Journal of Machine Learning Research 26 (2025) 1-49 Submitted 11/23; Revised 12/24; Published 3/25

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

Antoine Godichon-Baggioni ANTOINE.GODICHON_BAGGIONI@SORBONNE-UNIVERSITE.FR
Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université
Paris, France

Nicklas Werge WERGE@SDU.DK
Department of Mathematics and Computer Science
University of Southern Denmark
Odense, Denmark

Editor: Aurelien Garivier

Abstract

Stochastic optimization methods face new challenges in the realm of streaming data, characterized by a continuous flow of large, high-dimensional data. While first-order methods, like stochastic gradient descent, are the natural choice for such data, they often struggle with ill-conditioned problems. In contrast, second-order methods, such as Newton's method, offer a potential solution but are computationally impractical for large-scale streaming applications. This paper introduces adaptive stochastic optimization methods that effectively address ill-conditioned problems while functioning in a streaming context. Specifically, we present adaptive inversion-free stochastic quasi-Newton methods with 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 points. Theoretical analysis establishes their asymptotic efficiency, and empirical studies demonstrate their effectiveness in scenarios with complex covariance structures and poor initializations. In particular, we demonstrate that our adaptive quasi-Newton methods can outperform or match existing first- and second-order methods.

Keywords: stochastic optimization, stochastic gradient methods, second-order methods, online learning, large-scale

1. Introduction

This paper focuses on the stochastic optimization problem, where the objective is to minimize a convex function $F : \mathbb{R}^d \to \mathbb{R}$ with $d \in \mathbb{N}$. Formally, the goal is to estimate

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

where f is a loss function, ξ is a random variable following an unknown distribution Ξ, and θ is the parameter of interest. This formulation is common in many machine learning applications (Kushner and Yin, 2003; Bottou et al., 2018; Sutton and Barto, 2018). For instance, when $\xi = (X, Y)$ represents an input-output pair, the function f typically takes the form $f(\theta; \xi) = l(h_\theta(X), Y)$, where l is a loss function onto $\mathbb{R}$ and h_θ is a prediction model parameterized by θ.

We tackle the stochastic optimization problem in (1) within a streaming context, where data are large in size, high in dimensionality, and arrive continuously as time-varying mini-batches. Following the framework studied in Godichon-Baggioni et al. (2023b,a), we consider an infinite sequence of independent and identically distributed (i.i.d.) samples of the random variable ξ, denoted as (ξ_t). Each ξ_t represents a block of n_t data points, $\{\xi_{t,1}, \ldots, \xi_{t,n_t}\}$. This setup mirrors the incremental and block-based nature of real-world streaming data.

Our adaptive stochastic optimization methods advance beyond the conventional stochastic gradient-based methods by incorporating a Hessian matrix approximation, A_t, at each step to refine the descent direction. These 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 $\nabla_\theta f(\theta_t; \xi_{t+1}) = n_{t+1}^{-1} \sum_{i=1}^{n_{t+1}} \nabla_\theta f(\theta_t; \xi_{t+1,i})$. Here, (γ_t) is the learning rate, (A_t) is the sequence of random matrices in $\mathbb{R}^{d\times d}$, and $(\nabla_\theta f(\theta_t; \xi_{t+1,i}))$ is unbiased gradient estimates.

The update in (2) reduces to the classical Robbins-Monro method (Robbins and Monro, 1951), commonly known as Stochastic Gradient Descent (SGD), if we set $A_t = I_d$ and $n_t = 1$. When $A_t = I_d$ and $n_t \in \mathbb{N}$, (2) forms a streaming version of SGD with time-varying mini-batches (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 $(\nabla_\theta f(\theta_t; \xi_{t+1}))$. Furthermore, the update in (2) transforms into a stochastic quasi-Newton method, when A_t serves as an approximation of the inverse Hessian matrix $\nabla_\theta^2 F(\theta_t)$.

Given the streaming nature of the data, it is essential that A_t is updated directly (i.e., inversion-free) and sparsely to preserve low computational complexity. However, these infrequent updates could potentially degrade convergence. To counteract this, we incorporate acceleration techniques that enhance convergence. Specifically, we consider an iterative weighted Polyak-Ruppert averaging scheme initialized at $\theta_{0,w} = \theta_0$, defined recursively by

$$\theta_{t+1,w} = \theta_{t,w} + \frac{n_{t+1} \ln(t+1)^w}{\sum_{i=0}^{t} n_{i+1} \ln(i+1)^w} (\theta_t - \theta_{t,w}), \quad w \geq 0. \tag{3}$$

Setting $w = 0$ results in the usual Polyak-Ruppert averaging scheme (Ruppert, 1988; Polyak and Juditsky, 1992; Godichon-Baggioni et al., 2023b). However, this standard Polyak-Ruppert averaging scheme can be prone to bad initializations. Instead, the iterative weighted version in (3) assigns more weight to the newer estimates of (2), which limits the effect of poor initializations (Mokkadem and Pelletier, 2011; Boyer and Godichon-Baggioni, 2023).

Our goal is to develop adaptive stochastic optimization methods for streaming data that are: i) computationally efficient, ii) robust to ill-conditioned problems, and iii) exhibit optimal convergence in both theory and practice. Thus, the central question of this paper is:

> *Can we construct a sequence of Hessian approximations (A_t) that are both computationally efficient and ensure that our adaptive methods are robust to ill-conditioned problems while exhibiting optimal convergence properties?*

Contributions Our paper makes several contributions to the field of stochastic optimization within a streaming context. Firstly, we introduce adaptive stochastic optimization methods that effectively manage ill-conditioned problems while maintaining computational

efficiency. These methods dynamically adjust learning rates on a per-dimension basis by leveraging historical gradient and Hessian information. Secondly, we propose iterative weighted average versions of these adaptive methods, which provide variance reduction during learning and accelerated convergence. Our theoretical analysis establishes their strong consistency, rate of convergence, and asymptotic efficiency.

A key contribution of our work is our adaptive inversion-free stochastic quasi-Newton methods, which 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. Specifically, our adaptive quasi-Newton methods: i) use second-order information to better handle ill-conditioned problems, ii) maintain the computational efficiency of first-order methods, and iii) incorporate acceleration techniques. By leveraging acceleration techniques, our adaptive quasi-Newton method mitigate the need for frequent updates of the Hessian approximation, ensuring that performance/convergence rates are not compromised despite the lower computational costs.

In addition to this adaptive quasi-Newton method, we also apply our methodology to develop a streaming version of AdaGrad along with its iterative weighted average version.

Our adaptive methods can be applied to a wide range of models, e.g., linear, logistic, softmax, ridge, and non-linear regression, as well as the estimation of the geometric median and optimal transport (Bercu et al., 2020; Boyer and Godichon-Baggioni, 2023; Godichon-Baggioni et al., 2024; Cénac et al., 2020; Godichon-Baggioni and Lu, 2024; Bercu et al., 2023). To demonstrate the effectiveness of our methods, we provide several examples, specifically focusing on linear and logistic regression, as well as the estimation of the geometric median.

Organization In Section 2, we briefly review related work. Section 3 presents the theoretical framework for our adaptive methods. In Section 4, we establish their asymptotic efficiencies. Section 5 details our adaptive quasi-Newton methods and application examples. Finally, Section 6 demonstrates the efficiency of our proposed methods on both synthetic and real-world datasets.

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

2. Related Work

Stochastic optimization and adaptive methods have been extensively researched, as evident in works such as Bottou et al. (2018); Chau et al. (2024). Theoretical investigations into SGD cover a wide range of 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; Nocedal and Wright, 1999; Boyd and Vandenberghe, 2004). A noteworthy extension of SGD is the concept of averaging, known for its role in accelerating convergence. This technique, known 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 decay rates, 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 method can be sensitive to ill-conditioned problems, 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 (Tieleman and Hinton, 2012; Kingma and Ba, 2015; Zeiler, 2012; Dozat, 2016; 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 lead to sub-optimal outcomes in practice.

To address these issues, an alternative approach involves considering stochastic (inversion-free) Newton 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 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, 2024), non-linear regression (Cénac et al., 2020), and optimal transport (Bercu et al., 2023).

Our adaptive stochastic optimization methods integrate the strengths of first-order adaptive methods, acceleration techniques such as iterative weighted Polyak-Ruppert averaging, and second-order methods. This novel combination allows us to develop methods that are computationally efficient, robust to ill-conditioned problems, and achieve optimal convergence both in theory and practice.

3. Underlying Theoretical Framework

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

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

Our theoretical framework relies on three key assumptions. These assumptions, which depend on the differentiability of the function F, are standard in the realms of stochastic optimization, stochastic approximation, and adaptive methods (Bottou et al., 2018; Leluc

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

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' such that*

$$\mathbb{E}[\|\nabla_\theta f(\theta;\xi)\|^2] \leq C + C'(F(\theta) - F(\theta^*)), \quad \forall \theta \in \mathbb{R}^d. \tag{4}$$

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, $\mathbb{E}[\|\nabla_\theta f(\theta;\xi)\|^2]$ is not confined by a constant or the squared errors $\|\theta - \theta\|^2$. Instead, we use 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 $C = 0$, (4) is known as the weak growth condition (Vaswani et al., 2019; Nguyen et al., 2018). Notably, in the context of μ-strong convexity of the function F, the squared errors condition implies the functional error condition, as $\|\theta - \theta^*\|^2 \leq 2/\mu(F(\theta) - F(\theta^*))$ for any $\theta \in \mathbb{R}^d$.

To establish the strong consistency of our method's estimates, we introduce a second assumption that enables the use of a second-order Taylor expansion to the functional F.

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

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 convex and $\lambda_{\min} := \lambda_{\min}(\nabla^2_\theta F(\theta^*)) > 0$.*

Note that this convexity assumption is important when using stochastic (quasi-)Newton methods, particularly to ensure the positivity of each adaptive step, which aims to estimate the inverse of the Hessian.

4. Theoretical Analysis

In this section, we present the theoretical analysis of our adaptive stochastic optimization methods and their iterated weighted averaged versions, as described in equations (2) and (3), respectively. Specifically, we demonstrate strong consistency, rate of convergence, and asymptotic normality. For clarity, this section focuses on constant mini-batches of size n, leaving the discussion of time-varying mini-batches n_t to Appendix A.

Focusing on constant mini-batches provides a more clear presentation of the foundational properties of our methods, such as computational efficiency, robustness to ill-conditioned problems, and optimal convergence. Readers seeking a deeper understanding, particularly of extensions to time-varying mini-batches, are encouraged to refer to Appendix A and Appendix C, which provide detailed proofs and adaptations. The consideration of time-varying mini-batches builds on recent work by Godichon-Baggioni et al. (2023a), which highlights their potential to accelerate convergence and mitigate both long- and short-term dependence structures.

4.1 Adaptive Stochastic Optimization Methods

From this point forward, we focus on constant mini-batches of size n. At each time t, a mini-batch of n i.i.d. samples of ξ, represented by $\xi_t = \{\xi_{t,1}, \ldots, \xi_{t,n}\}$, arrives, and the total number of data points processed, N_t, is tn. Thus, the updates of our adaptive stochastic optimization methods, as described in (2), are given by:

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

where $\nabla_\theta f(\theta_t; \xi_{t+1}) = n^{-1} \sum_{i=1}^n \nabla_\theta f(\theta_t; \xi_{t+1,i})$. Let $(\mathcal{F}_t)$ be the filtration such that θ_t and A_t are $\mathcal{F}_t$-measurable, and the incoming mini-batch ξ_{t+1} is independent of $\mathcal{F}_t$.

Our goal is to recursively update θ_t at each time step t to integrate the most recent information ξ_{t+1}. Since the stochastic gradient estimates are approximations of the gradient of (1), it is essential to control the step lengths $(\gamma_{t+1} A_t)$ to guarantee convergence. For the subsequent discussion, we assume that the learning rate (γ_t) and the sequence of random matrices (A_t) satisfy the following conditions:

$$\sum_{t \geq 1} \gamma_t \lambda_{\min}(A_{t-1}) = +\infty \text{ a. s.}, \quad \text{and} \quad \sum_{t \geq 1} \gamma_t^2 \lambda_{\max}(A_{t-1})^2 < +\infty \text{ a. s.} \tag{5}$$

In Section 5, we will discuss the modifications needed to ensure these conditions are met. For example, if $A_t = I_d$ the conditions in (5) reduces to the usual conditions on the learning rate (γ_t), e.g., see Robbins and Monro (1951). Additional details about (5) can be found in works such as Boyer and Godichon-Baggioni (2023); Godichon-Baggioni and Tarrago (2023).

For simplicity, we set $\gamma_t = C_\gamma t^{-\gamma}$ with $C_\gamma > 0$ and $\gamma \in (1/2, 1)$. 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 valid.

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

Theorem 1 *Suppose Assumptions 1 to 3 hold, along with the conditions in* (5)*. Then, θ_t converges almost surely to θ^*.*

The proof is given in Appendix C.1. 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 some $A \succ 0$.

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, A 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). For more details, see Boyer and Godichon-Baggioni (2023); Leluc and Portier (2023) or the proofs of Corollaries 1 to 3.

Theorem 2 *Suppose Assumptions 1 to 4 hold, along with the conditions in* (5)*. In addition, assume there exist positive constants C_η and $\eta > \frac{1}{\gamma} - 1$ such that*

$$\mathbb{E}\left[\|\nabla_\theta f(\theta; \xi)\|^{2+2\eta}\right] \leq C_\eta \left(1 + F(\theta) - F(\theta^*)\right)^{1+\eta}, \quad \forall \theta \in \mathbb{R}^d. \tag{6}$$

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

The proof is provided in Appendix C.2. Observe that this type of inequality is usual to establish asymptotic rate of convergence of stochastic gradient algorithms (see, for example, Pelletier (1998, Theorem 2) among others) and it is satisfied as soon as $\nabla_\theta f$ admits fourth-order moment.

4.2 Weighted Averaged Adaptive Stochastic Optimization Methods

For constant mini-batches, the weighted average of our adaptive stochastic optimization methods in (3) simplifies to:

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

with $\theta_{0,w} = \theta_0$. The case involving time-varying mini-batches are discussed in Appendix A.

To establish the convergence rate and the optimal asymptotic normality, we make the following assumption about the Hessian of F.

Assumption 5 *There exist positive constants L_η and η such that*

$$\|\nabla_\theta F(\theta) - \nabla^2_\theta F(\theta^*)(\theta - \theta^*)\| \leq L_\eta \|\theta - \theta^*\|^2, \quad \forall \theta \in \mathcal{B}(\theta^*, \eta).$$

Assumption 5 is satisfied as soon as the Hessian of F is locally Lipschitz on a neighborhood around θ^*. Coupled with Assumption 2, this implies the existence of a positive constant L_δ such that

$$\|\nabla_\theta F(\theta) - \nabla^2_\theta F(\theta^*)(\theta - \theta^*)\| \leq L_\delta \|\theta - \theta^*\|^2, \quad \forall \theta \in \mathbb{R}^d.$$

Theorem 3 *Suppose Assumptions 1 to 5 hold, along with the conditions in* (5) *and* (6)*. In addition, assume there exists a positive constant ν such that*

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

Then, $\theta_{t,w}$ converges almost surely to θ^*, and

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

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

The proof can be found in Appendix C.3. 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) ensures the asymptotic efficiency of the weighted average estimates $(\theta_{t,w})$.

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

Theorem 4 *Suppose Assumptions 1 to 5 hold, along with the conditions in* (5) *and* (6)*. In addition, assume there exists a positive constant* $v' > 1/2$ *such that*

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

for some $\delta > 0$*. Then,* $\theta_{t,w}$ *converges almost surely to* θ^**,* $\|\theta_{t,w} - \theta^*\|^2 = \mathcal{O}\left(\ln(N_t) N_t^{-1}\right)$ a. s.*, and* $\sqrt{N_t}(\theta_{t,w} - \theta^*) \xrightarrow[t \to +\infty]{\mathcal{L}} \mathcal{N}(0, \nabla_\theta^2 F(\theta^*)^{-1} \Sigma \nabla_\theta^2 F(\theta^*)^{-1})$.

See Appendix C.4 for a proof. Note that, although the condition in (8) may seem unusual, it is straightforward to verify in practice. For example, the proofs of Theorems B.6 and B.9 provide insights into practical methods for verifying this condition.

5. Applications to quasi-Newton's methods

In this section, we apply our adaptive stochastic optimization methodology from Section 4 to quasi-Newton's methods. Specifically, in Section 5.1, we introduce an adaptive inversion-free stochastic quasi-Newton method and its iterative weighted average version, both designed to match the computational efficiency of first-order stochastic optimizations methods.

Next, in Section 5.2, we provide three specific examples: linear regression, logistic regression, and geometric median estimation. Nonetheless, our methods are also applicable to other models, such as softmax regression (Boyer and Godichon-Baggioni, 2023), ridge regression (Godichon-Baggioni et al., 2024), non-linear regression (Cénac et al., 2020), and optimal transport (Bercu et al., 2023).

In Appendix B, we present additional applications of our methodology to Newton's method and AdaGrad, with corresponding proofs in Appendix C. In particular, we introduce a novel streaming variant of AdaGrad, along with its iterative weighted Polyak-Ruppert average counterpart, in Appendix B.3.

5.1 Adaptive Inversion-Free Stochastic Quasi-Newton Methods with $\mathcal{O}(dN_t)$ Operations

To overcome the computational challenges associated with Hessian inversion, we propose a variant of the stochastic quasi-Newton's method that entirely avoids Hessian inversion, as seen in Bercu et al. (2020); Boyer and Godichon-Baggioni (2023). We further enhance this approach to develop Streaming Stochastic quasi-Newton (SSN) methods, which operate with only $\mathcal{O}(dN_t)$ operations.

The SSN and the Weighted Averaged SSN (WASSN) methods are defined recursively for all $t \geq 0$ as follows:

$$\theta_{t+1} = \theta_t - \gamma_{t+1} \bar{S}_{t,w'}^{-1} \nabla_\theta f(\theta_t; \xi_{t+1}), \quad \theta_0 \in \mathbb{R}^d, \tag{9}$$

$$\theta_{t+1,w} = \theta_{t,w} + \frac{\ln(t+1)^w}{\sum_{i=0}^{t} \ln(i+1)^w} (\theta_t - \theta_{t,w}), \quad \theta_{0,w} = \theta_0 \text{ and } w \geq 0, \tag{10}$$

where $\overline{S}_{t,w'}$ is a recursive estimate of the Hessian. Specifically, we suppose that there is a natural recursive estimate of the Hessian $\bar{H}_t = N_t^{-1} H_t$ 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 definite, $\alpha_{i,j} \in \mathbb{R}_+$ and $\Phi_{i,j} \in \mathbb{R}^d$, which may depend on θ_{i-1} or $\theta_{i-1,w}$.

Remark 1 *We would like to emphasize that this type of Hessian estimation is applicable to many machine learning problems, 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 employed in tasks such as geometric median estimation (Godichon-Baggioni and Lu, 2024), nonlinear regression (Cénac et al., 2020), and optimal transport (Bercu et al., 2023). In Section 5.2, we provide some examples in detail.*

To develop the Hessian estimate $\bar{S}_{t,w'}$ for SSN and WASSN with the same computational complexity as first-order stochastic gradient methods, we first need to derive a computationally efficient estimate of H_t^{-1} using the Riccati/Sherman-Morrison formula (Duflo, 2013; Sherman and Morrison, 1950), applied n times. Updating of H_t^{-1} requires $\mathcal{O}(d^2 n)$ operations. Nevertheless, the total computation cost at time t would be of order $\mathcal{O}(d^2 N_t)$ operations, instead of $\mathcal{O}(dN_t)$ for first-order stochastic gradient methods.

In addition, to apply Theorem 1, it is necessary to control the eigenvalues of the Hessian estimates. Therefore, we propose a modified version of $\overline{H}_t$ given by $\bar{S}_{t,w'} = N_{t,Z}^{-1} S_{t,w'}$, where

$$S_{t,w'} = S_{0,w'} + \sum_{i=1}^{t} \ln(i+1)^{w'} \sum_{j=1}^{n} Z_{i,j} \left(\iota_{i,j} e_{i,j} e_{i,j}^{\top} + \alpha_{i,j} \Phi_{i,j} \Phi_{i,j}^{\top} \right), \quad w' \geq 0, \tag{11}$$

with

- $S_{0,w'}$ symmetric and positive definite,
- $N_{t,Z} = 1 + \sum_{i=1}^{t} \ln(i+1)^{w'} \sum_{j=1}^{n} Z_{i,j}$, where $w' \geq 0$ and $Z_{i,j}$ are i.i.d. with $Z_{i,j} \sim \mathcal{B}(p)$ for some $p \in (0,1]$,
- $\iota_{i,j} = c_\iota N_{i,j,Z}^{-\iota}$ where $\iota \in (0, \gamma - 1/2)$, $c_\iota \geq 0$ and $N_{t,k,Z} = 1 + \sum_{i=1}^{t-1} \sum_{j=1}^{n} Z_{i,j} + \sum_{j=1}^{k} Z_{t,j}$,
- $e_{i,j}$ is the ($N_{i,j,Z}$ modulo d)-th component of the canonical basis.

Observe that the term $\iota_{i,j}$ enables to control the smallest eigenvalue of $\bar{S}_{t,w'}$ while $\ln(t+1)^{w'}$ enables us to give more weights to the latest updates $\alpha_{t,j} \Phi_{t,j} \Phi_{t,j}^{\top}$, which are supposed to be better since they often depends on θ_t which should converge to θ^* almost surely.

The random variables $(Z_{i,j})$ enables us to adjust the computational cost. More precisely, observe that with the help of Riccati's formula (Duflo, 2013), one can update the inverse of $S_{t+1,w'}$ as follows: for all $j = \{1, , \ldots, n\}$

$$S^{-1}_{t+\frac{j}{2n},w'} = S^{-1}_{t+\frac{j-1}{2n},w'} - \frac{Z_{t+1,j} \iota_{t+1,j}}{1 + \iota_{t+1,j} e_{t+1,j}^T S^{-1}_{t+\frac{j-1}{2n},w'} e_{t+1,j}} S^{-1}_{t+\frac{j-1}{2n},w'} e_{t+1,j} e_{t+1,j}^T S^{-1}_{t+\frac{j-1}{2n},w'},$$

and for all $j = \{n+1, \ldots, 2n\}$

$$S^{-1}_{t+\frac{j}{2n},w'} = S^{-1}_{t+\frac{j-1}{2n},w'} - \frac{\alpha_{t,j-n} Z_{t+1,j-n}}{1 + \alpha_{t,j-n} \Phi_{t,j-n}^\top S^{-1}_{t+\frac{j-1}{2n},w'} \Phi_{t,j-n}} S^{-1}_{t+\frac{j-1}{2n},w'} \Phi_{t,j-n} \Phi_{t,j-n}^\top S^{-1}_{t+\frac{j-1}{2n},w'}.$$

Thus, as each update is only made if $Z_{t+1,j} = 1$ (or $Z_{t+1,j-n} = 1$), the update of $S^{-1}_{t+1,w}$ only costs, on average, $\mathcal{O}(pd^2n)$ operations, leading to a total number of operations of order (on average):

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

Hence, adjusting the value of p can help reduce the computational cost of updating the inverse of the Hessian. Indeed, one can obtain an average 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 an adaptive stochastic quasi-Newton method with only $\mathcal{O}(dN_t)$ operations.

Next, we can ensure that these adaptive stochastic quasi-Newton methods are asymptotically efficient. For this, let us consider the σ-algebra $\mathcal{F}'_{t-1} = \sigma(\xi_{1,1}, \ldots, \xi_{t-1,n}, Z_{t,1}, \ldots, Z_{t,n})$.

Theorem 5 *Suppose Assumptions 1 to 3 and 5 hold and that $c_\iota > 0$. In addition, assume that there exist positive constants $C_{\eta'}$ and $\eta' > 1$ such that for any $t \geq 1$ and $j \in \{1, \ldots, n\}$,*

$$\mathbb{E}[\|\alpha_{t,j} \Phi_{t,j} \Phi_{t,j}^\top\|^{\eta'} | \mathcal{F}'_{t-1}] \leq C_{\eta'}^{\eta'}.$$

Then, θ_t and $\theta_{t,w}$ converges almost surely to θ^*. Moreover, if $\overline{S}_{t,w'}$ converges almost surely to H, 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

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

The details of the proof are included in Appendix C.7. To suppose that $\overline{S}_{t,w'}$ converges to H may seem unrealistic, but it is often the case in practice when θ_t converges to θ^* (see, for example, Cénac et al. (2020, Corollary 3.1), Boyer and Godichon-Baggioni (2023, Theorem A.1 to A.3), or the proofs of Corollaries 1 to 3). Indeed, these proofs are often constructed as follows: (i) demonstrate that θ_t converges almost surely to θ^*, (ii) deduce the consistency of $\overline{S}_{t,w}$, (iii) infer the convergence rates of the estimators θ_t and $\theta_{t,w}$, and (iv) derive asymptotic normality.

At last, note that for Newton methods, the asymptotic efficiency of the estimates can be achieved without averaging by setting the learning rate $\gamma_t = 1/t$, e.g., see Leluc and Portier (2023); Bercu et al. (2020); Boyer and Godichon-Baggioni (2023). A streaming version of this approach, with potentially $\mathcal{O}(dN_t)$ operations, is detailed in Appendix B as well.

In the following sections, we provide three examples of applications; linear regression, logistic regression and the estimation of the median. Nevertheless, there are still many applications where our methodology works as highlighted in Remark 1.

5.2 Examples

5.2.1 Linear Regression

Consider the linear regression problem, where $\xi = (X, Y) \in \mathbb{R}^d \times \mathbb{R}$ such that $Y = X^\top \theta^* + \epsilon$, with $\theta^* \in \mathbb{R}^d$ and ϵ a centered random variable independent of X with variance σ^2. Then, θ^* minimizes the convex function $F : \mathbb{R}^d \to \mathbb{R}_+$ defined as $F(\theta) = \frac{1}{2}\mathbb{E}[(Y - X^\top\theta)^2]$. Let $\{(X_{t,1}, Y_{t,1}), \dots, (X_{t,n}, Y_{t,n})\}_{t\geq 1}$ be i.i.d. pairs of variables arriving sequentially in blocks. Then, the SSN and WASSN methods are defined by (9) and (10), with

$$\nabla_\theta f(\theta_t; X_{t+1}, Y_{t+1}) = -\frac{1}{n}\sum_{i=1}^{n}(Y_{t+1,i} - X_{t+1,i}^\top \theta_t)X_{t+1,i},$$

and $\overline{S}_{t,w'}$ defined by (11) with $c_\iota = 0$, $\alpha_{t,j} = 1$, and $\Phi_{t,j} = X_{t,j}$ for $j \in \{1, \dots, n\}$.

Corollary 1 *Suppose that X and ϵ have moments of order 4 and 2, respectively, and $H = \mathbb{E}[XX^\top]$ is positive definite. Then, for any $p \in (0, 1]$, we have*

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

and

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

The proof is given in Appendix C.9.

5.2.2 Logistic Regression

Consider the logistic regression problem, where $\xi = (X, Y) \in \mathbb{R}^d \times \{0, 1\}$ such that $Y|X \sim \text{Ber}(\pi(X^\top\theta^*))$, with $\theta^* \in \mathbb{R}^d$ and $\pi(x) = e^x/(1 + e^x)$. Here, the parameter θ^* minimizes the convex function $F : \mathbb{R}^d \to \mathbb{R}$ defined as $F(\theta) = \mathbb{E}[\log(1 + \exp(X^\top\theta)) - YX^\top\theta]$. Let $\{(X_{t,1}, Y_{t,1}), \dots, (X_{t,n}, Y_{t,n})\}_{t\geq 1}$ be i.i.d. pairs of variables arriving sequentially in blocks. Then, the SSN and WASSN methods are defined by (9) and (10), with

$$\nabla_\theta f(\theta_t; X_{t+1}, Y_{t+1}) = -\frac{1}{n}\sum_{i=1}^{n}(Y_{t+1,i} - \pi(X_{t+1,i}^\top \theta_t))X_{t+1,i},$$

and $\overline{S}_{t,w'}$ is defined by (11) with $\alpha_{t,j} = \pi(X_{t,j}^\top\theta_{t-1})[1 - \pi(X_{t,j}^\top\theta_{t-1})]$ and $\Phi_{t,j} = X_{t,j}$ for $j \in \{1, \dots, n\}$.

Corollary 2 *Suppose that X have moment of order 4 and $H = \mathbb{E}[\pi(X^\top\theta^*)[1-\pi(X^\top\theta^*)]XX^\top]$ is positive definite. Then, if $c_\iota > 0$, we have for any $p \in (0, 1]$ that*

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

and

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

The proof is given in Appendix C.10. Observe that the results in Corollary 2 also holds when using $\theta_{t-1,w}$ in $\alpha_{t,j}$ instead of θ_{t-1}. Note that assumption of the positivity of H is not unrealistic, as it is satisfied, for instance, when X is elliptic (Gadat and Panloup, 2023, Proposition 3).

5.2.3 Geometric Median

Consider a random variable $\xi = X \in \mathbb{R}^d$. The geometric median of X is defined by $\theta^* = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}[\|X - \theta\| - \|X\|]$. Let $\{X_{t,1}, \ldots, X_{t,n}\}_{t \geq 1}$ be i.i.d variables arriving sequentially in blocks. Then, the SSN and WASSN methods are defined by (9) and (10), with

$$\nabla_\theta f(\theta_t; X_{t+1}) = -\frac{1}{n} \sum_{i=1}^{n} \frac{X_{t+1,i} - \theta_t}{\|X_{t+1,i} - \theta_t\|},$$

and $\overline{S}_{t,w'}$ is defined by (11) with

$$\alpha_{t,j} = \frac{\|X_{t,j} - \theta_{t-1}\|}{v_t^2} \qquad \text{and} \qquad \Phi_{t,j} = \frac{X_{t,j} + v_t U_{t,j} - \theta_{t-1}}{\|X_{t,j} + v_t U_{t,j} - \theta_{t-1}\|} - \frac{X_{t,j} - \theta_{t-1}}{\|X_{t,j} - \theta_{t-1}\|}$$

for $j \in \{1, \ldots, n\}$, where $v_t = \frac{1}{t \log(t+1)}$ and $U_{t,j}$ are zero-mean i.i.d vectors, with moments of any orders and simulated independently from $X_{i,j'}$ and with covariance equal to the identity (Godichon-Baggioni and Lu, 2024). Typically, $U_{t,j}$ are standard Gaussian vectors.

Corollary 3 *Suppose Godichon-Baggioni and Lu (2024, Assumption 1 and 2) hold and let*

$$H = \mathbb{E}\left[\frac{1}{\|X - \theta^*\|}\left(I_d - \frac{(X - \theta^*)(X - \theta^*)^T}{\|X - \theta^*\|^2}\right)\right], \quad \textit{and} \quad \Sigma = \mathbb{E}\left[\frac{(X - \theta^*)(X - \theta^*)^T}{\|X - \theta^*\|^2}\right].$$

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

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

The proof as well as the assumptions are given in Appendix C.11. Similarly to Corollary 2, the results in Corollary 3 remains true when using $\theta_{t-1,w}$ in $\alpha_{t,j}$ and $\Phi_{t,j}$ instead of θ_{t-1}.

6. Experiments

In this section, we empirically evaluate our adaptive stochastic optimization methods, focusing on three fundamental problems: linear regression (Section 5.2.1), logistic regression (Section 5.2.2), and the estimation of the geometric median (Section 5.2.3). First, in Section 6.1, we explore our methods under complex covariance structures and bad initializations using synthetic data. In particular, in Section 6.1.1 we compare the computational costs of our adaptive methods with both first- and second-order methods, demonstrating that the computational efficiency of our adaptive methods aligns closely with that of first-order methods. Next, in Section 6.2, we evaluate and compare the methods on classification tasks using UCI datasets.

6.1 Synthetic Experiments

We generate samples of d-dimensional Gaussian random vectors $X \sim \mathcal{N}(0, \Sigma)$. We consider two structures for the covariance matrix Σ: (1) simple covariance, where Σ is the identity matrix I_d, implying no correlations, and (2) complex covariance, where $\Sigma_{ii} = i$ and $\Sigma_{ij} = 0.9^{|i-j|}$ for $i, j = 1, \ldots, d$. These choices of covariance structures allow us to explore the adaptability of our methods under diverse conditions, including strong correlations between the coordinates of X.

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

We imposed several restrictions on hyperparameters to theoretically derive the convergence rates of the algorithms. However, in our experiments, we set $\gamma_t = C_\gamma(t + t_0)^{-\gamma}$ with $C_\gamma = 1$ for the linear and logistic regression, and $C_\gamma = \sqrt{d}$ for the geometric median, $t_0 = d$, and $\gamma = 3/4$. The weight parameters for both the estimates and Hessian approximations are set to 2, i.e., $w = w' = 2$. The initial Hessian for SSN and WASSN are set to λI_d with $\lambda = 0.001$ for linear regression and logistic regression, and $\lambda = 0.05$ for the geometric median. While these settings serve as a proof-of-concept, one could fine-tune these parameters further; for instance, higher values of w (and w') would enhance adaptability. In the figures, the SSN algorithm corresponds to the case where $C_\gamma = 1$ and $\gamma = 1$ without averaging (Appendix B.2).

We set the mini-batch size n equal to the dimension d, which allows our adaptive quasi-Newton's method to maintain the same computational cost of first-order methods while incorporating second-order information. Notable, our results clearly demonstrate that this approach significantly improves performance compared to AdaGrad and SGD. Specifically, when dealing with highly correlated data, the AdaGrad algorithm's adaptive step size becomes less effective, whereas quasi-Newton's methods excel. Particularly, in scenarios involving less-than-ideal initializations (as depicted on the right side of the figures), both quasi-Newton's methods show outstanding performance.

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

6.1.1 Computational Efficiency

In our experiments, we investigate various optimization methods, including SGD, AdaGrad, our streaming AdaGrad detailed in Appendix B.3, along with their iterated weighted Polyak-Ruppert averages. Additionally, we explore previous quasi-Newton algorithm (SN) and (WASN) (Boyer and Godichon-Baggioni, 2023), our streaming stochastic quasi-Newton (SSN) and our weighted Polyak-Ruppert averaged streaming stochastic quasi-Newton (WASSN) from Section 5.1. 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 whole Hessian at each step.

In Figure 1, we present the running times of the various algorithms considered. Here, we can see that AdaGrad, streaming AdaGrad, SSN ($p = 1/d$), and WASSN ($p = 1/d$) all have similar running times to SGD. This indicates that the more sophisticated algorithms incorporating second-order information and adaptive step sizes do not incur significant additional computational costs compared to the simpler SGD. Furthermore, the iterative weighted average does not add any significant computational overhead, maintaining efficiency while enhancing performance.

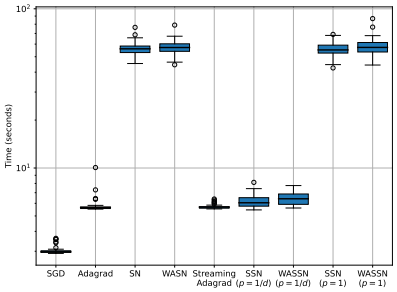

Figure 1: Running times (in seconds) for the algorithms considered. These running times are based on processing one million samples (i.e., $N_t = 1,000,000$), with a dimension of $d = 100$, and a mini-batch size of $n = 100$.

6.1.2 Linear Regression

In the context of linear 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 follow the approach Boyer and Godichon-Baggioni (2023) and set $\theta^* = (-d/2, -(d-1)/2, \ldots, (d-1)/2, d/2)^\top$.

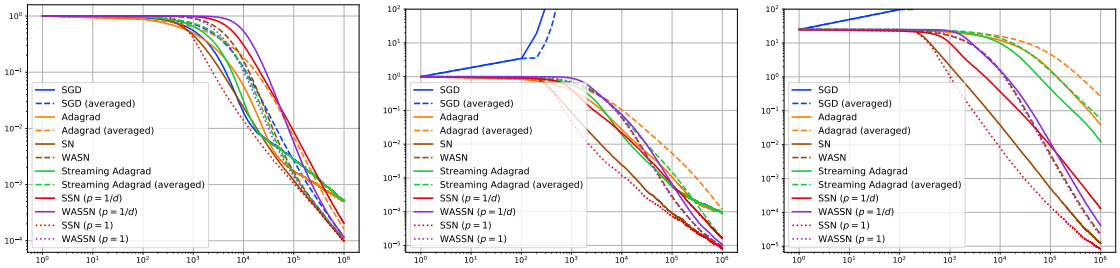

Figure 2: Linear regression. Each curve shows $\|\theta_t - \theta^*\|$ averaged over 50 epochs with different initial points and one million samples. Initial points θ_0 are generated as $\theta_0 = \theta^* + rU$, where U is a random variable on the unit sphere of $\mathbb{R}^d$ and $r \in \{1, 5\}$. Left: simple covariance with $r = 1$; middle: complex covariance with $r = 1$; right: complex covariance with $r = 5$.

In Figure 2, we observe that all algorithms perform well in the well-posed setting with no correlation and good initialization. However, introducing correlations causes SGD to diverge, while both AdaGrad and quasi-Newton's methods still manage to converge. This can be attributed to their innate capability to handle 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 and less-than-ideal initializations (as depicted on the right side of the figure). In contrast, our adaptive Quasi-Newton's methods demonstrate outstanding performance in all cases. Notably, the WASSN, with only $\mathcal{O}(dN_t)$ computational costs, performs close to the full Hessian versions (streaming, SN and WASN).

6.1.3 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. Inspired by Boyer and Godichon-Baggioni (2023), we choose $\theta^* \in \mathbb{R}^d$ with all components equal to 0.1. Unlike the linear regression setting, logistic regression exhibits intrinsic non-linearity, which makes the impact of the covariance structures less clear.

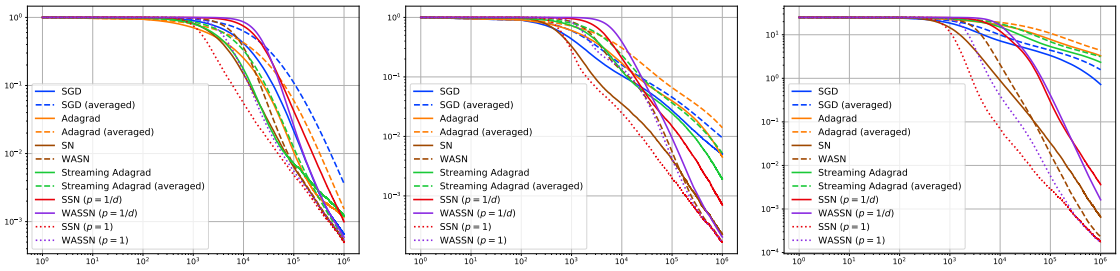

Figure 3: Logistic regression. Each curve shows $\|\theta_t - \theta^*\|$ averaged over 50 epochs with different initial points and one million samples. Initial points θ_0 are generated as $\theta_0 = \theta^* + rU$, where U is a random variable on the unit sphere of $\mathbb{R}^d$ and $r \in \{1, 5\}$. Left: simple covariance with $r = 1$; middle: complex covariance with $r = 1$; right: complex covariance with $r = 5$.

In Figure 3, our adaptive quasi-Newton methods consistently perform well across all configurations. The streaming AdaGrad algorithm also performs well as long as the initial point is not too far from the solution. In scenarios involving less-than-ideal initializations, as depicted on the right side of the figure, the best performance is achieved by the SSN/WASSN. This exceptional asymptotic behavior is enabled by the incorporation of weighted estimates, assigning greater significance to the most recent ones, distinguishing it from the usual Polyak-Ruppert averaging quasi-Newton's method, as elaborated in Cénac et al. (2020).

6.1.4 Geometric Median

In the context of geometric median estimation, we aim to evaluate the performance of our adaptive stochastic optimization methods for estimating the geometric median of a distribution with median $\theta^* = (-d/2, -(d-1)/2, \ldots, (d-1)/2, d/2)^\top$.

In Figure 4, we observe our adaptive quasi-Newton methods, particularly the averaged versions, demonstrate superior performance, especially in scenarios with suboptimal initializations. Similar to the previous experiments, our adaptive Quasi-Newton methods show remarkable robustness and efficiency. The SSN method, with its weighted estimates, WASSN, stands out by maintaining high performance even when starting points are far from the solution. This is particularly evident under complex covariance structures and larger initialization offsets, as shown on the right side of the figure. These results highlight the advantages of incorporating second-order information and adaptive weighting in optimization algorithms.

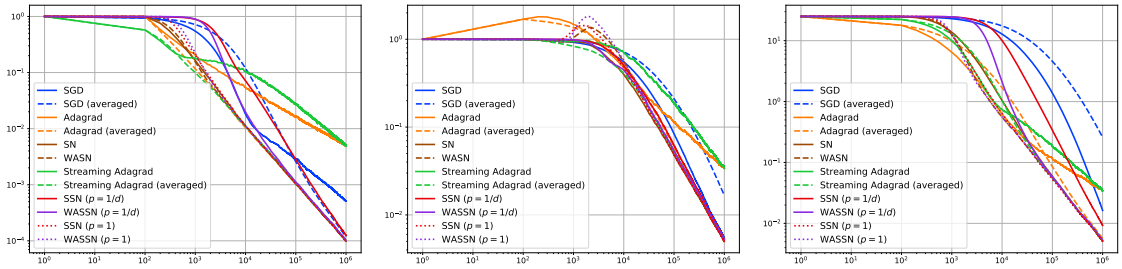

Figure 4: Geometric median. Each curve shows $\|\theta_t - \theta^*\|$ averaged over 50 epochs with different initial points and one million samples. Initial points θ_0 are generated as $\theta_0 = \theta^* + rU$, where U is a random variable on the unit sphere of $\mathbb{R}^d$ and $r \in \{1, 5\}$. Left: simple covariance with $r = 1$; middle: complex covariance with $r = 1$; right: complex covariance with $r = 5$.

6.2 Real-World Experiments

In this section, we evaluate the performance of our methods on logistic regression for binary classification tasks using three UCI datasets (Kelly et al.): Adult, Higgs, and Statlog. For consistency, we used the same hyperparameters as outlined in Section 6.1.

Figure 5 presents the averaged loss curves for each dataset under different initialization conditions. The columns, from left to right, correspond to the Adult, Higgs, and Statlog datasets. The top row represents narrow initializations, while the bottom row illustrates wider initializations, simulating more challenging starting points.

Across all datasets and initialization configurations, our methods, SSN and WASSN, consistently demonstrate strong performance, comparable to full Hessian Newton methods. In particular, the weighted averaged streaming stochastic quasi-Newton method (WASSN) closely matches the performance of the full Hessian method ($p = 1$), even under more difficult initialization conditions. Our streaming AdaGrad method also exhibits strong convergence behavior. Under wider initializations, while its performance falls short of the quasi-Newton methods, it remains superior to both standard AdaGrad and SGD.

Conclusion 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 Quasi-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.

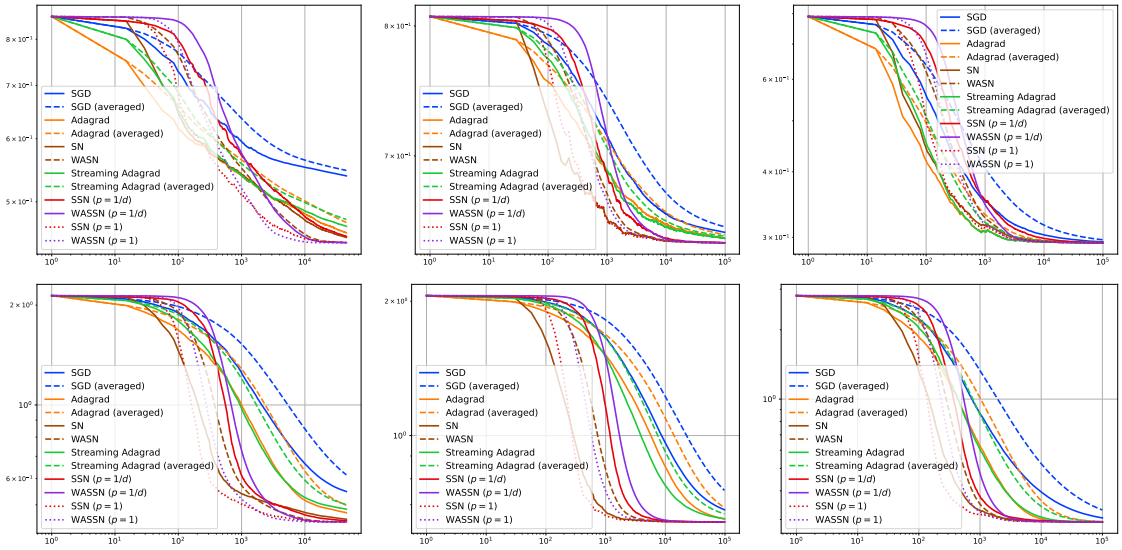

Figure 5: Logistic regression for classification. Each curve shows the loss averaged over ten epochs with different initial points. Initial points θ_0 are generated as $\theta_0 = rU$, where U is a random variable on the unit sphere of $\mathbb{R}^d$ and $r \in \{1, 5\}$. The first row represents $r = 1$, and the bottom row corresponds to $r = 5$. The columns, from left to right, correspond to the Adult, Higgs, and Statlog datasets.

One significant contribution is the inversion-free adaptive Quasi-Newton's method in Section 5.1, 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 Quasi-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 streaming AdaGrad algorithm in non-convex scenarios will contribute to a more comprehensive understanding of their applicability across diverse optimization landscapes such as Neural Networks. (b) Dimensionality effects. Although it is obvious that considering the case where the dimension is larger than the sample size in our streaming setting is unrealistic, an other important perspective would be to quantify theoretically the impact of the dimension on the quality of the estimates. (c) 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.

Acknowledgments

N. Werge is supported by the Novo Nordisk Foundation (grant no. NNF21OC0070621).

Appendix

Appendix A contains the theoretical analysis of Section 4 with increasing mini-batches. Appendix B presents additional applications of our methodology; for example, AdaGrad and stochastic quasi-Newton methods under different learning rates, both with and without weighted averaging. Appendix C contains the proofs of the theoretical analysis in Section 4 and Appendix A

Appendix A. Theoretical Analysis with Increasing Mini-Batches

In this appendix, we examine our adaptive methods in the context of increasing mini-batches and provide the corresponding translations of the various theorems. The motivation for considering increasing mini-batches comes from recent work by Godichon-Baggioni et al. (2023a), which demonstrated that this approach can accelerate convergence and break long- and short-term dependence structures.

Following, Godichon-Baggioni et al. (2023a,b), we consider mini-batches of the form $n_t = \lfloor C_\rho t^\rho \rfloor$ with $C_\rho \in \mathbb{N}$. Moreover, we suppose that the learning rate $\gamma_t = C_\gamma n_t^\beta t^{-\gamma}$, 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 put more weight 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

This section presents the results for increasing mini-batches corresponding to Section 4.1. When considering increasing mini-batches, our adaptive stochastic optimization methods are defined as in (2), namely 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 $\nabla_\theta f(\theta_t;\xi_{t+1}) = n_{t+1}^{-1}\sum_{i=1}^{n_{t+1}}\nabla_\theta f(\theta_t;\xi_{t+1,i})$. In this case, the conditions in (5) should be modified to

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

Under these conditions outlined in (A.1), we have the strong consistency of the estimates derived from (2), similar to the results in Theorem 1.

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

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

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

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

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 Section 4.2, we consider the weighted averaged version of our adaptive stochastic optimization methods. The weighted estimates of (3) (with time-varying mini-batches) are defined as follows:

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

which can be written recursively as $\theta_{t+1,w} = \theta_{t,w} + \frac{n_{t+1}\ln(t+1)^w}{\sum_{i=0}^{t} n_{i+1}\ln(i+1)^w}(\theta_t - \theta_{t,w})$.

Similarly 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.2)*. 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*

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

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.2). *In addition, assume there exists a positive constant* $v' > 1/2$ *such that*

$$\frac{1}{\sum_{i=0}^{t-1} n_{i+1}\ln(i+1)^w}\sum_{i=0}^{t-1} n_{i+1}\ln(i+1)^{w+1/2+\delta}\|A_{i+1}^{-1} - A_i^{-1}\|_{\mathrm{op}}(i+1)^{\frac{\gamma-\rho(\beta+1)}{2}} = \mathcal{O}\left(\frac{1}{t^{(1+\rho)v'}}\right) \text{ a. s.}, \tag{A.3}$$

for some $\delta > 0$. *Then,*

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

Appendix B. Applications to Newton's method and AdaGrad

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 Appendix B.2 and Section 5.1.

B.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:

$$\theta_{t+1} = \theta_t - \frac{1}{t+1}\bar{H}_t^{-1}\nabla_\theta f(\theta_t; \xi_{t+1}), \quad \theta_0 \in \mathbb{R}^d, \tag{B.4}$$

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.

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

Theorem B.5 *Suppose Assumptions 1, 2, 3, and 5 hold, along with inequality* (6). *Then,* θ_t *converges almost surely to* θ^*. *In addition, assume* $\bar{H}_t^{-1}$ *converges almost surely to* $\nabla^2_\theta F(\theta^*)^{-1}$. *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* $\|\bar{H}_t^{-1} - \nabla^2_\theta F(\theta^*)^{-1}\|_{\mathrm{op}} = \mathcal{O}(1/t^\nu)$ a. s.. *Then*

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

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

Remark B.1 *Observe that assuming the convergence of* $\overline{H}_t^{-1}$ *could be considered unrealistic. However, at this point, the strong consistency of* θ_t *is already established. Then, the consistency of* $\overline{H}_t$ *is verified, for example, if* $\mathbb{E}\left[\alpha_{i,j}\Phi_{i,j}\Phi_{i,j}^T|\mathcal{F}'_{i,j-1}\right]$ *converges to* H^{-1} *when* i *goes to infinity, and if for all* i, j

$$\mathbb{E}\left[\alpha_{i,j}^2\left\|\Phi_{i,j}\right\|^4|\mathcal{F}'_{i,j-1}\right]\mathbb{1}_{\{\|\theta_i-\theta^*\|\leq M_1\}}\leq M_2,$$

for some $M_1, M_2 > 0$. *This is satisfied, for instance (under weak conditions), for the case of linear and logistic regression and the estimation of the geometric median; see the proofs of Corollaries 1 to 3.*

B.2 Direct Streaming Stochastic Newton's methods with possibly $\mathcal{O}(dN_t)$ operations

The direct stochastic Newton's method presented in Appendix B.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 and following the idea given in Section 5.1. 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, $\alpha_{i,j}\in\mathbb{R}_+$ and $\Phi_{i,j}\in\mathbb{R}^d$. Observe that these quantities may depend on θ_{i-1} or $\theta_{i-1,w}$. We now introduce the direct streaming stochastic Newton's method using weighted Hessian estimates:

$$\theta_{t+1} = \theta_t - \frac{1}{t+1}\bar{H}_{t,w'}^{-1}\nabla_\theta f(\theta_t;\xi_{t+1}),\quad \theta_0\in\mathbb{R}^d, \tag{B.5}$$

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

$$H_{t,w'} = H_{0,w'} + \sum_{i=1}^{t}\ln(i+1)^{w'}\sum_{j=1}^{n}Z_{i,j}\left(\iota_{i,j}e_{i,j}e_{i,j}^\top + \alpha_{i,j}\Phi_{i,j}\Phi_{i,j}^\top\right), \tag{B.6}$$

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.

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: for all $j=\{1,,\ldots,n\}$

$$H_{t+\frac{j}{2n},w'}^{-1} = H_{t+\frac{j-1}{2n},w'}^{-1} - \frac{Z_{t+1,j}\iota_{t+1,j}}{1+\iota_{t+1,j}e_{t+1,j}^T\tilde{H}_{t+\frac{j-1}{2n},w'}^{-1}e_{t+1,j}}H_{t+\frac{j-1}{2n},w'}^{-1}e_{t+1,j}e_{t+1,j}^TH_{t+\frac{j-1}{2n},w'}^{-1}$$

and for all $j=\{n+1,\ldots,2n\}$

$$H_{t+\frac{j}{2n},w'}^{-1} = H_{t+\frac{j-1}{2n},w'}^{-1} - \frac{\alpha_{t,j-n}Z_{t+1,j-n}}{1+\alpha_{t,j-n}\Phi_{t,j-n}^\top H_{t+\frac{j-1}{2n},w'}^{-1}\Phi_{t,j-n}}H_{t+\frac{j-1}{2n},w'}^{-1}\Phi_{t,j-n}\Phi_{t,j-n}^\top H_{t+\frac{j-1}{2n},w'}^{-1}.$$

We now ensure that this method is still asymptotically efficient. In this aim, let us the following σ algebra: $\mathcal{F}'_{t-1} = \sigma\left(\xi_{1,1}, \ldots, \xi_{t-1,n}, Z_{t,1}, \ldots, Z_{t,n}\right)$.

Theorem B.6 *Suppose Assumptions 1 to 3 and 5 hold and $c_\iota > 0$. In addition, assume that there exist positive constants $C_{\eta'}$ and $\eta' \geq 2$ such that for any $t \geq 1$ and $j \in \{1, \ldots, n\}$,*

$$\mathbb{E}[\|\alpha_{t,j}\Phi_{t,j}\Phi_{t,j}^\top\|^{\eta'}|\mathcal{F}'_{t-1}] \leq C_{\eta'}^{\eta'}.$$

Then, θ_t and $\theta_{t,w}$ converges almost surely to θ^*. Suppose also that $\mathbb{E}\left[\alpha_{t,j}\Phi_{t,j}\Phi_{t,j}|\mathcal{F}'_{t-1}\right]$ converges almost surely to H, then

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

Finally, assuming also that $\left\|\mathbb{E}\left[\alpha_{t,j}\Phi_{t,j}\Phi_{t,j}|\mathcal{F}'_{t-1}\right] - H\right\| = O(t^{-v})$ for some $v > 0$, 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.

B.2.1 Streaming Stochastic Newton's Methods with possibly $\mathcal{O}(dN_t)$ operations

Expanding the mini-batch scenario from Appendix B.2 leads to the formulation of the streaming variant of stochastic Newton's method, as defined by:

$$\theta_{t+1} = \theta_t - \frac{1}{N_{t+1}}\bar{H}_{t,w'}^{-1}\sum_{i=1}^{n_{t+1}}\nabla_\theta f(\theta_t; \xi_{t+1,i}), \tag{B.7}$$

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 $(N_{Z,t'-1,i}$ modulo d $+1)$-th componant of the canonical basis.

As in Theorem B.6, we can establish the rate of convergence and the asymptotic normality of (B.7). In this aim, let us the following σ algebra: $\mathcal{F}'_{t-1} = \sigma\left(\xi_{1,1}, \ldots, \xi_{t-1,n_{t-1}}, Z_{t,1}, \ldots, Z_{t,n_t}\right)$.

Theorem B.7 *Suppose Assumptions 1 to 3 and 5 hold.In addition, assume that there exist positive constants $C_{\eta'}$ and $\eta' \geq 2$ such that for any $t \geq 1$ and $j \in \{1, \ldots, n\}$,*

$$\mathbb{E}[\|\alpha_{t,j}\Phi_{t,j}\Phi_{t,j}^\top\|^{\eta'}|\mathcal{F}_{t-1}] \leq C_{\eta'}^{\eta'}.$$

Then, θ_t and $\theta_{t,w}$ converge almost surely to θ^*. Suppose also that $\mathbb{E}\left[\alpha_{t,j}\Phi_{t,j}\Phi_{t,j}|\mathcal{F}'_{t-1}\right]$ converges almost surely to H, then

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

Finally, assuming also that $\left\|\mathbb{E}\left[\alpha_{t,j}\Phi_{t,j}\Phi_{t,j}|\mathcal{F}'_{t-1}\right] - H\right\| = O(t^{-v})$ for some $v > 0$, 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 $\eta' \geq 2$. Then,

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

B.2.2 Weighted Averaged Version of Streaming Stochastic Newton's Methods with possibly $\mathcal{O}(dN_t)$ operations

The weighted averaged version outlined in Section 5.1 can similarly be adapted to the increasing mini-batch case. The weighted averaged streaming stochastic Newton's method is defined as

$$\begin{aligned}
\theta_{t+1} &= \theta_t - \gamma_{t+1}\bar{S}^{-1}_{t,w'}\frac{1}{n_{t+1}}\sum_{i=1}^{n_{t+1}}\nabla_\theta f(\theta_t;\xi_{t+1,i}),\\
\theta_{t+1,w} &= \theta_{t,w} + \frac{n_{t+1}\ln(t+1)^w}{\sum_{i=0}^{t} n_{i+1}\ln(i+1)^w}(\theta_t - \theta_{t,w}),
\end{aligned} \tag{B.8}$$

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

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

with S_0 symmetric and positive, $\iota_{t',i} = C_\iota N^{-\iota}_{Z,t',i}$ 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 5, we have the following asymptotic optimality:

Theorem B.8 *Suppose Assumptions 1 to 3 and 5 hold, along the condition* (A.2)*. In addition, assume that there exist positive constants $C_{\eta'}$ and $\eta' \geq 2$ such that for any $t \geq 1$ and $j \in \{1, \ldots, n\}$,*

$$\mathbb{E}[\|\alpha_{t,j}\Phi_{t,j}\Phi^\top_{t,j}\|^{\eta'}|\mathcal{F}'_{t-1}] \leq C^{\eta'}_{\eta'}.$$

Then θ_t and $\theta_{t,w}$ converge almost surely to θ^*. Suppose also that $\overline{S}_{t,w'}$ converges almost surely to H, 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,

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

B.3 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.

B.3.1 Streaming AdaGrad with constant mini-batches

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

$$\theta_{t+1} = \theta_t - \gamma_{t+1} G_t \nabla_\theta f(\theta_t; \xi_{t+1}), \quad \theta_0 \in \mathbb{R}^d, \tag{B.9}$$

$$\theta_{t+1,w} = \theta_{t,w} + \frac{\ln(t+1)^w}{\sum_{i=0}^{t} \ln(i+1)^w}(\theta_t - \theta_{t,w}), \tag{B.10}$$

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 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 (5) are satisfied.

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

Theorem B.9 *Suppose Assumptions 1 to 3 and 5 hold, along with inequality* (6)*. 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

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

B.3.2 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.1). To simplify it, one can take $\beta' < \gamma - \beta\rho - 1/2$. Then, Theorem B.9 can be written as follows:

Theorem B.10 *Suppose Assumptions 1 to 3 and 5 hold, along with the conditions in* (A.2). *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

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

Appendix C. 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.

C.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.1). 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.1);

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

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.

C.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,

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

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

$$Q(\theta_{t+1} - \theta^*) = Q(\theta_t - \theta^*) - \gamma_{t+1}QA_tH(\theta_t - \theta^*) + \gamma_{t+1}QA_t\Xi_{t+1} - \gamma_{t+1}QA_t\delta_t,$$

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

$$\begin{aligned}Q\left(\theta_{t+1} - \theta^*\right) =& Q\left(\theta_t - \theta^*\right) - \gamma_{t+1}QAH\left(\theta_t - \theta^*\right) - \gamma_{t+1}Q\left(A_t - A\right)H\left(\theta_t - \theta^*\right)\\ &+ \gamma_{t+1}QA_t\Xi_{t+1} - \gamma_{t+1}QA_t\delta_t\\ =& \left(I_d - \gamma_{t+1}D\right)Q\left(\theta_t - \theta^*\right) - \gamma_{t+1}Q\left(A_t - A\right)H\left(\theta_t - \theta^*\right)\\ &+ \gamma_{t+1}QA_t\Xi_{t+1} - \gamma_{t+1}QA_t\delta_t. \qquad \text{(C.11)}\end{aligned}$$

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 (C.11) that

$$\begin{aligned}Q\left(\theta_T - \theta^*\right) =& \overbrace{\beta_{T,0}Q\left(\theta_0 - \theta^*\right)}^{R_{1,T}:=} \overbrace{- \sum_{t=0}^{T-1}\beta_{T,t+1}\gamma_{t+1}Q\left(A_t - A\right)H\left(\theta_t - \theta^*\right) - \sum_{t=0}^{T-1}\beta_{T,t+1}\gamma_{t+1}QA_t\delta_t}^{R_{2,T}:=}\\ &+ \underbrace{\sum_{t=0}^{T-1}\beta_{T,t+1}\gamma_{t+1}QA_t\Xi_{t+1}}_{M_T:=}, \qquad \text{(C.12)}\end{aligned}$$

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 (C.12) 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

$$\left\|\beta_{T,0}\right\|_{\text{op}} = \begin{cases} \mathcal{O}\left(\exp\left(-\lambda_{\min}(D)\frac{c_\gamma n^{\gamma-1}}{1-\gamma}N_T^{1-\gamma}\right)\right) & \text{if } n_t = n,\\ \mathcal{O}\left(\exp\left(-\lambda_{\min}(D)\frac{c_\gamma C_\rho^{\frac{\beta-1+\gamma}{1+\rho}}}{1+\beta\rho-\gamma}N_T^{\frac{1+\beta\rho-\gamma}{1+\rho}}\right)\right) & \text{if } n_t = \lfloor C_\rho t^\rho \rfloor\,. \end{cases} \tag{C.13}$$

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

A first rate of convergence of M_T First, remark that

$$\mathbb{E}\left[\left\|\Xi_{t+1}\right\|^2 |\mathcal{F}_t\right] \leq \mathbb{E}\left[\left\|f'_{t+1}\right\|^2 |\mathcal{F}_t\right] \leq \frac{1}{n_{t+1}}C\left(1+F\left(\theta_t\right)-F(\theta^*)\right)+\left\|\nabla F\left(\theta_t\right)\right\|^2. \tag{C.14}$$

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

$$\left\|M_T\right\|^2 = \mathcal{O}\left(\ln(T)T^{\beta\rho-\gamma}\right) \text{ a. s.} \tag{C.15}$$

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 $\|\delta_t\| = o\left(\|\theta_t - \theta^*\|\right)$ a.s and $\left\|A_t - A\right\|_{\text{op}}$ 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 $t \geq t_0$,

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

Then, with the help of (C.13) and (C.15), there exists a positive random variable C_1, such that

$$\left\|R_{2,t+1}\right\|^2 \leq \left(1-\gamma_{t+1}\right)\left\|R_{2,t}\right\| + 2\gamma_{t+1}r_{t+1}\left(\left\|R_{2,t}\right\| + C_1\ln(t+1)(t+1)^{\frac{\beta\rho-\gamma}{2}}\right), \tag{C.16}$$

so that

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

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 (C.14), 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

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

C.3 Proof of Theorems 3 and A.3

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

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

which can be written as

$$\theta_t - \theta^* = H^{-1} A^{-1} \frac{u_t - u_{t+1}}{\gamma_{t+1}} + H^{-1} A^{-1} A_t \Xi_{t+1} - H^{-1} A^{-1} A_t \delta_t - H^{-1} A^{-1} (A_t - A) H (\theta_t - \theta^*), \tag{C.17}$$

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

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

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

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

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

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

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,

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

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

$$\begin{aligned}
&\left|\frac{n_{t+1}\ln(t+1)^w}{\gamma_{t+1}} - \frac{n_t\ln(t)^w}{\gamma_t}\right| \\
&\quad \le C_\rho \max\left\{\rho(1-\beta)+\gamma, w\right\} \max\left\{t^{\rho(1-\beta)+\gamma-1}, (t+1)^{\rho(1-\beta)+\gamma-1}\right\}\ln(t+1)^w,
\end{aligned}$$

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

$$\frac{1}{s_T}\left\|\sum_{t=0}^{T-1}\left(\frac{n_{t+1}\ln(t+1)^w}{\gamma_{t+1}}-\frac{n_t\ln(t)^w}{\gamma_t}\right)(\theta_t-\theta^*)\right\|=\mathcal{O}\left(\sqrt{\ln\left(N_T\right)}N_T^{\frac{-2+\gamma-\rho(1+\beta)}{2(1+\rho)}}\right)\ \text{a. s.},\tag{C.18}$$

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

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

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

which can be written as

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

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

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

which can be written as

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

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

C.4 Proof of Theorems 4 and A.4

First, remark that one can rewrite decomposition (C.17) 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

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

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

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

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 (C.18)), 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

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

At last, one can conclude the proof with the help of equality (A.3).

C.5 Proof of Theorem B.5

Observe that the convergence of θ_T is obtained with the same calculus as in the proof of Theorem 1. We now give the proof here for a non-decreasing streaming batch size. Remark that decomposition (C.11) 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

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

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

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

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,

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

Convergence of the rest terms Since $\overline{H}_t^{-1}$ converges to H^{-1} and $\|\delta_t\| = o\left(\|\theta_t - \theta^*\|\right)$ 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

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

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

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

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

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

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

C.6 Proof of Theorems B.6 and B.7

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[3]

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

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

$$\begin{aligned}\frac{1}{N_{Z,T}}\sum_{t=1}^{T}\ln(t+1)^{w'}\sum_{i=1}^{n_t}Z_{t,i}\alpha_{t,i}\Phi_{t,i}\Phi_{t,i}^\top =&\frac{1}{N_{Z,T}}\sum_{t=1}^{T}\ln(t+1)^{w'}\mathbb{E}\left[\alpha_{t,i}\Phi_{t,i}\Phi_{t,i}^\top|\mathcal{F}'_{t-1}\right]\sum_{i=1}^{n_t}Z_{t,i}\\&+\frac{1}{N_{Z,T}}\sum_{t=1}^{T}\ln(t+1)^{w'}\Xi_{Z,t},\end{aligned}$$

where $\Xi_{Z,t}:=\sum_{i=1}^{n_t}Z_{t,i}\alpha_{t,i}\Phi_{t,i}\Phi_{t,i}^\top-\sum_{i=1}^{n_t}Z_{t,i}\mathbb{E}\left[\alpha_{t,i}\Phi_{t,i}\Phi_{t,i}^\top|\mathcal{F}'_{t-1}\right]$ is a sequence of martingale differences for the filtration $\mathcal{F}'_{t-1}=\sigma\left(\xi_{1,1},\ldots,\xi_{t-1,n_{t-1}},Z_{t,1},\ldots,Z_{t,n_t}\right)$. 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

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

In addition

$$\left\|\frac{1}{N_{Z,T}}\sum_{t=1}^{T}\ln(t+1)^{w'}\sum_{i=1}^{n_t}Z_{t,i}\right\|_{\text{op}}\mathbb{E}\left[\alpha_{t,i}\Phi_{t,i}\Phi_{t,i}^\top|\mathcal{F}'_{t-1}\right]\leq C_{\eta'}^{1/\eta'}.$$

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

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

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

and the conditions in (A.1) are satisfied as soon as $i < \frac{1-\rho}{2(1+\rho)}$. Then, according to Theorem B.5, θ_T converges almost surely to θ^*. Supposing that $\mathbb{E}\left[\alpha_{t,i}\Phi_{t,i}\Phi_{t,i}^\top|\mathcal{F}'_{t-1}\right]$ converges almost surely to H, one has

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

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

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

Since $\left\|\mathbb{E}\left[\alpha_{t,i}\Phi_{t,i}\Phi_{t,i}^\top|\mathcal{F}'_{t-1}\right]-H\right\|=\mathcal{O}(t^v)$ a.s., taking $v<1$;

$$\left\|\frac{1}{N_{Z,T}}\sum_{t=1}^{T}\ln(t+1)^{w'}\sum_{i=1}^{n_t}Z_{t,i}\mathbb{E}\left[\alpha_{t,i}\Phi_{t,i}\Phi_{t,i}^\top|\mathcal{F}'_{t-1}\right]-H\right\|_{\text{op}}=\mathcal{O}\left(\frac{\sqrt{\ln(N_T)}}{T^v}\right)\text{ a. s.}$$

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

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

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

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

In addition

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

Then, there is $\nu>0$ such that

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

Then, with the help of Theorem B.5, one has

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

C.7 Proof of Theorems 5 and B.8

As in the proof of Theorems B.6 and B.7, one can easily check that the conditions in (A.1) 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 B.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,

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

such that

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

which concludes the proof.

C.8 Proof of Theorems B.9 and B.10

First, since the conditions in (5) (or in (A.1)) 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.2) coupled with Duflo (2013, Proposition 1.III.19), we have

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

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

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

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

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

where $\text{diag}\left(\sum_{i=1}^{n_{t+1}}\left(\frac{\partial}{\partial j} f_{t+1,i}\left(\theta_{t,w}\right)\right)^2\right)_{j=1,\ldots,d}$ 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 $E_{t,1} := \{c_{ada} < \lambda_{\min}\left(G_t\right) \leq \lambda_{\max}\left(G_t\right) < C_{ada}\}$. Then,

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

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

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

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 (6), 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.

C.9 Proof of Corollary 1

Since X and ϵ admit moments of order 4 and 2, and since $\mathbb{E}[XX^T]$ is positive, Assumptions 1, 2, 3 and 5 hold (see Boyer and Godichon-Baggioni (2023) amoung others). Furthermore, one has, considering the filtration $\mathcal{F}_t=(X_{1,1},X_{t,n},Y_{1,1},\ldots,Y_{t,n},,Z_{t,1},\ldots,Z_{t,n})$,

$$\mathbb{E}\left[\left\|\alpha_{t,j}X_{t,j}X_{t,j}^T\right\|^2|\mathcal{F}'_{t-1}\right]\leq\mathbb{E}\left[\|X\|^4\right].$$

Then, thanks to Theorem 5, θ_t and $\theta_{t,w}$ converge almost surely to θ^*. In addition, one has

$$\overline{S}_{t,w} = \frac{1}{N_{t,Z}} \sum_{i=1}^{t} \ln(i+1)^{w'} \sum_{j=1}^{n} X_{i,j} X_{i,j}^T$$

with $N_{t,Z} := 1 + \sum_{i=1}^{t} \ln(i+1)^{w'} \sum_{j=1}^{n} Z_{t,i}$. One has, since $p > 0$,

$$N_{t,Z} \xrightarrow[t\to+\infty]{a.s.} +\infty,$$

so that one can use a law of large number to check that $\overline{S}_{t,w}$ converges almost surely to $\mathbb{E}[XX^T]$, which concludes the proof.

C.10 Proof of Corollary 2

Since X admits a moment of order 4 and since in this case, for all $\theta \in \mathbb{R}^d$,

$$\nabla^2 F(\theta) = \mathbb{E}\left[\pi\left(X^T\theta\right)\left(1 - \pi\left(X^T\theta\right)\right) XX^T\right]$$

and $\nabla^2 F\theta^*)$ is supposed to be positive, Assumptions 1, 2, 3 and 5 hold (see Boyer and Godichon-Baggioni (2023) amoung others). Furthermore, one has,

$$\mathbb{E}\left[\left\|\alpha_{t,j} X_{t,j} X_{t,j}^T\right\|^2 |\mathcal{F}'_{t-1}\right] \leq \frac{1}{16}\mathbb{E}\left[\|X\|^4\right].$$

Then, thanks to Theorem 5, θ_t and $\theta_{t,w}$ converge almost surely to θ^*. In addition, one has

$$\overline{S}_{t,w} = \frac{1}{N_{t,Z}} \sum_{i=1}^{t} \ln(i+1)^{w'} \sum_{j=1}^{n} \alpha_{i,j} X_{i,j} X_{i,j}^T + \frac{1}{N_{t,Z}} \sum_{i=1}^{t} \ln(i+1)^{w'} \sum_{j=1}^{n} \iota_{i,j} Z_{i,j} e_{i,j} e_{i,j}^T,$$

where $\alpha_{i,j} = \pi\left(X_{i,j}^T \theta_{t-1}\right)\left(1 - \pi\left(X_{i,j}^T\theta\right)\right)$. 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]

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

i.e that

$$\frac{1}{N_{t,Z}} \sum_{i=1}^{t} \ln(i+1)^{w'} \sum_{j=1}^{n} \iota_{i,j} Z_{i,j} e_{i,j} e_{i,j}^T \xrightarrow[t\to+\infty]{} 0 \quad a.s.$$

4. See, e.g. Godichon-Baggioni et al. (2024); Bercu et al. (2023) for more details.

In addition,

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

where $\xi_{i,Z} := \sum_{j=1}^{n} Z_{i,j}\alpha_{i,j}X_{i,j}X_{i,j}^{\top} - \sum_{j=1}^{n} Z_{i,j}\nabla_{\theta}^{2}F\left(\theta_{i-1}\right)$ is a sequence of martingale differences for the filtration $\mathcal{F}'_{i-1} = \sigma\left(X_{1,1},\ldots,X_{i-1,n},Z_{i,1},\ldots,Z_{i,n}\right)$. Thus, since

$$\mathbb{E}\left[\left\|\xi_{t,Z}\right\|_{F}^{2}|\mathcal{F}'_{t-1}\right] \leq 2\left(\sum_{j=1}^{n}Z_{t,j}\mathbb{E}\left[\left\|\alpha_{t,j}X_{t,j}X_{t,j}^{\top}\right\|^{2}|\mathcal{F}'_{t-1}\right]^{\frac{1}{2}}\right)^{2} \leq 2^{-3}n^{2}\mathbb{E}\left[\|X\|^{4}\right],$$

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

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

In addition, by continuity of the Hessian and since $N_{t,Z}$ tends to infinity almost surely, one has

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

which concludes the proof.

C.11 Proof of Corollary 3

Before giving the proof, let us recall assumptions from Godichon-Baggioni and Lu (2024):

- **Assumption 1'.** The random variable X is absolutely continuous and is not concentrated around single points. There exists $C_6 > 0$ such that for all $\theta \in \mathbb{R}^d$,

$$\mathbb{E}\left[\frac{1}{\|X-\theta\|^{6}}\right] \leq C_6.$$

- **Assumption 2'.** The random variable X is not concentrated on a straight line. For all $\theta \in \mathbb{R}^d$, there exists $\theta' \in \mathbb{R}^d$ such that $\langle \theta, \theta' \rangle \neq 0$ and

$$\mathbb{V}\left[\langle X, \theta' \rangle\right] > 0.$$

The proof construction can be made by adapting the calculus in Godichon-Baggioni and Lu (2024) to our context. If Assumptions 1' and 2' are fulfilled, then Assumptions 1 to 3 and 5 hold. In addition, thanks to Assumption 1 in Godichon-Baggioni and Lu (2024), and denoting by $\mathcal{F}'_t = \sigma\left(X_{1,1},\ldots X_{t,n},U_{1,1},\ldots,U_{t+1,n},Z_{t,1},\ldots,Z_{t,n}\right)$ and $\mathcal{F}_t = \sigma\left(X_{1,1},\ldots X_{t,n},U_{1,1},\ldots,U_{t,n},Z_{t,1},\ldots,Z_{t,n}\right)$, we have

$$\mathbb{E}\left[\left\|\alpha_{t,j}\Phi_{t,j}\Phi_{t,j}^{T}\right\||\mathcal{F}_{t-1}\right] \leq \mathbb{E}\left[\left\|U_{t,j}\right\|^{2}\mathbb{E}\left[\frac{1}{\left\|X_{t,j}-\theta_{t-1}\right\|}|\mathcal{F}'_{t-1}\right]|\mathcal{F}_{t-1}\right] \leq C_6^{1/6}\mathbb{E}\left[\|U\|^{2}\right],$$

so that θ_{t-1} converges almost surely to θ^*.

Let us now prove the convergence of the estimate of the Hessian. First, remark that for all $\theta \in \mathbb{R}^d$,

$$\nabla^2 F(\theta) = \mathbb{E}\left[\frac{1}{\|X-\theta\|}\left(I_d - \frac{(X-\theta)(X-\theta)^T}{\|X-\theta\|^2}\right)\right] =: \mathbb{E}\left[\nabla^2 f(X,\theta)\right].$$

For all $j = 1, \ldots, n$ and $t \in [0,1]$, let us denote $X_{t,j,u} = X_{t,j} - (\theta_{t-1} + tv_t U_{t,j})$ and

$$w_{t,j,u} = \frac{1}{\|X_{t,j,u}\|}\left(I_d - \frac{X_{t,j,u}X_{t,j,u}^T}{\|X_{t,j,u}\|^2}\right) = \nabla^2 f\left(X_{t,j,u}, \theta_{t-1}\right),$$

and let us remark that

$$\Phi_{t,j} = \nabla f\left(X_{t,j,1}, \theta_{t-1}\right) - \nabla f\left(X_{t,j,0}\right) = \int_0^1 w_{t,j,u} du v_t U_{t,j}.$$

Then,

$$\begin{aligned}
\alpha_{t,j}\Phi_{t,j}\Phi_{t,j}^T &= \frac{\|X_{t,j}-\theta_{t-1}\|}{v_t^2}\int_0^1 w_{t,j,u} du v_t^2 U_{t,j}U_{t,j}^T \int_0^1 w_{t,j,u} du \\
&= \|X_{t,j}-\theta_{t-1}\|\int_0^1 w_{t,j,u} - w_{t,j,0} du U_{t,j}U_{t,j}^T \int_0^1 w_{t,j,u} du \\
&\quad + \|X_{t,j}-\theta_{t-1}\| w_{t,j,0} U_{t,j}U_{t,j}^T \int_0^1 w_{t,j,u} - w_{t,j,0} du \\
&\quad + \|X_{t,j}-\theta_{t-1}\| w_{t,j0} U_{t,j}U_{t,j}^T w_{t,j,0}.
\end{aligned}$$

Remark that since $U_{t,j}$ is independent from $X_{t,j}$ and θ_{t-1}, and considering the filtration $\mathcal{F}_{t-1} = \sigma\left(X_{1,1}, \ldots, X_{t-1,n}, \ldots, U_{1,1}, \ldots, U_{t-1,n}, Z_{1,1}, \ldots, Z_{t,n}\right)$, and since $w_{t,j,0}^2 = \frac{1}{\|X_{t,j,0}-\theta_{t-1}\|} w_{t,j,0}$, it comes that

$$\mathbb{E}\left[\|X_{t,j}-\theta_{t-1}\| w_{t,j0} U_{t,j}U_{t,j}^T w_{t,j,0} | \mathcal{F}_{t-1}\right] = \mathbb{E}\left[w_{t,j,0}|\mathcal{F}_{t-1}\right] = \nabla^2 F\left(\theta_{t-1}\right).$$

Since $Z_{i,j}$ is $\mathcal{F}_{i-1}$-measurable, one has

$$\begin{aligned}
&\frac{1}{N_{t,Z}}\sum_{i=1}^t \log(i+1)^{w'} \sum_{j=1}^n Z_{i,j}\|X_{i,j}-\theta_{i-1}\| w_{i,j0} U_{i,j}U_{i,j}^T w_{i,j,0} \\
&\quad = \frac{1}{N_{t,Z}}\sum_{i=1}^t \log(i+1)^{w'} \sum_{j=1}^n Z_{i,j}\nabla^2 F\left(\theta_{i-1}\right) + \frac{1}{N_{t,Z}}\sum_{i=1}^t \log(i+1)^{w'} \sum_{j=1}^n Z_{i,j}\Xi_{i,j},
\end{aligned}$$

where $\Xi_{i,j} := \|X_{i,j}-\theta_{i-1}\| w_{i,j0} U_{i,j}U_{i,j}^T w_{i,j,0} - \nabla^2 F\left(\theta_{i-1}\right)$. Denoting, for all $\delta > w'$, $M_t = \frac{1}{N_{t,Z}\log(N_{t,Z})^{1+\delta}}\left\|\sum_{i=1}^t \log(i+1)^{w'}\sum_{j=1}^n Z_{i,j}\Xi_{i,j}\right\|^2$, one have

$$\begin{aligned}
\mathbb{E}\left[M_{t+1}|\mathcal{F}_t\right] &= \frac{N_{t,Z}\log(N_{t,Z})^{1+\delta}}{N_{t+1,Z}\log(N_{t,Z})^{1+\delta}} M_t + \frac{1}{N_{t+1,Z}\log(N_{t+1,Z})^{1+\delta}}\mathbb{E}\left[\left\|\log(i+1)^{w'}\sum_{j=1}^n Z_{i,j}\Xi_{t+1,j}\right\|^2 |\mathcal{F}_t\right] \\
&\leq M_t + \frac{1}{N_{t+1,Z}\log(N_{t+1,Z})^{1+\delta}}\log(t+1)^{2w'}\sum_{j=1}^n \mathbb{E}\left[\|\Xi_{t+1,j}\|^2|\mathcal{F}_t\right].
\end{aligned}$$

Then, thanks to Assumption 1 in Godichon-Baggioni and Lu (2024) (see the Proof of Theorems 1 and 2 for more details), one has

$$\mathbb{E}\left[\|\Xi_{t+1,j}\|^2 |\mathcal{F}_t\right] \leq C_6^{1/2}\mathbb{E}\left[\|U\|^4\right].$$

Then, remarking that $\log(t+1)^{w'}t/N_{t,Z} = \mathcal{O}(1)$ a.s. and applying Robbins Siegmund Theorem, M_t converges almost surely to a random finite variable, i.e.,

$$\left\|\frac{1}{N_{t,Z}}\sum_{i=1}^{t}\log(i+1)^{w'}\sum_{j=1}^{n}Z_{i,j}\Xi_{i,j}\right\|^2 = \mathcal{O}\left(\frac{\ln(1+N_{t,Z})^{1+\delta}}{N_{t,Z}}\right) \text{ a. s.},$$

which means that this term converges to zero. In addition, since θ_{t-1} converges almost surely to θ^* and by continuity of the Hessian of F, one can easily prove that

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

At last, observe that thanks to Godichon-Baggioni and Lu (2024, Lemma 1), for any $q \leq 3$, we have

$$\mathbb{E}\left[\|w_{t,j,u} - w_{t,j,0}\|^q |\mathcal{F}_{t-1}\right] \leq 6^q C_6^{q/3} v_t^q \|U_{t,j}\|^q$$

and this term converges fastly to zero. Then, following the proof in Godichon-Baggioni and Lu (2024), one can easily prove that

$$\frac{1}{N_{t,Z}}\sum_{i=1}^{t}\log(i+1)^{w'}\|X_{i,j} - \theta_{i-1}\|\int_0^1 w_{i,j,u} - w_{i,j,0}du U_{i,j}U_{i,j}^T\int_0^1 w_{i,j,u}du \xrightarrow[t\to+\infty]{a.s.} 0$$

$$\frac{1}{N_{t,Z}}\sum_{i=1}^{t}\log(i+1)^{w'}\|X_{i,j} - \theta_{t-1}\|w_{t,j,0}U_{t,j}U_{t,j}^T\int_0^1 w_{t,j,u} - w_{t,j,0}du \xrightarrow[t\to+\infty]{a.s.} 0,$$

which concludes the proof.

References

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

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

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

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

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

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

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

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

H. N. Chau, J. L. Kirkby, D. H. Nguyen, D. Nguyen, N. N. Nguyen, and T. Nguyen. On the inversion-free newton's method and its applications. *International Statistical Review*, 2024.

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

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

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

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

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

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

N. Gazagnadou, R. Gower, and J. Salmon. Optimal mini-batch and step sizes for saga. In *International Conference on Machine Learning*, pages 2142–2150. PMLR, 2019.

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

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

A. Godichon-Baggioni and W. Lu. Online stochastic newton methods for estimating the geometric median and applications. *Journal of Multivariate Analysis*, 2024.

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

A. Godichon-Baggioni, N. Werge, and O. Wintenberger. Learning from time-dependent streaming data with online stochastic algorithms. *Transactions on Machine Learning Research*, 2023a. ISSN 2835-8856.

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

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

R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik. Sgd: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR, 2019.

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

M. Kelly, R. Longjohn, and K. Nottingham. The UCI Machine Learning Repository.

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

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

S. Lacoste-Julien, M. Schmidt, and F. Bach. A simpler approach to obtaining an O (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

R. Leluc and F. Portier. Asymptotic analysis of conditioned stochastic gradient descent. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

A. Mokkadem and M. Pelletier. A generalization of the averaging procedure: The use of two-time-scale algorithms. *SIAM Journal on Control and Optimization*, 49(4):1523–1543, 2011.

E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems*, 24, 2011.

D. Needell, R. Ward, and N. Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Advances in Neural Information Processing Systems*, 27, 2014.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

L. Nguyen, P. H. Nguyen, M. Dijk, P. Richtárik, K. Scheinberg, and M. Takác. Sgd and hogwild! convergence without the bounded gradients assumption. In *International Conference on Machine Learning*, pages 3750–3758. PMLR, 2018.

J. Nocedal and S. J. Wright. *Numerical optimization*. Springer, 1999.

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

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

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

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

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

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

J. Sherman and W. J. Morrison. 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, 1950.

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

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

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

S. Vaswani, F. Bach, and M. Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR, 2019.

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