Journal of Machine Learning Research 23 (2022) 1-22 Submitted 5/21; Published 1/22

Beyond Sub-Gaussian Noises: Sharp Concentration Analysis for Stochastic Gradient Descent

Zhipeng Lou ZHIPENGPROB@GMAIL.COM
Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08544, USA

Wanrong Zhu WANRONGZHU@UCHICAGO.EDU
Wei Biao Wu WBWU@GALTON.UCHICAGO.EDU
Department of Statistics
University of Chicago
Chicago, IL 60637, USA

Editor: Prateek Jain

Abstract

In this paper, we study the concentration property of stochastic gradient descent (SGD) solutions. In existing concentration analyses, researchers impose restrictive requirements on the gradient noise, such as boundedness or sub-Gaussianity. We consider a much richer class of noise where only finitely-many moments are required, thus allowing heavy-tailed noises. In particular, we obtain Nagaev type high-probability upper bounds for the estimation errors of averaged stochastic gradient descent (ASGD) in a linear model. Specifically, we prove that, after T steps of SGD, the ASGD estimate achieves an $O(\sqrt{\log(1/\delta)/T} + (\delta T^{q-1})^{-1/q})$ error rate with probability at least $1-\delta$, where $q > 2$ controls the tail of the gradient noise. In comparison, one has the $O(\sqrt{\log(1/\delta)/T})$ error rate for sub-Gaussian noises. We also show that the Nagaev type upper bound is almost tight through an example, where the exact asymptotic form of the tail probability can be derived. Our concentration analysis indicates that, in the case of heavy-tailed noises, the polynomial dependence on the failure probability δ is generally unavoidable for the error rate of SGD.

Keywords: Stochastic gradient descent, high probability analysis, heavy-tailed noise, Nagaev inequality.

1. Introduction

Algorithms based on stochastic approximation (SA), especially the stochastic gradient descent (SGD) and its variants, are workhorses of modern statistical and machine learning (Robbins and Monro, 1951; Lai, 2003; Bottou et al., 2018). For a convex optimization problem $\min_{\theta\in\mathbb{R}^p} F(\theta)$, where $F : \mathbb{R}^d \to \mathbb{R}$, SGD updates the estimate of the minimum $\theta^\star$ based on the stochastic gradient $\widehat{g}(\theta)$ at some θ, which is a noisy measurement of the gradient/subgradient $g(\theta) = \nabla F(\theta)$. The algorithm is easy to implement and popular in applications for its effectiveness, computational efficiency, and versatility.

Given the huge success in applications, it is important to understand the theoretical properties of SA. There have been extensive studies on the theoretical properties since 1951, from early work on consistency to distributions/inference and from asymptotic to non-

asymptotic investigations (Blum, 1954; Dvoretzky, 1956; Moulines and Bach, 2011; Rakhlin et al., 2012; Bach and Moulines, 2013; Toulis and Airoldi, 2017; Anastasiou et al., 2019; Chen et al., 2020; Zhu et al., 2021). However, there are still gaps between the theory of SA and applications, especially with heavy-tailed stochastic gradient noises which commonly arise in practice. This paper focuses on the concentration property of the SGD estimates. We obtain a nearly sharp high-probability error bound for SGD estimates with heavy-tailed noises in the linear model. We show that the tail behaviors of SGD estimates are quite different in heavy-tailed noise cases and sub-Gaussian noise cases.

Most of the literature on the quality of SGD estimates focuses on the *expected* error rate. Polyak and Juditsky (1992) and Ruppert (1988) introduced the averaged SGD (ASGD), a simple modification where iterates are averaged, and established the asymptotic normality of the obtained estimate. It is known that ASGD estimates achieve the optimal rate $\mathcal{O}_P(1/\sqrt{T})$ due to the central limit theorem (CLT), after T steps of SGD, under certain regularity conditions. Further analyses on the error rate show that the *expected* squared error of the SGD estimate (with average if necessary) is $O(1/T)$ for strongly convex objective functions, and $O(1/\sqrt{T})$ for smooth convex and non-smooth Lipschitz objective functions (Nemirovski et al., 2009; Rakhlin et al., 2012; Shamir and Zhang, 2013; Lacoste-Julien et al., 2012).

Besides the guarantees in expectation, practitioners usually want to ensure that the output of a single trial of the algorithm is well behaved and may ask: how many iterations are needed in a single trial of the algorithm to achieve the desired accuracy? In other words, they would prefer high confidence guarantees, i.e., high-probability error bounds in the form of

$$\mathbb{P}(\|\widehat{\theta} - \theta^\star\|_2^2 \geq \epsilon) \leq \delta,$$

where $\epsilon > 0$, $\delta \in (0, 1)$ can be arbitrarily small, and $\widehat{\theta}$ is the estimate of $\theta^\star$. These high-probability guarantees are usually adopted in statistical learning theory (Valiant, 1984), where a tight sample complexity bound is of great interest. Note that bounds in expectation are generally too conservative to derive high-probability guarantees. Specifically, if one has $\mathbb{E}\|\widehat{\theta}_T - \theta^\star\|_2^q = O(T^{-2/q})$ (Chung, 1954), by Markov's inequality, one can only guarantee with probability at least $1 - \delta$,

$$\|\widehat{\theta}_T - \theta^\star\|_2^2 \leq O(\delta^{-2/q} T^{-1}).$$

Then, the resulting sample complexity

$$T(\epsilon, \delta) = O\left(\frac{\delta^{-2/q}}{\epsilon}\right) \tag{1}$$

can be very high for a small δ. Also, the confidence intervals obtained from the CLT only hold asymptotically when the number of samples goes to infinity and cannot be used to rigorously compute sample complexity when $\delta \to 0$. Thus, additional tail probability results (non-asymptotic) are needed.

High-probability bounds on SGD are much less explored than the bounds in expectation. Some known high-probability results under light-tailed noise assumptions include Rakhlin et al. (2012), who showed that for the strongly convex setting and suffix averaging $\widehat{\theta}_T$, with probability at least $1 - \delta$,

$$\|\widehat{\theta}_T - \theta^\star\|_2^2 \leq O\left(\log(\log(T)/\delta)/T\right).$$

Recently, Harvey et al. (2019a) improved the above bound to $O\left(\log(1/\delta)/T\right)$. Other similar results can be found in Hazan and Kale (2014); Cardot et al. (2017); Jain et al. (2019); Harvey et al. (2019b); Feldman and Vondrak (2019); Mou et al. (2020). These high-probability bounds depend logarithmically on $1/\delta$, and the resulting sample complexity is

$$T(\epsilon, \delta) = O\left(\frac{\log(1/\delta)}{\epsilon}\right),$$

substantially improving $T(\epsilon, \delta) = O(\delta^{-2/q}\epsilon^{-1})$ in (1) when δ is small. Such bounds with a dependence on $\log(1/\delta)$ are often called *sub-Gaussian bounds* or with *sub-Gaussian performance*. Harvey et al. (2019a) also remark that a dependence on $\log(1/\delta)$ is necessary, which indicates that SGD can not achieve a better performance than this one under sub-Gaussianity.

The aforementioned high-probability results all rely on the *light-tailed* assumption on the gradient noise $z = \widehat{g} - g$, such as boundedness or sub-Gaussianity. However, such assumptions can be violated in practice. The heavy-tailed phenomenon is not uncommon in applications (Simsekli et al., 2019). It is also more likely to get a bad output in a single trail of SGD due to the more frequent outliers with heavy-tailed stochastic gradients. Thus, a high-probability guarantee is especially needed. Then a natural question is: *Can SGD achieve the sub-Gaussian performance with* $\log(1/\delta)$ *tail behavior in the case of heavy-tailed stochastic noises?* This paper answers this question by delivering a nearly tight high-probability bound in a linear model with heavy-tailed stochastic noises. In particular, with probability at least $1 - \delta$, for any $\delta \in (0, 1)$,

$$\|\bar{\theta}_T - \theta^\star\|_2^2 \leq O\left(\frac{\log(1/\delta)}{T} + \frac{(1/\delta)^{2/q}}{T^{2-2/q}}\right),$$

where $\theta^\star$ is the true parameter and $\bar{\theta}_T$ is the ASGD estimate ($q > 2$ controls the tail of the stochastic noise). As a result, the sample complexity bound, with tolerance error $\epsilon > 0$ and failure probability $\delta \in (0, 1)$, is

$$T(\epsilon, \delta) = O\left(\frac{\log(1/\delta)}{\epsilon} + \left(\frac{\delta^{-2/q}}{\epsilon}\right)^{q/(2(q-1))}\right). \tag{2}$$

It is better than the $T(\epsilon, \delta) = O(\delta^{-2/q}\epsilon^{-1})$ in (1). Besides the advantage of the logarithmical term, the polynomial term $O((\delta^{-2/q}\epsilon^{-1})^{q/(2(q-1))})$ is sharper than $O(\delta^{-2/q}\epsilon^{-1})$ since $q > 2$. We also compare the logarithmical term and the polynomial term in (2) numerically. Figure 1 shows that, when δ is big, the logarithmical dependence dominates, and therefore the sample complexity is the same as that in the sub-Gaussian case. While when δ is small, which is more of interest in most cases, the polynomial dependence term dominates, showing that the polynomial dependence on δ is unavoidable. Thus, one cannot achieve the sub-Gaussian performance when the gradient noise exhibits heavy-tailed distribution.

There recently has been renewed interest in obtaining robust guarantees of SGD without the light-tailed assumption. Robust modifications of SGD (or GD), such as gradient clipping and using the geometric median of stochastic gradients, are studied to accommodate the

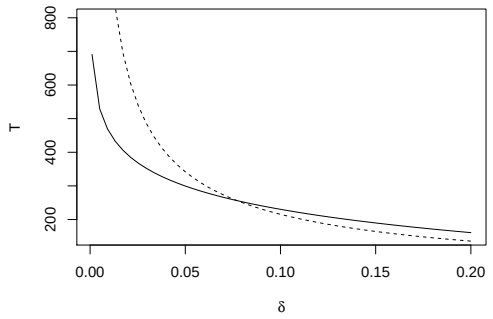

Figure 1: Compare the two terms in sample complexity (2). Here X axis represents failure probability δ; the solid line denotes $\epsilon^{-1}\log(1/\delta)$, the dashed line denotes $(\delta^{-2/q}\epsilon^{-1})^{q/(2(q-1))}$. We choose $\epsilon = 0.01$ and $q = 2.5$.

heavy-tailed noises (Nazin et al., 2019; Holland and Ikeda, 2019; Davis and Drusvyatskiy, 2020; Gorbunov et al., 2020). Our lower bound answers the question "whether that robust modification of SGD is necessary". This question is vital because using SGD is a very common heuristic in modern learning tasks, and it is easier to implement and more widely used than its modified versions.

Contribution. Our primary theoretical contribution is to develop sharp probability bounds for SGD with heavy-tailed noises for linear models. To this end, we introduce the Nagaev type inequality to the machine learning community where traditionally researchers generally use exponential inequalities for sub-exponential or sub-Gaussian random variables and the Markov inequality if only polynomial moments exist. In general, Nagaev type inequalities are much sharper than Markov inequalities; however, they are not very well-known in the machine learning community. The linear model analysis can provide useful insights into general models in view of the connection between linear models and more complex models such as neural networks (Chizat et al., 2019; Hastie et al., 2019). Also, the idea of decomposing the martingale differences and taking advantage of specific martingale structures in SGD may be useful for analyzing the theoretical properties of SA, especially for studying tail probabilities.

Organization. The remainder of this article is organized as follows. In Section 2, we introduce the linear model setting and specify assumptions on the noise. In Section 3, we present our main result, i.e., the Nagaev type high probability bound. We also provide technical innovations in this section. Then in Section 4, we show that the obtained high-probability bound is nearly sharp, and the polynomial dependence on δ cannot be avoided. Section 5 provides proofs for main theorems. Discussions and future directions are contained in Section 6.

2. Preliminaries

In this section, we first introduce some notations. Then, we describe the linear model setting and assumptions which will be used later.

2.1 Notations

For a vector $v = (v_1, \ldots, v_p)^\top$, we denote its Euclidean norm as $\|v\|_2 = (v_1^2 + \ldots + v_p^2)^{1/2}$. The operator norm of a matrix A is defined as $\|A\|_2 = \max_{\|\nu\|_2=1} \|A\nu\|_2$. When A is positive semi-definite, $\lambda_{\max}(A)$ denotes the largest eigenvalue of A, $\lambda_{\min}(A)$ denotes the smallest eigenvalue of A, and $\mathrm{tr}(A)$ denotes its trace. We use I_p to denote a $p \times p$ identity matrix. We use $\mathbb{S}^{p-1}$ to denote the p-dimensional unit sphere. For two positive sequences $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, we write $a_n \lesssim b_n$ if $a_n \leq Cb_n$ for some constant $C > 0$ that does not depend on n. Moreover, we write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Throughout the paper, we use C to denote generic constant whose value may change from line to line. For a sequence of *i.i.d.* sample $\{\xi_i\}_{i\geq 1}$ from some distribution Π, we define conditional expectation $\mathbb{E}_n(\cdot) = \mathbb{E}(\cdot|\mathcal{F}_n)$, where $\mathcal{F}_n$ is σ-algebra generated by $\{\xi_i\}_{i\leq n}$.

2.2 Linear model setting

Assume that we observe data $(X_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i \geq 1$, from the following linear regression model:

$$y_i = X_i^\top \theta^\star + \epsilon_i, \quad i \geq 1,$$

where $\theta^\star \in \mathbb{R}^p$ is the unknown true parameter. The random draws (X_i, ϵ_i) across $i = 1, 2, ...$ are *i.i.d.* from $P_X \times P_\epsilon$. Here we assume that P_X is a distribution on $\mathbb{R}^p$ such that $\mathbb{E}(X_i X_i^T) = \Sigma$, while P_ϵ is a distribution on $\mathbb{R}$ such that $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$. Note that, we consider a much richer class of gradient noise beyond sub-Gaussian, where only finitely-many moments are required allowing heavy-tailed noise. More detailed assumptions are included in Section 2.3.

To solve the above linear regression problem, we consider the optimization problem

$$\min_{\theta\in\mathbb{R}^p} F(\theta) = \mathbb{E}_{X,y}\frac{1}{2}(y - X^\top\theta)^2.$$

We apply the mini-batch SGD, which is a popular parallelization technique reducing the communication costs (Li et al., 2014; Reddi et al., 2016; Jain et al., 2017). Mini-batching is efficient in practice and brings convenience in later proofs. Initialized at θ_0, the t-th iteration with step size η_t is given by:

$$\begin{aligned} &\theta_t = \theta_{t-1} - \eta_t \widehat{g}_t(\theta_{t-1}), \quad t \geq 1, \\ &\widehat{g}_t(\theta_{t-1}) = \frac{1}{B}\sum_{i=(t-1)B+1}^{tB} X_i\left(X_i^\top\theta_{t-1} - y_i\right), \end{aligned} \tag{3}$$

where B is the mini-batch size and step sizes η_t will be discussed in later analysis. In this paper, we are interested in the the high-probability bound of the averaged iterate $\bar{\theta}_T = T^{-1}\sum_{t=1}^T \theta_t$ with T iterations ($n = TB$ samples) in total.

For $(X_i, y_i)_{i\geq 1}$ in above linear regression model, let

$$A_t = \frac{1}{B}\sum_{i=(t-1)B+1}^{tB} X_i X_i^\top \text{ and } b_t = \frac{1}{B}\sum_{i=(t-1)B+1}^{tB} X_i y_i.$$

We can rewrite the t-th iteration from the mini-batch SGD as:

$$\theta_t = \theta_{t-1} - \eta_t(A_t\theta_{t-1} - b_t), \ \ t \geq 1.$$

Note that $\mathbb{E}(A_t) = \mathbb{E}(XX^\top) = \Sigma$, and $b = \mathbb{E}(b_t) = \Sigma\theta^\star$. We can see that solving the linear regression problem through mini-batch SGD (3) is equivalent to solving the linear system of the form:

$$\Sigma\theta^\star = b,$$

through stochastic approximation (Mou et al., 2020).

2.3 Assumptions

Assumption 1 *For distribution P_ϵ, assume that for some constant $q > 2$,*

$$\mu_q = \mathbb{E}|\epsilon|^q < \infty.$$

Assumption 2 *Assume that $M_\psi := \max_{1\leq \ell\leq p}\left(\mathbb{E}|X_{i\ell}|^{2\psi}\right)^{1/2\psi} < \infty$, for some constant $\psi > \max\{4, q\}$. Let $\lambda_{\min}(\Sigma) > 0$, assume that the mini-batch size satisfy*

$$B \geq 16(\psi - 1)M_\psi^4 p^2/\lambda_{\min}(\Sigma)^2.$$

Remark 1 *In existing works, light-tailed assumptions of the gradient noises are required, i.e., finite exponential moments (e.g. bounded, sub-Gaussian, sub-exponential). While in our assumptions, the noise conditions are more general. We only require finite polynomial moments in Assumption 1, in which case heavy-tailed noises are allowed. Assumption 2 is a fairly mild condition on P_X and the mini-batch size B. It ensures that*

$$\left(\mathbb{E}\|A_t - \Sigma\|_2^\psi\right)^{1/\psi} \leq \lambda_{\min}(\Sigma)/2, \tag{4}$$

which is shown in Lemma 12 and is a useful condition for controlling the correlation between SGD iterates in later proofs. On the other hand, if X_i is Gaussian, Corollary 2 in Koltchinskii and Lounici (2017) implies that a weaker assumption for (4) is, for come constant C_ψ,

$$B \geq C_\psi r(\Sigma)\mathcal{K}(\Sigma)^2,$$

where $r(\Sigma) = \mathrm{tr}(\Sigma)/\lambda_{\max}(\Sigma)$ is the effective rank of Σ, and $\mathcal{K}(\Sigma) = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$ is the condition number. It is worth mentioning that the linear system $\Sigma\theta^\star = b$ becomes more unstable when the condition number of Σ grows. Therefore, it is reasonable to require a larger mini-batch size B when the condition number is larger. For more discussion about concentration inequality and expectation inequality of the operator norm $\|A_t - \Sigma\|_2$, we refer to Vershynin (2010); Koltchinskii and Lounici (2017); Tropp (2012) and the references therein.

3. Main Results

3.1 Nagaev type upper bound

The step size sequence $(\eta_t)_{t\geq 1}$ controls the convergence of the SGD algorithm. In this section, we focus on two commonly used step size regimes: polynomial decay step size $\eta_t = \eta_0 t^{-\alpha}$ with $\alpha \in (0,1)$ and constant step size with $\eta_t = \eta_0$ for any $t \geq 1$.

We analyze the tail probability of the error $\bar{\theta}_T - \theta^\star$, after T steps of SGD, in the linear model setting. In what follows, we denote $\lambda_0 = \lambda_{\min}(\Sigma)/2$, $\lambda^* = \lambda_{\max}(\Sigma)$,

$$\mathcal{K}_q = \sup_{\nu \in \mathbb{S}^{p-1}} \mathbb{E}|\nu^\top X_t|^q,$$

and

$$\Upsilon_{\varpi,\alpha} = \int_1^\infty \exp\left(-\varpi \int_1^z x^{-\alpha} dx\right) dz.$$

Theorem 2 (polynomial decay step size) *Let Assumptions 1 and 2 hold. Assume that* $\eta_0 \leq 1/\lambda^*$ *and*

$$\psi > \frac{2q - 4\alpha}{2 - \alpha}.$$

Then, for any $\omega \in \mathbb{S}^{p-1}$ *and* $x > 0$*, we have*

$$\mathbb{P}\left(|\omega^\top(\bar{\theta}_T - \theta^\star)| > x\right) \leq \frac{C_0 \|\theta_0 - \theta^\star\|_2^\psi}{(Tx)^\psi} + \frac{C_1 W_q}{T^{q-1}x^q} + C\exp\left(-\frac{C_2 T x^2}{W_2}\right), \tag{5}$$

where $W_q = \mu_q \mathcal{K}_q \lambda_0^{-q} B^{1-q}$, $W_2 = \sigma^2 \lambda^* \lambda_0^{-2} B^{-1}$, $C_0 = (2\Upsilon_{\lambda_0\eta_0,\alpha})^\psi$, C, C_1 *and* C_2 *are constants depending only on* q, ψ *and* α.

Remark 3 *The first term on the RHS of* (5) *characterizes the effect of the initial point* θ_0 *on the tail probability of* $\bar{\theta}_T - \theta^\star$*. The influence of* θ_0 *decays quickly, note that for any* $x \gtrsim T^{-1/2}$*, we have*

$$\frac{C_0 \|\theta_0 - \theta^\star\|_2^\psi}{(Tx)^\psi} \leq \frac{C_1 W_q}{T^{q-1}x^q},$$

as long as $\|\theta_0 - \theta^\star\|_2^\psi \lesssim C_1 C_0^{-1} W_q T^{1+(\psi-q)/2}$*, which is a fairly mild condition on* θ_0 *as* $\psi > q$*. Consequently, in this case, Theorem 2 implies that*

$$\mathbb{P}\left(|\omega^\top(\bar{\theta}_T - \theta^\star)| > x\right) \leq \frac{2C_1 W_q}{T^{q-1}x^q} + C\exp\left(-\frac{C_2 T x^2}{W_2}\right),$$

which, together with $\mathbb{P}(\|\bar{\theta}_T - \theta^\star\|_2 > x) \leq \sum_{j=1}^p \mathbb{P}(|\bar{\theta}_{T,j} - \theta_j^\star| > x/\sqrt{p})$*, imply that*

$$\mathbb{P}\left(\|\bar{\theta}_T - \theta^\star\|_2 > x\right) \leq \frac{2p^{1+q/2} C_1 W_q}{T^{q-1}x^q} + pC\exp\left(-\frac{C_2 T x^2}{pW_2}\right). \tag{6}$$

Theorem 4 (Constant step size) *Let Assumptions 1 and 2 hold. Assume that* $\eta_0 \leq 1/\lambda^*$ *and*

$$\eta_0 \gtrsim \frac{(\log T)^{(3\psi-4)/(\psi-4)}}{T^{2(\psi-q)/(\psi-4)}}.$$

Then, for any vector $\omega \in \mathbb{S}^{p-1}$ *and* $x > 0$*, we have*

$$\mathbb{P}\left(|\omega^\top(\bar{\theta}_T - \theta^\star)| > x\right) \leq \frac{C_0\|\theta_0 - \theta^\star\|_2^\psi}{(Tx)^\psi} + \frac{C_1 W_q}{T^{q-1}x^q} + C\exp\left(-\frac{C_2 T x^2}{W_2}\right),$$

where $W_q = \mu_q \mathcal{K}_q \lambda_0^{-q} B^{1-q}$ *,* $W_2 = \sigma^2\lambda^*\lambda_0^{-2}B^{-1}$*,*$C_0 = (2/\lambda_0\eta_0)^\psi$*,* C*,* C_1 *and* C_2 *are constants depending only on* q *and* ψ*.*

Remark 5 *Both inequalities with different step size regimes imply two types of bounds: Gaussian type tail and polynomial type tail. When* x *is small, i.e., for small deviations, the Gaussian type tail is the dominating term for the tail of the estimation error. While for large* x*, the polynomial type tail dominates. A combination of these two types of tail approximation calibrates the tail behavior of SGD solutions more accurately in the case of heavy-tailed noises.*

After elementary calculations, we can translate the tail probability results as following. For any $\delta \in (0,1)$*, with probability at least* $1-\delta$*,*

$$\|\bar{\theta}_T - \theta^\star\|_2 \leq O\left(\sqrt{\frac{\log(1/\delta)}{T}} + \frac{(1/\delta)^{1/q}}{T^{1-1/q}}\right).$$

For large or moderate failure probability $\delta > \delta^*$*,*

$$\|\bar{\theta}_T - \theta^\star\|_2 \leq O\left(\sqrt{\frac{\log(1/\delta)}{T}}\right),$$

where δ^* *is the solution of the equation* $\sqrt{T^{-1}\log(1/\delta)} = T^{-1+1/q}(1/\delta)^{1/q}$ *and has the asymptotic form* $\delta^* \asymp T^{1-q/2}(\log T)^{-q/2}$*. This high probability error rate matches existing results considering sub-Gaussian/bounded gradient noises (Mou et al., 2020). While for small failure probability* $\delta < \delta^*$*, which is more of interest in most applications, we have*

$$\|\bar{\theta}_T - \theta^\star\|_2 \leq O\left(\frac{(1/\delta)^{1/q}}{T^{1-1/q}}\right).$$

3.2 Technical overview and proof sketch for main results

Let $\Delta_t = \theta_t - \theta^\star$ and $\mathcal{E}_t = B^{-1}\sum_{i=(t-1)B+1}^{tB} X_i\epsilon_i$. At the t-th step, the gradient g and the stochastic gradient $\widehat{g}_t$ can be written with $\Sigma, A_t, \mathcal{E}_t, \Delta_t$ notations as follows:

$$g(\theta_{t-1}) = \Sigma\Delta_{t-1}, \quad \widehat{g}_t(\theta_{t-1}) = A_t\Delta_{t-1} - \mathcal{E}_t.$$

Let $z_t(\theta_{t-1}) = \widehat{g}_t(\theta_{t-1}) - g(\theta_{t-1})$ denote the gradient noise. Note that it is a martingale difference sequence since $\mathbb{E}_{t-1}(z_t(\theta_{t-1})) = 0$. The recursion of Δ_t is usually represented using martingales as follows

$$\Delta_t = (I_p - \eta_t \Sigma)\Delta_{t-1} - \eta_t z_t(\theta_{t-1}), \ \ t \geq 1. \tag{7}$$

Then, the classic analysis uses properties of martingales, such as Freedman and Azuma inequalities. The high-probability bounds obtained from those general martingale inequalities are sharp only when finite exponential moments of the noise $z_t|\mathcal{F}_{t-1}$ exists. Therefore, existing studies require the gradient noise z_t (or equivalently A_t and b_t in linear stochastic approximation) to be sub-Gaussian or to be bounded. In our work, we extend the noise condition to a more general case, where heavy-tailed noises are allowed. To obtain sharp high-probability bounds for heavy-tailed noises, we study the detailed structure of the martingale differences and use inequalities which are nearly sharp under polynomial moment conditions.

We can see that the martingale difference z_t at θ_{t-1} can be decomposed as

$$z_t(\theta_{t-1}) = (A_t - \Sigma)\Delta_{t-1} - \mathcal{E}_t, \ \ t \geq 1,$$

which is the sum of two parts, one related to the noise from A_t and the other part $\mathcal{E}_t$. Note that the dependence between $\{z_t(\theta_{t-1})\}_{t\geq 1}$ comes from the dependence between $(\theta_t)_{t\geq 1}$, and $(\mathcal{E}_t)_{t\geq 1}$ are independent. Then leveraging the structure of z_t, we study the recursion with a different representation:

$$\Delta_t = (I_p - \eta_t A_t)\Delta_{t-1} + \eta_t \mathcal{E}_t, \ \ t \geq 1. \tag{8}$$

Compared with the form in (7), although more considerations are needed for the correlation term $(I_p - \eta_t A_t)$ as variability is introduced (we now have $(I_p - \eta_t A_t)$ instead of $(I_p - \eta_t \Sigma)$), the remaining independent structure makes it possible to obtain a tight tail bound under heavy-tailed noise assumptions.

In the following, we sketch the proof of our main results under the step size regime $\eta_t = \eta_0 t^{-\alpha}$ with $\alpha \in (0, 1)$. Proof for the constant step size regime shares the same spirit with minor modifications. We defer the complete proof to Section 5. From (8) we can see that $(\Delta_t)_{t\geq 1}$ has a closed form expression

$$\Delta_t = \prod_{\ell=1}^{t}(I_p - \eta_\ell A_\ell)\Delta_0 + \sum_{m=1}^{t}\prod_{\ell=m+1}^{t}(I_p - \eta_\ell A_\ell)\eta_m \mathcal{E}_m.$$

Let $S_T = T(\bar{\theta}_T - \theta) = \sum_{t=1}^{T}\Delta_t$ which is further decomposed as $S_T = S_T^{\diamond} + S_T^{\star}$, where

$$S_T^{\diamond} = \sum_{t=1}^{T}\prod_{\ell=1}^{t}(I_p - \eta_\ell A_\ell)\Delta_0 \ \text{ and } \ S_T^{\star} = \sum_{t=1}^{T}\sum_{m=1}^{t}\prod_{\ell=m+1}^{t}(I_p - \eta_\ell A_\ell)\eta_m \mathcal{E}_m.$$

To bound the target $\omega^\top S_T/T$ for any $\omega \in \mathbb{S}^{p-1}$ in Section 3.1, we deal with $S_T^{\diamond}$ and $S_T^{\star}$ separately.

Lemma 6 *Under Assumption 2, for any vector $\omega \in \mathbb{S}^{p-1}$ and $x > 0$, we have*

$$\mathbb{P}\left(|\omega^\top S_T^\diamond| > x\right) \leq \frac{\|\theta_0 - \theta^\star\|_2^\psi \Upsilon_{\lambda_0\eta_0,\alpha}^\psi}{x^\psi}.$$

Next, we observe that for any $\omega \in \mathbb{S}^{p-1}$,

$$\omega^\top S_T^\star = \frac{1}{B}\sum_{m=1}^{T}\sum_{i=(m-1)B+1}^{mB} \eta_m \omega^\top H_m X_i \epsilon_i, \text{ where } H_m = \sum_{t=m}^{T}\prod_{\ell=m+1}^{t}(I_p - \eta_\ell A_\ell),$$

which means $\omega^\top S_T^\star$ is a sum of independent zero-mean random variables conditional on $\mathcal{F}_{X,n} = \sigma\{X_1, X_2, \ldots, X_n\}$. Hence, for $x > 0$, by Lemma 11 (Nagaev inequality),

$$\mathbb{P}\left(|\omega^\top S_T^\star| > x | \mathcal{F}_{X,n}\right) \leq \frac{C_{q,1} D_{T,q}}{(Bx)^q} + 2\exp\left(-\frac{C_{q,2} B^2 x^2}{D_{n,2}}\right),$$

where $C_{q,1}$ and $C_{q,2}$ are constants depending only on q and

$$D_{T,q} = \mu_q \sum_{m=1}^{T} \eta_m^q \sum_{i=(m-1)B+1}^{mB} |\omega^\top H_m X_i|^q.$$

We bound the conditional variance $D_{T,2}$ in Lemma 13, which is a main technical step, and obtain the following results for $S_T^\star$.

Lemma 7 *Under the conditions of Theorem 2, we have*

$$\mathbb{P}\left(|\omega^\top S_T^\star| > x\right) \leq \frac{C_1 W_q T}{x^q} + C\exp\left(-\frac{C_2 x^2}{T W_2}\right),$$

Consequently, Theorem 2 directly follows from Lemma 6 and Lemma 7.

4. Tightness of the Upper Bound

This section shows that the Nagaev type upper bound obtained in the above section is tight through the example of the mean estimation model. Therefore, the polynomial term in the upper bounds in Section 3.1 is unavoidable, and the sub-Gaussian performance with $\log(1/\delta)$ tail behavior cannot be achieved through SGD with heavy-tailed gradient noise. In particular, we consider the model

$$y_i = \theta^\star + \epsilon_i, \;\; i \geq 1, \tag{9}$$

where $\theta^\star \in R$ is the mean we want to estimate and $\{\epsilon_i\}_{i\geq 1}$ are i.i.d. generated from a t-distribution with degree of freedom $\nu > 2$. For initial value θ_0, the t-th iterate θ_t from SGD algorithm, with mini-batch size $B = 1$, takes the following form:

$$\theta_t = \theta_{t-1} + \eta_t(y_t - \theta_{t-1}), \;\; t \geq 1, \tag{10}$$

where η_t is the step size at the t-th iteration.

The gradient noise $z_t = \epsilon_t$ is *heavy-tailed*. The mean estimation model (9) is a special case of the linear regression model in Section 2. Assumptions 1 and 2 can be easily verified since $A_t = 1$ with no randomness here. Then we can apply theorems in Section 3 and get the upper bounds for the estimation error $\bar{\theta}_T - \theta^\star$ when there are T iterations in total. We focus on the polynomial decay step size regime, i.e., $\eta_t = \eta_0 t^{-\alpha}$, with $\eta_0 = 0.1, \alpha = 0.55$ in the rest of this section. We modify the upper bound (6) in Section 3.1 as follows.

Nagaev type upper bound: We have for all $x > 0$,

$$\mathbb{P}\left(|\bar{\theta}_T - \theta^\star| > x\right) \leq \frac{C_1}{x^q T^{q-1}} + \exp\left(-C_2 T x^2\right), \tag{11}$$

for some constant C_1, C_2. Then, with probability at least $1 - \delta$,

$$|\bar{\theta}_T - \theta^\star| \leq O\left(\frac{1}{(\delta T^{q-1})^{1/q}} + \sqrt{\frac{\log(1/\delta)}{T}}\right).$$

Next, we will show that the Nagaev type upper bound for the estimation error $\bar{\theta}_T - \theta^\star$ is tight by taking advantage of the simple structure of the mean estimation model. First, we introduce the following notation

$$\begin{aligned} V_i &= \prod_{k=1}^{i}(1-\eta_k),\ i \geq 1,\ V_0 = 1; \\ V_i^j &= \frac{V_j}{V_i},\ j \geq i. \end{aligned} \tag{12}$$

Then, $\bar{\theta}_T - \theta^\star$ has the closed form as follows:

$$\bar{\theta}_T - \theta^\star = \frac{1}{T}\sum_{i=1}^{T} V_i \Delta_0 + \frac{1}{T}\sum_{t=1}^{T}\sum_{i=t}^{T} V_t^i \eta_t \epsilon_t,$$

where Δ_0 is the initialization error $\theta_0 - \theta^\star$. Since $\{\epsilon_t\}_{t\geq 1}$ is a sequence of i.i.d. random errors, the estimation error above (deducted by the initialization error) can be view as the weighted sum of T i.i.d. random variables with mean 0. We can then further analyze the estimation error based on existing studies about deviations and tail probabilities of linear processes.

4.1 Upper bound from Nagaev inequality

The Nagaev inequality (Nagaev, 1979) for tail probability is a useful result in probability theory. It is known that the performance bounds obtained from Nagaev inequality are nearly sharp under polynomial moment conditions.

Proposition 8 *Consider the mean estimation model in* (9) *and the SGD iterates* $\{\theta_t\}_{t=1,\ldots,T}$ *defined in* (10). *For any* $x > 0$ *and* $2 < q < \nu$, *we have*

$$\mathbb{P}\left(\left|\bar{\theta}_T - \theta^\star\right| \geq \frac{C|\Delta_0|}{T} + x\right) \leq \frac{(1+2/q)^q \mathbb{E}|\epsilon|^q}{x^q T^{q-1}} + 2\exp\left(-c_q x^2 T\right), \tag{13}$$

where $c_q = 2e^{-q}(q+1)^{-2}/\mathbb{E}|\epsilon|^2$, *and* $C = \sum_{i=1}^{\infty} V_i$, V_i *is defined in* (12).

While the Nagaev inequality gives more precise constants, the upper bound in (13) is of the same order as that in (11). Thus, the tightness of Nagaev inequality implies that our proposed Nagaev type upper bound is also tight.

4.2 Exact deviation

Furthermore, instead of an upper bound, we give the exact asymptotic tail probability of the estimation error in the mean estimation model. Inspired by Peligrad et al. (2014), which studied the exact moderate and large deviation of linear processes, we obtain Proposition 9.

Proposition 9 *Consider the mean estimation model* (9) *and the SGD iterates* $\{\theta_t\}_{t=1,\dots,T}$ *defined in* (10). *Define*

$$\sigma_T^2 = \mathbb{E}|\epsilon|^2 \sum_{t=1}^{T} \left(\sum_{i=t}^{T} V_t^i \eta_t / T \right)^2 .$$

For $x \geq \sigma_T$,

$$\mathbb{P}\left(\left| \bar{\theta}_T - \theta^\star - \frac{\gamma_T \Delta_0}{T} \right| \geq x \right) = (2 + o(1)) \left(1 - \Phi(x/\sigma_T) + R(T, x) \right), \tag{14}$$

where $\gamma_T = \sum_{i=1}^{T} V_i, V_i$ *is defined in* (12), *and*

$$R(T, x) = \sum_{t=1}^{T} \mathbb{P}\left(\epsilon_t \geq Tx / \sum_{i=t}^{T} V_t^i \eta_t \right).$$

The right-hand-side (RHS) of (14) comprises two parts: Gaussian approximation $1 - \Phi(x/\sigma_T)$ and tail approximation $R(T, x)$. Note that $\sigma_T^2 \asymp 1/T$ as discussed in Section 5.6 . Then the Gaussian approximation refines the term $\exp(-C_2 T x^2)$ in (11). Also,

$$\mathbb{P}\left(\epsilon_t \geq y\right) \sim c_\nu / y^\nu, y \to \infty,$$

where $c_\nu = \nu^{-3/2} \pi^{-1/2} \Gamma((\nu+1)/2)/\Gamma(\nu/2)$ according to the property of t_ν distribution, and $\sum_{i=t}^{\infty} V_t^i \eta_t = O(1)$ as discussed in Section 5.6. Then the tail approximation

$$R(T, x) \asymp 1/(x^\nu T^{\nu-1}),$$

matching the polynomial term in our proposed Nagaev type upper bound (11). Therefore, we can see that the tail probability polynomial dependence on $1/\delta$ is necessary in the tail bound of SGD and sub-Gaussian tails cannot be achieved under heavy-tailed assumptions.

4.3 A numerical study

We conduct a numerical study of the accuracy of the exact tail probability in (14) for $\nu = 3$. The true tail probability of the estimation error (LHS of (14)) can be calculated through the inversion formula. Let

$$S_T = \bar{\theta}_T - \theta^* - \frac{1}{T} \sum_{i=1}^{T} V_i \Delta_0 = \sum_{t=1}^{T} \left(\sum_{i=t}^{T} V_t^i \eta_t / T \right) \epsilon_t.$$

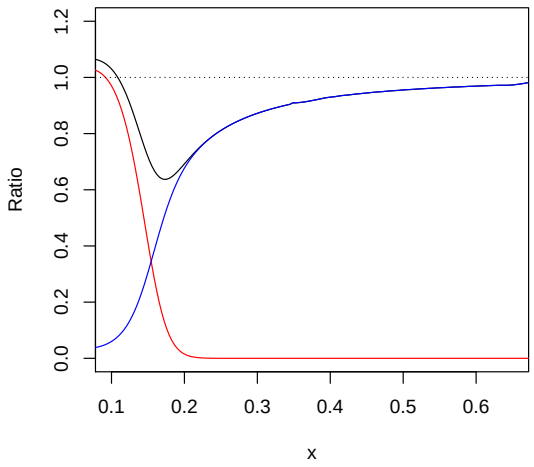

Figure 2: Ratio of approximated and true tail probability. Here X axis represents deviation x. Red curves represent Gaussian approximation: $\left(1-\Phi(x/\sqrt{\mu_{T,2}})\right)/\mathbb{P}(S_T \geq x)$; blue curves represent tail approximation: $R(T,x)/\mathbb{P}(S_T \geq x)$; black curves represent their sum: $\left(1-\Phi(x/\sqrt{\mu_{T,2}})+R(T,x)\right)/\mathbb{P}(S_T \geq x)$.

Then the characteristic function of S_T is

$$\phi_{S_T}(x) = \prod_{t=1}^{T} \phi\left(\left(\sum_{i=t}^{T} V_t^i \eta_t / T\right) x\right),$$

where ϕ is the characteristic function of a t_3-distribution. By the inversion formula,

$$\mathbb{P}\left(S_T \leq x\right) - \mathbb{P}\left(S_T \leq 0\right) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{\sqrt{-1}yx}-1}{\sqrt{-1}y}\phi_{S_T}(y)dy.$$

Since S_T is symmetric, $\mathbb{P}\left(S_T \leq 0\right) = \frac{1}{2}$. In our numerical study, we use the above formula to compute the probability $\mathbb{P}(S_T \geq x)$. In figure 2, we report the ratios $\left(1-\Phi(x/\sqrt{\mu_{T,2}})\right)/\mathbb{P}(S_T \geq x)$, $R(T,x)/\mathbb{P}(S_T \geq x)$ and $\left(1-\Phi(x/\sqrt{\mu_{T,2}})+R(T,x)\right)/\mathbb{P}(S_T \geq x)$. We can see that the Gaussian approximation is good for small deviations, while the tail approximation is better when the deviation is moderate or large. The numerical study confirms that the polynomial term in the upper bound (11) is necessary in the case of heavy-tailed gradient noise, especially for moderate and large deviations.

5. Proofs

We first introduce some notations. For a random variable X and $q > 0$, we write $\|X\|_q = (\mathbb{E}|X|^q)^{1/q}$ if $\mathbb{E}|X|^q < \infty$. Moreover, for any random matrix A, we write $\|A\|_q = (\mathbb{E}\|A\|_2^q)^{1/q}$

by convention. From this point on, abusing notation, depending on context we may write $\|\cdot\|_2$ to denote the matrix norm introduced in Section 2.1, or may also write $\|\cdot\|_2$ to denote the random matrix norm discussed here.

5.1 Some useful lemmas

In Lemma 10, the case $1 < q \leq 2$ follows from Burkholder (1988) and the other case $q > 2$ is due to Rio (2009). Lemma 11 follows from Corollary 1.8 of Nagaev (1979).

Lemma 10 (Burkholder) *Let $q > 1$ and $q' = \min\{q, 2\}$. Let $(D_t)_{t\in\mathbb{Z}}$ be martingale differences with $\mathbb{E}|D_t|^q < \infty$ for every $t \in \mathbb{Z}$. Write $M_n = \sum_{t=1}^n D_t$. Then*

$$\|M_n\|_q^{q'} \leq C_q^{q'} \sum_{t=1}^n \|D_t\|_q^{q'}, \text{ where } C_q = \begin{cases} (q-1)^{-1}, & 1 < q \leq 2, \\ \sqrt{q-1}, & q > 2. \end{cases}$$

Lemma 11 (Nagaev) *Let $(e_t)_{t\in\mathbb{Z}}$ be independent zero-mean random variables with $\sup_{t\in\mathbb{Z}} \mathbb{E}|e_t|^q < \infty$ for some $q > 2$. Let $S_n = \sum_{t=1}^n e_t$ and $c_q = 2e^{-q}(q+2)^{-2}$. Then, for $x > 0$, we have*

$$\mathbb{P}(|S_n| \geq x) \leq (1 + 2/q)^q \frac{\sum_{t=1}^n \mathbb{E}|e_i|^q}{x^q} + 2\exp\left(-\frac{c_q x^2}{\sum_{t=1}^n \mathbb{E}|e_i|^2}\right).$$

Lemma 12 (Moment bounds for sample covariance operators) *Under Assumption 2, we have*

$$\left(\mathbb{E}\|A_t - \Sigma\|_2^\psi\right)^{1/\psi} \leq \lambda_0.$$

Proof For simplicity, write

$$\Delta_{B,jk} = \sum_{i=1}^B (X_{ij}X_{ik} - \Sigma_{jk})$$

for $1 \leq j, k \leq p$. By Lemma 10, it follows that

$$\|\Delta_{B,jk}\|_\psi^2 \leq (\psi - 1)\sum_{i=1}^B \|X_{ij}X_{ik} - \Sigma\|_\psi^2 \leq 4B(\psi - 1)M_\psi^4.$$

Consequently, under Assumption 2, we have

$$\left(\mathbb{E}\|A_t - \Sigma\|_2^\psi\right)^{1/\psi} \leq \frac{1}{B}\left(\sum_{j,k=1}^p \|\Delta_{B,jk}\|_\psi^2\right)^{1/2} \leq \frac{2p(\psi-1)^{1/2}M_\psi^2}{B^{1/2}} \leq \lambda_0.$$

■

5.2 Proof of Lemma 6

By Lemma 12, triangle inequality and the fact that $\eta_0 \leq 1/\|\Sigma\|_2$, we have for each $\ell \geq 1$,

$$\|I_p - \eta_\ell A_\ell\|_\psi \leq \|I_p - \eta_\ell \Sigma\|_\psi + \eta_\ell \|A_\ell - \Sigma\|_\psi \leq 1 - 2\lambda_0\eta_\ell + \lambda_0\eta_\ell = 1 - \lambda_0\eta_\ell.$$

Consequently, by the triangle inequality, it follows that

$$\|\omega^\top S_T^\diamond\|_\psi \leq \|\theta_0 - \theta^\star\|_2 \sum_{t=1}^{T} \prod_{\ell=1}^{t} (1 - \lambda_0\eta_\ell) \leq \|\theta_0 - \theta^\star\|_2 \Upsilon_{\lambda_0\eta_0,\alpha}. \tag{15}$$

Then Lemma 6 is obtained through Markov's inequality.

5.3 Proof of Lemma 7

We first introduce the following lemma, providing a concentration inequality for $D_{T,2}$, where

$$D_{T,2} = B\sigma^2 \sum_{m=1}^{T} \eta_m^2 \omega^\top H_m A_m H_m^\top \omega =: B\sigma^2 \sum_{m=1}^{T} \eta_m^2 \xi_m.$$

Lemma 13 (Main Technical Lemma) *Under Assumption 2, for $z > 0$, we have*

$$\mathbb{P}(|D_{T,2} - \mathbb{E}(D_{T,2})| > z) \leq \frac{C_{\psi,\alpha} T L^{\psi/4-1} \|\Sigma\|_2^{\psi/2}}{\lambda_0^\psi (z/B)^{\psi/2}} + C \exp\left\{ -\frac{C'_{\psi,\alpha} (z/B)^2 \lambda_0^4}{T\|\Sigma\|_2^2} \right\},$$

where $C_{\psi,\alpha}$ and $C'_{\psi,\alpha}$ are positive constants depending only on ψ and α, and

$$L \asymp \frac{T^\alpha}{\lambda_0\eta_0} \log\left(\frac{B\|\Sigma\|_2 T^{1+\alpha}}{\lambda_0^2} \right). \tag{16}$$

Proof For any $k \geq 1$, define $\mathcal{F}_{A,k} = \sigma\{A_1, A_2, \ldots, A_k\}$ and the projection operator

$$\mathcal{P}_{A,k}(\cdot) = \mathbb{E}(\cdot|\mathcal{F}_{A,k}) - \mathbb{E}(\cdot|\mathcal{F}_{A,k-1}).$$

Denote $H_m = \mathcal{H}(A_{m+1}, A_{m+2}, \ldots, A_T)$. For any $h \geq 1$, define

$$H_{m,\{m+h\}} = \mathcal{H}(A_{m+1}, A_{m+2}, \ldots, A_{m+h-1}, A^\star_{m+h}, A_{m+h+1}, \ldots, A_T).$$

where $(A_t^\star)_{t\in\mathbb{Z}}$ are i.i.d. random matrix with $A_t^\star \overset{\mathcal{D}}{=} A_t$. Note that

$$H_m - H_{m,\{m+h\}} = \sum_{k=m+h}^{T} \prod_{\ell=m+h+1}^{k} (I_p - \eta_\ell A_\ell) \eta_{m+h} (A_{m+h} - A^\star_{m+h}) \prod_{\ell=m+1}^{m+h-1} (I_p - \eta_\ell A_\ell).$$

Hence, by Assumption 2, we have $\|A_{m+h} - \Sigma\|_\psi \leq \lambda_0$ and consequently

$$\begin{aligned}
\|H_m - H_{m,\{m+h\}}\|_\psi &\lesssim \sum_{k=m+h}^{T} \eta_{m+h} \|A_{m+h} - \Sigma\|_\psi \prod_{\ell=m+1}^{k} (1 - \lambda_0\eta_\ell) \\
&\leq \lambda_0 \eta_{m+h} \int_{m+h}^{\infty} \exp\left(-\lambda_0\eta_0 \int_{m+1}^{z} x^{-\alpha} dx \right) dz.
\end{aligned}$$

Therefore, together with the fact that $\|A_m\|_\psi \le \|A_m - \Sigma\|_\psi + \|\Sigma\|_2 \le 2\|\Sigma\|_2$, we have

$$\begin{aligned}\|\mathcal{P}_{A,m+h}(\xi_m)\|_{\psi/2} &\le 2\|A_m\|_\psi \|H_m - H_{m,\{m+h\}}\|_\psi \|H_m\|_\psi \\ &\lesssim \lambda_0 \|\Sigma\| \eta_{m+h} \|H_m\|_\psi \int_{m+h}^{\infty} \exp\left(-\lambda_0\eta_0 \int_{m+1}^{z} x^{-\alpha} dx\right) dz.\end{aligned}$$

Define the L-approximation of $D_{T,2}$ as

$$D_{T,2,L} = B\sigma^2 \sum_{m=1}^{T} \eta_m^2 \mathbb{E}(\xi_m | \mathcal{P}_{A,m+L}) = D_{T,2} - B\sigma^2 \sum_{m=1}^{T} \eta_m^2 \sum_{h=L+1}^{T-h} \mathcal{P}_{A,m+h}(\xi_m).$$

Note that $\mathbb{E}(D_{T,2}) = \mathbb{E}(D_{T,2,L})$. Hence, by Lemma 10 and (16),

$$\begin{aligned}\|D_{T,2} - D_{T,2,L}\|_{\psi/2} &\le C_\psi B\sigma^2 \sum_{h=L+1}^{T-1} \left\{ \sum_{m=1}^{T-h} \eta_m^4 \|\mathcal{P}_{A,m+h}(\xi_m)\|_{\psi/2}^2 \right\}^{1/2} \\ &\le \frac{C_{\psi,\alpha} B\sigma^2 \|\Sigma\| T^{1+\alpha}}{\lambda_0^2 L^\alpha} \exp\left(-\frac{\lambda_0\eta_0 L}{2^\alpha T^\alpha}\right) \le C_{\psi,\alpha} T^{-1/2}.\end{aligned}$$

Now we bound $|D_{T,2,L} - \mathbb{E}(D_{T,2,L})|$. By Lemma 11 and a similar argument as that of (17),

$$\mathbb{P}(|D_{T,2,L} - \mathbb{E}(D_{T,2,L})| > z) \le \frac{C_{\psi,\alpha} T L^{\psi/4-1} \|\Sigma\|_2^{\psi/2}}{\lambda_0^\psi (z/B)^{\psi/2}} + C \exp\left\{ -\frac{C'_{\psi,\alpha} (z/B)^2 \lambda_0^4}{T\|\Sigma\|_2^2} \right\}.$$

Consequently, Lemma 13 follows in view of

$$\mathbb{P}(|D_{T,2} - \mathbb{E}(D_{T,2})| > z) \le \mathbb{P}(|D_{T,2} - D_{T,2,L}| > z/2) + \mathbb{P}(|D_{T,2,L} - \mathbb{E}(D_{T,2,L})| > z/2).$$

■

Remaining proof: As discussed in Section 3.2, it suffices to bound $D_{T,q}$ and $D_{T,2}$. By Assumption 2 and a similar argument as (15),

$$\|H_m\|_q \le 1 + \int_{m+1}^{\infty} \exp\left(-\lambda_0\eta_0 \int_{m+1}^{z} x^{-\alpha} dx\right) dz,$$

which leads to

$$\mathbb{E}(D_{T,q}) = B\mu_q \sum_{m=1}^{T} \eta_m^q \mathbb{E}|\omega^\top H_m X_i|^q \le B\mu_q \mathcal{K}_q \sum_{m=1}^{T} \eta_m^q \|H_m\|_q^q \le \frac{C_{q,\alpha} n \mu_q \mathcal{K}_q}{\lambda_0^q}. \tag{17}$$

Hence, by Lemma 13, we have

$$\mathbb{P}\left(D_{T,2} > \mathbb{E}(D_{T,2}) + \frac{x^2}{\log x}\right) \le \frac{C_{\psi,\alpha} T L^{\psi/4-1} \|\Sigma\|_2^{\psi/2} B^{\psi/2}}{\lambda_0^\psi (x^2/\log x)^{\psi/2}} + C \exp\left\{ -\frac{C'_{\psi,\alpha} (x^2/\log x)^2 \lambda_0^4}{T\|\Sigma\|_2^2 B^2} \right\}.$$

As $\psi > (2q-4\alpha)/(2-\alpha)$, for any $x \gtrsim \sqrt{T}$, we have

$$\frac{TL^{\psi/4-1}(\log x)^{\psi/2}}{x^{\psi}} = o\left(\frac{T}{x^q}\right).$$

Consequently, as $\mathbb{E}(D_{T,2}) \leq C_{q,\alpha} n W_2$, we have

$$\mathbb{P}\left(|\omega^\top S_T^\star| > x\right) \leq \frac{C_1 T W_q}{x^q} + C\exp\left(-\frac{C_2 x^2}{TW_2 + x^2/\log x}\right) \leq \frac{C_1 T W_q}{x^q} + C\exp\left(-\frac{C_2 x^2}{TW_2}\right),$$

where C_1 and C_2 are positive constants depending only on q, α and ψ.

5.4 Proof of Theorems 2, 4

As discussed in Section 3.2, Theorem 2 directly follows from Lemma 6 and Lemma 7. The proof of Theorem 4 is similar to that of Theorem 2 and thus omitted.

5.5 Proof of Proposition 8

Proof Let

$$\mu_{T,q} = \sum_{t=1}^{T}\left(\sum_{i=t}^{T} V_t^i \eta_t / T\right)^q.$$

Note that

$$\bar{\theta}_T - \theta^* - \frac{1}{T}\sum_{i=1}^{T} V_i \Delta_0 = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=t}^{T} V_t^i \eta_t \epsilon_t.$$

Since $\{\epsilon_t\}_{t\geq 1}$ are i.i.d. , according to Corollary 1.8 in Nagaev (1979) we have

$$\mathbb{P}\left(\left|\bar{\theta}_T - \theta^* - \frac{1}{T}\sum_{i=1}^{T} V_i\Delta_0\right| \geq x\right) \leq (1+2/q)^q \frac{\mu_{T,q}\mathbb{E}|\epsilon|^q}{x^q} + 2\exp\left(-\frac{c_q x^2}{\mu_{T,2}\mathbb{E}|\epsilon|^2}\right).$$

Then all we need to show is that $\mu_{T,q} \asymp T^{1-q}$ for $2 \leq q < \nu$, and $\sum_{i=1}^{T} V_i = O(1)$. Since there's no randomness in $\mu_{T,q}$ and V_i, we can check the order through numerical computation; see figure 3. Also, according to Lemma $A.2$ in Zhu et al. (2021), $\sum_{i=t}^{T} V_t^i = O(t^\alpha)$ for $\alpha \in (1/2, 1)$, which implies that $\mu_{T,q} = O(T^{1-q})$. ■

5.6 Proof of Proposition 9

Proof Let

$$S_T = \bar{\theta}_T - \theta^* - \frac{1}{T}\sum_{i=1}^{T} V_i\Delta_0 = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=t}^{T} V_t^i \eta_t \epsilon_t.$$

To apply Theorem 1 in Peligrad et al. (2014), we need to verify the basic assumption, the uniform asymptotic negligibility of the variance of individual summands, that is

$$\max_t \left(\sum_{i=t}^{T} V_t^i \eta_t\right)^2 \Big/ \sum_{t=1}^{T}\left(\sum_{i=t}^{T} V_t^i \eta_t\right)^2 \to 0. \tag{18}$$

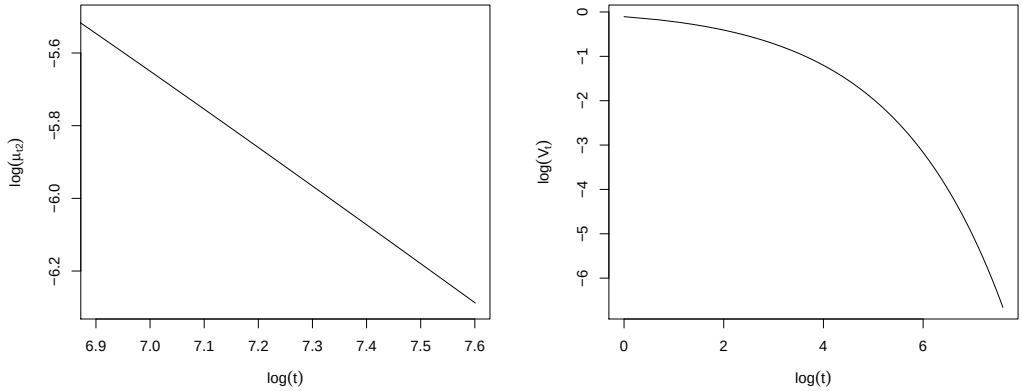

Figure 3: Left: Check the order of $\mu_{T,2}$. The X axis represents $\log(t)$; the Y axis represents $\log(\mu_{t,2})$. The slop of the log-log curve is about -1, which implies that $\mu_{T,2} \asymp T^{-1}$. Right: Check the order of V_t. The X axis represents $\log(t)$; the Y axis represents $\log(V_t)$. The slop of the log-log curve is much less than -1 when t is large, which means V_t is summable and $\sum_{i=1}^{T} V_i = O(1)$.

Since Lemma $A.2$ in Zhu et al. (2021) shows $\sum_{i=t}^{T} V_t^i \asymp t^{\alpha}$ as $T \to \infty$ and $\eta_t = \eta_0 t^{-\alpha}$, the above limit is of order T^{-1}. (Note that $\sigma_T^2 = \mathbb{E}|\epsilon|^2 \mu_{T,2} \asymp T^{-1}$.) We can also verify (18) from numerical computation; see Figure 4. Then, according to Theorem 1 in Peligrad et al. (2014), we have

$$\mathbb{P}\left(S_T \geq x\right) = (1 + o(1)) \left(1 - \Phi(x/\sigma_T) + \sum_{t=1}^{T} P\left(\sum_{i=t}^{T} V_t^i \eta_t \epsilon_t / T \geq x\right)\right),$$

which naturally yields (14).

■

6. Discussion and Future Directions

In this paper, we established nearly tight tail probabilities for SGD errors with heavy-tailed noises in linear models. The resulting high probability error bounds and sample complexity are quite different from those obtained in light-tailed noise cases. In particular, with probability at least $1 - \delta$, we have $\|\bar{\theta}_T - \theta^\star\|_2^2 \leq O\left(T^{-1}\log(1/\delta) + (\delta T^{q-1})^{-2/q}\right)$, where the polynomial dependence on the failure probability δ is generally unavoidable. For future directions, it is interesting to extend our concentration analysis under heavy-tailed noise assumptions to other examples of SA. Also, the robust modification of SGD can be a promising topic to accommodate the heavy-tailed noises.

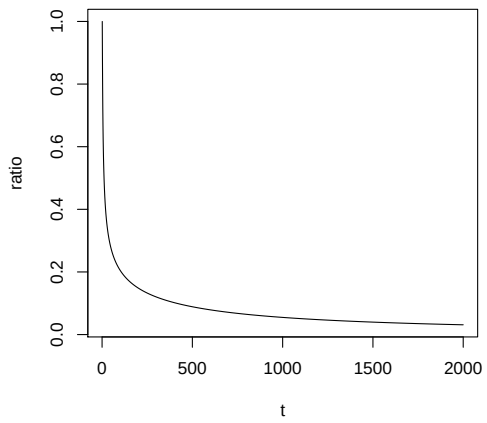

Figure 4: Check the uniform asymptotic negligibility of the variance of individual summands. The X axis represents t; the Y axis represents the ratio of the largest individual variance and variance of individual summands.

Acknowledgments

Wanrong Zhu and Wei Biao Wu would like to thank the support from NSF via NSF-DMS-1916351 and NSF-DMS-2027723

References

Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat A. Erdogdu. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale clt. In *Conference on Learning Theory*, 2019.

Francis R. Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $\mathcal{O}(1/n)$. In *Advances in Neural Information Processing Systems*, 2013.

Julius R. Blum. Approximation methods which converge with probability one. *The Annals of Mathematical Statistics*, 25(2):382–386, 1954.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Donald L Burkholder. Sharp inequalities for martingales and stochastic integrals. *Astérisque*, 157(158):75–94, 1988.

Hervé Cardot, Peggy Cénac, Antoine Godichon-Baggioni, et al. Online estimation of the geometric median in hilbert spaces: Nonasymptotic confidence balls. *The Annals of Statistics*, 45(2):591–614, 2017.

Xi Chen, Jason D. Lee, Xin T. Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 02 2020.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, 2019.

K. L. Chung. On a Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 25(3):463 – 483, 1954.

Damek Davis and Dmitriy Drusvyatskiy. High probability guarantees for stochastic convex optimization. In *Conference on Learning Theory*, 2020.

Aryeh Dvoretzky. On stochastic approximation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1956.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, 2019.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Advances in Neural Information Processing Systems*, 2020.

Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, 2019a.

Nicholas JA Harvey, Christopher Liaw, and Sikander Randhawa. Simple and optimal high-probability bounds for strongly-convex stochastic gradient descent. Preprint. Available at arXiv:1909.00843, 2019b.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. Preprint. Available at arXiv:1903.08560, 2019.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(71):2489–2512, 2014.

Matthew Holland and Kazushi Ikeda. Better generalization with less data using robust gradient descent. In *International Conference on Machine Learning*, 2019.

Prateek Jain, Praneeth Netrapalli, Sham M Kakade, Rahul Kidambi, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *The Journal of Machine Learning Research*, 18(1): 8258–8299, 2017.

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of sgd information theoretically optimal. In *Conference on Learning Theory*, 2019.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017.

Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an $\mathcal{O}(1/t)$ convergence rate for the projected stochastic subgradient method. Preprint. Available at arXiv:1212.2002, 2012.

Tze Leung Lai. Stochastic approximation. *The Annals of Statistics*, 31(2):391–406, 2003.

Mu Li, Tong Zhang, Yuqiang Chen, and Alexander J Smola. Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014.

Wenlong Mou, Chris Junchi Li, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. On linear stochastic approximation: Fine-grained polyak-ruppert and non-asymptotic concentration. In *Conference on Learning Theory*, 2020.

Eric Moulines and Francis R. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, 2011.

S. V. Nagaev. Large deviations of sums of independent random variables. *The Annals of Probability*, 7(5):745–789, 1979.

Alexander V Nazin, Arkadi S Nemirovsky, Alexandre B Tsybakov, and Anatoli B Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80(9):1607–1627, 2019.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Magda Peligrad, Hailin Sang, Yunda Zhong, and Wei Biao Wu. Exact moderate and large deviations for linear processes. *Statistica Sinica*, 24:957–969, 2014.

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

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *International Conference on Machine Learning*, 2012.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning*, 2016.

Emmanuel Rio. Moment inequalities for sums of dependent random variables under projective conditions. *Journal of Theoretical Probability*, 22(1):146–163, 2009.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, 2013.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, 2019.

Panos Toulis and Edoardo M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. Preprint. Available at arXiv:1011.3027, 2010.

Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association (To appear)*, 2021.