A Piecewise Lyapunov Analysis of Sub-quadratic SGD: Applications to Robust and Quantile Regression

Yixuan Zhang
University of Wisconsin-Madison
Madison, Wisconsin, USA
yzhang2554@wisc.edu

Dongyan (Lucy) Huo
Cornell University
Ithaca, New York, USA
dh622@cornell.edu

Yudong Chen
University of Wisconsin-Madison
Madison, Wisconsin, USA
yudong.chen@wisc.edu

Qiaomin Xie
University of Wisconsin-Madison
Madison, Wisconsin, USA
qiaomin.xie@wisc.edu

Abstract

Motivated by robust and quantile regression problems, we investigate the stochastic gradient descent (SGD) algorithm for minimizing an objective function f that is locally strongly convex with a sub–quadratic tail. This setting covers many widely used online statistical methods. We introduce a novel piecewise Lyapunov function that enables us to handle functions f with only first-order differentiability, which includes a wide range of popular loss functions such as Huber loss. Leveraging our proposed Lyapunov function, we derive finite-time moment bounds under general diminishing stepsizes, as well as constant stepsizes. We further establish the weak convergence, central limit theorem and bias characterization under constant stepsize, providing the first geometrical convergence result for sub–quadratic SGD. Our results have wide applications, especially in online statistical methods. In particular, we discuss two applications of our results. 1) Online robust regression: We consider a corrupted linear model with sub–exponential covariates and heavy–tailed noise. Our analysis provides convergence rates comparable to those for corrupted models with Gaussian covariates and noise. 2) Online quantile regression: Importantly, our results relax the common assumption in prior work that the conditional density is continuous and provide a more fine-grained analysis for the moment bounds.[1]

CCS Concepts

• **Mathematics of computing** → *Robust regression*; • **Theory of computation** → **Stochastic approximation**.

Keywords

Stochastic gradient descent, sub-quadratic tail, piecewise Lyapunov function, robust regression, quantile regression

[1]Extended Abstract. The full paper can be found at [23].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGMETRICS Abstracts '25, Stony Brook, NY, USA

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1593-8/2025/06
https://doi.org/10.1145/3726854.3727269

ACM Reference Format:
Yixuan Zhang, Dongyan (Lucy) Huo, Yudong Chen, and Qiaomin Xie. 2025. A Piecewise Lyapunov Analysis of Sub-quadratic SGD: Applications to Robust and Quantile Regression. In *Abstracts of the 2025 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS Abstracts '25), June 9–13, 2025, Stony Brook, NY, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3726854.3727269

1 Introduction

The problem of minimizing objective functions that are not strongly convex has garnered considerable attention across various domains, including modern statistical machine learning—such as matrix and tensor completion [9], deep neural networks [17], and robust statistics [13]—as well as in optimization [6] and stochastic approximation [1].

Specifically, in robust regression [2], the goal is to recover the underlying model when the data is contaminated by outliers and/or corruption. In particular, it aims to find the global optimizer θ^* of the population-level loss function:

$$f(\theta) = \mathbb{E}[l(y - x^\top \theta)], \tag{1}$$

when given observations of independent and identically distributed (i.i.d.) data $\{(x_n, y_n)\}_{n\geq 0}$ generated from a corrupted linear model: $y = x^\top \theta^* + \epsilon + s$, where ϵ represents the error and s denotes the corruption. In robust regression, the error ϵ is typically heavy–tailed without a second moment bound. It is well recognized that the classical squared loss function cannot handle heavy–tailed errors or corruption effectively. To encourage robustness, the field has been widely using loss functions l that have *sub–quadratic tails* to assign less weight to outliers. Common robust loss functions include Huber [12], Pseudo-Huber loss [10] and log-cosh loss [21]. Importantly, the population-level loss f inherits the sub–quadratic tail behavior of the loss function l. Meanwhile, we observe that f is strongly convex in a neighborhood of θ^*. The problem of minimizing objective functions with local strong convexity and sub–quadratic tails also arises in quantile regression [15]. Quantile regression estimates the conditional quantile of the response given covariates, and is widely used in various applications [22].

Motivated by the robust and quantile regression problems, we consider stochastic gradient descent (SGD) algorithms aimed at approximating the global minimum $\theta^* \in \mathbb{R}^d$ of a non-strongly convex function $f : \mathbb{R}^d \to \mathbb{R}$, using unbiased estimates of the

function's gradients:

$$\theta_{n+1} = \theta_n - \alpha_n \left(\nabla f(\theta_n) + w_n(\theta_n)\right), \tag{2}$$

where ∇f denotes the population-level gradient of objective function f, $\{\alpha_n\}_{n\geq 0}$ are the stepsize sequence, and $\{w_n(\cdot)\}_{n\geq 0}$ are i.i.d. copies of a random field $w(\cdot)$ with $\mathbb{E}[w(\theta)] = 0$ for all $\theta \in \mathbb{R}^d$. In this work, we focus on a class of objective functions that are locally strongly convex around θ^* and exhibit a **sub–quadratic tail**. Throughout the work, we refer to this function class as *sub–quadratic functions* and, for brevity, we call the corresponding SGD procedure *sub–quadratic SGD*.

Sub–quadratic SGD covers many online statistical algorithms, including robust regression [19] and quantile regression [4]. Classical studies on robust regression [2] and quantile regression [15] primarily focus on processing complete datasets, which can be computationally inefficient and memory intensive. In practice, data arrives sequentially or is too large for batch processing, making online methods more scalable. Prior online robust regression [19] and quantile regression [4] impose stronger assumptions (e.g., Gaussian covariates, continuous error densities). Our sub–quadratic SGD framework unifies and relaxes these requirements.

One line of prior work considers strongly convex objective functions f, where the behavior of the iterates $\{\theta_n\}_{n\geq 0}$ is well-understood under both diminishing stepsizes ($\alpha_n \to 0$) and constant stepsizes ($\alpha_n \equiv \alpha$). With diminishing stepsizes, studies have demonstrated that setting $\alpha_n = \iota/(n+\kappa)$ achieves the optimal convergence rate of $\mathbb{E}[\|\theta_n - \theta^*\|^2] = O(1/n)$ [3, 5]. Under a constant stepsize α, the iterates $\{\theta_n\}_{n\geq 0}$ form a Markov chain that converges geometrically to a limiting random variable $\theta_\infty^{(\alpha)}$, which oscillates around θ^* [7, 24]. Notably, the asymptotic bias $\mathbb{E}[\theta_\infty^{(\alpha)}] - \theta^*$ is typically nonzero. When ∇f is continuously differentiable at θ^*, this bias is proportional to the stepsize α [7, 14, 25]; when ∇f lacks continuous differentiability the bias is proportional to $\sqrt{\alpha}$ [24].

In contrast, our understanding of SGD for minimizing sub–quadratic functions f is very limited. Prior work [16] has primarily focused on deterministic gradient descent ($w(\cdot) \equiv 0$), providing convergence rates for $\|\theta_n - \theta^*\|$ with diminishing stepsizes. Recent work [8] considered sub–quadratic SGD with diminishing stepsizes and derived non-asymptotic moment bounds for $\mathbb{E}[\|\theta_n - \theta^*\|^p]$. However, their analysis requires twice differentiablility of the function f and restrictive assumption on the noise sequence $\{w_n(\cdot)\}_{n\geq 0}$, due to the limitation of their proposed Lyapunov function. Beyond diminishing stepsizes, the study of constant stepsize sub–quadratic SGD via a Markov chain perspective remains scarce. In recent developments, [20] introduced a novel technique to analyze the weak convergence of Markov chains, validating their approach with specific examples of sub–quadratic SGD involving solely additive (i.e., $w(\theta)$ is independent of θ) and heavy–tailed noise.

2 Preliminaries

In this work, we study the SGD of sub–quadratic objective functions f satisfying the following assumptions.

Assumption 1 (Smoothness). *There exists a constant $L > 0$ such that, for all $\theta, \theta' \in \mathbb{R}^d$, $\|\nabla f(\theta) - \nabla f(\theta')\| \leq L\|\theta - \theta'\|$.*

Assumption 2 (Local Strong Convexity and sub–quadratic Tail). *There exist $\theta^* \in \mathbb{R}^d$ and constants $\mu, a, b, \Delta > 0$ and $k \in [1, 2)$ such that:*

(1) $\forall \theta', \theta'' \in \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\| \leq \Delta\}$,

$$\langle \theta' - \theta'', \nabla f(\theta') - \nabla f(\theta'')\rangle \geq \mu\|\theta' - \theta''\|^2.$$

(2) $\forall \theta \in \mathbb{R}^d \in \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\| > \Delta\}$,

$$\|\nabla f(\theta)\| \leq a\|\theta - \theta^*\|^{k-1} \quad \text{and} \quad \langle \theta - \theta^*, \nabla f(\theta)\rangle \geq b\|\theta - \theta^*\|^k.$$

Assumption 3 (q). *$\|w(\theta)\|$ is in ψ_q -Orlicz Space for all $\theta \in \mathbb{R}^d$.*

We remark that many popular loss functions in robust statistics, such as Huber loss and Pseudo-Huber loss, satisfy Assumptions 1 and 2. In this work, we focus on objective functions satisfying Assumption 2 with a parameter $k \in [1, 2)$. Consequently, we require the noise sequence to fulfill Assumption 3(q) with $q = 2 - k$. We note that a sub-exponential random variable is in ψ_1-Orlicz space.

3 Main Results

In this work, we investigate sub–quadratic SGD, focusing on scenarios where the objective function f is only first-order differentiable and exhibits tails with at least linear growth ($k \in [1, 2)$). This function class covers many widely adopted loss functions in various domains. Building upon this class of SGD, our main contributions are summarized as follows.

- **A Novel Piecewise Lyapunov Function:** A central challenge in analyzing $\{\theta_n\}_{n\geq 0}$ is to construct an appropriate Lyapunov function. This task is complicated by the piecewise behavior of the objective function, which is locally strongly convex around θ^* but has a sub–quadratic tail. Prior work [8] develops a Lyapunov function for sub–quadratic SGD but imposes restrictive conditions on both the objective function and noise. In this work, we propose a novel piecewise Lyapunov function that effectively exploits the structure of sub–quadratic SGD without such restrictive assumptions:

$$V_{k,0}(\theta) = \begin{cases} \exp\left(\frac{k\|\theta-\theta^*\|^{2-k}}{(2-k)\Delta^{2-k}}\right) - \left(1 - \frac{k}{2}\right)\exp\left(\frac{k}{2-k}\right), & \text{if } \|\theta - \theta^*\| > \Delta, \\ \frac{k\exp(k/(2-k))\|\theta-\theta^*\|^2}{2\Delta^2}, & \text{if } \|\theta - \theta^*\| \leq \Delta; \end{cases}$$

and $V_{k,p}(\theta) = \|\theta - \theta^*\|^p \cdot V_{k,0}(\theta), \forall p \in \mathbb{N}$.

Our construction of the Lyapunov function V is guided by the intuition that V should satisfy the drift condition $\mathbb{E}[V(\theta_1) \mid \theta_0] \leq (1 - \alpha v)V(\theta_0) + O(\alpha^2)$ for establishing geometric convergence. Consequently, we aim to find V that satisfies the condition $V'f' \in \Omega(V)$. Using the family of $V_{k,p}$, we prove the following proposition.

Proposition 3.1 (Informal). *Under Assumptions 1–2 with $k \in [1, 2)$,*

$$\mathbb{E}[V_{k,0}(\theta_1) \mid \theta_0] \lesssim (1 - \alpha)V_{k,0}(\theta_0) + \alpha^2,$$

$$\mathbb{E}[V_{k,p}(\theta_1) \mid \theta_0] \lesssim (1 - \alpha)V_{k,p}(\theta_0) + \alpha^2 V_{k,p-2}(\theta_0) + \alpha^{p+2}, \forall p \geq 2.$$

Although the second inequality in Proposition 3.1 does not offer an exact one-step contraction, it establishes a recursive relationship among $V_{k,p}(\theta_{n+1})$, $V_{k,p}(\theta_n)$, and $V_{k,p-2}(\theta_n)$, which would allow us to upper bound $\mathbb{E}[V_{k,p}(\theta_n)]$ by employing induction on p and n. Proposition 3.1 allows us to derive the following results.

- **Finite-Time Moment Bounds:** We explore the finite-time moment bounds under constant and diminishing stepsizes.

  Theorem 3.2 (Informal). *Consider SGD dynamic* (2) *with* $\alpha_n = \frac{\iota}{(n+\kappa)^\xi}$, *under Assumptions 1–3 with* $k \in [1,2)$,

  (1) When $\xi \in (0,1]$, *for all* $n \geq 0$, *we have*

  $$\mathbb{E}[\|\theta_n - \theta^*\|^2] \lesssim 1/(n+\kappa)^\xi.$$

  (2) When $\xi = 0$, *i.e.,* $\alpha_n \equiv \alpha$, *for all* $p \in \mathbb{N}$ *and* $n \geq 0$, *we have*

  $$\mathbb{E}[\|\theta_n - \theta^*\|^{2p+2}] \lesssim \exp(-\alpha n) + \alpha^{p+1}.$$

  While the recent work [8] also obtained moment bounds, they only considered diminishing stepsizes $\alpha_n = \frac{\iota}{(n+\kappa)^\xi}$ with $\xi \in [1/2, 1)$. In contrast, with our new piecewise Lyapunov function, we are able to analyze a broader class of diminishing stepsizes with $\xi \in (0,1]$. In particular, we recover the results in [8], without requiring the objective function to be twice differential or imposing restrictive assumptions on the noise sequence. Theorem 3.2 states that with a constant stepsize α, the $2p$-th moment $\mathbb{E}[\|\theta_n - \theta^*\|^{2p}]$ is bounded by a geometrically decaying term (rate scaling with α) plus an $O(\alpha^p)$ bias. This result resembles the behavior of constant-stepsize SGD for strongly convex objectives [5, 24]—to the best of our knowledge, this is the first high-moment result under subquadratic objectives.
- **Weak Convergence, Central Limit Theorem and Bias Characterization:** We also investigate the fluctuations of $\{\theta_n\}_{n\geq 0}$ of sub–quadratic SGD with constant stepsize.

  Theorem 3.3 (Informal). *Consider SGD dynamic* (2) *with* $\alpha_n \equiv \alpha$, $\{\theta_n\}_{n\geq 0}$ *geometrically converges to a unique limit random variable* θ_∞ *in Wasserstein* 1*-distance. Moreover, we have*

  $$\mathbb{E}[\theta_\infty] - \theta^* = \alpha B + O(\alpha^{3/2}),$$

  where $B = f''(\theta^*)^{-1} f'''(\theta^*)(f''(\theta^*) \otimes I_d + I_d \otimes f''(\theta^*))^{-1} S(\theta^*)$ *and* $S(\theta) = \mathbb{E}[w(\theta) w(\theta)^T]$.

  Theorem 3.3 provides the first geometric convergence result for SGD applied to sub–quadratic functions. Having established weak convergence, we further prove a central limit theorem for the Markov chain $\{\theta_n\}_{n\geq 0}$, which is crucial for statistical inference [18]. Using a constant stepsize induces an asymptotic bias $\mathbb{E}[\theta_\infty^{(\alpha)}] - \theta^*$ that is in general not zero. As demonstrated in Theorem 3.3, this asymptotic bias is proportional to α up to higher-order terms. This finding has important algorithmic implications for reducing the bias to higher orders of α through Richardson-Romberg (RR) extrapolation technique [11]. Therefore, by applying RR extrapolation to constant stepsize sub–quadratic SGD, one can achieve fast convergence with a reduced bias term.
- **Applications in Robust Regression and Quantile Regression:** We apply our results on general sub–quadratic SGD to important statistical problems: online robust regression [19] and online quantile regression [4]. For both settings, our new Lyapunov analysis allows us to relax restrictive assumptions considered in prior work and thereby provide a more comprehensive analysis. Specifically, our results are applicable to online robust regression with sub–exponential covariates and heavy–tailed error, and similarly to online quantile regression with sub–exponential covariates without requiring the continuity of the conditional density of the error ϵ.

Acknowledgments

Q. Xie and Y. Zhang are supported in part by National Science Foundation (NSF) grants CNS-1955997 and EPCN-2339794 and EPCN-2432546. Y. Chen is supported in part by NSF grant CCF-2233152. Y. Zhang is also supported in part by NSF Award DMS-2023239.

References

[1] Francis Bach and Eric Moulines. 2013. Non-strongly-convex smooth stochastic approximation with convergence rate O (1/n). *Advances in neural information processing systems* 26 (2013).
[2] Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. 2017. Consistent robust regression. *Advances in Neural Information Processing Systems* 30 (2017).
[3] Léon Bottou, Frank E Curtis, and Jorge Nocedal. 2018. Optimization methods for large-scale machine learning. *SIAM review* 60, 2 (2018), 223–311.
[4] Xi Chen, Weidong Liu, and Yichen Zhang. 2019. Quantile regression under memory constraint. (2019).
[5] Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. 2020. Finite-sample analysis of stochastic approximation using smooth convex envelopes. *arXiv preprint arXiv:2002.00874* (2020).
[6] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. 2014. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems* 27 (2014).
[7] Aymeric Dieuleveut, Alain Durmus, and Francis Bach. 2018. Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains. arXiv:1707.06386 [stat.ML]
[8] Sébastien Gadat and Fabien Panloup. 2023. Optimal non-asymptotic analysis of the Ruppert–Polyak averaging stochastic algorithm. *Stochastic Processes and their Applications* 156 (2023), 312–348.
[9] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. 2015. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*. PMLR, 797–842.
[10] Richard Hartley and Andrew Zisserman. 2003. *Multiple view geometry in computer vision*. Cambridge university press.
[11] Francis Begnaud Hildebrand. 1987. *Introduction to numerical analysis*. Courier Corporation.
[12] Peter J Huber. 1992. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*. Springer, 492–518.
[13] Peter J Huber and Elvezio M Ronchetti. 2011. *Robust statistics*. John Wiley & Sons.
[14] Dongyan Huo, Yudong Chen, and Qiaomin Xie. 2023. Bias and extrapolation in Markovian linear stochastic approximation with constant stepsizes. In *Abstract Proceedings of the 2023 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*. 81–82.
[15] Roger Koenker. 2005. Quantile regression. *Cambridge Univ Pr* (2005).
[16] Krzysztof Kurdyka. 1998. On gradients of functions definable in o-minimal structures. In *Annales de l'institut Fourier*, Vol. 48. 769–783.
[17] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.
[18] Xiang Li, Wenhao Yang, Jiadong Liang, Zhihua Zhang, and Michael I Jordan. 2023. A statistical analysis of polyak-ruppert averaged q-learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2207–2261.
[19] Scott Pesme and Nicolas Flammarion. 2020. Online robust regression via sgd on the l1 loss. *Advances in Neural Information Processing Systems* 33 (2020), 2540–2552.
[20] Yanlin Qu, Jose Blanchet, and Peter Glynn. 2023. Computable Bounds on Convergence of Markov Chains in Wasserstein Distance. *arXiv preprint arXiv:2308.10341* (2023).
[21] Resve A Saleh and AK Saleh. 2022. Statistical properties of the log-cosh loss function used in machine learning. *arXiv preprint arXiv:2208.04564* (2022).
[22] Huixia Judy Wang and Lan Wang. 2014. Quantile regression analysis of length-biased survival data. *Stat* 3, 1 (2014), 31–47.
[23] Yixuan Zhang, Dongyan, Huo, Yudong Chen, and Qiaomin Xie. 2025. A Piecewise Lyapunov Analysis of sub–quadratic SGD: Applications to Robust and Quantile Regression. arXiv:2504.08178 [stat.ML] https://arxiv.org/abs/2504.08178
[24] Yixuan Zhang, Dongyan Huo, Yudong Chen, and Qiaomin Xie. 2024. Prelimit Coupling and Steady-State Convergence of Constant-stepsize Nonsmooth Contractive SA. *arXiv preprint arXiv:2404.06023* (2024).
[25] Yixuan Zhang and Qiaomin Xie. 2024. Constant Stepsize Q-learning: Distributional Convergence, Bias and Extrapolation. *arXiv preprint arXiv:2401.13884* (2024).