Stochastic Optimization with Heavy-Tailed Noise via Accelerated Gradient Clipping

Eduard Gorbunov[*]
MIPT and HSE, Russia

Marina Danilova[†]
ICS RAS and MIPT, Russia

Alexander Gasnikov[‡]
MIPT and HSE, Russia

Abstract

In this paper, we propose a new accelerated stochastic first-order method called `clipped-SSTM` for smooth convex stochastic optimization with heavy-tailed distributed noise in stochastic gradients and derive the first high-probability complexity bounds for this method closing the gap in the theory of stochastic optimization with heavy-tailed noise. Our method is based on a special variant of accelerated Stochastic Gradient Descent (`SGD`) and clipping of stochastic gradients. We extend our method to the strongly convex case and prove new complexity bounds that outperform state-of-the-art results in this case. Finally, we extend our proof technique and derive the first non-trivial high-probability complexity bounds for `SGD` with clipping without light-tails assumption on the noise.

1 Introduction

In this paper we focus on the following problem

$$\min_{x\in\mathbb{R}^n} f(x), \quad f(x) = \mathbb{E}_\xi\left[f(x,\xi)\right], \tag{1}$$

where $f(x)$ is a smooth convex function and the mathematical expectation in (1) is taken with respect to the random variable ξ defined on the probability space $(\mathcal{X}, \mathcal{F}, \mathbb{P})$ with some σ-algebra $\mathcal{F}$ and probability measure $\mathbb{P}$. Such problems appear in various applications of machine learning [21, 61, 64] and mathematical statistics [66]. Perhaps, the most popular method to solve problems like (1) is Stochastic Gradient Descent (`SGD`) [26, 50, 51, 59, 63]. There is a lot of literature on the convergence in expectation of `SGD` for (strongly) convex [20, 24, 25, 46, 48, 49, 55] and non-convex [6, 20, 34] problems under different assumptions on stochastic gradient. When the problem is good enough, i.e. when the distributions of stochastic gradients are *light-tailed*, this theory correlates well with the real behavior of trajectories of `SGD` in practice. Moreover, the existing *high-probability* bounds for `SGD` [9, 11, 49] coincide with its counterpart from the theory of convergence in expectation up to logarithmical factors depending on the confidence level.

However, there are a lot of important applications where the noise distribution in the stochastic gradient is significantly *heavy-tailed* [65, 71]. For such problems `SGD` is often less robust and shows poor performance in practice. Furthermore, existing results for the convergence with high-probability for `SGD` are also much worse in the presence of heavy-tailed noise than its "light-tailed counterparts". In this case, rates of the convergence in expectation can be insufficient to describe the behavior of the method.

To illustrate this phenomenon we consider a simple example of stochastic optimization problem and apply `SGD` with constant stepsize to solve it. After that, we present a natural and simple way to resolve the issue of `SGD` based on the *clipping* of stochastic gradients. However, we need to introduce some important notations and definitions before we start to discuss this example.

[*] `eduard.gorbunov@phystech.edu`, `eduardgorbunov.github.io`
[†] `danilovamarina15@gmail.com`, `marinadanya.github.io`
[‡] `gasnikov@yandex.ru`

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

1.1 Preliminaries

In this section we introduce the main part of notations, assumption and definitions. The rest is classical for optimization literature and stated in the appendix (see Section A). Throughout the paper we assume that at each point $x \in \mathbb{R}^n$ function f is accessible only via stochastic gradients $\nabla f(x, \xi)$ such that

$$\mathbb{E}_\xi[\nabla f(x,\xi)] \quad = \quad \nabla f(x), \quad \mathbb{E}_\xi\left[\|\nabla f(x,\xi) - \nabla f(x)\|_2^2\right] \leq \sigma^2, \tag{2}$$

i.e. we have an access to the unbiased estimator of $\nabla f(x)$ with uniformly bounded by σ^2 variance where σ is some non-negative number. These assumptions on the stochastic gradient are standard in the stochastic optimization literature [18, 20, 31, 38, 49]. Below we introduce one of the most important definitions in this paper.

Definition 1.1 (light-tailed random vector)**.** We say that random vector η has a light-tailed distribution, i.e. satisfies "light-tails" assumption, if there exist $\mathbb{E}[\eta]$ and $\mathbb{P}\left\{\|\eta - \mathbb{E}[\eta]\|_2 > b\right\} \leq 2\exp\left(-\frac{b^2}{2\sigma^2}\right)$ for all $b > 0$

Such distributions are often called sub-Gaussian ones (see [30] and references therein). One can show (see Lemma 2 from [30]) that this definition is equivalent to

$$\mathbb{E}\left[\exp\left(\|\eta - \mathbb{E}[\eta]\|_2^2/\sigma^2\right)\right] \leq \exp(1) \tag{3}$$

up to absolute constant difference in σ. Due to Jensen's inequality and convexity of $\exp(\cdot)$ one can easily show that inequality (3) implies $\mathbb{E}[\|\eta - \mathbb{E}[\eta]\|_2^2] \leq \sigma^2$. However, the reverse implication does not hold in general. Therefore, in the rest of the paper by stochastic gradient with heavy-tailed distribution, we mean such a stochastic gradient that satisfies (2) but not necessarily (3).

1.2 Simple Motivational Example: Convergence in Expectation and Clipping

In this section we consider SGD $x^{k+1} = x^k - \gamma \nabla f(x^k, \xi^k)$ applied to solve the problem (1) with $f(x, \xi) = \|x\|_2^2/2 + \langle \xi, x \rangle$, where ξ is a random vector with zero mean and the variance by σ^2 (see the details in Section H.1). The state-of-the-art theory (e.g. [24, 25]) says that convergence properties in expectation of SGD in this case depend only on the stepsize γ, condition number of f, initial suboptimality $f(x^0) - f(x^*)$ and the variance σ, but does not depend on distribution of ξ. However, the trajectory of SGD significantly depends on the distribution of ξ. To illustrate this we consider 3 different distributions of ξ with the same σ, i.e., Gaussian distribution, Weibull distribution [69] and Burr Type XII distribution [3, 42] with proper shifts and scales to get needed mean and variance for ξ (see the details in Section H.1). For each distribution, we run SGD several times from the same starting point, the same stepsize γ, and the same batchsize, see typical runs in Figure 1. This simple

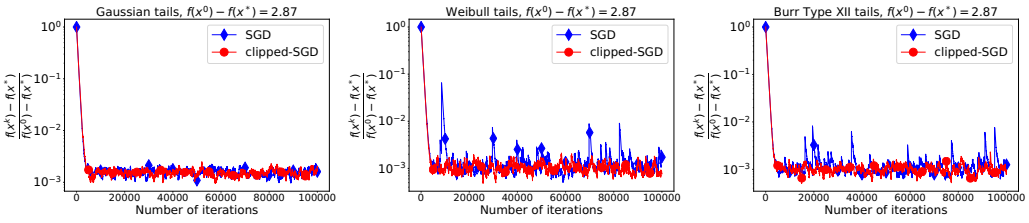

Figure 1: Typical trajectories of SGD and clipped-SGD applied to solve (130) with ξ having Gaussian, Weibull, and Burr Type XII tails.

example shows that SGD in all 3 cases rapidly reaches a neighborhood of the solution and then starts to oscillate there. However, these oscillations are significantly larger for the second and the third cases where stochastic gradients are heavy-tailed. Unfortunately, guarantees for the convergence in expectation cannot express this phenomenon, since in expectation the convergence guarantees for all 3 cases are identical.

Moreover, in practice, e.g., in training big machine learning models, it is often used only a couple runs of SGD or another stochastic method. The training process can take hours or even days, so, it is extremely important to obtain good accuracy of the solution *with high probability*. However,

as our simple example shows, SGD fails to converge robustly if the noise in stochastic gradients is heavy-tailed which was also noticed for several real-world problems like training AlexNet [37] on CIFAR10 [36] (see [65]) and training an attention model [68] via BERT [8] (see [71]).

Clearly, since the distributions of stochastic gradients in the second and the third cases are heavy tailed the probability of sampling too large ξ (in terms of the norm) and, as a consequence, too large $\nabla f(x, \xi)$ is high even if we are close to the solution. Once the current point x^k is not too far from the solution and SGD gets a stochastic gradient with too large norm the method jumps far from the solution. Therefore, we see large oscillations. Since the reason of such oscillations is large norm of stochastic gradient it is natural to *clip* it, i.e., update x^{k+1} according to $x^{k+1} = x^k - \gamma \min\{1, \lambda/\|\nabla f(x^k, \xi^k)\|_2\} \nabla f(x^k, \xi^k)$. The obtained method is known in literature as clipped-SGD (see [17, 21, 43, 44, 57, 70, 71] and references therein). Among the good properties of clipped-SGD we emphasize its robustness to the heavy-tailed noise in stochastic gradients (see also [71]). In our tests, trajectories of clipped-SGD oscillate not significantly even for heavy-tailed distributions, and clipping does not spoil the rate of convergence. These two factors make clipped-SGD preferable than SGD when we deal with heavy-tailed distributed stochastic gradients (see further discussion in Section B.2).

1.3 Related Work

1.3.1 Smooth Stochastic Optimization: Light-Tailed Noise

In the light-tailed case high-probability complexity bounds and complexity bounds in expectation for SGD and AC-SA differ only in logarithmical factors of $1/\beta$, see the details in Table 1. Such bounds were obtained in [9] for SGD in the convex case and then were extended to the μ-strongly convex case in [11] for modification of SGD called Stochastic Intermediate Gradient Method (SIGM). Finally, optimal complexities were derived in [18, 19, 38] for the method called AC-SA in the convex case and for Multi-Staged AC-SA (MS-AC-SA) in the strongly convex case.

1.3.2 Smooth Stochastic Optimization: Heavy-Tailed Noise

Without light tails assumption the most straightforward results lead to $O(1/\beta^2)$ and $O(1/\beta)$ dependency on β in the complexity bounds. Such bounds can be obtained from the complexity bounds for the convergence in expectation via Markov's inequality. However, for small β these bounds become unacceptably poor. Classical results [13, 53, 62] reduce these dependence to $O(\ln(\beta^{-1}))$ but they have worse dependence on ε than corresponding results relying on light tails assumption.

For a long time the following question was open: *is it possible to design stochastic methods having the same or comparable complexity bounds as in the light-tailed case but without light tails assumption on stochastic gradients?* In [47] and [7] the authors give a positive answer to this question *but only partially*. Let us discuss the results from these papers in detail.

In [47] Nazin et al. develop a new algorithm called Robust Stochastic Mirror Descent (RSMD) which is based on a special truncation of stochastic gradients and derive complexity guarantees similar to SGD in the convex case but without light assumption, see Table 1. This technique is very similar to gradient clipping. Moreover, in [47] authors consider also composite problems with non-smooth composite term. However, in [47] the optimization problem is defined on some *compact* convex set X with diameter $\Theta = \max\{\|x - y\|_2 \mid x, y \in X\} < \infty$ and the analysis depends substantially on the boundedness of X. Using special restarts technique together with iterative squeezing of the set X Nazin et al. extend their method to the μ-strongly convex case, see Table 2. Finally, in the discussion section of [47] authors formulate the following question: *is it possible to develop such* ***accelerated*** *stochastic methods that have the same or comparable complexity bounds as in the light-tailed case but do not require stochastic gradients to be light-tailed?*

In the strongly convex case the positive answer to this question was given by Davis et al. [7] where authors propose a new method called proxBoost that is based on robust distance estimation [29, 51] and proximal point method [40, 41, 60], see Table 2. However, this approach requires solving an auxiliary optimization problem at each iteration that can lead to poor performance in practice.

In our paper we close the gap in theory, i.e., we provide a positive answer to the following question: *Is it possible to develop such an accelerated stochastic method that have the same or comparable*

complexity bound as for `AC-SA` *in the convex case but do not require stochastic gradients to be light-tailed?*

1.4 Our Contributions

- One of the main contributions of our paper is a new method called Clipped Stochastic Similar Triangles Method (`clipped-SSTM`). For the case when the objective function f is convex and L-smooth we derive the following complexity bound *without light tails assumption on the stochastic gradients:* $O(\max\{\sqrt{LR_0^2/\varepsilon}, \sigma^2 R_0^2/\varepsilon^2\}\ln(LR_0^2/\varepsilon\beta))$. This bound outperforms all known bounds for this setting (see Table 1) and up to the difference in logarithmical factors recovers the complexity bound of `AC-SA` derived under light tails assumption. That is, in this paper we close the gap in theory theory of smooth convex stochastic optimization with heavy-tailed noise. Moreover, unlike in [47], we do not assume boundedness of the set where the optimization problem is defined, which makes our analysis more complicated. We also study different batchsize policies for `clipped-SSTM`.
- Using restarts technique we extend `clipped-SSTM` to the μ-strongly convex objectives and obtain a new method called Restarted `clipped-SSTM` (`R-clipped-SSTM`). For this method we prove the following complexity bound (again, *without light tails assumption on the stochastic gradients*): $O(\max\{\sqrt{L/\mu}\ln(\mu R^2/\varepsilon), \sigma^2/\mu\varepsilon\}\ln(L/\mu\beta\ln(\mu R^2/\varepsilon)))$. Our bound outperforms the state-of-the-art result from [7] in terms of the dependence on $\ln\frac{L}{\mu}$, see Table 2 for the details.
- We prove the first high-probability complexity guarantees for `clipped-SGD` in convex and strongly convex cases *without light tails assumption on the stochastic gradients*, see Tables 1 and 2. The complexity we prove for `clipped-SGD` in the convex case is comparable with corresponding bound for `SGD` derived under light tails assumption. In the μ-strongly convex case we derive a new complexity bound for the restarted version of `clipped-SGD` (`R-clipped-SGD`) which is comparable with its "light-tailed counterpart".
- We conduct several numerical experiments with the proposed methods in order to justify the theory we develop. In particular, we show that `clipped-SSTM` can outperform `SGD` and `clipped-SGD` in practice even without using large batchsizes. Moreover, in our experiments we illustrate how clipping makes the convergence of `SGD` and `SSTM` more robust and reduces their oscillations.

Table 1: Comparison of existing high-probability convergence results for stochastic optimization under assumptions (2) for convex and L-smooth objectives. The second column contains an overall number of stochastic first-order oracle calls needed to achieve ε-solution with probability at least $1-\beta$. In the third column "light" means that $\nabla f(x,\xi)$ satisfies (3) and "heavy" means that the result holds even in the case when (3) does not hold. Column "Domain" describes the set where the optimization problem is defined. For `RSMD` Θ is a diameter of the set where the optimization problem is defined. We use red color to emphasize the restrictions we eliminate.

Method	Complexity	Tails	Domain
`SGD` [9]	$O\left(\max\left\{\frac{LR_0{}^2}{\varepsilon}, \frac{\sigma^2 R_0{}^2}{\varepsilon^2}\ln^2(\beta^{-1})\right\}\right)$	light	bounded
`AC-SA` [18, 38]	$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\ln(\beta^{-1})\right\}\right)$	light	arbitrary
`RSMD` [47]	$O\left(\max\left\{\frac{L\Theta^2}{\varepsilon}, \frac{\sigma^2\Theta^2}{\varepsilon^2}\right\}\ln(\beta^{-1})\right)$	heavy	bounded
`clipped-SGD` [This work]	$O\left(\max\left\{\frac{LR_0{}^2}{\varepsilon}, \frac{\sigma^2 R_0{}^2}{\varepsilon^2}\right\}\ln(\beta^{-1})\right)$	heavy	$\mathbb{R}^n$
`clipped-SSTM` [This work]	$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\}\ln\frac{LR_0^2+\sigma R_0}{\varepsilon\beta}\right)$	heavy	$\mathbb{R}^n$

1.4.1 Relation to [71]

While Zhang et al. [71] consider different setup, [71] is highly relevant to our paper, and, in some sense, it complements our findings. In particular, it contains the analysis of several versions of

Table 2: Comparison of existing high-probability convergence results for stochastic optimization under assumptions (2) for μ-strongly convex and L-smooth objectives. The second column contains an overall number of stochastic first-order oracle calls needed to achieve ε-solution with probability at least $1-\beta$. In the third column "light" means that $\nabla f(x,\xi)$ satisfies (3) and "heavy" means that the result holds even in the case when (3) does not hold. Column "Domain" describes the set where the optimization problem is defined. For `RSMD` Θ is a diameter of the set where the optimization problem is defined and $R = \sqrt{2(f(x^0)-f(x^*))/\mu}$, $r_0 = f(x^0) - f(x^*)$. We use red color to emphasize the restrictions we eliminate.

Method	Complexity	Tails	Domain
`SIGM` [11]	$O\left(\max\left\{\frac{L}{\mu}\ln\frac{\mu R_0^2}{\varepsilon}, \frac{\sigma^2}{\mu\varepsilon}\ln\left(\beta^{-1}\ln\frac{\mu R_0^2}{\varepsilon}\right)\right\}\right)$	light	arbitrary
`MS-AC-SA` [19]	$O\left(\max\left\{\sqrt{\frac{L}{\mu}}\ln\frac{LR_0^2}{\varepsilon}, \frac{\sigma^2}{\mu\varepsilon}\ln\left(\beta^{-1}\ln\frac{LR_0^2}{\varepsilon}\right)\right\}\right)$	light	arbitrary
`restarted-RSMD` [47]	$O\left(\max\left\{\frac{L}{\mu}\ln\left(\frac{\mu\Theta^2}{\varepsilon}\right), \frac{\sigma^2}{\mu\varepsilon}\right\}\ln\left(\beta^{-1}\ln\frac{\mu\Theta^2}{\varepsilon}\right)\right)$	heavy	bounded
`proxBoost` [7]	$O\left(\max\left\{\sqrt{\frac{L}{\mu}}\ln\left(\frac{LR_0^2\ln\frac{L}{\mu}}{\varepsilon}\right), \frac{\sigma^2\ln\frac{L}{\mu}}{\mu\varepsilon}\right\}\cdot C\right)$, where $C = \ln\left(\frac{L}{\mu}\right)\ln\left(\frac{\ln\frac{L}{\mu}}{\beta}\right)$	heavy	arbitrary
`clipped-SGD` [This work]	$O\left(\max\left\{\frac{L}{\mu}, \frac{\sigma^2}{\mu\varepsilon}\cdot\frac{L}{\mu}\right\}\ln\left(\frac{r_0}{\varepsilon}\right)\ln\left(\frac{L}{\mu\beta}\ln\frac{r_0}{\varepsilon}\right)\right)$	heavy	$\mathbb{R}^n$
`R-clipped-SGD` [This work]	$O\left(\max\left\{\frac{L}{\mu}\ln\frac{\mu R^2}{\varepsilon}, \frac{\sigma^2}{\mu\varepsilon}\right\}\ln\left(\frac{L}{\mu\beta}\ln\frac{\mu R^2}{\varepsilon}\right)\right)$	heavy	$\mathbb{R}^n$
`R-clipped-SSTM` [This work]	$O\left(\max\left\{\sqrt{\frac{L}{\mu}}\ln\frac{\mu R^2}{\varepsilon}, \frac{\sigma^2}{\mu\varepsilon}\right\}\ln\left(\frac{L}{\mu\beta}\ln\frac{\mu R^2}{\varepsilon}\right)\right)$	heavy	$\mathbb{R}^n$

`clipped-SGD` establishing the rates of convergence *in expectation* while we focus on the *high-probability* complexity guarantees. Secondly, we consider convex and strongly convex cases while [71] provides an analysis for non-convex and strongly convex problems. Finally, [71] relies on the following assumption: there exist such $G > 0$ and $\alpha \in (1, 2]$ that the stochastic gradient $g(x)$ satisfies $\mathbb{E}\|g(x)\|_2^\alpha \leq G^\alpha$. This assumption implies the boundedness of the gradient of the objective function $f(x)$ which is quite restrictive and does not hold on the whole space for strongly convex functions. In our paper, we assume only boundedness of the variance. Moreover, we consider *smooth* problems that allows us to accelerate `clipped-SGD` and obtain `clipped-SSTM`, while Zhang et al. [71] provide non-accelerated rates.

1.5 Paper Organization

The remaining part of the paper is organized as follows. In Section 2 we present `clipped-SSTM` together with the main complexity result in the convex case that we prove for this method. Then, we present the first high-probability complexity bounds for `clipped-SGD` for for the convex problems. In Section 4 we provide our numerical experiments justifying our theoretical results. Finally, in Section 5 we provide some concluding remarks and discuss the limitations and possible extensions of the results developed in the paper. Due to the space limitations, we put the exact formulations of all theorems, results for the strongly convex problems and the full proofs in the Appendix (see Sections F and G), together with auxiliary and technical results and additional experiments (see Section H). Moreover, in Section F.1.2 we present a sketch of the proof of the main convergence result for `clipped-SSTM` and explain the intuition behind it.

2 Accelerated `SGD` with Clipping

In this section we consider the situation when $f(x)$ is convex and L-smooth on $\mathbb{R}^n$. For this problem we present a new method called Clipped Stochastic Similar Triangles Method (`clipped-SSTM`, see Algorithm 1). In our method we use a clipped stochastic gradient that is defined in the following way:

$$\text{clip}(\nabla f(x,\boldsymbol{\xi}),\lambda) = \min\left\{1, \lambda/\|\nabla f(x,\boldsymbol{\xi})\|_2\right\}\nabla f(x,\boldsymbol{\xi}) \tag{4}$$

Algorithm 1 Clipped Stochastic Similar Triangles Method (`clipped-SSTM`)

Input: starting point x^0, number of iterations N, batchsizes $\{m_k\}_{k=1}^N$, stepsize parameter a, clipping parameter B

Set $A_0 = \alpha_0 = 0$, $y^0 = z^0 = x^0$
for $k = 0, \ldots, N-1$ **do**
  Set $\alpha_{k+1} = \frac{k+2}{2aL}$, $A_{k+1} = A_k + \alpha_{k+1}$, $\lambda_{k+1} = \frac{B}{\alpha_{k+1}}$
  $x^{k+1} = {}^{(A_k y^k + \alpha_{k+1} z^k)}/_{A_{k+1}}$
  Draw fresh i.i.d. samples $\xi_1^k, \ldots, \xi_{m_k}^k$ and compute $\nabla f(x^{k+1}, \boldsymbol{\xi}^k) = \frac{1}{m_k}\sum_{i=1}^{m_k} \nabla f(x^{k+1}, \xi_i^k)$
  Compute $\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) = \text{clip}(\nabla f(x^{k+1}, \boldsymbol{\xi}^k), \lambda_{k+1})$ using (4)
  $z^{k+1} = z^k - \alpha_{k+1}\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)$
  $y^{k+1} = {}^{(A_k y^k + \alpha_{k+1} z^{k+1})}/_{A_{k+1}}$
end for

Output: y^N

where $\nabla f(x, \boldsymbol{\xi}) = \frac{1}{m}\sum_{i=1}^m \nabla f(x, \xi_i)$ is a mini-batched version of $\nabla f(x)$. That is, in order to compute $\text{clip}(\nabla f(x, \boldsymbol{\xi}), \lambda)$ one needs to get m i.i.d. samples $\nabla f(x, \xi_1), \ldots, \nabla f(x, \xi_m)$, compute its average and then project the result $\nabla f(x, \boldsymbol{\xi})$ on the Euclidean ball with radius λ and center at the origin. Next theorem summarizes the main convergence result for `clipped-SSTM`.

Theorem 2.1. Assume that function f is convex and L-smooth. Then for all $\beta \in (0, 1)$ and $N \geq 1$ such that $\ln({}^{4N}/_{\beta}) \geq 2$ we have that after N iterations of `clipped-SSTM` with $m_k = \Theta\left(\max\left\{1, {}^{\sigma^2 \alpha_{k+1}^2 N \ln({}^{N}/_{\beta})}/_{R_0^2}\right\}\right)$, $B = \Theta({}^{R_0}/_{\ln({}^{N}/_{\beta})})$ and $a = \Theta(\ln^2({}^{N}/_{\beta}))$ that $f(y^N) - f(x^*) = O({}^{aLR_0^2}/_{N^2})$ holds with probability at least $1 - \beta$ where $R_0 = \|x^0 - x^*\|_2$. In other words, if we choose a to be equal to the maximum from (27), then the method achieves $f(y^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ after $O(\sqrt{{}^{LR_0^2}/_{\varepsilon}} \ln({}^{LR_0^2}/_{\varepsilon\beta}))$ iterations and requires $O(\max\{\sqrt{{}^{LR_0^2}/_{\varepsilon}}, {}^{\sigma^2 R_0^2}/_{\varepsilon^2}\} \ln({}^{LR_0^2}/_{\varepsilon\beta}))$ oracle calls.

The theorem says that for any $\beta \in (0, 1)$ `clipped-SSTM` converges to ε-solution with probability at least $1 - \beta$ and requires exactly the same number of stochastic first-order oracle calls (up to the difference in constant and logarithmical factors) as optimal stochastic methods like `AC-SA` [18, 38] or Stochastic Similar Triangles Method [16, 22]. However, our method *achieves this rate under less restrictive assumption*. Indeed, Theorem 2.1 holds even in the case when the stochastic gradient $\nabla f(x, \xi)$ satisfies only (2) and can have *heavy-tailed* distribution. In contrast, all existing results that establish (30) and that are known in the literature hold only in the light-tails case, see Section 1.3.1.

Finally, when σ^2 is big then Theorem 2.1 says that at iteration k `clipped-SGD` requires large batchsizes $m_k \sim k^2 N$ (see (26)) which is proportional to $\varepsilon^{-3/2}$ for last iterates. It can make the cost of one iteration extremely high, therefore, we also consider different stepsize policies that remove this drawback in Section F.1.1. In particular, the following result shows that `clipped-SSTM` achieves the same oracle complexity even with constant batchsizes m_k when stepsize parameter a is chosen properly.

Corollary 2.2. Let the assumptions of Theorem F.1 hold and $a = \Theta\left(\max\{1, \ln^2({}^{N}/_{\beta}), \sqrt{{}^{\ln N/\beta}}\sigma N^{3/2}/_{LR_0}\}\right)$. Then $m_k = O(1)$ and `clipped-SSTM` achieves $f(y^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ after $O(\max\{\sqrt{{}^{LR_0^2}/_{\varepsilon}}, {}^{\sigma^2 R_0^2}/_{\varepsilon^2}\} \ln({}^{(LR_0^2 + \sigma R_0)}/_{\varepsilon\beta}))$ iterations/oracle calls.

3 `SGD` with Clipping

In this section we present our complexity results for `clipped-SGD` (see Algorithm 2) in the convex case. Next theorem summarizes the main convergence result for `clipped-SGD` in this case.

Algorithm 2 Clipped Stochastic Gradient Descent (`clipped-SGD`)

Input: starting point x^0, number of iterations N, batchsizes $\{m_k\}_{k=0}^{N-1}$, stepsize $\gamma > 0$, clipping level $\lambda > 0$

1: **for** $k = 0, \ldots, N-1$ **do**
2: Draw fresh i.i.d. samples $\xi_1^k, \ldots, \xi_{m_k}^k$ and compute $\nabla f(x^k, \boldsymbol{\xi}^k) = \frac{1}{m_k} \sum_{i=1}^{m_k} \nabla f(x^k, \xi_i^k)$
3: Compute $\widetilde{\nabla} f(x^k, \boldsymbol{\xi}^k) = \text{clip}(\nabla f(x^k, \boldsymbol{\xi}^k), \lambda)$ using (4)
4: $x^{k+1} = x^k - \gamma \widetilde{\nabla} f(x^k, \boldsymbol{\xi}^k)$
5: **end for**

Output: $\bar{x}^N = \frac{1}{N} \sum_{k=0}^{N-1} x^k$

Theorem 3.1. Assume that function f is convex and L-smooth. Then for all $\beta \in (0,1)$ and $N \geq 1$ such that $\ln(4N/\beta) \geq 2$ we have that after N iterations of `clipped-SGD` with $\lambda = \Theta(LR_0)$ and $m_k = m = \Theta(\max\{1, N\sigma^2/R_0^2 L^2 \ln(N/\beta)\})$ where $R_0 = \|x^0 - x^*\|_2$ and stepsize $\gamma = 1/80L\ln(4N/\beta)$ that $f(\bar{x}^N) - f(x^*) = O(LR_0^2 \ln(4N/\beta)/N)$ with probability at least $1-\beta$ where $\bar{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k$. In other words, the method achieves $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$ after $O\left(LR_0^2/\varepsilon \ln(LR_0^2/\varepsilon\beta)\right)$ iterations and requires $O(\max\{LR_0^2/\varepsilon, \sigma^2 R_0^2/\varepsilon^2\} \ln(LR_0^2/\varepsilon\beta))$ oracle calls.

To the best of our knowledge, it is the first result for `clipped-SGD` establishing non-trivial complexity guarantees for the convergence with high probability. Up to the difference in logarithmical factors our bound recovers the complexity bound for `SGD` which was obtained under light tails assumption and the complexity bound for `RSMD`. However, unlike in [47], we do not assume that the optimization problem is defined on the bounded set. The proof technique is similar to one we use to prove Theorem F.1. One can find the full proof in Section G.3.1.

4 Numerical Experiments

We have tested[4] `clipped-SSTM` and `clipped-SGD` on the logistic regression problem, the datasets were taken from LIBSVM library [4]. To implement methods we use Python 3.7 and standard libraries. One can find additional experiments and details in Section H.2.

First of all, using standard solvers from `scipy` library we find good enough approximation of the solution of the problem for each dataset. For simplicity, we denote this approximation by x^*. Then, we numerically study the distribution of $\|\nabla f_i(x^*)\|_2$ and plot corresponding histograms for each dataset, see Figure 2. These histograms hint that near the solution for `heart` dataset tails of stochastic

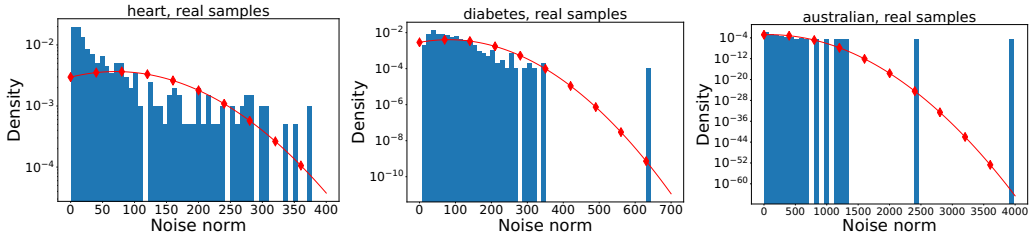

Figure 2: Histograms of $\|\nabla f_i(x^*)\|_2$ for different datasets. Red lines correspond to probability density functions of normal distributions with empirically estimated means and variances.

gradients are not heavy and the norm of the noise can be well-approximated by Gaussian distribution, whereas for `diabetes` and `australian` we see the presense of outliers that makes the distribution heavy-tailed.

Next, let us consider numerical results for `SGD` and `SSTM` with and without clipping applied to solve logistic regression problem on these 3 datasets, see Figures 3- 5. For all methods we used constant batchsizes m, stepsizes and clipping levels were tuned, see Section H.2 for the details. In our experiments we also consider `clipped-SGD` with periodically decreasing clipping level λ

[4] One can find the code here: `https://github.com/eduardgorbunov/accelerated_clipping`.

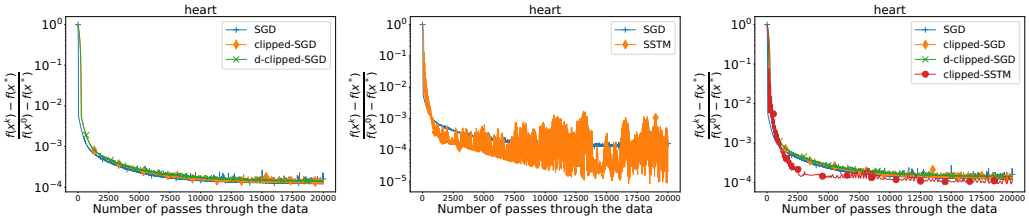

Figure 3: Trajectories of SGD, clipped-SGD, SSTM and clipped-SSTM applied to solve logistic regression problem on heart dataset.

(d-clipped-SGD in Figures), i.e. the method starts with some initial clipping level λ_0 and after every l epochs or, equivalently, after every $\lceil rl/m \rceil$ iterations the clipping level is multiplied by some constant $\alpha \in (0, 1)$.

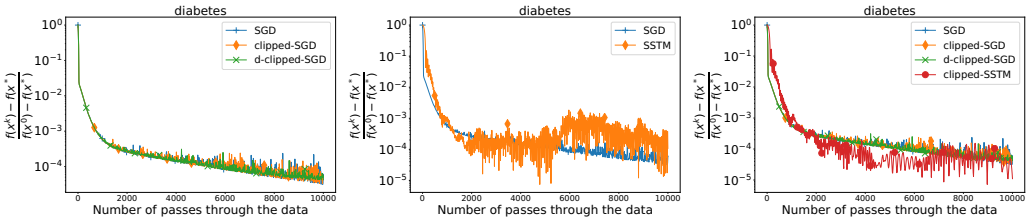

Figure 4: Trajectories of SGD, clipped-SGD, SSTM and clipped-SSTM applied to solve logistic regression problem on diabetes dataset.

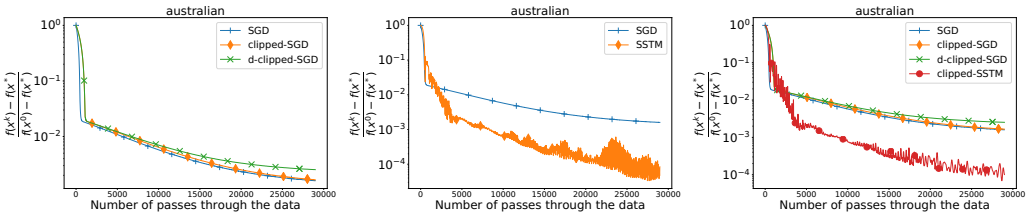

Figure 5: Trajectories of SGD, clipped-SGD, SSTM and clipped-SSTM applied to solve logistic regression problem on australian dataset.

Let us discuss the obtained numerical results. First of all, d-clipped-SGD stabilizes the oscillations of SGD even if the initial clipping level was high. In contrast, clipped-SGD with too large clipping level λ behaves similarly to SGD. Secondly, we emphasize that due to the fact that we used small bathcsizes SSTM has very large oscillations in comparison to SGD. Actually, fast error/noise accumulation is a typical drawback of accelerated SGD with small batchsizes [35]. Moreover, deterministic accelerated and momentum-based methods often have non-monotone behavior (see [5] and references therein). However, to some extent clipped-SSTM suffers from the first drawback less than SSTM and has comparable convergence rate with SSTM. Finally, in our experiments on heart and australian datasets clipped-SSTM converges faster than SGD and clipped-SGD and oscillates little, while on diabetes dataset it also converges faster than SGD, but oscillates more if parameter B is not fine-tuned.

We also want to mention that the behavior of SGD on heart and diabetes datasets correlates with the insights from Section 1.2 and our numerical study of the distribution of $\|\nabla f_i(x^*)\|_2$. Indeed, for heart dataset SGD has little oscillations since the distribution of $\|\nabla f_i(x^k) - \nabla f(x^k)\|_2$, where x^k is the last iterate, is well concentrated near its mean and can be approximated by Gaussian distribution (see the details in Section H.2). In contrast, Figure 4 shows that SGD oscillates more than in the previous example. One can explain such behavior using Figure 2 showing that the distribution of $\|\nabla f(x^*)\|_2$ has heavier tails than for heart dataset.

However, we do not see any oscillations of SGD for australian dataset despite the fact that according to Figure 2 the distribution of $\|\nabla f_i(x^*)\|_2$ in this case has heavier tails than in previous examples. Actually, there is no contradiction and in this case it simply means that SGD does not get close to the solution in terms of functional value, despite the fact that we used $\gamma = 1/L$. In Section H.2 we present the results of different tests where we tried to use bigger stepsize γ in order to reach oscillation region faster and show that in fact in that region SGD oscillates significantly more, but clipping fixes this issue without spoiling the convergence rate.

5 Discussion

In this paper we close the gap in the theory of high-probability complexity bounds for stochastic optimization with heavy-tailed noise. In particular, we propose a new accelerated stochastic method — clipped-SSTM — and prove the first accelerated high-probability complexity bounds for smooth convex stochastic optimization without light-tails assumption. Moreover, we extend our results to the strongly convex case and prove new complexity bounds outperforming the state-of-the-art results. Finally, we derive first high-probability complexity bounds for the popular method called clipped-SGD in convex and strongly convex cases and conduct a numerical study of the considered methods.

However, our approach has several limitations. In particular, it significantly relies on the assumption that the optimization problem is defined on $\mathbb{R}^n$. Moreover, we do not consider regularized or composite problems like in [47] and [7]. However, in [47] it is significant in the analysis that the set where the problem is defined is bounded and in [7] the analysis works only for the strongly convex problems. It would also be interesting to generalize our approach to generally non-smooth problems using the trick from [52].

Broader Impact

Our contribution is primarily theoretical. Therefore, a broader impact discussion is not applicable.

Acknowledgments and Disclosure of Funding

The research of E. Gorbunov and A. Gasnikov was partially supported by the Ministry of Science and Higher Education of the Russian Federation (Goszadaniye) 075-00337-20-03, project No. 0714-2020-0005. The research of Marina Danilova was funded by RFBR, project number 20-31-90073.

References

[1] George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

[2] Aleksandr Alekseevich Borovkov and Konstantin Aleksandrovich Borovkov. On probabilities of large deviations for random walks. i. regularly varying distribution tails. *Theory of Probability & Its Applications*, 46(2):193–213, 2002.

[3] Irving W Burr. Cumulative frequency functions. *The Annals of mathematical statistics*, 13(2):215–232, 1942.

[4] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

[5] Marina Danilova, Anastasiia Kulakova, and Boris Polyak. Non-monotone behavior of the heavy ball method. In Martin Bohner, Stefan Siegmund, Roman Šimon Hilscher, and Petr Stehlík, editors, *Difference Equations and Discrete Dynamical Systems with Applications*, pages 213–230, Cham, 2020. Springer International Publishing.

[6] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[7] Damek Davis, Dmitriy Drusvyatskiy, Lin Xiao, and Junyu Zhang. From low probability to high confidence in stochastic convex optimization. *arXiv preprint arXiv:1907.13307*, 2019.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Olivier Devolder et al. Stochastic first order methods in smooth convex optimization. Technical report, CORE, 2011.

[10] Pavel Dvurechenskii, Darina Dvinskikh, Alexander Gasnikov, Cesar Uribe, and Angelia Nedich. Decentralize and randomize: Faster algorithm for wasserstein barycenters. In *Advances in Neural Information Processing Systems*, pages 10760–10770, 2018.

[11] 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.

[12] Kacha Dzhaparidze and JH Van Zanten. On bernstein-type inequalities for martingales. *Stochastic processes and their applications*, 93(1):109–117, 2001.

[13] O Bousquet A Elisseeff and Olivier Bousquet. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[14] David A Freedman et al. On tail probabilities for martingales. *the Annals of Probability*, 3(1):100–118, 1975.

[15] Alexander Gasnikov, Pavel Dvurechensky, and Yurii Nesterov. Stochastic gradient methods with inexact oracle. *arXiv preprint arXiv:1411.4218*, 2014.

[16] Alexander Gasnikov and Yurii Nesterov. Universal fast gradient method for stochastic composit optimization problems. *arXiv:1604.05275*, 2016.

[17] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org, 2017.

[18] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[19] 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.

[20] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

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

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

[24] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. *arXiv preprint arXiv:1905.11261*, 2019.

[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, 2019.

[26] Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.

[27] 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.

[28] Elad Hazan, Kfir Levy, and Shai Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. In *Advances in Neural Information Processing Systems*, pages 1594–1602, 2015.

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

[30] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

[31] Anatoli Juditsky, Arkadi Nemirovski, et al. First order methods for nonsmooth convex large-scale optimization, i: general purpose methods. *Optimization for Machine Learning*, pages 121–148, 2011.

[32] Anatoli Juditsky and Yuri Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4(1):44–80, 2014.

[33] Sham M Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems*, pages 801–808, 2009.

[34] Ahmed Khaled and Peter Richtárik. Better theory for sgd in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

[35] Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

[36] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 and cifar-100 datasets. *URl: https://www. cs. toronto. edu/kriz/cifar. html*, 6, 2009.

[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[38] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

[39] Kfir Y Levy. The power of normalization: Faster evasion of saddle points. *arXiv preprint arXiv:1611.04831*, 2016.

[40] Bernard Martinet. Régularisation d'inéquations variationnelles par approximations successives. rev. française informat. *Recherche Opérationnelle*, 4:154–158, 1970.

[41] Bernard Martinet. Détermination approchée d'un point fixe d'une application pseudo-contractante. *CR Acad. Sci. Paris*, 274(2):163–165, 1972.

[42] Michael P McLaughlin. *A compendium of common probability distributions*. Michael P. McLaughlin, 2001.

[43] Aditya Krishna Menon, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Can gradient clipping mitigate label noise? In *International Conference on Learning Representations*, 2020.

[44] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing lstm language models. *arXiv preprint arXiv:1708.02182*, 2017.

[45] Tomáš Mikolov. Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*, 80, 2012.

[46] Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

[47] Aleksandr Viktorovich Nazin, AS Nemirovsky, Aleksandr Borisovich Tsybakov, and AB Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80(9):1607–1627, 2019.

[48] Deanna Needell, Nathan Srebro, and Rachel Ward. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Mathematical Programming*, 155(1-2):549–573, 2016.

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

[50] Arkadi S Nemirovski and David Berkovich Yudin. Cesari convergence of the gradient method of approximating saddle points of convex-concave functions. In *Doklady Akademii Nauk*, volume 239, pages 1056–1059. Russian Academy of Sciences, 1978.

[51] Arkadi Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[52] Yu Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1-2):381–404, 2015.

[53] Yu Nesterov and J-Ph Vial. Confidence level solutions for stochastic programming. *Automatica*, 44(6):1559–1568, 2008.

[54] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[55] Lam Nguyen, Phuong Ha Nguyen, Marten Dijk, Peter Richtarik, Katya Scheinberg, and Martin Takac. Sgd and hogwild! convergence without the bounded gradients assumption. In *International Conference on Machine Learning*, pages 3750–3758, 2018.

[56] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318, 2013.

[57] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

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

[59] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[60] R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

[61] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[62] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, 2009.

[63] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

[64] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

[65] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. *arXiv preprint arXiv:1901.06053*, 2019.

[66] Vladimir Spokoiny et al. Parametric estimation. finite sample theory. *The Annals of Statistics*, 40(6):2877–2909, 2012.

[67] Ilnura Usmanova. Robust solutions to stochastic optimization problems. *Master Thesis (MSIAM); Institut Polytechnique de Grenoble ENSIMAG, Laboratoire Jean Kuntzmann*, 2017.

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[69] Waloddi Weibull. A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, 18:293–297, 1951.

[70] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.

[71] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J Reddi, Sanjiv Kumar, and Suvrit Sra. Why adam beats sgd for attention models. *arXiv preprint arXiv:1912.03194*, 2019.

Appendix
Stochastic Optimization with Heavy-Tailed Noise via Accelerated Gradient Clipping

Contents

A Notations and Definitions

We use $\langle x, y\rangle$ to define standard inner product between two vectors $x, y \in \mathbb{R}^n$, i.e. $\langle x, y\rangle \stackrel{\text{def}}{=} \sum_{i=1}^n x_i y_i$, where x_i is i-th coordinate of vector x, $i = 1, \ldots, n$. Standard Euclidean norm of vector $x \in \mathbb{R}^n$ is defined as $\|x\|_2 \stackrel{\text{def}}{=} \sqrt{\langle x, x\rangle}$.

We use $\mathbb{P}\{\cdot\}$ to define probability measure which is always known from the context, $\mathbb{E}[\cdot]$ denotes mathematical expectation, $\mathbb{E}_\xi[\cdot]$ is used to define conditional mathematical expectation with respect to the randomness coming from ξ only and $\mathbb{E}\left[\xi \mid \eta\right]$ denotes mathematical expectation of ξ conditional on η. In our proofs, we also use $\mathbb{E}_k[\cdot]$ to denote conditional mathematical expectation with respect to all randomness coming from k-th iteration. For $\mathbb{P}$-measurable set X we use $\mathbb{1}_X$ to denote indicator of event X, i.e.

$$\mathbb{1}_X = \begin{cases} 1, & \text{if event } X \text{ holds,} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Next, we introduce some standard definitions.

Definition A.1 (L-smoothness)**.** Function f is called L-smooth on $\mathbb{R}^n$ with $L > 0$ when it is differentiable and its gradient is L-Lipschitz continuous on $\mathbb{R}^n$, i.e.

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y \in \mathbb{R}^n. \tag{6}$$

It is well-known that L-smoothness implies (see [54])

$$f(y) \quad \leq \quad f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2 \quad \forall x, y \in \mathbb{R}^n, \tag{7}$$

and if f is additionally convex, then

$$\|\nabla f(x) - \nabla f(y)\|_2^2 \quad \leq \quad 2L\left(f(x) - f(y) - \langle \nabla f(y), x - y\rangle\right) \quad \forall x, y \in \mathbb{R}^n. \tag{8}$$

Since in this paper we focus only on smooth optimization problems we introduce strong convexity in the following way.

Definition A.2 (μ-strong convexity)**.** Differentiable function f is called μ-strongly convex on $\mathbb{R}^n$ with $\mu \geq 0$ if for all $x, y \in \mathbb{R}^n$

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|_2^2. \tag{9}$$

In particular, μ-strong convexity implies that for all $x \in \mathbb{R}^n$

$$f(x) - f(x^*) \geq \frac{\mu}{2}\|x - x^*\|_2^2. \tag{10}$$

Throughout the paper, we use x^* to denote any solution of problem (1) assuming its existence. By the complexity of stochastic first-order method we always mean the total number of stochastic first-order oracle calls that the method needs in order to produce such a point $\hat{x}$ that $f(\hat{x}) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ for some $\varepsilon > 0$ and $\beta \in (0, 1)$. Finally, in the complexity bounds we often use R_0 to denote $\|x^0 - x^*\|_2$ where x^0 is the starting point of the method.

B Related Work: Additional Details

B.1 Related Work on Non-Smooth Stochastic Optimization

Here we present an overview of existing results in the convex non-smooth case, i.e. when f is still convex but not necessarily L-smooth and the stochastic gradients have a bounded second moment: $\mathbb{E}_\xi[\|\nabla f(x,\xi)\|_2^2] \leq M^2$ for all $x \in \mathbb{R}^n$. Under additional assumption that the stochastic gradients have light-tailed distribution it was shown that SGD [49] has $O\left(M^2R_0^2\ln(\beta^{-1})/\varepsilon^2\right)$ complexity and if additionally f is μ-strongly convex it was shown in [31, 32] that the restarted version of SGD has $O\left(M^2\ln\left(\beta^{-1}\ln(M^2\mu^{-1}\varepsilon^{-1})\right)/\mu\varepsilon\right)$ complexity (see also [27, 33, 58]). Moreover, removing logarithmical factors from these bounds we get the complexity bounds of these methods for the convergence in expectation, i.e. needed number of oracle calls to find such $\hat{x}$ that $\mathbb{E}[f(\hat{x})] - f(x^*) \leq \varepsilon$. That is, under light tails assumption high-probability complexity bounds and complexity bounds in expectation for SGD and `restarted-SGD` differ only in logarithmical factors of $1/\beta$.

Unfortunately, for these methods the situation changes dramatically when the stochastic gradients are heavy-tailed. To the best of our knowledge, the best know bounds in the literature with the same dependency on ε are $O\left(M^2R_0^2/\beta^2\varepsilon^2\right)$ and $O\left(M^2/\mu\beta\varepsilon\right)$. One can obtain these bounds using complexity results for the convergence in expectation and Markov's inequality. However, it leads to significantly worse dependence on β: instead of $O(\ln(\beta^{-1}))$ we get $O(\beta^{-2})$ and $O(\beta^{-1})$ dependence on the confidence level β. Furthermore, based on the well-known results on the distribution of sum of i.i.d. random variables (see Section D.2) in [15] authors consider the case when the tails of the distribution of stochastic gradient satisfy $\mathbb{P}\{\|\nabla f(x,\xi) - \nabla f(x)\|_2 > s\} = O(s^{-\alpha})$ for $\alpha > 2$ and give the following complexity bounds without formal proofs that SGD for convex problems and `restarted-SGD` for μ-strongly convex problems have following complexities:

$$O\left(M^2R^2\max\left\{\frac{\ln\left(\beta^{-1}\right)}{\varepsilon^2}, \left(\frac{1}{\beta\varepsilon^\alpha}\right)^{\frac{2}{3\alpha-2}}\right\}\right),$$

$$O\left(\max\left\{\frac{M^2\ln\left(\beta^{-1}\ln\frac{M^2}{\mu\varepsilon}\right)}{\mu\varepsilon}, \left(\frac{M^2}{\mu\varepsilon}\right)^{\frac{\alpha}{3\alpha-2}}\left(\beta^{-1}\ln\frac{M^2}{\mu\varepsilon}\right)^{\frac{2}{3\alpha-2}}\right\}\right).$$

The first terms in maximums above correspond to the Central Limit Theorem regime, while the second terms correspond to the heavy-tailed regime, see Section D.2. These bounds show that heavy tailed distributions of the stochastic gradients significantly spoil complexity bounds of SGD and `restarted-SGD` when the confidence level β is small enough.

B.2 Related Work on Gradient Clipping

As we mentioned Section 1.2 `clipped-SGD` [21, 45, 56, 67] is known to be robust to the noise in stochastic gradients and performs better than SGD in the vicinity of extremely steep cliffs. Zhang et al. [71] analyse the convergence of `clipped-SGD` *in expectation* for strongly convex and non-convex objectives under assumption that $\mathbb{E}[\|\nabla f(x,\xi)\|_2^\alpha]$ is bounded for some $\alpha \in (1,2]$. For $\alpha < 2$ this assumption covers some heavy-tailed distributions of stochastic gradients appearing in practice. Moreover, in [71] authors conduct several numerical tests showing that in some real-world problems where the noise in stochastic gradients is heavy-tailed `clipped-SGD` converges faster than SGD. In [70] Zhang et al. found that `clipped-GD` is able to converge in non-convex case to the stationary point under the relaxed smoothness assumption with $O(\varepsilon^{-2})$ rate while Gradient Descent (GD) can fail to converge with the same rate in this setting. A very similar approach based on the normalization of GD is studied in [28, 39].

C Basic Facts

In this section we enumerate for convenience basic facts that we use many times in our proofs.

Fenchel-Young inequality. For all $a, b \in \mathbb{R}^n$ and $\lambda > 0$

$$|\langle a, b\rangle| \leq \frac{\|a\|_2^2}{2\lambda} + \frac{\lambda\|b\|_2^2}{2}. \tag{11}$$

Squared norm of the sum. For all $a, b \in \mathbb{R}^n$

$$\|a+b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2. \tag{12}$$

Inner product representation. For all $a, b \in \mathbb{R}^n$

$$\langle a, b\rangle = \frac{1}{2}\left(\|a+b\|_2^2 - \|a\|_2^2 - \|b\|_2^2\right) \tag{13}$$

Variance decomposition. If ξ is a random vector in $\mathbb{R}^n$ with bounded second moment, then

$$\mathbb{E}\left[\|\xi + a\|_2^2\right] = \mathbb{E}\left[\|\xi - \mathbb{E}[\xi]\|_2^2\right] + \|\mathbb{E}[\xi] + a\|_2^2 \tag{14}$$

for any deterministic vector $a \in \mathbb{R}^n$. In particular, this implies

$$\mathbb{E}\left[\|\xi - \mathbb{E}[\xi]\|_2^2\right] \leq \mathbb{E}\left[\|\xi + a\|_2^2\right] \tag{15}$$

for any deterministic vector $a \in \mathbb{R}^n$.

D Auxiliary Results

D.1 Bernstein Inequality

Lemma D.1 (Bernstein inequality for martingale differences [1, 12, 14])**.** Let the sequence of random variables $\{X_i\}_{i\geq 1}$ form a martingale difference sequence, i.e. $\mathbb{E}\left[X_i \mid X_{i-1}, \ldots, X_1\right] = 0$ for all $i \geq 1$. Assume that conditional variances $\sigma_i^2 \stackrel{\text{def}}{=} \mathbb{E}\left[X_i^2 \mid X_{i-1}, \ldots, X_1\right]$ exist and are bounded and assume also that there exists deterministic constant $c > 0$ such that $\|X_i\|_2 \leq c$ almost surely for all $i \geq 1$. Then for all $b > 0$, $F > 0$ and $n \geq 1$

$$\mathbb{P}\left\{\Big|\sum_{i=1}^n X_i\Big| > b \text{ and } \sum_{i=1}^n \sigma_i^2 \leq F\right\} \leq 2\exp\left(-\frac{b^2}{2F + {}^{2cb}/_3}\right). \tag{16}$$

D.2 About the Sum of i.i.d. Random Variables with Heavy Tails

In this section we present some classical results about the distribution of sum of i.i.d. random variables $\sum_{k=1}^N \xi_k$ with heavy tails [2]. As one can see from our proofs of main results for `clipped-SSTM` and `clipped-SGD` such sums play a central role in the analysis of convergence with high probability.

Assume that $\{\xi_k\}$ is i.i.d. with $\mathbb{E}[\xi_k] = 0$ and $\text{Var}[\xi_k] \stackrel{\text{def}}{=} \mathbb{E}[(\xi_k - \mathbb{E}[\xi_k])^2] = \sigma^2$. Assume also that $V(s) = \mathbb{P}\{\xi_k \geq s\} = \Theta\left(s^{-\alpha}\right)$, where $\alpha > 2$. In this case

$$\mathbb{P}\left\{\sum_{k=1}^N \xi_k \geq s\right\} \simeq 1 - \Phi\left(\frac{s}{\sqrt{\sigma^2 N}}\right) + N \cdot V(s),$$

where $N \gg 1$ and $\Phi(x) = \frac{1}{2\pi}\int_{-\infty}^x \exp\left(-y^2/2\right) dy$. Since

$$0.2\exp\left(-\frac{2x^2}{\pi}\right) \leq 1 - \Phi(x) \leq \exp\left(-\frac{x^2}{2}\right),$$

we have[5]

$$\mathbb{P}\left\{\sum_{k=1}^N \xi_k \geq s\right\} \simeq 1 - \Phi\left(\frac{s}{\sqrt{\sigma^2 N}}\right), \quad s \leq \sqrt{(\alpha-2)\sigma^2 N \ln N} \quad \text{(CLT regime)} \tag{17}$$

[5] CLT = Central Limit Theorem.

and

$$\mathbb{P}\left\{\sum_{k=1}^{N}\xi_k \geq s\right\} \simeq N\cdot V(s), \quad s > \sqrt{(\alpha-2)\sigma^2 N\ln N} \quad \text{(heavy-tailed regime).} \tag{18}$$

This simple observation can play a significant role in deriving complexity results for non-smooth convex optimization under the assumption that stochastic gradients are heavy-tailed, see [15] for the details.

E Technical Results

Lemma E.1. Consider two sequences of non-negative numbers $\{\alpha_k\}_{k\geq 0}$ and $\{A_k\}_{k\geq 0}$ such that

$$\alpha_0 = A_0 = 0, \quad A_{k+1} = A_k + \alpha_{k+1}, \quad \alpha_{k+1} = \frac{k+2}{2aL} \quad \forall k \geq 0, \tag{19}$$

where $a, L > 0$. Then for all $k \geq 0$

$$A_{k+1} = \frac{(k+1)(k+4)}{4aL}, \tag{20}$$

$$A_{k+1} \geq aL\alpha_{k+1}^2. \tag{21}$$

Proof. By definition of A_{k+1} we have that

$$A_{k+1} = \sum_{l=1}^{k+1}\alpha_l = \frac{1}{2aL}\sum_{l=1}^{k+1}(l+1) = \frac{(k+1)(k+4)}{4aL}.$$

Using $(k+1)(k+4) \geq (k+2)^2$ together with the inequality above we derive (21). □

F Accelerated SGD with Clipping: Exact Formulations and Missing Proofs

In this section we provide exact formulations of all the results that we have for `clipped-SSTM` and `R-clipped-SSTM` together with the full proofs.

F.1 Convex Case

Recall that in order to compute $\text{clip}(\nabla f(x, \boldsymbol{\xi}), \lambda)$ one needs to get m i.i.d. samples $\nabla f(x, \xi_1), \ldots, \nabla f(x, \xi_m)$, compute its average

$$\nabla f(x, \boldsymbol{\xi}) = \frac{1}{m} \sum_{i=1}^{m} \nabla f(x, \xi_i), \tag{22}$$

and then project the result $\nabla f(x, \boldsymbol{\xi})$ on the Euclidean ball with radius λ and center at the origin. We also notice that

$$\mathbb{E}_{\xi}[\nabla f(x, \boldsymbol{\xi})] = \nabla f(x), \tag{23}$$

$$\mathbb{E}_{\xi}\left[\|\nabla f(x, \boldsymbol{\xi}) - \nabla f(x)\|_2^2\right] \leq \frac{\sigma^2}{m}. \tag{24}$$

F.1.1 Convergence Guarantees for `clipped-SSTM`

Next theorem summarizes the main convergence result for `clipped-SSTM`.

Theorem F.1. Assume that function f is convex and L-smooth. Then for all $\beta \in (0, 1)$ and $N \geq 1$ such that

$$\ln \frac{4N}{\beta} \geq 2 \tag{25}$$

we have that after N iterations of `clipped-SSTM` with

$$m_k = \max\left\{1, \frac{6000\sigma^2 \alpha_{k+1}^2 N \ln \frac{4N}{\beta}}{C^2 R_0^2}, \frac{10368\sigma^2 \alpha_{k+1}^2 N}{C^2 R_0^2}\right\}, \tag{26}$$

$$B = \frac{CR_0}{8 \ln \frac{4N}{\beta}}, \quad a \geq \max\left\{1, \frac{16 \ln \frac{4N}{\beta}}{C}, 36\left(2 \ln \frac{4N}{\beta} + \sqrt{4 \ln^2 \frac{4N}{\beta} + 2 \ln \frac{4N}{\beta}}\right)^2\right\}, \tag{27}$$

that with probability at least $1 - \beta$

$$f(y^N) - f(x^*) \leq \frac{2aLC^2 R_0^2}{N(N+3)}, \tag{28}$$

where $R_0 = \|x^0 - x^*\|_2$ and

$$C = \sqrt{5}. \tag{29}$$

In other words, if we choose a to be equal to the maximum from (27), then the method achieves $f(y^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ after $O\left(\sqrt{\frac{LR_0^2}{\varepsilon}} \ln \frac{LR_0^2}{\varepsilon\beta}\right)$ iterations and requires

$$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\} \ln \frac{LR_0^2}{\varepsilon\beta}\right) \text{ oracle calls.} \tag{30}$$

One can easily notice that multiplicative constant factors in formulas for m_k and a are too big and seem to be impractical, but in practice one can tune these constants to get good enough performance. That is, big constants in (26) and (27) are needed only in our analysis in order to get bound (30).

Finally, when σ^2 is big then Theorem F.1 says that at iteration k `clipped-SGD` requires large batchsizes $m_k \sim k^2 N$ (see (26)) which is proportional to $\varepsilon^{-3/2}$ for last iterates. It can make the cost of one iteration extremely high, therefore, we consider different stepsize policies that remove this drawback.

Corollary F.2. Let the assumptions of Theorem F.1 hold.

1. **(Medium batchsize).** If N and β are such that $N \ln \frac{4N}{\beta}$ is bigger than the maximum from (27), then for $a = N \ln \frac{4N}{\beta}$ we have

$$m_k = \max\left\{1, \frac{6000\sigma^2(k+2)^2}{4L^2NC^2R_0^2 \ln \frac{4N}{\beta}}, \frac{10368\sigma^2(k+2)^2}{4L^2C^2R_0^2N \ln^2 \frac{4N}{\beta}}\right\} \tag{31}$$

and the method achieves $f(y^N) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$ after $O\left(\frac{LR_0^2}{\varepsilon} \ln \frac{LR_0^2}{\varepsilon\beta}\right)$ iterations and requires

$$O\left(\max\left\{\frac{LR_0^2}{\varepsilon}, \frac{\sigma^2R_0^2}{\varepsilon^2}\right\} \ln \frac{LR_0^2}{\varepsilon\beta}\right) \text{ oracle calls.} \tag{32}$$

2. **(Constant batchsize).** If N and β are such that $a_0N^{3/2}\sqrt{\ln \frac{4N}{\beta}}$ is bigger than the maximum from (27) for some positive constant a_0, then for $a = a_0N^{3/2}\sqrt{\ln \frac{4N}{\beta}}$ we have

$$m_k = \max\left\{1, \frac{6000\sigma^2(k+2)^2}{4a_0^2L^2N^2C^2R_0^2}, \frac{10368\sigma^2(k+2)^2}{4a_0^2L^2C^2R_0^2N^2 \ln \frac{4N}{\beta}}\right\} \tag{33}$$

and the method achieves $f(y^N) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$ after $O\left(\frac{a_0^2L^2R_0^4}{\varepsilon^2} \ln \frac{a_0LR_0^2}{\varepsilon\beta}\right)$ iterations and requires

$$O\left(\max\left\{\frac{a_0^2L^2R_0^4}{\varepsilon^2}, \frac{\sigma^2R_0^2}{\varepsilon^2}\right\} \ln \frac{a_0LR_0^2}{\varepsilon\beta}\right) \text{ oracle calls.} \tag{34}$$

Finally, if $a_0 = \frac{\sigma}{LR_0}$, then $m_k = O(1)$ for $k = 0, 1, \ldots, N$ and `clipped-SSTM` finds ε-solution with probability at least $1-\beta$ after $O\left(\frac{\sigma^2R_0^2}{\varepsilon^2} \ln \frac{\sigma R_0}{\varepsilon\beta}\right)$ iterations and requires $O(1)$ oracle calls per iteration.

In the first case batchsizes increase from $O(1)$ for $k = 1$ to $O(\varepsilon^{-1})$ for $k = N$ and the overall complexity recovers the complexity of Robust Stochastic Mirror Descent (`RSMD`) from [47]. However, analysis from [47] works only for the optimization problems on *compact* convex sets, whereas our analysis handles an unconstrained optimization on $\mathbb{R}^n$. Despite the similarities of our approach and [47], it seems that the technique from [47] cannot be generalized to obtain the complexity like in (30) due to the fast bias accumulation that appears because of the special truncation of stochastic gradients that is used in `RSMD`.

In the second case the corollary establishes $\varepsilon^{-2}\ln(\varepsilon^{-1}\beta^{-1})$ rate for `clipped-SSTM` with constant batchsizes, i.e. $m_k = O(1)$ for all k. The ability of `clipped-SSTM` to converge with constant batchsizes makes it more practical and applicable for wider class of problems where it can be very expensive to compute large batchsizes, e.g. training deep neural networks. Moreover, when σ is not too small, i.e. $\sigma^2 \geq L\varepsilon$, this rate is optimal (up to logarithmical factors) and also recovers the rate of `RSMD`.

Finally, setting

$$\begin{aligned} a' &= \max\left\{1, \frac{16 \ln \frac{4N}{\beta}}{C}, 36\left(2\ln\frac{4N}{\beta} + \sqrt{4\ln^2\frac{4N}{\beta} + 2\ln\frac{4N}{\beta}}\right)^2\right\}, \\ a &= \max\left\{a', \frac{\sigma N^{3/2}}{LR_0}\sqrt{\ln\frac{4N}{\beta}}\right\} \end{aligned} \tag{35}$$

and m_k as in (26), we get $m_k = O(1)$ for $k = 0, 1, \ldots, N$ and derive the following result.

Corollary F.3. Let the assumptions of Theorem F.1 hold, a is chosen as in (35) and m_k is computed via (26). Then `clipped-SSTM` achieves $f(y^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ after

$$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\} \ln \frac{LR_0^2 + \sigma R_0}{\varepsilon\beta}\right) \text{ iterations/oracle calls.}$$

F.1.2 Sketch of the Proof of Theorem F.1

We start with the following lemma that is pretty standard in the analysis of Stochastic Similar Triangles Method, e.g. see the proof of Theorem 1 from [10].

Lemma F.4. Let f be a convex L-smooth function and let stepsize parameter a satisfy $a \geq 1$. Then after $N \geq 0$ iterations of `clipped-SSTM` for all $z \in \mathbb{R}^n$ we have

$$\begin{aligned} A_N\left(f(y^N) - f(z)\right) &\leq \frac{1}{2}\|z^0 - z\|_2^2 - \frac{1}{2}\|z^N - z\|_2^2 + \sum_{k=0}^{N-1} \alpha_{k+1}\left\langle \theta_{k+1}, z - z^k\right\rangle \\ &\quad + \sum_{k=0}^{N-1} \alpha_{k+1}^2 \left\|\theta_{k+1}\right\|_2^2 + \sum_{k=0}^{N-1} \alpha_{k+1}^2 \left\langle \theta_{k+1}, \nabla f(x^{k+1})\right\rangle, \quad (36) \\ \theta_{k+1} &\overset{\text{def}}{=} \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) - \nabla f(x^{k+1}). \quad (37) \end{aligned}$$

That is, if $z = x^*$, then the result above gives a preliminary upper bound for $A_N(f(y^N) - f(x^*))$. The first and the second terms in the r.h.s. of (36) come from the analysis of Similar Triangles Method [16] and three last terms have a stochastic nature. In particular, they explicitly depend on differences $\theta_{k+1} = \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) - \nabla f(x^{k+1})$ between clipped mini-batched stochastic gradients and full gradients at x^{k+1}, so, if $\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) = \nabla f(x^{k+1})$ with probability 1, then we easily get needed convergence rate. However, we are interested in the more general case and, as a consequence, to continue the proof, we need to find a good enough upper bound for the last three terms from (36). In other words, we need to show that choosing parameters a, m_k and λ_{k+1} properly we can upper bound these terms by something that coincides with $\|z^0 - x^*\|_2^2$ up to numerical multiplicative constant. The proof of convergence result for `RSMD` from [47] where authors provide upper bound for similar sums hints that Bernstein's inequality (see Lemma D.1) applied to estimate these terms can help us to reach our goal. In order to apply Bernstein's inequality one should derive tight bounds for such characteristics of $\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)$ as upper bounds for the magnitude, bias, variance and distortion and the next lemma provides us with this.

Lemma F.5. For all $k \geq 0$ the following inequality holds:

$$\left\|\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) - \mathbb{E}_{\boldsymbol{\xi}^k}\left[\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)\right]\right\|_2 \leq 2\lambda_{k+1}. \quad (38)$$

Moreover, if $\|\nabla f(x^{k+1})\|_2 \leq \frac{\lambda_{k+1}}{2}$ for some $k \geq 0$, then for this k we have:

$$\begin{aligned} \left\|\mathbb{E}_{\boldsymbol{\xi}^k}\left[\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)\right] - \nabla f(x^{k+1})\right\|_2 &\leq \frac{4\sigma^2}{m_k \lambda_{k+1}}, \quad (39) \\ \mathbb{E}_{\boldsymbol{\xi}^k}\left[\left\|\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) - \nabla f(x^{k+1})\right\|_2^2\right] &\leq \frac{18\sigma^2}{m_k}, \quad (40) \\ \mathbb{E}_{\boldsymbol{\xi}^k}\left[\left\|\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) - \mathbb{E}_{\boldsymbol{\xi}^k}\left[\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)\right]\right\|_2^2\right] &\leq \frac{18\sigma^2}{m_k}. \quad (41) \end{aligned}$$

Clearly, clipping introduces a bias in $\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)$ which influences the convergence of the method. Hence, the clipping level λ_{k+1} should be chosen in a very accurate way. Below we informally describe what does it mean and present the sketch of the remaining part of the proof.

Imagine the ideal situation: $\nabla f(x^{k+1}, \boldsymbol{\xi}^k) = \nabla f(x^{k+1})$ with probability 1 for all k, i.e. we have an access to the full gradients at points x^{k+1}. Then it is natural to choose λ_{k+1} in such a way that $\text{clip}(\nabla f(x^{k+1}), \lambda^{k+1}) = \nabla f(x^{k+1})$ in order to recover Similar Triangles Method (`STM`) that

converges with optimal rate in the deterministic case. In other words, one can pick λ_{k+1} such that $\|\nabla f(x^{k+1})\|_2 \leq \lambda_{k+1}$ and get an optimal method. Since we know that in this case the method should converge with $O(1/k^2)$ rate in terms of $f(x^k) - f(x^*)$ one can expect that the gradient's norm decays with $O(1/k)$ rate, so, one can choose λ_{k+1} to be proportional to $1/k$. It is exactly what we do when we define λ_{k+1} as B/α_{k+1}.

The ideal case described above gives a good insight on how to choose λ_{k+1} in the general case and can be described as follows: if we want to prevent our gradient estimator $\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)$ from large deviations from $\nabla f(x^{k+1})$ with high probability, then it is needed to choose λ_{k+1} such that $\|\nabla f(x^k)\|_2 \leq c\lambda_{k+1}$ with high probability where $c < 1$ is some positive number. This choice guarantees that with high probability clipped mini-batched gradient $\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)$ cannot deviates from $\nabla f(x^{k+1})$ significantly and, as a consequence, the convergence rate of `clipped-SSTM` in terms of the number of iterations needed to achieve the desired accuracy of the solution with high probability becomes similar to the convergence rate of `STM` up to some logarithmical factors depending on the confidence level.

In particular, we choose λ_{k+1} such that $\|\nabla f(x^{k+1})\|_2 \leq \lambda_{k+1}/2$ with high probability. Moreover, we derive this relation by induction via refined estimation of the three last terms from the r.h.s. of (36) that is based on the new variant of advanced recurrences technique from [22, 23]. The main trick there is in showing by induction that sequence $\|z^k - x^*\|_2$ is bounded by some constant multiplied by $\|x^0 - x^*\|_2$ and in deriving $\|\nabla f(x^{k+1})\|_2 \leq \lambda_{k+1}/2$ simultaneously for all $k = 0, 1, \ldots, N$. With such bounds and Lemma F.5 in hand, it is possible to apply Bernstein's inequality to three sums from the r.h.s. of (36) since all summands are bounded *with high probability*. After applying Bernstein's inequality we adjust parameters α_{k+1} and m_k in such a way that after rearranging the terms in the obtained upper bounds we get that r.h.s. in (36) (with $z = x^*$) is smaller than $\|x^0 - x^*\|_2^2$ up to some multiplicative numerical constant. This finishes the proof.

To conclude, the key tools in our analysis are Bernstein's inequality (see Lemma D.1) and advanced recurrences technique [22, 23] that helps us to show boundedness of $\|z^N - x^*\|_2$ and $\|\nabla f(x^{k+1})\|_2 \leq \lambda_{k+1}/2$ with high probability. We provide detailed proofs of presented result in the Appendix (see Section F.3).

F.2 Strongly Convex Case

In this section we assume additionally that $f(x)$ is μ-strongly convex. For this case we modify Algorithm 1 and propose a new method called Restarted Clipped Similar Triangles Method (`R-clipped-SSTM`), see Algorithm 3. At each iteration `R-clipped-SSTM` runs `clipped-SSTM` for

Algorithm 3 Restarted Clipped Stochastic Similar Triangles Method (`R-clipped-SSTM`)

Input: starting point x^0, number of iterations N_0 of `clipped-SSTM`, number of `clipped-SSTM` runs, batchsizes $\{m_k^0\}_{k=0}^{N_0-1}$, $\{m_k^1\}_{k=0}^{N_0-1}$, $\ldots$, $\{m_k^\tau\}_{k=0}^{N_0-1}$, stepsize parameter a, clipping parameters $\{B_t\}_{t=0}$

1: Set $\hat{x}^0 = x^0$
2: **for** $t = 0, 1, \ldots, \tau - 1$ **do**
3: Run `clipped-SSTM` (Algorithm 1) for N_0 iterations with batchsizes $\{m_k^t\}_{k=1}^{N_0}$, stepsize parameter a, clipping parameter B_t and starting point $\hat{x}^t$. Define the output of `clipped-SSTM` by $\hat{x}^{t+1}$.
4: **end for**

Output: $\hat{x}^\tau$

N_0 iterations from the current point $\hat{x}^k$ and use its output as next iterate $\hat{x}^{k+1}$. In literature this approach is known as the restarts technique [11, 31, 32, 51]. Choosing N_0 and parameters m_k, a and B in a proper way one can get an accelerated method for strongly convex objectives. Theorem below states the main convergence result for `R-clipped-SSTM`.

Theorem F.6. Assume that f is μ-strongly convex and L-smooth. If we choose $\beta \in (0,1)$, τ and $N_0 \geq 1$ such that

$$\ln \frac{4N_0\tau}{\beta} \geq 2, \quad N_0 \geq C\sqrt{\frac{8aL}{\mu}}, \tag{42}$$

and

$$m_k^t = \max\left\{1, \frac{6000 \cdot 2^t \sigma^2 \alpha_{k+1}^2 N_0 \ln \frac{4N_0\tau}{\beta}}{C^2R^2}, \frac{10368 \cdot 2^t \sigma^2 \alpha_{k+1}^2 N_0}{C^2R^2}\right\}, \tag{43}$$

$$B_t = \frac{CR}{8 \cdot 2^t \ln \frac{4N_0\tau}{\beta}}, \tag{44}$$

$$a \geq \max\left\{1, \frac{16 \ln \frac{4N_0\tau}{\beta}}{C}, 36\left(2\ln\frac{4N_0\tau}{\beta} + \sqrt{4\ln^2\frac{4N_0\tau}{\beta} + 2\ln\frac{4N_0\tau}{\beta}}\right)^2\right\}, \tag{45}$$

where $R = \sqrt{\frac{2(f(x^0)-f(x^*))}{\mu}}$ and $C = \sqrt{5}$, then we have that after τ runs of `clipped-SSTM` in `R-clipped-SSTM` the inequality

$$f(\hat{x}^\tau) - f(x^*) \leq 2^{-\tau}\left(f(x^0) - f(x^*)\right) \tag{46}$$

holds with probability at least $1-\beta$. That is, if we choose a to be equal to the maximum from (45) and $N_0 \leq C_1\sqrt{\frac{8aL}{\mu}}$ with some numerical constant $C_1 \geq C$, then the method achieves $f(\hat{x}^\tau) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$ after

$$O\left(\sqrt{\frac{L}{\mu}}\ln\left(\frac{\mu R^2}{\varepsilon}\right)\ln\left(\frac{L}{\mu\beta}\ln\frac{\mu R^2}{\varepsilon}\right)\right) \text{ iterations (in total)} \tag{47}$$

of `clipped-SSTM` and requires

$$O\left(\max\left\{\sqrt{\frac{L}{\mu}}\ln\frac{\mu R^2}{\varepsilon}, \frac{\sigma^2}{\mu\varepsilon}\right\}\ln\left(\frac{L}{\mu\beta}\ln\frac{\mu R^2}{\varepsilon}\right)\right) \text{ oracle calls.} \tag{48}$$

In other words, `R-clipped-SSTM` has the same convergence rate as optimal stochastic methods for strongly convex problems like Multi-Staged `AC-SA` (`MS-AC-SA`) [19] or Stochastic Similar Triangles Method for strongly convex problems (`SSTM_sc`) [16, 22]. Moreover, in Theorem F.6 we *do not assume* that stochastic gradients are sampled from sub-Gaussian distribution while corresponding results for `MS-AC-SA` and `SSTM_sc` are substantially based on the light tails assumption. Our bound outperforms the state-of-the-art result from [7] in terms of the dependence on $\ln\frac{L}{\mu}$. It is worth to mention here that using special restarts technique Nazin et al. [47] generalize their method (`RSMD`) for the strongly convex case, but since `RSMD` is not accelerated their approach gives only non-accelerated convergence rate.

We also emphasize that big numerical factors in formulas for m_k^t and a are needed only in our analysis and in practice they can be tuned. However, when σ^2 is big bathsizes m_k^t become of the order $k^2\varepsilon^{-1}$. It can make the cost of one iteration extremely high, therefore, as for `clipped-SSTM` we consider a different stepsize policy removing this drawback.

Corollary F.7. Let the assumptions of Theorem F.6 hold. Assume that conditions (42), (43), (44) and (45) are satisfied for

$$a = \Theta\left(\frac{\sigma^4 \ln^2\frac{N_0\tau}{\beta}}{L\mu\varepsilon^2}\right), \quad N_0 = \Theta\left(\sqrt{\frac{aL}{\mu}}\right). \tag{49}$$

Then after $\tau = \lceil \ln(\mu R^2/2\varepsilon) \rceil$ runs of `clipped-SSTM` in `R-clipped-SSTM` the method achieves $f(\hat{x}^\tau) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$. Moreover, the total number of iterations of

`clipped-SSTM` equals

$$O\left(\frac{\sigma^2}{\mu\varepsilon}\ln\left(\frac{\mu R^2}{\varepsilon}\right)\ln\left(\frac{\sigma^2}{\mu\varepsilon\beta}\ln\frac{\mu R^2}{\varepsilon}\right)\right) \tag{50}$$

with $m_k^t = O(1)$ for all $k = 0, 1, \ldots, N_0 - 1$, $t = 0, 1, \ldots, \tau - 1$.

When σ^2 is big the obtained bound is comparable with bounds for `restarted-RSMD` and `proxBoost`, see Table 2.

F.3 Proofs

F.3.1 Proof of Lemma F.4

Using $z^{k+1} = z^k - \alpha_{k+1}\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)$ we get that for all $z \in \mathbb{R}^n$

$$\begin{aligned}
\alpha_{k+1}\left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z^k - z\right\rangle &= \alpha_{k+1}\left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z^k - z^{k+1}\right\rangle \\
&\quad + \alpha_{k+1}\left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z^{k+1} - z\right\rangle \\
&= \alpha_{k+1}\left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z^k - z^{k+1}\right\rangle + \left\langle z^{k+1} - z^k, z - z^{k+1}\right\rangle \\
&\overset{(13)}{\leq} \alpha_{k+1}\left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z^k - z^{k+1}\right\rangle - \frac{1}{2}\|z^k - z^{k+1}\|_2^2 \\
&\quad + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2.
\end{aligned} \tag{51}$$

Next, we notice that

$$y^{k+1} = \frac{A_k y^k + \alpha_{k+1} z^{k+1}}{A_{k+1}} = \frac{A_k y^k + \alpha_{k+1} z^k}{A_{k+1}} + \frac{\alpha_{k+1}}{A_{k+1}}\left(z^{k+1} - z^k\right) = x^{k+1} + \frac{\alpha_{k+1}}{A_{k+1}}\left(z^{k+1} - z^k\right) \tag{52}$$

which implies:

$$\begin{aligned}
\alpha_{k+1}\left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z^k - z\right\rangle &\overset{(37),(51)}{\leq} \alpha_{k+1}\left\langle \nabla f(x^{k+1}), z^k - z^{k+1}\right\rangle - \frac{1}{2}\|z^k - z^{k+1}\|_2^2 \\
&\quad + \alpha_{k+1}\left\langle \theta_{k+1}, z^k - z^{k+1}\right\rangle + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 \\
&\overset{(52)}{=} A_{k+1}\left\langle \nabla f(x^{k+1}), x^{k+1} - y^{k+1}\right\rangle - \frac{1}{2}\|z^k - z^{k+1}\|_2^2 \\
&\quad + \alpha_{k+1}\left\langle \theta_{k+1}, z^k - z^{k+1}\right\rangle + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 \\
&\overset{(7)}{\leq} A_{k+1}\left(f(x^{k+1}) - f(y^{k+1})\right) + \frac{A_{k+1}L}{2}\|x^{k+1} - y^{k+1}\|_2^2 \\
&\quad - \frac{1}{2}\|z^k - z^{k+1}\|_2^2 + \alpha_{k+1}\left\langle \theta_{k+1}, z^k - z^{k+1}\right\rangle \\
&\quad + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 \\
&\overset{(52)}{=} A_{k+1}\left(f(x^{k+1}) - f(y^{k+1})\right) + \frac{1}{2}\left(\frac{\alpha_{k+1}^2 L}{A_{k+1}} - 1\right)\|z^k - z^{k+1}\|_2^2 \\
&\quad + \alpha_{k+1}\left\langle \theta_{k+1}, z^k - z^{k+1}\right\rangle + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2.
\end{aligned}$$

Since $A_{k+1} \geq aL\alpha_{k+1}^2$ (see Lemma E.1) and $a \geq 1$ we can continue our derivations:

$$\begin{aligned}
\alpha_{k+1}\left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z^k - z\right\rangle &\leq A_{k+1}\left(f(x^{k+1}) - f(y^{k+1})\right) + \alpha_{k+1}\left\langle \theta_{k+1}, z^k - z^{k+1}\right\rangle \\
&\quad + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2.
\end{aligned} \tag{53}$$

Next, due to convexity of f we have

$$\begin{aligned}\left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), y^k - x^{k+1} \right\rangle &\overset{(37)}{=} \left\langle \nabla f(x^{k+1}), y^k - x^{k+1} \right\rangle + \left\langle \theta_{k+1}, y^k - x^{k+1} \right\rangle \\ &\leq f(y^k) - f(x^{k+1}) + \left\langle \theta_{k+1}, y^k - x^{k+1} \right\rangle. \end{aligned} \tag{54}$$

By definition of x^{k+1} we have $x^{k+1} = \frac{A_k y^k + \alpha_{k+1} z^k}{A_{k+1}}$ which implies

$$\alpha_{k+1}\left(x^{k+1} - z^k\right) = A_k\left(y^k - x^{k+1}\right) \tag{55}$$

since $A_{k+1} = A_k + \alpha_{k+1}$. Putting all together we derive that

$$\begin{aligned}\alpha_{k+1}\left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), x^{k+1} - z \right\rangle &= \alpha_{k+1}\left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), x^{k+1} - z^k \right\rangle \\ &\quad + \alpha_{k+1}\left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z^k - z \right\rangle \\ &\overset{(55)}{=} A_k\left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), y^k - x^{k+1} \right\rangle \\ &\quad + \alpha_{k+1}\left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z^k - z \right\rangle \\ &\overset{(54),(53)}{\leq} A_k\left(f(y^k) - f(x^{k+1})\right) + A_k\left\langle \theta_{k+1}, y^k - x^{k+1} \right\rangle \\ &\quad + A_{k+1}\left(f(x^{k+1}) - f(y^{k+1})\right) + \alpha_{k+1}\left\langle \theta_{k+1}, z^k - z^{k+1} \right\rangle \\ &\quad + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 \\ &\overset{(55)}{=} A_k f(y^k) - A_{k+1} f(y^{k+1}) + \alpha_{k+1}\left\langle \theta_{k+1}, x^{k+1} - z^k \right\rangle \\ &\quad + \alpha_{k+1} f(x^{k+1}) + \alpha_{k+1}\left\langle \theta_{k+1}, z^k - z^{k+1} \right\rangle \\ &\quad + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 \\ &\leq A_k f(y^k) - A_{k+1} f(y^{k+1}) + \alpha_{k+1} f(x^{k+1}) \\ &\quad + \alpha_{k+1}\left\langle \theta_{k+1}, x^{k+1} - z^{k+1} \right\rangle \\ &\quad + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2. \end{aligned}$$

Rearranging the terms we get

$$\begin{aligned}A_{k+1} f(y^{k+1}) - A_k f(y^k) &\leq \alpha_{k+1}\left(f(x^{k+1}) + \left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z - x^{k+1} \right\rangle\right) + \frac{1}{2}\|z^k - z\|_2^2 \\ &\quad - \frac{1}{2}\|z^{k+1} - z\|_2^2 + \alpha_{k+1}\left\langle \theta_{k+1}, x^{k+1} - z^{k+1} \right\rangle \\ &\overset{(37)}{=} \alpha_{k+1}\left(f(x^{k+1}) + \left\langle \nabla f(x^{k+1}), z - x^{k+1} \right\rangle\right) \\ &\quad + \alpha_{k+1}\left\langle \theta_{k+1}, z - x^{k+1} \right\rangle + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 \\ &\quad + \alpha_{k+1}\left\langle \theta_{k+1}, x^{k+1} - z^{k+1} \right\rangle \\ &\leq \alpha_{k+1} f(z) + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 + \alpha_{k+1}\left\langle \theta_{k+1}, z - z^{k+1} \right\rangle \end{aligned}$$

where in the last inequality we use the convexity of f. Taking into account $A_0 = \alpha_0 = 0$ and $A_N = \sum_{k=0}^{N-1} \alpha_{k+1}$ we sum up these inequalities for $k = 0, \ldots, N-1$ and get

$$\begin{aligned} A_N f(y^N) &\leq A_N f(z) + \frac{1}{2}\|z^0 - z\|_2^2 - \frac{1}{2}\|z^N - z\|_2^2 + \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \theta_{k+1}, z - z^{k+1} \right\rangle \\ &= A_N f(z) + \frac{1}{2}\|z^0 - z\|_2^2 - \frac{1}{2}\|z^N - z\|_2^2 + \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \theta_{k+1}, z - z^{k} \right\rangle \\ &\quad + \sum_{k=0}^{N-1} \alpha_{k+1}^2 \left\langle \theta_{k+1}, \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) \right\rangle \\ &\overset{(37)}{=} A_N f(z) + \frac{1}{2}\|z^0 - z\|_2^2 - \frac{1}{2}\|z^N - z\|_2^2 + \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \theta_{k+1}, z - z^{k} \right\rangle \\ &\quad + \sum_{k=0}^{N-1} \alpha_{k+1}^2 \|\theta_{k+1}\|_2^2 + \sum_{k=0}^{N-1} \alpha_{k+1}^2 \left\langle \theta_{k+1}, \nabla f(x^{k+1}) \right\rangle \end{aligned}$$

which concludes the proof.

F.3.2 Proof of Lemma F.5

Proof of (38)**.** By definition of $\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)$ we have that $\|\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)\|_2 \leq \lambda_{k+1}$ and, as a consequence, $\left\|\mathbb{E}_{\boldsymbol{\xi}^k}[\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)]\right\|_2 \leq \lambda_{k+1}$. Using this we get

$$\left\|\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) - \mathbb{E}_{\boldsymbol{\xi}^k}\left[\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)\right]\right\|_2 \leq \left\|\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)\right\|_2 + \left\|\mathbb{E}_{\boldsymbol{\xi}^k}\left[\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)\right]\right\|_2 \leq 2\lambda_{k+1}.$$

Proof of (39)**.** In order to prove this bound we introduce following indicator random variables:

$$\chi_k \overset{\text{def}}{=} \mathbb{1}_{\|\nabla f(x^{k+1}, \boldsymbol{\xi}^k)\|_2 > \lambda_{k+1}}, \quad \eta_k \overset{\text{def}}{=} \mathbb{1}_{\|\nabla f(x^{k+1}, \boldsymbol{\xi}^k) - \nabla f(x^{k+1})\|_2 > \frac{1}{2}\lambda_{k+1}}. \tag{56}$$

From the assumptions of the lemma, we have that $\|\nabla f(x^{k+1})\|_2 \leq \frac{\lambda_{k+1}}{2}$ which implies

$$\begin{aligned} \left\|\nabla f(x^{k+1}, \boldsymbol{\xi}^k)\right\|_2 &\leq \left\|\nabla f(x^{k+1}, \boldsymbol{\xi}^k) - \nabla f(x^{k+1})\right\|_2 + \left\|\nabla f(x^{k+1})\right\|_2 \\ &\leq \left\|\nabla f(x^{k+1}, \boldsymbol{\xi}^k) - \nabla f(x^{k+1})\right\|_2 + \frac{\lambda_{k+1}}{2}, \end{aligned}$$

hence

$$\chi_k \leq \eta_k. \tag{57}$$

The introduced notation helps us to rewrite $\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)$ in the following way:

$$\begin{aligned} \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) &= \nabla f(x^{k+1}, \boldsymbol{\xi}^k)(1 - \chi_k) + \frac{\lambda_{k+1}}{\|\nabla f(x^{k+1}, \boldsymbol{\xi})\|_2} \nabla f(x^{k+1}, \boldsymbol{\xi}^k)\chi_k \qquad (58) \\ &= \nabla f(x^{k+1}, \boldsymbol{\xi}^k) + \left(\frac{\lambda_{k+1}}{\left\|\nabla f(x^{k+1}, \boldsymbol{\xi}^k)\right\|_2} - 1\right) \nabla f(x^{k+1}, \boldsymbol{\xi}^k)\chi_k. \qquad (59) \end{aligned}$$

We use this representation to obtain the following inequality:

$$
\begin{aligned}
\left\|\mathbb{E}_{\boldsymbol{\xi}^k}\left[\widetilde{\nabla} f(x^{k+1},\boldsymbol{\xi}^k)\right]-\nabla f(x^{k+1})\right\|_2 &\overset{(23),(59)}{=} \left\|\mathbb{E}_{\boldsymbol{\xi}^k}\left[\left(\frac{\lambda_{k+1}}{\left\|\nabla f(x^{k+1},\boldsymbol{\xi}^k)\right\|_2}-1\right)\nabla f(x^{k+1},\boldsymbol{\xi}^k)\chi_k\right]\right\|_2 \\
&\leq \mathbb{E}_{\boldsymbol{\xi}^k}\left[\left\|\nabla f(x^{k+1},\boldsymbol{\xi}^k)\right\|_2\cdot\left|\frac{\lambda_{k+1}}{\left\|\nabla f(x^{k+1},\boldsymbol{\xi}^k)\right\|_2}-1\right|\chi_k\right] \\
&\overset{(56)}{=} \mathbb{E}_{\boldsymbol{\xi}^k}\left[\left\|\nabla f(x^{k+1},\boldsymbol{\xi}^k)\right\|_2\cdot\left(1-\frac{\lambda_{k+1}}{\left\|\nabla f(x^{k+1},\boldsymbol{\xi}^k)\right\|_2}\right)\chi_k\right] \\
&\overset{(56)}{\leq} \mathbb{E}_{\boldsymbol{\xi}^k}\left[\left\|\nabla f(x^{k+1},\boldsymbol{\xi}^k)\right\|_2\chi_k\right] \\
&\overset{(57)}{\leq} \mathbb{E}_{\boldsymbol{\xi}^k}\left[\left\|\nabla f(x^{k+1},\boldsymbol{\xi}^k)\right\|_2\eta_k\right] \\
&\leq \mathbb{E}_{\boldsymbol{\xi}^k}\left[\left\|\nabla f(x^{k+1},\boldsymbol{\xi}^k)-\nabla f(x^{k+1})\right\|_2\eta_k\right] \\
&\quad + \left\|\nabla f(x^{k+1})\right\|_2\mathbb{E}_{\boldsymbol{\xi}^k}\left[\eta_k\right] \\
&\leq \sqrt{\mathbb{E}_{\boldsymbol{\xi}^k}\left[\left\|\nabla f(x^{k+1},\boldsymbol{\xi}^k)-\nabla f(x^{k+1})\right\|_2^2\right]\mathbb{E}_{\boldsymbol{\xi}^k}[\eta_k^2]} \\
&\quad + \left\|\nabla f(x^{k+1})\right\|_2\mathbb{E}_{\boldsymbol{\xi}^k}\left[\eta_k\right] \\
&\overset{(24)}{\leq} \frac{\sigma}{\sqrt{m_k}}\sqrt{\mathbb{E}_{\boldsymbol{\xi}^k}\left[\eta_k^2\right]}+\frac{\lambda_{k+1}}{2}\mathbb{E}_{\boldsymbol{\xi}^k}\left[\eta_k\right]. \qquad (60)
\end{aligned}
$$

Next, we derive an upper bound for the expectation of η_k using Markov's inequality:

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\xi}^k}\left[\eta_k\right] &= \mathbb{E}_{\boldsymbol{\xi}^k}\left[\eta_k^2\right]=\mathbb{P}_{\boldsymbol{\xi}^k}\{\eta_k=1\} \\
&\overset{(56)}{=} \mathbb{P}_{\boldsymbol{\xi}^k}\left\{\left\|\nabla f(x^{k+1},\boldsymbol{\xi}^k)-\nabla f(x^{k+1})\right\|_2>\frac{\lambda_{k+1}}{2}\right\} \\
&\leq \frac{4\mathbb{E}_{\boldsymbol{\xi}^k}\left[\left\|\nabla f(x^{k+1},\boldsymbol{\xi}^k)-\nabla f(x^{k+1})\right\|_2^2\right]}{\lambda_{k+1}^2}\overset{(24)}{\leq}\frac{4\sigma^2}{m_k\lambda_{k+1}^2}. \qquad (61)
\end{aligned}
$$

Putting all together we derive (39):

$$
\left\|\mathbb{E}_{\boldsymbol{\xi}^k}\left[\widetilde{\nabla} f(x^{k+1},\boldsymbol{\xi}^k)\right]-\nabla f(x^{k+1})\right\|_2 \overset{(60),(61)}{\leq} \frac{2\sigma^2}{m_k\lambda_{k+1}}+\frac{\lambda_{k+1}}{2}\cdot\frac{4\sigma^2}{m_k\lambda_{k+1}^2}=\frac{4\sigma^2}{m_k\lambda_{k+1}}.
$$

Proof of (40)**.** Recall that in the space of random variables with finite second moment, i.e. in L_2, one can introduce a norm as $\sqrt{\mathbb{E}|X|^2}$ for an arbitrary random variable X from this space. Using triangle

inequality for this norm we get

$$\begin{aligned}
\sqrt{\mathbb{E}_{\boldsymbol{\xi}^k}\left[\left\|\nabla f(x^{k+1},\boldsymbol{\xi}^k)-\nabla f(x^{k+1})\right\|_2^2\right]} &\overset{(58)}{\leq} \sqrt{\mathbb{E}_{\boldsymbol{\xi}^k}\left[\left\|\frac{\lambda_{k+1}\nabla f(x^{k+1},\boldsymbol{\xi}^k)}{\left\|\nabla f(x^{k+1},\boldsymbol{\xi}^k)\right\|_2}-\nabla f(x^{k+1})\right\|_2^2\chi_k^2\right]} \\
&\quad+\sqrt{\mathbb{E}_{\boldsymbol{\xi}^k}\left[\left\|\nabla f(x^{k+1},\boldsymbol{\xi}^k)-\nabla f(x^{k+1})\right\|_2^2(1-\chi_k)^2\right]} \\
&\overset{(12)}{\leq} \sqrt{\mathbb{E}_{\boldsymbol{\xi}^k}\left[\left(2\left\|\frac{\lambda_{k+1}\nabla f(x^{k+1},\boldsymbol{\xi}^k)}{\left\|\nabla f(x^{k+1},\boldsymbol{\xi}^k)\right\|_2}\right\|_2^2+2\left\|\nabla f(x^{k+1})\right\|_2^2\right)\chi_k^2\right]} \\
&\quad+\sqrt{\mathbb{E}_{\boldsymbol{\xi}^k}\left[\left\|\nabla f(x^{k+1},\boldsymbol{\xi}^k)-\nabla f(x^{k+1})\right\|_2^2\right]} \\
&\overset{(24)}{\leq} \sqrt{\frac{5}{2}}\lambda_{k+1}\sqrt{\mathbb{E}_{\boldsymbol{\xi}^k}\left[\chi_k^2\right]}+\frac{\sigma}{\sqrt{m_k}} \\
&\overset{(57),(61)}{\leq} \sqrt{\frac{5}{2}}\lambda_{k+1}\cdot\frac{2\sigma}{\sqrt{m_k}\lambda_{k+1}}+\frac{\sigma}{\sqrt{m_k}}=\left(\sqrt{10}+1\right)\frac{\sigma}{\sqrt{m_k}} \\
&\leq \frac{\sqrt{18}\sigma}{\sqrt{m_k}}.
\end{aligned}$$

Proof of (41)**.** To derive (41) we use (40):

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\xi}^k}\left[\left\|\widetilde{\nabla}f(x^{k+1},\boldsymbol{\xi}^k)-\mathbb{E}_{\boldsymbol{\xi}^k}\left[\widetilde{\nabla}f(x^{k+1},\boldsymbol{\xi}^k)\right]\right\|_2^2\right] &\overset{(15)}{\leq} \mathbb{E}_{\boldsymbol{\xi}^k}\left[\left\|\widetilde{\nabla}f(x^{k+1},\boldsymbol{\xi}^k)-\nabla f(x^{k+1})\right\|_2^2\right] \\
&\overset{(40)}{\leq} \frac{18\sigma^2}{m_k}.
\end{aligned}$$

F.3.3 Proof of Theorem F.1

Lemma F.4 implies that the inequality

$$\begin{aligned}
A_N\left(f(y^N)-f(x^*)\right) &\leq \frac{1}{2}\|z^0-x^*\|_2^2-\frac{1}{2}\|z^N-x^*\|_2^2+\sum_{k=0}^{N-1}\alpha_{k+1}\left\langle\theta_{k+1},x^*-z^k\right\rangle \\
&\quad+\sum_{k=0}^{N-1}\alpha_{k+1}^2\left\|\theta_{k+1}\right\|_2^2+\sum_{k=0}^{N-1}\alpha_{k+1}^2\left\langle\theta_{k+1},\nabla f(x^{k+1})\right\rangle, \qquad (62)
\end{aligned}$$

$$\theta_{k+1} \overset{\text{def}}{=} \widetilde{\nabla}f(x^{k+1},\boldsymbol{\xi}^k)-\nabla f(x^{k+1}) \qquad (63)$$

holds for all $N\geq 0$. Taking into account that $f(y^N)-f(x^*)\geq 0$ for all y^N and using new notation $R_k\overset{\text{def}}{=}\|z^k-x^*\|_2$, $\widetilde{R}_0=R_0$, $\widetilde{R}_{k+1}=\max\{\widetilde{R}_k,R_{k+1}\}$ we derive that for all $k\geq 0$

$$R_k^2\leq R_0^2+2\sum_{l=0}^{k-1}\alpha_{l+1}\left\langle\theta_{l+1},x^*-z^l\right\rangle+2\sum_{l=0}^{k-1}\alpha_{l+1}^2\left\langle\theta_{l+1},\nabla f(x^{l+1})\right\rangle+2\sum_{l=0}^{k-1}\alpha_{l+1}^2\|\theta_{l+1}\|_2^2. \qquad (64)$$

First of all, we notice that for each $k\geq 0$ iterates x^{k+1},z^k,y^k lie in the ball $B_{\widetilde{R}_k}(x^*)$. We prove it using induction. Since $y^0=z^0=x^0$, $\widetilde{R}_0=R_0=\|z^0-x^*\|_2$ and $x^1=\frac{A_0y^0+\alpha_1z^0}{A_1}=z^0$ we have that $x^1,z^0,y^0\in B_{\widetilde{R}_0}(x^*)$. Next, assume that $x^l,z^{l-1},y^{l-1}\in B_{\widetilde{R}_{l-1}}(x^*)$ for some $l\geq 1$. By definitions of R_l and $\widetilde{R}_l$ we have that $z^l\in B_{R_l}(x^*)\subseteq B_{\widetilde{R}_l}(x^*)$. Since y^l is a convex combination of $y^{l-1}\in B_{\widetilde{R}_{l-1}}(x^*)\subseteq B_{\widetilde{R}_l}(x^*)$, $z^l\in B_{\widetilde{R}_l}(x^*)$ and $B_{\widetilde{R}_l}(x^*)$ is a convex set we conclude that

$y^l \in B_{\widetilde{R}_l}(x^*)$. Finally, since x^{l+1} is a convex combination of y^l and z^l we have that x^{l+1} lies in $B_{\widetilde{R}_l}(x^*)$ as well.

The rest of the proof is based on the refined analysis of inequality (64). In particular, via induction we prove that for all $k = 0, 1, \ldots, N$ with probability at least $1 - \frac{k\beta}{N}$ the following statement holds: inequalities

$$\begin{aligned} R_t^2 &\overset{(64)}{\leq} R_0^2 + 2\sum_{l=0}^{t-1} \alpha_{l+1} \left\langle \theta_{l+1}, x^* - z^l \right\rangle + 2\sum_{l=0}^{t-1} \alpha_{l+1}^2 \left\langle \theta_{l+1}, \nabla f(x^{l+1}) \right\rangle + 2\sum_{l=0}^{t-1} \alpha_{k+1}^2 \|\theta_{l+1}\|_2^2 \\ &\leq C^2 R_0^2 \end{aligned} \tag{65}$$

hold for $t = 0, 1, \ldots, k$ simultaneously where C is defined in (29). Let us define the probability event when this statement holds as E_k. Then, our goal is to show that $\mathbb{P}\{E_k\} \geq 1 - \frac{k\beta}{N}$ for all $k = 0, 1, \ldots, N$. For $t = 0$ inequality (65) holds with probability 1 since $C \geq 1$, hence $\mathbb{P}\{E_0\} = 1$. Next, assume that for some $k = T - 1 \leq N - 1$ we have $\mathbb{P}\{E_k\} = \mathbb{P}\{E_{T-1}\} \geq 1 - \frac{(T-1)\beta}{N}$. Let us prove that $\mathbb{P}\{E_T\} \geq 1 - \frac{T\beta}{N}$. First of all, probability event E_{T-1} implies that

$$\begin{aligned} f(y^t) - f(x^*) &\overset{(62)}{\leq} \frac{1}{A_t}\left(\frac{1}{2}R_0^2 + \sum_{l=0}^{t-1} \alpha_{l+1} \left\langle \theta_{l+1}, x^* - z^l + \alpha_{l+1}\nabla f(x^{l+1}) \right\rangle + \sum_{l=0}^{t-1} \alpha_{k+1}^2 \|\theta_{l+1}\|_2^2\right) \\ &\overset{(65)}{\leq} \frac{C^2 R_0^2}{2A_t} \end{aligned} \tag{66}$$

hold for $t = 0, 1, \ldots, T-1$. Then, inequalities

$$\begin{aligned} \left\|\nabla f(x^1)\right\|_2 &= \left\|\nabla f(z^0)\right\|_2 \overset{(6)}{\leq} L\|z^0 - x^*\|_2 = \frac{1}{a} \cdot \frac{R_0}{\alpha_1}, \\ \left\|\nabla f(x^{t+1})\right\|_2 &\leq \left\|\nabla f(x^{t+1}) - \nabla f(y^t)\right\|_2 + \left\|\nabla f(y^t)\right\|_2 \\ &\overset{(6),(8)}{\leq} L\|x^{t+1} - y^t\|_2 + \sqrt{2L(f(y^t) - f(x^*))} \\ &\overset{(55),(66)}{\leq} \frac{\alpha_{t+1}L}{A_t}\|x^{t+1} - z^k\|_2 + \sqrt{\frac{LC^2R_0^2}{A_t}} \\ &\overset{(20)}{\leq} \frac{2L(t+2)}{t(t+3)}\left(\|x^{k+1} - x^*\|_2 + \|x^* - z^k\|_2\right) + \frac{2LCR_0\sqrt{a}}{\sqrt{t(t+3)}} \\ &\leq \frac{4L(t+2)\widetilde{R}_k}{t(t+3)} + \frac{2LCR_0\sqrt{a}}{\sqrt{t(t+3)}} \\ &\overset{(65)}{\leq} \frac{2aLCR_0}{t+2}\left(\frac{2(t+2)^2}{at(t+3)} + \frac{t+2}{\sqrt{at(t+3)}}\right) \\ &\leq \frac{CR_0}{\alpha_{t+1}}\left(\frac{9}{2a} + \frac{3}{2\sqrt{a}}\right) \end{aligned}$$

hold for $t = 1, \ldots, T-1$ where the last inequality follows from $\frac{(t+2)^2}{t(t+3)} \leq \frac{(1+2)^2}{1(1+3)} = \frac{9}{4}$. Taking a such that

$$a \geq \frac{2R_0}{B} \quad \text{and} \quad \frac{9}{2a} + \frac{3}{2\sqrt{a}} \leq \frac{B}{2CR_0}$$

we obtain that probability event E_{T-1} implies

$$\left\|\nabla f(x^{t+1})\right\|_2 \leq \frac{B}{2\alpha_{t+1}} = \frac{\lambda_{t+1}}{2} \tag{67}$$

for $t = 0, \ldots, T-1$. Since $B = \frac{CR_0}{8\ln\frac{4N}{\beta}}$ we have to choose such a that

$$a \geq \frac{16\ln\frac{4N}{\beta}}{C} \quad \text{and} \quad \frac{9}{a} + \frac{3}{\sqrt{a}} \leq \frac{1}{8\ln\frac{4N}{\beta}}.$$

Solving quadratic inequality

$$a - 24\sqrt{a}\ln\frac{4N}{\beta} - 72\ln\frac{4N}{\beta} \geq 0$$

w.r.t. $\sqrt{a}$ we get that a should satisfy

$$a \geq \max\left\{\frac{16\ln\frac{4N}{\beta}}{C}, 36\left(2\ln\frac{4N}{\beta} + \sqrt{4\ln^2\frac{4N}{\beta} + 2\ln\frac{4N}{\beta}}\right)^2\right\}.$$

Having inequalities (67) in hand we show in the rest of the proof that (65) holds for $t = T$ with big enough probability. First of all, we introduce new random variables:

$$\eta_l = \begin{cases} x^* - z^l, & \text{if } \|x^* - z^l\|_2 \leq CR_0, \\ 0, & \text{otherwise,} \end{cases} \quad \text{and} \quad \zeta_l = \begin{cases} \nabla f(x^{l+1}), & \text{if } \|\nabla f(x^{l+1})\|_2 \leq \frac{B}{2\alpha_{l+1}}, \\ 0, & \text{otherwise,} \end{cases} \tag{68}$$

for $l = 0, 1, \ldots T-1$. Note that these random variables are bounded with probability 1, i.e. with probability 1 we have

$$\|\eta_l\|_2 \leq CR_0 \quad \text{and} \quad \|\zeta_l\|_2 \leq \frac{B}{2\alpha_{l+1}}. \tag{69}$$

Secondly, we use the introduced notation and get that E_{T-1} implies

$$\begin{aligned} R_T^2 &\overset{(64),(65),(67),(68)}{\leq} R_0^2 + 2\sum_{l=0}^{T-1}\alpha_{l+1}\langle\theta_{l+1}, \eta_l\rangle + 2\sum_{l=0}^{T-1}\alpha_{l+1}^2\|\theta_{l+1}\|_2^2 + 2\sum_{l=0}^{T-1}\alpha_{l+1}^2\langle\theta_{l+1}, \zeta_l\rangle \\ &= R_0^2 + \sum_{l=0}^{T-1}\alpha_{l+1}\langle\theta_{l+1}, 2\eta_l + 2\alpha_{l+1}\zeta_l\rangle + 2\sum_{l=0}^{T-1}\alpha_{l+1}^2\|\theta_{l+1}\|_2^2. \end{aligned}$$

Finally, we do some preliminaries in order to apply Bernstein's inequality (see Lemma D.1) and obtain that E_{T-1} implies

$$\begin{aligned} R_T^2 \overset{(12)}{\leq} R_0^2 &+ \underbrace{\sum_{l=0}^{T-1}\alpha_{l+1}\left\langle\theta_{l+1}^u, 2\eta_l + 2\alpha_{l+1}\zeta_l\right\rangle}_{①} + \underbrace{\sum_{l=0}^{T-1}\alpha_{l+1}\left\langle\theta_{l+1}^b, 2\eta_l + 2\alpha_{l+1}\zeta_l\right\rangle}_{②} \\ &+ \underbrace{\sum_{l=0}^{T-1}4\alpha_{l+1}^2\left(\|\theta_{l+1}^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right]\right)}_{③} + \underbrace{\sum_{l=0}^{T-1}4\alpha_{l+1}^2\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right]}_{④} \\ &+ \underbrace{\sum_{l=0}^{T-1}4\alpha_{l+1}^2\|\theta_{l+1}^b\|_2^2}_{⑤} \end{aligned} \tag{70}$$

where we introduce new notations:

$$\theta_{l+1}^u \overset{\text{def}}{=} \widetilde{\nabla} f(x^{l+1}, \boldsymbol{\xi}^l) - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\widetilde{\nabla} f(x^{l+1}, \boldsymbol{\xi}^l)\right], \quad \theta_{l+1}^b \overset{\text{def}}{=} \mathbb{E}_{\boldsymbol{\xi}^l}\left[\widetilde{\nabla} f(x^{l+1}, \boldsymbol{\xi}^l)\right] - \nabla f(x^{l+1}), \tag{71}$$

$$\theta_{l+1} \overset{(37)}{=} \theta_{l+1}^u + \theta_{l+1}^b.$$

It remains to provide tight upper bounds for ①, ②, ③, ④ and ⑤, i.e. in the remaining part of the proof we show that ① + ② + ③ + ④ + ⑤ $\leq \delta C^2 R_0^2$ for some $\delta < 1$.

Upper bound for ①. First of all, since $\mathbb{E}_{\boldsymbol{\xi}^l}[\theta_{l+1}^u] = 0$ summands in ① are conditionally unbiased:

$$\mathbb{E}_{\boldsymbol{\xi}^l}\left[\alpha_{l+1}\left\langle\theta_{l+1}^u, 2\eta_l + 2\alpha_{l+1}\zeta_l\right\rangle\right] = 0.$$

Secondly, these summands are bounded with probability 1:

$$\begin{aligned}\left|\alpha_{l+1}\left\langle\theta^u_{l+1}, 2\eta_l + 2\alpha_{l+1}\zeta_l\right\rangle\right| &\leq \alpha_{l+1}\|\theta^u_{l+1}\|_2\,\|2\eta_l + 2\alpha_{l+1}\zeta_l\|_2 \\ &\overset{(38),(69)}{\leq} 2\alpha_{l+1}\lambda_{l+1}\left(2CR_0 + B\right) = 2B(2CR_0+B) \\ &= \frac{C^2R_0^2}{2\ln\frac{4N}{\beta}} + \frac{C^2R_0^2}{32\ln^2\frac{4N}{\beta}} \\ &\overset{(25)}{\leq} \frac{C^2R_0^2}{2\ln\frac{4N}{\beta}} + \frac{C^2R_0^2}{64\ln\frac{4N}{\beta}} \leq \frac{33C^2R_0^2}{64\ln\frac{4N}{\beta}}.\end{aligned}$$

Finally, one can bound conditional variances $\sigma_l^2 \overset{\text{def}}{=} \mathbb{E}_{\boldsymbol{\xi}^l}\left[\alpha_{l+1}^2\left\langle\theta^u_{l+1}, 2\eta_l + 2\alpha_{l+1}\zeta_l\right\rangle^2\right]$ in the following way:

$$\begin{aligned}\sigma_l^2 &\leq \mathbb{E}_{\boldsymbol{\xi}^l}\left[\alpha_{l+1}^2\left\|\theta^u_{l+1}\right\|_2^2\|2\eta_l + 2\alpha_{l+1}\zeta_l\|_2^2\right] \\ &\overset{(69)}{\leq} \alpha_{l+1}^2\mathbb{E}_{\boldsymbol{\xi}^l}\left[\left\|\theta^u_{l+1}\right\|_2^2\right](2CR_0+B)^2. \end{aligned} \tag{72}$$

In other words, sequence $\left\{\alpha_{l+1}\left\langle\theta^u_{l+1}, 2\eta_l + 2\alpha_{l+1}\zeta_l\right\rangle\right\}_{l\geq 0}$ is bounded martingale difference sequence with bounded conditional variances $\{\sigma_l^2\}_{l\geq 0}$. Therefore, we can apply Bernstein's inequality, i.e. we apply Lemma D.1 with $X_l = \alpha_{l+1}\left\langle\theta^u_{l+1}, 2\eta_l + 2\alpha_{l+1}\zeta_l\right\rangle$, $c = \frac{33C^2R_0^2}{64\ln\frac{4N}{\beta}}$ and $F = \frac{c^2\ln\frac{4N}{\beta}}{18}$ and get that for all $b > 0$

$$\mathbb{P}\left\{\left|\sum_{l=0}^{T-1} X_l\right| > b \text{ and } \sum_{l=0}^{T-1}\sigma_l^2 \leq F\right\} \leq 2\exp\left(-\frac{b^2}{2F + {}^{2cb}/_3}\right)$$

or, equivalently, with probability at least $1 - 2\exp\left(-\frac{b^2}{2F+{}^{2cb}/_3}\right)$

$$\text{either } \sum_{l=0}^{T-1}\sigma_l^2 > F \quad \text{or} \quad \underbrace{\left|\sum_{l=0}^{T-1} X_l\right|}_{|①|} \leq b.$$

The choice of F will be clarified further, let us now choose b in such a way that $2\exp\left(-\frac{b^2}{2F+{}^{2cb}/_3}\right) = \frac{\beta}{2N}$. This implies that b is the positive root of the quadratic equation

$$b^2 - \frac{2c\ln\frac{4N}{\beta}}{3}b - 2F\ln\frac{4N}{\beta} = 0,$$

hence

$$\begin{aligned} b &= \frac{c\ln\frac{4N}{\beta}}{3} + \sqrt{\frac{c^2\ln^2\frac{4N}{\beta}}{9} + 2F\ln\frac{4N}{\beta}} \leq \frac{c\ln\frac{4N}{\beta}}{3} + \sqrt{\frac{2c^2\ln^2\frac{4N}{\beta}}{9}} \\ &= \frac{1+\sqrt{2}}{3}c\ln\frac{4N}{\beta} \leq \frac{33C^2R_0^2}{64}. \end{aligned}$$

That is, with probability at least $1 - \frac{\beta}{2N}$

$$\underbrace{\text{either } \sum_{l=0}^{T-1}\sigma_l^2 > F \quad \text{or} \quad |①| \leq \frac{33C^2R_0^2}{64}}_{\text{probability event } E_①}.$$

Next, we notice that probability event E_{T-1} implies that

$$\begin{aligned}
\sum_{l=0}^{T-1}\sigma_l^2 &\overset{(72)}{\leq} (2CR_0+B)^2\sum_{l=0}^{T-1}\alpha_{l+1}^2\mathbb{E}_{\boldsymbol{\xi}^l}\left[\left\|\theta_{l+1}^u\right\|_2^2\right]\\
&\overset{(41),(67)}{\leq} 18\sigma^2C^2R_0^2\left(2+\frac{1}{8\ln\frac{4N}{\beta}}\right)^2\sum_{l=0}^{T-1}\frac{\alpha_{l+1}^2}{m_l}\\
&\overset{(25),(26)}{\leq} 18\sigma^2C^2R_0^2\left(2+\frac{1}{16}\right)^2\sum_{l=0}^{T-1}\frac{\alpha_{l+1}^2C^2R_0^2}{6000\sigma^2\alpha_{l+1}^2N\ln\frac{4N}{\beta}}\\
&\overset{T\leq N}{\leq} \frac{18\left(2+\frac{1}{16}\right)^2}{6000\ln\frac{4N}{\beta}}C^4R_0^4\sum_{l=0}^{N-1}\frac{1}{N}\leq\frac{c^2\ln\frac{4N}{\beta}}{18}=F,
\end{aligned}$$

where the last inequality follows from $c=\frac{33C^2R_0^2}{64\ln\frac{4N}{\beta}}$ and simple arithmetic.

Upper bound for ②. First of all, we notice that probability event E_{T-1} implies

$$\begin{aligned}
\alpha_{l+1}\left\langle\theta_{l+1}^b,2\eta_l+2\alpha_{l+1}\zeta_l\right\rangle &\leq \alpha_{l+1}\left\|\theta_{l+1}^b\right\|_2\left\|2\eta_l+2\alpha_{l+1}\zeta_l\right\|_2\\
&\overset{(39),(69)}{\leq} \alpha_{l+1}\cdot\frac{4\sigma^2}{m_l\lambda_{l+1}}(2CR_0+B)\\
&= \frac{32\alpha_{l+1}^2\sigma^2\ln\frac{4N}{\beta}}{m_lCR_0}\left(2CR_0+\frac{CR_0}{8\ln\frac{4N}{\beta}}\right)\\
&\overset{(25),(26)}{\leq} \frac{32\alpha_{l+1}^2\sigma^2C^2R_0^2\ln\frac{4N}{\beta}}{6000\alpha_{l+1}^2N\sigma^2\ln\frac{4N}{\beta}}\left(2+\frac{1}{16}\right)\\
&= \frac{11C^2R_0^2}{1000N}.
\end{aligned}$$

This implies that

$$② = \sum_{l=0}^{T-1}\alpha_{l+1}\left\langle\theta_{l+1}^b,2\eta_l+2\alpha_{l+1}\zeta_l\right\rangle \overset{T\leq N}{\leq}\frac{11C^2R_0^2}{1000}.$$

Upper bound for ③. We derive the upper bound for ③ using the same technique as for ①. First of all, we notice that the summands in ③ are conditionally independent:

$$\mathbb{E}_{\boldsymbol{\xi}^l}\left[4\alpha_{l+1}^2\left(\|\theta_{l+1}^u\|_2^2-\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right]\right)\right]=0.$$

Secondly, the summands are bounded with probability 1:

$$\begin{aligned}
\left|4\alpha_{l+1}^2\left(\|\theta_{l+1}^u\|_2^2-\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right]\right)\right| &\leq 4\alpha_{l+1}^2\left(\|\theta_{l+1}^u\|_2^2+\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right]\right)\\
&\overset{(38)}{\leq} 4\alpha_{l+1}^2\left(4\lambda_{l+1}^2+4\lambda_{l+1}^2\right)\\
&= 32B^2=\frac{C^2R_0^2}{2\ln^2\frac{4N}{\beta}}\overset{(25)}{\leq}\frac{C^2R_0^2}{4\ln\frac{4N}{\beta}}\overset{\text{def}}{=}c_1. \qquad (73)
\end{aligned}$$

Finally, one can bound conditional variances $\hat{\sigma}_l^2\overset{\text{def}}{=}\mathbb{E}_{\boldsymbol{\xi}^l}\left[\left|4\alpha_{l+1}^2\left(\|\theta_{l+1}^u\|_2^2-\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right]\right)\right|^2\right]$ in the following way:

$$\begin{aligned}
\hat{\sigma}_l^2 &\overset{(73)}{\leq} c_1\mathbb{E}_{\boldsymbol{\xi}^l}\left[\left|4\alpha_{l+1}^2\left(\|\theta_{l+1}^u\|_2^2-\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right]\right)\right|\right]\\
&\leq 4c_1\alpha_{l+1}^2\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2+\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right]\right]=8c_1\alpha_{l+1}^2\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right]. \qquad (74)
\end{aligned}$$

In other words, sequence $\left\{4\alpha_{l+1}^2\left(\|\theta_{l+1}^u\|_2^2-\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right]\right)\right\}_{l\geq0}$ is bounded martingale difference sequence with bounded conditional variances $\{\hat{\sigma}_l^2\}_{l\geq0}$. Therefore, we can apply Bernstein's inequality, i.e. we apply Lemma D.1 with $X_l=\hat{X}_l=4\alpha_{l+1}^2\left(\|\theta_{l+1}^u\|_2^2-\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right]\right)$, $c=c_1=\frac{C^2R_0^2}{4\ln\frac{4N}{\beta}}$

and $F = F_1 = \frac{c_1^2 \ln \frac{4N}{\beta}}{18}$ and get that for all $b > 0$

$$\mathbb{P}\left\{\left|\sum_{l=0}^{T-1} \hat{X}_l\right| > b \text{ and } \sum_{l=0}^{T-1} \hat{\sigma}_l^2 \le F_1\right\} \le 2\exp\left(-\frac{b^2}{2F_1 + {}^{2c_1 b}/_3}\right)$$

or, equivalently, with probability at least $1 - 2\exp\left(-\frac{b^2}{2F_1 + {}^{2c_1 b}/_3}\right)$

$$\text{either } \sum_{l=0}^{T-1} \hat{\sigma}_l^2 > F_1 \quad \text{or} \quad \underbrace{\left|\sum_{l=0}^{T-1} \hat{X}_l\right|}_{|③|} \le b.$$

As in our derivations of the upper bound for ① we choose such b that $2\exp\left(-\frac{b^2}{2F_1 + {}^{2c_1 b}/_3}\right) = \frac{\beta}{2N}$, i.e.

$$b = \frac{c_1 \ln \frac{4N}{\beta}}{3} + \sqrt{\frac{c_1^2 \ln^2 \frac{4N}{\beta}}{9} + 2F_1 \ln \frac{4N}{\beta}} \le \frac{1+\sqrt{2}}{3} c_1 \ln \frac{4N}{\beta} \le \frac{C^2 R_0^2}{4}.$$

That is, with probability at least $1 - \frac{\beta}{2N}$

$$\underbrace{\text{either } \sum_{l=0}^{T-1} \hat{\sigma}_l^2 > F_1 \quad \text{or} \quad |③| \le \frac{C^2 R_0^2}{4}}_{\text{probability event } E_③}.$$

Next, we notice that probability event E_{T-1} implies that

$$\begin{aligned}
\sum_{l=0}^{T-1} \hat{\sigma}_l^2 &\overset{(74)}{\le} 8c_1 \sum_{l=0}^{T-1} \alpha_{l+1}^2 \mathbb{E}_{\boldsymbol{\xi}^l}\left[\left\|\theta_{l+1}^u\right\|_2^2\right] \\
&\overset{(41),(67)}{\le} c_1 \sum_{l=0}^{T-1} \frac{144\sigma^2 \alpha_{l+1}^2}{m_l} \\
&\overset{(26)}{\le} c_1 \sum_{l=0}^{T-1} \frac{144\sigma^2 \alpha_{l+1}^2 C^2 R_0^2}{10368 \sigma^2 \alpha_{l+1}^2 N} \\
&\overset{T \le N}{\le} c_1 \cdot \underbrace{\frac{C^2 R_0^2}{4 \ln \frac{4N}{\beta}}}_{c_1} \cdot \frac{\ln \frac{4N}{\beta}}{18} = F_1.
\end{aligned}$$

Upper bound for ④. The probability event E_{T-1} implies

$$\begin{aligned}
④ &= \sum_{l=0}^{T-1} 4\alpha_{l+1}^2 \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right] \overset{(41),(67)}{\le} \sum_{l=0}^{T-1} \frac{72\alpha_{l+1}^2 \sigma^2}{m_l} \overset{(26)}{\le} \sum_{l=0}^{T-1} \frac{72\alpha_{l+1}^2 \sigma^2 C^2 R_0^2}{10368 \alpha_{l+1}^2 \sigma^2 N} \\
&\overset{T \le N}{\le} \frac{C^2 R_0^2}{144}.
\end{aligned}$$

Upper bound for ⑤. Again, we use corollaries of probability event E_{T-1}:

$$\begin{aligned}
⑤ &= \sum_{l=0}^{T-1} 4\alpha_{l+1}^2 \|\theta_{l+1}^b\|_2^2 \overset{(39),(67)}{\le} \sum_{l=0}^{T-1} \frac{64\alpha_{l+1}^2 \sigma^4}{m_l^2 \lambda_{l+1}^2} = \frac{1}{B^2} \sum_{l=0}^{T-1} \frac{64\alpha_{l+1}^4 \sigma^4}{m_l^2} \\
&\overset{(26)}{\le} \frac{64 \ln^2 \frac{4N}{\beta}}{C^2 R_0^2} \sum_{l=0}^{T-1} \frac{64\alpha_{l+1}^4 \sigma^4 C^4 R_0^4}{6000^2 \sigma^4 \alpha_{l+1}^4 N^2 \ln^2 \frac{4N}{\beta}} \\
&\overset{T \le N}{\le} \frac{16 C^2 R_0^2}{140625}.
\end{aligned}$$

Now we summarize all bound that we have: probability event E_{T-1} implies

$$
\begin{aligned}
R_T^2 &\overset{(64)}{\leq} R_0^2 + 2\sum_{l=0}^{T-1}\alpha_{l+1}\left\langle\theta_{l+1}, x^* - z^l\right\rangle + 2\sum_{l=0}^{k-1}\alpha_{l+1}^2\left\langle\theta_{l+1}, \nabla f(x^{l+1})\right\rangle + 2\sum_{l=0}^{T-1}\alpha_{l+1}^2\|\theta_{l+1}\|_2^2 \\
&\overset{(70)}{\leq} R_0^2 + ① + ② + ③ + ④ + ⑤, \\
② &\leq \frac{11C^2R_0^2}{1000}, \quad ④ \leq \frac{C^2R_0^2}{144}, \quad ⑤ \leq \frac{16C^2R_0^2}{140625}, \\
\sum_{l=0}^{T-1}\sigma_l^2 &\leq F, \quad \sum_{l=0}^{T-1}\hat{\sigma}_l^2 \leq F_1
\end{aligned}
$$

and

$$
\mathbb{P}\{E_{T-1}\} \geq 1 - \frac{(T-1)\beta}{N}, \quad \mathbb{P}\{E_①\} \geq 1 - \frac{\beta}{2N}, \quad \mathbb{P}\{E_③\} \geq 1 - \frac{\beta}{2N},
$$

where

$$
\begin{aligned}
E_① &= \left\{\text{either } \sum_{l=0}^{T-1}\sigma_l^2 > F \quad \text{or} \quad |①| \leq \frac{33C^2R_0^2}{64}\right\}, \\
E_③ &= \left\{\text{either } \sum_{l=0}^{T-1}\hat{\sigma}_l^2 > F_1 \quad \text{or} \quad |③| \leq \frac{C^2R_0^2}{4}\right\}.
\end{aligned}
$$

Taking into account these inequalities we get that probability event $E_{T-1} \cap E_① \cap E_③$ implies

$$
\begin{aligned}
R_T^2 &\overset{(64)}{\leq} R_0^2 + 2\sum_{l=0}^{T-1}\alpha_{l+1}\left\langle\theta_{l+1}, x^* - z^l\right\rangle + 2\sum_{l=0}^{k-1}\alpha_{l+1}^2\left\langle\theta_{l+1}, \nabla f(x^{l+1})\right\rangle + 2\sum_{l=0}^{T-1}\alpha_{l+1}^2\|\theta_{l+1}\|_2^2 \\
&\leq R_0^2 + \left(\frac{33}{64} + \frac{11}{1000} + \frac{1}{4} + \frac{1}{144} + \frac{16}{140625}\right)C^2R_0^2 \\
&\leq \left(1 + \frac{4}{5}C^2\right)R_0^2 \overset{(29)}{\leq} C^2R_0^2. \qquad (75)
\end{aligned}
$$

Moreover, using union bound we derive

$$
\mathbb{P}\left\{E_{T-1} \cap E_① \cap E_③\right\} = 1 - \mathbb{P}\left\{\overline{E}_{T-1} \cup \overline{E}_① \cup \overline{E}_③\right\} \geq 1 - \frac{T\beta}{N}. \qquad (76)
$$

That is, by definition of E_T and E_{T-1} we have proved that

$$
\mathbb{P}\{E_T\} \overset{(75)}{\geq} \mathbb{P}\left\{E_{T-1} \cap E_① \cap E_③\right\} \overset{(76)}{\geq} 1 - \frac{T\beta}{N},
$$

which implies that for all $k = 0, 1, \ldots, N$ we have $\mathbb{P}\{E_k\} \geq 1 - \frac{k\beta}{N}$. Then, for $k = N$ we have that with probability at least $1 - \beta$

$$
\begin{aligned}
A_N\left(f(y^N) - f(x^*)\right) &\overset{(62)}{\leq} \frac{1}{2}\|z^0 - z\|_2^2 - \frac{1}{2}\|z^N - z\|_2^2 + \sum_{k=0}^{N-1}\alpha_{k+1}\left\langle\theta_{k+1}, z - z^k\right\rangle \\
&\quad + \sum_{k=0}^{N-1}\alpha_{k+1}^2\left\|\theta_{k+1}\right\|_2^2 + \sum_{k=0}^{N-1}\alpha_{k+1}^2\left\langle\theta_{k+1}, \nabla f(x^{k+1})\right\rangle \\
&\overset{(65)}{\leq} \frac{C^2R_0^2}{2}.
\end{aligned}
$$

Since $A_N = \frac{N(N+3)}{4aL}$ (see Lemma E.1) we get that with probability at least $1 - \beta$

$$
f(y^N) - f(x^*) \leq \frac{2aLC^2R_0^2}{N(N+3)}.
$$

In other words, `clipped-SSTM` with $a = \max\left\{1, \frac{16\ln\frac{4N}{\beta}}{C}, 36\left(2\ln\frac{4N}{\beta} + \sqrt{4\ln^2\frac{4N}{\beta} + 2\ln\frac{4N}{\beta}}\right)^2\right\} = 36\left(2\ln\frac{4N}{\beta} + \sqrt{4\ln^2\frac{4N}{\beta} + 2\ln\frac{4N}{\beta}}\right)^2$ achieves $f(y^N) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$ after $O\left(\sqrt{\frac{LR_0^2}{\varepsilon}}\ln\frac{LR_0^2}{\varepsilon\beta}\right)$ iterations and requires

$$\begin{aligned}\sum_{k=0}^{N-1} m_k &\overset{(26)}{=} \sum_{k=0}^{N-1} O\left(\max\left\{1, \frac{\sigma^2\alpha_{k+1}^2 N\ln\frac{N}{\beta}}{R_0^2}\right\}\right)\\ &= O\left(\max\left\{N, \sum_{k=0}^{N-1}\frac{\sigma^2(k+2)^2 N\ln\frac{N}{\beta}}{a^2L^2R_0^2}\right\}\right)\\ &\overset{(27)}{=} O\left(\max\left\{N, \frac{\sigma^2N^4}{\ln^3\frac{N}{\beta}L^2R_0^2}\right\}\right)\\ &= O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2R_0^2}{\varepsilon^2}\right\}\ln\frac{LR_0^2}{\varepsilon\beta}\right).\end{aligned}$$

oracle calls.

F.3.4 Proof of Corollary F.2

Theorem F.1 implies that with probability at least $1-\beta$

$$f(y^N) - f(x^*) \overset{(28)}{\leq} \frac{2aLC^2R_0^2}{N(N+3)}, \tag{77}$$

where a satisfies

$$a \overset{(27)}{\geq} \max\left\{1, \frac{16\ln\frac{4N}{\beta}}{C}, 36\left(2\ln\frac{4N}{\beta} + \sqrt{4\ln^2\frac{4N}{\beta} + 2\ln\frac{4N}{\beta}}\right)^2\right\} \overset{\text{def}}{=} \hat{a}, \tag{78}$$

$\alpha_{k+1} = \frac{k+2}{2aL}$ and batchsizes m_k are chosen according to (26):

$$\begin{aligned}m_k &\overset{(26)}{=} \max\left\{1, \frac{1185\sigma^2\alpha_{k+1}^2N\ln\frac{4N}{\beta}}{C^2R_0^2}, \frac{10368\sigma^2\alpha_{k+1}^2N}{C^2R_0^2}\right\}\\ &= \max\left\{1, \frac{1185\sigma^2(k+2)^2N\ln\frac{4N}{\beta}}{4a^2L^2C^2R_0^2}, \frac{10368\sigma^2(k+2)^2N}{4a^2L^2C^2R_0^2}\right\}.\end{aligned} \tag{79}$$

We consider two different options for a.

1. If $N\ln\frac{4N}{\beta}$ is bigger than $\hat{a}$, then we take $a = N\ln\frac{4N}{\beta}$ which implies that

   $$m_k = \max\left\{1, \frac{1185\sigma^2(k+2)^2}{4L^2NC^2R_0^2\ln\frac{4N}{\beta}}, \frac{10368\sigma^2(k+2)^2}{4L^2C^2R_0^2N\ln^2\frac{4N}{\beta}}\right\} = O\left(\max\left\{1, \frac{\sigma^2(k+2)^2}{L^2R_0^2N\ln\frac{4N}{\beta}}\right\}\right)$$

   and with probability at least $1-\beta$

   $$f(y^N) - f(x^*) \leq \frac{2LC^2R_0^2\ln\frac{4N}{\beta}}{N+3}. \tag{80}$$

   That is, if ε is small enough to satisfy $\frac{LR_0^2}{\varepsilon}\ln\frac{LR_0^2}{\varepsilon\beta} \geq C_1\ln^2\frac{LR_0^2}{\varepsilon\beta}$ for some constant C_1, then due to (80) we have that after

   $$N = O\left(\frac{LR_0^2}{\varepsilon}\ln\frac{LR_0^2}{\varepsilon\beta}\right) \text{ iterations}$$

of clipped-SSTM we obtain such point y^N that with probability at least $1-\beta$ inequality $f(y^N)-f(x^*)\leq\varepsilon$ holds and the method requires

$$\begin{aligned}\sum_{k=0}^{N-1} m_k &= \sum_{k=0}^{N-1} O\left(\max\left\{1, \frac{\sigma^2(k+2)^2}{L^2R_0^2N\ln\frac{4N}{\beta}}\right\}\right)\\ &= O\left(\max\left\{N, \frac{\sigma^2N^2}{L^2R_0^2\ln\frac{4N}{\beta}}\right\}\right) = O\left(\max\left\{\frac{LR_0^2}{\varepsilon}, \frac{\sigma^2R_0^2}{\varepsilon^2}\right\}\ln\frac{LR_0^2}{\varepsilon\beta}\right)\end{aligned}$$

stochastic first-order oracle calls.

2. If $a_0N^{3/2}\sqrt{\ln\frac{4N}{\beta}}$ is bigger than $\hat{a}$ for some $a_0>0$, then we take $a=a_0N^{3/2}\sqrt{\ln\frac{4N}{\beta}}$ which implies that

$$m_k = \max\left\{1, \frac{1185\sigma^2(k+2)^2}{4a_0^2L^2N^2C^2R_0^2}, \frac{10368\sigma^2(k+2)^2}{4a_0^2L^2C^2R_0^2N^2\sqrt{\ln\frac{4N}{\beta}}}\right\} = O\left(\max\left\{1, \frac{\sigma^2(k+2)^2}{a_0^2L^2R_0^2N^2}\right\}\right)$$

and with probability at least $1-\beta$

$$f(y^N)-f(x^*)\leq\frac{2a_0LC^2R_0^2\sqrt{N\ln\frac{4N}{\beta}}}{N+3}. \tag{81}$$

That is, if ε is small enough to satisfy $\frac{a_0^3L^3R_0^6}{\varepsilon^3}\left(\ln\frac{LR_0^2}{\varepsilon\beta}\right)^{3/2}\geq C_2\ln^2\frac{LR_0^2}{\varepsilon\beta}$ for some constant C_2, then due to (81) we have that after

$$N = O\left(\frac{a_0^2L^2R_0^4}{\varepsilon^2}\ln\frac{a_0^2L^2R_0^4}{\varepsilon^2\beta}\right) = O\left(\frac{a_0^2L^2R_0^4}{\varepsilon^2}\ln\frac{a_0LR_0^2}{\varepsilon\beta}\right) \text{ iterations}$$

of clipped-SSTM we obtain such point y^N that with probability at least $1-\beta$ inequality $f(y^N)-f(x^*)\leq\varepsilon$ holds and the method requires

$$\begin{aligned}\sum_{k=0}^{N-1} m_k &= \sum_{k=0}^{N-1} O\left(\max\left\{1, \frac{\sigma^2(k+2)^2}{a_0^2L^2R_0^2N^2}\right\}\right)\\ &= O\left(\max\left\{N, \frac{\sigma^2N}{a_0^2L^2R_0^2}\right\}\right) = O\left(\max\left\{\frac{a_0^2L^2R_0^4}{\varepsilon^2}, \frac{\sigma^2R_0^2}{\varepsilon^2}\right\}\ln\frac{a_0LR_0^2}{\varepsilon\beta}\right)\end{aligned}$$

stochastic first-order oracle calls. Finally, if all assumptions on N, β and ε hold for $a_0=\frac{\sigma}{LR_0}$, then for all $k=0,1,\ldots,N-1$

$$m_k = O\left(\max\left\{1, \frac{\sigma^2(k+2)^2}{a_0^2L^2R_0^2N^2}\right\}\right) = O\left(\max\left\{1, \frac{(k+2)^2}{N^2}\right\}\right) = O(1),$$

i.e. one iteration of clipped-SSTM requires $O(1)$ oracle calls, and $f(y^N)-f(x^*)\leq\varepsilon$ with probability at least $1-\beta$ after

$$N = O\left(\frac{\sigma^2R_0^2}{\varepsilon^2}\ln\frac{\sigma R_0}{\varepsilon\beta}\right) \text{ iterations.}$$

F.3.5 Proof of Corollary F.3

Recall that

$$\begin{aligned}a' &= \max\left\{1, \frac{16\ln\frac{4N}{\beta}}{C}, 36\left(2\ln\frac{4N}{\beta}+\sqrt{4\ln^2\frac{4N}{\beta}+2\ln\frac{4N}{\beta}}\right)^2\right\},\\ a &= \max\left\{a', \frac{\sigma N^{3/2}}{LR_0}\sqrt{\ln\frac{4N}{\beta}}\right\}, \quad \alpha_{k+1}=\frac{k+2}{2aL},\\ m_k &= \max\left\{1, \frac{6000\sigma^2\alpha_{k+1}^2N\ln\frac{4N}{\beta}}{C^2R_0^2}, \frac{10368\sigma^2\alpha_{k+1}^2N}{C^2R_0^2}\right\}.\end{aligned}$$

Since $a\geq\frac{\sigma N^{3/2}}{LR_0}$ we have that $m_k=O(1)$. Next, there are two possible situations.

1. If $a = a'$, then we are in the settings of Theorem F.1. This means that `clipped-SSTM` achieves $f(y^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ after

$$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\} \ln \frac{LR_0^2}{\varepsilon\beta}\right) \text{ oracle calls.}$$

2. If $a = \frac{\sigma N^{3/2}}{LR_0}\sqrt{\ln \frac{4N}{\beta}}$, then we are in the settings of Corollary F.2 which implies that `clipped-SSTM` achieves $f(y^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ after

$$O\left(\frac{\sigma^2 R_0^2}{\varepsilon^2} \ln \frac{\sigma R_0}{\varepsilon\beta}\right) \text{ oracle calls.}$$

Finally, we combine these two cases and obtain that with $a = \max\left\{a', \frac{\sigma N^{3/2}}{LR_0}\sqrt{\ln \frac{4N}{\beta}}\right\}$ `clipped-SSTM` guarantees $f(y^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ after

$$O\left(\max\left\{\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\} \ln \frac{LR_0^2}{\varepsilon\beta}, \frac{\sigma^2 R_0^2}{\varepsilon^2} \ln \frac{\sigma R_0}{\varepsilon\beta}\right\}\right)$$
$$= O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\} \ln \frac{LR_0^2 + \sigma R_0}{\varepsilon\beta}\right)$$

iterations/oracle calls.

F.3.6 Proof of Theorem F.6

First of all, consider behavior of `clipped-SSTM` during the first run in `R-clipped-SSTM`. We notice that the proof of Theorem F.1 will be valid if we substitute R_0 everywhere by its upper bound R. From μ-strong convexity of f we have

$$R_0^2 = \|x^0 - x^*\|_2^2 \overset{(10)}{\leq} \frac{2}{\mu}\left(f(x^0) - f(x^*)\right),$$

therefore, one can choose $R = \sqrt{\frac{2}{\mu}\left(f(x^0) - f(x^*)\right)}$. It implies that after N_0 iterations of `clipped-SSTM` we have

$$f(y^{N_0}) - f(x^*) \leq \frac{2aC^2LR^2}{N_0(N_0+3)} = \frac{4aC^2L}{N_0^2\mu}(f(x^0) - f(x^*)).$$

with probability at least $1 - \frac{\beta}{\tau}$, hence with the same probability $f(y^{N_0}) - f(x^*) \leq \frac{1}{2}(f(x^0) - f(x^*))$ since $N_0 \geq C\sqrt{\frac{8aL}{\mu}}$. In other words, with probability at least $1 - \frac{\beta}{\tau}$

$$f(\hat{x}^1) - f(x^*) \leq \frac{1}{2}\left(f(x^0) - f(x^*)\right) = \frac{1}{4}\mu R^2.$$

Then, by induction one can show that for arbitrary $k \in \{0, 1, \ldots, \tau - 1\}$ the inequality

$$f(\hat{x}^{k+1}) - f(x^*) \leq \frac{1}{2}\left(f(\hat{x}^k) - f(x^*)\right)$$

holds with probability at least $1 - \frac{\beta}{\tau}$. Therefore, these inequalities hold simultaneously with probability at least $1 - \beta$. Using this we derive that inequality

$$f(\hat{x}^\tau) - f(x^*) \leq \frac{1}{2}\left(f(\hat{x}^{\tau-1}) - f(x^*)\right) \leq \frac{1}{2^2}\left(f(\hat{x}^{\tau-2}) - f(x^*)\right) \leq \ldots \leq \frac{1}{2^\tau}\left(f(x^0) - f(x^*)\right) = \frac{\mu R^2}{2^{\tau+1}}$$

holds with probability $\geq 1 - \beta$. That is, after $\tau = \left\lceil \log_2 \frac{\mu R^2}{2\varepsilon} \right\rceil$ restarts `R-clipped-SSTM` generates such a point $\hat{x}^\tau$ that $f(\hat{x}^\tau) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$. Moreover, if a equals

the maximum from (45) and $N_0 \leq C_1\sqrt{\frac{8aL}{\mu}}$ with some numerical constant $C_1 \geq C$, then $a \sim \left(\ln \frac{N_0\tau}{\beta}\right)^2$, the total number of iterations of `clipped-SSTM` equals

$$N_0\tau = O\left(\sqrt{\frac{L}{\mu}}\ln\left(\frac{\mu R^2}{\varepsilon}\right)\ln\left(\frac{L}{\mu\beta}\ln\frac{\mu R^2}{\varepsilon}\right)\right)$$

and the overall number of stochastic first-order oracle calls is

$$\begin{aligned}
\sum_{t=0}^{\tau-1}\sum_{k=0}^{N_0-1} m_k^t &= \sum_{t=0}^{\tau-1}\sum_{k=0}^{N_0-1} O\left(\max\left\{1, \frac{2^t\sigma^2\alpha_{k+1}^2 N_0 \ln\frac{4N_0\tau}{\beta}}{R^2}\right\}\right)\\
&= \sum_{t=0}^{\tau-1}\sum_{k=0}^{N_0-1} O\left(\max\left\{1, \frac{2^t\sigma^2(k+2)^2 N_0}{\ln^3\frac{4N_0\tau}{\beta} L^2R^2}\right\}\right)\\
&= O\left(\max\left\{N_0\tau, \frac{\sigma^2 2^\tau N_0^4}{\ln^3\frac{4N_0\tau}{\beta} L^2R^2}\right\}\right)\\
&= O\left(\max\left\{\sqrt{\frac{L}{\mu}}\ln\left(\frac{\mu R^2}{\varepsilon}\right), \frac{\sigma^2}{\mu\varepsilon}\right\}\ln\left(\frac{L}{\mu\beta}\ln\frac{\mu R^2}{\varepsilon}\right)\right).
\end{aligned}$$

F.3.7 Proof of Corollary F.7

Similarly to the proof of Theorem F.6 (see the previous subsection) we derive that under assumptions of the corollary after $\tau = \left\lceil \log_2 \frac{\mu R^2}{2\varepsilon}\right\rceil$ restarts `R-clipped-SSTM` generates such a point $\hat{x}^\tau$ that $f(\hat{x}^\tau) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$. Moreover, a and N_0 satisfy the following system of inequalities

$$a = \Theta\left(\frac{\sigma^4 \ln^2\frac{N_0\tau}{\beta}}{L\mu\varepsilon^2}\right), \quad N_0 = \Theta\left(\sqrt{\frac{aL}{\mu}}\right) \tag{82}$$

which is consistent and implies that

$$a = \Theta\left(\frac{\sigma^4}{L\mu\varepsilon}\ln^2\left(\frac{\sigma^2}{\mu\varepsilon\beta}\ln\frac{\mu R^2}{\varepsilon}\right)\right), \quad N_0 = \Theta\left(\frac{\sigma^2}{\mu\varepsilon}\ln\left(\frac{\sigma^2}{\mu\varepsilon\beta}\ln\frac{\mu R^2}{\varepsilon}\right)\right). \tag{83}$$

Then, for all $k = 0, 1, \ldots, N_0 - 1$ and $t = 0, 1, \ldots, \tau - 1$ batchsizes satisfy

$$\begin{aligned}
m_k^t \leq m_{N_0-1}^{\tau-1} &= O\left(\max\left\{1, \frac{2^\tau\sigma^2\alpha_{N_0}^2 N_0 \ln\frac{N_0\tau}{\beta}}{R^2}\right\}\right)\\
&= O\left(\max\left\{1, \frac{\mu R^2\sigma^2 N_0^3 \ln\frac{N_0\tau}{\beta}}{a^2L^2\varepsilon R^2}\right\}\right) \overset{(82),(83)}{=} O(1),
\end{aligned}$$

i.e. the algorithm requires $O(1)$ oracle calls per iteration. Finally, the total number of iterations is

$$N_0\tau = O\left(\frac{\sigma^2}{\mu\varepsilon}\ln\left(\frac{\mu R^2}{\varepsilon}\right)\ln\left(\frac{\sigma^2}{\mu\varepsilon\beta}\ln\frac{\mu R^2}{\varepsilon}\right)\right).$$

G SGD with Clipping: Exact Formulations and Missing Proofs

In this section we provide exact formulations of all the results that we have for clipped-SGD and R-clipped-SGD together with the full proofs.

G.1 Convex Case

We start with the case when $f(x)$ is convex and L-smooth and, as before, we assume that at each point $x \in \mathbb{R}^n$ function f is accessible only via stochastic gradients $\nabla f(x, \xi)$ such that (2) holds. Next theorem summarizes the main convergence result for clipped-SGD in this case.

Theorem G.1. Assume that function f is convex and L-smooth. Then for all $\beta \in (0, 1)$ and $N \geq 1$ such that

$$\ln \frac{4N}{\beta} \geq 2 \tag{84}$$

we have that after N iterations of clipped-SGD with

$$\lambda = 2LCR_0, \quad m_k = m = \max\left\{1, \frac{27N\sigma^2}{2(CR_0)^2L^2 \ln \frac{4N}{\beta}}\right\}, \tag{85}$$

where $R_0 = \|x^0 - x^*\|_2$ and stepsize

$$\gamma = \frac{1}{80L \ln \frac{4N}{\beta}}, \tag{86}$$

that with probability at least $1 - \beta$

$$f(\bar{x}^N) - f(x^*) \leq \frac{80LC^2R_0^2 \ln \frac{4N}{\beta}}{N}, \tag{87}$$

where $\bar{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k$ and

$$C = \sqrt{2}. \tag{88}$$

In other words, the method achieves $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ after $O\left(\frac{LR_0^2}{\varepsilon} \ln \frac{LR_0^2}{\varepsilon\beta}\right)$ iterations and requires

$$O\left(\max\left\{\frac{LR_0^2}{\varepsilon}, \frac{\sigma^2R_0^2}{\varepsilon^2}\right\} \ln \frac{LR_0^2}{\varepsilon\beta}\right) \text{ oracle calls.} \tag{89}$$

To the best of our knowledge, it is the first result for clipped-SGD establishing non-trivial complexity guarantees for the convergence with high probability. One can find the full proof in Section G.3.1.

G.2 Strongly Convex Case

Next, we consider the situation when f is additionally μ-strongly convex and propose a restarted version of clipped-SGD (R-clipped-SGD), see Algorithm 4. For this method we prove the following result.

Algorithm 4 Restarted Clipped Stochastic Gradient Descent (R-clipped-SGD)

Input: starting point x^0, number of iterations N_0 of clipped-SGD, number τ of clipped-SGD runs, batchsizes $m^0, m^1, \ldots, m^\tau$

1: Set $\hat{x}^0 = x^0$, stepsize $\gamma > 0$
2: **for** $t = 0, 1, \ldots, \tau - 1$ **do**
3: Run clipped-SGD (Algorithm 2) for N_0 iterations with constant batchsizes m^t, stepsize γ and starting point $\hat{x}^t$. Define the output of clipped-SGD by $\hat{x}^{t+1}$.
4: **end for**

Output: $\hat{x}^\tau$

Theorem G.2. Assume that f is μ-strongly convex and L-smooth. If we choose $\beta \in (0,1)$, τ and $N_0 \geq 1$ such that

$$\ln \frac{4N_0\tau}{\beta} \geq 2, \quad \frac{N_0}{\ln \frac{4N_0\tau}{\beta}} \geq \frac{320C^2L}{\mu}, \tag{90}$$

and

$$m^t = \max\left\{1, \frac{27 \cdot 2^t N_0 \sigma^2}{2(CR)^2 L^2 \ln \frac{4N_0\tau}{\beta}}\right\}, \tag{91}$$

where $R = \sqrt{\frac{2(f(x^0)-f(x^*))}{\mu}}$ and $C = \sqrt{2}$, then we have that after τ runs of `clipped-SGD` in `R-clipped-SGD` the inequality

$$f(\hat{x}^\tau) - f(x^*) \leq 2^{-\tau}\left(f(x^0) - f(x^*)\right) \tag{92}$$

holds with probability at least $1-\beta$. That is, if we choose $\frac{N_0}{\ln \frac{4N_0\tau}{\beta}} \leq \frac{C_1 L}{\mu}$ with some numerical constant $C_1 \geq 320C^2$, then the method achieves $f(\hat{x}^\tau) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$ after

$$O\left(\frac{L}{\mu}\ln\left(\frac{\mu R^2}{\varepsilon}\right)\ln\left(\frac{L}{\mu\beta}\ln\frac{\mu R^2}{\varepsilon}\right)\right) \text{ iterations (in total)} \tag{93}$$

of `clipped-SGD` and requires

$$O\left(\max\left\{\frac{L}{\mu}\ln\frac{\mu R^2}{\varepsilon}, \frac{\sigma^2}{\mu\varepsilon}\right\}\ln\left(\frac{L}{\mu\beta}\ln\frac{\mu R^2}{\varepsilon}\right)\right) \text{ oracle calls.} \tag{94}$$

This theorem implies that `R-clipped-SGD` has the same complexity as the restarted version of `RSMD` from [47] up to the difference in logarithmical factors. We notice that the main difference between our result and one from [47] is that we do not need to assume that the optimization problem is considered on the bounded set.

However, in order to get (94) `R-clipped-SGD` requires to know strong convexity parameter μ. In order to remove this drawback we analyse `clipped-SGD` for the strongly convex case and get the following result.

Theorem G.3. Assume that function f is μ-strongly convex and L-smooth. Then for all $\beta \in (0,1)$ and $N \geq 1$ such that

$$\ln \frac{4N}{\beta} \geq 2 \tag{95}$$

we have that after N iterations of `clipped-SGD` with

$$\lambda_l = 4\sqrt{L(1-\gamma\mu)^l r_0}, \quad m_k = \max\left\{1, \frac{27N\sigma^2}{16Lr_0(1-\gamma\mu)^k \ln\frac{4N}{\beta}}\right\}, \tag{96}$$

where $r_0 = f(x^0) - f(x^*)$ and stepsize

$$\gamma = \frac{1}{81L\ln\frac{4N}{\beta}}, \tag{97}$$

that with probability at least $1-\beta$

$$f(x^N) - f(x^*) \leq 2(1-\gamma\mu)^N(f(x^0) - f(x^*)). \tag{98}$$

In other words, the method achieves $f(x^N) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$ after $O\left(\frac{L}{\mu}\ln\left(\frac{r_0}{\varepsilon}\right)\ln\left(\frac{L}{\mu\beta}\ln\frac{r_0}{\varepsilon}\right)\right)$ iterations and requires

$$O\left(\max\left\{\frac{L}{\mu}, \frac{\sigma^2}{\mu\varepsilon}\cdot\frac{L}{\mu}\right\}\ln\left(\frac{r_0}{\varepsilon}\right)\ln\left(\frac{L}{\mu\beta}\ln\frac{r_0}{\varepsilon}\right)\right) \text{ oracle calls.} \tag{99}$$

Unfortunately, our approach leads to worse complexity bound than we have for `R-clipped-SGD`: in the second term of the maximum in (99) we get an extra factor L/μ that can be large. Nevertheless, to

the best of our knowledge it is the first non-trivial complexity result for `clipped-SGD` that guarantees convergence with high probability. One can find the full proof of Theorem G.3 in Section G.3.3.

G.3 Proofs

G.3.1 Proof of Theorem G.1

Since $f(x)$ is convex and L-smooth, we get the following inequality:

$$\begin{aligned}
\|x^{k+1}-x^*\|_2^2 &= \|x^k-\gamma\widetilde{\nabla}f(x^k,\boldsymbol{\xi}^k)-x^*\|_2^2 = \|x^k-x^*\|_2^2+\gamma^2\|\widetilde{\nabla}f(x^k,\boldsymbol{\xi}^k)\|_2^2-2\gamma\left\langle x^k-x^*,g^k\right\rangle \\
&= \|x^k-x^*\|_2^2+\gamma^2\|\nabla f(x^k)+\theta_k\|_2^2-2\gamma\left\langle x^k-x^*,\nabla f(x^k)+\theta_k\right\rangle \\
&\overset{(12)}{\leq} \|x^k-x^*\|_2^2+2\gamma^2\|\nabla f(x^k)\|_2^2+2\gamma^2\|\theta_k\|_2^2-2\gamma\left\langle x^k-x^*,\nabla f(x^k)+\theta_k\right\rangle \\
&\overset{(8)}{\leq} \|x^k-x^*\|_2^2+4\gamma^2L\left(f(x^k)-f(x^*)\right)+2\gamma^2\|\theta_k\|_2^2-2\gamma\left\langle x^k-x^*,\nabla f(x^k)+\theta_k\right\rangle \\
&\leq \|x^k-x^*\|_2^2+(4\gamma^2L-2\gamma)\left(f(x^k)-f(x^*)\right)+2\gamma^2\|\theta_k\|_2^2-2\gamma\left\langle x^k-x^*,\theta_k\right\rangle,
\end{aligned}$$

where $\theta_k=\widetilde{\nabla}f(x^k,\boldsymbol{\xi}^k)-\nabla f(x^k)$ and the last inequality follows from the convexity of f. Using notation $R_k\overset{\text{def}}{=}\|x^k-x^*\|_2$ we derive that for all $k\geq 0$

$$R_{k+1}^2\leq R_k^2+(4\gamma^2L-2\gamma)\left(f(x^k)-f(x^*)\right)+2\gamma^2\|\theta_k\|_2^2-2\gamma\left\langle x^k-x^*,\theta_k\right\rangle.$$

Let us define $A=\left(2\gamma-4\gamma^2L\right)$, then

$$A\left(f(x^k)-f(x^*)\right)\leq R_k^2-R_{k+1}^2+2\gamma^2\|\theta_k\|_2^2-2\gamma\left\langle x^k-x^*,\theta_k\right\rangle.$$

Summing up these inequalities for $k=0,\ldots,N-1$ we obtain

$$\begin{aligned}
\frac{A}{N}\sum_{k=0}^{N-1}\left[f(x^k)-f(x^*)\right] &\leq \frac{1}{N}\sum_{k=0}^{N-1}\left(R_k^2-R_{k+1}^2\right)+\frac{2\gamma^2}{N}\sum_{k=0}^{N-1}\|\theta_k\|_2^2-\frac{2\gamma^2}{N}\sum_{k=0}^{N-1}\left\langle x^k-x^*,\theta_k\right\rangle \\
&= \frac{1}{N}\left(R_0^2-R_N^2\right)+\frac{2\gamma^2}{N}\sum_{k=0}^{N-1}\|\theta_k\|_2^2-\frac{2\gamma^2}{N}\sum_{k=0}^{N-1}\left\langle x^k-x^*,\theta_k\right\rangle.
\end{aligned}$$

Noticing that for $\bar{x}^N=\frac{1}{N}\sum\limits_{k=0}^{N-1}x^k$ Jensen's inequality gives $f(\bar{x}^N)=f\left(\frac{1}{N}\sum\limits_{k=0}^{N-1}x^k\right)\leq\frac{1}{N}\sum\limits_{k=0}^{N-1}f(x^k)$ we have

$$AN\left(f(\bar{x}^N)-f(x^*)\right)\leq R_0^2-R_N^2+2\gamma^2\sum_{k=0}^{N-1}\|\theta_k\|_2^2-2\gamma\sum_{k=0}^{N-1}\left\langle x^k-x^*,\theta_k\right\rangle. \tag{100}$$

Taking into account that $f(\bar{x}^N)-f(x^*)\geq 0$ and changing the indices we get that for all $k\geq 0$

$$R_k^2\leq R_0^2+2\gamma^2\sum_{l=0}^{k-1}\|\theta_l\|_2^2-2\gamma\sum_{l=0}^{k-1}\left\langle x^l-x^*,\theta_k\right\rangle. \tag{101}$$

The remaining part of the proof is based on the analysis of inequality (101). In particular, via induction we prove that for all $k=0,1,\ldots,N$ with probability at least $1-\frac{k\beta}{N}$ the following statement holds: inequalities

$$R_t^2\overset{(101)}{\leq}R_0^2+2\gamma^2\sum_{l=0}^{t-1}\|\theta_k\|_2^2-2\gamma\sum_{l=0}^{t-1}\left\langle x^k-x^*,\theta_k\right\rangle\leq C^2R_0^2 \tag{102}$$

hold for $t=0,1,\ldots,k$ simultaneously where C is defined in (88). Let us define the probability event when this statement holds as E_k. Then, our goal is to show that $\mathbb{P}\{E_k\}\geq 1-\frac{k\beta}{N}$ for all $k=0,1,\ldots,N$. For $t=0$ inequality (102) holds with probability 1 since $C\geq 1$. Next, assume

that for some $k = T-1 \leq N-1$ we have $\mathbb{P}\{E_k\} = \mathbb{P}\{E_{T-1}\} \geq 1 - \frac{(T-1)\beta}{N}$. Let us prove that $\mathbb{P}\{E_T\} \geq 1 - \frac{T\beta}{N}$. First of all, probability event E_{T-1} implies that

$$f(\bar{x}^N) - f(x^*) \overset{(100)}{\leq} \frac{1}{AN}\left(R_0^2 + 2\gamma^2 \sum_{k=0}^{N-1} \|\theta_k\|_2^2 - 2\gamma \sum_{k=0}^{N-1} \left\langle x^k - x^*, \theta_k \right\rangle\right) \overset{(102)}{\leq} \frac{C^2 R_0^2}{AN}$$

hold for $t = 0, 1, \ldots, T-1$. Since f is L-smooth, we have that probability event E_{T-1} implies

$$\left\|\nabla f(x^t)\right\|_2 \leq L\|x^t - x^*\|_2 \leq LCR_0 = \frac{\lambda}{2} \tag{103}$$

for $t = 0, \ldots, T-1$, where the clipping level is defined as

$$\lambda = 2LCR_0. \tag{104}$$

Having inequalities (103) in hand we show in the rest of the proof that (102) holds for $t = T$ with big enough probability. First of all, we introduce new random variables:

$$\eta_l = \begin{cases} x^* - z^l, & \text{if } \|x^* - z^l\|_2 \leq CR_0, \\ 0, & \text{otherwise,} \end{cases} \tag{105}$$

for $l = 0, 1, \ldots T-1$. Note that these random variables are bounded with probability 1, i.e. with probability 1 we have

$$\|\eta_l\|_2 \leq CR_0. \tag{106}$$

Secondly, we use the introduced notation and get that E_{T-1} implies

$$R_T^2 \overset{(101),(102),(103),(105)}{\leq} R_0^2 + 2\gamma \sum_{l=0}^{T-1} \langle \theta_l, \eta_l \rangle + 2\gamma^2 \sum_{l=0}^{T-1} \|\theta_{l+1}\|_2^2.$$

Finally, we do some preliminaries in order to apply Bernstein's inequality (see Lemma D.1) and obtain that E_{T-1} implies

$$\begin{aligned} R_T^2 \overset{(12)}{\leq} R_0^2 &+ \underbrace{2\gamma \sum_{l=0}^{T-1} \langle \theta_l^u, \eta_l \rangle}_{①} + \underbrace{2\gamma \sum_{l=0}^{T-1} \left\langle \theta_l^b, \eta_l \right\rangle}_{②} + \underbrace{4\gamma^2 \sum_{l=0}^{T-1} \left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right)}_{③} \\ &\underbrace{+ 4\gamma^2 \sum_{l=0}^{T-1} \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]}_{④} + \underbrace{4\gamma^2 \sum_{l=0}^{T-1} \|\theta_l^b\|_2^2}_{⑤} \end{aligned} \tag{107}$$

where we introduce new notations:

$$\theta_l^u \overset{\text{def}}{=} \widetilde{\nabla} f(x^l, \boldsymbol{\xi}^l) - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\widetilde{\nabla} f(x^l, \boldsymbol{\xi}^l)\right], \quad \theta_l^b \overset{\text{def}}{=} \mathbb{E}_{\boldsymbol{\xi}^l}\left[\widetilde{\nabla} f(x^l, \boldsymbol{\xi}^l)\right] - \nabla f(x^l), \tag{108}$$

$$\theta_l = \theta_l^u + \theta_l^b.$$

It remains to provide tight upper bounds for ①, ②, ③, ④ and ⑤, i.e. in the remaining part of the proof we show that ① + ② + ③ + ④ + ⑤ $\leq \delta C^2 R_0^2$ for some $\delta < 1$.

Upper bound for ①. First of all, since $\mathbb{E}_{\boldsymbol{\xi}^l}[\theta_l^u] = 0$ summands in ① are conditionally unbiased:

$$\mathbb{E}_{\boldsymbol{\xi}^l}\left[2\gamma \left\langle \theta_l^u, \eta_l \right\rangle\right] = 0.$$

Secondly, these summands are bounded with probability 1:

$$\left|2\gamma \left\langle \theta_l^u, \eta_l \right\rangle\right| \leq 2\gamma \|\theta_l^u\|_2 \|\eta_l\|_2 \overset{(38),(106)}{\leq} 4\gamma\lambda CR_0 \overset{(104)}{=} 8\gamma (CR_0)^2 L.$$

Finally, one can bound conditional variances $\sigma_l^2 \overset{\text{def}}{=} \mathbb{E}_{\boldsymbol{\xi}^l}\left[4\gamma^2 \left\langle \theta_l^u, \eta_l \right\rangle^2\right]$ in the following way:

$$\sigma_l^2 \leq \mathbb{E}_{\boldsymbol{\xi}^l}\left[4\gamma^2 \|\theta_l^u\|_2^2 \|\eta_l\|_2^2\right] \overset{(106)}{\leq} 4\gamma^2 (CR_0)^2 \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right].$$

In other words, sequence $\{2\gamma \langle \theta_l^u, \eta_l \rangle\}_{l \geq 0}$ is a bounded martingale difference sequence with bounded conditional variances $\{\sigma_l^2\}_{l \geq 0}$. Therefore, we can apply Bernstein's inequality, i.e. we apply Lemma D.1 with $X_l = 2\gamma \langle \theta_l^u, \eta_l \rangle$, $c = 8\gamma(CR_0)^2 L$ and $F = \frac{c^2 \ln \frac{4N}{\beta}}{6}$ and get that for all $b > 0$

$$\mathbb{P}\left\{\left|\sum_{l=0}^{T-1} X_l\right| > b \text{ and } \sum_{l=0}^{T-1} \sigma_l^2 \leq F\right\} \leq 2\exp\left(-\frac{b^2}{2F + {}^{2cb}/_3}\right)$$

or, equivalently, with probability at least $1 - 2\exp\left(-\frac{b^2}{2F+{}^{2cb}/_3}\right)$

$$\text{either } \sum_{l=0}^{T-1} \sigma_l^2 > F \quad \text{or} \quad \underbrace{\left|\sum_{l=0}^{T-1} X_l\right|}_{|①|} \leq b.$$

The choice of F will be clarified further, let us now choose b in such a way that $2\exp\left(-\frac{b^2}{2F+{}^{2cb}/_3}\right) = \frac{\beta}{2N}$. This implies that b is the positive root of the quadratic equation

$$b^2 - \frac{2c\ln\frac{4N}{\beta}}{3} b - 2F\ln\frac{4N}{\beta} = 0,$$

hence

$$\begin{aligned} b &= \frac{c\ln\frac{4N}{\beta}}{3} + \sqrt{\frac{c^2\ln^2\frac{4N}{\beta}}{9} + 2F\ln\frac{4N}{\beta}} = \frac{c\ln\frac{4N}{\beta}}{3} + \sqrt{\frac{4c^2\ln^2\frac{4N}{\beta}}{9}} \\ &= c\ln\frac{4N}{\beta} = 8\gamma(CR_0)^2 L\ln\frac{4N}{\beta}. \end{aligned}$$

That is, with probability at least $1 - \frac{\beta}{2N}$

$$\underbrace{\text{either } \sum_{l=0}^{T-1} \sigma_l^2 > F \quad \text{or} \quad |①| \leq 8\gamma(CR_0)^2 L\ln\frac{4N}{\beta}}_{\text{probability event } E_①}.$$

Next, we notice that probability event E_{T-1} implies that

$$\begin{aligned} \sum_{l=0}^{T-1} \sigma_l^2 &\leq 4\gamma^2(CR_0)^2 \sum_{l=0}^{T-1} \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right] \overset{(41)}{\leq} 72\gamma^2(CR_0)^2\sigma^2\frac{T}{m} \\ &\overset{(85)}{\leq} 72\gamma^2(CR_0)^2\sigma^2\frac{2T(CR_0)^2L^2\ln\frac{4N}{\beta}}{27N\sigma^2} \\ &\overset{T\leq N}{\leq} \frac{16}{3}\gamma^2(CR_0)^4L^2\ln\frac{4N}{\beta} \leq \frac{c^2\ln\frac{4N}{\beta}}{6} = F, \end{aligned}$$

where the last inequality follows from $c = 8\gamma(CR_0)^2 L$ and simple arithmetic.

Upper bound for ②. First of all, we notice that probability event E_{T-1} implies

$$2\gamma\left\langle \theta_l^b, \eta_l \right\rangle \leq 2\gamma\left\|\theta_l^b\right\|_2 \|\eta_l\|_2 \overset{(39),(106)}{\leq} 2\gamma\frac{4\sigma^2}{m\lambda}CR_0 \overset{(104)}{=} \frac{4\gamma\sigma^2}{Lm}.$$

This implies that

$$② = 2\gamma\sum_{l=0}^{T-1}\left\langle \theta_l^b, \eta_l \right\rangle \overset{T\leq N}{\leq} \frac{4\gamma N\sigma^2}{mL}.$$

Upper bound for ③. We derive the upper bound for ③ using the same technique as for ①. First of all, we notice that the summands in ③ are conditionally independent:

$$\mathbb{E}_{\boldsymbol{\xi}^l}\left[4\gamma^2\left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right)\right] = 0.$$

Secondly, the summands are bounded with probability 1:

$$\begin{aligned}\left|4\gamma^2\left(\|\theta_l^u\|_2^2-\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right)\right| &\leq 4\gamma^2\left(\|\theta_l^u\|_2^2+\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right)\overset{(38)}{\leq}4\gamma^2\left(4\lambda^2+4\lambda^2\right)\\ &\overset{(104)}{=} 128\gamma^2(CR_0)^2L^2\overset{\text{def}}{=}c_1. \end{aligned}\tag{109}$$

Finally, one can bound conditional variances $\hat{\sigma}_l^2\overset{\text{def}}{=}\mathbb{E}_{\boldsymbol{\xi}^l}\left[\left|4\gamma^2\left(\|\theta_l^u\|_2^2-\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right)\right|^2\right]$ in the following way:

$$\begin{aligned}\hat{\sigma}_l^2 &\overset{(109)}{\leq} c_1\mathbb{E}_{\boldsymbol{\xi}^l}\left[\left|4\gamma^2\left(\|\theta_l^u\|_2^2-\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right)\right|\right]\\ &\leq 4\gamma^2c_1\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2+\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right]=8\gamma^2c_1\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right].\end{aligned}\tag{110}$$

In other words, sequence $\left\{4\gamma^2\left(\|\theta_l^u\|_2^2-\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right)\right\}_{l\geq 0}$ is a bounded martingale difference sequence with bounded conditional variances $\{\hat{\sigma}_l^2\}_{l\geq 0}$. Therefore, we can apply Bernstein's inequality, i.e. we apply Lemma D.1 with $X_l=\hat{X}_l=4\gamma^2\left(\|\theta_l^u\|_2^2-\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right)$, $c=c_1=128\gamma^2(CR_0)^2L^2$ and $F=F_1=\frac{c_1^2\ln\frac{4N}{\beta}}{6}$ and get that for all $b>0$

$$\mathbb{P}\left\{\left|\sum_{l=0}^{T-1}\hat{X}_l\right|>b \text{ and } \sum_{l=0}^{T-1}\hat{\sigma}_l^2\leq F_1\right\}\leq 2\exp\left(-\frac{b^2}{2F_1+{}^{2c_1b}/_3}\right)$$

or, equivalently, with probability at least $1-2\exp\left(-\frac{b^2}{2F_1+{}^{2c_1b}/_3}\right)$

$$\text{either } \sum_{l=0}^{T-1}\hat{\sigma}_l^2>F_1 \quad \text{or} \quad \underbrace{\left|\sum_{l=0}^{T-1}\hat{X}_l\right|}_{|③|}\leq b.$$

As in our derivations of the upper bound for ① we choose such b that $2\exp\left(-\frac{b^2}{2F_1+{}^{2c_1b}/_3}\right)=\frac{\beta}{2N}$, i.e.

$$b = \frac{c_1\ln\frac{4N}{\beta}}{3}+\sqrt{\frac{c_1^2\ln^2\frac{4N}{\beta}}{9}+2F_1\ln\frac{4N}{\beta}}=c_1\ln\frac{4N}{\beta}=128\gamma^2(CR_0)^2L^2\ln\frac{4N}{\beta}.$$

That is, with probability at least $1-\frac{\beta}{2N}$

$$\underbrace{\text{either } \sum_{l=0}^{T-1}\hat{\sigma}_l^2>F_1 \quad \text{or} \quad |③|\leq 128\gamma^2(CR_0)^2L^2\ln\frac{4N}{\beta}}_{\text{probability event } E_③}.$$

Next, we notice that probability event E_{T-1} implies that

$$\begin{aligned}\sum_{l=0}^{T-1}\hat{\sigma}_l^2 &\overset{(110)}{\leq} 8\gamma^2c_1\sum_{l=0}^{T-1}\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\overset{(41)}{\leq}144\gamma^2c_1\sigma^2\frac{T}{m}\\ &\overset{(85)}{\leq} \frac{32}{3}\gamma^2c_1(CR_0)^2L^2\frac{T}{N}\ln\frac{4N}{\beta}\\ &\overset{T\leq N}{\leq} \frac{c_1^2\ln\frac{4N}{\beta}}{6}\leq F_1.\end{aligned}$$

Upper bound for ④. The probability event E_{T-1} implies

$$④ = 4\gamma^2\sum_{l=0}^{T-1}\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\overset{(41)}{\leq}72\gamma^2\sigma^2\sum_{l=0}^{T-1}\frac{1}{m}\overset{T\leq N}{\leq}\frac{72\gamma^2N\sigma^2}{m}.$$

Upper bound for ⑤. Again, we use corollaries of probability event E_{T-1}:

$$⑤ \quad = \quad 4\gamma^2 \sum_{l=0}^{T-1} \|\theta_l^b\|_2^2 \overset{(39)}{\leq} 64\gamma^2\sigma^4 \frac{T}{m^2\lambda^2} \overset{(104)}{=} \frac{64\gamma^2\sigma^4}{4(CR_0)^2L^2} \cdot \frac{T}{m^2} \overset{T\leq N}{\leq} \frac{16\gamma^2 N\sigma^4}{(CR_0)^2L^2m^2}.$$

Now we summarize all bound that we have: probability event E_{T-1} implies

$$\begin{aligned} R_T^2 &\overset{(101)}{\leq} R_0^2 + 2\gamma^2 \sum_{l=0}^{T-1} \|\theta_l\|_2^2 - 2\gamma \sum_{l=0}^{T-1} \langle x^l - x^*, \theta_l \rangle \\ &\overset{(107)}{\leq} R_0^2 + ① + ② + ③ + ④ + ⑤, \\ ② &\leq \frac{4\gamma N\sigma^2}{mL}, \quad ④ \leq \frac{72\gamma^2 N\sigma^2}{m}, \quad ⑤ \leq \frac{16\gamma^2 N\sigma^4}{(CR_0)^2L^2m^2}, \\ \sum_{l=0}^{T-1} \sigma_l^2 &\leq F, \quad \sum_{l=0}^{T-1} \hat{\sigma}_l^2 \leq F_1 \end{aligned}$$

and

$$\mathbb{P}\{E_{T-1}\} \geq 1 - \frac{(T-1)\beta}{N}, \quad \mathbb{P}\{E_①\} \geq 1 - \frac{\beta}{2N}, \quad \mathbb{P}\{E_③\} \geq 1 - \frac{\beta}{2N},$$

where

$$\begin{aligned} E_① &= \left\{ \text{either } \sum_{l=0}^{T-1} \sigma_l^2 > F \quad \text{or} \quad |①| \leq 8\gamma(CR_0)^2 L \ln\frac{4N}{\beta} \right\}, \\ E_③ &= \left\{ \text{either } \sum_{l=0}^{T-1} \hat{\sigma}_l^2 > F_1 \quad \text{or} \quad |③| \leq 128\gamma^2(CR_0)^2 L^2 \ln\frac{4N}{\beta} \right\}. \end{aligned}$$

Taking into account these inequalities and our assumptions on m and γ (see (85) and (86)) we get that probability event $E_{T-1} \cap E_① \cap E_③$ implies

$$\begin{aligned} R_T^2 &\overset{(101)}{\leq} R_0^2 + 2\gamma^2 \sum_{l=0}^{T-1} \|\theta_l\|_2^2 - 2\gamma \sum_{l=0}^{T-1} \langle x^l - x^*, \theta_l \rangle \\ &\leq R_0^2 + \left(\frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10}\right) C^2 R_0^2 \leq \left(1 + \frac{1}{2}C^2\right) R_0^2 \overset{(88)}{\leq} C^2 R_0^2. \end{aligned} \tag{111}$$

Moreover, using union bound we derive

$$\mathbb{P}\left\{E_{T-1} \cap E_① \cap E_③\right\} = 1 - \mathbb{P}\left\{\overline{E}_{T-1} \cup \overline{E}_① \cup \overline{E}_③\right\} \geq 1 - \frac{T\beta}{N}. \tag{112}$$

That is, by definition of E_T and E_{T-1} we have proved that

$$\mathbb{P}\{E_T\} \overset{(111)}{\geq} \mathbb{P}\left\{E_{T-1} \cap E_① \cap E_③\right\} \overset{(112)}{\geq} 1 - \frac{T\beta}{N},$$

which implies that for all $k = 0, 1, \ldots, N$ we have $\mathbb{P}\{E_k\} \geq 1 - \frac{k\beta}{N}$. Then, for $k = N$ we have that with probability at least $1 - \beta$

$$AN f(\bar{x}^N) - f(x^*) \overset{(100)}{\leq} R_0^2 + 2\gamma^2 \sum_{k=0}^{N-1} \|\theta_k\|_2^2 - 2\gamma \sum_{k=0}^{N-1} \langle x^k - x^*, \theta_k \rangle \overset{(102)}{\leq} C^2 R_0^2.$$

Since $A = 2\gamma\left(1 - 2\gamma L\right)$ and $1 - \gamma L \geq \frac{1}{2}$ we get that with probability at least $1 - \beta$

$$f(\bar{x}^N) - f(x^*) \quad \leq \quad \frac{C^2 R_0^2}{AN} \leq \frac{C^2 R_0^2}{\gamma N} \overset{(86)}{\leq} \frac{80 C^2 R_0^2 L \ln\frac{4N}{\beta}}{N}.$$

In other words, clipped-SGD achieves $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$ after $O\left(\frac{LR_0^2}{\varepsilon}\ln\frac{LR_0^2}{\varepsilon\beta}\right)$ iterations and requires

$$\begin{aligned}\sum_{k=0}^{N-1} m_k &\overset{(85)}{=} \sum_{k=0}^{N-1} O\left(\max\left\{1, \frac{N\sigma^2}{C^2R_0^2L^2\ln\frac{N}{\beta}}\right\}\right) = O\left(\max\left\{N, \frac{N^2\sigma^2}{C^2R_0^2L^2\ln\frac{N}{\beta}}\right\}\right)\\ &= O\left(\max\left\{\frac{LR_0^2}{\varepsilon}, \frac{\sigma^2R_0^2}{\varepsilon^2}\right\}\ln\frac{LR_0^2}{\varepsilon\beta}\right)\end{aligned}$$

oracle calls.

G.3.2 Proof of Theorem G.2

First of all, consider behavior of clipped-SGD during the first run in R-clipped-SGD. We notice that the proof of Theorem G.1 will be valid if we substitute R_0 everywhere by its upper bound R. From μ-strong convexity of f we have

$$R_0^2 = \|x^0 - x^*\|_2^2 \overset{(10)}{\leq} \frac{2}{\mu}\left(f(x^0) - f(x^*)\right),$$

therefore, one can choose $R = \sqrt{\frac{2}{\mu}\left(f(x^0) - f(x^*)\right)}$. It implies that after N_0 iterations of clipped-SGD we have

$$f(\bar{x}^{N_0}) - f(x^*) \leq \frac{80LC^2R^2\ln\frac{4N_0\tau}{\beta}}{N_0} = \frac{160LC^2R^2\ln\frac{4N_0\tau}{\beta}}{N_0\mu}(f(x^0) - f(x^*)).$$

with probability at least $1-\frac{\beta}{\tau}$, hence with the same probability $f(\bar{x}^{N_0}) - f(x^*) \leq \frac{1}{2}(f(x^0) - f(x^*))$ since $\frac{N_0}{\ln\frac{4N_0\tau}{\beta}} \geq \frac{320C^2L}{\mu}$. In other words, with probability at least $1-\frac{\beta}{\tau}$

$$f(\hat{x}^1) - f(x^*) \leq \frac{1}{2}\left(f(x^0) - f(x^*)\right) = \frac{1}{4}\mu R^2.$$

Then, by induction one can show that for arbitrary $k \in \{0, 1, \ldots, \tau-1\}$ the inequality

$$f(\hat{x}^{k+1}) - f(x^*) \leq \frac{1}{2}\left(f(\hat{x}^k) - f(x^*)\right)$$

holds with probability at least $1-\frac{\beta}{\tau}$. Therefore, these inequalities hold simultaneously with probability at least $1-\beta$. Using this we derive that inequality

$$\begin{aligned}f(\hat{x}^\tau) - f(x^*) &\leq \frac{1}{2}\left(f(\hat{x}^{\tau-1}) - f(x^*)\right) \leq \frac{1}{2^2}\left(f(\hat{x}^{\tau-2}) - f(x^*)\right) \leq \ldots \leq \frac{1}{2^\tau}\left(f(x^0) - f(x^*)\right)\\ &= \frac{\mu R^2}{2^{\tau+1}}\end{aligned}$$

holds with probability $\geq 1-\beta$. That is, after $\tau = \left\lceil \log_2 \frac{\mu R^2}{2\varepsilon}\right\rceil$ restarts R-clipped-SGD generates such point $\hat{x}^\tau$ that $f(\hat{x}^\tau) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$. Moreover, if $\frac{N_0}{\ln\frac{4N_0\tau}{\beta}} \leq \frac{C_1L}{\mu}$ with some numerical constant $C_1 \geq 320C^2$, then the total number of iterations of clipped-SGD equals

$$N_0\tau = O\left(\frac{L}{\mu}\ln\left(\frac{\mu R^2}{\varepsilon}\right)\ln\left(\frac{L}{\mu\beta}\ln\frac{\mu R^2}{\varepsilon}\right)\right)$$

and the overall number of stochastic first-order oracle calls is

$$\begin{aligned}\sum_{t=0}^{\tau-1} N_0 m^t &= \sum_{t=0}^{\tau-1} O\left(\max\left\{N_0, \frac{2^tN_0^2\sigma^2}{R^2L^2\ln\frac{4N_0\tau}{\beta}}\right\}\right)\\ &= O\left(\max\left\{N_0\tau, \sum_{t=0}^{\tau-1}\frac{2^tN_0^2\sigma^2}{R^2L^2\ln\frac{4N_0\tau}{\beta}}\right\}\right)\\ &= O\left(\max\left\{\frac{L}{\mu}\ln\left(\frac{\mu R^2}{\varepsilon}\right), \frac{\sigma^2}{\mu\varepsilon}\right\}\ln\left(\frac{L}{\mu\beta}\ln\frac{\mu R^2}{\varepsilon}\right)\right).\end{aligned}$$

G.3.3 Proof of Theorem G.3

Since f is L-smooth we have

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) - \gamma \left\langle \nabla f(x^k), \widetilde{\nabla} f(x^k, \boldsymbol{\xi}^k) \right\rangle + \frac{L\gamma^2}{2} \|\widetilde{\nabla} f(x^k, \boldsymbol{\xi}^k)\|_2^2 \\ &\overset{(12)}{\leq} f(x^k) - \gamma \|\nabla f(x^k)\|_2^2 - \gamma \left\langle \nabla f(x^k), \theta_k \right\rangle + L\gamma^2 \|\nabla f(x^k)\|_2^2 + L\gamma^2 \|\theta_k\|_2^2 \\ &= f(x^k) - \gamma(1 - L\gamma)\|\nabla f(x^k)\|_2^2 - \gamma \left\langle \nabla f(x^k), \theta_k \right\rangle + L\gamma^2 \|\theta_k\|_2^2 \\ &\leq f(x^k) - \frac{\gamma}{2}\|\nabla f(x^k)\|_2^2 - \gamma \left\langle \nabla f(x^k), \theta_k \right\rangle + L\gamma^2 \|\theta_k\|_2^2, \\ \theta_k &\overset{\text{def}}{=} \widetilde{\nabla} f(x^k, \boldsymbol{\xi}^k) - \nabla f(x^k) \end{aligned} \tag{113}$$

where in the last inequality we use $1 - \gamma L \geq \frac{1}{2}$. Next, μ-strong convexity of f implies $\|\nabla f(x^k)\|_2^2 \geq 2\mu(f(x^k) - f(x^*))$ and

$$\begin{aligned} f(x^{k+1}) - f(x^*)) &\leq f(x^k) - f(x^*) - \gamma\mu(f(x^k) - f(x^*)) - \gamma \left\langle \nabla f(x^k), \theta_k \right\rangle + L\gamma^2 \|\theta_k\|_2^2 \\ &= (1 - \gamma\mu)(f(x^k) - f(x^*)) - \gamma \left\langle \nabla f(x^k), \theta_k \right\rangle + L\gamma^2 \|\theta_k\|_2^2. \end{aligned}$$

Unrolling the recurrence we obtain

$$\begin{aligned} f(x^N) - f(x^*)) &\leq (1 - \gamma\mu)^N (f(x^0) - f(x^*)) + \gamma \sum_{l=0}^{N-1} (1 - \gamma\mu)^{N-1-l} \left\langle -\nabla f(x^l), \theta_l \right\rangle \\ &\quad + L\gamma^2 \sum_{l=0}^{N-1} (1 - \gamma\mu)^{N-1-l} \|\theta_l\|_2^2, \end{aligned} \tag{114}$$

for all $N \geq 0$. Using notation $r_k \overset{\text{def}}{=} f(x^k) - f(x^*)$ we rewrite this inequality in the following form:

$$r_k \leq (1 - \gamma\mu)^k r_0 + \gamma \sum_{l=0}^{k-1} (1 - \gamma\mu)^{k-1-l} \left\langle -\nabla f(x^l), \theta_l \right\rangle + L\gamma^2 \sum_{l=0}^{k-1} (1 - \gamma\mu)^{k-1-l} \|\theta_l\|_2^2. \tag{115}$$

The rest of the proof is based on the refined analysis of inequality (115). In particular, via induction we prove that for all $k = 0, 1, \ldots, N$ with probability at least $1 - \frac{k\beta}{N}$ the following statement holds: inequalities

$$\begin{aligned} r_t &\overset{(115)}{\leq} (1 - \gamma\mu)^t r_0 + \gamma \sum_{l=0}^{t-1} (1 - \gamma\mu)^{t-1-l} \left\langle -\nabla f(x^l), \theta_l \right\rangle + L\gamma^2 \sum_{l=0}^{t-1} (1 - \gamma\mu)^{t-1-l} \|\theta_l\|_2^2 \\ &\leq 2(1 - \gamma\mu)^t r_0 \end{aligned} \tag{116}$$

hold for $t = 0, 1, \ldots, k$ simultaneously. Let us define the probability event when this statement holds as E_k. Then, our goal is to show that $\mathbb{P}\{E_k\} \geq 1 - \frac{k\beta}{N}$ for all $k = 0, 1, \ldots, N$. For $t = 0$ inequality (116) holds with probability 1 since $2(1 - \gamma\mu)^0 \geq 1$, hence $\mathbb{P}\{E_0\} = 1$. Next, assume that for some $k = T - 1 \leq N - 1$ we have $\mathbb{P}\{E_k\} = \mathbb{P}\{E_{T-1}\} \geq 1 - \frac{(T-1)\beta}{N}$. Let us prove that $\mathbb{P}\{E_T\} \geq 1 - \frac{T\beta}{N}$. First of all, probability event E_{T-1} implies that

$$f(x^t) - f(x^*) \overset{(116)}{\leq} 2(1 - \gamma\mu)^t r_0 \tag{117}$$

hold for $t = 0, 1, \ldots, T - 1$. Since f is L-smooth, we have that probability event E_{T-1} implies

$$\left\|\nabla f(x^l)\right\|_2 \leq \sqrt{2L(f(x^l) - f(x^*))} \leq \sqrt{4L(1 - \gamma\mu)^l r_0} = \frac{\lambda_l}{2} \tag{118}$$

for $t = 0, \ldots, T - 1$ and

$$\lambda_l = 4\sqrt{L(1 - \gamma\mu)^l r_0}. \tag{119}$$

Having inequalities (118) in hand we show in the rest of the proof that (116) holds for $t = T$ with big enough probability. First of all, we introduce new random variables:

$$\zeta_l = \begin{cases} -\nabla f(x^{l+1}), & \text{if } \|\nabla f(x^{l+1})\|_2 \leq \frac{\lambda_l}{2}, \\ 0, & \text{otherwise,} \end{cases} \tag{120}$$

for $l = 0, 1, \ldots T-1$. Note that these random variables are bounded with probability 1, i.e. with probability 1 we have

$$\|\zeta_l\|_2 \leq \frac{\lambda_l}{2}. \tag{121}$$

Secondly, we use the introduced notation and get that E_{T-1} implies

$$\begin{aligned} r_T &\overset{(115),(116),(118),(120)}{\leq} (1-\gamma\mu)^T r_0 + \gamma \sum_{l=0}^{T-1} (1-\gamma\mu)^{T-1-l} \langle \zeta_l, \theta_l \rangle \\ &\quad + L\gamma^2 \sum_{l=0}^{T-1} (1-\gamma\mu)^{T-1-l} \|\theta_l\|_2^2. \end{aligned}$$

Finally, we do some preliminaries in order to apply Bernstein's inequality (see Lemma D.1) and obtain that E_{T-1} implies

$$\begin{aligned} r_T &\overset{(12)}{\leq} (1-\gamma\mu)^T r_0 + \underbrace{\gamma \sum_{l=0}^{T-1} (1-\gamma\mu)^{T-1-l} \langle \theta_l^u, \zeta_l \rangle}_{①} + \underbrace{\gamma \sum_{l=0}^{T-1} (1-\gamma\mu)^{T-1-l} \langle \theta_l^b, \zeta_l \rangle}_{②} \\ &\quad \underbrace{+ 2L\gamma^2 \sum_{l=0}^{T-1} (1-\gamma\mu)^{T-1-l} \left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l} \left[\|\theta_l^u\|_2^2 \right] \right)}_{③} \\ &\quad \underbrace{+ 2L\gamma^2 \sum_{l=0}^{T-1} (1-\gamma\mu)^{T-1-l} \mathbb{E}_{\boldsymbol{\xi}^l} \left[\|\theta_l^u\|_2^2 \right]}_{④} + \underbrace{2L\gamma^2 \sum_{l=0}^{T-1} (1-\gamma\mu)^{T-1-l} \|\theta_l^b\|_2^2}_{⑤} \end{aligned} \tag{122}$$

where we introduce new notations:

$$\theta_l^u \overset{\text{def}}{=} \widetilde{\nabla} f(x^l, \boldsymbol{\xi}^l) - \mathbb{E}_{\boldsymbol{\xi}^l} \left[\widetilde{\nabla} f(x^l, \boldsymbol{\xi}^l) \right], \quad \theta_l^b \overset{\text{def}}{=} \mathbb{E}_{\boldsymbol{\xi}^l} \left[\widetilde{\nabla} f(x^l, \boldsymbol{\xi}^l) \right] - \nabla f(x^l), \tag{123}$$

$$\theta_l = \theta_l^u + \theta_l^b.$$

It remains to provide tight upper bounds for ①, ②, ③, ④ and ⑤, i.e. in the remaining part of the proof we show that ① + ② + ③ + ④ + ⑤ $\leq (1-\gamma\mu)^T r_0$.

Upper bound for ①. First of all, since $\mathbb{E}_{\boldsymbol{\xi}^l}[\theta_l^u] = 0$ summands in ① are conditionally unbiased:

$$\mathbb{E}_{\boldsymbol{\xi}^l} \left[\gamma (1-\gamma\mu)^{T-1-l} \langle \theta_l^u, \zeta_l \rangle \right] = 0.$$

Secondly, these summands are bounded with probability 1:

$$\begin{aligned} \left| \gamma (1-\gamma\mu)^{T-1-l} \langle \theta_l^u, \zeta_l \rangle \right| &\leq \gamma (1-\gamma\mu)^{T-1-l} \|\theta_l^u\|_2 \|\zeta_l\|_2 \\ &\overset{(38),(121)}{\leq} \gamma (1-\gamma\mu)^{T-1-l} \lambda_l^2 \overset{(119)}{=} 16\gamma L r_0 (1-\gamma\mu)^{T-1}. \end{aligned}$$

Finally, one can bound conditional variances $\sigma_l^2 \overset{\text{def}}{=} \mathbb{E}_{\boldsymbol{\xi}^l} \left[\gamma^2 (1-\gamma\mu)^{2(T-1-l)} \langle \theta_l^u, \zeta_l \rangle^2 \right]$ in the following way:

$$\begin{aligned} \sigma_l^2 &\leq \mathbb{E}_{\boldsymbol{\xi}^l} \left[\gamma^2 (1-\gamma\mu)^{2(T-1-l)} \|\theta_l^u\|_2^2, \|\zeta_l\|_2^2 \right] \overset{(121)}{\leq} \gamma^2 (1-\gamma\mu)^{2(T-1-l)} \frac{\lambda^2}{4} \mathbb{E}_{\boldsymbol{\xi}^l} \left[\|\theta_l^u\|_2^2 \right] \\ &\overset{(119)}{\leq} 4\gamma^2 L r_0 (1-\gamma\mu)^{2(T-1)-l} \mathbb{E}_{\boldsymbol{\xi}^l} \left[\|\theta_l^u\|_2^2 \right]. \end{aligned} \tag{124}$$

In other words, sequence $\left\{\gamma(1-\gamma\mu)^{T-1-l}\left\langle\theta_l^u,\zeta_l\right\rangle\right\}_{l\geq 0}$ is a bounded martingale difference sequence with bounded conditional variances $\{\sigma_l^2\}_{l\geq 0}$. Therefore, we can apply Bernstein's inequality, i.e. we apply Lemma D.1 with $X_l=\gamma(1-\gamma\mu)^{T-1-l}\left\langle\theta_l^u,\zeta_l\right\rangle$, $c=16\gamma Lr_0(1-\gamma\mu)^{T-1}$ and $F=\frac{c^2\ln\frac{4N}{\beta}}{6}$ and get that for all $b>0$

$$\mathbb{P}\left\{\left|\sum_{l=0}^{T-1}X_l\right|>b \text{ and } \sum_{l=0}^{T-1}\sigma_l^2\leq F\right\}\leq 2\exp\left(-\frac{b^2}{2F+{}^{2cb}/_3}\right)$$

or, equivalently, with probability at least $1-2\exp\left(-\frac{b^2}{2F+{}^{2cb}/_3}\right)$

$$\text{either } \sum_{l=0}^{T-1}\sigma_l^2>F \quad \text{or} \quad \underbrace{\left|\sum_{l=0}^{T-1}X_l\right|}_{|①|}\leq b.$$

The choice of F will be clarified further, let us now choose b in such a way that $2\exp\left(-\frac{b^2}{2F+{}^{2cb}/_3}\right)=\frac{\beta}{2N}$. This implies that b is the positive root of the quadratic equation

$$b^2-\frac{2c\ln\frac{4N}{\beta}}{3}b-2F\ln\frac{4N}{\beta}=0,$$

hence

$$\begin{aligned}b &= \frac{c\ln\frac{4N}{\beta}}{3}+\sqrt{\frac{c^2\ln^2\frac{4N}{\beta}}{9}+2F\ln\frac{4N}{\beta}}=\frac{c\ln\frac{4N}{\beta}}{3}+\sqrt{\frac{4c^2\ln^2\frac{4N}{\beta}}{9}}\\ &= c\ln\frac{4N}{\beta}=16\gamma Lr_0(1-\gamma\mu)^{T-1}\ln\frac{4N}{\beta}.\end{aligned}$$

That is, with probability at least $1-\frac{\beta}{2N}$

$$\underbrace{\text{either } \sum_{l=0}^{T-1}\sigma_l^2>F \quad \text{or} \quad |①|\leq 16\gamma Lr_0(1-\gamma\mu)^{T-1}\ln\frac{4N}{\beta}}_{\text{probability event } E_①}.$$

Next, we notice that probability event E_{T-1} implies that

$$\begin{aligned}\sum_{l=0}^{T-1}\sigma_l^2 &\overset{(124)}{\leq} 4\gamma^2Lr_0\sigma^2(1-\gamma\mu)^{2(T-1)}\sum_{l=0}^{T-1}\mathbb{E}_{\boldsymbol{\xi}^l}\left[\left\|\theta_l^u\right\|_2^2\right]\\ &\overset{(41)}{\leq} 72\gamma^2Lr_0\sigma^2(1-\gamma\mu)^{2(T-1)}\sum_{l=0}^{T-1}\frac{1}{m_l(1-\gamma\mu)^l}\\ &\overset{(96)}{\leq} \frac{128}{3}\gamma^2L^2r_0^2(1-\gamma\mu)^{2(T-1)}\ln\frac{4N}{\beta}=\frac{c^2\ln\frac{4N}{\beta}}{6}=F,\end{aligned}$$

where the last inequality follows from $c=16\gamma Lr_0(1-\gamma\mu)^{T-1}$ and simple arithmetic.

Upper bound for ②. First of all, we notice that probability event E_{T-1} implies

$$\begin{aligned}\gamma(1-\gamma\mu)^{T-1-l}\left\langle\theta_l^b,\zeta_l\right\rangle &\leq \gamma(1-\gamma\mu)^{T-1-l}\left\|\theta_l^b\right\|_2\left\|\zeta_l\right\|_2\\ &\overset{(39),(121)}{\leq} \gamma(1-\gamma\mu)^{T-1-l}\frac{4\sigma^2}{m_l\lambda_l}\frac{\lambda_l}{2}\\ &= \frac{2\sigma^2\gamma(1-\gamma\mu)^{T-1-l}\sigma^2}{m_l}\\ &\overset{(96)}{=} \frac{64}{27}\frac{\gamma Lr_0(1-\gamma\mu)^{T-1}\ln\frac{4N}{\beta}}{N}.\end{aligned}$$

This implies that

$$② \quad = \quad \sum_{l=0}^{T-1} \gamma(1-\gamma\mu)^{T-1-l} \left\langle \theta_l^b, \zeta_l \right\rangle \overset{T\leq N}{\leq} \frac{64}{27}\gamma L r_0 (1-\gamma\mu)^{T-1} \ln\frac{4N}{\beta}.$$

Upper bound for ③. We derive the upper bound for ③ using the same technique as for ①. First of all, we notice that the summands in ③ are conditionally independent:

$$\mathbb{E}_{\boldsymbol{\xi}^l}\left[2L\gamma^2(1-\gamma\mu)^{T-1-l}\left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right)\right] = 0.$$

Secondly, the summands are bounded with probability 1:

$$\begin{aligned}\left|2L\gamma^2(1-\gamma\mu)^{T-1-l}\left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right)\right| &\leq 2L\gamma^2(1-\gamma\mu)^{T-1-l}\left(\|\theta_l^u\|_2^2 + \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right)\\ &\overset{(38)}{\leq} 2L\gamma^2(1-\gamma\mu)^{T-1-l}\left(4\lambda_l^2 + 4\lambda_l^2\right)\\ &\overset{(119)}{=} 256\gamma^2L^2r_0(1-\gamma\mu)^{T-1} \overset{\text{def}}{=} c_1. \end{aligned} \tag{125}$$

Finally, one can bound conditional variances $\hat{\sigma}_l^2 \overset{\text{def}}{=} \mathbb{E}_{\boldsymbol{\xi}^l}\left[\left|2L\gamma^2(1-\gamma\mu)^{T-1-l}\left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right)\right|^2\right]$ in the following way:

$$\begin{aligned}\hat{\sigma}_l^2 &\overset{(125)}{\leq} c_1\mathbb{E}_{\boldsymbol{\xi}^l}\left[\left|2L\gamma^2(1-\gamma\mu)^{T-1-l}\left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right)\right|\right]\\ &\leq 2L\gamma^2(1-\gamma\mu)^{T-1-l}c_1\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2 + \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right]\\ &= 4L\gamma^2(1-\gamma\mu)^{T-1-l}c_1\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]. \end{aligned} \tag{126}$$

In other words, sequence $\left\{2L\gamma^2(1-\gamma\mu)^{T-1-l}\left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right)\right\}_{l\geq 0}$ is a bounded martingale difference sequence with bounded conditional variances $\{\hat{\sigma}_l^2\}_{l\geq 0}$. Therefore, we can apply Bernstein's inequality, i.e. we apply Lemma D.1 with $X_l = \hat{X}_l = 2L\gamma^2(1-\gamma\mu)^{T-1-l}\left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right)$, $c = c_1 = 256\gamma^2L^2r_0(1-\gamma\mu)^{T-1}$ and $F = F_1 = \frac{c_1^2\ln\frac{4N}{\beta}}{6}$ and get that for all $b > 0$

$$\mathbb{P}\left\{\left|\sum_{l=0}^{T-1}\hat{X}_l\right| > b \text{ and } \sum_{l=0}^{T-1}\hat{\sigma}_l^2 \leq F_1\right\} \leq 2\exp\left(-\frac{b^2}{2F_1 + {}^{2c_1b}/_3}\right)$$

or, equivalently, with probability at least $1 - 2\exp\left(-\frac{b^2}{2F_1+{}^{2c_1b}/_3}\right)$

$$\text{either } \sum_{l=0}^{T-1}\hat{\sigma}_l^2 > F_1 \quad \text{or} \quad \underbrace{\left|\sum_{l=0}^{T-1}\hat{X}_l\right|}_{|③|} \leq b.$$

As in our derivations of the upper bound for ① we choose such b that $2\exp\left(-\frac{b^2}{2F_1+{}^{2c_1b}/_3}\right) = \frac{\beta}{2N}$, i.e.

$$b \quad = \quad \frac{c_1\ln\frac{4N}{\beta}}{3} + \sqrt{\frac{c_1^2\ln^2\frac{4N}{\beta}}{9} + 2F_1\ln\frac{4N}{\beta}} = c_1\ln\frac{4N}{\beta} = 256\gamma^2L^2r_0(1-\gamma\mu)^{T-1}\ln\frac{4N}{\beta}.$$

That is, with probability at least $1 - \frac{\beta}{2N}$

$$\underbrace{\text{either } \sum_{l=0}^{T-1}\hat{\sigma}_l^2 > F_1 \quad \text{or} \quad |③| \leq 256\gamma^2L^2r_0(1-\gamma\mu)^{T-1}\ln\frac{4N}{\beta}}_{\text{probability event } E_③}.$$

Next, we notice that probability event E_{T-1} implies that

$$\begin{aligned}\sum_{l=0}^{T-1}\hat{\sigma}_l^2 &\overset{(126)}{\leq} 4L\gamma^2(1-\gamma\mu)^{T-1}c_1\sum_{l=0}^{T-1}\frac{1}{(1-\gamma\mu)^l}\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\\ &\overset{(41)}{\leq} 72L\gamma^2(1-\gamma\mu)^{T-1}c_1\sigma^2\sum_{l=0}^{T-1}\frac{1}{(1-\gamma\mu)^l}\frac{1}{m_l} \overset{(96),T\leq N}{\leq} \frac{c_1^2\ln\frac{4N}{\beta}}{6} = F_1.\end{aligned}$$

Upper bound for ④. The probability event E_{T-1} implies

$$\begin{aligned}④ &= 2L\gamma^2\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right] \overset{(41)}{\leq} 2L\gamma^2(1-\gamma\mu)^{T-1}\sum_{l=0}^{T-1}\frac{1}{(1-\gamma\mu)^l}\frac{18\sigma^2}{m_l}\\ &\overset{(96),T\leq N}{\leq} \frac{64}{3}\gamma^2L^2r_0(1-\gamma\mu)^{T-1}\ln\frac{4N}{\beta}.\end{aligned}$$

Upper bound for ⑤. Again, we use corollaries of probability event E_{T-1}:

$$\begin{aligned}⑤ &= 2L\gamma^2\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\|\theta_l^b\|_2^2 \overset{(39)}{\leq} 2L\gamma^2(1-\gamma\mu)^{T-1}\sum_{l=0}^{T-1}\frac{1}{(1-\gamma\mu)^l}\frac{16\sigma^4}{m_l^2\lambda_l^2}\\ &\overset{(119),(96)}{=} \frac{512}{729}\gamma^2L^2r_0(1-\gamma\mu)^{T-1}\ln^2\frac{4N}{\beta}\sum_{l=0}^{T-1}\frac{1}{N^2}\\ &\overset{T\leq N}{\leq} \frac{512}{729}\frac{\gamma^2L^2r_0(1-\gamma\mu)^{T-1}\ln^2\frac{4N}{\beta}}{N}.\end{aligned}$$

Now we summarize all bounds that we have: probability event E_{T-1} implies

$$\begin{aligned}r_T &\overset{(115)}{\leq} (1-\gamma\mu)^Tr_0+\gamma\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\left\langle-\nabla f(x^l),\theta_l\right\rangle+L\gamma^2\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\|\theta_l\|_2^2\\ &\overset{(122)}{\leq} (1-\gamma\mu)^Tr_0+①+②+③+④+⑤,\\ ② &\leq \frac{32}{27}\gamma Lr_0(1-\gamma\mu)^{T-1}\ln\frac{4N}{\beta},\quad ④\leq\frac{64}{3}\gamma^2L^2r_0(1-\gamma\mu)^{T-1}\ln\frac{4N}{\beta},\\ ⑤ &\leq \frac{512}{729}\frac{\gamma^2L^2r_0(1-\gamma\mu)^{T-1}\ln^2\frac{4N}{\beta}}{N},\quad \sum_{l=0}^{T-1}\sigma_l^2\leq F,\quad \sum_{l=0}^{T-1}\hat{\sigma}_l^2\leq F_1\end{aligned}$$

and

$$\mathbb{P}\{E_{T-1}\}\geq 1-\frac{(T-1)\beta}{N},\quad \mathbb{P}\{E_①\}\geq 1-\frac{\beta}{2N},\quad \mathbb{P}\{E_③\}\geq 1-\frac{\beta}{2N},$$

where

$$\begin{aligned}E_① &= \left\{\text{either } \sum_{l=0}^{T-1}\sigma_l^2>F \quad\text{or}\quad |①|\leq 16\gamma Lr_0(1-\gamma\mu)^{T-1}\ln\frac{4N}{\beta}\right\},\\ E_③ &= \left\{\text{either } \sum_{l=0}^{T-1}\hat{\sigma}_l^2>F_1 \quad\text{or}\quad |③|\leq 256\gamma^2L^2r_0(1-\gamma\mu)^{T-1}\ln\frac{4N}{\beta}\right\}.\end{aligned}$$

Taking into account these inequalities and our assumptions on m_k and γ (see (96) and (97)) we get that probability event $E_{T-1}\cap E_①\cap E_③$ implies

$$\begin{aligned}r_T &\overset{(115)}{\leq} (1-\gamma\mu)^Tr_0+\gamma\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\left\langle-\nabla f(x^l),\theta_l\right\rangle+L\gamma^2\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\|\theta_l\|_2^2\\ &\leq (1-\gamma\mu)^Tr_0+\left(\frac{1}{5}+\frac{1}{5}+\frac{1}{5}+\frac{1}{5}+\frac{1}{5}\right)(1-\gamma\mu)^Tr_0=2(1-\gamma\mu)^Tr_0.\end{aligned}\tag{127}$$

Moreover, using union bound we derive

$$\mathbb{P}\left\{E_{T-1}\cap E_①\cap E_③\right\}=1-\mathbb{P}\left\{\overline{E}_{T-1}\cup\overline{E}_①\cup\overline{E}_③\right\}\geq 1-\frac{T\beta}{N}.\tag{128}$$

That is, by definition of E_T and E_{T-1} we have proved that

$$\mathbb{P}\{E_T\}\overset{(127)}{\geq}\mathbb{P}\left\{E_{T-1}\cap E_①\cap E_③\right\}\overset{(128)}{\geq}1-\frac{T\beta}{N},$$

which implies that for all $k = 0, 1, \ldots, N$ we have $\mathbb{P}\{E_k\} \geq 1 - \frac{k\beta}{N}$. Then, for $k = N$ we have that with probability at least $1 - \beta$

$$f(x^N) - f(x^*)) \overset{(114)}{\leq} (1-\gamma\mu)^N (f(x^0) - f(x^*)) + \gamma \sum_{l=0}^{N-1} (1-\gamma\mu)^{N-1-l} \left\langle -\nabla f(x^l), \theta_l \right\rangle + L\gamma^2 \sum_{l=0}^{N-1} (1-\gamma\mu)^{N-1-l} \|\theta_l\|_2^2 \overset{(116)}{\leq} 2(1-\gamma\mu)^N (f(x^0) - f(x^*)) \quad (129)$$

As a result, we get that with probability at least $1 - \beta$

$$\begin{aligned} f(x^N) - f(x^*) &\leq 2(1-\gamma\mu)^N (f(x^0) - f(x^*)) \leq 2\exp\left(-\gamma\mu N\right)(f(x^0) - f(x^*)) \\ &\overset{(97)}{\leq} 2\exp\left(-\frac{\mu N}{80L\ln\frac{4N}{\beta}}\right)(f(x^0) - f(x^*)). \end{aligned}$$

In other words, `clipped-SGD` achieves $f(x^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ after

$$O\left(\frac{L}{\mu}\ln\left(\frac{r_0}{\varepsilon}\right)\ln\left(\frac{L}{\mu\beta}\ln\left(\frac{r_0}{\varepsilon}\right)\right)\right)$$

iterations, where $r_0 = f(x^0) - f(x^*)$ and requires

$$\begin{aligned} \sum_{k=0}^{N-1} m_k &\overset{(96)}{=} \sum_{k=0}^{N-1} O\left(\max\left\{1, \frac{N\sigma^2}{Lr_0(1-\gamma\mu)^k \ln\frac{4N}{\beta}}\right\}\right) \\ &\overset{(97)}{=} O\left(\max\left\{N, \frac{N\sigma^2}{\mu r_0(1-\gamma\mu)^{N-1}}\right\}\right) = O\left(\max\left\{N, \frac{N\sigma^2}{\mu\varepsilon}\right\}\right) \\ &= O\left(\max\left\{\frac{L}{\mu}, \frac{\sigma^2}{\mu\varepsilon}\cdot\frac{L}{\mu}\right\}\ln\left(\frac{r_0}{\varepsilon}\right)\ln\left(\frac{L}{\mu\beta}\ln\left(\frac{r_0}{\varepsilon}\right)\right)\right). \end{aligned}$$

oracle calls.

H Extra Experiments

H.1 Detailed Description of Experiments from Section 1.2

In this section we provide a detailed description of experiments from Section 1.2 together with additional experiments. In these experiments we consider the following problem:

$$\min_{x\in\mathbb{R}^n} f(x), \quad f(x) = \|x\|_2^2/2 = \mathbb{E}_\xi\left[f(x,\xi)\right], \quad f(x,\xi) = \|x\|_2^2/2 + \langle \xi, x\rangle \tag{130}$$

where ξ is a random vector with zero mean and bounded variance. Clearly, $f(x)$ is μ-strongly convex and L-smooth with $\mu = L = 1$. We assume that $\mathbb{E}\left[\|\xi\|_2^2\right] \leq \sigma^2$ for some non-negative number σ. Then, the stochastic gradient $\nabla f(x,\xi) = x + \xi$ satisfies conditions (2) and the state-of-the-art theory (e.g. [24, 25]) says that after k iterations of SGD with constant stepsize $\gamma \leq 1/L = 1$ we have $\mathbb{E}\left[\|x^k - x^*\|_2^2\right] \leq (1-\gamma\mu)^k\|x^0 - x^*\|_2^2 + \gamma\sigma^2/\mu$. Taking into account that for our problem $x^* = 0$, $f(x) = \frac{1}{2}\|x\|_2^2$, $f(x^*) = 0$ and $\mu = 1$ we derive

$$\mathbb{E}\left[f(x^k) - f(x^*)\right] \leq (1-\gamma)^k\left(f(x^0) - f(x^*)\right) + \gamma\sigma^2/2. \tag{131}$$

That is, for given k the r.h.s. of the formula above depends only on the stepsize γ, initial suboptimality $f(x^0) - f(x^*)$ and the variance σ.

We emphasize that the obtained bound and the convergence in expectation itself does not imply non-trivial upper bound for $f(x^k) - f(x^*)$ with high-probability without additional assumptions on the distribution of random vector ξ. In fact, the trajectory of SGD significantly depends on the distribution of ξ. To illustrate this we consider 3 different distributions of ξ with the same σ.

1. In the first case we consider ξ from standard normal distribution, i.e. ξ is a Gaussian random vector with zero mean and covariance matrix I. Clearly, in this situation $\sigma^2 = n$.
2. Next, we consider a random vector ξ with i.i.d. components having Weibull distribution [69]. The cumulative distribution function (CDF) for Weibull distribution with parameters $c > 0$ and $\alpha > 0$ is

$$\text{CDF}_W(x) = \begin{cases} 1 - \exp\left(-\left(\frac{x}{\alpha}\right)^c\right), & \text{if } x \geq 0, \\ 0, & \text{if } x < 0. \end{cases} \tag{132}$$

   There are explicit formulas for mean and variance for Weibull distribution:

$$\text{mean} = \alpha\Gamma\left(1 + \frac{1}{c}\right), \quad \text{variance} = \alpha^2\left(\Gamma\left(1 + \frac{2}{c}\right) - \left(\Gamma\left(1 + \frac{1}{c}\right)\right)^2\right),$$

   where Γ denotes the gamma function. Having these formulas one can easily shift and scale the distribution in order to get a random variable with zero mean and the variance equal 1. In our experiments, we take $c = 0.2$,

$$\alpha = \frac{1}{\sqrt{\Gamma\left(1 + \frac{2}{c}\right) - \left(\Gamma\left(1 + \frac{1}{c}\right)\right)^2}},$$

   shift the distribution by $-\alpha\Gamma\left(1 + \frac{1}{c}\right)$ and sample from the obtained distribution n i.i.d. random variables to form ξ. Such a choice of parameters implies that $\mathbb{E}[\xi] = 0$ and $\mathbb{E}[\|\xi\|_2^2] = n$.
3. Finally, we consider a random vector ξ with i.i.d. components having Burr Type XII distribution [3] having the following cumulative distribution function

$$\text{CDF}_B(x) = \begin{cases} 1 - (1 + x^c)^{-d}, & \text{if } x > 0, \\ 0, & \text{if } x \leq 0, \end{cases} \tag{133}$$

   where $c > 0$ and $d > 0$ are the positive parameters. There are explicit formulas for mean and variance for Burr distribution:

$$\text{mean} = \mu_1, \quad \text{variance} = -\mu_1^2 + \mu_2,$$

where the r-th moment (if exists) is defined as follows [42]:

$$\mu_r = d\mathrm{B}\left(\frac{cd-r}{c}, \frac{c+r}{c}\right),$$

where B denotes the beta function.

In our experiments, we take $c = 1$ and $d = 2.3$ and then apply shifts and scales similarly to the case with Weibull distribution. Again, such a choice of parameters implies that $\mathbb{E}[\xi] = 0$ and $\mathbb{E}[\|\xi\|_2^2] = n$.

For all experiments we considered the dimension $n = 100$, the stepsize $\gamma = 0.001$ and for `clipped-SGD` we set $\lambda = 100$. The result of 10 independent runs of `SGD` and `clipped-SGD` are presented in Figures 6-10. These numerical tests show that for Weibull and Burr Type XII distributions `SGD` have significantly larger oscillations than for Gaussian distribution in all 10 tests. In contrast, `clipped-SGD` behaves much more robust in all 3 cases during all 10 runs without significant oscillations.

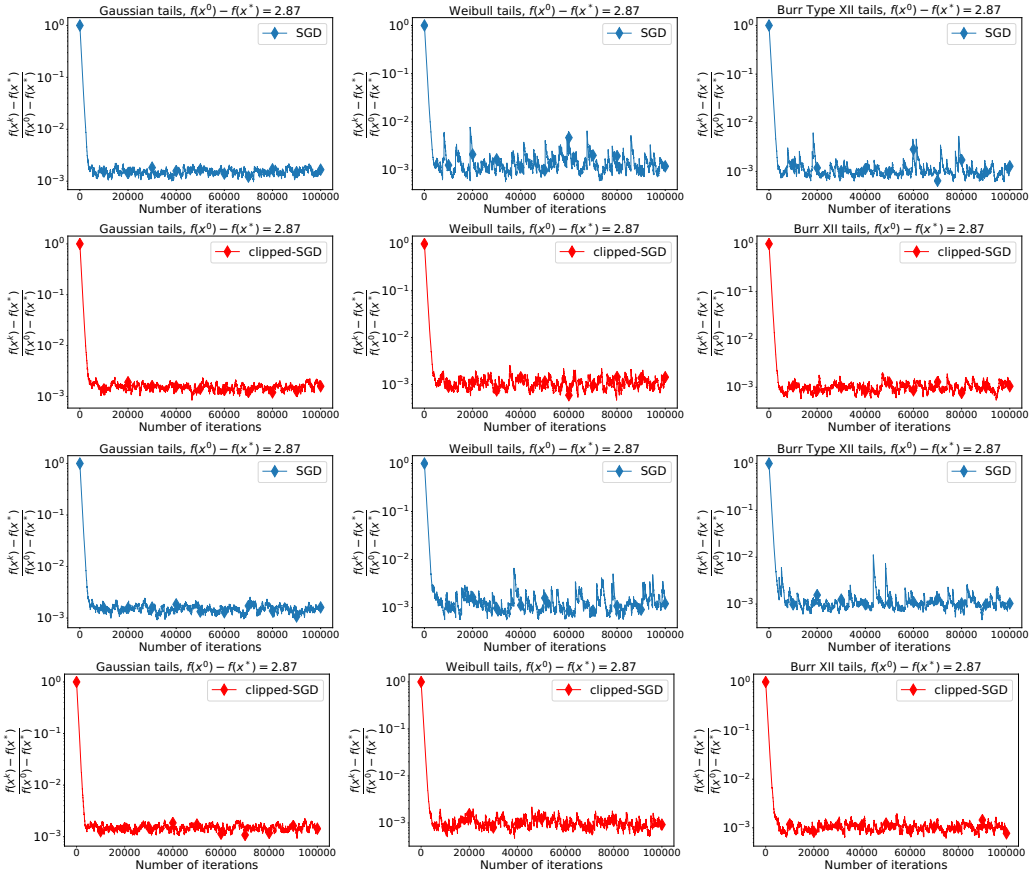

Figure 6: 2 independent runs of `SGD` (blue) and `clipped-SGD` (red) applied to solve (130) with ξ having Gaussian (left column), Weibull (central column) and Burr Type XII (right column) tails.

H.2 Additional Details and Experiments with Logistic Regression

In this section, we provide additional details of the experiments presented in Section 4 together with extra numerical results. In particular, we consider the logistic regression problem:

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{r}\sum_{i=1}^{r} \underbrace{\log\left(1 + \exp\left(-y_i \cdot (Ax)_i\right)\right)}_{f_i(x)} \tag{134}$$

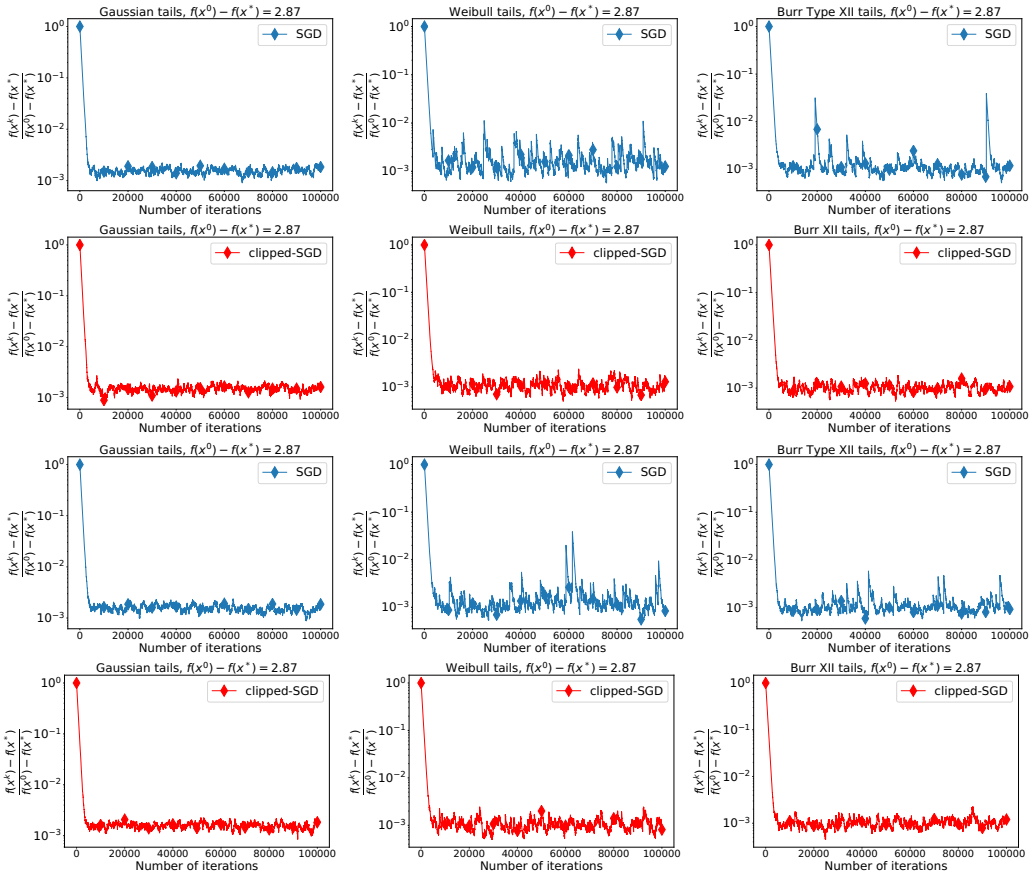

Figure 7: 2 independent runs of SGD (blue) and clipped-SGD (red) applied to solve (130) with ξ having Gaussian (left column), Weibull (central column) and Burr Type XII (right column) tails.

where $A \in \mathbb{R}^{r \times n}$ is matrix of instances and $y \in \{0,1\}^r$ is vector of labels. It is well-known that $f(x)$ from (134) is convex and L-smooth with $L = {}^{\lambda_{\max}(A^\top A)}/_{4r}$ where $\lambda_{\max}(A^\top A)$ denotes the maximal eigenvalue of $A^\top A$. One can consider problem (134) as a special case of (1) where ξ is a random index uniformly distributed on $\{1, \ldots, r\}$ and $f(x, \xi) = f_\xi(x)$. We take the datasets from LIBSVM library [4]: see Table 3 with the summary of the datasets we used.

Table 3: Summary of used datasets.

	heart	diabetes	australian	a9a	w8a
Size	270	768	690	32561	49749
Dimension	13	8	13	123	300

We notice that in all experiments that we did with logistic regression the initial suboptimality $f(x^0) - f(x^*)$ was of order 10. Moreover, as it was mentioned in the main part of the paper the parameters for the methods were tuned. One can find parameters that we used in the experiments from Section 4 in Table 4.

Next, we provide our numerical study of the distribution of $\|\nabla f_i(x^k) - \nabla f(x^k)\|_2$, where x^k is the last iterate produced by SGD in experiments presented in Section 4, see Figure 11. As we mentioned in the main part of the paper these histograms are very similar to ones presented in Figure 2, so, the insights that we got from Figure 2 are right. However, in our experiments with australian dataset SGD with the stepsize $\gamma = {}^1/_L$ did not reach needed suboptimality in order to oscillate.

Therefore, we run SGD along with its clipped variants with the same batchsize $m = 50$ for bigger number of epochs and also tuned their parameters. One can find the results of these runs in Figure 12.

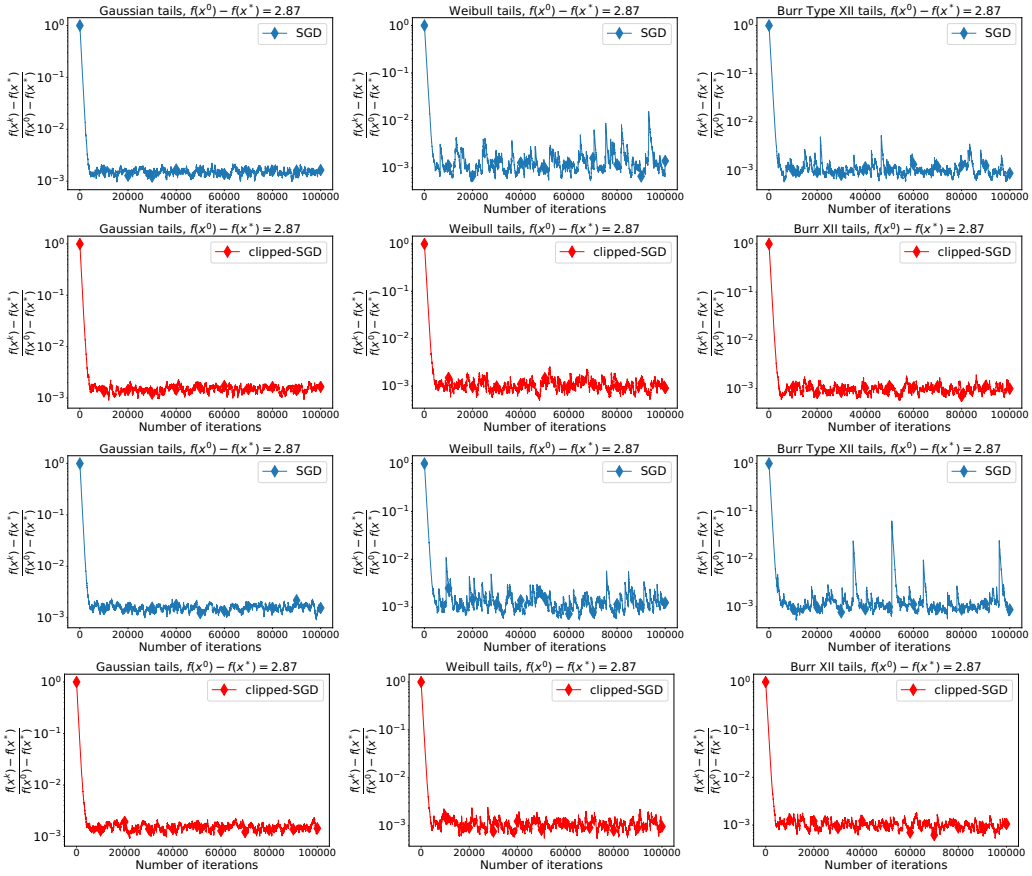

Figure 8: 2 independent runs of SGD (blue) and clipped-SGD (red) applied to solve (130) with ξ having Gaussian (left column), Weibull (central column) and Burr Type XII (right column) tails.

Table 4: Parameters that are used to produce plots presented in Figures 3-5. In the first contains the name of the dataset and the batchsize m that was used for all methods tested on the dataset. For d-clipped-SGD λ_0 is an initial clipping level, l is a period (in terms of epochs) of decreasing the clipping level and α is a coefficient of decrease, i.e. every l epochs the clipping level is multiplied by α. For SSTM parameter a was picked the same as for clipped-SSTM in order to emphasize the effect of clipping.

	SGD	clipped-SGD	d-clipped-SGD	SSTM	clipped-SSTM
heart $m = 20$	$\gamma = \frac{1}{2L}$	$\gamma = \frac{1}{2L}, \lambda = 2.72$	$\gamma = \frac{1}{2L}, \lambda_0 = 2.72,$ $l = 10^3, \alpha = 0.9$	$a = 10^4$	$a = 10^4,$ $B = 2 \cdot 10^{-4}$
diabetes $m = 100$	$\gamma = \frac{1}{10L}$	$\gamma = \frac{1}{10L}, \lambda = 68.86$	$\gamma = \frac{1}{10L}, \lambda_0 = 68.86,$ $l = 10^3, \alpha = 0.7$	$a = 5 \cdot 10^3$	$a = 5 \cdot 10^3,$ $B = 7 \cdot 10^{-4}$
australian $m = 50$	$\gamma = \frac{1}{L}$	$\gamma = \frac{1}{L}, \lambda = 74.47$	$\gamma = \frac{1}{L}, \lambda_0 = 74.47,$ $l = 1000, \alpha = 0.9$	$a = 10^3$	$a = 5 \cdot 10^3,$ $B = 2 \cdot 10^{-4}$
a9a $m = 100$	$\gamma = \frac{1}{2L}$	$\gamma = \frac{1}{2L}, \lambda = 0.025$	$\gamma = \frac{1}{L}, \lambda_0 = 4.9,$ $l = 5, \alpha = 0.5$	$a = 1$	$a = 1,$ $B = 3 \cdot 10^{-2}$
w8a $m = 1000$	$\gamma = \frac{1}{L}$	$\gamma = \frac{1}{L}, \lambda = 1.3$	$\gamma = \frac{1}{L}, \lambda_0 = 64.78,$ $l = 50, \alpha = 0.9$	$a = 1$	$a = 1,$ $B = 19 \cdot 10^{-2}$

We see that SGD with this stepsize achieves better suboptimality but it also oscillates significantly more. In contrast, clipped-SGD and d-clipped-SGD do not have significant oscillations and converge with the same rate as SGD. Moreover, clipped-SSTM shows slightly better performance in this case. Finally, we numerically studied the distribution of $\|\nabla f_i(x^k) - \nabla f(x^k)\|_2$, where x^k is the last iterate produced by SGD, see Figure 13. These histograms imply that the noise in stochastic gradients is heavy-tailed and explain an unstable behavior of SGD in this case.

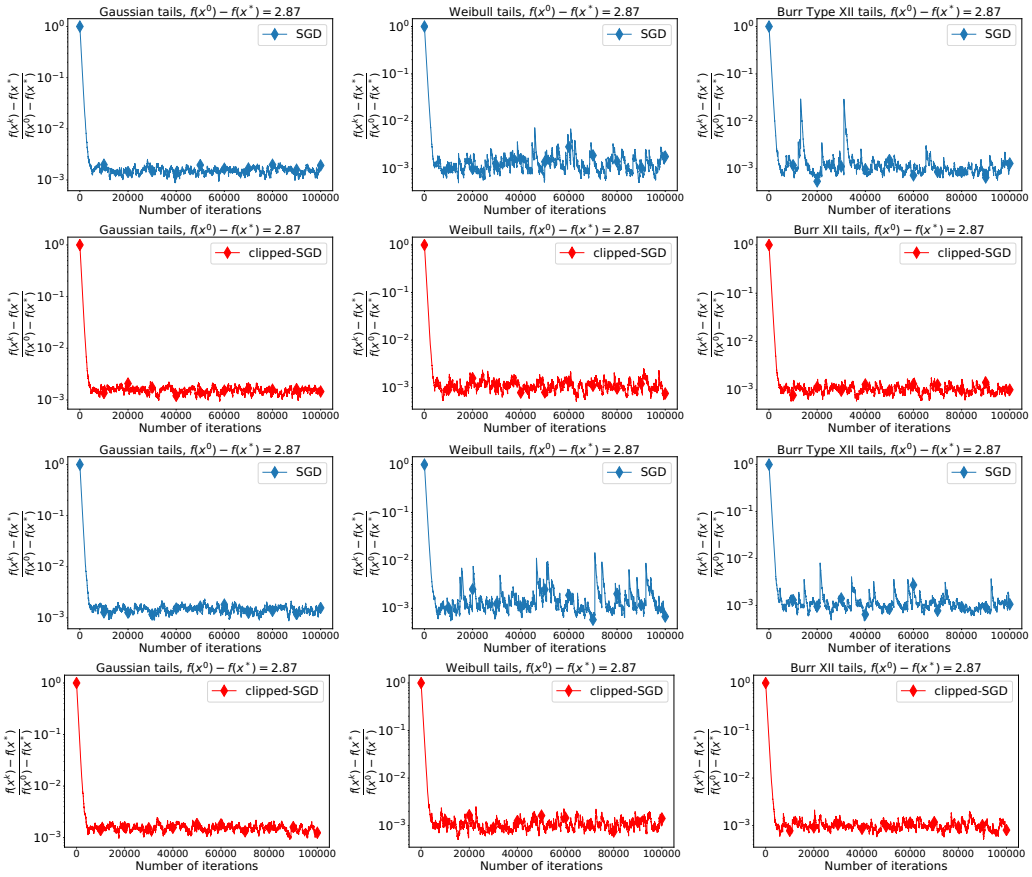

Figure 9: 2 independent runs of `SGD` (blue) and `clipped-SGD` (red) applied to solve (130) with ξ having Gaussian (left column), Weibull (central column) and Burr Type XII (right column) tails.

Finally, we conducted experiments on larger datasets: `a9a` and `w8a`. The results of our numerical test are reported on Figures 14 and 15. We notice that `SSTM` with given stepsize and batchsize suffers from noise accumulation, while `clipped-SSTM` does not have this drawback and shows comparable performance with `SGD` on `a9a` and much better performance on `w8a`.

Figure 15 shows the gradient's noise distributions for both datasets. While the distribution of stochastic gradients at the optimum for `a9a` have sub-Gaussian-like distribution, for `w8a` they have heavy-tailed distribution.

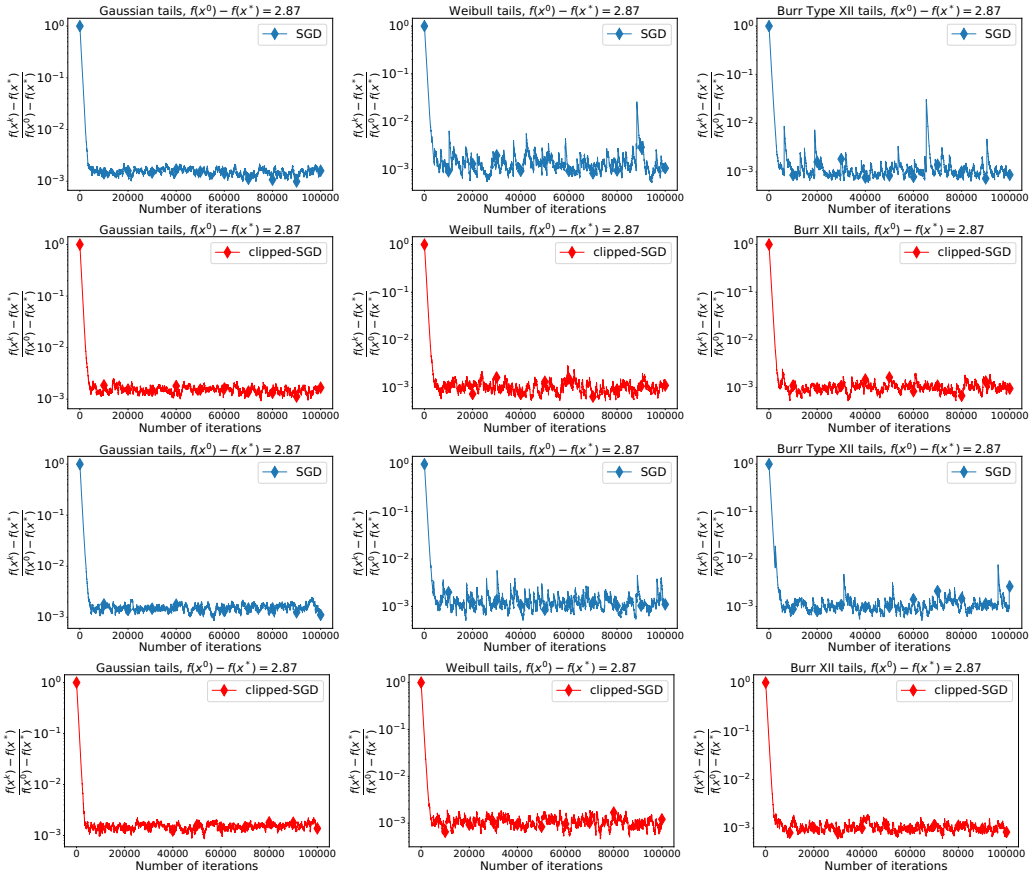

Figure 10: 2 independent runs of SGD (blue) and clipped-SGD (red) applied to solve (130) with ξ having Gaussian (left column), Weibull (central column) and Burr Type XII (right column) tails.

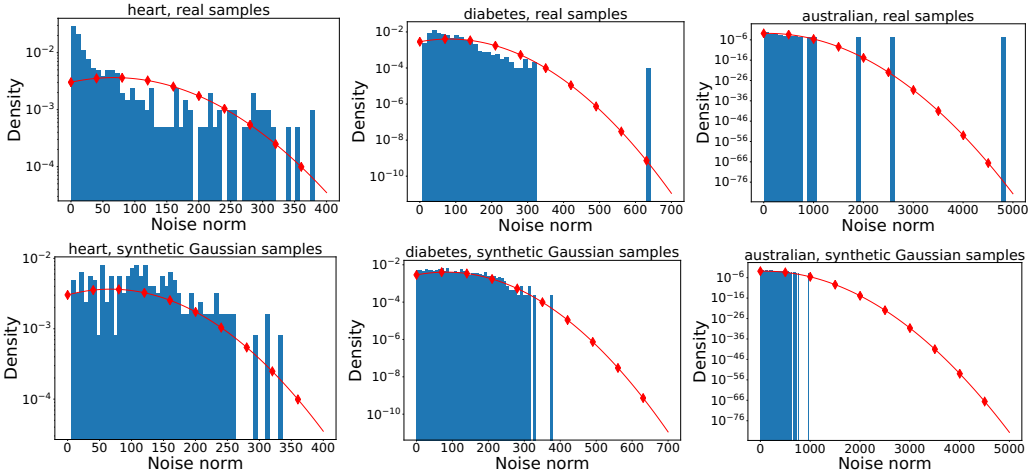

Figure 11: Histograms of $\|\nabla f_i(x^k) - \nabla f(x^k)\|_2$ for different datasets (the first row) and synthetic Gaussian samples with mean and variance estimated via empirical mean and variance of real samples $\|\nabla f_1(x^k) - \nabla f(x^k)\|_2, \ldots, \|\nabla f_r(x^k) - \nabla f(x^k)\|_2$ (the second row) where x^k is the last point produced by SGD. Red lines correspond to probability density functions of normal distributions with empirically estimated means and variances.

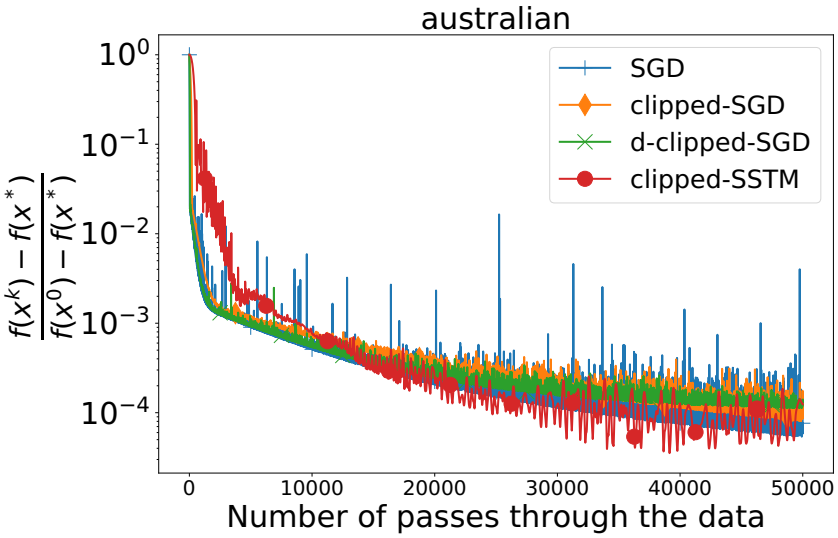

Figure 12: Trajectories of `SGD`, `clipped-SGD`, `d-clipped-SGD` and `clipped-SSTM` applied to solve logistic regression problem on `australian` dataset. For `SGD` and its clipped variants stepsize $\gamma = \frac{20}{L}$ was used. For `clipped-SGD` we used $\lambda = 18.62$ and for `d-clipped-SGD` the parameters are as follows: $\lambda_0 = 74.47$, $l = 1500$, $\alpha = 0.9$. Parameters for `clipped-SSTM` are the same as in the corresponding cell in Table 4.

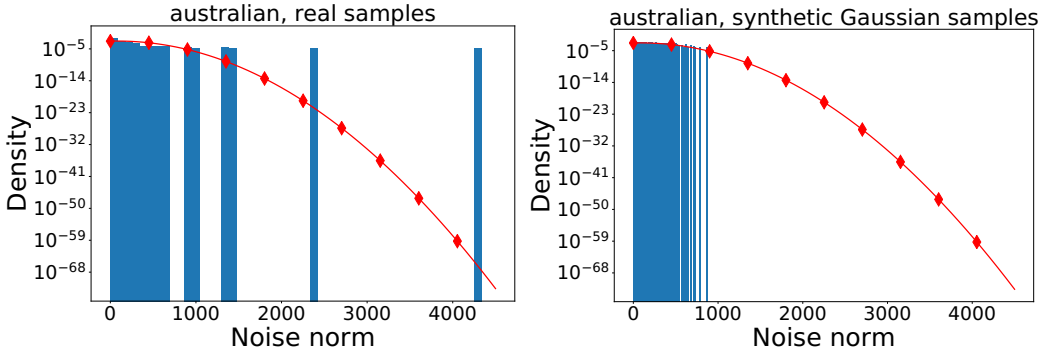

Figure 13: Histograms of $\|\nabla f_i(x^k) - \nabla f(x^k)\|_2$ for `australian` dataset and synthetic Gaussian samples with mean and variance estimated via empirical mean and variance of real samples $\|\nabla f_1(x^k) - \nabla f(x^k)\|_2, \ldots, \|\nabla f_r(x^k) - \nabla f(x^k)\|_2$ where x^k is the last point produced by `SGD` with $\gamma = \frac{20}{L}$. Red lines correspond to probability density functions of normal distributions with empirically estimated means and variances.

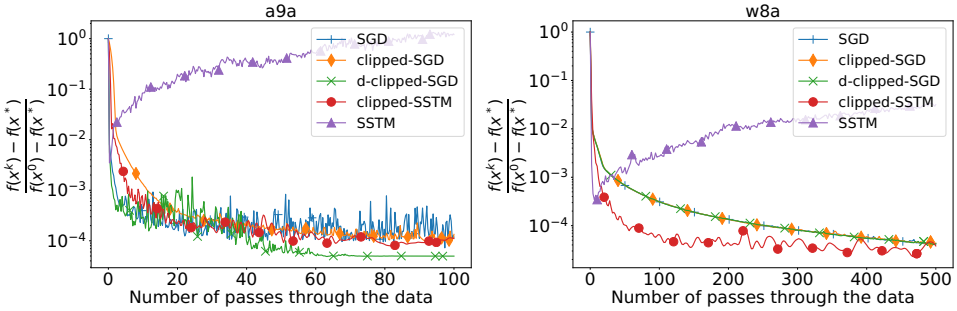

Figure 14: Trajectories of `SGD`, `clipped-SGD`, `d-clipped-SGD` and `clipped-SSTM` applied to solve logistic regression problem on `a9a` and `w8a` datasets. Parameters of the methods used in experiments are presneted in Table 4.

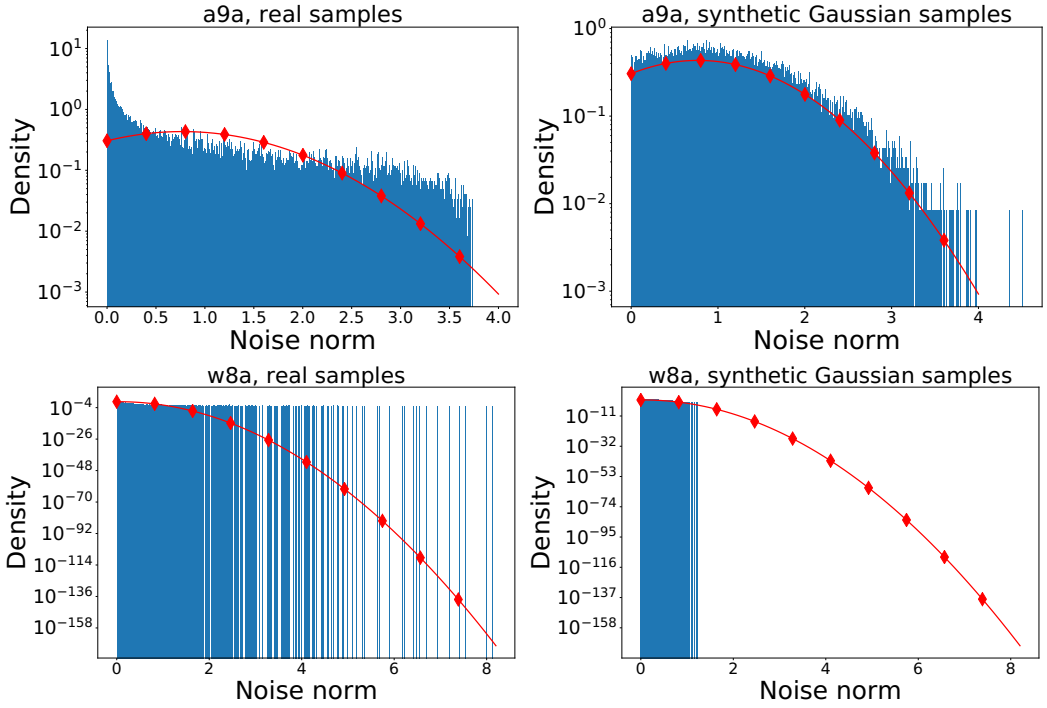

Figure 15: Histograms of $\|\nabla f_i(x^*)\|_2$ for `a9a` and `w8a` dataset and synthetic Gaussian samples with mean and variance estimated via empirical mean and variance of real samples $\|\nabla f_1(x^*)\|_2, \ldots, \|\nabla f_r(x^*)\|_2$. Red lines correspond to probability density functions of normal distributions with empirically estimated means and variances.