Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Streaming Data

Antoine Godichon-Baggioni, Nicklas Werge, Olivier Wintenberger

Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Streaming Data

Antoine Godichon-Baggioni[a], Nicklas Werge[a], Olivier Wintenberger[a]

[a]*LPSM, Sorbonne Université, 4 place Jussieu, 75005 Paris, France*

Abstract

Motivated by the high-frequency data streams continuously generated, real-time learning is becoming increasingly important. These data streams should be processed sequentially with the property that the stream may change over time. In this streaming setting, we propose techniques for minimizing a convex objective through unbiased estimates of its gradients, commonly referred to as stochastic approximation problems. Our methods rely on stochastic approximation algorithms due to their computationally advantage as they only use the previous iterate as a parameter estimate. The reasoning includes iterate averaging that guarantees optimal statistical efficiency under classical conditions. Our non-asymptotic analysis shows accelerated convergence by selecting the learning rate according to the expected data streams. We show that the average estimate converges optimally and robustly to any data stream rate. In addition, noise reduction can be achieved by processing the data in a specific pattern, which is advantageous for large-scale machine learning. These theoretical results are illustrated for various data streams, showing the effectiveness of the proposed algorithms.

Keywords: machine learning, large-scale, stochastic approximation, stochastic optimization, streaming data

1. Introduction

Machine learning and artificial intelligence have become an integral part of modern society. This massive utilization of intelligent systems generates an endless sequence of data, many of which come as *streaming data* such as internet traffic data, financial investments, self-driving cars, or sensor data. It requires robust and time-efficient algorithms to analyze and process such data without compromising on accuracy. This problem has attracted a lot of attention in the machine learning community (Bottou and LeCun, 2004; Zhang, 2004; Bottou and Bousquet, 2008; Xiao, 2010; Shalev-Shwartz et al., 2011).
Even after over 70 years, stochastic approximation algorithms are still the most popular method for handling large amounts of data (Robbins and Monro, 1951); the most well-known is presumably the stochastic gradient (SG) method, which has led to many extensions (Rumelhart et al., 1986; Duchi et al., 2011; Tieleman et al., 2012; Zeiler, 2012; Kingma and Ba, 2014; Dozat, 2016; Reddi et al., 2018). One essential extension is the Polyak-Ruppert averaging (ASG) proposed by Polyak and Juditsky (1992); Ruppert (1988), which guarantees optimal statistical efficiency without jeopardizing the computational cost. Bottou et al. (2018) reviews stochastic approximation methods for large-scale machine learning, including noise reduction methods and second-order methods, among others.

A fundamental aspect of our streaming setting is to consider the *streams* to which the data arrives, meaning we are concerned about how the data stream may evolve. We examine two different streams for the data: constant and varying *streaming-batches*. These streaming-batch streams mean we are considering everything from vanilla SG and ASG, mini-batch SG and ASG to more exotic learning designs.
Our main theoretical contribution is deriving explicit upper-bounds of the modified SG and ASG in a non-asymptotic

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

way. This contribution is a direct extension of Bach and Moulines (2011) to a streaming setting. These boundaries are derived by thoroughly analyzing the convex objective using unbiased estimates of its gradients. Our non-asymptotic analysis in this streaming setting is presented in Section 2 for the (modified) SG method, followed by the (modified) ASG method in Section 3. The theoretical results are illustrated in Section 4 for a variety of data streams. Our findings and experiments show a noticeable improvement in the convergence rates by selecting the hyper-parameters (in the *learning rate*) according to the expected data streams. In particular, we show how to obtain *optimal* convergence rates *robust* to any data streaming rate.

1.1. Strongly Convex Objectives

We consider minimizing convex functions $L : \mathbb{R}^d \to \mathbb{R}$ with $d \geq 1$, given by $L(\theta) = \mathbb{E}[l_t(\theta)]$, where $\theta \in \mathbb{R}^d$ is the predictor and $l_t : \mathbb{R}^d \to \mathbb{R}$ some random functions, e.g., see Kushner (2010) for a historical survey. Let $(l_t)_{t\geq 1}$ be differentiable (possibly non-convex), and their gradients unbiased estimates of the gradient of L, e.g, see Nesterov (2018) for definitions and properties of such functions. The principles for biased (non-random) functions are rather different; we refer to d'Aspremont (2008); Schmidt et al. (2011) for this.

Let $\mathbb{R}^d$ be equipped with inner product $\langle \cdot, \cdot \rangle$ and denote $\|\cdot\|$ the associated norm and the operator norm on bounded linear operators from $\mathbb{R}^d$ to $\mathbb{R}^d$, defined by $\|A\| = \sup_{\|x\|\leq 1} \|Ax\|$. Following Sridharan et al. (2008), we make the following assumptions on L: the function $L : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex for some $\mu > 0$, that is, for all $\theta, \theta' \in \mathbb{R}^d$ the following inequality holds,

$$L(\theta) \geq L(\theta') + \langle \nabla_\theta L(\theta'), \theta - \theta' \rangle + \frac{\mu}{2} \left\|\theta - \theta'\right\|^2. \tag{1}$$

Teo et al. (2007) provides a comprehensive record of various convex loss functions L used in machine learning applications. The *true* optimizer of L is denoted by θ^*, defined as $\theta^* = \arg\min_{\theta\in\mathbb{R}^d} L(\theta)$. Next, the function $\nabla_\theta L$ is C_∇-Lipschitz continuous, i.e., there is a constant $C_\nabla > 0$ such that for all $\theta, \theta' \in \mathbb{R}^d$,

$$\left\|\nabla_\theta L(\theta) - \nabla_\theta L(\theta')\right\|^2 \leq C_\nabla^2 \left\|\theta - \theta'\right\|^2. \tag{2}$$

1.2. Problem Formulation

We now outline our streaming setting in which we want to minimize L: at each time $t \in \mathbb{N}$, we consider $n_t \in \mathbb{N}$ random functions $l_t = (l_{t,1}, \ldots, l_{t,n_t})$. One can think of these random functions $(l_{t,i})$ as loss functions depending on the true minimizer θ^* and some noise sequences. The accumulated sum is denoted by $N_t = \sum_{i=1}^t n_i$. The *stochastic streaming gradient* (SSG) is defined as

$$\theta_t = \theta_{t-1} - \frac{\gamma_t}{n_t} \sum_{i=1}^{n_t} \nabla_\theta l_{t,i}(\theta_{t-1}), \tag{3}$$

where $(\gamma_t)_{t\geq 1}$ is a decreasing sequence of positive numbers also referred to as the *learning rate* satisfying $\sum_{i=1}^t \gamma_i = \infty$ and $\sum_{i=1}^t \gamma_i^2 < \infty$ for $t \to \infty$. To ease the notation, we let $\nabla_\theta l_t(\theta) = n_t^{-1} \sum_{i=1}^{n_t} \nabla_\theta l_{t,i}(\theta)$ for all $\theta \in \mathbb{R}^d$. To guarantee optimal convergence, we introduce the *averaged stochastic streaming gradient* (ASSG), defined for all $t \in \mathbb{N}$ by

$$\bar{\theta}_t = \frac{1}{N_t} \sum_{i=0}^{t-1} n_{i+1} \theta_i, \tag{4}$$

with $\bar{\theta}_0 = 0$. As we handle data sequentially, we will make use of the rewritten formula $\bar{\theta}_t = (N_{t-1}/N_t)\bar{\theta}_{t-1} + (n_t/N_t)\theta_{t-1}$.

2. Stochastic Streaming Gradient

In this section, we consider SSG with streaming-batches arriving in constant and varying streams. Let $(\mathcal{F}_t)_{t\geq 0}$ be an increasing family of σ-fields, namely $\mathcal{F}_t = \sigma(l_1, \ldots, l_t)$. Furthermore, we expand this notation with $\mathcal{F}_{t-1,i} = \sigma(l_{1,1}, \ldots, l_{t-1,n_{t-1}}, l_{t,1}, \ldots, l_{t,i})$ with $\mathcal{F}_{t-1,0} = \mathcal{F}_{t-1}$. Meaning, for all $0 \leq i < j$, we have $\mathcal{F}_{t-1} \subseteq \mathcal{F}_{t-1,i} \subset \mathcal{F}_{t-1,j}$. Our aim is to provide a bound on the expectation $\mathbb{E}[\|\theta_t - \theta^*\|^2]$, which depends explicitly upon the problem's parameters. In order to do this, we assume the following about $(l_t)_{t\geq 1}$ functions:

Assumption A1. Let θ_0 be $\mathcal{F}_0$-measurable. For each $t \geq 1$, the random variable $\nabla_\theta l_{t,i}(\theta)$ is square-integrable, $\mathcal{F}_{t,i}$-measurable, and $\mathbb{E}[\nabla_\theta l_{t,i}(\theta)|\mathcal{F}_{t-1,i-1}] = \nabla_\theta L(\theta)$ for all $\theta \in \mathbb{R}^d$ and $i = 1, \ldots, n_t$.

Assumption A2. For each $t \geq 1$, the function $l_{t,i}$ is differentiable, and there exists $C_l > 0$ such that $\mathbb{E}[\|\nabla_\theta l_{t,i}(\theta) - \nabla_\theta l_{t,i}(\theta')\|^2|\mathcal{F}_{t-1}] \leq C_l^2 \|\theta - \theta'\|^2$ a.s. for all $\theta, \theta' \in \mathbb{R}^d$ and $i = 1, \ldots, n_t$.

Assumption A3. For each $t \geq 1$, there exists $\sigma^2 > 0$ such that $\mathbb{E}[\|\nabla_\theta l_{t,i}(\theta^*)\|^2|\mathcal{F}_{t-1}] \leq \sigma^2$ a.s. for $i = 1, \ldots, n_t$.

These assumptions are a modified version (as they hold for any $i = 1, \ldots, n_t$) of the standard assumptions for stochastic approximations, e.g., see Kushner and Yin (2003); Bach and Moulines (2011). They include classic examples such as stochastic approximation and learning from i.i.d. data (e.g., linear regression and ridge regressions) under regularity conditions.

In the following theorem, we derive an explicit (non-asymptotic) upper bound on the t-th estimate of (3) for any decreasing step sequence $(\gamma_t)_{t\geq 1}$ using classical techniques from stochastic approximations (Benveniste et al., 1990; Kushner and Yin, 2003).

Theorem 1. *Denote $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ with some $\delta_0 \geq 0$. Under Assumption (A1,A2,A3), we have for any decreasing step sequence $(\gamma_t)_{t\geq 1}$ that*

$$\delta_t \leq \exp\left(-\mu \sum_{i=t/2}^{t} \gamma_i\right) \exp\left(4C_l^2 \sum_{i=1}^{t} \frac{\gamma_i^2}{n_i}\right) \exp\left(2C_\nabla^2 \sum_{i=1}^{t} \mathbb{1}_{\{n_i>1\}} \gamma_i^2\right) \left(\delta_0 + \frac{2\sigma^2}{C_l^2}\right) + \frac{2\sigma^2}{\mu} \max_{t/2\leq i\leq t} \frac{\gamma_i}{n_i}, \tag{5}$$

where $n_t \geq 1$ for all $t \geq 1$.

To prove Theorem 5, we show that (δ_t) satisfies the recursive relation in (15), which can be bounded by Proposition A.5. Note that by setting $n_t = 1$ for all $t \geq 1$ in (15), the error from C_∇^2 will disappear, giving us the same results as in Bach and Moulines (2011).
By the conditions imposed on the learning rate (Robbins and Monro, 1951), we have $\sum_{i=1}^t \gamma_i = \infty$ and $\sum_{i=1}^t \gamma_i^2/n_i \leq \sum_{i=1}^t \gamma_i^2 < \infty$ for $t \to \infty$. 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 $\exp(-\mu \sum_{i=t/2}^t \gamma_i)$. Furthermore, Theorem 5 provides an upper bound on the function values, namely, $\mathbb{E}\left[L(\theta_t) - L(\theta^*)\right] \leq C_l\delta_t/2$; this follows directly by Cauchy-Schwarz inequality and Assumption A2 since

$$\begin{aligned}
\mathbb{E}\left[l_t(\theta_{t-1}) - l_t(\theta^*)|\mathcal{F}_{t-1}\right] &= \frac{1}{n_t} \sum_{i=1}^{n_t} \mathbb{E}\left[l_{t,i}(\theta_{t-1}) - l_{t,i}(\theta^*)\middle|\mathcal{F}_{t-1}\right] \\
&= \frac{1}{n_t} \sum_{i=1}^{n_t} \int_0^1 \mathbb{E}\left[\left\langle \nabla_\theta l_{t,i}(\nu\theta_{t-1} + (1-\nu)\theta^*) - \nabla_\theta l_{t,i}(\theta^*), \theta_{t-1} - \theta^* \right\rangle \middle|\mathcal{F}_{t-1}\right] d\nu \\
&\leq \frac{1}{n_t} \sum_{i=1}^{n_t} \int_0^1 \left(\mathbb{E}\left[\left\|\nabla_\theta l_{t,i}(\nu\theta_{t-1} + (1-\nu)\theta^*) - \nabla_\theta l_{t,i}(\theta^*)\right\|^2 \middle|\mathcal{F}_{t-1}\right]\right)^{\frac{1}{2}} \|\theta_{t-1} - \theta^*\| \, d\nu \\
&\leq \frac{C_l}{n_t} \sum_{i=1}^{n_t} \|\theta_{t-1} - \theta^*\| \int_0^1 \nu \, d\nu = \frac{C_l}{2} \|\theta_{t-1} - \theta^*\|.
\end{aligned}$$

Throughout this paper, we will consider learning rates on the form $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ with $C_\gamma > 0$, $\beta \in [0, 1]$, and α chosen accordingly to the expected streaming-batches denoted by n_t. To obtain sub-linear convergence of $O(t^{-1})$, we generally need to choose $\gamma_t = O(t^{-1})$. We start by considering constant streaming-batches where n_t follows the constant streaming-batch size $C_\rho \in \mathbb{N}$:

Corollary 1 (SSG with constant streaming-batches)**.** *Suppose $\gamma_t = C_\gamma C_\rho^\beta t^{-\alpha}$ such that $\alpha \in (1/2, 1)$. Under Assumption (A1,A2,A3), we have*

$$\delta_t \leq \exp\left(-\frac{\mu C_\gamma N_t^{1-\alpha}}{2^{1-\alpha} C_\rho^{1-\alpha-\beta}}\right) \left(\delta_0 + \frac{2\sigma^2}{C_l^2}\right) \pi_c + \frac{2^{1+\alpha}\sigma^2 C_\gamma}{\mu C_\rho^{1-\alpha-\beta} N_t^\alpha}, \tag{6}$$

where $\pi_c = \exp\left(\frac{4\alpha C_\gamma^2\left(2C_l^2 + C_\rho \mathbb{1}_{\{C_\rho > 1\}} C_\nabla^2\right)}{(2\alpha-1)C_\rho^{1-2\beta}}\right)$ *is a finite constant.*

The bound in Corollary 1 depends on the initial condition $\delta_0 = \mathbb{E}[\|\theta_0 - \theta^*\|^2]$ and the variance σ^2 in the noise term. The initial condition δ_0 vanish sub-exponentially fast for $\alpha \in (1/2, 1)$. Thus, the asymptotic term is $2^{1+\alpha}\sigma^2 C_\gamma / \mu C_\rho^{1-\alpha-\beta} N_t^\alpha$, i.e., $\delta_t = O(N_t^{-\alpha})$. Moreover, the bound in (6) is optimal (up to some constants) for quadratic functions ($l_{t,i}$), since the deterministic recursion equation (15) would be with equality. This matches the findings made by Bach and Moulines (2011) for $C_\rho = 1$. It is worth noting that if $C_\gamma C_l$ or $C_\gamma C_\nabla$ is chosen too large, they may produce a large π_c constant. To control π_c for any C_ρ, setting $\beta = 0$ seems to be a suitable compromise. Obviously, the hyper-parameter β only comes into play if the streaming-batch size 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. Next, the asymptotic term is divided by $C_\rho^{1-\alpha-\beta}$, implying we could achieve noise reduction by taking $\alpha + \beta \leq 1$ (when C_ρ is large). Taking a large streaming-batch size, e.g., $C_\rho = t$, one accelerates the vanilla SG convergence rate to $O(N_t^{1-\beta})$. However, this large streaming batch size would be unsuitable in practice, and it would mean that we would take few steps until convergence is achieved.

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

Corollary 2 (SSG with varying streaming-batches)**.** *Suppose* $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ *where* $n_t = C_\rho t^\rho$ *with* $C_\rho \geq 1$ *and* $\rho \in (-1, 1)$*, such that* $\alpha - \beta\tilde{\rho} \in (1/2, 1)$*. Under Assumption (A1,A2,A3), we have*

$$\delta_t \leq \exp\left(-\frac{\mu C_\gamma N_t^{1-\phi}}{2^{(2+\rho)(1-\phi)} C_\rho^{1-\beta-\phi}}\right)\left(\delta_0 + \frac{2\sigma^2}{C_l^2}\right)\pi_v + \frac{2^{1+(2+\rho)\phi}\sigma^2 C_\gamma}{\mu C_\rho^{(1-\beta)\mathbb{1}_{\{\rho\geq 0\}}-\phi} N_t^\phi}, \tag{7}$$

where $\phi = \frac{(1-\beta)\tilde{\rho}+\alpha}{1+\tilde{\rho}}$ *and* $\pi_v = \exp\left(\frac{4(\alpha-\beta\tilde{\rho})C_\gamma^2 C_\rho^{2\beta}\left(2C_l^2 + C_\nabla^2\right)}{2(\alpha-\beta\tilde{\rho})-1}\right)$ *is a finite constant.*

As mentioned for Corollary 1, the condition of having $\alpha - \beta\tilde{\rho} \in (1/2, 1)$ is a natural restriction coming from Robbins and Monro (1951), which relaxes the usual condition of having $\alpha \in (1/2, 1)$ for ρ non-negative. Moreover, for $\rho \geq 0$, the sub-exponential and asymptotic term is scaled by $C_\rho^{1-\beta-\phi}$, implying we should take $\alpha + \beta \leq 1$ for having $1 - \beta - \phi$ positive. We could accelerate the convergence by, e.g., setting $\alpha = 1$ and $\beta = 1/2$, whatever the streaming-batch rate $\rho \in (0, 1)$, would give us the rate of convergence $\delta_t = O(N_t^{-(1+\rho/2)/(1+\rho)})$. Likewise, setting $\alpha = 2/3$ and $\beta = 1/3$ would give convergence rate $\delta_t = O(N_t^{-2/3})$ for any $\rho \in (-1, 1/2)$. These conclusions will change when we consider the (averaging estimate) ASSG in Section 3.

The reasoning in Corollary 2 could be expanded to include *random* streaming-batches where n_t is given such that $C_L t^{\rho_L} \leq n_t \leq C_H t^{\rho_H}$ with $\rho_L, \rho_H \in (-1, 1)$ and $C_L, C_H \geq 1$. This yields the modified $\phi' = ((1-\beta)\rho_L + \alpha)/(1 + \rho_H)$; nevertheless, we will leave the proof to the reader.

3. Averaged Stochastic Streaming Gradient

In what follows, we consider the averaging estimate $\bar{\theta}_n$ given in (4). We expand assumptions A2 and A3 with the three following assumptions:

Assumption A4. For each $t \geq 1$, the function $l_{t,i}$ is differentiable, and there exists $C_l > 0$ such that $\mathbb{E}[\|\nabla_\theta l_{t,i}(\theta) - \nabla_\theta l_{t,i}(\theta')\|^p | \mathcal{F}_{t-1}] \leq C_l^p \|\theta - \theta'\|^p$ a.s. for all $\theta, \theta' \in \mathbb{R}^d$ with $p \in \{1, 2, 3, 4\}$ and $i = 1, \ldots, n_t$.

Assumption A5. There exists $\tau > 0$ such that for all $t \geq 1$ and $p \in \{1, 2, 3, 4\}$, $\mathbb{E}[\|\nabla_\theta l_{t,i}(\theta^*)\|^p | \mathcal{F}_{t-1}] \leq \tau^p$ a.s. for all $\theta \in \mathbb{R}^d$ and $i = 1, \ldots, n_t$.

Assumption A6. There exists a non-negative self-adjoint operator Σ such that $\mathbb{E}[\nabla_\theta l_t(\theta^*) \nabla_\theta l_t(\theta^*)^\top | \mathcal{F}_{t-1}] \leq \Sigma$ a.s. for all $t \geq 1$.

Moreover, an additional assumption is needed for bounding the *rest* term of the averaging estimate ASSG.

Assumption A7. The function L is almost surely twice differentiable with Lipschitz-continuous Hessian operator $\nabla^2_\theta L$, and Lipschitz constant $C_\delta > 0$, that is, for all $\theta, \theta' \in \mathbb{R}^d$, $\|\nabla_\theta L(\theta) - \nabla^2_\theta L(\theta')(\theta - \theta')\| \leq C_\delta \|\theta - \theta'\|^2$.

As in Section 2, we conduct a general study for any decreasing step $(\gamma_t)_{t\geq 1}$ when applying the Polyak-Ruppert averaging estimate (4). First, let us give the fourth-order rate of convergence of the estimates by use of Proposition A.5:

Lemma 1. *Denote $\Delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^4]$ with some $\Delta_0 \geq 0$. Under Assumption (A1,A4,A5,A6,A7), we have for any decreasing step sequence $(\gamma_t)_{t\geq 1}$ that*

$$\Delta_t \leq \exp\left(-\mu \sum_{i=t/2}^{t} \gamma_i\right)\left(\Delta_0 + \frac{2\tau^4}{C_l^4} + \frac{4\tau^4 \gamma_1}{\mu C_l^2 n_1}\right)\Pi + \frac{32\tau^4}{\mu^2} \max_{t/2\leq i\leq t} \frac{\gamma_i^2}{n_i^2} + \frac{48\tau^4}{\mu} \max_{t/2\leq i\leq t} \frac{\gamma_i^3}{n_i^3} + \frac{114\tau^4}{\mu} \max_{t/2\leq i\leq t} \frac{\gamma_i^3 \mathbb{1}_{\{n_i>1\}}}{n_i^2}, \tag{8}$$

with Π given in (26).

Theorem 2. *Denote $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t - \theta^*\|^2]$ with $\bar{\theta}_n$ given by (4). Under Assumption (A1,A4,A5,A6,A7), we have for any decreasing step sequence $(\gamma_t)_{t\geq 1}$ that*

$$\bar{\delta}_t^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{1}{N_t\mu} \sum_{i=1}^{t-1} \left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| \delta_i^{1/2} + \frac{n_t}{N_t\gamma_t\mu}\delta_t^{1/2} + \frac{n_1}{N_t\mu}\left(\frac{1}{\gamma_1} + C_l\right)\delta_0^{1/2} + \frac{C_l}{N_t\mu}\left(\sum_{i=1}^{t-1} n_{i+1}\delta_i\right)^{1/2} + \frac{C_\delta}{N_t\mu}\sum_{i=0}^{t-1} n_{i+1}\Delta_i^{1/2}, \tag{9}$$

where $\Lambda = \mathrm{Tr}(\nabla^2_\theta L(\theta^*)^{-1}\Sigma\nabla^2_\theta L(\theta^*)^{-1})$.

As noticed in Polyak and Juditsky (1992), the leading term Λ/N_t achieves the Cramer-Rao bound (Murata and Amari, 1999; Gadat and Panloup, 2017). Note that the leading term Λ/N_t is independent of the learning rate (γ_t). Next, the processes (δ_t) and (Δ_t) can be bounded by the recursive relations in (5) and (8). There are no sub-exponential decaying terms for the initial conditions in Theorem 2, which is a common problem for averaging. However, as described in the previous section corollaries, we are more interested in advancing the decay of the asymptotic terms.

Corollary 3 (ASSG with constant streaming-batches)**.** *Suppose $\gamma_t = C_\gamma C_\rho^\beta t^{-\alpha}$ such that $\alpha \in (1/2, 1)$. Under Assumption (A1,A4,A5,A6,A7), we have*

$$\begin{aligned}\bar{\delta}_t^{1/2} \leq &\frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{6\sigma C_\rho^{(1-\alpha-\beta)/2}}{\sqrt{C_\gamma}\mu^{3/2}N_t^{1-\alpha/2}} + \frac{2^\alpha 6 C_\delta \tau^2 C_\gamma}{C_\rho^{1-\alpha-\beta}\mu^2 N_t^\alpha} + \frac{C_\rho^{1-\alpha-\beta}\sqrt{\pi_c' A_\infty}}{C_\gamma \mu N_t^{1-\alpha}} + \frac{2C_l\sigma\sqrt{C_\gamma}}{C_\rho^{(1-\alpha-\beta)/2}\mu^{3/2}N_t^{(1+\alpha)/2}} + \frac{C_\rho \Gamma_c}{\mu N_t} \\ &+ \frac{(6 + 7\mathbb{1}_{\{C_\rho>1\}})2^{3\alpha/2} C_\delta \tau^2 C_\gamma^{3/2} C_\rho^{3\beta/2}}{\mu^{3/2} N_t}\left(\frac{C_\rho^{3\alpha/2-1}}{N_t^{3\alpha/2-1}}\mathbb{1}_{\{\alpha<2/3\}} + \log(N_t)\,\mathbb{1}_{\{\alpha=2/3\}} + \frac{3\alpha}{3\alpha-2}\mathbb{1}_{\{\alpha>2/3\}}\right),\end{aligned}$$

with Γ_c given by $\left(\frac{1}{C_\gamma C_\rho^\beta} + C_l\right)\delta_0^{1/2} + \frac{C_l\sqrt{\pi_c'}\sqrt{A_\infty}}{C_\rho^{1/2}} + \frac{\sqrt{\pi_c' A_\infty}}{C_\gamma C_\rho^\beta} + C_\delta\sqrt{\Pi_c' A_\infty}$, consisting of the finite constants π_c', Π_c' and A_∞, given in (37).

By averaging, we have increased the rate of convergence from $O(N_t^{-\alpha})$ to the optimal rate $O(N_t^{-1})$. The two subsequent terms are the main remaining terms decaying at the rate $O(N_t^{\alpha-2})$ and $O(N_t^{-2\alpha})$, which suggests setting $\alpha = 2/3$ would be *optimal*. The remaining terms are negligible. It is worth noting that having $\alpha + \beta = 1$ in Corollary 3, we would give no impact in the main remaining terms from the streaming-batch size C_ρ. Moreover, taking $\alpha = 2/3$ and $\beta \leq 1/3$ would be an *optimal* choice of hyper-parameters such that the streaming-batch size C_ρ have a positive or no impact. At last, as we do not rely on sub-exponentially decaying terms, we need to be more careful when picking our hyper-parameters, e.g., taking $C_\gamma C_l$ too large may cause Γ_c to be significant. In practice, Kushner and Yin (2003) suggests that performing some iterations before beginning to average may help with the initial slow convergence. Nevertheless, the term consisting of Γ_c decay at a rate of at least $O(N_t^{-2})$.

Corollary 4 (ASSG with varying streaming-batches)**.** *Suppose* $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ *where* $n_t = C_\rho t^\rho$ *with* $C_\rho \geq 1$ *and* $\rho \in (-1, 1)$*, such that* $\alpha - \beta\tilde{\rho} \in (1/2, 1)$*. Under Assumption (A1,A4,A5,A6,A7), we have*

$$\bar{\delta}_t^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{2^{3+\phi(1+\tilde{\rho})}\sigma C_\rho^{(1-\phi-\beta)/2\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma}N_t^{1-\phi/2}} + \frac{2^{(1+\phi)(1+\tilde{\rho})-2}C_\delta\tau^2 C_\gamma}{\mu^2 C_\rho^{1-\phi-\beta}N_t^\phi} + \frac{C_\rho^{1-\phi-\beta}\sqrt{\pi_v'}A_\infty'}{\mu C_\gamma N_t^{1-\phi}} + \frac{2^{\phi(1+\tilde{\rho})/2}C_l\sigma\sqrt{C_\gamma}}{\mu^{3/2}C_\rho^{(1-\phi-\beta)/2\mathbb{1}_{\{\rho\geq 0\}}}N_t^{(1+\phi)/2}}$$
$$+ \frac{C_\rho\Gamma_v}{\mu N_t} + \frac{2^{3(1+\phi)(1+\tilde{\rho})/2}C_\delta\tau^2 C_\gamma^{3/2}C_\rho^{1+3\beta/2}}{\mu^{3/2}C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}}N_t}\left(\frac{N_t^{3(1-\phi)/2}}{C_\rho^{3(1-\phi)/2}}\mathbb{1}_{\{\alpha-\beta\tilde{\rho}<2/3\}} + \log(N_t)\mathbb{1}_{\{\alpha-\beta\tilde{\rho}=2/3\}} + \frac{3(\alpha-\beta\tilde{\rho})}{3(\alpha-\beta\tilde{\rho})-2}\mathbb{1}_{\{\alpha-\beta\tilde{\rho}>2/3\}}\right),$$

with Γ_v *given by* $\left(\frac{1}{C_\gamma C_\rho^\beta} + C_l\right)\delta_0^{1/2} + \frac{2^{\tilde{\rho}}C_l\sqrt{\pi_v'}\sqrt{A_\infty'}}{C_\rho^{1/2}} + \frac{2\sqrt{\pi_v'}A_\infty'}{C_\gamma C_\rho^\beta} + 2^{\tilde{\rho}}C_\delta\sqrt{\Pi_v'}A_\infty'$*, consisting of the finite constants* π_v'*,* Π_v' *and* A_∞'*, given in* (41).

Following the arguments above, the two main remainder terms reveal 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 3, namely $\alpha = 2/3$. However, these hyper-parameter choices are not resilient against any arrival schedule 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 *optimal* convergence for any data stream we may encounter by having $\alpha = 2/3$ and $\beta = 1/3$.

4. Experiments

To demonstrate the theoretical results presented in Sections 2 and 3, we consider linear regression for various data streams. Specifically, we assume $y_t = X_t^T\theta + \epsilon_t$ for $t \geq 1$ where the features $X_t \in \mathbb{R}^d$ is a random vector, $\theta \in \mathbb{R}^d$ is the parameters vector, and ϵ_t is a random variable with zero mean, independent from x_t. Moreover, $(X_i, \epsilon_t)_{t\geq 1}$ are independent and identically distributed. In this section, 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 come from a standard Gaussian distribution. For measuring the performance of our methods, we calculate the quadratic mean error of the parameter estimates for 100 replications, given by $(\mathbb{E}[\|\theta_{N_t} - \theta\|^2])_{t\geq 1}$. Note that averaging over several iterations gives a reduction in variability, which mainly benefits SSG. To compare our experiments with a focus on the various data streams, we fix the following parameters: $C_\gamma = 1$ and $\alpha = 2/3$.
In Figure 1, we consider constant data streams to illustrate the results in Corollaries 1 and 3. The figures show a solid decay rate proportional to $\alpha = 2/3$ for any streaming-batch size $C_\rho \in \{1, 8, 64, 128\}$ with $\beta = 0$, as shown in Corollary 1. In addition, we see an acceleration in decay by averaging (ASSG), as explained in Corollary 3. Both methods show a noticeable reduction in variance when C_ρ increases which are particularly beneficial in the beginning. Moreover, as mentioned in Remark 1, the *stationary* phase may also commence earlier when we raise the streaming-batch size C_ρ.

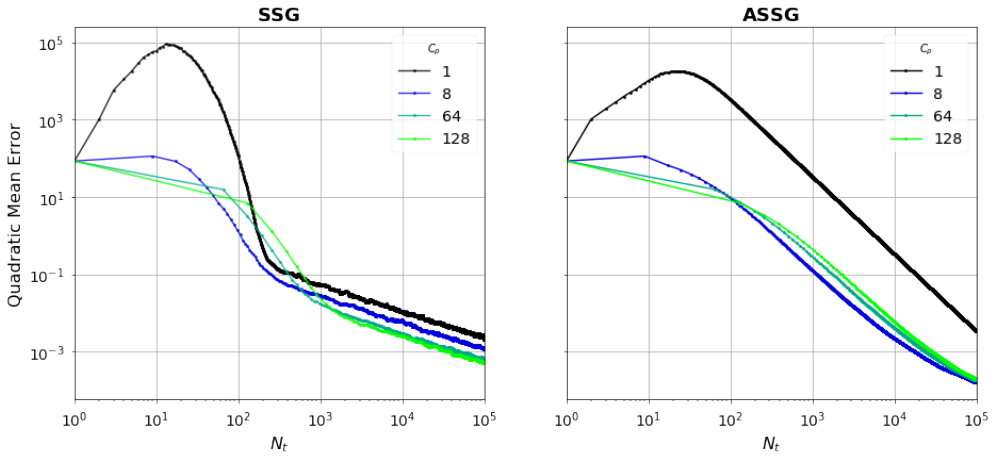

Figure 1: Trajectories of $(\mathbb{E}[\|\theta_{N_t} - \theta^*\|^2])_{t\geq 1}$ for constant streaming-batch sizes.

Next, in Figures 2, 3, 4, and 5, we vary the streaming rate ρ for streaming-batch sizes $C_\rho = 1$, 8, 64, and 128, respectively, with $\beta = 0$. These figures shows an increase of the decay rate for SSG when we take ρ positive while having $\beta = 0$. We know this from Corollary 2, as $\phi = (\tilde{\rho} + \alpha)/(1 + \tilde{\rho}) \geq \alpha$ for $\beta = 0$. Despite this, we still achieve better convergence for the ASSG method, which seems more immune to the different choices of streaming rate ρ, e.g., see the discussion after Corollary 4. In addition, we see that C_ρ has a positive effect on the noise, but if C_ρ becomes too large, it may slow down convergence (as seen in Figure 5).

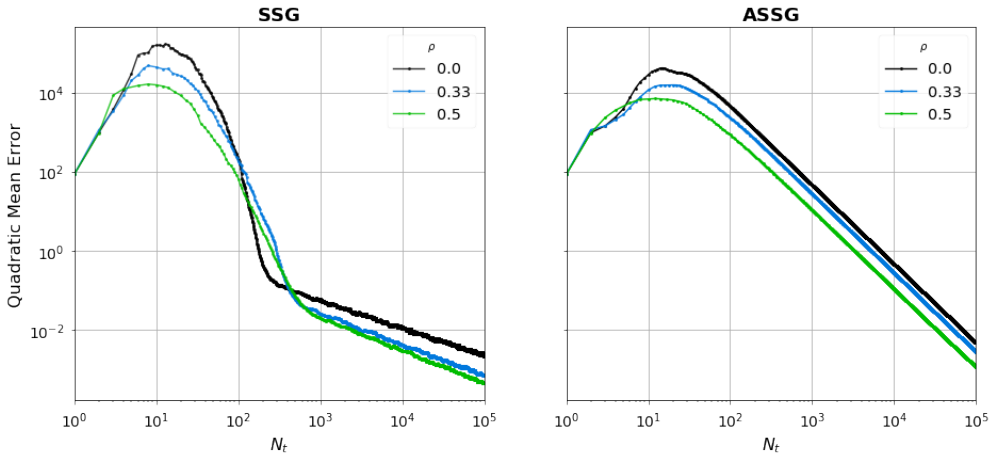

Figure 2: Trajectories of $(\mathbb{E}[\|\theta_{N_t} - \theta^*\|^2])_{t\geq 1}$ for varying streaming rates with $C_\rho = 1$.

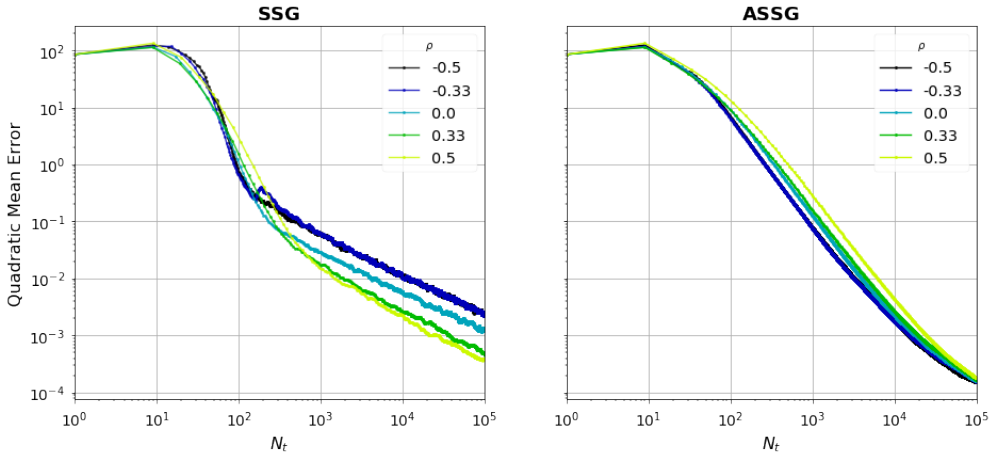

Figure 3: Trajectories of $(\mathbb{E}[\|\theta_{N_t} - \theta^*\|^2])_{t\geq 1}$ for varying streaming rates with $C_\rho = 8$.

Alternatively, we could think around the problem in another way; how can we choose α and β such that we have *optimal* 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 optimal decay of $\phi = 2/3$. As discussed after Corollary 4, one example of this could be done by setting $\alpha = 2/3$ and $\beta = 1/3$ such that $\phi = 2/3$ for any ρ. Figure 6 shows an example of this where we achieve the same decay rate for any streaming rate ρ.

5. Conclusion

We considered the stochastic approximation problem in a streaming setting where we had to minimize a convex objective using only unbiased estimates of its gradients. We introduced and studied the convergence rates of the SSG

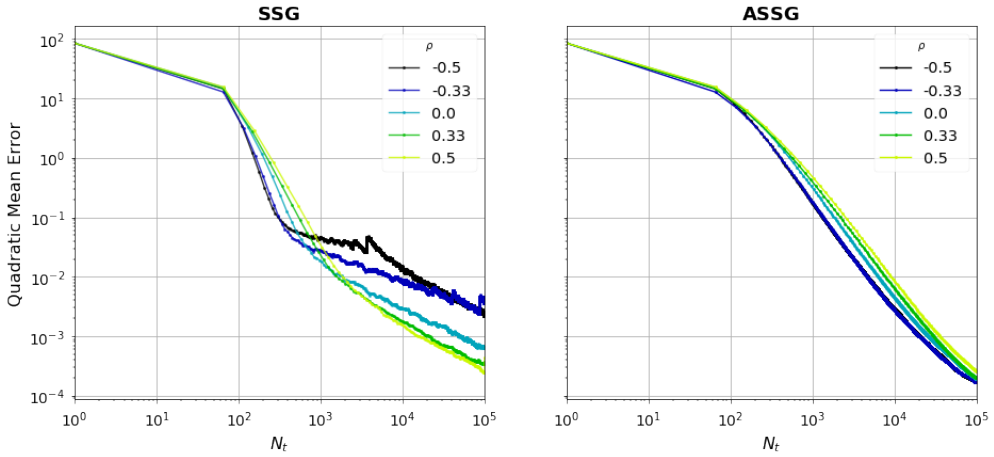

Figure 4: Trajectories of $(\mathbb{E}[\|\theta_{N_t} - \theta^*\|^2])_{t\geq 1}$ for varying streaming rates with $C_\rho = 64$.

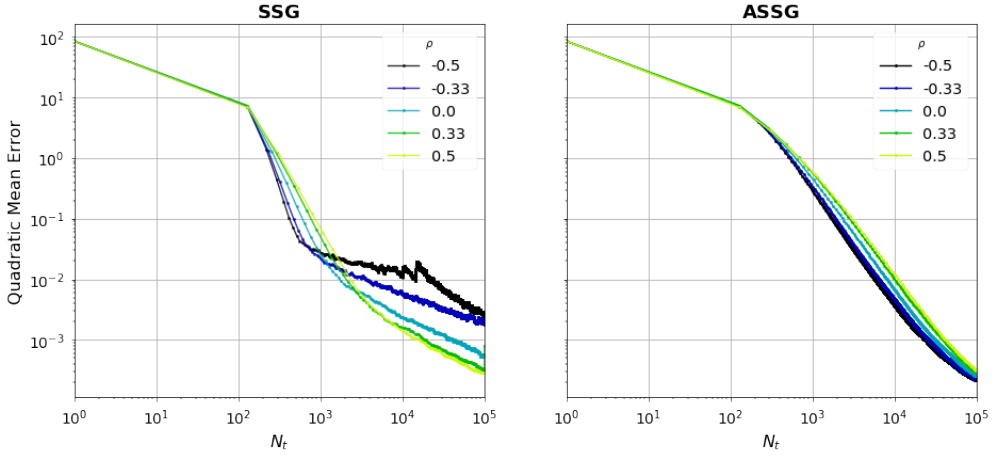

Figure 5: Trajectories of $(\mathbb{E}[\|\theta_{N_t} - \theta^*\|^2])_{t\geq 1}$ for varying streaming rates with $C_\rho = 128$.

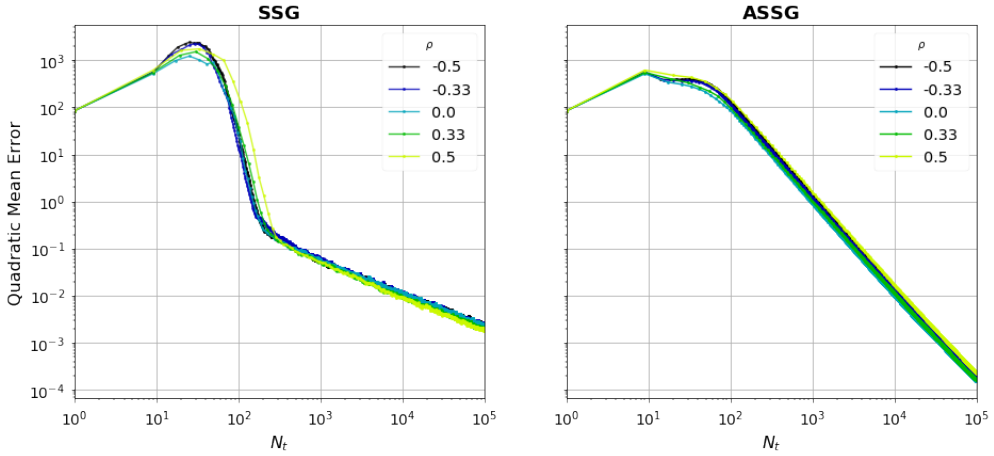

Figure 6: Trajectories of $(\mathbb{E}[\|\theta_{N_t} - \theta^*\|^2])_{t\geq 1}$ for varying streaming rates with $C_\rho = 8$ and $\phi = 2/3$.

and ASSG algorithms in a non-asymptotic manner. This investigation was derived using learning rates of the form $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ under varying data streams of n_t, which includes classic algorithms such as vanilla SG and ASG and mini-batch SG and ASG. The theoretical results and our experiments showed a noticeable improvement in the convergence rate by choosing the learning rate (hyper-parameters) according to the expected data streams. For ASSG, we showed that this choice of learning rate led to optimal convergence rates and was robust to any data stream rate we may encounter. Moreover, in large-scale learning problems, we know how to accelerate convergence and reduce noise through the learning rate and the treatment pattern of the data.

There are several ways to expand our work but let us give some examples: first, we can extend our analysis to include streaming-batches of any size in the spirit of the discussion after Corollary 2. Second, many machine learning problems encounter correlated variables and high-dimensional data, making an extension to non-strongly convex objectives advantageous. Third, Assumption A1 requires random functions with minimal dependency structure; thus, an obvious extension could incorporate a more realistic dependency assumption, thereby increasing the applicability for more models. Moreover, studying dependence may give insight into how to process dependent information *optimally*. Next, a natural extension would be to modify our averaging estimate (4) to a weighted averaged version (WASSG) proposed by Mokkadem and Pelletier (2011); Boyer and Godichon-Baggioni (2020), given as

$$\bar{\theta}_{t,\lambda} = \frac{1}{\sum_{i=1}^{t} n_i \log(1+i)^\lambda} \sum_{i=1}^{t} n_i \log(1+i)^\lambda \theta_{i-1}, \tag{10}$$

with $\bar{\theta}_{0,\lambda} = 0$ and $\lambda > 0$. By giving more importance to the latest estimates, we should improve convergence and limit the effect of bad initializations. Following the demonstrations in Section 4, an example of this WASSG estimate $(\bar{\theta}_{t,\lambda})$ can be found in Figure 7 with use of $\lambda = 2$. Here we see that although the WASSG estimate in (10) may not achieve a better final error (compared to the ASSG estimate in Figure 6), it still achieves a better decay along the way, often referred to as *parameter tracking*.

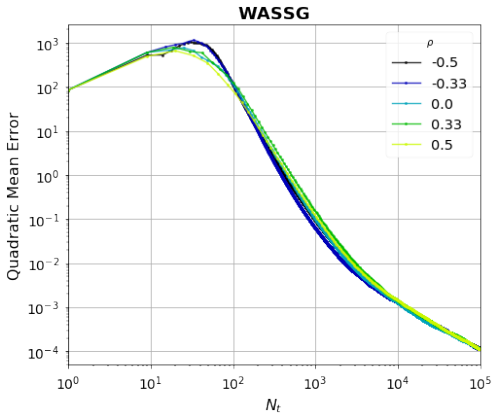

Figure 7: Trajectories of $(\mathbb{E}[\|\theta_{N_t} - \theta^*\|^2])_{t\geq 1}$ for varying streaming rates with $C_\rho = 8$ and $\phi = 2/3$.

6. Proofs

In this section, we provide detailed proofs of the results presented in the manuscript. Purely technical results used in the proofs can be found in Appendix A.

6.1. Proofs for Section 2

Proof of Theorem 1. By equation (3), we have

$$\|\theta_t - \theta^*\|^2 = \|\theta_{t-1} - \theta^*\|^2 + \gamma_t^2 \left\| \frac{1}{n_t} \sum_{i=1}^{n_t} \nabla_\theta l_{t,i}(\theta_{t-1}) \right\|^2 - \frac{2\gamma_t}{n_t} \sum_{i=1}^{n_t} \langle \nabla_\theta l_{t,i}(\theta_{t-1}), \theta_{t-1} - \theta^* \rangle . \tag{11}$$

Taking the conditional expectation of equality (11) yields

$$\mathbb{E}\left[\|\theta_t-\theta^*\|^2\middle|\mathcal{F}_{t-1}\right]=\|\theta_{t-1}-\theta^*\|^2+\gamma_t^2\mathbb{E}\left[\left\|\frac{1}{n_t}\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2\middle|\mathcal{F}_{t-1}\right]-\frac{2\gamma_t}{n_t}\sum_{i=1}^{n_t}\mathbb{E}\left[\left\langle\nabla_\theta l_{t,i}\left(\theta_{t-1}\right),\theta_{t-1}-\theta^*\right\rangle\middle|\mathcal{F}_{t-1}\right]. \quad (12)$$

To bound the second term in (12), we first expand it as follows:

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

Utilizing the Lipschitz continuity of $\nabla_\theta l_{t,i}$, together with Assumption A2 and A3, and that θ_{t-1} is $\mathcal{F}_{t-1}$-measurable (Assumption A1), we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2\middle|\mathcal{F}_{t-1}\right]=&\mathbb{E}\left[\left\|\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)-\nabla_\theta l_{t,i}\left(\theta^*\right)+\nabla_\theta l_{t,i}\left(\theta^*\right)\right\|^2\middle|\mathcal{F}_{t-1}\right]\\ \leq&2\mathbb{E}\left[\left\|\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)-\nabla_\theta l_{t,i}\left(\theta^*\right)\right\|^2\middle|\mathcal{F}_{t-1}\right]+2\mathbb{E}\left[\left\|\nabla_\theta l_{t,i}\left(\theta^*\right)\right\|^2\middle|\mathcal{F}_{t-1}\right]\\ \leq&2C_l^2\left\|\theta_{t-1}-\theta^*\right\|^2+2\sigma^2,\end{aligned}$$

using the bound $\|x+y\|^p\leq 2^{p-1}\left(\|x\|^p+\|y\|^p\right)$. Next, we note that for all $0\leq i<j$ that $\mathcal{F}_{t-1}\subseteq\mathcal{F}_{t-1,i}\subset\mathcal{F}_{t-1,j}$. Thus

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

Since θ_{t-1} and $l_{t,i}$ are $\mathcal{F}_{t-1,j-1}$-measurable for all $0\leq i<j$, then

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[\left\langle\nabla_\theta l_{t,i}\left(\theta_{t-1}\right),\nabla_\theta l_{t,j}\left(\theta_{t-1}\right)\right\rangle\middle|\mathcal{F}_{t-1,j-1}\right]\middle|\mathcal{F}_{t-1,i-1}\right]\middle|\mathcal{F}_{t-1}\right]=&\mathbb{E}\left[\mathbb{E}\left[\left\langle\nabla_\theta l_{t,i}\left(\theta_{t-1}\right),\mathbb{E}\left[\nabla_\theta l_{t,j}\left(\theta_{t-1}\right)\middle|\mathcal{F}_{t-1,j-1}\right]\right\rangle\middle|\mathcal{F}_{t-1,i-1}\right]\middle|\mathcal{F}_{t-1}\right]\\ =&\mathbb{E}\left[\mathbb{E}\left[\left\langle\nabla_\theta l_{t,i}\left(\theta_{t-1}\right),\nabla_\theta L\left(\theta_{t-1}\right)\right\rangle\middle|\mathcal{F}_{t-1,i-1}\right]\middle|\mathcal{F}_{t-1}\right].\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[\left\langle\nabla_\theta l_{t,i}\left(\theta_{t-1}\right),\nabla_\theta L\left(\theta_{t-1}\right)\right\rangle\middle|\mathcal{F}_{t-1,i-1}\right]\middle|\mathcal{F}_{t-1}\right]=&\mathbb{E}\left[\left\langle\mathbb{E}\left[\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\middle|\mathcal{F}_{t-1,i-1}\right],\nabla_\theta L\left(\theta_{t-1}\right)\right\rangle\middle|\mathcal{F}_{t-1}\right]\\ =&\mathbb{E}\left[\left\langle\nabla_\theta L\left(\theta_{t-1}\right),\nabla_\theta L\left(\theta_{t-1}\right)\right\rangle|\mathcal{F}_{t-1}\right]\\ =&\mathbb{E}\left[\left\|\nabla_\theta L\left(\theta_{t-1}\right)\right\|^2\middle|\mathcal{F}_{t-1}\right]\\ =&\left\|\nabla_\theta L\left(\theta_{t-1}\right)\right\|^2.\end{aligned}$$

As $\nabla_\theta L$ is C_∇-Lipschitz continuous, then $\|\nabla_\theta L\left(\theta_{t-1}\right)\|^2\leq C_\nabla^2\left\|\theta_{t-1}-\theta^*\right\|^2$ as $\nabla_\theta L\left(\theta^*\right)=0$. Thus, we obtain a bound of the second term of (12):

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla_\theta l_t\left(\theta_{t-1}\right)\right\|^2\middle|\mathcal{F}_{t-1}\right]=&\frac{1}{n_t^2}\sum_{i=1}^{n_t}\mathbb{E}\left[\left\|\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2\middle|\mathcal{F}_{t-1}\right]+\frac{1}{n_t^2}\sum_{i\neq j}^{n_t}\mathbb{E}\left[\left\langle\nabla_\theta l_{t,i}\left(\theta_{t-1}\right),\nabla_\theta l_{t,j}\left(\theta_{t-1}\right)\right\rangle\middle|\mathcal{F}_{t-1}\right]\\ \leq&\frac{1}{n_t^2}\sum_{i=1}^{n_t}\left(2C_l^2\left\|\theta_{t-1}-\theta^*\right\|^2+2\sigma^2\right)+\frac{1}{n_t^2}\sum_{i\neq j}^{n_t}C_\nabla^2\left\|\theta_{t-1}-\theta^*\right\|^2\\ =&\frac{1}{n_t^2}n_t\left(2C_l^2\left\|\theta_{t-1}-\theta^*\right\|^2+2\sigma^2\right)+\frac{1}{n_t^2}n_t(n_t-1)C_\nabla^2\left\|\theta_{t-1}-\theta^*\right\|^2\\ =&\left(2C_l^2n_t^{-1}+C_\nabla^2(n_t-1)n_t^{-1}\right)\left\|\theta_{t-1}-\theta^*\right\|^2+2\sigma^2n_t^{-1}. \quad (13)\end{aligned}$$

Since L is $\mu-$strongly convex and θ_{t-1} is $\mathcal{F}_{t-1}$-measurable, we can bound the third term in (12) as

$$\mathbb{E}\left[\left\langle\nabla_\theta l_{t,i}\left(\theta_{t-1}\right),\theta_{t-1}-\theta^*\right\rangle\middle|\mathcal{F}_{t-1}\right]=\left\langle\mathbb{E}\left[\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\middle|\mathcal{F}_{t-1}\right],\theta_{t-1}-\theta^*\right\rangle=\left\langle\nabla_\theta L\left(\theta_{t-1}\right),\theta_{t-1}-\theta^*\right\rangle\geq\mu\left\|\theta_{t-1}-\theta^*\right\|^2, \quad (14)$$

by Assumption A1. Combining inequality (14) and (13) into (12), gives us

$$\mathbb{E}\left[\|\theta_t - \theta^*\|^2 \middle| \mathcal{F}_{t-1}\right] \leq \left[1 - 2\mu\gamma_t + \left(2C_l^2 + (n_t - 1)\, C_\nabla^2\right) n_t^{-1}\gamma_t^2\right] \|\theta_{t-1} - \theta^*\|^2 + 2\sigma^2 n_t^{-1}\gamma_t^2.$$

Thus, setting $\delta_t = \mathbb{E}[\|\theta_t - \theta^*\|^2]$ and taking the expectation on both sides of the inequality above, yields the recursive relation:

$$\delta_t \leq \left[1 - 2\mu\gamma_t + \left(2C_l^2 + (n_t - 1)\, C_\nabla^2\right) n_t^{-1}\gamma_t^2\right] \delta_{t-1} + 2\sigma^2 n_t^{-1}\gamma_t^2. \tag{15}$$

By Proposition A.5, we achieve the (upper) bound of δ_t, given as

$$\delta_t \leq \exp\left(-\mu \sum_{i=t/2}^{t} \gamma_i\right) \exp\left(2 \sum_{i=1}^{t} \eta_i \gamma_i\right) \left(\delta_0 + 2 \max_{1 \leq i \leq t} \frac{\nu_i}{\eta_i}\right) + \frac{1}{\mu} \max_{t/2 \leq i \leq t} \nu_i,$$

by setting $\omega = \mu$, $\eta_t = (2C_l^2 + (n_t - 1)C_\nabla^2) n_t^{-1}\gamma_t$ and $\nu_t = 2\sigma^2 n_t^{-1}\gamma_t$.

Remark 1. The decrease of η_t determines when the *stationary* phase occurs. This is more clearly seen in Proposition A.4, where the inner terms directly depend on the inception of the stationary phase. Thus, by increasing n_t, we decrease η_t, and especially it dominates the constant C_l.

Before plugging these terms into the bound, we note that

$$\exp\left(2 \sum_{i=1}^{t} \eta_i \gamma_i\right) = \exp\left(4C_l^2 \sum_{i=1}^{t} n_i^{-1}\gamma_i^2\right) \exp\left(2C_\nabla^2 \sum_{i=1}^{t} (n_i - 1) n_i^{-1}\gamma_i^2\right) \leq \exp\left(4C_l^2 \sum_{i=1}^{t} n_i^{-1}\gamma_i^2\right) \exp\left(2C_\nabla^2 \sum_{i=1}^{t} \mathbb{1}_{\{n_i > 1\}}\gamma_i^2\right),$$

as $(n_t - 1)n_t^{-1} \leq \mathbb{1}_{\{n_t > 1\}}$. Next, as $n_t \geq 1$, we can simplify the term $\max_{1 \leq i \leq t} \frac{\nu_i}{\eta_i}$ as follows:

$$\max_{1 \leq i \leq t} \frac{\nu_i}{\eta_i} = \max_{1 \leq i \leq t} \frac{2\sigma^2}{2C_l^2 + (n_i - 1)C_\nabla^2} \leq \max_{1 \leq i \leq t} \frac{2\sigma^2}{2C_l^2} = \frac{\sigma^2}{C_l^2}.$$

Combining what we have above, we obtain the inequality in (5), namely

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

□

Proof of Corollary 1. By Theorem 1, we have the upper bound in (5) which can be simplified as $n_t = C_\rho$, giving us

$$\delta_t \leq \exp\left(-\mu \sum_{i=t/2}^{t} \gamma_i\right) \exp\left(\frac{4C_l^2}{C_\rho} \sum_{i=1}^{t} \gamma_i^2\right) \exp\left(2C_\nabla^2 \mathbb{1}_{\{C_\rho > 1\}} \sum_{i=1}^{t} \gamma_i^2\right) \left(\delta_0 + \frac{2\sigma^2}{C_l^2}\right) + \frac{2\sigma^2}{\mu C_\rho} \max_{t/2 \leq i \leq t} \gamma_i. \tag{16}$$

Plugging $\gamma_t = C_\gamma C_\rho^\beta t^{-\alpha}$ into (16), we can bound each term as follows:

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

using the integral test for convergence. Likewise, with the help of integral tests for convergence, we can bound the sum term $\sum_{i=1}^{t} \gamma_i^2 = C_\gamma^2 C_\rho^{2\beta} \sum_{i=1}^{t} i^{-2\alpha}$, using

$$\sum_{i=1}^{t} i^{-2\alpha} = 1 + \sum_{i=2}^{t} i^{-2\alpha} \leq 1 + \int_1^t x^{-2\alpha}\, dx = 1 + \frac{1}{2\alpha - 1}\left[-x^{1-2\alpha}\right]_{x=1}^{t} = 1 + \frac{1}{2\alpha - 1}\left(1 - t^{1-2\alpha}\right) \leq 1 + \frac{1}{2\alpha - 1} = \frac{2\alpha}{2\alpha - 1},$$

as $\alpha \in (1/2, 1)$. 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 (16), gives us

$$\delta_t \leq \exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1-\alpha}}{2^{1-\alpha}}\right)\exp\left(\frac{4\alpha C_\gamma^2\left(2C_l^2 + C_\rho \mathbb{1}_{\{C_\rho>1\}} C_\nabla^2\right)}{(2\alpha-1)\, C_\rho^{1-2\beta}}\right)\left(\delta_0 + \frac{2\sigma^2}{C_l^2}\right) + \frac{2^{1+\alpha}\sigma^2 C_\gamma}{\mu C_\rho^{1-\beta} t^\alpha}. \tag{17}$$

Converting (17) into terms of N_t (using $N_t = C_\rho t$), we obtain:

$$\delta_t \leq \exp\left(-\frac{\mu C_\gamma N_t^{1-\alpha}}{2^{1-\alpha} C_\rho^{1-\alpha-\beta}}\right)\exp\left(\frac{4\alpha C_\gamma^2\left(2C_l^2 + C_\rho \mathbb{1}_{\{C_\rho>1\}} C_\nabla^2\right)}{(2\alpha-1)\, C_\rho^{1-2\beta}}\right)\left(\delta_0 + \frac{2\sigma^2}{C_l^2}\right) + \frac{2^{1+\alpha}\sigma^2 C_\gamma}{\mu C_\rho^{1-\alpha-\beta} N_t^\alpha}. \tag{18}$$

□

Proof of Corollary 2. By Theorem 1 we have the upper bound in (5) under Assumption A1, A2 and A3. For convenience, we divided our proof into two cases to comprehend that $n_t \geq 1$ for all t: We bound each term of (5) after inserting, $\gamma_t = C_\gamma n_t^\beta t^{-\alpha} = C_\gamma C_\rho^\beta t^{\beta\rho-\alpha}$ if $\rho \geq 0$, or $\gamma_t \geq C_\gamma t^{-\alpha}$ if $\rho < 0$ (using that $\beta \geq 0$) into the inequality. If $\rho \geq 0$, the first term of (5) can be bounded, as follows:

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

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

$$\sum_{i=1}^{t} \frac{\gamma_i^2}{n_i} \leq \sum_{i=1}^{t} \gamma_i^2 \leq \frac{2\,(\alpha-\beta\rho)\, C_\gamma^2 C_\rho^{2\beta}}{2\,(\alpha-\beta\rho)-1},$$

as $n_t \geq 1$ and $\alpha - \rho\beta > 1/2$. If $\rho < 0$, one has

$$\sum_{i=1}^{t} \frac{\gamma_i^2}{n_i} \leq \sum_{i=1}^{t} \gamma_i^2 \leq \frac{2\alpha C_\gamma^2 C_\rho^{2\beta}}{2\alpha-1},$$

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 (5) by

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

Likewise, if $\rho < 0$, we have

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

since $n_t \geq 1$ and $\beta \leq 1$. Combining all these findings into (5), gives us (in terms of t):

$$\delta_t \leq \exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1+\beta\rho-\alpha}}{2^{1+\beta\rho-\alpha}}\right)\exp\left(\frac{4\,(\alpha-\beta\rho)\, C_\gamma^2 C_\rho^{2\beta}\left(2C_l^2 + C_\nabla^2\right)}{2\,(\alpha-\beta\rho)-1}\right)\left(\delta_0 + \frac{2\sigma^2}{C_l^2}\right) + \frac{2^{1+(1-\beta)\rho+\alpha}\sigma^2 C_\gamma}{\mu C_\rho^{1-\beta} t^{(1-\beta)\rho+\alpha}},$$

if $\rho \geq 0$. For $\rho < 0$, we have

$$\delta_t \leq \exp\left(-\frac{\mu C_\gamma t^{1-\alpha}}{2^{1-\alpha}}\right)\exp\left(\frac{4\alpha C_\gamma^2 C_\rho^{2\beta}\left(2C_l^2+C_\nabla^2\right)}{2\alpha-1}\right)\left(\delta_0+\frac{2\sigma^2}{C_l^2}\right)+\frac{2^{1+\alpha}\sigma^2 C_\gamma}{\mu t^\alpha}.$$

We can write this in terms of N_t instead of t using

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

for $\rho \geq 0$. Thus, $t \geq \left(\frac{N_t}{2C_\rho}\right)^{\frac{1}{1+\rho}}$ if $\rho \geq 0$. Hence, in terms of N_t, we have the following:

$$\delta_t \leq \exp\left(-\frac{\mu C_\gamma N_t^{\frac{1+\beta\rho-\alpha}{1+\rho}}}{2^{\frac{(2+\rho)(1+\beta\rho-\alpha)}{1+\rho}} C_\rho^{\frac{1-\alpha-\beta}{1+\rho}}}\right)\exp\left(\frac{4\left(\alpha-\beta\rho\right)C_\gamma^2 C_\rho^{2\beta}\left(2C_l^2+C_\nabla^2\right)}{2\left(\alpha-\beta\rho\right)-1}\right)\left(\delta_0+\frac{2\sigma^2}{C_l^2}\right)+\frac{2^{1+\frac{(2+\rho)((1-\beta)\rho+\alpha)}{1+\rho}}\sigma^2 C_\gamma}{\mu C_\rho^{\frac{1-\alpha-\beta}{1+\rho}} N_t^{\frac{(1-\beta)\rho+\alpha}{1+\rho}}}.$$

Similarly, for $\rho < 0$, as $n_t \leq C_\rho$, we have $N_t \leq C_\rho t$, i.e, $t \geq N_t/C_\rho$, giving us

$$\delta_t \leq \exp\left(-\frac{\mu C_\gamma N_t^{1-\alpha}}{2^{1-\alpha}C_\rho^{1-\alpha}}\right)\exp\left(\frac{4\alpha C_\gamma^2 C_\rho^{2\beta}\left(2C_l^2+C_\nabla^2\right)}{2\alpha-1}\right)\left(\delta_0+\frac{2\sigma^2}{C_l^2}\right)+\frac{2^{1+\alpha}\sigma^2 C_\gamma C_\rho^\alpha}{\mu N_t^\alpha}.$$

Combining the two ρ-cases (using the notation $\tilde{\rho}=\rho\mathbb{1}_{\{\rho\geq 1\}}$), yields

$$\delta_t \leq \exp\left(-\frac{\mu C_\gamma N_t^{1-\phi}}{2^{(2+\rho)(1-\phi)}C_\rho^{1-\beta-\phi}}\right)\left(\delta_0+\frac{2\sigma^2}{C_l^2}\right)\pi_v+\frac{2^{1+(2+\rho)\phi}\sigma^2 C_\gamma}{\mu C_\rho^{(1-\beta)\mathbb{1}_{\{\rho\geq 0\}}-\phi}N_t^\phi},$$

where $\phi=\frac{(1-\beta)\tilde{\rho}+\alpha}{1+\tilde{\rho}}$ and $\pi_v=\exp\left(\frac{4(\alpha-\beta\tilde{\rho})C_\gamma^2 C_\rho^{2\beta}\left(2C_l^2+C_\nabla^2\right)}{2(\alpha-\beta\tilde{\rho})-1}\right)$ is a finite constant. □

6.2. *Proofs for Section 3*

Proof of Lemma 1. We will now derive the recursive step sequence for the fourth-order moment using the same arguments as for the second-order moments in (15) (See proof of Theorem 1). Taking the square on both sides of (11) yields,

$$\begin{aligned}
\|\theta_t-\theta^*\|^4 &= \left(\|\theta_{t-1}-\theta^*\|^2+\frac{\gamma_t^2}{n_t^2}\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2-\frac{2\gamma_t}{n_t}\left\langle\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right),\theta_{t-1}-\theta^*\right\rangle\right)^2\\
&=\|\theta_{t-1}-\theta^*\|^4+\frac{\gamma_t^4}{n_t^4}\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^4+\frac{4\gamma_t^2}{n_t^2}\left\langle\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right),\theta_{t-1}-\theta^*\right\rangle^2\\
&\quad+\frac{2\gamma_t^2}{n_t^2}\|\theta_{t-1}-\theta^*\|^2\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2-\frac{4\gamma_t}{n_t}\|\theta_{t-1}-\theta^*\|^2\left\langle\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right),\theta_{t-1}-\theta^*\right\rangle\\
&\quad-\frac{4\gamma_t^3}{n_t^3}\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2\left\langle\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right),\theta_{t-1}-\theta^*\right\rangle,
\end{aligned}$$

using that $(x+y-z)^2 = x^2+y^2+z^2+2xy-2xz-2yz$. Taking $\mathbb{E}\left[\cdot|\mathcal{F}_{t-1}\right]$ on both sides of the equality gives us

$$\begin{aligned}
\mathbb{E}\left[\|\theta_t-\theta^*\|^4\middle|\mathcal{F}_{t-1}\right] &= \|\theta_{t-1}-\theta^*\|^4 + \frac{\gamma_t^4}{n_t^4}\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{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 l_{t,i}\left(\theta_{t-1}\right),\theta_{t-1}-\theta^*\right\rangle^2\middle|\mathcal{F}_{t-1}\right] \\
&\quad + \frac{2\gamma_t^2}{n_t^2}\|\theta_{t-1}-\theta^*\|^2\,\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{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\,\mathbb{E}\left[\left\langle\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right),\theta_{t-1}-\theta^*\right\rangle\middle|\mathcal{F}_{t-1}\right] \\
&\quad - \frac{4\gamma_t^3}{n_t^3}\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2\left\langle\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right),\theta_{t-1}-\theta^*\right\rangle\middle|\mathcal{F}_{t-1}\right] \\
&\le \|\theta_{t-1}-\theta^*\|^4 + \frac{\gamma_t^4}{n_t^4}\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{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 l_{t,i}\left(\theta_{t-1}\right),\theta_{t-1}-\theta^*\right\rangle^2\middle|\mathcal{F}_{t-1}\right] \\
&\quad + \frac{2\gamma_t^2}{n_t^2}\|\theta_{t-1}-\theta^*\|^2\,\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{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}\left\langle\mathbb{E}\left[\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\middle|\mathcal{F}_{t-1}\right],\theta_{t-1}-\theta^*\right\rangle \\
&\quad + \frac{4\gamma_t^3}{n_t^3}\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2\left\langle\sum_{i=1}^{n_t}\nabla_\theta l_{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 A1, we have

$$\left\langle\mathbb{E}\left[\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\middle|\mathcal{F}_{t-1}\right],\theta_{t-1}-\theta^*\right\rangle = \left\langle\nabla_\theta L\left(\theta_{t-1}\right),\theta_{t-1}-\theta^*\right\rangle \ge \mu\|\theta_{t-1}-\theta^*\|^2,$$

as L is μ-strongly convex. Combining this with the Cauchy-Schwarz inequality $\langle x,y\rangle \le \|x\|\,\|y\|$, we obtain the simplified expression:

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

Next, recall Young's Inequality, i.e., for any $a_t, b_t, c_t > 0$ we have $b_t c_t \le \frac{1}{2}a_t b_t^2 + \frac{1}{2a_t}c_t^2$, giving us

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

with $a_t = \frac{\gamma_t}{n_t\|\theta_{t-1}-\theta^*\|}$, $b_t = \left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2$ and $c_t = \left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|$. Using this inequality, we obtain:

$$\mathbb{E}\left[\|\theta_t-\theta^*\|^4\middle|\mathcal{F}_{t-1}\right] \le (1-4\mu\gamma_t)\|\theta_{t-1}-\theta^*\|^4 + \frac{3\gamma_t^4}{n_t^4}\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^4\middle|\mathcal{F}_{t-1}\right] + \frac{8\gamma_t^2}{n_t^2}\|\theta_{t-1}-\theta^*\|^2\,\mathbb{E}\left[\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2\middle|\mathcal{F}_{t-1}\right]. \tag{19}$$

To bound the second and fourth-order terms in (19), we would need to study the recursive sequences: firstly, utilizing the Lipschitz continuity of $\nabla_\theta l_{t,i}$, together with Assumption A4, A5, and that θ_{t-1} is $\mathcal{F}_{t-1}$-measurable (Assumption A1), we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^p\middle|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\left\|\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)-\nabla_\theta l_{t,i}\left(\theta^*\right)+\nabla_\theta l_{t,i}\left(\theta^*\right)\right\|^p\middle|\mathcal{F}_{t-1}\right] \\
&\le 2^{p-1}\left[\mathbb{E}\left[\left\|\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)-\nabla_\theta l_{t,i}\left(\theta^*\right)\right\|^p\middle|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[\left\|\nabla_\theta l_{t,i}\left(\theta^*\right)\right\|^p\middle|\mathcal{F}_{t-1}\right]\right] \\
&\le 2^{p-1}\left[C_l^p\|\theta_{t-1}-\theta^*\|^p + \tau^p\right],
\end{aligned} \tag{20}$$

for any $p \in [1,4]$ using the bound $\|x+y\|^p \leq 2^{p-1}\left(\|x\|^p + \|y\|^p\right)$. Thus, we can bound the second-order term in (19) by

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{i=1}^{t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2\middle|\mathcal{F}_{t-1}\right] &\leq \left[2C_l^2 n_t + C_\nabla^2(n_t-1)n_t\right]\|\theta_{t-1}-\theta^*\|^2 + 2\tau^2 n_t \\ &\leq \left[2C_l^2 n_t + C_\nabla^2 n_t^2 \mathbb{1}_{\{n_t>1\}}\right]\|\theta_{t-1}-\theta^*\|^2 + 2\tau^2 n_t, \end{aligned} \tag{21}$$

following the same steps in the proof of Theorem 1 (but with use of (20)). Bounding the fourth-order term is a bit heavier computationally, but let us recall that $\|\sum_i x_i\|^2 = \sum_i \|x_i\|^2 + \sum_{i\neq j}\langle x_i, x_j\rangle$. Then, we have, since $(x+y)^2 \leq 2x^2 + 2y^2$, that

$$\begin{aligned}\left\|\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^4 &= \left(\sum_{i=1}^{n_t}\left\|\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2 + \sum_{i\neq j}^{n_t}\left\langle\nabla_\theta l_{t,i}\left(\theta_{t-1}\right), \nabla_\theta l_{t,j}\left(\theta_{t-1}\right)\right\rangle\right)^2 \\ &\leq 2\left(\sum_{i=1}^{n_t}\left\|\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2\right)^2 + 2\left(\sum_{i\neq j}^{n_t}\left\langle\nabla_\theta l_{t,i}\left(\theta_{t-1}\right), \nabla_\theta l_{t,j}\left(\theta_{t-1}\right)\right\rangle\right)^2 \\ &\leq 2\left(\sum_{i=1}^{n_t}\left\|\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2\right)^2 + 4\left(\sum_{i<j}^{n_t}\left\langle\nabla_\theta l_{t,i}\left(\theta_{t-1}\right), \nabla_\theta l_{t,j}\left(\theta_{t-1}\right)\right\rangle\right)^2. \end{aligned} \tag{22}$$

For the first term of (22), we have

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{n_1}\left\|\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2\right)^2\middle|\mathcal{F}_{t-1}\right] &= \sum_{i=1}^{n_t}\mathbb{E}\left[\left\|\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^4\middle|\mathcal{F}_{t-1}\right] + \sum_{i\neq j}^{n_t}\mathbb{E}\left[\left\|\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^2\left\|\nabla_\theta l_{t,j}\left(\theta_{t-1}\right)\right\|^2\middle|\mathcal{F}_{t-1}\right] \\ &\leq 8n_t\left[C_l^4\|\theta_{t-1}-\theta^*\|^4 + \tau^4\right] + 4n_t\left(n_t-1\right)\left[C_l^2\|\theta_{t-1}-\theta^*\|^2 + \tau^2\right]^2 \\ &\leq 8n_t\left[C_l^4\|\theta_{t-1}-\theta^*\|^4 + \tau^4\right] + 4n_t^2\mathbb{1}_{\{n_t>1\}}\left[C_l^2\|\theta_{t-1}-\theta^*\|^2 + \tau^2\right]^2, \end{aligned}$$

using the bound from (20), and that $\mathcal{F}_{t-1} \subseteq \mathcal{F}_{t-1,i} \subset \mathcal{F}_{t-1,j}$ for all $0 \leq i < j$. To bound the second term of (22), we ease notation by denoting $\nabla_\theta l_{t,i}\left(\theta_{t-1}\right)$ by v_i, giving us

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

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

$$\mathbb{E}\left[A|\mathcal{F}_{t-1}\right] \leq \sum_{i<j}^{n_t}\mathbb{E}\left[\|v_i\|^2\left\|v_j\right\|^2\middle|\mathcal{F}_{t-1}\right] \leq 2n_t\left(n_t-1\right)\left[C_l^2\|\theta_{t-1}-\theta^*\|^2 + \tau^2\right]^2 \leq 2n_t^2\mathbb{1}_{\{n_t>1\}}\left[C_l^2\|\theta_{t-1}-\theta^*\|^2 + \tau^2\right]^2,$$

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

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

using the Cauchy-Schwarz inequality and the bound in (20). In the same way, as $j \neq l$ includes $(i, j) \neq (k, l)$, we can rewrite C as

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

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

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

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

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

Thus, the fourth-order term of (19), is bounded by

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

which can be simplified to

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{i=1}^{n_t} \nabla_\theta l_{t,i}\left(\theta_{t-1}\right)\right\|^4 \middle| \mathcal{F}_{t-1}\right] \leq& \left[16 C_l^4 n_t + 16 C_l^4 n_t^2 \mathbb{1}_{\{n_t>1\}} + 12 C_\nabla^2 C_l^2 n_t^3 \mathbb{1}_{\{n_t>1\}} + 4 C_\nabla^4 n_t^4 \mathbb{1}_{\{n_t>1\}}\right] \left\|\theta_{t-1} - \theta^*\right\|^4 \\
&+ \left[32 C_l^2 \tau^2 n_t^2 \mathbb{1}_{\{n_t>1\}} + 12 C_\nabla^2 \tau^2 n_t^3 \mathbb{1}_{\{n_t>1\}}\right] \left\|\theta_{t-1} - \theta^*\right\|^2 + 16 \tau^4 n_t + 16 \tau^4 n_t^2 \mathbb{1}_{\{n_t>1\}}. \qquad (23)
\end{aligned}$$

Combining the bound from (21) and (23) into (19), we obtain the recursive relation for the fourth-order moment:

$$\begin{aligned}\mathbb{E}\left[\|\theta_t-\theta^*\|^4\middle|\mathcal{F}_{t-1}\right] \leq &\Big[1-4\mu\gamma_t+8C_\nabla^2\mathbb{1}_{\{n_t>1\}}\gamma_t^2+16C_l^2n_t^{-1}\gamma_t^2+48C_l^4n_t^{-3}\gamma_t^4+48C_l^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4+36C_\nabla^2C_l^2n_t^{-1}\mathbb{1}_{\{n_t>1\}}\gamma_t^4\\ &+12C_\nabla^4\mathbb{1}_{\{n_t>1\}}\gamma_t^4\Big]\|\theta_{t-1}-\theta^*\|^4+\Big[16\tau^2n_t^{-1}\gamma_t^2+96C_l^2\tau^2n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4+36C_\nabla^2\tau^2n_t^{-1}\mathbb{1}_{\{n_t>1\}}\gamma_t^4\Big]\|\theta_{t-1}-\theta^*\|^2\\ &+48\tau^4n_t^{-3}\gamma_t^4+48\tau^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4.\end{aligned}$$

By Young's inequality, we have $2C_\nabla^2C_l^2 \leq n_tC_\nabla^4+n_t^{-1}C_l^4$, such that

$$\begin{aligned}\mathbb{E}\left[\|\theta_t-\theta^*\|^4\middle|\mathcal{F}_{t-1}\right] \leq &\Big[1-4\mu\gamma_t+8C_\nabla^2\mathbb{1}_{\{n_t>1\}}\gamma_t^2+16C_l^2n_t^{-1}\gamma_t^2+48C_l^4n_t^{-3}\gamma_t^4+66C_l^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4+30C_\nabla^4\mathbb{1}_{\{n_t>1\}}\gamma_t^4\Big]\|\theta_{t-1}-\theta^*\|^4\\ &+\Big[16\tau^2n_t^{-1}\gamma_t^2+96C_l^2\tau^2n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4+36C_\nabla^2\tau^2n_t^{-1}\mathbb{1}_{\{n_t>1\}}\gamma_t^4\Big]\|\theta_{t-1}-\theta^*\|^2\\ &+48\tau^4n_t^{-3}\gamma_t^4+48\tau^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4.\end{aligned}$$

Likewise, utilising Young's inequality several times, we have $16\tau^2n_t^{-1}\gamma_t^2\|\theta_{t-1}-\theta^*\|^2 \leq 2\mu\gamma_t\|\theta_t-\theta^*\|^4+32\tau^4\mu^{-1}n_t^{-2}\gamma_t^3$, $2C_l^2\tau^2n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4\|\theta_{t-1}-\theta^*\|^2 \leq C_l^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4\|\theta_t-\theta^*\|^4+\tau^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4$, and that $2C_\nabla^2\tau^2n_t^{-1}\mathbb{1}_{\{n_t>1\}}\gamma_t^4\|\theta_{t-1}-\theta^*\|^2 \leq C_\nabla^4\mathbb{1}_{\{n_t>1\}}\gamma_t^4\|\theta_t-\theta^*\|^4+\tau^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4$, giving us

$$\begin{aligned}\mathbb{E}\left[\|\theta_t-\theta^*\|^4\middle|\mathcal{F}_{t-1}\right] \leq &\Big[1-2\mu\gamma_t+8C_\nabla^2\mathbb{1}_{\{n_t>1\}}\gamma_t^2+16C_l^2n_t^{-1}\gamma_t^2+48C_l^4n_t^{-3}\gamma_t^4+114C_l^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4+48C_\nabla^4\mathbb{1}_{\{n_t>1\}}\gamma_t^4\Big]\|\theta_{t-1}-\theta^*\|^4\\ &+\frac{32\tau^4n_t^{-2}\gamma_t^3}{\mu}+48\tau^4n_t^{-3}\gamma_t^4+114\tau^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4. \qquad (24)\end{aligned}$$

Taking, the expectation on both sides of the inequality (24) yields the recursive relation for the fourth-order moment:

$$\begin{aligned}\Delta_t \leq &\Big[1-2\mu\gamma_t+8C_\nabla^2\mathbb{1}_{\{n_t>1\}}\gamma_t^2+16C_l^2n_t^{-1}\gamma_t^2+48C_l^4n_t^{-3}\gamma_t^4+114C_l^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4+48C_\nabla^4\mathbb{1}_{\{n_t>1\}}\gamma_t^4\Big]\Delta_{t-1}\\ &+\frac{32\tau^4n_t^{-2}\gamma_t^3}{\mu}+48\tau^4n_t^{-3}\gamma_t^4+114\tau^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^4. \qquad (25)\end{aligned}$$

with $\Delta_t = \mathbb{E}\left[\|\theta_t-\theta^*\|^4\right]$ for some $\Delta_0 \geq 0$.

Upper bounding the fourth-order moment: Before inserting a specific step sequence, we first write a general upper bound for Δ_t. By Proposition A.5, we achieve the (upper) bound of Δ_t, given as

$$\Delta_t \leq \exp\left(-\mu\sum_{i=t/2}^{t}\gamma_i\right)\exp\left(2\sum_{i=1}^{t}\eta_i\gamma_i\right)\left(\Delta_0+2\max_{1\leq i\leq t}\frac{\nu_i}{\eta_i}\right)+\frac{1}{\mu}\max_{t/2\leq i\leq t}\nu_i,$$

by setting $\eta_t = 8C_\nabla^2\mathbb{1}_{\{n_t>1\}}\gamma_t+16C_l^2n_t^{-1}\gamma_t+48C_l^4n_t^{-3}\gamma_t^3+114C_l^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^3+48C_\nabla^4\mathbb{1}_{\{n_t>1\}}\gamma_t^3$ and $\nu_t = \frac{32\tau^4n_t^{-2}\gamma_t^2}{\mu}+48\tau^4n_t^{-3}\gamma_t^3+114\tau^4n_t^{-2}\mathbb{1}_{\{n_t>1\}}\gamma_t^3$. Before plugging these terms into the bound, we first note that $\exp\left(2\sum_{i=1}^{t}\eta_i\gamma_i\right)$ can be given by

$$\underbrace{\exp\left(32C_l^2\sum_{i=1}^{t}\frac{\gamma_i^2}{n_i}\right)\exp\left(96C_l^4\sum_{i=1}^{t}\frac{\gamma_i^4}{n_i^3}\right)\exp\left(228C_l^4\sum_{i=1}^{t}\frac{\mathbb{1}_{\{n_i>1\}}\gamma_i^4}{n_i^2}\right)\exp\left(16C_\nabla^2\sum_{i=1}^{t}\mathbb{1}_{\{n_i>1\}}\gamma_i^2\right)\exp\left(96C_\nabla^4\sum_{i=1}^{t}\mathbb{1}_{\{n_i>1\}}\gamma_i^4\right)}_{=\Pi}. \qquad (26)$$

Next, we can simplify the term $\max_{1\leq i\leq t}\frac{\nu_i}{\eta_i}$ as follows:

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

Combining what we have above, we obtain the inequality:

$$\Delta_t \leq \exp\left(-\mu\sum_{i=t/2}^{t}\gamma_i\right)\left(\Delta_0+\frac{2\tau^4}{C_l^4}+\frac{4\tau^4\gamma_1}{\mu C_l^2n_1}\right)\Pi+\frac{32\tau^4}{\mu^2}\max_{t/2\leq i\leq t}\frac{\gamma_i^2}{n_i^2}+\frac{48\tau^4}{\mu}\max_{t/2\leq i\leq t}\frac{\gamma_i^3}{n_i^3}+\frac{114\tau^4}{\mu}\max_{t/2\leq i\leq t}\frac{\gamma_i^3\mathbb{1}_{\{n_i>1\}}}{n_i^2}. \qquad (27)$$

□

Proof of Theorem 2. Following Polyak and Juditsky (1992), we rewrite (3) to

$$\theta_t = \theta_{t-1} - \frac{\gamma_t}{n_t}\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right) \iff \frac{1}{n_t}\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right) = \frac{1}{\gamma_t}\left(\theta_{t-1}-\theta_t\right) \iff \nabla_\theta l_t\left(\theta_{t-1}\right) = \frac{1}{\gamma_t}\left(\theta_{t-1}-\theta_t\right), \tag{28}$$

as $n_t^{-1}\sum_{i=1}^{n_t}\nabla_\theta l_{t,i}\left(\theta_{t-1}\right) = \nabla_\theta l_t\left(\theta_{t-1}\right)$. Note $\nabla_\theta l_t\left(\theta_{t-1}\right) \approx \nabla_\theta l_t\left(\theta^*\right) + \nabla_\theta^2 l_t\left(\theta^*\right)\left(\theta_{t-1}-\theta^*\right)$, and that $\nabla_\theta l_t\left(\theta^*\right)$ and $\nabla_\theta l_t\left(\theta\right) - \nabla_\theta L\left(\theta\right)$ behaves almost like an i.i.d. sequences with zero mean. Thus, $\bar{\theta}_t - \theta^*$ behaves like $-\nabla_\theta L\left(\theta^*\right)^{-1} N_t^{-1}\sum_{i=1}^{t} n_i \nabla_\theta l_i\left(\theta^*\right)$ leading to a bound in $O(\sqrt{N_t})$. Observe that

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

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

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

As $\left(\nabla_\theta l_{t,i}\left(\theta^*\right)\right)$ is a square-integrable martingale increment sequences on $\mathbb{R}^d$ (Assumption A1), we have

$$\mathbb{E}\left[\left\|\nabla_\theta^2 L\left(\theta^*\right)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i \nabla_\theta l_i\left(\theta^*\right)\right\|^2\right] = \mathbb{E}\left[\left\|\nabla_\theta^2 L\left(\theta^*\right)^{-1}\frac{1}{N_t}\sum_{i=1}^{t}\sum_{j=1}^{n_i}\nabla_\theta l_{i,j}\left(\theta^*\right)\right\|^2\right] \leq \frac{\mathrm{Tr}\left[\nabla_\theta^2 L(\theta^*)^{-1}\Sigma\nabla_\theta^2 L(\theta^*)^{-1}\right]}{N_t}, \tag{29}$$

using Assumption A6. To ease notation, we denote $\mathrm{Tr}[\nabla_\theta^2 L(\theta^*)^{-1}\Sigma\nabla_\theta^2 L(\theta^*)^{-1}]$ by Λ. Next, note that for all $t \geq 1$, we have the relation in (28), giving us

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

leading to

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

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

$$\left(\mathbb{E}\left[\left\|\nabla_\theta^2 L\left(\theta^*\right)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i \nabla_\theta l_i\left(\theta_{i-1}\right)\right\|^2\right]\right)^{\frac{1}{2}} \leq \frac{1}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{\frac{1}{2}}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| + \frac{n_t}{N_t\gamma_t\mu}\delta_t^{\frac{1}{2}} + \frac{n_1}{N_t\gamma_1\mu}\delta_0^{\frac{1}{2}}. \tag{30}$$

For the martingale term, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla_\theta^2 L(\theta^*)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i\left[\nabla_\theta l_i(\theta_{i-1})-\nabla_\theta l_i(\theta^*)-\nabla_\theta L(\theta_{i-1})\right]\right\|^2\right] &\leq \frac{1}{N_t^2\mu^2}\sum_{i=1}^{t} n_i^2\mathbb{E}\left[\|\nabla_\theta l_i(\theta_{i-1})-\nabla_\theta l_i(\theta^*)\|^2\right]\\ &= \frac{1}{N_t^2\mu^2}\sum_{i=1}^{t}\mathbb{E}\left[\left\|\sum_{j=1}^{n_i}\nabla_\theta l_{i,j}(\theta_{i-1})-\nabla_\theta l_{i,j}(\theta^*)\right\|^2\right]\\ &\leq \frac{1}{N_t^2\mu^2}\sum_{i=1}^{t}\sum_{j=1}^{n_i}\left(\mathbb{E}\left[\left\|\nabla_\theta l_{i,j}(\theta_{i-1})-\nabla_\theta l_{i,j}(\theta^*)\right\|^2\right]\right)^{\frac{1}{2}}\\ &\leq \frac{C_l^2}{N_t^2\mu^2}\sum_{i=1}^{t} n_i\delta_{i-1}, \end{aligned} \tag{31}$$

by the Cauchy-Schwarz inequality and Assumption A4. For all $t \geq 1$, the rest term is directly bounded by Assumption A7:

$$\left(\mathbb{E}\left[\left\|\nabla_\theta^2 L(\theta^*)^{-1}\frac{1}{N_t}\sum_{i=1}^{t} n_i\left[\nabla_\theta L(\theta_{i-1})-\nabla_\theta^2 L(\theta^*)(\theta_{i-1}-\theta^*)\right]\right\|^2\right]\right)^{\frac{1}{2}} \leq \frac{C_\delta}{N_t\mu}\sum_{i=1}^{t} n_i\Delta_{i-1}^{\frac{1}{2}}, \tag{32}$$

with the notion $\Delta_t = \mathbb{E}[\|\theta_t-\theta^*\|^4]$. Finally, combining the terms (29)-(32), gives us

$$\bar{\delta}_t^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{1}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{1/2}\left|\frac{n_{i+1}}{\gamma_{i+1}}-\frac{n_i}{\gamma_i}\right| + \frac{n_t}{N_t\gamma_t\mu}\delta_t^{1/2} + \frac{n_1}{N_t\gamma_1\mu}\delta_0^{1/2} + \frac{C_l}{N_t\mu}\left(\sum_{i=1}^{t} n_i\delta_{i-1}\right)^{1/2} + \frac{C_\delta}{N_t\mu}\sum_{i=1}^{t} n_i\Delta_{i-1}^{1/2}, \tag{33}$$

where $\bar{\delta}_t = \mathbb{E}[\|\bar{\theta}_t-\theta^*\|^2]$, which can be simplified into (9) by shifting the indices and collecting the δ_0 terms. □

Proof of Corollary 3. As $n_t = C_\rho$ for all $t \geq 1$, we simplify the bound for $\bar{\delta}_t$ in (9) to

$$\bar{\delta}_t^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{C_\rho}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{1/2}\left|\frac{1}{\gamma_{i+1}}-\frac{1}{\gamma_i}\right| + \frac{C_\rho}{N_t\gamma_t\mu}\delta_t^{1/2} + \frac{C_\rho}{N_t\mu}\left(\frac{1}{\gamma_1}+C_l\right)\delta_0^{1/2} + \frac{C_l C_\rho^{\frac{1}{2}}}{N_t\mu}\left(\sum_{i=1}^{t-1}\delta_i\right)^{1/2} + \frac{C_\delta C_\rho}{N_t\mu}\sum_{i=0}^{t-1}\Delta_i^{1/2}. \tag{34}$$

Second-order moment: The second-order moment δ_t is bounded by Corollary 1. As we work in terms of t, we use inequality (17), given as

$$\delta_t \leq \exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1-\alpha}}{2^{1-\alpha}}\right)\pi_c' + \frac{2^{1+\alpha}\sigma^2 C_\gamma}{\mu C_\rho^{1-\beta}t^\alpha},$$

where $\pi_c' = \left(\delta_0 + \frac{2\sigma^2}{C_l^2}\right)\pi_c$ is a finite constant independent of t.

Fourth-order moment: The fourth-order moment Δ_t from Lemma 1 can be simplified to:

$$\begin{aligned}\Delta_t &\leq \exp\left(-\mu\sum_{i=t/2}^{t}\gamma_i\right)\left(\Delta_0 + \frac{2\tau^4}{C_l^4} + \frac{4\tau^4 C_\gamma}{\mu C_l^2 C_\rho^{1-\beta}}\right)\Pi + \frac{1}{\mu}\left(\frac{32\tau^4}{\mu C_\rho^2}\max_{t/2\leq i\leq t}\gamma_i^2 + \frac{48\tau^4}{C_\rho^3}\max_{t/2\leq i\leq t}\gamma_i^3 + \frac{114\tau^4\mathbb{1}_{\{C_\rho>1\}}}{C_\rho^2}\max_{t/2\leq i\leq t}\gamma_i^3\right)\\ &\leq \exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1-\alpha}}{2^{1-\alpha}}\right)\left(\Delta_0 + \frac{2\tau^4}{C_l^4} + \frac{4\tau^4 C_\gamma}{\mu C_l^2 C_\rho^{1-\beta}}\right)\Pi + \frac{1}{\mu}\left(\frac{2^{2\alpha}32\tau^4 C_\gamma^2 C_\rho^{2\beta}}{\mu C_\rho^2 t^{2\alpha}} + \frac{2^{3\alpha}48\tau^4 C_\gamma^3 C_\rho^{3\beta}}{C_\rho^3 t^{3\alpha}} + \frac{2^{3\alpha}114\tau^4 C_\gamma^3 C_\rho^{3\beta}\mathbb{1}_{\{C_\rho>1\}}}{C_\rho^2 t^{3\alpha}}\right),\end{aligned}$$

using that $\gamma_t = C_\gamma C_\rho^\beta t^{-\alpha}$ is decreasing as $\alpha \in (1/2, 1)$. Regarding Π in (26), we obtain

$$\Pi = \exp\left(\frac{32C_l^2}{C_\rho}\sum_{i=1}^{t}\gamma_i^2\right)\exp\left(\frac{96C_l^4}{C_\rho^3}\sum_{i=1}^{t}\gamma_i^4\right)\exp\left(\frac{228C_l^4\mathbb{1}_{\{C_\rho>1\}}}{C_\rho^2}\sum_{i=1}^{t}\gamma_i^4\right)\exp\left(16C_\nabla^2\mathbb{1}_{\{C_\rho>1\}}\sum_{i=1}^{t}\gamma_i^2\right)\exp\left(96C_\nabla^4\mathbb{1}_{\{C_\rho>1\}}\sum_{i=1}^{t}\gamma_i^4\right)$$
$$\leq \underbrace{\exp\left(\frac{64\alpha C_l^2C_\gamma^2C_\rho^{2\beta}}{(2\alpha-1)\,C_\rho}\right)\exp\left(\frac{192C_l^4C_\gamma^4C_\rho^{4\beta}}{C_\rho^3}\right)\exp\left(\frac{456C_l^4C_\gamma^4C_\rho^{4\beta}\mathbb{1}_{\{C_\rho>1\}}}{C_\rho^2}\right)\exp\left(\frac{32\alpha C_\nabla^2C_\gamma^2C_\rho^{2\beta}\mathbb{1}_{\{C_\rho>1\}}}{2\alpha-1}\right)\exp\left(192C_\nabla^4C_\gamma^4C_\rho^{4\beta}\mathbb{1}_{\{C_\rho>1\}}\right)}_{=\Pi_c},$$

using $\sum_{i=1}^{t} i^{-2\alpha} \leq \frac{2\alpha}{2\alpha-1}$ and $\sum_{i=1}^{t} i^{-4\alpha} \leq 2$. Note that Π_c is a finite constant, independent of t opposite Π.

Bounding $\frac{C_\rho}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{\frac{1}{2}}\left|\frac{1}{\gamma_{i+1}}-\frac{1}{\gamma_i}\right|$: Remarking that $|\gamma_{t+1}^{-1}-\gamma_t^{-1}| \leq C_\gamma^{-1}C_\rho^{-\beta}\alpha t^{\alpha-1}$, one has (since $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$),

$$\frac{C_\rho}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{\frac{1}{2}}\left|\frac{1}{\gamma_{i+1}}-\frac{1}{\gamma_i}\right| \leq \frac{C_\rho^{1-\beta}\alpha}{C_\gamma\mu N_t}\sum_{i=1}^{t} i^{\alpha-1}\left(\exp\left(-\frac{\mu C_\gamma C_\rho^\beta i^{1-\alpha}}{2^{2-\alpha}}\right)\sqrt{\pi_c'}+\frac{2^{\frac{1+\alpha}{2}}\sigma\sqrt{C_\gamma}}{\sqrt{\mu}C_\rho^{\frac{1-\beta}{2}}i^{\alpha/2}}\right). \tag{35}$$

For simplicity, let us denote

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

as $\alpha < 1$. Thus, the first part of (35) is bounded as follows:

$$\frac{C_\rho^{1-\beta}\alpha\sqrt{\pi_c'}}{C_\gamma\mu N_t}\sum_{i=1}^{t} i^{\alpha-1}\exp\left(-\frac{\mu C_\gamma C_\rho^\beta i^{1-\alpha}}{2^{2-\alpha}}\right) \leq \frac{C_\rho^{1-\beta}\alpha\sqrt{\pi_c'}A_\infty}{C_\gamma\mu N_t}.$$

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 (35) 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

$$\frac{C_\rho}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{\frac{1}{2}}\left|\frac{1}{\gamma_{i+1}}-\frac{1}{\gamma_i}\right| \leq \frac{C_\rho^{1-\beta}\alpha\sqrt{\pi_c'}A_\infty}{C_\gamma\mu N_t}+\frac{2^{\frac{3+\alpha}{2}}\sigma C_\rho^{\frac{1-\alpha-\beta}{2}}}{\sqrt{C_\gamma}\mu^{3/2}N_t^{1-\alpha/2}}. \tag{36}$$

Bounding $\frac{C_\rho}{N_t\gamma_t\mu}\delta_t^{\frac{1}{2}}$: Similarly, we have

$$\frac{C_\rho}{N_t\gamma_t\mu}\delta_t^{\frac{1}{2}} \leq \frac{C_\rho^{1-\alpha-\beta}}{C_\gamma\mu N_t^{1-\alpha}}\left(\exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1-\alpha}}{2^{2-\alpha}}\right)\sqrt{\pi_c'}+\frac{2^{\frac{1+\alpha}{2}}\sigma\sqrt{C_\gamma}}{\sqrt{\mu}C_\rho^{\frac{1-\beta}{2}}t^{\alpha/2}}\right)$$
$$= \frac{C_\rho^{1-\alpha-\beta}\sqrt{\pi_c'}A_\infty}{C_\gamma\mu N_t^{1-\alpha}}+\frac{2^{\frac{1+\alpha}{2}}C_\rho^{\frac{1-\alpha-\beta}{2}}\sigma}{\sqrt{C_\gamma}\mu^{3/2}N_t^{1-\alpha/2}}.$$

Bounding $\frac{C_l C_\rho^{\frac{1}{2}}}{N_t\mu}\left(\sum_{i=1}^{t-1}\delta_i\right)^{\frac{1}{2}}$: In a same way, one has

$$\begin{aligned}\frac{C_l C_\rho^{\frac{1}{2}}}{N_t\mu}\left(\sum_{i=1}^{t-1}\delta_i\right)^{\frac{1}{2}} \leq& \frac{C_l C_\rho^{\frac{1}{2}}}{N_t\mu}\left(\sum_{i=1}^{t-1}\exp\left(-\frac{\mu C_\gamma C_\rho^\beta i^{1-\alpha}}{2^{1-\alpha}}\right)\pi_c' + \sum_{i=1}^{t-1}\frac{2^{1+\alpha}\sigma^2 C_\gamma}{\mu C_\rho^{1-\beta} i^\alpha}\right)^{1/2}\\ \leq& \frac{C_l C_\rho^{\frac{1}{2}}}{N_t\mu}\left(A_\infty\pi_c' + \frac{2^{1+\alpha}\sigma^2 C_\gamma t^{1-\alpha}}{(1-\alpha)\mu C_\rho^{1-\beta}}\right)^{1/2}\\ \leq& \frac{C_l C_\rho^{\frac{1}{2}}\sqrt{\pi_c'}\sqrt{A_\infty}}{N_t\mu} + \frac{2^{\frac{1+\alpha}{2}} C_l\sigma\sqrt{C_\gamma}}{C_\rho^{\frac{1-\alpha-\beta}{2}}\mu^{3/2}N_t^{\frac{1+\alpha}{2}}}.\end{aligned}$$

Bounding $\frac{C_\delta C_\rho}{N_t\mu}\sum_{i=0}^{t-1}\Delta_i^{\frac{1}{2}}$: To ease notation, let us denote $\Pi_c' = \left(\Delta_0 + \frac{2\tau^4}{C_l^4} + \frac{4\tau^4 C_\gamma}{\mu C_l^2 C_\rho^{1-\beta}}\right)\Pi_c$, then

$$\begin{aligned}\frac{C_\delta C_\rho}{N_t\mu}\sum_{i=0}^{t-1}\Delta_i^{\frac{1}{2}} \leq& \frac{C_\delta C_\rho}{N_t\mu}\sum_{i=0}^{t-1}\exp\left(-\frac{\mu C_\gamma C_\rho^\beta i^{1-\alpha}}{2^{2-\alpha}}\right)\sqrt{\Pi_c'} + \frac{2^\alpha 6 C_\delta\tau^2 C_\gamma C_\rho^\beta}{N_t\mu^2}\sum_{i=1}^{t-1} i^{-\alpha} + \frac{(6+7\mathbb{1}_{\{C_\rho>1\}})2^{3\alpha/2}C_\delta\tau^2 C_\gamma^{3/2}C_\rho^{3\beta/2}}{N_t\mu^{3/2}}\sum_{i=1}^{t-1} i^{-3\alpha/2}\\ \leq& \frac{C_\delta C_\rho\sqrt{\Pi_c'}A_\infty}{N_t\mu} + \frac{2^\alpha 6 C_\delta\tau^2 C_\gamma}{C_\rho^{1-\alpha-\beta}\mu^2 N_t^\alpha} + \begin{cases}\frac{(6+7\mathbb{1}_{\{C_\rho>1\}})2^{3\alpha/2}C_\delta\tau^2 C_\gamma^{3/2}C_\rho^{3\beta/2+3\alpha/2-1}}{\mu^{3/2}N_t^{3\alpha/2}}, & \text{if } 3\alpha/2<1,\\ \frac{(6+7\mathbb{1}_{\{C_\rho>1\}})2^{3\alpha/2}C_\delta\tau^2 C_\gamma^{3/2}C_\rho^{3\beta/2}\log(N_t)}{\mu^{3/2}N_t}, & \text{if } 3\alpha/2=1,\\ \frac{3\alpha(6+7\mathbb{1}_{\{C_\rho>1\}})2^{3\alpha/2}C_\delta\tau^2 C_\gamma^{3/2}C_\rho^{3\beta/2}}{(3\alpha-2)\mu^{3/2}N_t}, & \text{if } 3\alpha/2>1.\end{cases}\end{aligned}$$

Final Bound: Thus, as a conclusion, we obtain:

$$\begin{aligned}\bar{\delta}_t^{1/2} \leq& \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{6\sigma C_\rho^{\frac{1-\alpha-\beta}{2}}}{\sqrt{C_\gamma}\mu^{3/2}N_t^{1-\alpha/2}} + \frac{C_\rho^{1-\alpha-\beta}\sqrt{\pi_c'}A_\infty}{C_\gamma\mu N_t^{1-\alpha}} + \frac{2^\alpha 6 C_\delta\tau^2 C_\gamma}{C_\rho^{1-\alpha-\beta}\mu^2 N_t^\alpha} + \frac{2^{\frac{1+\alpha}{2}}C_l\sigma\sqrt{C_\gamma}}{C_\rho^{\frac{1-\alpha-\beta}{2}}\mu^{3/2}N_t^{\frac{1+\alpha}{2}}} + \frac{C_\rho\Gamma_c}{\mu N_t}\\ &+ \frac{(6+7\mathbb{1}_{\{C_\rho>1\}})2^{3\alpha/2}C_\delta\tau^2 C_\gamma^{3/2}C_\rho^{3\beta/2}}{\mu^{3/2}}\left(\frac{C_\rho^{3\alpha/2-1}}{N_t^{3\alpha/2}}\mathbb{1}_{\{3\alpha/2<1\}} + \frac{\log(N_t)}{N_t}\mathbb{1}_{\{3\alpha/2=1\}} + \frac{3\alpha}{(3\alpha-2)N_t}\mathbb{1}_{\{3\alpha/2>1\}}\right),\end{aligned}$$

where $\Gamma_c = \left(\frac{1}{C_\gamma C_\rho^\beta} + C_l\right)\delta_0^{1/2} + \frac{C_l\sqrt{\pi_c'}\sqrt{A_\infty}}{C_\rho^{1/2}} + \frac{\sqrt{\pi_c'}A_\infty}{C_\gamma C_\rho^\beta} + C_\delta\sqrt{\Pi_c'}A_\infty$ is a finite constant, consisting of π_c', Π_c' and A_∞, given as

$$\pi_c' = \left(\delta_0 + \frac{2\sigma^2}{C_l^2}\right)\pi_c,\ \Pi_c' = \left(\Delta_0 + \frac{2\tau^4}{C_l^4} + \frac{4\tau^4 C_\gamma}{\mu C_l^2 C_\rho^{1-\beta}}\right)\Pi_c,\ \text{and } A_\infty = \sum_{i=0}^{\infty}\exp\left(-\frac{\mu C_\gamma C_\rho^\beta i^{1-\alpha}}{2^{2-\alpha}}\right). \tag{37}$$

□

Proof of Corollary 4. The steps of the proof follows the ones of Corollary 3 with the smart notation of ϕ and $\tilde{\rho}$: The bound for $\bar{\delta}_t$ in (9) is given by

$$\bar{\delta}_t^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{1}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{1/2}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| + \frac{n_t}{N_t\gamma_t\mu}\delta_t^{1/2} + \frac{n_1}{N_t\mu}\left(\frac{1}{\gamma_1} + C_l\right)\delta_0^{1/2} + \frac{C_l}{N_t\mu}\left(\sum_{i=1}^{t-1} n_{i+1}\delta_i\right)^{1/2} + \frac{C_\delta}{N_t\mu}\sum_{i=0}^{t-1} n_{i+1}\Delta_i^{1/2}, \tag{38}$$

where the learning rate is on the form $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ with $n_t = C_\rho t^\rho$.
Second-order moment: Upper bounding of the second-order moment δ_t follows by Corollary 2:

$$\delta_t \leq \exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} t^{(1-\phi)(1+\tilde{\rho})}}{2^{(1-\phi)(1+\tilde{\rho})}}\right)\pi_v' + \frac{2^{1+\phi(1+\tilde{\rho})}\sigma^2 C_\gamma}{\mu C_\rho^{(1-\beta)\mathbb{1}_{\{\rho\geq 0\}}} t^{\phi(1+\tilde{\rho})}}, \tag{39}$$

where $\pi'_v = \left(\delta_0 + \frac{2\sigma^2}{C_l^2}\right)\pi_v$.

Fourth-order moment: The fourth-order moment Δ_t from Lemma 1 can be simplified as follows:

$$\begin{aligned}\Delta_t &\leq \exp\left(-\mu\sum_{i=t/2}^{t}\gamma_i\right)\left(\Delta_0 + \frac{2\tau^4}{C_l^4} + \frac{4\tau^4\gamma_1}{\mu C_l^2 n_1}\right)\Pi + \frac{1}{\mu}\left(\frac{32\tau^4}{\mu}\max_{t/2\leq i\leq t}\frac{\gamma_i^2}{n_i^2} + 48\tau^4\max_{t/2\leq i\leq t}\frac{\gamma_i^3}{n_i^3} + 114\tau^4\max_{t/2\leq i\leq t}\frac{\gamma_i^3\mathbb{1}_{\{n_i>1\}}}{n_i^2}\right)\\ &\leq \exp\left(-\mu\sum_{i=t/2}^{t}\gamma_i\right)\left(\Delta_0 + \frac{2\tau^4}{C_l^4} + \frac{4\tau^4 C_\gamma}{\mu C_l^2}\right)\Pi + \frac{32\tau^4}{\mu^2}\max_{t/2\leq i\leq t}\frac{\gamma_i^2}{n_i^2} + \frac{162\tau^4}{\mu}\max_{t/2\leq i\leq t}\frac{\gamma_i^3}{n_i^2},\end{aligned}$$

as $n_t \geq 1$ for any $t \geq 1$ and $\beta \leq 1$. Likewise, for Π in (26), we have

$$\begin{aligned}\Pi &\leq \exp\left(16(2C_l^2 + C_\nabla^2)\sum_{i=1}^{t}\gamma_i^2\right)\exp\left(96(4C_l^4 + C_\nabla^4)\sum_{i=1}^{t}\gamma_i^4\right)\\ &\leq \underbrace{\exp\left(\frac{32(\alpha-\beta\tilde{\rho})C_\gamma^2 C_\rho^{2\beta}(2C_l^2 + C_\nabla^2)}{2(\alpha-\beta\tilde{\rho})-1}\right)\exp\left(192C_\gamma^4 C_\rho^{4\beta}(4C_l^4 + C_\nabla^4)\right)}_{=\Pi_v},\end{aligned}$$

using that $\sum_{i=1}^{t} i^{-a} \leq 2$ for $a \geq 2$. To ease notation, let Π'_v denote the finite constants $\left(\Delta_0 + \frac{2\tau^4}{C_l^4} + \frac{4\tau^4 C_\gamma}{\mu C_l^2}\right)\Pi_v$. Next, for $\rho \geq 0$, we have

$$\Delta_t \leq \exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1+\beta\rho-\alpha}}{2^{1+\beta\rho-\alpha}}\right)\Pi'_v + \frac{2^{2\alpha-2\beta\rho+2\rho}32\tau^4 C_\gamma^2 C_\rho^{2\beta}}{\mu^2 C_\rho^2 t^{2\alpha-2\beta\rho+2\rho}} + \frac{2^{3\alpha-3\beta\rho+2\rho}162\tau^4 C_\gamma^3 C_\rho^{3\beta}}{\mu C_\rho^2 t^{3\alpha-3\beta\rho+2\rho}},$$

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

$$\Delta_t \leq \exp\left(-\frac{\mu C_\gamma C_\rho^\beta t^{1-\alpha}}{2^{1-\alpha}}\right)\Pi'_v + \frac{2^{2\alpha}32\tau^4 C_\gamma^2 C_\rho^{2\beta}}{\mu^2 t^{2\alpha}} + \frac{2^{3\alpha}162\tau^4 C_\gamma^3 C_\rho^{3\beta}}{\mu t^{3\alpha}}.$$

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

$$\Delta_t \leq \exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} t^{(1-\phi)(1+\tilde{\rho})}}{2^{(1-\phi)(1+\tilde{\rho})}}\right)\Pi'_v + \frac{2^{2\phi(1+\tilde{\rho})}32\tau^4 C_\gamma^2 C_\rho^{2\beta}}{\mu^2 C_\rho^{2\mathbb{1}_{\{\rho\geq 0\}}} t^{2\phi(1+\tilde{\rho})}} + \frac{2^{3\phi(1+\tilde{\rho})-\tilde{\rho}}162\tau^4 C_\gamma^3 C_\rho^{3\beta}}{\mu C_\rho^{2\mathbb{1}_{\{\rho\geq 0\}}} t^{3\phi(1+\tilde{\rho})-\tilde{\rho}}}.$$

Going from t to N_t: 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 $t \geq \left(N_t/2C_\rho\right)^{1/(1+\rho)}$ since $N_t = C_\rho\left(t^\rho + \sum_{i=1}^{t-1} i^\rho\right) \leq C_\rho\left(t^\rho + \int_1^t x^\rho\,dx\right) \leq C_\rho\left(t^\rho + t^\rho\int_1^t dx\right) = C_\rho\left(t^\rho + t^{1+\rho}\right) \leq 2C_\rho t^{1+\rho}$ by use the integral test for convergence. Similarly, $N_t = C_\rho\sum_{i=1}^{t} i^\rho \geq C_\rho\int_0^t x^\rho\,dx = C_\rho t^{\rho+1}$, thus, $t \leq \left(N_t/C_\rho\right)^{1/(1+\rho)}$. 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 $\frac{1}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{\frac{1}{2}}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right|$: Note $n_t/\gamma_t = C_\gamma^{-1}C_\rho^{1-\beta}t^{(1-\beta)\rho+\alpha}$ for $\rho \geq 0$. Thus, by the mean value theorem, we obtain:

$$\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| \leq ((1-\beta)\rho+\alpha)\frac{C_\rho^{1-\beta}}{C_\gamma}\sup_{v\in(i,i+1)}\left|v^{(1-\beta)\rho+\alpha-1}\right| \leq \frac{((1-\beta)\rho+\alpha)\,C_\rho^{1-\beta}}{C_\gamma i^{1-(1-\beta)\rho-\alpha}}, \tag{40}$$

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

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

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

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

By using this, we obtain:

$$\frac{1}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{\frac{1}{2}}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| \leq \frac{\phi(1+\tilde{\rho})C_\rho^{1-\beta}}{N_t\mu C_\gamma}\sum_{i=1}^{t} i^{\phi(1+\tilde{\rho})-1}\left(\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} i^{(1-\phi)(1+\tilde{\rho})}}{2^{1+(1-\phi)(1+\tilde{\rho})}}\right)\sqrt{\pi'_v} + \frac{2^{\frac{1+\phi(1+\tilde{\rho})}{2}}\sigma\sqrt{C_\gamma}}{\sqrt{\mu}C_\rho^{\frac{(1-\beta)}{2}\mathbb{1}_{\{\rho\geq 0\}}} i^{\frac{\phi(1+\tilde{\rho})}{2}}}\right).$$

Next, let us denote

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

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

$$\frac{\phi(1+\tilde{\rho})C_\rho^{1-\beta}\sqrt{\pi'_v}}{N_t\mu C_\gamma}\sum_{i=1}^{t} i^{\phi(1+\tilde{\rho})-1}\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} i^{(1-\phi)(1+\tilde{\rho})}}{2^{1+(1-\phi)(1+\tilde{\rho})}}\right) \leq \frac{\phi(1+\tilde{\rho})C_\rho^{1-\beta}\sqrt{\pi'_v}A'_\infty}{N_t\mu C_\gamma}.$$

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

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

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

$$\begin{aligned}\frac{1}{N_t\mu}\sum_{i=1}^{t-1}\delta_i^{\frac{1}{2}}\left|\frac{n_{i+1}}{\gamma_{i+1}} - \frac{n_i}{\gamma_i}\right| &\leq \frac{\phi(1+\tilde{\rho})C_\rho^{1-\beta}\sqrt{\pi'_v}A'_\infty}{N_t\mu C_\gamma} + \frac{2^{\frac{3+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma}N_t^{1-\phi/2}} \\ &\leq \frac{2C_\rho^{1-\beta}\sqrt{\pi'_v}A'_\infty}{\mu C_\gamma N_t} + \frac{2^{\frac{3+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma}N_t^{1-\phi/2}}.\end{aligned}$$

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

$$\begin{aligned}\frac{n_t}{N_t\gamma_t\mu}\delta_t^{\frac{1}{2}} &\leq \frac{C_\rho^{1-\beta}t^{\phi(1+\tilde{\rho})}}{N_tC_\gamma\mu}\left(\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} t^{(1-\phi)(1+\tilde{\rho})}}{2^{1+(1-\phi)(1+\tilde{\rho})}}\right)\sqrt{\pi'_v} + \frac{2^{\frac{1+\phi(1+\tilde{\rho})}{2}}\sigma\sqrt{C_\gamma}}{\mu^{1/2}C_\rho^{\frac{1-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}} t^{\frac{\phi(1+\tilde{\rho})}{2}}}\right) \\ &= \frac{C_\rho^{1-\beta}\sqrt{\pi'_v}t^{\phi(1+\tilde{\rho})}}{N_tC_\gamma\mu}\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}} t^{(1-\phi)(1+\tilde{\rho})}}{2^{1+(1-\phi)(1+\tilde{\rho})}}\right) + \frac{2^{\frac{1+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}} t^{\frac{\phi(1+\tilde{\rho})}{2}}}{\mu^{3/2}\sqrt{C_\gamma}N_t} \\ &\leq \frac{C_\rho^{1-\phi-\beta}\sqrt{\pi'_v}A'_\infty}{\mu C_\gamma N_t^{1-\phi}} + \frac{2^{\frac{1+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma}N_t^{1-\phi/2}}.\end{aligned}$$

Bounding $\frac{n_1}{N_t\mu}\left(\frac{1}{\gamma_1} + C_l\right)\delta_0^{\frac{1}{2}}$: Inserting the definition of our learning functions, gives us

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

Bounding $\frac{C_l}{N_t\mu}\left(\sum_{i=1}^{t-1} n_{i+1}\delta_i\right)^{\frac{1}{2}}$: Following the ideas from above and using that $n_{t+1} \le 2^{\tilde{\rho}} n_t$, we get

$$\begin{aligned}
\frac{C_l}{N_t\mu}\left(\sum_{i=1}^{t-1} n_{i+1}\delta_i\right)^{\frac{1}{2}} \le& \frac{2^{\tilde{\rho}/2}C_l}{N_t\mu}\left(C_\rho \sum_{i=1}^{t} i^{\tilde{\rho}}\left(\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\ge 0\}}} i^{(1-\phi)(1+\tilde{\rho})}}{2^{(1-\phi)(1+\tilde{\rho})}}\right)\pi_v' + \frac{2^{1+\phi(1+\tilde{\rho})}\sigma^2 C_\gamma}{\mu C_\rho^{(1-\beta)\mathbb{1}_{\{\rho\ge 0\}}} i^{\phi(1+\tilde{\rho})}}\right)\right)^{\frac{1}{2}} \\
=& \frac{2^{\tilde{\rho}/2}C_l}{N_t\mu}\left(C_\rho\pi_v' \sum_{i=1}^{t} i^{\tilde{\rho}}\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\ge 0\}}} i^{(1-\phi)(1+\tilde{\rho})}}{2^{(1-\phi)(1+\tilde{\rho})}}\right) + \frac{2^{1+\phi(1+\tilde{\rho})}\sigma^2 C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\ge 0\}}}}{\mu}\sum_{i=1}^{t} i^{\beta\tilde{\rho}-\alpha}\right)^{\frac{1}{2}} \\
\le& \frac{2^{\tilde{\rho}/2}C_l}{N_t\mu}\left(C_\rho\pi_v' A_\infty' + \frac{2^{\phi(1+\tilde{\rho})}\sigma^2 C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\ge 0\}}} t^{(1-\phi)(1+\tilde{\rho})}}{\mu}\right)^{\frac{1}{2}} \\
\le& \frac{2^{\tilde{\rho}/2}C_l\sqrt{C_\rho}\sqrt{\pi_v'}\sqrt{A_\infty'}}{\mu N_t} + \frac{2^{\frac{\phi(1+\tilde{\rho})}{2}} C_l\sigma\sqrt{C_\gamma} C_\rho^{\beta/2\mathbb{1}_{\{\rho\ge 0\}}} t^{\frac{(1-\phi)(1+\tilde{\rho})}{2}}}{\mu^{3/2}N_t} \\
\le& \frac{2^{\tilde{\rho}/2}C_l\sqrt{C_\rho}\sqrt{\pi_v'}\sqrt{A_\infty'}}{\mu N_t} + \frac{2^{\frac{\phi(1+\tilde{\rho})}{2}} C_l\sigma\sqrt{C_\gamma}}{\mu^{3/2}C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\ge 0\}}} N_t^{\frac{1+\phi}{2}}}.
\end{aligned}$$

Bounding $\frac{C_\delta}{N_t\mu}\sum_{i=0}^{t-1} n_{i+1}\Delta_i^{\frac{1}{2}}$: Likewise, we get

$$\begin{aligned}
\frac{C_\delta}{N_t\mu}\sum_{i=0}^{t-1} n_{i+1}\Delta_i^{\frac{1}{2}} \le& \frac{2^{\tilde{\rho}}C_\delta C_\rho}{N_t\mu}\sum_{i=1}^{t-1} i^{\tilde{\rho}}\left(\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\ge 0\}}} i^{(1-\phi)(1+\tilde{\rho})}}{2^{(1-\phi)(1+\tilde{\rho})}}\right)\Pi_v' + \frac{2^{2\phi(1+\tilde{\rho})}32\tau^4 C_\gamma^2 C_\rho^{2\beta}}{\mu^2 C_\rho^{2\mathbb{1}_{\{\rho\ge 0\}}} i^{2\phi(1+\tilde{\rho})}} + \frac{2^{3\phi(1+\tilde{\rho})-\tilde{\rho}}162\tau^4 C_\gamma^3 C_\rho^{3\beta}}{\mu C_\rho^{2\mathbb{1}_{\{\rho\ge 0\}}} i^{3\phi(1+\tilde{\rho})-\tilde{\rho}}}\right)^{\frac{1}{2}} \\
\le& \frac{2^{\tilde{\rho}}C_\delta C_\rho}{N_t\mu}\sum_{i=1}^{t-1} i^{\tilde{\rho}}\left(\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\ge 0\}}} i^{(1-\phi)(1+\tilde{\rho})}}{2^{1+(1-\phi)(1+\tilde{\rho})}}\right)\sqrt{\Pi_v'} + \frac{2^{\phi(1+\tilde{\rho})}6\tau^2 C_\gamma C_\rho^{\beta}}{\mu C_\rho^{\mathbb{1}_{\{\rho\ge 0\}}} i^{\phi(1+\tilde{\rho})}} + \frac{2^{3\phi(1+\tilde{\rho})/2-\tilde{\rho}/2}13\tau^2 C_\gamma^{3/2} C_\rho^{3\beta/2}}{\mu^{1/2} C_\rho^{\mathbb{1}_{\{\rho\ge 0\}}} i^{3\phi(1+\tilde{\rho})/2}}\right) \\
\le& \frac{2^{\tilde{\rho}}C_\delta C_\rho\sqrt{\Pi_v'}A_\infty'}{\mu N_t} + \frac{2^{\phi(1+\tilde{\rho})+\tilde{\rho}}C_\delta\tau^2 C_\gamma C_\rho^{1+\beta}}{\mu^2 C_\rho^{\mathbb{1}_{\{\rho\ge 0\}}} N_t}\sum_{i=1}^{t-1} i^{\beta\tilde{\rho}-\alpha} + \frac{2^{3\phi(1+\tilde{\rho})/2+\tilde{\rho}/2}C_\delta\tau^2 C_\gamma^{3/2} C_\rho^{1+3\beta/2}}{\mu^{3/2} C_\rho^{\mathbb{1}_{\{\rho\ge 0\}}} N_t}\sum_{i=1}^{t-1} i^{(\beta\tilde{\rho}-\alpha)/2},
\end{aligned}$$

where the second term can be bounded as

$$\frac{2^{(1+\phi)(1+\tilde{\rho})-1}C_\delta\tau^2 C_\gamma C_\rho^{1+\beta}}{\mu^2 C_\rho^{\mathbb{1}_{\{\rho\ge 0\}}} N_t}\sum_{i=1}^{t-1} i^{\beta\tilde{\rho}-\alpha} \le \frac{2^{(1+\phi)(1+\tilde{\rho})-1}C_\delta\tau^2 C_\gamma C_\rho^{1+\beta} t^{1+\beta\tilde{\rho}-\alpha}}{(1+\beta\tilde{\rho}-\alpha)\mu^2 C_\rho^{\mathbb{1}_{\{\rho\ge 0\}}} N_t} \le \frac{2^{(1+\phi)(1+\tilde{\rho})-2}C_\delta\tau^2 C_\gamma}{\mu^2 C_\rho^{1-\phi-\beta} N_t^{\phi}},$$

and the third term by

$$\frac{2^{3(1+\phi)(1+\tilde{\rho})/2}C_\delta\tau^2 C_\gamma^{3/2} C_\rho^{1+3\beta/2}}{\mu^{3/2} C_\rho^{\mathbb{1}_{\{\rho\ge 0\}}} N_t}\sum_{i=1}^{t-1} i^{3(\beta\tilde{\rho}-\alpha)/2} \le \frac{2^{3(1+\phi)(1+\tilde{\rho})/2}C_\delta\tau^2 C_\gamma^{3/2} C_\rho^{1+3\beta/2}}{\mu^{3/2} C_\rho^{\mathbb{1}_{\{\rho\ge 0\}}} N_t}\begin{cases} \frac{t^{1+3(\beta\tilde{\rho}-\alpha)/2}}{1+3(\beta\tilde{\rho}-\alpha)/2}, & \text{if } 3(\alpha-\beta\tilde{\rho})/2 < 1, \\ \log(t), & \text{if } 3(\alpha-\beta\tilde{\rho})/2 = 1, \\ \frac{3(\alpha-\beta\tilde{\rho})}{3(\alpha-\beta\tilde{\rho})-2}, & \text{if } 3(\alpha-\beta\tilde{\rho})/2 > 1, \end{cases}$$

By collecting these bounds, we get

$$\begin{aligned}
\frac{C_\delta}{N_t\mu}\sum_{i=0}^{t-1} n_{i+1}\Delta_i^{\frac{1}{2}} \le& \frac{2^{\tilde{\rho}}C_\delta C_\rho\sqrt{\Pi_v'}A_\infty'}{\mu N_t} + \frac{2^{(1+\phi)(1+\tilde{\rho})-2}C_\delta\tau^2 C_\gamma}{\mu^2 C_\rho^{1-\phi-\beta} N_t^{\phi}} \\
&+ \frac{2^{3(1+\phi)(1+\tilde{\rho})/2}C_\delta\tau^2 C_\gamma^{3/2} C_\rho^{1+3\beta/2}}{\mu^{3/2} C_\rho^{\mathbb{1}_{\{\rho\ge 0\}}} N_t}\left(\frac{N_t^{3(1-\phi)/2}}{C_\rho^{3(1-\phi)/2}}\mathbb{1}_{\{\alpha-\beta\tilde{\rho}<2/3\}} + \log(N_t)\mathbb{1}_{\{\alpha-\beta\tilde{\rho}=2/3\}} + \frac{3(\alpha-\beta\tilde{\rho})}{3(\alpha-\beta\tilde{\rho})-2}\mathbb{1}_{\{\alpha-\beta\tilde{\rho}>2/3\}}\right).
\end{aligned}$$

Final bound: Combining our findings from above, we have

$$\begin{aligned}\bar{\delta}_t^{1/2} \leq & \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{2C_\rho^{1-\beta}\sqrt{\pi_v'}A_\infty'}{\mu C_\gamma N_t} + \frac{2^{\frac{3+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma}N_t^{1-\phi/2}} + \frac{C_\rho^{1-\phi-\beta}\sqrt{\pi_v'}A_\infty'}{\mu C_\gamma N_t^{1-\phi}} + \frac{2^{\frac{1+\phi(1+\tilde{\rho})}{2}}\sigma C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma}N_t^{1-\phi/2}} + \frac{C_\rho}{N_t\mu}\left(\frac{1}{C_\gamma C_\rho^\beta} + C_l\right)\delta_0^{\frac{1}{2}} \\ & + \frac{2^{\tilde{\rho}/2}C_l\sqrt{C_\rho}\sqrt{\pi_v'}\sqrt{A_\infty'}}{\mu N_t} + \frac{2^{\frac{\phi(1+\tilde{\rho})}{2}}C_l\sigma\sqrt{C_\gamma}}{\mu^{3/2}C_\rho^{\frac{1-\phi-\beta}{2}\mathbb{1}_{\{\rho\geq 0\}}}N_t^{\frac{1+\phi}{2}}} + \frac{2^{\tilde{\rho}}C_\delta C_\rho\sqrt{\Pi_v'}A_\infty'}{\mu N_t} + \frac{2^{(1+\phi)(1+\tilde{\rho})-2}C_\delta\tau^2 C_\gamma}{\mu^2 C_\rho^{1-\phi-\beta}N_t^\phi} \\ & + \frac{2^{3(1+\phi)(1+\tilde{\rho})/2}C_\delta\tau^2 C_\gamma^{3/2}C_\rho^{1+3\beta/2}}{\mu^{3/2}C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}}N_t}\left(\frac{N_t^{3(1-\phi)/2}}{C_\rho^{3(1-\phi)/2}}\mathbb{1}_{\{\alpha-\beta\tilde{\rho}<2/3\}} + \log(N_t)\mathbb{1}_{\{\alpha-\beta\tilde{\rho}=2/3\}} + \frac{3(\alpha-\beta\tilde{\rho})}{3(\alpha-\beta\tilde{\rho})-2}\mathbb{1}_{\{\alpha-\beta\tilde{\rho}>2/3\}}\right).\end{aligned}$$

This can be simplified to

$$\begin{aligned}\bar{\delta}_t^{1/2} \leq & \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{2^{3+\phi(1+\tilde{\rho})}\sigma C_\rho^{(1-\phi-\beta)/2\mathbb{1}_{\{\rho\geq 0\}}}}{\mu^{3/2}\sqrt{C_\gamma}N_t^{1-\phi/2}} + \frac{2^{(1+\phi)(1+\tilde{\rho})-2}C_\delta\tau^2 C_\gamma}{\mu^2 C_\rho^{1-\phi-\beta}N_t^\phi} + \frac{C_\rho^{1-\phi-\beta}\sqrt{\pi_v'}A_\infty'}{\mu C_\gamma N_t^{1-\phi}} + \frac{2^{\phi(1+\tilde{\rho})/2}C_l\sigma\sqrt{C_\gamma}}{\mu^{3/2}C_\rho^{(1-\phi-\beta)/2\mathbb{1}_{\{\rho\geq 0\}}}N_t^{(1+\phi)/2}} \\ & + \frac{C_\rho\Gamma_v}{\mu N_t} + \frac{2^{3(1+\phi)(1+\tilde{\rho})/2}C_\delta\tau^2 C_\gamma^{3/2}C_\rho^{1+3\beta/2}}{\mu^{3/2}C_\rho^{\mathbb{1}_{\{\rho\geq 0\}}}N_t}\left(\frac{N_t^{3(1-\phi)/2}}{C_\rho^{3(1-\phi)/2}}\mathbb{1}_{\{\alpha-\beta\tilde{\rho}<2/3\}} + \log(N_t)\mathbb{1}_{\{\alpha-\beta\tilde{\rho}=2/3\}} + \frac{3(\alpha-\beta\tilde{\rho})}{3(\alpha-\beta\tilde{\rho})-2}\mathbb{1}_{\{\alpha-\beta\tilde{\rho}>2/3\}}\right).\end{aligned}$$

with Γ_v given by $\left(\frac{1}{C_\gamma C_\rho^\beta} + C_l\right)\delta_0^{1/2} + \frac{2^{\tilde{\rho}}C_l\sqrt{\pi_v'}\sqrt{A_\infty'}}{C_\rho^{1/2}} + \frac{2\sqrt{\pi_v'}A_\infty'}{C_\gamma C_\rho^\beta} + 2^{\tilde{\rho}}C_\delta\sqrt{\Pi_v'}A_\infty'$, consisting of the finite constants π_v', Π_v' and A_∞', given as

$$\pi_v' = \left(\delta_0 + \frac{2\sigma^2}{C_l^2}\right)\pi_v,\ \Pi_v' = \left(\Delta_0 + \frac{2\tau^4}{C_l^4} + \frac{4\tau^4 C_\gamma}{\mu C_l^2}\right)\Pi_v, \text{ and } A_\infty' = \sum_{i=0}^{\infty} i^{\tilde{\rho}}\exp\left(-\frac{\mu C_\gamma C_\rho^{\beta\mathbb{1}_{\{\rho\geq 0\}}}i^{(1-\phi)(1+\tilde{\rho})}}{2^{1+(1-\phi)(1+\tilde{\rho})}}\right). \tag{41}$$

□

A. Technical Proofs

Appendix A contains purely technical results used in the proofs presented in Section 6. In what follows, we use the convention $\inf\emptyset = 0$, $\sum_{t=1}^{0} = 0$, and $\prod_{t=1}^{0} = 1$.

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

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

Proof of Proposition A.1. We begin with considering the first inequality in (A.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 (A.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 (A.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 A.2. *Let $(\gamma_t)_{t\geq1}$ be a positive sequence. Let $\omega>0$ and $k\leq t$ such that for all $i\geq k$, $\omega\gamma_i\leq1$, then*

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

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

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

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 $\prod_{j=k}^{t}\left[1-\omega\gamma_j\right]\geq0$, showing the inequality in (A.2). □

Proposition A.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:*

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

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

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

Proof of Proposition A.3. We obtain the inequality in (A.3) directly by Proposition A.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 (A.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 A.2. □

Proposition A.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 and decreasing sequences satisfying the following:*

- *The sequence δ_t follows the recursive relation:*

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

  with $\delta_0 \geq 0$ and $\omega > 0$.

- *Let γ_t and η_t converge to 0.*

- *Let $t_0 = \inf\{t \geq 1 \ : \eta_t \leq \omega\}$, and let us suppose that for all $t \geq t_0 + 1$, one has $\omega\gamma_t \leq 1$.*

Then, for all $t \in \mathbb{N}$, we have the upper bound:

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

with the convention that $\sum_{t_0}^{t/2} = 0$ if $t/2 < t_0$.

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

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

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,

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

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

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

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

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

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

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

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

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

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

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

as $1 + t \leq \exp(t)$ for all $t \in \mathbb{R}$. Thus, for all $t \in \mathbb{R}$,

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

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 A.1, as follows:

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

by the definition of t_0, thus

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

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

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

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

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

Proposition A.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 following:*

- *Suppose δ_t follows the recursive relation in (A.5) with $\delta_0 \geq 0$ and $\omega > 0$.*
- *Let $t_0 = \inf\{t \geq 1 \,:\, \eta_t \leq \omega\}$, and let us suppose that for all $t \geq t_0 + 1$, one has $\omega\gamma_t \leq 1$.*

Then, for all $t \in \mathbb{N}$, we have the upper bound:

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

Proof of Proposition A.5. The resulting (upper) bound in (A.11) follows directly from (A.6) by noting that $t_0 \leq t$, giving us

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

with use of

$$\sum_{i=t_0+1}^{t/2-1} \nu_i\gamma_i \leq \sum_{i=1}^{t} \nu_i\gamma_i \leq \max_{1\leq i\leq t} \frac{\nu_i}{\eta_i} \sum_{i=1}^{t} \eta_i\gamma_i \leq \max_{1\leq i\leq t} \frac{\nu_i}{\eta_i} \exp\left(2\sum_{i=1}^{t} \eta_i\gamma_i\right),$$

as (ν_t) and (γ_t) are positive sequences. □

References

Bach, F., Moulines, E., 2011. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. Advances in Neural Information Processing Systems 24, 451–459.

Benveniste, A., Métivier, M., Priouret, P., 1990. Adaptive algorithms and stochastic approximations. Springer-Verlag.

Bottou, L., Bousquet, O., 2008. The tradeoffs of large scale learning, in: Advances in Neural Information Processing Systems, pp. 161–168.

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

Bottou, L., LeCun, Y., 2004. Large scale online learning. Advances in Neural Information Processing Systems 16, 217–224.

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

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

Dozat, T., 2016. Incorporating nesterov momentum into adam. Proceedings of 4th International Conference on Learning Representations .

Duchi, J., Hazan, E., Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. Journal of machine learning research 12.

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

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Kushner, H., 2010. Stochastic approximation: a survey. WIREs Computational Statistics 2, 87–96.

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

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

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

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

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

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

Robbins, H., Monro, S., 1951. A stochastic approximation method. Annals of Mathematical Statistics 22, 400–407.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. nature 323, 533–536.

Ruppert, D., 1988. Efficient Estimations from a Slowly Convergent Robbins-Monro Process. Cornell University Operations Research and Industrial Engineering.

Schmidt, M., Roux, N., Bach, F., 2011. Convergence rates of inexact proximal-gradient methods for convex optimization. Advances in Neural Information Processing Systems 24, 1458–1466.

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

Sridharan, K., Shalev-Shwartz, S., Srebro, N., 2008. Fast rates for regularized objectives. Advances in Neural Information Processing Systems 21, 1545–1552.

Teo, C.H., Smola, A., Vishwanathan, S., Le, Q.V., 2007. A scalable modular convex solver for regularized risk minimization, in: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 727–736.

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

Xiao, L., 2010. Dual averaging methods for regularized stochastic learning and online optimization. Journal of Machine Learning Research 11, 2543–2596.

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

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