Learning from time-dependent streaming data with online stochastic algorithms

Antoine Godichon-Baggioni, Nicklas Werge, Olivier Wintenberger

Learning from time-dependent streaming data with online stochastic algorithms

Antoine Godichon-Baggioni, Nicklas Werge*, Olivier Wintenberger

LPSM, Sorbonne Université, 4 place Jussieu, 75005 Paris, France

Abstract

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

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

1. Introduction

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

Streaming data arrives as an endless sequence of samples (data points), which means that at any given time, the model must be able to adapt to the samples observed (so far) to predict/label new samples accurately. Such (streaming) models can never be seen as complete but must be updated continuously as newer samples arrive. Methods that recalculate the model from scratch on the arrival of new samples are impractical due to their high computational cost. Therefore we need procedures that effectively update the model as more samples arrive. This computational efficiency should not be at the expense of accuracy; the model's accuracy should be close to that achieved if we built a model from scratch using all the samples (Bottou and Cun, 2003).

Stochastic approximation algorithms have proven effective in overcoming the drawbacks of traditional (batch/offline) machine learning methods as they only use samples one by one without knowing their number in advance, especially the Stochastic Gradient (SG) method (Robbins and Monro, 1951). These SG methods have proven scalable and robust in many areas ranging from smooth and strongly convex problems to complex non-convex ones, which makes them applicable in many large-scale machine learning tasks for real-world applications where data are large in size (and dimension) and arrive at a high velocity. Such first-order methods have been intensively studied in theory and practice in recent years (Bottou et al., 2018).

*Corresponding author

Email addresses: antoine.godichon_baggioni@upmc.fr (Antoine Godichon-Baggioni), nicklas.werge@upmc.fr (Nicklas Werge), olivier.wintenberger@upmc.fr (Olivier Wintenberger)

The classical analyses for SG methods typically require unbiased gradients drawn independently and identically distributed (i.i.d.) from some underlying (and unknown) data generation process (Cesa-Bianchi et al., 2004). However, in practice, learning often happens with non-i.i.d. (and biased) data, e.g., network traffic, meteorological, financial time series, or other sensor data. We go beyond these standard assumptions by allowing dependent and biased gradients. SG methods can converge even when they only have access to biased gradients, but most analysis has been developed with specific applications in mind (Ajalloeian and Stich, 2020; Bertsekas, 2016; d'Aspremont, 2008; Devolder et al., 2011; Schmidt et al., 2011). Stochastic learning algorithms for non-i.i.d. data are not as well understood as for i.i.d. data; however, some researchers have examined the convergence of statistical learning algorithm in non-i.i.d. settings (Agarwal and Duchi, 2012; Mohri and Rostamizadeh, 2010; Yu, 1994).

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

Contributions. In this paper, we investigate SG methods in a streaming framework (Godichon-Baggioni et al., 2021), where the data comes from dependent stochastic processes. We provide non-asymptotic analysis and quantify the magnitude of achievable convergence rates under various dependency structures and convexity levels. Our framework covers many applications with dependence and biased gradients under weak gradient assumptions. Our results construct a heuristic between the level of dependency, noise, and convexity and the achievable learning rate to obtain optimal convergence. Generally, SG methods can achieve convergence using non-decreasing (streaming) batch sizes, which counteract the long-range (and short-range) dependence and model misspecification. We show that biased SG methods converge with the same accuracy as unbiased SG methods if the bias is not too large. More surprisingly, these heuristics can be used in practice to help increase the stability of SG-based methods.

Organization. Section 2 presents the streaming framework on which the non-asymptotic analysis relies; we introduce some key concepts, definitions, and assumptions. In particular, Section 2.2 contains the assumptions about dependency structures and gradients, with some examples of how these could be verified using mixing conditions. Our convergence results are presented in Section 3, with and without averaging (Sections 3.1 and 3.2). Each result is followed by a thorough discussion that relates to other work. All our convergence analysis depends on the assumptions in Section 2 and some additional conditions for the averaged case (Section 3.2). At last, experimentations of our findings are illustrated in Section 4, with some final remarks in Section 5.

2. Problem Formulation

We consider the Stochastic Optimization (SO) problem $\min_{\theta\in\Theta} L(\theta) = \mathbb{E}_t[l_t(\theta)]$, where $\Theta \subseteq \mathbb{R}^d$ is a convex body[1] and $l_t : \mathbb{R}^d \to \mathbb{R}$ is some differentiable random functions (possibly non-convex), e.g, see Nesterov et al. (2018). We solve the SO problem in a streaming framework, where a *block* $l_t = (l_{t,1}, \dots, l_{t,n_t})$ of $n_t \in \mathbb{N}$ random functions arrives at any given time $t \in \mathbb{N}$. In solving the SO problem, we use the Stochastic Streaming Gradient (SSG) estimate proposed by Godichon-Baggioni et al. (2021), given as

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

where γ_t is the learning rate satisfying the conditions $\sum_{i=1}^{\infty} \gamma_i = \infty$ and $\sum_{i=1}^{\infty} \gamma_i^2 < \infty$ (Robbins and Monro, 1951). Note that if $\forall t, n_t = 1$, SSG becomes the well-known SG method, which has attracted a lot of attention (Bousquet and Elisseeff, 2002; Hardt et al., 2016; Shalev-Shwartz et al., 2011; Xiao, 2009; Zhang, 2004). Almost surely convergence of SO algorithms were shown in Pelletier (1998). In many models, there may be constraints on the parameter space, which would require a projection of the parameters; therefore, we also introduce the Projected Stochastic Streaming

[1] A convex body in $\mathbb{R}^d$ is a compact convex set with non-empty interior.

Gradient (PSSG) estimate, defined by

$$\theta_t = \mathcal{P}_\Theta\left(\theta_{t-1} - \frac{\gamma_t}{n_t}\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}(\theta_{t-1})\right), \quad \theta_0 \in \Theta, \tag{2}$$

where $\mathcal{P}_\Theta$ denotes the Euclidean projection onto Θ, i.e., $\mathcal{P}_\Theta(\theta) = \arg\min_{\theta'\in\Theta}\|\theta - \theta'\|_2$. To shorten notation, we let $\nabla_\theta l_t(\theta) = n_t^{-1}\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}(\theta)$. An essential extension is the Polyak-Ruppert averaging (Polyak and Juditsky, 1992; Ruppert, 1988), which guarantees optimal statistical efficiency without jeopardizing the computational cost; the Averaged Stochastic Streaming Gradient (ASSG) is given by

$$\bar{\theta}_t = \frac{1}{N_t}\sum_{i=0}^{t-1} n_{i+1}\theta_i, \quad \bar{\theta}_0 = 0, \tag{3}$$

where $N_t = \sum_{i=1}^t n_i$ is the accumulated sum of observations. Likewise, let PASSG denote the (Polyak-Ruppert) averaged estimate of PSSG (2).

2.1. Quasi-strong Convex Objectives

Following Gower et al. (2019); Moulines and Bach (2011), we assume that L has a unique global minimizer $\theta^* \in \Theta$ such that $\nabla_\theta L(\theta^*) = 0$, and it is μ-quasi-strongly convex (Karimi et al., 2016; Necoara et al., 2019), i.e, there exists $\mu > 0$ such that $\forall\theta \in \Theta$,

$$L(\theta^*) \geq L(\theta) + \langle\nabla_\theta L(\theta), \theta - \theta^*\rangle + \frac{\mu}{2}\|\theta - \theta^*\|^2. \tag{4}$$

The μ-quasi-strongly convexity assumption is a non-strongly convex relaxation of the SO problem, which is more conservative than μ-strongly convexity. Relaxations of convexity is crucial in practice to ensure robustness and adaptiveness of the algorithms, e.g., for non-strongly convex SO, see Bach and Moulines (2013); Necoara et al. (2019); Nemirovski et al. (2009).

2.2. Stochastic Streaming Gradient Assumptions: Dependence, Biased, Expected Smoothness, and Gradient Noise

We go beyond the classical assumptions that require unbiased (uniformly bounded) gradients by allowing the gradients to be dependent and biased estimates. Our aim is to non-asymptotically bound the SSG estimates (1) to (3) explicitly using the SO problem parameters. In order to do this, we let the natural filtration of the SO problem $\mathcal{F}_t = \sigma(l_i : i \leq t)$, and assume the following about the gradients $(\nabla_\theta l_t)$:

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

$$\mathbb{E}[\|\mathbb{E}[\nabla_\theta l_t(\theta)|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta)\|^p] \leq \nu_t^p(D_\nu^p\mathbb{E}[\|\theta - \theta^*\|^p] + B_\nu^p), \tag{5}$$

for some positive sequence $(\nu_t)_{t\geq1}$ with $D_\nu, B_\nu \geq 0$.

In the classical convergence analysis of SG methods, one assumes that the SGs are uniformly bounded. However, this assumption is too restrictive as it only may hold for some losses (Bottou et al., 2018; Nguyen et al., 2018). Instead, we follow the same ideas as in Gower et al. (2019); Moulines and Bach (2011), to make the following assumption about the expected smoothness of the stochastic gradients $(\nabla_\theta l_t)$.

Assumption 2-p (κ_t-expected smoothness)**.** *There exists a positive integer p such that $\forall\theta, \theta' \in \Theta$, $\mathbb{E}[\|\nabla_\theta l_t(\theta) - \nabla_\theta l_t(\theta')\|^p] \leq \kappa_t^p\mathbb{E}[\|\theta - \theta'\|^p]$ for some positive sequence $(\kappa_t)_{t\geq1}$.*

Assumption 2-p can be seen as an assumption about the smoothness properties of (l_t). The last fundamental assumption (Assumption 3-p) is a very weak assumption, and should be seen as an assumption on Θ rather than on (l_t):

Assumption 3-p (σ_t-gradient noise)**.** *There exists a positive integer p such that* $\mathbb{E}[\|\nabla_\theta l_t(\theta^*)\|^p] \le \sigma_t^p$ *for some positive sequence* $(\sigma_t)_{t\ge1}$.

These assumptions (Assumptions 1-p to 3-p) are milder than the standard assumptions for stochastic approximations, e.g., see Benveniste et al. (2012); Godichon-Baggioni et al. (2021); Kushner and Yin (2003); Moulines and Bach (2011). They include classic examples such as stochastic approximation and learning from 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 ν_t. It is possible to verify Assumption 1-p by using 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

$$\mathbb{E}[\|\mathbb{E}[\nabla_\theta l_t(\theta)|\mathcal{F}_{t-1}] - \nabla_\theta L(\theta)\|^p] \le \mathbb{E}[\|\nabla_\theta l_t(\theta) - \nabla_\theta L(\theta)\|^p] = n_t^{-p}\mathbb{E}[\|S_t\|^p], \tag{6}$$

using Jensen's inequality, where $S_t = \sum_{i=1}^{n_t}(\nabla_\theta l_{t,i}(\theta) - \nabla_\theta L(\theta))$ is a d-dimensional vector. Let $(\nabla_\theta l_{t,i})$ be a strictly stationary sequence and assume that there exists some $r > p$ such that $\sup_{x>0}(x^r Q(x))^{1/r} < \infty$, where $Q(x)$ denotes the quantile function of $\|\nabla_\theta l_{t,i}\|$. Suppose that $(\nabla_\theta l_{t,i})$ is strongly α-mixing in the sense of Rosenblatt (1956), with strong mixing coefficients $(\alpha_t)_{t\ge1}$ satisfying $\alpha_t = O(t^{-pr/(2r-2p)})$. Then by Rio (2017, Corollary 6.1), we have that $\mathbb{E}[\|S_t\|^p] = O(n_t^{p/2})$, meaning, (6) is at most $O(n_t^{-p/2})$; 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 well-specified examples, which we will see later in Section 4. Note that Assumptions 2-p and 3-p can be verified using α-mixing conditions by analogues arguments as for Assumption 1-p such that κ_t^p and σ_t^p is $O(n_t^{-p/2})$.

3. Convergence Analysis

In this section, we consider the stochastic streaming estimates in (1) to (3) with streaming-batches (n_t) arriving in non-decreasing streams. We aim to non-asymptotically bound $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ and $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$, such that they only depend on the parameters of the problem.

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

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

3.1. Stochastic Streaming Gradients

Theorem 1. *Denote* $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ *for some* $\delta_0 \ge 0$*, where* (θ_t) *follows the recursion in* (1) *or* (2)*. Assume that Assumptions 1-p to 3-p hold true for* $p = 2$*. Suppose* $n_t = C_\rho t^\rho$ *with* $\rho \in [0,1)$ *and* $C_\rho \in \mathbb{N}$*, such that* $\mu_\nu = \mu - \mathbb{1}_{\{\rho=0\}} 2D_\nu C_\rho^{-\nu} > 0$*. For* $\alpha - \rho\beta \in (1/2,1)$*, we have*

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

with π_t *given in* (22) *such that* $\pi_t = O(\exp(-N_t^{(1+\rho\beta-\alpha)/(1+\rho)}))$.

[2] You can't beat the system.

Sketch of proof. Under Assumptions 1-p to 3-p with $p = 2$, it can be shown that (δ_t) satisfies the recursive relation (20),

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

for any γ_t, ν_t, κ_t, σ_t, and n_t. This recursive relation can be explicitly upper bounded in a non-asymptotic way (by Proposition 1) using classical techniques from stochastic approximations (Benveniste et al., 2012; Kushner and Yin, 2003). As mentioned in Zinkevich (2003), bounding the projected estimate in (2) follows directly from that $\mathbb{E}[\|\mathcal{P}_\Theta(\theta) - \theta^*\|^2] \leq \mathbb{E}[\|\theta - \theta^*\|^2]$, $\forall \theta \in \mathbb{R}^d$, $\forall \theta^* \in \Theta$, as Θ is a convex body.

Related work. Theorem 1 replicate the results of the well-specified i.i.d. case (with $B_\nu = 0$ and $\kappa = \sigma = 1/2$) considered in Godichon-Baggioni et al. (2021). Our findings also reproduce the results of Moulines and Bach (2011), where they considered the well-specified i.i.d. case (under slightly different assumptions) using the vanilla SG method, namely, when $C_\rho = 1$ and $\rho = 0$. Moreover, if the function L has C_∇-Lipschitz continuous gradients[3], then (7) implies the bound on the objective function values of L, $\mathbb{E}[L(\theta_t) - L(\theta^*)] \leq C_\nabla \delta_t/2$ by Cauchy–Schwarz's inequality.

Decay of the initial conditions. The initial conditions that π_t contains will be forgotten sub-exponentially fast, since $\pi_t = O(\exp(-N_t^{(1+\rho\beta-\alpha)/(1+\rho)}))$ as long as $\mu_\nu = \mu - \mathbb{1}_{\{\rho=0\}} 2D_\nu C_\rho^{-\nu} > 0$. Note that the positivity of the dependence penalised convexity constant μ_ν is essential in all terms of (7). Having $\mu_\nu > 0$ depends solely on the level of dependence D_ν but it is scaled by $C_\rho^{-\nu}$, meaning if D_ν is so large that μ_ν is no longer positive, then we should take C_ρ large enough such that μ_ν becomes positive again; this is illustrated in Sections 4.2 and 4.3. The streaming constant C_ρ contributes positively to all terms in (7), either directly or though μ_ν.

The last term of (7) can be seen as the noise term decaying with $O(N_t^{-(\rho(2\sigma-\beta)+\alpha)/(1+\rho)})$ for $\alpha - \rho\beta \in (1/2, 1)$, e.g., for any $\rho \in [0, 1)$, $\delta_t = O(N_t^{-2/3})$ when $\alpha = 2/3$, $\beta = 1/3$, and $\sigma = 1/2$. In addition, the noise term is positively affected by large streaming constants C_ρ when $\alpha + \beta < 2\sigma$, which will be expressed as a variance reduction, e.g., see Section 4. In well-specified cases ($B_\nu = 0$) the noise term would also be the asymptotic term.

Behavior for B_ν. The second term of (7) can be seen as an dependency term as it is determined solely by the level of dependence ν, the bias (misspecification error) B_ν, and the convexity constant μ_ν; It is remarkable that the dependence term is unconnected from the choice of the learning rate (γ_t) but instead by the streaming rate through C_ρ and ρ. The dependence term decay with $O(N_t^{-2\rho\nu/(1+\rho)})$ which requires ρ positive to decay since $\nu \in (0, \infty)$, e.g., to obtain $O(N_t^{-1/2})$ we would need $\rho = 1$ and $\nu = 1/2$. It is surprising that Theorem 1 allows both long-range and short-range dependence. Indeed, long-range dependence leads to slow convergence (slower than $O(N_t^{-1/2})$) but it will still converge. Obviously, this only matters if $B_\nu \neq 0$. Overall, $\delta_t = O(\max\{\mathbb{1}_{\{B_\nu \neq 0\}} N_t^{-2\rho\nu/(1+\rho)}, N_t^{-(\rho(2\sigma-\beta)+\alpha)/(1+\rho)}\})$.

3.2. Averaged Stochastic Streaming Gradients

In what follows, we consider the averaging estimate $\bar{\theta}_n$ given in (3) with (θ_t) following the SSG estimate in (1) or the PSSG estimate in (2). Some additional assumptions is needed for bounding the *rest* terms of the averaging estimate: let the function L have C_∇-Lipschitz continuous gradients, i.e., there exists a constant $C_\nabla > 0$, $\forall \theta, \theta' \in \Theta \subseteq \mathbb{R}^d$,

$$\|\nabla_\theta L(\theta) - \nabla_\theta L(\theta')\| \leq C_\nabla \|\theta - \theta'\|. \tag{8}$$

As discussed in Bottou et al. (2018), this assumption ensures that $\nabla_\theta L$ does not vary arbitrarily, making the gradient $\nabla_\theta L$ a useful indicator on how to decrease L. Next, assume that the Hessian of L is C'_∇-Lipschitz-continuous, that is, there exists $C'_\nabla > 0$ such that $\forall \theta, \theta' \in \Theta \subseteq \mathbb{R}^d$,

$$\|\nabla_\theta^2 L(\theta) - \nabla_\theta^2 L(\theta')\| \leq C'_\nabla \|\theta - \theta'\|. \tag{9}$$

Note that (8) and (9) only needs to hold true for $\theta' = \theta^*$. Moreover, 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:

[3] Later, in Section 3.2 for the averaged estimate (3), we assume in (8) that the function L has C_∇-Lipschitz continuous gradients.

Assumption 4. *There exists a non-negative self-adjoint operator* Σ *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* Σ_t *is a positive symmetric matrix with* $Tr(\Sigma_t) = C'_\sigma n_t^{-2\sigma'}$, $C'_\sigma \geq 0$, *and* $\sigma' \in (0, 1/2]$.

Remark that in the independent or some well specified cases such as in Section 4.1.1, Assumption 4 is verified with $\sigma = 1/2$ and $C'_\sigma = 0$ (Godichon-Baggioni et al., 2021). The short-range dependence cases is when $\sigma = 1/2$, as in Section 4.1.2, whereas, the long-range dependence case is for $\sigma < 1/2$. Moreover, Assumption 4 allows us to obtain leading term Λ/N_t with $\Lambda = \mathrm{Tr}(\nabla_\theta^2 L(\theta^*)^{-1}\Sigma\nabla_\theta^2 L(\theta^*)^{-1})$, which attains the Cramer-Rao bound; we will see this in Theorem 2.

To consider the averaging estimate $\bar{\theta}_n$ given in (3) but with the use of the projected estimate PSSG from (2), which we will denote PASSG. An additional assumption is needed in order to avoid calculating the six-order moment, we make the unnecessary assumption that $(\nabla_\theta l_t)$ is uniformly bounded; the derivation of the six-order moment can be found in Godichon-Baggioni (2016).

Assumption 5. *Let* $D_\Theta = \inf_{\theta\in\partial\Theta}\|\theta - \theta^*\| > 0$ *with* $\partial\Theta$ *denoting the frontier of* Θ. *Moreover, there exists* $G_\Theta > 0$ *such that* $\forall t \geq 1$, $\sup_{\theta\in\Theta}\|\nabla_\theta l_t(\theta)\|^2 \leq G_\Theta^2$ *a.s.*

Theorem 2. *Denote* $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ *with* $\bar{\theta}_n$ *given by* (3), *where* (θ_t) *follows the recursion in* (1) *or* (2). *Assume that Assumptions 1-p to 3-p for* $p = 4$ *and Assumption 4 hold true. In addition, Assumption 5 must hold true if* (θ_t) *follows the recursion in* (2). *Suppose* $n_t = C_\rho t^\rho$ *with* $\rho \in [0, 1)$ *and* $C_\rho \in \mathbb{N}$, *such that* $\mu_\nu = \mu - \mathbb{1}_{\{\rho=0\}}2D_\nu C_\rho^{-\nu} > 0$. *For* $\alpha - \rho\beta \in (1/2, 1)$, *we have*

$$\bar{\delta}_t^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}}\mathbb{1}_{\{\sigma=1/2\}} + \frac{2^{1/2}\Lambda^{1/2}C_\rho^{\frac{1-2\sigma}{2(1+\rho)}}}{N_t^{\frac{1+2\rho\sigma}{2(1+\rho)}}}\mathbb{1}_{\{\sigma\neq1/2\}} + \frac{2^{1/2}C_\sigma'^{1/2}C_\rho^{\frac{1-2(\sigma+\sigma')}{2(1+\rho)}}}{\mu N_t^{\frac{1+2\rho(\sigma+\sigma')}{2(1+\rho)}}} \tag{10}$$

$$+ O\left(\max\left\{N_t^{-\frac{2+\rho(2\sigma+\beta)-\alpha}{2(1+\rho)}}, N_t^{-\frac{\rho(2\sigma-\beta)+\alpha}{1+\rho}}\right\}\right) + \tilde{O}\left(N_t^{-\frac{\delta+\rho\nu}{2(1+\rho)}}\right) + \mathbb{1}_{\{B_\nu\neq0\}}\Psi_t, \tag{11}$$

with $\delta = \mathbb{1}_{\{B_\nu=0\}}(\rho(2\sigma-\beta)+\alpha) + \mathbb{1}_{\{B_\nu\neq0\}}\min\{\rho(2\sigma-\beta)+\alpha, 2\rho\nu\}$ *and* Ψ_t *given in* (36), *such that*

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

An explicit version of the bound is given in (37).

Sketch of proof. In Lemma 3, we conduct a general study of the Polyak-Ruppert averaging estimate $(\bar{\theta}_t)$ defined in (3) for (γ_t), (ν_t), (κ_t), (σ_t) and (n_t) on any form. Thus, Theorem 2 follows by Lemma 3 using the (specific) bounds of $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ and $\Delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^4]$ in Theorem 1 (eq. (21)) and Lemma 2.

Related work. Similarly to the well-specified i.i.d. case (Godichon-Baggioni et al., 2021), the leading term of (10) is Λ/N_t, which obtain the (asymptotically optimal) Cramer-Rao bound (Murata and Amari, 1999). Each term of (10) is a direct consequence of Assumption 4 and they are all independent of the choice of learning rate (γ_t). Moreover, as discussed in Gadat and Panloup (2017), the bound of $\bar{\delta}_t$ can be seen as a bias-variance decomposition between the leading terms (10) and the remaining terms in (11).

Accelerated decay. By averaging it is possible (in some specific cases) to achieve the desirable Cramer-Rao bound, namely, the leading term Λ/N_t could obtain the optimal and incorrigible rate of $O(N_t^{-1})$. This is always achieved in the well-specified case with $\sigma = 1/2$, even under short-range dependence (i.e., when $\nu > 1/2$).

As for Theorem 1, the positivity of μ_ν is essential for all terms in (11) even if it does not appear directly. In case of lack of convexity μ or high levels of dependence constant D_ν, we can only ensure convergence by increasing C_ρ, i.e., ensuring positivity of μ_ν; this is illustrated in Sections 4.2 and 4.3 for ARCH models.

The first term of (11) decay at the rate $O(N_t^{-(2+\rho(\beta+2\sigma)-\alpha)/(1+\rho)})$ or $O(N_t^{-2(\rho(2\sigma-\beta)+\alpha)/(1+\rho)})$, which suggests choosing α, β such that $\alpha + \rho(2\sigma/3 - \beta) = 2/3$, e.g., $\alpha = 2/3, \beta = 1/3$ and $\sigma = 1/2$ yields a decay of $O(N_t^{-4/3})$ for any ρ. Thus, we can robustly achieve $O(N_t^{-4/3})$ for any streaming rate ρ by setting $\alpha = 2/3$ and $\beta = 1/3$ if $\sigma = 1/2$. In general, the convergence is resilient to any streaming rate ρ by having $\alpha = 2/3$ and $\beta = 2\sigma/3$. But taking $\beta > 0$ would damage the variance reduction effect from having C_ρ large (e.g., see discussion after Theorem 1). Thus, there is a trade-off between accelerating the convergence by taking $\beta = 2\sigma/3 > 0$ or taking $\beta = 0$ to favor from variance reduction. In practice, an immediate choice would be to take $\beta = 0$, but if the data or model contains a low amount of noise, it can be advantageous to raise β to improve convergence (Godichon-Baggioni et al., 2021).

Next, the decay of the second term in (11) is tricky to interpret in a simple manner as it is a mixture of the learning rate, streaming rate, dependence, and bias (misspecification error). Nevertheless, some observations can be made: first, having $\beta = 0$ is beneficial for the decay rate δ in all cases. Similarly, increasing streaming rate ρ would also increase the decay. The most important thing to mention is that if $B_\nu \neq 0$ then we would at most have a decay of $O(N_t^{-2/3})$, which is the same as for the SSG in Theorem 1.

Behavior for B_ν. The influence of B_ν is exclusively contained in Ψ_t, with the exception of the second term of (11). Also, increasing ρ will always diminish the bad influence of this bias term. Surprisingly, $\Psi_t \to 0$ as $t \to \infty$ for any ν, but long-range dependence is excluded if we wish to obtain the desired rate of $\bar{\delta}_t = O(N^{-1})$. However, it does not seem to have any major influence in our experiments in Section 4.

4. Experiments

A way to illustrate our findings is by use of classical methods that aims to construct models for time-series analysis, modeling, and prediction of the underlying sequences of real-valued signals (X_t). These methods have been successfully used in a wide range of applications such as statistics, econometrics, and signal processing because of their ability to describe or predict time-varying (dependent) processes, e.g., the AutoRegressive (AR), Moving-Average (MA), and AutoRegressive Moving-Average (ARMA) models are the most well-known models for time-series (Box et al., 2015; Brockwell and Davis, 2009; 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 and Duchi (2012); Anava et al. (2013); Wintenberger (2021).

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

4.1. *AR model*

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

4.1.1. *Well-specified case*

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

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

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

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

meaning that Assumption 1-p is fulfilled if X_t has bounded moments of order p. Moreover, from this we can directly deduce that $B_\nu = 0$. Likewise, the remaining assumptions can be verified, in particular Assumption 4 is satisfied with $\Sigma_t = 0$.

4.1.2. Misspecified case

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

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

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

4.2. 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 (ϵ_t) is called an ARCH(1) process with parameters α_0 and α_1 if it satisfies

$$\begin{cases}\epsilon_t = \sigma_t z_t, \\ \sigma_t^2 = \alpha_0 + \alpha_1\epsilon_{t-1}^2,\end{cases} \tag{12}$$

where $\alpha_0 > 0$ and $\alpha_1 \geq 0$ ensures the non-negativity of the conditional variance process (σ_t^2), and the innovations (z_t) is white noise. The ARCH process parameters are known to be challenging to estimate in empirical applications as the optimization algorithms can quickly fail or converge to irregular solutions. Therefore, projecting the estimates is vital for the optimization procedure. A well-discussed problem for the ARCH models is that small values of α_0 are tricky to estimate. Stabilizing the estimation of α_0 would not only improve the α_0 estimate but also have a positive impact on the other model parameters. One way to deal with small values of α_0 is by the using the models homogeneity, i.e., scaling an ARCH process (X_t) with parameters (α_0, α_1) gives us an ARCH process $(\sqrt{\lambda}X_t)$ with parameters $(\lambda\alpha_0, \alpha_1)$ with same innovations. To simplify our analysis we consider a stationary ARCH(1) model, where we fix α_0 at 1 and initialize it at $1/2$. We employ the quasi-maximum likelihood procedure for the statistical inference as outlined in Werge and Wintenberger (2022); the quasi likelihood losses is given by $l_t(\theta) = 2^{-1}(X_t^2/\sigma_t^2(\theta) + \log(\sigma_t^2(\theta))$ with first-order derivative

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

where $\nabla_\theta \sigma_t^2(\theta) = (1, X_{t-1}^2)^T$. Observe that the loss function (l_t) itself is not strongly convex but only the objective function L may be strongly convex; convexity conditions of ARCH was investigated in Wintenberger (2021). There are different ways to overcome lack of convexity: first, projecting the estimates such that the (conditional) variance process (σ_t^2) stays away from zero (and close to the unconditional variance). Second, in the specific example of ARCH model, one could also recover convexity by implementing variance targeting techniques; an example using Generalized ARCH (GARCH) models can be found in Werge and Wintenberger (2022).

4.3. AutoRegressive (AR)-AutoRegressive Conditional Heteroskedasticity (ARCH) Model

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

$$\begin{cases}X_t = \theta X_{t-1} + \epsilon_t, \\ \epsilon_t = \sigma_t z_t, \\ \sigma_t^2 = \alpha_0 + \alpha_1\epsilon_{t-1}^2.\end{cases} \tag{13}$$

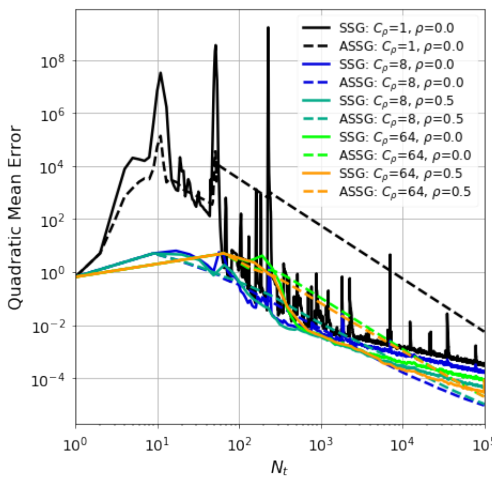
where the innovations (z_t) is white noise. The statistical inference of this model is done using the squared loss for the AR-part and the QMLE for the ARCH part, e.g., see Sections 4.1.1 and 4.2.

4.4. Discussion of experiments

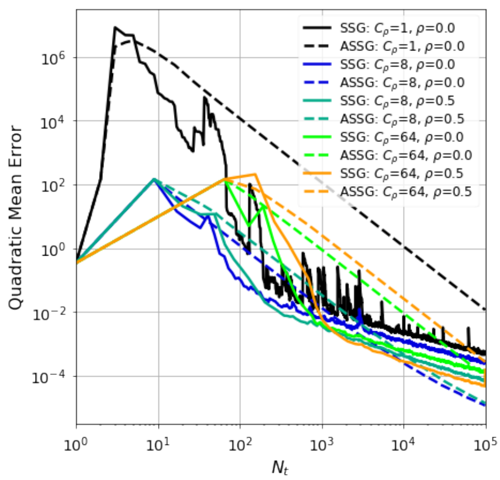
The experiments described earlier in Sections 4.1 to 4.3 can be found in Figure 1; here $\{C_\rho = 1, \rho = 0\}$ corresponds to the classical SG method and its (Polyak-Ruppert) average estimate, $\{C_\rho = 64, \rho = 0\}$ is a mini-batch SSG/ASSG, and $\{C_\rho = 64, \rho = 1/2\}$ is an increasing SSG/ASSG with initial batch size of $C_\rho = 64$.

Figure 1: Simulation of various data streams $n_t = C_\rho t^\rho$. See Section 4 for details.

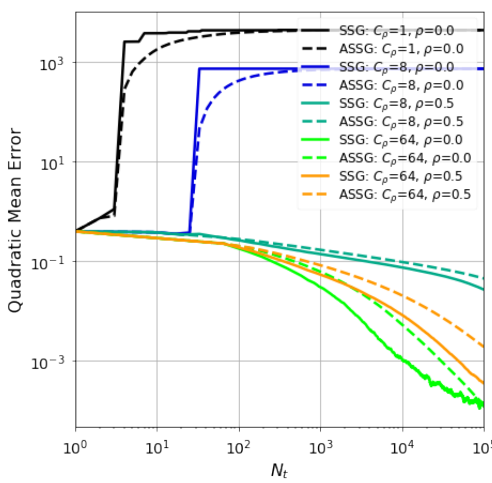
(a) AR(1): well-specified case. See Section 4.1.1 for details.



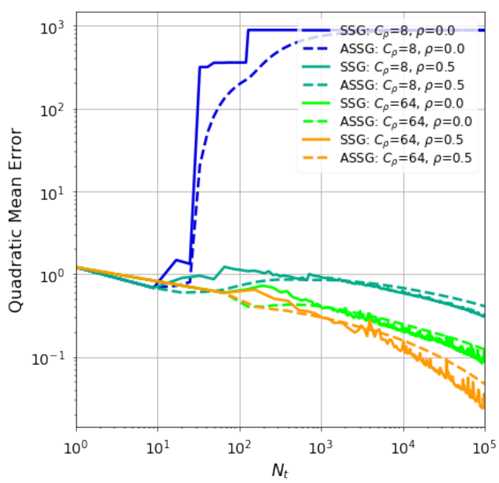
(b) AR(1): misspecified case. See Section 4.1.2 for details.



(c) ARCH(1). See Section 4.2 for details.



(d) AR(1)-ARCH(1). See Section 4.3 for details.



First consider the AR illustration in Figures 1a and 1b: each pair of data streams converges, but it is clear that the traditional SG method experiences a large amount of noise initially, particularly affecting the average estimate period but not its decay rate.[4] Both methods show a noticeable reduction in variance when C_ρ increases, which is particularly beneficial in the beginning. Nevertheless, too large streaming batch sizes C_ρ may hinder the convergence as this leads

[4] A modification of our average estimate to a weighted average version could improve convergence as it could limit the effect of poor initializations (Boyer and Godichon-Baggioni, 2020; Mokkadem and Pelletier, 2011). But despite this, we still achieve better convergence for the ASSG method.

to too few iterations. Moreover, Figures 1a and 1b indicates improving decay for SSG when increasing the streaming rate ρ. Conversely, ASSG does not see improvements in the same way, as we do not exploit the potential of using multiple observations through the β parameter, which could accelerate convergence, e.g., see Godichon-Baggioni et al. (2021) for a discussion in the (well-specified) i.i.d. case. It is surprising that we do not see any effect from Σ_t in Assumption 4, but this seems to be an artifact effect in the proof as we need fourth-order moments.

In Figures 1c and 1d, we have the experiments for the stationary ARCH(1) models, with and without the AR-part, respectively, as outlined in Sections 4.2 and 4.3. These figures illustrate the lack of convexity when using small streaming batch sizes, e.g., the traditional SG method and its average estimate, $\{C_\rho = 1, \rho = 0\}$ diverges. Remark that the lack of convexity is expressed through the lack positively of μ_ν, which only larger streaming batch sizes C_ρ can counteract. Moreover, the traditional SG method, $\{C_\rho = 1, \rho = 0\}$ is omitted in Figure 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. Conclusions

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

There are several ways to expand our work: first, we can extend our analysis to include streaming batches of any size (not in terms of streaming batch size and streaming rates). Second, an extension to non-strongly convex goals could be beneficial as it will provide more insight into how we can choose robust learning rates (Bach and Moulines, 2013; Necoara et al., 2019; Nemirovski et al., 2009). At the same time, this learning rate could be made adaptive such that it is robust to poor initialization and requires less fine-tuning. This last objective is the most important for practitioners as it builds a universality across applications.

6. Proofs

Let us start by giving a short sketch of how our proofs section is structured: we start by deriving recursive relations to the desired quantities. Next, we derive a general bounds to the recursive relationship for any (γ_t), (ν_t), (κ_t), (σ_t), and (n_t). Finally, we insert the specific functions forms of (γ_t), (ν_t), (κ_t), (σ_t), and (n_t), which yield the results seen in Theorems 1 and 2. Before doing the proofs, we recall a repeating argument used to non-asymptotically bound recursive relations of form (14):

Proposition 1 (Godichon-Baggioni et al. (2021))**.** *Suppose (ω_t), (α_t), (η_t), and (β_t) to be some non-negative sequences satisfying the recursive relation,*

$$\omega_t \leq [1 - 2\lambda\alpha_t + \eta_t\alpha_t]\omega_{t-1} + \beta_t\alpha_t, \tag{14}$$

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

$$\omega_t \leq \tau_t + \frac{1}{\lambda}\max_{t/2\leq i\leq t}\beta_i, \quad \textit{with} \quad \tau_t = \exp\left(-\lambda\sum_{i=t/2}^{t}\alpha_i\right)\left[\exp\left(C_\omega\sum_{i=1}^{t}\eta_i\alpha_i\right)\left(\omega_0 + \frac{1}{\lambda}\max_{1\leq i\leq t}\beta_i\right) + \sum_{i=1}^{t/2-1}\beta_i\alpha_i\right]. \tag{15}$$

Proposition 1 shows a simple way to bound (ω_t) in (14); the bound in (15) consists of a sub-exponential term τ_t and a noise term $\lambda^{-1}\max_{t/2\leq i\leq t}\beta_i$. Thus, our attention is on reducing the noise term without damaging the natural decay of the sub-exponential term where $\tau_t \to 0$ exponentially fast as $t \to \infty$.

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

$$\psi_x(t) = \begin{cases} t^{1-x}/(1-x) & \text{if } x < 1, \\ 1 + \log(t) & \text{if } x = 1, \\ x/(x-1) & \text{if } x > 1, \end{cases} \quad \text{and} \quad \psi_x^y(t) = \begin{cases} t^{(1-x)/(1+y)}/(1-x) & \text{if } x < 1, \\ 1 + \log(t^{1/(1+y)}) & \text{if } x = 1, \\ x/(x-1) & \text{if } x > 1, \end{cases} \tag{16}$$

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

6.1. Proofs for Section 3.1

In the following lemma, we derive an explicit recursive relation of $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ to non-asymptotically bound the t-th estimate of (1) for any (γ_t), (ν_t), (κ_t), (σ_t), and (n_t) using classical techniques from stochastic approximations (Benveniste et al., 2012; Kushner and Yin, 2003). As mentioned in Zinkevich (2003), bounding the projected estimate in (2) follows directly from that $\mathbb{E}[\|\mathcal{P}_\Theta(\theta) - \theta^*\|^2] \leq \mathbb{E}[\|\theta - \theta^*\|^2]$, $\forall \theta \in \mathbb{R}^d$, $\forall \theta^* \in \Theta$, as Θ is a convex body.

Lemma 1 (Second-order moment)**.** *Assume that Assumptions 1-p to 3-p hold true for $p = 2$. Suppose that $\mu_\nu = \mu - \mathbb{1}_{\{\nu_t = C\}} 2D_\nu \nu_t > 0$. Let $\mathbb{1}_{\{\nu_t = C\}}$ and $\mathbb{1}_{\{\nu_t \neg C\}}$ indicate whether (ν_t) is constant or not. Denote $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ for some $\delta_0 \geq 0$, where (θ_t) follows the recursion in (1) or (2). For any learning rate (γ_t), we have*

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

with

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

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

$$\delta_t = \delta_{t-1} + \gamma_t^2 \mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1})\|^2] - 2\gamma_t \mathbb{E}[\langle \nabla_\theta l_t(\theta_{t-1}), \theta_{t-1} - \theta^* \rangle], \tag{17}$$

where $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ with $\delta_0 \geq 0$. To bound the second term in (17), we use Assumptions 2-p and 3-p for $p = 2$, to obtain that

$$\begin{aligned}\mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1})\|^2] =& \mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1}) - \nabla_\theta l_t(\theta^*) + \nabla_\theta l_t(\theta^*)\|^2] \\ \leq& 2\mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1}) - \nabla_\theta l_t(\theta^*)\|^2] + 2\mathbb{E}[\|\nabla_\theta l_t(\theta^*)\|^2] \leq 2\kappa_t^2 \delta_{t-1} + 2\sigma_t^2,\end{aligned} \tag{18}$$

as $\|x+y\|^p \leq 2^{p-1}(\|x\|^p + \|y\|^p)$. As noted in Bottou et al. (2018); Nesterov et al. (2018), (4) implies that $\langle \nabla_\theta L(\theta), \theta - \theta^* \rangle \geq \mu\|\theta - \theta^*\|^2$ for all $\theta \in \Theta \subseteq \mathbb{R}^d$. Next, since L is $\mu-$strongly convex (4) and θ_{t-1} is $\mathcal{F}_{t-1}$-measurable (Assumption 1-p), we can bound the third term on the right-hand side of (17) by

$$\begin{aligned}\mathbb{E}[\langle \nabla_\theta l_t(\theta_{t-1}), \theta_{t-1} - \theta^* \rangle] =& \mathbb{E}[\langle \nabla_\theta L(\theta_{t-1}), \theta_{t-1} - \theta^* \rangle] + \mathbb{E}[\langle \mathbb{E}[\nabla_\theta l_t(\theta_{t-1}) | \mathcal{F}_{t-1}] - \nabla_\theta L(\theta_{t-1}), \theta_{t-1} - \theta^* \rangle] \\ \geq& \mu\delta_{t-1} - D_\nu \nu_t \delta_{t-1} - B_\nu \nu_t \delta_{t-1}^{1/2},\end{aligned} \tag{19}$$

since

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

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

by Jensen's inequality, Cauchy–Schwarz inequality, Hölder's inequality, and Assumption 1-p with $p = 2$. Hence, applying the inequalities (18) and (19) to (17), yields

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

using Young's inequality[5]; $2B_\nu\nu_t\gamma_t\delta_{t-1}^{1/2} \leq \mu\gamma_t\delta_{t-1} + B_\nu^2\nu_t^2\gamma_t/\mu$. Next, we introduce the indicator function for whether (ν_t) is constant ($= C$) or not ($\neg C$), such that

$$\delta_t \leq [1 - (\mu_\nu - \mathbb{1}_{\{\nu_t \neg C\}}2D_\nu\nu_t)\gamma_t + 2\kappa_t^2\gamma_t^2]\delta_{t-1} + \frac{B_\nu^2}{\mu}\nu_t^2\gamma_t + 2\sigma_t^2\gamma_t^2, \tag{20}$$

with $\mu_\nu = \mu - \mathbb{1}_{\{\nu_t = C\}}2D_\nu\nu_t > 0$. Let C_δ be the constant fulfilling the conditions of Proposition 1 such that C_δ is chosen larger than 1 verifying $C_\delta(\mathbb{1}_{\{\nu_t \neg C\}}2D_\nu\nu_t + 2\kappa_t^2\gamma_t) \leq \mu_\nu/2$ such that it's imply $\mu_\nu\gamma_t/2 \leq 1$, which is possible as the sequence (ν_t) is non-increasing, and (κ_t) and (γ_t) is decreasing. At last, bounding (20) 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

$$\begin{aligned}\delta_t \leq& \pi_t + \frac{2^{1+2\rho\nu}B_\nu^2}{\mu\mu_\nu C_\rho^{2\nu}t^{2\rho\nu}} + \frac{2^{2+\rho(2\sigma-\beta)+\alpha}C_\sigma^2 C_\gamma C_\rho^\beta}{\mu_\nu C_\rho^{2\sigma}t^{\rho(2\sigma-\beta)+\alpha}} \\ \leq& \pi_t + \frac{2^{(2+6\rho\nu)/(1+\rho)}B_\nu^2}{\mu\mu_\nu C_\rho^{2\nu/(1+\rho)}N_t^{2\rho\nu/(1+\rho)}} + \frac{2^{(7+6\rho\sigma)/(1+\rho)}C_\sigma^2 C_\gamma}{\mu_\nu C_\rho^{(2\sigma-\beta-\alpha)/(1+\rho)}N_t^{(\rho(2\sigma-\beta)+\alpha)/(1+\rho)}},\end{aligned} \tag{21}$$

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

$$\begin{aligned}&\exp\left(-\frac{\mu_\nu C_\gamma C_\rho^\beta}{2}\sum_{i=t/2}^t i^{\rho\beta-\alpha}\right)\left[\exp\left(\frac{\mathbb{1}_{\{\rho\neq0\}}2C_\delta D_\nu C_\gamma C_\rho^\beta}{C_\rho^\nu}\sum_{i=1}^t i^{\rho(\beta-\nu)-\alpha}\right)\exp\left(\frac{2C_\delta C_\kappa^2 C_\gamma^2 C_\rho^{2\beta}}{C_\rho^{2\kappa}}\sum_{i=1}^t i^{2\rho(\beta-\kappa)-2\alpha}\right)\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}{\mu C_\rho^{2\nu}}\sum_{i=1}^{t/2-1} i^{\rho(\beta-2\nu)-\alpha} + \frac{2C_\sigma^2 C_\gamma^2 C_\rho^{2\beta}}{C_\rho^{2\sigma}}\sum_{i=1}^{t/2-1} i^{2\rho(\beta-\sigma)-2\alpha}\right]\\
&\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\neq0\}}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\neq0\}}2C_\delta D_\nu C_\gamma C_\rho^\beta\psi_{\alpha-\rho(\beta-\nu)}^\rho(2N_t/C_\rho)}{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)}^\rho(N_t/C_\rho)}{\mu C_\rho^{2\nu}} + \frac{4(\alpha-\rho(\beta-\sigma))C_\sigma^2 C_\gamma^2 C_\rho^{2\beta}}{(2\alpha-2\rho(\beta-\sigma)-1)C_\rho^{2\sigma}}\right],\end{aligned} \tag{22}$$

with help of an integral test for convergence[6], $\psi_x(t)$ and $\psi_x^y(t)$ from (16), and by use of $(N_t/2C_\rho)^{1/(1+\rho)} \leq t \leq (2N_t/C_\rho)^{1/(1+\rho)}$. □

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

[6] 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 $\alpha - \rho\beta \in (1/2, 1)$.

6.2. Proofs for Section 3.2

As in Section 6.1, we begin the following sections by conducting a general study for any (γ_t), (ν_t), (κ_t), (σ_t), and (n_t), when applying the Polyak-Ruppert averaging estimate $(\bar{\theta}_t)$ from (3). Moreover, we need to study fourth-order rate $\Delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^4]$ of the recursive estimates (1) and (2).

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

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

with

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

Proof of Lemma 2. The derivation of the recursive step sequence for the fourth-order moment Δ_t of (1) follows the same methodology as for the second-order moment in Lemma 1. In the same way we deduced (17), we can take the quadratic norm on (1), expand the norm, and take the square on both sides, to derive the equation

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

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

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

by use of Cauchy-Schwarz inequality. Next, Young's inequality yields

$$\begin{aligned}
4\gamma_t^3 \|\theta_{t-1} - \theta^*\| \|\nabla_\theta l_t(\theta_{t-1})\|^3 \leq& 2\gamma_t^4 \|\nabla_\theta l_t(\theta_{t-1})\|^4 + 2\gamma_t^2 \|\theta_{t-1} - \theta^*\|^2 \|\nabla_\theta l_t(\theta_{t-1})\|^2, \\
8\gamma_t^2 \|\theta_{t-1} - \theta^*\|^2 \|\nabla_\theta l_t(\theta_{t-1})\|^2 \leq& (\mu\gamma_t/2) \|\theta_{t-1} - \theta^*\|^4 + 32\mu^{-1}\gamma_t^3 \|\nabla_\theta l_t(\theta_{t-1})\|^4,
\end{aligned}$$

which helps us to obtain the simplified expression,

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

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 θ_{t-1} is $\mathcal{F}_{t-1}$-measurable (Assumption 1-p), to have that

$$\mathbb{E}[\|\nabla_\theta l_t(\theta_{t-1})\|^4] \leq 8\kappa_t^4 \Delta_{t-1} + 8\sigma_t^4, \tag{23}$$

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

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

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

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

such that the last term of (24) can be bounded as follows,

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

Indeed, inserting this into (24) gives us

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

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

$$\Delta_t \leq \left[1-\left(\frac{\mu_\nu}{2}-\frac{\mathbb{1}_{\{\nu_t\neg C\}}D_\nu^4\nu_t^4}{\mu^3}\right)\gamma_t+\frac{256\kappa_t^4\gamma_t^3}{\mu}+24\kappa_t^4\gamma_t^4\right]\Delta_{t-1}+\frac{B_\nu^4\nu_t^4\gamma_t}{\mu^3}+\frac{256\sigma_t^4\gamma_t^3}{\mu}+24\sigma_t^4\gamma_t^4, \tag{25}$$

with $\mu'_\nu = \mu - \mathbb{1}_{\{\nu_t=C\}}2D_\nu^4\nu_t^4/\mu^3 > 0$. Note that μ_ν from Lemma 1 is lower bounded by μ'_ν, and strictly lower bounded for (ν_t) constant, i.e., $\mu_\nu > \mu'_\nu > 0$. Let $C_\Delta \geq 1$ fulfill the conditions of Proposition 1; the C_Δ constant is chosen such that $C_\Delta(\mathbb{1}_{\{\nu_t\neg C\}}D_\nu^4\nu_t^4/\mu^3+256\kappa_t^4\gamma_t^2/\mu+24\kappa_t^4\gamma_t^3) \leq \mu'_\nu/2$ implying $\mu'_\nu\gamma_t/2 \leq 1$, which is possible as the sequence (ν_t) is non-increasing, and (κ_t) and (γ_t) decrease. Hence, by applying Proposition 1 on (25), we obtain the desired bound for Δ_t. □

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

$$\Delta_t \leq \Pi_t + \frac{2^{2+4\rho\nu}B_\nu^4}{\mu^3\mu'_\nu C_\rho^{4\nu}t^{4\rho\nu}} + \frac{2^{2\rho(2\sigma-\beta)+2\alpha}(2^{10}\mu^{-1}+2^7C_\gamma C_\rho^\beta)C_\sigma^4C_\gamma^2C_\rho^{2\beta}}{\mu'_\nu C_\rho^{4\sigma}t^{2\rho(2\sigma-\beta)+2\alpha}}, \tag{26}$$

with Π_t given in (27) 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 (26) with $\mu'_\nu = \mu - \mathbb{1}_{\{\rho=0\}}2D_\nu^4/\mu^3C_\rho^{4\nu} > 0$, where Π_t can be bounded as follows,

$$\exp\left(-\frac{\mu'_\nu C_\gamma C_\rho^\beta}{4}\sum_{i=t/2}^{t} i^{\rho\beta-\alpha}\right)\left[\exp\left(\frac{\mathbb{1}_{\{\rho\neq0\}}C_\Delta D_\nu^4C_\gamma C_\rho^\beta}{\mu^3C_\rho^{4\nu}}\sum_{i=1}^{t} i^{\rho(\beta-4\nu)-\alpha}\right)\exp\left(\frac{2^8C_\Delta C_\kappa^4C_\gamma^3C_\rho^{3\beta}}{\mu C_\rho^{4\kappa}}\sum_{i=1}^{t} i^{\rho(3\beta-4\kappa)-3\alpha}\right)\right.$$
$$\exp\left(\frac{24C_\Delta C_\kappa^4C_\gamma^4C_\rho^{4\beta}}{C_\rho^{4\kappa}}\sum_{i=1}^{t} i^{4\rho(\beta-\kappa)-4\alpha}\right)\left(\Delta_0+\frac{4B_\nu^4}{\mu^3\mu'_\nu C_\rho^{4\nu}}+\frac{1024C_\sigma^4C_\gamma^2C_\rho^{2\beta}}{\mu\mu'_\nu C_\rho^{4\sigma}}+\frac{96C_\sigma^4C_\gamma^3C_\rho^{3\beta}}{\mu'_\nu C_\rho^{4\sigma}}\right)$$
$$\left.+\frac{B_\nu^4C_\gamma C_\rho^\beta}{\mu^3C_\rho^{4\nu}}\sum_{i=1}^{t/2-1} i^{\rho(\beta-4\nu)-\alpha}+\frac{256C_\sigma^4C_\gamma^3C_\rho^{3\beta}}{\mu C_\rho^{4\sigma}}\sum_{i=1}^{t/2-1} i^{\rho(3\beta-4\sigma)-3\alpha}+\frac{24C_\sigma^4C_\gamma^4C_\rho^{4\beta}}{C_\rho^{4\sigma}}\sum_{i=1}^{t/2-1} i^{4\rho(\beta-\sigma)-4\alpha}\right]$$
$$\leq \exp\left(-\frac{\mu'_\nu C_\gamma C_\rho^\beta t^{1+\rho\beta-\alpha}}{2^3}\right)\left[\exp\left(\frac{\mathbb{1}_{\{\rho\neq0\}}C_\Delta D_\nu^4C_\gamma C_\rho^\beta\psi^0_{\alpha-\rho(\beta-4\nu)}(t)}{\mu^3C_\rho^{4\nu}}\right)\exp\left(\frac{2^{10}C_\Delta C_\kappa^4C_\gamma^3C_\rho^{3\beta}}{\mu C_\rho^{4\kappa}}\right)\exp\left(\frac{2^6C_\Delta C_\kappa^4C_\gamma^4C_\rho^{4\beta}}{C_\rho^{4\kappa}}\right)\right.$$
$$\left.\left(\Delta_0+\frac{2^2B_\nu^4}{\mu^3\mu'_\nu C_\rho^{4\nu}}+\frac{2^{10}C_\sigma^4C_\gamma^2C_\rho^{2\beta}}{\mu\mu'_\nu C_\rho^{4\sigma}}+\frac{2^7C_\sigma^4C_\gamma^3C_\rho^{3\beta}}{\mu'_\nu C_\rho^{4\sigma}}\right)+\frac{B_\nu^4C_\gamma C_\rho^\beta\psi^0_{\alpha-\rho(\beta-4\nu)}(t/2)}{\mu^3C_\rho^{4\nu}}+\frac{2^{10}C_\sigma^4C_\gamma^3C_\rho^{3\beta}}{\mu C_\rho^{4\sigma}}+\frac{2^6C_\sigma^4C_\gamma^4C_\rho^{4\beta}}{C_\rho^{4\sigma}}\right], \tag{27}$$

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

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

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

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, (θ_t) follows (1), and the second part considers (2). Assume that (θ_t) is derived from the recursion in (1): following Polyak and Juditsky (1992), we rewrite (1) to

$$\frac{1}{\gamma_t}(\theta_{t-1} - \theta_t) = \nabla_\theta l_t(\theta_{t-1}), \tag{28}$$

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

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

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

$$\left(\mathbb{E}\left[\left\|\bar{\theta}_t - \theta^*\right\|^2\right]\right)^{1/2} \leq \left(\mathbb{E}\left[\left\|\nabla_\theta^2 L(\theta^*)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i\nabla_\theta l_i(\theta^*)\right\|^2\right]\right)^{1/2} + \left(\mathbb{E}\left[\left\|\nabla_\theta^2 L(\theta^*)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i\nabla_\theta l_i(\theta_{i-1})\right\|^2\right]\right)^{1/2}$$
$$+ \left(\mathbb{E}\left[\left\|\nabla_\theta^2 L(\theta^*)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i[\nabla_\theta l_i(\theta_{i-1}) - \nabla_\theta l_i(\theta^*) - \nabla_\theta L(\theta_{i-1})]\right\|^2\right]\right)^{1/2}$$
$$+ \left(\mathbb{E}\left[\left\|\nabla_\theta^2 L(\theta^*)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i\left[\nabla_\theta L(\theta_{i-1}) - \nabla_\theta^2 L(\theta^*)(\theta_{i-1} - \theta^*)\right]\right\|^2\right]\right)^{1/2}. \tag{29}$$

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

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

where the first term can be bounded by Assumption 4,

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

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

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

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

$$\begin{aligned}\left(\mathbb{E}\left[\left\|\nabla_\theta^2 L(\theta^*)^{-1}\frac{1}{N_t}\sum_{i=1}^t n_i\nabla_\theta l_i(\theta^*)\right\|^2\right]\right)^{1/2} \le& \frac{\Lambda^{1/2}}{N_t}\left(\sum_{i=1}^t n_i^{2(1-\sigma)}\right)^{1/2} + \frac{C_\sigma'^{1/2}}{\mu N_t^{1/2}}\left(\sum_{i=1}^t n_i^{2(1-\sigma-\sigma')}\right)^{1/2} \\ &+ \frac{2^{1/2}B_\nu^{1/2}}{\mu N_t}\left(\sum_{j=2}^t\left(n_j\nu_j\sum_{i=1}^{j-1} n_i\sigma_i\right)\right)^{1/2}.\end{aligned} \tag{30}$$

Next, by the relation in (28), we have

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

leading to

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

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

$$\left(\mathbb{E}\left[\left\|\nabla_\theta^2 L(\theta^*)^{-1}\frac{1}{N_t}\sum_{i=1}^t n_i\nabla_\theta l_i(\theta_{i-1})\right\|^2\right]\right)^{1/2} \le \frac{1}{\mu N_t}\sum_{i=1}^{t-1}\delta_i^{\frac12}\left|\frac{n_{i+1}}{\gamma_{i+1}}-\frac{n_i}{\gamma_i}\right| + \frac{n_t}{\mu\gamma_t N_t}\delta_t^{\frac12} + \frac{n_1}{\mu\gamma_1 N_t}\delta_0^{\frac12}. \tag{31}$$

For the third term, we can derive it as

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

where

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

by the Cauchy-Schwarz inequality, Assumption 2-p and (8). For the other term, we note that

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

using $\mathcal{F}_{i-1} \subset \mathcal{F}_{j-1}$ since $i < j$, Cauchy–Schwarz inequality, Hölder's inequality, $\|a+b\|^p \leq 2^{p-1}(\|a\|^p + \|b\|^p)$ with $p \in \mathbb{N}$, Assumptions 1-p and 2-p, and (8). Thus,

$$
\begin{aligned}
\left(\mathbb{E}\left[\left\|\nabla_\theta^2 L(\theta^*)^{-1}\frac{1}{N_t}\sum_{i=1}^t n_i\left[\nabla_\theta l_i(\theta_{i-1}) - \nabla_\theta l_i(\theta^*) - \nabla_\theta L(\theta_{i-1})\right]\right\|^2\right]\right)^{1/2} \leq& \frac{2^{1/2}}{\mu N_t}\left(\sum_{i=1}^t n_i^2\kappa_i^2\delta_{i-1}\right)^{1/2} + \frac{2^{1/2}C_\nabla}{\mu N_t}\left(\sum_{i=1}^t n_i^2\delta_{i-1}\right)^{1/2} \\
&+ \frac{2^{3/4}}{\mu N_t}\left(\sum_{j=2}^t\left((D_\nu\delta_{j-1}^{1/2} + 2^{1/2}B_\nu)n_j\nu_j\sum_{i=1}^{j-1}(C_\nabla + \kappa_i)n_i\delta_{i-1}^{1/2}\right)\right)^{1/2}.
\end{aligned} \tag{32}
$$

The last term is directly bounded by (9), using that (9) implies $\forall\theta$, $\|\nabla_\theta L(\theta) - \nabla_\theta^2 L(\theta^*)(\theta - \theta^*)\| \leq C'_\nabla\|\theta - \theta^*\|^2/2$ (Nesterov and Polyak, 2006), giving us

$$
\left(\mathbb{E}\left[\left\|\nabla_\theta^2 L(\theta^*)^{-1}\frac{1}{N_t}\sum_{i=1}^t n_i\left[\nabla_\theta L(\theta_{i-1}) - \nabla_\theta^2 L(\theta^*)(\theta_{i-1} - \theta^*)\right]\right\|^2\right]\right)^{\frac{1}{2}} \leq \frac{C'_\nabla}{2\mu N_t}\sum_{i=1}^t n_i\Delta_{i-1}^{1/2}, \tag{33}
$$

with the notion $\Delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^4]$. Combining the terms (30) to (33) into (29), gives us

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

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

$$
\begin{aligned}
\bar{\delta}_t^{1/2} \leq& \frac{\Lambda^{1/2}}{N_t}\left(\sum_{i=1}^t n_i^{2(1-\sigma)}\right)^{1/2} + \frac{C_\sigma'^{1/2}}{\mu N_t}\left(\sum_{i=1}^t n_i^{2(1-\sigma-\sigma')}\right)^{1/2} + \frac{2^{1/2}B_\nu^{1/2}}{\mu N_t}\left(\sum_{j=2}^t\left(n_j\nu_j\sum_{i=1}^{j-1}n_i\sigma_i\right)\right)^{1/2} + \frac{1}{\mu N_t}\sum_{i=1}^{t-1}\delta_i^{1/2}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| \\
&+ \frac{n_t}{\mu\gamma_t N_t}\delta_t^{1/2} + \frac{n_1}{\mu N_t}\left(\frac{1}{\gamma_1} + 2^{1/2}(C_\nabla + \kappa_1)\right)\delta_0^{1/2} + \frac{2^{1/2}}{\mu N_t}\left(\sum_{i=1}^{t-1}n_{i+1}^2(C_\nabla^2 + \kappa_{i+1}^2)\delta_i\right)^{1/2} \\
&+ \frac{2^{3/4}}{\mu N_t}\left(\sum_{j=1}^{t-1}\left((D_\nu\delta_j^{1/2} + 2^{1/2}B_\nu)n_{j+1}\nu_{j+1}\sum_{i=0}^{j-1}(C_\nabla + \kappa_{i+1})n_{i+1}\delta_i^{1/2}\right)\right)^{1/2} + \frac{C'_\nabla}{2\mu N_t}\sum_{i=0}^{t-1}n_{i+1}\Delta_i^{1/2}.
\end{aligned} \tag{34}
$$

Now, assume that (θ_t) is derived from the recursion in (2): as above, we follow the steps of Polyak and Juditsky (1992), in which, we can rewrite (2) to

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

where $\Omega_t=\mathcal{P}_\Theta(\theta_{t-1}-\gamma_t\nabla_\theta l_t(\theta_{t-1}))-(\theta_{t-1}-\gamma_t\nabla_\theta l_t(\theta_{t-1}))$. Thus, summing the parts, taking the norm and expectation, and using the Minkowski's inequality, yields the same terms as in (29), but with an additional term regarding Ω_t, namely

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

using Godichon-Baggioni (2016, Lemma 4.3). Next, we note that $\mathbb{E}[\|\Omega_t\|^2\mathbb{1}_{\{\theta_{t-1}-\gamma_t\nabla_\theta l_t(\theta_{t-1})\notin\Theta\}}]=4\gamma_t^2G_\Theta^2\mathbb{P}[\theta_{t-1}-\gamma_t\nabla_\theta l_t(\theta_{t-1})\notin\Theta]$, since

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

as $\mathcal{P}_\Theta$ is Lipschitz and $\|\nabla_\theta l_t(\theta)\|^2\leq G_\Theta^2$ for any $\theta\in\Theta$. Moreover, as in Godichon-Baggioni and Portier (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 Θ. Thus, (35) can then be bounded by

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

since the sequence (n_t) is either constant or increasing, meaning $\forall t, n_t/n_{t+1}\leq 1$. At last, this term can be combined into (34) with use of $C_\nabla''=C_\nabla'/2+\mathbb{1}_{\{D_\Theta<\infty\}}2G_\Theta/D_\Theta^2$, which indicates whether (θ_t) follows (2) 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

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

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

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

using the ψ-function defined in (16), such that $\sqrt{\psi^{\rho}_{\rho(\sigma-1)}(2N_t/C_\rho)\psi^{\rho}_{\rho(\nu-1)}(2N_t/C_\rho)}/N_t \leq \tilde{O}(N_t^{-\rho(\sigma+\nu)/2(1+\rho)})$. Let D^{κ}_{∇} denote $C_\nabla + 2^{\rho\kappa}C_\kappa/C^{\kappa}_{\rho}$ with $\kappa \in [0, 1/2]$, such that

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

and, likewise, we have that

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

From (21) we know that $\delta_t \leq D_\delta/t^{\delta}$ with

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

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

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

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

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

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

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

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

$$
\begin{aligned}
& \frac{2^{1/2} B_\nu^{1/2} C_\sigma^{1/2} C_\rho \sqrt{\psi^\rho_{\rho(\sigma-1)}(2N_t/C_\rho)\psi^\rho_{\rho(\nu-1)}(2N_t/C_\rho)}}{\mu C_\rho^{(\sigma+\nu)/2} N_t} + \frac{2^{3/2+\rho\nu} B_\nu C_\rho \psi^\rho_{1+\rho(\beta+\nu-1)-\alpha}(2N_t/C_\rho)}{\mu^{3/2}\mu_\nu^{1/2} C_\gamma C_\rho^{\beta+\nu} N_t} \\
& + \frac{2^{3/2+\rho(1+\nu)} B_\nu D_\nabla^\kappa C_\rho \sqrt{\psi^\rho_{2\rho(\nu-1)}(2N_t/C_\rho)}}{\mu^{3/2}\mu_\nu^{1/2} C_\rho^\nu N_t} + \frac{2^{3(1+\rho\nu)} B_\nu C_\rho^{(1-\beta-\nu-\alpha)/(1+\rho)}}{\mu^{3/2}\mu_\nu^{1/2} C_\gamma N_t^{(1+\rho(\beta+\nu)-\alpha)/(1+\rho)}} \\
& + \frac{2^{1+\rho(2-\nu)/2} B_\nu^{1/2} \sqrt{D_\nabla^\kappa} D_\delta^{1/4} C_\rho \sqrt{\psi^\rho_{\delta/2-\rho}(2N_t/C_\rho)\psi^\rho_{\rho(\nu-1)}(2N_t/C_\rho)}}{\mu C_\rho^{\nu/2} N_t} + \frac{2^{2(1+\rho\nu)} B_\nu^2 C_\nabla'' C_\rho \psi^\rho_{\rho(2\nu-1)}(2N_t/C_\rho)}{\mu^{5/2}\sqrt{\mu_\nu'} C_\rho^{2\nu} N_t} \\
& = \tilde{O}(N_t^{-\rho(\sigma+\nu)/2(1+\rho)}) + \tilde{O}(N_t^{-(1+\rho(\beta+\nu)-\alpha)/(1+\rho)}) + \tilde{O}(N_t^{-(1+2\rho\nu)/2(1+\rho)}) \\
& + O(N_t^{-(1+\rho(\beta+\nu)-\alpha)/(1+\rho)}) + \tilde{O}(N_t^{-(\delta/2+\rho\nu)/2(1+\rho)}) + \tilde{O}(N_t^{-2\rho\nu/(1+\rho)}),
\end{aligned}
\tag{36}
$$

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

$$
\begin{aligned}
\bar{\delta}_t^{1/2} \leq & \frac{\Lambda^{1/2}}{N_t^{1/2}} \mathbb{1}_{\{\sigma=1/2\}} + \frac{2^{1/2}\Lambda^{1/2} C_\rho^{(1-2\sigma)/2(1+\rho)}}{N_t^{(1+2\rho\sigma)/2(1+\rho)}} \mathbb{1}_{\{\sigma\neq 1/2\}} + \frac{2^{1/2} C_\sigma'^{1/2} C_\rho^{(1-2(\sigma+\sigma'))/2(1+\rho)}}{\mu N_t^{(1+2\rho(\sigma+\sigma'))/2(1+\rho)}} + \frac{2^6 C_\sigma C_\rho^{(2-2\sigma-\beta-\alpha)/2(1+\rho)}}{\mu \mu_\nu^{1/2} C_\gamma^{1/2} N_t^{(2+\rho(\beta+2\sigma)-\alpha)/2(1+\rho)}} \\
& + \mathbb{1}_{\{B_\nu\neq 0\}} \Psi_t + \frac{2^7 (\mu^{-1/2} + C_\gamma^{1/2} C_\rho^{\beta/2}) C_\nabla'' C_\sigma^2 C_\gamma}{\mu \sqrt{\mu_\nu'} C_\rho^{(1-2\rho\sigma-\alpha)/(1+\rho)} N_t^{(\rho(2\sigma-\beta)+\alpha)/(1+\rho)}} + \frac{2^2 D_\nabla^\kappa C_\sigma C_\gamma^{1/2} C_\rho^{(1+\beta-2\sigma+\alpha)/2(1+\rho)}}{\mu \mu_\nu^{1/2} N_t^{(1+\rho(2\sigma-\beta)+\alpha)/(2(1+\rho))}} + \frac{\Gamma C_\rho}{\mu N_t} \\
& + \frac{2^2 C_\rho^{(2+\beta-\alpha)/(1+\rho)} \bar{\pi}_\infty}{\mu C_\gamma N_t^{(2+\rho\beta-\alpha)/(1+\rho)}} + \tilde{O}(N_t^{-(\delta+\rho\nu)/2(1+\rho)}),
\end{aligned}
\tag{37}
$$

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

References

Agarwal, A., Duchi, J.C., 2012. The generalization ability of online algorithms for dependent data. IEEE Transactions on Information Theory 59, 573–587.

Ajalloeian, A., Stich, S.U., 2020. On the convergence of sgd with biased gradients. arXiv preprint arXiv:2008.00051 .

Anava, O., Hazan, E., Mannor, S., Shamir, O., 2013. Online learning for time series prediction, in: Conference on learning theory, PMLR. pp. 172–184.

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

Benveniste, A., Métivier, M., Priouret, P., 2012. Adaptive algorithms and stochastic approximations. volume 22. Springer Science & Business Media.

Bertsekas, D., 2016. Nonlinear Programming. volume 3. Athena Scientific.

Bottou, L., Cun, Y.L., 2003. Large scale online learning. Advances in neural information processing systems 16.

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

Bousquet, O., Elisseeff, A., 2002. Stability and generalization. The Journal of Machine Learning Research 2, 499–526.

Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. Time series analysis: forecasting and control. John Wiley & Sons.

Boyer, C., Godichon-Baggioni, A., 2020. On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. arXiv preprint arXiv:2011.09706 .

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

Brockwell, P.J., Davis, R.A., 2009. Time series: theory and methods. Springer Science & Business Media.

Cesa-Bianchi, N., Conconi, A., Gentile, C., 2004. On the generalization ability of on-line learning algorithms. IEEE Transactions on Information Theory 50, 2050–2057.

d'Aspremont, A., 2008. Smooth optimization with approximate gradient. SIAM Journal on Optimization 19, 1171–1183.

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

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

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

Gadat, S., Panloup, F., 2017. Optimal non-asymptotic bound of the ruppert-polyak averaging without strong convexity. arXiv preprint arXiv:1709.03342 .

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

Godichon-Baggioni, A., Portier, B., 2017. An averaged projected robbins-monro algorithm for estimating the parameters of a truncated spherical distribution. Electronic Journal of Statistics 11, 1890–1927.

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

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

Hamilton, J.D., 2020. Time series analysis. Princeton university press.

Hardt, M., Recht, B., Singer, Y., 2016. Train faster, generalize better: Stability of stochastic gradient descent, in: International conference on machine learning, PMLR. pp. 1225–1234.

Karimi, H., Nutini, J., Schmidt, M., 2016. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer. pp. 795–811.

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

Mohri, M., Rostamizadeh, A., 2010. Stability bounds for stationary φ-mixing and β-mixing processes. Journal of Machine Learning Research 11.

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

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

Murata, N., Amari, S., 1999. Statistical analysis of learning dynamics. Signal Processing 74, 3–28.

Necoara, I., Nesterov, Y., Glineur, F., 2019. Linear convergence of first order methods for non-strongly convex optimization. Mathematical Programming 175, 69–107.

Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A., 2009. Robust stochastic approximation approach to stochastic programming. SIAM Journal on optimization 19, 1574–1609.

Nesterov, Y., Polyak, B.T., 2006. Cubic regularization of newton method and its global performance. Mathematical Programming 108, 177–205.

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

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

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

Polyak, B.T., Juditsky, A.B., 1992. Acceleration of stochastic approximation by averaging. SIAM journal on control and optimization 30, 838–855.

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

Robbins, H., Monro, S., 1951. A stochastic approximation method. The annals of mathematical statistics , 400–407.

Rosenblatt, M., 1956. A central limit theorem and a strong mixing condition. Proceedings of the National Academy of Sciences of the United States of America 42, 43.

Ruppert, D., 1988. Efficient estimations from a slowly convergent Robbins-Monro process. Technical Report. Cornell University Operations Research and Industrial Engineering.

Schmidt, M., Roux, N., Bach, F., 2011. Convergence rates of inexact proximal-gradient methods for convex optimization. Advances in neural information processing systems 24.

Shalev-Shwartz, S., Singer, Y., Srebro, N., Cotter, A., 2011. Pegasos: Primal estimated sub-gradient solver for svm. Mathematical programming 127, 3–30.

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

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

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

Yu, B., 1994. Rates of convergence for empirical processes of stationary mixing sequences. The Annals of Probability , 94–116.

Zhang, T., 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms, in: Proceedings of the twenty-first international conference on Machine learning, p. 116.

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