mu (F(\theta) - F(\theta^*))$ for any $\theta \in \mathbb{R}^d$.

In order to ensure the strong consistency of our method's estimates, we invoke a second key assumption. This assumption allows the use of a second-order Taylor expansion of the functional G and is based on the following hypothesis:

[1] Sherman-Morrison's formula is also known as Riccati's equation for matrix inversion (Duflo, 2013).

Assumption 2 *The functional F is twice-continuously differentiable with uniformly bounded Hessian, i.e., there exists $L_{\nabla F}$ such that $\|\nabla^2_\theta F(\theta)\|_{\text{op}} \leq L_{\nabla F}$.*

Note that this implies, among other things, that the gradient of F is $L_{\nabla F}$-Lipschitz. The third assumption pertains to the uniqueness of the minimizer θ^* of the functional F.

Assumption 3 *The functional F is locally strongly convex; $\lambda_{\min} := \lambda_{\min}(\nabla^2_\theta F(\theta^*)) > 0$.*

3 Adaptive Stochastic Optimization Methods

For clarity, our main discussion revolves around constant mini-batches of size n (instead of time-varying mini-batches n_t). This approach enables us to intricately address the nuances of the streaming data setting while upholding the conceptual clarity of our core findings. However, it's crucial to emphasize that we offer translations and adaptations of these discussions for the scenario of time-varying mini-batches in Appendix A. The motivation for considering time-varying mini-batches stems from recent work by Godichon-Baggioni et al. (2023a,b), which demonstrated that increasing mini-batches can accelerate convergence and break long- and short-term dependence structures.

Throughout the paper, we consider constant mini-batches of size n, i.e., at each time t, n i.i.d copies of ξ denoted by $\xi_t = \{\xi_{t,1}, \ldots, \xi_{t,n}\}$ arrives. Our adaptive stochastic optimization methods, as defined in (2), can recursively be written as

$$\theta_{t+1} = \theta_t - \gamma_{t+1} A_t \nabla_\theta f(\theta_t; \xi_{t+1}), \quad \theta_0 \in \mathbb{R}^d,$$

where $\nabla_\theta f(\theta_t; \xi_{t+1}) = n^{-1} \sum_{i=1}^n \nabla_\theta f(\theta_t; \xi_{t+1,i})$. We assume the construction of a filtration $(\mathcal{F}_t)$ such that θ_t and A_t are $\mathcal{F}_t$-measurable, and $\xi_{t+1} = \{\xi_{t+1,1}, \ldots, \xi_{t+1,n}\}$ are independent from $\mathcal{F}_t$. Let N_t denote the total number of data processed up to time t, i.e., $N_t = nt$.

Our goal is to recursively update θ_t at each time step t to integrate the most recent information. For the subsequent discussion, we assume that the learning rate (γ_t) and the sequence of random matrices (A_t) satisfy the following conditions:

$$\sum_{t \geq 1} \gamma_t \lambda_{\min}(A_{t-1}) = +\infty \text{ a. s.}, \quad \text{and} \quad \sum_{t \geq 1} \gamma_t^2 \lambda_{\max}(A_{t-1})^2 < +\infty \text{ a. s.} \tag{4}$$

In Section 5, we will delve into the modifications required in the methods to fulfill these conditions. Additional insights can be found in works such as Boyer and Godichon-Baggioni (2023); Godichon-Baggioni and Tarrago (2023). In all the sequel, we take $\gamma_t = C_\gamma t^{-\gamma}$ with $C_\gamma > 0$ and $\gamma \in (1/2, 1)$ for the sake of simplicity. However, one can also take $\gamma_t = C_\gamma (t + t_0)^{-\gamma}$ with $t_0 \in \mathbb{N}$, and all the theoretical results remain true.

The following theorem establishes the strong consistency of our adaptive stochastic gradient estimates (θ_t) derived from (2).

Theorem 1 *Suppose Assumptions 1 to 3 hold, along with the conditions in* (4)*. Then, θ_t converges almost surely to θ^*.*

To ascertain the rate of convergence of our adaptive stochastic gradient estimates (θ_t), we assume that the sequence of random matrices A_t converges to $A \succ 0$.

Assumption 4 *The random matrix A_t converges almost surely to a positive definite matrix A.*

For instance, in Newton's methods, the matrix A represents the inverse Hessian, and in the case of Adagrad, it corresponds to the inverse of the square root of the diagonal of the gradient's variance. Note that once Theorem 1 is fulfilled, the strong consistency of θ_t often implies the consistency of A_t.

Theorem 2 *Suppose Assumptions 1 to 4 hold, along with the conditions in* (4)*. In addition, assume there exist positive constants C_η and $\eta > \frac{1}{\gamma} - 1$ such that for all $\theta \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\|\nabla_\theta f(\theta; \xi)\|^{2+2\eta}\right] \leq C_\eta \left(1 + F(\theta) - F(\theta^*)\right)^{1+\eta}. \tag{5}$$

Then,

$$\|\theta_t - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t^\gamma}\right) \text{ a. s.}$$

4 The Weighted Averaged Version

The weighted averaged version of our adaptive stochastic optimization methods in (2) is defined for $w \geq 0$ as follows:

$$\theta_{t,w} = \frac{1}{\sum_{i=0}^{t-1} \ln(i+1)^w} \sum_{i=0}^{t-1} \ln(i+1)^w \theta_i, \tag{6}$$

which can be written recursively as

$$\theta_{t+1,w} = \theta_{t,w} + \frac{\ln(t+1)^w}{\sum_{i=0}^{t} \ln(i+1)^w}(\theta_t - \theta_{t,w}).$$

This weighted averaging in (6) enhances the optimization by adaptively assigning more weight to the latest estimates of (θ_t). The logarithmic weighting emphasizes recent estimates, presumed to be more accurate, while providing robustness against sub-optimal initializations (Mokkadem and Pelletier, 2011; Boyer and Godichon-Baggioni, 2023). Observe that taking $w = 0$ leads to the usual Polyak-Ruppert averaging scheme (Ruppert, 1988; Polyak and Juditsky, 1992; Godichon-Baggioni et al., 2023b).

To establish the convergence rate of the weighted averaged version of our adaptive stochastic optimization methods in (2), we begin by introducing a new assumption.

Assumption 5 *There exists positive constants L_η and η such that for all $\theta \in \mathcal{B}(\theta^*, \eta)$,*

$$\|\nabla_\theta F(\theta) - \nabla^2_\theta F(\theta^*)(\theta - \theta^*)\| \leq L_\eta \|\theta - \theta^*\|^2.$$

Assumption 5 is satisfied as soon as the Hessian of F is locally Lipschitz on a neighborhood around θ^*. Coupled with Assumption 2, this imply there exists a positive constant L_δ such that for all $\theta \in \mathbb{R}^d$,

$$\|\nabla_\theta F(\theta) - \nabla^2_\theta F(\theta^*)(\theta - \theta^*)\| \leq L_\delta \|\theta - \theta^*\|^2.$$

The following result establish the rate of convergence and the optimal asymptotic normality of the weighted averaged estimates $(\theta_{t,w})$.

Theorem 3 *Suppose Assumptions 1 to 5 hold, along with inequality* (5)*. In addition, assume there exists a positive constant ν such that*

$$\|A_t - A\|_{\text{op}} = \mathcal{O}\left(\frac{1}{t^\nu}\right) \text{ a. s.} \tag{7}$$

Then,

$$\|\theta_{t,w} - \theta^*\|^2 = \begin{cases} \mathcal{O}\left(\frac{\ln(N_t)}{N_t^{\gamma+2\nu}}\right) \text{ a. s.} & \text{if } 2\nu + \gamma \leq 1, \\ \mathcal{O}\left(\frac{\ln(N_t)}{N_t}\right) \text{ a. s.} & \text{if } 2\nu + \gamma > 1. \end{cases}$$

Moreover, if $2\nu + \gamma > 1$, then

$$\sqrt{N_t}(\theta_{t,w} - \theta^*) \xrightarrow[t\to+\infty]{\mathcal{L}} \mathcal{N}(0, \nabla^2_\theta F(\theta^*)^{-1} \Sigma \nabla^2_\theta F(\theta^*)^{-1}).$$

To establish strong results, such as the asymptotic efficiency of the weighted average estimates $(\theta_{t,w})$, the sequence of random matrices A_t should exhibit a (weak) rate of convergence, as outlined in (7). In simpler terms, achieving a satisfactory rate of convergence of A_t leads to the asymptotic efficiency of the weighted average estimates $(\theta_{t,w})$.

However, to establish asymptotic efficiency without relying on a (weak) rate of convergence of A_t, one can also consider the following theorem:

Theorem 4 *Suppose Assumptions 1 to 5 hold, along with inequality* (5)*. In addition, assume there exists a positive constant $v' > 1/2$ such that*

$$\frac{1}{\sum_{i=0}^{t-1} \ln(i+1)^w} \sum_{i=0}^{t-1} \ln(i+1)^{w+1/2+\delta} \|A_{i+1}^{-1} - A_i^{-1}\|_{\text{op}} (i+1)^{\frac{\gamma}{2}} = \mathcal{O}\left(\frac{1}{t^{v'}}\right) \text{ a. s.}, \tag{8}$$

for some $\delta > 0$. Then

$$\|\theta_{t,w} - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t}\right) \text{ a. s.} \quad \text{and} \quad \sqrt{N_t}(\theta_{t,w} - \theta^*) \xrightarrow[t\to+\infty]{\mathcal{L}} \mathcal{N}(0, \nabla^2_\theta F(\theta^*)^{-1} \Sigma \nabla^2_\theta F(\theta^*)^{-1}).$$

Note that, while condition (8) may appear unusual, it is straightforward to verify in practice. The proofs of Theorems 6 and A.7 provide insights into practical methods for checking this condition.

5 Applications to Newton's Method

In this section, we apply our adaptive stochastic optimization methodology, as detailed in Sections 3 and 4, to (stochastic) Newton's methods. Here, we present inversion-free adaptive Newton's methods explicitly designed to align with the computational complexity of first-order optimization methods. Specifically, we present a weighted average inversion-free adaptive Newton method that seamlessly integrates the strengths of both approaches. Additionally, it's worth noting that we propose a novel streaming variant of Adagrad, along with its weighted average counterpart, in Appendix A.4.

To overcome the computational challenges linked to Hessian inversion, we propose a variant of the stochastic Newton's method that entirely avoids Hessian inversion. In Section 5.1, our adaptive stochastic optimization methodology is applied to the stochastic Newton's method, resulting in a direct streaming stochastic Newton's method requiring $\mathcal{O}(d^2 N_t)$ operations. Next, in Section 5.2, we introduce a weighted Hessian estimate demanding only $\mathcal{O}(dN_t)$ operations. Finally, in Section 5.3, we accelerate this stochastic Newton's method through weighted Polyak-Ruppert averaging.

5.1 Direct Streaming Stochastic Newton's Method

In the special case of stochastic Newton's methods, one can obtain the asymptotic efficiency without averaging by taking a step sequence of the form $\gamma_t = 1/t$.[2] The streaming stochastic Newton algorithm is defined by the update:

$$\theta_{t+1} = \theta_t - \frac{1}{t+1} \bar{H}_t^{-1} \nabla_\theta f(\theta_t; \xi_{t+1}), \quad \theta_0 \in \mathbb{R}^d, \tag{9}$$

where $\nabla_\theta f(\theta_t; \xi_{t+1}) = n^{-1} \sum_{i=1}^n \nabla_\theta f(\theta_t; \xi_{t+1,i})$ and $\bar{H}_t$ is an approximation of the Hessian of F. Specifically, we consider Hessian estimates $\bar{H}_t = N_t^{-1} H_t$ of the form

$$H_t = H_0 + \sum_{i=1}^{t} \sum_{j=1}^{n} \alpha_{i,j} \Phi_{i,j} \Phi_{i,j}^\top,$$

with H_0 symmetric and positive, $\alpha_{i,j} = \alpha(\theta_{i-1}; \xi_{i,j})$, and $\Phi_{i,j} = \Phi(\theta_{i-1}; \xi_{i,j})$. A computationally-efficient estimate of H_t inverse can be derived using the Riccati/Sherman-Morrisson's formula (Duflo, 2013; Sherman and Morrison, 1950):

$$H_t^{-1} = H_{t-1}^{-1} - \sum_{j=1}^{n} \alpha_{t,j} \left(1 + \alpha_{t,j} \Phi_{t,j}^\top H_{t-1}^{-1} \Phi_{t,j}\right)^{-1} H_{t-1}^{-1} \Phi_{t,j} \Phi_{t,j}^\top H_{t-1}^{-1}.$$

The explicit construction of the recursive estimates of the inverse Hessian is detailed in various applications, including linear, logistic, softmax, and ridge regressions (Bercu et al., 2020; Boyer and Godichon-Baggioni, 2023; Godichon-Baggioni et al., 2024). Additionally, these methods are applied to tasks such as the estimation of the geometric median (Godichon-Baggioni and Lu, 2023), non-linear regression (Cénac et al., 2020), and optimal transport (Bercu et al., 2023).

The asymptotic efficiency of the streaming version of the stochastic Newton's method can now be articulated as follows:

Theorem 5 *Suppose Assumptions 1, 2, 3, and 5 hold, along with inequality* (5)*. Then, θ_t converges almost surely to θ^*. In addition, assume $\bar{H}_t^{-1}$ converges almost surely to $\nabla_\theta^2 F(\theta^*)^{-1}$. Then,*

$$\|\theta_t - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t}\right) \text{ a. s.}$$

Moreover, assume there exists a positive constant ν such that $\|\bar{H}_t^{-1} - \nabla_\theta^2 F(\theta^*)^{-1}\|_{\text{op}} = \mathcal{O}(1/t^\nu)$ a. s.. Then

$$\sqrt{N_t}(\theta_t - \theta^*) \xrightarrow[t \to +\infty]{\mathcal{L}} \mathcal{N}(0, \nabla_\theta^2 F(\theta^*)^{-1} \Sigma \nabla_\theta^2 F(\theta^*)^{-1}).$$

[2] Observe, in the increasing batch-size case in Appendix A, one should take $\gamma_t = n_t / N_t$.

5.2 Streaming Stochastic Newton's methods with possibly $\mathcal{O}(dN_t)$ operations

The direct stochastic Newton's method presented in Section 5.1 is associated with computational costs of $\mathcal{O}(d^2N_t)$, which can be computationally expensive, especially in high-dimensional streaming settings. To address this challenge, we introduce the streaming stochastic Newton's method using weighted Hessian estimates:

$$\theta_{t+1} = \theta_t - \frac{1}{t+1}\bar{H}_{t,w'}^{-1}\nabla_\theta f(\theta_t;\xi_{t+1}), \quad \theta_0 \in \mathbb{R}^d, \tag{10}$$

where $\nabla_\theta f(\theta_t;\xi_{t+1}) = n^{-1}\sum_{i=1}^n \nabla_\theta f(\theta_t;\xi_{t+1,i})$ and $\bar{H}_{t,w'} = N_{t,Z}^{-1}H_{t,w'}$ with

$$H_{t,w'} = H_{0,w'} + \sum_{i=1}^{t}\ln(i+1)^{w'}\sum_{j=1}^{n} Z_{i,j}\left(\iota_{i,j}\tilde{e}_{i,j}\tilde{e}_{i,j}^\top + \alpha_{i,j}\Phi_{i,j}\Phi_{i,j}^\top\right), \tag{11}$$

with $N_{t,Z} = 1+\sum_{i=1}^t \ln(i+1)^{w'}\sum_{j=1}^n Z_{i,j}$, H_0 symmetric and positive, $w' \geq 0$, and $Z_{i,j}$ are i.i.d with $Z_{i,j} \sim \mathcal{B}(p)$ for some $p \in (0,1]$. In addition, let $N_{t,k,Z} = (1+\sum_{i=1}^{t-1}\sum_{j=1}^n Z_{i,j} + \sum_{j=1}^k Z_{t,j})$, $\iota_{i,j} = c_\iota N_{i,j,Z}^{-\iota}$ for $\iota \in (0,1/2)$, and $e_{i,j}$ be the $(N_{i,j,Z}$ modulo $d+1)$-th component of the canonical basis. Observe that the term ι_t enables to control the smallest eigenvalue of $\bar{H}_{t,w'}$ while $\ln(t+1)^{w'}$ enables us to give more weights to the lastest updates $\alpha_{t,j}\Phi_{t,j}\Phi_{t,j}^\top$, which are supposed to be better since (θ_t) converges to θ^*. The add of the random variables $Z_{i,j}$ enables us to play with the computational cost. As explain later, taking $p = 1$ leads to a natural recursive estimate of the Hessian in Section 5.1.

We now discuss about $Z_{i,j}$. Let us recall that with the help of Riccati's formula (Duflo, 2013), one can update the inverse of $H_{t+1,w'}$ as follows:

$$\begin{cases}
H_{t+\frac{1}{n},w'}^{-1} &= H_{t,w'}^{-1} - \frac{Z_{t+1,1}\iota_{t+1,1}}{1+\iota_{t+1,1}e_{t+1,1}\bar{H}_{t,w'}^{-1}e_{t+1,1}^T}H_{t,w'}^{-1}e_{t+1,1}e_{t+1,1}^T H_{t,w'}^{-1} \\
&\vdots \\
H_{t+1,w'}^{-1} &= H_{t+\frac{n-1}{n},w'}^{-1} - \frac{Z_{t+1,n}\iota_{t+1,n}}{1+\iota_{t+1,n}e_{t+1,n}H_{t,w'}^{-1}e_{t+1,n}^T}H_{t,w'}^{-1}e_{t+1,n}e_{t+1,n}^T H_{t,w'}^{-1} \\
H_{t+\frac{1}{n},w'}^{-1} &= H_{t+1,w'}^{-1} - \frac{Z_{t+1,1}\ln(t+1)^{w'}\alpha_{t+1,1}}{1+\ln(t+1)^{w'}\alpha_{t+1,1}\Phi_{t+1,1}^T H_{t+1,w'}^{-1}\Phi_{t+1,1}}H_{t+1,w'}^{-1}\Phi_{t+1,1}\Phi_{t+1,1}^T H_{t+1,w'}^{-1} \\
H_{t+\frac{2}{n},w'}^{-1} &= H_{t+\frac{1}{n},w'}^{-1} - \frac{Z_{t+1,2}\ln(t+1)^{w'}\alpha_{t+1,2}}{1+\ln(t+1)^{w'}\alpha_{t+1,2}\Phi_{t+1,2}^T H_{t+\frac{1}{n},w'}^{-1}\Phi_{t+1,2}}H_{t+\frac{1}{n},w'}^{-1}\Phi_{t+1,2}\Phi_{t+1,2}^T H_{t+\frac{1}{n},w'}^{-1} \\
&\vdots \\
H_{t+1,w'}^{-1} &= H_{t+\frac{n-1}{n},w'}^{-1} - \frac{Z_{t+1,n}\ln(t+1)^{w'}\alpha_{t+1,n}}{1+\ln(t+1)^{w'}\alpha_{t+1,n}\Phi_{t+1,n}^T H_{t+\frac{n-1}{n},w'}^{-1}\Phi_{t+1,n}}H_{t+\frac{n-1}{n},w'}^{-1}\Phi_{t+1,n}\Phi_{t+1,n}^T H_{t+\frac{n-1}{n},w'}^{-1}.
\end{cases}$$

Then, the update of H_{t+1}^{-1} only costs, in average, $\mathcal{O}(pd^2n)$ operations leading to a total number of operations of order (in average);

$$\underbrace{pd^2N_t}_{\text{estimating the inverse Hessian}} + \underbrace{dN_t}_{\text{estimating the gradient}} + \underbrace{\frac{d^2N_t}{n}}_{\text{multiplication of Hessian and gradient estimates}}.$$

Thus, one can play with the value of p to reduce the cost of the update of the inverse of the Hessian. Indeed, one can obtain an averaged computational cost at time t of order $\mathcal{O}(dN_t)$ operations taking $p = d^{-1}$ and $n = d$. In other words, it is possible to obtain a stochastic Newton's method with only $\mathcal{O}(dN_t)$ operations, which still is asymptotically efficient. In all the sequel, for the sake of simplicity, we suppose that for any $\theta \in \mathbb{R}^d$,

$$\nabla_\theta^2 F(\theta) = \mathbb{E}\left[\alpha(\theta;\xi)\Phi(\theta;\xi)\Phi(\theta;\xi)^T\right]. \tag{12}$$

Theorem 6 *Suppose Assumptions 1 to 3 and 5 hold, along with inequalities* (5) *and* (12)*. In addition, assume that for almost any ξ, there exists positive constants $C_{\eta'}$ and $\eta' > 1$ such that for all $\theta \in \mathbb{R}^d$,*

$$\mathbb{E}[\|\alpha(\theta;\xi)\Phi(\theta;\xi)\Phi(\theta;\xi)^\top\|^{\eta'}] \leq C_{\eta'}^{\eta'}.$$

Then,

$$\|\theta_t - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t}\right) \text{ a. s.}$$

Moreover, suppose that the Hessian of F is locally $L_{\nabla^2 F}$-Lipschitz on a neighborhood around θ^* and that $\eta' \geq 2$. Then

$$\sqrt{N_t}(\theta_t - \theta^*) \xrightarrow[t\to+\infty]{\mathcal{L}} \mathcal{N}(0, \nabla_\theta^2 F(\theta^*)^{-1} \Sigma \nabla_\theta^2 F(\theta^*)^{-1}).$$

Observe that contrary to Theorem 3, no restriction on ν is necessary.

5.3 Weighted Averaged Version of Streaming Stochastic Newton's methods with possibly $\mathcal{O}(dN_t)$ operations

Although the streaming Newton's methods is very performant, it can be quite sensitive to bad initialization since the learning rate can be too small (Boyer and Godichon-Baggioni, 2023). In order to overcome this, one can consider the weighted averaged version, given by

$$\theta_{t+1} = \theta_t - \gamma_{t+1} \bar{S}_{t,w'}^{-1} \nabla_\theta f(\theta_t; \xi_{t+1}),\ \theta_0 \in \mathbb{R}^d, \tag{13}$$

$$\theta_{t+1,w} = \theta_{t,w} + \frac{\ln(t+1)^w}{\sum_{i=0}^t \ln(i+1)^w}(\theta_t - \theta_{t,w}), \tag{14}$$

where $\nabla_\theta f(\theta_t; \xi_{t+1}) = n^{-1} \sum_{i=1}^n \nabla_\theta f(\theta_t; \xi_{t+1,i})$, $\gamma_t = C_\gamma t^{-\gamma}$, and $\bar{S}_{t,w'} = N_{t,Z}^{-1} S_{t,w'}$ with

$$S_{t,w'} = S_{0,w'} + \sum_{i=1}^t \ln(i+1)^{w'} \sum_{j=1}^n Z_{i,j} \left(\iota_i e_{i,j} e_{i,j}^\top + \alpha_{i,j} \Phi_{i,j} \Phi_{i,j}^\top\right),$$

with $S_{0w'}$ symmetric and positive, $N_{t,Z} = 1 + \sum_{i=1}^t \ln(i+1)^{w'} \sum_{j=1}^n Z_{i,j}$, $\iota_i = c_\iota N_{i,j,Z}^{-\iota}$, and $\iota \in (0, \gamma - 1/2)$. One can follow the same recursive scheme as for $H_{t,w'}^{-1}$ to update the inverse of $S_{t,w'}^{-1}$. Indeed, the only difference between $S_{t,w'}^{-1}$ and $H_{t,w'}^{-1}$ is the choice of the estimate of θ^*.

Theorem 7 *Suppose Assumptions 1 to 3 and 5 hold, along with inequalites* (5) *and* (12)*. In addition, assume that for almost any ξ, there exists positive constants $C_{\eta'}$ and $\eta' > 1$ such that for all $\theta \in \mathbb{R}^d$,*

$$\mathbb{E}[\|\alpha(\theta;\xi)\Phi(\theta;\xi)\Phi(\theta;\xi)^\top\|^{\eta'}] \leq C_{\eta'}^{\eta'}.$$

Then,

$$\|\theta_t - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t^\gamma}\right) \text{ a. s.}, \qquad \|\theta_{t,w} - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t}\right) \text{ a. s.},$$

and

$$\sqrt{N_t}(\theta_{t,w} - \theta^*) \xrightarrow[t\to+\infty]{\mathcal{L}} \mathcal{N}(0, \nabla_\theta^2 F(\theta^*)^{-1} \Sigma \nabla_\theta^2 F(\theta^*)^{-1}).$$

6 Experiments

In this section, we empirically evaluate our adaptive stochastic optimization methods, focusing on two fundamental problems in statistical optimization: least-squares regression and logistic regression. For least-squares regression, our data points are represented as $\xi = (x, y) \in \mathbb{R}^d \times \mathbb{R}$, and we employ the functional $F(\theta) = \frac{1}{2}\mathbb{E}[(y - x^\top\theta)^2]$. In the case of logistic regression, our data points are $\xi = (x, y) \in \mathbb{R}^d \times \{-1, 1\}$, and the corresponding functional is $F(\theta) = \mathbb{E}[\ln(1 + \exp(x^\top\theta)) - yx^\top\theta]$.

To introduce complex covariance structures into our datasets, we adopt the experimental framework detailed in Boyer and Godichon-Baggioni (2023). This involves modeling the covariance of our feature vector x as follows:

$$x \sim \mathcal{N}\left(0, M \operatorname{diag}\left(\frac{i^2}{d^2}\right)_{i=1,\dots,d} M^\top\right).$$

Here, M represents a random orthogonal matrix. This choice of covariate distribution, influenced by the action of M, allows us to introduce strong correlations between the coordinates of x. This variation in data structure enables us to assess the adaptability of our method under diverse conditions.

For our experiments, we deliberately set $d = 100$ to emphasize the challenges posed by high dimensionality. This choice serves to highlight the scalability and robustness of our proposed methods when handling large-dimensional datasets. Within this setting, the Hessian associated with the model exhibits a wide range of eigenvalues, with the largest eigenvalue being a thousand times larger than the smallest one.

The weight parameters for both the estimates and Hessian approximations are set to 2, i.e., $w = w' = 2$. While higher values of w and w' would enhance adaptability, the chosen setting serves as a proof-of-concept.

We set the mini-batch size n equal to the dimension d, which has the implication that our adaptive Newton's method, incorporating second-order information, updates the Hessian less frequently than the first-order SGD and Adagrad algorithms.[3] Surprisingly, our results clearly demonstrate that despite fewer Hessian updates, the Newton's method perform exceptionally well. Notably, when dealing with highly correlated data, the Adagrad algorithm's adaptive step-size becomes less effective, whereas the Newton's methods excel. Particularly, in scenarios involving less-than-ideal initializations (as depicted on the right side of the figures), both Newton's methods demonstrate outstanding performance.

Leveraging this setup, we demonstrate the adaptability of our methods when dealing with high-dimensional datasets featuring complex covariance structures. Our experiments underscore the efficiency of our adaptive Newton's method in comparison to first-order gradient methods and highlight our methods' state-of-the-art performance in terms of both convergence speed and accuracy when compared to existing methods.

In our applications, we investigate various optimization methods, encompassing SGD, Adagrad, our streaming Adagrad detailed in Appendix A.4, along with their weighted Polyak-Ruppert averages. Additionally, we explore our streaming stochastic Newton (SSN) from Section 5.2 and our weighted Polyak-Ruppert averaged streaming stochastic Newton (WASSN) from Section 5.3; for SNN and WASSN, we set $p = 1$ (i.e., $\mathcal{O}(d^2 N_t)$ computations) and $p = 1/d$ (i.e., $\mathcal{O}(dN_t)$ computations) to explore the loss of not updating the hole Hessian at each step.

6.1 Least-Squares Regression

In the context of least-squares regression, we aim to evaluate the performance of our adaptive stochastic optimization methods for fitting linear models to the data. This entails modeling a linear relationship where the dependent variable y is expressed as a linear combination of the feature vector x and a parameter vector θ^*. We adopt θ^* as our target parameter vector, specifically defined as $\theta^* = (-d/2, \ldots, d/2)^\top$ (Boyer and Godichon-Baggioni, 2023).

In Figure 1, we present the quadratic mean error of different estimates, considering two types of initializations. Notably, we observe that both the Adagrad and Newton's methods exhibit faster convergence rates when compared to the standard Stochastic Gradient Descent (SGD). This enhanced performance is attributed to their innate capability to manage the diagonal structure of the Hessian matrix, which comprises eigenvalues at different scales. It's important to note that while the Adagrad algorithm adapts its step size, it may be less effective when confronted with highly correlated data. Intriguingly, for scenarios involving less-than-ideal initializations (as depicted on the right side of the figure), both Newton's methods demonstrate outstanding performance.

6.2 Logistic Regression

In logistic regression, our focus shifts to evaluating the performance of our adaptive stochastic optimization methods within the realm of binary classification. Logistic regression models the probability of a data point belonging to one of two classes based on predictor variables. We utilize a sigmoid function to transform a linear combination of the feature vector x and the parameter vector θ^* into class probabilities. Consistent with Boyer and Godichon-Baggioni (2023), we choose $\theta^* \in \mathbb{R}^d$ with all components equal to one. Unlike the linear regression setting, logistic regression exhibits intrinsic non-linearity, which makes the impact of the covariance structures less clear.

[3]The streaming variant of Adagrad, along with its weighted average counterpart, is detailed in Appendix A.4.

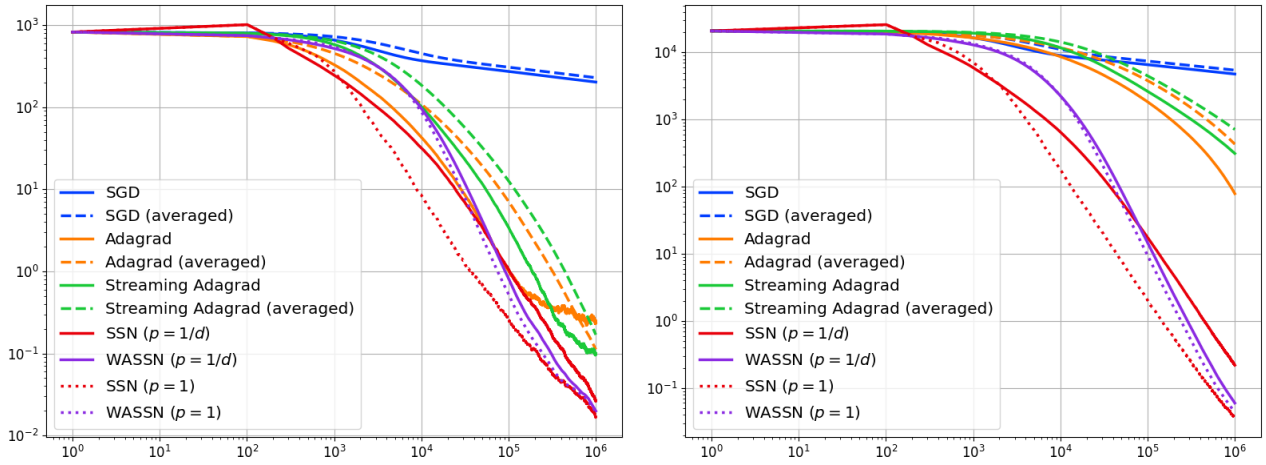

Figure 1: Least-Squares Regression: Mean-squared error of the distance to the optimum θ^*, plotted against the sample size of 1.000.000, for various initializations. The initial points θ_0 are generated as $\theta_0 = \theta^*(1 + rU)$, where U is a uniform random variable on the unit sphere of $\mathbb{R}^d$, and r takes values of 1 (left) or 5 (right). Each curve reports $\|\theta_t - \theta^*\|$ averaged over 50 different epochs, with a different initial point drawn for each sample.

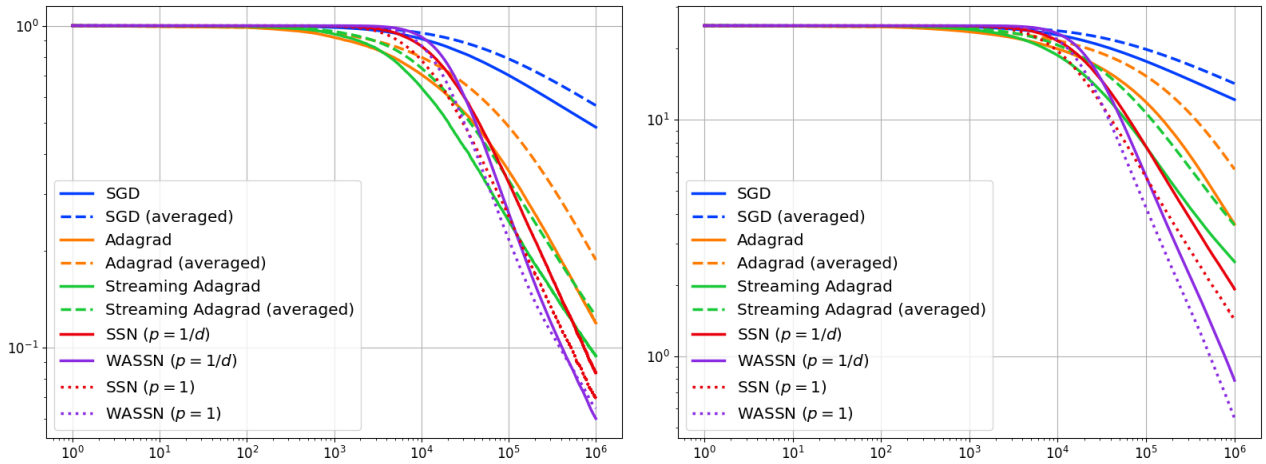

Figure 2: Logistic Regression: Mean-squared error of the distance to the optimum θ^*, plotted against the sample size of 1.000.000, for various initializations. The initial points θ_0 are generated as $\theta_0 = \theta^*(1 + rU)$, where U is a uniform random variable on the unit sphere of $\mathbb{R}^d$, and r takes values of 1 (left) or 5 (right). Each curve reports $\|\theta_t - \theta^*\|$ averaged over 50 different epochs, with a different initial point drawn for each sample.

In Figure 2, we display the evolution of the quadratic mean error of different estimates, for two different initializations. Across all initial configurations, the stochastic Newton's method consistently emerges as a strong competitor. Conversely, the effectiveness of the Adagrad algorithm diminishes as the initial starting point moves further away from the solution. In scenarios involving less-than-ideal initializations, as depicted on the right side of the figure, the best performances are achieved by the averaged Newton's method. This exceptional asymptotic behavior is enabled by the incorporation of weighted estimates, assigning greater significance to the most recent ones, distinguishing it from the "standard" averaging Newton's method, as elaborated in (Bercu et al., 2020).

Conclusions and Future Work

In this work, we addressed the unique challenges posed by streaming data in the context of stochastic optimization. The continuous influx of large, high-dimensional data necessitates adaptive approaches that can effectively handle ill-conditioned problems while maintaining computational efficiency. Our contributions lie in the development of adaptive stochastic optimization methods, particularly an inversion-free adaptive Newton's method with a

computational complexity matching that of first-order methods, $\mathcal{O}(dN_t)$, where d represents the number of dimensions/features, and N_t denotes the quantity of data up to time t.

Theoretical analyses have confirmed the asymptotic efficiency of our proposed methods. By dynamically adjusting learning rates per-dimension and incorporating historical gradient or Hessian information, our methods exhibit adaptability and efficiency in navigating through the complexities of ill-conditioned problems. Notably, the introduction of a weighted averaged version enhances the adaptability and robustness of our methods, particularly in scenarios involving complex covariance structures and challenging initializations.

One significant contribution is the inversion-free adaptive Newton's method in Section 5.3, which strikes a balance between addressing ill-conditioned problems and meeting the computational demands of streaming data. This innovation allows us to harness the advantages of second-order information while aligning with the computational complexity of first-order methods. Empirical evidence demonstrates the effectiveness of our adaptive methods, showcasing superior performance, especially in challenging scenarios.

In conclusion, our adaptive stochastic optimization methods offer a versatile solution for streaming data settings, providing an efficient and adaptive framework for handling ill-conditioned problems. The inversion-free adaptive Newton's method, in particular, stands out as a computationally efficient alternative that bridges the gap between first-order and second-order methods. As we look ahead, further exploration of real-world applications, theoretical advancements, and extensions to non-convex settings will be key directions for future research in this evolving field.

Future work. Looking ahead, there are several promising directions for future research: (a) Non-convex analysis: Extending our methodologies to non-convex optimization problems is a crucial next step. Analyzing the behavior and convergence properties of our adaptive methods in non-convex scenarios will contribute to a more comprehensive understanding of their applicability across diverse optimization landscapes. (b) Time-dependent observations: The streaming context often involves time-dependent observations, and our current work assumes independence among the data points. Investigating extensions of our methods to handle dependent observations will be essential for real-world applications where temporal or spatial dependencies are prevalent. Recently, Godichon-Baggioni et al. (2023a) showed that increasing mini-batches can break both short- and long-term dependence structures. These future research directions aim to refine the versatility and robustness of our adaptive stochastic optimization methods, ensuring their effectiveness across a broader spectrum of optimization challenges.

Aknowledgements

N. Werge acknowledges the support of the Novo Nordisk Foundation (NNF) through grant NNF21OC0070621.

Appendix

Appendix A contains the statements for increasing mini-batches, while Appendix B contains the mathematical proofs of the main results.

A Statements for Increasing Mini-Batches

In this appendix, we investigate our adaptive methods, in the case, where the mini-batches are increasing, and give the translation of the different theorems in this case. Following, Godichon-Baggioni et al. (2023a,b), we consider mini-batch sizes of the form $n_t = \lfloor C_\rho t^\rho \rfloor$ with $C_\rho \in \mathbb{N}$. In this case, we take $\gamma_t = C_\gamma n_t^\beta t^{-\gamma}$, which roughly means that $\gamma_t \sim C_\gamma C_\rho^\beta t^{-\gamma+\beta\rho}$. Adding the term n_t^β to the learning rate enables us to give more weights on (presumably) more precise gradient steps, as they are estimated with larger mini-batches n_t. We suppose that $\gamma - \beta\rho \in (1/2, 1)$ and $\gamma > \frac{\rho(2\beta-1)+1}{2}$.

A.1 Adaptive Stochastic Optimization Methods

When considering increasing mini-batches, our adaptive stochastic optimization methods are defined as in (2):

$$\theta_{t+1} = \theta_t - \gamma_{t+1} A_t \nabla_\theta f(\theta_t; \xi_{t+1}),\ \theta_0 \in \mathbb{R}^d, \tag{A.1}$$

where $\nabla_\theta f(\theta_t;\xi_{t+1}) = n_{t+1}^{-1}\sum_{i=1}^{n_{t+1}} \nabla_\theta f(\theta_t;\xi_{t+1,i})$. In this case, the conditions in (4) should be rewritten to

$$\sum_{t\geq 1}\gamma_t\lambda_{\min}(A_{t-1}) = +\infty \text{ a. s.}, \quad \sum_{t\geq 1}\frac{\gamma_t^2}{n_t}\lambda_{\max}(A_{t-1})^2 < +\infty \text{ a. s.}, \quad \frac{\lambda_{\max}(A_t)^2\gamma_{t+1}}{\lambda_{\min}(A_t)} \xrightarrow[t\to+\infty]{\text{a.s.}} 0. \tag{A.2}$$

Under these conditions in (A.2), we can show the strongly consistency of the estimates derived from (A.1). This, in the increasing mini-batch case, Theorem 1 can be rewritten as follows:

Theorem A.1 *Suppose Assumptions 1 to 3 hold, along with the conditions in* (A.2)*. Then θ_t converges almost surely to θ.*

Similarly, we have the rate of convergence for (A.1) as in Theorem 2:

Theorem A.2 *Suppose Assumptions 1 to 4 hold, along with the conditions in* (A.2)*. In addition, assume there exists positive constants C_η and $\eta > \frac{1}{\gamma-\beta\rho} - 1$ such that for all $\theta \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\|\nabla_\theta f(\theta;\xi)\|^{2+2\eta}\right] \leq C_\eta(1 + F(\theta_t) - F(\theta^*))^{1+\eta}. \tag{A.3}$$

Then,

$$\|\theta_t - \theta^*\|^2 = \mathcal{O}\left(\ln(N_t)N_t^{\frac{-\rho-\gamma+\beta\rho}{1+\rho}}\right) \text{ a. s.}$$

Note that the rate of convergence in Theorem A.2 reproduce the results of the constant mini-batch case in Theorem 2 when $n_t = n = C_\rho$, $\beta = 0$, and $\rho = 0$.

A.2 The Weighted Averaged Version

As in Appendix A.1, we consider the weighted Polyak-Ruppert averaged version of our adaptive stochastic optimization methods for increasing mini-batches; the constant mini-batch case is in Section 4. These weighted estimates are defined for $w \geq 0$ as follows:

$$\theta_{t,w} = \frac{1}{\sum_{i=0}^{t-1} n_{i+1}\ln(i+1)^w}\sum_{i=0}^{t-1} n_{i+1}\ln(i+1)^w\theta_i,$$

which can be written recursively as

$$\theta_{t+1,w} = \theta_{t,w} + \frac{n_{t+1}\ln(t+1)^w}{\sum_{i=0}^{t} n_{i+1}\ln(i+1)^w}(\theta_t - \theta_{t,w}).$$

Likewise to Theorem 3, we have the rate of convergence and the optimal asymptotic normality of these weighted estimates:

Theorem A.3 *Suppose Assumptions 1 to 5 hold, along with inequality* (A.3)*. In addition, assume there exists a positive constant ν such that*

$$\|A_t - A\|_{\text{op}} = \mathcal{O}\left(\frac{1}{t^\nu}\right) \text{ a. s.}$$

Then,

$$\|\theta_{t,w} - \theta^*\|^2 = \begin{cases} \mathcal{O}\left(\frac{\ln(N_t)^{\frac{1}{2}+\mathbb{1}\left\{\nu+\frac{\rho(1-\beta)+\gamma}{2}=1\right\}}}{N_t^{\frac{2\nu+\rho(1-\beta)+\gamma}{1+\rho}}}\right) \text{ a. s.}, & \text{if } 2\nu+\rho(1-\beta)+\gamma \leq 1+\rho, \\ \mathcal{O}\left(\frac{\ln(N_t)}{N_t}\right) \text{ a. s.}, & \text{if } 2\nu+\rho(1-\beta)+\gamma > 1+\rho. \end{cases}$$

Moreover, if $2\nu+\rho(1-\beta)+\gamma > 1+\rho$, then

$$\sqrt{N_t}(\theta_{t,w} - \theta^*) \xrightarrow[t\to+\infty]{\mathcal{L}} \mathcal{N}(0, \nabla_\theta^2 F(\theta^*)^{-1}\Sigma\nabla_\theta^2 F(\theta^*)^{-1}).$$

Similarly to Theorem 4, we can establish the asymptotic efficiency without relying on a (weak) rate of convergence of A_t:

Theorem A.4 *Suppose Assumptions 1 to 5 hold, along with inequality* (A.3)*. In addition, assume there exists a positive constant $v' > 1/2$ such that*

$$\frac{1}{\sum_{i=0}^{t-1} n_{i+1}\ln(i+1)^w}\sum_{i=0}^{t-1} n_{i+1}\ln(i+1)^{w+1/2+\delta}\|A_{i+1}^{-1}-A_i^{-1}\|_{\mathrm{op}}(i+1)^{\frac{\gamma-\rho(\beta+1)}{2}} = \mathcal{O}\left(\frac{1}{t^{(1+\rho)v'}}\right) \text{ a. s.}, \quad \text{(A.4)}$$

for some $\delta > 0$. Then,

$$\|\theta_{t,w}-\theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t}\right) \text{ a. s.}, \quad \textit{and} \quad \sqrt{N_t}(\theta_{t,w}-\theta^*) \xrightarrow[t\to+\infty]{\mathcal{L}} \mathcal{N}(0, \nabla_\theta^2 F(\theta^*)^{-1}\Sigma\nabla_\theta^2 F(\theta^*)^{-1}).$$

A.3 Applications to Newton's Methods

As in Section 5, we apply our adaptive stochastic optimization methodology to (stochastic) Newton's methods (but with increasing mini-batches). In particular, we consider the Newton's methods with the possibly $\mathcal{O}(dN_t)$ operations, analogues to Sections 5.2 and 5.3.

A.3.1 Streaming Stochastic Newton's Methods with possibly $\mathcal{O}(dN_t)$ operations

Expanding the mini-batch scenario from Section 5.2 leads to the formulation of the streaming variant of stochastic Newton's method, as defined by:

$$\theta_{t+1} = \theta_t - \frac{1}{N_{t+1}}\bar{H}_{t,w'}^{-1}\sum_{i=1}^{n_{t+1}}\nabla_\theta f(\theta_t;\xi_{t+1,i}), \quad \text{(A.5)}$$

where $\bar{H}_{t,w'} = N_{t,Z}^{-1}H_{t,w'}$ with

$$H_{t,w'} = H_{0,w'} + \sum_{t'=1}^{t}\ln(t'+1)^{w'}\sum_{i=1}^{n_t'} Z_{t',i}\left(\iota_{t',i}\tilde{e}_{t',i}\tilde{e}_{t',i}^T + \alpha_{t',i}\Phi_{t',i}\Phi_{t',i}^\top\right),$$

with $N_{Z,t} = 1 + \sum_{t'=1}^{t}\ln(t'+1)^{w'}\sum_{i=1}^{n_{t'}} Z_{t',i}$. In addition, let $N_{Z,t,i} = (1 + \sum_{t'=1}^{t-1}\sum_{j=1}^{n_{t'}} Z_{t',j} + \sum_{j=1}^{i} Z_{t,j})$, $\iota_{t',i} = c_\iota N_{Z,t',i}^{-\iota}$, $\iota \in (0\frac{1-\rho}{2(1+\rho)})$, $e_{t',i}$ is the $(N_{t'-1}+i$ modulo $d+1)$-th componant of the canonical basis.

As in Theorem 6, we can establish the rate of convergence and the asymptotic normality of (A.5):

Theorem A.5 *Suppose Assumptions 1 to 3 and 5 hold, along with the conditions in* (12) *and* (A.3)*. In addition, assume there are positive constants $C_{\eta'}$ and $\eta' > 1$ such that for all $\theta \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\|\alpha(\theta;\xi)\Phi(\theta;\xi)\Phi(\theta;\xi)^\top\|^{\eta'}\right] \le C_{\eta'}^{\eta'}.$$

Then,

$$\|\theta_t-\theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t}\right) \text{ a. s.}$$

Moreover, suppose that the Hessian of F is locally Lipschitz on a neighborhood around θ^* and that $\eta' \ge 2$. Then,

$$\sqrt{N_t}\,(\theta_t-\theta^*) \xrightarrow[t\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \nabla_\theta^2 F(\theta^*)^{-1}\Sigma\nabla_\theta^2 F(\theta^*)^{-1}\right).$$

A.3.2 Weighted Averaged Version of Streaming Stochastic Newton's Methods with possibly $\mathcal{O}(dN_t)$ operations

The weighted averaged version outlined in Section 5.3 can similarly be adapted to the increasing mini-batch case. The weighted averaged streaming stochastic Newton's method is defined as

$$\begin{aligned}\theta_{t+1} &= \theta_t - \gamma_{t+1}\bar{S}_{t,w'}^{-1}\frac{1}{n_{t+1}}\sum_{i=1}^{n_{t+1}}\nabla_\theta f(\theta_t;\xi_{t+1,i}),\\ \theta_{t+1,w} &= \theta_{t,w} + \frac{n_{t+1}\ln(t+1)^w}{\sum_{i=0}^{t} n_{i+1}\ln(i+1)^w}(\theta_t-\theta_{t,w}),\end{aligned} \quad \text{(A.6)}$$

where $\gamma_t = C_\gamma n_t^\beta t^{-\gamma}$ and $\bar{S}_{t,w'} = N_{t,Z}^{-1} S_{t,w'}$ with

$$S_{t,w'} = S_{0,w'} + \sum_{t'=1}^{t} \ln(t'+1)^{w'} \sum_{i=1}^{n_{t'}} Z_{t',i} \left(\iota_{t',i} e_{t',i} e_{t',i}^T + \alpha_{t',i} \Phi_{t',i} \Phi_{t',i}^\top \right),$$

with S_0 symmetric and positive, $\iota_{t',i} = C_\iota N_{Z,t',i}^{-\iota}$ with $\iota \in \left(0, \frac{\min\{\gamma-\rho\beta, 2\gamma-2\rho\beta-1+\rho\}}{2(1+\rho)}\right)$, which is possible since $\gamma - \rho\beta \in (1/2, 1)$.

Like in Theorem 7, we have the following asymptotic optimality:

Theorem A.6 *Suppose Assumptions 1 to 3 and 5 hold, along the conditions in* (12) *and* (A.3)*. In addition, assume there exists positive constants $C_{\eta'}$ and $\eta' > 1$ such that for all $\theta \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\|\alpha(\theta;\xi)\Phi(\theta;\xi)\Phi(\theta;\xi)^\top\|^{\eta'} \right] \leq C_{\eta'}^{\eta'}.$$

Then,

$$\|\theta_t - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t^{\frac{\gamma+\rho(1-\beta)}{1+\rho}}} \right) \text{ a. s.} \qquad \textit{and} \qquad \|\theta_{t,w} - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t} \right).$$

In addition,

$$\sqrt{N_t}\left(\theta_{t,w} - \theta^*\right) \xrightarrow[t\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \nabla_\theta^2 F(\theta^*)^{-1} \Sigma \nabla_\theta^2 F(\theta^*)^{-1}\right).$$

A.4 Streaming Adagrad and its Weighted Averaged Version

In this section, we apply our adaptive stochastic optimization methodology to Adagrad (Duchi et al., 2011). Our adaptation results in a streaming version of Adagrad, specifically tailored for efficient handling of evolving data streams. Additionally, we introduce the weighted averaged version of streaming Adagrad, enhancing adaptability and accelerating convergence.

A.4.1 Streaming Adagrad with constant mini-batches

The recursive definitions for streaming Adagrad and its weighted averaged version are as follows:

$$\theta_{t+1} = \theta_t - \gamma_{t+1} G_t \nabla_\theta f(\theta_t; \xi_{t+1}), \quad \theta_0 \in \mathbb{R}^d, \tag{A.7}$$

$$\theta_{t+1,w} = \theta_{t,w} + \frac{\ln(t+1)^w}{\sum_{i=0}^t \ln(i+1)^w}(\theta_t - \theta_{t,w}), \tag{A.8}$$

where $\nabla_\theta f(\theta_t;\xi_{t+1}) = n^{-1}\sum_{i=1}^n \nabla_\theta f(\theta_t;\xi_{t+1,i})$ and G_t is a diagonal matrix with k-th element $G_t^{(k)}$ for $k = 1,\ldots,d$, given as

$$G_t^{(k)} = \left(\frac{1}{N_t} \left(G_0^{(k)} + \sum_{i=1}^t \sum_{j=1}^n \frac{\partial}{\partial k} f(\theta_{t-1};\xi_{i,j})^2 \right) \right)^{-1/2},$$

with $\nabla_{\theta^{(k)}}$ denoting the partial derivative with respect to k-th element of θ, i.e., $\theta^{(k)}$.

To mitigate the potential divergence of the eigenvalues of G_t, we employ a technique introduced by Godichon-Baggioni and Tarrago (2023), resulting in a mild modification of the standard random matrix G_t. The modification is expressed as:

$$G_t^{(k)} = \max\left\{ C_{\beta''} t^{\beta''}, \min\left\{ C_{\beta'} t^{\beta'}, \left(\frac{1}{N_t} \left(G_0^{(k)} + \sum_{i=1}^t \sum_{j=1}^n \frac{\partial}{\partial k} f(\theta_{t-1};\xi_{i,j})^2 \right) \right)^{-1/2} \right\} \right\},$$

with $C_{\beta'}, C_{\beta''} > 0$. In this formulation, the addition of the min-term in G_t aids in controlling the potential divergence of its largest eigenvalue, while the max-term ensures a lower bound for the smallest eigenvalue. Precisely, selecting $\gamma \in (1/2, 1)$, $\beta' \in (0, \gamma - 1/2)$, and $\beta'' \in (\gamma - 1, 0)$ satisfies $2\beta' - \gamma - \beta'' < 0$, which ensures the conditions in (4) are satisfied.

With these modifications in place, we can now establish the rate of convergence and asymptotic normality.

Theorem A.7 *Suppose Assumptions 1 to 3 and 5 hold, along with inequality* (5)*. In addition, assume that the variance* $\mathbb{V}[\frac{\partial}{\partial k} f(\theta^*;\xi)] > 0$ *for* $k = 1, \ldots, d$*. Then,*

$$\|\theta_t - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t^{\gamma}}\right) \text{ a. s.}, \qquad \|\theta_{t,w} - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t}\right) \text{ a. s.},$$

and

$$\sqrt{N_t}(\theta_{t,w} - \theta^*) \xrightarrow[t\to+\infty]{\mathcal{L}} \mathcal{N}(0, \nabla_\theta^2 F(\theta^*)^{-1} \Sigma \nabla_\theta^2 F(\theta^*)^{-1}).$$

A.4.2 Streaming Adagrad with increasing mini-batches

For the increasing mini-batch case, the streaming Adagrad variant and its weighted averaged version is defined recursively by

$$\theta_{t+1} = \theta_t - \gamma_{t+1} G_t \nabla_\theta f(\theta_t; \xi_{t+1}), \theta_0 \in \mathbb{R}^d,$$
$$\theta_{t+1,w} = \theta_{t,w} + \frac{n_{t+1} \ln(t+1)^w}{\sum_{i=0}^t n_{i+1} \ln(i+1)^w}(\theta_t - \theta_{t,w}),$$

where $\nabla_\theta f(\theta_t; \xi_{t+1}) = n_{t+1}^{-1} \sum_{i=1}^{n_{t+1}} \nabla_\theta f(\theta_t; \xi_{t+1,i})$ and G_t is a diagonal matrix with, denoting by $G_t^{(k)}$ the k-th element of the diagonal of G_t,

$$G_t^{(k)} = \max\left\{ C_{\beta''} t^{\beta''}, \min\left\{ C_{\beta'} t^{\beta'}, \left(\frac{1}{N_t} \left(G_0^{(k)} + \sum_{i=1}^{t} \sum_{j=1}^{n_t} \left(\frac{\partial}{\partial k} f(\theta_{t-1}; \xi_{i,j}) \right)^2 \right) \right)^{-1/2} \right\} \right\}.$$

Remark that the add of the minimum in the expression of G_t enables to control the possible divergence of the larges eigenvalue of G_t while the max term enables to lower bound the smallest eigenvalue. More precisely, taking $\gamma - \beta\rho \in (1/2, 1)$, $\beta' \in (0, \gamma + \rho(\frac{1}{2} - \beta) - 1/2)$ and $\beta'' \in (\gamma - \beta\rho - 1, 0)$ satisfying $2\beta' - \gamma + \beta\rho - \beta'' < 0$ enables to verify the conditions in (A.2). To simplify it, one can take $\beta' < \gamma - \beta\rho - 1/2$. Then, Theorem A.7 can be written as follows:

Theorem A.8 *Suppose Assumptions 1 to 3 and 5 hold, along with the conditions in* (A.3)*. In addition, assume that the variance* $\mathbb{V}\left[\frac{\partial}{\partial k} f(\theta^*;\xi)\right] > 0$ *for* $k = 1, \ldots, d$*. Then,*

$$\|\theta_t - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t^{\frac{\gamma+\rho(1-\beta)}{1+\rho}}}\right) \text{ a. s.}, \qquad \|\theta_{t,w} - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t}\right) \text{ a. s.},$$

and

$$\sqrt{N_t}(\theta_{t,w} - \theta^*) \xrightarrow[t\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \nabla_\theta^2 F(\theta^*)^{-1} \Sigma \nabla_\theta^2 F(\theta^*)^{-1}\right).$$

B Proofs

The proof are solely presented for the increasing mini-batch case outlined in Appendix A, as the constant mini-batch case corresponds to $n_t = n = C_\rho$, $\beta = 0$, and $\rho = 0$.

For the sake of simplicity, in all the sequel, since $n_t \sim C_\rho t^\rho$, we will make the abuse that $n_t = C_\rho t^\rho$. To lighten the notation, we let H denote $\nabla_\theta^2 F(\theta^*)$. In addition, let f'_{t+1} and $f'_{t+1,i}$ denote $\nabla_\theta f(\theta_t; \xi_{t+1})$ and $\nabla_\theta f(\theta_t; \xi_{t+1,i})$, respectively.

B.1 Proof of Theorems 1 and A.1

Let V_t denote $F(\theta_t) - F(\theta^*)$. Observe that with the help of a Taylor's expansion of the objective function F and since the Hessian is uniformly bounded (Assumption 2), then one has

$$\begin{aligned} V_{t+1} &\leq V_t + \nabla_\theta F(\theta_t)^\top (\theta_{t+1} - \theta_t) + \frac{L_{\nabla F}}{2} \|\theta_{t+1} - \theta_t\|^2 \\ &\leq V_t - \gamma_{t+1} \nabla_\theta F(\theta_t)^\top A_t f'_{t+1} + \frac{L_{\nabla F}}{2} \gamma_{t+1}^2 \lambda_{\max}(A_t)^2 \|f'_{t+1}\|^2. \end{aligned}$$

Before taking the conditional expectation, recall from Godichon-Baggioni et al. (2023b) that

$$\mathbb{E}[\|f'_{t+1}\|^2|\mathcal{F}_t] = \frac{1}{n_{t+1}}\mathbb{E}[\|f'_{t+1,i}\|^2|\mathcal{F}_t] + \|\nabla_\nabla F(\theta_t)\|^2 \leq \frac{1}{n_{t+1}}C(1+F(\theta_t)-F(\theta^*)) + \|\nabla_\theta F(\theta_t)\|^2.$$

Thus, we obtain that

$$\begin{aligned}
\mathbb{E}[V_{t+1}|\mathcal{F}_t] \leq & V_t - \gamma_{t+1}\nabla_\theta F(\theta_t)^\top A_t \nabla_\theta F(\theta_t) \\
& + \frac{L_{\nabla F}}{2}\gamma_{t+1}^2\lambda_{\max}(A_t)^2\left(\frac{C}{n_{t+1}}\left(1+F(\theta_t)-F(\theta^*)\right) + \|\nabla_\theta F(\theta_t)\|^2\right) \\
\leq & \left(1+\frac{L_{\nabla F}C}{2}\frac{\gamma_{t+1}^2\lambda_{\max}(A_t)^2}{n_{t+1}}\right)V_t - \gamma_{t+1}\|\nabla_\nabla F(\theta_t)\|^2\left(\lambda_{\min}(A_t) - \frac{L_{\nabla F}}{2}\gamma_{t+1}\lambda_{\max}(A_t)^2\right) \\
& + \frac{L_{\nabla F}C}{2}\frac{\gamma_{t+1}^2\lambda_{\max}(A_t)^2}{n_{t+1}}.
\end{aligned}$$

Observe that as $\frac{\gamma_{t+1}\lambda_{\max}(A_t)^2}{\lambda_{\min}(A_t)}$ converges almost surely to zero for any constant $C \in (0,1)$ due to the conditions in (A.2). Then, $\mathbb{1}_{\left\{\frac{L_{\nabla F}}{2}\gamma_{t+1}\lambda_{\max}(A_t)^2 \geq C\lambda_{\min}(A_t)\right\}}$ converges almost surely to zero as well. Thus, we have that

$$\begin{aligned}
\mathbb{E}[V_{t+1}|\mathcal{F}_t] \leq & \left(1+\frac{L_{\nabla F}C}{2}\frac{\gamma_{t+1}^2\lambda_{\max}(A_t)^2}{n_{t+1}}\right)V_t - (1-C)\gamma_{t+1}\lambda_{\min}(A_t)\|\nabla_\theta F(\theta_t)\|^2 \\
& + \frac{L_{\nabla F}C}{2}\frac{\gamma_{t+1}^2\lambda_{\max}(A_t)^2}{n_{t+1}} \\
& + \frac{L_{\nabla F}C}{2}\gamma_{t+1}^2\lambda_{\max}(A_t)^2\|\nabla_\theta F(\theta_t)\|^2\mathbb{1}_{\left\{\frac{L_{\nabla F}}{2}\gamma_{t+1}\lambda_{\max}(A_t)^2 \geq C\lambda_{\min}(A_t)\right\}}.
\end{aligned}$$

Next, since $\mathbb{1}_{\left\{\frac{L_{\nabla F}}{2}\gamma_{t+1}\lambda_{\max}(A_t)^2 \geq C\lambda_{\min}(A_t)\right\}}$ converges almost surely to zero and by the conditions in (A.2);

$$\sum_{t\geq 0}\frac{\gamma_{t+1}^2\lambda_{\max}\left(A_t\right)^2}{n_{t+1}} < +\infty \text{ a. s.},$$

and

$$\sum_{t\geq 0}\gamma_{t+1}^2\lambda_{\max}\left(A_t\right)^2\mathbb{1}_{\left\{\frac{L_{\nabla F}C}{2}\gamma_{t+1}\|\nabla F(\theta_t)\|^2\lambda_{\max}(A_t)^2 \geq c\lambda_{\min}(A_t)\right\}} < +\infty \text{ a. s.},$$

then, applying Robbins-Siegmund's theorem gives that V_t converges almost surely to a finite random variable and

$$\sum_{t\geq 0}\gamma_{t+1}\lambda_{\min}(A_t)\|\nabla_\theta F(\theta_t)\|^2 < +\infty \text{ a. s.},$$

meaning, that $\liminf_t \|\nabla_\theta F(\theta_t)\|^2 = 0$ a.s., such that $\liminf_t V_t = 0$ a.s., i.e., V_t converges almost surely to zero, which concludes the proof.

B.2 Proof of Theorems 2 and A.2

Following the reasoning of Antonakopoulos et al. (2022, page 11), AH and $A^{1/2}HA^{1/2}$ have the same eigenvalues. Indeed, for any $\lambda \in \mathbb{R}$,

$$\begin{aligned}
\det\left(A^{1/2}HA^{1/2} - \lambda I_d\right) &= \det\left(A^{-1/2}\left(AHA^{1/2} - \lambda A^{1/2}\right)\right) \\
&= \det\left(A^{-1/2}\left(AH - \lambda I_d\right)A^{1/2}\right) \\
&= \det\left(AH - \lambda I_d\right).
\end{aligned}$$

Then, there exists matrix Q and a positive diagonal matrix D, such that $AH = Q^{-1}DQ$. Thus,

$$Q(\theta_{t+1}-\theta^*) = Q(\theta_t-\theta^*) - \gamma_{t+1}QA_tf'_{t+1} = Q(\theta_t-\theta^*) - \gamma_{t+1}QA_t\nabla_\theta F(\theta_t) + \gamma_{t+1}QA_t\xi_{t+1},$$

where $\xi_{t+1} = \nabla_\theta F(\theta_t) - f'_{t+1}$. By linearizing the gradient one has

$$Q(\theta_{t+1} - \theta^*) = Q(\theta_t - \theta^*) - \gamma_{t+1} Q A_t H(\theta_t - \theta^*) + \gamma_{t+1} Q A_t \xi_{t+1} - \gamma_{t+1} Q A_t \delta_t,$$

where $\delta_t = \nabla_\theta F(\theta_t) - H(\theta_t - \theta^*)$ is the remainder term of the Taylor's expansion of the gradient. Next, we have

$$\begin{aligned} Q\left(\theta_{t+1} - \theta^*\right) =& Q\left(\theta_t - \theta^*\right) - \gamma_{t+1} Q A H\left(\theta_t - \theta^*\right) - \gamma_{t+1} Q\left(A_t - A\right) H\left(\theta_t - \theta^*\right) \\ &+ \gamma_{t+1} Q A_t \xi_{t+1} - \gamma_{t+1} Q A_t \delta_t \\ =& \left(I_d - \gamma_{t+1} D\right) Q\left(\theta_t - \theta^*\right) - \gamma_{t+1} Q\left(A_t - A\right) H\left(\theta_t - \theta^*\right) \\ &+ \gamma_{t+1} Q A_t \xi_{t+1} - \gamma_{t+1} Q A_t \delta_t. \end{aligned} \tag{B.9}$$

Observe that in the case where $A = H^{-1}$, i.e., in the stochastic Newton's method, one has $D = Q = I_d$. With the help of induction, one has by (B.9) that

$$\begin{aligned} Q\left(\theta_T - \theta^*\right) =& \overbrace{\beta_{T,0} Q\left(\theta_0 - \theta^*\right)}^{R_{1,T}:=} \overbrace{- \sum_{t=0}^{T-1} \beta_{T,t+1}\gamma_{t+1} Q\left(A_t - A\right) H\left(\theta_t - \theta^*\right) - \sum_{t=0}^{T-1} \beta_{T,t+1}\gamma_{t+1} Q A_t \delta_t}^{R_{2,T}:=} \\ &+ \underbrace{\sum_{t=0}^{T-1} \beta_{T,t+1}\gamma_{t+1} Q A_t \xi_{t+1}}_{M_T:=}, \end{aligned} \tag{B.10}$$

where $\beta_{T,t} = \prod_{j=t+1}^{T} \left(I_d - \gamma_j D\right)$ and $\beta_{T,T} = I_d$. The rest of the proof consists in giving the rate of convergence of each term on the right-hand side of decomposition (B.10) for both cases, i.e., for the constant mini-batches and increasing mini-batches.

Rate of convergence for $R_{1,T}$. Since D is a positive diagonal matrix, and since γ_t is decreasing, there is a rank t_0 such that for all $t \geq t_0$, $\|I_d - \gamma_t D\|_{\text{op}} \leq 1 - \lambda_{\min}(D)\gamma_t$. Then, for all $T \geq t_0$,

$$\begin{aligned} \left\|\beta_{T,0}\right\|_{\text{op}} &\leq \prod_{t=1}^{t_0-1} \left(1 + \gamma_t \lambda_{\max}(D)\right) \prod_{t=t_0}^{T} \left(1 - \gamma_t \lambda_{\min}(D)\right) \\ &\leq \exp\left(2\lambda_{\max}(D) \frac{c_\gamma C_\rho^\beta}{1 + \beta\rho - \gamma} t_0^{1+\beta\rho-\gamma}\right) \exp\left(-\lambda_{\min}(D) \frac{c_\gamma C_\rho^\beta}{1 + \beta\rho - \gamma} T^{1+\beta\rho-\gamma}\right). \end{aligned}$$

With N_T denoting $\sum_{t=1}^{T} n_t$, one has $T = \frac{N_T}{n}$ in the case of the constant mini-batch size, and $T \sim \left(\frac{1+\rho}{C_\rho} N_T\right)^{\frac{1}{1+\rho}}$ for the increasing mini-batch size. Then, one has

$$\left\|\beta_{T,0}\right\|_{\text{op}} = \begin{cases} \mathcal{O}\left(\exp\left(-\lambda_{\min}(D) \frac{c_\gamma n^{\gamma-1}}{1-\gamma} N_T^{1-\gamma}\right)\right) & \text{if } n_t = n, \\ \mathcal{O}\left(\exp\left(-\lambda_{\min}(D) \frac{c_\gamma C_\rho^{\frac{\beta-1+\gamma}{1+\rho}}}{1+\beta\rho-\gamma} N_T^{\frac{1+\beta\rho-\gamma}{1+\rho}}\right)\right) & \text{if } n_t = \lfloor C_\rho t^\rho \rfloor. \end{cases} \tag{B.11}$$

Then, in both cases, this term converges exponentially fast to zero.

A first rate of convergence of M_T. First, remark that

$$\mathbb{E}\left[\left\|\xi_{t+1}\right\|^2 |\mathcal{F}_t\right] \leq \mathbb{E}\left[\left\|f'_{t+1}\right\|^2 |\mathcal{F}_t\right] \leq \frac{1}{n_{t+1}} C\left(1 + F\left(\theta_t\right) - F(\theta^*)\right) + \left\|\nabla F\left(\theta_t\right)\right\|^2. \tag{B.12}$$

Then, applying Cénac et al. (2020, Theorem 6.1), one has, since A_t converges almost surely to A, that

$$\left\|M_T\right\|^2 = \mathcal{O}\left(\ln(T) T^{\beta\rho-\gamma}\right) \text{ a. s.} \tag{B.13}$$

Observe that for the constant mini-batch size, we already have the good rate of convergence for this term, but not for the increasing case. We will come back later to this term below when we find the first rate of convergence of θ_T.

A first rate of convergence of $M_{2,T}$. As $\|\delta_t\| = o(\|\theta_t - \theta^*\|)$ a.s and $\|A_t - A\|_{\text{op}}$ converge almost surely to 0, there exists a sequence of random positive variables r_t which converges to 0 almost surely, such that for all $t \geq t_0$,

$$\begin{aligned}\|R_{2,t+1}\| &\leq (1-\gamma_{t+1})\|R_{2,t}\| + \gamma_{t+1} r_{t+1}\|\theta_t - \theta^*\| \\ &\leq (1-\gamma_{t+1})\|R_{2,t}\| + 2\gamma_{t+1} r_{t+1}\left(\|R_{2,t}\|^2 + \|M_t + R_{1,t}\|\right).\end{aligned}$$

Then, with the help of (B.11) and (B.13), there exists a positive random variable C_1, such that

$$\|R_{2,t+1}\|^2 \leq (1-\gamma_{t+1})\|R_{2,t}\| + 2\gamma_{t+1} r_{t+1}\left(\|R_{2,t}\| + C_1 \ln(t+1)(t+1)^{\frac{\beta\rho-\gamma}{2}}\right), \tag{B.14}$$

so that

$$\|R_{2,T}\|^2 = \mathcal{O}\left(\ln(T) T^{\beta\rho-\gamma}\right) \text{ a. s.}$$

This concludes the proof for the constant mini-batch size case. For the non constant case, we need to return to the martingale term.

A good rate of convergence for M_T and $R_{2,T}$. Let $k_0 = \inf\{k, k(\gamma - \beta\rho) > \rho\}$. Then, let us prove by induction that for any non negative integer $k \leq k_0$,

$$\|\theta_T - \theta^*\|^2 = O\left(\ln(T)^k T^{-k(\gamma-\beta\rho)}\right) \text{ a. s.}$$

If $k_0 = 0$, this is satisfied. Let us suppose from now on that $k_0 \geq 1$ and prove this result by induction: Suppose it is true for $k-1$. Then, thanks to inequality (B.12), one has

$$\mathbb{E}\left[\|\xi_{t+1}\|^2 |\mathcal{F}_t\right] = \mathcal{O}\left(\ln(T)^{k-1} T^{-(k-1)(\gamma-\beta\rho)}\right) \text{ a. s.},$$

and with the help of Cénac et al. (2020, Theorem 6.1), we have

$$\|M_T\|^2 = \mathcal{O}\left(\ln(T)^k T^{-k(\gamma-\beta\rho)}\right) \text{ a. s.},$$

and $\|R_{2,T}\|^2 = \mathcal{O}\left(\ln(T)^k T^{-k(\gamma-\beta\rho)}\right)$ a.s., which concludes the induction proof.

As a particular case, one has

$$\|\theta_T - \theta^*\|^2 = O\left(\ln(T)^{k_0} T^{-k_0(\gamma-\beta\rho)}\right) \text{ a. s.},$$

so that by definition of k_0, $\mathbb{E}[\|\xi_{t+1}\|^2|\mathcal{F}_t] = \mathcal{O}(t^{-\rho})$ a.s., and we obtain with the help of Cénac et al. (2020, Theorem 6.1), that

$$\|M_T\|^2 = \mathcal{O}\left(\ln(T) T^{-\rho-\gamma+\beta\rho}\right) \text{ a. s.} \qquad \text{and} \qquad \|R_{2,T}\|^2 = \mathcal{O}\left(\ln(T) T^{-\rho-\gamma+\beta\rho}\right) \text{ a. s.}$$

Then, since $T \sim \left(\frac{1+\rho}{C_\rho} N_T\right)^{\frac{1}{1+\rho}}$, one has

$$\|\theta_T - \theta^*\|^2 = O\left(\ln(N_T) N_T^{\frac{-\rho-\gamma+\beta\rho}{1+\rho}}\right) \text{ a. s.}$$

B.3 Proof of Theorems 3 and A.3

Observe that one has for all $t \geq 0$,

$$\theta_{t+1} - \theta^* = \theta_t - \theta^* - \gamma_{t+1} AH(\theta_t - \theta^*) - \gamma_{t+1}(A_t - A) H(\theta_t - \theta^*) + \gamma_{t+1} A_t \xi_{t+1} - \gamma_{t+1} A_t \delta_t,$$

which can be written as

$$\theta_t - \theta^* = H^{-1}A^{-1}\frac{u_t - u_{t+1}}{\gamma_{t+1}} + H^{-1}A^{-1}A_t\xi_{t+1} - H^{-1}A^{-1}A_t\delta_t - H^{-1}A^{-1}(A_t - A)H(\theta_t - \theta^*), \tag{B.15}$$

where $u_t = \theta_t - \theta^*$. Summing these equalities and dividing by $s_T = \sum_{t=0}^{T-1} n_{t+1} \ln(t+1)^w$, we have

$$\theta_{T,w} - \theta^* = H^{-1}A^{-1}\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w \frac{u_t - u_{t+1}}{\gamma_{t+1}} + H^{-1}A^{-1}\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w A_t\xi_{t+1}$$
$$- H^{-1}A^{-1}\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w A_t\delta_t - \frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w H^{-1}A^{-1}\left(A_t - A\right)H\left(\theta_t - \theta^*\right). \tag{B.16}$$

The rest of this proof consists in giving the rate of convergence of each term on the right-hand side of previous decomposition.

Rate of convergence of $H^{-1}A^{-1}\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w A_t\xi_{t+1}$. Remark that $M'_T = \sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w A_t\xi_{t+1}$ is a martingale term and that

$$\begin{aligned}\langle M'_T\rangle &= \sum_{t=0}^{T-1} n_{t+1}^2\ln(t+1)^{2w}A_t\mathbb{E}\left[\xi_{t+1}\xi_{t+1}^T|\mathcal{F}_t\right]A_t\\ &= \sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^{2w}A_t\mathbb{E}\left[f'_{t+1,i}f'^{\top}_{t+1,i}|\mathcal{F}_t\right]A_t\\ &\quad - \sum_{t=0}^{T-1} n_{t+1}^2\ln(t+1)^{2w}A_t\nabla_\theta F\left(\theta_t\right)\nabla_\theta F\left(\theta_t\right)^T A_t.\end{aligned}$$

Since

$$n_{t+1}\left\|\nabla F\left(\theta_t\right)\right\|^2 = \mathcal{O}\left(\frac{\ln t}{t^{\gamma-\beta\rho}}\right)\ \text{a. s.},$$

then this converges to 0. Next, since θ_t and A_t converge to θ^* and A, we have

$$\frac{1}{\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^{2\omega}}\langle M'_T\rangle \xrightarrow[T\to+\infty]{\text{a.s.}} A\Sigma A.$$

Then, with the help of a law of large numbers for martingales, we obtain that

$$\frac{1}{s_T^2}\left\|M'_T\right\|^2 = \mathcal{O}\left(\frac{\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^{2\omega}\ln\left(\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^{2\omega}\right)}{s_T^2}\right)\ \text{a. s.},$$

which can be written as

$$\frac{1}{s_T^2}\left\|M'_T\right\|^2 = O\left(\frac{\ln(T+1)}{T^{\rho+1}}\right)\ \text{a. s.}.$$

This, can also be written as

$$\frac{1}{s_T^2}\left\|M'_T\right\|^2 = \mathcal{O}\left(\frac{\ln(N_T)}{N_T}\right)\ \text{a. s.}$$

In addition, Central Limit Theorem for martingales yields,

$$\frac{1}{\sqrt{\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^{2\omega}}}M'_T \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, A\Sigma A\right).$$

Thus, as

$$\frac{\sqrt{N_T}\sqrt{\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^{2\omega}}}{s_T} \xrightarrow[T\to+\infty]{\text{a.s.}} 1,$$

we have

$$\sqrt{N_T}\frac{1}{s_T}H^{-1}A^{-1}M'_T \xrightarrow[T\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right).$$

Rate of convergence of $H^{-1}A^{-1}\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w \frac{u_t-u_{t+1}}{\gamma_{t+1}}$. With the help of Abel's transformation, one have

$$\begin{aligned}&\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w \frac{u_t-u_{t+1}}{\gamma_{t+1}}\\&= -\frac{u_T n_T \ln(T)^w}{\gamma_T s_T} + \frac{u_0 n_1 \mathbb{1}_{\{w=0\}}}{\gamma_1 s_T} + \frac{1}{s_T}\sum_{t=1}^{T-1} u_t\left(\frac{n_{t+1}\ln(t+1)^w}{\gamma_{t+1}} - \frac{n_t \ln(t)^w}{\gamma_t}\right).\end{aligned}$$

One has thanks to Theorems 2 and A.2, we have

$$\left\|\frac{u_T n_T \ln(T)^w}{\gamma_T s_T}\right\| = \mathcal{O}\left(\frac{\sqrt{\ln T}}{T^{\frac{2+\rho-\gamma+\beta\rho}{2}}}\right) \text{ a. s.},$$

which can be written as

$$\left\|\frac{u_T n_T \ln(T)^w}{\gamma_T s_T}\right\| = \mathcal{O}\left(\frac{\sqrt{\ln N_T}}{N_T^{\frac{2+\rho-\gamma+\beta\rho}{2(1+\rho)}}}\right) \text{ a. s.},$$

which is negligible as soon as $\gamma-\beta\rho<1$. In addition, it is obvious that $\frac{u_0 n_1 \mathbb{1}_{\{w=0\}}}{\gamma_1 s_T}$ is negligible too. Furthermore, observe that

$$\left|\frac{n_{t+1}\ln(t+1)^w}{\gamma_{t+1}} - \frac{n_t\ln(t)^w}{\gamma_t}\right| \leq C_\rho \max\{\rho(1-\beta)+\gamma, w\}\max\left\{t^{\rho(1-\beta)+\gamma-1}, (t+1)^{\rho(1-\beta)+\gamma-1}\right\}\ln(t+1)^w,$$

which with the help of Theorems 2 and A.2 yields,

$$\left\|\sum_{t=0}^{T-1}\left(\frac{n_{t+1}\ln(t+1)^w}{\gamma_{t+1}} - \frac{n_t\ln(t)^w}{\gamma_t}\right)(\theta_t-\theta^*)\right\| = \mathcal{O}\left(\ln(T)^{w+1/2}T^{\frac{\rho(1-\beta)+\gamma}{2}}\right) \text{ a. s.}$$

From this, we have

$$\frac{1}{s_T}\left\|\sum_{t=0}^{T-1}\left(\frac{n_{t+1}\ln(t+1)^w}{\gamma_{t+1}} - \frac{n_t\ln(t)^w}{\gamma_t}\right)(\theta_t-\theta^*)\right\| = \mathcal{O}\left(\sqrt{\ln(T)}T^{\frac{-2+\gamma-\rho(1+\beta)}{2}}\right) \text{ a. s.},$$

which can be written as

$$\frac{1}{s_T}\left\|\sum_{t=0}^{T-1}\left(\frac{n_{t+1}\ln(t+1)^w}{\gamma_{t+1}} - \frac{n_t\ln(t)^w}{\gamma_t}\right)(\theta_t-\theta^*)\right\| = \mathcal{O}\left(\sqrt{\ln(N_T)}N_T^{\frac{-2+\gamma-\rho(1+\beta)}{2(1+\rho)}}\right) \text{ a. s.}, \tag{B.17}$$

which is negligible as soon as $\gamma-\beta\rho<1$.

Rate of convergence of $H^{-1}A^{-1}\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w (A_t-A)H(\theta_t-\theta^*)$. Since $\|A_t-A\|_{\text{op}} = \mathcal{O}(t^{-\nu})$ a.s and with the help of Theorem 2, we have

$$\left\|\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w (A_t-A)H(\theta_t-\theta^*)\right\| = \begin{cases}\mathcal{O}\left(\frac{(\ln T)^{\frac{1}{2}+\mathbb{1}_{\nu+\frac{\rho(1-\beta)+\gamma}{2}=1}}}{T^{\nu+\frac{\rho(1-\beta)+\gamma}{2}}}\right) \text{ a. s.} & \text{if } \nu+\frac{\rho(1-\beta)+\gamma}{2}\leq 1\\ \mathcal{O}\left(\frac{1}{T^{1+\rho}}\right) \text{ a. s.} & \text{else}\end{cases}$$

which can be written as

$$\left\|\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w (A_t-A)H(\theta_t-\theta^*)\right\| = \begin{cases}\mathcal{O}\left(\frac{\ln(N_T)^{\left(\frac{1}{2}+\mathbb{1}_{\nu+\frac{\rho(1-\beta)+\gamma}{2}=1}\right)}}{N_T^{\frac{2\nu+\rho(1-\beta)+\gamma}{2(1+\rho)}}}\right) \text{ a. s.} & \text{if } \nu+\frac{\rho(1-\beta)+\gamma}{2}\leq 1\\ \mathcal{O}\left(\frac{1}{N_T}\right) \text{ a. s.} & \text{else}\end{cases}$$

Rate of convergence of $H^{-1}A^{-1}\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w A_t\delta_t$. As $\|\delta_t\| \leq L_\delta \|\theta_t - \theta^*\|^2$ and with the help of Theorem 2, we have

$$\left\| H^{-1}A^{-1}\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w A_t\delta_t \right\| = O\left(\frac{\ln T}{T^{\rho(1-\beta)+\gamma}}\right) \text{ a. s.},$$

which can be written as

$$\left\| H^{-1}A^{-1}\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w A_t\delta_t \right\| = O\left(\frac{\ln N_T}{N_T^{\frac{\rho(1-\beta)+\gamma}{1+\rho}}}\right) \text{ a. s.},$$

which is negligible as soon as $\gamma > \frac{\rho(2\beta-1)+1}{2}$.

B.4 Proof of Theorems 4 and A.4

Fisrt, remark that one can rewrite decomposition (B.15) to

$$\theta_t - \theta^* = H^{-1}A_t^{-1}\frac{u_t - u_{t+1}}{\gamma_{t+1}} + H^{-1}\xi_{t+1} - H^{-1}\delta_t,$$

meaning that

$$\begin{aligned}\theta_{T,w} - \theta^* = &\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w H^{-1}A_t^{-1}\frac{u_t - u_{t+1}}{\gamma_{t+1}} + \frac{1}{S_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w H^{-1}\xi_{t+1}\\ &- \frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w H^{-1}\delta_t.\end{aligned}$$

Analogously to the proof of Theorem 3, one can easily check that

$$\left\| \frac{1}{S_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w H^{-1}\xi_{t+1} \right\|^2 = \mathcal{O}\left(\frac{\ln N_T}{N_T}\right) \text{ a. s.},$$

and

$$\sqrt{N_T}\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w H^{-1}\xi_{t+1} \xrightarrow[T\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right).$$

In the same way, we have

$$\left\| \frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w H^{-1}\delta_t \right\| = O\left(\frac{\ln N_T}{N_T^{\frac{\rho(1-\beta)+\gamma}{1+\rho}}}\right) \text{ a. s.}$$

In addition, note that

$$\begin{aligned}\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w H^{-1}A_t^{-1}\frac{u_t - u_{t+1}}{\gamma_{t+1}} = &\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w H^{-1}\frac{A_t^{-1}u_t - A_{t+1}^{-1}u_{t+1}}{\gamma_{t+1}}\\ &+ \frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w H^{-1}\left(A_{t+1}^{-1} - A_t^{-1}\right)\frac{u_{t+1}}{\gamma_{t+1}}.\end{aligned}$$

With the help of Abel's transformation and since A_t converges almost surely to the positive matrix A (Assumption 4), following the lines of the proof for Theorem 3 (e.g., see (B.17)), one can show that

$$\left\| \frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w H^{-1}\frac{A_t^{-1}u_t - A_{t+1}^{-1}u_{t+1}}{\gamma_{t+1}} \right\| = O\left(\frac{\sqrt{\ln(N_T)}}{N_T^{\frac{2+\gamma-\rho(1+\beta)}{2(1+\rho)}}}\right) \text{ a. s.}$$

In addition, since $\|\theta_t - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(t)}{t^{\gamma-\beta\rho+\rho}}\right)$ a.s., with E_t denoting the event $\{\|\theta_t-\theta\|^2 \leq \frac{(\ln(t))^{1+\delta}}{t^{\gamma+\rho(1-\beta)}}, \|\theta_{t,w}-\theta\|^2 \leq \frac{(\ln(t))^{1+\delta}}{t^{\gamma+\rho(1-\beta)}}\}$, $\mathbb{1}_{\{E_t^C\}}$ converges almost surely to 0, then, we have

$$\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w H^{-1}\left\|A_{t+1}^{-1}-A_t^{-1}\right\|_{\text{op}}\frac{\|u_{t+1}\|}{\gamma_{t+1}}\mathbb{1}_{\{E_{t+1}^C\}} = \mathcal{O}\left(\frac{1}{N_T}\right) \text{ a. s.},$$

and

$$\begin{aligned}\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w H^{-1}\left\|A_{t+1}^{-1}-A_t^{-1}\right\|_{\text{op}}&\frac{\|u_{t+1}\|}{\gamma_{t+1}}\mathbb{1}_{\{E_{t+1}\}}\\ &\leq \frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^{w+1/2+\delta}H^{-1}\left\|A_{t+1}^{-1}-A_t^{-1}\right\|_{\text{op}} c_\gamma^{-1}(t+1)^{\frac{\gamma-\rho(\beta+1)}{2}}.\end{aligned}$$

At last, one can conclude the proof with the help of equality (A.4).

B.5 Proof of Theorem 5

Observe that the convergence of θ_T is obtained with the same calculus as in the proof of Theorem 1. Remark that decomposition (B.9) can now be written as

$$\theta_{t+1}-\theta^* = \left(1-\frac{n_{t+1}}{N_{t+1}}\right)(\theta_t-\theta^*) - \frac{n_{t+1}}{N_{t+1}}\left(\overline{H}_t^{-1}-H^{-1}\right)H(\theta_t-\theta^*) + \frac{n_{t+1}}{N_{t+1}}\overline{H}_t^{-1}\xi_{t+1} - \frac{n_{t+1}}{N_{t+1}}\overline{H}_t^{-1}\delta_t.$$

Then, with the help of induction, one has

$$\begin{aligned}\theta_T-\theta^* =& \frac{1}{N_T}(\theta_0-\theta^*) \underbrace{-\frac{1}{N_T}\sum_{t=0}^{T-1} n_{t+1}\left(\overline{H}_t^{-1}-H^{-1}\right)H(\theta_t-\theta^*) - \frac{1}{N_T}\sum_{t=0}^{T-1} n_{t+1}\overline{H}_t^{-1}\delta_t}_{=:\Delta_T}\\ &+ \underbrace{\frac{1}{N_T}\sum_{t=0}^{T-1} n_{t+1}\overline{H}_t^{-1}\xi_{t+1}}_{=:M_T}.\end{aligned}$$

Convergence of the martingale term M_T. Observe that M_T is a martingale term and that

$$\begin{aligned}\langle M\rangle_T = \sum_{t=0}^{T-1} n_{t+1}^2\overline{H}_t^{-1}\mathbb{E}\left[\xi_{t+1}\xi_{t+1}^T\right]A_t &= \sum_{t=0}^{T-1} n_{t+1}\overline{H}_t^{-1}\mathbb{E}\left[\nabla_\theta f(X_{t+1},\theta_t)\nabla_\theta f(X_{t+1},\theta_t)^T|\mathcal{F}_t\right]\overline{H}_t^{-1}\\ &\quad - \sum_{t=0}^{T-1} n_{t+1}\overline{H}_t^{-1}\nabla F(\theta_t)\nabla F(\theta_t)^T\overline{H}_t^{-1}.\end{aligned}$$

Then, since θ_t and $\overline{H}_t^{-1}$ converge almost surely to θ^* and H^{-1} and by continuity (Assumption 1), one obtain that

$$\frac{1}{N_T}\langle M\rangle_T \xrightarrow[T\to+\infty]{\text{a.s.}} H^{-1}\Sigma H^{-1}.$$

Thus, with the help of a law of large numbers for martingales, we have

$$\left\|\frac{1}{N_T}M_T\right\|^2 = \mathcal{O}\left(\frac{\ln N_T}{N_T}\right) \text{ a. s.},$$

and with the help of Central Limit Theorem for martingales,

$$\frac{1}{\sqrt{N_T}}M_T \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right).$$

Convergence of the rest terms. Since $\overline{H}_t^{-1}$ converges to H^{-1} and $\|\delta_t\| = o(\|\theta_t - \theta^*\|)$ a.s., there is a sequence of positive random variables (r'_t) converging to 0, such that

$$\begin{aligned}\|\Delta_{T+1}\| &\leq \left(1 - \frac{n_{T+1}}{N_{T+1}}\right)\|\Delta_T\| + \frac{n_{T+1}}{N_{T+1}} r'_T \|\theta_T - \theta^*\| \\ &\leq \left(1 - \frac{n_{T+1}}{N_{T+1}}\right)\|\Delta_T\| + \frac{n_{T+1}}{N_{T+1}} r'_T \left\| \frac{1}{N_T}(\theta_0 - \theta^*) + \frac{1}{N_T} M_T + \Delta_T \right\|.\end{aligned}$$

Then, there is a positive random variable C_M, such that

$$\|\Delta_{T+1}\| \leq \left(1 - \frac{n_{T+1}}{N_{T+1}}\right)\|\Delta_T\| + \frac{n_{T+1}}{n_{T+1}} r'_T \left(\|\Delta_T\| + C_M \frac{\sqrt{\ln T}}{T^{\frac{1+\rho}{2}}}\right),$$

which can also be written for any $c \in (0,1)$ as

$$\|\Delta_{T+1}\| \leq \left(1 - c\frac{n_{T+1}}{N_{T+1}}\right)\|\Delta_T\| + \frac{n_{T+1}}{n_{T+1}} C_M \frac{\sqrt{\ln(T+1)}}{(T+1)^{\frac{1+\rho}{2}}} + r''_T,$$

with $r''_T = \frac{n_{T+1}}{n_{T+1}} r'_T \left(\|\Delta_T\| + C_M \frac{\sqrt{\ln(T+1)}}{(T+1)^{\frac{1+\rho}{2}}}\right) \mathbb{1}_{r'_T > c}$. Then, with the help of an induction, one has

$$\|\Delta_T\| \leq \tilde{\beta}_{T,0}\|\Delta_0\| + \sum_{t=0}^{T-1} \tilde{\beta}_{T,t+1} \frac{n_{t+1}}{N_{t+1}} C_M \frac{\sqrt{\ln(t+1)}}{(t+1)^{\frac{1+\rho}{2}}} + \sum_{t=0}^{T-1} \tilde{\beta}_{T,t+1} r''_t,$$

with $\tilde{\beta}_{T,t} = \prod_{j=t+1}^{T} \left(1 - c\frac{n_j}{N_j}\right)$ and $\beta_{T,T} = 1$. In addition, since for any t, one has $N_t \leq \frac{C_\rho}{1+\rho}\left((t+1)^{1+\rho} - 1\right)$, one has for any $t \leq T$,

$$\begin{aligned}\tilde{\beta}_{T,t} &\leq \exp\left(-c \sum_{j=t+1}^{T} \frac{n_j}{N_j}\right) \leq \exp\left(-c(1+\rho) \sum_{j=t+1}^{T} \frac{j^\rho}{(j+1)^{1+\rho}}\right) \\ &\leq \exp\left(-c(1+\rho)\left(\frac{t+1}{t+2}\right)^\rho \sum_{j=t+1}^{T} \frac{1}{j+1}\right) \leq \left(\frac{t+1}{T+1}\right)^{c_t},\end{aligned}$$

with $c_t = c(1+\rho)\left(\frac{t+1}{t+2}\right)^\rho \geq c(1+\rho)2^{-\rho}$. Taking $1 > c > 2^{\rho-1}$ and denoting $c_\rho = c2^{-\rho} > 1/2$, one has

$$\tilde{\beta}_{T,t} \leq \left(\frac{t+1}{T+1}\right)^{c_\rho(1+\rho)}.$$

Then, as a particular case, $\tilde{\beta}_{T,0} \leq \frac{1}{(T+1)^{c_\rho(1+\rho)}}$ and this term is so negligible. In addition, since

$$\sum_{t=0}^{T-1} \tilde{\beta}_{T,t+1} r''_t = \tilde{\beta}_{T,0} \sum_{t=0}^{T-1} \tilde{\beta}^{-1}_{t+1,0} r''_t,$$

and since $\mathbb{1}_{\{r'_T > c\}}$ converges almost surely to 0, one has

$$\sum_{t=0}^{T-1} \tilde{\beta}_{T,t+1} r''_t = \mathcal{O}\left(\tilde{\beta}_{T,0}\right) = \mathcal{O}\left(\frac{1}{(T+1)^{c_\rho(1+\rho)}}\right) \text{ a. s.},$$

and this term is so negligible as $c_\rho > 1/2$. Finally,

$$\sum_{t=0}^{T-1} \tilde{\beta}_{T,t+1} \frac{n_{t+1}}{N_{t+1}} C_M \frac{\sqrt{\ln(t+1)}}{(t+1)^{\frac{1+\rho}{2}}} \leq \sum_{t=0}^{T-1} \left(\frac{t+1}{T+1}\right)^{c_\rho(1+\rho)} \frac{n_{t+1}}{N_{t+1}} C_M \frac{\sqrt{\ln(t+1)}}{(t+1)^{\frac{1+\rho}{2}}} = \mathcal{O}\left(\frac{\sqrt{\ln T}}{T^{\frac{1+\rho}{2}}}\right) \text{ a. s.},$$

leading to $\|\Delta_T\| = \mathcal{O}\left(\sqrt{\frac{\ln T}{T^{1+\rho}}}\right)$ a.s., and

$$\|\theta_T - \theta^*\|^2 = O\left(\frac{\ln T}{T^{1+\rho}}\right) \text{ a. s.}, \qquad \text{and} \qquad \|\theta_T - \theta^*\|^2 = O\left(\frac{\ln N_T}{N_T}\right) \text{ a. s.}$$

Asymptotic efficiency. In order to get the asymptotic normality, we now have to give a better rate of convergence of $\|\Delta_T\|$. First, since $\bar{H}_t^{-1}$ converges to H^{-1}, $\|\delta_t\| \leq L_\delta \|\theta_t - \theta^*\|^2$, and with the help of the rate of convergence of θ_t, one has

$$\frac{1}{N_T}\left\|\sum_{t=0}^{T-1} n_{t+1}\bar{H}_t^{-1}\delta_t\right\| \leq \frac{L_\delta}{N_T}\sum_{t=0}^{T-1} n_{t+1}\left\|\bar{H}_t^{-1}\right\|_{\mathrm{op}}\|\theta_t-\theta^*\|^2 = \mathcal{O}\left(\frac{(\ln T)^2}{T^{1+\rho}}\right) \text{ a. s.},$$

which is a negligible term. In addition, since $\|\bar{H}_t^{-1} - H^{-1}\|_{\mathrm{op}} = \mathcal{O}\left(t^{-\nu}\right)$ a.s., one has

$$\begin{aligned}\frac{1}{N_T}\left\|\sum_{t=0}^{T-1} n_{t+1}\left(\bar{H}_t^{-1}-H^{-1}\right)H\left(\theta_t-\theta^*\right)\right\| &\leq \frac{1}{N_T}\|H\|_{\mathrm{op}}\sum_{t=0}^{T-1} n_{t+1}\left\|\bar{H}_t^{-1}-H^{-1}\right\|_{\mathrm{op}}\|\theta_t-\theta^*\| \\ &= \mathcal{O}\left(\frac{\ln(T)^{1/2+\mathbb{1}_{\{(1+\rho)/2+\nu=1\}}}}{T^{\rho+\min\{1,(1-\rho)/2+\nu\}}}\right) \text{ a. s.}\end{aligned}$$

Hence, as $\nu > 0$, this term is negligible, which thereby concludes the proof.

B.6 Proof of Theorems 6 and A.5

Let us first check that the assumptions on the learning rate (step-sequence) are satisfied: First, since for all $t \geq 1$ and $i = 1, \ldots, n_t$,

$$\frac{N_{Z,t,i}}{N_t} \xrightarrow[t\to+\infty]{\text{a.s.}} p,$$

we can observe that[4]

$$\frac{d(1-\iota)}{p\ln(t+1)^{w'}N_t^{1-\iota}}\sum_{t'=1}^{t}\ln(t'+1)^{w'}\sum_{i=1}^{n_{t'}} Z_{t',i}\iota_{t',i}e_{t',i}e_{t',i}^T \xrightarrow[t\to+\infty]{\text{a.s.}} I_d,$$

such that

$$\frac{1+\sum_{t'=1}^{t}\ln(t'+1)^{w'}\sum_{i=1}^{n'_t} Z_{t',i}}{\ln(t+1)^{w'}N_t} \xrightarrow[t\to+\infty]{\text{a.s.}} p,$$

Next, by definition of ι, we have

$$\lambda_{\max}\left(\bar{H}_{t,w'}^{-1}\right) = \mathcal{O}\left(t^{\iota(1+\rho)}\right) \text{ a. s.}, \qquad \text{and} \qquad \sum_{t\geq 1}\frac{\gamma_t^2}{n_t}\lambda_{\max}\left(\bar{H}_{t-1,w'}^{-1}\right)^2 < +\infty \text{ a. s.}$$

In addition, with $N_{Z,T} := 1 + \sum_{t=1}^{T}\ln(t+1)^{w'}\sum_{i=1}^{n_{t'}} Z_{t,i}$, one has

$$\begin{aligned}\frac{1}{N_{Z,T}}\sum_{t=1}^{T}\ln(t+1)^{w'}\sum_{i=1}^{n_t} Z_{t,i}\alpha_{t,i}\Phi_{t,i}\Phi_{t,i}^\top =& \frac{1}{N_{Z,T}}\sum_{t=1}^{T}\ln(t+1)^{w'}\nabla_\theta^2 F\left(\theta_{t-1}\right)\sum_{i=1}^{n_t} Z_{t,i} \\ &+ \frac{1}{N_{Z,T}}\sum_{t=1}^{T}\ln(t+1)^{w'}\xi_{Z,t},\end{aligned}$$

where $\xi_{Z,t} := \sum_{i=1}^{n_t} Z_{t,i}\alpha_{t,i}\Phi_{t,i}\Phi_{t,i}^\top - \sum_{i=1}^{n_t} Z_{t,i}\nabla_\theta^2 F\left(\theta_{t-1}\right)$ is a sequence of martingale differences for the filtration $\mathcal{F}'_{t-1} = \sigma\left(X_{1,1}, \ldots, X_{t-1,n_{t-1}}, Z_{t,1}, \ldots, Z_{t,n_t}\right)$. Thus,

$$\mathbb{E}\left[\|\xi_{Z,t}\|_F^{\eta'}|\mathcal{F}'_{t-1}\right] \leq 2^{\eta'-1}\left(\sum_{i=1}^{n_t} Z_{t,i}\mathbb{E}\left[\left\|\alpha_{t,i}\Phi_{t,i}\Phi_{t,i}^\top\right\|^{\eta'}|\mathcal{F}_{t-1}\right]^{\frac{1}{\eta'}}\right)^{\eta'} \leq 2^{\eta'-1}C_{\eta'}^{\eta'}\left(\sum_{i=1}^{n_t} Z_{t,i}\right)^{\eta'},$$

and with the help of a law of large numbers for martingales, one has

$$\left\|\sum_{t=1}^{T}\ln(t+1)^{w'}\xi_{Z,t}\right\|_F = o\left(N_{Z,T}\right) \text{ a. s.}$$

[4]E.g., see Godichon-Baggioni et al. (2024); Bercu et al. (2023) for more details.

Since for all $\theta \in \mathbb{R}^d$, $\left\|\nabla_\theta^2 F(\theta)\right\|_{\text{op}} \leq L_{\nabla G}$,

$$\left\| \frac{1}{N_{Z,T}} \sum_{t=1}^{T} \ln(t+1)^{w'} \nabla_\theta^2 F\left(\theta_{t-1}\right) \sum_{i=1}^{n_t} Z_{t,i} \right\|_{\text{op}} \leq L_{\nabla F}.$$

Then, $\lambda_{\max}\left(\bar{H}_{t,w'}\right) = \mathcal{O}(1)$ a.s., such that

$$\sum_{t\geq 1} \gamma_t \lambda_{\min}\left(\bar{H}_{t-1,w'}^{-1}\right) = +\infty \text{ a. s.}, \qquad \text{and} \qquad \frac{\lambda_{\max}\left(\bar{H}_{t,w'}^{-1}\right)^2 \gamma_{t+1}}{\lambda_{\min}\left(\bar{H}_{t,w'}^{-1}\right)} = \mathcal{O}\left(t^{2\iota(1+\rho)-1}\right) \text{ a. s.},$$

and the conditions in (A.2) are satisfied as soon as $i < \frac{1-\rho}{2(1+\rho)}$. Then, according to Theorem 5, θ_T converges almost surely to θ^*. By continuity, this implies that

$$\frac{1}{N_{Z,T}} \sum_{t=1}^{T} \ln(t+1)^{w'} \nabla^2 F\left(\theta_{t-1}\right) \sum_{i=1}^{n_t} Z_{t,i} \xrightarrow[T\to+\infty]{\text{a.s.}} \nabla_\theta^2 F(\theta^*)$$

meaning that $\bar{H}_{T,w'}$ and $\bar{H}_{T,w'}^{-1}$ converge almost surely to H and H^{-1}. Then, thanks to Theorem 5, one has that

$$\left\|\theta_{T,w} - \theta^*\right\|^2 = \mathcal{O}\left(\frac{\ln N_T}{N_T}\right) \text{ a. s.},$$

and since the Hessian is locally Lipschitz, $\left\|\nabla_\theta^2 F\left(\theta_t\right) - H\right\|_{\text{op}} = \mathcal{O}\left(\sqrt{\ln N_t} N_t^{-1/2}\right)$ a.s., and

$$\left\| \frac{1}{N_{Z,T}} \sum_{t=1}^{T} \ln(t+1)^{w'} \nabla^2 F\left(\theta_{t-1}\right) \sum_{i=1}^{n_t} Z_{t,i} - H \right\|_{\text{op}} = \mathcal{O}\left(\frac{1}{\ln(T+1)^{w'} N_T} \sum_{t=1}^{T} \ln(t+1)^{w'+1/2} n_t t^{-\frac{1+\rho}{2}} \right) \text{ a. s.},$$

one has

$$\left\| \frac{1}{N_{Z,T}} \sum_{t=1}^{T} \ln(t+1)^{w'} \nabla^2 F\left(\theta_{t-1}\right) \sum_{i=1}^{n_t} Z_{t,i} - H \right\|_{\text{op}} = \mathcal{O}\left(\frac{\sqrt{\ln(N_T)}}{\sqrt{N_T}} \right) \text{ a. s.}$$

In addition, since $\eta' \geq 2$ and

$$\mathbb{E}\left[\left\|\xi_{Z,t}\right\|_F^2 |\mathcal{F}_{t-1}\right] \leq \sum_{i=1}^{n_t} Z_{t,i}^2 \mathbb{E}\left[\left\|a\left(X_{t,i}, \theta_{t-1}\right) \Phi_{t,i} \Phi_{t,i}^T\right\|_F^2 |\mathcal{F}'_{t-1}\right] \leq n_t C_{\eta'}^{\frac{2}{\eta'}},$$

one has, with the help of a law of large numbers for martingales, that for all $\delta > 0$,

$$\left\| \frac{1}{N_{Z,T}} \sum_{t=1}^{T} \ln(t+1)^{w'} \xi_{Z,t} \right\|_F^2 = O\left(\frac{(\ln N_T)^{1+\delta}}{N_T} \right) \text{ a. s.}$$

Meaning, that

$$\left\| \frac{1}{N_{Z,T}} \sum_{t=1}^{T} \ln(t+1)^{w'} \sum_{i=1}^{n_t} Z_{t,i} \iota_{t,i} e_{t,i} e_{t,i}^T \right\|_{\text{op}} = \mathcal{O}\left(\frac{1}{T^{\iota(1+\rho)}} \right) \text{ a. s.},$$

and by definition of ι, one has

$$\left\|\bar{H}_{T,w'} - H\right\|^2 = \mathcal{O}\left(\frac{1}{N_T^{2\iota}}\right) \text{ a. s.}$$

Then, with the help of Theorem 5, one has

$$\sqrt{N_T}\left(\theta_T - \theta^*\right) \xrightarrow[T\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1} \Sigma H^{-1}\right).$$

B.7 Proof of Theorems 7 and A.6

As in the proof of Theorems 6 and A.5, one can easily check that the conditions in (A.2) are satisfied, such that Theorem A.1 hold, i.e., θ_T and $\theta_{T,w}$ converges almost surely to θ^*. In a same way, as in the proof of Theorem 6, one can easily get the consistency of $\bar{S}_{T,w'}$, leading with the help of Theorem A.2 to

$$\|\theta_T - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_T)}{N_T^{\frac{\gamma+\rho(1-\beta)}{1+\rho}}}\right) \text{ a. s.,} \qquad \text{and} \qquad \|\theta_{T,w} - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln N_T}{N_T^{\frac{\gamma+\rho(1-\beta)}{1+\rho}}}\right) \text{ a. s.}$$

In order to conclude the proof, we will now check that equality (8) is satisfied, i.e., that

$$\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^{w+1/2+\delta}\left(r_{t+1}+r'_{t+1}\right)t^{\frac{\gamma-\rho(\beta+1)}{2}} = \mathcal{O}\left(\frac{1}{T^{(1+\rho)v'}}\right) \text{ a. s.,}$$

with

$$r'_{t+1} = \frac{\ln(t+1)^{w'}}{N_{Z,t+1}}\sum_{i=1}^{n_{t+1}} Z_{t+1,i}\iota_{t+1,i} \quad \text{and} \quad r_{t+1} = \frac{\ln(t+1)^{w'}}{N_{Z,t+1}}\sum_{i=1}^{n_{t+1}} Z_{t+1,i}\left\|\alpha_{t+1,i}\Phi_{t+1,i}\right\|.$$

First, since $\sum_{i=1}^{n_{t+1}}\iota_{t+1,i} = \mathcal{O}\left(t^{-\iota(1+\rho)+\rho}\right)$, and since $\iota < \frac{\gamma-\rho\beta}{2(1+\rho)}$, one has

$$\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^{w+1/2+\delta}r'_{t+1}t^{\frac{\gamma-\rho(\beta+1)}{2}} = \mathcal{O}\left(\frac{\ln(T+1)^{3/2+\delta}}{T^{\iota(1+\rho)+1+\frac{\beta(1+\rho)-\gamma}{2}}}\right) \text{ a. s.,}$$

and since $\gamma - \beta\rho < 1$, it comes that $\iota(1+\rho)+1+\frac{\beta(1+\rho)-\gamma}{2} > \frac{1+\rho}{2}$. Considering the sequence of martingale differences $\Xi_{Z,t+1} = \sum_{i=1}^{n_{t+1}} Z_{t+1,i}\alpha_{t+1,i}\left\|\Phi_{t+1,i}\right\|^2 - \sum_{i=1}^{n_{t+1}} Z_{t+1,i}\mathbb{E}[\alpha_{t+1,i}\left\|\Phi_{t+1,i}\right\|^2|\mathcal{F}'_{t-1}]$, one has

$$\begin{aligned}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^{w+1/2+\delta}r_{t+1}t^{\frac{\gamma-\rho(\beta+1)}{2}} \leq & C_{\eta'}^{\frac{1}{\eta'}}\sum_{t=0}^{T-1}\frac{n_{t+1}\ln(t+1)^{w+w'+1/2+\delta}t^{\frac{\gamma-\rho(\beta+1)}{2}}}{N_{Z,t+1}}\sum_{i=1}^{n_{t+1}} Z_{t+1,i} \\ & + \sum_{t=0}^{T-1}\frac{n_{t+1}\ln(t+1)^{w+w'+1/2+\delta}t^{\frac{\gamma-\rho(\beta+1)}{2}}}{N_{Z,t+1}}\Xi_{Z,t+1}.\end{aligned}$$

Furthermore,

$$\frac{1}{s_T}\sum_{t=0}^{T-1}\frac{n_{t+1}\ln(t+1)^{w+w'+1/2+\delta}t^{\frac{\gamma-\rho(\beta+1)}{2}}}{N_{Z,t+1}}\sum_{i=1}^{n_{t+1}} Z_{t+1,i} = \mathcal{O}\left(\frac{\ln(T+1)^{3/2+\delta}}{T^{\min\left\{1+\frac{\rho(\beta+1)-\gamma}{2},1+\rho\right\}}}\right) \text{ a. s.,}$$

and since $\gamma - \beta\rho < 1$, one has that $1+\frac{\rho(\beta+1)-\gamma}{2} > \frac{1+\rho}{2}$. In addition, since

$$\mathbb{E}\left[\left\|\Xi_{Z,t+1}\right\|_F^{\eta'}\right] \leq \left(\sum_{i=1}^{n_{t+1}} Z_{t+1,i}\right)^{\eta'} C_{\eta'}^{\eta'},$$

and with the help of a law of large numbers for martingales,

$$\begin{aligned}&\left|\sum_{t=0}^{T-1}\frac{n_{t+1}\ln(t+1)^{w+w'+1/2+\delta}t^{\frac{\gamma-\rho(\beta+1)}{2}}}{N_{Z,t+1}}\Xi_{Z,t+1}\right| \\ &= o\left(\sum_{t=0}^{T-1}\frac{n_{t+1}\ln(t+1)^{w+w'+1/2+\delta}t^{\frac{\gamma-\rho(\beta+1)}{2}}}{N_{Z,t+1}}\sum_{i=1}^{n_{t+1}} Z_{t+1,i}\right) \text{ a. s.,}\end{aligned}$$

such that

$$\begin{aligned}&\frac{1}{\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^w}\left|\sum_{t=0}^{T-1}\frac{n_{t+1}\ln(t+1)^{w+w'+1/2+\delta}t^{\frac{\gamma-\rho(\beta+1)}{2}}}{N_{Z,t+1}}\Xi_{Z,t+1}\right| \\ &\qquad = o\left(\frac{\ln(T+1)^{3/2+\delta}}{T^{\min\left\{1+\frac{\rho(\beta+1)-\gamma}{2},1+\rho\right\}}}\right) \text{ a. s.,}\end{aligned}$$

which concludes the proof.

B.8 Proof of Theorems A.7 and A.8

First, since the conditions in (4) (or in (A.2)) are satisfied, one has that θ_t and $\theta_{t,w}$ converge almost surely to θ^*. Let us now prove that it implies the convergence of G_t.

Convergence of G_t. For all coordinate j, let us now consider

$$\tilde{G}_T^{(j)} := \frac{1}{N_T}\sum_{t=1}^{T}\sum_{i=1}^{n_t}\left(\frac{\partial}{\partial j}f\left(\theta_{t-1,w};\xi_{t,i}\right)\right)^2.$$

Then, denoting

$$V_j := \mathbb{E}\left[\left(\frac{\partial}{\partial j}f\left(\theta^*;\xi\right)\right)^2\right] = \mathbb{V}\left[\frac{\partial}{\partial j}f\left(\theta^*;\xi\right)\right],$$

one has

$$\tilde{G}_T^{(j)} - V_j = \frac{1}{N_T}\sum_{t=1}^{T} n_t\left(\mathbb{E}\left[\left(\frac{\partial}{\partial j}f\left(\theta_{t-1,w};\xi_{t,1}\right)\right)^2|\mathcal{F}_{t-1}\right] - V_j\right) + \frac{1}{N_T}\sum_{t=1}^{T}\Xi_t$$

where $\Xi_t = \sum_{i=1}^{n_t}\left(\frac{\partial}{\partial j}f\left(\theta_{t-1,w};X_{t,i}\right)\right)^2 - n_t\mathbb{E}\left[\left(\frac{\partial}{\partial j}f\left(\theta_{t-1,w},\xi\right)\right)^2\right]$ is a martingale difference. Then, thanks to (A.3) coupled with Duflo (2013, Proposition 1.III.19), we have

$$\frac{1}{N_T}\sum_{t=1}^{T}\Xi_t \xrightarrow[n\to+\infty]{\text{a.s.}} 0.$$

In addition, since the functional $\theta \longmapsto \mathbb{E}\left[\nabla_\theta f(\theta;\xi)\nabla_\theta f(\theta;\xi)^\top\right]$ is continuous at θ^*, one has for all j

$$\frac{1}{N_T}\sum_{t=1}^{T} n_t\left(\mathbb{E}\left[\left(\frac{\partial}{\partial j}f\left(\theta_{t-1,w},\xi_{t,1}\right)\right)^2|\mathcal{F}_{t-1}\right] - V_j\right) \xrightarrow[T\to+\infty]{\text{a.s.}} 0,$$

such that, for all j,

$$\tilde{G}_T^{(j)} = \frac{1}{N_T}\sum_{t=1}^{T}\sum_{i=1}^{n_t}\left(\frac{\partial}{\partial j}f\left(\theta_{t-1,w};\xi_{t,i}\right)\right)^2 \xrightarrow[T\to+\infty]{a.s} \mathbb{V}\left[\frac{\partial}{\partial j}f\left(\theta^*;\xi\right)\right] > 0.$$

Then, G_t converges almost surely to the diagonal matrix G, whose diagonal elements are given by $G^{(j)} = \mathbb{V}\left[\frac{\partial}{\partial j}f(\theta^*;\xi)\right]^{-1/2}$.

Rate of convergence of θ_T. With the help of Theorem A.2, one has that

$$\|\theta_T - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(T)}{T^{\gamma+\rho(1-\beta)}}\right) \text{ a. s.}, \qquad \text{and} \qquad \|\theta_{T,w} - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(T)}{T^{\gamma+\rho(1-\beta)}}\right) \text{ a. s.},$$

which can also be written as

$$\|\theta_T - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_T)}{N_T^{\frac{\gamma+\rho(1-\beta)}{1+\rho}}}\right) \text{ a. s.}, \qquad \text{and} \qquad \|\theta_{T,w} - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_T)}{N_T^{\frac{\gamma+\rho(1-\beta)}{1+\rho}}}\right) \text{ a. s.}$$

Rate of convergence of $\theta_{T,w}$. Let us consider the event:

$$E_t = \left\{\exists j, G_t^{(j)} \neq \left(\frac{1}{N_T}\left(G_0^{(j)} + \sum_{t=1}^{T}\sum_{i=1}^{n_t}\left(\frac{\partial}{\partial j}f_{t,i}\left(\theta_{t-1,w}\right)\right)^2\right)\right)^{-1/2}\right\},$$

where $f_{t,i}\left(\theta_{t-1,w}\right) := f\left(\theta_{t-1,w};\xi_{t,i}\right)$. Observe that since G_t converges to G, $\mathbb{1}_{\{E_t\}}$ converges almost surely to 0, such that

$$\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^{w+1/2+\delta}\left\|G_{t+1}^{-1}-G_t^{-1}\right\|_{\mathrm{op}}\mathbb{1}_{\{E_t\bigcup E_{t+1}\}}(t+1)^{\frac{\gamma-\rho(\beta+1)}{2}} = \mathcal{O}\left(\frac{1}{T^{(1+\rho)}\ln(T)^w}\right) \text{ a. s.}$$

In addition, on $\{E_t^C\bigcap E_{t+1}^C\}$, one has

$$\begin{aligned}
&\left(G_{t+1}^{-1}-G_t^{-1}\right)\mathbb{1}_{\{E_t^C\bigcap E_{t+1}^C\}} = \left(G_{t+1}^{-1}+G_t^{-1}\right)^{-1}\left(G_{t+1}^{-2}-G_t^{-2}\right)\mathbb{1}_{E_t^C\bigcap E_{t+1}^C}\\
&= \left(G_t^{-1}+G_{t+1}^{-1}\right)^{-1}\frac{1}{N_{t+1}}\left(\mathrm{diag}\left(\sum_{i=1}^{n_{t+1}}\left(\frac{\partial}{\partial j}f_{t+1,i}\left(\theta_{t,w}\right)\right)^2\right)_{j=1,\ldots,d} - n_{t+1}G_t^{-2}\right)\mathbb{1}_{\{E_t^C\bigcap E_{t+1}^C\}},
\end{aligned}$$

where $\mathrm{diag}\left(\sum_{i=1}^{n_{t+1}}\left(\frac{\partial}{\partial j}f_{t+1,i}\left(\theta_{t,w}\right)\right)^2\right)_{j=1,\ldots,d}$ is the diagonal matrix whose elements are $\sum_{i=1}^{n_{t+1}}\left(\frac{\partial}{\partial j}f_{t+1,i}\left(\theta_{t,w}\right)\right)^2$. Observe that since G_t converges almost surely to a positive matrix, there are positive constants c_{ada}, C_{ada} such that $\mathbb{1}_{\{E_{t,1}\}}$ converges almost surely to 1, where $E_{t,1} := \{c_{ada} < \lambda_{\min}\left(G_t\right) \le \lambda_{\max}\left(G_t\right) < C_{ada}\}$. Then,

$$\begin{aligned}
&\|(G_{t+1}-G_t^{-1})^{-1}\|_{\mathrm{op}}\mathbb{1}_{\{E_t^C\bigcap E_{t+1}^C\}}\\
&\le \|G_{t+1}^{-1}+G_t^{-1}\|_{\mathrm{op}}\frac{1}{N_{t+1}}\left\|\mathrm{diag}\left(\sum_{i=1}^{n_{t+1}}\left(\frac{\partial}{\partial j}f_{t+1,i}\left(\theta_{t,w}\right)\right)^2\right) - n_{t+1}G_t^{-2}\right\|_{\mathrm{op}}\mathbb{1}_{\{E_{t,1}^C\bigcup E_{t+1,1}^C\}}\\
&+ 2C_{ada}\frac{1}{N_{t+1}}\left(\sum_{j=1}^{d}\sum_{i=1}^{n_{t+1}}\left(\frac{\partial}{\partial j}f_{t+1,i}\left(\theta_{t,w}\right)\right)^2 + n_{t+1}c_{ada}^{-2}\right).
\end{aligned}$$

In addition, since $\mathbb{1}_{\{E_{t,1}^C\}}$ converges almost surely to 0, one can easily check that

$$\begin{aligned}
&\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^{w+1/2+\delta}(t+1)^{\frac{\gamma-(\beta+1)\rho}{2}}\\
&\times\left\|G_t^{-1}+G_{t+1}^{-1}\right\|_{\mathrm{op}}\frac{1}{N_{t+1}}\left\|\mathrm{diag}\left(\sum_{i=1}^{n_{t+1}}\left(\frac{\partial}{\partial j}f_{t+1,i}\left(\theta_{t,w}\right)\right)^2\right) - n_{t+1}G_t^{-2}\right\|_{\mathrm{op}}\mathbb{1}_{\{E_{t,1}^C\bigcup E_{t+1,1}^C\}}\\
&= \mathcal{O}\left(\frac{1}{T^{(1+\rho)}\ln(T)^w}\right) \text{ a. s.}
\end{aligned}$$

In addition,

$$\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^{w+1/2+\delta}(t+1)^{\frac{\gamma-(\beta+1)\rho}{2}}2C_{ada}c_{ada}^{-2}\frac{n_{t+1}}{N_{t+1}} = O\left(\frac{\ln(T)^{\delta+1/2-w}}{T^{\frac{2-\gamma+(\beta+1)\rho}{2}}}\right) \text{ a. s.},$$

which is negligible as soon as $\gamma-\beta\rho<1$. In addition, remark that

$$\frac{1}{N_{t+1}}\sum_{j=1}^{d}\sum_{i=1}^{n_{t+1}}\left(\frac{\partial}{\partial j}f_{t+1,i}\left(\theta_{t,w}\right)\right)^2 = \frac{1}{N_{t+1}}\sum_{j=1}^{d}\mathbb{E}\left[\sum_{i=1}^{n_{t+1}}\left(\frac{\partial}{\partial j}f_{t+1,i}\left(\theta_{t,w}\right)\right)^2\middle|\mathcal{F}_t\right] + \frac{n_{t+1}}{N_{t+1}}\tilde{\xi}_{t+1},$$

with $\tilde{\xi}_{t+1} = \sum_{j=1}^{d}\left(\sum_{i=1}^{n_{t+1}}\left(\frac{\partial}{\partial j}f_{t+1,i}\left(\theta_{t,w}\right)\right)^2 - \mathbb{E}\left[\sum_{i=1}^{n_{t+1}}\left(\frac{\partial}{\partial j}f_{t+1,i}\left(\theta_{t,w}\right)\right)^2\middle|\mathcal{F}_t\right]\right)$. Since $\theta_{t,w}$ converges almost surely to θ^* and with the help of inequality (5), one has

$$\begin{aligned}
&\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^{w+1/2+\delta}(t+1)^{\frac{\gamma-(\beta+1)\rho}{2}}\frac{1}{N_{t+1}}\sum_{j=1}^{d}\mathbb{E}\left[\sum_{i=1}^{n_{t+1}}\left(\frac{\partial}{\partial j}f_{t+1,i}\left(\theta_{t,w}\right)\right)^2|\mathcal{F}_t\right]\\
&\quad= \mathcal{O}\left(\frac{\ln(T)^{\delta+1/2-w}}{T^{\frac{2-\gamma+(\beta+1)\rho}{2}}}\right) \text{ a. s.},
\end{aligned}$$

while with the help of a law of large numbers for martingales (e.g., see Duflo (2013)), one has

$$\frac{1}{s_T}\sum_{t=0}^{T-1} n_{t+1}\ln(t+1)^{w+1/2+\delta}(t+1)^{\frac{\gamma-(\beta+1)\rho}{2}}\frac{n_{t+1}}{N_{t+1}}\tilde{\xi}_{t+1} = o\left(\frac{\ln(T)^{\delta+1/2-w}}{T^{\frac{2-\gamma+(\beta+1)\rho}{2}}}\right) \text{ a. s.},$$

which concludes the proof.

References

Antonakopoulos, K., Mertikopoulos, P., Piliouras, G., and Wang, X. (2022). Adagrad avoids saddle points. In *International Conference on Machine Learning*, pages 731–771. PMLR.

Benveniste, A., Métivier, M., and Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximations*, volume 22 of *Applications of Mathematics*. Springer-Verlag, New York.

Bercu, B., Bigot, J., Gadat, S., and Siviero, E. (2023). A stochastic gauss–newton algorithm for regularized semi-discrete optimal transport. *Information and Inference: A Journal of the IMA*, 12(1):390–447.

Bercu, B., Godichon, A., and Portier, B. (2020). An efficient stochastic newton algorithm for parameter estimation in logistic regressions. *SIAM Journal on Control and Optimization*, 58(1):348–367.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311.

Boyer, C. and Godichon-Baggioni, A. (2023). On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. *Computational Optimization and Applications*, 84(3):921–972.

Cénac, P., Godichon-Baggioni, A., and Portier, B. (2020). An efficient averaged stochastic Gauss-Newton algorithm for estimating parameters of non linear regressions models. *arXiv preprint arXiv:2006.12920*.

Chau, H., Kirkby, J., Nguyen, D., Nguyen, D., Nguyen, N., and Nguyen, T. (2022). On the inversion-free newton's method and its applications. Technical report, Working paper.

Dozat, T. (2016). Incorporating nesterov momentum into adam. In *International Conference on Learning Representations (ICLR)*. Workshop Track.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Duflo, M. (2013). *Random iterative models*, volume 34. Springer Science & Business Media.

Fabian, V. (1968). On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, pages 1327–1332.

Gadat, S. and Gavra, I. (2022). Asymptotic study of stochastic adaptive algorithms in non-convex landscape. *The Journal of Machine Learning Research*, 23(1):10357–10410.

Gadat, S. and Panloup, F. (2023). Optimal non-asymptotic analysis of the ruppert–polyak averaging stochastic algorithm. *Stochastic Processes and their Applications*, 156:312–348.

Gazagnadou, N., Gower, R., and Salmon, J. (2019). Optimal mini-batch and step sizes for saga. In *International conference on machine learning*, pages 2142–2150. PMLR.

Godichon-Baggioni, A. (2019a). Lp and almost sure rates of convergence of averaged stochastic gradient algorithms: locally strongly convex objective. *ESAIM: Probability and Statistics*, 23:841–873.

Godichon-Baggioni, A. (2019b). Online estimation of the asymptotic variance for averaged stochastic gradient algorithms. *Journal of Statistical Planning and Inference*, 203:1–19.

Godichon-Baggioni, A. and Lu, W. (2023). Online stochastic newton methods for estimating the geometric median and applications. *arXiv preprint arXiv:2304.00770*.

Godichon-Baggioni, A., Lu, W., and Portier, B. (2024). Recursive ridge regression using second-order stochastic algorithms. *Computational Statistics & Data Analysis*, 190:107854.

Godichon-Baggioni, A. and Tarrago, P. (2023). Non asymptotic analysis of adaptive stochastic gradient algorithms and applications. *arXiv preprint arXiv:2303.01370*.

Godichon-Baggioni, A., Werge, N., and Wintenberger, O. (2023a). Learning from time-dependent streaming data with online stochastic algorithms. *Transactions on Machine Learning Research*.

Godichon-Baggioni, A., Werge, N., and Wintenberger, O. (2023b). Non-asymptotic analysis of stochastic approximation algorithms for streaming data. *ESAIM: Probability and Statistics*, 27:482–514.

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.

Gower, R. M., Richtárik, P., and Bach, F. (2021). Stochastic quasi-gradient methods: Variance reduction via jacobian sketching. *Mathematical Programming*, 188:135–192.

Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Kushner, H. and Yin, G. (2003). *Stochastic approximation and recursive algorithms*. Springer-Verlag NY.

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*.

Leluc, R. and Portier, F. (2023). Asymptotic analysis of conditioned stochastic gradient descent. *Transactions on Machine Learning Research*.

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.

Moulines, E. and Bach, F. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24.

Needell, D., Ward, R., and Srebro, N. (2014). Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Advances in neural information processing systems*, 27.

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.

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.

Pelletier, M. (1998). On the almost sure asymptotic behaviour of stochastic algorithms. *Stochastic processes and their applications*, 78(2):217–244.

Pelletier, M. (2000). Asymptotic almost sure efficiency of averaged stochastic algorithms. *SIAM Journal on Control and Optimization*, 39(1):49–72.

Polyak, B. and Juditsky, A. (1992). Acceleration of stochastic approximation. *SIAM J. Control and Optimization*, 30:838–855.

Reddi, S. J., Kale, S., and Kumar, S. (2018). On the convergence of adam and beyond. In *International Conference on Learning Representations*.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Ruppert, D. (1988). Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. (2021). *Lectures on stochastic programming: modeling and theory*. SIAM.

Sherman, J. and Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tieleman, T. and Hinton, G. (2012). Rmsprop: Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. *COURSERA Neural Networks Mach. Learn*, 17.

Toulis, P. and Airoldi, E. (2017). Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727.

Vaswani, S., Bach, F., and Schmidt, M. (2019). Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pages 1195–1204. PMLR.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.