Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Streaming Data

Antoine Godichon-Baggioni, Nicklas Werge, Olivier Wintenberger

ESAIM: PS 27 (2023) 482–514
https://doi.org/10.1051/ps/2023006

ESAIM: Probability and Statistics
www.esaim-ps.org

NON-ASYMPTOTIC ANALYSIS OF STOCHASTIC APPROXIMATION ALGORITHMS FOR STREAMING DATA*

Antoine Godichon-Baggioni[1], Nicklas Werge[1,**] and Olivier Wintenberger[1,2]

Abstract. We introduce a streaming framework for analyzing stochastic approximation/optimization problems. This streaming framework is analogous to solving optimization problems using time-varying mini-batches that arrive sequentially. We provide non-asymptotic convergence rates of various gradient-based algorithms; this includes the famous Stochastic Gradient (SG) descent (a.k.a. Robbins-Monro algorithm), mini-batch SG and time-varying mini-batch SG algorithms, as well as their iterated averages (a.k.a. Polyak-Ruppert averaging). We show (i) how to accelerate convergence by choosing the learning rate according to the time-varying mini-batches, (ii) that Polyak-Ruppert averaging achieves optimal convergence in terms of attaining the Cramer-Rao lower bound, and (iii) how time-varying mini-batches together with Polyak-Ruppert averaging can provide variance reduction and accelerate convergence simultaneously, which is advantageous for many learning problems, such as online, sequential, and large-scale learning. We further demonstrate these favorable effects for various time-varying mini-batches.

Mathematics Subject Classification. 62L12, 62L20, 68W27, 90C25.

Received March 16, 2022. Accepted March 3, 2023.

1. Introduction

Machine learning-based intelligent systems are becoming more and more widespread in modern society [18, 24]. A crucial component of machine learning is optimization, which, in this context, involves estimating parameters for the intelligent systems to make decisions about future data. A growing challenge is that these future data will arrive in an endless stream, for example through sensors from real-time measurement of weather, traffic and e-commerce, to name a few; we call these *streaming data*. Such streaming data arrives sequentially in time-varying mini-batches. This places wide demands on computational efficiency and the robustness of the underlying optimization algorithms, which must be updated sequentially as more data becomes available.

Stochastic approximation/optimization algorithms have proven effective in handling large amounts of data and perform well across many fields ranging from smooth and strongly convex problems to complex non-convex

* This work was supported by Région Île-de-France project number 19006497.

Keywords and phrases: Stochastic algorithms, stochastic optimization, machine learning, online learning, mini-batch, streaming.

[1] Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation (LPSM), 75005 Paris, France.
[2] Wolfgang Pauli Institut, c/o Fakultät für Mathematik, Universität Wien, 1090 Vienna, Austria.

** Corresponding author: nicklas.werge@sorbonne-universite.fr

ones; [4] reviews such algorithms for large-scale problems in machine learning. Among these, the most well-known is probably the Stochastic Gradient (SG) descent introduced in [33], which forms the basis of many optimization algorithms used in machine learning [22, 23, 30, 36]. In a nutshell, these SG-based algorithms minimize the objective (a.k.a. loss or risk) of a model by iteratively updating the model parameters using stochastic approximations of its gradient. Traditionally, these gradients are processed individually or in (fixed) mini-batches taken from a (fixed) dataset. However, in our streaming framework, these gradients must be computed as a sequential stream of time-varying mini-batches.

1.1. Contributions

The objective of this paper is to solve stochastic approximation/optimization problems in a streaming framework. Our main theoretical contribution is the non-asymptotic analysis of SG-based algorithms in this streaming framework, extending the work of [1]. This means that we investigate everything from the classical SG descent to time-varying mini-batch SG-based algorithms, as well as their Polyak-Ruppert extensions. Our results show how to accelerate convergence by choosing the learning rate according to the time-varying mini-batches. In addition, we show that Polyak-Ruppert averaging [32, 34] achieves optimal convergence in terms of achieving the Cramer-Rao lower bound in this streaming framework. In particular, we show how time-varying mini-batches together with Polyak-Ruppert averaging can provide variance reduction and accelerate convergence simultaneously, without jeopardizing the computational complexity. These theoretical findings are demonstrated for various streaming settings of time-varying mini-batches.

1.2. Organization

Section 2 presents the streaming framework, in which we will analyze the stochastic algorithms. In Section 3, we introduce our stochastic streaming-gradient algorithms, their projected versions and Polyak-Ruppert extensions. The main results, namely the non-asymptotic convergence analysis, are presented in Section 4. These results are illustrated in Section 5 for various time-varying mini-batches. In Section 6, we provide some concluding remarks with related future perspectives.

2. Problem formulation

Our objective is to solve stochastic approximation/optimization problems in a streaming framework, where data arrives sequentially in time-varying mini-batches; we consider problems on the form

$$\min_{\theta\in\mathbb{R}^d}\{F(\theta) := \mathbb{E}[f(\theta)]\}. \tag{2.1}$$

We will refer to $F : \mathbb{R}^d \to \mathbb{R}$ as the objective function, but in the literature, F is also known as the expected loss (and risk); See for instance [4]. Let θ^* denote the global minimum of F, and assume that $\theta^* \in \Theta$, where Θ is a closed convex set in $\mathbb{R}^d$. Typical convergence results measure how quickly some estimate θ_t approaches θ^* (or the function value $L(\theta_t)$ approaches $F(\theta^*)$). In this paper we are interested in bounding the quantity $\mathbb{E}[\|\theta_t - \theta^*\|^2]$. As in [1, 16], we make the analysis more convenient through convexity and smoothness assumptions on F in (2.1).[1] The following assumptions are frequently referenced.

Assumption 2.1 (μ-quasi-strong convex [19, 28])**.** The objective function $F : \mathbb{R}^d \to \mathbb{R}$ is differentiable with $\nabla_\theta F(\theta^*) = 0$ and there exists a constant $\mu > 0$ such that $\forall\theta \in \Theta$,

$$F(\theta^*) \geq F(\theta) + \langle \nabla_\theta F(\theta), \theta^* - \theta\rangle + \frac{\mu}{2}\|\theta^* - \theta\|^2. \tag{2.2}$$

[1] Milder degrees of convexity have been investigated; [13] studied SG algorithms under local strongly convexity, [19] studied SG algorithms under the Polyak-Łojasiewicz condition [25, 31], and [10] studied the Ruppert-Polyak averaging estimate under some Kurdyka-Łojasiewicz-type condition [21, 25].

Assumption 2.2 (C_∇-Lipschitz smoothness)**.** The function $\nabla_\theta F$ is C_∇-Lipschitz continuous around θ^*, *i.e.*, there exists $C_\nabla > 0$ such that $\forall \theta \in \Theta$,

$$\|\nabla_\theta F(\theta) - \nabla_\theta F(\theta^*)\| \leq C_\nabla \|\theta - \theta^*\|. \tag{2.3}$$

2.1. Streaming framework and notation

Let each $(f_t(\theta))$ constitute a sequence of independent differentiable random functions (possibly non-convex) and let their gradients be unbiased estimates of $\nabla_\theta F(\theta)$, see *e.g.* [30] for definitions and properties of such functions. The shorthand notation of

$$(f_t(\theta)) \text{ represents the sequence of time-varying mini-batches parameterized by } \theta.$$

We say that each f_t consist of $n_t \in \mathbb{N}$ data points, which we denote by the set $\{f_{t,1}, \ldots, f_{t,n_t}\}$. For example, for a class of models $\{h_\theta\}_{\theta\in\Theta}$ parameterized by θ, a loss function l, and a regularizer Ω, then $f_{t,i}(\theta)$ can be seen as the composition:

$$f_{t,i}(\theta) = l(y_{t,i}, h_\theta(x_{t,i})) + \Omega(\theta). \tag{2.4}$$

where $\{(x_{t,i}, y_{t,i})\}_{i=1}^{n_t}$ is a time-varying mini-batch of i.i.d. input-output data points with generic element $(x, y) \in \mathcal{X} \times \mathcal{Y}$. The associated objective function from (2.1) thus corresponds to having $F(\theta) = \mathbb{E}[f(\theta)]$ with $f(\theta) = l(y, h_\theta(x)) + \Omega(\theta)$.

Our streaming framework includes many machine learning problems, from classification, and regression to ranking; this includes stochastic approximation (Robbins-Monro setting [33]), learning from i.i.d. data with linear, logistic, softmax, quantile and general ridge regression, and p-means and geometric median under regularity conditions [8, 18, 23, 30, 37, 38]. More specifically, (2.4) could be the l_2-regularized least squares regression model, $f(\theta) = (\langle x, \theta\rangle - y) + \frac{\lambda}{2}\|\theta\|^2$ with $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$, or the l_2-regularized logistic regression for binary classification, $f(\theta) = \log(1 + \exp(-y\langle x, \theta\rangle)) + \frac{\lambda}{2}\|\theta\|^2$ with $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{-1, 1\}$; here, we used $h_\theta(x) = \langle x, \theta\rangle$.

3. Stochastic streaming gradient algorithms

SG-based algorithms, which dates back to the seminal work of [33], have become the predominant optimization algorithm for solving these stochastic approximation/optimization problems. To solve problem (2.1) in our streaming framework, we introduce the Stochastic Streaming Gradient (SSG) algorithm, defined as the recursion

(SSG)
$$\theta_t = \theta_{t-1} - \frac{\gamma_t}{n_t}\sum_{i=1}^{n_t} \nabla_\theta f_{t,i}(\theta_{t-1}), \quad \theta_0 \in \mathbb{R}^d, \tag{3.1}$$

where (γ_t) is the learning rate satisfying $\sum_{i=1}^t \gamma_i = \infty$ and $\sum_{i=1}^t \gamma_i^2 < \infty$ for $t \to \infty$. This SSG algorithm sequentially processes the time-varying mini-batches. Note that if for all $t \geq 1$, $n_t = 1$, then the SSG algorithm is an online version of the well-known SG descent.

In many machine learning models, there may be restrictions on the parameter space of θ. We embrace this by defining a projected version of SSG, given as

(PSSG)
$$\theta_t = \mathcal{P}_\Theta\left(\theta_{t-1} - \frac{\gamma_t}{n_t}\sum_{i=1}^{n_t} \nabla_\theta f_{t,i}\left(\theta_{t-1}\right)\right), \quad \theta_0 \in \Theta, \tag{3.2}$$

where $\mathcal{P}_\Theta$ denotes the Euclidean projection onto the closed convex set Θ in $\mathbb{R}^d$, *i.e.*, $\mathcal{P}_\Theta(\theta) = \arg\min_{\theta'\in\Theta}\|\theta - \theta'\|_2$. It is worth noting that SG-based algorithms are not gradient descent in the sense that the objective function

values often increase, but only decrease on average; examples of this are illustrated in Section 5. Therefore, it makes intuitive sense to use sets of stochastic gradient $\{\nabla_\theta f_{t,i}\}_{i=1}^{n_t}$ in each iteration, as it naturally reduces the variance and makes it easier to adjust the learning rate (γ_t), which (on average) improves the convergence.

Next, let's consider a streaming variant of the celebrated Polyak-Ruppert averaging procedure [32, 34]:

(ASSG)/(APSSG)
$$\bar{\theta}_t = \frac{1}{N_t}\sum_{i=0}^{t-1} n_{i+1}\theta_i, \tag{3.3}$$

where $N_t = \sum_{i=1}^{t} n_i$ denotes the accumulated number of data points processed at each $t \in \mathbb{N}$. This averaging procedure sequentially aggregates the estimates of (3.1) and (3.2), which stabilizes and accelerates convergence [29, 32]. In particular, this average allows us to obtain the optimal Cramer-Rao lower bound. Note that (3.3) does not actually change the estimates produced by the SSG or PSSG algorithms, but instead simply keeps track of a running average over the estimates. Practically, as we handle data sequentially, we will make use of the recursive formula, $\bar{\theta}_t = (N_{t-1}/N_t)\bar{\theta}_{t-1} + (n_t/N_t)\theta_{t-1}$. A detailed overview of our stochastic streaming gradient algorithms (defined in (3.1) to (3.3)) is presented in Algorithm 1.

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

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

4. Non-asymptotic convergence analysis

Throughout this paper, we consider learning rates (γ_t) of the form

$$\gamma_t := C_\gamma n_t^\beta t^{-\alpha},$$

with $C_\gamma > 0$, $\beta \in [0, 1]$, and α chosen according to the time-varying mini-batches n_t. This learning rate allows us to add more weight to larger mini-batches (n_t) through the β parameter. Note that [1] considered learning rates of the same form, but with $\beta = 0$ (and $n_t = 1$). For simplicity, we let the time-varying mini-batches (n_t) be given as

$$n_t := \lceil C_\rho t^\rho \rceil,$$

with $C_\rho \in \mathbb{N}$ and $\rho \in (-1, 1)$ such that $n_t \geq 1$ for all $t \in \mathbb{N}$. This setting includes classical (online) SG descent algorithms (*i.e.*, $\{C_\rho = 1, \rho = 0\}$) and (online) mini-batch procedures of both constant and time-varying size (*i.e.* $\{C_\rho \in \mathbb{N}, \rho = 0\}$ and $\{C_\rho \in \mathbb{N}, \rho \in (-1, 1)\}$), as well as the Polyak-Ruppert average of (online) time-varying mini-batches. We will refer to C_ρ as the *mini-batch size* and ρ as the *mini-batch rate*.

Our goal is to non-asymptotic bound the quantities $\delta_t := \mathbb{E}[\|\theta_t - \theta^*\|^2]$ and $\bar{\delta}_t := \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$, such that they solely depend on the parameters of the problem. To our knowledge, this is the first work that studies the non-asymptotic convergence behavior of SG-based algorithms and their Polyak-Ruppert averaging in a streaming framework. To do this, we assume for each $t \in \mathbb{N}$ the following about the (stochastic) gradients of $\{f_{t,i}\}_{i=1}^{n_t}$.

Assumption 4.1 (unbiased gradients)**.** For each $\theta \in \Theta$, the random variable $\nabla_\theta f_{t,i}(\theta)$ is square-integrable and $\forall \theta \in \Theta$, $\nabla_\theta F(\theta) = \mathbb{E}[\nabla_\theta f_{t,i}(\theta)]$.

Assumption 4.2-p (C_f-expected smoothness). For a positive integer p, there exists $C_f > 0$ such that $\forall \theta \in \Theta$, $\mathbb{E}[\|\nabla_\theta f_{t,i}(\theta) - \nabla_\theta f_{t,i}(\theta^*)\|^p] \leq C_f^p \mathbb{E}[\|\theta - \theta^*\|^p]$.

Assumption 4.3-p (σ-gradient noise). For a positive integer p, there exists $\sigma > 0$ such that $\mathbb{E}[\|\nabla_\theta f_{t,i}(\theta^*)\|^p] \leq \sigma^p$.

Discussion of Assumptions 4.1 to 4.3-p. These assumptions are standard for analyzing stochastic approximation/optimization problems with SG algorithms, *e.g.*, see [3, 22]. Assumption 4.1 concerns the access to unbiased stochastic approximations of the gradient $\nabla_\theta F$, which are common when SG algorithms are used to solve problem (2.1).[2] Another common assumption for SG algorithms is that they are uniformly bounded. But such an assumption is often too restrictive, as it can only hold for some loss functions [4, 14]. Instead, we make the weaker expected smoothness assumption of the gradients of $\{f_{t,i}\}_{i=1}^{n_t}$ in Assumption 4.2-p [1, 16]. The last key assumption concerns the finiteness of the gradient noise $\{f_{t,i}\}_{i=1}^{n_t}$ at θ^* (Asm. 4.3-p). It is worth noting that Assumptions 4.2-p and 4.3-p can be verified explicitly, *e.g.*, see [16]. For SSG and PSSG, Assumptions 4.2-p and 4.3-p only needs to hold for $p = 2$, where for ASSG and APSSG, we need $p = 4$ to bound the fourth order moment.

4.1. Stochastic streaming gradients

In this section, we analysis the SSG and PSSG algorithms from (3.1) and (3.2). To do this, we first derive an explicit upper bound on the tth estimate of (3.1) and (3.2) for any learning rate (γ_t) and time-varying mini-batch (n_t) using classical techniques from stochastic approximations [3, 22].

Theorem 4.4 (SSG/PSSG)**.** *Let $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ for $\delta_0 \geq 0$, where (θ_t) either follows the recursion in (3.1) or (3.2). Suppose Assumptions 2.1 to 4.3-p hold with $p = 2$. Then, for any learning rate (γ_t) and time-varying mini-batch (n_t), we have*

$$\delta_t \leq \exp\left(-\mu \sum_{i=t/2}^{t} \gamma_i\right) \pi_t^\delta + \frac{2\sigma^2}{\mu} \max_{t/2 \leq i \leq t} \frac{\gamma_i}{n_i}, \tag{4.1}$$

with $\pi_t^\delta = \exp(4C_f^2 \sum_{i=1}^t \gamma_i^2/n_i) \exp(2C_\nabla^2 \sum_{i=1}^t \mathbb{1}_{\{n_i>1\}}\gamma_i^2)(\delta_0 + 2\sigma^2/C_f^2)$.

Sketch of proof. Under Assumptions 2.1 to 4.3-p with $p = 2$, we show that (δ_t) (derived using (3.1)) satisfies the recursive relation

$$\delta_t \leq [1 - 2\mu\gamma_t + (2C_f^2 + (n_t - 1)C_\nabla^2)n_t^{-1}\gamma_t^2]\delta_{t-1} + 2\sigma^2 n_t^{-1}\gamma_t^2, \tag{4.2}$$

for any (γ_t) and (n_t) fulfilling the conditions imposed on the learning rate [33]. This recursive relation is then explicitly upper bounded in a non-asymptotic manner using Proposition B.5 in Appendix B. Bounding the projected estimate in (3.2) follows directly from the fact that $\mathbb{E}[\|\mathcal{P}_\Theta(\theta) - \theta^*\|^2] \leq \mathbb{E}[\|\theta - \theta^*\|^2]$, $\forall \theta \in \Theta$ [40]. Alternatively, the projected estimate can also be shown without Assumptions 4.2-p and 4.3-p, but instead with a bounded gradient assumption, *e.g.*, see [1].

[2] The principles for biased gradients are rather different, *e.g.*, see [9, 35].

Related work. When $n_t = 1$ in (4.1), we obtain (an online version of) the usual SG descent studied in [1]. As mentioned in [1], Theorem 4.4 forms an upper bound on the function values, $\mathbb{E}[F(\theta_t) - F(\theta^*)] \leq C_f \delta_t/2$; this follows from the Cauchy-Schwarz inequality and Assumption 4.2-p.

Decay of the initial conditions. The learning rate (γ_t) should satisfy the conditions $\sum_{i=1}^t \gamma_i = \infty$ and $\sum_{i=1}^t \gamma_i^2 < \infty$ as $t \to \infty$ of [33]. These conditions directly imply that $\pi_t^\delta < \infty$. Thus, our attention is on reducing the *noise term* $\max_{t/2 \leq i \leq t} \gamma_i/n_i$ without damaging the natural decay of the *sub-exponential term* $\exp(-\mu \sum_{i=t/2}^t \gamma_i)$. In particular, the non-asymptotic bound in (4.1) holds for any learning rate fulfilling these conditions. In addition, the scaling of n_t in the noise term shows an obvious possibility of variance reduction.

Before considering time-varying mini-batches, we consider the constant case where n_t follows the constant $C_\rho \in \mathbb{N}$, *i.e.*, an online (projected) mini-batch SG variant.

Corollary 4.5 (SSG/PSSG with constant mini-batches)**.** *Let $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ for $\delta_0 \geq 0$, where (θ_t) either follows the recursion in (3.1) or (3.2). Suppose Assumptions 2.1 to 4.3-p hold with $p = 2$. Then, if $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ with $n_t = C_\rho$, for $\alpha \in (1/2, 1)$, we have*

$$\delta_t \leq \exp\left(-\frac{\mu C_\gamma N_t^{1-\alpha}}{2^{1-\alpha} C_\rho^{1-\alpha-\beta}}\right) \pi_\infty^c + \frac{2^{1+\alpha}\sigma^2 C_\gamma}{\mu C_\rho^{1-\alpha-\beta} N_t^\alpha}, \tag{4.3}$$

where $\pi_\infty^c = \exp(4\alpha C_\gamma^2 (2C_f^2 + C_\rho \mathbb{1}_{\{C_\rho > 1\}} C_\nabla^2)/(2\alpha - 1) C_\rho^{1-2\beta})(\delta_0 + 2\sigma^2/C_f^2)$ is a finite constant.

Decay of the initial conditions. The bound in Corollary 4.5 depends on the initial condition $\delta_0 = \|\theta_0 - \theta^*\|^2$ and the variance σ^2 in the noise term. The initial condition δ_0 vanish sub-exponentially fast for $\alpha \in (1/2, 1)$; the condition of having $\alpha \in (1/2, 1)$ is a natural restriction from [33]. Thus, the asymptotic term is $2^{1+\alpha}\sigma^2 C_\gamma / \mu C_\rho^{1-\alpha-\beta} N_t^\alpha$, *i.e.*, $\delta_t = \mathcal{O}(N_t^{-\alpha})$. Moreover, the bound in (4.3) is optimal (up to some constants) for quadratic functions $(f_{t,i})$, since the deterministic recursion in (4.2) would be with equality. It is worth noting that if $C_\gamma C_f$ or $C_\gamma C_\nabla$ is chosen too large, they may produce a large π_∞^c constant. In addition, π_∞^c is positively affected by C_ρ when $\beta < 1/2$. Obviously, the hyper-parameter β only comes into play if the mini-batch size C_ρ is larger than one, *i.e.*, $C_\rho > 1$. Nonetheless, the effect of π_∞^c will decrease exponentially fast due to the sub-exponentially decaying factor in front.

Variance reduction from larger mini-batches. Not surprisingly, larger mini-batches C_ρ cause a variance reducing effect, *e.g.*, see the illustrations in Section 5. Nevertheless, (4.3) explicitly shows the variance reducing effect in each term, which can help us better understand how to optimally tune the learning rate. In particular, the asymptotic term is divided by $C_\rho^{1-\alpha-\beta}$, implying we should take $\alpha + \beta \leq 1$ when C_ρ is large. However, this must be done with moderation as larger mini-batches C_ρ simultaneously damage the sub-exponential term. Another important point from this is that mini-batches do not provide a better convergence rate, but simply scale, *i.e.*, the slope of the rate of convergence is unchanged, but the intercept is lowered (*e.g.*, see Fig. 1a).

Having fixed size mini-batches is not the most realistic streaming framework, these mini-batches are much more likely to vary in size over time. For the convenience of notation, let $\tilde{\rho} = \rho \mathbb{1}_{\{\rho \geq 0\}}$.

Corollary 4.6 (SSG/PSSG with time-varying mini-batches)**.** *Let $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ for $\delta_0 \geq 0$, where (θ_t) either follows the recursion in (3.1) or (3.2). Suppose Assumptions 2.1 to 4.3-p hold with $p = 2$. Then, if $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ with $n_t = \lceil C_\rho t^\rho \rceil$, for $\alpha - \beta\tilde{\rho} \in (1/2, 1)$, we have*

$$\delta_t \leq \exp\left(-\frac{\mu C_\gamma N_t^{1-\phi}}{2^{(2+\rho)(1-\phi)} C_\rho^{1-\beta-\phi}}\right) \pi_\infty^v + \frac{2^{1+(2+\rho)\phi}\sigma^2 C_\gamma}{\mu C_\rho^{(1-\beta)\mathbb{1}_{\{\rho \geq 0\}} - \phi} N_t^\phi}, \tag{4.4}$$

where $\phi = ((1-\beta)\tilde{\rho} + \alpha)/(1 + \tilde{\rho})$ and $\pi_\infty^v = \exp(4(\alpha - \beta\tilde{\rho}) C_\gamma^2 C_\rho^{2\beta} (2C_f^2 + C_\nabla^2)/(2(\alpha - \beta\tilde{\rho}) - 1))(\delta_0 + 2\sigma^2/C_f^2)$ is a finite constant.

Accelerated decay with increasing mini-batches. As mentioned for Corollary 4.5, $\alpha - \beta\tilde{\rho} \in (1/2, 1)$ is a natural condition from [33]; this relaxes the condition of having $\alpha \in (1/2, 1)$ for $\rho \geq 0$. In particular, this shows that we can accelerate convergence by taking increasing mini-batches, *e.g.*, taking $\alpha = 2/3$ and $\beta = 0$ yields $\delta_t = \mathcal{O}(N_t^{-(2/3+\rho)/(1+\rho)})$ when $\rho > 0$. Conversely, when $\rho < 0$, we obtain the same decay as in Corollary 4.5, namely, $\delta_t = \mathcal{O}(N_t^{-\alpha})$. These effects are illustrated in Figures 1b to 1e for $C_\rho \in \{1, 8, 64, 128\}$.

Variance reduction from larger mini-batches. Similarly to Corollary 4.5, the sub-exponential and asymptotic term is scaled by $C_\rho^{1-\beta-\phi}$ for $\rho \geq 0$, implying we should take $\alpha + \beta \leq 1$ to obtain variance reduction. However, as discussed above, this variance reduction is only beneficial in the beginning and does not contribute to a better convergence rate (relative to the slope). Thus, large mini-batch sizes C_ρ and negative mini-batch rates ρ will give (an initial) variance reduction but the same convergence rate as in Corollary 4.5.

4.2. Polyak-Ruppert averaging

In what follows, we consider the Polyak-Ruppert averaging estimate $(\bar{\theta}_n)$ given in (3.3), where (θ_t) follows the recursion in (3.1) or (3.2). Besides having Assumptions 4.2-p and 4.3-p to hold for $p = 4$, additional assumptions are needed for bounding the Polyak-Ruppert averaging estimate. First, we make an additional smoothness assumption on the objective function F.

Assumption 4.7 (C'_∇-Lipschitz continuous Hessian operator)**.** The function F is twice differentiable with C'_∇-Lipschitz continuous Hessian operator $\nabla^2_\theta F$, meaning, there exists $C'_\nabla \geq 0$ such that $\forall \theta \in \Theta$,

$$\|\nabla^2_\theta F(\theta) - \nabla^2_\theta F(\theta^*)\| \leq C'_\nabla \|\theta - \theta^*\|. \tag{4.5}$$

Next, in continuation of Assumption 4.3-p, we make the following assumption about covariance of $(\nabla_\theta f_{t,i}(\theta^*))$, which we interpret as the sequence of score vectors with respect to the parameter vector θ^*.

Assumption 4.8 (Covariance of the scores)**.** There exists a non-negative self-adjoint operator Σ such that $\mathbb{E}[\nabla_\theta f_{t,i}(\theta^*) \nabla_\theta f_{t,i}(\theta^*)^\top] \preceq \Sigma$.

Note that the operator Σ always exists when σ is finite for order $p = 4$ in Assumption 4.3-p.

4.2.1. Polyak-Ruppert averaging of stochastic streaming gradients (ASSG)

As in Section 4.1, we conduct a general study for any learning rate (γ_t) and time-varying mini-batch (n_t) when applying the Polyak-Ruppert averaging estimate $(\bar{\theta}_n)$ from (3.3), where (θ_t) follows the recursion in (3.1), *i.e.*, the ASSG.

Theorem 4.9 (ASSG)**.** *Let $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ with $(\bar{\theta}_t)$ given by (3.3), where (θ_t) follows the recursion in (3.1). Suppose Assumptions 2.1 to 4.8 hold with $p = 4$. Then, for any learning rate (γ_t) and time-varying mini-batch (n_t), we can upper bound $\bar{\delta}_t^{1/2}$ by*

$$\frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{1}{\mu N_t} \sum_{i=1}^{t-1} \left| \frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i} \right| \delta_i^{1/2} + \frac{n_t}{\mu \gamma_t N_t} \delta_t^{1/2} + \frac{n_1}{\mu N_t} \left(\frac{1}{\gamma_1} + C_f \right) \delta_0^{1/2} + \frac{C_f}{\mu N_t} \left(\sum_{i=1}^{t-1} n_{i+1} \delta_i \right)^{1/2} + \frac{C'_\nabla}{\mu N_t} \sum_{i=0}^{t-1} n_{i+1} \Delta_i^{1/2}, \tag{4.6}$$

where $\Lambda = \mathrm{Tr}(\nabla^2_\theta F(\theta^*)^{-1} \Sigma \nabla^2_\theta F(\theta^*)^{-1})$ and $\Delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^4]$ for some $\Delta_0 \geq 0$.

As noticed in [32], the leading term Λ/N_t achieves the Cramer-Rao lower bound [10, 27]. Note that the leading term Λ/N_t is invariant of the learning rate (γ_t) and the time-varying mini-batches (n_t). Moreover, the bound is $\mathcal{O}(N_t^{-1})$ without inverting the Hessian. Next, the processes (δ_t) and (Δ_t) can be bounded by the recursive relations in (4.1) and (A.9). There are no sub-exponential decaying terms for the initial conditions in Theorem 4.9, 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 $\psi_x^y(t) : \mathbb{R} \to \mathbb{R}$, given as

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

with $y \in \mathbb{R}_+$, such that $\sum_{i=1}^t i^{-x} \leq \psi_x^0(t)$ for any $x \in \mathbb{R}_+$. Note that $\psi_x^y(t)/t = \mathcal{O}(t^{-(x+y)/(1+y)})$ if $x < 1$, $\psi_x^y(t)/t = \mathcal{O}(\log(t)t^{-1})$ if $x = 1$, and $\psi_x^y(t)/t = \mathcal{O}(t^{-1})$ if $x > 1$. Hence, for any $x, y \in \mathbb{R}_+$, $\psi_x^y(t)/t = \tilde{\mathcal{O}}(t^{-(x+y)/(1+y)})$, where the $\tilde{\mathcal{O}}(\cdot)$ notation hides logarithmic factors.

Corollary 4.10 (ASSG with constant mini-batches)**.** *Let $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ with $(\bar{\theta}_t)$ given by (3.3), where (θ_t) follows the recursion in (3.1). Suppose Assumptions 2.1 to 4.8 hold with $p = 4$. Then, if $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ with $n_t = C_\rho$, for $\alpha \in (1/2, 1)$, we have*

$$\bar{\delta}_t^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{6\sigma C_\rho^{(1-\alpha-\beta)/2}}{\mu^{3/2} C_\gamma^{1/2} N_t^{1-\alpha/2}} + \frac{2^\alpha 6 C'_\nabla \sigma^2 C_\gamma}{\mu^2 C_\rho^{1-\alpha-\beta} N_t^\alpha} + \frac{2 C_f \sigma C_\gamma^{1/2}}{\mu^{3/2} C_\rho^{(1-\alpha-\beta)/2} N_t^{(1+\alpha)/2}} + \frac{C_\rho \Gamma_c}{\mu N_t}$$
$$+ \frac{C_\rho^{2-\alpha-\beta}\sqrt{\pi_\infty^c} A_\infty^c}{\mu C_\gamma N_t^{2-\alpha}} + \frac{(6 + 7\mathbb{1}_{\{C_\rho > 1\}}) 2^{3\alpha/2} C'_\nabla \sigma^2 C_\gamma^{3/2} C_\rho^{3\beta/2} \psi_{3\alpha/2}^0(N_t/C_\rho)}{\mu^{3/2} N_t},$$

with Γ_c given by $(1/C_\gamma C_\rho^\beta + C_f)\delta_0^{1/2} + C_f\sqrt{\pi_\infty^c A_\infty^c / C_\rho} + \sqrt{\pi_\infty^c} A_\infty^c / C_\gamma C_\rho^\beta + C'_\nabla \sqrt{\Pi_\infty^c} A_\infty^c$, consisting of the finite constants π_∞^c, Π_∞^c and A_∞^c, that only depends on μ, δ_0, Δ_0, C_f, σ, C_∇, C'_∇, C_γ, C_ρ, β and α.

Accelerated decay the initial conditions. By averaging, we have increased the rate of convergence from $\mathcal{O}(N_t^{-\alpha})$ to the optimal rate $\mathcal{O}(N_t^{-1})$ (when we compare to SSG with constant mini-batches in Corollary 4.5). The two subsequent terms are the main remaining terms decaying at the rate $\mathcal{O}(N_t^{\alpha-2})$ and $\mathcal{O}(N_t^{-2\alpha})$, which suggest taking $\alpha = 2/3$. The remaining terms are negligible. Next, it is worth noting that having $\alpha + \beta = 1$ in Corollary 4.10, we would give no impact in the main remaining terms from the mini-batch size C_ρ. 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 $C_\gamma C_f$ too large may cause Γ_c to be significant. Nevertheless, the term consisting of Γ_c decay at a rate of at least $\mathcal{O}(N_t^{-2})$.

Corollary 4.11 (ASSG with time-varying mini-batches)**.** *Let $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ with $(\bar{\theta}_t)$ given by (3.3), where (θ_t) follows the recursion in (3.1). Suppose Assumptions 2.1 to 4.8 hold with $p = 4$. Then, if $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ with $n_t = \lceil C_\rho t^\rho \rceil$, for $\alpha - \beta\tilde{\rho} \in (1/2, 1)$, we have*

$$\bar{\delta}_t^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{2^{3+\phi(1+\tilde{\rho})}\sigma C_\rho^{(1-\phi-\beta)/2\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2} C_\gamma^{1/2} N_t^{1-\phi/2}} + \frac{2^{(1+\phi)(1+\tilde{\rho})-2} C'_\nabla \sigma^2 C_\gamma}{\mu^2 C_\rho^{1-\phi-\beta} N_t^\phi} + \frac{2^{\phi(1+\tilde{\rho})/2} C_f \sigma C_\gamma^{1/2}}{\mu^{3/2} C_\rho^{(1-\phi-\beta)/2\mathbb{1}_{\{\rho\geq 0\}}} N_t^{(1+\phi)/2}} + \frac{C_\rho \Gamma_v}{\mu N_t}$$
$$+ \frac{C_\rho^{2-\phi-\beta}\sqrt{\pi_\infty^v} A_\infty^v}{\mu C_\gamma N_t^{2-\phi}} + \frac{2^{3(1+\phi)(1+\tilde{\rho})/2} C'_\nabla \sigma^2 C_\gamma^{3/2} C_\rho^{1+3\beta/2} \psi_{3(\alpha-\beta\tilde{\rho})/2}^{\tilde{\rho}}(N_t/C_\rho)}{\mu^{3/2} C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} N_t},$$

with Γ_v given by $(1/C_\gamma C_\rho^\beta + C_f)\delta_0^{1/2} + 2^{\tilde{\rho}} C_f \sqrt{\pi_\infty^v A_\infty^v / C_\rho} + 2\sqrt{\pi_\infty^v} A_\infty^v / C_\gamma C_\rho^\beta + 2^{\tilde{\rho}} C'_\nabla \sqrt{\Pi_\infty^v} A_\infty^v$, consisting of the finite constants π_∞^v, Π_∞^v and A_∞^v, that only depends on μ, δ_0, Δ_0, C_f, σ, C_∇, C'_∇, C_γ, C_ρ, β and α.

Robustness towards mini-batch rate ρ: Following the arguments above, the two main remainder terms suggest that $\phi = 2/3 \Leftrightarrow \alpha - \beta\tilde{\rho} = (2 - \tilde{\rho})/3$, *e.g.*, by setting $\beta = 0$, we should pick $\alpha = (2 - \tilde{\rho})/3$. Likewise, if $\rho = 0$, we yield the same conclusion as in Corollary 4.10, namely $\alpha = 2/3$. However, these hyper-parameter

choices are not resilient against any time-varying streaming rate ρ. 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 the same convergence for any time-varying mini-batch rate by having $\alpha = 2/3$ and $\beta = 1/3$; this is illustrated in Figures 1f and 2f.

4.2.2. Polyak-Ruppert averaging of projected stochastic streaming gradients (APSSG)

In this section, we analyze the projected Polyak-Ruppert averaging estimate (a.k.a. APSSG), where (θ_t) follows the recursion in (3.2). To avoid calculating the six-order moment, we make the unnecessary assumption that $\|\nabla_\theta f_{t,i}(\theta)\|$ is uniformly bounded on Θ; the derivation of the six-order moment can be found in [12].

Assumption 4.12 (G_Θ-bounded stochastic gradients)**.** Let $D_\theta = \inf_{\theta \in \partial\Theta} \|\theta - \theta^*\| > 0$ with $\partial\Theta$ denoting the frontier of Θ. Assume there exists $G_\Theta > 0$ such that $\forall t \geq 1$, $\sup_{\theta\in\Theta} \|\nabla_\theta f_{t,i}(\theta)\|^2 \leq G_\Theta^2$ a.s., with $i = 1, \ldots, n_t$.

Corollary 4.13 (APSSG with constant mini-batches)**.** *Let $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ with $(\bar{\theta}_t)$ given by (3.3), where (θ_t) follows the recursion in (3.2). Suppose Assumptions 2.1 to 4.12 hold with $p = 4$. Then, if $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ with $n_t = C_\rho$, for $\alpha \in (1/2, 1)$, we have*

$$\bar{\delta}_t^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{6\sigma C_\rho^{(1-\alpha-\beta)/2}}{\mu^{3/2} C_\gamma^{1/2} N_t^{1-\alpha/2}} + \frac{2^\alpha 6 C'_\nabla \sigma^2 C_\gamma}{\mu^2 C_\rho^{1-\alpha-\beta} N_t^\alpha} + \frac{2 C_f \sigma C_\gamma^{1/2}}{\mu^{3/2} C_\rho^{(1-\alpha-\beta)/2} N_t^{(1+\alpha)/2}} + \frac{C_\rho \Gamma_c}{\mu N_t}$$
$$+ \frac{C_\rho^{2-\alpha-\beta} \sqrt{\pi_\infty^c} A_\infty^c}{\mu C_\gamma N_t^{2-\alpha}} + \frac{(6 + 7\mathbb{1}_{\{C_\rho > 1\}}) 2^{3\alpha/2} C''_\nabla \sigma^2 C_\gamma^{3/2} C_\rho^{3\beta/2} \psi^0_{3\alpha/2}(N_t/C_\rho)}{\mu^{3/2} N_t},$$

with $C''_\nabla = C'_\nabla + 2^2 G_\Theta / D_\theta^2$ and Γ_c given by $(1/C_\gamma C_\rho^\beta + C_f)\delta_0^{1/2} + C_f \sqrt{\pi_\infty^c A_\infty^c / C_\rho} + \sqrt{\pi_\infty^c} A_\infty^c / C_\gamma C_\rho^\beta + C'_\nabla \sqrt{\Pi_\infty^c} A_\infty^c$, consisting of the finite constants π_∞^c, Π_∞^c and A_∞^c, that only depends on μ, δ_0, Δ_0, C_f, σ, C_∇, C'_∇, C_γ, C_ρ, β and α.

Corollary 4.14 (APSSG with time-varying mini-batches)**.** *Let $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ with $(\bar{\theta}_t)$ given by (3.3), where (θ_t) follows the recursion in (3.2). Suppose Assumptions 2.1 to 4.12 hold with $p = 4$. Then, if $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ with $n_t = \lceil C_\rho t^\rho \rceil$, for $\alpha - \beta\tilde{\rho} \in (1/2, 1)$, we have*

$$\bar{\delta}_t^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{2^{3+\phi(1+\tilde{\rho})} \sigma C_\rho^{(1-\phi-\beta)/2\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2} C_\gamma^{1/2} N_t^{1-\phi/2}} + \frac{2^{(1+\phi)(1+\tilde{\rho})-2} C'_\nabla \sigma^2 C_\gamma}{\mu^2 C_\rho^{1-\phi-\beta} N_t^\phi} + \frac{2^{\phi(1+\tilde{\rho})/2} C_f \sigma C_\gamma^{1/2}}{\mu^{3/2} C_\rho^{(1-\phi-\beta)/2\mathbb{1}_{\{\rho\geq 0\}}} N_t^{(1+\phi)/2}} + \frac{C_\rho \Gamma_v}{\mu N_t}$$
$$+ \frac{C_\rho^{2-\phi-\beta} \sqrt{\pi_\infty^v} A_\infty^v}{\mu C_\gamma N_t^{2-\phi}} + \frac{2^{3(1+\phi)(1+\tilde{\rho})/2} C''_\nabla \sigma^2 C_\gamma^{3/2} C_\rho^{1+3\beta/2} \psi^{\tilde{\rho}}_{3(\alpha-\beta\tilde{\rho})/2}(N_t/C_\rho)}{\mu^{3/2} C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} N_t},$$

with $C''_\nabla = C'_\nabla + 2^2 G_\Theta / D_\theta^2$ and Γ_v given by $(1/C_\gamma C_\rho^\beta + C_f)\delta_0^{1/2} + 2^{\tilde{\rho}} C_f \sqrt{\pi_\infty^v A_\infty^v / C_\rho} + 2\sqrt{\pi_\infty^v} A_\infty^v / C_\gamma C_\rho^\beta + 2^{\tilde{\rho}} C'_\nabla \sqrt{\Pi_\infty^v} A_\infty^v$, consisting of the finite constants π_∞^v, Π_∞^v and A_∞^v, that only depends on μ, δ_0, Δ_0, C_f, σ, C_∇, C'_∇, C_γ, C_ρ, β and α.

5. Experiments

In this section, we demonstrate the theoretical results presented in Section 4 for various time-varying mini-batches. The performance is measured over one-hundred replications of the quadratic mean error, *i.e.*, $(\mathbb{E}[\|\theta_{N_t} - \theta^*\|^2])_{t\geq 0}$ and $(\mathbb{E}[\|\bar{\theta}_{N_t} - \theta^*\|^2])_{t\geq 0}$. Note that averaging over several replications gives a reduction in variability, which mainly benefits the SSG and PSSG. All metrics are shown in log-scale and normalized such that the first iteration is one, namely, $\mathbb{E}[\|\theta_0 - \theta^*\|^2] = \mathbb{E}[\|\bar{\theta}_0 - \theta^*\|^2] = 1$.

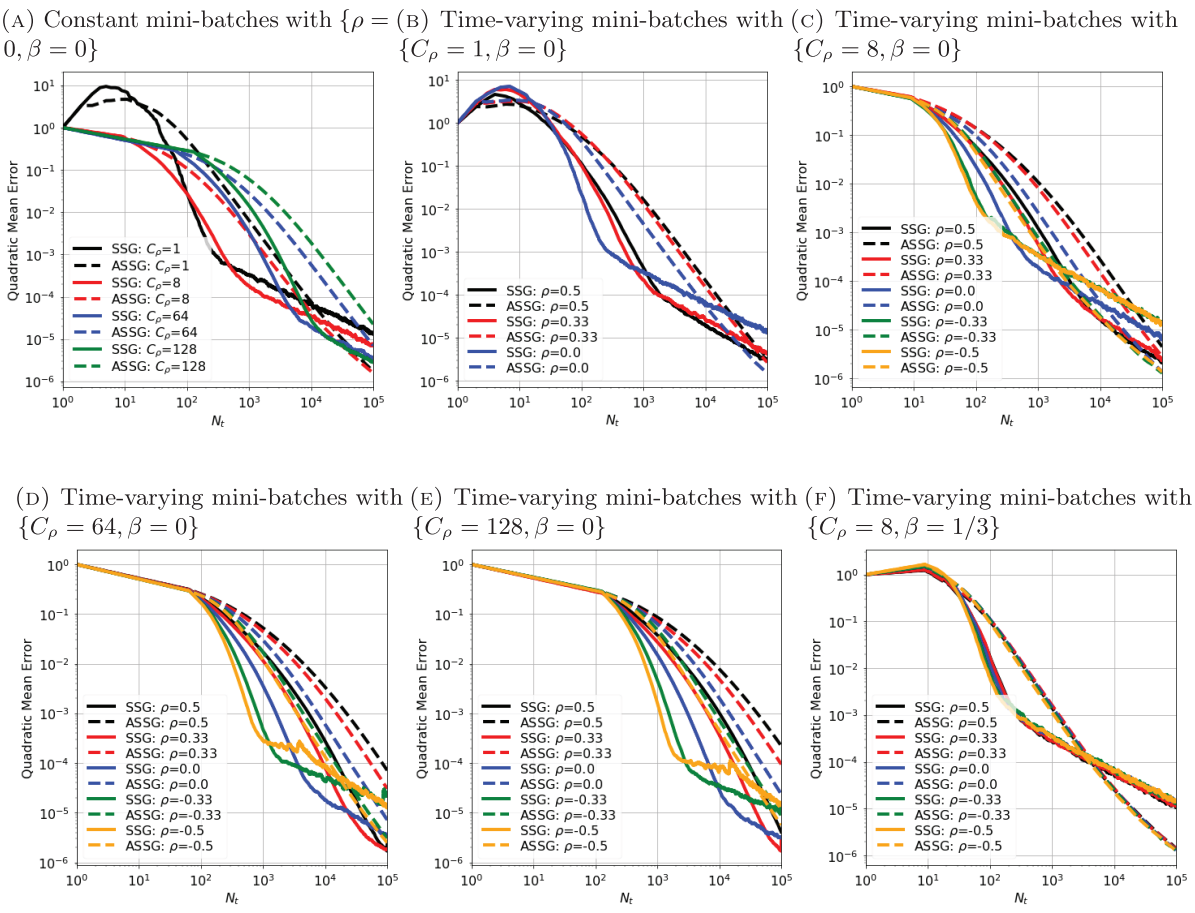

FIGURE 1. Linear regression with learning rate $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ and time-varying mini-batch $n_t = \lceil C_\rho t^\rho \rceil$. See Section 5.1 for details.

5.1. Linear regression

We continue the generic notation from Section 2, where the linear regression is defined by $y = x^T\theta + \epsilon$, where $y \in \mathbb{R}$ is the measure, $x \in \mathbb{R}^d$ is a random feature vector, $\theta \in \mathbb{R}^d$ is the parameters vector, and ϵ is a random variable with zero mean, and x and ϵ are independent and identically distributed. Thus, θ^* is the minimizer of $F(\theta) = \mathbb{E}[(y - x^T\theta)^2]$. 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 and ϵ be standard Gaussian. It is well-known that C_γ can substantially impact convergence; when C_γ is too large, instability can occur, leading to an explosion during the first iterations. If C_γ is too small, the convergence can become very slow and destroy the desired learning rate. To focus on the various time-varying mini-batches, we set $C_\gamma = 1/2$ and $\alpha = 2/3$.

Discussion. In Figure 1a, we consider constant mini-batches to illustrate the results in Corollary 4.5 and 4.10. This figure show a solid decay rate proportional to $\alpha = 2/3$ for any mini-batch size $C_\rho \in \{1, 8, 64, 128\}$ with $\beta = 0$, as shown in Corollary 4.5. In particular, the mini-batches does not provide better convergence rates, but simply scales the error, *i.e.* the slope of the rate of convergence is unchanged, but the intercept is lowered. As explained after Corollary 4.10, we see an acceleration in decay by averaging. Both algorithms show a noticeable reduction in variance when C_ρ increases which are particularly beneficial in the beginning. Next, in Figures 1b to 1e, we vary the mini-batch rate ρ for (fixed) mini-batch sizes $C_\rho = 1$, 8, 64, and 128, respectively, with $\beta = 0$. These figures shows an increase in decay of the SSG when the mini-batch rate ρ increase. Despite

this, we still achieve better convergence for the ASSG algorithm, which seems more immune to the different choices of mini-batch rate ρ, *e.g.*, see the discussion after Corollary 4.11. We know this from Corollary 4.6, as $\phi = (\tilde{\rho} + \alpha)/(1 + \tilde{\rho}) \geq \alpha$ for $\beta = 0$. In addition, we see that C_ρ has a positive effect on the noise (*i.e.*, variance reduction), but if C_ρ becomes too large, it may slow down convergence (as seen in Fig. 1e). Alternatively, we could think around the problem in another way: how can we choose α and β such that we have obtain decay of $\phi = 2/3$ for any ρ. In other words, for any arrival schedule that may occur, how should we choose our hyper-parameters such that we achieve decay of $\phi = 2/3$. As discussed after Corollary 4.11, one example of this could be achieved by setting $\alpha = 2/3$ and $\beta = 1/3$ such that $\phi = 2/3$ for any ρ. Figure 1f shows an example of this where we (indeed) achieve the same decay rate for any mini-batch rate ρ.

5.2. Geometric median

Robust estimators such as the geometric median may be preferred over the mean when the data is noisy; the geometric median is a generalization of the real median introduced by [17]. In addition, SG-based algorithms are preferred in our streaming framework, as they can process large samples of high-dimensional data efficiently [6, 8, 12]. The geometric median of $x \in \mathbb{R}^d$ is found by minimizing the objective $F(\theta) = \mathbb{E}[\|x - \theta\| - \|x\|]$ using gradients of the form $\nabla_\theta f(\theta) = -(x - \theta)/\|x - \theta\|$. Properties of this geometric median, such as existence, uniqueness and robustness, can be found in, *e.g.*, [11, 20]. Note that this objective function only possesses locally strong convexity properties [8]. But by projecting the gradients, one could adapt the proof of [10] to a streaming setting. Otherwise, if x is bounded, one can adapt [7] to the streaming setting showing that the estimates are bounded, and there is no use to project it in this case. Similarly to Section 5.1, we fix $d = 10$ and let x be standard Gaussian centered at $\theta = (-4, -3, 2, 1, 0, 1, 2, 3, 4, 5)^T \in \mathbb{R}^{10}$. Moreover, following the reasoning of [8], we set $C_\gamma = \sqrt{d} = \sqrt{10}$, and let $\alpha = 2/3$.

Discussion. Figure 2a shows the variance reduction effect for different constant mini-batches C_ρ with $\beta = 0$. However, the robustness of the geometric median leaves only a small positive impact for further variance reduction. Thus, too large (constant) mini-batch sizes C_ρ hinders the convergence as we make too few iterations. These findings can be extended to Figures 2b to 2e, where we vary the mini-batch rate ρ for mini-batch sizes $C_\rho = 1$, 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 improved convergence robust to any mini-batch rate ρ. Choosing a proper $\beta > 0$ is particularly important when C_ρ is large, as robustness is an integral part of the geometric median.

6. Conclusions

We introduced a streaming framework for analyzing stochastic approximation/optimization problems. This streaming framework was analogous to solving optimization problems using time-varying mini-batches that arrive sequentially. We provided non-asymptotic convergence rates for different gradient-based algorithms; this included the famous Stochastic Gradient (SG) descent (a.k.a. Robbins-Monro algorithm), mini-batch SG, and time-varying mini-batch SG algorithms, as well as their iterated averages (a.k.a. Polyak-Ruppert averaging). We showed how time-varying mini-batches together with Polyak-Ruppert averaging can provide variance reduction and accelerate convergence simultaneously. We further demonstrated the beneficial effect of adapting learning to the time-varying mini-batches under different streaming settings.

Future perspectives. There are several ways to expand our work: first, we can extend our analysis to include time-varying mini-batches of any size. Second, many machine learning problems encounter correlated variables and high-dimensional data, thus an extension to non-strongly convex objectives would be advantageous [2], *e.g.*, in [39], they use SG-based algorithms to make adaptive volatility predictions through optimization of the GARCH model. Third, Assumption 4.1 requires unbiased (and independent) gradient estimates, thus, an obvious extension could incorporate a more realistic dependency assumption, thereby increasing the applicability. Moreover, studying dependence may give insight into how to process dependent information *optimally*. Next, a natural extension would be to modify our Polyak-Ruppert averaging estimate from (3.3) to a weighted averaged

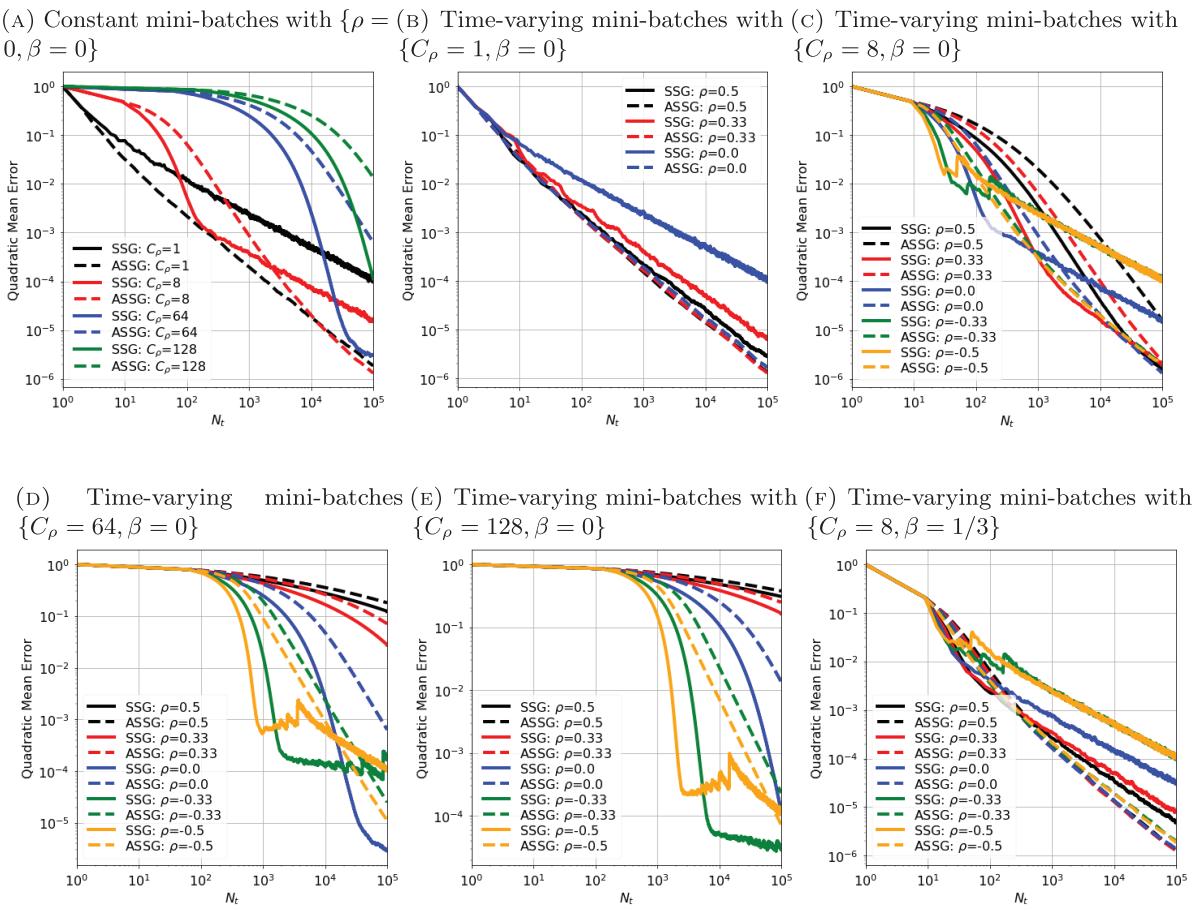

FIGURE 2. Geometric median with learning rate $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ and time-varying mini-batch $n_t = \lceil C_\rho t^\rho \rceil$. See Section 5.2 for details.

version [5, 26]:

(WASSG)
$$\bar{\theta}_{t,\lambda} = \frac{1}{\sum_{i=1}^t n_i \log(1+i)^\lambda} \sum_{i=1}^t n_i \log(1+i)^\lambda \theta_{i-1}, \tag{6.1}$$

for $\lambda > 0$ with (θ_t) following (3.1) or (3.2). One 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 $(\bar{\theta}_{t,\lambda})$ can be found in Figure 3 with use of $\lambda = 2$. Here we see that although the WASSG estimate in (6.1) 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*.

(A) Linear regression: Time-varying mini-batches with $\{C_\rho = 8, \beta = 1/3\}$

(B) Geometric median: Time-varying mini-batches with $\{C_\rho = 8, \beta = 1/3\}$

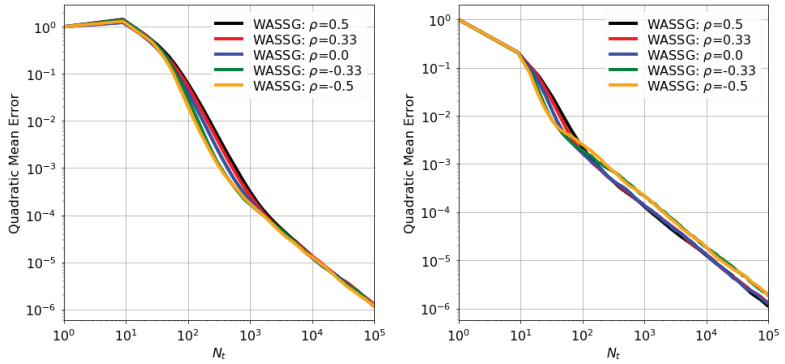

FIGURE 3. WASSG with learning rate $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ and time-varying mini-batch $n_t = \lceil C_\rho t^\rho \rceil$. See Section 6 for details.

APPENDIX A. PROOFS

In this appendix, we provide detailed proofs of the results. Purely technical results used in the proofs can be found in Appendix B. Let $(\mathcal{F}_t)_{t\geq 1}$ be an increasing family of σ-fields, namely $\mathcal{F}_t = \sigma(f_1, \ldots, f_t)$ with $f_t := \{f_{t,1}, \ldots, f_{t,n_t}\}$. Furthermore, we expand the notation with $\mathcal{F}_{t-1,i} = \sigma(f_{1,1}, \ldots, f_{t-1,n_{t-1}}, f_{t,1}, \ldots, f_{t,i})$ such that $\mathcal{F}_{t-1,0} = \mathcal{F}_{t-1}$. Meaning, $\forall 0 \leq i < j$, we have $\mathcal{F}_{t-1} \subseteq \mathcal{F}_{t-1,i} \subset \mathcal{F}_{t-1,j}$. Thus, by the independence of the differentiable random functions $\{f_{t,i}\}$, Assumption 4.1 yields that $\forall t \geq 1$, $\mathbb{E}[\nabla_\theta f_{t,i}(\theta_{t-1})|\mathcal{F}_{t-1,i-1}] = \nabla_\theta F(\theta_{t-1})$ with $i = 1, \ldots, n_t$.

A.1 Proofs for Section 4

The section is structured such that we start by analyzing the recursive relations and bounding them for every choice of learning rate (γ_t) and time-varying mini-batch (n_t). Next, we look at specific choices of (γ_t) and (n_t).

Proof of Theorem 4.4. Taking the quadratic norm on both sides of (3.1), expanding it, and take the conditional expectation, yields

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

To bound the second term (on the right-hand side) of (A.1), we first expand it as follows,

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

For first term of (A.2), we utilize the Lipschitz continuity of $\nabla_\theta f_{t,i}$, together with Assumptions 4.1 to 4.3-p, to obtain

$$\begin{aligned}\mathbb{E}[\|\nabla_\theta f_{t,i}(\theta_{t-1})\|^2|\mathcal{F}_{t-1}] &\leq 2\mathbb{E}[\|\nabla_\theta f_{t,i}(\theta_{t-1}) - \nabla_\theta f_{t,i}(\theta^*)\|^2|\mathcal{F}_{t-1}] + 2\mathbb{E}[\|\nabla_\theta f_{t,i}(\theta^*)\|^2|\mathcal{F}_{t-1}] \\ &\leq 2C_f^2\|\theta_{t-1} - \theta^*\|^2 + 2\sigma^2,\end{aligned} \tag{A.3}$$

using $\|x+y\|^2 \leq 2(\|x\|^2+\|y\|^2)$. Next, for the second term in (A.2): as $\mathcal{F}_{t-1} \subseteq \mathcal{F}_{t-1,i} \subset \mathcal{F}_{t-1,j}$ for all $0 \leq i < j$, we have

$$\mathbb{E}[\langle \nabla_\theta f_{t,i}(\theta_{t-1}), \nabla_\theta f_{t,j}(\theta_{t-1})\rangle | \mathcal{F}_{t-1}] = \mathbb{E}[\mathbb{E}[\langle \nabla_\theta f_{t,i}(\theta_{t-1}), \nabla_\theta F(\theta_{t-1})\rangle | \mathcal{F}_{t-1,i-1}] | \mathcal{F}_{t-1}],$$

since θ_{t-1} and $f_{t,i}$ are $\mathcal{F}_{t-1,j-1}$-measurable for all $0 \leq i < j$, and similarly, as θ_{t-1} is $\mathcal{F}_{t-1}$-measurable and $\mathcal{F}_{t-1,i-1}$-measurable for all $i \geq 0$, we also have

$$\mathbb{E}[\mathbb{E}[\langle \nabla_\theta f_{t,i}(\theta_{t-1}), \nabla_\theta F(\theta_{t-1})\rangle | \mathcal{F}_{t-1,i-1}] | \mathcal{F}_{t-1}] = \mathbb{E}[\langle \mathbb{E}[\nabla_\theta f_{t,i}(\theta_{t-1}) | \mathcal{F}_{t-1,i-1}], \nabla_\theta F(\theta_{t-1})\rangle | \mathcal{F}_{t-1}] = \|\nabla_\theta F(\theta_{t-1})\|^2,$$

where $\|\nabla_\theta F(\theta_{t-1})\|^2 \leq C_\nabla^2 \|\theta_{t-1}-\theta^*\|^2$ as $\nabla_\theta F$ is C_∇-Lipschitz continuous and $\nabla_\theta F(\theta^*) = 0$. Thus, we obtained a bound for the second term (on the right-hand side) of (A.1) using the bounds of the two terms in (A.2):

$$\sum_{i=1}^{n_t} (2C_f^2 \|\theta_{t-1}-\theta^*\|^2 + 2\sigma^2) + \sum_{i \neq j}^{n_t} C_\nabla^2 \|\theta_{t-1}-\theta^*\|^2 = (2C_f^2 n_t + C_\nabla^2 (n_t-1) n_t)\|\theta_{t-1}-\theta^*\|^2 + 2\sigma^2 n_t. \tag{A.4}$$

For the third term (on the right-hand side) of (A.1) we use that F is μ-quasi-strong convex and θ_{t-1} is $\mathcal{F}_{t-1}$-measurable,

$$\mathbb{E}[\langle \nabla_\theta f_{t,i}(\theta_{t-1}), \theta_{t-1}-\theta^*\rangle | \mathcal{F}_{t-1}] = \langle \mathbb{E}[\nabla_\theta f_{t,i}(\theta_{t-1}) | \mathcal{F}_{t-1}], \theta_{t-1}-\theta^*\rangle = \langle \nabla_\theta F(\theta_{t-1}), \theta_{t-1}-\theta^*\rangle \geq \mu \|\theta_{t-1}-\theta^*\|^2, \tag{A.5}$$

by Assumption 4.1. Combining the inequalities from (A.4) and (A.5) into (A.1) and taking the expectation on both sides of the inequality, yields the recursive relation (4.2):

$$\delta_t \leq [1 - 2\mu\gamma_t + (2C_f^2 + (n_t-1)C_\nabla^2) n_t^{-1} \gamma_t^2] \delta_{t-1} + 2\sigma^2 n_t^{-1} \gamma_t^2,$$

with $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ with some $\delta_0 \geq 0$. At last, by Proposition B.5, we obtain the desired inequality in (4.1), namely

$$\delta_t \leq \exp\left(-\mu \sum_{i=t/2}^{t} \gamma_i\right) \exp\left(4C_f^2 \sum_{i=1}^{t} \frac{\gamma_i^2}{n_i}\right) \exp\left(2C_\nabla^2 \sum_{i=1}^{t} \mathbb{1}_{\{n_i>1\}} \gamma_i^2\right) \left(\delta_0 + \frac{2\sigma^2}{C_f^2}\right) + \frac{2\sigma^2}{\mu} \max_{t/2 \leq i \leq t} \frac{\gamma_i}{n_i}.$$

using that $(n_t-1)n_t^{-1} \leq \mathbb{1}_{\{n_t>1\}}$, $n_t \geq 1$, and that $\max_{1\leq i\leq t} 2\sigma^2/(2C_f^2 + (n_i-1)C_\nabla^2) \leq \max_{1\leq i\leq t} 2\sigma^2/2C_f^2 = \sigma^2/C_f^2$. □

Proof of Corollary 4.5. By Theorem 4.4, we have the upper bound giving as

$$\delta_t \leq \exp\left(-\mu \sum_{i=t/2}^{t} \gamma_i\right) \pi_t^c + \frac{2\sigma^2}{\mu C_\rho} \max_{t/2 \leq i \leq t} \gamma_i. \tag{A.6}$$

as $n_t = C_\rho$, with $\pi_t^c = \exp((4C_f^2/C_\rho)\sum_{i=1}^t \gamma_i^2) \exp(2C_\nabla^2 \mathbb{1}_{\{C_\rho>1\}} \sum_{i=1}^t \gamma_i^2)(\delta_0 + \sigma^2/C_f^2)$. The sum term $\sum_{i=1}^t \gamma_i^2 = C_\gamma^2 C_\rho^{2\beta} \sum_{i=1}^t i^{-2\alpha}$ in π_t^c can be bounded with the help of integral tests for convergence, $\sum_{i=1}^t i^{-2\alpha} = 1 + \sum_{i=2}^t i^{-2\alpha} \leq 1 + \int_1^t x^{-2\alpha}\, dx \leq 1 + 1/(2\alpha-1) = 2\alpha/(2\alpha-1)$, as $\alpha \in (1/2, 1)$. Likewise, plugging $\gamma_t = C_\gamma C_\rho^\beta t^{-\alpha}$

into the first term of (A.6), gives

$$\exp\left(-\mu\sum_{i=t/2}^{t}\gamma_i\right)=\exp\left(-\mu C_\gamma C_\rho^\beta\sum_{i=t/2}^{t}i^{-\alpha}\right)\leq\exp\left(-\mu C_\gamma C_\rho^\beta\int_{t/2}^{t}x^{-\alpha}\,dx\right)\leq\exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1-\alpha}}{2^{1-\alpha}}\right),$$

using the integral test for convergence. Next, as $(\gamma_t)_{t\geq 1}$ is decreasing, then $\max_{t/2\leq i\leq t}\gamma_t=\gamma_{t/2}$. Combining all these findings into (A.6), gives us

$$\delta_t\leq\exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1-\alpha}}{2^{1-\alpha}}\right)\pi_\infty^c+\frac{2^{1+\alpha}\sigma^2C_\gamma}{\mu C_\rho^{1-\beta}t^\alpha},\tag{A.7}$$

with $\pi_\infty^c=\exp(4\alpha C_\gamma^2(2C_f^2+C_\rho\mathbb{1}_{\{C_\rho>1\}}C_\nabla^2)/(2\alpha-1)C_\rho^{1-2\beta})(\delta_0+2\sigma^2/C_f^2)$. At last, converting (A.7) into terms of N_t using $N_t\geq C_\rho t$, yields the desired. □

Proof of Corollary 4.6. For convenience, we divided the proof into two cases to comprehend that $n_t\geq 1$ for all t. First, we bound each term of (4.1) (from Thm. 4.4) after inserting $\gamma_t=C_\gamma n_t^\beta t^{-\alpha}$ and $n_t=\lceil C_\rho t^\rho\rceil$ into the inequality. Here, we use $\gamma_t\geq C_\gamma C_\rho^\beta t^{\beta\rho-\alpha}$ if $\rho\geq 0$, $\gamma_t\geq C_\gamma t^{-\alpha}$ if $\rho<0$, and $x\leq\lceil x\rceil\leq x+1$ for $x\in\mathbb{R}_+$. Thus, for $\rho\geq 0$, the first term of (4.1) can be bounded, as follows:

$$\exp\left(-\mu\sum_{i=t/2}^{t}\gamma_i\right)\leq\exp\left(-\mu C_\gamma C_\rho^\beta\sum_{i=t/2}^{t}i^{\beta\rho-\alpha}\right)\leq\exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1+\beta\rho-\alpha}}{2^{1+\beta\rho-\alpha}}\right),$$

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

$$\exp\left(-\mu\sum_{i=t/2}^{t}\gamma_i\right)\leq\exp\left(-\mu C_\gamma\sum_{i=t/2}^{t}i^{-\alpha}\right)\leq\exp\left(-\frac{\mu C_\gamma t^{1-\alpha}}{2^{1-\alpha}}\right).$$

Likewise, with the help of integral tests for convergence, we have for $\rho\geq 0$, that $\sum_{i=1}^t\gamma_i^2/n_i\leq\sum_{i=1}^t\gamma_i^2\leq 2(\alpha-\beta\rho)C_\gamma^2C_\rho^{2\beta}/(2(\alpha-\beta\rho)-1)$, as $n_t\geq 1$ and $\alpha-\rho\beta>1/2$. For $\rho<0$, one has $\sum_{i=1}^t\gamma_i^2/n_i\leq\sum_{i=1}^t\gamma_i^2\leq 2\alpha C_\gamma^2C_\rho^{2\beta}/(2\alpha-1)$ since $C_\rho\geq n_t\geq 1$. Next, as $(1-\beta)\rho+\alpha>0$ for $\rho\geq 0$, then we can bound the last term of (4.1) by

$$\frac{2\sigma^2}{\mu}\max_{t/2\leq i\leq t}\frac{\gamma_i}{n_i}\leq\frac{2\sigma^2C_\gamma}{\mu C_\rho^{1-\beta}}\max_{t/2\leq i\leq t}\frac{1}{i^{(1-\beta)\rho+\alpha}}\leq\frac{2^{1+(1-\beta)\rho+\alpha}\sigma^2C_\gamma}{\mu C_\rho^{1-\beta}t^{(1-\beta)\rho+\alpha}}.$$

using $n_t=\lceil C_\rho t^\rho\rceil\geq C_\rho t^\rho$. Likewise, if $\rho<0$, we have

$$\frac{2\sigma^2}{\mu}\max_{t/2\leq i\leq t}\frac{\gamma_i}{n_i}=\frac{2\sigma^2C_\gamma}{\mu}\max_{t/2\leq i\leq t}\frac{1}{n_i^{1-\beta}i^\alpha}\leq\frac{2^{1+\alpha}\sigma^2C_\gamma}{\mu t^\alpha},$$

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

$$\delta_t\leq\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}}t^{(1-\phi)(1+\tilde\rho)}}{2^{(1-\phi)(1+\tilde\rho)}}\right)\pi_\infty^v+\frac{2^{1+\phi(1+\tilde\rho)}\sigma^2C_\gamma}{\mu C_\rho^{(1-\beta)\mathbb{1}_{\{\rho\geq 0\}}}t^{\phi(1+\tilde\rho)}},\tag{A.8}$$

where $\pi_\infty^v = \exp(4(\alpha-\beta\tilde{\rho})C_\gamma^2 C_\rho^{2\beta}(2C_f^2+C_\nabla^2)/2(\alpha-\beta\tilde{\rho})-1)$ with $\tilde{\rho}=\rho\mathbb{1}_{\{\rho\geq 0\}}$ and $\phi=((1-\beta)\tilde{\rho}+\alpha)/(1+\tilde{\rho})$. To write this in terms of N_t, we use the bounds following bounds: for $\rho\geq 0$, we have that

$$\begin{aligned}N_t=\sum_{i=1}^{t} n_i \leq \sum_{i=1}^{t}(C_\rho i^\rho+1) = t + C_\rho t^\rho + C_\rho\sum_{i=1}^{t-1} i^\rho \leq t + C_\rho t^\rho + C_\rho\int_1^t x^\rho\, dx\\ \leq t + C_\rho t^\rho + C_\rho(t^{1+\rho}-1) \leq t + C_\rho(t^\rho-1) + C_\rho t^{1+\rho} \leq 2C_\rho t^{1+\rho},\end{aligned}$$

thus, $t\geq (N_t/2C_\rho)^{1/(1+\rho)}$. Similarly, for $\rho<0$, we have that $N_t\leq C_\rho t$, i.e, $t\geq N_t/C_\rho$. □

A.2 Proofs for Section 4.2

Lemma A.1 (ASSG/APSSG). *Let $\Delta_t=\mathbb{E}[\|\theta_t-\theta^*\|^4]$ for $\Delta_0\geq 0$, where (θ_t) either follows the recursion in (3.1) or (3.2). Suppose Assumptions 2.1 to 4.8 hold with $p=4$. Then, for any learning rate (γ_t) and time-varying mini-batch (n_t), we have*

$$\Delta_t \leq \exp\left(-\mu\sum_{i=t/2}^{t}\gamma_i\right)\Pi_t^\Delta + \frac{32\sigma^4}{\mu^2}\max_{t/2\leq i\leq t}\frac{\gamma_i^2}{n_i^2} + \frac{48\sigma^4}{\mu}\max_{t/2\leq i\leq t}\frac{\gamma_i^3}{n_i^3} + \frac{114\sigma^4}{\mu}\max_{t/2\leq i\leq t}\frac{\gamma_i^3\mathbb{1}_{\{n_i>1\}}}{n_i^2}, \tag{A.9}$$

with Π_t^Δ given in (A.17).

Proof of Lemma A.1. We will now derive the recursive step sequence for the fourth-order moment using the same arguments as in proof for Theorem 4.4. Thus, one can show that

$$\begin{aligned}\mathbb{E}[\|\theta_t-\theta^*\|^4|\mathcal{F}_{t-1}] \leq& \|\theta_{t-1}-\theta^*\|^4 + \frac{\gamma_t^4}{n_t^4}\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta f_{t,i}\left(\theta_{t-1}\right)\right\|^4\middle|\mathcal{F}_{t-1}\right] + \frac{4\gamma_t^2}{n_t^2}\mathbb{E}\left[\left\langle\sum_{i=1}^{n_t}\nabla_\theta f_{t,i}\left(\theta_{t-1}\right),\theta_{t-1}-\theta^*\right\rangle^2\middle|\mathcal{F}_{t-1}\right]\\ &+\frac{2\gamma_t^2}{n_t^2}\|\theta_{t-1}-\theta^*\|^2\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta f_{t,i}\left(\theta_{t-1}\right)\right\|^2\middle|\mathcal{F}_{t-1}\right] - \frac{4\gamma_t}{n_t}\|\theta_{t-1}-\theta^*\|^2\sum_{i=1}^{n_t}\langle\mathbb{E}[\nabla_\theta f_{t,i}(\theta_{t-1})|\mathcal{F}_{t-1}],\theta_{t-1}-\theta^*\rangle\\ &+\frac{4\gamma_t^3}{n_t^3}\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta f_{t,i}\left(\theta_{t-1}\right)\right\|^2\left\langle\sum_{i=1}^{n_t}\nabla_\theta f_{t,i}\left(\theta_{t-1}\right),\theta_{t-1}-\theta^*\right\rangle\middle|\mathcal{F}_{t-1}\right],\end{aligned}$$

using θ_{t-1} is $\mathcal{F}_{t-1}-$measurable. Note, by Assumption 4.1, we have

$$\langle\mathbb{E}[\nabla_\theta f_{t,i}(\theta_{t-1})|\mathcal{F}_{t-1}],\theta_{t-1}-\theta^*\rangle = \langle\nabla_\theta F(\theta_{t-1}),\theta_{t-1}-\theta^*\rangle \geq \mu\|\theta_{t-1}-\theta^*\|^2,$$

as F is μ-quasi-strong convex. Combining this with Cauchy-Schwarz inequality (*i.e.*, $\langle x,y\rangle\leq\|x\|\|y\|$), we obtain the simplified expression:

$$\begin{aligned}\mathbb{E}[\|\theta_t-\theta^*\|^4|\mathcal{F}_{t-1}] \leq& \|\theta_{t-1}-\theta^*\|^4 + \frac{\gamma_t^4}{n_t^4}\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta f_{t,i}\left(\theta_{t-1}\right)\right\|^4\middle|\mathcal{F}_{t-1}\right] + \frac{6\gamma_t^2}{n_t^2}\|\theta_{t-1}-\theta^*\|^2\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta f_{t,i}\left(\theta_{t-1}\right)\right\|^2\middle|\mathcal{F}_{t-1}\right]\\ &-4\mu\gamma_t\|\theta_{t-1}-\theta^*\|^4 + \frac{4\gamma_t^3}{n_t^3}\|\theta_{t-1}-\theta^*\|\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta f_{t,i}\left(\theta_{t-1}\right)\right\|^3\middle|\mathcal{F}_{t-1}\right].\end{aligned}$$

Next, recall Young's inequality for products, *i.e.*, for any $a_t,b_t,c_t>0$, we have $a_tb_t\leq a_t^2c_t^2/2+b_t^2/2c_t^2$,

$$\left\|\sum_{i=1}^{n_t}\nabla_\theta f_{t,i}\left(\theta_{t-1}\right)\right\|^3 \leq \frac{\gamma_t}{2n_t\|\theta_{t-1}-\theta^*\|}\left\|\sum_{i=1}^{n_t}\nabla_\theta f_{t,i}\left(\theta_{t-1}\right)\right\|^4 + \frac{2n_t\|\theta_{t-1}-\theta^*\|}{\gamma_t}\left\|\sum_{i=1}^{n_t}\nabla_\theta f_{t,i}\left(\theta_{t-1}\right)\right\|^2,$$

giving us

$$\mathbb{E}[\|\theta_t-\theta^*\|^4|\mathcal{F}_{t-1}] \leq (1-4\mu\gamma_t)\|\theta_{t-1}-\theta^*\|^4 + \frac{3\gamma_t^4}{n_t^4}\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta f_{t,i}\left(\theta_{t-1}\right)\right\|^4\middle|\mathcal{F}_{t-1}\right] + \frac{8\gamma_t^2}{n_t^2}\|\theta_{t-1}-\theta^*\|^2\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta f_{t,i}\left(\theta_{t-1}\right)\right\|^2\middle|\mathcal{F}_{t-1}\right]. \tag{A.10}$$

To bound the second and fourth-order terms in (A.10), we would need to study the recursive sequences: firstly, utilizing the Lipschitz continuity of $\nabla_\theta f_{t,i}$, together with Assumptions 4.2-p and 4.3-p, and that θ_{t-1} is $\mathcal{F}_{t-1}$-measurable (Asm. 4.1), we obtain

$$\begin{aligned}\mathbb{E}[\|\nabla_\theta f_{t,i}(\theta_{t-1})\|^p|\mathcal{F}_{t-1}] &\leq 2^{p-1}[\mathbb{E}[\|\nabla_\theta f_{t,i}(\theta_{t-1})-\nabla_\theta f_{t,i}(\theta^*)\|^p|\mathcal{F}_{t-1}] + \mathbb{E}[\|\nabla_\theta f_{t,i}(\theta^*)\|^p|\mathcal{F}_{t-1}]] \\ &\leq 2^{p-1}[C_f^p\|\theta_{t-1}-\theta^*\|^p + \sigma^p],\end{aligned} \tag{A.11}$$

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 (A.10) by

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta f_{t,i}\left(\theta_{t-1}\right)\right\|^2\middle|\mathcal{F}_{t-1}\right] &\leq [2C_f^2 n_t + C_\nabla^2(n_t-1)n_t]\|\theta_{t-1}-\theta^*\|^2 + 2\sigma^2 n_t \\ &\leq [2C_f^2 n_t + C_\nabla^2 n_t^2\mathbb{1}_{\{n_t>1\}}]\|\theta_{t-1}-\theta^*\|^2 + 2\sigma^2 n_t,\end{aligned} \tag{A.12}$$

following the same steps in the proof of Theorem 4.4, but with use of (A.11). Bounding the fourth-order term is a bit heavier computationally, but let us recall that $\|\sum_i x_i\|^2 = \sum_i\|x_i\|^2 + \sum_{i\neq j}\langle x_i,x_j\rangle = \sum_i\|x_i\|^2 + 2\sum_{i<j}\langle x_i,x_j\rangle$. Then, we have that

$$\begin{aligned}\left\|\sum_{i=1}^{n_t}\nabla_\theta f_{t,i}(\theta_{t-1})\right\|^4 &= \left(\sum_{i=1}^{n_t}\|\nabla_\theta f_{t,i}(\theta_{t-1})\|^2 + \sum_{i\neq j}^{n_t}\langle\nabla_\theta f_{t,i}(\theta_{t-1}),\nabla_\theta f_{t,j}(\theta_{t-1})\rangle\right)^2 \\ &\leq 2\left(\sum_{i=1}^{n_t}\|\nabla_\theta f_{t,i}(\theta_{t-1})\|^2\right)^2 + 4\left(\sum_{i<j}^{n_t}\langle\nabla_\theta f_{t,i}(\theta_{t-1}),\nabla_\theta f_{t,j}(\theta_{t-1})\rangle\right)^2,\end{aligned} \tag{A.13}$$

as $(x+y)^2\leq 2x^2+2y^2$. For the first term of (A.13), we have

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{n_t}\|\nabla_\theta f_{t,i}(\theta_{t-1})\|^2\right)^2\middle|\mathcal{F}_{t-1}\right] &= \sum_{i=1}^{n_t}\mathbb{E}[\|\nabla_\theta f_{t,i}(\theta_{t-1})\|^4|\mathcal{F}_{t-1}] + \sum_{i\neq j}^{n_t}\mathbb{E}[\|\nabla_\theta f_{t,i}(\theta_{t-1})\|^2\|\nabla_\theta f_{t,j}(\theta_{t-1})\|^2|\mathcal{F}_{t-1}] \\ &\leq 8n_t[C_f^4\|\theta_{t-1}-\theta^*\|^4+\sigma^4] + 4n_t^2\mathbb{1}_{\{n_t>1\}}[C_f^2\|\theta_{t-1}-\theta^*\|^2+\sigma^2]^2,\end{aligned}$$

using the bound from (A.11) , $n_t(n_t-1) \le n_t^2 \mathbb{1}_{\{n_t>1\}}$, and that $\mathcal{F}_{t-1} \subseteq \mathcal{F}_{t-1,i} \subset \mathcal{F}_{t-1,j}$ for all $0 \le i < j$. To bound the second term of (A.13), we ease notation by denoting $\nabla_\theta f_{t,i}(\theta_{t-1})$ by v_i, giving us

$$\begin{aligned}\left(\sum_{i<j}^{n_t}\langle v_i, v_j\rangle\right)^2 &= \sum_{i<j}^{n_t}\langle v_i, v_j\rangle^2 + \sum_{\substack{i<j,k<l\\(i,j)\neq(k,l)}}^{n_t}\langle v_i, v_j\rangle\langle v_k, v_l\rangle \\ &= \underbrace{\sum_{i<j}^{n_t}\langle v_i, v_j\rangle^2}_{A} + \underbrace{\sum_{\substack{i<j,k<l\\(i,j)\neq(k,l),j=l}}^{n_t}\langle v_i, v_j\rangle\langle v_k, v_l\rangle}_{B} + \underbrace{\sum_{\substack{i<j,k<l\\(i,j)\neq(k,l),j\neq l}}^{n_t}\langle v_i, v_j\rangle\langle v_k, v_l\rangle}_{C}.\end{aligned}$$

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

$$\mathbb{E}[A|\mathcal{F}_{t-1}] \le \sum_{i<j}^{n_t}\mathbb{E}[\|v_i\|^2\|v_j\|^2|\mathcal{F}_{t-1}] \le 2n_t(n_t-1)[C_f^2\|\theta_{t-1}-\theta^*\|^2+\sigma^2]^2 \le 2n_t^2\mathbb{1}_{\{n_t>1\}}[C_f^2\|\theta_{t-1}-\theta^*\|^2+\sigma^2]^2,$$

using that $\mathcal{F}_{t-1} \subseteq \mathcal{F}_{t-1,i} \subset \mathcal{F}_{t-1,j}$ for all $0 \le i < j$. Next, since $l = j$ implies $i \neq k$, we have

$$\begin{aligned}\mathbb{E}[B|\mathcal{F}_{t-1}] &= \sum_{i<j,k<l,i\neq k,j=l}^{n_t}\mathbb{E}[\langle v_i, v_j\rangle\langle v_k, v_l\rangle|\mathcal{F}_{t-1}] \\ &= \sum_{i<j,k<l,i\neq k,j=l}^{n_t}\mathbb{E}[\mathbb{E}[\langle\mathbb{E}[v_i|\mathcal{F}_{t-1,i-1}], v_j\rangle\langle\mathbb{E}[v_k|\mathcal{F}_{t-1,k-1}], v_l\rangle|\mathcal{F}_{t-1,l-1}]|\mathcal{F}_{t-1}] \\ &= \sum_{i<j,k<l,i\neq k,j=l}^{n_t}\mathbb{E}[\mathbb{E}[\langle\nabla_\theta F(\theta_{t-1}), v_l\rangle^2|\mathcal{F}_{t-1,l-1}]|\mathcal{F}_{t-1}] \\ &\le \sum_{i<j,k<l,i\neq k,j=l}^{n_t}\mathbb{E}[\|\nabla_\theta F(\theta_{t-1})\|^2\mathbb{E}[\|v_l\|^2|\mathcal{F}_{t-1,l-1}]|\mathcal{F}_{t-1}] \\ &\le \sum_{i<j,k<l,i\neq k,j=l}^{n_t}2C_\nabla^2\|\theta_{t-1}-\theta^*\|^2[C_f^2\|\theta_{t-1}-\theta^*\|^2+\sigma^2] \\ &= n_t(n_t-1)(n_t-2)C_\nabla^2\|\theta_{t-1}-\theta^*\|^2[C_f^2\|\theta_{t-1}-\theta^*\|^2+\sigma^2] \\ &\le n_t^3\mathbb{1}_{\{n_t>1\}}C_\nabla^2\|\theta_{t-1}-\theta^*\|^2[C_f^2\|\theta_{t-1}-\theta^*\|^2+\sigma^2],\end{aligned}$$

using Cauchy-Schwarz inequality and the bound in (A.11). In the same way, as $j \neq l$ includes $(i,j) \neq (k,l)$, we can rewrite C as

$$C = \sum_{i<j,k<l,j\neq l}^{n_t}\langle v_i, v_j\rangle\langle v_k, v_l\rangle = \underbrace{\sum_{i<j,k<l,i=k,j\neq l}^{n_t}\langle v_i, v_j\rangle\langle v_k, v_l\rangle}_{C_1} + \underbrace{\sum_{i<j,k<l,i\neq k,j\neq l}^{n_t}\langle v_i, v_j\rangle\langle v_k, v_l\rangle}_{C_2},$$

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

$$C_2 = \underbrace{\sum_{i<j,k<l,i\neq k,j\neq l,i=l,j\neq k}^{n_t} \langle v_i v_j\rangle\langle v_k v_l\rangle}_{C_{2,1}} + \underbrace{\sum_{i<j,k<l,i\neq k,j\neq l,i\neq l,j=k}^{n_t} \langle v_i v_j\rangle\langle v_k v_l\rangle}_{C_{2,2}} + \underbrace{\sum_{i<j,k<l,i\neq j\neq k\neq l}^{n_t} \langle v_i v_j\rangle\langle v_k v_l\rangle}_{C_{2,3}},$$

where $\mathbb{E}[C_{2,1}|\mathcal{F}_{t-1}] = \mathbb{E}[C_{2,2}|\mathcal{F}_{t-1}] = \mathbb{E}[B|\mathcal{F}_{t-1}]$, and

$$\begin{aligned}\mathbb{E}[C_{2,3}|\mathcal{F}_{t-1}] &= \sum_{i<j,k<l,i\neq j\neq k\neq l}^{n_t} \mathbb{E}[\|\nabla_\theta F(\theta_{t-1})\|^4|\mathcal{F}_{t-1}]\\ &\leq n_t(n_t-1)(n_t-2)(n_t-3)C_\nabla^4\|\theta_{t-1}-\theta^*\|^4\\ &\leq n_t^4\mathbb{1}_{\{n_t>1\}}C_\nabla^4\|\theta_{t-1}-\theta^*\|^4.\end{aligned}$$

Thus, the fourth-order term of (A.10), is bounded by

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta f_{t,i}\left(\theta_{t-1}\right)\right\|^4\middle|\mathcal{F}_{t-1}\right] \leq& 16n_t[C_f^4\|\theta_{t-1}-\theta^*\|^4+\sigma^4] + 16n_t^2\mathbb{1}_{\{n_t>1\}}[C_f^2\|\theta_{t-1}-\theta^*\|^2+\sigma^2]^2\\ &+12n_t^3\mathbb{1}_{\{n_t>1\}}C_\nabla^2\|\theta_{t-1}-\theta^*\|^2[C_f^2\|\theta_{t-1}-\theta^*\|^2+\sigma^2] + 4n_t^4\mathbb{1}_{\{n_t>1\}}C_\nabla^4\|\theta_{t-1}-\theta^*\|^4\\ \leq&[16C_f^4n_t + 16C_f^4n_t^2\mathbb{1}_{\{n_t>1\}} + 12C_\nabla^2C_f^2n_t^3\mathbb{1}_{\{n_t>1\}} + 4C_\nabla^4n_t^4\mathbb{1}_{\{n_t>1\}}]\|\theta_{t-1}-\theta^*\|^4\\ &+[32C_f^2\sigma^2n_t^2\mathbb{1}_{\{n_t>1\}} + 12C_\nabla^2\sigma^2n_t^3\mathbb{1}_{\{n_t>1\}}]\|\theta_{t-1}-\theta^*\|^2 + 16\sigma^4n_t + 16\sigma^4n_t^2\mathbb{1}_{\{n_t>1\}}.\end{aligned} \tag{A.14}$$

Combining the bound from (A.12) and (A.14) into (A.10), we can bound the fourth-order moment $\mathbb{E}[\|\theta_t-\theta^*\|^4|\mathcal{F}_{t-1}]$ by the recursive relation:

$$\begin{aligned}&[1-4\mu\gamma_t + 8C_\nabla^2\mathbb{1}_{\{n_t>1\}}\gamma_t^2 + 16C_f^2n_t^{-1}\gamma_t^2 + 48C_f^4n_t^{-3}\gamma_t^4 + 48C_f^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4 + 36C_\nabla^2C_f^2n_t^{-1}\mathbb{1}_{\{n_t>1\}}\gamma_t^4\\ &+12C_\nabla^4\mathbb{1}_{\{n_t>1\}}\gamma_t^4]\|\theta_{t-1}-\theta^*\|^4 + [16\sigma^2n_t^{-1}\gamma_t^2 + 96C_f^2\sigma^2n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4 + 36C_\nabla^2\sigma^2n_t^{-1}\mathbb{1}_{\{n_t>1\}}\gamma_t^4]\|\theta_{t-1}-\theta^*\|^2\\ &+48\sigma^4n_t^{-3}\gamma_t^4 + 48\sigma^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4.\end{aligned}$$

By Young's inequality for products, one have

$$\begin{aligned}2C_\nabla^2C_f^2 &\leq n_tC_\nabla^4 + n_t^{-1}C_f^4,\\ 16\sigma^2n_t^{-1}\gamma_t^2\|\theta_{t-1}-\theta^*\|^2 &\leq 2\mu\gamma_t\|\theta_t-\theta^*\|^4 + 32\sigma^4\mu^{-1}n_t^{-2}\gamma_t^3,\\ 2C_f^2\sigma^2n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4\|\theta_{t-1}-\theta^*\|^2 &\leq C_f^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4\|\theta_t-\theta^*\|^4 + \sigma^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4,\\ 2C_\nabla^2\sigma^2n_t^{-1}\mathbb{1}_{\{n_t>1\}}\gamma_t^4\|\theta_{t-1}-\theta^*\|^2 &\leq C_\nabla^4\mathbb{1}_{\{n_t>1\}}\gamma_t^4\|\theta_t-\theta^*\|^4 + \sigma^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4,\end{aligned}$$

which yields the bound on $\mathbb{E}[\|\theta_t-\theta^*\|^4|\mathcal{F}_{t-1}]$,

$$\begin{aligned}&[1-2\mu\gamma_t + 8C_\nabla^2\mathbb{1}_{\{n_t>1\}}\gamma_t^2 + 16C_f^2n_t^{-1}\gamma_t^2 + 48C_f^4n_t^{-3}\gamma_t^4 + 114C_f^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4 + 48C_\nabla^4\mathbb{1}_{\{n_t>1\}}\gamma_t^4]\|\theta_{t-1}-\theta^*\|^4\\ &+32\mu^{-1}\sigma^4n_t^{-2}\gamma_t^3 + 48\sigma^4n_t^{-3}\gamma_t^4 + 114\sigma^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4.\end{aligned} \tag{A.15}$$

Taking, the expectation on both sides of the inequality in (A.15) yields the recursive relation for the fourth-order moment:

$$\Delta_t \leq [1-2\mu\gamma_t + 8C_\nabla^2\mathbb{1}_{\{n_t>1\}}\gamma_t^2 + 16C_f^2n_t^{-1}\gamma_t^2 + 48C_f^4n_t^{-3}\gamma_t^4 + 114C_f^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4 + 48C_\nabla^4\mathbb{1}_{\{n_t>1\}}\gamma_t^4]\Delta_{t-1}$$

$$+ 32\mu^{-1}\sigma^4 n_t^{-2}\gamma_t^3 + 48\sigma^4 n_t^{-3}\gamma_t^4 + 114\sigma^4 n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4. \tag{A.16}$$

with $\Delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^4]$ for some $\Delta_0 \geq 0$. By Proposition B.5, we achieve the (upper) bound of Δ_t in (A.16), given as

$$\Delta_t \leq \exp\left(-\mu \sum_{i=t/2}^{t} \gamma_i\right) \Pi_t^{\Delta} + \frac{32\sigma^4}{\mu^2} \max_{t/2\leq i\leq t} \frac{\gamma_i^2}{n_i^2} + \frac{48\sigma^4}{\mu} \max_{t/2\leq i\leq t} \frac{\gamma_i^3}{n_i^3} + \frac{114\sigma^4}{\mu} \max_{t/2\leq i\leq t} \frac{\gamma_i^3 \mathbb{1}_{\{n_i>1\}}}{n_i^2}.$$

where Π_t^{Δ} is given by

$$\begin{aligned} &\exp\left(32C_f^2 \sum_{i=1}^{t} \frac{\gamma_i^2}{n_i}\right) \exp\left(96C_f^4 \sum_{i=1}^{t} \frac{\gamma_i^4}{n_i^3}\right) \exp\left(228C_f^4 \sum_{i=1}^{t} \frac{\mathbb{1}_{\{n_i>1\}}\gamma_i^4}{n_i^2}\right) \\ &\exp\left(16C_\nabla^2 \sum_{i=1}^{t} \mathbb{1}_{\{n_i>1\}}\gamma_i^2\right) \exp\left(96C_\nabla^4 \sum_{i=1}^{t} \mathbb{1}_{\{n_i>1\}}\gamma_i^4\right) \left(\Delta_0 + \frac{2\sigma^4}{C_f^4} + \frac{4\sigma^4\gamma_1}{\mu C_f^2 n_1}\right), \end{aligned} \tag{A.17}$$

with use of

$$\max_{1\leq i\leq t} \frac{32\mu^{-1}\sigma^4 n_i^{-2}\gamma_i + 48\sigma^4 n_i^{-3}\gamma_i^2 + 114\sigma^4 n_i^{-2}\mathbb{1}_{\{n_i>1\}}\gamma_i^2}{8C_\nabla^2\mathbb{1}_{\{n_i>1\}} + 16C_f^2 n_i^{-1} + 48C_f^4 n_i^{-3}\gamma_i^2 + 114C_f^4 n_i^{-2}\mathbb{1}_{\{n_i>1\}}\gamma_i^2 + 48C_\nabla^4\mathbb{1}_{\{n_i>1\}}\gamma_i^2} \leq \frac{\sigma^4}{C_f^4} + \frac{2\sigma^4\gamma_1}{\mu C_f^2 n_1}.$$

At last, bounding the projected estimate (3.2) follows from that $\mathbb{E}[\|\mathcal{P}_\Theta(\theta) - \theta^*\|^2] \leq \mathbb{E}[\|\theta - \theta^*\|^2]$, $\forall \theta \in \Theta$. □

A.2.1 Proofs for Section 4.2.1

Proof of Theorem 4.9. Following [32], we rewrite (3.1) to

$$\theta_t = \theta_{t-1} - \frac{\gamma_t}{n_t}\sum_{i=1}^{n_t} \nabla_\theta f_{t,i}(\theta_{t-1}) \iff \frac{1}{\gamma_t}(\theta_{t-1} - \theta_t) = \nabla_\theta f_t(\theta_{t-1}), \tag{A.18}$$

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

$$\nabla_\theta^2 F(\theta^*)(\theta_{t-1} - \theta^*) = \nabla_\theta f_t(\theta_{t-1}) - \nabla_\theta f_t(\theta^*) - \underbrace{[\nabla_\theta f_t(\theta_{t-1}) - \nabla_\theta f_t(\theta^*) - \nabla_\theta F(\theta_{t-1})]}_{\text{martingale term}} - \underbrace{[\nabla_\theta F(\theta_{t-1}) - \nabla_\theta^2 F(\theta^*)(\theta_{t-1} - \theta^*)]}_{\text{rest term}},$$

where $\nabla_\theta^2 F(\theta^*)$ is invertible with lowest eigenvalue greater than μ, *i.e.*, $\nabla_\theta^2 F(\theta^*) \geq \mu$. Thus, summing the parts and using the Minkowski's inequality, we obtain the inequality:

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

$$+\left(\mathbb{E}\left[\left\|\nabla_\theta^2 F\left(\theta^*\right)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i\left[\nabla_\theta F\left(\theta_{i-1}\right)-\nabla_\theta^2 F\left(\theta^*\right)\left(\theta_{i-1}-\theta^*\right)\right]\right\|^2\right]\right)^{\frac{1}{2}}.$$

As $(\nabla_\theta f_{t,i}(\theta^*))$ is a square-integrable martingale increment sequences on $\mathbb{R}^d$ (Asm. 4.1), we have

$$\mathbb{E}\left[\left\|\nabla_\theta^2 F\left(\theta^*\right)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i\nabla_\theta f_i\left(\theta^*\right)\right\|^2\right]\leq\frac{1}{N_t^2}\sum_{i=1}^{t}\sum_{j=1}^{n_i}\mathbb{E}\left[\left\|\nabla_\theta^2 F\left(\theta^*\right)^{-1}\nabla_\theta f_{i,j}\left(\theta^*\right)\right\|^2\right]\leq\frac{\mathrm{Tr}\left[\nabla_\theta^2 F(\theta^*)^{-1}\Sigma\nabla_\theta^2 F(\theta^*)^{-1}\right]}{N_t},\tag{A.19}$$

using Assumption 4.8. To ease notation, we denote $\mathrm{Tr}[\nabla_\theta^2 F(\theta^*)^{-1}\Sigma\nabla_\theta^2 F(\theta^*)^{-1}]$ by Λ. Next, note that for all $t\geq 1$, we have the relation in (A.18), giving us

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

leading to

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

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

$$\left(\mathbb{E}\left[\left\|\nabla_\theta^2 F\left(\theta^*\right)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i\nabla_\theta f_i\left(\theta_{i-1}\right)\right\|^2\right]\right)^{\frac{1}{2}}\leq\frac{1}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{\frac{1}{2}}\left|\frac{n_{i+1}}{\gamma_{i+1}}-\frac{n_i}{\gamma_i}\right|+\frac{n_t}{N_t\gamma_t\mu}\delta_t^{\frac{1}{2}}+\frac{n_1}{N_t\gamma_1\mu}\delta_0^{\frac{1}{2}}.\tag{A.20}$$

For the martingale term, we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\nabla_\theta^2 F\left(\theta^*\right)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i\left[\nabla_\theta f_i\left(\theta_{i-1}\right)-\nabla_\theta f_i\left(\theta^*\right)-\nabla_\theta F\left(\theta_{i-1}\right)\right]\right\|^2\right]\leq\frac{1}{N_t^2\mu^2}\sum_{i=1}^{t} n_i^2\mathbb{E}\left[\left\|\nabla_\theta f_i\left(\theta_{i-1}\right)-\nabla_\theta f_i\left(\theta^*\right)\right\|^2\right]\\&=\frac{1}{N_t^2\mu^2}\sum_{i=1}^{t}\mathbb{E}\left[\left\|\sum_{j=1}^{n_i}\nabla_\theta f_{i,j}\left(\theta_{i-1}\right)-\nabla_\theta f_{i,j}\left(\theta^*\right)\right\|^2\right]\leq\frac{1}{N_t^2\mu^2}\sum_{i=1}^{t}\sum_{j=1}^{n_i}\left(\mathbb{E}\left[\left\|\nabla_\theta f_{i,j}\left(\theta_{i-1}\right)-\nabla_\theta f_{i,j}\left(\theta^*\right)\right\|^2\right]\right)^{\frac{1}{2}}\\&\leq\frac{C_f^2}{N_t^2\mu^2}\sum_{i=1}^{t} n_i\delta_{i-1},\end{aligned}\tag{A.21}$$

by Cauchy-Schwarz inequality and Assumption 4.2-p. For all $t\geq 1$, the rest term is directly bounded by (4.5):

$$\left(\mathbb{E}\left[\left\|\nabla_\theta^2 F\left(\theta^*\right)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i\left[\nabla_\theta F\left(\theta_{i-1}\right)-\nabla_\theta^2 F\left(\theta^*\right)\left(\theta_{i-1}-\theta^*\right)\right]\right\|^2\right]\right)^{\frac{1}{2}}\leq\frac{C'_\nabla}{N_t\mu}\sum_{i=1}^{t} n_i\Delta_{i-1}^{\frac{1}{2}},\tag{A.22}$$

with the notion $\Delta_t=\mathbb{E}[\|\theta_t-\theta^*\|^4]$. Finally, combining the terms from (A.19) to (A.22), gives us

$$\bar{\delta}_t^{1/2}\leq\frac{\Lambda^{1/2}}{N_t^{1/2}}+\frac{1}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{1/2}\left|\frac{n_{i+1}}{\gamma_{i+1}}-\frac{n_i}{\gamma_i}\right|+\frac{n_t}{N_t\gamma_t\mu}\delta_t^{1/2}+\frac{n_1}{N_t\gamma_1\mu}\delta_0^{1/2}+\frac{C_f}{N_t\mu}\left(\sum_{i=1}^{t} n_i\delta_{i-1}\right)^{1/2}+\frac{C'_\nabla}{N_t\mu}\sum_{i=1}^{t} n_i\Delta_{i-1}^{1/2},\tag{A.23}$$

where $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$, which can be simplified into (4.6) by shifting the indices and collecting the δ_0 terms. □

Proof of Corollary 4.10. As $n_t = C_\rho$ for all $t \geq 1$, we simplify the bound for $\bar{\delta}_t^{1/2}$ in (4.6) to

$$\frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{C_\rho}{N_t \mu} \sum_{i=1}^{t-1} \delta_i^{1/2} \left| \frac{1}{\gamma_{i+1}} - \frac{1}{\gamma_i} \right| + \frac{C_\rho}{N_t \gamma_t \mu} \delta_t^{1/2} + \frac{C_\rho}{N_t \mu} \left(\frac{1}{\gamma_1} + C_f \right) \delta_0^{1/2} + \frac{C_f C_\rho^{\frac{1}{2}}}{N_t \mu} \left(\sum_{i=1}^{t-1} \delta_i \right)^{1/2} + \frac{C'_\nabla C_\rho}{N_t \mu} \sum_{i=0}^{t-1} \Delta_i^{1/2}. \tag{A.24}$$

The second-order moment δ_t is bounded by Corollary 4.5 but with use of (A.7) as we work in terms of t. The fourth-order moment Δ_t from Lemma A.1 can be simplified to:

$$\begin{aligned}
\Delta_t &\leq \exp\left(-\mu \sum_{i=t/2}^{t} \gamma_i \right) \Pi_\infty^c + \frac{1}{\mu} \left(\frac{32\sigma^4}{\mu C_\rho^2} \max_{t/2 \leq i \leq t} \gamma_i^2 + \frac{48\sigma^4}{C_\rho^3} \max_{t/2 \leq i \leq t} \gamma_i^3 + \frac{114\sigma^4 \mathbb{1}_{\{C_\rho > 1\}}}{C_\rho^2} \max_{t/2 \leq i \leq t} \gamma_i^3 \right) \\
&\leq \exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1-\alpha}}{2^{1-\alpha}} \right) \Pi_\infty^c + \frac{1}{\mu} \left(\frac{2^{2\alpha} 32\sigma^4 C_\gamma^2 C_\rho^{2\beta}}{\mu C_\rho^2 t^{2\alpha}} + \frac{2^{3\alpha} 48\sigma^4 C_\gamma^3 C_\rho^{3\beta}}{C_\rho^3 t^{3\alpha}} + \frac{2^{3\alpha} 114\sigma^4 C_\gamma^3 C_\rho^{3\beta} \mathbb{1}_{\{C_\rho > 1\}}}{C_\rho^2 t^{3\alpha}} \right),
\end{aligned}$$

using that $\gamma_t = C_\gamma C_\rho^\beta t^{-\alpha}$ is decreasing as $\alpha \in (1/2, 1)$. Regarding Π_t^Δ defined in (A.17), we can bound it by

$$\begin{aligned}
\Pi_\infty^c = \exp\left(\frac{64\alpha C_f^2 C_\gamma^2 C_\rho^{2\beta}}{(2\alpha - 1) C_\rho} \right) \exp\left(\frac{(192 + 456 C_\rho \mathbb{1}_{\{C_\rho > 1\}}) C_f^4 C_\gamma^4 C_\rho^{4\beta}}{C_\rho^3} \right) \exp\left(\frac{32\alpha C_\nabla^2 C_\gamma^2 C_\rho^{2\beta} \mathbb{1}_{\{C_\rho > 1\}}}{2\alpha - 1} \right) \\
\exp\left(192 C_\nabla^4 C_\gamma^4 C_\rho^{4\beta} \mathbb{1}_{\{C_\rho > 1\}} \right) \left(\Delta_0 + \frac{2\sigma^4}{C_f^4} + \frac{4\sigma^4 C_\gamma}{\mu C_f^2 C_\rho^{1-\beta}} \right),
\end{aligned}$$

using $\sum_{i=1}^t i^{-2\alpha} \leq 2\alpha/(2\alpha - 1)$ and $\sum_{i=1}^t i^{-4\alpha} \leq 2$. Note that Π_∞^c is a finite constant, independent of t. To bound the first term of (A.24), namely $\frac{C_\rho}{N_t \mu} \sum_{i=1}^{t-1} \delta_i^{1/2} |\gamma_{i+1}^{-1} - \gamma_i^{-1}|$, we remark that $|\gamma_{t+1}^{-1} - \gamma_t^{-1}| \leq C_\gamma^{-1} C_\rho^{-\beta} \alpha t^{\alpha - 1}$, one has (since $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$),

$$\frac{C_\rho}{N_t \mu} \sum_{i=1}^{t-1} \delta_i^{\frac{1}{2}} \left| \frac{1}{\gamma_{i+1}} - \frac{1}{\gamma_i} \right| \leq \frac{C_\rho^{1-\beta} \alpha}{C_\gamma \mu N_t} \sum_{i=1}^{t} i^{\alpha - 1} \left(\exp\left(-\frac{\mu C_\gamma C_\rho^\beta i^{1-\alpha}}{2^{2-\alpha}} \right) \sqrt{\pi_\infty^c} + \frac{2^{\frac{1+\alpha}{2}} \sigma \sqrt{C_\gamma}}{\sqrt{\mu} C_\rho^{\frac{1-\beta}{2}} i^{\alpha/2}} \right). \tag{A.25}$$

For simplicity, let us denote

$$A_\infty^c = \sum_{i=0}^{\infty} \exp\left(-\frac{\mu C_\gamma C_\rho^\beta i^{1-\alpha}}{2^{2-\alpha}} \right) \geq \sum_{i=0}^{\infty} i^{\alpha - 1} \exp\left(-\frac{\mu C_\gamma C_\rho^\beta i^{1-\alpha}}{2^{2-\alpha}} \right),$$

as $\alpha < 1$. Thus, the first part of (A.25) is bounded as follows:

$$\frac{C_\rho^{1-\beta} \alpha \sqrt{\pi_\infty^c}}{C_\gamma \mu N_t} \sum_{i=1}^{t} i^{\alpha - 1} \exp\left(-\frac{\mu C_\gamma C_\rho^\beta i^{1-\alpha}}{2^{2-\alpha}} \right) \leq \frac{C_\rho^{1-\beta} \alpha \sqrt{\pi_\infty^c} A_\infty^c}{C_\gamma \mu N_t}.$$

Furthermore, with the help of an integral test for convergence, one has $\sum_{i=1}^{t} i^{\alpha/2-1} \leq 1 + \int_1^t s^{\alpha/2-1}\, ds = 1 + (2/\alpha)t^{\alpha/2} - (2/\alpha) \leq (2/\alpha)t^{\alpha/2}$, such that the second part of (A.25) can be bounded by

$$\frac{2^{\frac{1+\alpha}{2}}\sigma C_\rho^{\frac{1-\beta}{2}}\alpha}{C_\gamma^{1/2}\mu^{3/2}N_t}\sum_{i=1}^{t} i^{\alpha/2-1} \leq \frac{2^{\frac{3+\alpha}{2}}\sigma C_\rho^{\frac{1-\beta}{2}}t^{\alpha/2}}{C_\gamma^{1/2}\mu^{3/2}N_t} = \frac{2^{\frac{3+\alpha}{2}}\sigma C_\rho^{\frac{1-\alpha-\beta}{2}}}{C_\gamma^{1/2}\mu^{3/2}N_t^{1-\alpha/2}}.$$

By combining this, we get

$$\frac{C_\rho}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{\frac{1}{2}}\left|\frac{1}{\gamma_{i+1}} - \frac{1}{\gamma_i}\right| \leq \frac{C_\rho^{1-\beta}\alpha\sqrt{\pi_\infty^c}A_\infty^c}{C_\gamma\mu N_t} + \frac{2^{\frac{3+\alpha}{2}}\sigma C_\rho^{\frac{1-\alpha-\beta}{2}}}{\sqrt{C_\gamma}\mu^{3/2}N_t^{1-\alpha/2}}. \tag{A.26}$$

Similarly, second term of (A.24), can be bounded by

$$\frac{C_\rho}{N_t\gamma_t\mu}\delta_t^{\frac{1}{2}} \leq \frac{C_\rho^{1-\alpha-\beta}}{C_\gamma\mu N_t^{1-\alpha}}\left(\exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1-\alpha}}{2^{2-\alpha}}\right)\sqrt{\pi_\infty^c} + \frac{2^{\frac{1+\alpha}{2}}\sigma\sqrt{C_\gamma}}{\sqrt{\mu}C_\rho^{\frac{1-\beta}{2}}t^{\alpha/2}}\right) \leq \frac{C_\rho^{2-\alpha-\beta}\sqrt{\pi_\infty^c}A_\infty^c}{C_\gamma\mu N_t^{2-\alpha}} + \frac{2^{\frac{1+\alpha}{2}}C_\rho^{\frac{1-\alpha-\beta}{2}}\sigma}{\sqrt{C_\gamma}\mu^{3/2}N_t^{1-\alpha/2}},$$

using $\exp(-\mu C_\gamma C_\rho^\beta t^{1-\alpha}/2^{2-\alpha}) = A_t^c \leq t^{-1}\sum_{i=1}^{t} A_i^c \leq t^{-1}A_\infty^c$ as A_t^c is decreasing. In a same way, one has

$$\frac{C_f C_\rho^{\frac{1}{2}}}{N_t\mu}\left(\sum_{i=1}^{t-1}\delta_i\right)^{\frac{1}{2}} \leq \frac{C_f C_\rho^{\frac{1}{2}}}{N_t\mu}\left(A_\infty^c\pi_\infty^c + \frac{2^{1+\alpha}\sigma^2 C_\gamma t^{1-\alpha}}{(1-\alpha)\,\mu C_\rho^{1-\beta}}\right)^{1/2} \leq \frac{C_f C_\rho^{\frac{1}{2}}\sqrt{\pi_\infty^c}\sqrt{A_\infty^c}}{N_t\mu} + \frac{2^{\frac{1+\alpha}{2}}C_f\sigma\sqrt{C_\gamma}}{C_\rho^{\frac{1-\alpha-\beta}{2}}\mu^{3/2}N_t^{\frac{1+\alpha}{2}}}.$$

Bound the last term of (A.24), is done as follows,

$$\begin{aligned}
\frac{C'_\nabla C_\rho}{N_t\mu}\sum_{i=0}^{t-1}\Delta_i^{\frac{1}{2}} \leq& \frac{C'_\nabla C_\rho}{N_t\mu}\sum_{i=0}^{t-1}\exp\left(-\frac{\mu C_\gamma C_\rho^\beta i^{1-\alpha}}{2^{2-\alpha}}\right)\sqrt{\Pi_\infty^c} + \frac{2^\alpha 6C'_\nabla\sigma^2 C_\gamma C_\rho^\beta}{N_t\mu^2}\sum_{i=1}^{t-1} i^{-\alpha} \\
&+ \frac{(6+7\mathbb{1}_{\{C_\rho>1\}})2^{3\alpha/2}C'_\nabla\sigma^2 C_\gamma^{3/2}C_\rho^{3\beta/2}}{N_t\mu^{3/2}}\sum_{i=1}^{t-1} i^{-3\alpha/2} \\
\leq& \frac{C'_\nabla C_\rho\sqrt{\Pi_\infty^c}A_\infty^c}{N_t\mu} + \frac{2^\alpha 6C'_\nabla\sigma^2 C_\gamma}{C_\rho^{1-\alpha-\beta}\mu^2 N_t^\alpha} + \frac{(6+7\mathbb{1}_{\{C_\rho>1\}})2^{3\alpha/2}C'_\nabla\sigma^2 C_\gamma^{3/2}C_\rho^{3\beta/2}\psi_{3\alpha/2}^0(N_t/C_\rho)}{\mu^{3/2}N_t}.
\end{aligned}$$

Thus, by collecting the terms above, we obtain:

$$\begin{aligned}
\bar{\delta}_t^{1/2} \leq& \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{6\sigma C_\rho^{\frac{1-\alpha-\beta}{2}}}{\sqrt{C_\gamma}\mu^{3/2}N_t^{1-\alpha/2}} + \frac{2^\alpha 6C'_\nabla\sigma^2 C_\gamma}{C_\rho^{1-\alpha-\beta}\mu^2 N_t^\alpha} + \frac{C_\rho^{2-\alpha-\beta}\sqrt{\pi_\infty^c}A_\infty^c}{C_\gamma\mu N_t^{2-\alpha}} \\
&+ \frac{2^{\frac{1+\alpha}{2}}C_f\sigma\sqrt{C_\gamma}}{C_\rho^{\frac{1-\alpha-\beta}{2}}\mu^{3/2}N_t^{\frac{1+\alpha}{2}}} + \frac{C_\rho\Gamma_c}{\mu N_t} + \frac{(6+7\mathbb{1}_{\{C_\rho>1\}})2^{3\alpha/2}C'_\nabla\sigma^2 C_\gamma^{3/2}C_\rho^{3\beta/2}}{\mu^{3/2}\psi_{3\alpha/2}^0(N_t/C_\rho)^{-1}N_t},
\end{aligned}$$

where $\Gamma_c = (1/C_\gamma C_\rho^\beta + C_f)\delta_0^{1/2} + C_f\sqrt{\pi_\infty^c A_\infty^c}/C_\rho^{1/2} + \sqrt{\pi_\infty^c}A_\infty^c/C_\gamma C_\rho^\beta + C'_\nabla\sqrt{\Pi_\infty^c}A_\infty^c$. □

Proof of Corollary 4.11. The steps of the proof follows the ones of Corollary 4.10 with the smart notation of ϕ and $\tilde{\rho}$: The bound for $\bar{\delta}_t^{1/2}$ in (4.6) is given by

$$\frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{1}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{1/2}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| + \frac{n_t}{N_t\gamma_t\mu}\delta_t^{1/2} + \frac{n_1}{N_t\mu}\left(\frac{1}{\gamma_1} + C_f\right)\delta_0^{1/2} + \frac{C_f}{N_t\mu}\left(\sum_{i=1}^{t-1} n_{i+1}\delta_i\right)^{1/2} + \frac{C'_\nabla}{N_t\mu}\sum_{i=0}^{t-1} n_{i+1}\Delta_i^{1/2}, \tag{A.27}$$

where the learning rate and time-varying mini-batches are on the form $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ and $n_t = \lceil C_\rho t^\rho \rceil$. The second-order moment δ_t is upper bounded by (A.8) from Corollary 4.6. The fourth-order moment Δ_t from Lemma A.1 can be simplified as follows,

$$\Delta_t \leq \exp\left(-\mu\sum_{i=t/2}^{t}\gamma_i\right)\Pi_\infty^v + \frac{32\sigma^4}{\mu^2}\max_{t/2\leq i\leq t}\frac{\gamma_i^2}{n_i^2} + \frac{162\sigma^4}{\mu}\max_{t/2\leq i\leq t}\frac{\gamma_i^3}{n_i^2},$$

as $n_t \geq 1$ for any $t \geq 1$ and $\beta \leq 1$, and

$$\Pi_\infty^v = \exp\left(\frac{32(\alpha-\beta\tilde{\rho})C_\gamma^2 C_\rho^{2\beta}(2C_f^2 + C_\nabla^2)}{2(\alpha-\beta\tilde{\rho})-1}\right)\exp\left(192C_\gamma^4 C_\rho^{4\beta}(4C_f^4 + C_\nabla^4)\right)\left(\Delta_0 + \frac{2\sigma^4}{C_f^4} + \frac{4\sigma^4 C_\gamma}{\mu C_f^2 C_\rho^{1-\beta}},\right)$$

using that $\sum_{i=1}^t i^{-a} \leq 2$ for $a \geq 2$. Next, for $\rho \geq 0$, we have

$$\Delta_t \leq \exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1+\beta\rho-\alpha}}{2^{1+\beta\rho-\alpha}}\right)\Pi_\infty^v + \frac{2^{2\alpha-2\beta\rho+2\rho}32\sigma^4 C_\gamma^2 C_\rho^{2\beta}}{\mu^2 C_\rho^2 t^{2\alpha-2\beta\rho+2\rho}} + \frac{2^{3\alpha-3\beta\rho+2\rho}162\sigma^4 C_\gamma^3 C_\rho^{3\beta}}{\mu C_\rho^2 t^{3\alpha-3\beta\rho+2\rho}},$$

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

$$\Delta_t \leq \exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1-\alpha}}{2^{1-\alpha}}\right)\Pi_\infty^v + \frac{2^{2\alpha}32\sigma^4 C_\gamma^2 C_\rho^{2\beta}}{\mu^2 t^{2\alpha}} + \frac{2^{3\alpha}162\sigma^4 C_\gamma^3 C_\rho^{3\beta}}{\mu t^{3\alpha}}.$$

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

$$\Delta_t \leq \exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} t^{(1-\phi)(1+\tilde{\rho})}}{2^{(1-\phi)(1+\tilde{\rho})}}\right)\Pi_\infty^v + \frac{2^{2\phi(1+\tilde{\rho})}32\sigma^4 C_\gamma^2 C_\rho^{2\beta}}{\mu^2 C_\rho^{2\mathbb{1}_{\{\rho\geq 0\}}} t^{2\phi(1+\tilde{\rho})}} + \frac{2^{3\phi(1+\tilde{\rho})-\tilde{\rho}}162\sigma^4 C_\gamma^3 C_\rho^{3\beta}}{\mu C_\rho^{2\mathbb{1}_{\{\rho\geq 0\}}} t^{3\phi(1+\tilde{\rho})-\tilde{\rho}}}.$$

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 $t \geq (N_t/2C_\rho)^{1/(1+\rho)}$, *e.g.*, see the proof of Corollary 4.6. Similarly, $N_t \geq C_\rho\sum_{i=1}^t i^\rho \geq C_\rho\int_0^t x^\rho\,dx = C_\rho t^{\rho+1}$, thus, $t \leq (N_t/C_\rho)^{1/(1+\rho)}$. If $\rho < 0$, one has $t \leq N_t$ and $N_t \leq C_\rho t$, *i.e.*, $t \geq N_t/C_\rho$.

Bounding $\frac{1}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{1/2}|n_{i+1}/\gamma_{i+1} - n_i/\gamma_i|$, we first note that $n_t/\gamma_t = C_\gamma^{-1}\lceil C_\rho t^\rho\rceil^{1-\beta}t^\alpha$. Thus, by the mean value inequality, we obtain for $\rho \geq 0$:

$$\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| \leq \frac{2C_\rho^{1-\beta}}{C_\gamma}\sup_{\nu\in(i,i+1)}\left|\nu^{(1-\beta)\rho+\alpha-1}\right| \leq \frac{2C_\rho^{1-\beta}}{C_\gamma i^{1-(1-\beta)\rho-\alpha}}, \tag{A.28}$$

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

$$\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| \leq \frac{2C_\rho^{1-\beta}}{C_\gamma}\sup_{\nu\in(i,i+1)}\left|\nu^{\alpha-1}\right| \leq \frac{2C_\rho^{1-\beta}}{C_\gamma i^{1-\alpha}},$$

as $(n_t)_{t\geq 1}$ is a decreasing sequence and $\beta \leq 1$. Thus, for any $\rho \in (-1,1)$, we have

$$\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| \leq \frac{2C_\rho^{1-\beta}}{C_\gamma i^{1-\phi(1+\tilde{\rho})}}.$$

By using this, we obtain a bound on $\frac{1}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{\frac{1}{2}}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right|$ given as

$$\frac{2C_\rho^{1-\beta}}{N_t\mu C_\gamma}\sum_{i=1}^{t} i^{\phi(1+\tilde{\rho})-1}\left(\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} i^{(1-\phi)(1+\tilde{\rho})}}{2^{1+(1-\phi)(1+\tilde{\rho})}}\right)\sqrt{\pi_\infty^v} + \frac{2^{\frac{1+\phi(1+\tilde{\rho})}{2}}\sigma\sqrt{C_\gamma}}{\sqrt{\mu}C_\rho^{\frac{(1-\beta)}{2}\mathbb{1}_{\{\rho\geq 0\}}} i^{\frac{\phi(1+\tilde{\rho})}{2}}}\right).$$

Next, let us denote

$$A_\infty^v = \sum_{i=0}^{\infty} i^{\tilde{\rho}}\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} i^{(1-\phi)(1+\tilde{\rho})}}{2^{1+(1-\phi)(1+\tilde{\rho})}}\right) \geq \sum_{i=0}^{\infty} i^{\phi(1+\tilde{\rho})-1}\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} i^{(1-\phi)(1+\tilde{\rho})}}{2^{1+(1-\phi)(1+\tilde{\rho})}}\right),$$

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

$$\frac{2C_\rho^{1-\beta}\sqrt{\pi_\infty^v}}{N_t\mu C_\gamma}\sum_{i=1}^{t} i^{\phi(1+\tilde{\rho})-1}\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} i^{(1-\phi)(1+\tilde{\rho})}}{2^{1+(1-\phi)(1+\tilde{\rho})}}\right) \leq \frac{2C_\rho^{1-\beta}\sqrt{\pi_\infty^v}A_\infty^v}{N_t\mu C_\gamma}.$$

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

$$\frac{2^{\frac{3+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma}N_t}\sum_{i=1}^{t} i^{\frac{\phi(1+\tilde{\rho})}{2}-1} \leq \frac{2^{\frac{3+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}} t^{\frac{\phi(1+\tilde{\rho})}{2}}}{\mu^{3/2}\sqrt{C_\gamma}N_t} \leq \frac{2^{\frac{3+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma}N_t^{1-\phi/2}}.$$

Summarising, we obtain

$$\frac{1}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{\frac{1}{2}}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| \leq \frac{2C_\rho^{1-\beta}\sqrt{\pi_\infty^v}A_\infty^v}{N_t\mu C_\gamma} + \frac{2^{\frac{3+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma}N_t^{1-\phi/2}}.$$

Similarly, for $\frac{n_t}{N_t\gamma_t\mu}\delta_t^{1/2}$, one have

$$\begin{aligned}\frac{n_t}{N_t\gamma_t\mu}\delta_t^{\frac{1}{2}} &\leq \frac{C_\rho^{1-\beta}\sqrt{\pi_\infty^v}t^{\phi(1+\tilde{\rho})}}{N_tC_\gamma\mu}\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} t^{(1-\phi)(1+\tilde{\rho})}}{2^{1+(1-\phi)(1+\tilde{\rho})}}\right) + \frac{2^{\frac{1+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}} t^{\frac{\phi(1+\tilde{\rho})}{2}}}{\mu^{3/2}\sqrt{C_\gamma}N_t} \\ &\leq \frac{C_\rho^{2-\phi-\beta}\sqrt{\pi_\infty^v}A_\infty^v}{\mu C_\gamma N_t^{2-\phi}} + \frac{2^{\frac{1+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma}N_t^{1-\phi/2}}.\end{aligned}$$

For $\frac{n_1}{N_t\mu}(\gamma_1^{-1} + C_f)\delta_0^{1/2}$, we insert the definition of our learning functions, giving us

$$\frac{n_1}{N_t\mu}\left(\frac{1}{\gamma_1} + C_f\right)\delta_0^{1/2} = \frac{C_\rho}{N_t\mu}\left(\frac{1}{C_\gamma C_\rho^\beta} + C_f\right)\delta_0^{1/2}.$$

Bounding $\frac{C_f}{N_t\mu}(\sum_{i=1}^{t-1} n_{i+1}\delta_i)^{1/2}$, follows the ideas from above, using that $n_{t+1} \leq 2^{\tilde{\rho}} n_t$; it can be upper bounded by

$$
\begin{aligned}
&\frac{2^{\tilde{\rho}/2}C_f}{N_t\mu}\left(C_\rho\sum_{i=1}^{t} i^{\tilde{\rho}}\left(\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} i^{(1-\phi)(1+\tilde{\rho})}}{2^{(1-\phi)(1+\tilde{\rho})}}\right)\pi_\infty^v + \frac{2^{1+\phi(1+\tilde{\rho})}\sigma^2 C_\gamma}{\mu C_\rho^{(1-\beta)\mathbb{1}_{\{\rho\geq 0\}}} i^{\phi(1+\tilde{\rho})}}\right)\right)^{\frac{1}{2}} \\
=&\frac{2^{\tilde{\rho}/2}C_f}{N_t\mu}\left(C_\rho\pi_\infty^v\sum_{i=1}^{t} i^{\tilde{\rho}}\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} i^{(1-\phi)(1+\tilde{\rho})}}{2^{(1-\phi)(1+\tilde{\rho})}}\right) + \frac{2^{1+\phi(1+\tilde{\rho})}\sigma^2 C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}}}{\mu}\sum_{i=1}^{t} i^{\beta\tilde{\rho}-\alpha}\right)^{\frac{1}{2}} \\
\leq&\frac{2^{\tilde{\rho}/2}C_f}{N_t\mu}\left(C_\rho\pi_\infty^v A_\infty^v + \frac{2^{\phi(1+\tilde{\rho})}\sigma^2 C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} t^{(1-\phi)(1+\tilde{\rho})}}{\mu}\right)^{\frac{1}{2}} \\
\leq&\frac{2^{\tilde{\rho}/2}C_f\sqrt{C_\rho}\sqrt{\pi_\infty^v}\sqrt{A_\infty^v}}{\mu N_t} + \frac{2^{\frac{\phi(1+\tilde{\rho})}{2}} C_f\sigma\sqrt{C_\gamma} C_\rho^{\beta/2\mathbb{1}_{\{\rho\geq 0\}}} t^{\frac{(1-\phi)(1+\tilde{\rho})}{2}}}{\mu^{3/2}N_t} \\
\leq&\frac{2^{\tilde{\rho}/2}C_f\sqrt{C_\rho}\sqrt{\pi_\infty^v}\sqrt{A_\infty^v}}{\mu N_t} + \frac{2^{\frac{\phi(1+\tilde{\rho})}{2}} C_f\sigma\sqrt{C_\gamma}}{\mu^{3/2}C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}} N_t^{\frac{1+\phi}{2}}}.
\end{aligned}
$$

Likewise, for $\frac{C'_\nabla}{N_t\mu}\sum_{i=0}^{t-1} n_{i+1}\Delta_i^{1/2}$, we can bound by

$$
\begin{aligned}
&\frac{2^{\tilde{\rho}}C'_\nabla C_\rho}{N_t\mu}\sum_{i=1}^{t-1} i^{\tilde{\rho}}\left(\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} i^{(1-\phi)(1+\tilde{\rho})}}{2^{(1-\phi)(1+\tilde{\rho})}}\right)\Pi_\infty^v + \frac{2^{2\phi(1+\tilde{\rho})}32\sigma^4 C_\gamma^2 C_\rho^{2\beta}}{\mu^2 C_\rho^{2\mathbb{1}_{\{\rho\geq 0\}}} i^{2\phi(1+\tilde{\rho})}} + \frac{2^{3\phi(1+\tilde{\rho})-\tilde{\rho}}162\sigma^4 C_\gamma^3 C_\rho^{3\beta}}{\mu C_\rho^{2\mathbb{1}_{\{\rho\geq 0\}}} i^{3\phi(1+\tilde{\rho})-\tilde{\rho}}}\right)^{\frac{1}{2}} \\
\leq&\frac{2^{\tilde{\rho}}C'_\nabla C_\rho}{N_t\mu}\sum_{i=1}^{t-1} i^{\tilde{\rho}}\left(\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} i^{(1-\phi)(1+\tilde{\rho})}}{2^{1+(1-\phi)(1+\tilde{\rho})}}\right)\sqrt{\Pi_\infty^v} + \frac{2^{\phi(1+\tilde{\rho})}6\sigma^2 C_\gamma C_\rho^{\beta}}{\mu C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} i^{\phi(1+\tilde{\rho})}} + \frac{2^{3\phi(1+\tilde{\rho})/2-\tilde{\rho}/2}13\sigma^2 C_\gamma^{3/2} C_\rho^{3\beta/2}}{\mu^{1/2} C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} i^{3\phi(1+\tilde{\rho})/2}}\right) \\
\leq&\frac{2^{\tilde{\rho}}C'_\nabla C_\rho\sqrt{\Pi_\infty^v}A_\infty^v}{\mu N_t} + \frac{2^{\phi(1+\tilde{\rho})+\tilde{\rho}}C'_\nabla\sigma^2 C_\gamma C_\rho^{1+\beta}}{\mu^2 C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} N_t}\sum_{i=1}^{t-1} i^{\beta\tilde{\rho}-\alpha} + \frac{2^{3\phi(1+\tilde{\rho})/2+\tilde{\rho}/2}C'_\nabla\sigma^2 C_\gamma^{3/2} C_\rho^{1+3\beta/2}}{\mu^{3/2} C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} N_t}\sum_{i=1}^{t-1} i^{3(\beta\tilde{\rho}-\alpha)/2},
\end{aligned}
$$

where the second term can be bounded as

$$
\frac{2^{(1+\phi)(1+\tilde{\rho})-1}C'_\nabla\sigma^2 C_\gamma C_\rho^{1+\beta}}{\mu^2 C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} N_t}\sum_{i=1}^{t-1} i^{\beta\tilde{\rho}-\alpha} \leq \frac{2^{(1+\phi)(1+\tilde{\rho})-1}C'_\nabla\sigma^2 C_\gamma C_\rho^{1+\beta} t^{1+\beta\tilde{\rho}-\alpha}}{(1+\beta\tilde{\rho}-\alpha)\mu^2 C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} N_t} \leq \frac{2^{(1+\phi)(1+\tilde{\rho})-2}C'_\nabla\sigma^2 C_\gamma}{\mu^2 C_\rho^{1-\phi-\beta} N_t^\phi},
$$

and the third term by

$$
\frac{2^{3(1+\phi)(1+\tilde{\rho})/2}C'_\nabla\sigma^2 C_\gamma^{3/2} C_\rho^{1+3\beta/2}}{\mu^{3/2} C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} N_t}\sum_{i=1}^{t-1} i^{3(\beta\tilde{\rho}-\alpha)/2} \leq \frac{2^{3(1+\phi)(1+\tilde{\rho})/2}C'_\nabla\sigma^2 C_\gamma^{3/2} C_\rho^{1+3\beta/2}\psi_{3(\alpha-\beta\tilde{\rho})/2}^{\tilde{\rho}}(N_t/C_\rho)}{\mu^{3/2} C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} N_t}.
$$

By collecting these bounds, we get

$$
\frac{C'_\nabla}{N_t\mu}\sum_{i=0}^{t-1} n_{i+1}\Delta_i^{\frac{1}{2}} \leq \frac{2^{\tilde{\rho}}C'_\nabla C_\rho\sqrt{\Pi_\infty^v}A_\infty^v}{\mu N_t} + \frac{2^{(1+\phi)(1+\tilde{\rho})-2}C'_\nabla\sigma^2 C_\gamma}{\mu^2 C_\rho^{1-\phi-\beta} N_t^\phi} + \frac{2^{3(1+\phi)(1+\tilde{\rho})/2}C'_\nabla\sigma^2 C_\gamma^{3/2} C_\rho^{1+3\beta/2}\psi_{3(\alpha-\beta\tilde{\rho})/2}^{\tilde{\rho}}(N_t/C_\rho)}{\mu^{3/2} C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}} N_t}.
$$

Combining our findings from above, we have

$$
\bar{\delta}_t^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{2C_\rho^{1-\beta}\sqrt{\pi_\infty^v}A_\infty^v}{\mu C_\gamma N_t} + \frac{2^{\frac{3+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma} N_t^{1-\phi/2}} + \frac{C_\rho^{2-\phi-\beta}\sqrt{\pi_\infty^v}A_\infty^v}{\mu C_\gamma N_t^{2-\phi}} + \frac{2^{\frac{1+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma} N_t^{1-\phi/2}}
$$

$$+\frac{C_\rho}{N_t\mu}\left(\frac{1}{C_\gamma C_\rho^\beta}+C_f\right)\delta_0^{\frac{1}{2}}+\frac{2^{\tilde\rho/2}C_f\sqrt{C_\rho}\sqrt{\pi_\infty^v}\sqrt{A_\infty^v}}{\mu N_t}+\frac{2^{\frac{\phi(1+\tilde\rho)}{2}}C_f\sigma\sqrt{C_\gamma}}{\mu^{3/2}C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}N_t^{\frac{1+\phi}{2}}}+\frac{2^{\tilde\rho}C'_\nabla C_\rho\sqrt{\Pi_\infty^v}A_\infty^v}{\mu N_t}$$
$$+\frac{2^{(1+\phi)(1+\tilde\rho)-2}C'_\nabla\sigma^2C_\gamma}{\mu^2C_\rho^{1-\phi-\beta}N_t^\phi}+\frac{2^{3(1+\phi)(1+\tilde\rho)/2}C'_\nabla\sigma^2C_\gamma^{3/2}C_\rho^{1+3\beta/2}\psi^{\tilde\rho}_{3(\alpha-\beta\tilde\rho)/2}(N_t/C_\rho)}{\mu^{3/2}C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}}N_t}.$$

This can be simplified to the desired using Γ_v given by $(1/C_\gamma C_\rho^\beta+C_f)\delta_0^{1/2}+2^{\tilde\rho}C_f\sqrt{\pi_\infty^vA_\infty^v}/C_\rho^{1/2}+2\sqrt{\pi_\infty^v}A_\infty^v/C_\gamma C_\rho^\beta+2^{\tilde\rho}C'_\nabla\sqrt{\Pi_\infty^v}A_\infty^v$, consisting of the finite constants π_∞^v, Π_∞^v and A_∞^v. □

A.2.2 Proofs for Section 4.2.2

Theorem A.2 (APSSG)**.** *Let $\bar\delta_t=\mathbb{E}[\|\bar\theta_t-\theta^*\|^2]$ with $(\bar\theta_t)$ given by (3.3), where (θ_t) follows the recursion in (3.2). Suppose Assumptions 2.1 to 4.8 hold with $p=4$. Then, for any learning rate (γ_t) and time-varying mini-batch (n_t), we can upper bound $\bar\delta_t^{1/2}$ by*

$$\frac{\Lambda^{1/2}}{N_t^{1/2}}+\frac{1}{N_t\mu}\sum_{i=1}^{t-1}\left|\frac{n_{i+1}}{\gamma_{i+1}}-\frac{n_i}{\gamma_i}\right|\delta_i^{1/2}+\frac{n_t}{N_t\gamma_t\mu}\delta_t^{1/2}+\frac{n_1}{N_t\mu}\left(\frac{1}{\gamma_1}+C_f\right)\delta_0^{1/2}+\frac{C_f}{N_t\mu}\left(\sum_{i=1}^{t-1}n_{i+1}\delta_i\right)^{1/2}+\frac{C''_\nabla}{N_t\mu}\sum_{i=0}^{t}n_{i+1}\Delta_i^{1/2}$$

where $\Lambda=\mathrm{Tr}(\nabla^2_\theta F(\theta^*)^{-1}\Sigma\nabla^2_\theta F(\theta^*)^{-1})$ and $C''_\nabla=C'_\nabla+2^2G_\Theta/D_\theta^2$.

Proof of Theorem A.2. Denote $\mathbb{E}[\|\bar\theta_t-\theta^*\|^2]$ by $\bar\delta_t$ with $(\bar\theta_t)$ given by (3.3) using (θ_t) from (3.2). As in the proof Theorem 4.9, we follow the steps of [32], in which, we can rewrite (3.2) to

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

where $\nabla_\theta f_t(\theta_{t-1})=n_t^{-1}\sum_{i=1}^{n_t}\nabla_\theta f_{t,i}(\theta_{t-1})$ and $\Omega_t=\mathcal{P}_\Theta(\theta_{t-1}-\gamma_t\nabla_\theta f_t(\theta_{t-1}))-(\theta_{t-1}-\gamma_t\nabla_\theta f_t(\theta_{t-1}))$. Thus, summing the parts, using the Minkowski's inequality, and bounding each term gives us the same bound as in Theorem 4.9, but with an additional term regarding Ω_t, namely

$$\left(\mathbb{E}\left[\left\|\nabla^2_\theta F(\theta^*)^{-1}\frac{1}{N_t}\sum_{i=1}^{t}\frac{n_i}{\gamma_i}\Omega_i\right\|^2\right]\right)^{\frac{1}{2}}\leq\frac{1}{\mu N_t}\sum_{i=1}^{t}\frac{n_i}{\gamma_i}\sqrt{\mathbb{E}\left[\|\Omega_i\|^2\right]}=\frac{1}{\mu N_t}\sum_{i=1}^{t}\frac{n_i}{\gamma_i}\sqrt{\mathbb{E}\left[\|\Omega_i\|^2\mathbb{1}_{\{\theta_{i-1}-\gamma_i\nabla_\theta f_i(\theta_{i-1})\notin\Theta\}}\right]},\tag{A.29}$$

using Lemma 4.3 of [12]. Next, we note that $\mathbb{E}[\|\Omega_t\|^2\mathbb{1}_{\{\theta_{t-1}-\gamma_t\nabla_\theta f_t(\theta_{t-1})\notin\Theta\}}]=4\gamma_t^2G_\Theta^2\mathbb{P}[\theta_{t-1}-\gamma_t\nabla_\theta f_t(\theta_{t-1})\notin\Theta]$, since

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

as $\mathcal{P}_\Theta$ is Lipschitz and $\|\nabla_\theta f_{t,i}(\theta)\|^2\leq G_\Theta^2$ for any $\theta\in\Theta$. Moreover, as in Theorem 4.2 of [15], we know that $\mathbb{P}[\theta_{t-1}-\gamma_t\nabla_\theta f_t(\theta_{t-1})\notin\Theta]\leq\Delta_t/D_\theta^4$, where $D_\theta=\inf_{\theta\in\partial\Theta}\|\theta-\theta^*\|$ with $\partial\Theta$ denoting the frontier of Θ. Thus,

(A.29) can then be bounded by

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

using that the sequence (n_t) is either constant or time-varying, meaning $n_{t+1}/n_t \leq 2$. □

Proof of Corollary 4.13. The proof follows directly from Corollary 4.10 with use of Theorem A.2. □

Proof of Corollary 4.14. The proof follows directly from Corollary 4.11 with use of Theorem A.2. □

Appendix B. Technical propositions

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

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

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

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

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

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

$$\begin{aligned}\frac{1}{\omega}\sum_{i=k}^{t}\left[\prod_{j=i}^{t}\left[1+\omega\gamma_j\right]-\prod_{j=i+1}^{t}\left[1+\omega\gamma_j\right]\right] =& \frac{1}{\omega}\left[\prod_{j=k}^{t}\left[1+\omega\gamma_j\right]-\prod_{j=k+1}^{t}\left[1+\omega\gamma_j\right]+\cdots-\prod_{j=t+1}^{t}\left[1+\omega\gamma_j\right]\right]\\ =& \frac{1}{\omega}\left[\prod_{j=k}^{t}\left[1+\omega\gamma_j\right]-\prod_{j=t+1}^{t}\left[1+\omega\gamma_j\right]\right]\\ =& \frac{1}{\omega}\left[\prod_{j=k}^{t}\left[1+\omega\gamma_j\right]-1\right] \leq \frac{1}{\omega}\prod_{j=k}^{t}\left[1+\omega\gamma_j\right],\end{aligned}$$

as $\prod_{t+1}^{t} = 1$ 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 (B.1):

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

□

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

$$\sum_{i=k}^{t}\prod_{j=i+1}^{t}\left[1-\omega\gamma_j\right]\gamma_i \leq \frac{1}{\omega}. \tag{B.2}$$

Proof of Proposition B.2. We start with expanding the sums of products term in (B.2), given us

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

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

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

using $\prod_{j=k}^{t}[1-\omega\gamma_j] \geq 0$, showing the inequality in (B.2). □

Proposition B.3. *Let $(\gamma_t)_{t\geq 1}$ and $(\eta_t)_{t\geq 1}$ be positive sequences. For any $k \leq t$, we can obtain the (upper) bounds:*

$$\sum_{i=k}^{t}\prod_{j=i+1}^{t}\left[1+\omega\gamma_j\right]\eta_i\gamma_i \leq \frac{1}{\omega}\max_{k\leq i\leq t}\eta_i\exp\left(\omega\sum_{j=k}^{t}\gamma_j\right), \tag{B.3}$$

with $\omega > 0$. Furthermore, suppose that for all $i \geq k$, $\omega\gamma_i \leq 1$, then

$$\sum_{i=k}^{t}\prod_{j=i+1}^{t}\left[1-\omega\gamma_j\right]\eta_i \leq \frac{1}{\omega}\max_{k\leq i\leq t}\eta_i. \tag{B.4}$$

Proof of Proposition B.3. We obtain the inequality in (B.3) directly by Proposition B.1:

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

Similarly, for the inequality in (B.4), we have

$$\sum_{i=k}^{t}\prod_{j=i+1}^{t}\left[1-\omega\gamma_j\right]\eta_i\gamma_i \leq \max_{k\leq i\leq t}\eta_i\sum_{i=k}^{t}\prod_{j=i+1}^{t}\left[1-\omega\gamma_j\right]\gamma_i \leq \frac{1}{\omega}\max_{k\leq i\leq t}\eta_i,$$

by Proposition B.2. □

Proposition B.4. *Let $(\delta_t)_{t\geq 0}$, $(\gamma_t)_{t\geq 1}$, $(\eta_t)_{t\geq 1}$, and $(\nu_t)_{t\geq 1}$ be some positive sequences satisfying the recursive relation:*

$$\delta_t \leq \left(1-2\omega\gamma_t+\eta_t\gamma_t\right)\delta_{t-1}+\nu_t\gamma_t, \tag{B.5}$$

with $\delta_0 \geq 0$ and $\omega > 0$. Denote $t_0 = \inf\{t \geq 1 : \eta_t \leq \omega\}$, and suppose that for all $t \geq t_0 + 1$, one has $\omega\gamma_t \leq 1$. Then, for γ_t and η_t decreasing, we have the upper bound on (δ_t):

$$\delta_t \leq \exp\left(-\omega \sum_{i=t/2}^{t} \gamma_i\right)\left[\exp\left(\sum_{i=1}^{t_0} \eta_i\gamma_i\right)\left(\delta_0 + \max_{1\leq i\leq t_0} \frac{\nu_i}{\eta_i}\right) + \sum_{i=t_0+1}^{t/2-1} \nu_i\gamma_i\right] + \frac{1}{\omega} \max_{t/2\leq i\leq t} \nu_i, \tag{B.6}$$

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

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

$$\delta_t \leq \underbrace{\prod_{i=1}^{t} [1 - 2\omega\gamma_i + \eta_i\gamma_i]}_{B_t} \delta_0 + \underbrace{\sum_{i=1}^{t} \prod_{j=i+1}^{t} [1 - 2\omega\gamma_j + \eta_j\gamma_j]\, \nu_i\gamma_i}_{A_t},$$

where B_t can be seen as a transient term only depending on the initialisation δ_0, and a stationary term A_t. The transient term B_t can be divided into two products, before and after t_0,

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

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

$$\begin{aligned} B_t &\leq \left(\prod_{i=1}^{t_0} [1 - 2\omega\gamma_i + \eta_i\gamma_i]\right)\left(\prod_{i=t_0+1}^{t} [1 - \omega\gamma_i]\right) \leq \left(\prod_{i=1}^{t_0} \exp\left(-2\omega\gamma_i + \eta_i\gamma_i\right)\right)\left(\prod_{i=t_0+1}^{t} \exp\left(-\omega\gamma_i\right)\right) \\ &= \exp\left(-2\omega \sum_{i=1}^{t_0} \gamma_i\right) \exp\left(\sum_{i=1}^{t_0} \eta_i\gamma_i\right) \exp\left(-\omega \sum_{i=t_0+1}^{t} \gamma_i\right) \leq \exp\left(-\omega \sum_{i=1}^{t} \gamma_i\right) \exp\left(\sum_{i=1}^{t_0} \eta_i\gamma_i\right) \end{aligned}$$

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

$$B_t \leq \exp\left(-\omega \sum_{i=t/2}^{t} \gamma_i\right) \exp\left(\sum_{i=1}^{t_0} \eta_i\gamma_i\right). \tag{B.7}$$

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

$$A_t = \underbrace{\sum_{i=t_0+1}^{t} \prod_{j=i+1}^{t} [1 - 2\omega\gamma_j + \eta_j\gamma_j]\, \nu_i\gamma_i}_{A_{t,1}} + \underbrace{\sum_{i=1}^{t_0} \prod_{j=i+1}^{t} [1 - 2\omega\gamma_j + \eta_j\gamma_j]\, \nu_i\gamma_i}_{A_{t,2}}.$$

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

$$A_{t,1} \leq \max_{t_0+1\leq i\leq t} \nu_i \sum_{i=t_0+1}^{t} \prod_{j=i+1}^{t} [1 - \omega\gamma_j]\, \gamma_i = \frac{1}{\omega} \max_{t_0+1\leq i\leq t} \nu_i \leq \frac{1}{\omega} \max_{t/2\leq i\leq t} \nu_i,$$

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

$$A_{t,1} \leq \sum_{i=t_0+1}^{t} \prod_{j=i+1}^{t} [1-\omega\gamma_j]\,\nu_i\gamma_i = \sum_{i=t_0+1}^{t/2-1} \prod_{j=i+1}^{t} [1-\omega\gamma_j]\,\nu_i\gamma_i + \sum_{i=t/2}^{t} \prod_{j=i+1}^{t} [1-\omega\gamma_j]\,\nu_i\gamma_i$$
$$\leq \sum_{i=t_0+1}^{t/2-1} \prod_{j=t/2}^{t} [1-\omega\gamma_j]\,\nu_i\gamma_i + \max_{t/2\leq i\leq t} \nu_i \sum_{i=t/2}^{t} \prod_{j=i+1}^{t} [1-\omega\gamma_j]\,\gamma_i = \prod_{j=t/2}^{t} [1-\omega\gamma_j] \sum_{i=t_0+1}^{t/2-1} \nu_i\gamma_i + \frac{1}{\omega} \max_{t/2\leq i\leq t} \nu_i,$$

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

$$A_{t,1} \leq \exp\left(-\omega \sum_{j=t/2}^{t} \gamma_j\right) \sum_{i=t_0+1}^{t/2-1} \nu_i\gamma_i + \frac{1}{\omega} \max_{t/2\leq i\leq t} \nu_i, \tag{B.8}$$

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

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

by the definition of t_0, thus

$$A_{t,2} \leq \exp\left(-\omega \sum_{j=1}^{t} \gamma_j\right) \max_{1\leq i\leq t_0} \frac{\nu_i}{\eta_i} \exp\left(2\sum_{i=1}^{t_0} \eta_i\gamma_i\right) \leq \exp\left(-\omega \sum_{j=t/2}^{t} \gamma_j\right) \max_{1\leq i\leq t_0} \frac{\nu_i}{\eta_i} \exp\left(2\sum_{i=1}^{t_0} \eta_j\gamma_j\right). \tag{B.9}$$

Then, using the bound for $A_{t,1}$ in (B.8) and $A_{t,2}$ in (B.9), we can bound A_t by

$$A_t \leq \exp\left(-\omega \sum_{j=t/2}^{t} \gamma_j\right) \left[\exp\left(2\sum_{i=1}^{t_0} \eta_j\gamma_j\right) \max_{1\leq i\leq t_0} \frac{\nu_i}{\eta_i} + \sum_{i=t_0+1}^{t/2-1} \nu_i\gamma_i\right] + \frac{1}{\omega} \max_{t/2\leq i\leq t} \nu_i. \tag{B.10}$$

Finally, combining the bound for B_t in (B.7) and A_t in (B.10), we achieve the bound for $\delta_t \leq B_t\delta_0 + A_t$, namely the upper bound in (B.6). □

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

Proposition B.5. *Let $(\delta_t)_{t\geq 0}$, $(\gamma_t)_{t\geq 1}$, $(\eta_t)_{t\geq 1}$, and $(\nu_t)_{t\geq 1}$ be some positive sequences satisfying the recursive relation in (B.5). Denote $t_0 = \inf\{t \geq 1 : \eta_t \leq \omega\}$, and suppose that for all $t \geq t_0 + 1$, one has $\omega\gamma_t \leq 1$. Then,*

for γ_t and η_t decreasing, we have for all $t \in \mathbb{N}$,

$$\delta_t \leq \exp\left(-\omega \sum_{i=t/2}^{t} \gamma_i\right) \exp\left(2 \sum_{i=1}^{t} \eta_i \gamma_i\right) \left(\delta_0 + 2 \max_{1 \leq i \leq t} \frac{\nu_i}{\eta_i}\right) + \frac{1}{\omega} \max_{t/2 \leq i \leq t} \nu_i. \tag{B.11}$$

Proof of Proposition B.5. The resulting (upper) bound in (B.11) follows directly from (B.6) by noting that $t_0 \leq t$, giving us $\sum_{i=t_0+1}^{t/2-1} \nu_i \gamma_i \leq \sum_{i=1}^{t} \nu_i \gamma_i \leq \max_{1 \leq i \leq t}(\nu_i/\eta_i) \sum_{i=1}^{t} \eta_i \gamma_i \leq \max_{1 \leq i \leq t}(\nu_i/\eta_i) \exp(2 \sum_{i=1}^{t} \eta_i \gamma_i)$, as (ν_t) and (γ_t) are positive sequences. □

Acknowledgements. The authors thank the anonymous reviewer for the valuable and helpful comments.

References

[1] F. Bach and E. Moulines, Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Adv. Neural Inf. Process. Syst.* **24** (2011).

[2] F. Bach and E. Moulines, Non-strongly-convex smooth stochastic approximation with convergence rate O (1/n). *Adv. Neural Inf. Process. Syst.* **26** (2013).

[3] A. Benveniste, M. Métivier and P. Priouret, vol. 22 of *Adaptive algorithms and stochastic approximations*. Springer Science & Business Media (2012).

[4] L. Bottou, F.E. Curtis and J. Nocedal, Optimization methods for large-scale machine learning. *Siam Rev.* **60** (2018) 223–311.

[5] C. Boyer and A. Godichon-Baggioni, On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. *Comput. Optim. Appl.* (2022) 1–52.

[6] H. Cardot, P. Cénac and A. Godichon-Baggioni, Online estimation of the geometric median in Hilbert spaces: nonasymptotic confidence balls. *Ann. Stat.* (2017) 591–614.

[7] H. Cardot, P. Cénac and J.-M. Monnez, A fast and recursive algorithm for clustering large datasets with k-medians. *Comput. Stat. Data Anal.* **56** (2012) 1434–1449.

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

[9] A. d'Aspremont, Smooth optimization with approximate gradient. *SIAM J. Optim.* **19** (2008) 1171–1183.

[10] S. Gadat and F. Panloup, Optimal non-asymptotic analysis of the Ruppert–Polyak averaging stochastic algorithm. *Stoch. Process. Appl.* **156** (2023) 312–348.

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

[12] A. Godichon-Baggioni, Estimating the geometric median in Hilbert spaces with stochastic gradient algorithms: Lp and almost sure rates of convergence. *J. Multivariate Anal.* **146** (2016) 209–222.

[13] A. Godichon-Baggioni, Lp and almost sure rates of convergence of averaged stochastic gradient algorithms: locally strongly convex objective. *ESAIM: Probab. Stat.* **23** (2019) 841–873.

[14] A. Godichon-Baggioni, Convergence in quadratic mean of averaged stochastic gradient algorithms without strong convexity nor bounded gradient. Preprint arXiv:2107.12058 (2021).

[15] A. Godichon-Baggioni and B. Portier, An averaged projected Robbins-Monro algorithm for estimating the parameters of a truncated spherical distribution. *Electr. J. Stat.* **11** (2017) 1890–1927.

[16] R.M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin and P. Richtárik, SGD: General analysis and improved rates, in International conference on machine learning, PMLR (2019) 5200–5209.

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

[18] T. Hastie, R. Tibshirani, J.H. Friedman and J.H. Friedman, The elements of statistical learning: data mining, inference, and prediction, vol. 2. Springer (2009).

[19] H. Karimi, J. Nutini and M. Schmidt, Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition, in Joint European conference on machine learning and knowledge discovery in databases. Springer (2016) 795–811.

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

[21] K. Kurdyka, On gradients of functions definable in o-minimal structures. *Ann. l'institut Fourier* **48** (1998) 769–783.

[22] H. Kushner and G.G. Yin, vol. 35 of Stochastic approximation and recursive algorithms and applications. Springer Science & Business Media (2003).

[23] G. Lan, First-order and stochastic optimization methods for machine learning. Springer (2020).

[24] Y. LeCun, Y. Bengio and G. Hinton, Deep learning. *Nature* **521** (2015) 436–444.

[25] S. Lojasiewicz, A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles* **117** (1963) 2.

[26] A. Mokkadem and M. Pelletier, A generalization of the averaging procedure: the use of two-time-scale algorithms. *SIAM J. Control Optim.* **49** (2011) 1523–1543.
[27] N. Murata and S.-i. Amari, Statistical analysis of learning dynamics. *Signal Process.* **74** (1999) 3–28.
[28] I. Necoara, Y. Nesterov and F. Glineur, Linear convergence of first order methods for non-strongly convex optimization. *Math. Program.* **175** (2019) 69–107.
[29] A. Nemirovski, A. Juditsky, G. Lan and A. Shapiro, Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* **19** (2009) 1574–1609.
[30] Y. Nesterov *et al.*, Lectures on convex optimization, vol. 137. Springer (2018).
[31] B.T. Polyak, Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki* **3** (1963) 643–653.
[32] B.T. Polyak and A.B. Juditsky, Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **30** (1992) 838–855.
[33] H. Robbins and S. Monro, A stochastic approximation method. *Ann. Math. Stat.* (1951) 400–407.
[34] D. Ruppert, Efficient estimations from a slowly convergent Robbins-Monro process. Tech. rep., Cornell University Operations Research and Industrial Engineering (1988).
[35] M. Schmidt, N. Roux and F. Bach, Convergence rates of inexact proximal-gradient methods for convex optimization. *Adv. Neural Inf. Process. Syst.* **24** (2011) 1458–1466.
[36] S. Shalev-Shwartz *et al.*, Online learning and online convex optimization. *Found. Trends Mach. Learn.* **4** (2012) 107–194.
[37] I. Steinwart and A. Christmann, Estimating conditional quantiles with the help of the pinball loss. *Bernoulli* **17** (2011) 211–225.
[38] C.H. Teo, A. Smola, S. Vishwanathan and Q.V. Le, A scalable modular convex solver for regularized risk minimization, in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (2007) 727–736.
[39] N. Werge and O. Wintenberger, AdaVol: An adaptive recursive volatility prediction method. *Econometr. Stat.* **23** (2022) 19–35.
[40] M. Zinkevich, Online convex programming and generalized infinitesimal gradient ascent, in Proceedings of the 20th International Conference on Machine Learning (ICML-03) (2003) 928–936.