# Acceleration of stochastic gradient descent with momentum by averaging: finite-sample rates and asymptotic normality

Kejie Tang * Weidong Liu † Yichen Zhang‡

**Abstract**

Stochastic gradient descent with momentum (SGDM) has been widely used in many machine learning and statistical applications. Despite the observed empirical benefits of SGDM over traditional SGD, the theoretical understanding of the role of momentum for different learning rates in the optimization process remains widely open. We analyze the finite-sample convergence rate of SGDM under the strongly convex settings and show that, with a large batch size, the mini-batch SGDM converges faster than mini-batch SGD to a neighborhood of the optimal value. Furthermore, we analyze the Polyak-averaging version of the SGDM estimator, establish its asymptotic normality, and justify its asymptotic equivalence to the averaged SGD.

## 1 Introduction

In this paper, we are interested in solving the following stochastic optimization problem:

$$x^* = \operatorname{argmin}_{x\in\mathbb{R}^d} \mathrm{E}[f_\xi(x)]; \tag{1.1}$$

where $f_\xi(x)$ is a stochastic convex loss function, $\xi$ is a random element, and $\mathrm{E}[\cdot]$ is the expectation respected to $\xi$. Stochastic optimization plays an important role in many statistics and machine learning problems. For a wide range of applications, the objective function is strongly convex,

*School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, 200240, China, tangkj00@sjtu.edu.cn

†School of Mathematical Sciences, MoE Key Lab of Artifcial Intelligence, Shanghai Jiao Tong University, Shanghai, 200240, China, weidongl@sjtu.edu.cn

‡Daniels School of Business, Purdue University, West Lafayette, IN, 47907, USA, zhang@purdue.edu

and stochastic gradient descent (SGD) is widely preferred over gradient descent (GD) (See, e.g., Nesterov, 2003; Nocedal and Wright, 2006) due to its high computational advantage. SGD is a first-order optimization algorithm that approximates the expected loss by averaging the loss function over a mini-batch of training examples. At each iteration, the algorithm updates the model parameters in the direction of the negative gradient of the mini-batch loss, scaled by a learning rate parameter.

While SGD is simple and easy to implement, it may suffer from slow convergence rates or oscillations in high-dimensional optimization problems, particularly when the loss function is ill-conditioned or noisy. Momentum-based methods enhance SGD by introducing an exponentially weighted moving average of the past gradients to the update rule, which serves to dampen oscillations and accelerate convergence. In particular, the momentum term introduces a form of inertia to the update process, allowing the algorithm to maintain a more consistent direction of movement even in the presence of noisy gradients. Several variants of momentum-based SGD have been proposed, such as Nesterov's accelerated gradient (NAG), Adagrad, and Adam, each with its own strengths and weaknesses.

SGD with momentum has become increasingly popular in deep learning, where large-scale neural networks are trained using mini-batch stochastic gradient descent. The use of momentum-based optimization methods has been shown to improve the convergence rate and the generalization performance of neural networks, as well as to reduce the sensitivity to the choice of hyperparameters. However, the theoretical analysis of momentum-based SGD under strong convexity, a property that ensures the existence of a unique and well-defined minimizer, is still an active area of research.

In this paper, we focus on the theoretical analysis of momentum-based SGD from an optimization perspective under the assumption of strong convexity, a property that ensures the existence of a unique and well-defined minimizer. Strong convexity is a natural assumption in many machine learning applications, such as logistic regression and support vector machines, where the loss function has a unique global minimum. Our goal is to establish the convergence properties of momentum-based SGD under strong convexity and to provide insights into the role of momentum in the optimization process.

SGD with momentum is one of the most common algorithms for solving (1.1). The SGDM

updates the target estimator by a weighted combination with historical gradients

$$\begin{aligned} m_{t+1} &= \gamma m_t + (1-\gamma) \circ f_{\xi_t}(x_t); \\ x_{t+1} &= x_t - \alpha m_{t+1}; \end{aligned} \tag{1.2}$$

where $\xi_t$, $t \geq 1$ are i.i.d. samplings from $\xi$, $\gamma$ is the momentum weight and $\alpha$ is the learning rate. The classical SGD is a special case of SGDM with $\gamma = 0$ and $m_{t+1} = r f_{\xi_t}(x_t)$.

SGDM is often also called as stochastic heavy-ball method, originated from Polyak's heavy-ball method for deterministic optimization (Polyak, 1964). We refer to Section 1.1 for a detailed literature review. In practice, a large momentum weight is often placed to accelerate the algorithm, for example, $\gamma = 0.9$. That being said, it is still an open problem to investigate the theoretical properties of SGDM with a general specification of $\gamma$. The theoretical analysis in existing studies has not given an affirmative answer to the open problem by Liu et al. (2020) and the assertion by Kidambi et al. (2018).

We establish finite sample rates for SGDM and the averaging of its trajectories under smooth and strongly convex loss functions. By our results, we can give some partial answers to these open questions in several aspects.

1. We provide rigorous theoretical support for the assertion by Kidambi et al. (2018) in comparing the performance of SGD and SGDM. Our finite-sample analysis shows that the mini-batch SGDM with appropriate momentum weights and batch sizes converges faster than mini-batch SGD to a given neighborhood of the true solution. We further provide an adaptive choice of the momemtum weight which theoretically attains the optimal convergence rate. Our experiments support the theoretical results of convergence and the optimal momentum weight.

2. We provide a non-asymptotic analysis of the Polyak-Ruppert averaging version of SGDM. The averaging of SGDM trajectories $\bar{x}_t = \frac{1}{t-n_0} \sum_{i=n_0+1}^{t} x_i$ can accelerate SGDM with a wide range of learning rates $\alpha$ to the rate $O(\frac{1}{\sqrt{t}\sigma})$, where $\sigma^2$ is the variance of stochastic gradient. Furthermore, we show theoretically that averaged SGDM converges faster than averaged SGD in the early phases of the iteration process.

3. We further establish the asymptotic normality for the averaging $\bar{x}_t$ as $t \to \infty$ under certain

scenarios of a decaying learning rate $\alpha$ or a diverging batch size $s$. Particularly, as $\alpha = \Theta(t^{-\epsilon})$ for $\epsilon \in (\frac{1}{2}; 1)$, the asymptotic normality holds for averaged SGDM under strongly convex loss functions. The asymptotic covariance matrix depends only on the Hessian and Gram matrix at $x^*$, and matches the one in Polyak and Juditsky (1992) of averaged SGD. Interestingly, mini-batch SGDM is asymptotically equivalent to mini-batch SGD as $t \to \infty$.

## 1.1 Related Works

Since the seminal work by Robbins and Monro (1951), the convergence properties of SGD have been extensively studied in the literature (See, e.g., Moulines and Bach, 2011; Bottou et al., 2018; Nguyen et al., 2018). On the contrary, the theory for SGDM has rarely been explored until recently, although it is very popular in training many modern machine learning models such as neural networks to improve the training speed and accuracy of various models.

Kidambi et al. (2018) showed that (1.2) cannot achieve any improvement over SGD for a specially constructed linear regression problem. Together with some numerical experiments, they asserted that the only reason for the superiority of stochastic momentum methods in practice is mini-batching. In a recent paper, Liu et al. (2020) proved SGDM can be as fast as SGD. They established the identical convergence bound as SGD. They also posed an open problem whether it is possible to show that SGDM converges faster than SGD for special objectives such as quadratic ones. Other studies on the theoretical analysis of SGDM can be found in Loizou and Richtárik (2017), Loizou and Richtárik (2020), Gitman et al. (2019) for linear system and quadratic loss; Sebbouh et al. (2021) for almost sure convergence rates under smooth and convex loss functions with a time-varying momentum weight. Mai and Johansson (2020) studied a class of general convex loss, and obtain convergence rates of time averages regardless of the momentum weight. We refer the readers to the latter paper and Liu et al. (2020) for a few more papers on the convergence rate for SGDM.

Richtárik and Takác (2020) considered the problem of solving a consistent linear system $Ax = b$ on least square regression. Some works (Loizou and Richtárik, 2017; Kidambi et al., 2018; Loizou and Richtárik, 2020; Bollapragada et al., 2022) based on the stochastic heavy ball method(SHB)

achieved the linear convergence rate. Yang et al. (2016); Defazio (2020); Jin et al. (2022); Li et al. (2022) investigate the impact of momentum on the convergence properties of non-convex optimization problems and provides insights into its practical use. Gitman et al. (2019) focus on the noise reduction properties of momentum.

## 2 Preliminaries

In this section, we present the mini-batch SGDM settings considered in this paper. Particularly, we define a mini-batch stochastic loss

$$g_{\eta_t}(x) = \frac{1}{s}\sum_{i=1}^{s} f_{\xi_{ti}}(x); \tag{2.3}$$

where $\eta_t = \{\xi_{t1}, \ldots, \xi_{ts}\}$ is a mini-batch of size $s$ and elements $\xi_{ti}$ are i.i.d. sampled from the distribution of $\xi$. The update of $m_t$ in (1.2) uses the mini-batch stochastic gradient $\nabla g_{\eta_t}(x_t)$ instead of the individual stochastic gradient $\nabla f_{\xi_{ti}}(x_t)$.

We first introduce some regularity assumptions and discuss their use in the theoretical results. We denote the $\ell_2$-norm of a vector as $\|\cdot\|$ and the operator norm as $\|A\| = \max_{\|x\|=1}\|Ax\|$.

(A1). Assume that the loss function $f_\xi(x)$ is twice differentiable. Let $\kappa_1 \le \cdots \le \kappa_d$ be the eigenvalues of Hessian matrix $\Sigma = \mathrm{E}[\nabla^2 f_\xi(x^*)]$. Assume that $\mu := \kappa_1 > 0$, and $L := \kappa_d < 1$.

(A2). There exists a constant $\bar{L} \ge 0$ such that the Hessian matrix of the loss $f(x) := \mathrm{E}[f_\xi(x)]$, $\Sigma(x) = \nabla^2 f(x)$ satisfies $\|\Sigma(x) - \Sigma(x^*)\| \le \bar{L}\|x - x^*\|$, $\forall x$.

(A3). Define $L_\xi = \sup_x \frac{\|\nabla f_\xi(x) - \nabla f_\xi(x^*)\|}{\|x - x^*\|}$. Assume that $\mathrm{E}[\|\nabla f_\xi(x^*)\|^2] \le \sigma^2$ and $\mathrm{E}[L_\xi^2] \le L_f^2$.

(A3'). Assume that $\nabla f_\xi(x^*)$ and $L_\xi$ satisfy

$$\sup_{\|v\|=1} \mathrm{E}\left[\exp\left(v^\top \nabla f_\xi(x^*)/\sigma\right)\right] \le 2; \quad \mathrm{E}\left[\exp\left(L_\xi/L_f\right)\right] \le 2.$$

A few discussions follow concerning the assumptions above. Assumption (A1) is a regularity condition on the smoothness and strong convexity of the loss function $f_\xi(x)$ at $x^*$. Beyond that, (A2) assumes the Lipschitz condition on the Hessian matrix of the loss function. Notably, the $\Sigma$ defined in (A1) is a brief notation for $\mathrm{E}[\Sigma_\xi(x^*)]$ defined in (A2). For quadratic losses, the Hessian matrix $\Sigma(x)$ is identical for different $x$, and therefore (A2) holds with $\bar{L} = 0$. There are a few important

applications that fit in the settings of non-quadratic but more general strongly convex losses, such as logistic regression, Poisson regression, etc. In the following, we consider two scenarios, quadratic losses, and general losses. Particularly, we illustrate (A2) under a logistic regression setting in Example 1 below. This assumption is weaker than many in the literature, such as those in Yan et al. (2018); Liu et al. (2020) that $\mathrm{E}[\|\nabla f_\xi(x) - \nabla f(x)\|^2] \le \sigma^2$ for all $x$. For quadratic losses, such an assumption does not hold for some $x$ when $\|x - x^*\| \neq 1$. On the other hand, Bottou et al. (2018); Wang and Johansson (2022) assumed that $\mathrm{E}[\|\nabla f_\xi(x) - \nabla f(x)\|^2] \le \rho\|\nabla f(x)\|^2 + \sigma^2$ for $\rho > 0$. Meanwhile, our assumptions lead to

$$\mathrm{E}[\|\nabla f_\xi(x) - \nabla f(x)\|^2] \le \frac{3(L_f^2 + L^2)}{\mu^2}\|\nabla f(x)\|^2 + 3\sigma^2;$$

due to the triangle inequality and strongly convexity.

Assumptions (A3) and (A3') are two separate conditions on the smoothness of quadratic and general stochastic loss functions, respectively. For any bounded domain $\|x\| \le r$, (A3) and (A3') are easily satisfied for bounded gradients and smoothness. For unbounded problems, they are usually satisfied given certain design properties in many popular statistical models. These assumptions are relatively innocuous for practical problems. For instance, we illustrate the assumptions under logistic regression in the following example.

**Example 1** *Consider logistic regression with samples $\{a_\xi; b_\xi\}$ such that $a_\xi \in \mathbb{R}^d$ and $b_\xi \in \{0; 1\}$ are generated by $b_\xi = 1$ with probability $p_x(a_\xi)$ and $b_\xi = 0$ otherwise, where $p_x(a) = 1/(1 + \exp(-x^\top a))$. We consider the loss function, i.e., the negative log-likelihood, is defined as:*

$$f_\xi(x) = -b_\xi \log(p_x(a_\xi)) - (1 - b_\xi)\log(1 - p_x(a_\xi)):$$

*The gradient and the Hessian matrix are*

$$\nabla f_\xi(x) = (p_x(a_\xi) - b_\xi)a_\xi; \quad \Sigma(x) = \mathrm{E}[p_x(a_\xi)(1 - p_x(a_\xi))a_\xi a_\xi^\top]:$$

*By the fact that $|p_x(a) - p_y(a)| \le \frac{\sqrt{3}}{18}\|a\|\|x - y\|$, (A2) is satisfied with $\bar{L} = \frac{\sqrt{3}}{6}\mathrm{E}\|a_\xi\|^3$: Suppose $\{a_\xi\}$ in logistic regression satisfy $\mathrm{E}[a_\xi] = 0_d$ and*

$$\sup_{\|v\|=1} \mathrm{E}\ \exp\left(v^\top a_\xi\right)^2/\sigma^2 \le 2;$$

*we have that* $|p_x(a) - b| \le 2$, $\|(p_x(a) - p_y(a))a\| \le \frac{\sqrt{3}}{18}\|a\|^2\|x - y\|$ *and therefore,*

$$\sup_{\|v\|=1} \mathrm{E}\left[\exp\left(v^\top \nabla f_\xi(x^*)/(c\sigma)\right)\right] \le 2;\ \mathrm{E}\left[\exp\left(L_\xi/(cL_f)\right)\right] \le 2:$$

*for some absolute constant* $c > 0$ *and* $L_f = \mathrm{E}\|a_\xi\|^2$.

# 3 Finite-sample Convergence Rates for SGDM

In this section, we first present the finite-sample convergence results for SGDM with general momentum weight $\gamma$. We consider the two cases separately: $\underline{L} = 0$ corresponds to the quadratic losses, and $\underline{L} > 0$ corresponds to general strongly convex losses.

## 3.1 The finite-sample rates for SGDM on quadratic losses

We first establish the finite-sample rates under quadratic losses, where

$$f_\xi(x) = \frac{1}{2}x^\top A_\xi x - b_\xi^\top x + c: \tag{3.4}$$

The Hessian matrix of the loss function $\Sigma = \mathrm{E}[A_\xi]$. The stochastic loss is $g_{\eta_t}(x) = \frac{1}{s}\sum_{i=1}^{s} f_{\xi_{ti}}(x)$ in the mini-batch setting (2.3). The following theorem shows the convergence rate of the last iterate $x_{t+1}$.

**Theorem 1** *Under (A1)-(A3) and* $\underline{L} = 0$, *for any momentum* $\gamma \in [0, 1)$ *and fixed* $\delta \in (0, 1]$, *assume the learning rate* $\alpha > 0$ *satisfies* $\alpha L < 2(1+\gamma)/(1-\gamma)$ *and* $16M^2\alpha^2 L_f^2 \le s\delta\lambda^{2(1-\delta)}(1-\lambda)$, *where*

$$M = \frac{4}{\Delta}\left(2(1-\gamma)(1+\alpha L + L) + 3\alpha\gamma\right);\ \Delta = \min_k\left\{(\gamma + 1 - \alpha(1-\gamma)\kappa_k)^2 - 4\gamma\right\} > 0; \tag{3.5}$$

*and* $\lambda$ *is the spectral radius of the matrix*

$$\Gamma = \begin{pmatrix} \gamma I, & (1-\gamma)\Sigma \\ -\alpha\gamma I, & I - \alpha(1-\gamma)\Sigma \end{pmatrix}: \tag{3.6}$$

*Let* $\varepsilon_{t+1} = (1-\gamma)\sum_{j=1}^{t}\gamma^{t-j}\Sigma(x_j - x^*)$, *we have for* $t \ge 1$,

$$\mathrm{E}[\|\varepsilon_{t+1}\|^2 + \|x_{t+1} - x^*\|^2] \le 2M^2\left(\frac{4}{s(1-\lambda)}\alpha^2\sigma^2 + \|x_1 - x^*\|^2\right)\lambda^{2(1-\delta)t}: \tag{3.7}$$

Theorem 1 provides a finite-sample bound simultaneously for the error of the last iterate $x_{t+1} - x^*$, and a weighted average of the stochastic gradients $\bar{m}_{t+1}$. The first term in the error bound (3.7), $\frac{8M^2}{s(1-\lambda)}\alpha^2\sigma^2$, corresponds to a non-decaying bias, due to the noisy observation in the stochastic gradient, which is the same as the one in SGD. The second term is exponentially decaying when $\lambda < 1$ and establishes a linear convergence of the SGDM algorithm to the neighborhood of the true solution with size with range $\sqrt{\frac{8M^2}{s(1-\lambda)}}\alpha\sigma$.

We provide some intuition in deriving the second term. Considering the noiseless setting where the stochastic gradient is the same as the true gradient, i.e., $\nabla f_\xi(x) = \nabla f(x)$ for any $x, \xi$. We have $\nabla g(x) = \Sigma(x - x^*)$ and the SGDM updating rule (1.2) can be rewritten as

$$\begin{pmatrix} m_{t+1} \\ x_{t+1} - x^* \end{pmatrix} = \Gamma \begin{pmatrix} m_t \\ x_t - x^* \end{pmatrix} = \Gamma^t \begin{pmatrix} m_1 \\ x_1 - x^* \end{pmatrix},$$

where $\Gamma$ is the matrix in (3.6). When the spectral radius of $\Gamma$ is less than 1, the full-batch gradient descent with momentum enjoys linear convergence. In Theorem 1, the matrix $\Gamma$ satisfies $\|\Gamma^t\| \le M\lambda^t$ with $\lambda < 1$, which is proved in the appendix.

**Remark 2** *The assumption that $\Delta > 0$ in (3.5) is placed for the diagonalization of the matrix $\Gamma$, which is required in our analysis to provide a last-iterate convergence analysis and can be relaxed if we aim for a time-average convergence analysis of the sum $\sum_t \mathbb{E}[\|\bar{m}_t\|^2 + \|x_t - x^*\|^2]$. More particularly, a time-average convergence analysis computes $\sum_{t=0}^{\infty} \Gamma^t = (I - \Gamma)^{-1}$ and requires only that the spectral radius $\lambda < 1$.*

In Theorem 1, the convergence rate mainly depends on the spectral radius $\lambda$ of the matrix (3.6). We characterize its explicit form in the following theorem.

**Theorem 3** *For any momentum weight $\gamma \in [0, 1)$, let $\alpha, \gamma$ satisfy $\alpha L < 2(1+\gamma)/(1-\gamma)$ and define $\phi = \min\{\alpha\mu, 2(1+\gamma)/(1-\gamma) - \alpha L\}$. If the momentum $\gamma < (1-\phi)^2/(1+\phi)^2$, we have that the spectral radius $\lambda$ of $\Gamma$ defined by (3.6) satisfies*

$$\lambda = \frac{\gamma + 1 - (1-\gamma)\phi + \sqrt{(\gamma + 1 - (1-\gamma)\phi)^2 - 4\gamma}}{2}. \tag{3.8}$$

*On the other hand, if the momentum $\gamma \ge (1-\phi)^2/(1+\phi)^2$, we have*

$$\lambda = \sqrt{\gamma}.$$

Theorem 3 reveals that the behavior of $\gamma$ is essentially different in the two ranges of momentum weights $\gamma \in [0, 1)$, segmented by $(1-\phi)^2/(1+\phi)^2$. We now explicitly characterize the convergence rates of SGDM in Theorem 1 with certain specifications of the momentum weight $\gamma$ that is either close to 0 or 1, in the following two corollaries.

**Corollary 4 (Small momentum weight $\gamma$)** *Under the assumptions in Theorem 1, for $0 \le \gamma \le (1-\alpha L)/(4+4\alpha^2 L^2)$, we have that $\Delta \ge (1-\alpha L)^2/2$, $\alpha L^2 L_f^2 = O(s\mu)$ and*

$$\mathrm{E}[\|e_{t+1}\|^2 + \|x_{t+1} - x^*\|^2] = O\left(\frac{L^2}{s\mu}\alpha\sigma^2 + L^2\lambda^{2(1-\delta)t}\right),$$

*where $\lambda$ is defined in (3.8) above and*

$$1 - \phi - \frac{\phi(1+\phi)}{1-\phi}\gamma \le \lambda \le 1 - \phi - \frac{\phi^2}{1-\phi}\gamma.$$

Corollary 4 establishes the explicit finite-sample rate of convergence of SGDM under small $\gamma$, and shows that, the spectral radius $\lambda$ decreases slowly as $\gamma$ increases from 0 to $(1-\alpha L)/(4+4\alpha^2 L^2)$.

**Corollary 5 (Large momentum weight $\gamma$)** *Under the assumptions in Theorem 1, for $0 < 1-\gamma \le 2\alpha\mu/(1+\alpha\mu)^2$, we have that $\Delta \ge 2(1-\gamma)\alpha\mu$, $\alpha L_f^2 \left(L^2 + \frac{\alpha^2}{(1-\gamma)^2}\right) = O(s\mu)$; and*

$$\mathrm{E}[\|e_{t+1}\|^2 + \|x_{t+1} - x^*\|^2] = O\left(\frac{1}{s\mu}\left(L^2 + \frac{\alpha^2}{(1-\gamma)^2}\right)\alpha\sigma^2 + \left(L^2 + \frac{\alpha}{(1-\gamma)\mu}\right)\lambda^{2(1-\delta)t}\right),$$

*where $\lambda$ is defined in (3.8) above. Furthermore, if $\alpha < \frac{2(1+\gamma)}{(\mu+L)(1-\gamma)}$, we have $\lambda = \sqrt{\gamma}$.*

Corollary 5 demonstrates that the spectral radius $\lambda$ increases with $\gamma$ when it approaches 1, exhibiting an opposite behavior compared to the small $\gamma$ settings that in Corollary 4.

**Remark 6** *As one increases $\gamma$ from 0 to 1, the corresponding spectral radius $\lambda$ first decreases and then increases. The minimal spectral radius in Theorem 3, given $\phi$, is $\lambda^* = (1-\phi)/(1+\phi)$, when momentum weight is $\gamma^* = (1-\phi)^2/(1+\phi)^2$. Particularly, if one specifies $\gamma$ and $\alpha$ as*

$$\gamma = \frac{(1-\phi)^2}{(1+\phi)^2} + \epsilon, \quad \text{and} \quad \alpha \le \sqrt{\frac{1}{\mu L}},$$

*for sufficiently small $\epsilon > 0$, we have*

$$\mathrm{E}[\|e_{t+1}\|^2 + \|x_{t+1} - x^*\|^2] = O\left(\frac{(L+1/\mu)^2}{s\epsilon}\alpha^2\sigma^2 + \frac{\alpha\mu(L+1/\mu)^2}{\epsilon}\gamma^{(1-\delta)t}\right).$$

We now compare our rate of convergence with several results in the existing literature.

**Comparison to SGD** Theorem 4.6 in Bottou et al. (2018) proved that the convergence rate of SGD under strong convexity is $O(\frac{L}{\mu}\alpha\sigma^2 + (1-\alpha\mu)^t)$, where they assume $\mathrm{E}[\|\nabla f_\xi(x) - \nabla f(x)\|^2] \le \sigma^2$ for all $x$. Comparing with their result, our analysis in Theorem 3 can apply with $\phi = \alpha\mu$ for $\gamma = 0$ and $\alpha < 2/(\mu + L)$, which is consistent with Bottou et al. (2018). Our result also implies that SGDM with momentum weight $0 < \gamma < (1-\phi)^2/(1+\phi)^2$ can accelerate the convergence rate in Bottou et al. (2018).

**Comparison to existing results for SGDM** Liu et al. (2020) proved that the convergence rate for strongly convex loss is $O(\frac{L}{\mu}\alpha\sigma^2 + \max\{1-\alpha\mu, \gamma\}^t)$ for $t \ge t_0 = O(-1/\log(\gamma))$. Comparing that with Corollary 4 above, our result improves the rate when the momentum weight $\gamma$ is small. For large $\gamma$ when $\gamma \ge (1-\phi)^2/(1+\phi)^2$, our rate is $\gamma^{(1-\delta)t}$ for $t \ge 1$, which can be arbitrarily close to $\gamma^t$.

Then SHB method is equivalent to SGDM for $\alpha' = \alpha(1-\gamma)$ and $\gamma' = \gamma$. Under the noiseless setting, for $\alpha' = 1/L$ and $\gamma' = (\sqrt{L/\mu} - 1)/(\sqrt{L/\mu} + 1)$, Nesterov (1983) gave the $O((1 - \sqrt{\mu/L})^t)$ convergence rate. For $\alpha' = O(1/L)$ and $\alpha = O(\sqrt{1/\mu L})$ in Remark 6, our convergence factor is $\gamma = O(1 - \sqrt{\mu/L})$, which improves the factor $O(1-\mu/L)$ in SGD and is consistent with Nesterov (1983).

Theorem 3.5 in Bollapragada et al. (2022) proved that for

$$\sqrt{\alpha'} = \frac{2}{\sqrt{L+\epsilon} + \sqrt{\mu-\epsilon}} \quad \text{and} \quad \sqrt{\gamma'} = \frac{\sqrt{(L+\epsilon)/(\mu-\epsilon)} - 1}{\sqrt{(L+\epsilon)/(\mu-\epsilon)} + 1}; \tag{3.9}$$

for $\epsilon \in (0, \mu)$, in the SHB method, the convergence rate on a consistent linear system $Ax = b$ is $O(\frac{L^2}{\epsilon}t^2(\gamma')^t + \frac{L^4}{\epsilon\mu^2}\sigma^2)$; see Bollapragada et al. (2022) for the definition of $\sigma$. Their convergence rate is the same as ours in Remark 6 under a different setting, and their results are more restrictive on the choice of learning rates and momentum weights.

## 3.2 The finite-sample rates for SGDM on general losses

For the general strongly convex loss where $\bar{L} > 0$, we provide the following convergence rate.

**Theorem 7** *Under (A1)–(A2) and (A3') and $\bar{L} > 0$, for any momentum $\gamma \in [0, 1)$, assume the*

*learning rate* $\alpha > 0$ *satisfies* $\alpha L < 2(1+\gamma)/(1-\gamma)$ *and*

$$6\sqrt{2c}(2\log T + 4d)M\left(\frac{2L_f}{\lambda^{(1-\delta)}\sqrt{\delta}(1-\lambda)^{1/2}} + \frac{3\bar{L}M\alpha\sigma}{(1-\lambda)^{3/2}}\right)\alpha \le \sqrt{s};$$

*for fixed* $\delta \in (0,1]$ *and an absolute constant* $c > 0$, *where* $M$ *is defined in (3.5). In addition, the initialization* $x_1$ *satisfies* $\|x_1 - x^*\| \le \frac{\delta(1-\lambda)\lambda^{1-\delta}}{36M^2\alpha\bar{L}}$. *Then with probability* $1 - 2T^{-1}$,

$$\sqrt{\|x_{t+1}-x_t\|^2 + \|x_{t+1}-x^*\|^2} \le \frac{3\sqrt{2c}(2\log T + 4d)M}{\sqrt{s(1-\lambda)}}\alpha\sigma + 3M\|x_1 - x^*\|\lambda^{(1-\delta)t}; \text{ for } 1 \le t \le T:$$

Compared to Theorem 1, the convergence rate in Theorem 7 is worse but with only up to a logarithm term, which indicates that SGDM almost achieves the same rate when $\bar{L} > 0$. Mini-batch SGDM for general convex loss still converges faster than SGD with high probability, as discussed in Theorem 3. In contrast to the quadratic settings $\bar{L} = 0$, the general convex setting requires additionally that the initialization $x_1$ is close to $x^*$, i.e., $\|x_1 - x^*\|$ is small. This assumption is expected since the loss function is not guaranteed to be convex everywhere under the weak assumptions in (A1)–(A2). Particularly, for $\gamma$ close to 0, we assume $\|x_1 - x^*\| = O(\mu/\bar{L}L^2)$. For $\gamma$ close to 1, we assume $\|x_1 - x^*\| = O(\mu/\bar{L}(L^2 + \alpha^2/(1-\gamma)^2))$.

Based on Theorem 7, we have the following corollaries of the convergence rates for small and large specifications of $\gamma$ under general losses, similar to Corollaries 4–5 for quadratic losses.

**Corollary 8 (Small momentum weight $\gamma$)** *Following Theorem 7, for* $0 \le \gamma \le (1-\alpha L)/(4+4\alpha^2L^2)$, *we have that* $\Delta \ge (1-\alpha L)^2/2$, $\sqrt{\alpha\bar{L}L_f L(\mu + \bar{L}L)} = O(s^{1/2}\mu^{3/2})$ *and with high probability*

$$\sqrt{\|x_{t+1}-x_t\|^2 + \|x_{t+1}-x^*\|^2} = O\left(\frac{L\log T}{\sqrt{s\mu}}\sqrt{\alpha\bar{\alpha}}\sigma + L\lambda^{(1-\delta)t}\right);$$

*where* $\lambda$ *is defined in (3.8) above.*

**Corollary 9 (Large momentum weight $\gamma$)** *Following Theorem 7, for* $0 < 1-\gamma \le 2\alpha\mu/(1+\alpha\mu)^2$, *we have that* $\Delta \ge 2(1-\gamma)\alpha\mu$,

$$\sqrt{\alpha\bar{\alpha}L_f\left(L + \frac{\alpha}{1-\gamma}\right)\left(1 + \bar{L}\left(L + \frac{\alpha}{1-\gamma}\right)\sqrt{\frac{\alpha}{(1-\gamma)\mu}}\right)} = O(s\mu);$$

*and with high probability*

$$\sqrt{\|x_{t+1}-x_t\|^2 + \|x_{t+1}-x^*\|^2} = O\left(\frac{\log T}{\sqrt{s\mu}}\left(L + \frac{\alpha}{1-\gamma}\right)\sqrt{\alpha\bar{\alpha}}\sigma + \left(L + \sqrt{\frac{\alpha}{(1-\gamma)\mu}}\right)\lambda^{(1-\delta)t}\right);$$

*where* $\lambda$ *is defined in (3.8) above. Furthermore, if* $\alpha < \frac{2(1+\gamma)}{(\mu+L)(1-\gamma)}$, *we have* $\lambda = \sqrt{\gamma}$.

In the following remark, we choose the optimal momentum weights $\gamma$ to show the explicit convergence results with respect to $L$, $\mu$, and $s$.

**Remark 10** *For $\gamma$ and $\alpha$ in Theorem 1 satisfying*

$$\gamma = \frac{(1-\phi)^2}{(1+\phi)^2} + \epsilon; \quad \text{and} \quad \alpha \leq \sqrt{\frac{1}{\mu L}};$$

*for sufficiently small $\epsilon > 0$, we have that $\Delta = 4\alpha\mu\epsilon + O(\epsilon^2)$ and*

$$\sqrt{k\mathbb{E}\|x_{t+1}\|^2 + k x_{t+1} - x^*\|^2} = O\left(\frac{(L + 1/\mu)\log T}{\sqrt{s\epsilon}}\alpha\sigma + \frac{\sqrt{\alpha\mu}(L + 1/\mu)}{\sqrt{\epsilon}}(\sqrt{\gamma})^{(1-\delta)t}\right).$$

*We arrive at the same conclusions as we had in quadratic settings that SGDM can accelerate convergence over SGD under general strongly convex losses when $\bar{L} > 0$.*

# 4 Acceleration with Averaging

In this section, we study the Polyak-averaging SGDM (referred to as *averaged SGDM*), following Ruppert (1988) and Polyak and Juditsky (1992). Particularly, we aim at building the convergence result of $\frac{1}{n-n_0}\sum_{t=n_0+1}^{n} x_t$, which is an average of all iterates $x_t$ starting from a constant period $n_0$. We show that the averaging leads to acceleration for SGDM, compared to the last-iterate bounds established in the previous section.

## 4.1 Averaged SGDM under quadratic losses

For the mini-batch model (2.3) under quadratic losses where $\bar{L} = 0$, we first define the stochastic gradients evaluated at $x^*$,

$$z_t^* = \frac{1}{s}\sum_{i=1}^{s}(A_{\xi_{ti}}x^* - b_{\xi_{ti}}). \tag{4.10}$$

Here $\{z_t^*\}_{t\geq 1}$ are i.i.d. random vectors when $\xi_{ti}$ are i.i.d. sampled. We now characterize a decomposition of the error of the averaged SGDM $\frac{1}{n-n_0}\sum_{t=n_0+1}^{n}(x_t - x^*)$, which leads to an acceleration over the last iterate of SGDM $x_n$.

**Theorem 11** *Under (A1)-(A3) and* $\overline{L} = 0$*, suppose that the conditions in Theorem 1 hold, and the* $t$*-th iteration* $(w_t; x_t)$ *satisfies*

$$\mathbb{E}[\|w_t\|^2 + \|x_t - x^*\|^2] \le \frac{C_1}{s(1-\lambda)}\alpha^2\sigma^2 + C_2\|x_1 - x^*\|^2\lambda^{2(1-\delta)(t-1)}; \quad t \ge 1. \tag{4.11}$$

*Then for* $n \ge 2n_0$ *where* $\lambda^{2n_0} = s^{-1}(1-\lambda)$*, we have*

$$\frac{\sum_{t=n_0+1}^{n}(x_t - x^*)}{n - n_0} = -\frac{\sum_{t=n_0+1}^{n}\sum_{i=1}^{s}\Sigma^{-1}(A_{\xi_{ti}}x^* - b_{\xi_{ti}})}{s(n-n_0)} + R_n; \tag{4.12}$$

*where*

$$\mathbb{E}\|R_n\|^2 \le \frac{\bar{C}_1}{sn^2} + \frac{\bar{C}_2}{s^2 n};$$

$$\bar{C}_1 = 40\alpha^2 L_f^2 M^2 \frac{C_2\|x_1 - x^*\|^2}{(1-\delta)(1-\lambda)^3} + 120\alpha^2\sigma^2 M^2\frac{1}{(1-\lambda)^3} + \frac{4M^2\|x_1 - x^*\|^2}{1-\lambda};$$

$$\bar{C}_2 = 30\alpha^2\sigma^2 L_f^2 M^2 \frac{C_1\alpha^2}{(1-\lambda)^3};$$

*and* $M$ *is defined in (3.5).*

In the above theorem, we mainly obtain the convergence rate of averaged SGDM in (4.12), while (4.11) is nothing but a rephrasing of the last-iterate bounds established in (3.7), for the purpose of explicitly establishing the dependence of $\bar{C}_1$, $\bar{C}_2$ to the constants $C_1$, $C_2$ in (4.11). Comparing the convergence rates in Theorem 11, the averaged SGDM converges to $x^*$ without a bias as $n$ increases. The leading term $-\frac{\sum_{t=n_0+1}^{n}\Sigma^{-1}z_t^*}{n-n_0}$ is an average of $(n-n_0)$ i.i.d. vectors (as in (4.10)). Therefore the leading term is bounded by $O(\frac{1}{sn})$ in squared expectation, while the remainder term $R_n$ in (4.12) is bounded by $\frac{\bar{C}_1}{sn^2} + \frac{\bar{C}_2}{s^2 n}$.

**Remark 12** *The convergence rate of the remainder terms in Theorem 11 depends on* $\alpha$ *and* $\lambda$*, which implicitly depends on the momentum* $\gamma$*. When* $\gamma$ *increases in the range* $0 \le \gamma \le (1-\phi)^2/(1+\phi)^2$*, the spectral radius* $\lambda$ *decreases as we proved in Theorem 3 and Corollary 4. As a consequence, the term* $\bar{C}_2$ *in Theorem 11 decreases due to the factor* $(1-\lambda)^{-1}$*, which indicates a faster convergence. On the other hand, as* $\gamma$ *increases when* $\gamma$ *is close to* $1$*, we can see the increase of* $\lambda$ *and* $\bar{C}_2$*, and therefore the convergence is slower. In summary, as one increases* $\gamma$ *from* $0$ *to* $1$*, the convergence rate of averaged SGDM is first faster and then slower. Except for* $\gamma$ *being too large, averaged SGDM converges faster than the averaged SGD (*$\gamma = 0$*).*

**Remark 13** *It is noteworthy to mention that, for averaged SGDM, the initialization error* $\|x_1 - x^*\|$ *is forgotten at the rate of* $n^{-2}$*, slower than the exponential initialization-forgetting for the last-iterate convergence established in Theorem 1.*

In addition, the leading term in (4.12) is independent of momentum weight $\gamma$ and learning rate $\alpha$, which indeed indicates that averaged SGDM has the same rate of convergence as the averaged SGD. The following Corollary 14 further establishes that the asymptotic distribution of averaged SGDM is the same regardless of $\gamma$.

**Corollary 14** *Following Theorem 11, we have*

$$\frac{\sqrt{s}}{\sigma}\frac{\sum_{t=n_0+1}^{n}(x_t - x^*)}{\sqrt{n-n_0}} \Rightarrow N(0; \Sigma^{-1}\Omega\Sigma^{-1}); \quad \text{as} \quad \frac{s}{\alpha}; \alpha n \to \infty;$$

*where*

$$\Omega = \frac{1}{\sigma^2}\mathrm{E}[(A_\xi x^* - b_\xi)(A_\xi x^* - b_\xi)^\top]. \tag{4.13}$$

The asymptotic covariance matrix of the averaged SGDM depends on the error of loss function at $x^*$. When $x_t$ converges to a neighborhood of $x^*$, the stochastic gradient at $x_t$ is close to $z_t^*$, the stochastic gradient evaluated at $x^*$ defned in (4.10), due to Assumptions (A2){(A3). As the averaged SGDM is close to $x^*$, the limiting distribution is established with covariance matrix $\Sigma^{-1}\Omega\Sigma^{-1}$ where $\Sigma$ and $\Omega$ are respectively the Hessian and Gram matrix of $z_t^*$. The asymptotic covariance matrix matches the one in the existing literature of averaged SGD (Polyak and Juditsky, 1992). Furthermore, the asymptotic distribution established in Corollary 14 can be used to perform an online statistical inference for SGDM (Chen et al., 2020; Zhu et al., 2023), which we leave as an interesting future direction.

**Remark 15** *Asymptotic normality in Corollary 14 hold under asymptotics* $s/\alpha; \alpha n \to \infty$*. We specify two common scenarios of the learning rate* $\alpha$*, batch size* $s$ *and sample size* $n$*.*

- *For a constant learning rate* $\alpha$*, Corollary 14 holds as* $n$ *and* $s$ *tend to infinity. As the batch size* $s$ *increases, the bias term in (4.11) approaches zero, reducing the difference between the stochastic gradient at* $x_t$ *and* $x^*$*.*

- *For decaying learning rate* $\alpha$*, the asymptotic normality holds with a fixed batch size* $s$ *as long as* $\alpha n$ *diverges. The idea is intuitive: as alpha decreases, the random effect reduces. For instance, the condition of Corollary 14 is met when the batch size* $s$ *is fixed and* $\alpha = \Theta(n^{-\epsilon})$ *with* $\epsilon \in (0,1)$ *as* $n \to \infty$*.*

## 4.2 Averaged SGDM under general losses

For the general strongly convex loss function, we provide the convergence rate of the averaged SGDM in the following theorem.

**Theorem 16** *For* $\bar{L} > 0$*, suppose that the conditions in Theorem 7 hold, and the* $t$*-th iteration* $(w_t, x_t)$ *satisfies for fixed* $\delta \in (0, \frac{1}{2}]$*,*

$$\sqrt{\|w_t\|^2 + \|x_t - x^*\|^2} \le \frac{C_1}{\sqrt{s(1-\lambda)}}\alpha\sigma + C_2\|x_1 - x^*\|\lambda^{(1-\delta)(t-1)}; \quad 1 \le t \le T. \tag{4.14}$$

*Then for* $2n_0 \le n \le T$*, where* $\lambda^{n_0} = s^{-1/2}(1-\lambda)^{1/2}$*, with probability of at least* $1 - 4T^{-1}$*, we have*

$$\frac{\sum_{t=n_0+1}^{n}(x_t - x^*)}{n - n_0} = -\frac{\sum_{t=n_0+1}^{n}\sum_{i=1}^{s}\Sigma^{-1}\nabla f_{\xi_{ti}}(x^*)}{s(n-n_0)} + R_n; \tag{4.15}$$

*where*

$$\|R_n\| \le \frac{C_1 + C_3}{\sqrt{sn}} + \frac{C_2^2}{s\sqrt{n}} + \frac{C_4}{s}; \tag{4.16}$$

$$C_1 = \alpha L_f M\frac{16c(2\log T + 4d)C_2\|x_1 - x^*\|}{(1-\lambda)^{3/2}} + \alpha\sigma M\frac{4\sqrt{3c(2\log T + 4d)}}{(1-\lambda)^{3/2}} + \frac{2M\|x_1 - x^*\|}{(1-\lambda)^{1/2}};$$

$$C_2 = \alpha L_f \sigma M\frac{8\sqrt{2c(2\log T + 4d)C_1\alpha}}{(1-\lambda)^{3/2}};$$

$$C_3 = 8\alpha M\bar{L}\frac{C_2^2\|x_1 - x^*\|^2(n_0(1-\lambda)+1)}{(1-\lambda)^{3/2}}; \quad C_4 = 4\alpha\sigma^2 M\bar{L}\frac{C_1^2\alpha^2}{(1-\lambda)^2};$$

*where* $M$ *is defined in* (3.5).

Using the same arguments as in Remark 12, we can also see that SGDM converges faster than SGD, and the error bounds first decrease then increase when we moves $\gamma$ from 0 to 1, due to the factor $(1-\lambda)^{-1}$ in the error bounds.

For non-quadratic settings of $\bar{L} > 0$, there are two additional terms $C_3=(\sqrt{s}n)$ and $C_4=s$ in (4.16) compared to the case of $\bar{L} = 0$. As $n \to \infty$, the term $C_4=s$ appears as a non-vanishing bias, if the batch size $s$ and learning rate $\alpha$ both stays fixed, due to which the asymptotic normality does not hold. Nonetheless, if one decays $\alpha$ or increases $s$ appropriately as $n$ increases, the asymptotic normality result remains to hold in the following corollary.

**Corollary 17** *Following Theorem 16, we have*

$$\frac{\sqrt{s}}{\sigma}\frac{\sum_{t=n_0+1}^{n}(x_t - x^*)}{\sqrt{n-n_0}} \Rightarrow N(0; \Sigma^{-1}\Omega\Sigma^{-1}); \quad \text{as} \quad \frac{\sqrt{\alpha n}}{\log(s/\alpha)}; \frac{\sqrt{s}}{\alpha\sqrt{n}} \to \infty;$$

*where*

$$\Omega = \frac{1}{\sigma^2}\mathrm{E}[\nabla f_\xi(x^*)\nabla f_\xi(x^*)^\top]. \tag{4.17}$$

**Remark 18** *The conditions on $n$, $s$, and $\alpha$ in this corollary are more restrictive than those in Corollary 14 due to the presence of the approximation errors $\bar{L}\|x - x^*\|^2$. For a decaying learning rate $\alpha$, the condition on the batch size is much relaxed since a small learning rate reduces the error caused by randomness, as shown in the $C_1$ term in (4.14). Particularly, when the batch size $s$ is fixed, the condition of Corollary 17 is met for $\alpha = \Theta(n^{-\epsilon})$ with $\epsilon \in (\frac{1}{2}; 1)$ as $n \to \infty$.*

# 5 Experiments

In this section, we support our theoretical results with simulations in a quadratic example and a logistic loss example of the general strongly convex loss, and real-data experiments of a multinomial logistic regression on MNIST hand-written digit classification.

In the numerical experiments, we verify the convergence results built in the paper on a training set $\{1; 2; \ldots; N\}$. Setting $\xi$ to be a uniform random variable with values in $\{1; 2; \ldots; N\}$, we consider to minimize $\min_x \frac{1}{N}\sum_{i=1}^{N} f_i(x)$, as an example of the stochastic optimization model (1.1), where the objective is either the empirical risk $\mathrm{E}[f_\xi(x)] = \frac{1}{N}\sum_{i=1}^{N} f_i(x)$, or more specifically in a maximum likelihood estimation, the negative log-likelihood.

In all the simulation results below, we repeat the experiment 200 times. For the real-data analysis on MNIST, we repeat the experiment in a replicable setting while we fix the random seeds to 1, 2, and 3 in training.

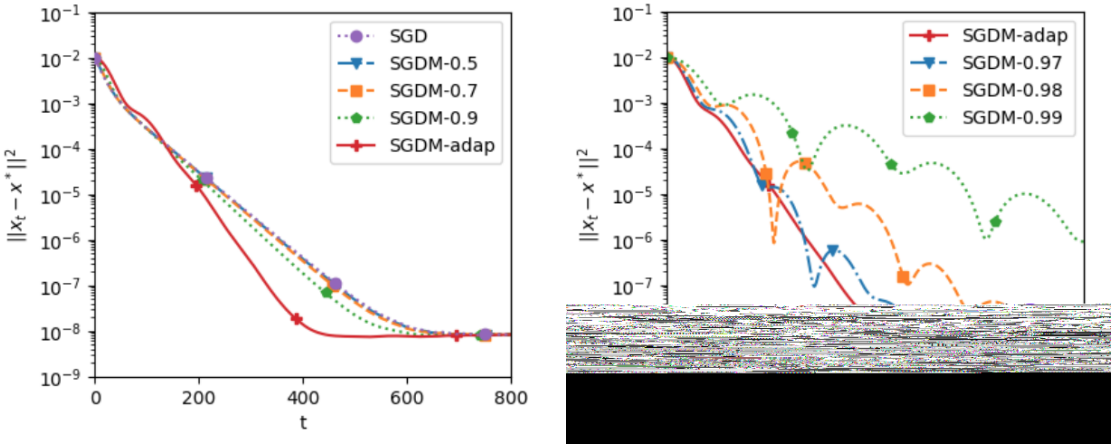
## 5.1 Simulation: quadratic loss

In a quadratic loss model (3.4), we first conduct a simulation study with a sample of size $N = 20{,}000$ and dimension $d = 10$. We generate the parameters $\{(A_i, b_i)\}_{i=1}^N$ in (3.4) i.i.d and $b_i \sim N(0_d, I_d)$, and the positive definitive matrix $A_i = \rho V_i^\top V_i + 10 I_d$. Here $V_i$ is a matrix in $\mathbb{R}^{d\times d}$ and each row of $V_i$ generates from the normal distribution $N(0_d, I_d)$, where $\rho > 0$ affect the conditional number $L/\mu$. We choose $\rho = 1$ and the average conditional number $L/\mu$ is 35/10. The exact minimizer $x^*$ is computed by $x^* = \left(\sum_{i=1}^N A_i\right)^{-1} \left(\sum_{i=1}^N b_i\right)$.

We consider the mini-batch SGDM with batch size $s = 0.2N$ with replacement. The learning rate is fixed at $\alpha = 0.001$. According to Theorem 3, we choose the momentum weight following

$$\gamma = \left(\frac{1-\mu\alpha}{1+\mu\alpha}\right)^2. \tag{5.18}$$

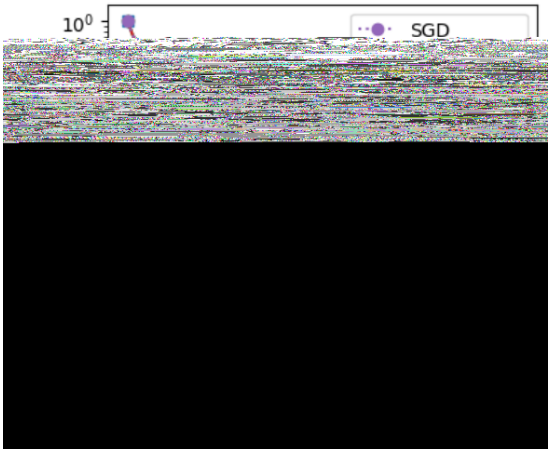
We refer to the above $\gamma$ as the adaptive momentum weight in SGDM (`SGDM-adap`), while we also compare it with other fixed $\gamma$, as well as SGD as a special case of SGDM with $\gamma = 0$ in Figure 1.




(a) Small $\gamma$ (b) Large $\gamma$

Figure 1: Performance of SGD and SGDM on the quadratic loss. The average value of the adaptive momentum weight is 0.96.

Figure 1 supports the theoretical results in Theorems 1 and 3. Both SGD and SGDM enjoy linear convergence at the beginning and have the same order of error at the end. The adaptive momemtum `SGDM-adap` with $\gamma = (1-\phi)^2/(1+\phi)^2$ in (5.18) converges fastest, where the average




(a) Small γ (b) Large γ

Figure 4: Performance of SGD and SGDM on the logistic loss, and the average value of the adaptive momentum weight is 0.75.

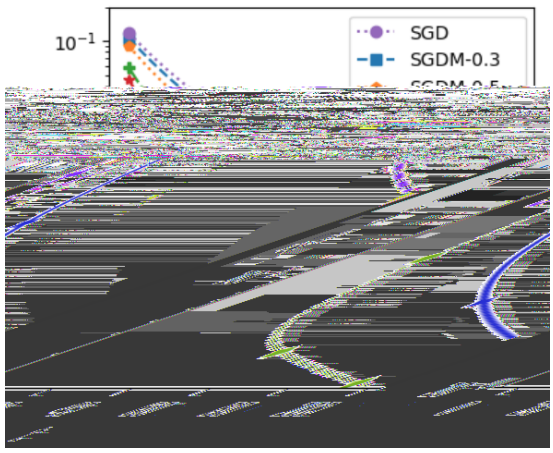
one-hot indicators corresponding the digits $0; 1; \cdots; 9$, the loss function is

$$f_i(X) = H(X^\top a_i; b_i);$$

where $X \in \mathrm{R}^{784\times 10}$ and $H$ is the multi-class cross entropy. We specify the batch size $s = 256$ and the learning rate $\alpha = 1.0$. Due to the difficulty of computing the condition number $L/\mu$, we fix the momentum weight $\gamma = 0.1; 0.3; 0.5; 0.7; 0.9; 0.99$ in training.

Figure 6 shows the convergence of the training loss over iterations $t$. For the purpose of clear representation, the reported training loss for each iteration $t$ is based on an average of mini-batch stochastic losses $g_{\eta_t}$ of the past $bN/sc$ batches, with the batch size $s$.

From Figure 6, we see that SGDM with momentum weight γ ranging from 0.1 to 0.9 greatly outperforms SGD. SGDM with $\gamma = 0.5$ converges fastest in the earlier iterations, while the training losses are almost identical for different momentum weights in the later iterations, except for $\gamma = 0.99$. The convergence rate is not very sensitive to the momentum weights and a wide range of $\gamma \in [0.3; 0.9]$ leads to similar performance.




(a) Small $\gamma$ (b) Large $\gamma$

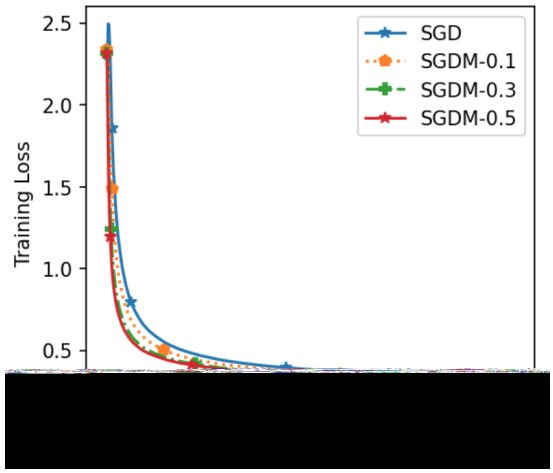
Figure 5: Performance of Averaged SGD and Averaged SGDM on the logistic loss. The average value of the adaptive momentum weight is 0.75.

## 6 Conclusion

Our study sheds light on the performance of mini-batch SGD with momentum (SGDM) in solving optimization problems under smooth and strongly convex loss functions. The convergence rate of mini-batch SGDM is influenced by several factors, including those of the batch size, the momentum weight, and the learning rate. Our analysis showed that certain choices of momentum weight chosen with a reasonably large batch size can lead to faster convergence compared to mini-batch SGD. This finding is consistent with previous studies on the numerical performance of SGDM. We further establish the asymptotic normality of averaged SGDM in the quadratic settings and reveal the non-vanishing bias of that in the general settings.

Furthermore, our investigation of the impact of averaging mini-batch SGDM trajectories on the convergence rate of the algorithm revealed that the averaging of trajectories accelerate the convergence rate of the algorithm, particularly in the early phases of the iteration process. This result suggests that averaging trajectories is a useful technique for enhancing the performance of mini-batch SGDM in practice.

In summary, our study contributes to the theoretical understanding of mini-batch SGDM and has practical implications for developing efficient optimization algorithms in machine learning. It




(a) Small γ (b) Large γ

Figure 6: Performance of SGD and SGDM on MNIST.

is noteworthy to mention that our study assumed that the loss function was smooth and strongly convex. In future work, our work can be extended in several ways. It would be interesting to investigate the performance of mini-batch SGDM under non-convex loss functions, since this would have important implications for the application of SGDM in deep learning. It is also of interest to extend the analysis to the case when the learning rate decays over time.

## Acknowledgments

Weidong Liu's research is supported by NSFC Grant No. 11825104, Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

## References

Bollapragada, R., T. Chen, and R. Ward (2022). On the fast convergence of minibatch heavy ball momentum. *arXiv preprint arXiv:2206.07553*.

Bottou, L., F. E. Curtis, and J. Nocedal (2018). Optimization methods for large-scale machine learning. *SIAM Review 60*(2), 223{311.

Chen, X., J. D. Lee, X. T. Tong, and Y. Zhang (2020). Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics 48*(1), 251 { 273.

Defazio, A. (2020). Understanding the role of momentum in non-convex optimization: Practical insights from a lyapunov analysis. *arXiv preprint arXiv:2010.00406*.

Gitman, I., H. Lang, P. Zhang, and L. Xiao (2019). Understanding the role of momentum in stochastic gradient methods. *Advances in Neural Information Processing Systems 32*.

Jin, R., Y. Xing, and X. He (2022). On the convergence of msgd and adagrad for stochastic optimization. *arXiv preprint arXiv:2201.11204*.

Kidambi, R., P. Netrapalli, P. Jain, and S. Kakade (2018). On the insu f ciency of existing momentum schemes for stochastic optimization. In *Information Theory and Applications Workshop*, pp. 1{9. IEEE.

Li, X., M. Liu, and F. Orabona (2022). On the last iterate convergence of momentum methods. In *International Conference on Algorithmic Learning Theory*, pp. 699{717. PMLR.

Liu, Y., Y. Gao, and W. Yin (2020). An improved analysis of stochastic gradient descent with momentum. *Advances in Neural Information Processing Systems 33*, 18261{18271.

Loizou, N. and P. Richtárik (2017). Linearly convergent stochastic heavy ball method for minimizing generalization error. *arXiv preprint arXiv:1710.10737*.

Loizou, N. and P. Richtárik (2020). Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *Computational Optimization and Applications 77*(3), 653{710.

Mai, V. and M. Johansson (2020). Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization. In *International Conference on Machine Learning*, pp. 6630{6639. PMLR.

Moulines, E. and F. Bach (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems 24*.

Nesterov, Y. (1983). A method for unconstrained convex minimization problem with the rate of convergence $o(1=k^2)$. In *Doklady AN USSR*, Volume 269, pp. 543{547.

Nesterov, Y. (2003). *Introductory lectures on convex optimization: A basic course*, Volume 87. Springer Science & Business Media.

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

Nocedal, J. and S. J. Wright (2006). Numerical optimization. springer series in operations research. *SIAM J Optimization*.

Polyak, B. T. (1964). Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics 4*(5), 1{17.

Polyak, B. T. and A. B. Juditsky (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization 30*(4), 838{855.

Richtárik, P. and M. Takác (2020). Stochastic reformulations of linear systems: algorithms and convergence theory. *SIAM Journal on Matrix Analysis and Applications 41*(2), 487{524.

Robbins, H. and S. Monro (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 400{407.

Ruppert, D. (1988). E f cient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering.

Sebbouh, O., R. M. Gower, and A. Defazio (2021). Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Conference on Learning Theory*, pp. 3935{3971. PMLR.

Wang, X. and M. Johansson (2022). On uniform boundedness properties of sgd and its momentum variants. *arXiv preprint arXiv:2201.10245*.

Yan, Y., T. Yang, Z. Li, Q. Lin, and Y. Yang (2018). A unifed analysis of stochastic momentum methods for deep learning. *arXiv preprint arXiv:1808.10396*.

Yang, T., Q. Lin, and Z. Li (2016). Unifed convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv preprint arXiv:1604.03257*.

Zhu, W., X. Chen, and W. B. Wu (2023). Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association 118*(541), 393{404.

# A Proofs of finite-time convergence rates of SGDM

## A.1 Proof of Theorem 1

Recall that

$$m_{t+1} = \gamma m_t + (1-\gamma)\nabla g_{\eta_t}(x_t); \quad x_{t+1} = x_t - \alpha m_{t+1}.$$

Put

$$\Gamma = \begin{pmatrix} \gamma I; & (1-\gamma)\Sigma \\ -\alpha\gamma I; & I - \alpha(1-\gamma)\Sigma \end{pmatrix}.$$

Let $\varepsilon_t = x_t - x^*$ and $\varepsilon m_{t+1} = (1-\gamma)\sum_{j=1}^{t}\gamma^{t-j}\Sigma\varepsilon_j$, we can write

$$\begin{aligned}
\begin{pmatrix} \varepsilon m_{t+1} \\ \varepsilon_{t+1} \end{pmatrix} &= \Gamma \begin{pmatrix} \varepsilon m_t \\ \varepsilon_t \end{pmatrix} - \alpha(1-\gamma)\begin{pmatrix} 0 \\ \sum_{j=1}^{t}\gamma^{t-j}(\nabla g_{\eta_j}(x_j) - \Sigma\varepsilon_j) \end{pmatrix} \\
&= \Gamma^t \begin{pmatrix} \varepsilon m_1 \\ \varepsilon_1 \end{pmatrix} - \alpha(1-\gamma)\sum_{j=1}^{t}\Gamma^{t-j}\begin{pmatrix} 0 \\ \sum_{k=1}^{j}\gamma^{j-k}(\nabla g_{\eta_k}(x_k) - \Sigma\varepsilon_k) \end{pmatrix}. \qquad \text{(A.20)}
\end{aligned}$$

Define

$$q_t := \mathrm{E}[\|\varepsilon m_t\|^2 + \|\varepsilon_t\|^2].$$

Recall for $\overline{L} = 0$, there hold that $\nabla g(x) = \mathrm{E}[\nabla g_\eta(x)] = \Sigma(x - x^*)$ and

$$\nabla g_{\eta_k}(x_k) - \Sigma\varepsilon_k = (\nabla g_{\eta_k}(x_k) - \nabla g_{\eta_k}(x^*) + \nabla g(x^*) - \nabla g(x_k)) + \nabla g_{\eta_k}(x^*).$$

By the $L_f$-smooth of the individual function $f_\xi(x)$ in (A3) and the independence of $L_\xi$ and $\varepsilon_k$, we have

$$\mathrm{E}\|\nabla f_\xi(x_k) - \nabla f_\xi(x^*) + \nabla f(x^*) - \nabla f(x_k)\|^2 \le \mathrm{E}\|\nabla f_\xi(x_k) - \nabla f_\xi(x^*)\|^2 \le L_f^2\mathrm{E}\|\varepsilon_k\|^2;$$

and

$$\begin{aligned}
q_{t+1} \le\ & \|\Gamma^t\|^2 q_1 + 2\alpha^2(1-\gamma)^2\frac{\sigma^2}{s}\sum_{k=1}^{t}\left(\sum_{j=k}^{t}\|\Gamma^{t-j}\|\gamma^{j-k}\right)^2 \\
& + 2\alpha^2(1-\gamma)^2\frac{L_f^2}{s}\sum_{k=1}^{t}\left(\sum_{j=k}^{t}\|\Gamma^{t-j}\|\gamma^{j-k}\right)^2 q_k. \qquad \text{(A.21)}
\end{aligned}$$

In the remaining part of proof, we use the inductive method to derive the bound of $q_t$. Let $C_1 \geq 1$ and $C_2 \geq 1$ be some constants which will be specifed later. We will prove that, for any $t \geq 1$, the inequalities

$$q_j \leq \frac{C_1}{s(1-\lambda)}\alpha^2\sigma^2 + C_2 q_1 \lambda^{2(1-\delta)(j-1)}; \quad 1 \leq j \leq t \tag{A.22}$$

with fxed $\delta \in (0,1]$, imply that

$$q_{t+1} \leq \frac{C_1}{s(1-\lambda)}\alpha^2\sigma^2 + C_2 q_1 \lambda^{2(1-\delta)t}. \tag{A.23}$$

By Lemma 20 below, we have $\Gamma^j \leq M\lambda^j$. From Theorem 3, we have $\lambda \geq \sqrt{\gamma}$. To handle the sum in (A.21), we have the following lemma:

**Lemma 19** *For $0 \leq \lambda, \gamma < 1$, we have*

$$\sum_{k=1}^{\infty}\left(\sum_{j=k}^{\infty}\lambda^{t-j}\gamma^{j-k}A\right)^2 \leq \frac{2}{(1-\gamma)^2(1-\lambda)}.$$

*In addition, if $\lambda \geq \sqrt{\gamma}$, for fixed $\delta \in (0,1]$, we have*

$$\sum_{k=1}^{\infty}\left(\sum_{j=k}^{\infty}\lambda^{t-j}\gamma^{j-k}A\right)^2\lambda^{2(1-\delta)(k-1)} \leq \frac{4}{\delta(1-\gamma)^2(1-\lambda)}\lambda^{2(1-\delta)(t-1)};$$

*and*

$$\sum_{k=1}^{\infty}\sum_{j=k}^{\infty}\lambda^{t-j}\gamma^{j-k}\lambda^{(1-\delta)(k-1)} \leq \frac{2}{\delta(1-\gamma)(1-\lambda)}\lambda^{(1-\delta)(t-1)}.$$

**Proof.**

$$\begin{aligned}
\sum_{k=1}^{\infty}\left(\sum_{j=k}^{\infty}\lambda^{t-j}\gamma^{j-k}A\right)^2 &= \sum_{k=1}^{\infty}\sum_{i=k}^{\infty}\sum_{j=k}^{\infty}\lambda^{2t-i-j}\gamma^{i+j-2k} = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\lambda^{2t-i-j}\sum_{k=1}^{\min\{i,j\}}\gamma^{i+j-2k} \\
&\leq \frac{1}{1-\gamma}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\lambda^{2t-i-j}\gamma^{|i-j|} \leq \frac{2}{1-\gamma}\sum_{k=0}^{t-1}\sum_{j=1}^{t-k}\lambda^{2t-2j-k}\gamma^{k} \\
&\leq \frac{2}{1-\gamma}\sum_{k=0}^{t-1}\frac{\lambda^k}{1-\lambda^2}\gamma^k \leq \frac{2}{(1-\gamma)(1-\lambda)(1-\lambda\gamma)} \leq \frac{2}{(1-\gamma)^2(1-\lambda)}.
\end{aligned}$$

If $\lambda \geq \sqrt{\gamma}$, we have

$$\lambda^{-2(1-\delta)(t-1)} \sum_{k=1}^{t} \Big( \sum_{j=k}^{t} \lambda^{t-j} \gamma^{j-k} \Big)^2 \lambda^{2(1-\delta)(k-1)} = \sum_{k=1}^{t} \sum_{i=k}^{t} \sum_{j=k}^{t} \lambda^{2k-i-j} \gamma^{i+j-2k} \lambda^{2\delta(t-k)}$$

$$\leq \sum_{i=1}^{t} \sum_{j=1}^{t} \lambda^{2\delta(t-\min\{i,j\})} \sum_{k=1}^{\min\{i,j\}} \Big( \frac{\gamma}{\lambda} \Big)^{i+j-2k} \leq \frac{1}{1-\gamma^2/\lambda^2} \sum_{i=1}^{t} \sum_{j=1}^{t} \lambda^{2\delta(t-\min\{i,j\})} \big(\sqrt{\gamma}\big)^{|i-j|}$$

$$\leq \frac{2}{1-\gamma} \sum_{i=1}^{t} \sum_{k=0}^{t-i} \lambda^{2\delta(t-i)} \big(\sqrt{\gamma}\big)^{k} \leq \frac{4}{(1-\gamma)^2(1-\lambda^{2\delta})} \leq \frac{4}{\delta(1-\gamma)^2(1-\lambda)};$$

where the last inequality is due to the fact that $\frac{1-x}{1-x^\delta} \leq \frac{1}{\delta}$ for $\delta \in (0,1]$. Similarly, we have

$$\lambda^{-(1-\delta)(t-1)} \sum_{k=1}^{t} \sum_{j=k}^{t} \lambda^{t-j} \gamma^{j-k} \lambda^{(1-\delta)(k-1)} = \sum_{k=1}^{t} \sum_{j=k}^{t} \lambda^{\delta(t-k)} \Big( \frac{\gamma}{\lambda} \Big)^{j-k}$$

$$\leq \sum_{j=1}^{t} \lambda^{\delta(t-j)} \sum_{k=1}^{j} \big(\sqrt{\gamma}\big)^{j-k} \leq \frac{2}{1-\gamma} \frac{1}{1-\lambda^{\delta}} \leq \frac{2}{\delta(1-\gamma)(1-\lambda)}.$$

Then from (A.21) and (A.22), it holds that

$$q_{t+1} \leq M^2 \lambda^{2t} q_1 + 2\alpha^2 \frac{\sigma^2}{s} M^2 \frac{2}{1-\lambda}$$

$$+ 2\alpha^2 (1-\gamma)^2 \frac{L_f^2}{s} M^2 \left( \frac{2C_1 \alpha^2 \sigma^2}{s(1-\gamma)^2(1-\lambda)^2} + C_2 q_1 \lambda^{2(1-\delta)t} \frac{1}{\lambda^{2(1-\delta)}} \frac{4}{\delta(1-\gamma)^2(1-\lambda)} \right)$$

$$\leq \frac{C_1}{s(1-\lambda)} \alpha^2 \sigma^2 M^2 \left( \frac{4}{C_1} + \frac{4\alpha^2 L_f^2}{s(1-\lambda)} \right) + C_2 q_1 \lambda^{2(1-\delta)t} M^2 \left( \frac{\lambda^{2\delta t}}{C_2} + \frac{8\alpha^2 L_f^2}{s\lambda^{2(1-\delta)}\delta(1-\lambda)} \right).$$

Let the learning rate $\alpha$ and the batch size $s$ satisfy

$$8M^2 \frac{1}{\lambda^{2(1-\delta)}} \frac{1}{\delta(1-\lambda)} \alpha^2 \frac{L_f^2}{s} \leq \frac{1}{2};$$

we can take

$$C_1 = 8M^2; \quad C_2 = 2M^2.$$

Then it is easy to see that (A.23) holds, and the induction is completed.

## A.2 Proof of Theorem 3

Put

$$\Gamma = \begin{pmatrix} U, & 0_d \\ 0_d, & U \end{pmatrix} \begin{pmatrix} \gamma I, & (1-\gamma)A \\ -\alpha\gamma I, & I - \alpha(1-\gamma)A \end{pmatrix} \begin{pmatrix} U^\top, & 0_d \\ 0_d, & U^\top \end{pmatrix}$$

where $\Sigma = U diag(\kappa_1, \ldots, \kappa_d) U^T =: UAU^T$, $\mu = \kappa_1 \leq \cdots \leq \kappa_d = L$, $U$ is an orthogonal matrix. Now we can define

$$\lambda_k^{\pm} = \frac{\gamma + 1 - \alpha(1-\gamma)\kappa_k \pm \sqrt{(\alpha(1-\gamma)\kappa_k - \gamma - 1)^2 - 4\gamma}}{2},$$

for $1 \leq k \leq d$, and the diagonal matrix

$$\Lambda = \begin{pmatrix} diag(\lambda_1^+, \ldots, \lambda_d^+) & 0_d \\ 0_d & diag(\lambda_1^-, \ldots, \lambda_d^-) \end{pmatrix}. \tag{A.24}$$

Notice that if $(\alpha(1-\gamma)\kappa_k - \gamma - 1)^2 < 4\gamma$, $\lambda_k^{\pm}$ is complex and the definition is

$$\lambda_k^{\pm} = \gamma + 1 - \alpha(1-\gamma)\kappa_k \pm \sqrt{}$$

where

$$\phi = \min\left\{\alpha\mu, \frac{2(1+\gamma)}{1-\gamma} - \alpha L\right\}.$$

In this condition, there holds that $\gamma + 1 - (1-\gamma)\phi > 2\sqrt{\gamma}$. Define $\eta = (\gamma + 1 - (1-\gamma)\phi)^2 - 4\gamma > 0$,

$$h'(\gamma) = \frac{(1+\phi)h(\gamma) - 1}{\sqrt{\eta}}, \quad h''(\gamma) = \frac{h'(\gamma)\left((1+\phi)\sqrt{\eta} + 2 - (1+\phi)(\gamma + 1 - (1-\gamma)\phi)\right)}{2\eta}.$$

By the fact that $h(0) = 1 - \phi, h'(0) < 0$ and $2 - (1+\phi)(\gamma + 1 - (1-\gamma)\phi) \geq 0$, where $0 \leq \gamma < (1-\phi)^2/(1+\phi)^2$, we have that $h$ is concave and

$$1 - \phi - \frac{\phi(1+\phi)}{1-\phi}\gamma \leq \lambda \leq 1 - \phi - \frac{\phi^2}{1-\phi}\gamma.$$

For complex eigenvalues, which means that $|\gamma + 1 - \alpha(1-\gamma)\kappa_k| \leq 2\sqrt{\gamma}$ for all $k$. Then $\gamma \geq (1-\phi)^2/(1+\phi)^2$ and

$$\lambda = \max_k \lambda_k^{\pm} = \max_k \left(\frac{\gamma + 1 - \alpha(1-\gamma)\kappa_k \pm \sqrt{-1}\sqrt{4\gamma - (\alpha(1-\gamma)\kappa_k - \gamma - 1)^2}}{2}\right) = \sqrt{\gamma}.$$

## A.3 Proof of Theorem 7

Recall that

$$\begin{pmatrix} \mathbf{e}_{t+1} \\ \mathbf{x}_{t+1} \end{pmatrix} = \Gamma^t \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{x}_1 \end{pmatrix} - \alpha(1-\gamma)\sum_{j=1}^{t} \Gamma^{t-j} \begin{pmatrix} 0 \\ \sum_{k=1}^{j} \gamma^{j-k}(\mathbf{g}_{\eta_k}(x_k) - \Sigma\mathbf{x}_k) \end{pmatrix}.$$

For $\bar{L} > 0$, it holds that

$$\begin{aligned} \mathbf{g}_{\eta_k}(x_k) - \Sigma\mathbf{x}_k &= (\mathbf{g}_{\eta_k}(x_k) - \mathbf{g}_{\eta_k}(x^*) + \mathbf{g}(x^*) - \mathbf{g}(x_k)) + \mathbf{g}_{\eta_k}(x^*) + \mathbf{g}(x_k) - \Sigma\mathbf{x}_k \\ &= (\mathbf{g}_{\eta_k}(x_k) - \mathbf{g}_{\eta_k}(x^*) + \mathbf{g}(x^*) - \mathbf{g}(x_k)) + \mathbf{g}_{\eta_k}(x^*) + V_k\mathbf{x}_k; \end{aligned} \tag{A.28}$$

where $\|V_k\| = \left\|\int_0^1 \Sigma(x^* + y\mathbf{x}_k) - \Sigma(x^*)dy\right\| \leq \bar{L}\|\mathbf{x}_k\|$.

The purpose of the following proof is to provide bounds for three parts in (A.28). To this end, we begin by introducing some useful lemmas concerning sub-exponential vectors.

**Lemma 21** *Suppose $x_k \in \mathbb{R}^d$ are independent, zero-mean $\sigma_k$-sub-exponential random vector, we have*

$$\mathbb{P}\left(\sum_{k=1}^{n} x_k > c\left(\log(1/\epsilon) + 2d\right)\sqrt{\sum_{k=1}^{n}\sigma_k^2}\right) \leq \epsilon;$$

*where $c > 0$ is a absolute constant.*

**Proof.** Let $N$ be the 1/2-net of the unit ball $\{z \in \mathbb{R}^d : \|z\| \le 1\}$ with respect to the Euclidean norm that satisfes $|N| \le 6^d$. By the inequality that

$$\max_{\|v\|=1} v^\top x \le \max_{z\in\mathcal{N}} z^\top x + \max_{\|v\|=\frac{1}{2}} v^\top x = \max_{z\in\mathcal{N}} z^\top x + \frac{1}{2}\max_{\|v\|=1} v^\top x;$$

we have

$$P[\max_{\|v\|=1} v^\top \sum_{k=1}^n x_k > \delta] \le P[2\max_{z\in\mathcal{N}} z^\top \sum_{k=1}^n x_k > \delta] \le 6^d \sup_{\|z\|=1} P[2z^\top \sum_{k=1}^n x_k > \delta]:$$

According to Bernstein's inequality, there holds

$$P[2z^\top \sum_{k=1}^n x_k > \delta] \le \exp\left(-c_0 \min\left(\frac{\delta^2}{4\sum_{k=1}^n \sigma_k^2}; \frac{\delta}{2\max_k \sigma_k}\right)\right);$$

for some absolute constant $c_0$. To bound $P[\max_{\|v\|=1} v^\top \sum_{k=1}^n x_k > \delta] \le \epsilon$, we fnd $\delta$ such that

$$\delta = 2\max\left\{\sqrt{\sum_{k=1}^n \sigma_k^2}\sqrt{\frac{1}{c_0}\log(1/\epsilon) + 2d}; \max_k \sigma_k \frac{1}{c_0}(\log(1/\epsilon) + 2d)\right\} \le c\sqrt{\sum_{k=1}^n \sigma_k^2}(\log(1/\epsilon) + 2d);$$

where $c = 2\max\left\{\frac{1}{c_0}; \frac{1}{\sqrt{c_0}}\right\}$:

**Lemma 22** *Under (A1)-(A2) and (A3'), we have*

$$P\left(\sum_{j=1}^n \Gamma^{t-j}\left(\sum_{k=1}^j \gamma^{j-k} 0 g_{\eta_k}(x^*)\right)\right) \ge M\frac{\sigma}{\bar{s}}\frac{\sqrt{2c}(2\log T + 4d)}{(1-\lambda)^{1/2}(1-\gamma)}\Bigg) \le \frac{1}{T^2};$$

*for* $1 \le t \le T$.

**Proof.** Defne

$$Y_{t;k} = \sum_{j=k}^n \Gamma^{t-j}\gamma^{j-k}\begin{pmatrix}0\\ 0 g_{\eta_k}(x^*)\end{pmatrix}:$$

By the defnition of the sub-exponential, $Y_{t;k}$ is $\sum_{j=k}^t \Gamma^{t-j}\gamma^{j-k}\frac{c\sigma}{\sqrt{s}}$-sub-exponential random vector for some absolute constant $c > 0$, and $\{Y_{t;k}\}_{k=1}^t$ are independent. Then by Lemmas 19 and 20, we have

$$\sum_{k=1}^n \left\|\sum_{j=k}^n \Gamma^{t-j}\gamma^{j-k}\right\|^2 \le M^2\frac{2}{(1-\lambda)(1-\gamma)^2}:$$

By applying Lemma 21 on $\{Y_{t;k}\}_{k=1}^t$, with probability $1 - 1/T^2$, we can obtain the following result:

$$\sum_{k=1}^n Y_{t;k} \le M\frac{\sigma}{\bar{s}}\frac{\sqrt{2c}(2\log T + 4d)}{(1-\lambda)^{1/2}(1-\gamma)}:$$

**Lemma 23** *Under (A1)-(A2) and (A3'), suppose that the $t$-th iteration $(\boldsymbol{\theta}_t, \boldsymbol{x}_t)$ satisfies*

$$\sqrt{\|\boldsymbol{\theta}_j\|^2 + \|\boldsymbol{x}_j\|^2} \le \frac{C_1}{\sqrt{s(1-\lambda)}}\alpha\sigma + C_2 q_1^{1/2}\lambda^{(1-\delta)(j-1)},$$

*for $1 \le j \le t$, we have*

$$\mathbb{P}\left(\sum_{j=1}^{t}\Gamma^{t-j}\left(\sum_{k=1}^{j}\gamma^{j-k}\left(\nabla g_{\eta_k}(x_k) - \nabla g_{\eta_k}(x^*) + \nabla g(x^*) - \nabla g(x_k)\right)\right) \ge C\right) \le \frac{1}{T^2},$$

*where*

$$C = c(2\log T + 4d)M\left(\frac{L_f}{\sqrt{s}} + \frac{4C_1\alpha\sigma}{\sqrt{s(1-\lambda)(1-\gamma)}} + \frac{4\sqrt{2}C_2 q_1^{1/2}\lambda^{(1-\delta)(t-1)}}{\sqrt{\delta(1-\lambda)^{1/2}(1-\gamma)}}\right).$$

**Proof.** Define

$$Y_{t,k} = \sum_{j=k}^{t}\Gamma^{t-j}\gamma^{j-k}\begin{pmatrix}0\\ \nabla g_{\eta_k}(x_k) - \nabla g_{\eta_k}(x^*) + \nabla g(x^*) - \nabla g(x_k)\end{pmatrix}.$$

According to the iteration of SGDM, we have $\xi_k$ and $x_k$ are independent and $\{Y_{t,k}\}_{k=1}^{t}$ are martingale difference. Due to the $L_f$-smooth of the individual gradient $\nabla f_\xi(x)$, it holds that

$$\max_{\|v\|=1}\mathbb{E}\left[\exp\left(v^\top Y_{t,k}\right)\middle|\sigma_k\right] \le 2,$$

where

$$\sigma_k = c\frac{L_f}{\sqrt{s}}\sum_{j=k}^{t}\Gamma^{t-j}\gamma^{j-k}\|\boldsymbol{x}_k\|,$$

and $c > 0$ is an absolute constant. Then according to Lemmas 19 and 20, we have

$$\sum_{k=1}^{t}\left(\sum_{j=k}^{t}\Gamma^{t-j}\gamma^{j-k}\right)^2\|\boldsymbol{x}_k\|^2 \le M^2\left(2C_1^2\alpha^2\frac{\sigma^2}{s}\frac{2}{(1-\gamma)^2(1-\lambda)^2} + 2C_2^2 q_1\frac{4\lambda^{2(1-\delta)(t-1)}}{\delta(1-\gamma)^2(1-\lambda)}\right).$$

By applying Lemma 21 on $\{Y_{t,k}\}_{k=1}^{t}$, with probability $1 - 1/T^2$, we can obtain the following result:

$$\sum_{j=1}^{t}Y_{t,j} \le c(2\log T + 4d)M\left(\frac{L_f}{\sqrt{s}} + \frac{4C_1\alpha\sigma}{\sqrt{s(1-\lambda)(1-\gamma)}} + \frac{4\sqrt{2}C_2 q_1^{1/2}\lambda^{(1-\delta)(t-1)}}{\sqrt{\delta(1-\lambda)^{1/2}(1-\gamma)}}\right).$$

**Lemma 24** *Under (A1)-(A2) and (A3'), suppose that the $t$-th iteration $(\boldsymbol{\theta}_t, \boldsymbol{x}_t)$ satisfies*

$$\sqrt{\|\boldsymbol{\theta}_j\|^2 + \|\boldsymbol{x}_j\|^2} \le \frac{C_1}{\sqrt{s(1-\lambda)}}\alpha\sigma + C_2 q_1^{1/2}\lambda^{(1-\delta)(j-1)},$$

*for* $1 \le j \le t$, *for* $\bar{L} > 0$, *we have*

$$\sum_{j=1}^{\infty} \Gamma^{t-j} \left( \sum_{k=1}^{j} \gamma^{j-k} V_k \right) \le M\bar{L} \left( \frac{2C_1^2\alpha^2\sigma^2}{s(1-\lambda)^2(1-\gamma)} + \frac{4C_2^2 q_1 \lambda^{(1-\delta)(t-1)}}{\delta(1-\lambda)(1-\gamma)} \right) :$$

**Proof.** By the fact that $\sum_{j=1}^{t} \sum_{k=1}^{j} \lambda^{t-j}\gamma^{j-k} \le (1-\lambda)^{-1}(1-\gamma)^{-1}$ and Lemma 19, we have

$$\begin{aligned}
\sum_{j=1}^{\infty} \Gamma^{t-j} \left( \sum_{k=1}^{j} \gamma^{j-k} V_k \right) &\le \sum_{j=1}^{\infty} M\lambda^{t-j} \sum_{k=1}^{\infty} \gamma^{j-k} \bar{L} \|e_k\|^2 \\
&\le M\bar{L} \sum_{j=1}^{t} \lambda^{t-j} \sum_{k=1}^{j} \gamma^{j-k} \left( \frac{2C_1^2}{s(1-\lambda)} \alpha^2\sigma^2 + 2C_2^2 q_1 \lambda^{2(1-\delta)(k-1)} \right) \\
&\le M\bar{L} \left( \frac{2C_1^2\alpha^2\sigma^2}{s(1-\lambda)^2(1-\gamma)} + 2C_2^2 q_1 \sum_{k=1}^{t} \sum_{j=k}^{t} \lambda^{t-j}\gamma^{j-k}\lambda^{(1-\delta)(k-1)} \right) \\
&\le M\bar{L} \left( \frac{2C_1^2\alpha^2\sigma^2}{s(1-\lambda)^2(1-\gamma)} + 2C_2^2 q_1 \frac{2\lambda^{(1-\delta)(t-1)}}{\delta(1-\gamma)(1-\lambda)} \right) :
\end{aligned}$$

**Proof.** [Theorem 7] Define the events

$$E_j = \left\{ \left( \sum_{k=1}^{j} \gamma^{j-k} e_k \right) \le \frac{C_1}{\sqrt{s(1-\lambda)}} \alpha\sigma + C_2 q_1^{1/2} \lambda^{(1-\delta)(j-1)} \right\}; \quad 1 \le j \le t:$$

Fork2