# Learning from time-dependent streaming data with online stochastic algorithms

**Anonymous authors**
**Paper under double-blind review**

## Abstract

This paper addresses the problem of stochastic optimization in a streaming setting, where the objective function must be minimized using only time-dependent and biased estimates of its gradients. The study presents a non-asymptotic analysis of various Stochastic Gradient (SG) based methods, including the well-known SG descent, mini-batch SG, and time-varying mini-batch SG methods, as well as their iterated averages (Polyak-Ruppert averaging). The analysis establishes novel heuristics that link dependence, biases, and convexity levels, which allow for the acceleration of convergence of SG-based methods. Specifically, the heuristics demonstrate how SG-based methods can overcome long- and short-range dependencies and biases. In particular, the use of time-varying mini-batches counteracts dependency structures and biases while ensuring convexity, and combining this with Polyak-Ruppert averaging further accelerates convergence. These heuristics are particularly useful for learning problems with highly dependent data, noisy variables, and lacking convexity. Our results are validated through experiments using simulated and real-life data.

## 1 Introduction

In recent decades, intelligent systems, such as machine learning and artificial intelligence, have become mainstream in many parts of society, e.g., through online, deep, reinforcement, and supervised learning (Goodfellow et al., 2016; Sutton & Barto, 2018; Hastie et al., 2009; Hazan et al., 2016; Shalev-Shwartz et al., 2012). These systems predominantly follow traditional learning schemes, also known as batch or offline learning, where the model is trained on an entire dataset before making predictions on new data. However, such methods have significant drawbacks, such as expensive re-training costs, as the model needs to be retrained from scratch when new data arrives, leading to poor scalability in large and real-world applications.

At the same time, these intelligent systems generate vast amounts of data, much of which arrives as a continuous stream of samples, known as streaming data Examples of streaming data include network traffic (tweets, search engines, advertisements), self-driving cars, financial investments, weather data, and sensor data (Kushner & Yin, 2003; Abu-Mostafa et al., 2012). Streaming data presents unique challenges as the algorithm must adapt to the data observed so far to predict new samples accurately. Such streaming algorithms can never be seen as complete but must be updated continuously as newer samples arrive. Methods that re-calculate the model from scratch on the arrival of new samples are impractical due to their high computational cost. Therefore we need procedures that effectively update as more samples arrive. This computational efficiency should not be at the expense of accuracy; the model's accuracy should be close to that achieved if we built a model from scratch using all the samples (Bottou & Cun, 2003).

A hallmark of learning from streaming data (or large-scale learning) is the uncertainty from limited (or no) access to accurate information about incoming data. We assume that the optimization's objectives are stochastic, leading to Stochastic Optimization (SO) (Nemirovskij & Yudin, 1983). Solving the SO problem in a streaming framework means we approach the objective using the gradually arriving samples drawn according to an unknown time-dependent process. Next, the computational complexity of an algorithm is a limited element when handling streaming data. Stochastic algorithms, such as the Stochastic Gradient (SG) method (Robbins & Monro, 1951), have proven effective in overcoming the drawbacks of traditional

(batch/offline) learning methods, as they use the samples one by one without knowing their total number. These SG methods are scalable and robust in many areas ranging from smooth and strongly convex problems to complex non-convex ones (Bach & Moulines, 2013; Moulines & Bach, 2011). They solve many large-scale machine learning tasks for real-world applications where data are large in size (and dimension) and arrive at a high velocity. Such first-order methods have been intensively studied in theory and practice in recent years (Bottou et al., 2018).

The classical analyses of SO problems typically require unbiased gradients drawn independently and identically distributed (i.i.d.) from some underlying (and unknown) data generation process (Cesa-Bianchi et al., 2004). But in practice, the inference often involves data-generating processes that produce highly dependent data examples, which are known to heavily bias the problem and slow down convergence. Such time-dependent streaming data can, for example, be meteorological or financial time series. Nevertheless, SG-based methods can converge even if they only have access to biased gradients (Bertsekas, 2016; d'Aspremont, 2008; Devolder et al., 2011; Schmidt et al., 2011). However, many of these studies have been conducted with specific applications in mind. Yet, some researchers have examined the convergence of SG-based methods under these difficult settings, e.g., see Agarwal & Duchi (2012); Karimi et al. (2019); Ma et al. (2022); Ajalloeian & Stich (2020); Chen & Luss (2018); Schmidt et al. (2011).

While the above works utilized concepts of data dependence to characterize different SG-based methods for dependent data, there is still a lack of theoretical understanding of how different levels of data dependence affect these algorithms. In particular, the learning scheme of SG-based methods critically affects the bias and variance of the learning process. In fact, under i.i.d. data and convexity, SG-based methods achieve only scaled improvements in their convergence by using constant mini-batch vs. single batch (Godichon-Baggioni et al., 2021). However, these learning schemes may lead to substantially different convergence behaviors over highly dependent data, as the gradients are no longer unbiased estimates. Therefore, it is vital to understand the interplay between data dependence and SG-based methods.

**Contributions.** This paper goes beyond standard assumptions by allowing for dependent and biased gradient estimates in stochastic optimization problems in a streaming framework. The study provides non-asymptotic convergence rates of time-varying mini-batch SG-based methods, which can be applied to a wide range of applications with dependence and biased inputs. The results establish a connection between the dependency and convexity levels to the model parameters, enabling us to achieve and improve convergence.

The study shows that SG-based methods can overcome long- and short-range dependence by using time-varying mini-batches, which counteract the dependency structures. Additionally, the results demonstrate that biased SG-based methods converge and can achieve the same accuracy as unbiased SG-based methods if the bias is not too large.

Furthermore, the study presents an explicit heuristic that can be used in practice to increase the stability of SG-based methods and improve convergence. Specifically, the findings show that increasing mini-batches is essential to break dependence and ensure convexity, while the use of Polyak-Ruppert averaging can accelerate convergence (Polyak & Juditsky, 1992; Ruppert, 1988). These contributions are particularly valuable for learning problems with highly dependent data, noisy variables, and lacking convexity.

**Organization.** Section 2 provides an overview of the streaming framework on which our non-asymptotic analysis is based. We introduce key concepts, definitions, and assumptions, including those related to dependency structures for gradient estimates. In particular, section 2.3 outlines our assumptions regarding $\alpha$-mixing conditions for verifying dependency structures.

In section 3, we present our convergence results for SG-based algorithms, both with and without averaging. The proofs of our results are provided in appendix A. We discuss each result in detail, drawing connections to previous work in the field. Our convergence analysis depends on the assumptions outlined in section 2 and some additional conditions for the averaged case (section 3.3).

In section 4, we provide experimental results that validate our findings on synthetic and real-life time-dependent streaming data. Finally, we conclude our paper in section 5 with some closing remarks.

## 2 Problem formulation

In statistics and machine learning, one often wants to describe the behavior of a real system of interest, usually in the form of a parameterized mathematical model (Hastie et al., 2009; Bottou et al., 2018). Therefore, we set up a mathematical function representing how well the model describes the system of interest with the model parameters as arguments: we consider the Stochastic Optimization (SO) problem

$$\min_{\theta\in\mathbb{R}^d}\{L(\theta)=\mathbb{E}[l_t(\theta)]\}, \tag{1}$$

where $l_t:\mathbb{R}^d\to\mathbb{R}$ are some differentiable random functions (possibly non-convex), e.g, see Nesterov et al. (2018); Nemirovskij & Yudin (1983). Throughout this paper, we refer to $L$ as the objective function.

An example of such a SO problem (1) can be given as follows (Kushner & Yin, 2003): there is an unknown one-to-one mapping $L:\mathbb{R}^d\to\mathbb{R}$ embedded into the system by nature, which we are interested in. Thus, in order to approximate $L$ (and recover $\theta$ from it), we use the stochastic gradient estimates $(\nabla_\theta l_t)$, where $l_t$ is e.g., the loss between the predicted $h_\theta(X_t)$ and true $Y_t$ outputs, respectively; here we assume that the prediction function $h_\theta$ has a fixed form and is parameterized by a vector $\theta\in\mathbb{R}^d$ over which the optimization is to be performed, e.g., see Teo et al. (2007) for examples of scalar and vectorial loss functions and their derivatives. Hence, the aim is to find $\theta$ such that the prediction function $h_\theta$ minimizes the objective $L$.

### 2.1 Stochastic streaming gradient methods

We solve the SO problem (1) in a streaming framework, where a time-varying mini-batch $l_t=(l_{t,1},\dots,l_{t,n_t})$ of $n_t\in\mathbb{N}$ random functions arrives at any given time $t\in\mathbb{N}$. To solve the SO problem (1), we use the Stochastic Streaming Gradient (SSG) method proposed by Godichon-Baggioni et al. (2021), given as

$$\textbf{(SSG)}\qquad \theta_t=\theta_{t-1}-\frac{\gamma_t}{n_t}\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right), \tag{2}$$

where $\gamma_t$ is the learning rate satisfying the conditions $\sum_{i=1}^{\infty}\gamma_i=\infty$ and $\sum_{i=1}^{\infty}\gamma_i^2<\infty$ (Robbins & Monro, 1951). Note that if $\forall t$, $n_t=1$, SSG becomes the well-known SG descent, which has attracted a lot of attention (Bousquet & Elisseeff, 2002; Hardt et al., 2016; Shalev-Shwartz et al., 2011; Zhang, 2004; Xiao, 2009). In many models, there may be constraints on the parameter space, which would require a projection of the parameters; therefore, we also introduce the Projected SSG (PSSG) estimate, defined by

$$\textbf{(PSSG)}\qquad \theta_t=\mathcal{P}_\Theta\left(\theta_{t-1}-\frac{\gamma_t}{n_t}\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right), \tag{3}$$

where $\mathcal{P}_\Theta$ denotes the Euclidean projection onto a closed convex set $\Theta$ in $\mathbb{R}^d$, i.e., $\mathcal{P}_\Theta(\theta)=\arg\min_{\theta'\in\Theta}\|\theta-\theta'\|_2$. If one examined the trajectory of the stochastic gradients $(\nabla_\theta l_{t,i})$, one would be surprised; they contain high noise levels and lack robustness which can lead to slow convergence or prevent convergence altogether. Therefore, it intuitively makes sense to use sets of stochastic gradient estimates in each iteration as it reduces the variance and makes it easier to adjust the learning rate $(\gamma_t)$, which improves the quality of each iteration.

In this streaming framework, we are also interested in acceleration approaches to the existing algorithms in (2) and (3). An essential extension is the Polyak-Ruppert procedure (Polyak & Juditsky, 1992; Ruppert, 1988), which guarantees optimal statistical efficiency without jeopardizing the computational cost: the Averaged SSG (ASSG) is given by

$$\textbf{(ASSG)/(APSSG)}\qquad \bar{\theta}_t=\frac{1}{N_t}\sum_{i=0}^{t-1}n_{i+1}\theta_i, \tag{4}$$

where $N_t=\sum_{i=1}^{t}n_i$ is the accumulated sum of observations; in order to compare our streaming methods fairly, we should always compare in terms of the number of observations used, namely using $N_t$. Likewise, let APSSG denote the (Polyak-Ruppert) averaged estimate of PSSG in (3). These averaging methods sequentially

aggregate past estimates, which leads to smoother curves (i.e., variance reduction in the estimation trajectories), and accelerates the convergence (Polyak & Juditsky, 1992). Practically, as we handle data sequentially, we will make use of the rewritten formula, $\bar{\theta}_t = (N_{t-1}/N_t)\bar{\theta}_{t-1} + (n_t/N_t)\theta_{t-1}$ with $\bar{\theta}_0 = 0$. Pseudo-code of these streaming estimates (2) to (4) are presented in algorithm 1. In addition, (4) can be modified to a weighted average version, giving greater weight to the latest estimates and thereby improving the convergence while limiting the effect of poor initializations; examples of such algorithms can be found in Boyer & Godichon-Baggioni (2022); Mokkadem & Pelletier (2011).

**Algorithm 1:** Stochastic streaming gradient estimates (SSG/PSSG/ASSG/APSSG)

**Inputs** : $\theta_0 \in \Theta \subseteq \mathbb{R}^d$, project: **True** or **False**, average: **True** or **False**
**Outputs** : $\theta_t$, $\bar{\theta}_t$ (resulting estimates)
Initialization: $\bar{\theta}_0 \in \mathbb{R}^d$
**for** *each* $t \geq 1$*, a time-varying mini-batch of* $n_t$ *data arrives,* **do**
  $\theta_t \leftarrow \theta_{t-1} - \frac{\gamma_t}{n_t} \sum_{i=1}^{n_t} \nabla_\theta l_{t,i}(\theta_{t-1})$ `/* update */`
  **if** *project* **then**
    $\theta_t \leftarrow \mathcal{P}_\Theta(\theta_t)$ `/* project */`
  **if** *average* **then**
    $\bar{\theta}_t \leftarrow (N_{t-1}/N_t)\bar{\theta}_{t-1} + (n_t/N_t)\theta_{t-1}$ `/* average */`

Due to the massive popularity of SG-based methods, it is obvious to ask how we can make SG-based algorithms even more efficient, robust, and user-friendly for several different optimization problem. This question has led to very many variants, e.g., see Boyd & Vandenberghe (2004); Nesterov et al. (2018); Byrd et al. (2016) for more details on second-order methods (such as Newton's method), or other extensions. The choice of learning rate ($\gamma_t$) has a significant impact on the convergence of SG methods; if it is too small, it will slow down the convergence, while too high a learning rate may prevent convergence as the loss function will fluctuate around the minimum. Thus, an adaptive learning rate would be much more effortless to adjust and more user-friendly, as it requires less fine-tuning. In addition, it would be preferable to have a learning rate per-dimension, which thereby adjust learning individually as convergence evolves.[1] Bottou et al. (2018) gives an overview of various SG-based methods for (convex and non-convex) optimization, including how to parallelize and distribute the SG updates.

### 2.2 Quasi-strong convex objectives

The analysis of SO algorithms requires assumptions on the objective function $L$: the SO problem (1) is specified over a convex domain $\Theta$, which in this paper we always take to be a closed subset of $\mathbb{R}^d$, $d \geq 1$, and an objective function $L : \Theta \to \mathbb{R}$ which is convex with respect to its argument $\theta \in \Theta$. This is a closely related branch of optimization tools for (online) convex optimization (Boyd & Vandenberghe, 2004; Nesterov et al., 2018; Hazan et al., 2016; Shalev-Shwartz et al., 2012). Following Moulines & Bach (2011); Gower et al. (2019); Nguyen et al. (2019), we assume that $L$ has a unique global minimizer $\theta^* \in \Theta$ such that $\nabla_\theta L(\theta^*) = 0$, and it is $\mu$-quasi-strongly convex (Karimi et al., 2016; Necoara et al., 2019), i.e, there exists $\mu > 0$ such that $\forall \theta \in \Theta$,

$$L(\theta^*) \geq L(\theta) + \langle \nabla_\theta L(\theta), \theta^* - \theta \rangle + \frac{\mu}{2}\|\theta^* - \theta\|^2. \tag{5}$$

The $\mu$-quasi-strongly convexity assumption is a non-strongly convex relaxation of the SO problem, which is more conservative than $\mu$-strongly convexity. Teo et al. (2007) provides a comprehensive record of various objectives $L$ used in machine learning applications. Milder degrees of convexity have been studied by, e.g., Karimi et al. (2016), which studied SG-based methods under the Polyak-Łojasiewicz condition (Polyak, 1963; Lojasiewicz, 1963), or Gadat & Panloup (2023), which studied the Ruppert-Polyak averaging estimate under some Kurdyka-Łojasiewicz-type condition (Kurdyka, 1998; Lojasiewicz, 1963). Relaxations of convexity is

[1] Some of the most common adaptive learning algorithms for SG optimization is Momentum (Qian, 1999), Nesterov accelerated gradient (Nesterov, 1983), Adagrad (Duchi et al., 2011), Adadelta (Zeiler, 2012), RMSprop (Tieleman et al., 2012), and Adam (Kingma & Ba, 2014).

crucial in practice to ensure robustness and adaptiveness of the algorithms, e.g., for non-strongly convex SO, see Bach & Moulines (2013); Nemirovski et al. (2009); Necoara et al. (2019).

### 2.3 Assumptions; dependence, bias, expected smoothness, and gradient noise

We go beyond the classical assumptions that require unbiased (uniformly bounded) gradient estimates by allowing the gradients to be dependent and biased estimates (Godichon-Baggioni et al., 2021). Our aim is to non-asymptotically bound the SSG estimates (2) to (4) explicitly using the SO problem parameters. To shorten notation, we let $\nabla_\theta l_t(\theta) = n_t^{-1} \sum_{i=1}^{n_t} \nabla_\theta l_{t,i}(\theta)$. Next, let the natural filtration of the SO problem $\mathcal{F}_t = \sigma(l_i : i \leq t)$ with $l_t = (l_{t,1}, \ldots, l_{t,n_t})$, and assume the following about the stochastic gradients $(\nabla_\theta l_t)$:

**Assumption 1-p** ($D_\nu \nu_t$-dependence and $B_\nu \nu_t$-bias)**.** *Let $\theta_0$ be $\mathcal{F}_0$-measurable. For each $t \geq 1$, the random function $\nabla_\theta l_t(\theta)$ is square-integrable, $\mathcal{F}_t$-measurable, and for a positive integer $p$, there exists some positive sequence $(\nu_t)_{t\geq 1}$ and constants $D_\nu, B_\nu \geq 0$ such that*

$$\mathbb{E}[\|\mathbb{E}[\nabla_\theta l_t(\theta)|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta)\|^p] \leq \nu_t^p (D_\nu^p \mathbb{E}[\|\theta - \theta^*\|^p] + B_\nu^p), \tag{6}$$

*for all $\mathcal{F}_{t-1}$-measurable $\theta \in \Theta$.*

**Assumption 2-p** ($\kappa_t$-expected smoothness)**.** *For a positive integer $p$, there exists some positive sequence $(\kappa_t)_{t\geq 1}$ such that $\forall \theta \in \Theta$, $\mathbb{E}[\|\nabla_\theta l_t(\theta) - \nabla_\theta l_t(\theta^*)\|^p] \leq \kappa_t^p \mathbb{E}[\|\theta - \theta^*\|^p]$.*

**Assumption 3-p** ($\sigma_t$-gradient noise)**.** *For a positive integer $p$, there exists some positive sequence $(\sigma_t)_{t\geq 1}$ such that $\mathbb{E}[\|\nabla_\theta l_t(\theta^*)\|^p] \leq \sigma_t^p$.*

**Discussion of assumptions.** Assumptions 1-p to 3-p are milder assumptions than the standard assumptions for stochastic approximations, e.g., see Benveniste et al. (2012); Kushner & Yin (2003); Moulines & Bach (2011); Godichon-Baggioni et al. (2021). They include classic examples such as stochastic approximation but also more complex models such as learning from time-dependent data, which we will demonstrate later in section 4.

Assumption 1-p is on the form of mixing conditions for weakly dependence sequences, implying that dependence dilutes with the rate of $\nu_t$. It is possible to verify Assumption 1-p by moment inequalities for partial sums of strongly mixing sequences (Rio, 2017); we will refer to this as short-range dependence. Note that for any positive integer $p$, Assumption 1-p can be upper bounded by

$$\mathbb{E}[\|\mathbb{E}[\nabla_\theta l_t(\theta)|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta)\|^p] \leq \mathbb{E}[\|\nabla_\theta l_t(\theta) - \nabla_\theta L(\theta)\|^p] = n_t^{-p} \mathbb{E}[\|S_t\|^p], \tag{7}$$

using Jensen's inequality, where $S_t = \sum_{i=1}^{n_t} (\nabla_\theta l_{t,i}(\theta) - \nabla_\theta L(\theta))$ is a $d$-dimensional vector. Let $(\nabla_\theta l_{t,i})$ be a strictly stationary sequence and assume that there exists some $r > p$ such that $\sup_{x>0}(x^r Q(x))^{1/r} < \infty$, where $Q(x)$ denotes the quantile function of $\|\nabla_\theta l_{t,i}\|$. Suppose that $(\nabla_\theta l_{t,i})$ is strongly $\alpha$-mixing in the sense of Rosenblatt (1956), with strong mixing coefficients $(\alpha_t)_{t\geq 1}$ satisfying $\alpha_t = \mathcal{O}(t^{-pr/(2r-2p)})$. Then by Rio (2017, Corollary 6.1), we have that $\mathbb{E}[\|S_t\|^p] = \mathcal{O}(n_t^{p/2})$, meaning, (7) is at most $\mathcal{O}(n_t^{-p/2})$; this includes several linear, non-linear, and Markovian time series, e.g., see Bradley (2005); Doukhan (2012) for more examples, other mixing coefficients of weak dependence and the relations between them. In relation to the form of Assumption 1-p, this means that $B_\nu \neq 0$ in this case. However, having $B_\nu = 0$ is possible in unbiased examples, which we will see later in section 4.

Regarding Assumptions 2-p and 3-p, the classical convergence analysis of SG methods is conducted under uniformly bounded gradients $(\nabla_\theta l_t)$. However, this assumption is too restrictive as it only may hold for some losses (Bottou et al., 2018; Nguyen et al., 2018). Instead, we follow the same ideas as in Moulines & Bach (2011); Gower et al. (2019), to make an assumption about the expected smoothness of $(\nabla_\theta l_t)$ (Assumption 2-p) and an assumption about the expected finitude of $(\nabla_\theta l_t(\theta^*))$ (Assumption 3-p). Note that Assumptions 2-p and 3-p can be verified using $\alpha$-mixing conditions by analogues arguments as for Assumption 1-p such that $\kappa_t^p$ and $\sigma_t^p$ is $\mathcal{O}(n_t^{-p/2})$.

## 3 Convergence analysis

In this section, we consider the SSG estimates in (2) to (4) with streaming batches $(n_t)$ arriving in non-decreasing streams. We aim to non-asymptotically bound $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ and $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$, such

that they only depend on the parameters of the problem (Sections 3.2 and 3.3). But before we do this, we will in section 3.1 specify the function forms of the learning rate $(\gamma_t)$, the uncertain terms $(\nu_t)$, $(\kappa_t)$, $(\sigma_t)$, and the streaming batch $(n_t)$.

### 3.1 Learning rates and function forms

Throughout this paper, we consider learning rates on the form $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ with $C_\gamma > 0$, $\beta \in [0, 1]$, and $\alpha$ chosen accordingly to the expected streaming batches $n_t$. This learning rate gives us the opportunity to add more weight to larger streaming batches through the hyperparameter $\beta$. Obviously, $(\nu_t)$, $(\kappa_t)$, and $(\sigma_t)$ may be considered as uncertain terms depending on the streaming batches $(n_t)$. Thus, let $\nu_t = n_t^{-\nu}$, $\kappa_t = C_\kappa n_t^{-\kappa}$, and $\sigma_t = C_\sigma n_t^{-\sigma}$ with $\nu \in (0, \infty)$, $\kappa, \sigma \in [0, 1/2]$, and $C_\kappa, C_\sigma > 0$. Having, $\sigma, \kappa \in [0, 1/2]$ follows directly from Godichon-Baggioni et al. (2021), since $\sigma = \kappa = 1/2$ corresponds to the i.i.d. case[2], whereas $\sigma, \kappa < 1/2$ allows noisier outputs. Similarly, $v_t = 0$ corresponds to the classical i.i.d. setting (Godichon-Baggioni et al., 2021). Having $\nu_t = n_t^{-\nu}$ means Assumption 1-p, allow so-called long-range dependence when $\nu \in (0, 1/2)$ (also known as long memory or long-range persistence) and short-range dependence when $\nu \in [1/2, \infty)$. Thus, the i.i.d. case is when $\nu \to \infty$.

Next, for the sake of simplicity, we consider (time-varying) streaming batches $(n_t)$ on the form $C_\rho t^\rho$ with $C_\rho \in \mathbb{N}$ and $\rho \in [0, 1)$ such that $n_t \in \mathbb{N}$; this form includes classical (online) SG descent methods when $\{C_\rho = 1, \rho = 0\}$ and (online) mini-batch procedures of both constant and time-varying size when $\{C_\rho \in \mathbb{N}, \rho = 0\}$ and $\{C_\rho \in \mathbb{N}, \rho \in [0, 1)\}$, respectively, as well as the Polyak-Ruppert average of (online) time-varying mini-batches. We will refer to $C_\rho$ as the *streaming batch* size and $\rho$ as the *streaming rate*.

### 3.2 Stochastic streaming gradients

**Theorem 1** (SSG/PSSG)**.** *Let $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$, where $(\theta_t)$ either follows the recursion in (2) or (3). Suppose Assumptions 1-p to 3-p hold for $p = 2$. If $\mu_\nu = \mu - \mathbb{1}_{\{\rho=0\}} 2 D_\nu C_\rho^{-\nu} > 0$, then there exist explicit constants $C_\delta, C'_\delta, C''_\delta > 0$ such that for $\alpha - \rho\beta \in (1/2, 1)$, we have*

$$\delta_t \leq \mathcal{O}\left(\exp\left(-\frac{\mu C_\gamma N_t^{\frac{1+\rho\beta-\alpha}{1+\rho}}}{C_\delta C_\rho^{\frac{1-\beta-\alpha}{1+\rho}}}\right)\right) + \frac{C'_\delta B_\nu^2}{\mu\mu_\nu C_\rho^{\frac{2\nu}{1+\rho}} N_t^{\frac{2\rho\nu}{1+\rho}}} + \frac{C''_\delta C_\sigma^2 C_\gamma}{\mu_\nu C_\rho^{\frac{2\sigma-\beta-\alpha}{1+\rho}} N_t^{\frac{\rho(2\sigma-\beta)+\alpha}{1+\rho}}}. \tag{8}$$

*An explicit version of this bound is given in appendix A.*

**Sketch of proof.** Under Assumptions 1-p to 3-p with $p = 2$, it can be shown that $(\delta_t)$ satisfies the recursive relation,

$$\delta_t \leq [1 - (\mu - 2D_\nu\nu_t)\gamma_t + 2\kappa_t^2\gamma_t^2]\delta_{t-1} + \mu^{-1}B_\nu^2\nu_t^2\gamma_t + 2\sigma_t^2\gamma_t^2, \tag{9}$$

for $(\gamma_t)$, $(\nu_t)$, $(\kappa_t)$, $(\sigma_t)$, and $(n_t)$ on any form. This recursive relation (9) can be explicitly upper bounded in a non-asymptotic way using classical techniques from stochastic approximations (Benveniste et al., 2012; Kushner & Yin, 2003). As mentioned in Zinkevich (2003), bounding the projected estimate in (3) follows directly from that $\mathbb{E}[\|\mathcal{P}_\Theta(\theta) - \theta^*\|^2] \leq \mathbb{E}[\|\theta - \theta^*\|^2]$, as $\Theta$ is a closed convex set. Note that the projected estimate (3) could also be proved through a bounded gradient assumption that replaces Assumptions 2-p and 3-p, e.g., see Moulines & Bach (2011); Godichon-Baggioni et al. (2021).

**Related work.** Theorem 1 replicates the results of the unbiased i.i.d. case considered in Godichon-Baggioni et al. (2021), i.e., with $B_\nu = 0$ and $\sigma = 1/2$. Furthermore, our findings also reproduce the results of Moulines & Bach (2011), where they considered the unbiased i.i.d. case (under slightly different assumptions) using the SG descent, i.e., when $C_\rho = 1$ and $\rho = 0$.

**Bound on function values.** If the objective $L$ has $C_\nabla$-Lipschitz continuous gradients, then (8) implies a bound on the function values of $L$, namely, $\mathbb{E}[L(\theta_t) - L(\theta^*)] \leq C_\nabla \delta_t/2$ by Cauchy–Schwarz's inequality. Indeed, when we consider the average estimate (4) in section 3.3, we assume that the objectives $L$ has $C_\nabla$-Lipschitz continuous gradients (Assumption 4).

[2]You can't beat the system.

**Ensuring convexity through non-decreasing streaming batches.** The positivity of the dependence penalised convexity constant $\mu_\nu = \mu - \mathbb{1}_{\{\rho=0\}} 2D_\nu C_\rho^{-\nu}$ is essential in all terms of (8). If the streaming rate $\rho = 0$, then $\mu_\nu$ solely depends on the convexity constant $\mu$, streaming batch size $C_\rho$, and the dependency quantities imposed by Assumption 1-p, e.g., if the dependence $D_\nu$ is so large that $\mu_\nu$ is no longer positive, then we should take streaming batch size $C_\rho$ large enough such that $\mu_\nu$ becomes positive again; this is illustrated in sections 4.1.2 and 4.1.3 for ARCH models (Werge & Wintenberger, 2022). Another way to ensure convexity is to take increasing streaming batches, i.e., streaming rate $\rho > 0$, which is more robust since we do not have to find $C_\rho$ such that $\mu_\nu$ remains positive. However, combining both more ideal, as we ensure convexity while a large $C_\rho$ reduces variance.

**Variance reduction from larger streaming batches.** Not surprisingly, larger streaming batches $C_\rho$ have a variance-reducing effect, e.g., see the illustrations in section 4. Nevertheless, (8) explicitly shows the variance-reducing effect in each term, which can help us better understand how to optimally tune the hyperparameters.

**Decay of the initial conditions.** The initial conditions in the first term of (8) will be forgotten sub-exponentially fast; an explicit version of this term can be found in appendix A. The last term of (8) can be seen as the noise term, which depends on the gradient noise (Assumption 3-p). For $\alpha - \rho\beta \in (1/2, 1)$, the noise term is $\mathcal{O}(N_t^{-(\rho(2\sigma-\beta)+\alpha)/(1+\rho)})$, e.g., if $\alpha = 2/3$, $\beta = 1/3$, and $\sigma = 1/2$, we have $\mathcal{O}(N_t^{-2/3})$ for any streaming rate $\rho \in [0, 1)$; this is illustrated in Godichon-Baggioni et al. (2021). In unbiased cases ($B_\nu = 0$), the noise term would also be the asymptotic term. In addition, the noise/asymptotic term is positively affected by large streaming batches $C_\rho$ when $\alpha + \beta < 2\sigma$, e.g., see section 4.

**Behavior for bias $B_\nu > 0$.** The second term of (8) is a pure bias term determined by bias $B_\nu$, level of dependence $\nu$, and (dependence penalised) convexity constant $\mu_\nu$. It is noteworthy that the bias term is independent of the learning rate but depends on the streaming batch $C_\rho$ and streaming rate $\rho$. The dependence term is $\mathcal{O}(N_t^{-2\rho\nu/(1+\rho)})$, which requires $\rho$ positive since $\nu \in (0, \infty)$, e.g., to obtain $\mathcal{O}(N_t^{-1/2})$, we would need $\rho = 1$ and $\nu = 1/2$. It is surprising that Theorem 1 allows both long- and short-range dependence. Indeed, long-range dependence leads to slow convergence (slower than $\mathcal{O}(N_t^{-1/2})$) but a positive streaming rate $\rho$ will break long-range dependence. To conclude, increasing streaming batches ensure convexity and breaks long- and short-term dependence, by which we can obtain $\delta_t = \mathcal{O}(\max\{\mathbb{1}_{\{B_\nu>0\}} N_t^{-2\rho\nu/(1+\rho)}, N_t^{-(\rho(2\sigma-\beta)+\alpha)/(1+\rho)}\})$.

### 3.3 Averaged stochastic streaming gradients

In what follows, we consider the averaging estimate $\bar{\theta}_n$ given in (4) with $(\theta_t)$ following the SSG estimate in (2) or the PSSG estimate in (3). Some additional smoothness assumptions on $L$ is needed for the averaging estimate:

**Assumption 4** ($C_\nabla$- and $C'_\nabla$-smoothness)**.** *The objective function $L$ have $C_\nabla$-Lipschitz continuous gradients around $\theta^*$, i.e., there exists a constant $C_\nabla > 0$ such that $\forall \theta \in \Theta$,*

$$\|\nabla_\theta L(\theta) - \nabla_\theta L(\theta^*)\| \leq C_\nabla \|\theta - \theta^*\|. \tag{10}$$

*Next, the Hessian of $L$ is $C'_\nabla$-Lipschitz-continuous around $\theta^*$, that is, there exists a constant $C'_\nabla \geq 0$ such that $\forall \theta \in \Theta$,*

$$\|\nabla^2_\theta L(\theta) - \nabla^2_\theta L(\theta^*)\| \leq C'_\nabla \|\theta - \theta^*\|. \tag{11}$$

As discussed in Bottou et al. (2018), Assumption 4 ensures that $\nabla_\theta L$ does not vary arbitrarily, making the gradient $\nabla_\theta L$ a useful indicator on how to decrease $L$. For deterministic optimization, convergence rates for smooth optimization are better than for non-smooth optimization, but for SO, smoothness only leads to improvements in constants (Nesterov et al., 2018).

Next, in continuation of Assumption 3-p with $\sigma_t = C_\sigma n_t^{-\sigma}$ for $\sigma \in [0, 1/2]$, we make the following assumption about covariance of $(\nabla_\theta l_t(\theta^*))$, which we interpret as the sequence of score vectors with respect to the parameter vector $\theta^*$:

**Assumption 5** (Covariance of the scores)**.** *There exists a non-negative self-adjoint operator* $\Sigma$ *such that* $\forall t \geq 1$, $n_t^{2\sigma}\mathbb{E}[\nabla_\theta l_t(\theta^*)\nabla_\theta l_t(\theta^*)^\top] \preceq \Sigma + \Sigma_t$, *where* $\Sigma_t$ *is a positive symmetric matrix with* $Tr(\Sigma_t) = C'_\sigma n_t^{-2\sigma'}$, $C'_\sigma \geq 0$, *and* $\sigma' \in (0, 1/2]$.

Remark that in the independent or some unbiased cases, such as in section 4.1.1, Assumption 5 is verified with $\sigma = 1/2$ and $C'_\sigma = 0$ (Godichon-Baggioni et al., 2021). The short-range dependence cases is when $\sigma = 1/2$ (and $C'_\sigma > 0$), as in section 4.1.1, whereas, the long-range dependence case is for $\sigma < 1/2$ (and $C'_\sigma > 0$). Assumption 5 allows us to obtain leading term $\Lambda/N_t$ with $\Lambda = \mathrm{Tr}(\nabla^2_\theta L(\theta^*)^{-1}\Sigma\nabla^2_\theta L(\theta^*)^{-1})$. Also, Assumption 5 makes it possible to acquire the Cramer-Rao lower bound in the (unbiased) i.i.d. case, i.e., a non-asymptotic bound of the averaged estimate $(\bar{\theta}_t)$ of the form $\mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2] \leq \mathcal{O}(\Lambda N_t^{-1}) + \mathcal{O}(N_t^{-b})$ with $b > 1$.

To consider the averaged estimate $\bar{\theta}_n$ (4) of the projected estimate PSSG from (3), called APSSG, an additional assumption is necessary to avoid to calculate the sixth-order moment; we make the unnecessary assumption that $(\nabla_\theta l_t)$ is uniformly bounded but the derivation of the six-order moment can be found in Godichon-Baggioni (2016).

**Assumption 6** ($G_\Theta$-bounded gradients)**.** *Let* $D_\Theta = \inf_{\theta\in\partial\Theta}\|\theta - \theta^*\| > 0$ *with* $\partial\Theta$ *denoting the boundary of* $\Theta$. *Moreover, there exists* $G_\Theta > 0$ *such that* $\forall t \geq 1$, $\sup_{\theta\in\Theta}\|\nabla_\theta l_t(\theta)\|^2 \leq G^2_\Theta$ *a.s.*

**Theorem 2** (ASSG/PASSG)**.** *Let* $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ *with* $\bar{\theta}_n$ *given by* (4)*, where* $(\theta_t)$ *either follows the recursion in* (2) *or* (3). *Suppose Assumptions 1-p to 5 hold for* $p = 4$. *In addition, Assumption 6 must hold if* $(\theta_t)$ *follows the recursion in* (3). *If* $\mu_\nu = \mu - \mathbb{1}_{\{\rho=0\}}2D_\nu C_\rho^{-\nu} > 0$, *then for* $\alpha - \rho\beta \in (1/2, 1)$, *we have*

$$\bar{\delta}_t^{\frac{1}{2}} \leq \frac{\Lambda^{\frac{1}{2}}}{N_t^{\frac{1}{2}}}\mathbb{1}_{\{\sigma=1/2\}} + \frac{2^{\frac{1}{2}}\Lambda^{\frac{1}{2}}C_\rho^{\frac{1-2\sigma}{2(1+\rho)}}}{N_t^{\frac{1+2\rho\sigma}{2(1+\rho)}}}\mathbb{1}_{\{\sigma<1/2\}} + \frac{2^{\frac{1}{2}}C_\sigma'^{\frac{1}{2}}C_\rho^{\frac{1-2(\sigma+\sigma')}{2(1+\rho)}}}{\mu N_t^{\frac{1+2\rho(\sigma+\sigma')}{2(1+\rho)}}} \tag{12}$$

$$+\tilde{\mathcal{O}}\left(\max\left\{N_t^{-\frac{2+\rho(2\sigma+\beta)-\alpha}{2(1+\rho)}}, N_t^{-\frac{\rho(2\sigma-\beta)+\alpha}{1+\rho}}\right\}\right) + \mathbb{1}_{\{B_\nu\neq 0\}}\Psi_t, \tag{13}$$

*with*

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

*where* $\delta = \mathbb{1}_{\{B_\nu=0\}}(\rho(2\sigma-\beta)+\alpha) + \mathbb{1}_{\{B_\nu\neq 0\}}\min\{\rho(2\sigma-\beta)+\alpha, 2\rho\nu\}$. *An explicit version of this bound is given in appendix A.*

**Related work.** Note that the bound in Theorem 2 is on the root mean square error, but our discussions are about $\bar{\delta}_t$ (without the root). Similarly to the unbiased i.i.d. case (Godichon-Baggioni et al., 2021), the leading term of $\bar{\delta}_t$ in (12) and (13) is $\Lambda/N_t$, which obtain the (asymptotically optimal) Cramer-Rao lower bound (Murata & Amari, 1999). Each term of (12) is a direct consequence of Assumption 5 and they are all independent of the choice of learning rate $(\gamma_t)$. Moreover, as discussed in Gadat & Panloup (2023), the bound of $\bar{\delta}_t$ can be seen as a bias-variance decomposition between the leading terms (12) and the remaining terms in (13).

**Ensuring convexity through non-decreasing streaming batches.** As for Theorem 1, the positivity of $\mu_\nu$ is essential for all terms in (13) even if it does not appear directly. In case of lack of convexity $\mu$ or high levels of dependence constant $D_\nu$, we can only ensure convergence by increasing $C_\rho$, i.e., ensuring positivity of $\mu_\nu$; this is illustrated in sections 4.1.2 and 4.1.3 for ARCH models.

**Accelerated decay.** By averaging it is possible (in some specific cases) to achieve the leading term $\Lambda/N_t$, which is known to obtain the asymptotically optimal Cramer-Rao bound in the i.i.d. case (Godichon-Baggioni et al., 2021). Thus, we could obtain the optimal and incorrigible rate of $\bar{\delta}_t = \mathcal{O}(N_t^{-1})$. This is always achieved in the unbiased case (i.e., $B_\nu = 0$) with $\sigma = 1/2$, even under short-range dependence. More specifically, (12) and (13) simplifies significantly in the case of $\sigma = 1/2$: the last two terms of (12) become negligible as $\sigma' > 0$. The first term of (13) decay at the rate $\mathcal{O}(N_t^{-(2+\rho(2\sigma+\beta)-\alpha)/(1+\rho)})$ or $\mathcal{O}(N_t^{-2(\rho(2\sigma-\beta)+\alpha)/(1+\rho)})$, which suggests choosing $\alpha, \beta$ such that $\alpha + \rho(2\sigma/3 - \beta) = 2/3$, e.g., $\alpha = 2/3$, $\beta = 1/3$ and $\sigma = 1/2$ yields a decay of $\mathcal{O}(N_t^{-4/3})$ for any $\rho$. Thus, the first term of (13) robustly achieve $\mathcal{O}(N_t^{-4/3})$ for any streaming rate

$\rho$ by setting $\alpha = 2/3$ and $\beta = 1/3$ if $\sigma = 1/2$. Likewise, the last term of (13) is $\mathcal{O}(N_t^{-\rho(1/2+\nu)/(1+\rho)})$ for any $\rho$ when $\alpha = 2/3$ and $\beta = 1/3$. To summarize, Theorem 2 with $\sigma = 1/2$ simplifies into the bound,

$$\bar{\delta}_t^{\frac{1}{2}} \leq \frac{\Lambda^{\frac{1}{2}}}{N_t^{\frac{1}{2}}} + \tilde{\mathcal{O}}\left(N_t^{-\frac{2}{3}}\right) + \mathbb{1}_{\{B_\nu \neq 0\}} \tilde{\mathcal{O}}\left(N_t^{-\frac{\rho(1/2+\nu)}{2(1+\rho)}}\right), \tag{14}$$

using $\alpha = 2/3$ and $\beta = 1/3$.

**Variance reduction from larger streaming batches.** Taking $\beta > 0$ would damage the variance reduction effect from having $C_\rho$ large (e.g., see discussion after Theorem 1). Thus, there is a trade-off between accelerating the convergence by taking $\beta > 0$ or taking $\beta = 0$ to favor from variance reduction. In practice, an immediate choice would be to take $\beta = 0$, but if the data contains a low amount of noise or the model is robust, it can be advantageous to raise $\beta$ to improve convergence (Godichon-Baggioni et al., 2021).

**Behavior for $B_\nu$.** The influence of $B_\nu$ is exclusively contained in $\Psi_t$, with the exception of the second term of (13) via the decay rate $\delta$. Also, increasing $\rho$ will always diminish the bad influence of this bias term. Surprisingly, $\Psi_t \to 0$ as $t \to \infty$ for any $\nu$, but long-range dependence is excluded if we wish to obtain the desired rate of $\bar{\delta}_t = \mathcal{O}(N^{-1})$. However, it does not seem to have any major influence in our experiments in section 4.

## 4 Experiments

### 4.1 Synthetic data

A way to illustrate our findings is by use of classical methods that aim to model and predict an underlying sequence of real-valued time-series $(X_s)$; here $s$ is short notation for indexing the sequence of observations, $(X_{N_t}, X_{N_t-1}, \dots, X_{N_t-n_t} \equiv X_{N_{t-1}}, X_{N_{t-1}-1}, \dots)$ with $N_t = \sum_{i=1}^t n_t$. The AutoRegressive (AR), Moving-Average (MA), and AutoRegressive Moving-Average (ARMA) models are the most well-known models for time-series (Brockwell & Davis, 2009; Box et al., 2015; Hamilton, 2020). The standard time-series analysis often relies on independence and constant noise, but it can be relaxed by, e.g., the AutoRegressive Conditional Heteroskedasticity (ARCH) model (Engle, 1982). Online learning algorithms of (both stationary and non-stationary) dependent time-series have been studied in Agarwal & Duchi (2012); Anava et al. (2013); Wintenberger (2021).

#### 4.1.1 AR model

A process $(X_s)$ is called a (zero-mean) AR(1) process, if there exists real-valued parameter $\theta$ such that $X_s = \theta X_{s-1} + \epsilon_s$, where $(\epsilon_s)$ is weak white noise with zero mean and variance $\sigma_\epsilon^2$. To illustrate the versatility of our results, we construct some (heavy-tailed) noise processes with long-range dependence: the noisiness is integrated using a Student's $t$-distribution with degrees of freedom above four, denoted by $(z_s)$. The long-range dependence is incorporated by multiplying $(z_s)$ with the fractional Gaussian noise $G_s(H) = B_{s+1}(H) - B_s(H)$, where $(B_s(H))$ is a fractional Brownian motion with Hurst index $H \in (0, 1)$. $(G_s(H))$ can also be seen as a (zero-mean) Gaussian process with stationary and self-similar increments (Nualart, 2006).

**Well-specified case.** Consider the well-specified case, in which, we estimate an AR(1) model from the underlying stationary AR(1) process $X_s = \theta^* X_{s-1} + \epsilon_s$ with $|\theta^*| < 1$. The squared loss function $l_t(\theta) = n_t^{-1} \sum_{i=1}^{n_t} (X_{N_{t-1}+i} - \theta X_{N_{t-1}+i-1})^2$ with gradient $\nabla_\theta l_t(\theta) = -2n_t^{-1} \sum_{i=1}^{n_t} X_{N_{t-1}+i-1}(X_{N_{t-1}+i} - \theta X_{N_{t-1}+i-1})$. Thus, the objective function is $L(\theta) = (\sigma_\epsilon^2(\theta^* - \theta)^2)/(1-(\theta^*)^2) + \sigma_\epsilon^2$, as $\mathbb{E}[X_s] = 0$ and $\mathbb{E}[X_s^2] = \sigma_\epsilon^2/(1-(\theta^*)^2)$, yielding $\nabla_\theta L(\theta) = 2\sigma_\epsilon^2(\theta - \theta^*)/(1-(\theta^*)^2)$. Assumption 1-p can for $p = 2$, we written as

$$\mathbb{E}[\|\mathbb{E}[\nabla_\theta l_t(\theta)|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta)\|^2] = \frac{4(\theta - \theta^*)^2(1-(\theta^*)^{2n_t})^2\sigma_\epsilon^2}{(1-(\theta^*)^2)^4 n_t^2}\left(\sigma_\epsilon^2 + \frac{1}{1-(\theta^*)^2}\right),$$

meaning that Assumption 1-p is verified if $(X_s)$ has bounded moments; this is fulfilled by the natural constraint that $|\theta^*| < 1$.[3] Thus, we can deduce that $D_\nu > 0$, $B_\nu = 0$, and $\nu_t$ is $\mathcal{O}(n_t^{-1})$. The remaining assumptions can

[3]The verification is available in a longer version in appendix B.

be verified in the same way, in particular, Assumptions 2-p and 3-p is satisfied with $\kappa_t$ and $\sigma_t$ is $\mathcal{O}(n_t^{-1/2})$, Assumption 4 with $C_\nabla = 2\sigma_\epsilon^2/(1-(\theta^*)^2)$ and $C'_\nabla = 0$, and Assumption 5 with $\Sigma = 4\sigma_\epsilon^4/(1-(\theta^*)^2)$ and $\Sigma_t = 0$. Furthermore, for an AR(1) process $X_s$ constructed using the noise process $\epsilon_s = \sqrt{G_s(H)}z_s$ with Hurst index $H \geq 1/2$, one can verify that $\nu_t^4, \kappa_t^4, \sigma_t^4$ is $\mathcal{O}(n_t^{H-1})$ in Assumptions 1-p to 3-p using the self-similarty property (Nourdin, 2012).

**Misspecified case.** Next, assume that the underlying data generating process follows the MA(1)-process, $X_s = \epsilon_s + \phi^* \epsilon_{s-1}$, with $\phi^* \in \mathbb{R}$. The misspecification error of fitting an AR(1) model to a MA(1) process can be found by minimizing $L(\theta) = \mathbb{E}[(X_s - \theta X_{s-1})^2] = \sigma_\epsilon^2(1 + (\phi^* - \theta)^2 + \theta^2(\phi^*)^2)$, where $\nabla_\theta L(\theta) = 2(\theta - \phi^*)\sigma_\epsilon^2 + 2\theta(\phi^*)^2\sigma_\epsilon^2$. Thus, as $\theta^* = \arg\min_\theta L(\theta) \equiv \arg\min_\theta(\phi^* - \theta)^2 + \theta^2(\phi^*)^2$ is a strictly convex function in $\theta$, we have $\nabla_\theta L(\theta) = 0 \Leftrightarrow 2(\theta - \phi^*) + 2\theta(\phi^*)^2 = 0 \Leftrightarrow 2\theta(1 + (\phi^*)^2) = 2\phi^* \Leftrightarrow \theta = \phi^*/(1 + (\phi^*)^2)$. This means for any $\phi^* \in \mathbb{R}$ then $\theta \in (-1/2, 1/2)$. With this in mind, we can conduct our study of fitting an AR(1) model to the MA(1) process with $\phi^*$ drawn randomly from $\mathbb{R}$ (figure 1b). Furthermore, this reparametrization trick can be used to verify Assumption 1-p in the following way:

$$\mathbb{E}[\|\mathbb{E}[\nabla_\theta l_t(\theta)|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta)\|^2] = \frac{4(\theta - \theta^*)^2}{n_t^2} f_{\phi^*}(\epsilon_{N_{t-1}}),$$

where $f_{\phi^*}(\epsilon_{N_{t-1}})$ is finite function depending on the moments of $(\epsilon_{N_{t-1}})$ and $\phi^*$.[4] Hence, we have $D_\nu > 0$ and $B_\nu = 0$ with $\nu_t$ being $\mathcal{O}(n_t^{-1})$. Similarly, it can be verified that $\kappa_t$ and $\sigma_t$ are $\mathcal{O}(n_t^{-1/2})$ by use of the reparametrization trick (Assumptions 2-p and 3-p).

#### 4.1.2 ARCH model

A key element of time series analysis is modeling heteroscedasticity of the conditional variance, e.g., volatility clustering in financial time-series; ARCH models are some of the most well-known models that incorporate this feature. A process $(\epsilon_s)$ is called an ARCH(1) process with parameters $\alpha_0$ and $\alpha_1$ if it satisfies

$$\begin{cases} \epsilon_s = \sigma_s z_s, \\ \sigma_s^2 = \alpha_0 + \alpha_1 \epsilon_{s-1}^2, \end{cases} \tag{15}$$

where $\alpha_0 > 0$ and $\alpha_1 \geq 0$ ensures the non-negativity of the conditional variance process $(\sigma_s^2)$, and the innovations $(z_s)$ is white noise. We employ the quasi-maximum likelihood procedure for the statistical inference as outlined in Werge & Wintenberger (2022); the quasi likelihood losses is given by $l_s(\theta) = 2^{-1}(\epsilon_s^2/\sigma_s^2(\theta) + \log(\sigma_s^2(\theta))$ with first-order derivative

$$\nabla_\theta l_s(\theta) = \nabla_\theta \sigma_s^2(\theta) \left( \frac{\sigma_s^2(\theta) - \epsilon_s^2}{2\sigma_s^4(\theta)} \right),$$

where $\nabla_\theta \sigma_s^2(\theta) = (1, \epsilon_{s-1}^2)^T$. Verification of Assumptions 1-p to 3-p can be done using mixing conditions; Francq & Zakoian (2019, Theorem 3.5) showed that stationary ARCH processes are geometrically $\beta$-mixing, which implies $\alpha$-mixing as well. Observe that the loss function $(l_s)$ itself is not strongly convex but only the objective function $L$ may be strongly convex; convexity conditions of ARCH processes was investigated in Wintenberger (2021). This makes the parameters challenging to estimate in empirical applications as the optimization algorithms can quickly fail or converge to irregular solutions. There are different ways to overcome lack of convexity: first, projecting the estimates such that the (conditional) variance process $(\sigma_s^2)$ stays away from zero (and close to the unconditional variance). Second, in the specific example of ARCH model, one could also recover convexity by implementing variance targeting techniques; an example using Generalized ARCH (GARCH) models can be found in Werge & Wintenberger (2022). To simplify our analysis we consider stationary ARCH(1) processes, where we fix $\alpha_0$ at 1 and initialize it at 1/2.

[4]The verification is available in a longer version in appendix B.

#### 4.1.3 AR-ARCH model

We complete our experiments by considering an AR models with ARCH noise: the process $(X_s)$ is called an AR(1)-ARCH(1) process with parameters $\theta$, $\alpha_0$ and $\alpha_1$ if it satisfies

$$\begin{cases} X_s = \theta X_{s-1} + \epsilon_s, \\ \epsilon_s = \sigma_s z_s, \\ \sigma_s^2 = \alpha_0 + \alpha_1 \epsilon_{s-1}^2. \end{cases} \tag{16}$$

where the innovations $(z_s)$ is weak white noise. The statistical inference of this model is done using the squared loss for the AR-part and the QMLE for the ARCH part, e.g., see sections 4.1.1 and 4.1.2. Assumptions 1-p to 3-p can be verified by Doukhan (1994, Proposition 6), which showed that ARMA-ARCH processes are $\beta$-mixing.

#### 4.1.4 Discussion

Our experiment measures the performance by the mean quadratic error $\mathbb{E}[\|\theta_{N_t} - \theta^*\|^2]$ over one thousand replications with $\theta_0$ and $\theta^*$ drawn randomly according to the models' specifications. We have omitted to project our estimates as this would hide the errors we want to explore, namely the errors from the dependence and the lack of convexity. It should be noted that averaging over several replications gives a reduction in variability, that mainly benefits the SSG. The experiments will demonstrate how the choice of $C_\rho$ and $\rho$ affects the dependence $D_\nu$, bias $B_\nu$, and the (dependence) penalised convexity constant $\mu_\nu$. To compare different data streams through the selection of $C_\rho$ and $\rho$, we fix the parameters $C_\gamma = 1$, $\alpha = 2/3$, and $\beta = 0$.

The experiments described in sections 4.1.1 to 4.1.3 can be found in figure 1; here $\{C_\rho = 1, \rho = 0\}$ corresponds to the classical SG descent and its (Polyak-Ruppert) average estimate, $\{C_\rho = 64, \rho = 0\}$ is a mini-batch SSG/ASSG, and $\{C_\rho = 64, \rho = 1/2\}$ is an increasing SSG/ASSG with initial batch size of $C_\rho = 64$.

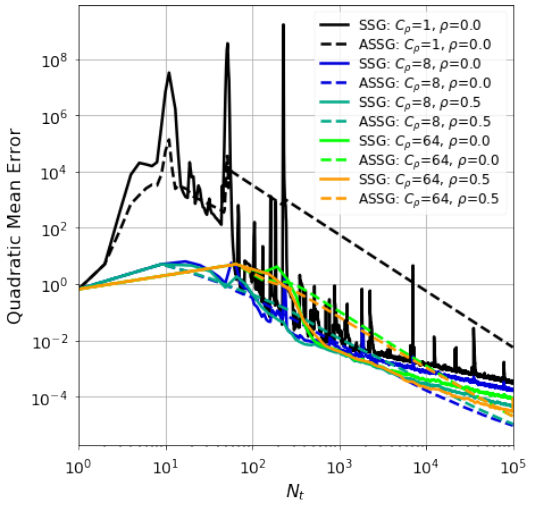
First consider the AR (well- and misspecified) cases in figures 1a and 1b; these figures shows the results for long-range dependent white noise processes with Hurst index $H = 3/4$. Each pair of data streams converges, but it is clear that the traditional SG method experiences a large amount of noise initially, particularly affecting the average estimate period but not its decay rate.[5] But despite this, we still achieve better convergence for the ASSG method. Both methods show a noticeable reduction in variance when $C_\rho$ increases, which is particularly beneficial in the beginning. Nevertheless, too large streaming batch sizes $C_\rho$ may hinder the convergence as this leads to too few iterations. Furthermore, figures 1a and 1b indicates an improved decay of the SSG methods when the streaming rate $\rho$ is increased. Conversely, improvements to the ASSG method do not occur as we do not exploit the potential of using more observations through parameter $\beta$ parameter, which could accelerate convergence, e.g., see section 4.2. It is surprising that we do not see any effect from the long-range dependent white noise processes, but this seems to be an artifact effect in the proof as we need fourth-order moments (i.e., Assumptions 1-p to 3-p with $p = 4$). Therefore, we conjecture that only second-order moment properties are responsible for the behavior of our simulations and that $\sigma = 1/2$ even for long-range dependent white noise processes, as proved in section 4.1.1.

In figures 1c and 1d, we have the experiments for the stationary ARCH(1) models, with and without an AR-part, respectively, as outlined in sections 4.1.2 and 4.1.3. These figures illustrate the lack of convexity when using small streaming batch sizes $C_\rho$, e.g., the traditional SG descent and its average estimate, $\{C_\rho = 1, \rho = 0\}$ diverges. Remark that the lack of convexity is expressed through the lack positively of $\mu_\nu$, which only larger streaming batch sizes $C_\rho$ can counteract. Moreover, the traditional SG descent $\{C_\rho = 1, \rho = 0\}$ is omitted in figure 1d due to lack of convexity. Figure 1d shows that large ($C_\rho = 64$) and non-decreasing ($\rho \geq 0$) streaming batches can converge under difficult settings.

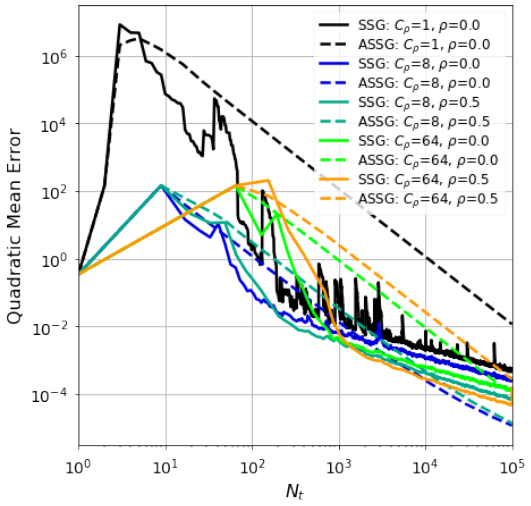
[5] A modification of our average estimate to a weighted average version could improve convergence as it could limit the effect of poor initializations Mokkadem & Pelletier (2011); Boyer & Godichon-Baggioni (2022).

Figure 1: Simulation of various data streams $n_t = C_\rho t^\rho$. See section 4.1 for details.

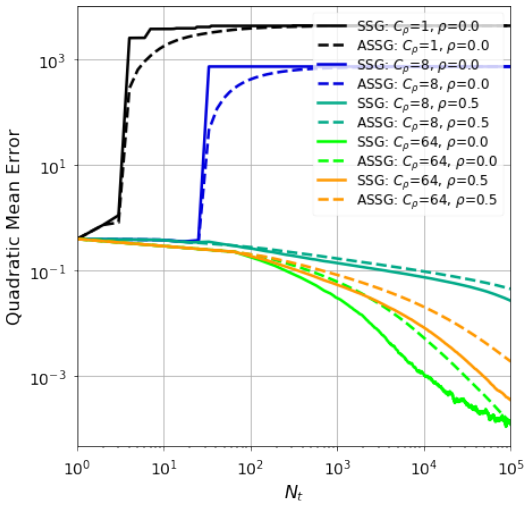
(a) AR(1): well-specified case. See section 4.1.1 for details.



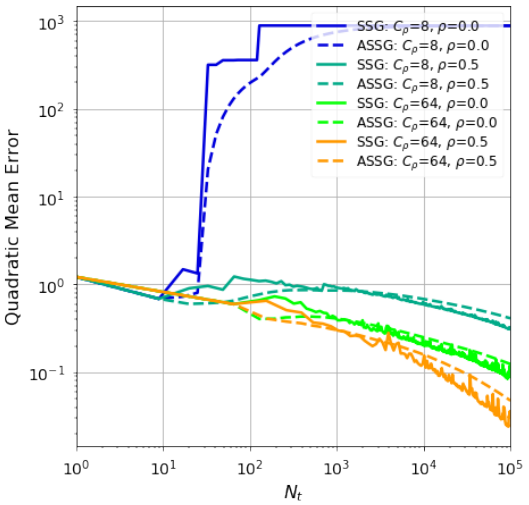

(b) AR(1): misspecified case. See section 4.1.1 for details.




(c) ARCH(1). See section 4.1.2 for details.




(d) AR(1)-ARCH(1). See section 4.1.3 for details.




### 4.2 Historical hourly weather data

To illustrate our methodology on real-life time-dependent streaming data, we consider some historical hourly weather data.[6] The dataset contains around five years (roughly 45000 data points) of high temporal resolution hourly measurements over various weather attributes, such as temperature, humidity, and air pressure. These measurements are available for thirty-six cities, i.e., the dimension $d = 36$. In our study, we consider the hourly temperature measurements, which we filter for monthly and annual seasonality by subtracting the monthly and annual averages.

#### 4.2.1 Geometric median

When the data are noisy, robust estimators such as the geometric median are preferred; Haldane (1948) introduced the geometric median as a generalization of the real median. The efficiency of the geometric median

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

makes it suitable for handling high-dimensional streaming data (Cardot et al., 2013; Godichon-Baggioni, 2016). We estimate the geometric median of $X_s \in \mathbb{R}^d$ by minimizing the objective $L(\theta) = \mathbb{E}[\|X_s - \theta\| - \|X_s\|]$ using the stochastic gradient estimates $\nabla_\theta l_t(\theta) = -n_t^{-1} \sum_{i=1}^{n_t} (X_{N_{t-1}+i} - \theta)/\|X_{N_{t-1}+i} - \theta\|$; properties such as existence, uniqueness, and robustness (breakdown point) can be found in Kemperman (1987); Gervini (2008). We omit to project our estimates as this would hide the errors we want to explore. Instead of projecting the estimates, one could adapt the proof of Gadat & Panloup (2023) to a streaming setting. Otherwise, if $X_s$ is bounded, one can adapt Cardot et al. (2012) to the streaming setting showing that the streaming estimates are bounded.

#### 4.2.2 Discussion

To measure the performance, we use the mean quadratic error of the parameter estimates over one-hundred replications, given by $\mathbb{E}[\|\theta_{N_t} - \theta^*\|^2]$; here we compare our estimates to the geometric median estimate calculated by the Weiszfeld's algorithm (Weiszfeld & Plastria, 2009). Suppose $(X_s)$ is standard Gaussian centered at $(\theta_i)_{1 \leq i \leq d}$ with $\theta_i$ taken randomly in the range $[-d, d]$. Moreover, following the reasoning of Cardot et al. (2013), we set $C_\gamma = \sqrt{d}$, and let $\alpha = 2/3$.

Figure 2 shows the results of the geometric median estimated in the same way as described in section 4.2.1. Although the geometric median is a robust metric, we still see a considerable amount of fluctuations in figure 2, which comes from the time-dependency and the noise in the weather measurements. Figure 2a shows that it is essential to use a mini-batch $C_\rho$ of a certain size to stabilize the optimization, i.e., ensure convexity through larger streaming batches $C_\rho$. But to achieve reasonable convergence, we need to have increasing streaming batches, i.e., positive streaming rates $\rho > 0$; this is illustrated in figures 2b and 2c. These figures shows an increase in decay of the SSG when the streaming rate $\rho$ increase. However, the lack of convergence improvements in figure 2c comes from $\beta = 0$, which means we do not exploit the potential of using more observations to accelerate convergence, e.g., see Godichon-Baggioni et al. (2021) for a discussion in the unbiased i.i.d. case. As discussed after Theorem 2, one example of this could be achieved by setting $\alpha = 2/3$ and $\beta = 1/3$. As shown in figure 2d, we can achieve this acceleration by simply taking $\beta = 1/3$. In addition, $\beta = 1/3$ provides optimal convergence robust to any streaming rate $\rho$. Choosing a proper $\beta > 0$ is particularly important when $C_\rho$ is large, as robustness is an integral part of the geometric median method. Most surprising is that we can achieve excellent convergence with a final error of only $10^{-5}$ by combining increasing streaming batches with averaging, e.g., see figure 2d with $C_\rho = 64$, $\rho > 0$ and $\beta = 1/3$. Following the discussion in section 4.1.4, together with these real-life experiments, we suspect that the sequence of scores $(\nabla_\theta l_t(\theta^*))$ is a martingale difference sequence that generalizes beyond our examples. In particular, our findings indicate that $\sigma = 1/2$ even under long-range dependence. Thus, the complexity of Theorem 2 seems to be an artifact of the proof, which relies on Assumptions 1-p to 3-p with $p = 4$.

## 5 Conclusions

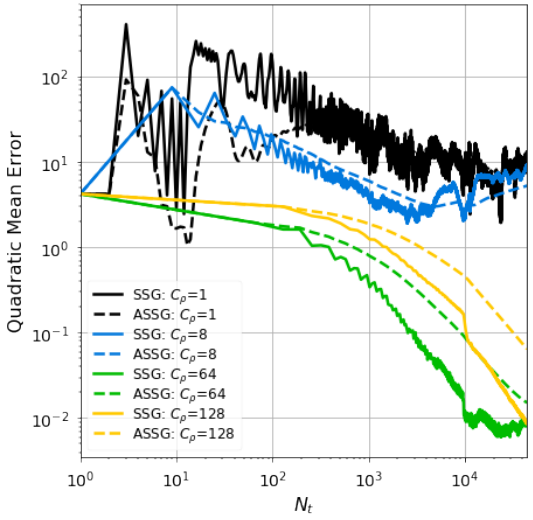
This study examined the robustness and convergence guarantees of SG-based methods under different settings, covering a broad range of applications with dependence and biased gradients. The non-asymptotic convergence rates of SG-based algorithms were explored, and theoretical results were used to develop heuristics that link the level of dependency and convexity to the model parameters. These heuristics provided new insights into determining optimal learning rates, which can increase the stability of SG-based methods.

The findings demonstrate that SG-based methods can break short- and long-term dependence by using non-decreasing batch sizes, which counteracts the dependency structures. Specifically, mini-batch is essential to break dependence and ensure convexity, and convergence can be accelerated by simultaneously averaging. The experimentation results validate these conclusions, suggesting large non-decreasing mini-batches for highly dependent data sources.

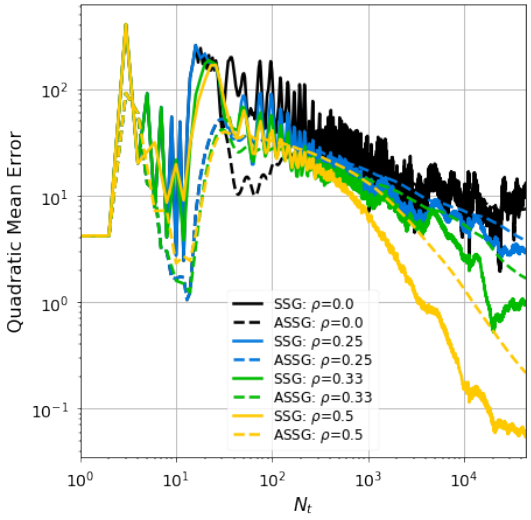
Furthermore, in large-scale learning problems with dependence, noisy variables, and lack of convexity, we now know how to accelerate convergence while reducing variance through the learning rate and the treatment pattern of the data. Overall, this study offers valuable insights into the use of SG-based methods in complex learning problems and provides practical guidelines for optimizing convergence rates.

Figure 2: Geometric median for various data streams $n_t = C_\rho t^\rho$. See section 4.2 for details.

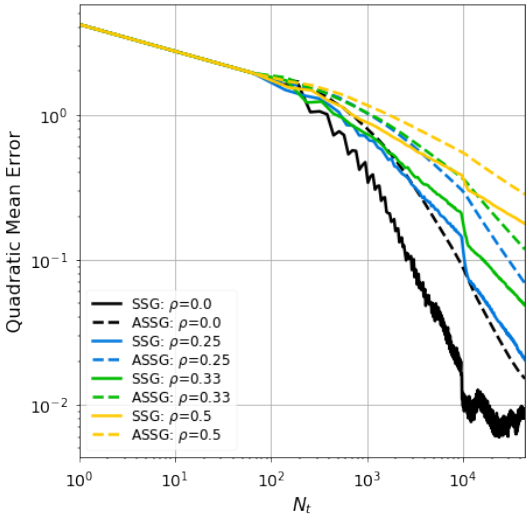
(a) Varying $C_\rho$, $\rho = 0$, $\beta = 0$



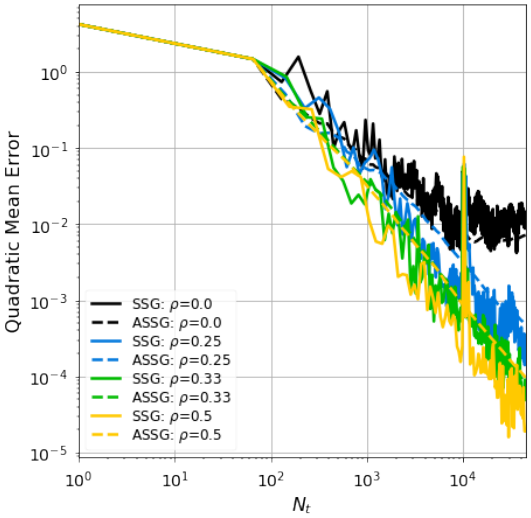

(b) Varying $\rho$, $C_\rho = 1$, $\beta = 0$




(c) Varying $\rho$, $C_\rho = 64$, $\beta = 0$




(d) Varying $\rho$, $C_\rho = 64$, $\beta = 1/3$




**Future perspectives.** There are several ways to expand our work about stochastic streaming algorithms: (a) we can extend our analysis to include streaming batches of any size (and not as a function of streaming batch size $C_\rho$ and streaming rates $\rho$), e.g., Godichon-Baggioni et al. (2021) discuss random streaming batches with negative and positive drift. (b) an extension to non-strongly convex objectives could be advantageous as it will provide more insight into how we should choose our learning rates (Bach & Moulines, 2013; Nemirovski et al., 2009; Necoara et al., 2019; Gadat & Panloup, 2023). (c) learning rates should be made adaptive so they are robust to poor initialization and require less tuning; an adaptive learning rate is essential for practitioners as it builds a form of universality across applications, e.g., see Duchi et al. (2011); Kingma & Ba (2014). (d) non-parametric analysis could improve our theoretical results for large values of $d$. (e) we have focused on results in quadratic mean but another way to strengthen our non-asymptotic guarantees could be high probability bounds (Durmus et al., 2021; 2022); for any $\delta \in (0, 1)$, we could obtain bounds on the sequence $\{\|\theta_t - \theta^*\| : t \in \mathbb{N}\}$ that holds with probability at least $1 - \delta$.

## References


Yaser S Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from data*, volume 4. AMLBook New York, 2012.

Alekh Agarwal and John C Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2012.

Ahmad Ajalloeian and Sebastian U Stich. On the convergence of sgd with biased gradients. *arXiv preprint arXiv:2008.00051*, 2020.

Oren Anava, Elad Hazan, Shie Mannor, and Ohad Shamir. Online learning for time series prediction. In *Conference on learning theory*, pp. 172–184. PMLR, 2013.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). *Advances in neural information processing systems*, 26, 2013.

Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.

Dimitri Bertsekas. *Nonlinear Programming*, volume 3. Athena Scientific, 2016.

Léon Bottou and Yann Le Cun. Large scale online learning. *Advances in neural information processing systems*, 16, 2003.

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

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Claire Boyer and Antoine Godichon-Baggioni. On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. *Computational Optimization and Applications*, pp. 1–52, 2022.

Richard C Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability surveys*, 2:107–144, 2005.

Peter J Brockwell and Richard A Davis. *Time series: theory and methods*. Springer Science & Business Media, 2009.

Richard H Byrd, Samantha L Hansen, Jorge Nocedal, and Yoram Singer. A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.

Hervé Cardot, Peggy Cénac, and Jean-Marie Monnez. A fast and recursive algorithm for clustering large datasets with k-medians. *Computational Statistics & Data Analysis*, 56(6):1434–1449, 2012.

Hervé Cardot, Peggy Cénac, and Pierre-André Zitt. Efficient and fast estimation of the geometric median in hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19(1):18–43, 2013.

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Jie Chen and Ronny Luss. Stochastic gradient descent with biased but consistent gradient estimators. *arXiv preprint arXiv:1807.11880*, 2018.

Alexandre d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, 2008.

Olivier Devolder et al. Stochastic first order methods in smooth convex optimization. Technical report, CORE, 2011.

Paul Doukhan. Mixing. In *Mixing*, pp. 15–23. Springer, 1994.

Paul Doukhan. *Mixing: properties and examples*, volume 85. Springer Science & Business Media, 2012.

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

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, Kevin Scaman, and Hoi-To Wai. Tight high probability bounds for linear stochastic approximation with fixed stepsize. *Advances in Neural Information Processing Systems*, 34:30063–30074, 2021.

Alain Durmus, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Finite-time high-probability bounds for polyak-ruppert averaged iterates of linear stochastic approximation. *arXiv preprint arXiv:2207.04475*, 2022.

Robert F Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the econometric society*, pp. 987–1007, 1982.

Christian Francq and Jean-Michel Zakoian. *GARCH models: structure, statistical inference and financial applications*. John Wiley & Sons, 2019.

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

Daniel Gervini. Robust functional estimation using the median and spherical principal components. *Biometrika*, 95(3):587–600, 2008.

Antoine Godichon-Baggioni. Estimating the geometric median in hilbert spaces with stochastic gradient algorithms: Lp and almost sure rates of convergence. *Journal of Multivariate Analysis*, 146:209–222, 2016.

Antoine Godichon-Baggioni and Bruno Portier. An averaged projected robbins-monro algorithm for estimating the parameters of a truncated spherical distribution. *Electronic Journal of Statistics*, 11(1):1890–1927, 2017.

Antoine Godichon-Baggioni, Nicklas Werge, and Olivier Wintenberger. Non-asymptotic analysis of stochastic approximation algorithms for streaming data. *arXiv preprint arXiv:2109.07117*, 2021.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International Conference on Machine Learning*, pp. 5200–5209. PMLR, 2019.

JBS Haldane. Note on the median of a multivariate distribution. *Biometrika*, 35(3-4):414–417, 1948.

James Douglas Hamilton. *Time series analysis*. Princeton university press, 2020.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Trevor Hastie, Robert Tibshirani, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2 (3-4):157–325, 2016.

Belhal Karimi, Blazej Miasojedow, Eric Moulines, and Hoi-To Wai. Non-asymptotic analysis of biased stochastic approximation scheme. In *Conference on Learning Theory*, pp. 1944–1974. PMLR, 2019.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

JHB Kemperman. The median of a finite measure on a banach space. *Statistical data analysis based on the L1-norm and related methods (Neuchâtel, 1987)*, pp. 217–230, 1987.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. In *Annales de l'institut Fourier*, volume 48, pp. 769–783, 1998.

H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer-Verlag, 2003.

Stanislaw Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963.

Shaocong Ma, Ziyi Chen, Yi Zhou, Kaiyi Ji, and Yingbin Liang. Data sampling affects the complexity of online sgd over dependent data. In *Uncertainty in Artificial Intelligence*, pp. 1296–1305. PMLR, 2022.

Abdelkader Mokkadem and Mariane 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.

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

Noboru Murata and Shun-ichi Amari. Statistical analysis of learning dynamics. *Signal Processing*, 74(1): 3–28, 1999. ISSN 0165-1684.

Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1):69–107, 2019.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2). In *Doklady an ussr*, volume 269, pp. 543–547, 1983.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Lam Nguyen, Phuong Ha Nguyen, Marten Dijk, Peter Richtárik, Katya Scheinberg, and Martin Takác. Sgd and hogwild! convergence without the bounded gradients assumption. In *International Conference on Machine Learning*, pp. 3750–3758. PMLR, 2018.

Lam M. Nguyen, Phuong Ha Nguyen, Peter Richtárik, Katya Scheinberg, Martin Takáč, and Marten van Dijk. New convergence aspects of stochastic gradient algorithms. *Journal of Machine Learning Research*, 20(176):1–49, 2019. URL `http://jmlr.org/papers/v20/18-759.html`.

Ivan Nourdin. *Selected aspects of fractional Brownian motion*, volume 4. Springer, 2012.

David Nualart. *The Malliavin calculus and related topics*, volume 1995. Springer, 2006.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

Emmanuel Rio. *Asymptotic theory of weakly dependent random processes*, volume 80. Springer, 2017.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Murray Rosenblatt. A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America*, 42(1):43, 1956.

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

Mark Schmidt, Nicolas Roux, and Francis Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. *Advances in neural information processing systems*, 24, 2011.

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

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

Choon Hui Teo, Alex Smola, SVN Vishwanathan, and Quoc Viet Le. A scalable modular convex solver for regularized risk minimization. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 727–736, 2007.

Tijmen Tieleman, Geoffrey Hinton, et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

Endre Weiszfeld and Frank Plastria. On the point for which the sum of the distances to n given points is minimum. *Annals of Operations Research*, 167(1):7–41, 2009.

Nicklas Werge and Olivier Wintenberger. Adavol: An adaptive recursive volatility prediction method. *Econometrics and Statistics*, 23:19–35, 2022. ISSN 2452-3062.

Olivier Wintenberger. Stochastic online convex optimization; application to probabilistic time series forecasting. *arXiv preprint arXiv:2102.00729*, 2021.

Lin Xiao. Dual averaging method for regularized stochastic learning and online optimization. *Advances in Neural Information Processing Systems*, 22, 2009.

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

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 116, 2004.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pp. 928–936, 2003.

# A Proofs

We start by deriving recursive relations to the desired quantities $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ and $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$. These, are derived for any $(\gamma_t)$, $(\nu_t)$, $(\kappa_t)$, $(\sigma_t)$, and $(n_t)$. Finally, we insert the specific functions forms of these, which yield the results seen in Theorems 1 and 2. Before doing the proofs, we recall a repeating argument used to non-asymptotically bound $\delta_t$ and $\bar{\delta}_t$:

**Proposition 1** (Godichon-Baggioni et al. (2021))**.** *Suppose $(\omega_t)$, $(\alpha_t)$, $(\eta_t)$, and $(\beta_t)$ to be some non-negative sequences satisfying the recursive relation,*

$$\omega_t \leq [1 - 2\lambda\alpha_t + \eta_t\alpha_t]\omega_{t-1} + \beta_t\alpha_t, \tag{17}$$

*with $\omega_0 \geq 0$ and $\lambda > 0$. Let $C_\omega \geq 1$ be such that $\lambda\alpha_t \leq 1$ for all $t \geq t_\omega$ with $t_\omega = \inf\{t \geq 1 : C_\omega\eta_t \leq \lambda\}$. Then, for $(\alpha_t)$ and $(\eta_t)$ decreasing, we have the upper bound on $(\omega_t)$ given by*

$$\omega_t \leq \tau_t + \frac{1}{\lambda}\max_{t/2\leq i\leq t}\beta_i, \tag{18}$$

*with*

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

Proposition 1 shows a simple way to bound $(\omega_t)$ in (17); the bound in (18) consists of a sub-exponential term $\tau_t$ and a noise term $\lambda^{-1}\max_{t/2\leq i\leq t}\beta_i$. Thus, our attention is on reducing the noise term without damaging the natural decay of the sub-exponential term where $\tau_t \to 0$ exponentially fast as $t \to \infty$.

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

$$\psi_x(t) = \begin{cases} t^{1-x}/(1-x) & \text{if } x < 1, \\ 1 + \log(t) & \text{if } x = 1, \\ x/(x-1) & \text{if } x > 1, \end{cases} \quad \text{and} \quad \psi_x^y(t) = \begin{cases} t^{(1-x)/(1+y)}/(1-x) & \text{if } x < 1, \\ 1 + \log(t^{1/(1+y)}) & \text{if } x = 1, \\ x/(x-1) & \text{if } x > 1, \end{cases} \tag{19}$$

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

In the following lemma, we derive an explicit recursive relation of $\delta_t$ to non-asymptotically bound the $t$-th estimate of (2) for any $(\gamma_t)$, $(\nu_t)$, $(\kappa_t)$, $(\sigma_t)$, and $(n_t)$ using classical techniques from stochastic approximations (Benveniste et al., 2012; Kushner & Yin, 2003). As mentioned in Zinkevich (2003), bounding the projected estimate in (3) follows directly from that $\mathbb{E}[\|\mathcal{P}_\Theta(\theta) - \theta^*\|^2] \leq \mathbb{E}[\|\theta - \theta^*\|^2]$, $\forall\theta \in \mathbb{R}^d$, $\forall\theta^* \in \Theta$, as $\Theta$ is a convex body.

**Lemma 1.** *Let $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$, where $(\theta_t)$ either follows the recursion in (2) or (3). Suppose Assumptions 1-p to 3-p hold for $p = 2$. Let $\mathbb{1}_{\{\nu_t = \mathcal{C}\}}$ and $\mathbb{1}_{\{\nu_t \neg \mathcal{C}\}}$ indicate whether $(\nu_t)$ is constant or not. If $\mu_\nu = \mu - \mathbb{1}_{\{\nu_t = \mathcal{C}\}}2D_\nu\nu_t > 0$, then for any learning rate $(\gamma_t)$, we have*

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

*with*

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

*Proof of Lemma 1.* By taking the quadratic norm on (2), expanding the norm, and taking the expectation, we can derive the equation,

$$\delta_t = \delta_{t-1}+\gamma_t^2\mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1})\|^2]-2\gamma_t\mathbb{E}[\langle\nabla_\theta l_t(\theta_{t-1}),\theta_{t-1}-\theta^*\rangle], \tag{20}$$

with $\delta_0\ge 0$. To bound the second term in (20), we use Assumptions 2-p and 3-p for $p=2$, to obtain that

$$\mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1})\|^2]\le 2\mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1})-\nabla_\theta l_t(\theta^*)\|^2]+2\mathbb{E}[\|\nabla_\theta l_t(\theta^*)\|^2]\le 2\kappa_t^2\delta_{t-1}+2\sigma_t^2, \tag{21}$$

as $\|x+y\|^p\le 2^{p-1}(\|x\|^p+\|y\|^p)$. As noted in Bottou et al. (2018); Nesterov et al. (2018), (5) implies that $\langle\nabla_\theta L(\theta),\theta-\theta^*\rangle\ge\mu\|\theta-\theta^*\|^2$ for all $\theta\in\Theta$. Next, since $L$ is $\mu-$strongly convex (5) and $\theta_{t-1}$ is $\mathcal{F}_{t-1}$-measurable (Assumption 1-p), we can bound the third term on the right-hand side of (20) as follows:

$$\mathbb{E}[\langle\nabla_\theta l_t(\theta_{t-1}),\theta_{t-1}-\theta^*\rangle]\ge\mu\delta_{t-1}-D_\nu\nu_t\delta_{t-1}-B_\nu\nu_t\delta_{t-1}^{\frac{1}{2}}, \tag{22}$$

since

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

by Jensen's, Cauchy–Schwarz, and Hölder's inequality, and Assumption 1-p with $p=2$. Hence, applying the inequalities (21) and (22) to (20), yields

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

using Young's inequality[7]; $2B_\nu\nu_t\gamma_t\delta_{t-1}^{\frac{1}{2}}\le\mu\gamma_t\delta_{t-1}+B_\nu^2\nu_t^2\gamma_t/\mu$. Next, we introduce the indicator function for whether $(\nu_t)$ is constant $(=\mathcal{C})$ or not $(\neg\mathcal{C})$, such that

$$\delta_t\le[1-(\mu_\nu-\mathbb{1}_{\{\nu_t\neg\mathcal{C}\}}2D_\nu\nu_t)\gamma_t+2\kappa_t^2\gamma_t^2]\delta_{t-1}+\frac{B_\nu^2}{\mu}\nu_t^2\gamma_t+2\sigma_t^2\gamma_t^2, \tag{23}$$

with $\mu_\nu=\mu-\mathbb{1}_{\{\nu_t=\mathcal{C}\}}2D_\nu\nu_t>0$. Let $C_\delta$ be the constant fulfilling the conditions of Proposition 1 such that $C_\delta$ is chosen larger than 1 verifying $C_\delta(\mathbb{1}_{\{\nu_t\neg\mathcal{C}\}}2D_\nu\nu_t+2\kappa_t^2\gamma_t)\le\mu_\nu/2$ such that it's imply $\mu_\nu\gamma_t/2\le 1$, which is possible as the sequence $(\nu_t)$ is non-increasing, and $(\kappa_t)$ and $(\gamma_t)$ is decreasing. At last, bounding (23) by Proposition 1 concludes the proof. □

*Proof of Theorem 1.* Inserting the functions $\gamma_t=C_\gamma n_t^\beta t^{-\alpha}$, $\nu_t=n_t^{-\nu}$, $\kappa_t=C_\kappa n_t^{-\kappa}$, $\sigma_t=C_\sigma n_t^{-\sigma}$, and $n_t=C_\rho t^\rho$ into the bound of Lemma 1 yields

$$\delta_t\le\pi_t+\frac{2^{1+2\rho\nu}B_\nu^2}{\mu\mu_\nu C_\rho^{2\nu}t^{2\rho\nu}}+\frac{2^{2+\rho(2\sigma-\beta)+\alpha}C_\sigma^2C_\gamma C_\rho^\beta}{\mu_\nu C_\rho^{2\sigma}t^{\rho(2\sigma-\beta)+\alpha}} \tag{24}$$

$$\le\pi_t+\frac{2^{(2+6\rho\nu)/(1+\rho)}B_\nu^2}{\mu\mu_\nu C_\rho^{2\nu/(1+\rho)}N_t^{2\rho\nu/(1+\rho)}}+\frac{2^{(7+6\rho\sigma)/(1+\rho)}C_\sigma^2C_\gamma}{\mu_\nu C_\rho^{(2\sigma-\beta-\alpha)/(1+\rho)}N_t^{(\rho(2\sigma-\beta)+\alpha)/(1+\rho)}}, \tag{25}$$

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

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

$$\begin{aligned}
\pi_t \leq & \exp\left(-\frac{\mu_\nu C_\gamma C_\rho^\beta t^{1+\rho\beta-\alpha}}{2^2}\right)\left[\exp\left(\frac{\mathbb{1}_{\{\rho\neq 0\}} 2C_\delta D_\nu C_\gamma C_\rho^\beta \psi_{\alpha-\rho(\beta-\nu)}(t)}{C_\rho^\nu}\right)\exp\left(\frac{4(\alpha-\rho(\beta-\kappa))C_\delta C_\kappa^2 C_\gamma^2 C_\rho^{2\beta}}{(2\alpha-2\rho(\beta-\kappa)-1)C_\rho^{2\kappa}}\right)\right. \\
& \left.\left(\delta_0 + \frac{2B_\nu^2}{\mu\mu_\nu C_\rho^{2\nu}} + \frac{4C_\sigma^2 C_\gamma C_\rho^\beta}{\mu_\nu C_\rho^{2\sigma}}\right) + \frac{B_\nu^2 C_\gamma C_\rho^\beta \psi_{\alpha-\rho(\beta-2\nu)}(t/2)}{\mu C_\rho^{2\nu}} + \frac{4(\alpha-\rho(\beta-\sigma))C_\sigma^2 C_\gamma^2 C_\rho^{2\beta}}{(2\alpha-2\rho(\beta-\sigma)-1)C_\rho^{2\sigma}}\right] \\
\leq & \exp\left(-\frac{\mu C_\gamma N_t^{(1+\rho\beta-\alpha)/(1+\rho)}}{2^{(3+\rho(2+\beta)-\alpha)/(1+\rho)} C_\rho^{(1-\beta-\alpha)/(1+\rho)}}\right)\left[\exp\left(\frac{\mathbb{1}_{\{\rho\neq 0\}} 2C_\delta D_\nu C_\gamma C_\rho^\beta \psi^\rho_{\alpha-\rho(\beta-\nu)}(2N_t/C_\rho)}{C_\rho^\nu}\right)\right. \\
& \exp\left(\frac{4(\alpha-\rho(\beta-\kappa))C_\delta C_\kappa^2 C_\gamma^2 C_\rho^{2\beta}}{(2\alpha-2\rho(\beta-\kappa)-1)C_\rho^{2\kappa}}\right)\left(\delta_0 + \frac{2B_\nu^2}{\mu\mu_\nu C_\rho^{2\nu}} + \frac{4C_\sigma^2 C_\gamma C_\rho^\beta}{\mu_\nu C_\rho^{2\sigma}}\right) \\
& \left. + \frac{B_\nu^2 C_\gamma C_\rho^\beta \psi^\rho_{\alpha-\rho(\beta-2\nu)}(N_t/C_\rho)}{\mu C_\rho^{2\nu}} + \frac{4(\alpha-\rho(\beta-\sigma))C_\sigma^2 C_\gamma^2 C_\rho^{2\beta}}{(2\alpha-2\rho(\beta-\sigma)-1)C_\rho^{2\sigma}}\right], \qquad (26)
\end{aligned}$$

with help of an integral test for convergence[8], the functions $\psi_x(t)$ and $\psi_x^y(t)$ from (19), and by use of $(N_t/2C_\rho)^{1/(1+\rho)} \leq t \leq (2N_t/C_\rho)^{1/(1+\rho)}$. □

Next, we need to study fourth-order rate $\Delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^4]$ of the recursive estimates (2) and (3). As in Lemma 1, we begin the following sections by conducting a general study for any $(\gamma_t)$, $(\nu_t)$, $(\kappa_t)$, $(\sigma_t)$, and $(n_t)$, when applying the Polyak-Ruppert averaging estimate $(\bar{\theta}_t)$ from (4).

**Lemma 2.** *Let $\Delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^4]$, where $(\theta_t)$ either follows the recursion in (2) or (3). Suppose Assumptions 1-p to 3-p hold for $p = 4$. Let $\mathbb{1}_{\{\nu_t=\mathcal{C}\}}$ and $\mathbb{1}_{\{\nu_t\neg\mathcal{C}\}}$ indicate whether $(\nu_t)$ is constant or not. If $\mu'_\nu = \mu - \mathbb{1}_{\{\nu_t=\mathcal{C}\}} 2D_\nu^4 \nu_t^4/\mu^3 > 0$, then for any learning rate $(\gamma_t)$, we have*

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

*with $\Pi_t$ given as*

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

*Proof of Lemma 2.* The derivation of the recursive step sequence for the fourth-order moment $\Delta_t$ of (2) follows the same methodology as for the second-order moment in Lemma 1. In the same way we deduced (20), we can take the quadratic norm on (2), expand the norm, take the square on both sides, and the conditional expectation on both sides of the equality;

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

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

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

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

To bound the fourth-order term $\mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1})\|^4]$, we make use of the Lipschitz continuity of $\nabla_\theta l_t$ (Assumption 2-p), Assumption 3-p, and that $\theta_{t-1}$ is $\mathcal{F}_{t-1}$-measurable (Assumption 1-p), to show that $\mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1})\|^4] \leq 8\kappa_t^4\Delta_{t-1} + 8\sigma_t^4$ as $\|x+y\|^p \leq 2^{p-1}(\|x\|^p + \|y\|^p)$ for any $p \in \mathbb{N}$. Thus,

$$\begin{aligned}\Delta_t \leq &[1 + \mu\gamma_t/2 + 256\mu^{-1}\kappa_t^4\gamma_t^3 + 24\kappa_t^4\gamma_t^4]\Delta_{t-1} + 256\mu^{-1}\sigma_t^4\gamma_t^3 + 24\sigma_t^4\gamma_t^4 \\ &- 4\gamma_t\mathbb{E}[\|\theta_{t-1} - \theta^*\|^2\langle\nabla_\theta l_t(\theta_{t-1}), \theta_{t-1} - \theta^*\rangle]. \end{aligned} \tag{27}$$

Next, using the same arguments as in the proof of Lemma 1, Young's inequality, and Assumption 1-p with $p = 4$, we have

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

such that the last term of (27) can be bounded as follows,

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

Indeed, inserting this into (27) together with using the indicator function that determines whether $(\nu_t)$ is constant $(= \mathcal{C})$ or not $(\neg\mathcal{C})$, gives us

$$\Delta_t \leq \left[1 - \left(\frac{\mu_\nu}{2} - \frac{\mathbb{1}_{\{\nu_t\neg\mathcal{C}\}}D_\nu^4\nu_t^4}{\mu^3}\right)\gamma_t + \frac{256\kappa_t^4\gamma_t^3}{\mu} + 24\kappa_t^4\gamma_t^4\right]\Delta_{t-1} + \frac{B_\nu^4\nu_t^4\gamma_t}{\mu^3} + \frac{256\sigma_t^4\gamma_t^3}{\mu} + 24\sigma_t^4\gamma_t^4, \tag{28}$$

with $\mu'_\nu = \mu - \mathbb{1}_{\{\nu_t=\mathcal{C}\}}2D_\nu^4\nu_t^4/\mu^3 > 0$. Note that $\mu_\nu$ from Lemma 1 is lower bounded by $\mu'_\nu$, and strictly lower bounded for $(\nu_t)$ constant, i.e., $\mu_\nu > \mu'_\nu > 0$. Let $C_\Delta \geq 1$ fulfill the conditions of Proposition 1; the $C_\Delta$ constant is chosen such that $C_\Delta(\mathbb{1}_{\{\nu_t\neg\mathcal{C}\}}D_\nu^4\nu_t^4/\mu^3 + 256\kappa_t^4\gamma_t^2/\mu + 24\kappa_t^4\gamma_t^3) \leq \mu'_\nu/2$ implying $\mu'_\nu\gamma_t/2 \leq 1$, which is possible as the sequence $(\nu_t)$ is non-increasing, and $(\kappa_t)$ and $(\gamma_t)$ decrease. Hence, by applying Proposition 1 on (28), we obtain the desired bound for $\Delta_t$. □

**Corollary 1.** *Let $\Delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^4]$, where $(\theta_t)$ either follows the recursion in (2) or (3). Suppose Assumptions 1-p to 3-p hold for $p = 4$. If $\mu'_\nu = \mu - \mathbb{1}_{\{\rho=0\}}2D_\nu^4/\mu^3C_\rho^{4\nu} > 0$, then for $\alpha - \rho\beta \in (1/2, 1)$, we have*

$$\Delta_t \leq \Pi_t + \frac{2^{2+4\rho\nu}B_\nu^4}{\mu^3\mu'_\nu C_\rho^{4\nu}t^{4\rho\nu}} + \frac{2^{2\rho(2\sigma-\beta)+2\alpha}(2^{10}\mu^{-1} + 2^7C_\gamma C_\rho^\beta)C_\sigma^4C_\gamma^2C_\rho^{2\beta}}{\mu'_\nu C_\rho^{4\sigma}t^{2\rho(2\sigma-\beta)+2\alpha}}, \tag{29}$$

*with $\Pi_t$ given in (30) such that $\Pi_t = \mathcal{O}(\exp(-N_t^{(1+\rho\beta-\alpha)/(1+\rho)}))$.*

*Proof of Corollary 1.* Inserting the functions $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$, $\nu_t = n_t^{-\nu}$, $\kappa_t = C_\kappa n_t^{-\kappa}$, $\sigma_t = C_\sigma n_t^{-\sigma}$, and $n_t = C_\rho t^\rho$ into the bound of Lemma 2 and using $\gamma_t^3 \leq C_\gamma C_\rho^\beta\gamma_t^2$ as $\alpha - \rho\beta \in (1/2, 1)$, yields (29) with

$\mu'_\nu = \mu - \mathbb{1}_{\{\rho=0\}} 2D_\nu^4/\mu^3 C_\rho^{4\nu} > 0$, where $\Pi_t$ can be bounded as follows:

$$\begin{aligned}
\Pi_t \leq & \exp\left(-\frac{\mu'_\nu C_\gamma C_\rho^\beta}{4}\sum_{i=t/2}^{t} i^{\rho\beta-\alpha}\right)\left[\exp\left(\frac{\mathbb{1}_{\{\rho\neq 0\}} C_\Delta D_\nu^4 C_\gamma C_\rho^\beta}{\mu^3 C_\rho^{4\nu}}\sum_{i=1}^{t} i^{\rho(\beta-4\nu)-\alpha}\right)\right. \\
& \exp\left(\frac{2^8 C_\Delta C_\kappa^4 C_\gamma^3 C_\rho^{3\beta}}{\mu C_\rho^{4\kappa}}\sum_{i=1}^{t} i^{\rho(3\beta-4\kappa)-3\alpha}\right)\exp\left(\frac{24 C_\Delta C_\kappa^4 C_\gamma^4 C_\rho^{4\beta}}{C_\rho^{4\kappa}}\sum_{i=1}^{t} i^{4\rho(\beta-\kappa)-4\alpha}\right) \\
& \left(\Delta_0 + \frac{4B_\nu^4}{\mu^3\mu'_\nu C_\rho^{4\nu}} + \frac{1024 C_\sigma^4 C_\gamma^2 C_\rho^{2\beta}}{\mu\mu'_\nu C_\rho^{4\sigma}} + \frac{96 C_\sigma^4 C_\gamma^3 C_\rho^{3\beta}}{\mu'_\nu C_\rho^{4\sigma}}\right) + \frac{B_\nu^4 C_\gamma C_\rho^\beta}{\mu^3 C_\rho^{4\nu}}\sum_{i=1}^{t/2-1} i^{\rho(\beta-4\nu)-\alpha} \\
& \left. + \frac{256 C_\sigma^4 C_\gamma^3 C_\rho^{3\beta}}{\mu C_\rho^{4\sigma}}\sum_{i=1}^{t/2-1} i^{\rho(3\beta-4\sigma)-3\alpha} + \frac{24 C_\sigma^4 C_\gamma^4 C_\rho^{4\beta}}{C_\rho^{4\sigma}}\sum_{i=1}^{t/2-1} i^{4\rho(\beta-\sigma)-4\alpha}\right] \\
\leq & \exp\left(-\frac{\mu'_\nu C_\gamma C_\rho^\beta t^{1+\rho\beta-\alpha}}{2^3}\right)\left[\exp\left(\frac{\mathbb{1}_{\{\rho\neq 0\}} C_\Delta D_\nu^4 C_\gamma C_\rho^\beta \psi^0_{\alpha-\rho(\beta-4\nu)}(t)}{\mu^3 C_\rho^{4\nu}}\right)\exp\left(\frac{2^{10} C_\Delta C_\kappa^4 C_\gamma^3 C_\rho^{3\beta}}{\mu C_\rho^{4\kappa}}\right)\right. \\
& \exp\left(\frac{2^6 C_\Delta C_\kappa^4 C_\gamma^4 C_\rho^{4\beta}}{C_\rho^{4\kappa}}\right)\left(\Delta_0 + \frac{2^2 B_\nu^4}{\mu^3\mu'_\nu C_\rho^{4\nu}} + \frac{2^{10} C_\sigma^4 C_\gamma^2 C_\rho^{2\beta}}{\mu\mu'_\nu C_\rho^{4\sigma}} + \frac{2^7 C_\sigma^4 C_\gamma^3 C_\rho^{3\beta}}{\mu'_\nu C_\rho^{4\sigma}}\right) \\
& \left. + \frac{B_\nu^4 C_\gamma C_\rho^\beta \psi^0_{\alpha-\rho(\beta-4\nu)}(t/2)}{\mu^3 C_\rho^{4\nu}} + \frac{2^{10} C_\sigma^4 C_\gamma^3 C_\rho^{3\beta}}{\mu C_\rho^{4\sigma}} + \frac{2^6 C_\sigma^4 C_\gamma^4 C_\rho^{4\beta}}{C_\rho^{4\sigma}}\right],
\end{aligned} \tag{30}$$

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

**Lemma 3.** *Let $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ with $\bar{\theta}_n$ given by* (4)*, where $(\theta_t)$ either follows the recursion in* (2) *or* (3)*. Suppose Assumptions 1-p to 3-p and 5 hold for $p = 4$. In addition, Assumption 6 must hold true if $(\theta_t)$ follows the recursion in* (3)*, which is indicated by $\mathbb{1}_{\{D_\Theta<\infty\}}$. Then, for any learning rate $(\gamma_t)$, we have*

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

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

*Proof of Lemma 3.* The proof is divided into two parts; in the first part, $(\theta_t)$ follows (2), and the second part considers (3). Assume that $(\theta_t)$ is derived from the recursion in (2). Following Polyak & Juditsky (1992), we observe that

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

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

$$\begin{aligned}
\left(\mathbb{E}\left[\left\|\bar{\theta}_t - \theta^*\right\|^2\right]\right)^{\frac{1}{2}} \leq & \left(\mathbb{E}\left[\left\|\nabla^2_\theta L\left(\theta^*\right)^{-1} \frac{1}{N_t}\sum_{i=1}^{t} n_i \nabla_\theta l_i\left(\theta^*\right)\right\|^2\right]\right)^{\frac{1}{2}} \\
& + \left(\mathbb{E}\left[\left\|\nabla^2_\theta L\left(\theta^*\right)^{-1} \frac{1}{N_t}\sum_{i=1}^{t} n_i \nabla_\theta l_i\left(\theta_{i-1}\right)\right\|^2\right]\right)^{\frac{1}{2}} \\
& + \left(\mathbb{E}\left[\left\|\nabla^2_\theta L\left(\theta^*\right)^{-1} \frac{1}{N_t}\sum_{i=1}^{t} n_i \left[\nabla_\theta l_i\left(\theta_{i-1}\right) - \nabla_\theta l_i\left(\theta^*\right) - \nabla_\theta L\left(\theta_{i-1}\right)\right]\right\|^2\right]\right)^{\frac{1}{2}} \\
& + \left(\mathbb{E}\left[\left\|\nabla^2_\theta L\left(\theta^*\right)^{-1} \frac{1}{N_t}\sum_{i=1}^{t} n_i \left[\nabla_\theta L\left(\theta_{i-1}\right) - \nabla^2_\theta L\left(\theta^*\right)\left(\theta_{i-1} - \theta^*\right)\right]\right\|^2\right]\right)^{\frac{1}{2}}. \qquad (31)
\end{aligned}$$

First term of (31): As $(\nabla_\theta l_t(\theta^*))$ is a square-integrable sequences on $\mathbb{R}^d$ (Assumption 1-p), we have

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

where the first term can be bounded by Assumption 5, namely

$$\frac{1}{N_t^2}\sum_{i=1}^{t} n_i^2 \mathbb{E}\left[\left\|\nabla^2_\theta L\left(\theta^*\right)^{-1} \nabla_\theta l_i\left(\theta^*\right)\right\|^2\right] \leq \frac{\Lambda}{N_t^2}\sum_{i=1}^{t} n_i^{2(1-\sigma)} + \frac{C'_\sigma}{\mu^2 N_t^2}\sum_{i=1}^{t} n_i^{2(1-\sigma-\sigma')},$$

where $\Lambda$ denotes $\mathrm{Tr}[\nabla^2_\theta L(\theta^*)^{-1}\Sigma\nabla^2_\theta L(\theta^*)^{-1}]$. For the next term, we use

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

by Cauchy-Schwarz and Hölder's inequality, and Assumptions 1-p and 3-p. Thus,

$$\begin{aligned}
& \left(\mathbb{E}\left[\left\|\nabla^2_\theta L\left(\theta^*\right)^{-1} \frac{1}{N_t}\sum_{i=1}^{t} n_i \nabla_\theta l_i\left(\theta^*\right)\right\|^2\right]\right)^{\frac{1}{2}} \leq \frac{\Lambda^{\frac{1}{2}}}{N_t}\left(\sum_{i=1}^{t} n_i^{2(1-\sigma)}\right)^{\frac{1}{2}} \\
& + \frac{C'^{1/2}_\sigma}{\mu N_t^{1/2}}\left(\sum_{i=1}^{t} n_i^{2(1-\sigma-\sigma')}\right)^{\frac{1}{2}} + \frac{2^{1/2} B_\nu^{1/2}}{\mu N_t}\left(\sum_{j=2}^{t}\left(n_j \nu_j \sum_{i=1}^{j-1} n_i \sigma_i\right)\right)^{\frac{1}{2}}. \qquad (32)
\end{aligned}$$

Second term of (31): Next, using that $\frac{1}{N_t}\sum_{i=1}^t n_i \nabla_\theta l_i(\theta_{i-1}) = \frac{1}{N_t}\sum_{i=1}^t \frac{n_i}{\gamma_i}(\theta_{i-1}-\theta_i) = \frac{1}{N_t}\sum_{i=1}^{t-1}(\theta_i - \theta^*)(\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}) - \frac{1}{N_t}(\theta_t - \theta^*)\frac{n_t}{\gamma_t} + \frac{1}{N_t}(\theta_0 - \theta^*)\frac{n_1}{\gamma_1}$ leads to an upper bound on normed quantity $\|\nabla^2_\theta L(\theta^*)^{-1}\frac{1}{N_t}\sum_{i=1}^t n_i \nabla_\theta l_i(\theta_{i-1})\|$ given by

$$\frac{1}{\mu N_t}\sum_{i=1}^{t-1}\|\theta_i - \theta^*\|\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| + \frac{1}{\mu N_t}\|\theta_t - \theta^*\|\frac{n_t}{\gamma_t} + \frac{1}{\mu N_t}\|\theta_0 - \theta^*\|\frac{n_1}{\gamma_1}.$$

Hence, with the notation of $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$, the second term of (31), namely $(\mathbb{E}[\|\nabla^2_\theta L(\theta^*)^{-1}\frac{1}{N_t}\sum_{i=1}^t n_i \nabla_\theta l_i(\theta_{i-1})\|^2])^{1/2}$, can be bounded by

$$\frac{1}{\mu N_t}\sum_{i=1}^{t-1}\delta_i^{\frac{1}{2}}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| + \frac{n_t}{\mu\gamma_t N_t}\delta_t^{\frac{1}{2}} + \frac{n_1}{\mu\gamma_1 N_t}\delta_0^{\frac{1}{2}}. \tag{33}$$

Third term of (31): Here, we use that $\mathbb{E}[\|\nabla^2_\theta L(\theta^*)^{-1}\frac{1}{N_t}\sum_{i=1}^t n_i[\nabla_\theta l_i(\theta_{i-1}) - \nabla_\theta l_i(\theta^*) - \nabla_\theta L(\theta_{i-1})]\|^2]$ can be derived as

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

Here, we use Cauchy-Schwarz inequality, Assumption 2-p and (10) to show that

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

and for the other term, we note that

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

using $\mathcal{F}_{i-1}\subset\mathcal{F}_{j-1}$ since $i<j$, Cauchy–Schwarz and Hölder's inequality, $\|a+b\|^p \le 2^{p-1}(\|a\|^p + \|b\|^p)$ with $p\in\mathbb{N}$, Assumptions 1-p and 2-p, and (10). Thus, the third term of (31) can be upper bounded by

$$\begin{aligned}
&\frac{2^{1/2}}{\mu N_t}\left(\sum_{i=1}^t n_i^2\kappa_i^2\delta_{i-1}\right)^{1/2} + \frac{2^{1/2}C_\nabla}{\mu N_t}\left(\sum_{i=1}^t n_i^2\delta_{i-1}\right)^{1/2} \\
+&\frac{2^{3/4}}{\mu N_t}\left(\sum_{j=2}^t\left((D_\nu\delta_{j-1}^{1/2} + 2^{1/2}B_\nu)n_j\nu_j\sum_{i=1}^{j-1}(C_\nabla + \kappa_i)n_i\delta_{i-1}^{1/2}\right)\right)^{1/2}.
\end{aligned} \tag{34}$$

Fourth term of (31): Here, we use that (11) imply $\forall\theta$, $\|\nabla_\theta L(\theta) - \nabla^2_\theta L(\theta^*)(\theta-\theta^*)\| \le C'_\nabla\|\theta-\theta^*\|^2/2$ (Nesterov et al., 2018), which gives the upper bound $\frac{C'_\nabla}{2\mu N_t}\sum_{i=1}^t n_i\Delta_{i-1}^{1/2}$ using the notion $\Delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^4]$. Combining

the terms (32) to (34) into (31), together with shifting the indices and collecting the $\delta_0$ terms, gives use the desired bound for when $(\theta_t)$ follows (2).

Now, assume that $(\theta_t)$ is derived from the recursion in (3). As above, we follow the steps of Polyak & Juditsky (1992), in which, we can rewrite (3) to $\frac{1}{\gamma_t}(\theta_{t-1}-\theta_t) = \nabla_\theta l_t(\theta_{t-1}) - \frac{1}{\gamma_t}\Omega_t$, where $\Omega_t = \mathcal{P}_\Theta(\theta_{t-1} - \gamma_t\nabla_\theta l_t(\theta_{t-1})) - (\theta_{t-1} - \gamma_t\nabla_\theta l_t(\theta_{t-1}))$. Thus, summing the parts, taking the norm and expectation, and using the Minkowski's inequality, yields the same terms as in (31), but with an additional term regarding $\Omega_t$, namely

$$\left(\mathbb{E}\left[\left\|\nabla_\theta^2 L\left(\theta^*\right)^{-1}\frac{1}{N_t}\sum_{i=1}^{t}\frac{n_i}{\gamma_i}\Omega_i\right\|^2\right]\right)^{\frac{1}{2}} \leq \frac{1}{\mu N_t}\sum_{i=1}^{t}\frac{n_i}{\gamma_i}\sqrt{\mathbb{E}\left[\left\|\Omega_i\right\|^2 \mathbb{1}_{\{\theta_{i-1}-\gamma_i\nabla_\theta l_i(\theta_{i-1})\notin\Theta\}}\right]}, \tag{35}$$

using (Godichon-Baggioni, 2016, Lemma 4.3). Next, we note that

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

as $\mathcal{P}_\Theta$ is Lipschitz and $\|\nabla_\theta l_t(\theta)\|^2 \leq G_\Theta^2$ for any $\theta\in\Theta$. This means that the inner expectation of (35), $\mathbb{E}[\|\Omega_t\|^2\mathbb{1}_{\{\theta_{t-1}-\gamma_t\nabla_\theta l_t(\theta_{t-1})\notin\Theta\}}] = 4\gamma_t^2 G_\Theta^2\mathbb{P}[\theta_{t-1}-\gamma_t\nabla_\theta l_t(\theta_{t-1})\notin\Theta]$. Moreover, as in (Godichon-Baggioni & Portier, 2017, Theorem 4.2) with use of Lemma 2, we know that $\mathbb{P}[\theta_{t-1}-\gamma_t\nabla_\theta l_t(\theta_{t-1})\notin\Theta] \leq \Delta_t/D_\Theta^4$, where $D_\Theta = \inf_{\theta\in\partial\Theta}\|\theta-\theta^*\|$ with $\partial\Theta$ denoting the frontier of $\Theta$. Thus, (35) can then be bounded by

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

since the sequence $(n_t)$ is either constant or increasing, meaning $\forall t, n_t/n_{t+1}\leq 1$. At last, let $C''_\nabla = C'_\nabla/2 + \mathbb{1}_{\{D_\Theta<\infty\}}2G_\Theta/D_\Theta^2$ indicate whether $(\theta_t)$ follows (3) or not. □

*Proof of Theorem 2.* The result follows by simplifying and bounding each term of Lemma 3, with use of Theorem 1 and Lemma 2. Thus, by inserting $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$, $\nu_t = n_t^{-\nu}$, $\kappa_t = C_\kappa n_t^{-\kappa}$, $\sigma_t = C_\sigma n_t^{-\sigma}$, and $n_t = C_\rho t^\rho$ into the bound of Lemma 3, we obtain

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

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

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

using the $\psi$-function defined in (19). Hence, $\sqrt{\psi^\rho_{\rho(\sigma-1)}(2N_t/C_\rho)\psi^\rho_{\rho(\nu-1)}(2N_t/C_\rho)}/N_t$ is $\tilde{\mathcal{O}}(N_t^{-\rho(\sigma+\nu)/2(1+\rho)})$. Let $D^\kappa_\nabla$ denote $C_\nabla + 2^{\rho\kappa}C_\kappa/C_\rho^\kappa$ with $\kappa \in [0,1/2]$, such that

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

and, likewise,

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

From (24) we know that $\delta_t \leq D_\delta/t^\delta$ with

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

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

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

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

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

as $\alpha-\rho\beta \in (1/2,1)$, where $\mu'_\nu = \mu - \mathbb{1}_{\{\rho=0\}} 2D_\nu^4/\mu^3 C_\rho^{4\nu}$, $D_\nabla^\kappa = C_\nabla + 2^\kappa C_\kappa/C_\rho^\kappa$, $C''_\nabla = C'_\nabla + \mathbb{1}_{\{D_\Theta<\infty\}} 2G_\Theta/D_\Theta^2$, $\Gamma = 2\bar{\pi}_\infty/C_\gamma C_\rho^\beta + (1/C_\gamma C_\rho^\beta + 2^{1/2} D_\nabla^\kappa)\delta_0^{1/2} + 2^{1/2+\rho} D_\nabla^\kappa \bar{\pi}_\infty^{1/2} + 2^\rho C''_\nabla \bar{\Pi}_\infty$, $D_\nabla^\kappa = C_\nabla + 2^\kappa C_\kappa/C_\rho^\kappa$, $\delta = \mathbb{1}_{\{B_\nu=0\}}(\rho(2\sigma-\beta)+\alpha) + \mathbb{1}_{\{B_\nu\neq 0\}} \min\{\rho(2\sigma-\beta)+\alpha, 2\rho\nu\}$, and $\Psi_t$ given as

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

Furthermore, using the $\tilde{\mathcal{O}}$-notation one can show that

$$
\begin{aligned}
\bar{\delta}_t^{1/2} \leq\ & \frac{\Lambda^{1/2}}{N_t^{1/2}} \mathbb{1}_{\{\sigma=1/2\}} + \frac{2^{1/2}\Lambda^{1/2} C_\rho^{(1-2\sigma)/2(1+\rho)}}{N_t^{(1+2\rho\sigma)/2(1+\rho)}} \mathbb{1}_{\{\sigma\neq 1/2\}} + \frac{2^{1/2} C_\sigma'^{1/2} C_\rho^{(1-2(\sigma+\sigma'))/2(1+\rho)}}{\mu N_t^{(1+2\rho(\sigma+\sigma'))/2(1+\rho)}} \\
&+ \frac{2^6 C_\sigma C_\rho^{(2-2\sigma-\beta-\alpha)/2(1+\rho)}}{\mu\mu_\nu^{1/2} C_\gamma^{1/2} N_t^{(2+\rho(\beta+2\sigma)-\alpha)/2(1+\rho)}} + \frac{2^7(\mu^{-1/2} + C_\gamma^{1/2} C_\rho^{\beta/2}) C''_\nabla C_\sigma^2 C_\gamma}{\mu\sqrt{\mu'_\nu} C_\rho^{(1-2\rho\sigma-\alpha)/(1+\rho)} N_t^{(\rho(2\sigma-\beta)+\alpha)/(1+\rho)}} \\
&+ \frac{2^2 D_\nabla^\kappa C_\sigma C_\gamma^{1/2} C_\rho^{(1+\beta-2\sigma+\alpha)/2(1+\rho)}}{\mu\mu_\nu^{1/2} N_t^{(1+\rho(2\sigma-\beta)+\alpha)/(2(1+\rho))}} + \frac{\Gamma C_\rho}{\mu N_t} + \frac{2^2 C_\rho^{(2+\beta-\alpha)/(1+\rho)} \bar{\pi}_\infty}{\mu C_\gamma N_t^{(2+\rho\beta-\alpha)/(1+\rho)}} \\
&+ \tilde{\mathcal{O}}(N_t^{-(\delta+\rho\nu)/2(1+\rho)}) + \mathbb{1}_{\{B_\nu\neq 0\}} \Psi_t,
\end{aligned} \tag{36}
$$

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

## B Verifications of Assumptions 1-p to 3-p for the AR model

**Well-specified case.** Consider the well-specified case, in which, we estimate an AR(1) model from the underlying stationary AR(1) process $X_s = \theta^* X_{s-1} + \epsilon_s$ with $|\theta^*| < 1$. The squared loss function $l_t(\theta) = n_t^{-1} \sum_{i=1}^{n_t} (X_{N_{t-1}+i} - \theta X_{N_{t-1}+i-1})^2$ with gradient $\nabla_\theta l_t(\theta) = -2n_t^{-1} \sum_{i=1}^{n_t} X_{N_{t-1}+i-1}(X_{N_{t-1}+i} - \theta X_{N_{t-1}+i-1})$. Thus, the objective function is

$$
L(\theta) = \mathbb{E}\left[\frac{1}{n_t}\sum_{i=1}^{n_t}(X_{N_{t-1}+i} - \theta X_{N_{t-1}+i-1})^2\right] = \frac{\sigma_\epsilon^2(\theta^*-\theta)^2}{1-(\theta^*)^2} + \sigma_\epsilon^2,
$$

using $\mathbb{E}[X_s] = 0$ and $\mathbb{E}[X_s^2] = \sigma_\epsilon^2/(1-(\theta^*)^2)$, yielding $\nabla_\theta L(\theta) = 2\sigma_\epsilon^2(\theta-\theta^*)/(1-(\theta^*)^2)$. Next, to verify Assumption 1-p for $p=2$, we first note that

$$
\begin{aligned}
\mathbb{E}[\nabla_\theta l_t(\theta)|\mathcal{F}_{t-1}] &= \frac{2\theta}{n_t}\sum_{i=1}^{n_t}\mathbb{E}\left[X^2_{N_{t-1}+i-1}\middle|\mathcal{F}_{t-1}\right] - \frac{2}{n_t}\sum_{i=1}^{n_t}\mathbb{E}\left[X_{N_{t-1}+i-1}X_{N_{t-1}+i}\middle|\mathcal{F}_{t-1}\right] \\
&= \frac{2(\theta-\theta^*)}{n_t}\sum_{i=1}^{n_t}\mathbb{E}\left[X^2_{N_{t-1}+i-1}\middle|\mathcal{F}_{t-1}\right] - \frac{2}{n_t}\sum_{i=1}^{n_t}\mathbb{E}\left[X_{N_{t-1}+i-1}\epsilon_{N_{t-1}+i}\middle|\mathcal{F}_{t-1}\right],
\end{aligned} \tag{37}
$$

as $X_{N_{t-1}+i} = \theta^* X_{N_{t-1}+i-1} + \epsilon_{N_{t-1}+i}$. For the first term of (37), we use that $\mathbb{E}[X_{s+i}|\mathcal{F}_s] = (\theta^*)^i X_s$ and $\mathrm{Var}[X_{s+i}|\mathcal{F}_s] = \sigma_\epsilon^2(1-(\theta^*)^{2i})/(1-(\theta^*)^2)$, yielding

$$\begin{aligned}\sum_{i=1}^{n_t} \mathbb{E}[X^2_{N_{t-1}+i-1}|\mathcal{F}_{t-1}] &= X^2_{N_{t-1}} \sum_{i=1}^{n_t} (\theta^*)^{2(i-1)} - \frac{\sigma_\epsilon^2}{(1-(\theta^*)^2)} \sum_{i=1}^{n_t} (\theta^*)^{2(i-1)} + \frac{\sigma_\epsilon^2 n_t}{1-(\theta^*)^2} \\ &= \frac{(1-(\theta^*)^{2n_t})X^2_{N_{t-1}}}{(1-(\theta^*)^2)} - \frac{(1-(\theta^*)^{2n_t})\sigma_\epsilon^2}{(1-(\theta^*)^2)^2} + \frac{\sigma_\epsilon^2 n_t}{1-(\theta^*)^2}.\end{aligned}$$

Next, the second term of (37) is zero by utilising that $(\epsilon_s)$ is a Martingale difference sequence, i.e., $\mathbb{E}[\epsilon_{s+i}|\mathcal{F}_s] = 0$ and $\mathbb{E}[\epsilon_{s+i}\epsilon_{s+j}|\mathcal{F}_s] = 0$ for $i \neq j$. Thus,

$$\mathbb{E}[\|\mathbb{E}[\nabla_\theta l_t(\theta)|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta)\|^2] = \frac{4(\theta-\theta^*)^2(1-(\theta^*)^{2n_t})^2\sigma_\epsilon^2}{(1-(\theta^*)^2)^4 n_t^2}\left(\sigma_\epsilon^2 + \frac{1}{1-(\theta^*)^2}\right),$$

meaning that Assumption 1-p is verified for $p = 2$ if $(X_s)$ has bounded moments; this is fulfilled by the natural constraint that $|\theta^*| < 1$. Thus, we can deduce that $D_\nu > 0$, $B_\nu = 0$, and $\nu_t$ is $\mathcal{O}(n_t^{-1})$. The remaining assumptions can be verified in the same way, in particular, Assumptions 2-p and 3-p is satisfied with $\kappa_t$ and $\sigma_t$ is $\mathcal{O}(n_t^{-1/2})$, Assumption 4 with $C_\nabla = 2\sigma_\epsilon^2/(1-(\theta^*)^2)$ and $C'_\nabla = 0$, and Assumption 5 with $\Sigma = 4\sigma_\epsilon^4/(1-(\theta^*)^2)$ and $\Sigma_t = 0$. Furthermore, for an AR(1) process $X_s$ constructed using the noise process $\epsilon_s = \sqrt{G_s(H)}z_s$ with Hurst index $H \geq 1/2$, one can verify that $\nu_t^4, \kappa_t^4, \sigma_t^4$ is $\mathcal{O}(n_t^{H-1})$ in Assumptions 1-p to 3-p using the self-similarty property (Nourdin, 2012).

**Misspecified case.** Next, assume that the underlying data generating process follows the MA(1)-process, $X_s = \epsilon_s + \phi^*\epsilon_{s-1}$, with $\phi^* \in \mathbb{R}$. The misspecification error of fitting an AR(1) model to a MA(1) process can be found by minimizing

$$\begin{aligned}L(\theta) =& \mathbb{E}[(X_s - \theta X_{s-1})^2] = \mathbb{E}[(\epsilon_s + \phi^*\epsilon_{s-1} - \theta(\epsilon_{s-1} + \phi^*\epsilon_{s-2}))^2] \\ =& \mathbb{E}[(\epsilon_s + (\phi^* - \theta)\epsilon_{s-1} - \theta\phi^*\epsilon_{s-2})^2] = \sigma_\epsilon^2(1 + (\phi^* - \theta)^2 + \theta^2(\phi^*)^2),\end{aligned}$$

where $\nabla_\theta L(\theta) = 2(\theta - \phi^*)\sigma_\epsilon^2 + 2\theta(\phi^*)^2\sigma_\epsilon^2$. Thus, as $\theta^* = \arg\min_\theta L(\theta) \equiv \arg\min_\theta(\phi^* - \theta)^2 + \theta^2(\phi^*)^2$ is a strictly convex function in $\theta$, we have $\nabla_\theta L(\theta) = 0 \Leftrightarrow 2(\theta - \phi^*) + 2\theta(\phi^*)^2 = 0 \Leftrightarrow 2\theta(1 + (\phi^*)^2) = 2\phi^* \Leftrightarrow \theta = \phi^*/(1 + (\phi^*)^2)$. This means for any $\phi^* \in \mathbb{R}$ then $\theta \in (-1/2, 1/2)$. With this in mind, we can conduct our study of fitting an AR(1) model to the MA(1) process with $\phi^*$ drawn randomly from $\mathbb{R}$ (figure 1b). Furthermore, this reparametrization trick can be used to verify Assumption 1-p: first, we can reparameterize $\nabla_\theta L(\theta) = 2\sigma_\epsilon^2(\theta - \theta^*)(1 + (\phi^*)^2)$ using $\theta^* = \phi^*/(1 + (\phi^*)^2)$. Next, for $\mathbb{E}[\nabla_\theta l_t(\theta)|\mathcal{F}_{t-1}]$ one have that

$$\mathbb{E}[\nabla_\theta l_t(\theta)|\mathcal{F}_{t-1}] = \frac{2\theta}{n_t}\sum_{i=1}^{n_t} \mathbb{E}[X^2_{N_{t-1}+i-1}|\mathcal{F}_{t-1}] - \frac{2}{n_t}\sum_{i=1}^{n_t} \mathbb{E}[X_{N_{t-1}+i-1}X_{N_{t-1}+i}|\mathcal{F}_{t-1}],$$

where

$$\begin{aligned}\sum_{i=1}^{n_t} \mathbb{E}[X^2_{N_{t-1}+i-1}|\mathcal{F}_{t-1}] =& X^2_{N_{t-1}} + \mathbb{E}[X^2_{N_{t-1}+1}|\mathcal{F}_{t-1}] + \cdots + \mathbb{E}[X^2_{N_{t-1}+n_t-1}|\mathcal{F}_{t-1}] \\ =& X^2_{N_{t-1}} + \sigma_\epsilon^2 + (\phi^*)^2\epsilon^2_{N_{t-1}} + \cdots + \sigma_\epsilon^2 + (\phi^*)^2\sigma_\epsilon^2 \\ =& X^2_{N_{t-1}} + (\phi^*)^2\epsilon^2_{N_{t-1}} + \sigma_\epsilon^2(n_t - 1) + (\phi^*)^2\sigma_\epsilon^2(n_t - 2) \\ =& X^2_{N_{t-1}} + (\phi^*)^2(\epsilon^2_{N_{t-1}} - \sigma_\epsilon^2) + (1 + (\phi^*)^2)\sigma_\epsilon^2(n_t - 1),\end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^{n_t} \mathbb{E}[X_{N_{t-1}+i-1}X_{N_{t-1}+i}|\mathcal{F}_{t-1}] =& \phi^* X_{N_{t-1}}\epsilon_{N_{t-1}} + \phi^*\sigma_\epsilon^2(n_t - 1) \\ =& \theta^*(1 + (\phi^*)^2)X_{N_{t-1}}\epsilon_{N_{t-1}} + \theta^*(1 + (\phi^*)^2)\sigma_\epsilon^2(n_t - 1),\end{aligned}$$

using the same white noise properties as for the well-specified case above. This yields,

$$\mathbb{E}[\|\mathbb{E}[\nabla_\theta l_t(\theta)|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta)\|^2] = \frac{4(\theta-\theta^*)^2}{n_t^2} f_{\phi^*}(\epsilon_{N_{t-1}}),$$

where $f_{\phi^*}(\epsilon_{N_{t-1}})$ is finite function depending on the moments of $(\epsilon_{N_{t-1}})$ and $\phi^*$. Hence, we have $D_\nu > 0$ and $B_\nu = 0$ with $\nu_t$ being $\mathcal{O}(n_t^{-1})$. Similarly, it can be verified that $\kappa_t$ and $\sigma_t$ are $\mathcal{O}(n_t^{-1/2})$ by use of the reparametrization trick (Assumptions 2-p and 3-p).