Heavy-tailed Streaming Statistical Estimation

Che-Ping Tsai[†], Adarsh Prasad[†], Sivaraman Balakrishnan[‡], Pradeep Ravikumar[†]

chepingt@cs.cmu.edu, adarshp@andrew.cmu.edu
siva@stat.cmu.edu, pradeepr@cs.cmu.edu

[†]Department of Machine Learning
[‡]Department of Statistics and Data Science
Carnegie-Mellon University
Pittsburgh, PA 15213

February 28, 2022

Abstract

We consider the task of heavy-tailed statistical estimation given streaming p-dimensional samples. This could also be viewed as stochastic optimization under heavy-tailed distributions, with an additional $O(p)$ space complexity constraint. We design a clipped stochastic gradient descent algorithm and provide an improved analysis, under a more nuanced condition on the noise of the stochastic gradients, which we show is critical when analyzing stochastic optimization problems arising from general statistical estimation problems. Our results guarantee convergence not just in expectation but with exponential concentration, and moreover does so using $O(1)$ batch size. We provide consequences of our results for mean estimation and linear regression. Finally, we provide empirical corroboration of our results and algorithms via synthetic experiments for mean estimation and linear regression.

1 Introduction

Statistical estimators are typically random, since they depend on a random training set; their statistical guarantees are typically stated in terms of the expected loss between estimated and true parameters [35, 14, 54, 29]. A bound on expected loss however might not be sufficient in higher stakes settings, such as autonomous driving, and risk-laden health care, among others, since the deviation of the estimator from its expected behavior could be large. In such settings, we might instead prefer a bound on the loss that holds with high probability. Such high-probability bounds are however often stated only under strong assumptions (e.g. sub-Gaussianity or boundedness) on the tail of underlying distributions [28, 27, 48, 21]; conditions which often do not hold in real-world settings. There has also been a burgeoning line of recent work that relaxes these assumptions and allows for heavy-tailed underlying distributions [5, 32, 39], but the resulting algorithms are often not only complex, but are also specifically batch learning algorithms that require storing the entire dataset, which limits their scalability. For instance, many popular polynomial time algorithms on heavy-tailed mean estimation [11, 7, 8, 36, 13, 10] and heavy-tailed linear regression algorithms [32, 49, 46] need to store the dataset to take polylogarithmic passes over data.

On the other hand, most successful practical modern learning algorithms are iterative, light-weight and access data in a "streaming" fashion. As a consequence, we focus on designing and analyzing iterative statistical estimators which only use constant storage in each step. To summarize, motivated by practical considerations, we have three desiderata: (1) allowing for heavy-tailed underlying distributions (weak modeling assumptions),

(2) high probability bounds on the loss between estimated and true parameters instead of just its expectation (strong statistical guarantees), and (3) estimators that access data in a streaming fashion while only using constant storage (scalable, simple algorithms).

A useful alternative viewpoint of the statistical estimation problem above is that of stochastic optimization: where we have access to the optimization objective function (which in the statistical estimation case is simply the population risk of the estimator) only via samples of the objective function or its higher order derivatives (typically just the gradient). Here again, most of the literature on stochastic optimization typically provides bounds in expectation [29, 54, 25], or places a strong assumptions on the tail behavior of the distributions of the derivatives of the stochastic objective, such as the distributions being bounded [27, 48] or sub-Gaussian [28, 38]. Figure 1 shows that even for the simple stochastic optimization task of mean estimation, the deviation of stochastic gradient descent (SGD) is much worse for heavy-tailed distributions than sub-Gaussian ones. Therefore, bounds on expected behavior, or strong assumptions on the tails of the stochastic noise distribution are no longer sufficient.

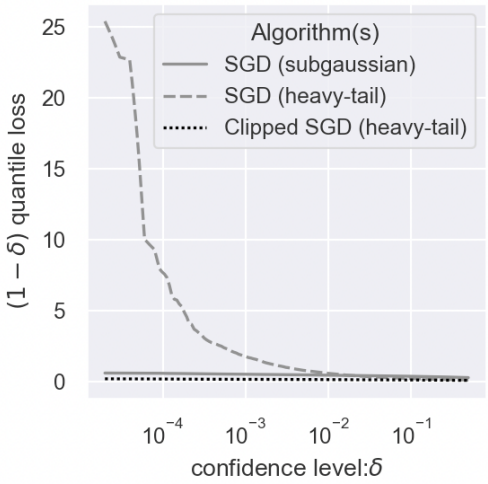


Figure 1: Tail performance of SGD and clipped-SGD of mean estimation. The underlying distributions are zero-mean sub-Gaussian or heavy-tailed distributions. This figure shows ℓ_2-loss between estimated and true mean against different confidence levels δ. See more details in Section C.1.

While there has been a line of work on heavy-tailed stochastic optimization, these require non-trivial storage complexity or batch sizes, making them unsuitable for streaming settings or large-scale problems [19, 41]. Specifically these existing works require at least $O(1/\epsilon)$ batch size to obtain a ϵ-approximate solution under heavy-tailed noise [47, 9, 24, 43] (See Section B in the Appendix for further discussion). In other words, to achieve a typical $O(1/N)$ convergence rate (on the squared error), where N is the number of samples, they would need a batch-size of nearly the entire dataset.

Therefore, we investigate the following question:

Can we develop a stochastic optimization method that satisfy our three desiderata?

Our answer is that a simple algorithm suffices: *stochastic gradient descent with clipping* (`clipped-SGD`). In particular, we first prove a high probability bound for `clipped-SGD` under heavy-tailed noise, with a decaying $O(1/t)$ step-size sequence for strongly convex objectives. By using a decaying step size, we improve the analysis of [24] and develop the first robust stochastic optimization algorithm *in a fully streaming setting* - i.e. with $O(1)$ batch size. We then consider corollaries for statistical estimation where the optimization objective is the population risk of the estimator, and derive the first streaming robust mean estimation and linear regression algorithm that satisfy all three desiderata above.

We summarize our contributions as follows:

- We prove the first high-probability bounds for `clipped-SGD` with a $O(1/t)$ step size and a constant batch size for strongly convex and smooth objectives *without sub-Gaussian assumption on stochastic gradients*. To the best of our knowledge, this is the first stochastic optimization algorithm that uses constant batch size in this setting. See Section 1.1 and Table 1 for more details and comparisons. A critical ingredient is a nuanced condition on the stochastic gradient noise.
- We show that our proposed stochastic optimization algorithm can be used for a broad class of statistical estimation problems. As corollaries of this framework, we present a new class of robust heavy-tailed estimators for streaming mean estimation and linear regression.

- Lastly, we conduct synthetic experiments to corroborate our theoretical and methodological developments. We show that `clipped-SGD` not only outperforms SGD and a number of baselines in average performance but also has a well-controlled tail performance.

1.1 Related Work

Batch Heavy-tailed mean estimation. In the batch-setting, [31] proposed the first polynomial-time algorithm that matches the error guarantees achieved by the empirical mean on Gaussian data. After this work, efficient algorithms with improved asymptotic runtimes were proposed: Hopkins [31], Cherapanamjeri et al. [8] proposed optimal estimators based on semi-definite programming (SDP). Diakonikolas et al. [12], Hopkins et al. [30], Cheng et al. [7], Lei et al. [36], Dong et al. [13] constructed more practical algorithms via spectral techniques. Estimators [42, 10] relied on the median-of-means framework. However, these approaches are not designed for the streaming setting and requires taking polylogarithmic passes over data. We discuss the difficulties in applying these approaches in the streaming setting in Section 4.1.

Batch Heavy-tailed regression. For the setting where the regression noise w is heavy-tailed with bounded variance, Huber's estimator is known to have exponential deviation bounds [16, 50] in high dimensional setting. For the case where both the covariates and the noise are both heavy-tailed, several recent works have proposed computationally efficient estimators that achieve exponential deviation bounds based on the median-of-means framework [40, 32, 42], thresholding techniques [49], and covariate filtering [46]. However, as we noted before, computing all of these estimators require storing the entire dataset.

Heavy-tailed stochastic optimization. A line of work in stochastic convex optimization have proposed bounds that achieve sub-Gaussian concentration around their mean (a common step towards providing sharp high-probability bounds), while only assuming that the variance of stochastic gradients is bounded (i.e. allowing for heavy-tailed stochastic gradients). Davis et al. [9] proposed `proxBoost` that is based on robust distance estimation and proximal operators. Prasad et al. [47] utilized the geometric median-of-means to robustly estimate gradients in each mini-batch. Gorbunov et al. [24] and Nazin et al. [43] proposed `clipped-SSTM` and `RSMD` respectively based on truncation of stochastic gradients for stochastic mirror/gradient descent. Zhang et al. [54] analyzed the convergence of `clipped-SGD` *in expectation* but focus on a different noise regime where the distribution of stochastic gradients has bounded $1+\alpha$ moments for some $0 < \alpha \leq 1$. However, all the above works [9, 47, 24, 43] have an unfavorable $O(n)$ dependency on the batch size to get the typical $O(1/n)$ convergence rate (on the squared error). We note that our bound is comparable to the above approaches while using a constant batch size. See Appendix B for more details.

2 Background and Problem Formulation

In this paper, we consider the following statistical estimation/stochastic optimization setting: we assume that there are some class of functions $\{f_\theta\}_{\theta\in\Theta}$ parameterized by θ, where Θ is a convex subset of $\mathbb{R}^p$; some random vector x with distribution P; and a loss function $\overline{\mathcal{L}}$ which takes x, θ and outputs the loss of f_θ at point x. In this setting, we want to recover the true parameter θ^* defined as the minimizer of the population risk function $\mathcal{R}(\theta)$:

$$\theta^* = \operatorname*{argmin}_{\theta\in\Theta} \mathcal{R}(\theta) = \operatorname*{argmin}_{\theta\in\Theta} \mathbb{E}_{x\sim P}[\overline{\mathcal{L}}(\theta, x)]. \tag{1}$$

We assume that $\overline{\mathcal{L}}$ is differentiable and convex, and further impose two regularity conditions on the population risk: there exist τ_ℓ and τ_u such that

$$\frac{\tau_\ell}{2}\|\theta_1-\theta_2\|_2^2 \leq \mathcal{R}(\theta_1) - \mathcal{R}(\theta_2) - \langle \nabla\mathcal{R}(\theta_2), \theta_1-\theta_2\rangle \leq \frac{\tau_u}{2}\|\theta_1-\theta_2\|_2^2 \tag{2}$$

with $\tau_\ell, \tau_u > 0$ and it holds for all $\theta_1, \theta_2 \in \Theta$. The parameters τ_ℓ, τ_u are called the strong-convexity and smoothness parameters of the function $\mathcal{R}(\theta)$.

To solve the minimization problem defined in Eq. (1), we assume that we can access the stochastic gradient of the population risk, $\nabla\overline{\mathcal{L}}(\theta, x)$, at any point $\theta \in \Theta$ given a sample x. We note that this is a unbiased gradient estimator, i.e.

$$\mathbb{E}_{x\sim P}[\nabla\overline{\mathcal{L}}(\theta, x)] = \nabla\mathcal{R}(\theta).$$

Our goal in this work is to develop robust statistical methods under heavy-tailed distributions. The specific characterization of heavy-tailed distributions we consider in this paper is the common notion of distributions where only very low order moments may be finite, e.g. student-t distribution or Pareto distribution. In this work, we assume the stochastic gradient distribution only has bounded second moment. Formally, for any $\theta \in \Theta$, we assume that there exists $\alpha(P, \overline{\mathcal{L}})$ and $\beta(P, \overline{\mathcal{L}})$ such that

$$\mathbb{E}_{x\sim P}[\|\nabla\overline{\mathcal{L}}(\theta, x) - \nabla\mathcal{R}(\theta)\|_2^2] \le \alpha(P, \overline{\mathcal{L}})\|\theta - \theta^*\|_2^2 + \beta(P, \overline{\mathcal{L}}). \tag{3}$$

In other words, the variance of the ℓ_2-norm of the gradient distribution depends on a uniform constant $\beta(P, \overline{\mathcal{L}})$ and a position-dependent variable, $\alpha(P, \overline{\mathcal{L}})$, which allows the variance of gradient noise to be large when θ is far from the true parameter θ^*. We note that this is a more general assumption compared to prior works which assumed that the variance is uniformly bounded by σ^2. It can be seen that our condition is more nuanced and can be weaker: even if our condition holds, we would allow for a uniform bound on variance to be large: $\alpha(P, \overline{\mathcal{L}}) \sup_{\theta\in\Theta} \|\theta - \theta^*\|^2 + \beta(P, \overline{\mathcal{L}})$. Whereas, a uniform bound on the variance could always be cast as $\alpha(P, \overline{\mathcal{L}}) = 0$ and $\beta(P, \overline{\mathcal{L}}) = \sigma^2$. We will show that this more nuanced assumption is essential to obtain tight bounds for linear regression problems.

We next provide some running examples to instantiate the above:

1. **Mean estimation:** Given observations $x_1, \cdots, x_n \sim P$ where the distribution P with mean μ and a bounded covariance matrix Σ. The minimizer of the following square loss is the mean μ of distribution P:

$$\overline{\mathcal{L}}(\theta, x) = \frac{1}{2}\|x - \theta\|_2^2 \quad \text{and} \quad \mu = \operatorname*{argmin}_{\theta\in\mathbb{R}^p} \mathbb{E}_{x\sim P}[\overline{\mathcal{L}}(\theta, x)]. \tag{4}$$

   In this case, $\tau_\ell = \tau_u = 1$, $\alpha(P, \overline{\mathcal{L}}) = 0$ and $\beta(P, \overline{\mathcal{L}}) = \operatorname{trace}(\Sigma)$ satisfy the assumption in Eq.(3).

2. **Linear regression:** Given covariate-response pairs (x, y), where x, y are sampled from P and have a linear relationship, i.e. $y = \langle x, \theta^*\rangle + w$, where θ^* is the true parameter we want to estimate and w is drawn from a zero-mean distribution. Suppose that under distribution P the covariate $x \in \mathbb{R}^p$ have mean 0 and non-singular covariance matrix Σ. In this setting, we consider the squared loss:

$$\overline{\mathcal{L}}(\theta, (x, y)) = \frac{1}{2}(y - \langle x, \theta\rangle)^2, \text{ and } \quad \mathcal{R}(\theta) = \frac{1}{2}(\theta - \theta^*)^\top\Sigma(\theta - \theta^*). \tag{5}$$

   The true parameter θ^* is the minimizer of $\mathcal{R}(\theta)$. We also note that $\tau_\ell = \lambda_{\min}(\Sigma)$ and $\tau_u = \lambda_{\max}(\Sigma)$ satisfies the assumption in Eq.(2) with $\alpha(P, \overline{\mathcal{L}}) = O(p\|\Sigma\|_2^2)$, and $\beta(P, \overline{\mathcal{L}}) = p\sigma^2\|\Sigma\|_2$ satisfy the assumption in Eq.(3). Note that if we had to uniformly bound the variance of the gradients as in previous stochastic optimization work, that bound would need to scale as: $O(p\|\Sigma\|_2^2 R^2 + p\sigma^2\|\Sigma\|_2)$, where $R = \sup_{\theta\in\Theta} \|\theta - \theta^*\|$, which will yield much looser bounds.

3 Main Results

In this section, we introduce our clipped stochastic gradient descent algorithm. We begin by formally defining clipped stochastic gradients. For a clipping parameter $\lambda \ge 0$:

$$\operatorname{clip}(\nabla\overline{\mathcal{L}}(\theta, x), \lambda) = \min\left(1, \frac{\lambda}{\|\nabla\overline{\mathcal{L}}(\theta, x)\|_2}\right)\nabla\overline{\mathcal{L}}(\theta, x), \tag{6}$$

where $\theta \in \Theta$, $x \sim P$ and $\nabla\overline{\mathcal{L}}(\theta, x)$ is the stochastic gradient. The overall algorithm is summarized in Algorithm 1, where we use $\mathcal{P}_\Theta$ to denote the (Euclidean) projection onto the domain Θ.

Algorithm 1 Clipped stochastic gradient descent (`clipped-SGD`) algorithm.

Input: loss function $\overline{\mathcal{L}}$, initial point θ^1, step size η_t, clipping level λ, samples $x_1, \cdots, x_N \sim P$.

1: **for** $t = 1, 2, ..., N$ **do**

2: $\quad \theta^{t+1} \leftarrow \mathcal{P}_\Theta\left(\theta^t - \eta_t \text{clip}(\nabla\overline{\mathcal{L}}(\theta^t, x_t), \lambda)\right)$.

Output: θ^{N+1}.

Next, we state our main convergence result for `clipped-SGD` in Theorem 1.

Theorem 1. *(Streaming heavy-tailed stochastic optimization) Suppose that the population risk satisfies the regularity conditions in Eq.* (2) *and stochastic gradient noise satisfies the condition in Eq.* (3)*. Let* $\delta \in (0, 2e^{-1})$ *and*

$$\gamma \stackrel{\text{def}}{=} 144 \max\left\{\frac{\tau_u}{\tau_\ell}, \frac{96\alpha(P, \overline{\mathcal{L}})}{\tau_\ell^2}\right\} \log(2/\delta) + 1. \tag{7}$$

Given N *samples* $x_1, \cdots, x_N$*, the Algorithm 1 initialized at* θ^1 *with* $\eta_t \stackrel{\text{def}}{=} \frac{1}{\tau_\ell(t+\gamma)}$ *and*

$$\lambda \stackrel{\text{def}}{=} C_1 \sqrt{\frac{\tau_\ell^2 \gamma(\gamma-1)\|\theta^1 - \theta^*\|_2^2}{\log(2/\delta)^2} + \frac{(N+\gamma)\beta(P, \overline{\mathcal{L}})}{\log(2/\delta)}}, \tag{8}$$

where $C_1 \geq 1$ *is a scaling constant can be chosen by users, returns* θ^{N+1} *such that with probability at least* $1 - \delta$*, we have*

$$\|\theta^{N+1} - \theta^*\|_2 \leq 100 C_1 \left(\frac{\gamma\|\theta^1 - \theta^*\|_2}{N+\gamma} + \frac{1}{\tau_\ell}\sqrt{\frac{\beta(P, \overline{\mathcal{L}})\log(2/\delta)}{N+\gamma}}\right). \tag{9}$$

We explain our theoretical contribution and provide a proof sketch in Section 5. The complete proof can be found in Appendix E.

Remarks: a) This theorem says that with a properly chosen clipping level λ, `clipped-SGD` has an asymptotic convergence rate of $O(1/\sqrt{N})$ and enjoys *sub-Gaussian style concentration* around the true minimizer (alternatively, its high probability bound scales logarithmically in the confidence parameter δ). The first term in the error bound is related to the *initialization error* and the second term is governed by the *stochastic gradient noise*. These two terms have different convergence rates: at early iterations when the initialization error is large, the first term dominates but quickly decreases at the rate of $O(\gamma/N)$. At later iterations the second term dominates and decreases at the usual statistical rate of convergence of $O(1/\sqrt{N})$.

b) Note that $\eta_t = O(\frac{1}{\tau_\ell t})$ is a common choice for optimizing τ_ℓ-strongly convex functions [27, 48]. The only difference is that we add a delay parameter γ to "slow down" the learning rate η_t and to stablize the training process. The delay parameter γ depends on the position-dependent variance term $\alpha(P, \overline{\mathcal{L}})$ and the condition number τ_u/τ_ℓ. From a theoretical viewpoint, the delay parameter ensures that the true gradient is within the clipping region with high-probability, i.e. $\|\nabla\mathcal{R}(\theta^t)\|_2 \leq \lambda$ for $t = 1, ..., N$ with high probability and this in turn allows us to control the variance and the bias incurred by clipped gradients. Moreover, it controls the position-dependent variance term $\alpha(P, \overline{\mathcal{L}})\|\theta^t - \theta^*\|_2^2$, especially during the initial iterations when the error (and the variance of stochastic gradients) is large.

c) We choose the clipping level to be proportional to $\sqrt{N}$ to balance the variance and bias of clipped gradients. Roughly speaking, the bias is inversely proportional to the clipping level (Lemma 2). As the error $\|\theta^N - \theta^*\|_2$ converges at the rate $O(1/\sqrt{N})$, and this in turn suggests that we should choose the clipping level to be $O(\sqrt{N})$.

d) Note that previous stochastic optimization algorithms use $O(n)$ batch sizes for strongly convex objective [9, 22, 43, 47]. To address this issue, the critical ingredients are the use of $O(1/t)$ *decayed learning rate* and *a*

delayed parameter γ to prevent it from diverging. Also, we explicitly control the variance of the gradient noise by clipping the gradients, and are able to provide a more careful analysis. Whereas in previous algorithms, they use a constant step size throughout the training process. Consequently, when getting close to the minimizer, they must use *an exponential growing batch size* to reduce the gradient noise and to prevent oscillations.

We also provide an error bound and sample complexity where, as in prior work, we assume the variance of stochastic gradients are *uniformly bounded* by σ^2, i.e. $\alpha(P, \overline{\mathcal{L}}) = 0$ and $\beta(P, \overline{\mathcal{L}}) = \sigma^2$ (as before, we assume that the population loss $\mathcal{R}(\cdot)$ is strongly-convex and smooth). We have the following corollary:

Corollary 2. *Under the same assumptions and with the same hyper-parameters in Theorem 1 and letting* $C_1 = 1$*, with the probability at least* $1 - \delta$*, we have the following error bound:*

$$\mathcal{R}(\theta^{N+1}) - \mathcal{R}(\theta^*) \leq O\left(\frac{\tau_u^3}{\tau_\ell^3} \cdot \frac{r_0 \log(1/\delta)^2}{N^2} + \frac{\tau_u}{\tau_\ell^2} \cdot \frac{\sigma^2 \log(1/\delta)}{N}\right), \tag{10}$$

where $r_0 = \mathcal{R}(\theta^1) - \mathcal{R}(\theta^*)$ *is the initialization error. In other words, to achieve* $\mathcal{R}(\theta^{N+1}) - \mathcal{R}(\theta^*) \leq \epsilon$ *with probability at least* $1 - \delta$*, we need* $O\left(\max\left(\sqrt{\frac{\tau_u^3}{\tau_\ell^3} \cdot \frac{r_0}{\epsilon}}, \frac{\tau_u \sigma^2}{\tau_\ell^2 \epsilon}\right) \log\left(\frac{1}{\delta}\right)\right)$ *samples.*

With our general results in place we now turn our attention to deriving some important consequences for mean estimation and linear regression.

4 Consequences for Heavy-tailed Parameter Estimation

In this section, we investigate the consequences of Theorem 1 for statistical estimation in the presence of heavy-tailed noise. We plug in the respective loss functions $\overline{\mathcal{L}}$, the terms $\alpha(P, \overline{\mathcal{L}})$ and $\beta(P, \overline{\mathcal{L}})$ capturing the underlying stochastic gradient distribution, in Theorem 1 to obtain high-probability bounds for the respective statistical estimators.

4.1 Heavy-tailed Mean Estimation

We assume that the distribution P has bounded covariance matrix Σ. Then `clipped-SGD` for mean estimation has the following guarantee.

Corollary 3. *(Streaming Heavy-tailed Mean Estimation) Given samples* $x_1, \cdots, x_N \in \mathbb{R}^p$ *from a distribution* P *and confidence level* $\delta \in (0, 2e^{-1})$*, the Algorithm 1 instantiated with a loss function* $\overline{\mathcal{L}}(\theta, x) = \frac{1}{2}\|x - \theta\|_2^2$*, an initial point* $\theta^1 \in \mathbb{R}^p$*,* $\gamma = 144 \log(2/\delta) + 1$*, a learning rate* $\eta_t = \frac{1}{t+\gamma}$*, and a clipping level*

$$\lambda = C_1 \sqrt{\frac{\gamma(\gamma - 1)\|\theta^1 - \theta^*\|_2^2}{\log(2/\delta)^2} + \frac{(N + \gamma) trace\,(\Sigma)}{\log(2/\delta)}},$$

where $C_1 \geq 1$ *is a scaling constant can be chosen by users, returns* θ^{N+1} *such that with probability at least* $1 - \delta$*, we have*

$$\|\theta^{N+1} - \theta^*\|_2 \leq 100\left(\frac{\gamma\|\theta^1 - \theta^*\|_2}{N + \gamma} + \sqrt{\frac{trace\,(\Sigma)\log(2/\delta)}{N + \gamma}}\right). \tag{11}$$

Remarks: a) The proposed mean estimator matches the error bound of the well-known geometric-median-of-means estimator [42], achieving $\|\theta^N - \theta^*\|_2 \lesssim \sqrt{\frac{\mathrm{trace}(\Sigma)\log(1/\delta)}{N}}$. This guarantee is still sub-optimal compared to the optimal *sub-Gaussian rate* [39]. Existing polynomial time algorithms having optimal performance are for the batch setting and require either storing the entire dataset [31, 8, 12, 30, 7, 36, 13] or have $O(p \log(1/\delta))$ storage complexity [10]. On the other hand, we argue that trading off some statistical accuracy for a large savings in memory and computation is favorable in practice.

Moreover, we claim that these algorithms are hard to be implemented in the streaming setting: Hopkins [31], Cherapanamjeri et al. [8] use the semi-definite programming method, which is not yet practical. Algorithms Diakonikolas et al. [12], Hopkins et al. [30], Cheng et al. [7], Lei et al. [36], Dong et al. [13] rely on analyzing the spectrum of the covariance matrix. These approaches require polylogarithmic passes over data to remove potential outliers in d orthogonal directions, making it unsuitable in our setting since computing the covariance matrix already requires taking one pass over data.

b) One may argue that median-based approaches, e.g. coordinate-wise/geometric median-of-means or much simpler coordinate-wise/geometric medians, can be implemented in a streaming fashion while retaining the same rate as in the batch setting. Specifically, coordinate-wise/geometric median-of-means first divide the data into b buckets of roughly equal size, compute the mean in each bucket, and then takes the coordinate-wise/geometric median of these bucketed means. We argue that they are not favorable for the following reasons.

First of all, simply using coordinate-wise/geometric medians of these N samples is inconsistent. These medians are consistent estimators for the medians of underlying distributions. To use them in the mean estimation, it incurs a large bias when the underlying distribution is asymmetric. For instance, the distance between the mean and the median for a one-dimensional Pareto distribution is a constant.

Second, it's not obvious how to turn these algorithms into the steaming setting. Current analyses for geometric median are asymptotic [3] or only applied after a large number of iterations [4] because estimating streaming geometric median incurs a large bias in the early iterations. Moreover, in practice, it might require one to wait for a bucket of samples to calculate bucketed means, which is not an 'any-time' algorithm as our algorithm.

Coordinate-wise median-of-means has similar issues. Also, even in the batch setting, the confidence parameter delta in their guarantee gets scaled by dimension, which can not be applied to very high-dimensional spaces [42].

4.2 Heavy-tailed Linear Regression

We consider the linear regression model described in Eq.(5). Assume that the covariates $x \in \mathbb{R}^p$ have bounded 4^{th} moments and a non-singular covariance matrix Σ with bounded operator norm, and the noise w has bounded 2^{nd} moments. We denote the minimum and maximum eigenvalue of Σ by τ_ℓ and τ_u. More formally, we say a random variable $x \in \mathbb{R}^p$ has a bounded 4^{th} moment if there exists a constant C_4 such that for every unit vector $v \in \mathcal{S}^{p-1}$, we have

$$\mathbb{E}[\langle x - \mathbb{E}[x], v\rangle^4] \le C_4 \left(\mathbb{E}[\langle x - \mathbb{E}[x], v\rangle]\right)^2. \tag{12}$$

Corollary 4. *(Streaming Heavy-tailed Regression) Given samples $(x_1, y_1), \cdots, (x_N, y_N) \in \mathbb{R}^p \times \mathbb{R}$ and confidence level $\delta \in (0, 2e^{-1})$, the Algorithm 1 instantiated with loss function in Eq. (5), initial point $\theta^1 \in \mathbb{R}^p$,*

$$\gamma = 144 \max\left\{\frac{\tau_u}{\tau_\ell}, \frac{192(C_4+1)p\tau_u^2}{\tau_\ell^2}\right\}\log(2/\delta) + 1,$$

learning rate $\eta_t = \frac{1}{\tau_\ell(t+\gamma)}$, and clipping level

$$\lambda = C_1\sqrt{\frac{\tau_\ell^2\gamma(\gamma-1)\|\theta^1 - \theta^*\|_2^2}{\log(2/\delta)^2} + \frac{(N+\gamma)\sigma^2 p\tau_u}{\log(2/\delta)}},$$

where C_1 is a scaling constant can be chosen by users and C_4 is the constant in Eq.(12), returns θ^{N+1} such that with probability at least $1-\delta$, we have

$$\|\theta^{N+1} - \theta^*\|_2 \le 100C_1\left(\frac{\gamma\|\theta^1 - \theta^*\|_2}{N+\gamma} + \frac{\sigma}{\tau_\ell}\sqrt{\frac{p\tau_u\log(2/\delta)}{N+\gamma}}\right). \tag{13}$$

5 Proof Sketch of Theorem 1

In this section, we provide an overview of the arguments that constitute the proof of Theorem 1 and explain our theoretical contribution. The full details of the proof can be found in Section E in Appendix. Our proof is improved upon previous analysis of clipped-SGD with constant learning rate and $O(n)$ batch size [24] and high probability bounds for SGD with $O(1/t)$ step size in the sub-Gaussian setting [28]. Our analysis consists of three steps: (i) Expansion of the update rule. (ii) Selection of the clipping level. (iii) Concentration of bounded martingale sequences.

Notations: Recall that we use step size $\eta_t = \frac{1}{\tau_\ell(t+\gamma)}$. We will write $\epsilon_t = \nabla\mathcal{R}(\theta^t) - \text{clip}(\nabla\overline{\mathcal{L}}(\theta^t, x_t), \lambda)$ is the noise indicating the difference between the stochastic gradient and the true gradient at step t. Let $\mathcal{F}_t = \sigma(x_1, \cdots, x_t)$ be the σ-algebra generated by the first t steps of `clipped-SGD`. We note that clipping introduce bias so that $\mathbb{E}_{x_t}[\epsilon_t|\mathcal{F}_{t-1}]$ is no longer zero, so we decompose the noise term $\epsilon_t = \epsilon_t^b + \epsilon_t^v$ into a bias term ϵ_t^b and a zero-mean variance term ϵ_t^v, i.e.

$$\epsilon_t = \epsilon_t^b + \epsilon_t^v, \text{ where } \epsilon_t^b = \mathbb{E}_{x_t}[\epsilon_t|\mathcal{F}_{t-1}] \text{ and } \epsilon_t^v = \epsilon_t - \mathbb{E}_{x_t}[\epsilon_t|\mathcal{F}_{t-1}]$$

(i) Expansion of the update rule: We start with the following lemma that is pretty standard in the analysis of SGD for strongly convex functions. It can be obtained by unrolling the update rules $\theta^{t+1} = \mathcal{P}_\Theta\left(\theta^t - \eta_t \text{clip}(\nabla\overline{\mathcal{L}}(\theta^t, x_t), \lambda)\right)$ and using properties of τ_ℓ-strongly-convex and τ_u-smooth functions.

Lemma 1. *Under the conditions in Theorem 1, for any $1 \leq i \leq N$, we have*

$$\|\theta^{i+1} - \theta^*\|_2^2 \leq \underbrace{\frac{\gamma(\gamma-1)\|\theta^1 - \theta^*\|_2^2}{(i+\gamma)(i+\gamma-1)}}_{\textit{the initialization error}} + \underbrace{\frac{\sum_{t=1}^i (t+\gamma-1)\left\langle \epsilon_t^b + \epsilon_t^v, \theta^t - \theta^*\right\rangle}{\tau_\ell(i+\gamma)(i+\gamma-1)}}_{\textit{the first noise term}} + \underbrace{\frac{2\sum_{t=1}^i \left(\|\epsilon_t^v\|_2^2 + \|\epsilon_t^b\|_2^2\right)}{\tau_\ell^2(i+\gamma)(i+\gamma-1)}}_{\textit{the second noise term}}.$$

(ii) Selection of the clipping level: Now, to upper bound *the noise terms*, we need to choose the clipping level λ properly to balance the variance term ϵ_t^v and the bias term ϵ_t^b. Specifically, we use the inequalities of Gorbunov et al. [24], which provides us upper bounds for the magnitude and variance of these noise terms.

Lemma 2. *(Lemma F.5, [24]) For any $t = 1, 2, .., N$, we have*

$$\|\epsilon_t^v\|_2 \leq 2\lambda. \tag{14}$$

Moreover, for all $t = 1, 2, .., N$, assume that the variance of stochastic gradients is bounded by σ_t^2, i.e. $\mathbb{E}_{x_t}[\|\nabla\overline{\mathcal{L}}(\theta^t, x_t) - \nabla\mathcal{R}(\theta^t)\|_2^2|\mathcal{F}_{t-1}] \leq \sigma_t^2$ and assume that the norm of the true gradient is less than $\lambda/2$, i.e. $\|\nabla\mathcal{R}(\theta^t)\|_2 \leq \lambda/2$. Then we have

$$\|\epsilon_t^b\|_2 \leq \frac{4\sigma_t^2}{\lambda} \quad \textit{and} \quad \mathbb{E}_{x_t}[\|\epsilon_t^v\|_2^2|\mathcal{F}_{t-1}] \leq 18\sigma_t^2 \quad \textit{for all } t = 1, 2, ..., N. \tag{15}$$

This lemma gives us the dependencies between the variance, bias and clipping level: a larger clipping level leads to a smaller bias but the magnitude of the variance term $\|\epsilon_t^v\|_2$ is larger, while the variance of the variance term $\mathbb{E}_{x_t}[\|\epsilon_t^v\|_2^2|\mathcal{F}_{t-1}]$ remains constant. These three inequalities are also essential for us to use concentration inequalities for martingales. However, we highlight the *necessity* for these inequalities to hold: *the true gradient lies in the clipping region up to a constant*, i.e. $\|\nabla\mathcal{R}(\theta^t)\|_2 \leq \lambda/2$. This condition is necessary since without this, we could not have upper bounds of the bias and variance terms. Therefore, the clipping level should be chosen in a very accurate way. Below we informally describe how do we choose it.

We note that $\|\theta^t - \theta^*\|_2$ should converge with $1/\sqrt{t}$ rate for strongly convex functions with $O(1/t)$ step size [48, 28]. To make sure *the first noise term* upper bound by $O(1/i)$, one should expect each summand

$t\left\langle \epsilon_t^b, \theta^t - \theta^* \right\rangle = O(1)$ for $1 \le t \le i$, which implies $\|\epsilon_t^b\|_2 = O(1/\sqrt{t})$. This motivates us to choose the clipping level to be proportional to $\sqrt{t}$ by Eq.(15). Also, from the detailed proof in Section E, we will show that the delay parameter γ makes sure $\|\nabla \mathcal{R}(\theta^t)\|_2 \le \lambda/2$ holds with high probabilities and the position dependent noise $\alpha(P, \overline{\mathcal{L}})$ is controlled.

(iii) Concentration of bounded martingale sequences: A significant technical step of our analysis is the use of the following Freedman's inequality.

Lemma 3. *(Freedman's inequality [18]) Let $d_1, d_2, \cdots, d_T$ be a martingale difference sequence with a uniform bound b on the steps d_i. Let V_s denote the sum of conditional variances, i.e. $V_s = \sum_{i=1}^{s} \mathrm{Var}(d_i | d_1, \cdots, d_{i-1})$. Then, for every $a, v > 0$,*

$$\Pr\left(\sum_{i=1}^{s} d_i \ge a \quad \textit{and} \quad V_s \le v \quad \textit{for some} \quad s \le T \right) \le \exp\left(\frac{-a^2}{2(v + ba)} \right).$$

Freedman's inequality says that if we know the boundness and variance of the martingale difference sequence, the summation of them has *exponential concentration* around its expected value for *all subsequence* $\sum_{i=1}^{s} d_i$.

Now we turn our attention to the variance term in *the first noise term*, i.e. $\sum_{t=1}^{i} (t + \gamma - 1) \left\langle \epsilon_t^v, \theta^t - \theta^* \right\rangle$. It is the summation of a martingale difference sequence since $\mathbb{E}[\epsilon_t^v | \mathcal{F}_{t-1}] = 0$. Note that Lemma 2 has given us upper bounds for boundness/variance for ϵ_t^v. However, the main technical difficulty is that each summand involves the error of past sequences, i.e. $\|\theta^t - \theta^*\|_2$.

Our solution is the use of Freedman's inequality, which gives us a *loose control* of all past error terms with high probabilities, i.e. $\|\theta^t - \theta^*\|_2^2 \le O\left(N/t^2\right)$ for $1 \le t \le N$. On the contrary, a recurrences technique used in the past works [22, 23, 24] uses an increasing clipping levels λ_t and calls the Bernstein inequality (, which only provides an upper bound for the *entire sequence*,) N times in order to control $\|\theta^t - \theta^*\|_2$ for every t. As a result, it incurs an extra factor $\log(N)$ on their bound since it imposes a too strong control over past error terms.

Finally, we describe why `clipped-SGD` allows a $O(1)$ batch size at a high level. For previous works of stochastic optimization with strongly convex and smooth objective, they use a constant step size throughout their training process [22, 47, 9, 43]. However, to ensure their approach make a constant progress at each iteration, they should use *an exponential growing batch size* to reduce variance of gradients. Whereas in our approach, we explicitly control the variance by using *a decayed learning rate and clipping the gradients*. Therefore, we are able to provide a careful analysis of the resulted bounded martingale sequences.

6 Experiments

To corroborate our theoretical findings, we conduct experiments on mean estimation and linear regression to study the performance of `clipped-SGD`. Another experiment on logistic regression is presented in the Section D.3 in the Appendix.

Methods. We compare `clipped-SGD` with vanilla SGD, which takes stochastic gradient to update without a clipping function. For linear regression, we also compare `clipped-SGD` with Lasso regression [44], Ridge regression and Huber regression [50, 33]. All methods use the same step size $\frac{1}{t+\gamma}$ at step t.

To simulate heavy-tailed samples, we draw from a scaled standardized Pareto distribution with tail-parameter β for which the k^{th} moment only exists for $k < \beta$. The smaller the β, the more heavy-tailed the distribution. Due to space constraints, we defer other results with different setups to the Appendix.

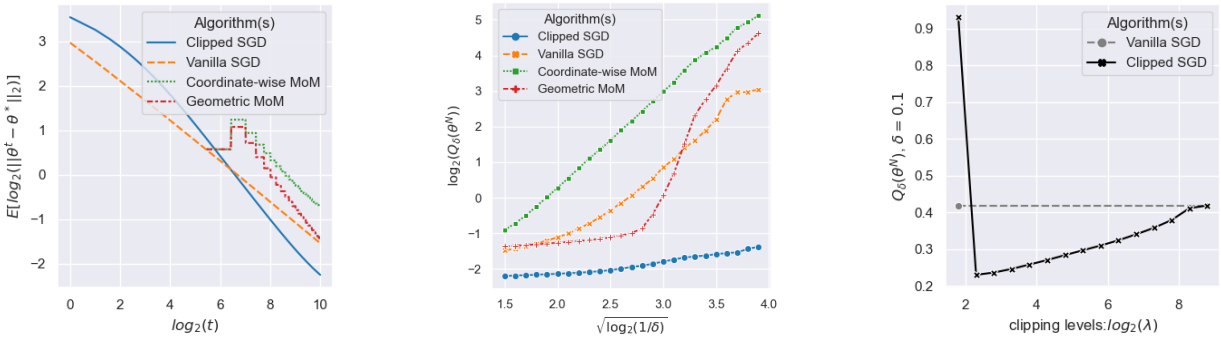


Figure 2: Results for robust mean estimation for $N = 1024$ and $p = 256$. Smaller $Q_\delta(\widehat{\theta})$ is better.

Choice of Hyper-parameter. Note that in Theorem 1, the clipping level λ depends on the initialization error, i.e. $\|\theta^1 - \theta^*\|_2$, which is not known in advance. Moreover, we found that the suggested value of γ has a too large of a constant factor and substantially decreases the convergence rate especially for small N. However, standard hyper-parameter selection techniques such as cross-validation and hold-out validation require storing the entire validation set, which are not suitable for streaming settings.

Consequently, we use a sequential validation method [34], where we do "training" and "evaluation" on the last q percent of the data. Formally, given candidate solutions $\{\hat{\theta}_1, \cdots, \hat{\theta}_m\}$ trained from samples $x_1, \cdots, x_N$, let θ_i^t be the estimated parameter for candidate i at iteration t. Then we choose the candidate that minimizes the empirical mean of the risk of the last q percents of samples, i.e.

$$j^* = \operatorname*{argmin}_{1 \le j \le m} \frac{1}{qN} \sum_{t=(1-q)N+1}^{N} \overline{\mathcal{L}}(\widehat{\theta}_j^t, x_t) \tag{16}$$

Specifically, at the last q percents of iterations, when a sample x_t comes, we first calculate the risk induced by θ_j^t and x_t and then use this sample to update the parameter θ_j^t. Therefore, instead of storing the entire validation set, we only need $O(mp)$ space to store the candidate parameters $\{\hat{\theta}_1, \cdots, \hat{\theta}_m\}$ and the validation losses of the candidates.

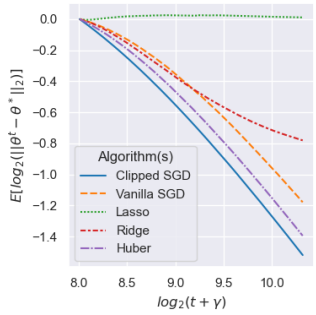
In our experiment, we choose $q = 0.2$ to tune the delay parameter γ, the clipping level λ, regularization factors for Lasso, Ridge and Huber regression, and the step size for streaming coordinate-wise/geometric median-of-means.

Metric. For any estimator $\widehat{\theta}$, we use the ℓ_2 loss $\ell(\widehat{\theta}) = \|\widehat{\theta} - \theta^*\|_2$ as our primary metric. To measure the tail performance of estimators (not just their expected loss), we also use $Q_\delta(\widehat{\theta}) = \inf(\alpha : \Pr(\ell(\widehat{\theta}) > \alpha) \le \delta)$, which is the bound on the loss that we wish to hold with probability $1 - \delta$. This could also be viewed as the $100(1-\delta)$ percentile of the loss (e.g. if $\delta = 0.05$, then this would be the 95th percentile of the loss).

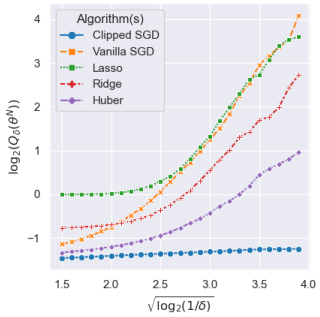
6.1 Synthetic Experiments: Mean estimation

Setup. We obtain samples $\{x_i\}_{i=1}^N \subseteq \mathbb{R}^p$ from a scaled standardized Pareto distribution with tail-parameter $\beta = 2.1$. We initialize the mean parameter estimate as $\theta^1 = [1, \cdots, 1] \in \mathbb{R}^p$, $\gamma = 0$, and fix the step size to $\eta_t = 1/t$ for `clipped-SGD` and Vanilla SGD. We note that in this setting, it can be seen that $\widehat{\theta}_{SGD}^t = \sum_{i=1}^t z_i/t$ is the empirical mean over t samples by some simple algebra.

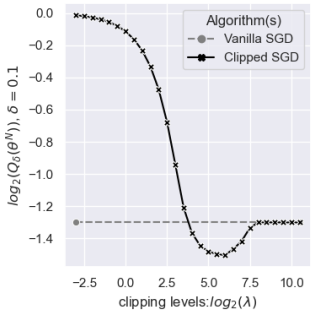
We also compare our approach with streaming coordinate-wise/geometric median-of-means(MoM), where we use the number of buckets $b = \lceil \log(1/\delta) \rceil$ with $\delta = 0.05$ as in the batch setting [39] and a step size of $\eta_t = \frac{c}{t_b}$



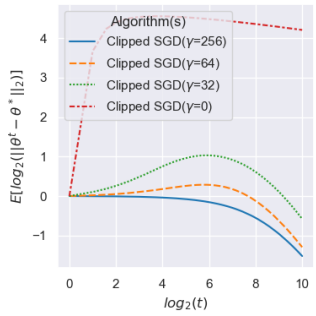
(a) Expected convergence curve.



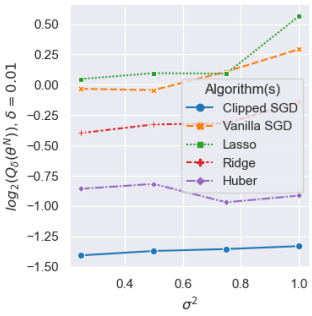
(b) Last iteration quantile loss.



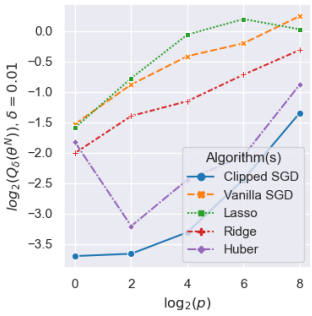
(c) $\log_2(Q_\delta(\theta^N))$ v.s. λ.



(d) Expected convergence curve for different γ.



(e) $\log_2(Q_\delta(\theta^N))$ v.s. σ^2 for $\delta = 0.01$.



(f) $\log_2(Q_\delta(\theta^N))$ v.s. p for $\delta = 0.01$ and $N = 1024$.

Figure 3: Results for robust linear regression for $N = 1024, p = 256$, $\gamma = 256$ and $\sigma^2 = 0.75$.

[4], where c is a constant selected by the validation method and t_b is a counter for bucketed means. Specifically, we first wait for $\lfloor \frac{n}{b} \rfloor$ points and calculate their running mean as a initial point. Then we calculate bucketed means for the remaining points in the same way and use them to update the coordinate-wise/geometric median with an averaged stochastic gradient algorithm. Each metric is reported over 50000 trials. See more implementation details in Appendix C.

Results. Figure 2 shows the performance of all algorithms. In the first panel, we can see `clipped-SGD` has *expected* error that converges as $O(1/N)$ as our theory indicates. Also, the 99.9 percent quantile loss of Vanilla SGD is over 10 times worse than the expected error as in the second panel while the tail performance of `clipped-SGD` is similar to its expected performance. Moreover, streaming coordinate-wise/geometric median-of-means have a much worse expected performance and their tail performance is not well-controlled.

In the third panel, we can see that `clipped-SGD` performs better across different λ. When the clipping level λ is too small, it takes too small a step size so that the final error is high. While if we use a very large clipping level, the performance of `clipped-SGD` is similar to Vanilla SGD.

6.2 Synthetic Experiments: Linear Regression

Setup We generate covariate $x \in \mathbb{R}^p$ from an scaled standardized Pareto distribution with tail-parameter $\beta = 4.1$. The true regression parameter is set to be $\theta^* = [\frac{1}{\sqrt{p}}, \cdots, \frac{1}{\sqrt{p}}] \in \mathbb{R}^p$ and the initial parameter is set to $\theta^1 = [0, 0, \cdots, 0]$. The response is generated by $y = \langle x, \theta^* \rangle + w$, where w is sampled from scaled rescaled Pareto distribution with a zero mean, a variance σ^2 and a tail-parameter $\beta = 2.1$. We select $\tau_\ell = \lambda_{min}(\Sigma) = 1$.

Each metric is reported over 50000 trials.

Results We note that in this experiment, our hyperparameter selection technique yields $\gamma = 256$. Figure 3(a), 3(b) show that `clipped-SGD` performs the best among all baselines in terms of average error and quantile errors across different probability levels δ. Also, $\sqrt{\log_2(1/\delta)}$ has linear relation to $Q_\delta(\widehat{\theta})$ as Corollary 4 indicates. In Figure 3(c), we plot quantile loss against different clipping levels. It shows a similar trend to the mean estimation.

Next, in Figure 3(d), we plot the averaged convergence curve for different delay parameters γ. This shows that it is necessary to use a large enough γ to prevent it from diverging. This phenomenon can also be observed for different baseline models. Figure 3(e), 3(f) shows that `clipped-SGD` performs the best across different noise level σ^2 and different dimension p.

7 Conclusion and Future Direction

In this paper, we provide a streaming algorithm for statistical estimation under heavy-tailed distribution. In particular, we close the gap in theory of clipped stochastic gradient descent with heavy-tailed noise. We show that `clipped-SGD` can not only be used in parameter estimation tasks, such as mean estimation and linear regression, but also a more general stochastic optimization problem with heavy-tailed noise. There are several avenues for future work, including a better understanding of `clipped-SGD` under different distributions, such as having higher bounded moments or symmetric distributions, where the clipping technique incurs less bias. Finally, it would also be of interest to extend our results to different robustness setting such as Huber's ϵ-contamination model [33], where there are a constant portion of arbitrary outliers in observed samples.

Acknowledgements

We acknowledge the support of ARL, NSF via OAC-1934584, and DARPA via HR00112020006.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

[3] Hervé Cardot, Peggy Cénac, and Pierre-André Zitt. Efficient and fast estimation of the geometric median in hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19(1):18–43, 2013.

[4] 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.

[5] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

[6] Xiangyi Chen, Steven Z Wu, and Mingyi Hong. Understanding gradient clipping in private sgd: A geometric perspective. *Advances in Neural Information Processing Systems*, 33, 2020.

[7] Yu Cheng, Ilias Diakonikolas, Rong Ge, and Mahdi Soltanolkotabi. High-dimensional robust mean estimation via gradient descent. In *International Conference on Machine Learning*, pages 1768–1778. PMLR, 2020.

[8] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L Bartlett. Fast mean estimation with sub-gaussian rates. In *Conference on Learning Theory*, pages 786–806. PMLR, 2019.

[9] Damek Davis, Dmitriy Drusvyatskiy, Lin Xiao, and Junyu Zhang. From low probability to high confidence in stochastic convex optimization. *Journal of Machine Learning Research*, 22(49):1–38, 2021.

[10] Jules Depersin and Guillaume Lecué. Robust subgaussian estimation of a mean vector in nearly linear time. *Annals of Statistics*, 2019.

[11] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pages 999–1008. PMLR, 2017.

[12] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2): 742–864, 2019.

[13] Yihe Dong, Samuel Hopkins, and Jerry Li. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. *Advances in Neural Information Processing Systems*, 32, 2019.

[14] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[15] Pavel Dvurechensky and Alexander Gasnikov. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. *Journal of Optimization Theory and Applications*, 171(1):121–145, 2016.

[16] Jianqing Fan, Quefeng Li, and Yuyan Wang. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 79(1):247, 2017.

[17] Dima Feldman and Yuval Shavitt. An optimal median calculation algorithm for estimating internet link delays from active measurements. In *2007 Workshop on End-to-End Monitoring Techniques and Services*, pages 1–7. IEEE, 2007.

[18] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[19] Virgile Fritsch, Benoit Da Mota, Eva Loth, Gaël Varoquaux, Tobias Banaschewski, Gareth J Barker, Arun LW Bokde, Rüdiger Brühl, Brigitte Butzek, Patricia Conrod, et al. Robust regression for large-scale neuroimaging studies. *Neuroimage*, 111:431–441, 2015.

[20] Saurabh Garg, Joshua Zhanson, Emilio Parisotto, Adarsh Prasad, Zico Kolter, Zachary Lipton, Sivaraman Balakrishnan, Ruslan Salakhutdinov, and Pradeep Ravikumar. On proximal policy optimization's heavy-tailed gradients. In *International Conference on Machine Learning*, pages 3610–3619. PMLR, 2021.

[21] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4): 2061–2089, 2013.

[22] Eduard Gorbunov, Pavel Dvurechensky, and Alexander Gasnikov. An accelerated method for derivative-free smooth stochastic convex optimization. *arXiv preprint arXiv:1802.09022*, 2018.

[23] Eduard Gorbunov, Darina Dvinskikh, and Alexander Gasnikov. Optimal decentralized distributed algorithms for stochastic convex optimization. *arXiv preprint arXiv:1911.07363*, 2019.

[24] Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *arXiv preprint arXiv:2005.10785*, 2020.

[25] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR, 2019.

[26] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

[27] Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613. PMLR, 2019.

[28] Nicholas JA Harvey, Christopher Liaw, and Sikander Randhawa. Simple and optimal high-probability bounds for strongly-convex stochastic gradient descent. *arXiv preprint arXiv:1909.00843*, 2019.

[29] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

[30] Sam Hopkins, Jerry Li, and Fred Zhang. Robust and heavy-tailed mean estimation made simple, via regret minimization. *Advances in Neural Information Processing Systems*, 33:11902–11912, 2020.

[31] Samuel B Hopkins. Mean estimation with sub-gaussian rates in polynomial time. *Annals of Statistics*, 48(2): 1193–1213, 2020.

[32] Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016.

[33] Peter J Huber et al. Robust regression: asymptotics, conjectures and monte carlo. *Annals of statistics*, 1(5): 799–821, 1973.

[34] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

[35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[36] Zhixian Lei, Kyle Luh, Prayaag Venkat, and Fred Zhang. A fast spectral algorithm for mean estimation with sub-gaussian rates. In *Conference on Learning Theory*, pages 2598–2612. PMLR, 2020.

[37] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 4313–4324. PMLR, 2020.

[38] Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 983–992. PMLR, 2019.

[39] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

[40] Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 22(3):925–965, 2019.

[41] H Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, et al. Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1222–1230, 2013.

[42] Stanislav Minsker et al. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

[43] 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.

[44] Nam H Nguyen and Trac D Tran. Robust lasso with missing and grossly corrupted observations. *IEEE transactions on information theory*, 59(4):2036–2058, 2012.

[45] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. Understanding the exploding gradient problem. *CoRR, abs/1211.5063*, 2(417):1, 2012.

[46] Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint arXiv:2009.12976*, 2020.

[47] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society Series B (JRSSB)*, 2020.

[48] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

[49] Arun Sai Suggala, Kush Bhatia, Pradeep Ravikumar, and Prateek Jain. Adaptive hard thresholding for near-optimal consistent robust regression. In *Conference on Learning Theory*, pages 2892–2897. PMLR, 2019.

[50] Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

[51] H Victor et al. A general class of exponential inequalities for martingales and ratios. *Annals of probability*, 27(1): 537–564, 1999.

[52] Yang You, Igor Gitman, and Boris Ginsburg. Scaling sgd batch size to 32k for imagenet training. *arXiv preprint arXiv:1708.03888*, 6:12, 2017.

[53] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

[54] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

A Organization

The Appendices contain additional technical content and are organized as follows. In Appendix B, we provide additional details for related work, which contain a detailed comparison of previous work on stochastic optimization. In Appendix C, we detail hyperparameters used for experiments in Section 6. In Appendix D, we present supplementary experimental results for different setups for heavy-tailed mean estimation and linear regression. Additionally, we show a synthetic experiment on logistic regression. Finally, in Appendix E and F, we give the proofs for Theorem 1 and corollaries respectively.

B Related work : Additional details

Heavy-tailed stochastic optimization. In this paragraph, we present a detailed comparison of existing results of stochastic optimization. In Table 1, we compare existing high probability bounds of stochastic optimization for strongly convex and smooth objectives.

Since our work focus on large-scale setting (where we need to access data in a streaming fashion), we assume the number of samples N is large so that the required ϵ is small. In such setting, $O(\frac{1}{\epsilon})$ is the dominating term and the error is driven by the stochastic noise term σ^2. If ignoring the difference in logarithmic factors and assuming τ_u/τ_ℓ is small, all methods for heavy-tailed noise in Table 1 achieve $O(\frac{\sigma^2}{\tau_\ell \epsilon})\log(\frac{1}{\delta})$ and are comparable to algorithms derived under the sub-Gaussian noise assumption.

However, we can see that all of the existing methods require $O(\frac{1}{\epsilon})$ batch size except ours. Their batch sizes are not constants because they use a constant step size throughout their training process. Although they can achieve linear convergence rates for initialization error r_0, they should use *an exponential growing batch size* to reduce variance induced by gradient noise. Additionally, in large scale setting where the noise term is the dominating term, the linear convergence of initial error is not important. On the contrary, we choose a $O(\frac{1}{T})$ step size, which is widely used in stochastic optimization for strongly convex objective [48, 28]. Our proposed `clipped-SGD` can therefore enjoy $O(\frac{1}{T})$ convergence rate while using a constant batch size.

Finally, our analysis improves the dependency on the confidence level term $\log(\frac{1}{\delta})$: it does not have extra logarithmic terms and does not depend on ϵ. Although our bound has a worse dependency on the condition number τ_u/τ_ℓ, we argue that our bound has an extra τ_u/τ_ℓ term because our bound is derived under the square error, i.e. $\|\theta^t - \theta^*\|_2^2$ instead of the difference between objective values $\mathcal{R}(\theta^t) - \mathcal{R}(\theta^*)$. As a result, we believe the dependency on τ_u/τ_ℓ of our bounds can be improved by slightly revising Lemma 5.

Gradient clipping. Gradient clipping is a well-known optimization technique for both convex/non-convex optimization problems [26, 52, 45]. It has been shown to accelerate neural network training [53], stablize the policy gradient algorithms [20] and design different private optimization algorithms [6, 1]. Gradient clipping has also been shown to be robust to label noise [37].

Method	Sample complexity	Batch size
Sub-Gaussian noise		
`SIGMA2` [15]	$O\left(\max\left(\frac{\tau_u}{\tau_\ell}\log(\frac{r_0}{\epsilon}), \frac{\sigma^2}{\tau_\ell\epsilon}\log\left(\frac{1}{\delta}\log(\frac{r_0}{\epsilon})\right)\right)\right)$	$O\left(\frac{1}{\epsilon}\cdot\frac{\sigma^2}{\tau_\ell}\log\left(\frac{1}{\delta}\log(\frac{r_0}{\epsilon})\right)\right)$
`MS-AC-SA` [21]	$O\left(\max\left(\sqrt{\frac{\tau_u}{\tau_\ell}}\log(\frac{\tau_u r_0}{\tau_\ell\epsilon}), \frac{\sigma^2}{\tau_\ell\epsilon}\log\left(\frac{1}{\delta}\log(\frac{\tau_u r_0}{\tau_\ell\epsilon})\right)\right)\right)$	$O\left(\frac{1}{\epsilon}\cdot\frac{\sigma^2}{\tau_\ell\epsilon}\log\left(\frac{1}{\delta}\log(\frac{\tau_u r_0}{\tau_\ell\epsilon})\right)\right)$
Heavy-tailed noise		
`restarted-RSMD` [43]	$O\left(\max\left(\frac{\tau_u}{\tau_\ell}\log\left(\frac{\tau_\ell R^2}{\epsilon}\right), \frac{\sigma^2}{\tau_\ell\epsilon}\right)\cdot C\right)$, where $C=\log\left(\frac{1}{\delta}\log(\frac{\tau_\ell R^2}{\epsilon})\right)$	$O\left(\frac{1}{\epsilon}\cdot\frac{\sigma^2}{\tau_\ell}\log\left(\frac{\log(\frac{\tau_\ell R^2}{\epsilon})}{\delta}\right)\right)$
`proxBoost` [9]	$O\left(\max\left(\sqrt{\frac{\tau_u}{\tau_\ell}}\log\left(\frac{\tau_u^2 r_0^2\log(\frac{\tau_u}{\tau_\ell})}{\tau_\ell\epsilon}\right), \frac{\sigma^2\log\left(\frac{\tau_u}{\tau_\ell}\right)}{\tau_\ell\epsilon}\right)C'\right)$, where $C'=\log(\frac{\tau_u}{\tau_\ell})\log\left(\frac{1}{\delta}\cdot\log(\frac{\tau_u}{\tau_\ell})\right)$	$O\left(\frac{1}{\epsilon}\cdot\frac{\sigma^2\log\left(\frac{\tau_u}{\tau_\ell}\right)}{\tau_\ell}C'\right)$
`RGD` [47]	$O\left(\frac{\phi\tau_u\sigma^2}{\epsilon}\log\left(\frac{\phi}{\delta}\right)\right)$ with $\phi=\log(\frac{\tau_u r_0}{\tau_\ell\epsilon})/\log\left(\frac{\tau_u-\tau_\ell}{\tau_u+\tau_\ell}\right)$	$O\left(\frac{1}{\epsilon}\cdot\tau_u\sigma^2\log\left(\frac{\phi}{\delta}\right)\right)$
`clipped-SGD` (constant step size) [24]	$O\left(\max\left(\frac{\tau_u}{\tau_\ell}, \frac{\tau_u\sigma^2}{\tau_\ell^2\epsilon}\right)\cdot C\right)$, where $C=\log\left(\frac{r_0}{\epsilon}\right)\log\left(\frac{\tau_u}{\tau_\ell\delta}\cdot\log(\frac{r_0}{\epsilon})\right)$	$O\left(\frac{1}{\epsilon}\cdot\frac{\sigma^2\tau_u}{\tau_\ell^2}\log(\frac{r_0}{\epsilon})\cdot\log(\frac{\tau_u}{\delta\tau_\ell}\log(\frac{r_0}{\epsilon}))\right)$
`R-clipped- SGD` [24]	$O\left(\max\left(\frac{\tau_u}{\tau_\ell}\log(\frac{r_0}{\epsilon}), \frac{\sigma^2}{\tau_\ell\epsilon}\right)\cdot C\right)$, where $C=\log\left(\frac{\tau_u}{\tau_\ell\delta}\right)+\log\left(\log(\frac{r_0}{\epsilon})\right)$	$O\left(\frac{1}{\epsilon}\cdot\frac{\sigma^2}{\tau_\ell}\log(\frac{\tau_u}{\delta\tau_\ell}\log(\frac{r_0}{\epsilon}))\right)$
`R-clipped- SSTM` [24]	$O\left(\max\left(\sqrt{\frac{\tau_u}{\tau_\ell}}\log(\frac{r_0}{\epsilon}), \frac{\sigma^2}{\tau_\ell\epsilon}\right)\cdot C\right)$, where $C=\log\left(\frac{\tau_u}{\tau_\ell\delta}\right)+\log\left(\log(\frac{r_0}{\epsilon})\right)$	$O\left(\frac{1}{\epsilon}\cdot\frac{\sigma^2}{\tau_\ell}\log(\frac{\tau_u}{\delta\tau_\ell}\log(\frac{r_0}{\epsilon}))\right)$
`clipped-SGD` ($O(1/T)$ step size) [This work]	$O\left(\max\left(\sqrt{\frac{\tau_u^3}{\tau_\ell^3}\cdot\frac{r_0}{\epsilon}}, \frac{\tau_u\sigma^2}{\tau_\ell^2\epsilon}\right)\log\left(\frac{1}{\delta}\right)\right)$	$O(1)$

Table 1: Comparison of existing high probability upper bound for stochastic optimization for any τ_ℓ-strongly convex and τ_u-smooth objective function $\mathcal{R}(\cdot)$ with sub-Gaussian/heavy-tailed noise. The second column provides number of samples needed to achieved an ϵ-approximated solution $\hat{\theta}$ such that $\mathcal{R}(\hat{\theta})-\mathcal{R}(\theta^*)<\epsilon$ with probability at least $1-\delta$. In this table, we assume a gradient distribution has a uniformly bounded variance σ^2, i.e. $\alpha(P,\overline{\mathcal{L}})=0$ and $\beta(P,\overline{\mathcal{L}})=\sigma^2$ in Eq.(3). We use $r_0=\mathcal{R}(\theta^0)-\mathcal{R}(\theta^*)$. For `RSMD`, R is the diameter of the domain where the optimization problem is defined. The third column indicates the batch size, which is the number of samples used in a single step.

C Experimental Details

C.1 Experimental Details of Figure 1

In Figure 1, we consider the mean estimation task with a loss function $\mathcal{R}(\theta) = \mathbb{E}_x\left[\|x-\theta\|_2^2\right]$, where x is either from a sub-Gaussian distribution or a Pareto distribution with tail parameter 2.1. Both distributions are 10-dimensional and have zero mean and an identity covariance matrix. We use $N = 100$ samples and run $100,000$ trials to estimate each confidence level δ. We choose the clipping level to be $\lambda = 1.5$.

C.2 Experimental Details of Mean Estimation

In this section, we describe the algorithms of streaming coordinate-wise/geometric median-of-means and present details of the synthetic experiment of mean estimation in Section 6.1 and Appendix D.1.

C.2.1 Streaming coordinate-wise/geometric median-of-means algorithms

Given points $z_1, \cdots, z_n \in \mathbb{R}^d$, coordinate-wise/geometric medians of these n points are defined as the minimizers of the following convex objective.

$$\text{coordinate-wise median: } m_c \stackrel{\text{def}}{=} \operatorname*{argmin}_{u\in\mathbb{R}^d} \sum_{i=1}^n \|z_i - u\|_1. \tag{17}$$

$$\text{geometric median: } m_g \stackrel{\text{def}}{=} \operatorname*{argmin}_{u\in\mathbb{R}^d} \sum_{i=1}^n \|z_i - u\|_2. \tag{18}$$

These objectives are convex and therefore can be minimized via stochastic gradient descent [3, 4, 17]. In the experiment, we use the step sizes of $\frac{c}{t_b}$, where c is a constant selected by using our sequential validation method and t_b is the number of steps. In Algorithm 2 and 3 ,we describe the streaming coordinate-wise/geometric median-of-means algorithm.

Algorithm 2 Streaming coordinate-wise median-of-means algorithm.

Input: step size η_{t_b}, number of buckets b, samples $x_1, \cdots, x_N \sim P$.

Streaming calculate the bucketed mean $\theta^1 = \text{Average}(x_1, \cdots, x_b)$.
for $t_b = 1, 2, ..., \lfloor \frac{N}{b} \rfloor$ **do**
  Streaming calculate the bucketed mean $\bar{z}_{t_b} = \text{Average}(x_{bt_b+1}, \cdots, x_{b(t_b+1)})$.
  $\theta^{t_b+1} \leftarrow \theta^{t_b} - \eta_{t_b}\text{sign}(\theta^{t_b} - \bar{z}_{t_b})$.

Output: $\theta^{\lfloor \frac{N}{b} \rfloor+1}$.

Algorithm 3 Streaming geometric median-of-means algorithm.

Input: initial point θ^1, step size η_{t_b}, number of buckets b, samples $x_1, \cdots, x_N \sim P$.

Streaming calculate the bucketed mean $\theta^1 = \text{Average}(x_1, \cdots, x_b)$.
for $t_b = 1, 2, ..., \lfloor \frac{N}{b} \rfloor$ **do**
  Streaming calculate the bucketed mean $\bar{z}_{t_b} = \text{Average}(x_{bt_b+1}, \cdots, x_{b(t_b+1)})$.
  $\theta^{t_b+1} \leftarrow \theta^{t_b} - \eta_{t_b}\frac{\theta^{t_b} - \bar{z}_{t_b}}{\|\theta^{t_b} - \bar{z}_{t_b}\|_2}$.

Output: $\theta^{\lfloor \frac{N}{b} \rfloor+1}$.

C.2.2 Hyperparameter Selection

For each setup (N, p), we use the hyper-parameter selection technique described in Section 6 and choose the hyperparameters from the candidate sets in Table 2.

Hyper-parameters	Candidate sets
clipping level λ	$\{c\sqrt{Np} \mid c \in \{0.01, 0.06, \cdots, 1.01\}\}$
Number of buckets for streaming geometric/coordinate-wise median-of-means	$\lceil 8\log(1/\delta)\rceil$ with $\delta = 0.05$ [39]
Step sizes for streaming geometric/coordinate-wise median-of-means	$\{\frac{c}{t_b} \mid c \in \{10^{-1}, 10^{-\frac{3}{4}}, 10^{-\frac{1}{2}}, 10^{-\frac{1}{4}}, 1, 10^{\frac{1}{4}}, 10^{\frac{1}{2}}, 10^{\frac{3}{4}}\}\}$

Table 2: Candidate set for different hyper-parameters for mean estimation.

C.3 Experimental Details of Linear Regression

In this section, we present details of the synthetic experiment of linear regression in Section 6.2 and D.2. For each setup (N, p), we use the hyper-parameter selection technique described in Section 6 and choose the hyperparameters from the candidate sets in Table 3.

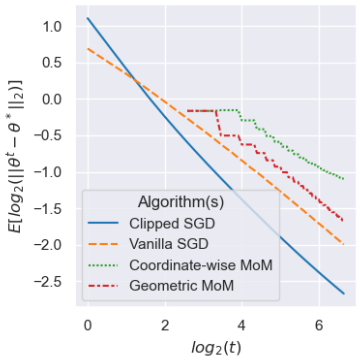
Hyper-parameters	Candidate sets
delay parameter γ	$\{0.1p, p, 10p\}$
clipping level λ	$\{c\sqrt{Np} \mid c \in \{0.01, 0.06, \cdots, 1.01\}\}$
regularization parameter for Lasso	$\{c\sqrt{Np} \mid c \in \{0.001, 0.006, \cdots, 0.101\}\}$
regularization parameter for Ridge	$\{c\sqrt{Np} \mid c \in \{0.001, 0.006, \cdots, 0.101\}\}$
regularization parameter for Huber	$\{c\sqrt{Np} \mid c \in \{0.001, 0.006, \cdots, 0.101\}\}$

Table 3: Candidate sets for different hyper-parameters for linear regression.

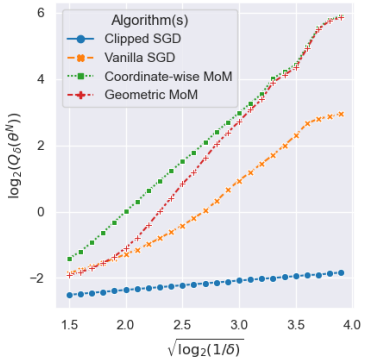
D Extra experiments

D.1 Extra Experiments on Mean Estimation

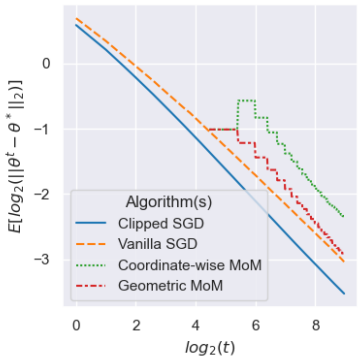
Figure 4 shows extra experimental results for $p = 20$ and $N = 100, 500, 1000$. The experimental setting is the same as in Section 6.1. We can see `clipped-SGD` consistently outperforms all other baselines in expected performance and tail performance.



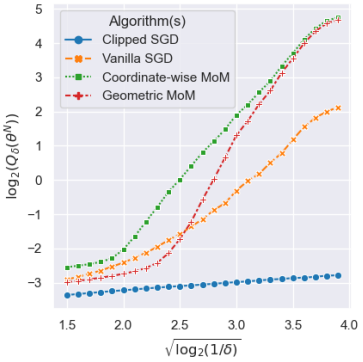
(a) Expected convergence curve for $N = 100$ and $p = 20$.



(b) Last iteration quantile loss for $N = 100$ and $p = 20$.

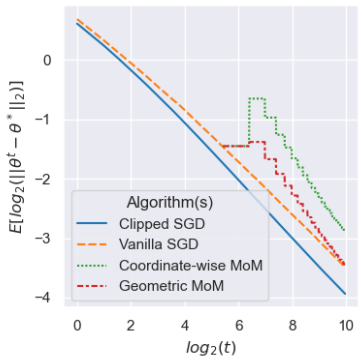


(c) Expected convergence curve for $N = 500$ and $p = 20$.

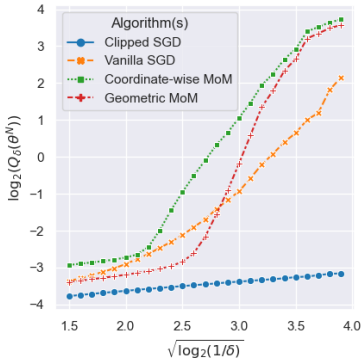


(d) Last iteration quantile loss for $N = 500$ and $p = 20$.

hfill



(e) Expected convergence curve for $N = 1000$ and $p = 20$.

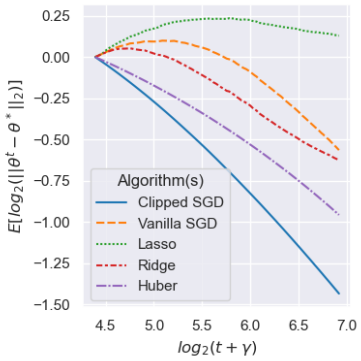


(f) Last iteration quantile loss for $N = 1000$ and $p = 20$.

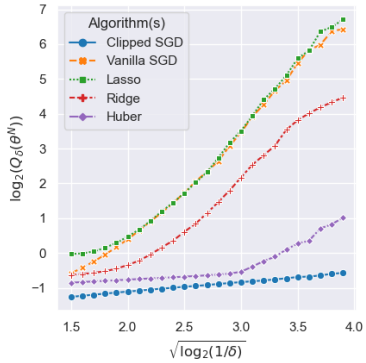
Figure 4: Results for robust mean estimation for different settings.

D.2 Extra Experiments on Linear Regression

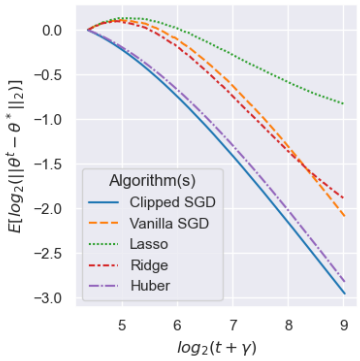
Figure 5 shows extra experimental results for $p = 20$ and $N = 100, 500, 1000$. The experimental setting is the same as in Section 6.2. We can see `clipped-SGD` consistently outperforms other baselines.



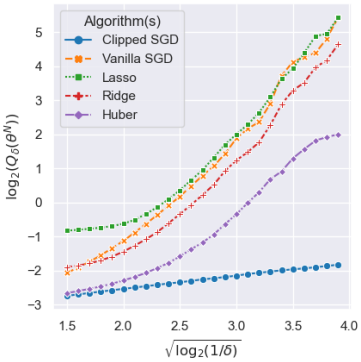
(a) Expected convergence curve for $N = 100$ and $p = 20$.



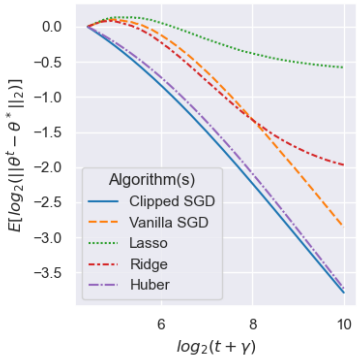
(b) Last iteration quantile loss for $N = 100$ and $p = 20$.



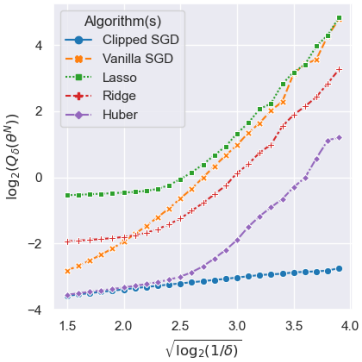
(c) Expected convergence curve for $N = 500$ and $p = 20$.



(d) Last iteration quantile loss for $N = 500$ and $p = 20$.



(e) Expected convergence curve for $N = 1000$ and $p = 20$.

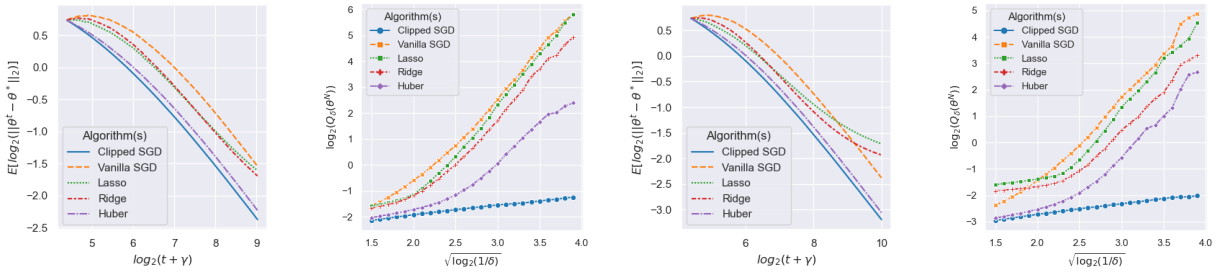


(f) Last iteration quantile loss for $N = 1000$ and $p = 20$.

Figure 5: Results for robust linear regression for $p = 20$, $\gamma = 20$, $\sigma^2 = 0.75$ and $N = 100, 500, 1000$.

Moreover, we compare our algorithm when the true parameter θ^* is sparse, which is favorable for regularized linear regression (esp. Lasso). Specifically, we set the true parameter $\theta^* = [\frac{1}{\sqrt{r}}, \cdots, \frac{1}{\sqrt{r}}, 0, \cdots, 0] \in \mathbb{R}^p$, where r is a sparsity parameter indicating the last $p - r$ dimensions of θ^* are zero. We generate covariate $x \in \mathbb{R}^p$ from an scaled standardized Pareto distribution with tail-parameter $\beta = 4.1$. The initial parameter is set to $\theta^1 = [\frac{-1}{\sqrt{p}}, \frac{-1}{\sqrt{p}}, \cdots, \frac{-1}{\sqrt{p}}]$. The response is generated by $y = \langle x, \theta^* \rangle + w$, where w is sampled from scaled rescaled Pareto distribution with mean 0, variance σ^2 and tail-parameter $\beta = 2.1$.

Figure 6 shows the results of sparse linear regression when $p = 20$,$r = 3$, $N = 500, 1000$. In sparse setting, Lasso performs better than in the dense setting but is still worse than our `clipped-SGD`.



(a) Expected convergence curve for $N = 500$ and $p = 20$.

(b) Last iteration quantile loss for $N = 500$ and $p = 20$.

(c) Expected convergence curve for $N = 1000$ and $p = 20$.

(d) Last iteration quantile loss for $N = 500$ and $p = 20$.

Figure 6: Results for robust sparse linear regression for $p = 20$, $r = 3$, $\sigma^2 = 0.75$, $\gamma = 20$ and $N = 500, 1000$.

D.3 Synthetic Experiments: Logistic regression

In this section, we present an extra experimental results on logistic regression.

Logistic regression model. In this model, we observe covariate-response pairs $(x_i, y_i) \in \mathbb{R}^p \times \{0, 1\}$ for $1 \le i \le N$, where each pair (x, y) is sampled from the true distribution P. The conditional distribution of the response y given the covariate x is

$$\Pr(y|x) = \begin{cases} \frac{1}{1+\exp(-\langle x,\theta^*\rangle)}, \text{ if } y = 1. \\ \frac{1}{1+\exp(\langle x,\theta^*\rangle)}, \text{ if } y = 0. \end{cases} \tag{19}$$

We focus on the random design setting where the covariates $x \in \mathbb{R}^p$ have mean 0 and covariance matrix Σ. The loss function we used is negative log-likelihood function:

$$\overline{\mathcal{L}}(\theta, (x, y)) = -y \langle x, \theta\rangle + \log(1 + \exp(\langle x, \theta\rangle)). \tag{20}$$

The true parameter is the minimizer of the resulting population risk. The gradient of the loss function is given by

$$\nabla\overline{\mathcal{L}}(\theta, (x, y)) = \left(y - \frac{1}{1 + \exp(-\langle x, \theta\rangle)}\right) x. \tag{21}$$

The hessian matrix of the population risk is

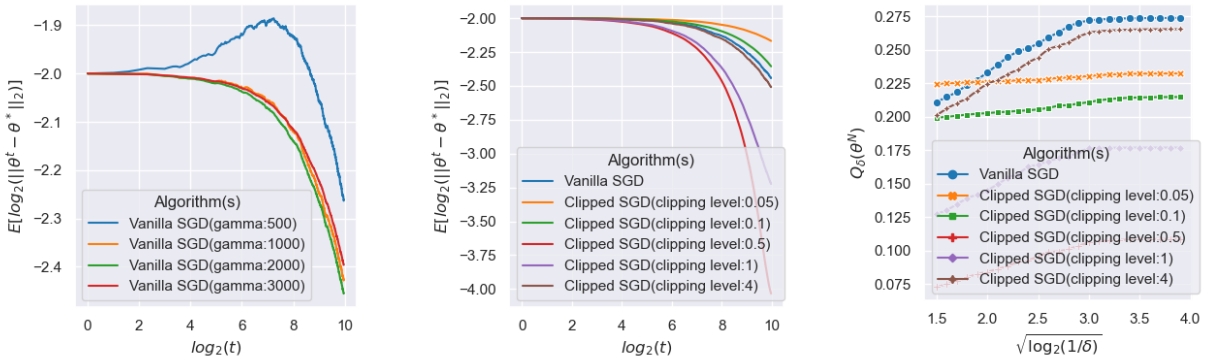
$$\nabla^2\mathcal{R}(\theta) = \mathbb{E}\left[\frac{\exp(\langle x, \theta\rangle)}{(1 + \exp(\langle x, \theta\rangle))^2} xx^\top\right]. \tag{22}$$

We note that $\lambda_{min}(\nabla^2\mathcal{R}(\theta))$ approaches to 0 as θ diverges and the objective function is no longer strongly convex. Therefore, in this case, we restrict the domain of θ to be a bounded convex set Θ.

Setup. We generate covariate $x \in \mathbb{R}^p$ from an scaled standardized Pareto distribution with tail-parameter $\beta = 4.1$. The true regression parameter is set to be $\theta^* = [\frac{1}{\sqrt{p}}, \cdots, \frac{1}{\sqrt{p}}] \in \mathbb{R}^p$ and the initial parameter is set to $\theta^1 = 0.75\theta^*$. The response is generated by Eq.(19). We select $\tau_\ell = \lambda_{min}(\Sigma) = 1$. To ensure $\lambda_{min}(\nabla^2\mathcal{R}(\theta))$ is lower bounded, we restrict the domain Θ to a unit ball centered at θ^*, i.e. $\Theta = \{v \mid \|v - \theta^*\|_2 \le 1\}$. We set $\tau_\ell = 0.1$. Each metric are reported over 5000 trials.

Results. Figure 7 shows the results of heavy-tailed logistic regression. In Figure 7(a), we plot expected convergence curves for different γ for SGD algorithm. We can see that $\gamma = 2000$ yields the best performance. Therefore, we fix $\gamma = 2000$ and compare SGD algorithm with `clipped-SGD`. Figure 7(b), shows expected convergence curves for different clipping levels λ. We can see that the red curve ($\lambda = 0.5$) clearly outperforms Vaniila SGD. The tail performance of $\lambda = 0.5$ is also the best as in Figure 7(c).

We note that the tail performance of Vanilla SGD is well-controlled for logistic regression, as can be seen in Figure 7(c). The reason may be that the distribution of stochastic gradient is not as heavy as the distribution of covariate x. If we see the formula of stochastic gradients in Eq.(21), when $\|x\|_2$ is large, the response y has high probability to be exponentially close to $1/(1+\exp(-\langle x, \theta\rangle))$. Therefore, the term inside the bracket is exponentially small with high probability. The stochastic gradient may not be as heavy-tailed as the covariate x. However, our results show that using clipped gradient is helpful in logistic regression.



(a) Expected convergence curves for different γ.

(b) Expected convergence curves for different λ ($\gamma = 2000$).

(c) Quantile loss for different clipping level λ.

Figure 7: Results for heavy-tailed logistic regression for $p = 10$ and $N = 1000$.

E Proof of Theorem 1

Proof. First of all, we let the clipped gradient at step t be $g_t = \text{clip}(\nabla\overline{\mathcal{L}}(\theta^t, x_t), \lambda)$ and let $\epsilon_t = \nabla\mathcal{R}(\theta^t) - g_t$ be the difference between the stochastic gradient and the true gradient at step t. Also, we let $\mathcal{F}_t = \sigma(x_1, \cdots, x_t)$ be the σ-algebra generated by the first t steps of `clipped-SGD`. Our first step is unrolling the update rule: $\theta^{t+1} = \mathcal{P}_\Theta\left(\theta^t - \eta_t \text{clip}(\nabla\overline{\mathcal{L}}(\theta^t, x_t), \lambda)\right)$.

Lemma 4. *[Lemma 3.11, [2]] Let f be M-smooth and m-strongly convex in $\mathbb{R}^p$, then for all $x, y \in \mathbb{R}^p$, we have*

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{mM}{m+M}\|x-y\|_2^2 + \frac{1}{m+M}\|\nabla f(y) - \nabla f(x)\|_2^2.$$

Lemma 5. *Under the conditions in theorem 1, for any $1 \leq t \geq N$, we have*

$$\|\theta^{t+1} - \theta^*\|_2^2 \leq \frac{\gamma(\gamma-1)}{(t+\gamma)(t+\gamma-1)}\|\theta^1 - \theta^*\|_2^2 + \frac{\sum_{i=1}^t (i+\gamma-1)\left\langle \epsilon_i, \theta^i - \theta^*\right\rangle}{\tau_\ell(t+\gamma)(t+\gamma-1)} + \frac{\sum_{i=1}^t \|\epsilon_i\|_2^2}{2\tau_\ell^2(t+\gamma)(t+\gamma-1)}. \quad (23)$$

Proof. By the strong convexity of $\mathcal{R}(\theta)$ and the fact that θ^* minimizes $\mathcal{R}(\theta)$ in Θ, we have

$$\left\langle \nabla\mathcal{R}(\theta^t), \theta^t - \theta^*\right\rangle \geq \mathcal{R}(\theta^t) - \mathcal{R}(\theta^*) + \frac{\tau_\ell}{2}\|\theta^t - \theta^*\|_2^2, \quad (24)$$

and

$$\mathcal{R}(\theta^t) - \mathcal{R}(\theta^*) \geq \frac{\tau_\ell}{2}\|\theta^t - \theta^*\|_2^2. \quad (25)$$

Putting these two inequality together, we have

$$\left\langle \nabla\mathcal{R}(\theta^t), \theta^t - \theta^*\right\rangle \geq \tau_\ell\|\theta^t - \theta^*\|_2^2. \quad (26)$$

Also, since Θ is a convex set, we have $\|\mathcal{P}_\Theta(\theta) - \theta^*\|_2 \leq \|\theta - \theta^*\|_2$ since $\theta^* \in \Theta$. By rewinding the update rule of clipped-SGD algorithm, we have the following:

$$\begin{aligned}
&\|\theta^{t+1} - \theta^*\|_2^2 \quad (27)\\
&= \|\mathcal{P}_\Theta(\theta^t - \eta_t g_t) - \theta^*\|_2^2\\
&\leq \|\theta^t - \eta_t g_t - \theta^*\|_2^2\\
&= \|\theta^t - \theta^*\|_2^2 - 2\eta_t\left\langle \theta^t - \theta^*, g_t\right\rangle + \eta_t^2\|g_t\|_2^2\\
&= \|\theta^t - \theta^*\|_2^2 - 2\eta_t\left\langle \nabla\mathcal{R}(\theta^t), \theta^t - \theta^*\right\rangle + 2\eta_t\left\langle \epsilon_t, \theta^t - \theta^*\right\rangle + \eta_t^2\|\nabla\mathcal{R}(\theta^t) + \epsilon_t\|_2^2\\
&\leq \|\theta^t - \theta^*\|_2^2 - 2\eta_t\left\langle \nabla\mathcal{R}(\theta^t), \theta^t - \theta^*\right\rangle + 2\eta_t\left\langle \epsilon_t, \theta^t - \theta^*\right\rangle + 2\eta_t^2\|\nabla\mathcal{R}(\theta^t)\|_2^2 + 2\eta_t^2\|\epsilon_t\|_2^2, \quad (28)
\end{aligned}$$

where the last inequality is by Cauchy-Schwarz inequality. Then we apply Lemma 4 by initiating with $f = \mathcal{R}(\cdot)$, $x = \theta^t$ and $y = \theta^*$. By the regularity assumption in Eq. (2), we have

$$(\tau_\ell + \tau_u)\left\langle \nabla\mathcal{R}(\theta^t), \theta^t - \theta^*\right\rangle - \tau_\ell\tau_u\|\theta^t - \theta^*\|_2^2 \geq \|\nabla\mathcal{R}(\theta^t)\|_2^2.$$

Combining the above inequality and Eq. (28), we have

$$
\begin{aligned}
&\|\theta^{t+1}-\theta^*\|_2^2 \\
&\leq (1-2\eta_t^2\tau_\ell\tau_u)\|\theta^t-\theta^*\|_2^2-(2\eta_t-2\eta_t^2(\tau_\ell+\tau_u))\left\langle\nabla\mathcal{R}(\theta^t),\theta^t-\theta^*\right\rangle+2\eta_t\left\langle\epsilon_t,\theta^t-\theta^*\right\rangle+2\eta_t^2\|\epsilon_t\|_2^2 \\
&\overset{(i)}{\leq} (1-2\eta_t^2\tau_\ell\tau_u)\|\theta^t-\theta^*\|_2^2-(2\eta_t-2\eta_t^2(\tau_\ell+\tau_u))\tau_\ell\|\theta^t-\theta^*\|_2^2+2\eta_t\left\langle\epsilon_t,\theta^t-\theta^*\right\rangle+2\eta_t^2\|\epsilon_t\|_2^2 \\
&= (1-2\eta_t\tau_\ell-2\eta_t^2\tau_\ell^2)\|\theta^t-\theta^*\|_2^2+2\eta_t\left\langle\epsilon_t,\theta^t-\theta^*\right\rangle+2\eta_t^2\|\epsilon_t\|_2^2 \\
&\leq (1-2\eta_t\tau_\ell)\|\theta^t-\theta^*\|_2^2+2\eta_t\left\langle\epsilon_t,\theta^t-\theta^*\right\rangle+2\eta_t^2\|\epsilon_t\|_2^2 \\
&\leq \left(1-\frac{2}{t+\gamma}\right)\|\theta^t-\theta^*\|_2^2+\frac{2}{\tau_\ell(t+\gamma)}\left\langle\epsilon_t,\theta^t-\theta^*\right\rangle+\frac{2\|\epsilon_t\|_2^2}{\tau_\ell^2(t+\gamma)^2},
\end{aligned}
\tag{29}
$$

where (i) is due to $\forall\, t\geq 0, 2\eta_t\geq 2\eta_t^2(\tau_\ell+\tau_u)\Leftrightarrow\gamma\geq\tau_u/\tau_\ell$ and Eq. (26). The last inequality follows from the definition of η_t. Then we unwind this formula till $t=1$ and we get that for any $t>1$,

$$
\begin{aligned}
&\|\theta^{t+1}-\theta^*\|_2^2 \\
&\leq \left(\prod_{j=1}^{t}\frac{j+\gamma-2}{j+\gamma}\right)\|\theta^1-\theta^*\|_2^2+\sum_{i=1}^{t}\frac{\left\langle\epsilon_i,\theta^i-\theta^*\right\rangle}{\tau_\ell(i+\gamma)}\left(\prod_{j=i+1}^{t}\frac{j+\gamma-2}{j+\gamma}\right)+\sum_{i=1}^{t}\frac{\|\epsilon_i\|_2^2}{2\tau_\ell^2(i+\gamma)^2}\left(\prod_{j=i+1}^{t}\frac{j+\gamma-2}{j+\gamma}\right) \\
&\leq \frac{\gamma(\gamma-1)}{(t+\gamma)(t+\gamma-1)}\|\theta^1-\theta^*\|_2^2+\frac{\sum_{i=1}^{t}(i+\gamma-1)\left\langle\epsilon_i,\theta^i-\theta^*\right\rangle}{\tau_\ell(t+\gamma)(t+\gamma-1)}+\frac{\sum_{i=1}^{t}\|\epsilon_i\|_2^2}{2\tau_\ell^2(t+\gamma)(t+\gamma-1)},
\end{aligned}
\tag{30}
$$

where the last equation is due to

$$
\prod_{j=i+1}^{t}\frac{j+\gamma-2}{j+\gamma}=\frac{(i+\gamma)(i+\gamma-1)}{(t+\gamma)(t+\gamma-1)}.
$$

□

Now come back to the proof of Theorem 1. We note that clipping introduces bias, which influences the convergence of this method. Hence, we decompose the noise term $\epsilon_i=\mathrm{clip}(\nabla\overline{\mathcal{L}}(\theta^i,x_i),\lambda)-\nabla\mathcal{R}(\theta^i)$ into a bias term ϵ_i^b and a variance term ϵ_i^v, i.e.

$$
\epsilon_i=\epsilon_i^b+\epsilon_i^v,\quad\text{where}\quad\epsilon_i^b=\mathbb{E}_{x_i}[\epsilon_i|\mathcal{F}_{i-1}]\quad\text{and}\quad\epsilon_i^v=\epsilon_i-\mathbb{E}_{x_i}[\epsilon_i|\mathcal{F}_{i-1}]
\tag{31}
$$

since ϵ_i is $\mathcal{F}_{i-1}$-measurable. Putting the definition and the result of Lemma 5, we have

$$
\begin{aligned}
&\|\theta^{t+1}-\theta^*\|_2^2 \\
&\leq \frac{\gamma(\gamma-1)}{(t+\gamma)(t+\gamma-1)}\|\theta^1-\theta^*\|_2^2+\frac{\sum_{i=1}^{t}(i+\gamma-1)\left\langle\epsilon_i,\theta^i-\theta^*\right\rangle}{\tau_\ell(t+\gamma)(t+\gamma-1)}+\frac{\sum_{i=1}^{t}\|\epsilon_i\|_2^2}{2\tau_\ell^2(i+\gamma)(i+\gamma-1)} \\
&\leq \frac{\gamma(\gamma-1)}{(t+\gamma)(t+\gamma-1)}\|\theta^1-\theta^*\|_2^2+\frac{\sum_{i=1}^{t}(i+\gamma-1)\left\langle\epsilon_i^v,\theta^i-\theta^*\right\rangle}{\tau_\ell(t+\gamma)(t+\gamma-1)}+\frac{\sum_{i=1}^{t}(i+\gamma-1)\left\langle\epsilon_i^b,\theta^i-\theta^*\right\rangle}{\tau_\ell(t+\gamma)(t+\gamma-1)} \\
&+\frac{2\sum_{i=1}^{t}\left(\|\epsilon_i^v\|_2^2-\mathbb{E}_{x_i}[\|\epsilon_i^v\|_2^2|\mathcal{F}_{i-1}]\right)}{\tau_\ell^2(t+\gamma)(t+\gamma-1)}+\frac{2\sum_{i=1}^{t}\mathbb{E}_{x_i}[\|\epsilon_i^v\|_2^2|\mathcal{F}_{i-1}]}{\tau_\ell^2(t+\gamma)(t+\gamma-1)}+\frac{2\sum_{i=1}^{t}\|\epsilon_i^b\|_2^2}{\tau_\ell^2(t+\gamma)(t+\gamma-1)}.
\end{aligned}
\tag{32}
$$

where the inequality follows from $\|a+b\|_2^2\leq 2\|a\|_2^2+2\|a\|_2^2$. The rest of the proof is based on the analysis of inequality (32). To bound it, we first introduce the Freedman's inequality for martingale differences. The following version of Freedman's inequality is in Theorem 1.2A in Victor et al. [51].

Lemma 6. *(Freedman's inequality) Let $d_1, d_2, \cdots, d_T$ be a martingale difference sequence with a uniform bound b on the steps d_i. Let V_s denote the sum of conditional variances, i.e.*

$$V_s = \sum_{i=1}^{s} \mathrm{Var}(d_i | d_1, \cdots, d_{i-1}).$$

Then, for every $a, v > 0$,

$$\Pr\left(\sum_{i=1}^{s} d_i \geq a \quad \textit{and} \quad V_s \leq v \quad \textit{for some} \quad s \leq T \right) \leq \exp\left(\frac{-a^2}{2(v+ba)} \right).$$

Next, to apply Freedman's inequality to bound the martingale difference sequence, e.g. $\left\langle \epsilon_i^v, \theta^i - \theta^* \right\rangle$ in the second term in Eq. (32), we should control the conditional variance $\mathbb{E}_{x_i}[\|\epsilon_i^v\|_2^2 | \mathcal{F}_{i-1}]$ and the upper bound of L2-norm $\|\epsilon_i^v\|_2$. Also, as in the third and sixth term in Eq. (32), we should control the magnitude of the bias term, $\|\epsilon_i^v\|_2$ for all $0 \leq i \leq t$. We introduce the following lemma to control these noise terms:

Lemma 7. *(Lemma F.5, [24]) For any $i = 0, 2, .., t$, we have*

$$\|\epsilon_i^v\|_2 \leq 2\lambda. \tag{33}$$

Moreover, for all $i = 1, 2, .., N$, assume that the variance of stochastic gradients is bounded by σ_i^2, i.e. $\mathbb{E}_{x_i}[\|\nabla \overline{\mathcal{L}}(\theta^i, x_i) - \nabla \mathcal{R}(\theta^i)\|_2^2 | \mathcal{F}_{i-1}] \leq \sigma_i^2$ and assume that the norm of the true gradient is less than $\lambda/2$, i.e. $\|\nabla \mathcal{R}(\theta^i)\|_2 \leq \lambda/2$. Then we have

$$\|\epsilon_i^b\|_2 \leq \frac{4\sigma_i^2}{\lambda}, \textit{ and} \tag{34}$$

$$\mathbb{E}_{x_i}[\|\epsilon_i^v\|_2^2 | \mathcal{F}_{i-1}] \leq 18\sigma_i^2 \quad \textit{for all } i = 1, 2, ..., N. \tag{35}$$

Also, recall that we have an assumption about the variance of stochastic gradient in Eq. (3): there exist $\alpha(P, \overline{\mathcal{L}})$ and $\beta(P, \overline{\mathcal{L}})$ such that for every $\theta \in \Theta$:

$$\mathbb{E}_{x \sim P}[\|\nabla \overline{\mathcal{L}}(\theta, x) - \nabla \mathcal{R}(\theta)\|_2^2] \leq \alpha(P, \overline{\mathcal{L}}) \|\theta - \theta^*\|_2^2 + \beta(P, \overline{\mathcal{L}}).$$

For brevity, in the rest of the proof, let $\alpha = \alpha(P, \overline{\mathcal{L}})$ and $\beta = \beta(P, \overline{\mathcal{L}})$. Also, to apply Lemma 7, we let $\sigma_i^2 = \alpha \|\theta^i - \theta^*\|_2^2 + \beta$. Then, we have

$$\Rightarrow \mathbb{E}_{x_i}[\|\nabla \overline{\mathcal{L}}(\theta^i, z_i) - \nabla \mathcal{R}(\theta^i)\|_2^2 | \mathcal{F}_{i-1}] \leq \sigma_i^2 = \alpha \|\theta^i - \theta^*\|_2^2 + \beta.$$

Now, with these two lemmas in hands, we start to analyze Eq. (32). We first define a new constant A and C:

$$A = C_1^2 \left(\gamma(\gamma - 1) \|\theta^1 - \theta^*\|_2^2 + \frac{(N + \gamma)\beta \log(2/\delta)}{\tau_\ell^2} \right) \quad \text{and} \quad C = 5000. \tag{36}$$

Recall that $C_1 \geq 1$ is a scaling constant in Theorem 1. We note that the clipping level λ can be written in the following form:

$$\lambda = \frac{\tau_\ell \sqrt{A}}{\log(2/\delta)}. \tag{37}$$

Then we introduce new random variables: for $1 \leq i \leq N$.

$$\zeta_i = \begin{cases} \theta^i - \theta^* & \text{, if } \|\theta^i - \theta^*\|_2^2 \leq \frac{CA}{(i+\gamma-1)(i+\gamma-2)}. \\ 0 & \text{, otherwise.} \end{cases} \tag{38}$$

We note that these random variables are bounded almost surely, i.e.

$$\Pr\left(\|\zeta_i\|_2 \le \sqrt{\frac{CA}{(i+\gamma-1)(i+\gamma-2)}}\right) = 1.$$

Next, we introduce the following claim to control these two martingale difference sequences:$\{(i+\gamma-1)\langle\epsilon_i^v, \zeta_i\rangle\}_{1\le i\le N}$ and $\{\|\epsilon_i^v\|_2^2 - \mathbb{E}[\|\epsilon_i^v\|_2^2|\mathcal{F}_{i-1}]\}_{1\le i\le N}$, which appeared in the second and forth terms in Eq.(32) respectively.

Claim 1. *Define* $X_i = (i+\gamma-1)\langle\epsilon_i^v, \zeta_i\rangle$ *and* $Y_i = \|\epsilon_i^v\|_2^2 - \mathbb{E}_{x_i}[\|\epsilon_i^v\|_2^2|\mathcal{F}_{i-1}]$ *for* $1 \le i \le N$ *be two sequence and let*

$$v_i = \mathrm{Var}[X_i|\mathcal{F}_{i-1}] \quad \textit{and} \quad u_i = \mathrm{Var}[Y_i|\mathcal{F}_{i-1}] \quad \textit{for } 1 \le i \le N$$

be its conditionally variances. Then with probability at least $1-\delta$*, the following event holds: for all* $1 \le s \le N$*,*

$$\sum_{i=1}^{s} X_i < 20\tau_\ell A\sqrt{C} + 2A\tau_\ell C^{3/4} \quad \textit{or} \quad \sum_{i=1}^{s} v_i > \frac{36A^2C\tau_\ell^2}{\log(2/\delta)} + \frac{3A^2C^{3/2}\tau_\ell^2}{4\log(2/\delta)}, \tag{39}$$

and

$$\sum_{i=1}^{s} Y_i < 118\tau_\ell^2 A + 15A\tau_\ell^2 C^{1/4} \quad \textit{or} \quad \sum_{i=1}^{s} u_i > \frac{108A^2\tau_\ell^4\sqrt{C}}{\log(2/\delta)} + \frac{5184\tau_\ell^4A^2}{\log(2/\delta)}. \tag{40}$$

We will explain the choice of these parameters in Claim 1 later. Now, we denote E be the event that Eq. (39) and (40) holds for all $1 \le s \le N$. We note that by Claim 1, E holds with probability $1-\delta$, i.e.

$$\Pr(E) \ge 1-\delta.$$

Then we prove that if E holds,

$$\|\theta^t - \theta^*\|_2^2 \le \frac{AC}{(t+\gamma-1)(t+\gamma-2)}, \tag{41}$$

for $1 \le t \le N+1$ by induction.

First of all, we prove that it holds for $t=1$. From our definition of constant A in Eq.(36) and the fact that $C_1 \ge 1$, this case holds trivially, i.e.

$$\|\theta^1 - \theta^*\|_2^2 \le \frac{C}{\gamma(\gamma-1)}\left(\gamma(\gamma-1)\|\theta^1-\theta^*\|_2^2 + \frac{(N+\gamma)\beta\log(2/\delta)}{\tau_\ell^2}\right) = \frac{AC}{\gamma(\gamma-1)}.$$

Next, we assume that Eq. (41) holds for $t = 1, \cdots, n$. When $t = n+1$, by Eq. (32), we have

$$\begin{aligned}
&(n+\gamma)(n+\gamma-1)\|\theta^{n+1} - \theta^*\|_2^2 \\
&\le \underbrace{\gamma(\gamma-1)\|\theta^1-\theta^*\|_2^2}_{①} + \underbrace{\frac{\sum_{i=1}^n (i+\gamma-1)\langle\epsilon_i^v, \theta^i-\theta^*\rangle}{\tau_\ell}}_{②} + \underbrace{\frac{\sum_{i=1}^n (i+\gamma-1)\langle\epsilon_i^b, \theta^i-\theta^*\rangle}{\tau_\ell}}_{③} \\
&+ \underbrace{\frac{2\sum_{i=1}^n \|\epsilon_i^v\|_2^2 - \mathbb{E}_{z_i}[\|\epsilon_i^v\|_2^2|\mathcal{F}_{i-1}]}{\tau_\ell^2}}_{④} + \underbrace{\frac{2\sum_{i=1}^n \mathbb{E}_{z_i}[\|\epsilon_i^v\|_2^2|\mathcal{F}_{i-1}]}{\tau_\ell^2}}_{⑤} + \underbrace{\frac{2\sum_{i=1}^n \|\epsilon_i^b\|_2^2}{\tau_\ell^2}}_{⑥}.
\end{aligned} \tag{42}$$

Check conditions in Lemma 7: Before we upper bound ① $\sim$ ⑥, we first prove that $\|\nabla\mathcal{R}(\theta^t)\|_2 \le 2\lambda$. Since $\mathcal{R}(\cdot)$ is τ_u-smooth by assumption in Eq.2, or its gradient is τ_u-Lipschitz, for $1 \le t \le n$ we have

$$\begin{aligned}\|\nabla\mathcal{R}(\theta^t)\|_2 \le \tau_u\|\theta^t-\theta^*\|_2 &\overset{(41)}{\le} \tau_u\sqrt{\frac{AC}{(t+\gamma-1)(t+\gamma-2)}}\\ &\overset{t\ge1}{\le} \tau_u\sqrt{\frac{AC}{\gamma(\gamma-1)}}\\ &\le \frac{\tau_u}{\gamma-1}\sqrt{AC}\\ &\overset{(8)}{\le} \frac{2\tau_\ell C_1\sqrt{A}}{\log(2/\delta)} = 2\lambda,\end{aligned} \tag{43}$$

where the last inequality is due to the definition of $\gamma - 1 > \frac{\sqrt{C}}{2}\frac{\tau_u}{\tau_\ell}\log(2/\delta)$. Therefore, the condition in Lemma 7 holds for $1 \le t \le n$, which means we could use Eq.(34) and (35) to control the bias and variance terms for $1 \le t \le n$.

Upper bounds for ①:

$$① \le C_1^2\left(\gamma(\gamma-1)\|\theta^1-\theta^*\|_2^2 + \frac{(N+\gamma)\beta\log(2/\delta)}{\tau_\ell^2}\right) = A. \tag{44}$$

Upper bounds for ②: Recall that our definition of bounded variable ζ_i in Eq. (38) and martingale difference sequence $\{X_i\}_{1\le i\le n}$ in Claim 1. Under the induction hypothesis, we have

$$\zeta_i = \theta^i - \theta^* \quad \text{for } 1 \le i \le n \text{ and } \quad ② = \frac{\sum_{i=1}^n X_i}{\tau_\ell}.$$

We first show that the sum of its conditional variances are upper bounded, i.e.

$$\sum_{i=1}^n \mathrm{Var}[X_i^2|\mathcal{F}_{i-1}] = \sum_{i=1}^n v_i \le \frac{36A^2C\tau_\ell^2}{\log(2/\delta)} + \frac{3A^2C^{3/2}\tau_\ell^2}{4\log(2/\delta)}.$$

Since $X_i = (i+\gamma-1)\langle \epsilon_i^v, \zeta_i \rangle$ for $1 \le i \le n$, we have

$$\begin{aligned}
\sum_{i=1}^{n} v_i &\le \sum_{i=1}^{n} (i+\gamma-1)^2 \mathbb{E}[\|\zeta_i\|_2^2 \|\epsilon_i^v\|_2^2 | \mathcal{F}_{i-1}] \\
&\overset{(38)}{\le} \sum_{i=1}^{n} \frac{(i+\gamma-1)^2 CA}{(i+\gamma-1)(i+\gamma-2)} \mathbb{E}[\|\epsilon_i^v\|_2^2 | \mathcal{F}_{i-1}] \\
&\le 2CA \sum_{i=1}^{n} \mathbb{E}[\|\epsilon_i^v\|_2^2 | \mathcal{F}_{i-1}] \\
&\overset{(33),(3)}{\le} 2CA \sum_{i=1}^{n} 18(\alpha \|\theta^i - \theta^*\|_2^2 + \beta) \\
&\overset{(41)}{\le} 36CA \left(n\beta + \alpha \sum_{i=1}^{n} \frac{AC}{(i+\gamma-1)(i+\gamma-2)} \right) \\
&\le 36CA \left(n\beta + \alpha AC \sum_{i=1}^{n} \left(\frac{1}{i+\gamma-2} - \frac{1}{i+\gamma-1} \right) \right) \\
&\le 36CA \left(n\beta + \frac{\alpha AC}{\gamma-1} \right) \\
&\overset{n \le N}{\le} 36\beta NAC + \frac{36\alpha A^2 C^2}{\gamma-1}. \qquad (45)
\end{aligned}$$

Since we have $A \overset{(36)}{\ge} \frac{(N+\gamma)\beta \log(2/\delta)}{\tau_\ell^2} \ge \frac{\beta N \log(2/\delta)}{\tau_\ell^2}$ and $\gamma - 1 \overset{(8)}{\ge} \frac{48\alpha\sqrt{C}\log(2/\delta)}{\tau_\ell^2}$, we have

$$\begin{aligned}
\sum_{i=1}^{n} v_i &\le 36(\beta N)AC + \frac{36\alpha A^2 C^2}{\gamma-1} \\
&\le \frac{36A^2 C\tau_\ell^2}{\log(2/\delta)} + \frac{3A^2 C^{3/2}\tau_\ell^2}{4\log(2/\delta)}. \qquad (46)
\end{aligned}$$

Therefore, we know the second inequality in Eq. (39) does not hold, so the first one must hold for $t = n$, i.e.

$$\sum_{i=1}^{n} X_i < 20\tau_\ell A\sqrt{C} + 2A\tau_\ell C^{3/4}.$$

Then we can upper bound ②:

$$② = \frac{\sum_{i=1}^{n} X_i}{\tau_\ell} \le A(20\sqrt{C} + 2C^{3/4}) \qquad (47)$$

Upper bound ③: By Eq.(34) in Lemma 7, we have $\|\epsilon_t^b\|_2 \le \frac{4(\alpha\|\theta^t-\theta^*\|_2^2+\beta)}{\lambda}$. Then we have

$$
\begin{aligned}
③ &= \frac{\sum_{i=1}^n (i+\gamma-1)\left\langle \epsilon_i^b, \theta^i-\theta^*\right\rangle}{\tau_\ell} \\
&\le \frac{\sum_{i=1}^n (i+\gamma-1)\|\epsilon_i^b\|_2\|\theta^i-\theta^*\|_2}{\tau_\ell} \\
&\overset{(34)}{\le} \frac{\sum_{i=1}^n 4(i+\gamma-1)(\alpha\|\theta^i-\theta^*\|_2^3+\beta\|\theta^i-\theta^*\|_2)}{\lambda\tau_\ell} \\
&\overset{(37),(41)}{\le} \frac{\sum_{i=1}^n 4(i+\gamma-1)\sqrt{\frac{CA}{(i+\gamma-1)(i+\gamma-2)}}(\frac{CA\alpha}{(i+\gamma-1)(i+\gamma-2)}+\beta)\log(2/\delta)}{\tau_\ell^2\sqrt{A}} \\
&\le \frac{\sum_{i=1}^n 8\sqrt{CA}(\frac{CA\alpha}{(i+\gamma-1)(i+\gamma-2)}+\beta)\log(2/\delta)}{\tau_\ell^2\sqrt{A}} \\
&\le \frac{\sum_{i=1}^n 8\beta\sqrt{C}\log(2/\delta)+8\alpha AC\sqrt{C}\log(2/\delta)\left(\frac{1}{(i+\gamma-2)}-\frac{1}{(i+\gamma-1)}\right)}{\tau_\ell^2} \\
&\overset{n\le N}{\le} \frac{8N\beta\sqrt{C}\log(2/\delta)}{\tau_\ell^2}+\frac{8\alpha AC\sqrt{C}\log(2/\delta)}{\tau_\ell^2(\gamma-1)}.
\end{aligned} \tag{48}
$$

Next, since

$$
A \ge \frac{(N+\gamma)\beta\log(2/\delta)}{\tau_\ell^2} \ge \frac{N\beta\log(2/\delta)}{\tau_\ell^2} \quad \text{and} \quad \gamma-1 \overset{(8)}{\ge} \frac{48\alpha\sqrt{C}\log(2/\delta)}{\tau_\ell^2},
$$

We have

$$
③ \le 8\sqrt{C}\frac{N\beta\log(2/\delta)}{\tau_\ell^2}+\frac{AC}{6}\times\frac{48\alpha\sqrt{C}\log(2/\delta)}{\tau_\ell^2(\gamma-1)} \le A\left(8\sqrt{C}+\frac{C}{6}\right). \tag{49}
$$

Upper bound of ④: Recall that our definition of martingale difference sequence $\{Y_i\}_{1\le i\le n}$ in Claim 1. We have

$$
④ = \frac{2\sum_{i=1}^n\left(\|\epsilon_i^v\|_2^2-\mathbb{E}_{z_i}[\|\epsilon_i^v\|_2^2|\mathcal{F}_{i-1}]\right)}{\tau_\ell^2} = \frac{2\sum_{i=1}^n Y_i}{\tau_\ell^2}.
$$

We first show that the sum of its conditional variances is upper bounded, i.e.

$$
\sum_{i=1}^n \mathrm{Var}[Y_i^2|\mathcal{F}_{i-1}] = \sum_{i=1}^n u_i \le \frac{108A^2\tau_\ell^4\sqrt{C}}{\log(2/\delta)}+\frac{5184\tau_\ell^4A^2}{\log(2/\delta)}.
$$

Since we have $\|\epsilon_i^v\|_2 \le 2\lambda$ by Eq. (33), we get $\left|\|\epsilon_i^v\|_2^2 - \mathbb{E}_{z_i}[\|\epsilon_i^v\|_2^2|\mathcal{F}_{i-1}]\right| \overset{(33)}{\le} 4\lambda^2 + 4\lambda^2 = 8\lambda^2$. Then,

$$\begin{aligned}
\sum_{i=1}^n u_i &\le \sum_{i=1}^n \mathbb{E}_{z_i}\left[\left(\|\epsilon_i^v\|_2^2 - \mathbb{E}_{z_i}[\|\epsilon_i^v\|_2^2|\mathcal{F}_{i-1}]\right)^2|\mathcal{F}_{i-1}\right] \\
&\le \sum_{i=1}^n 8\lambda^2 \mathbb{E}_{z_i}\left[\left|\|\epsilon_i^v\|_2^2 - \mathbb{E}_{z_i}[\|\epsilon_i^v\|_2^2|\mathcal{F}_{i-1}]\right||\mathcal{F}_{i-1}\right] \\
&\le \sum_{i=1}^n 8\lambda^2 \times 2\mathbb{E}_{z_i}[\|\epsilon_i^v\|_2^2|\mathcal{F}_{i-1}] \\
&\overset{(35)}{\le} 16\lambda^2 \sum_{i=1}^n 18(\alpha\|\theta^i - \theta^*\|_2^2 + \beta) \\
&\overset{(37),(41)}{\le} \frac{288\tau_\ell^2 A}{\log(2/\delta)^2}\sum_{i=1}^n 18\left(\alpha AC\left(\frac{1}{i+\gamma-2} - \frac{1}{i+\gamma-1}\right) + \beta\right) \\
&\overset{n\le N}{\le} \frac{5184\alpha\tau_\ell^2 A^2 C}{\log(2/\delta)^2(\gamma-1)} + \frac{5184N\beta\tau_\ell^2 A}{\log(2/\delta)^2}.
\end{aligned} \tag{50}$$

Next, since $\delta \le 2e^{-1}$ and

$$A \ge \frac{(N+\gamma)\beta\log(2/\delta)}{\tau_\ell^2} \ge \frac{N\beta}{\tau_\ell^2} \quad \text{and} \quad \gamma - 1 \overset{(8)}{\ge} \frac{48\alpha\sqrt{C}\log(2/\delta)}{\tau_\ell^2},$$

We have

$$\begin{aligned}
\sum_{i=1}^n u_i &\le \frac{5184\alpha\tau_\ell^2 A^2 C}{\log(2/\delta)^2(\gamma-1)} + \frac{5184N\beta\tau_\ell^2 A}{\log(2/\delta)^2} \\
&\overset{\delta\le 2e^{-1}}{\le} \frac{108A^2\tau_\ell^4\sqrt{C}}{\log(2/\delta)} + \frac{5184\tau_\ell^4 A^2}{\log(2/\delta)}.
\end{aligned}$$

Therefore, we know that the second inequality in Eq. (40) does not hold, so that the first one must be satisfied, i.e.

$$\sum_{i=1}^n Y_i < 118\tau_\ell^2 A + 15A\tau_\ell^2 C^{1/4}.$$

Finally, we have

$$④ = \frac{2\sum_{i=1}^n Y_i}{\tau_\ell^2} \le A(236 + 30C^{1/4}). \tag{51}$$

Upper bound of ⑤: By equation Eq.(35) in Lemma 7, we have $\mathbb{E}_{z_t}[\|\epsilon_t^v\|_2^2|\mathcal{F}_{t-1}] \le 18(\alpha\|\theta^t - \theta^*\|_2^2 + \beta)$. Then we have

$$\begin{aligned}
⑤ &= \frac{2\sum_{i=1}^n \mathbb{E}_{z_i}[\|\epsilon_i^v\|_2^2|\mathcal{F}_{i-1}]}{\tau_\ell^2} \\
&\overset{(35),(41)}{\le} \frac{36n\beta}{\tau_\ell^2} + \frac{36\alpha AC\sum_{i=1}^n\left(\frac{1}{i+\gamma-2} - \frac{1}{i+\gamma-1}\right)}{\tau_\ell^2} \\
&\overset{n\le N}{\le} \frac{36N\beta}{\tau_\ell^2} + \frac{36\alpha AC}{(\gamma-1)\tau_\ell^2}.
\end{aligned} \tag{52}$$

Next, since

$$A \geq \frac{(N+\gamma)\beta\log(2/\delta)}{\tau_\ell^2} \overset{\delta\leq 2e^{-1}}{\geq} \frac{N\beta\log(2/\delta)}{\tau_\ell^2} \quad \text{and} \quad \gamma - 1 \geq \frac{48\alpha\sqrt{C}\log(2/\delta)}{\tau_\ell^2} \overset{\delta\leq 2e^{-1}}{\geq} \frac{48\alpha\sqrt{C}}{\tau_\ell^2}.$$

We have

$$⑤ \leq A(36 + \frac{3\sqrt{C}}{4}) \leq A(36 + \sqrt{C}). \tag{53}$$

Upper bound of ⑥: By equation Eq.(34) in Lemma 7, we have $\|\epsilon_t^b\|_2 \leq \frac{4(\alpha\|\theta^t - \theta^*\|_2^2 + \beta)}{\lambda}$. Then we have

$$\begin{aligned}
⑥ &= \frac{2\sum_{i=1}^n \|\epsilon_i^b\|_2^2}{\tau_\ell^2} \\
&\overset{(34)}{\leq} \frac{32\sum_{i=1}^n \left(\beta + \alpha\|\theta^i - \theta^*\|_2^2\right)^2}{\lambda^2\tau_\ell^2} \\
&\leq \frac{64\sum_{i=1}^n \left(\beta^2 + \alpha^2\|\theta^i - \theta^*\|_2^4\right)}{\lambda^2\tau_\ell^2} \quad \text{(Cauchy-Schwarz)} \\
&\overset{n\leq N,(41)}{\leq} \frac{64N\beta^2 + 64A^2C^2\alpha^2\sum_{i=1}^n \frac{1}{(i+\gamma-2)^2(i+\gamma-1)^2}}{\lambda^2\tau_\ell^2} \\
&\overset{(i)}{\leq} \frac{64N\beta^2 + \frac{64A^2C^2\alpha^2}{(\gamma-1)^2}}{\lambda^2\tau_\ell^2} \\
&\overset{(37)}{=} \frac{64N\beta^2\log(2/\delta)^2}{A\tau_\ell^4} + \frac{64AC^2\alpha^2\log(2/\delta)^2}{\tau_\ell^4(\gamma-1)^2},
\end{aligned} \tag{54}$$

where (i) is due to

$$\begin{aligned}
\sum_{i=1}^n \frac{1}{(i+\gamma-2)^2(i+\gamma-1)^2} &\leq \left(\sum_{i=1}^n \frac{1}{(i+\gamma-2)(i+\gamma-1)}\right)^2 \\
&\leq \left(\sum_{i=1}^n \frac{1}{i+\gamma-2} - \frac{1}{i+\gamma-1}\right)^2 \\
&\leq \frac{1}{(\gamma-1)^2}.
\end{aligned}$$

$$A \geq \frac{(N+\gamma)\beta\log(2/\delta)}{\tau_\ell^2} \overset{\delta\leq 2e^{-1}}{\geq} \frac{N\beta\log(2/\delta)}{\tau_\ell^2} \quad \text{and} \quad \gamma - 1 \geq \frac{48\alpha\sqrt{C}\log(2/\delta)}{\tau_\ell^2}.$$

We have

$$\begin{aligned}
⑥ &\leq \frac{64}{AN} \times \frac{N^2\beta^2\log(2/\delta)^2}{\tau_\ell^4} + \frac{64AC^2\alpha^2\log(2/\delta)^2}{\tau_\ell^4(\gamma-1)^2} \\
&\leq A(\frac{64}{N} + \frac{C}{36}) \overset{N\geq 1}{\leq} A(64 + \frac{C}{36}).
\end{aligned} \tag{55}$$

Upper bounds for ① + ② + ③ + ④ + ⑤ + ⑥: Now, we have derived the upper bounds for ① ∼ ⑥. By combining Eq. (44),(47),(49),(51),(53),(55), we have

$$
\begin{aligned}
(n+\gamma)(n+\gamma-1)\|\theta^{n+1}-\theta^*\|_2^2 &\le ① + ② + ③ + ④ + ⑤ + ⑥ \\
&\le A(1+20\sqrt{C}+2C^{3/4}+8\sqrt{C}+\frac{C}{6}+236+30C^{1/4}+36+\sqrt{C}+64+\frac{C}{36}) \\
&= A(337+30C^{1/4}+29\sqrt{C}+2C^{3/4}+\frac{7C}{36}) \\
&\le AC, \qquad (56)
\end{aligned}
$$

where the last inequality follows from the definition of $C = 5000$. Therefore, we can conclude our induction proof.

Lastly, by plugging in $n = N$ and the definition of A in the above equation, we have, with probability at least $1-\delta$ (so that E holds),

$$
\begin{aligned}
\|\theta^{N+1}-\theta^*\|_2^2 &\le \frac{CC_1^2\left(\gamma(\gamma-1)\|\theta^1-\theta^*\|_2^2+\frac{(N+\gamma)\beta\log(2/\delta)}{\tau_\ell^2}\right)}{(N+\gamma)(N+\gamma-1)} \\
&\le 2CC_1^2\left(\frac{\gamma^2\|\theta^1-\theta^*\|_2^2}{(N+\gamma)^2}+\frac{\beta(P,\overline{\mathcal{L}})\log(2/\delta)}{(N+\gamma)\tau_\ell^2}\right), \qquad (57)
\end{aligned}
$$

where the last inequality is due to $(N+\gamma-1) \ge \frac{N+\gamma}{2}$. If we take square root on the both sides and use the inequality $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$ for $a, b \in \mathbb{R}^+$, we have

$$
\|\theta^{N+1}-\theta^*\|_2 \le 100C_1\left(\frac{\gamma\|\theta^1-\theta^*\|_2}{N+\gamma}+\frac{1}{\tau_\ell}\sqrt{\frac{\beta(P,\overline{\mathcal{L}})\log(2/\delta)}{N+\gamma}}\right). \qquad (58)
$$

□

E.1 Proof of Claim 1

Proof. **(Bounds for the first martingale difference sequence):** for martingale difference sequence $\{(i+\gamma-1)\langle\epsilon_i^v,\zeta_i\rangle\}_{0\le i\le T-1}$,

(i) we first check that it is conditionally unbiased, i.e.

$$
\mathbb{E}_{z_i\sim P}[\epsilon_i^v|\mathcal{F}_{i-1}] = 0 \Rightarrow \mathbb{E}_{z_i\sim P}[(i+\gamma-1)\langle\epsilon_i^v,\zeta_i\rangle|\mathcal{F}_{i-1}] = 0,
$$

since ζ_i is determinant conditioned on $\mathcal{F}_{i-1}$.

(ii) We check each summand is bounded, i.e.

$$
\|(i+\gamma-1)\langle\epsilon_i^v,\zeta_i\rangle\|_2 \overset{(33),(38)}{\le} (i+\gamma-1)(2\lambda)\sqrt{\frac{CA}{(i+\gamma-1)(i+\gamma-2)}} \le 4\lambda\sqrt{A} = \frac{4A\tau_\ell\sqrt{C}}{\log(2/\delta)}
$$

for $1 \le i \le N$.

(iii) Let $V_s = \sum_{i=1}^s v_i$ be the sum of its conditional variance. Then we apply the Freedman's inequality in Lemma 6 instantiated with the following parameters

$$
b = \frac{4A\tau_\ell\sqrt{C}}{\log(2/\delta)}\ , \quad v = \frac{36A^2C\tau_\ell^2}{\log(2/\delta)}+\frac{3A^2C^{3/2}\tau_\ell^2}{4\log(2/\delta)} \text{ and } a = 2b\log(2/\delta)+\sqrt{2v\log(2/\delta)}.
$$

We will specify our choices of parameters later. Then we have

$$\Pr\left(\sum_{i=1}^{s}(i+\gamma-1)\left\langle\epsilon_i^v,\zeta_i\right\rangle \geq a \text{ and } V_s \leq v \text{ for some } s \leq N\right) \leq \exp\left(\frac{-a^2}{2(v+ba)}\right) \leq \frac{\delta}{2}, \tag{59}$$

where the last inequality is due to

$$\begin{aligned}
\exp\left(\frac{-a^2}{2(v+ba)}\right) \leq \frac{\delta}{2} &\Leftrightarrow \frac{-a^2}{2(v+ba)} \leq \log(\frac{\delta}{2}) \\
&\Leftrightarrow a^2 - 2b\log(\frac{\delta}{2})a - 2v\log\left(\frac{\delta}{2}\right) \geq 0 \\
&\Leftrightarrow a \geq b\log(\frac{\delta}{2}) + \sqrt{(b\log(\frac{\delta}{2}))^2 + 2v\log(\frac{\delta}{2})}.
\end{aligned} \tag{60}$$

The choice of a satisfies the above inequality due to the fact that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for $x, y \in \mathbb{R}^+$. Also, we have

$$\begin{aligned}
a &= 2b\log(2/\delta) + \sqrt{2v\log(2/\delta)} \\
&= 8A\tau_\ell\sqrt{C} + \sqrt{72A^2C\tau_\ell^2 + \frac{3A^2C^{3/2}\tau_\ell^2}{2}} \\
&\leq 8A\tau_\ell\sqrt{C} + 12A\tau_\ell\sqrt{C} + 2AC^{3/4}\tau_\ell \\
&= 20\tau_\ell A\sqrt{C} + 2A\tau_\ell C^{3/4}.
\end{aligned} \tag{61}$$

Therefore, by Eq. (59), we have

$$\begin{aligned}
&1 - \frac{\delta}{2} \\
&\geq \Pr\left(\sum_{i=1}^{s}(i+\gamma-1)\left\langle\epsilon_i^v,\zeta_i\right\rangle \geq a \text{ and } V_s \leq v \text{ for some } s \leq N\right) \\
&= 1 - \Pr\left(\sum_{i=1}^{s}(i+\gamma-1)\left\langle\epsilon_i^v,\zeta_i\right\rangle < a \text{ or } V_s > v \text{ for all } 1 \leq s \leq N\right) \\
&\geq 1 - \Pr\left(\sum_{i=1}^{s}(i+\gamma-1)\left\langle\epsilon_i^v,\zeta_i\right\rangle < 20\tau_\ell A\sqrt{C} + 2A\tau_\ell C^{3/4} \text{ or } V_s > \frac{36A^2C\tau_\ell^2}{\log(2/\delta)} + \frac{3A^2C^{3/2}\tau_\ell^2}{4\log(2/\delta)} \text{ for all } 1 \leq s \leq N\right).
\end{aligned} \tag{62}$$

2.(Bounds for the second martingale difference sequence): For martingale difference sequence $\{\|\epsilon_i^v\|_2^2 - \mathbb{E}[\|\epsilon_i^v\|_2^2|\mathcal{F}_{i-1}]\}_{1\leq i\leq N}$, (i) we first check that it is conditionally unbiased, i.e.

$$\mathbb{E}\left[\|\epsilon_i^v\|_2^2 - \mathbb{E}[\|\epsilon_i^v\|_2^2|\mathcal{F}_{i-1}] \ |\mathcal{F}_{i-1}\right] = 0$$

(ii) We check each summand is bounded, i.e.

$$\left|\|\epsilon_i^v\|_2^2 - \mathbb{E}[\|\epsilon_i^v\|_2^2|\mathcal{F}_{i-1}]\right| \leq \|\epsilon_i^v\|_2^2 + \mathbb{E}[\|\epsilon_i^v\|_2^2|\mathcal{F}_{i-1}] \overset{(33)}{\leq} 4\lambda^2 + 4\lambda^2 = 8\lambda^2 = \frac{8A\tau_\ell^2}{\log^2(2/\delta)} \leq \frac{8A\tau_\ell^2}{\log(2/\delta)}$$

for $1 \leq i \leq N$ since $\|\epsilon_i^v\|_2 \leq 2\lambda$.

(iii) Let $U_s = \sum_{i=1}^{s} u_i^2$ be the sum of its conditional variance. Then we apply the Freedman's inequality instantiated with parameters

$$b = \frac{8A\tau_\ell^2}{\log(2/\delta)}\ , \quad v = \frac{108A^2\tau_\ell^4\sqrt{C}}{\log(2/\delta)} + \frac{5184\tau_\ell^4A^2}{\log(2/\delta)} \text{ and } a = 2b\log(2/\delta) + \sqrt{2v\log(2/\delta)}.$$

We will specify our choices of parameters later. Then we have

$$\Pr\left(\sum_{i=1}^{s}\|\epsilon_i^v\|_2^2-\mathbb{E}[\|\epsilon_i^v\|_2^2\geq a \ \text{ and } \ U_s\leq v \text{ for some } s\leq N\right)\leq \exp\left(\frac{-a^2}{2(v+ba)}\right)\overset{(60)}{\leq}\frac{\delta}{2} \tag{63}$$

for all $1\leq i\leq N$. The choice of a satisfies the above inequality due to the fact that $\sqrt{x+y}\leq\sqrt{x}+\sqrt{y}$ for $x,y\in\mathbb{R}^+$. Also, we have

$$\begin{aligned}
a&=2b\log(2/\delta)+\sqrt{2v\log(2/\delta)}\\
&=8A\tau_\ell^2+\sqrt{216A^2\tau_\ell^4\sqrt{C}+10368\tau_\ell^4A^2}\\
&\leq 8A\tau_\ell^2+15A\tau_\ell^2C^{1/4}+110\tau_\ell^2A\\
&=118A\tau_\ell^2+15A\tau_\ell^2C^{1/4}.
\end{aligned}\tag{64}$$

Therefore, by Eq. (63), we have

$$\begin{aligned}
&1-\frac{\delta}{2}\\
&\geq\Pr\left(\sum_{i=1}^{s}\|\epsilon_i^v\|_2^2-\mathbb{E}[\|\epsilon_i^v\|_2^2\geq a \ \text{ and } \ U_s\leq v \text{ for some } s\leq N\right)\\
&=1-\Pr\left(\sum_{i=1}^{s}\|\epsilon_i^v\|_2^2-\mathbb{E}[\|\epsilon_i^v\|_2^2<a \ \text{ or } \ U_s>v \text{ for all } 1\leq s\leq N\right)\\
&\geq 1-\Pr\left(\sum_{i=1}^{s}\|\epsilon_i^v\|_2^2-\mathbb{E}[\|\epsilon_i^v\|_2^2<118A\tau_\ell^2+15A\tau_\ell^2C^{1/4} \ \text{ or } \ U_s>\frac{108A^2\tau_\ell^4\sqrt{C}}{\log(2/\delta)}+\frac{5184\tau_\ell^4A^2}{\log(2/\delta)} \text{ for all } 1\leq s\leq N\right).
\end{aligned}\tag{65}$$

Therefore, by combining Eq.(62) and (65), we have, with probability at least $1-\delta$, for all $1\leq s\leq N$,

$$\sum_{i=1}^{s}X_i<20\tau_\ell A\sqrt{C}+2A\tau_\ell C^{3/4} \ \text{ or } \ \sum_{i=1}^{s}v_i>\frac{36A^2C\tau_\ell^2}{\log(2/\delta)}+\frac{3A^2C^{3/2}\tau_\ell^2}{4\log(2/\delta)}, \tag{66}$$

and

$$\sum_{i=1}^{s}Y_i<118\tau_\ell^2A+15A\tau_\ell^2C^{1/4} \ \text{ or } \ \sum_{i=1}^{s}u_i>\frac{108A^2\tau_\ell^4\sqrt{C}}{\log(2/\delta)}+\frac{5184\tau_\ell^4A^2}{\log(2/\delta)}. \tag{67}$$

□

F Proofs of Corollaries

In this section, we provide proofs for Corollary 2, 3 and 4.

F.1 Proof of Corollary 2

Since $\mathcal{R}(\cdot)$ is τ_ℓ-strongly convex and τ_u-smooth, we have, for all $\theta \in \Theta$

$$\frac{\tau_\ell}{2}\|\theta - \theta^*\|_2^2 \leq \mathcal{R}(\theta) - \mathcal{R}(\theta^*) \leq \frac{\tau_u}{2}\|\theta - \theta^*\|_2^2. \tag{68}$$

Also, by Theorem 1, we have

$$\begin{aligned}\|\theta^{N+1} - \theta^*\|_2^2 &\leq O\left(\frac{\gamma^2\|\theta^1 - \theta^*\|_2^2}{(N+\gamma)^2} + \frac{1}{\tau_\ell^2}\frac{\beta(P,\overline{\mathcal{L}})\log(1/\delta)}{N+\gamma}\right) \\ &= O\left(\frac{\tau_u^2\|\theta^1 - \theta^*\|_2^2\log(1/\delta)^2}{\tau_\ell^2 N^2} + \frac{1}{\tau_\ell^2}\frac{\sigma^2\log(1/\delta)}{N}\right).\end{aligned}$$

Therefore, by combining these equations, we obtain

$$\begin{aligned}\mathcal{R}(\theta^{N+1}) - \mathcal{R}(\theta^*) &\leq \frac{\tau_u}{2}\|\theta^{N+1} - \theta^*\|_2^2 \\ &\leq O\left(\frac{\tau_u^3\|\theta^1 - \theta^*\|_2^2\log(1/\delta)^2}{\tau_\ell^2 N^2} + \frac{\tau_u}{\tau_\ell^2}\frac{\sigma^2\log(1/\delta)}{N}\right) \\ &= O\left(\frac{\tau_u^3\left(\tau_\ell\|\theta^1 - \theta^*\|_2^2\right)\log(1/\delta)^2}{\tau_\ell^3 N^2} + \frac{\tau_u}{\tau_\ell^2}\frac{\sigma^2\log(1/\delta)}{N}\right) \\ &\leq O\left(\frac{\tau_u^3(\mathcal{R}(\theta^1) - \mathcal{R}(\theta^*))\log(1/\delta)^2}{\tau_\ell^3 N^2} + \frac{\tau_u}{\tau_\ell^2}\frac{\sigma^2\log(1/\delta)}{N}\right) \\ &= O\left(\frac{\tau_u^3 r_0\log(1/\delta)^2}{\tau_\ell^3 N^2} + \frac{\tau_u}{\tau_\ell^2}\frac{\sigma^2\log(1/\delta)}{N}\right).\end{aligned}$$

F.2 Proof of Corollary 3

Since $\nabla_\theta\overline{\mathcal{L}}(\theta, x) = \theta - x$, we have

$$\mathbb{E}_{z\sim P}\|\nabla_\theta\overline{\mathcal{L}}(\theta, x) - \nabla_\theta\mathcal{R}(\theta)\|_2^2 = \mathbb{E}_{z\sim P}\|x - \mathbb{E}[x]\|_2^2 = \operatorname{trace}(\Sigma). \tag{69}$$

Therefore, the corresponding $\alpha(P,\overline{\mathcal{L}})$ and $\beta(P,\overline{\mathcal{L}})$ in Eq. (3) are 0 and trace (Σ). Also, we note that $\tau_\ell = \tau_u = 1$ for the loss function $\overline{\mathcal{L}}(\theta, x) = \frac{1}{2}\|x - \theta\|_2^2$. By plugging these parameters to Theorem 1, we get the desired clipping level λ and the upper bound for $\|\theta^{N+1} - \theta^*\|_2$.

F.3 Proof of Corollary 4

From Lemma 7 of Prasad et al. [47], we have

$$\mathbb{E}[\nabla\overline{\mathcal{L}}(\theta, (x, y))] = \Sigma(\theta - \theta^*), \quad \text{and} \tag{70}$$

$$\|Cov(\nabla\overline{\mathcal{L}}(\theta, (x, y)))\|_2 \leq 2(C_4 + 1)\|\Sigma\|_2^2\|\theta - \theta^*\|_2^2 + \sigma^2\|\Sigma\|_2, \tag{71}$$

where $\Sigma = \mathbb{E}[xx^\top]$ is the covariance matrix of random variable X, $Cov(\nabla\overline{\mathcal{L}}(\theta, (x, y)))$ denotes the covariance matrix of $\nabla\overline{\mathcal{L}}(\theta, (x, y))$ and C_4 is the constant related to 4^{th} bounded moment defined in Eq. (12). Since we have

$$\mathbb{E}\|\nabla_\theta\overline{\mathcal{L}}(\theta, (x, y)) - \nabla_\theta\mathcal{R}(\theta)\|_2^2 = \operatorname{trace}\left(Cov(\nabla\overline{\mathcal{L}}(\theta, (x, y)))\right) \leq p\|Cov(\nabla\overline{\mathcal{L}}(\theta, (x, y)))\|_2.$$

We obtain $\alpha(P, \overline{\mathcal{L}}) = 2p(C_4 + 1)\|\Sigma\|_2^2$ and $\beta(P, \overline{\mathcal{L}}) = p\sigma^2\|\Sigma\|_2$ in Eq. (3). Also, by calculating the hessian matrix of population loss function $\mathcal{R}(\theta)$, we have $\tau_\ell = \lambda_{min}(\Sigma)$ and $\tau_u = \|\Sigma\|_2$. Finally, by plugging these values to Theorem 1, we got the desired bound and hyper-parameters.