Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Streaming Data
Antoine Godichon-Baggioni, Nicklas Werge, Olivier Wintenberger

To cite this version:
Antoine Godichon-Baggioni, Nicklas Werge, Olivier Wintenberger. Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Streaming Data. 2022. hal-03343481v3

HAL Id: hal-03343481
https://hal.science/hal-03343481v3
Preprint submitted on 11 Aug 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Streaming Data

Antoine Godichon-Baggioni, Nicklas Werge*, Olivier Wintenberger

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

Abstract

We consider the stochastic approximation problem in a streaming framework where an objective is minimized through unbiased estimates of its gradients. In this streaming framework, we consider time-varying data streams that must be processed sequentially. Our methods are Stochastic Gradient (SG) based due to their applicability and computational advantages. We provide a non-asymptotic analysis of the convergence of various SG-based methods; this includes the famous SG descent (a.k.a. Robbins-Monro algorithm), constant and time-varying mini-batch SG methods, and their averaged estimates (a.k.a. Polyak-Ruppert averaging). Our analysis suggests choosing the learning rate according to the expected data streams, which can speed up the convergence. In addition, we show how the averaged estimate can achieve optimal convergence in terms of attaining Cramer-Rao’s lower bound while being robust to any data stream rate. In particular, our analysis shows how Polyak-Ruppert averaging of time-varying mini-batches can provide variance reduction and accelerate convergence simultaneously, which is advantageous for large-scale learning problems. These theoretical results are illustrated for various data streams, showing the effectiveness of the proposed algorithms.

Keywords: stochastic optimization, machine learning, stochastic algorithms, online learning, streaming, mini-batch

1. Introduction

Machine learning and artificial intelligence have become an integral part of modern society. This massive utilization of intelligent systems generates an endless sequence of data, many of which come as streaming data such as weather, traffic, stock trade, or other real-time sensor data. These continuously generated data should be processed sequentially with the property that the data stream may change over time. Such a streaming framework requires computationally efficient and robust algorithms that can quickly update the model as more data arrives.

Stochastic approximation algorithms have proven effective in handling large amounts of data; Bottou et al. [3] reviews such stochastic algorithms for large-scale machine learning, including noise reduction and second-order methods, among others. Among these, the most well-known is presumably the Stochastic Gradient (SG) descent proposed by [27], which is used for many models within machine learning. Since its introduction, much work has been spent on analyzing, developing, and improving various SG-based methods, e.g., see Kushner and Yin [16], Lan [17], Shalev-Shwartz et al. [30]. An essential extension is the Polyak-Ruppert procedure (ASG) proposed by Polyak and Juditsky [26], Ruppert [28], which guarantees optimal statistical efficiency without jeopardizing the computational complexity; this average aggregates the estimates sequentially, which reduces the estimate variance while accelerating convergence [26].

Contributions. A fundamental aspect of this paper is to explore how changing data streams affect these stochastic algorithms. Our analysis extends the work of Moulines and Bach [21] to a streaming framework. We investigate several kinds of data streams, from vanilla SG descent and ASG to more exotic learning designs such as time-varying mini-batch SG and ASG. Our main theoretical contribution is the non-asymptotic analysis of SG-based methods in
this streaming framework. Our results show a noticeable improvement in convergence rates by having learning rates that adapt to the expected data streams. In particular, we show how to obtain optimal convergence rates robust to any data streaming rate.

**Organization.** Section 2 presents the streaming framework on which the non-asymptotic analysis relies. Our convergence results are presented in Sections 3 and 4, with and without averaging. Both sections includes analysis of unbounded and uniform bounded gradients. These theoretical results are illustrated in Section 5 for a variety of data streams. At last, some final remarks are done in Section 6.

2. Problem Formulation

The analysis of statistical and machine learning models often involves some form of optimization [3, 11, 18]. Many of these optimization problems aim to minimize functions of the form,

\[ L(\theta) = E[l_i(\theta)], \]  

with respect to \( \theta \in \Theta \), where \( \Theta \subseteq \mathbb{R}^d \) is a convex body; a convex body in \( \mathbb{R}^d \) is a compact convex set with non-empty interior. For streaming (or large-scale) problems, it would be too expensive to compute the full gradient \( \nabla \theta L(\theta) \). Instead, the minimization of \( L \) in (1) is achieved without evaluating it directly but by unbiased estimates of its gradients, namely, through \( \nabla \theta l_i : \Theta \to \mathbb{R} \). Observe that the principles for biased functions \( l_i \) are rather different [6, 29]. Thus, we let \( l_i \) constitute a sequence of independent differentiable random (possibly non-convex) functions and their gradients unbiased estimates of \( \nabla \theta L \), e.g., see Nesterov et al. [24] for definitions and properties of such functions.

Many problems, from classification, and regression to ranking, can be written on this form (1), e.g., see Teo et al. [33] for examples of scalar and vectorial loss functions and their derivatives. For example, consider the simple case of examples of scalar and vectorial loss functions and their derivatives. For example, consider the simple case where we have some samples \((X_t, Y_t)\), \( t = 1, \ldots, n \). Our interest is to find predictor \( h_0 \) over some parameterization \( \{h_0\}_{h_0 \in \Theta} \), by minimizing (1) with \( l_i(\theta) = l_i(h_0(X_t), Y_t) + A\Omega(\theta) \), where \( l \) is some loss function, \( A > 0 \) a regularizer parameter, and \( \Omega : \Theta \to \mathbb{R} \) some regularizer, e.g., the \( l_1 \) or \( l_2 \) regularization; here the loss \( l \) could be the quadratic, logistic, (squared) hinge, or Huber’s (robust) loss, but it depends on the experiments that one wants to perform [3, 17, 24].

**Streaming framework.** Let us now describe our streaming framework in which we will solve our problem in (1): at each time \( t \in \mathbb{N} \), a block consisting of \( n_t \in \mathbb{N} \) random functions \( l_t = (l_{t,1}, \ldots, l_{t,n_t}) \) arrive. To solve this, we introduce the Stochastic Streaming Gradient (SSG), defined as

\[ \theta_t = \theta_{t-1} - \frac{\gamma_t}{n_t} \sum_{i=1}^{n_t} \nabla l_{t,i}(\theta_{t-1}), \quad \theta_0 \in \Theta, \]  

(2)

where \((\gamma_t)\) is a decreasing sequence of positive numbers also referred to as the learning rate, satisfying \( \sum_{i=1}^{\infty} \gamma_i = \infty \) and \( \sum_{i=1}^{\infty} \gamma_i^2 < \infty \) for \( t \to \infty \) [27]. In the same way, we introduce the Projected SSG (PSSG), defined by

\[ \hat{\theta}_t = P_\Theta \left( \theta_{t-1} - \frac{\gamma_t}{n_t} \sum_{i=1}^{n_t} \nabla l_{t,i}(\theta_{t-1}) \right), \quad \theta_0 \in \Theta, \]  

(3)

where \( P_\Theta \) denotes the Euclidean projection onto the convex body \( \Theta \subseteq \mathbb{R}^d \), i.e., \( P_\Theta(\theta) = \arg \min_{\theta' \in \Theta} \|\theta - \theta'\|_2 \). The PSSG estimate in (3) is very convenient for models with conditions on the parameters space, and thereby, requires a projection of the parameters. Next, to guarantee optimal convergence properties [26, 28], we introduce the Polyak-Ruppert average of (2), called Averaged SSG (ASSG), given as

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

(4)

where \( N_t = \sum_{i=0}^{t-1} n_i \) denotes the accumulated sum of observations. Similarly, we define the (Polyak-Ruppert) Average PSSG (APSSG) estimate as when \( (\bar{\theta}_i) \) (in (4)) is derived using (3).
2.1. Quasi-strong Convex and Lipschitz Smooth Objectives

Following Moulines and Bach [21], Sridharan et al. [31], we make the following assumptions about the objective function $L$: assume that $\theta^* \in \Theta$ is the unique global minimizer of $L$ with $\nabla \theta L(\theta^*) = 0$. Also, let $L$ be $\mu$-quasi-strong convex [13, 23] with convexity constant $L$ function

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

(5)

Teo et al. [33] provides a comprehensive record of various convex functions $L$ used in machine learning applications. Milder degrees of convexity have been studied by, e.g., Karimi et al. [13], which studied stochastic gradient methods under the Polyak-Łojasiewicz condition [19, 25], or Gadat and Panloup [7], which studied the Ruppert-Polyak averaging estimate under some Kurdyka-Łojasiewicz-type condition [15, 19]. Next, let the function $\nabla \theta L$ be $C_\varphi$-Lipschitz continuous, i.e., there exists $C_\varphi > 0$ such that $\forall \theta, \theta' \in \Theta,$

$$||\nabla \theta L(\theta) - \nabla \theta L(\theta')|| \leq C_\varphi||\theta - \theta'||.$$  

(6)

Furthermore, for the averaging estimate in (4), we need the function $L$ to be twice differentiable with $C_\varphi^2$-Lipschitz continuous Hessian operator $\nabla^2 \theta L$, meaning, there exists $C_\varphi^2 \geq 0$ such that $\forall \theta, \theta' \in \Theta,$

$$||\nabla^2 \theta L(\theta) - \nabla^2 \theta L(\theta')|| \leq C_\varphi^2||\theta - \theta'||.$$  

(7)

Note that (6) and (7) only needs to hold for $\theta' = \theta^*$.

3. Stochastic Streaming Gradients

This section considers the SSG and PSSG methods with streaming batches arriving in constant and time-varying streams. Our aim is to provide bounds on the quadratic mean $E[||\theta_t - \theta^*||^2]$, which depends explicitly upon the problem’s parameters. In order to do this, we assume the following about the function $l_i$ for each $i \in \mathbb{N}$ with $i = 1, \ldots, n$.

**Assumption 1.** For each $\theta \in \Theta$, the random variable $\nabla \theta l_i(\theta)$ is square-integrable and $\forall \theta \in \Theta$, $E[|\nabla \theta l_i(\theta)|] = \nabla \theta L(\theta)$.

**Assumption 2-p (Cp-expected smoothness).** There exists a positive integer $p$ such that $\forall \theta, \theta' \in \Theta$, $E[|\nabla \theta l_i(\theta) - \nabla \theta l_i(\theta')|^p] \leq C_i^p E[|\theta - \theta'|^p]$ with $C_i \in \mathbb{R}_+.$

**Assumption 3-p (σ-gradient noise).** There exists a positive integer $p$ such that $E[|\nabla \theta l_i(\theta')|^p] \leq \sigma^p$ with $\sigma \in \mathbb{R}_+$.

These assumptions are modified versions of the standard assumptions for stochastic approximations [2, 16] as they hold for any $i = 1, \ldots, n$. Note that Assumption 2-p only needs to hold for $\theta' = \theta^*$. By the smoothness assumption (Assumption 2-p), we avoid the unfavorable uniformly bounded gradients assumption, which is too restrictive and only holds for a few losses. Assumption 3-p is a weak assumption that should be seen as an assumption on $\Theta$ rather than on $(l_i)$. For SSG and PSSG, we only need Assumptions 2-p and 3-p to hold for $p = 2$, whereas, for ASSG and APSSG, we need $p = 4$ in order to bound the fourth-order moment.

Our streaming framework include classic examples: stochastic approximation (Robbins-Monro setting [27]) and learning from i.i.d. data, such as linear regression, logistic regression, general ridge regressions and quantile regression, p-means, and softmax regression, under regularity conditions [5, 32, 33].

In the following theorem, we derive an explicit upper bound on the $t$-th estimate of (2) and (3) for any learning rate $(\gamma_t)$ using classical techniques from stochastic approximations [2, 16].

**Theorem 1 (SSG/PSSG).** Denote $\delta_t = E[|\theta_t - \theta^*|^2]$ for some $\delta_0 \geq 0$, where (6) follows (2) or (3). Under Assumption 1, Assumptions 2-p and 3-p with $p = 2$, we have for any learning rate $(\gamma_t)$ that

$$\delta_t \leq \exp\left(-\mu \sum_{i=1}^{t} \gamma_i \right)^{\pi_t^2} + \frac{2\sigma^2}{\mu} \max_{i \leq t} \gamma_i,$$

(8)

with $\pi_t^2 = \exp(4C_i^2 \sum_{i=1}^{t} \gamma_i^2/\nu_i) \exp(2C_i^2 \sum_{i=1}^{t} \frac{\nu_i}{\gamma_i}) (\delta_0 + 2\sigma^2/C_i^2)$. 

3
Sketch of proof. Under Assumption 1, Assumptions 2-p and 3-p with \( p = 2 \), we derive from (2) that (\( \delta_t \)) satisfies the recursive relation

\[
\delta_t \leq \left[ 1 - 2\mu \gamma_t + (2C_1^2 + (n_t - 1)C_2^2)\mu^{-1} \gamma_t^2 \right] \delta_{t-1} + 2\sigma^2 n_t^{-1} \gamma_t^2 ,
\]

(9)

for any (\( n_t \)) and (\( \gamma_t \)) fulfilling the conditions imposed on the learning rate [27]. This recursive relation is then bounded in a non-asymptotic manner using Proposition 5. Bounding the projected estimate in (3) follows directly from the fact that \( \mathbb{E}[\| \mathcal{P}_{\theta} \mathcal{L}(\theta) - \theta'^{\top} \|^2] \leq \mathbb{E}[\| \theta - \theta' \|^2] \), \( \forall \theta \in \Theta \) [35]. Alternatively, the projected estimate can also be shown without Assumptions 2-p and 3-p but instead with a bounded gradient assumption (Assumption 5), e.g., see Moulines and Bach [21].

Related work. When \( n_t = 1 \) in (8), we obtain the usual SG descent studied in Moulines and Bach [21]. Similarly, Theorem 1 provides an upper bound on the function values, \( \mathbb{E}[L(\theta) - L(\theta')] \leq C \delta_t / \pi ; \) this follows by Cauchy-Schwarz inequality and Assumption 2-p.

Natural decay imposed by Robbins and Monro [27]. The learning rate (\( \gamma_t \)) should satisfy the following requirements: \( \sum_{t=1}^{\infty} \gamma_t = \infty \) and \( \sum_{t=1}^{\infty} \gamma_t^2 / n_t \leq \sum_{t=1}^{\infty} \gamma_t^2 < \infty \) for \( t \to \infty \). These conditions directly imply that \( \pi_\infty^\alpha < \infty \) as \( t \to \infty \). Thus, our attention is on reducing the noise term \( \max_{0 < \sigma < \gamma_t} \gamma_t / n_t \), without damaging the natural decay of the sub-exponential term \( \exp(-\mu \gamma_t / n_t) \). In particular, this non-asymptotic bound shows convergence in quadratic mean for any learning rate, fulfilling these conditions. In addition, the scaling with (\( n_t \)) in the noise term shows an apparent variance reduction when we increase the streaming batches (\( n_t \)).

Throughout this paper, we will consider learning rates on the form \( \gamma_t = C_\gamma n_t^{-\beta} \alpha^{-\alpha} \) with hyper-parameters \( C_\gamma > 0, \beta \in [0, 1] \), and \( \alpha \) chosen accordingly to the expected streaming batches denoted by \( n_t \). We start by considering constant streaming batches (i.e., mini-batch SG), where \( n_t \) follows the constant streaming batch size \( C_\gamma \in \mathbb{N} \):

Corollary 1 (SSG/PSSG, constant streaming batches). Denote \( \delta_t = \mathbb{E}[\| \theta_t - \theta' \|^2] \), where (\( \theta_t \)) follows (2) or (3). Suppose \( \gamma_t = C_\gamma n_t^{-\beta} \alpha^{-\alpha} \) with \( n_t = C_\rho \) for \( C_\rho \in \mathbb{N} \), such that \( \alpha \in (1/2, 1) \). Under Assumption 1, Assumptions 2-p and 3-p with \( p = 2 \), we have

\[
\delta_t \leq \exp \left( - \frac{\mu C_\gamma n_t^{-1 - \alpha}}{2^{1 - \alpha} C_\rho^{1 - \beta}} \gamma_t^2 \right) \pi_\infty^\alpha + \frac{2^{1 - \alpha} \alpha^2 C_\gamma}{\mu C_\rho^{1 - \beta}} n_t^{-\beta} \gamma_t^2 ,
\]

(10)

where \( \pi_\infty^\alpha = \exp(4\alpha C_1^2 (2C_1^2 + C_\rho^{-1} C_2^2) / (2\alpha - 1) C_\rho^{-1 - 2\beta}) \delta_0 + 2\sigma^2 / C_1^2 \) is a finite constant.

Decay of the initial conditions. The bound in Corollary 1 depends on the initial condition \( \delta_0 = \mathbb{E}[\| \theta_0 - \theta' \|^2] \) and the variance \( \sigma^2 \) in the noise term. The initial condition \( \delta_0 \) vanish sub-exponentially fast for \( \alpha \in (1/2, 1) \). Thus, the asymptotic term is \( 2^{1 + \beta} \alpha^2 C_\gamma / \mu C_\rho^{1 - \beta} n_t^{-\beta} \), i.e., \( \delta_t = O(N_t^{-\beta}) \). Moreover, the bound in (10) is optimal (up to some constants) for quadratic functions (\( f_b \)), since the deterministic recursion in (9) would be with equality. It is worth noticing that if \( C_\gamma C_\rho > C_\gamma C_\rho \) is chosen too large, they may produce a large \( \pi_\infty^\alpha \) constant. In addition, \( \pi_\infty^\alpha \) is positively affected by \( C_\rho \) when \( \beta < 1/2 \). Obviously, the hyper-parameter \( \beta \) only comes into play if the streaming batch size is larger than one, i.e., \( C_\rho > 1 \). Nonetheless, the effect of \( \pi_\infty^\alpha \) will decrease exponentially fast due to the sub-exponentially decaying factor in front.

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

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

Corollary 2 (SSG/PSSG, time-varying streaming batches). Denote \( \delta_t = \mathbb{E}[\| \theta_t - \theta' \|^2] \), where (\( \theta_t \)) follows (2) or (3). Suppose \( \gamma_t = C_\gamma n_t^{-\beta} \alpha^{-\alpha} \) with \( n_t = C_\rho \rho^\alpha \) for \( C_\rho \in \mathbb{N} \) and \( \rho \in (-1, 1) \), such that \( \alpha = \beta \bar{\rho} \in (1/2, 1) \). Under Assumption 1,
4.1. Unbounded Gradients

Assumptions 2-p and 3-p with \( p = 2 \), we have

\[
\delta_t \leq \exp \left( -\frac{\mu C_\gamma N_{t}^{1-\phi}}{2(2+p)(1-\phi)C_{p}^{1-\phi}-\phi} \right) n_{\infty}^\phi + \frac{2(1+p)^{\phi} C_\gamma^2 C_{p}^{2(2+p)(1-\phi)}(2C_{\gamma}^2 + C_{p}^2)/((2(\alpha - \beta \rho) - 1))}{\mu C_{\rho}^{1-\phi}} n_{\infty}^{1-\phi},
\]

where \( \phi = ((1-\beta)\rho + \alpha)/(1+\rho) \) and \( n_{\infty} = \exp(4(\alpha - \beta \rho)C_{\gamma}^2 C_{p}^{2(2+p)(1-\phi)}((2(\alpha - \beta \rho) - 1))(\delta_0 + 2N_{\infty}^2/C_{\gamma}^2)) \) is a finite constant.

Decay of the initial conditions. As mentioned for Corollary 1, the condition of having \( \alpha - \beta \rho \in (1/2, 1) \) is a natural restriction from Robbins and Monro [27], which relaxes the usual condition of having \( \alpha \in (1/2, 1) \) for \( \rho \) non-negative. For \( \rho \in (-1, 1/2) \), setting \( \alpha = 2/3 \) and \( \beta = 1/3 \) would give same decay rate, \( \delta_t = O(N_{t}^{-2/3}) \) as we saw for Corollary 1 when \( \alpha = 2/3 \). However, accelerated convergence could be achieved by, e.g., setting \( \alpha = 1 \) and \( \beta = 1/2 \) for streaming rate \( \rho \in (0, 1) \), giving us \( \delta_t = O(N_{t}^{-(1+p)/(1+p)}) \).

Variance reduction. Similarly to Corollary 1, the sub-exponential and asymptotic term is scaled by \( C_{\rho}^{1-\rho} \) for \( \rho \geq 0 \), implying we should take \( \alpha + \beta \leq 1 \) to obtain variance reduction. These conclusions will change when we consider the averaging estimate in Section 4.

The reasoning in Corollary 2 could be expanded to include random streaming batches where \( n_i \) is given such that \( C_{t}^p \mu_{l}^p \leq n_i \leq C_{H}^p \mu_{l}^p \) with \( \mu_{l}, \mu_{H} \in (-1, 1) \) and \( C_{t}, C_{H} \geq 1 \). This yields the modified rate \( \phi' = ((1-\beta)\rho_{l} + \alpha)/(1+\rho_{H}) \); nevertheless, we will leave the proof to the reader.

4. Averaged Stochastic Gradient Descent

In what follows, we consider the averaging estimate (\( \bar{\delta}_t \)) given in (4) derived with use of (\( \bar{\theta}_t \)) from (2) (Section 4.1) or (3) (Section 4.2). Besides having Assumptions 2-p and 3-p to hold for \( p = 4 \), an additional assumption is needed for bounding the rest term of the averaging estimate.

Assumption 4. There exists a non-negative self-adjoint operator \( \Sigma \) such that \( \mathbb{E}[\nabla_{\theta l_{12}}(\theta^*)\nabla_{\theta l_{12}}(\theta^*)^T] \leq \Sigma \).

Note that the operator \( \Sigma \) always exists when \( \sigma \) is finite for order \( p = 4 \) in Assumption 3-p.

4.1. Unbounded Gradients

As in Section 3, we conduct a general study for any learning rate (\( \gamma_{l} \)) when applying the Polyak-Ruppert averaging estimate from (4):

Theorem 2 (ASSG). Denote \( \bar{\delta}_t = \mathbb{E}[||\bar{\theta}_t - \theta^*||^2] \) with \( \bar{\theta}_t \) given by (4), where \( \bar{\theta}_t \) follows (2). Under Assumption 1, Assumptions 2-p and 3-p with \( p = 4 \), and Assumption 4, we have for any learning rate (\( \gamma_{l} \)) that

\[
\frac{\delta_t^{1/2}}{\sqrt{N_t}} \leq \frac{\Lambda^{1/2}}{\sqrt{N_t}} + \frac{1}{\mu N_t} \sum_{i=1}^{t-1} \frac{n_{i+1}}{\gamma_{l_{i+1}}} \frac{n_{i+1}^{1/2}}{\gamma_{l_{i+1}}^{1/2}} + \frac{n_i}{\mu N_t} \left( \frac{1}{\gamma_{l}} + C_{t} \right) \frac{\delta_t^{1/2}}{\gamma_{l}} + \frac{C_{t}^{2}}{\mu N_t} \sum_{i=1}^{t-1} n_{i+1} \delta_{i}^{1/2} + C_{t}^{2} \frac{\sum_{i=0}^{t-1} n_{i+1} \Lambda_{i}^{1/2}}{\mu N_t},
\]

where \( \Lambda = \text{Tr}(\nabla_{\theta}^2 L(\theta^*)^{-1} \Sigma \nabla_{\theta}^2 L(\theta^*)^{-1}) \) and \( \Delta_{t} = \mathbb{E}[||\bar{\theta}_t - \theta^*||^4] \) for some \( \Delta_{0} \geq 0 \).

As noticed in Polyak and Juditsky [26], the leading term \( \Lambda/N_t \) achieves the Cramer-Rao lower bound [7, 22]. Note that the leading term \( \Lambda/N_t \) is invariant of the learning rate (\( \gamma_{l} \)). Moreover, this bound of \( O(N_{t}^{-1}) \) is achieved without inverting the Hessian. Next, the processes (\( \delta_t \)) and (\( \Delta_t \)) can be bounded by the recursive relations in (8) and (22). There are no sub-exponential decaying terms for the initial conditions in Theorem 2, which is a common problem for averaging. However, as mentioned previously, we are more interested in advancing the decay of the asymptotic terms.

To ease notation, we make use of the functions \( \psi_{x}(t) : \mathbb{R} \to \mathbb{R} \), given as

\[
\psi_{x}(t) = \begin{cases} 
(1-x)^{(1+y)/(1-x)} & \text{if } x < 1, \\
(1+y) \log(t) & \text{if } x = 1, \\
x/(x-1) & \text{if } x > 1,
\end{cases}
\]
with $y \in \mathbb{R}_+$, such that $\sum_{i=1}^t f^x \leq \psi_0(t)$ for any $x \in \mathbb{R}_+$. Note that $\psi_1(t)/t = O(t^{-1+\gamma}/(1+\gamma))$ if $x < 1$, $\psi_2(t)/t = O(\log(t)/t^\gamma)$ if $x = 1$, and $\psi_3(t)/t = O(t^{-1})$ if $x > 1$. Hence, for any $x, y \in \mathbb{R}_+$, $\psi_4(t)/t = O(t^{-1+\gamma}/(1+\gamma))$, where the $O(\cdot)$ notation hides logarithmic factors.

**Corollary 3 (ASSG, constant streaming batches).** Denote $\tilde{\delta}_t = E[||\tilde{\theta}_t - \theta||^2]$ with $(\tilde{\theta}_t)$ given by (4), where $(\tilde{\theta}_t)$ follows (2). Suppose $\gamma_1 = C_\rho^1\rho^{-\alpha}$ with $n_c = C_\rho$ for $C_\rho \in \mathbb{N}$, such that $\alpha \in (1/2, 1)$. Under Assumption 1, Assumptions 2-p and 3-p with $p = 4$, and Assumption 4, we have

$$
\delta_t^{3/2} \leq \frac{\Lambda t^{1/2}}{N_t^{1/2}} + \frac{6\sigma C_\rho^{2(\alpha-\beta)/2}}{\mu t^{1/2}C_\gamma^{1/2}N_t^{1/2}} + \frac{2^{\sigma C_\rho^2C_\gamma^2C_\rho^3C_\gamma^3/2}2^{\beta C_\rho C_\gamma^3}}{\mu^{3/2}C_\rho^{1+\gamma}(1+\sigma)^{1/2}N_t^{1/2}} + \frac{C_\rho^{\Gamma_\rho}}{\mu N_t}
$$

with $\Gamma_\rho$ given by $(1/C_\rho C_\rho^2 + C_\rho)\delta_t^{1/2} + C_\rho \sqrt{\pi_\rho^{\omega}}/C_\rho + \sqrt{\pi_\rho^{\omega}}/C_\rho + C_\rho \sqrt{\pi_\rho^{\omega}}/C_\rho$, consisting of the finite constants $\pi_\rho^{\omega}$, $\Pi_\rho^{\omega}$ and $A_\rho^{\omega}$, that only depends on $\mu, \delta_0, \Delta_0, C_\rho, \gamma, C_\gamma, C_\rho, \rho, \beta$ and $\alpha$.

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

**Corollary 4 (ASSG, time-varying streaming batches).** Denote $\delta_t = E[||\tilde{\theta}_t - \theta||^2]$ with $(\tilde{\theta}_t)$ given by (4), where $(\tilde{\theta}_t)$ follows (2). Suppose $\gamma_1 = C_\rho^1\rho^{-\alpha}$ with $n_c = C_\rho^p$ for $C_\rho \in \mathbb{N}$ and $\rho \in (-1, 1)$, such that $\alpha - \beta \rho \in (1/2, 1)$. Under Assumption 1, Assumptions 2-p and 3-p with $p = 4$, and Assumption 4, we have

$$
\delta_t^{1/2} \leq \frac{\Lambda t^{1/2}}{N_t^{1/2}} + \frac{2^{\sigma(\alpha+1)(\beta+1)p^2/2}}{\mu^{1/2}C_\rho^{1+\gamma}} + \frac{2^{(\alpha+\gamma+1)p^2/2}}{\mu^{1/2}C_\rho^{1+\gamma}} + \frac{2^{\sigma C_\rho C_\gamma^3}}{\mu^{3/2}C_\rho^{1+\gamma}(1+\sigma)^{1/2}N_t^{1/2}} + \frac{C_\rho^{\Gamma_\rho}}{\mu N_t}
$$

with $\Gamma_\rho$ given by $(1/C_\rho C_\rho^2 + C_\rho)\delta_t^{1/2} + C_\rho \sqrt{\pi_\rho^{\omega}}/C_\rho + \sqrt{\pi_\rho^{\omega}}/C_\rho + C_\rho \sqrt{\pi_\rho^{\omega}}/C_\rho$, consisting of the finite constants $\pi_\rho^{\omega}$, $\Pi_\rho^{\omega}$ and $A_\rho^{\omega}$, that only depends on $\mu, \delta_0, \Delta_0, C_p, \gamma, C_\gamma, C_\rho, \rho, \beta$ and $\alpha$.

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

4.2. Bounded Gradients

In what follows, we consider the averaging estimate $\tilde{\theta}_n$ given in (4) but with the use of the projected estimate PSSG from (3). To avoid calculating the six-order moment, we make the unnecessary assumption that $\|\nabla \delta_1(\theta_t)\|$ is uniformly bounded for any $\theta_t \in \Theta$; the derivation of the six-order moment can be found in Godichon-Baggioni [9].

**Assumption 5.** Let $D_\rho = \inf_{t \in \Theta} ||\theta - \theta_t|| > 0$ with $\partial \Theta$ denoting the frontier of $\Theta$. Moreover, there exists $G_\Theta > 0$ such that $\forall t \geq 1, \sup_{t \in \Theta} ||\nabla \delta_1(\theta_t)||^2 \leq G_\Theta^2 a.s.$, with $i = 1, \ldots, n$.
Corollary 5 (APSSG, constant streaming batches). Denote $\delta_i = \mathbb{E}[[\tilde{\theta}_t - \theta^*]^2]$ with $\tilde{\theta}_t$ given by (4), where $(\theta_t)$ follows (3). Suppose $\gamma_t = C_r n_t \sigma^a$ with $n_t = C_r n$ for $C_r \in \mathbb{N}$, such that $\alpha \in (1/2, 1)$. Under Assumption 1, Assumptions 2-p and 3-p with $p = 4$, Assumptions 4 and 5, we have

$$\delta_1^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{6\sigma C_r^{1-a-\beta/2}}{N_t^{1/2}} + \frac{2\sigma^6 C_r^2 C_y}{\mu^2 C_p^{1-a-\beta} N_t^{2-\beta}} + \frac{2\sigma C_r^2 C_y^2}{\mu^{3/2} C_p^{1-a-\beta} N_t^{4(1+\alpha)/2}} + \frac{C_p \Gamma_c}{\mu N_t},$$

with $C_r^\gamma = C_y + 2G_\theta / D_2$ and $\Gamma_c$ given by $(1/C_y C_r^\theta + C_y) \sigma_0^{1/2} + C_y \sqrt{\sigma_0^{1/2} C_y^2 + \sigma_0^{1/2}} C_y^2 + C_y^{1/2} \Gamma_0^{1/2}$, consisting of the finite constants $\pi_0$, $\Pi_0^{1/2}$, and $\lambda^{1/2}$, that only depends on $\mu$, $\delta_0$, $\Delta_0$, $C_r$, $\sigma$, $C_y$, $C_r$, $\gamma$, $C_p$, $\beta$, and $\alpha$.

Corollary 6 (APSSG, time-varying streaming batches). Denote $\delta_i = \mathbb{E}[[\tilde{\theta}_t - \theta^*]^2]$ with $\tilde{\theta}_t$ given by (4), where $(\theta_t)$ follows (3). Suppose $\gamma_t = C_r n_t \sigma^a$ with $n_t = C_r n$ for $C_r \in \mathbb{N}$ and $\rho \in (-1, 1)$, such that $\alpha \in (1/2, 1)$. Under Assumption 1, Assumptions 2-p and 3-p with $p = 4$, Assumptions 4 and 5, we have

$$\delta_1^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{2^{(1+\alpha)(1+\beta)/2} C_y^2 C_r^{1/2}}{\mu^2 C_p^{1-a-\beta} N_t^{2-\beta}} + \frac{2^{(1+\alpha)(1+\beta)/2} C_y^2 C_r^{1/2}}{\mu^{3/2} C_p^{1-a-\beta} N_t^{1+\beta/2}} + \frac{C_p \Gamma_c}{\mu N_t},$$

with $C_r^\gamma = C_y + 2G_\theta / D_2$ and $\Gamma_c$ given by $(1/C_y C_r^\theta + C_y) \sigma_0^{1/2} + C_y \sqrt{\sigma_0^{1/2} C_y^2 + \sigma_0^{1/2}} C_y^2 + C_y^{1/2} \Gamma_0^{1/2}$, consisting of the finite constants $\pi_0$, $\Pi_0^{1/2}$, and $\lambda^{1/2}$, that only depends on $\mu$, $\delta_0$, $\Delta_0$, $C_r$, $\sigma$, $C_y$, $C_r$, $\gamma$, $C_p$, $\beta$, and $\alpha$.

5. Experiments

In this section, we demonstrate the theoretical results presented in Sections 3 and 4 for various data streams. In Section 5.1, we illustrate the unbounded gradient case (Sections 3 and 4) using linear regression. Where in Section 5.2, we present the bounded gradient case (Sections 3 and 4.2) by considering the geometric median. To measure the performance, we use the quadratic mean error of the parameter estimates over one-hundred replications, given by $\mathbb{E}[[\tilde{\theta}_t - \theta^*]^2]$. Note that averaging over several iterations gives a reduction in variability, which mainly benefits the SGG and PSSG.

5.1. Linear Regression

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

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

Next, in Figures 1b to 1e, we vary the streaming rate $\rho$ for streaming batch sizes $C_p = 1, 8, 64, 128$, respectively, with $\beta = 0$. These figures show an increase in decay of the SGG when the streaming rate $\rho$ increase. Despite this, we still achieve better convergence for the ASSG method, which seems more immune to the different choices of...
Figure 1: Linear regression for various data streams $n_t = C_t \rho^t$. See Section 5.1 for details.

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

(b) Time-varying streaming batches, $C_t = 1, \beta = 0$

(c) Time-varying streaming batches, $C_t = 8, \beta = 0$

(d) Time-varying streaming batches, $C_t = 64, \beta = 0$

(e) Time-varying streaming batches, $C_t = 128, \beta = 0$

(f) Time-varying streaming batches, $C_t = 8, \beta = 1/3$
streaming rate $\rho$, e.g., see the discussion after Corollary 4. We know this from Corollary 2, as $\phi = (\hat{\rho} + \alpha)/(1 + \hat{\rho}) \geq \alpha$ for $\beta = 0$. In addition, we see that $C_p$ has a positive effect on the noise (i.e., variance reduction), but if $C_p$ becomes too large, it may slow down convergence (as seen in Figure 1e). Alternatively, we could think around the problem in another way: how can we choose $\alpha$ and $\beta$ such that we have optimal decay of $\phi = 2/3$ for any $\rho$. In other words, for any arrival schedule that may occur, how should we choose our hyper-parameters such that we achieve optimal decay of $\phi = 2/3$. As discussed after Corollary 4, one example of this could be achieved by setting $\alpha = 2/3$ and $\beta = 1/3$ such that $\phi = 2/3$ for any $\rho$. Figure 1f shows an example of this where we (indeed) achieve the same decay rate for any streaming rate $\rho$.

5.2. Geometric Median

The geometric median is a generalization of the real median introduced by Haldane [12]. Robust estimators such as the geometric median may be preferred over the mean when the data is noisy. Moreover, in our streaming framework, stochastic algorithms are preferred as they efficiently handle large samples of high-dimensional data [5, 9].

The geometric median of $X \in \mathbb{R}^d$ is defined by $\theta^\ast \in \mathbb{R}^d$ which minimizes the convex function $L(\theta) = E[||X - \theta|| - ||X||]$, e.g., see Gervini [8], Kemperman [14] for properties such as existence, uniqueness, and robustness (breakdown point). Thus, the gradient $\nabla L(\theta) = E[\nabla f_i(\theta)\mid \theta = X]$ is bounded for $\beta = 0$. However, the robustness of the geometric median leaves only a small positive impact for further variance reduction. Thus, too large (constant) streaming batch sizes $C_p$ hinders the convergence as we make too few iterations. These findings can be extended to Figures 2b to 2e, where we vary the streaming rate $\rho$ for streaming batch sizes $C_p = 1, 8, 64, and 128$, respectively, with $\beta = 0$. The lack of convergence improvements comes from $\beta = 0$, which means we do not exploit the potential of using more observations to accelerate convergence. As shown in Figure 2f, we can achieve this acceleration by simply taking $\beta = 1/3$. In addition, $\beta = 1/3$ provides optimal convergence robust to any streaming rate $\rho$. Choosing a proper $\beta > 0$ is particularly important when $C_p$ is large, as robustness is an integral part of the geometric median method.

6. Conclusions

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

There are several ways to expand our work but let us give some examples: first, we can extend our analysis to include streaming batches of any size in the spirit of the discussion after Corollary 2. Second, many machine learning problems encounter correlated variables and high-dimensional data, making an extension to non-strongly convex objectives advantageous Bach and Moulos [1], e.g., in Werger and Wintenberger [34], they use SG-based optimization methods for volatility prediction through GARCH modeling. Third, Assumption 1 requires unbiased (and independent) gradient estimates, thus, an obvious extension could incorporate a more realistic dependency assumption, thereby increasing the applicability for more models. Moreover, studying dependence may give insight into how to process dependent data optimally. Next, a natural extension would be to modify our averaging estimate from (4) to a weighted averaged version (WASSG) proposed by Mokkadem and Pelletier [20] and Boyer and Godichon-Baggioni [4], given as

$$\tilde{\theta}_{t+1} = \frac{1}{\sum_{i=1}^{t} n_i \log(1 + i)^{\lambda}} \sum_{l=1}^{t} n_l \log(1 + \bar{r}_{l-1}) \tilde{\theta}_{l-1}, \quad \tilde{\theta}_{0+1} = 0,$$

(13)
Figure 2: Geometric median for various data streams $n_t = C_p \rho^t$. See Section 5.2 for details.

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

(b) Time-varying streaming batches, $C_p = 1, \beta = 0$

(c) Time-varying streaming batches, $C_p = 8, \beta = 0$

(d) Time-varying streaming batches, $C_p = 64, \beta = 0$

(e) Time-varying streaming batches, $C_p = 128, \beta = 0$

(f) Time-varying streaming batches, $C_p = 8, \beta = 1/3$
for $\lambda > 0$ with $(\theta_t)$ following (2) or (3). We can limit the effect of bad initializations by placing more weight on the newest estimates. Following the demonstrations in Section 5, an example of this WASSG estimate $(\hat{\theta}_t)_{t=1}^n$ can be found in Figure 3 with use of $\lambda = 2$. Here we see that although the WASSG estimate in (13) may not achieve a better final error (compared to the ASSG and APSSG estimates in Figures 1f and 2f), it still achieves a better decay along the way, often referred to as parameter tracking.

![Figure 3: WASSG for various data streams $n_t = C_n \rho^t$. See Section 6 for details.](image)

(a) Linear regression, time-varying streaming batches, $C_n = 8, \beta = 1/3$  
(b) Geometric median, time-varying streaming batches, $C_n = 8, \beta = 1/3$

7. Proofs

In this section, we provide detailed proofs of the results presented in the manuscript. Purely technical results used in the proofs can be found in Appendix A. Let $(\mathcal{F}_t)_{t=1}^T$ be an increasing family of $\sigma$-fields, namely $\mathcal{F}_t = \sigma(l_1, \ldots, l_t)$ with $l_t = (l_{t,1}, \ldots, l_{t,n_t})$. Furthermore, we expand the notation with $\mathcal{F}_{t-1,j} = \sigma(l_{1,1}, \ldots, l_{i-1,j-1}, l_{t,1}, \ldots, l_{t,j})$ such that $\mathcal{F}_{t-1,0} = \mathcal{F}_{t-1}$. Meaning, $\forall 0 \leq i < j$, we have $\mathcal{F}_{t-1} \subseteq \mathcal{F}_{t-1,j} \subseteq \mathcal{F}_{t-1,j}$. Thus, by the independence of the random (differentiable) functions $(l_{i,j})$, Assumption 1 yields that $\forall t \geq 1$, $\mathbb{E}[\|\nabla_{l_{t,i}}(\theta_{t-1})\|_{\mathcal{F}_{t-1,j-1}}] = \mathbb{E}[\nabla_{l_{t,i}}(\theta_{t-1})]$ with $i = 1, \ldots, n_t$.

7.1. Proofs for Section 3

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

**Proof of Theorem 1.** Taking the quadratic norm on both sides of (2), expanding it, and take the conditional expectation, yields

$$\mathbb{E}[\|\theta_t - \theta^*\|^2 | \mathcal{F}_{t-1}] = \mathbb{E}[\|	heta_{t-1} - \theta^*\|^2] + \frac{\gamma_t^2}{n_t^2} \mathbb{E} \left[ \left\| \sum_{i=1}^{n_t} \nabla_{l_{t,i}}(\theta_{t-1}) \right\|^2 | \mathcal{F}_{t-1} \right] - \frac{2\gamma_t}{n_t} \mathbb{E} \left[ \left\| \sum_{i=1}^{n_t} \nabla_{l_{t,i}}(\theta_{t-1}) \right\|^2 | \mathcal{F}_{t-1} \right]. \tag{14}$$

To bound the second term (on the right-hand side) of (14), we first expand it as follows,

$$\sum_{i=1}^{n_t} \mathbb{E}[\|\nabla_{l_{t,i}}(\theta_{t-1})\|^2 | \mathcal{F}_{t-1}] + \sum_{i \neq j} \mathbb{E}[\langle \nabla_{l_{t,i}}(\theta_{t-1}), \nabla_{l_{t,j}}(\theta_{t-1}) \rangle | \mathcal{F}_{t-1}]. \tag{15}$$

For first term of (15), we utilize the Lipschitz continuity of $\nabla_{l_{t,i}}$, together with Assumptions 1 to 3-p, to obtain

$$\mathbb{E}[\|\nabla_{l_{t,i}}(\theta_{t-1})\|^2 | \mathcal{F}_{t-1}] \leq 2\mathbb{E}[\|\nabla_{l_{t,i}}(\theta_{t-1}) - \nabla_{l_{t,i}}(\theta^*)\|^2 | \mathcal{F}_{t-1}] + 2\mathbb{E}[\|\nabla_{l_{t,i}}(\theta^*)\|^2 | \mathcal{F}_{t-1}] \leq 2C_t^2 \|\theta_{t-1} - \theta^*\|^2 + 2\sigma^2, \tag{16}$$

11
using $\|x + y\|^2 \leq 2(\|x\|^2 + \|y\|^2)$. Next, for the second term in (15) as $F_{i-1} \subseteq F_{i-1,j} \subset F_{i-1,i}$ for all $0 \leq i < j$, we have
$$
\mathbb{E}[\langle \nabla_d l_{i,j}(\theta_{i-1}), \nabla_d l_{i,j}(\theta_{i-1}) \rangle] = \mathbb{E}[\mathbb{E}[\langle \nabla_d l_{i,j}(\theta_{i-1}), \nabla_d L(\theta_{i-1}) \rangle|F_{i-1,i}]|F_{i-1}]
$$
since $\theta_{i-1}$ and $l_{i,j}$ are $F_{i-1,j-1}$-measurable for all $0 \leq i < j$, and similarly, as $\theta_{i-1}$ is $F_{i-1}$-measurable and $F_{i-1,j-1}$-measurable for all $0 \leq i \geq 0$, we also have
$$
\mathbb{E}[\langle \nabla_d l_{i,j}(\theta_{i-1}), \nabla_d L(\theta_{i-1}) \rangle] = \mathbb{E}[\langle \mathbb{E}[\nabla_d l_{i,j}(\theta_{i-1})|F_{i-1,i}], \nabla_d L(\theta_{i-1}) \rangle|F_{i-1}] = \mathbb{E}[\mathbb{E}[\nabla_d L(\theta_{i-1})]|F_{i-1}]
$$
where $\|\nabla_d L(\theta_{i-1})\|^2 \leq C_5^2\|\theta_{i-1} - \theta^*\|^2$ as $\nabla_d L$ is $C_5$-Lipschitz continuous and $\mathbb{E}[\nabla_d L(\theta^*)] = 0$. Thus, we obtained a bound for the second term (on the right-hand side) of (14) using the bounds of the two terms in (15):
$$
\sum_{i=1}^{n_t} (2C_i^2\|\theta_{i-1} - \theta^*\|^2 + 2\sigma^2) + \sum_{i=1}^{n_t} C_i^2\|\theta_{i-1} - \theta^*\|^2 = (2C_i^2n_t + C_i^2(n_t - 1)n_t)\|\theta_{i-1} - \theta^*\|^2 + 2\sigma^2n_t. \tag{17}
$$
For the third term (on the right-hand side) of (14) we use that $L$ is $\mu$-quasi-strong convex and $\theta_{i-1}$ is $F_{i-1}$-measurable,
$$
\mathbb{E}[\langle \nabla_d l_{i,j}(\theta_{i-1}), \theta_{i-1} - \theta^* \rangle|F_{i-1}] = \mathbb{E}[\mathbb{E}[\langle \nabla_d l_{i,j}(\theta_{i-1}), \theta_{i-1} - \theta^* \rangle|F_{i-1,i}]|F_{i-1}] = \mathbb{E}[\mathbb{E}[\langle \nabla_d L(\theta_{i-1}), \theta_{i-1} - \theta^* \rangle|F_{i-1}]] \geq \mu\|\theta_{i-1} - \theta^*\|^2, \tag{18}
$$
by Assumption 1. Combining inequalities from (17) and (18) into (14) and taking the expectation on both sides of the inequality, yields the recursive relation (9):
$$
\delta_t \leq [1 - 2\mu\gamma_t + (2C_i^2 + (n_t - 1)C_i^2)n_t^{-1}\gamma_t^2]\delta_{t-1} + 2\sigma^2n_t^{-1}\gamma_t^2, \tag{19}
$$
with $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ with some $\delta_0 \geq 0$. At last, by Proposition 5, we obtain the desired inequality in (8), namely
$$
\delta_t \leq \exp\left(-\mu \sum_{i=1}^{n_t} \gamma_t \right) \exp\left(4C_i^2 \sum_{i=1}^{n_t} \gamma_t^2 \right) \exp\left[2C_i^2 \sum_{i=1}^{n_t} \frac{\gamma_t^2}{n_t} \right] \left(\delta_0 + \frac{2\sigma^2}{C_i^2} \right) + \frac{2\sigma^2}{C_i^2} \max_{i \leq t \leq s} \gamma_s.
$$
using that $(n_t - 1)n_t^{-1} \leq \|\gamma_t\|, n_t \geq 1$, and that $\max_{i \leq t \leq s} 2\sigma^2/(2C_i^2 + (n_t - 1)C_i^2) \leq \max_{1 \leq i \leq s} 2\sigma^2/(2C_i^2) = \sigma^2/C_i^2$. \hfill $\square$

**Remark 1.** The decrease of $(2C_i^2 + (n_t - 1)C_i^2)n_t^{-1}\gamma_t$ determines when the stationary phase occurs. This is more clearly seen in Proposition 4, where the inner terms directly depend on the inception of the stationary phase. Thus, by increasing $n_t$, we decrease $(2C_i^2 + (n_t - 1)C_i^2)n_t^{-1}\gamma_t$, and especially it dominates the constant $C_i$.

**Proof of Corollary 1.** By Theorem 1, we have the upper bound giving as
$$
\delta_t \leq \exp\left(-\mu \sum_{i=1}^{n_t} \gamma_t \right) \frac{2\sigma^2}{C_i^2} \max_{i \leq t \leq s} \gamma_s, \tag{19}
$$
as $n_t = C_p$, with $\pi_t = \exp((4C_i^2/C_p) \sum_{i=1}^{n_t} \gamma_t^2 \exp(2C_i^2 \|\gamma_t\| \sum_{i=1}^{n_t} \gamma_t^2 (\delta_0 + \sigma^2/C_i^2))$. The sum term $\sum_{i=1}^{n_t} \gamma_t^2 = 2C_i^2 \sum_{i=1}^{n_t} \gamma_t^2 \sum_{i=1}^{n_t} \gamma_t - 2\alpha$ in $\pi_t$ can be bounded with the help of integral tests for convergence, $\sum_{i=1}^{n_t} \gamma_t^2 - 2\alpha = 1 + \sum_{i=2}^{n_t} \gamma_t - 2\alpha \leq 1 + \int_{\gamma_t}^{\gamma_{t+1}} x^{-\alpha} dx \leq 1 + 1/(2\alpha - 1) = 2\alpha/(2\alpha - 1)$, as $\alpha \in (1/2, 1)$. Likewise, plugging $\gamma_t = C_p\gamma_t^{-\alpha}$ into the first term of (19), gives
$$
\exp\left(-\mu \sum_{i=1}^{n_t} \gamma_t \right) = \exp\left(-\mu C_p C_i^2 \sum_{i=1}^{n_t} \gamma_t^2 \right) \leq \exp\left(-\mu C_p C_i^2 \gamma_t^{-\alpha} \int_{\gamma_t}^{\gamma_{t+1}} x^{-\alpha} dx \right) = \exp\left(-\frac{\mu C_p C_i^2 \gamma_t^{-\alpha} \int_{\gamma_t}^{\gamma_{t+1}} x^{-\alpha} dx}{2^{-\alpha}} \right),
$$
using the integral test for convergence. Next, as $(\gamma_t)_{t \geq 1}$ is decreasing, then $\max_{i \leq t \leq s} \gamma_s = \gamma_{t/2}$. Combining all these findings into (19), gives us
$$
\delta_t \leq \exp\left(-\mu C_p C_i^2 \gamma_t^{-\alpha} \int_{\gamma_t}^{\gamma_{t+1}} x^{-\alpha} dx \right) \frac{2\sigma^2}{C_i^2} \max_{i \leq t \leq s} \gamma_s + \frac{2\sigma^2}{C_i^2} \gamma_t^{-\alpha} C_p C_i^2 \frac{2\sigma^2}{C_i^2} \gamma_t^{-\alpha} C_p C_i^2.
$$
with $\pi_{\gamma} = \exp(4\alpha C_i^2/(2C_i^2 + C_p\|\gamma_t\|C_i^2))/(2\alpha - 1)\gamma_t^{1-2\alpha}(\delta_0 + \sigma^2/C_i^2)$. At last, converting (20) into terms of $N_t$ using $N_t = C_p\gamma_t$, yields the desired. \hfill $\square$
Proof of Corollary 2. For convenience, we divided the proof into two cases to comprehend how that \( n_t \geq 1 \) for all \( t \): first, we bound each term of (8) (from Theorem 1) after inserting, \( \gamma_t = C_p n_t^{1/\gamma_t} = C_p \gamma_t^{1+\beta_t} \) if \( \rho < 0 \), or \( \gamma_t \geq C_p^{1/\gamma_t} \) if \( \rho > 0 \) (using that \( \beta \geq 0 \)) into the inequality. If \( \rho \geq 0 \), the first term of (8) can be bounded, as follows:

\[
\exp \left( -\mu \sum_{i=1}^t \gamma_i \right) = \exp \left( -\mu C_p \gamma_t \sum_{i=1}^t \gamma_i^{1+\beta_t} \right) \leq \exp \left( -\mu C_p \gamma_t \sum_{i=1}^t \gamma_i^{1+\beta_t} \right),
\]

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

\[
\exp \left( -\mu \sum_{i=1}^t \gamma_i \right) \leq \exp \left( -\mu C_p \gamma_t \sum_{i=1}^t \gamma_i^{1+\beta_t} \right) \leq \exp \left( -\mu C_p \gamma_t \sum_{i=1}^t \gamma_i^{1+\beta_t} \right).
\]

Likewise, with the help of integral tests for convergence, we have bounded the last term of (8) as follows:

\[
\frac{2\sigma^2}{\mu} \max_{i \leq t} \frac{\gamma_i}{n_i} \leq \frac{2\sigma^2 C_p^{1+\beta}\gamma_t^{1+\beta_t}}{\mu \gamma_t^{1+\beta_t}} \leq \frac{2\sigma^2 C_p^{1+\beta}\gamma_t^{1+\beta_t}}{\mu \gamma_t^{1+\beta_t}}.
\]

Likewise, if \( \rho < 0 \), we have

\[
\frac{2\sigma^2}{\mu} \max_{i \leq t} \frac{\gamma_i}{n_i} \leq \frac{2\sigma^2 C_p^{1+\beta}\gamma_t^{1+\beta_t}}{\mu \gamma_t^{1+\beta_t}} \leq \frac{2\sigma^2 C_p^{1+\beta}\gamma_t^{1+\beta_t}}{\mu \gamma_t^{1+\beta_t}}.
\]

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

\[
\delta_t \leq \exp \left( -\mu C_p \gamma_t \sum_{i=1}^t \gamma_i^{1+\beta_t} \right) \pi_{\infty}^0 \leq \frac{2\sigma^2 C_p^{1+\beta}\gamma_t^{1+\beta_t}}{\mu \gamma_t^{1+\beta_t}} \pi_{\infty}^0, \quad \text{for } \rho \geq 0,
\]

where \( \pi_{\infty}^0 = \exp(4(\alpha - \beta \rho)C_p^{1+\beta}/(2C_p^2 + C_p^{1+\beta}))/2(\alpha - \beta \rho) - 1 \) with \( \rho = \rho_{\infty} \) and \( \phi = ((1 - \beta \rho + \alpha)/(1 + \rho)) \). To write this in terms of \( N_t \), we use that \( N_t = \sum_{i=1}^t n_i = C_p \sum_{i=1}^t \gamma_i^{1+\beta_t} \leq C_p \gamma_t^{1+\beta_t} \leq C_p \gamma_t^{1+\beta_t} \leq C_p \gamma_t^{1+\beta_t} \leq 2C_p \gamma_t^{1+\beta_t} \), for \( \rho \geq 0 \), thus, \( t \geq (N_t/2C_p)^{1/(1+\beta)} \). For \( \rho < 0 \), we have \( N_t \leq C_p t, i.e., t \geq N_t/\rho \).

7.2. Proofs for Section 4

Lemma 1 (ASSG/APPSS). Denote \( \Delta_t = \mathbb{E}[\|\theta_t - \theta\|^4] \) for some \( \Delta_0 \geq 0 \), where (6) follows (2) or (3). Under Assumption 1, Assumptions 2-p and 3-p with \( p = 4 \) and Assumption 4, we have for any learning rate (\( \gamma_t \)) that

\[
\Delta_t \leq \exp \left( -\mu \sum_{i=1}^t \gamma_i \right) \Pi_t^0 \leq \frac{32\sigma^4}{\mu^2} \max_{i \leq t} \frac{\gamma_i^2}{n_i^2} + \frac{48\sigma^4}{\mu} \max_{i \leq t} \frac{\gamma_i^3}{n_i^3} + \frac{114\sigma^4}{\mu} \max_{i \leq t} \frac{\gamma_i^4}{n_i^4},
\]

with \( \Pi_t^0 \) given in (30).

Proof of Lemma 1. We will now derive the recursive step sequence for the fourth-order moment using the same arguments as in proof of Theorem 1. Thus, one can show that

\[
\mathbb{E}[\|\theta_t - \theta\|^4 | \mathcal{F}_{t-1}] \leq \|\theta_{t-1} - \theta\|^4 + \gamma_t^4 \mathbb{E} \left[ \left\| \sum_{i=1}^n \nabla^2 \ell_i((\theta_{t-1})) \right\|^4 | \mathcal{F}_{t-1} \right] + \frac{4\gamma_t^2}{n_t} \mathbb{E} \left[ \left\| \sum_{i=1}^n \nabla \ell_i((\theta_{t-1}), \theta_{t-1} - \theta') \right\|^2 | \mathcal{F}_{t-1} \right] + \frac{2\gamma_t^2}{n_t} \mathbb{E} \left[ \left\| \sum_{i=1}^n \nabla \ell_i((\theta_{t-1})) \right\|^2 | \mathcal{F}_{t-1} \right] - \frac{2\gamma_t^2}{n_t} \mathbb{E} \left[ \left\| \sum_{i=1}^n \nabla \ell_i((\theta_{t-1})) \right\|^2 | \mathcal{F}_{t-1} \right] - \frac{\gamma_t^2}{n_t} \mathbb{E} \left[ \left\| \sum_{i=1}^n \nabla^2 \ell_i((\theta_{t-1})) \right\|^2 | \mathcal{F}_{t-1} \right].
\]

13
using $\theta_{t-1}$ is $\mathcal{F}_{t-1}$-measurable. Note, by Assumption 1, we have
\[\langle E[\nabla_{\theta_{t-1}}(\theta_{t})]f_{t-1}, \theta_{t-1} - \theta^* \rangle = \langle \nabla_{\theta}L(\theta_{t-1}), \theta_{t-1} - \theta^* \rangle \geq \mu \|\theta_{t-1} - \theta^*\|^2,\]
as $L$ is $\mu$-quasi-strong convex. Combining this with the Cauchy-Schwarz inequality $\langle x, y \rangle \leq \|x\|\|y\|$, we obtain the simplified expression:
\[
E[\|\theta_{t} - \theta^*\|^4_{\mathcal{F}_{t-1}}] \leq \|\theta_{t-1} - \theta^*\|^4 + \frac{\gamma_t^4}{n_t^2} \left( \sum_{i=1}^{n_t} \|\nabla_{\theta_{t-1}}(\theta_{t-1})\|^4_{\mathcal{F}_{t-1}} \right) + 4\gamma_t^3 \|\theta_{t-1} - \theta^*\|^2_{\mathcal{F}_{t-1}} \left( \sum_{i=1}^{n_t} \|\nabla_{\theta_{t-1}}(\theta_{t-1})\|^2_{\mathcal{F}_{t-1}} \right)
-
4\mu \gamma_t \|\theta_{t-1} - \theta^*\|^4 + \frac{4\gamma_t^3}{n_t^2} \left( \sum_{i=1}^{n_t} \|\nabla_{\theta_{t-1}}(\theta_{t-1})\|^3_{\mathcal{F}_{t-1}} \right).
\]
Next, recall Young’s Inequality, i.e., for any $a, b, c > 0$ we have $ab \leq a^2 c^2 / 2 + b^2 / 2c^2$,
\[
\left( \sum_{i=1}^{n_t} \|\nabla_{\theta_{t-1}}(\theta_{t-1})\|^3 \right) \leq \frac{\gamma_t}{2n_t} \|\theta_{t-1} - \theta^*\|^3 + \frac{2\gamma_t^3}{n_t^2} \left( \sum_{i=1}^{n_t} \|\nabla_{\theta_{t-1}}(\theta_{t-1})\|^2_{\mathcal{F}_{t-1}} \right)
,
\]
giving us
\[
E[\|\theta_{t} - \theta^*\|^4_{\mathcal{F}_{t-1}}] \leq (1 - 4\mu \gamma_t) \|\theta_{t-1} - \theta^*\|^4 + \frac{3\gamma_t^4}{n_t^2} \left( \sum_{i=1}^{n_t} \|\nabla_{\theta_{t-1}}(\theta_{t-1})\|^4_{\mathcal{F}_{t-1}} \right) + \frac{8\gamma_t^3}{n_t^2} \|\theta_{t-1} - \theta^*\|^2_{\mathcal{F}_{t-1}} \left( \sum_{i=1}^{n_t} \|\nabla_{\theta_{t-1}}(\theta_{t-1})\|^2_{\mathcal{F}_{t-1}} \right).
\]
(23)

To bound the second and fourth-order terms in (23), we would need to study the recursive sequences: firstly, utilizing the Lipschitz continuity of $\nabla_{\theta_{t-1}}$, together with Assumptions 2-p and 3-p, and that $\theta_{t-1}$ is $\mathcal{F}_{t-1}$-measurable (Assumption 1), we obtain
\[
E[\|\nabla_{\theta_{t-1}}(\theta_{t-1})\|^p_{\mathcal{F}_{t-1}}] \leq 2^{1-p} E[\|\nabla_{\theta_{t-1}}(\theta_{t-1})\|^p_{\mathcal{F}_{t-1}}] + E[\|\nabla_{\theta_{t-1}}(\theta_{t-1})\|^p_{\mathcal{F}_{t-1}}] \leq 2^{1-p} C^1_p \|\theta_{t-1} - \theta^*\|^p + \sigma^p
\]
for any $p \in [1, 4]$ using the bound $\|x + y\|^p \leq 2^{p-1}(\|x\|^p + \|y\|^p)$. Thus, we can bound the second-order term in (23) by
\[
E\left[ \left( \sum_{i=1}^{n_t} \|\nabla_{\theta_{t-1}}(\theta_{t-1})\|_{\mathcal{F}_{t-1}} \right)^2 \right] \leq 2C^2 n_t + C^2 (n_t - 1)n_t \|\theta_{t-1} - \theta^*\|^2 + 2\sigma^2 n_t \leq 2C^2 n_t + C^2 (n_t - 1)n_t \|\theta_{t-1} - \theta^*\|^2 + 2\sigma^2 n_t,
\]
(25)
following the same steps in the proof of Theorem 1, but with use of (24). Bounding the fourth-order term is a bit heavier computationally, but let us recall that $\|\sum x_i\|^2 = \sum \|x_i\|^2 + \sum_{i < j} \langle x_i, x_j \rangle = \sum \|x_i\|^2 + 2\sum_{i < j} \langle x_i, x_j \rangle$. Then, we have that
\[
\left( \sum_{i=1}^{n_t} \|\nabla_{\theta_{t-1}}(\theta_{t-1})\|^2 \right)^2 \leq 2 \left( \sum_{i=1}^{n_t} \|\nabla_{\theta_{t-1}}(\theta_{t-1})\|^2 \right)^2 + 4 \left( \sum_{i<j} \langle \nabla_{\theta_{t-1}}(\theta_{t-1}), \nabla_{\theta_{t-1}}(\theta_{t-1}) \rangle \right)^2,
\]
(26)
as $(x + y)^2 \leq 2x^2 + 2y^2$. For the first term of (26), we have
\[
E\left[ \left( \sum_{i=1}^{n_t} \|\nabla_{\theta_{t-1}}(\theta_{t-1})\|^2 \right)^2 \right] = \sum_{i=1}^{n_t} E[\|\nabla_{\theta_{t-1}}(\theta_{t-1})\|^2_{\mathcal{F}_{t-1}}] + \sum_{i < j} E[\|\nabla_{\theta_{t-1}}(\theta_{t-1})\|^2_{\mathcal{F}_{t-1}} \|\nabla_{\theta_{t-1}}(\theta_{t-1})\|^2_{\mathcal{F}_{t-1}}]
\leq 8n_t C^2 \|\theta_{t-1} - \theta^*\|^4 + 4n_t^2 \sum_{i < j} (C^2 \|\theta_{t-1} - \theta^*\|^2 + \sigma^2) \]
using the bound from (24), \( n_i(n_i - 1) \leq n_i^2 \|v_i\|_3^2 \), and that \( \mathcal{F}_{T-1} \subseteq \mathcal{F}_{T-1,i} \) for all \( 0 \leq i < j \). To bound the second term of (26), we ease notation by denoting \( \nabla \varphi_i, (\theta_{i-1}) \) by \( v_i \), giving us

\[
\left( \sum_{i < j} \langle v_i, v_j \rangle \right)^2 = \sum_{i, j} \langle v_i, v_j \rangle^2 + \sum_{i, j \neq k} \langle v_i, v_j \rangle \langle v_k, v_i \rangle = \sum_{i, j, k} \langle v_i, v_j \rangle^2 + \sum_{i, j \neq k} \langle v_i, v_j \rangle \langle v_k, v_i \rangle + \sum_{i, j \neq k,l} \langle v_i, v_j \rangle \langle v_k, v_l \rangle.
\]

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

\[
\mathbb{E}[A|\mathcal{F}_{T-1}] \leq \sum_{i < j} \mathbb{E}[\|v_i\|_3^3 \|v_j\|_3^3 | \mathcal{F}_{T-1}] \leq 2n_i(n_i - 1)(\mathbb{C}_2^2 \|\theta_{i-1} - \theta^*\| + \sigma^2)^2 \leq 2n_i^2 \|v_i\|_3^2 \|\mathbb{C}_2^2\| \|\theta_{i-1} - \theta^*\| + \sigma^2)^2,
\]

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

\[
\mathbb{E}[B|\mathcal{F}_{T-1}] = \sum_{i < j < k \neq l} \mathbb{E}[\|v_i\|_3^2 \|v_j\|_3^2 \|v_k\|_3 \|v_l\|_3 | \mathcal{F}_{T-1}] = \sum_{i < j < k \neq l} \mathbb{E}[\|\nabla \varphi_i, (\theta_{i-1})\|_2^2 | \mathcal{F}_{T-1}] \leq \sum_{i < j < k \neq l} \mathbb{E}[\|\nabla \varphi_i, (\theta_{i-1})\|_2^2 \|v_i\|_3^2 | \mathcal{F}_{T-1}, \mathcal{F}_{T-1}]
\]

\[
\leq \sum_{i < j < k \neq l} 2\mathbb{C}_2^2 \|\theta_{i-1} - \theta^*\|^2 \mathbb{C}_2^2 \|\theta_{i-1} - \theta^*\|^2 + \sigma^2) = n_i(n_i - 1)(n_i - 2)(\mathbb{C}_2^2 \|\theta_{i-1} - \theta^*\|^2 + \sigma^2)^2 \leq n_i^2 \|v_i\|_3^2 \mathbb{C}_2^2 \|\theta_{i-1} - \theta^*\|^2 + \sigma^2),
\]

using the Cauchy-Schwarz inequality and the bound in (24). In the same way, as \( j \neq l \) includes \((i, j) \neq (k, l)\), we can rewrite \( C \) as

\[
C = \sum_{i < j < k \neq l} \langle v_i, v_j \rangle \langle v_k, v_l \rangle = \sum_{i < j < k \neq l} \langle v_i, v_j \rangle \langle v_k, v_l \rangle + \sum_{i < j < k \neq l} \langle v_i, v_j \rangle \langle v_k, v_l \rangle,
\]

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

\[
C_2 = \sum_{i < j < k \neq l} \langle v_i(v_j) \rangle \langle v_k \rangle \langle v_l \rangle + \sum_{i < j < k \neq l} \langle v_i(v_j) \rangle \langle v_k \rangle \langle v_l \rangle + \sum_{i < j < k \neq l} \langle v_i(v_j) \rangle \langle v_k \rangle \langle v_l \rangle,
\]

where \( \mathbb{E}[C_2|\mathcal{F}_{T-1}] = \mathbb{E}[C_2|\mathcal{F}_{T-1}] = \mathbb{E}[B|\mathcal{F}_{T-1}] \), and

\[
\mathbb{E}[C_3|\mathcal{F}_{T-1}] = \sum_{i < j < k \neq l} \mathbb{E}[\|\nabla \varphi_i, (\theta_{i-1})\|_2^4 | \mathcal{F}_{T-1}] \leq n_i(n_i - 1)(n_i - 2)(n_i - 3) \mathbb{C}_4^2 \|\theta_{i-1} - \theta^*\|^4 \leq n_i^3 \|v_i\|_3^2 \mathbb{C}_4^2 \|\theta_{i-1} - \theta^*\|^4.
\]

Thus, the fourth-order term of (23), is bounded by

\[
\mathbb{E} \left[ \left\| \sum_{i=1}^n \nabla \varphi_i, (\theta_{i-1}) \right\|_2^4 | \mathcal{F}_{T-1} \right] \leq 16n_i[\mathbb{C}_2^2 \|\theta_{i-1} - \theta^*\|^2 + \sigma^2] + 16n_i^2 \|v_i\|_3 \|\mathbb{C}_2^2\| \|\theta_{i-1} - \theta^*\|^2 + \sigma^2)^2
\]

\[
+ 12n_i^3 \|v_i\|_3 \mathbb{C}_2^2 \|\theta_{i-1} - \theta^*\|^2 \|\mathbb{C}_2^2\| \|\theta_{i-1} - \theta^*\|^2 + \sigma^2)^2 + 4n_i^4 \|v_i\|_3^2 \mathbb{C}_2^2 \|\theta_{i-1} - \theta^*\|^4
\]

\[
\leq [16\mathbb{C}_2^2]n_i + 16\mathbb{C}_2^2 n_i^2 \|v_i\|_3 + 12\mathbb{C}_2^2 n_i^3 \|v_i\|_3^2 + 4\mathbb{C}_2^2 n_i^4 \|v_i\|_3^3] \|\theta_{i-1} - \theta^*\|^2 + 32\mathbb{C}_2^2 \sigma^2 n_i^2 \|v_i\|_3^2 \|\theta_{i-1} - \theta^*\|^2 + 16\sigma^2 n_i + 16\sigma^2 n_i^2 \|v_i\|_3^2. \tag{27}
\]
Combining the bound from (25) and (27) into (23), we obtain the recursive relation for the fourth-order moment:

\[ \mathbb{E}[||\hat{\theta}_i - \theta^*||^4 | \mathcal{F}_{t-1}] \leq [1 - 4\mu \gamma_{i} + 8\sigma^2 \int_{[0,v]} \gamma_{i}^2 + 16C^2 \nu_{1} \gamma_{i}^2 + 48C^2 \nu_{1} \gamma_{i}^4 + 48C^2 \nu_{1} \gamma_{i}^6 + 36C^2 \nu_{1} \gamma_{i}^8 ] + 12C^2 \int_{[0,v]} \gamma_{i}^4 ||\theta_{t-1} - \theta^*||^4 + [16\sigma^2 \nu_{1} \gamma_{i}^2 + 96C^2 \sigma^2 \nu_{1} \gamma_{i}^4 + 36C^2 \sigma^2 \nu_{1} \gamma_{i}^6 ] ||\theta_{t-1} - \theta^*||^2 + 48\sigma^4 \nu_{1} \gamma_{i}^4 + 48\sigma^4 \nu_{1} \gamma_{i}^6 \int_{[0,v]} \gamma_{i}^4 . \]

Using the Young’s inequalities, \( 2C^2 \nu_{1} \leq \nu_{1} C_{2} + \nu_{1}^{-1} C_{1}^4, 16\sigma^2 \nu_{1} \gamma_{i}^2 ||\theta_{t-1} - \theta^*||^2 \leq 2\mu \gamma_{i} ||\theta_{t-1} - \theta^*||^4 + 32\sigma^4 \mu \nu_{1} \gamma_{i}^2 \), \( 2C^2 \sigma^2 \nu_{1} \gamma_{i}^4 ||\theta_{t-1} - \theta^*||^4 + \sigma^4 \nu_{1} \gamma_{i}^2 \), and \( 2C^2 \sigma^2 \nu_{1} \gamma_{i}^6 ||\theta_{t-1} - \theta^*||^2 \leq C^2 \int_{[0,v]} \gamma_{i}^2 \), yields:

\[ \mathbb{E}[||\hat{\theta}_i - \theta^*||^4 | \mathcal{F}_{t-1}] \leq [1 - 2\mu \gamma_{i} + 8\sigma^2 \int_{[0,v]} \gamma_{i}^2 + 16C^2 \nu_{1} \gamma_{i}^2 + 48C^2 \nu_{1} \gamma_{i}^4 + 48C^2 \nu_{1} \gamma_{i}^6 ] ||\theta_{t-1} - \theta^*||^4 + 32\mu^2 \nu_{1} \gamma_{i}^4 + 48\sigma^4 \nu_{1} \gamma_{i}^4 + 114\sigma^4 \nu_{1} \gamma_{i}^4 . \]

Taking, the expectation on both sides of the inequality in (28) yields the recursive relation for the fourth-order moment:

\[ \Delta_{i} \leq [1 - 2\mu \gamma_{i} + 8\sigma^2 \int_{[0,v]} \gamma_{i}^2 + 16C^2 \nu_{1} \gamma_{i}^2 + 48C^2 \nu_{1} \gamma_{i}^4 + 48C^2 \nu_{1} \gamma_{i}^6 ] \Delta_{i-1} + 32\mu^2 \nu_{1} \gamma_{i}^4 + 48\sigma^4 \nu_{1} \gamma_{i}^4 + 114\sigma^4 \nu_{1} \gamma_{i}^4 . \]

with \( \Delta_{i} = \mathbb{E}[||\hat{\theta}_i - \theta^*||^4] \) for some \( \Delta_{0} \geq 0 \). By Proposition 5, we achieve the (upper) bound of \( \Delta_{i} \) in (29), given as

\[ \Delta_{i} \leq \mathbb{E}[\mu \sum_{l=1}^{i} \gamma_{l} \Pi_{l}^i + 32\sigma^{2} \max_{l<2i} \frac{\gamma_{l}^{2}}{n_{l}^{2}} + 48\sigma^{4} \max_{l<2i} \frac{\gamma_{l}^{4}}{n_{l}^{4}} + 114\sigma^{4} \max_{l<2i} \frac{\gamma_{l}^{6}}{n_{l}^{6}} ] , \]

where \( \Pi_{l}^i \) is given by

\[ \exp \left( 32C_{2} \sum_{l=1}^{i} \frac{\gamma_{l}^{2}}{n_{l}} \right) \exp \left( 96C_{2} \sum_{l=1}^{i} \frac{\gamma_{l}^{4}}{n_{l}} \right) \exp \left( 228C_{4} \sum_{l=1}^{i} \frac{\int_{[0,v]} \gamma_{l}^{2}}{n_{l}^{2}} \right) \exp \left( 16C_{2} \sum_{l=1}^{i} \frac{\int_{[0,v]} \gamma_{l}^{4}}{n_{l}^{4}} \right) , \]

with use of

\[ \max_{l<2i} 8C_{2} \int_{[0,v]} \gamma_{l}^{2} + 16C_{2} \nu_{1} \gamma_{l}^{2} + 48C_{2} \nu_{1} \gamma_{l}^{4} + 114C_{2} \nu_{1} \gamma_{l}^{6} + 48C_{2} \mu \gamma_{l} \leq \frac{\sigma^{4}}{C_{4}} + \frac{2\sigma^{4} \gamma_{l}}{\mu C_{4}^{2} n_{l}} , \]

At last, bounding the projected estimate (3) follows from that \( \mathbb{E}[||P_{\theta}(\theta) - \theta^*||^2] \leq \mathbb{E}[||\theta - \theta^*||^2], \forall \theta \in \Theta. \)

\[ \square \]

7.2.1. Proofs for Section 4.1

Proof of Theorem 2. Following Polyak and Juditsky [26], we rewrite (2) to

\[ \theta_{i} = \theta_{i-1} - \frac{n_{i} \sum_{l=1}^{n_{i}} \nabla_{\theta} \ell_{l}(\theta_{t-1})}{n_{i} \sum_{l=1}^{n_{i}} \nabla_{\theta} \ell_{l}(\theta_{t-1})} \approx \frac{1}{n_{i}} (\theta_{t-1} - \theta_{i}) = \nabla_{\theta} l_{i}(\theta_{t-1}) , \]

where \( \nabla_{\theta} l_{i}(\theta_{t-1}) \) denotes \( n_{i} \sum_{l=1}^{n_{i}} \nabla_{\theta} \ell_{l}(\theta_{t-1}) \). Note \( \nabla_{\theta} l_{i}(\theta_{t-1}) \approx \nabla_{\theta} l_{i}(\theta^*) + \nabla_{\theta} L_{L}(\theta^*) (\theta_{t-1} - \theta^*) \), and that \( \nabla_{\theta} l_{i}(\theta^*) \) and \( \nabla_{\theta} l_{i}(\theta) \) behaves almost like an i.i.d. sequences with zero mean. Thus, \( \theta_{i} - \theta^* \) behaves like \( -\nabla_{\theta} L_{L}(\theta^*)^{-1} N_{i} \sum_{l=1}^{n_{i}} n_{l} \nabla_{\theta} l_{i}(\theta^*) \) leading to a bound in \( O(\sqrt{N_{i}}) \). Observe that

\[ \nabla_{\theta} L_{L}(\theta^*) (\theta_{t-1} - \theta^*) = \nabla_{\theta} l_{i}(\theta_{t-1}) - \nabla_{\theta} l_{i}(\theta^*) - \nabla_{\theta} L_{L}(\theta_{t-1}) = \nabla_{\theta} L_{L}(\theta_{t-1}) - \nabla_{\theta} L_{L}(\theta_{t-1} - \theta^*) , \]

marignale term  rest term
where $\nabla^2 \theta$ is invertible with lowest eigenvalue greater than $\mu$, i.e., $\nabla^2 \theta \geq \mu$. Thus, summing the parts and using the Minkowski’s inequality, we obtain the inequality:

$$\left( \mathbb{E} \left[ \|\hat{\theta}_t - \theta\|^2 \right] \right)^{\frac{1}{2}} \leq \left( \mathbb{E} \left[ \left\| \nabla^2 \theta L(\theta) \right\| \mathbb{E} \left[ \| n_i \nabla d_i(\theta) \| \right] \right]^2 \right)^{\frac{1}{2}} + \left( \mathbb{E} \left[ \left\| \nabla^2 \theta L(\theta) \mathbb{E} \left[ \| n_i \nabla d_i(\theta_{t-1}) \| \right] \right]^2 \right)^{\frac{1}{2}} + \left( \mathbb{E} \left[ \left\| \nabla^2 \theta L(\theta) \mathbb{E} \left[ \| n_i \nabla d_i(\theta_{t-1}) \| \right] \right]^2 \right)^{\frac{1}{2}}$$

As ($\nabla d_i(\theta)$) is a square-integrable martingale increment sequences on $\mathbb{R}^d$ (Assumption 1), we have

$$\mathbb{E} \left[ \left\| \nabla^2 \theta L(\theta) \right\| \mathbb{E} \left[ \| n_i \nabla d_i(\theta) \| \right] \right] \leq \frac{1}{N_i^2} \sum_{i=1}^{n} \left\| n_i \nabla d_i(\theta) \right\|^2 \leq \frac{\text{Tr} \left[ \nabla^2 \theta L(\theta) \right]}{N_i}$$

using Assumption 4. To ease notation, we denote $\text{Tr} \left[ \nabla^2 \theta L(\theta) \right]$ by $\Lambda$. Next, note that for all $t \geq 1$, we have the relation in (31), giving us

$$\frac{1}{N_i} \sum_{i=1}^{n} n_i \nabla d_i(\theta_{t-1}) = \frac{1}{N_i} \sum_{i=1}^{n} n_i \left( \theta_{t-1} - \theta_i \right) = \frac{1}{N_i} \sum_{i=1}^{n} \left( \theta_{i} - \theta \right) \left( n_{i+1} - n_i \right) \frac{n_i}{\gamma_i} + \frac{1}{N_i} \left( \theta - \theta \right) n_i \frac{1}{\gamma_i}$$

leading to

$$\nabla^2 \theta L(\theta)^{-1} \frac{1}{N_i} \sum_{i=1}^{n} n_i \nabla d_i(\theta_{t-1}) \leq \left( \frac{1}{N_i} \sum_{i=1}^{n} \left\| \nabla d_i(\theta_{t-1}) \right\|^2 \right)^{\frac{1}{2}} \leq \left( \frac{1}{N_i} \sum_{i=1}^{n} \left\| \nabla d_i(\theta_{t-1}) \right\| \right)^{\frac{1}{2}}$$

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

$$\left( \mathbb{E} \left[ \left\| \nabla^2 \theta L(\theta) \right\| \right] \right)^{\frac{1}{2}} \leq \frac{1}{N_i} \sum_{i=1}^{n} \left\| \nabla d_i(\theta_{t-1}) \right\|^2 \leq \frac{1}{N_i} \sum_{i=1}^{n} \left\| \nabla d_i(\theta_{t-1}) \right\|^2 \leq \left( \frac{1}{N_i^2} \sum_{i=1}^{n} \left\| \nabla d_i(\theta_{t-1}) \right\|^2 \right)^{\frac{1}{2}} \leq \left( \frac{1}{N_i^2} \sum_{i=1}^{n} \left\| \nabla d_i(\theta_{t-1}) \right\|^2 \right)^{\frac{1}{2}} \leq \frac{\Lambda}{N_i^2}$$

(33)

For the martingale term, we have

$$\mathbb{E} \left[ \left\| \nabla^2 \theta L(\theta) \right\| \mathbb{E} \left[ \| n_i \nabla d_i(\theta_{t-1}) \| \right] \right] \leq \frac{1}{N_i^2} \sum_{i=1}^{n} \left\| \nabla d_i(\theta_{t-1}) \right\|^2 \leq \left( \frac{1}{N_i^2} \sum_{i=1}^{n} \left\| \nabla d_i(\theta_{t-1}) \right\|^2 \right)^{\frac{1}{2}} \leq \frac{\Lambda}{N_i^2} \sum_{i=1}^{n} \left\| \nabla d_i(\theta_{t-1}) \right\|^2 \leq \frac{C \Lambda}{N_i^2} \sum_{i=1}^{n} \left\| \nabla d_i(\theta_{t-1}) \right\|^2$$

(34)

by the Cauchy-Schwarz inequality and Assumption 2-p. For all $t \geq 1$, the rest term is directly bounded by (7):

$$\left( \mathbb{E} \left[ \left\| \nabla^2 \theta L(\theta) \right\| \mathbb{E} \left[ \| n_i \nabla d_i(\theta_{t-1}) \| \right] \right] \right)^{\frac{1}{2}} \leq \frac{\Lambda}{N_i^2} \sum_{i=1}^{n} \left\| \nabla d_i(\theta_{t-1}) \right\|^2$$

(35)

with the notion $\Delta_t = \mathbb{E}[\|\theta_t - \theta\|^2]$. Finally, combining the terms from (32) to (35), gives us

$$\delta_t \leq \frac{\Lambda}{N_i^2} \sum_{i=1}^{n} \left\| \nabla d_i(\theta_{t-1}) \right\|^2$$

(36)

where $\delta_t = \mathbb{E}[\|\theta_t - \theta\|^2]$, which can be simplified into (12) by shifting the indices and collecting the $\delta_0$ terms. □
Proof of Corollary 3. As \( n_r = C_p \) for all \( t \geq 1 \), we simplify the bound for \( \bar{\tilde{\delta}_i} \) in (12) to

\[
\bar{\tilde{\delta}_i}^{1/2} \leq \frac{A_{1/2}^{1/2}}{N_{1/2}^{1/2}} + \frac{C_p}{N_{1/2}^{1/2}} \sum_{i=0}^{t-1} \delta_{1/2}^{1/2} \left| \frac{1}{\gamma_i+1} - \frac{1}{\gamma_i} \right| + \frac{C_p}{N_{1/2}^{1/2}} \left( \frac{1}{\gamma_i} + C_1 \right) \delta_{1/2}^{1/2} + \frac{C_p}{N_{1/2}^{1/2}} \left( \sum_{i=1}^{t-1} \rho \right)^{1/2} + \frac{C_p}{N_{1/2}^{1/2}} \sum_{i=0}^{t-1} \Delta_i^{1/2}. \tag{37}
\]

The second-order moment \( \delta_i \) is bounded by Corollary 1 but with use of (20) as we work in terms of \( t \). The fourth-order moment \( \Delta_i \) from Lemma 1 can be simplified to:

\[
\Delta_i \leq \exp \left( -\mu \sum_{i=1}^{t} \gamma_i \right) \Pi_\infty + \frac{1}{\mu} \left( \frac{2\sigma^4}{C_p} \max \gamma_i^2 + \frac{48\sigma^4}{C_p} \max \gamma_i^3 + \frac{114\sigma^4}{C_p} \max \gamma_i^4 \right),
\]

using that \( \gamma_i = C_p C_p^{1/2} \). Decreasing as \( \alpha \in (1/2, 1) \). Regarding \( \Pi_\infty \) defined in (30), we can bound it by

\[
\Pi_\infty = \exp \left( \frac{64\sigma^4 \gamma_i^2}{(2\alpha - 1) C_p} \right) \left( \frac{(192 + 456C_p\bar{\delta}_{[\gamma_i-1]}C_p^{1/2})}{C_p^{1/2}} \right) \left( \frac{528C_p^{3/2}}{2\alpha - 1} \right),
\]

using \( \sum_{i=1}^{t} \gamma_i^2 \leq \alpha \gamma_i/(2\alpha - 1) \) and \( \sum_{i=1}^{t} \gamma_i^3 \leq 2 \). Note that \( \Pi_\infty \) is a finite constant, independent of \( t \). To bound the first term of (37), namely \( \frac{C_p}{N_{1/2}} \sum_{i=1}^{t-1} \left| \frac{1}{\gamma_i+1} - \frac{1}{\gamma_i} \right| \), we remark that \( \sum_{i=1}^{t-1} \frac{1}{\gamma_i+1} \gamma_i^1 = \gamma_i^1 \), one has (since \( \sqrt{a+b} \leq \sqrt{a} + \sqrt{b} \)).

\[
\frac{C_p}{N_{1/2}} \sum_{i=1}^{t-1} \delta_{1/2}^{1/2} \left| \frac{1}{\gamma_i+1} - \frac{1}{\gamma_i} \right| \leq C_p \frac{\gamma_i^{1/2} \gamma_i^1}{C_p \mu N_1} \sum_{i=1}^{t-1} \rho - 1 \exp \left( -\frac{\mu C_p \gamma_i^{1/2} \gamma_i^1}{2\gamma_i^2} \right) \frac{\gamma_i^{1/2}}{\sqrt{\gamma_0}} + \frac{2^{1/2} \sigma \sqrt{C_p^{1/2}}}{\sqrt{\gamma_0 \gamma_0}}. \tag{38}
\]

For simplicity, let us denote

\[
\mathcal{A}_{\infty}^{c} = \sum_{i=0}^{\infty} \exp \left( -\frac{\mu C_p \gamma_i^{1/2} \gamma_i^1}{2\gamma_i^2} \right) \geq \sum_{i=0}^{\infty} \rho - 1 \exp \left( -\frac{\mu C_p \gamma_i^{1/2} \gamma_i^1}{2\gamma_i^2} \right),
\]

as \( \alpha < 1 \). Thus, the first part of (38) is bounded as follows:

\[
\frac{C_p^{1/2}}{C_p \gamma_i^{1/2} \mu N_1} \sum_{i=1}^{t-1} \rho - 1 \exp \left( -\frac{\mu C_p \gamma_i^{1/2} \gamma_i^1}{2\gamma_i^2} \right) \leq C_p^{1/2} \frac{\gamma_i^{1/2}}{\sqrt{\gamma_0 \gamma_0}} \mathcal{A}_{\infty}^{c},
\]

Furthermore, with the help of an integral test for convergence, one has \( \sum_{i=1}^{t} \gamma_i^{\alpha/2} \leq 1 + \int_1^{t} \gamma_i^{\alpha/2} ds = 1 + (2/\alpha) \gamma_i^{\alpha/2} - (2/\alpha) \gamma_i^{\alpha/2} \), such that the second part of (38) can be bounded by

\[
\frac{2^{1/2} \sigma C_p^{1/2}}{C_p^{1/2} \mu N_1} \sum_{i=1}^{t} \rho - 1 \exp \left( -\frac{\mu C_p \gamma_i^{1/2} \gamma_i^1}{2\gamma_i^2} \right) \leq 2^{1/2} \sigma C_p^{1/2} \gamma_i^{1/2} / \mu N_1 = C_p^{1/2} \mu N_1 \gamma_i^{1/2} / \mu N_1 \gamma_i^{1/2}.
\]

By combining this, we get

\[
\frac{C_p}{N_{1/2}} \sum_{i=1}^{t-1} \delta_{1/2}^{1/2} \left| \frac{1}{\gamma_i+1} - \frac{1}{\gamma_i} \right| \leq C_p^{1/2} \frac{\gamma_i^{1/2}}{\sqrt{\gamma_0 \gamma_0}} \mathcal{A}_{\infty}^{c} + \frac{2^{1/2} \sigma C_p^{1/2}}{\sqrt{\gamma_0 \gamma_0}} \gamma_i^{1/2} / \mu N_1 \gamma_i^{1/2}. \tag{39}
\]
Similarly, second term of (37), can be bounded by
\[
\frac{C_p}{N_\gamma^2 \mu} \delta_1^2 \leq \frac{C_p}{C_p} \left( \exp \left( -\frac{\mu C_p \delta_1^{1-\alpha}}{2^{\alpha-\beta}} \right) \right) \sqrt{\pi t} + \frac{2^{\alpha-\beta} \sigma \sqrt{C_p}}{\sqrt{\mu C_p}} \leq \frac{C_p^{2-\alpha-\beta} \sqrt{\pi \alpha} \delta_1^{1-\alpha}}{C_p \mu^2 N_\gamma^{1-\alpha/2}},
\]
using \( \exp(-\mu C_p \delta_1^{1-\alpha}/2^{\alpha-\beta}) = A_\delta \leq \sum_{i=1}^t A_\delta \leq t^{-1} A_\delta \), as \( A_\delta \) is decreasing. In a same way, one has
\[
\frac{C_p C_\gamma^2}{N_\mu} \left( \sum_{i=1}^t \delta_i \right)^2 \leq \frac{C_p C_\gamma^2}{N_\mu} \left( \Lambda_\delta \sigma_{\infty} + \frac{2^{\alpha-\beta} C_p \mu C_p}{(1-\alpha) \mu C_p} \right)^{\frac{1}{2}} \leq \frac{C_p C_\gamma^2}{N_\mu} \sqrt{\pi t} \Lambda_\delta \sigma_{\infty} + \frac{2^{\alpha-\beta} C_p \mu \sqrt{C_p}}{C_p} \leq \frac{C_p}{C_p^{3/2} \mu^{3/2} N_\gamma^{1/2}}.
\]
Bound the last term of (37), is done as follows,
\[
\frac{C_p C_\gamma^2}{N_\mu} \sum_{i=1}^t \delta_i \leq \frac{C_p C_\gamma^2}{N_\mu} \sum_{i=1}^t \exp \left( -\frac{\mu C_p \delta_1^{1-\alpha}}{2^{\alpha-\beta}} \right) \sqrt{\pi t} + \frac{2^{\alpha+\beta} C_p \sigma \sqrt{C_p}}{C_p \mu^2 N_\gamma^{1-\alpha/2}} \sum_{i=1}^t \delta_i^{1-\alpha/2} \leq \frac{C_p C_\gamma^2}{N_\mu} \sqrt{\pi t} \Lambda_\delta \sigma_{\infty} + \frac{2^{\alpha+\beta} C_p \sigma \sqrt{C_p}}{C_p \mu^2 N_\gamma^{1-\alpha/2}} + \frac{C_p C_\gamma^2}{N_\mu} \frac{C_\mu}{N_\gamma^2} \left( \frac{6 + 7 \|\psi\|_2 \mu^{3/2} C_p \alpha \beta \Gamma}{\mu^{3/2} \Lambda_\delta} \right) \left( \frac{N_\gamma}{C_p} \right),
\]
Thus, by following the terms above, we obtain
\[
\delta_1 \leq \frac{A_\delta^{1/2}}{N_\gamma^{1/2}} + \frac{1 + \frac{6 \sigma_\infty^{1-\alpha/2}}{\sqrt{C_p} \mu^{3/2} N_\gamma^{1-\alpha/2}} + \frac{2^{\alpha+\beta} \sigma \sqrt{C_p} \Lambda_\delta \sigma_{\infty} + 2^{\alpha+\beta} C_p \sigma \sqrt{C_p}}{C_p \mu^2 N_\gamma^{1-\alpha/2}} + \frac{2^{\alpha-\beta} C_p \mu \sqrt{C_p}}{C_p} + \frac{C_p C_\gamma^2}{N_\mu} \frac{C_\mu}{N_\gamma^2} \left( \frac{6 + 7 \|\psi\|_2 \mu^{3/2} C_p \alpha \beta \Gamma}{\mu^{3/2} \Lambda_\delta} \right) \left( \frac{N_\gamma}{C_p} \right) + \frac{C_p C_\gamma^2}{N_\mu} \frac{C_\mu}{N_\gamma^2} \Lambda_\delta^{1/2}.
\]
Proof of Corollary 4. The steps of the proof follows the ones of Corollary 3 with the same notation of \( \phi \) and \( \tilde{\phi} \). The bound for \( \delta_1 \) in (12) is given by
\[
\delta_1 \leq \frac{A_\delta^{1/2}}{N_\gamma^{1/2}} + \frac{1 + \frac{6 \sigma_\infty^{1-\alpha/2}}{\sqrt{C_p} \mu^{3/2} N_\gamma^{1-\alpha/2}} + \frac{2^{\alpha+\beta} \sigma \sqrt{C_p} \Lambda_\delta \sigma_{\infty} + 2^{\alpha+\beta} C_p \sigma \sqrt{C_p}}{C_p \mu^2 N_\gamma^{1-\alpha/2}} + \frac{2^{\alpha-\beta} C_p \mu \sqrt{C_p}}{C_p} + \frac{C_p C_\gamma^2}{N_\mu} \frac{C_\mu}{N_\gamma^2} \left( \frac{6 + 7 \|\psi\|_2 \mu^{3/2} C_p \alpha \beta \Gamma}{\mu^{3/2} \Lambda_\delta} \right) \left( \frac{N_\gamma}{C_p} \right) + \frac{C_p C_\gamma^2}{N_\mu} \frac{C_\mu}{N_\gamma^2} \Lambda_\delta^{1/2},
\]
where the learning rate is on the form \( \gamma_1 = C_p \rho^{\beta} \gamma_1^{1-\alpha} \) with \( n_t = C_p \rho^{\beta} \). The second-order moment \( \delta_1 \) is upper bounded by (21) from Corollary 2. The fourth-order moment \( \Lambda_\delta \) from Lemma 1 can be simplified as follows,
\[
\Lambda_\delta \leq \exp \left( -\mu \sum_{i=1}^t \gamma_i \right) \Pi_\infty^\prime + \frac{32 \sigma^4}{\mu^2} \max_{t \geq 2} \frac{\gamma_2^2}{n_t^{1/2}} + \frac{16 \sigma^4}{\mu} \max_{t \geq 2} \frac{\gamma_3}{n_t^{1/2}},
\]
as \( n_t \geq 1 \) for any \( t \geq 1 \) and \( \beta \leq 1 \), and
\[
\Pi_\infty^\prime = \exp \left( \frac{32 (\alpha - \beta \rho) C_p C_\rho^{2 \beta} (2 C_1^2 + C_2^2)}{2 (\alpha - \rho \beta)} - 1 \right) \exp \left( 192 C_p^2 C_\rho^{2 \beta} (4 C_1^2 + C_2^2) \right) \left( \Lambda_0 + \frac{2 \sigma^4}{C_1^2} + \frac{4 \sigma^4 C_p}{C_1^2} \right)
\]
using that \( \sum_{i=1}^t \gamma_i \leq 2 \) for \( \alpha \geq 2 \). Next, for \( \rho \geq 0 \), we have
\[
\Lambda_\delta \leq \exp \left( -\frac{\mu C_p C_\rho^{2 \beta} (1 + \beta \rho \alpha - \alpha)}{2^{1+\beta \rho \alpha}} \right) \Pi_\infty + \frac{32 \sigma^4}{\mu} \max_{t \geq 2} \frac{\gamma_2^2}{n_t^{1/2}} + \frac{16 \sigma^4}{\mu} \max_{t \geq 2} \frac{\gamma_3}{n_t^{1/2}},
\]

19
using that $\alpha - \beta \rho \in (1/2, 1)$. If $\rho < 0$, one directly has

$$\Delta_t \leq \exp \left( -\frac{\mu C_r C_\rho^{1-\beta} t^{1-\alpha}}{2^{1-\alpha}} \right) \Pi_\infty^t + \frac{2^{20} 32 \alpha^4 C_r^2 C_\rho^{2\beta}}{\mu^2 t^{2\alpha}} + \frac{2^{3\alpha} 16 \alpha^2 C_r^2 C_\rho^{3\beta}}{\mu t^{3\alpha}}.$$  

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

$$\Delta_t \leq \exp \left( -\frac{\mu C_r C_\rho^{\tilde{\rho}(1-\theta)(1-\phi(1+\tilde{\rho}))}}{2^{1-\theta}(1-\phi(1+\tilde{\rho}))} \right) \Pi_\infty^t + \frac{2^{2\beta(1+\tilde{\rho})} 32 \alpha^4 C_r^2 C_\rho^{2\beta}}{\mu^2 t^{2\alpha}} + \frac{2^{3\beta(1+\tilde{\rho})-\beta} 16 \alpha^2 C_r^2 C_\rho^{3\beta}}{\mu t^{3\alpha}}.$$  

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

Bounding $\frac{1}{N_t \rho} \sum_{i=1}^{t-1} \delta_i^{1/2} |n_{i+1}/y_{i+1} - n_i/y_i|$, we first note $n_i/y_i = C_r^{-1} C_\rho^{1-\beta} t^{1-\beta \rho \alpha} + \rho \alpha$ for $\rho \geq 0$. Thus, by the mean value theorem, we obtain:

$$\left| \frac{n_{i+1}}{y_{i+1}} - \frac{n_i}{y_i} \right| = \left| (1 - \beta \rho + \alpha) \frac{C_\rho^{1-\beta}}{C_r} \sup_{v \in (i,i+1)} |v^{1-\beta \rho \alpha} - 1| \right| \leq \frac{((1 - \beta \rho + \alpha) C_\rho^{1-\beta}}{C_r} \sup_{v \in (i,i+1)} |v^{1-\beta \rho \alpha} - 1| \leq \frac{C_\rho^{1-\beta}}{C_r} \sup_{v \in (i,i+1)} |v^{\beta \rho \alpha} - 1| \leq C_\rho^{1-\beta}.$$  

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

$$\left| \frac{n_{i+1}}{y_{i+1}} - \frac{n_i}{y_i} \right| = \left| \phi(1 + \tilde{\rho}) C_r^{1-\beta} \frac{C_\rho^{(1-\phi(1+\tilde{\rho}))}}{C_r^{1-\phi(1+\tilde{\rho})}} \right| \leq \frac{C_\rho^{1-\beta}}{C_r} \sup_{v \in (i,i+1)} |v^{\beta \rho \alpha} - 1| \leq C_\rho^{1-\beta}.$$  

as $(n_i)_{i \geq 1}$ is a decreasing sequence and $\beta \leq 1$. Thus, for any $\rho \in (1/2, 1)$, we have

$$\left| \frac{n_{i+1}}{y_{i+1}} - \frac{n_i}{y_i} \right| = \phi(1 + \tilde{\rho}) C_r^{1-\beta} \frac{C_\rho^{(1-\phi(1+\tilde{\rho}))}}{C_r^{1-\phi(1+\tilde{\rho})}}.$$  

By using this, we obtain:

$$\frac{1}{N_t \mu} \sum_{i=1}^{t-1} \delta_i^{1/2} \left| \frac{n_{i+1}}{y_{i+1}} - \frac{n_i}{y_i} \right| \leq \phi(1 + \tilde{\rho}) C_r^{1-\beta} \frac{C_\rho^{(1-\phi(1+\tilde{\rho}))}}{C_r^{1-\phi(1+\tilde{\rho})}} \sum_{i=1}^{t-1} \delta_i^{(1-\phi(1+\tilde{\rho}))} \left( \exp \left( -\frac{\mu C_r C_\rho^{\tilde{\rho}(1-\phi)(1+\tilde{\rho})}}{2^{1-\phi(1+\tilde{\rho})}} \right) \right)^{\sqrt{\Pi_\infty^t}} + \frac{2^{2\alpha(1+\alpha)} + \sqrt{\mu C_r C_\rho^{\tilde{\rho}(1-\phi)(1+\tilde{\rho})}}}{\mu^{3/2} C_r^{1/2}}.$$  

Next, let us denote

$$A_{\infty} = \sum_{i=0}^{\infty} \rho^{(i+1)-1} \exp \left( -\frac{\mu C_r C_\rho^{\tilde{\rho}(1-\phi(1+\tilde{\rho}))}}{2^{1-\phi(1+\tilde{\rho})}} \right) \geq \sum_{i=0}^{\infty} \rho^{(i+1)-1} \exp \left( -\frac{\mu C_r C_\rho^{\tilde{\rho}(1-\phi(1+\tilde{\rho}))}}{2^{1-\phi(1+\tilde{\rho})}} \right).$$  

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

$$\phi(1 + \tilde{\rho}) C_r^{1-\beta} \sqrt{\Pi_\infty^t} \sum_{i=1}^{t-1} \delta_i^{(1-\phi(1+\tilde{\rho}))} \exp \left( -\frac{\mu C_r C_\rho^{\tilde{\rho}(1-\phi(1+\tilde{\rho}))}}{2^{1-\phi(1+\tilde{\rho})}} \right) \leq \phi(1 + \tilde{\rho}) C_r^{1-\beta} \sqrt{\Pi_\infty^t} A_{\infty} \frac{\phi(1 + \tilde{\rho}) C_r^{1-\beta} \sqrt{\Pi_\infty^t}}{N_t \mu C_r}.$$  

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

$$\phi(1 + \tilde{\rho}) \frac{1}{\mu^{3/2} (C_r)^{1/2}} \sum_{i=1}^{t-1} \delta_i^{1/2} \leq 2 \frac{2^{2\alpha(1+\alpha)} + \sqrt{\mu C_r C_\rho^{\tilde{\rho}(1-\phi(1+\tilde{\rho)))}}}{\mu^{3/2} C_r^{1/2}}.$$  

and

$$\phi(1 + \tilde{\rho}) \frac{1}{\mu^{3/2} (C_r)^{1/2}} \sum_{i=1}^{t-1} \delta_i^{1/2} \leq 2 \frac{2^{2\alpha(1+\alpha)} + \sqrt{\mu C_r C_\rho^{\tilde{\rho}(1-\phi(1+\tilde{\rho)))}}}{\mu^{3/2} C_r^{1/2}}.$$  

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

$$
\frac{1}{N\mu} \sum_{j=1}^{\gamma_0} \sum_{i=1}^{n_j} \frac{n_j}{y_j \gamma_j} \leq \frac{\phi(1+\tilde{\rho}) C_\gamma^{1-\beta} \sqrt{\tau/\gamma_0} A_{\gamma_0}^{\beta}}{N_\mu C_\gamma} + \frac{2^{1+\alpha}(1+\beta) \sigma C_\rho^{1+\beta} \sqrt{\tau/\gamma_0} A_{\gamma_0}^{\beta}}{\mu^{3/2} \sqrt{\gamma_0 N_\gamma^{1-\phi/2}}} \leq \frac{2 C_\gamma^{1-\beta} \sqrt{\tau/\gamma_0} A_{\gamma_0}^{\beta}}{\mu C_\gamma N_\gamma} + \frac{2^{1+\alpha}(1+\beta) \sigma C_\rho^{1+\beta} \sqrt{\tau/\gamma_0} A_{\gamma_0}^{\beta}}{\mu^{3/2} \sqrt{\gamma_0 N_\gamma^{1-\phi/2}}}.
$$

Similarly, for $\frac{m_{y_0}}{N\mu \gamma_0} \delta_{0,1/2}$, one have

$$
n_1 \frac{1}{N\mu \gamma_0} \leq \frac{C_\rho^{1-\beta} \sqrt{\tau/\gamma_0} \phi(1+\tilde{\rho})}{N_\mu C_\gamma} \exp \left( -\frac{\mu C_\rho^{1\phi/\gamma_0} \phi(1+\tilde{\rho})}{2^{1+\phi}(1+\tilde{\rho})} \right) \frac{1}{\sqrt{\gamma_0 N_\gamma^{1-\phi/2}}} \leq \frac{C_\rho^{1-\beta} \sqrt{\tau/\gamma_0} A_{\gamma_0}^{\beta}}{\mu C_\gamma N_\gamma^{2-\phi}} + \frac{2^{1+\alpha}(1+\beta) \sigma C_\rho^{1+\beta} \sqrt{\tau/\gamma_0} A_{\gamma_0}^{\beta}}{\mu^{3/2} \sqrt{\gamma_0 N_\gamma^{1-\phi/2}}}.
$$

For $\frac{m_{y_0}}{N\mu}(y_1^{-1} + C_1)\delta_{0,1/2}$, we insert the definition of our learning functions, giving us

$$
n_1 \frac{1}{N\mu} \left( \frac{1}{\gamma_j} + C_1 \right) \delta_{0,1/2} = \frac{C_\rho}{N\mu} \left( \frac{1}{C_\gamma C_\rho^{1-\beta}} + C_1 \right) \delta_{0,1/2}.
$$

Bounding $\frac{C_\rho}{N\mu}(\sum_{i=1}^{n_j} \delta_i^{1/2})$, follows the ideas from above, using that $n_{i+1} \leq 2^\beta n_i$, to obtain

$$
\frac{C_\rho}{N\mu} \left( \sum_{i=1}^{n_j} \delta_i^{1/2} \right)^{1/2} \leq \frac{2^{\beta/2} C_\rho}{N\mu} \sum_{i=1}^{n_j} \delta_i \left( \exp \left( -\frac{\mu C_\rho^{1\phi/\gamma_0} \phi(1+\tilde{\rho})}{2^{1+\phi}(1+\tilde{\rho})} \right) \frac{1}{\sqrt{\gamma_0 N_\gamma^{1-\phi/2}}} \right)^{1/2} \leq \frac{2^{\beta/2} C_\rho}{N\mu} \left( C_\rho^{1\phi/\gamma_0} A_{\gamma_0}^{\beta} \right)^{1/2} \frac{1}{\sqrt{\gamma_0 N_\gamma^{1-\phi/2}}} \frac{2^{1+\phi}(1+\tilde{\rho}) \sigma C_\rho^{1+\beta} \sqrt{\tau/\gamma_0} A_{\gamma_0}^{\beta}}{\mu^{3/2} \sqrt{\gamma_0 N_\gamma^{1-\phi/2}}}.
$$

Likewise, for $\frac{C_\rho}{N\mu} \sum_{i=1}^{n_j} n_j A_{\gamma_0}^{1/2}$, we get

$$
\frac{C_\rho}{N\mu} \sum_{i=1}^{n_j} n_j A_{\gamma_0}^{1/2} \leq \frac{2 C_\rho C_\gamma^{1-\beta} \sqrt{\tau/\gamma_0} A_{\gamma_0}^{\beta}}{\mu N_\gamma} \sum_{i=1}^{n_j} \delta_i \left( \exp \left( -\frac{\mu C_\rho^{1\phi/\gamma_0} \phi(1+\tilde{\rho})}{2^{1+\phi}(1+\tilde{\rho})} \right) \frac{1}{\sqrt{\gamma_0 N_\gamma^{1-\phi/2}}} \right)^{1/2} \leq \frac{2 C_\rho C_\gamma^{1-\beta} \sqrt{\tau/\gamma_0} A_{\gamma_0}^{\beta}}{\mu N_\gamma} \left( C_\rho^{1\phi/\gamma_0} A_{\gamma_0}^{\beta} \right)^{1/2} \frac{1}{\sqrt{\gamma_0 N_\gamma^{1-\phi/2}}} \frac{2 \sigma C_\rho^{1+\beta} \sqrt{\tau/\gamma_0} A_{\gamma_0}^{\beta}}{\mu^{3/2} \sqrt{\gamma_0 N_\gamma^{1-\phi/2}}}.
$$

where the second term can be bounded as

$$
\frac{2^{1+\phi}(1+\tilde{\rho}) \sigma C_\rho^{1+\beta} \sqrt{\tau/\gamma_0} A_{\gamma_0}^{\beta}}{\mu^{3/2} \sqrt{\gamma_0 N_\gamma^{1-\phi/2}}} \leq \frac{2^{1+\phi}(1+\tilde{\rho}) \sigma C_\rho^{1+\beta} \sqrt{\tau/\gamma_0} A_{\gamma_0}^{\beta}}{\mu^{3/2} \sqrt{\gamma_0 N_\gamma^{1-\phi/2}}} \leq \frac{2^{1+\phi}(1+\tilde{\rho}) \sigma C_\rho^{1+\beta} \sqrt{\tau/\gamma_0} A_{\gamma_0}^{\beta}}{\mu^{3/2} \sqrt{\gamma_0 N_\gamma^{1-\phi/2}}}.
$$
and the third term by
\[
\frac{2^{3(1+\rho)(1+\rho)/2}C_\rho \sigma^2 C_{\gamma'}^{1/2} C_{\rho'}^{1+\rho/2}}{\mu^{1/2} C_{\rho'}^{1+\rho/2} N_i} \sum_{i=1}^{t-1} \beta(\theta_{i-1}) (1/2) \leq \frac{2^{3(1+\rho)(1+\rho)/2}C_\rho \sigma^2 C_{\gamma'}^{1/2} C_{\rho'}^{1+\rho/2} \psi_3(\sigma-\beta\rho)/2(N_i/C_{\rho'})}{\mu^{1/2} C_{\rho'}^{1+\rho/2} N_i}.
\]

By collecting these bounds, we get
\[
\frac{C_\rho}{N_i \mu} \sum_{i=0}^{t-1} n_i \Delta_t^{1/2} \leq \frac{2^{2(1+\rho)(1+\rho)/2}C_\rho \sigma^2 C_{\gamma'}^{2/2} C_{\rho'}^{1/2} \psi_3(\sigma-\beta\rho)/2(N_i/C_{\rho'})}{\mu^{1/2} C_{\rho'}^{1+\rho/2} N_i}.
\]

Combining our findings from above, we have
\[
\frac{\delta_l^{1/2}}{N_i^{1/2}} \leq \frac{2^{2(1+\rho)/2} C_{\rho} \sigma C_{\rho'}^{1/2} \psi_3(\sigma-\beta\rho)}{\mu C_{\rho} N_i^{1/2}} + \frac{2^{2(1+\rho)/2} C_{\rho} \sigma C_{\rho'}^{1/2} \psi_3(\sigma-\beta\rho)}{\mu C_{\rho} N_i^{1/2}} + \frac{2^{2(1+\rho)/2} C_{\rho} \sigma C_{\rho'}^{1/2} \psi_3(\sigma-\beta\rho)}{\mu C_{\rho} N_i^{1/2}} + \frac{2^{2(1+\rho)/2} C_{\rho} \sigma C_{\rho'}^{1/2} \psi_3(\sigma-\beta\rho)}{\mu C_{\rho} N_i^{1/2}}.
\]

This can be simplified to the desired using \(\Gamma_l\) given by \((1/C_{\rho} C_{\rho'}^{1/2} + C_i) t^{1/2} + 2^{2(1+\rho)/2} C_{\rho} \sigma C_{\rho'}^{1/2} \psi_3(\sigma-\beta\rho)/2 N_i/C_{\rho'} + 2^{2(1+\rho)/2} C_{\rho} \sigma C_{\rho'}^{1/2} \psi_3(\sigma-\beta\rho)/2 N_i/C_{\rho'} + 2^{2(1+\rho)/2} C_{\rho} \sigma C_{\rho'}^{1/2} \psi_3(\sigma-\beta\rho)/2 N_i/C_{\rho'}\), consisting of the finite constants \(c_{\rho}^{1/2}, \Gamma_l,\) and \(A_{\rho'}^{1/2}\).

7.2.2. **Proofs for Section 4.2**

**Theorem 3 (APSSG).** Denote \(\delta_i = \mathbb{E}[||\theta_i - \theta''||^2]\) with \((\theta_i)\) given by \((4)\) using \((\theta_i)\) from \((3)\). Under Assumptions 1, Assumptions 2-p and 3-p with \(p = 4\), Assumptions 4 and 5, we have for any learning rate \((\gamma_l)\) that
\[
\frac{\delta_l^{1/2}}{N_i^{1/2}} \leq \frac{2^{2(1+\rho)/2} C_{\rho} \sigma C_{\rho'}^{1/2} \psi_3(\sigma-\beta\rho)}{\mu C_{\rho} N_i^{1/2}} + \frac{2^{2(1+\rho)/2} C_{\rho} \sigma C_{\rho'}^{1/2} \psi_3(\sigma-\beta\rho)}{\mu C_{\rho} N_i^{1/2}} + \frac{2^{2(1+\rho)/2} C_{\rho} \sigma C_{\rho'}^{1/2} \psi_3(\sigma-\beta\rho)}{\mu C_{\rho} N_i^{1/2}} + \frac{2^{2(1+\rho)/2} C_{\rho} \sigma C_{\rho'}^{1/2} \psi_3(\sigma-\beta\rho)}{\mu C_{\rho} N_i^{1/2}}.
\]

where \(\Lambda = \text{Tr}(\Sigma_l^2 L(\theta')^{-1} \Sigma_l^2 L(\theta')^{-1})\) and \(C_{\rho'} = C_{\rho}' + 2^{2} C_{\rho}/D_{\rho}^2\).

**Proof of Theorem 3.** Denote \(\mathbb{E}[||\theta_i - \theta''||^2]\) by \(\delta_i\) with \((\theta_i)\) given by \((4)\) using \((\theta_i)\) from \((3)\). As in the proof Theorem 2, we follow the steps of Polyak and Juditsky [26], in which, we can rewrite \((3)\) to
\[
\frac{1}{\gamma_l} (\theta_{l+1} - \theta_l) = \nabla_{\theta_l} (\theta_{l+1}) - \frac{1}{\gamma_l} \Omega_l,
\]
where \(\nabla_{\theta_l} (\theta_{l+1}) = n_t^{-1} \Sigma_{n_t} \nabla_{\theta_l} (\theta_{l+1})\) and \(\Omega_l = \mathcal{P}_{\theta_l} (\theta_{l+1} - \gamma_l \nabla_{\theta_l} (\theta_{l+1})) - (\theta_{l+1} - \gamma_l \nabla_{\theta_l} (\theta_{l+1})).\) Thus, summing the parts, using the Minkowski's inequality, and bounding each term gives us the same bound as in Theorem 2, but with an additional term \(\Theta_l\), namely
\[
\left( \mathbb{E} \left[ \left[ \nabla_{\theta_l}^2 L(\theta')^{-1} \frac{1}{N_i} \sum_{i=0}^{t} n_i \Omega_i \right]^2 \right] \right)^{1/2} \leq \frac{1}{\mu N_i} \sum_{i=1}^{t} n_i \sqrt{\mathbb{E} \left[ \left[ \Omega_i \right]^2 \right]} = \frac{1}{\mu N_i} \sum_{i=1}^{t} n_i \sqrt{\mathbb{E} \left[ \left[ \Omega_i \right]^2 \right]} \mathbb{E} \left[ [\theta_{l+1} - \gamma_l \nabla_{\theta_l} (\theta_{l+1})] \right],
\]
using Godichon-Baggioni [9, Lemma 4.3]. Next, we note that \(\mathbb{E}[||\Omega_i||^2] [\theta_{l+1} - \gamma_l \nabla_{\theta_l} (\theta_{l+1})] \in \Theta_l\), since
\[
||\Omega_i||^2 \leq 2 ||\mathcal{P}_{\theta_l} (\theta_{l+1} - \gamma_l \nabla_{\theta_l} (\theta_{l+1}) - \theta_{l+1}) ||^2 + 2 \gamma_l^2 ||\nabla_{\theta_l} (\theta_{l+1}) ||^2 = 2 ||\mathcal{P}_{\theta_l} (\theta_{l+1} - \gamma_l \nabla_{\theta_l} (\theta_{l+1}) - \theta_{l+1}) ||^2 + 2 \gamma_l^2 ||\nabla_{\theta_l} (\theta_{l+1}) ||^2 \leq 2 ||\theta_{l+1} - \gamma_l \nabla_{\theta_l} (\theta_{l+1}) ||^2 + 2 \gamma_l^2 ||\nabla_{\theta_l} (\theta_{l+1}) ||^2 = 4 \gamma_l^2 ||\nabla_{\theta_l} (\theta_{l+1}) ||^2 \leq 4 \gamma_l^2 G_{\theta}^2.
\]
as $P_\Theta$ is Lipschitz and $\|\nabla P_{\theta_i}(\theta)\|^2 \leq G_\theta^2$ for any $\theta \in \Theta$. Moreover, as in Godichon-Baggioni and Portier [10, Theorem 4.2], we know that $P[\theta_{t-1} - \gamma \nabla d_i(\theta_{t-1}) \notin \Theta] \leq \Delta_i/D_\theta$, where $D_\theta = \inf_{\|\theta - \theta\|} \|\theta - \theta\|_\theta$ with $\partial \Theta$ denoting the frontier of $\Theta$. Thus, (42) can then be bounded by

$$\frac{1}{\mu N_i} \sum_{i=1}^t n_i \frac{1}{\gamma} \sqrt{E[\|\Omega\|^2 \Pi_{i=1}^{n_i} \Pi_{i=1}^{N_i}] \leq \frac{2G_\Theta}{\mu D_\theta^2 N_i} \sum_{i=1}^t n_i \Delta_i^{1/2} \leq \frac{2^2G_\Theta}{\mu D_\theta^2 N_i} \sum_{i=1}^t n_i \Delta_i^{1/2},$$

using that the sequence $(n_i)$ is either constant or time-varying, meaning $n_{i+1}/n_i \leq 2$. \hfill \Box

**Proof of Corollary 5.** The proof follows directly from Corollary 3 but with use of Theorem 3. \hfill \Box

**Proof of Corollary 6.** The proof follows directly from Corollary 4 but with use of Theorem 3. \hfill \Box

### A. Technical Proofs

Appendix A contains purely technical results used in the proofs presented in Section 7. In what follows, we use the convention $\inf \Theta = 0$, $\sum_{i=1}^0 = 0$, and $\prod_{i=1}^0 = 1$.

**Proposition 1.** Let $(\gamma_i)_{i \geq 1}$ be a positive sequence. For any $k \leq t$, and $\omega > 0$, we have

$$\sum_{i=k}^t \prod_{j=i+1}^t [1 + \omega \gamma_j] \gamma_i \leq \frac{1}{\omega} \sum_{i=k}^t \prod_{j=i+1}^t [1 + \omega \gamma_j] \leq \frac{1}{\omega} \exp \left( \omega \sum_{i=k}^t \gamma_i \right). \quad (A.1)$$

**Proof of Proposition 1.** We begin with considering the first inequality in (A.1), which follows by expanding the sum of product:

$$\sum_{i=k}^t \prod_{j=i+1}^t [1 + \omega \gamma_j] \gamma_i = \frac{1}{\omega} \sum_{i=k}^t \prod_{j=i+1}^t [1 + \omega \gamma_j] \omega \gamma_i = \frac{1}{\omega} \sum_{i=k}^t \prod_{j=i+1}^t [1 + \omega \gamma_j] [1 + \omega \gamma_i - 1]
\begin{align*}
= \frac{1}{\omega} \sum_{i=k}^t \left[ \prod_{j=i+1}^t [1 + \omega \gamma_j] [1 + \omega \gamma_i] - \prod_{j=i+1}^t [1 + \omega \gamma_j] \right]
= \frac{1}{\omega} \sum_{i=k}^t \prod_{j=i+1}^t [1 + \omega \gamma_j] - \prod_{j=k+1}^t [1 + \omega \gamma_j].
\end{align*}$$

As the (positive) terms cancel out, we end up with the first inequality in (A.1):

$$\frac{1}{\omega} \sum_{i=k}^t \prod_{j=i+1}^t [1 + \omega \gamma_j] - \frac{1}{\omega} \prod_{j=k+1}^t [1 + \omega \gamma_j] \leq \frac{1}{\omega} \prod_{j=k}^t [1 + \omega \gamma_j] - \prod_{j=k+1}^t [1 + \omega \gamma_j]
\begin{align*}
= \frac{1}{\omega} \prod_{j=k}^t [1 + \omega \gamma_j] - \prod_{j=k+1}^t [1 + \omega \gamma_j]
= \frac{1}{\omega} \prod_{j=k}^t [1 + \omega \gamma_j] - 1 \leq \frac{1}{\omega} \prod_{j=k}^t [1 + \omega \gamma_j].
\end{align*}$$

as $\prod_{i=1}^t = 1$ for all $t \in \mathbb{N}$. Using the (simple) bound of $1 + t \leq \exp(t)$ for all $t \in \mathbb{R}$, we obtain the second inequality of (A.1):

$$\frac{1}{\omega} \prod_{j=k}^t [1 + \omega \gamma_j] \leq \frac{1}{\omega} \prod_{j=k}^t \exp(\omega \gamma_j) \leq \frac{1}{\omega} \exp \left( \omega \sum_{j=k}^t \gamma_j \right).$$

\hfill \Box

**Proposition 2.** Let $(\gamma_i)_{i \geq 1}$ be a positive sequence. Let $\omega > 0$ and $k \leq t$ such that for all $i \geq k$, $\omega \gamma_i \leq 1$, then

$$\sum_{i=k}^t \prod_{j=i+1}^t [1 - \omega \gamma_j] \gamma_i \leq \frac{1}{\omega}. \quad (A.2)$$
Proof of Proposition 2. We start with expanding the sums of products term in (A.2), given us

\[
\frac{1}{\omega} \sum_{j=1}^{t} \prod_{k=1}^{j-1} [1 - \omega y_j] \gamma_i = \frac{1}{\omega} \sum_{j=1}^{t} \prod_{k=1}^{j-1} [1 - \omega y_j] [1 - \omega y_j] - 1 = \frac{1}{\omega} \sum_{j=1}^{t} \prod_{k=j+1}^{i} [1 - \omega y_j] [1 - \omega y_j] - 1 = \frac{1}{\omega} \sum_{j=1}^{t} \prod_{k=j+1}^{i} [1 - \omega y_j] - \prod_{j=1}^{i} [1 - \omega y_j].
\]

As we only have positive terms, we can upper bound the term:

\[
\frac{1}{\omega} \sum_{j=1}^{t} \prod_{k=j+1}^{i} [1 - \omega y_j] - \prod_{j=1}^{i} [1 - \omega y_j] \leq \frac{1}{\omega} \left[ 1 - \prod_{j=1}^{i} [1 - \omega y_j] \right] \leq \frac{1}{\omega},
\]

using \( \prod_{j=1}^{i} [1 - \omega y_j] \geq 0 \), showing the inequality in (A.2).

Proof of Proposition 3. Let \((\gamma_i)_{i \geq 1}\) and \((\eta_i)_{i \geq 1}\) be positive sequences. For any \(k \leq t\), we can obtain the (upper) bounds:

\[
\sum_{j=1}^{t} \prod_{k=1}^{j} [1 + \omega y_j] \eta_j y_i \leq \frac{1}{\omega} \max_{k \leq j} \eta_j \exp\left( \omega \sum_{j=1}^{t} y_j \right),
\]

with \(\omega > 0\). Furthermore, suppose that for all \(i \geq k\), \(\omega y_i \leq 1\), then

\[
\sum_{j=1}^{t} \prod_{k=1}^{j} [1 - \omega y_j] \eta_j y_i \leq \frac{1}{\omega} \max_{k \leq j} \eta_j.
\]

Proof of Proposition 4. Let \((\delta_i)_{i \geq 1}\), \((\gamma_i)_{i \geq 1}\), \((\eta_i)_{i \geq 1}\), and \((\nu_i)_{i \geq 1}\) be some positive sequences satisfying the recursive relation:

\[
\delta_i = (1 - 2\omega y_i + \eta_i \gamma_i) \delta_{i-1} + \nu_i \gamma_i,
\]

with \(\delta_0 \geq 0\) and \(\omega > 0\). Denote \(t_0 = \inf \{t \geq 1 : \eta_i \leq \omega\}\), and suppose that for all \(t \geq t_0 + 1\), one has \(\omega y_i \leq 1\). Then, for \(\gamma_i\) and \(\eta_i\) decreasing, we have the upper bound on \(\delta_i\):

\[
\delta_i \leq \exp\left( -\omega \sum_{i=1}^{t} \gamma_i \right) \left[ \exp\left( \sum_{i=1}^{t_0} \eta_i \gamma_i \right) \left( \delta_0 + \max_{1 \leq i \leq t_0} \eta_i \right) \right] + \frac{1}{\omega} \max_{1 \leq i \leq t_0} \gamma_i.
\]

for all \(t \in \mathbb{N}\) with the convention that \(\sum_{i=0}^{t_0} = 0\) if \(t/2 < t_0\).

Proof of Proposition 4. Applying the recursive relation from (A.5) \(t\) times, we derive:

\[
\delta_t \leq \prod_{i=1}^{t} [1 - 2\omega y_i + \eta_i \gamma_i] \delta_0 + \sum_{i=1}^{t} \prod_{j=i+1}^{t} [1 - 2\omega y_j + \eta_j \gamma_j] \nu_i \gamma_i.
\]
where $B_t$ can be seen as a transient term only depending on the initialisation $\delta_0$, and a stationary term $A_t$. The transient term $B_t$ can be divided into two products, before and after $t_0$,

$$B_t = \prod_{i=1}^{t_0} \left[1 - 2\omega \gamma_i + \eta \gamma_i \right] \left( \prod_{i=t_0+1}^{t} \left[1 - 2\omega \gamma_i + \eta \gamma_i \right] \right).$$

Using that $t_0 = \inf \{ t \geq 1 : \gamma_i \leq \omega \}$, and since for all $t \geq t_0 + 1$, we have $2\omega \gamma_i - \eta \gamma_i \geq \omega \gamma_i$, it comes

$$B_t \leq \left( \prod_{i=1}^{t_0} \left[1 - 2\omega \gamma_i + \eta \gamma_i \right] \right) \left( \prod_{i=t_0+1}^{t} \left[1 - \omega \gamma_i \right] \right) \leq \left( \prod_{i=1}^{t_0} \exp(-2\omega \gamma_i + \eta \gamma_i) \right) \left( \prod_{i=t_0+1}^{t} \exp(-\omega \gamma_i) \right)$$

$$= \exp(-2\omega \sum_{i=1}^{t_0} \gamma_i) \exp \left( \sum_{i=t_0+1}^{t} \eta \gamma_i \right) \leq \exp(-\omega \sum_{i=1}^{t_0} \gamma_i) \exp \left( \sum_{i=t_0+1}^{t} \eta \gamma_i \right)$$

by applying the (simple) bound $1 + t \leq \exp(t)$ for all $t \in \mathbb{R}$. We derive that

$$B_t \leq \exp(-\omega \sum_{i=1}^{t_0} \gamma_i) \exp \left( \sum_{i=t_0+1}^{t} \eta \gamma_i \right).$$

Next, the stationary term $A_t$ can (similarly) be divided into two sums (after and before $t_0$):

$$A_t = \sum_{i=t_0+1}^{t} \left( \prod_{j=i+1}^{t} \left[1 - 2\omega \gamma_j + \eta \gamma_j \right] \right) v_i \gamma_i + \sum_{i=t_0+1}^{t} \left( \prod_{j=i+1}^{t} \left[1 - 2\omega \gamma_j + \eta \gamma_j \right] \right) v_i \gamma_i .$$

The first stationary term $A_{t,1}$ (with $t > t_0$) can be bounded as follows: if $t/2 < t_0 + 1$, we have

$$A_{t,1} \leq \max_{2 \leq s \leq t} v_s \sum_{i=t_0+1}^{t} \prod_{j=i+1}^{t} \left[1 - \omega \gamma_j \right] \gamma_i = \max_{2 \leq s \leq t} v_s \prod_{i=t_0+1}^{t} \left[1 - \omega \gamma_i \right] \gamma_i,$$

by Proposition 3. Furthermore, if $t/2 > t_0 + 1$, we get

$$A_{t,1} \leq \prod_{i=t_0+1}^{t} \left[1 - 2\omega \gamma_i + \eta \gamma_i \right] v_i \gamma_i + \sum_{i=t/2}^{t} \prod_{j=i+1}^{t} \left[1 - \omega \gamma_j \right] \gamma_i$$

$$\leq \prod_{i=t_0+1}^{t} \left[1 - 2\omega \gamma_i + \eta \gamma_i \right] v_i \gamma_i + \prod_{j=t/2}^{t} \left[1 - \omega \gamma_j \right] \gamma_i = \prod_{j=t/2}^{t} \left[1 - \omega \gamma_j \right] \sum_{i=t_0+1}^{t} v_i \gamma_i + \frac{1}{\omega} \max v_s,$$

where $\prod_{j=t/2}^{t} \left[1 - \omega \gamma_j \right] \leq \exp(-\omega \sum_{i=t/2}^{t} \gamma_i)$ as $1 + t \leq \exp(t)$ for all $t \in \mathbb{R}$. Thus, for all $t \in \mathbb{R}$,

$$A_{t,1} \leq \exp(-\omega \sum_{j=t/2}^{t} \gamma_j) \sum_{i=t_0+1}^{t} v_i \gamma_i + \frac{1}{\omega} \max v_s,$$

(8)

where $\sum_{i=t_0}^{t} = 0$ if $t/2 < t_0$. The second stationary term $A_{t,2}$ can be bounded, thanks to Proposition 1, as follows:

$$A_{t,2} = \sum_{i=t_0+1}^{t} \left( \prod_{j=i+1}^{t} \left[1 - 2\omega \gamma_j + \eta \gamma_j \right] \right) v_i \gamma_i \leq \left( \prod_{i=t_0+1}^{t} \left[1 - 2\omega \gamma_i + \eta \gamma_i \right] \right) \max_{i \leq t/2} \frac{1}{\eta} \sum_{i=t_0+1}^{t} \left[1 + \eta \gamma_i \right] \gamma_i$$

$$\leq \left( \prod_{i=t_0+1}^{t} \left[1 - \omega \gamma_i \right] \right) \sum_{i=t_0+1}^{t} \left[1 + \eta \gamma_i \right] \gamma_i \leq \exp(-\omega \sum_{i=t_0+1}^{t} \gamma_i) \max_{i \leq t/2} \frac{1}{\eta} \sum_{i=t_0+1}^{t} \left[1 + \eta \gamma_i \right] \gamma_i$$

$$\leq \exp(-\omega \sum_{j=t_0+1}^{t} \gamma_j) \sum_{i=t_0+1}^{t} \left[1 + \eta \gamma_i \right] \gamma_i \leq \exp(-\omega \sum_{j=t_0+1}^{t} \gamma_j) \max_{i \leq t/2} \frac{1}{\eta} \sum_{i=t_0+1}^{t} \left[2 \sum_{j=t_0+1}^{t} \gamma_j \right].$$
by the definition of $t_0$, thus

$$A_{i,2} \leq \exp\left(-\omega \sum_{j=1}^{t_1} \gamma_j \right) \max_{1 \leq i \leq N_h} \nu_i \exp\left(2 \sum_{j=1}^{t_0} \eta_i \gamma_j \right) \leq \exp\left(-\omega \sum_{j=1}^{t_1/2} \gamma_j \right) \max_{1 \leq i \leq N_h} \nu_i \exp\left(2 \sum_{j=1}^{t_0} \eta_i \gamma_j \right). \quad (A.9)$$

Then, using the bound for $A_{i,1}$ in (A.8) and $A_{i,2}$ in (A.9), we can bound $A_i$ by

$$A_i \leq \exp\left(-\omega \sum_{j=1}^{t_1/2} \gamma_j \right) \exp\left(2 \sum_{j=1}^{t_0} \eta_i \gamma_j \right) \max_{1 \leq i \leq N_h} \nu_i \left(\delta_0 + \max_{1 \leq i \leq N_h} \eta_i \right) + \frac{1}{\omega} \max_{1 \leq i \leq N_h} \nu_i. \quad (A.10)$$

Finally, combining the bound for $B_i$ in (A.7) and $A_i$ in (A.10), we achieve the bound for $\delta_i \leq B_i \delta_0 + A_i$, namely the upper bound in (A.6). \qed

The following proposition is a more simplistic but rougher version of the bound in Proposition 4.

**Proposition 5.** Let $(\delta_i)_{i \geq 0}$, $(\gamma_i)_{i \geq 1}$, $(\eta_i)_{i \geq 1}$, and $(\nu_i)_{i \geq 1}$ be some positive sequences satisfying the recursive relation in (A.5). Denote $t_0 = \inf \{ t \geq 1 : \eta_i \leq \omega \}$, and suppose that for all $t \geq t_0 + 1$, one has $\omega \gamma_t \leq 1$. Then, for $\gamma_t$ and $\eta_i$ decreasing, we have for all $t \in \mathbb{N}$,

$$\delta_t \leq \exp\left(-\omega \sum_{j=1}^{t_1/2} \gamma_j \right) \exp\left(2 \sum_{i=1}^{t_0} \eta_i \gamma_i \right) \left(\delta_0 + \max_{1 \leq i \leq N_h} \eta_i \gamma_i \right) + \frac{1}{\omega} \max_{1 \leq i \leq N_h} \nu_i. \quad (A.11)$$

**Proof of Proposition 5.** The resulting (upper) bound in (A.11) follows directly from (A.6) by noting that $t_0 \leq t$, giving us $\sum_{i=1}^{t_1/2} \nu_i \gamma_i \leq \sum_{i=1}^{t_0} \nu_i \gamma_i \leq \max_{i \leq \|t\|_\omega} (\nu_i / \eta_i) \sum_{i=1}^{t_0} \eta_i \gamma_i \leq \max_{i \leq \|t\|_\omega} (\nu_i / \eta_i) \exp(2 \sum_{i=1}^{t_0} \eta_i \gamma_i)$, as $(\nu_i)$ and $(\gamma_i)$ are positive sequences. \qed

**Acknowledgement**

This work was supported by grants from Région Île-de-France.

**References**

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

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

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

[4] 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.

[5] Cardot, H., Cénac, P., Zitt, P.A., 2013. Efficient and fast estimation of the geometric median in hilbert spaces with an averaged stochastic gradient algorithm. Bernoulli 19, 18–43.

[6] d’Aspremont, A., 2008. Smooth optimization with approximate gradient. SIAM Journal on Optimization 19, 1171–1183.

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

[8] Gervini, D., 2008. Robust functional estimation using the median and spherical principal components. Biometrika 95, 587–600.

[9] 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.

[10] 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.

[11] Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT press.

[12] Haldane, J., 1948. Note on the median of a multivariate distribution. Biometrika 35, 414–417.

[13] 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.

[14] Kempenaar, J., 1987. The median of a finite measure on a banach space. Statistical data analysis based on the L1-norm and related methods (Neuchâtel, 1987) , 217–230.

[15] Kurdyka, K., 1998. On gradients of functions definable in o-minimal structures, in: Annales de l’institut Fourier, pp. 769–783.
[16] Kushner, H., Yin, G.G., 2003. Stochastic approximation and recursive algorithms and applications. volume 35. Springer Science & Business Media.
[17] Lan, G., 2020. First-order and stochastic optimization methods for machine learning. Springer.
[18] LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521, 436–444.
[19] Lojasiewicz, S., 1963. A topological property of real analytic subsets. Coll. du CNRS, Les équations aux dérivées partielles 117, 2.
[20] 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.
[21] Moulines, E., Bach, F., 2011. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. Advances in neural information processing systems 24.
[22] Murata, N., Amari, S.i., 1999. Statistical analysis of learning dynamics. Signal Processing 74, 3–28.
[23] Necoara, I., Nesterov, Y., Glineur, F., 2019. Linear convergence of first order methods for non-strongly convex optimization. Mathematical Programming 175, 69–107.
[24] Nesterov, Y., et al., 2018. Lectures on convex optimization. volume 137. Springer.
[25] Polyak, B.T., 1963. Gradient methods for minimizing functionals. Zhurnal Vychislitel’noi Matematiki i Matematicheskoi Fiziki 3, 643–653.
[26] Polyak, B.T., Juditsky, A.B., 1992. Acceleration of stochastic approximation by averaging. SIAM journal on control and optimization 30, 838–855.
[27] Robbins, H., Monro, S., 1951. A stochastic approximation method. The annals of mathematical statistics , 400–407.
[28] Ruppert, D., 1988. Efficient estimations from a slowly convergent Robbins-Monro process. Technical Report. Cornell University Operations Research and Industrial Engineering.
[29] Schmidt, M., Roux, N., Bach, F., 2011. Convergence rates of inexact proximal-gradient methods for convex optimization. Advances in Neural Information Processing Systems 24, 1458–1466.
[30] Shalev-Shwartz, S., et al., 2012. Online learning and online convex optimization. Foundations and Trends® in Machine Learning 4, 107–194.
[31] Sridharan, K., Shalev-Shwartz, S., Srebro, N., 2008. Fast rates for regularized objectives. Advances in neural information processing systems 21.
[32] Steinwart, I., Christmann, A., 2011. Estimating conditional quantiles with the help of the pinball loss. Bernoulli 17, 211–225.
[33] Teo, C., Smola, A., Vishwanathan, S., Le, Q.V., 2007. A scalable modular convex solver for regularized risk minimization, in: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 727–736.
[34] Werge, N., Wintenberger, O., 2022. Adavol: An adaptive recursive volatility prediction method. Econometrics and Statistics 23, 19–35.
[35] 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.

27