High Probability Complexity Bounds for Non-Smooth Stochastic Optimization with Heavy-Tailed Noise

Eduard Gorbunov[1] Marina Danilova[2,3] Innokentiy Shibaev[2,4]
Pavel Dvurechensky[5] Alexander Gasnikov[6,2,7]

[1] Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
[2] Moscow Institute of Physics and Technology, Russian Federation
[3] Institute of Control Sciences RAS, Russian Federation
[4] National Research University Higher School of Economics, Russian Federation
[5] Weierstrass Institute for Applied Analysis and Stochastics, Germany
[6] Innopolis University, Russian Federation
[7] Institute for Information Transmission Problems RAS, Russian Federation

Abstract

Stochastic first-order methods are standard for training large-scale machine learning models. Random behavior may cause a particular run of an algorithm to result in a highly suboptimal objective value, whereas theoretical guarantees are usually proved for the expectation of the objective value. Thus, it is essential to theoretically guarantee that algorithms provide small objective residuals with high probability. Existing methods for non-smooth stochastic convex optimization have complexity bounds with the dependence on the confidence level that is either negative-power or logarithmic but under an additional assumption of sub-Gaussian (light-tailed) noise distribution that may not hold in practice. In our paper, we resolve this issue and derive the first high-probability convergence results with logarithmic dependence on the confidence level for non-smooth convex stochastic optimization problems with non-sub-Gaussian (heavy-tailed) noise. To derive our results, we propose novel stepsize rules for two stochastic methods with gradient clipping. Moreover, our analysis works for generalized smooth objectives with Hölder-continuous gradients, and for both methods, we provide an extension for strongly convex problems. Finally, our results imply that the first (accelerated) method we consider also has optimal iteration and oracle complexity in all the regimes, and the second one is optimal in the non-smooth setting.

1 Introduction

Stochastic first-order optimization methods like SGD (Robbins and Monro, 1951), Adam (Kingma and Ba, 2015), and their various modifications are extremely popular in solving a number of different optimization problems, especially those appearing in statistics (Spokoiny, 2012), machine learning, and deep learning (Goodfellow et al., 2016). The success of these methods in real-world applications motivates the researchers to investigate the theoretical properties of the methods and to develop new ones with better convergence guarantees. Typically, stochastic methods are analyzed in terms of the convergence in expectation (see (Ghadimi and Lan, 2013; Gower et al., 2019; Moulines and Bach, 2011) and references therein), whereas high-probability complexity results are established more rarely. However, as illustrated in (Gorbunov et al., 2020), guarantees in terms of the convergence in

expectation have a much worse correlation with the real behavior of the methods than high-probability convergence guarantees when the noise in the stochastic gradients has *heavy-tailed distribution*.

Recent studies (Şimşekli et al., 2019; Simsekli et al., 2019; Zhang et al., 2020b) show that in several popular problems such as training `BERT` (Devlin et al., 2019) on the `Wikipedia` dataset the noise in the stochastic gradients is heavy-tailed. Moreover, (Zhang et al., 2020b) justify empirically that in such cases, SGD works significantly worse than clipped-SGD (Pascanu et al., 2013) and Adam. Therefore, it is important to theoretically study the methods' convergence when the noise is heavy-tailed.

For convex and strongly convex problems with Lipschitz continuous gradient, i.e., smooth convex and strongly convex problems, this question was properly addressed in (Davis et al., 2021; Gorbunov et al., 2020; Nazin et al., 2019), where the first high-probability complexity bounds with logarithmic dependence on the confidence level were derived for the stochastic problems with heavy-tailed noise. However, a number of practically important problems are non-smooth *on the whole space* (Mai and Johansson, 2021; Zhang et al., 2020a). For example, in deep neural network training, the loss function often grows polynomially fast when the norm of the network's weights goes to infinity. Moreover, non-smoothness of the activation functions such as ReLU or loss functions such as hinge loss implies the non-smoothness of the whole problem. While being well-motivated by practical applications, the existing high-probability convergence guarantees for stochastic first-order methods applied to solve non-smooth convex optimization problems with heavy-tailed noise have a drawback. Namely, the existing complexity bounds depend on the negative power of the confidence level. This dramatically increases the number of iterations required to obtain high accuracy of the solution with probability close to one. Such a discrepancy in the theory between algorithms for stochastic smooth and non-smooth problems leads us to the natural question: *is it possible to obtain high-probability complexity bounds with logarithmic dependence on the confidence level for* ***non-smooth*** *convex stochastic problems with heavy-tailed noise?* In this paper, we give a positive answer to this question. Moreover, we derive the corresponding bounds under much weaker assumptions than the ones used in the previous works. To achieve this we focus on gradient clipping methods, as in (Gehring et al., 2017; Mai and Johansson, 2021; Menon et al., 2020; Pascanu et al., 2013; Zhang et al., 2020a,b).

1.1 Preliminaries

Before we describe our contributions in detail, we formally state the considered setup.

Notation and standard definitions. We use standard notation for stochastic optimization literature. For all $x \in \mathbb{R}^n$ we use $\|x\|_2 = \sqrt{\langle x, x\rangle}$ to denote standard Euclidean norm, where $\langle x, y\rangle = x_1y_1 + x_2y_2 + \ldots + x_ny_n$, $x = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$. Next, we use $\mathbb{E}[\xi]$ and $\mathbb{E}[\xi \mid \eta]$ to denote expectation of ξ and expectation of ξ conditioned on η respectively. In some places of the paper, we also use $\mathbb{E}_\xi[\cdot]$ to denote conditional expectation taken w.r.t. the randomness coming from ξ. The probability of event E is defined as $\mathbb{P}\{E\}$. Finally, we use the following definition.

Definition 1.1. *A differentiable function $f : Q \subseteq \mathbb{R}^n \to \mathbb{R}$ is called μ-strongly convex for some $\mu \geq 0$ if for all $x, y \in Q$*

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

When $\mu = 0$, the function f is called convex.

Stochastic optimization. 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 is a convex but possibly non-smooth function. Next, we assume that at each point $x \in \mathbb{R}^n$ we have an access to the unbiased estimator $\nabla f(x,\xi)$ of $\nabla f(x)$ such that $\mathbb{E}_\xi\left[\|\nabla f(x,\xi) - \nabla f(x)\|_2^2\right] < \infty$ and if additionally $x \in Q \subseteq \mathbb{R}^n$, then

$$\mathbb{E}_\xi[\nabla f(x,\xi)] = \nabla f(x), \quad \mathbb{E}_\xi\left[\|\nabla f(x,\xi) - \nabla f(x)\|_2^2\right] \le \sigma^2, \quad \sigma > 0. \tag{2}$$

This assumption with $Q = \mathbb{R}^n$ on the stochastic oracle is widely used in stochastic optimization literature (Ghadimi and Lan, 2012, 2013; Juditsky and Nemirovski, 2011; Lan, 2012; Nemirovski et al., 2009). In contrast, in our theoretical results, we assume that Q is the ball centered at some[1] solution x^* of (1) with radius $\sim R_0 \ge \|x^0 - x^*\|_2$, where x^0 is a starting point of the method, i.e., *our analysis does not require* (2) *to hold on* $\mathbb{R}^n$. That is, our assumption on the noise is much more general than those used in previous works in the area. Finally, we emphasize that we do not assume that the stochastic gradients have so-called "light tails" (Lan, 2012), i.e., sub-Gaussian noise distribution meaning that $\mathbb{P}\{\|\nabla f(x,\xi) - \nabla f(x)\|_2 > b\} \le 2\exp(-b^2/(2\sigma^2))$ for all $b > 0$.

Level of smoothness. Finally, we assume that function f has (ν, M_ν)-Hölder continuous gradients[2] on a compact set $Q \subseteq \mathbb{R}^n$ for some $\nu \in [0,1]$, $M_\nu > 0$ meaning that

$$\|\nabla f(x) - \nabla f(y)\|_2 \le M_\nu \|x - y\|_2^\nu \quad \forall x, y \in Q. \tag{3}$$

When $\nu = 1$ inequality (3) implies M_1-smoothness of f, and when $\nu = 0$ we have that $\nabla f(x)$ has bounded variation which is equivalent to being uniformly bounded. Moreover, when $\nu = 0$ differentiability of f is not needed: one can assume uniform boundedness of the subgradients of f throughout the proofs. Linear regression in the case when the noise has generalized Gaussian distribution (Example 4.4 from (Chaux et al., 2007)) serves as a natural example of the situation with $\nu \in (0,1)$. Moreover, when (3) holds for $\nu = 0$ and $\nu = 1$ simultaneously then it holds for all $\nu \in [0,1]$ with $M_\nu \le M_0^{1-\nu} M_1^\nu$ (Nesterov, 2015). As we show in our results, it is sufficient to assume that the set Q is the ball centered at the solution x^* of (1) with radius $\sim R_0 \ge \|x^0 - x^*\|_2$, where x^0 is a starting point of the method, i.e., *our analysis does not require* (3) *to hold on* $\mathbb{R}^n$.

In addition to inequality (3), we also assume that

$$\|\nabla f(x)\|_2 \le \left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}} \left(f(x) - f(x^*)\right)^{\frac{\nu}{1+\nu}}, \quad \forall x \in Q, \tag{4}$$

where for $\nu = 0$ we use $\left[\left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}}\right]_{\nu=0} := \lim_{\nu\to 0}\left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}} = 1$, and

$$\|\nabla f(x)\|_2^2 \le 2\left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} \left(f(x) - f(x^*)\right) + \delta^{\frac{2\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}, \quad \forall x \in Q. \tag{5}$$

As we prove in Lemmas A.4 and A.5, inequalities (4) and (5) follow from (3) when $Q = \mathbb{R}^n$. However, when $Q \ne \mathbb{R}^n$ minimal value of M_ν such that (3) holds on Q can be smaller than the minimal value of M_ν such that (4) and (5) hold (see also the discussion in Appendix B from Sadiev et al. (2023)).

[1] Our proofs work for any x^*. In particular, one can choose x^* being a projection of x^0 on the solutions set.

[2] By default, we always write "gradients", though our analysis works for non-differentiable convex functions as well (when $\nu = 0$): at any point where the gradient is now calculated, it is sufficient to use any subgradient at this point. This remark is valid for Definition 1.1 as well.

Table 1: Summary of known and new high-probability complexity bounds for solving (1) with f being **convex** and having (ν, M_ν)-Hölder continuous gradients. Columns: "Complexity" = high-probability complexity (ε – accuracy, β – confidence level, numerical constants, and logarithmic factors are omitted), "HT" = heavy-tailed noise, "UD" = unbounded domain, "CS" = bound on the variance of the stochastic gradient and Hölder continuity of the gradient is required only on the compact set. Notation: $R_0 = \|x^0 - x^*\|_2$, where x^* is the closest solution to x^0; σ^2 = upper bound on the variance (see (2)); D = diameter of the set where optimization problem is defined. The results labeled by ♣ are obtained from the convergence guarantees in expectation via Markov's inequality. Negative-power dependencies on the confidence level β are colored in red. Our results are highlighted in green.

Method	Complexity	ν	HT?	UD?	CS?
SGD (Nemirovski et al., 2009)	$\max\left\{\frac{M_0^2 D^2}{\varepsilon^2}, \frac{\sigma^2 D^2}{\varepsilon^2}\right\}$	0	✗	✗	✗
AC-SA (Ghadimi and Lan, 2012; Lan, 2012)	$\max\left\{\sqrt{\frac{M_1 R_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\}$	1	✗	✗	✗
SIGMA (Dvurechensky and Gasnikov, 2016)	$\max\left\{\frac{M_\nu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\}$	$[0,1]$	✗	✗	✗
SGD (Nemirovski et al., 2009)♣	$\max\left\{\frac{M_0^2 R_0^2}{\beta^2\varepsilon^2}, \frac{\sigma^2 R_0^2}{\beta^2\varepsilon^2}\right\}$	0	✓	✗	✗
AC-SA (Ghadimi and Lan, 2012; Lan, 2012)♣	$\max\left\{\sqrt{\frac{M_1 R_0^2}{\beta\varepsilon}}, \frac{\sigma^2 R_0^2}{\beta^2\varepsilon^2}\right\}$	1	✓	✓	✗
SIGMA (Dvurechensky and Gasnikov, 2016)♣	$\max\left\{\frac{M_\nu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}}}{\beta^{\frac{2}{1+3\nu}}\varepsilon^{\frac{2}{1+3\nu}}}, \frac{\sigma^2 R_0^2}{\beta^2\varepsilon^2}\right\}$	$[0,1]$	✓	✓	✗
clipped-SSTM (Gorbunov et al., 2020)	$\max\left\{\sqrt{\frac{M_1 R_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\}$	1	✓	✓	✗
clipped-SGD (Gorbunov et al., 2020)	$\max\left\{\frac{M_1 R_0^2}{\varepsilon}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\}$	1	✓	✓	✗
clipped-SSTM (Theorem 2.1)	$\max\left\{\frac{M_\nu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\}$	$[0,1]$	✓	✓	✓
clipped-SGD (Theorem 3.1)	$\max\left\{\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{2}{1+\nu}}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\}$	$[0,1]$	✓	✓	✓

High-probability convergence. For a given accuracy $\varepsilon > 0$ and confidence level $\beta \in (0,1)$ we are interested in finding ε-solutions of problem (1) with probability at least $1-\beta$, i.e., such $\widehat{x}$ that $\mathbb{P}\{f(\widehat{x}) - f(x^*) \leq \varepsilon\} \geq 1 - \beta$. For brevity, we will call such (in general, random) points $\widehat{x}$ as (ε, β)-solution of (1). Moreover, by the high-probability iteration/oracle complexity of a stochastic method $\mathcal{M}$ we mean a sufficient number of iterations/oracle calls (number of $\nabla f(x, \xi)$ computations) needed to guarantee that $\mathcal{M}$ returns an (ε, β)-solution of (1).

1.2 Contributions

We summarize our main contributions below.

- ◇ **The first near-optimal high-probability bounds for non-smooth problems with heavy-tailed noise.** We propose novel stepsize rules for clipped-SSTM (Gorbunov et al., 2020) to handle problems with the objective having a (ν, M_ν)-Hölder continuous gradient and derive in this setting high-probability complexity guarantees for convex stochastic optimization problems

Table 2: Summary of known and new high-probability complexity bounds for solving (1) with f being **μ-strongly convex** and having (ν, M_ν)-Hölder continuous gradients. Columns: "Complexity" = high-probability complexity (ε – accuracy, β – confidence level, numerical constants, and logarithmic factors are omitted), "HT" = heavy-tailed noise, "UD" = unbounded domain, "CS" = bound on the variance of the stochastic gradient and Hölder continuity of the gradient is required only on the compact set. Notation: $R_0 = \|x^0 - x^*\|_2$, where x^* is the closest solution to x^0; σ^2 = upper bound on the variance (see (2)). The results labeled by ♣ are obtained from the convergence guarantees in expectation via Markov's inequality. Negative-power dependencies on the confidence level β are colored in red. Our results are highlighted in green.

Method	Complexity	ν	HT?	UD?	CS?
SGD (Nemirovski et al., 2009)	$\max\left\{\frac{M_0^2}{\mu\varepsilon}, \frac{\sigma^2}{\mu\varepsilon}\right\}$	0	✗	✗	✗
AC-SA (Ghadimi and Lan, 2012; Lan, 2012)	$\max\left\{\sqrt{\frac{M_1}{\mu}}, \frac{\sigma^2}{\mu\varepsilon}\right\}$	1	✗	✗	✗
SIGMA (Dvurechensky and Gasnikov, 2016)	$\max\left\{\hat{N}, \frac{\sigma^2}{\mu\varepsilon}\right\}$, $\hat{N} = \max\left\{\left(\frac{M_\nu}{\mu R_0^{1-\nu}}\right)^{\frac{2}{1+3\nu}}, \left(\frac{M_\nu^2}{\mu^{1+\nu}\varepsilon^{1-\nu}}\right)^{\frac{1}{1+3\nu}}\right\}$	$[0,1]$	✗	✗	✗
SGD (Nemirovski et al., 2009)♣	$\max\left\{\frac{M_0^2}{\mu\beta\varepsilon}, \frac{\sigma^2}{\mu\beta\varepsilon}\right\}$	0	✓	✗	✗
AC-SA (Ghadimi and Lan, 2012; Lan, 2012)♣	$\max\left\{\sqrt{\frac{M_1}{\mu}}, \frac{\sigma^2}{\mu\beta\varepsilon}\right\}$	1	✓	✓	✗
SIGMA (Dvurechensky and Gasnikov, 2016)♣	$\max\left\{\hat{N}, \frac{\sigma^2}{\mu\hat{\varepsilon}}\right\}$, $\hat{\varepsilon} = \beta\varepsilon$, $\hat{N} = \left(\frac{M_\nu}{\mu R_0^{1-\nu}}\right)^{\frac{2}{1+3\nu}} + \left(\frac{M_\nu^2}{\mu^{1+\nu}\varepsilon^{1-\nu}}\right)^{\frac{1}{1+3\nu}}$	$[0,1]$	✓	✓	✗
R-clipped-SSTM (Gorbunov et al., 2020)	$\max\left\{\sqrt{\frac{M_1}{\mu}}, \frac{\sigma^2}{\mu\varepsilon^2}\right\}$	1	✓	✓	✗
R-clipped-SGD (Gorbunov et al., 2020)	$\max\left\{\frac{M_1}{\mu}, \frac{\sigma^2}{\mu\varepsilon^2}\right\}$	1	✓	✓	✗
R-clipped-SSTM (Theorem 2.2)	$\max\left\{\hat{N}, \frac{\sigma^2}{\mu\varepsilon}\right\}$, $\hat{N} = \max\left\{\left(\frac{M_\nu}{\mu R_0^{1-\nu}}\right)^{\frac{2}{1+3\nu}}, \left(\frac{M_\nu^2}{\mu^{1+\nu}\varepsilon^{1-\nu}}\right)^{\frac{1}{1+3\nu}}\right\}$	$[0,1]$	✓	✓	✓
R-clipped-SGD (Theorem 3.2)	$\max\left\{\frac{M_\nu^{\frac{2}{1+\nu}}}{\mu^{\frac{2}{1+\nu}} R_0^{\frac{2(1-\nu)}{1+\nu}}}, \frac{M_\nu^{\frac{2}{1+\nu}}}{\mu\varepsilon^{\frac{1-\nu}{1+\nu}}}, \frac{\sigma^2}{\mu\varepsilon}\right\}$	$[0,1]$	✓	✓	✓

without using the "light tails" assumption, i.e., we prove that our version of clipped-SSTM has

$$\mathcal{O}\left(\max\left\{D\ln^{\frac{2(1+\nu)}{1+3\nu}}\frac{D}{\beta}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\ln\frac{D}{\beta}\right\}\right), \quad D = \frac{M_\nu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}}$$

high-probability complexity. Unlike all previous high-probability complexity results in this setup with $\nu < 1$ (see Table 1), our result depends only logarithmically on the confidence level β that is highly important when β is small. Moreover, up to the difference in logarithmic factors, the derived complexity guarantees meet the known lower bounds (Guzmán and Nemirovski, 2015; Lan, 2012) obtained for problems with light-tailed noise. In particular, when $\nu = 1$, we recover the accelerated convergence rate (Lan, 2012; Nesterov, 1983). That is, neglecting the logarithmic factors, our results are unimprovable and, surprisingly, coincide with the best-known results in the "light-tailed case".

- ◇ **New high-probability bounds for clipped-SGD.** We derive the first high-probability complexity bounds for clipped-SGD when the objective function is convex with (ν, M_ν)-Hölder

continuous gradient and the noise is heavy tailed., i.e., we derive

$$\mathcal{O}\left(\max\left\{D^2, \max\left\{D^{1+\nu}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\}\ln\frac{D^2+D^{1+\nu}}{\beta}\right\}\right), \quad D = \frac{M_\nu^{\frac{1}{1+\nu}} R_0}{\varepsilon^{\frac{1}{1+\nu}}}$$

the high-probability complexity bound. Interestingly, when $\nu = 0$, the derived bound for clipped-SGD has better dependence on the logarithms than the corresponding one for clipped-SSTM. Moreover, neglecting the dependence on ε under the logarithm, our bound for clipped-SGD has the same dependence on the confidence level as the tightest known result in this case under the "light tails" assumption (Guigues et al., 2017).

- ◇ **Extensions to the strongly convex case.** Using the restarts technique we extend the obtained results for clipped-SSTM and clipped-SGD to the strongly convex case (see Table 2). As in the convex case, the obtained results are superior to all previously known results in the general setup we consider. Moreover, the results derived for clipped-SSTM are optimal up to logarithmic factors (Guzmán and Nemirovski, 2015; Lan, 2012).

- ◇ **Generality of the results.** As one of the key contributions of this work, we emphasize that in our theoretical results it is sufficient to assume boundedness of the variance of the stochastic gradient (22) and Hölder continuity of the gradients of f only on the ball with radius $\sim R_0 = \|x^0 - x^*\|_2$ and centered at the closest to the starting point solution of the problem. This makes our results applicable to a much wider class of problems than functions with Hölder continuous gradients on $\mathbb{R}^n$, e.g., our analysis works even for polynomially growing objectives. Moreover, this feature of our analysis allows us to consider strongly convex functions. Indeed, the class of strongly convex functions on $\mathbb{R}^n$ with Hölder continuous gradients on $\mathbb{R}^n$ with $\nu < 1$ is empty. Therefore, it is crucial to assume Hölder continuity of gradients only on a bounded set[3]. Next, we do not require the variance of the stochastic gradient to be uniformly bounded on the whole space, e.g., we allow the variance at point x to grow when $\|x - x^*\|_2 \to \infty$. We emphasize that even for smooth problems ($\nu = 1$) all previous works in the area rely on the uniform boundedness of the variance on the whole space (see Tables 1 and 2). Next, in the works focusing on the "light tails" case, the uniform boundedness of sub-Gaussian variance and Hölder continuity of the gradients are also assumed on $\mathbb{R}^n$. All of these facts emphasize the generality of our results.

- ◇ **Experiments.** To test the performance of the considered methods, we conduct several numerical experiments on image classification and NLP tasks and observe that 1) clipped-SSTM and clipped-SGD show comparable performance with SGD on the image classification task, when the noise distribution is almost sub-Gaussian, 2) converge much faster than SGD on the NLP task, when the noise distribution is heavy-tailed, and 3) clipped-SSTM achieves a comparable performance with Adam on the NLP task enjoying both the best known theoretical guarantees and good practical performance. We also compare clipped-SSTM, clipped-SGD, SGD, and Adam on solving the convex problem, corresponding to the linear regression with the noise having generalized Gaussian distribution. Our codes are publicly available: https://github.com/ClippedStochasticMethods/clipped-SSTM.

[3]It is also worth mentioning that some functions have Hölder continuous gradients for multiple ν simultaneously (Nesterov, 2015). Therefore, if constants M_ν are available, one can choose the best ν in terms of the iteration/oracle complexity of a method.

1.3 Related Work

Light-tailed noise. The theory of high-probability complexity bounds for convex stochastic optimization with light-tailed noise is well-developed. Lower bounds and optimal methods for the problems with (ν, M_ν)-Hölder continuous gradients are obtained in (Nemirovski et al., 2009) for $\nu = 0$, and in (Ghadimi and Lan, 2012) for $\nu = 1$. Up to the logarithmic dependencies, these high-probability convergence bounds coincide with the corresponding results for the convergence in expectation (see first two rows of Table 1). While not being directly derived in the literature, the lower bound for the case when $\nu \in (0, 1)$ can be obtained as a combination of lower bounds in the deterministic (Guzmán and Nemirovski, 2015; Nemirovski and Yudin, 1983) and smooth stochastic settings (Ghadimi and Lan, 2012). The corresponding optimal methods are analyzed in (Devolder, 2013; Dvurechensky and Gasnikov, 2016) based on the concept of inexact oracle.

Heavy-tailed noise. Unlike in the "light-tailed" case, the first theoretical guarantees with reasonable dependence on both the accuracy ε and the confidence level β appeared just recently. In (Nazin et al., 2019), the first such results without the kind of acceleration appearing in (Nesterov, 1983) were derived for Mirror Descent with special truncation technique for smooth ($\nu = 1$) convex problems on a bounded domain, and then were accelerated and extended in (Gorbunov et al., 2020). For strongly convex problems, the first accelerated high-probability convergence guarantees were obtained in (Davis et al., 2021) for the special method called proxBoost requiring the solving of nontrivial auxiliary problems at each iteration. These bounds were tightened in (Gorbunov et al., 2020).

In contrast, for the case when $\nu < 1$ and, in particular, when $\nu = 0$ the best-known high-probability complexity bounds suffer from the negative-power dependence on the confidence level β, i.e., have a factor $1/\beta^\alpha$ for some $\alpha > 0$, that affects the convergence rate dramatically for small enough β. Without additional assumptions on the tails these results are obtained via Markov's inequality $\mathbb{P}\{f(x) - f(x^*) > \varepsilon\} < \mathbb{E}[f(x)-f(x^*)]/\varepsilon$ from the guarantees for the convergence in expectation to the accuracy $\varepsilon\beta$, see the results labeled by ♣ in Table 1. Under an additional assumption on noise tails that $\mathbb{P}\{\|\nabla f(x,\xi) - \nabla f(x)\|_2^2 > s\sigma^2\} = \mathcal{O}(s^{-\alpha})$ for $\alpha > 2$ these results can be tightened (Gasnikov et al., 2015) when $\nu = 0$ as $\sim \max\left\{\ln(\beta^{-1})/\varepsilon^2, (1/\beta\varepsilon^\alpha)^{2/(3\alpha-2)}\right\}$ without removing the negative-power dependence on the confidence level β. Different stepsize policies allow to change the last term in max to $\beta^{-\frac{1}{2\alpha-1}}\varepsilon^{-\frac{2\alpha}{2\alpha-1}}$ without removing the negative-power dependence on β.

Comparison with (Gorbunov et al., 2020). Although our results and proof technique are based on the ones proposed in (Gorbunov et al., 2020), our work extends and significantly differs from (Gorbunov et al., 2020). First of all, we consider problems with Hölder continuous gradients, while the authors of (Gorbunov et al., 2020) obtain their results only for the smooth functions. To derive a proper generalization of the results from (Gorbunov et al., 2020), we propose different stepsizes for clipped-SSTM and clipped-SGD and we also modify the proofs significantly to circumvent the additional issues arising due to the partial smoothness of the problem, especially in the part where we prove high-probability bound for the norm of the gradient (see the derivation of inequality (36)). Since this part is one of the most important ones in the proof, this fact highlights the difference between two approaches. Moreover, (Gorbunov et al., 2020) assume that the variance is uniformly upper bounded and the gradient is Lipschitz-continuous on $\mathbb{R}^n$, while our analysis relies on much weaker assumptions that (2) and (3) hold on a ball around the solution, i.e., on a compact set. Thus, our results are proven for a much wider class of problems, including ones with polynomially growing

objective function/variance $\mathbb{E}_\xi\left[\|\nabla f(x,\xi)-\nabla f(x)\|_2^2\right]$ when $\|x-x^*\|_2\to\infty$.

Gradient clipping. The methods based on gradient clipping (Pascanu et al., 2013) and normalization (Hazan et al., 2015) are popular in different machine learning and deep learning tasks due to their robustness in practice to the noise in the stochastic gradients and rapid changes of the objective function (Goodfellow et al., 2016). In (Mai and Johansson, 2021; Zhang et al., 2020a), clipped-GD and clipped-SGD are theoretically studied in applications to non-smooth problems with an objective that can grow polynomially fast when $\|x-x^*\|_2\to\infty$ showing the superiority of gradient clipping methods to the methods without clipping. The results from (Zhang et al., 2020a) are obtained for non-convex problems with almost surely bounded noise, and in (Mai and Johansson, 2021), the authors derive the stability and expectation convergence guarantees for strongly convex objectives under an assumption that the central p-th moment of the stochastic gradient is bounded for $p\geq 2$. Since (Mai and Johansson, 2021) do not provide convergence guarantees with explicit dependencies on all important parameters of the problem, it complicates direct comparison with our results. Nevertheless, convergence guarantees from (Mai and Johansson, 2021) are sub-linear and are given for the convergence in expectation, and, as a consequence, the corresponding high-probability convergence results obtained via the Markov's inequality also suffer from negative-power dependence on the confidence level. Next, the authors of (Zhang et al., 2020b) establish several expectation convergence guarantees for clipped-SGD and prove their optimality in the non-convex case under the assumption that the central α-moment of the stochastic gradient is uniformly bounded, where $\alpha\in(1,2]$. It turns out that clipped-SGD is able to converge even when $\alpha<2$, whereas vanilla SGD can diverge in this setting.

1.4 Paper Organization

We present and discuss the simplified versions of our main results on clipped-SSTM (Theorems 2.1 and 2.2) and clipped-SGD (Theorems 3.1 and 3.2) in Sections 2 and 3 respectively. The detailed statements of the main results, complete proofs, and the corollaries for unit batch sizes (Corollaries 4.1 and 5.1) are given in Sections 4 and 5. Finally, Section 6 contains the results of our numerical experiments. In Appendix A, we give some useful auxiliary lemmas and prove few technical results. In Appendix B, we provide a detailed description of the setup for numerical experiments and extra numerical results.

2 Clipped Stochastic Similar Triangles Method

In this section, we propose a novel variation of Clipped Stochastic Similar Triangles Method (Gorbunov et al., 2020) adjusted to the class of objectives with Hölder continuous gradients (clipped-SSTM, see Algorithm 1).

The method is based on the clipping of the stochastic gradients:

$$\text{clip}(\nabla f(x,\boldsymbol{\xi}),\lambda)=\min\left\{1,\frac{\lambda}{\|\nabla f(x,\boldsymbol{\xi})\|_2}\right\}\nabla f(x,\boldsymbol{\xi}) \tag{6}$$

where $\nabla f(x,\boldsymbol{\xi})=\frac{1}{m}\sum_{i=1}^m\nabla f(x,\xi_i)$ is a mini-batched stochastic gradient and for shortness we denote by $\boldsymbol{\xi}$ the collection $\{\xi_i\}_{i=1}^m$ of samples. Gradient clipping ensures that the resulting vector has a norm bounded by the clipping level λ. Since the clipped stochastic gradient cannot have an

arbitrary large norm, the clipping helps to avoid unstable behavior of the method when the noise is heavy-tailed, and the clipping level λ is properly adjusted.

However, unlike the stochastic gradient, the clipped stochastic gradient is a *biased* estimate of $\nabla f(x)$: the smaller the clipping level, the larger the bias. The biasedness of the clipped stochastic gradient complicates the analysis of the method. On the other hand, to circumvent the negative effect of the heavy-tailed noise on the high-probability convergence, one should choose λ to be not too large. Therefore, the question on the appropriate choice of the clipping level is highly non-trivial.

Fortunately, there exists a simple but insightful observation that helps us to obtain the right formula for the clipping level λ_k at iteration k in clipped-SSTM: if λ_k is chosen in such a way that $\|\nabla f(x^k)\|_2 \leq \lambda_k/2$ with high probability, then for the realizations $\nabla f(x^{k+1}, \boldsymbol{\xi}^k)$ of the mini-batched stochastic gradient such that $\|\nabla f(x^{k+1}, \boldsymbol{\xi}^k) - \nabla f(x^{k+1})\|_2 \leq \lambda_k/2$ the clipping is an identity operator. Next, if the probability mass of such realizations is big enough, then the bias of the clipped stochastic gradient is properly bounded, which helps derive the needed convergence guarantees. It turns out that the choice $\lambda_k \sim 1/\alpha_k$ ensures the method convergence with the needed rate and high enough probability.

Algorithm 1 Clipped Stochastic Similar Triangles Method (clipped-SSTM): case $\nu \in [0, 1]$

Input: starting point x^0, number of iterations N, batch sizes $\{m_k\}_{k=1}^N$, stepsize parameter α, clipping parameter B, Hölder exponent $\nu \in [0, 1]$.

Set $A_0 = \alpha_0 = 0$, $y^0 = z^0 = x^0$
for $k = 0, 1, \ldots, N-1$ **do**
  Set $\alpha_{k+1} = \alpha(k+1)^{\frac{2\nu}{1+\nu}}$, $A_{k+1} = A_k + \alpha_{k+1}$, $\lambda_{k+1} = \frac{B}{\alpha_{k+1}}$
  $x^{k+1} = (A_k y^k + \alpha_{k+1} z^k)/A_{k+1}$
  Draw mini-batch m_k of 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 (6)
  $z^{k+1} = z^k - \alpha_{k+1} \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)$
  $y^{k+1} = (A_k y^k + \alpha_{k+1} z^{k+1})/A_{k+1}$
end for

Output: y^N

Guided by this observation, we derive the precise expressions for all the parameters of clipped-SSTM and derive high-probability complexity bounds for the method. Below, we provide a simplified version of the main result for clipped-SSTM in the convex case. The complete formulation and the full proof of the theorem are deferred to Section 4.1 (see Theorem 4.1).

Theorem 2.1 (Simplified version of Theorem 4.1)**.** *Assume that function f is convex, its stochastic gradient and its gradient satisfy (2) and (3) respectively with $\sigma > 0$, $\nu \in [0, 1]$, $M_\nu > 0$ on $Q = B_{3R_0}(x^*) = \{x \in \mathbb{R}^n \mid \|x - x^*\|_2 \leq 3R_0\}$, where $R_0 \geq \|x^0 - x^*\|_2$. Then there exists such a choice of parameters that* clipped-SSTM *achieves $f(y^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ after $\mathcal{O}\left(D \ln^{\frac{2(1+\nu)}{1+3\nu}} \frac{D}{\beta}\right)$ iterations with $D = \frac{M_\nu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}}$ and requires*

$$\mathcal{O}\left(\max\left\{D \ln^{\frac{2(1+\nu)}{1+3\nu}} \frac{D}{\beta}, \frac{\sigma^2 R_0^2}{\varepsilon^2} \ln \frac{D}{\beta}\right\}\right) \text{ oracle calls.} \tag{7}$$

The obtained result has only logarithmic dependence on the confidence level β and optimal dependence on the accuracy ε up to logarithmic factors (Guzmán and Nemirovski, 2015; Lan, 2012) and matches the state-of-the-art results in the light-tailed case (Dvurechensky and Gasnikov, 2016; Ghadimi and Lan, 2012; Nemirovski et al., 2009) for all $\nu \in [0, 1]$. Moreover, the complexity bounds from (Dvurechensky and Gasnikov, 2016; Ghadimi and Lan, 2012; Nemirovski et al., 2009) are proportional to $\mathcal{O}\left(\ln^2 \frac{1}{\beta}\right)$ (neglecting the dependence on M_ν and R_0), while our bound has better dependence on the power of the logarithm when $\nu > 0$. In particular, when $\nu = 1$, our bound is proportional to $\mathcal{O}\left(\ln \frac{1}{\sqrt{\varepsilon}\beta}\right)$. When β is small enough (of the same order with ε or smaller), our logarithmic factor is much smaller than $\mathcal{O}\left(\ln^2 \frac{1}{\beta}\right)$.

Next, we emphasize that our result does not require f to have (ν, M_ν)-Hölder continuous gradient and the variance of the stochastic gradient to be uniformly bounded *on the whole space*. To achieve this, we prove that for the proposed choice of parameters the iterates of clipped-SSTM stay inside the ball $B_{3R_0} = \{x \in \mathbb{R}^n \mid \|x - x^*\|_2 \leq 3R_0\}$ with probability at least $1 - \beta$, and, as a consequence, it is sufficient to assume that (2) and (3) hold only inside this ball. In particular, this means that the better starting point leads not only to the reduction of R_0, but also it can reduce M_ν and σ. Moreover, our result is applicable to a much wider class of functions than the standard class of functions with Hölder continuous gradients in $\mathbb{R}^n$, e.g., to the problems with polynomial growth in both the gradient and the variance of the stochastic estimator.

For the strongly convex problems, we consider a restarted version of Algorithm 1 (R-clipped-SSTM, see Algorithm 2) and derive high-probability complexity result for this version. Below we provide a simplified version of the result. The complete formulation and the full proof of the theorem are deferred to Section 4.2 (see Theorem 4.2).

Theorem 2.2 (Simplified version of Theorem 4.2)**.** *Assume that function f is μ-strongly convex, its stochastic gradient and its gradient satisfy (2) and (3) respectively with $\sigma > 0$, $\nu \in [0, 1]$, $M_\nu > 0$ on $Q = B_{3R_0}(x^*) = \{x \in \mathbb{R}^n \mid \|x - x^*\|_2 \leq 3R_0\}$, where $R_0 \geq \|x^0 - x^*\|_2$. Then there exists such a choice of parameters that* R-clipped-SSTM *achieves $f(\hat{x}^\tau) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ after*

$$\hat{N} = \mathcal{O}\left(D \ln^{\frac{2(1+\nu)}{1+3\nu}} \frac{D}{\beta}\right), \; D = \max\left\{\left(\frac{M_\nu}{\mu R_0^{1-\nu}}\right)^{\frac{2}{1+3\nu}} \ln \frac{\mu R_0^2}{\varepsilon}, \left(\frac{M_\nu^2}{\mu^{1+\nu}\varepsilon^{1-\nu}}\right)^{\frac{1}{1+3\nu}}\right\} \tag{8}$$

iterations of Algorithm 1 in total and requires

$$\mathcal{O}\left(\max\left\{D \ln^{\frac{2(1+\nu)}{1+3\nu}} \frac{D}{\beta}, \frac{\sigma^2}{\mu\varepsilon} \ln \frac{D}{\beta}\right\}\right) \text{ oracle calls.} \tag{9}$$

Again, the obtained result has only logarithmic dependence on the confidence level β and, as our result in the convex case, it has optimal dependence on the accuracy ε up to logarithmic factors depending on β (Guzmán and Nemirovski, 2015; Lan, 2012) for all $\nu \in [0, 1]$.

3 SGD with Clipping

In this section, we present a new variant of clipped-SGD (Pascanu et al., 2013) properly adjusted to the class of objectives with (ν, M_ν)-Hölder continuous gradients (see Algorithm 3).

We emphasize that as for clipped-SSTM we use clipping level λ inversely proportional to the stepsize α. Below, we provide a simplified version of the main result for clipped-SGD in the convex

Algorithm 2 Restarted clipped-SSTM (R-clipped-SSTM): case $\nu \in [0,1]$

Input: starting point x^0, number of restarts τ, number of steps of clipped-SSTM in restarts $\{N_t\}_{t=1}^{\tau}$, batch sizes $\{m_k^1\}_{k=1}^{N_1-1}, \{m_k^2\}_{k=1}^{N_2-1}, \ldots, \{m_k^\tau\}_{k=1}^{N_\tau-1}$, stepsize parameters $\{\alpha^t\}_{t=1}^{\tau}$, clipping parameters $\{B_t\}_{t=1}^{\tau}$, Hölder exponent $\nu \in [0,1]$.

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

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

case. The complete formulation and the full proof of the theorem are deferred to Section 5.1 (see Theorem 5.1).

Theorem 3.1 (Simplified version of Theorem 5.1)**.** *Assume that function f is convex, its stochastic gradient and its gradient satisfy (2) and (3) respectively with $\sigma > 0$, $\nu \in [0,1]$, $M_\nu > 0$ on $Q = B_{7R_0}(x^*) = \{x \in \mathbb{R}^n \mid \|x - x^*\|_2 \leq 7R_0\}$, where $R_0 \geq \|x^0 - x^*\|_2$. Then there exists such a choice of parameters that* clipped-SGD *achieves $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ after*

$$\mathcal{O}\left(\max\left\{D^2, D^{1+\nu}\ln\frac{D^2 + D^{1+\nu}}{\beta}\right\}\right), \quad D = \frac{M_\nu^{\frac{1}{1+\nu}} R_0}{\varepsilon^{\frac{1}{1+\nu}}} \tag{10}$$

iterations and requires

$$\mathcal{O}\left(\max\left\{D^2, \max\left\{D^{1+\nu}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\}\ln\frac{D^2 + D^{1+\nu}}{\beta}\right\}\right) \text{ oracle calls.} \tag{11}$$

Algorithm 3 Clipped Stochastic Gradient Descent (clipped-SGD): case $\nu \in [0,1]$

Input: starting point x^0, number of iterations N, batch size m, stepsize γ, clipping parameter $B > 0$.

1: **for** $k = 0, 1, \ldots, N-1$ **do**
2: Draw mini-batch of m fresh i.i.d. samples $\xi_1^k, \ldots, \xi_m^k$ and compute $\nabla f(x^{k+1}, \boldsymbol{\xi}^k) = \frac{1}{m}\sum_{i=1}^m \nabla f(x^{k+1}, \xi_i^k)$
3: Compute $\widetilde{\nabla} f(x^k, \boldsymbol{\xi}^k) = \text{clip}(\nabla f(x^k, \boldsymbol{\xi}^k), \lambda)$ using (6) with $\lambda = B/\gamma$
4: $x^{k+1} = x^k - \gamma \widetilde{\nabla} f(x^k, \boldsymbol{\xi}^k)$
5: **end for**

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

As all our results in the paper, this result for clipped-SGD has two important features: 1) the dependence on the confidence level β is logarithmic and 2) Hölder continuity and uniformly bounded variance assumptions are required only on the ball $B_{7R_0}(x^*)$ centered at the solution. Moreover, up to the difference in the expressions under the logarithm, the dependence on ε in the result for clipped-SGD is the same as in the tightest known results for non-accelerated SGD-type methods

(Devolder, 2013; Guigues et al., 2017). Finally, we emphasize that for $\nu < 1$ the logarithmic factors appearing in the complexity bound for clipped-SSTM are worse than the corresponding factor in the complexity bound for clipped-SGD. Therefore, clipped-SGD has the best-known high-probability complexity results in the case when $\nu = 0$ and f is convex. Furthermore, when $\nu = 0$, our result has $\mathcal{O}(\ln \frac{1}{\varepsilon^2\beta})$ logarithmic factor, while the best-known high-probability results under "light tails" assumption are proportional to $\mathcal{O}(\ln^2 \frac{1}{\beta})$ (Dvurechensky and Gasnikov, 2016; Nemirovski et al., 2009). When β is small enough (of the same order with ε or smaller), our logarithmic factor is much smaller than $\mathcal{O}\left(\ln^2 \frac{1}{\beta}\right)$.

For the strongly convex problems, we consider a restarted version of Algorithm 3 (R-clipped-SGD, see Algorithm 4) and derive high-probability complexity result for this version. Below we provide

Algorithm 4 Restarted clipped-SGD (R-clipped-SGD): case $\nu \in [0, 1]$

Input: starting point x^0, number of restarts τ, number of steps of clipped-SGD in restarts $\{N_t\}_{t=1}^{\tau}$, batch sizes $\{m_t\}_{k=1}^{\tau}$, stepsizes $\{\gamma_t\}_{t=1}^{\tau}$, clipping parameters $\{B_t\}_{t=1}^{\tau}$

1: $\hat{x}^0 = x^0$
2: **for** $t = 1, \ldots, \tau$ **do**
3: Run clipped-SGD (Algorithm 3) for N_t iterations with batch size m_t, stepsize γ_t, clipping parameter B_t, and starting point $\hat{x}^{t-1}$. Define the output of clipped-SGD by $\hat{x}^t$.
4: **end for**

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

a simplified version of the result. The complete formulation and the full proof of the theorem are deferred to Section 5.2 (see Theorem 5.2).

Theorem 3.2 (Simplified version of Theorem 5.2)**.** *Assume that function f is μ-strongly convex, its stochastic gradient and its gradient satisfy (2) and (3) respectively with $\sigma > 0$, $\nu \in [0, 1]$, $M_\nu > 0$ on $Q = B_{7R_0}(x^*) = \{x \in \mathbb{R}^n \mid \|x - x^*\|_2 \leq 7R_0\}$, where $R_0 \geq \|x^0 - x^*\|_2$. Then there exists such a choice of parameters that* R-clipped-SGD *achieves $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ after*

$$\mathcal{O}\left(\max\left\{D_1^{\frac{2}{1+\nu}} \ln \frac{\mu R_0^2}{\varepsilon}, D_2^{\frac{2}{1+\nu}}, \max\left\{D_1 \ln \frac{\mu R_0^2}{\varepsilon}, D_2\right\} \ln \frac{D}{\beta}\right\}\right)$$

iterations of Algorithm 3 in total and requires

$$\mathcal{O}\left(\max\left\{D_1^{\frac{2}{1+\nu}} \ln \frac{\mu R_0^2}{\varepsilon}, D_2^{\frac{2}{1+\nu}}, \max\left\{D_1 \ln \frac{\mu R_0^2}{\varepsilon}, D_2, \frac{\sigma^2}{\mu\varepsilon}\right\} \ln \frac{D}{\beta}\right\}\right)$$

oracle calls, where

$$D_1 = \frac{M_\nu}{\mu R_0^{1-\nu}}, \quad D_2 = \frac{M_\nu}{\mu^{\frac{1+\nu}{2}} \varepsilon^{\frac{1-\nu}{2}}}, \quad D = (D_1^{\frac{2}{1+\nu}} + D_1) \ln \frac{\mu R_0^2}{\varepsilon} + D_2 + D_2^{\frac{2}{1+\nu}}.$$

As in the convex case, for $\nu < 1$, the log-factors appearing in the complexity bound for R-clipped-SSTM are worse than the corresponding factor in the bound for R-clipped-SGD. Thus, R-clipped-SGD has the best-known high-probability complexity results for strongly convex f and $\nu = 0$.

4 Clipped Similar Triangles Method: Missing Details and Proofs

4.1 Convergence in the Convex Case

In this section, we provide the full proof of Theorem 2.1 together with a complete statement of the result.

4.1.1 Two lemmas

The analysis of clipped-SSTM consists of three main steps. The first one is an "optimization lemma" – a modification of a standard lemma for Similar Triangles Method (see (Gasnikov and Nesterov, 2018) and Lemma F.4 from (Gorbunov et al., 2020)). This result helps to estimate the progress of the method after N iterations.

Lemma 4.1. *Let f be a convex function with a minimum at some[4] point x^*, its gradient be (ν, M_ν)-Hölder continuous on a ball $B_{3R_0}(x^*)$, where $R_0 \geq \|x^0 - x^*\|_2$, and let stepsize parameter α have the form $\alpha = \frac{(\varepsilon/2)^{\frac{1-\nu}{1+\nu}}}{2^{\frac{2\nu}{1+\nu}} a M_\nu^{\frac{2}{1+\nu}}}$, where $a \geq 1$. If $x^k, y^k, z^k \in B_{3R_0}(x^*)$ for all $k = 0, 1, \ldots, N$, $N \geq 0$, then after N iterations of clipped-SSTM for all $z \in \mathbb{R}^n$ we have*

$$A_N\left(f(y^N) - f(z)\right) \leq \frac{1}{2}\|z^0 - z\|_2^2 - \frac{1}{2}\|z^N - z\|_2^2 + \sum_{k=0}^{N-1} \alpha_{k+1}\left\langle \theta_{k+1}, z - z^k\right\rangle$$
$$+ \sum_{k=0}^{N-1} \alpha_{k+1}^2 \|\theta_{k+1}\|_2^2 + \sum_{k=0}^{N-1} \alpha_{k+1}^2 \left\langle \theta_{k+1}, \nabla f(x^{k+1})\right\rangle + \frac{A_N \varepsilon}{4}, \tag{12}$$

$$\theta_{k+1} \overset{def}{=} \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) - \nabla f(x^{k+1}). \tag{13}$$

Proof. Consider an arbitrary $k \in \{0, 1, \ldots, N-1\}$. Using $z^{k+1} = z^k - \alpha_{k+1}\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)$ we get that for all $z \in \mathbb{R}^n$

$$\begin{aligned}
& \alpha_{k+1}\left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z^k - z\right\rangle \\
= \; & \alpha_{k+1}\left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z^k - z^{k+1}\right\rangle + \alpha_{k+1}\left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z^{k+1} - z\right\rangle \\
= \; & \alpha_{k+1}\left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z^k - z^{k+1}\right\rangle + \left\langle z^{k+1} - z^k, z - z^{k+1}\right\rangle \\
\overset{(86)}{=} \; & \alpha_{k+1}\left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z^k - z^{k+1}\right\rangle - \frac{1}{2}\|z^k - z^{k+1}\|_2^2 \\
& + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2.
\end{aligned} \tag{14}$$

Next, we notice that

$$\begin{aligned}
y^{k+1} &= \frac{A_k y^k + \alpha_{k+1} z^{k+1}}{A_{k+1}} = \frac{A_k y^k + \alpha_{k+1} z^k}{A_{k+1}} + \frac{\alpha_{k+1}}{A_{k+1}}\left(z^{k+1} - z^k\right) \\
&= x^{k+1} + \frac{\alpha_{k+1}}{A_{k+1}}\left(z^{k+1} - z^k\right)
\end{aligned} \tag{15}$$

[4] Our proofs are valid for any solution x^* and, for example, one can take as x^* the closest solution to the starting point x^0.

implying

$$
\begin{aligned}
\alpha_{k+1} & \left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z^k - z \right\rangle \\
\overset{(13),(14)}{\leq} & \alpha_{k+1} \left\langle \nabla f(x^{k+1}), z^k - z^{k+1} \right\rangle - \frac{1}{2}\|z^k - z^{k+1}\|_2^2 \\
& + \alpha_{k+1} \left\langle \theta_{k+1}, z^k - z^{k+1} \right\rangle + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 \\
\overset{(15)}{=} & A_{k+1} \left\langle \nabla f(x^{k+1}), x^{k+1} - y^{k+1} \right\rangle - \frac{1}{2}\|z^k - z^{k+1}\|_2^2 \\
& + \alpha_{k+1} \left\langle \theta_{k+1}, z^k - z^{k+1} \right\rangle + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 \\
\overset{(88)}{\leq} & A_{k+1} \left(f(x^{k+1}) - f(y^{k+1}) \right) + \frac{A_{k+1} L_{k+1}}{2} \|x^{k+1} - y^{k+1}\|_2^2 \\
& + \frac{\alpha_{k+1}\varepsilon}{4} - \frac{1}{2}\|z^k - z^{k+1}\|_2^2 + \alpha_{k+1} \left\langle \theta_{k+1}, z^k - z^{k+1} \right\rangle \\
& + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 \\
\overset{(15)}{=} & A_{k+1} \left(f(x^{k+1}) - f(y^{k+1}) \right) + \frac{1}{2} \left(\frac{\alpha_{k+1}^2 L_{k+1}}{A_{k+1}} - 1 \right) \|z^k - z^{k+1}\|_2^2 \\
& + \alpha_{k+1} \left\langle \theta_{k+1}, z^k - z^{k+1} \right\rangle + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 + \frac{\alpha_{k+1}\varepsilon}{4},
\end{aligned}
$$

where in the third inequality we used $x^{k+1}, y^{k+1} \in B_{3R_0}(x^*)$ and (88) with $\delta = \frac{\alpha_{k+1}}{2A_{k+1}}\varepsilon$ and $L(\delta, \nu) = L_{k+1} = \left(\frac{2A_{k+1}}{\varepsilon \alpha_{k+1}}\right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}$. Since $A_{k+1} \geq a L_{k+1} \alpha_{k+1}^2$ (Lemma A.3) and $a \geq 1$ we can continue our derivations as follows:

$$
\begin{aligned}
\alpha_{k+1} & \left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z^k - z \right\rangle \\
\leq & A_{k+1} \left(f(x^{k+1}) - f(y^{k+1}) \right) + \alpha_{k+1} \left\langle \theta_{k+1}, z^k - z^{k+1} \right\rangle \\
& + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 + \frac{\alpha_{k+1}\varepsilon}{4}.
\end{aligned} \tag{16}
$$

Next, due to the convexity of f, we have

$$
\begin{aligned}
\left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), y^k - x^{k+1} \right\rangle & \overset{(13)}{=} \left\langle \nabla f(x^{k+1}), y^k - x^{k+1} \right\rangle + \left\langle \theta_{k+1}, y^k - x^{k+1} \right\rangle \\
& \leq f(y^k) - f(x^{k+1}) + \left\langle \theta_{k+1}, y^k - x^{k+1} \right\rangle.
\end{aligned} \tag{17}
$$

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

$$
\alpha_{k+1} \left(x^{k+1} - z^k \right) = A_k \left(y^k - x^{k+1} \right) \tag{18}
$$

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

$$
\begin{aligned}
\alpha_{k+1} & \left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), x^{k+1} - z \right\rangle \\
= & \alpha_{k+1} \left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), x^{k+1} - z^k \right\rangle + \alpha_{k+1} \left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z^k - z \right\rangle \\
\overset{(18)}{=} & A_k \left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), y^k - x^{k+1} \right\rangle + \alpha_{k+1} \left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z^k - z \right\rangle \\
\overset{(17),(16)}{\leq} & A_k \left(f(y^k) - f(x^{k+1}) \right) + A_k \left\langle \theta_{k+1}, y^k - x^{k+1} \right\rangle \\
& + A_{k+1} \left(f(x^{k+1}) - f(y^{k+1}) \right) + \alpha_{k+1} \left\langle \theta_{k+1}, z^k - z^{k+1} \right\rangle \\
& + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 + \frac{\alpha_{k+1}\varepsilon}{4} \\
\overset{(18)}{=} & A_k f(y^k) - A_{k+1} f(y^{k+1}) + \alpha_{k+1} \left\langle \theta_{k+1}, x^{k+1} - z^k \right\rangle \\
& + \alpha_{k+1} f(x^{k+1}) + \alpha_{k+1} \left\langle \theta_{k+1}, z^k - z^{k+1} \right\rangle \\
& + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 + \frac{\alpha_{k+1}\varepsilon}{4} \\
= & A_k f(y^k) - A_{k+1} f(y^{k+1}) + \alpha_{k+1} f(x^{k+1}) \\
& + \alpha_{k+1} \left\langle \theta_{k+1}, x^{k+1} - z^{k+1} \right\rangle \\
& + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 + \frac{\alpha_{k+1}\varepsilon}{4}.
\end{aligned}
$$

Rearranging the terms, we get

$$
\begin{aligned}
A_{k+1} & f(y^{k+1}) - A_k f(y^k) \\
\leq & \alpha_{k+1} \left(f(x^{k+1}) + \left\langle \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k), z - x^{k+1} \right\rangle \right) + \frac{1}{2}\|z^k - z\|_2^2 \\
& - \frac{1}{2}\|z^{k+1} - z\|_2^2 + \alpha_{k+1} \left\langle \theta_{k+1}, x^{k+1} - z^{k+1} \right\rangle + \frac{\alpha_{k+1}\varepsilon}{4} \\
\overset{(13)}{=} & \alpha_{k+1} \left(f(x^{k+1}) + \left\langle \nabla f(x^{k+1}), z - x^{k+1} \right\rangle \right) \\
& + \alpha_{k+1} \left\langle \theta_{k+1}, z - x^{k+1} \right\rangle + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 \\
& + \alpha_{k+1} \left\langle \theta_{k+1}, x^{k+1} - z^{k+1} \right\rangle + \frac{\alpha_{k+1}\varepsilon}{4} \\
\leq & \alpha_{k+1} f(z) + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 + \alpha_{k+1} \left\langle \theta_{k+1}, z - z^{k+1} \right\rangle + \frac{\alpha_{k+1}\varepsilon}{4}
\end{aligned}
$$

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 $k = 0, 1, \ldots, N-1$ and get

$$
\begin{aligned}
A_N \quad & f(y^N) \\
\leq \quad & A_N f(z) + \frac{1}{2}\|z^0 - z\|_2^2 - \frac{1}{2}\|z^N - z\|_2^2 + \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \theta_{k+1}, z - z^{k+1} \right\rangle + \frac{A_N \varepsilon}{4} \\
= \quad & A_N f(z) + \frac{1}{2}\|z^0 - z\|_2^2 - \frac{1}{2}\|z^N - z\|_2^2 + \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \theta_{k+1}, z - z^{k} \right\rangle \\
& + \sum_{k=0}^{N-1} \alpha_{k+1}^2 \left\langle \theta_{k+1}, \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) \right\rangle + \frac{A_N \varepsilon}{4} \\
\overset{(13)}{=} \quad & A_N f(z) + \frac{1}{2}\|z^0 - z\|_2^2 - \frac{1}{2}\|z^N - z\|_2^2 + \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \theta_{k+1}, z - z^{k} \right\rangle \\
& + \sum_{k=0}^{N-1} \alpha_{k+1}^2 \|\theta_{k+1}\|_2^2 + \sum_{k=0}^{N-1} \alpha_{k+1}^2 \left\langle \theta_{k+1}, \nabla f(x^{k+1}) \right\rangle + \frac{A_N \varepsilon}{4}
\end{aligned}
$$

that concludes the proof. □ □

From Lemma A.3 we know that $A_N \sim \frac{N^{\frac{1+3\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+\nu}}}{M_\nu^{\frac{2}{1+\nu}}}$. Therefore, in view of Lemma 4.1 (inequality (12) with $z = x^*$), to derive the desired complexity bound from Theorem 2.1 it is sufficient to show that

$$
\sum_{k=0}^{N-1} \left(\alpha_{k+1} \left\langle \theta_{k+1}, z - z^k \right\rangle + \alpha_{k+1}^2 \|\theta_{k+1}\|_2^2 + \alpha_{k+1}^2 \left\langle \theta_{k+1}, \nabla f(x^{k+1}) \right\rangle \right) + \frac{A_N \varepsilon}{4} \lesssim R_0^2
$$

with probability at least $1 - \beta$. One possible way to achieve this goal is to apply some concentration inequality to these three sums. Since we use clipped stochastic gradients, under a proper choice of the clipping parameter, random vector $\theta_{k+1} = \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) - \nabla f(x^{k+1})$ is bounded in ℓ_2-norm by $2\lambda_{k+1}$ with high probability as well. Taking into account the assumption on the stochastic gradients (see (2)), it is natural to apply Bernstein's inequality (see Lemma A.2). Despite the seeming simplicity, this part of the proof is the trickiest one.

First of all, it is useful to derive tight enough upper bounds for bias, variance, and distortion of $\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)$ – this is the second step of the whole proof. Fortunately, Lemma F.5 from Gorbunov et al. (2020) does exactly what we need in our proof and holds without any changes.

Lemma 4.2 (Lemma F.5 from Gorbunov et al. (2020).)**.** *For all $k \geq 0$, the following inequality holds:*

$$
\left\| \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) - \mathbb{E}_{\boldsymbol{\xi}^k} \left[\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) \right] \right\|_2 \leq 2\lambda_{k+1}. \tag{19}
$$

Moreover, if the stochastic gradient satisfies (2) on $Q = B_{3R_0}(x^*)$ and

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

$$\left\|\mathbb{E}_{\boldsymbol{\xi}^k}\left[\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)\right] - \nabla f(x^{k+1})\right\|_2 \leq \frac{4\sigma^2}{m_k \lambda_{k+1}}, \tag{20}$$

$$\mathbb{E}_{\boldsymbol{\xi}^k}\left[\left\|\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) - \nabla f(x^{k+1})\right\|_2^2\right] \leq \frac{18\sigma^2}{m_k}, \tag{21}$$

$$\mathbb{E}_{\boldsymbol{\xi}^k}\left[\left\|\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) - \mathbb{E}_{\boldsymbol{\xi}^k}\left[\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)\right]\right\|_2^2\right] \leq \frac{18\sigma^2}{m_k}. \tag{22}$$

4.1.2 Proof of the Main Result

The final, third step of the proof consists of providing explicit formulas and bounds for the parameters of the method and derivation of the desired result using induction and Bernstein's inequality. Below, we provide the complete statement of Theorem 2.1.

Theorem 4.1. *Assume that function* f *is convex, achieves minimum value at some*[5] x^**, its stochastic gradient and its gradient satisfy* (2) *and* (3) *respectively with* $\sigma > 0$*,* $\nu \in [0, 1]$*,* $M_\nu > 0$ *on* $Q = B_{3R_0}(x^*)$*, where* $R_0 \geq \|x^0 - x^*\|_2$*. Let* $\beta \in (0, 1)$ *and* $N \geq 1$ *be arbitrary such that*

$$\ln \frac{4N}{\beta} \geq 2 \tag{23}$$

and let the parameters of clipped-SSTM *satisfy*[6]

$$\alpha = \frac{\varepsilon^{\frac{1-\nu}{1+\nu}}}{2aM_\nu^{\frac{2}{1+\nu}}}, \quad m_k = \max\left\{1, \frac{20736N\sigma^2\alpha_{k+1}^2 \ln\frac{4N}{\beta}}{C^2R_0^2}\right\}, \tag{24}$$

$$B = \frac{CR_0}{16\ln\frac{4N}{\beta}}, \quad a \geq 16384\ln^2\frac{4N}{\beta}, \tag{25}$$

$$\varepsilon^{\frac{1-\nu}{1+\nu}} \leq \frac{aCM_\nu^{\frac{1-\nu}{1+\nu}}R_0^{1-\nu}}{16\ln\frac{4N}{\beta}}, \quad \varepsilon \leq \frac{2^{\frac{1+\nu}{2}}a^{\frac{1+\nu}{2}}C^{1+\nu}R_0^{1+\nu}M_\nu}{100^{\frac{1+3\nu}{2}}}, \tag{26}$$

$$\varepsilon^{\frac{1-\nu}{1+3\nu}} \leq \min\left\{\frac{a^{\frac{2+3\nu-\nu^2}{2(1+3\nu)}}}{2^{4+4\nu+\frac{3+8\nu-5\nu^2-6\nu^3}{(1+\nu)(1+3\nu)}}\ln\frac{4N}{\beta}}, \frac{a^{\frac{(1+\nu)^2}{1+3\nu}}}{2^{5+8\nu+\frac{2+9\nu+7\nu^2-3\nu^3+\nu^4}{(1+\nu)(1+3\nu)}}\ln^{1+\nu}\frac{4N}{\beta}}\right\}C^{\frac{1-\nu^2}{1+3\nu}}R_0^{\frac{1-\nu^2}{1+3\nu}}M_\nu^{\frac{1-\nu}{1+3\nu}}, \tag{27}$$

[5] Our proofs are valid for any solution x^* and, for example, one can take x^* as the closest solution to the starting point x^0.

[6] The choice of the parameters (in this and the following results) is dictated by the need to estimate and control the stochastic error in the proofs. If some of the parameters (such as ν, R_0, M_ν, σ) are unknown, one can directly tune parameters α, a, m_k. To satisfy (26) and (27) it sufficient to choose sufficiently large a (or, alternatively, sufficiently small ε).

$$\varepsilon \leq \frac{2^{\frac{1+\nu}{2}} a^{\frac{1+\nu}{2}} C^{1+\nu} R_0^{1+\nu} M_\nu}{N^{\frac{1+3\nu}{2}}} + \frac{1}{N^{\frac{1+3\nu}{2}}}. \tag{28}$$

Then, after N iterations of clipped-SSTM*, with probability at least $1-\beta$, it holds that*

$$f(y^N) - f(x^*) \leq \frac{4aC^2 R_0^2 M_\nu^{\frac{2}{1+\nu}}}{N^{\frac{1+3\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+\nu}}}, \tag{29}$$

where

$$C = \sqrt{7}. \tag{30}$$

In other words, if we choose $a = 16384 \ln^2 \frac{4N}{\beta}$, then the method achieves $f(y^N) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$ after $\mathcal{O}\left(D \ln^{\frac{2(1+\nu)}{1+3\nu}} \frac{D}{\beta}\right)$ iterations with $D = \frac{M_\nu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}}$ and requires

$$\mathcal{O}\left(\max\left\{D \ln^{\frac{2(1+\nu)}{1+3\nu}} \frac{D}{\beta}, \frac{\sigma^2 R_0^2}{\varepsilon^2} \ln \frac{D}{\beta}\right\}\right) \text{ oracle calls.} \tag{31}$$

Proof. The proof of this result (and the following ones) is induction-based: we will show by induction that the iterates stay in some bounded ball around x^* with high probability. This will allow us to apply Bernstein-type concentration inequality to estimate the stochastic sums appearing in the upper bounds.

First of all, we notice that for each $k \geq 0$ iterates x^{k+1}, z^k, y^k lie in the ball $B_{\widetilde{R}_k}(x^*)$, where $R_k = \|z^k - x^*\|_2$, $\widetilde{R}_0 = R_0$, $\widetilde{R}_{k+1} = \max\{\widetilde{R}_k, R_{k+1}\}$. We prove it using induction. Since $y^0 = z^0 = x^0$, $\widetilde{R}_0 = R_0 \geq \|z^0 - x^*\|_2$ and $x^1 = \frac{A_0 y^0 + \alpha_1 z^0}{A_1} = z^0$ 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 $y^{l-1} \in B_{\widetilde{R}_{l-1}}(x^*) \subseteq B_{\widetilde{R}_l}(x^*)$, and $z^l \in B_{\widetilde{R}_l}(x^*)$, 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.

Next, our goal is to prove via induction that for all $k = 0, 1, \ldots, N$ we have $\mathbb{P}\{E_k\} = 1 - \frac{k\beta}{N}$ for probability event E_k defined as follows:

Event E_k:

Inequalities

$$\begin{aligned} R_t^2 \leq\ & R_0^2 + 2\sum_{l=0}^{t-1} \alpha_{l+1} \left\langle \theta_{l+1}, x^* - z^l \right\rangle + 2\sum_{l=0}^{t-1} \alpha_{l+1}^2 \left\langle \theta_{l+1}, \nabla f(x^{l+1}) \right\rangle \\ & + 2\sum_{l=0}^{t-1} \alpha_{k+1}^2 \|\theta_{l+1}\|_2^2 + \frac{A_N \varepsilon}{2} \leq C^2 R_0^2 \end{aligned} \tag{32}$$

hold for $t = 0, 1, \ldots, k$ simultaneously where C is defined in (30).

For $t = 0$ inequality (32) 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, since R_{T-1} implies $R_t \leq CR_0$ for all $t = 0, 1, \ldots, T-1$ we have that $\widetilde{R}_{T-1} \leq CR_0$, and, as a consequence, $z^{T-1} \in B_{CR_0}(x^*)$. Therefore, probability event E_{T-1} implies

$$
\begin{aligned}
&\|z^T - x^*\|_2 \\
&= \|z^{T-1} - x^* - \alpha_T \widetilde{\nabla} f(x^T, \boldsymbol{\xi}^{T-1})\|_2 \leq \|z^{T-1} - x^*\|_2 + \alpha_T \|\widetilde{\nabla} f(x^T, \boldsymbol{\xi}^{T-1})\|_2 \\
&\leq CR_0 + \alpha_T \lambda_T = \left(1 + \frac{1}{16 \ln \frac{4N}{\beta}}\right) CR_0 \overset{(23),(30)}{\leq} \left(1 + \frac{1}{32}\right) \sqrt{7} R_0 \leq 3R_0,
\end{aligned}
$$

hence $\widetilde{R}_T \leq 3R_0$. Then, one can apply Lemma 4.1 and get that probability event E_{T-1} implies

$$
\begin{aligned}
A_t \left(f(y^t) - f(x^*)\right) &\leq \frac{1}{2}\|z^0 - x^*\|_2^2 - \frac{1}{2}\|z^t - x^*\|_2^2 + \sum_{k=0}^{t-1} \alpha_{k+1} \left\langle \theta_{k+1}, x^* - z^k \right\rangle \\
&\quad + \sum_{k=0}^{t-1} \alpha_{k+1}^2 \|\theta_{k+1}\|_2^2 + \sum_{k=0}^{t-1} \alpha_{k+1}^2 \left\langle \theta_{k+1}, \nabla f(x^{k+1}) \right\rangle + \frac{A_t \varepsilon}{4}, \qquad (33) \\
\theta_{k+1} &\overset{\text{def}}{=} \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) - \nabla f(x^{k+1}) \qquad (34)
\end{aligned}
$$

for all $t = 0, 1, \ldots, T$. Taking into account that $f(y^t) - f(x^*) \geq 0$ for all y^t we derive that probability event E_{T-1} implies

$$
\begin{aligned}
R_t^2 &\leq R_0^2 + 2\sum_{l=0}^{t-1} \alpha_{l+1} \left\langle \theta_{l+1}, x^* - z^l \right\rangle + 2\sum_{l=0}^{t-1} \alpha_{l+1}^2 \left\langle \theta_{l+1}, \nabla f(x^{l+1}) \right\rangle \\
&\quad + 2\sum_{l=0}^{t-1} \alpha_{l+1}^2 \|\theta_{l+1}\|_2^2 + \frac{A_t \varepsilon}{2}. \qquad (35)
\end{aligned}
$$

for all $t = 0, 1, \ldots, T$.

The rest of the proof is based on the refined analysis of inequality (35). First of all, when $\nu = 0$ from (4) for all $t \geq 0$ we have

$$
\left\|\nabla f(x^{t+1})\right\|_2 \leq M_0 \overset{(25)}{=} \frac{16 M_0 B \ln \frac{4N}{\beta}}{CR_0} \overset{(26)}{\leq} \frac{a M_0^2 B}{\varepsilon} = \frac{B}{2\alpha_{t+1}} = \frac{\lambda_{t+1}}{2}.
$$

Next, we prove that $\|\nabla f(x^{t+1})\|_2 \leq \frac{\lambda_{t+1}}{2}$ when $\nu > 0$. For $t = 0$ we have

$$
\begin{aligned}
\|\nabla f(x^1)\|_2 &= \|\nabla f(z^0)\|_2 \overset{(3)}{\leq} M_\nu \|z^0 - x^*\|_2^\nu \leq M_\nu R_0^\nu \overset{(26)}{\leq} \frac{2^{\frac{1-\nu}{1+\nu}} a C R_0 M_\nu^{\frac{2}{1+\nu}}}{32 \varepsilon^{\frac{1-\nu}{1+\nu}} \ln \frac{4N}{\beta}} \\
&\overset{(25),(26)}{\leq} \frac{B}{2\alpha_1} = \frac{\lambda_1}{2}.
\end{aligned}
$$

For $0 < t \leq T-1$ probability event E_{T-1} implies

$$
\begin{aligned}
&\|\nabla f(x^{t+1})\|_2 \\
&\leq \quad \|\nabla f(x^{t+1}) - \nabla f(y^t)\|_2 + \|\nabla f(y^t)\|_2 \\
&\overset{(3),\text{ Lemma A.4}}{\leq} \quad M_\nu \|x^{t+1} - y^t\|_2^\nu + \left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}} \left(f(y^t) - f(x^*)\right)^{\frac{\nu}{1+\nu}} \\
&\overset{(12),(18),(32)}{\leq} \quad M_\nu \left(\frac{\alpha_{t+1}}{A_t}\right)^\nu \|x^{t+1} - z^t\|_2^\nu + \left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}} \left(\frac{C^2R_0^2}{2A_t}\right)^{\frac{\nu}{1+\nu}} \\
&= \quad \underbrace{\frac{\lambda_{t+1}}{2} \cdot \frac{2M_\nu}{\lambda_{t+1}} \left(\frac{\alpha_{t+1}}{A_t}\right)^\nu \|x^{t+1} - z^t\|_2^\nu}_{D_1} \\
&\qquad + \underbrace{\frac{\lambda_{t+1}}{2} \cdot \left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}} \frac{2M_\nu^{\frac{1}{1+\nu}}}{\lambda_{t+1}} \left(\frac{C^2R_0^2}{2A_t}\right)^{\frac{\nu}{1+\nu}}}_{D_2}.
\end{aligned}
$$

Next, we show that $D_1 + D_2 \leq 1$. Using the definition of λ_{t+1}, the definition of $\alpha_{t+1} = \alpha(t+1)^{\frac{2\nu}{1+\nu}}$, triangle inequality $\|x^{t+1} - z^t\|_2 \leq \|x^{t+1} - x^*\|_2 + \|z^t - x^*\|_2 \leq 2CR_0$, and lower bound (91) for A_t (see Lemma A.3) we derive

$$
\begin{aligned}
D_1 \quad &= \quad \frac{2^{\nu+5} M_\nu \alpha_{t+1}^{1+\nu} \ln\frac{4N}{\beta}}{C^{1-\nu}R_0^{1-\nu}A_t^\nu} \overset{(24)}{=} \frac{2^{\nu+5} M_\nu (t+1)^{2\nu} (\varepsilon/2)^{1-\nu} \ln\frac{4N}{\beta}}{2^{2\nu} a^{1+\nu} C^{1-\nu} R_0^{1-\nu} M_\nu^2 A_t^\nu} \\
&\overset{(91)}{\leq} \quad \frac{2^4 (t+1)^{2\nu} \varepsilon^{1-\nu} \ln\frac{4N}{\beta}}{a^{1+\nu} C^{1-\nu} R_0^{1-\nu} M_\nu} \cdot \frac{2^{\frac{(1+3\nu)\nu}{1+\nu}} a^\nu M_\nu^{\frac{2\nu}{1+\nu}}}{t^{\frac{(1+3\nu)\nu}{1+\nu}} (\varepsilon/2)^{\frac{\nu(1-\nu)}{1+\nu}}} \\
&= \quad \frac{(t+1)^{2\nu}}{t^{\frac{\nu(1+3\nu)}{1+\nu}}} \cdot \frac{2^{4+2\nu} \varepsilon^{\frac{1-\nu}{1+\nu}} \ln\frac{4N}{\beta}}{a M_\nu^{\frac{1-\nu}{1+\nu}} C^{1-\nu} R_0^{1-\nu}} \overset{\frac{t+1}{t} \leq 2}{\leq} \frac{2^{4+4\nu} t^{\frac{\nu(1-\nu)}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+\nu}} \ln\frac{4N}{\beta}}{a M_\nu^{\frac{1-\nu}{1+\nu}} C^{1-\nu} R_0^{1-\nu}} \\
&\overset{t \leq N-1, (28)}{\leq} \quad \frac{2^{4+4\nu} \varepsilon^{\frac{1-\nu}{1+\nu}} \ln\frac{4N}{\beta}}{a M_\nu^{\frac{1-\nu}{1+\nu}} C^{1-\nu} R_0^{1-\nu}} \cdot \frac{2^{\frac{2\nu(1-\nu)(1+2\nu)}{(1+\nu)(1+3\nu)}} a^{\frac{\nu(1-\nu)}{1+3\nu}} C^{\frac{2\nu(1-\nu)}{1+3\nu}} R_0^{\frac{2\nu(1-\nu)}{1+3\nu}} M_\nu^{\frac{2\nu(1-\nu)}{(1+\nu)(1+3\nu)}}}{\varepsilon^{\frac{2\nu(1-\nu)}{(1+\nu)(1+3\nu)}}} \\
&= \quad \frac{2^{4+4\nu+\frac{2\nu(1-\nu)(1+2\nu)}{(1+\nu)(1+3\nu)}} \varepsilon^{\frac{1-\nu}{1+3\nu}} \ln\frac{4N}{\beta}}{a^{\frac{(1+\nu)^2}{1+3\nu}} M_\nu^{\frac{1-\nu}{1+3\nu}} C^{\frac{(1-\nu)(1+\nu)}{1+3\nu}} R_0^{\frac{(1-\nu)(1+\nu)}{1+3\nu}}} \overset{(27)}{\leq} \frac{1}{2^{\frac{3+6\nu-7\nu^2-2\nu^3}{(1+\nu)(1+3\nu)}} a^{\frac{\nu}{2}}}.
\end{aligned}
$$

Applying the same inequalities and $\left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}} \leq 2$ we estimate D_2:

$$
\begin{aligned}
D_2 &= \left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}} \frac{2^{5-\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}} \alpha_{t+1} \ln\frac{4N}{\beta}}{C^{\frac{1-\nu}{1+\nu}} R_0^{\frac{1-\nu}{1+\nu}} A_t^{\frac{\nu}{1+\nu}}} \\
&\leq 2 \cdot \frac{2^{5-\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}} \ln\frac{4N}{\beta}}{C^{\frac{1-\nu}{1+\nu}} R_0^{\frac{1-\nu}{1+\nu}} A_t^{\frac{\nu}{1+\nu}}} \cdot \frac{(t+1)^{\frac{2\nu}{1+\nu}} (\varepsilon/2)^{\frac{1-\nu}{1+\nu}}}{2^{\frac{2\nu}{1+\nu}} a M_\nu^{\frac{2}{1+\nu}}} \\
&\leq \frac{2^{5-\frac{\nu}{1+\nu}} \cdot 2^{\frac{2\nu}{1+\nu}} t^{\frac{2\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+\nu}} \ln\frac{4N}{\beta}}{a C^{\frac{1-\nu}{1+\nu}} R_0^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}} A_t^{\frac{\nu}{1+\nu}}} \\
&\overset{(91)}{\leq} \frac{2^{5+\frac{\nu}{1+\nu}} t^{\frac{2\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+\nu}} \ln\frac{4N}{\beta}}{a C^{\frac{1-\nu}{1+\nu}} R_0^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}}} \cdot \frac{2^{\frac{\nu(1+3\nu)}{(1+\nu)^2}} a^{\frac{\nu}{1+\nu}} M_\nu^{\frac{2\nu}{(1+\nu)^2}}}{t^{\frac{\nu(1+3\nu)}{(1+\nu)^2}} (\varepsilon/2)^{\frac{\nu(1-\nu)}{(1+\nu)^2}}} \\
&= \frac{2^{5+\frac{3\nu}{1+\nu}} t^{\frac{\nu(1-\nu)}{(1+\nu)^2}} \varepsilon^{\frac{1-\nu}{(1+\nu)^2}} \ln\frac{4N}{\beta}}{a^{\frac{1}{1+\nu}} C^{\frac{1-\nu}{1+\nu}} R_0^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{1-\nu}{(1+\nu)^2}}} \\
&\overset{t\leq N-1,(28)}{\leq} \frac{2^{5+\frac{3\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{(1+\nu)^2}} \ln\frac{4N}{\beta}}{a^{\frac{1}{1+\nu}} C^{\frac{1-\nu}{1+\nu}} R_0^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{1-\nu}{(1+\nu)^2}}} \\
&\qquad \cdot \frac{2^{\frac{2\nu(1+2\nu)(1-\nu)}{(1+\nu)^2(1+3\nu)}} a^{\frac{\nu(1-\nu)}{(1+\nu)(1+3\nu)}} C^{\frac{2\nu(1-\nu)}{(1+\nu)(1+3\nu)}} R_0^{\frac{2\nu(1-\nu)}{(1+\nu)(1+3\nu)}} M_\nu^{\frac{2\nu(1-\nu)}{(1+\nu)^2(1+3\nu)}}}{\varepsilon^{\frac{2\nu(1-\nu)}{(1+\nu)^2(1+3\nu)}}} \\
&= \frac{2^{5+\frac{3\nu}{1+\nu}+\frac{2\nu(1+2\nu)(1-\nu)}{(1+\nu)^2(1+3\nu)}} \varepsilon^{\frac{1-\nu}{(1+\nu)(1+3\nu)}} \ln\frac{4N}{\beta}}{a^{\frac{1+\nu}{1+3\nu}} C^{\frac{1-\nu}{1+3\nu}} R_0^{\frac{1-\nu}{1+3\nu}} M_\nu^{\frac{1-\nu}{(1+\nu)(1+3\nu)}}} \overset{(27)}{\leq} \frac{1}{2^{\frac{2+5\nu+\nu^3}{(1+\nu)^2(1+3\nu)}}}.
\end{aligned}
$$

Combining the upper bounds for D_1 and D_2 we get

$$
D_1 + D_2 \leq \frac{1}{2^{\frac{3+6\nu-7\nu^2-2\nu^3}{(1+\nu)(1+3\nu)}} a^{\frac{\nu}{2}}} + \frac{1}{2^{\frac{2+5\nu+\nu^3}{(1+\nu)^2(1+3\nu)}}}.
$$

Since $\frac{2+5\nu+\nu^3}{(1+\nu)^2(1+3\nu)}$ is a decreasing function of ν for $\nu \in [0,1]$ we continue as

$$
D_1 + D_2 \leq \frac{1}{2^{\frac{3+6\nu-7\nu^2-2\nu^3}{(1+\nu)(1+3\nu)}} a^{\frac{\nu}{2}}} + \frac{1}{\sqrt{2}}.
$$

Next, we use $a \overset{(25)}{\geq} 16384 \ln^2\frac{4N}{\beta} \overset{(23)}{\geq} 2^{10}$ and obtain

$$
D_1 + D_2 \leq \frac{1}{2^{\frac{3+11\nu+13\nu^2+13\nu^3}{(1+\nu)(1+3\nu)}}} + \frac{1}{\sqrt{2}}.
$$

One can numerically verify that $\frac{1}{2^{\frac{3+11\nu+13\nu^2+13\nu^3}{(1+\nu)(1+3\nu)}}} + \frac{1}{\sqrt{2}}$ is smaller than 1 for $\nu \in [0,1]$. Putting all together, we conclude that probability event E_{T-1} implies

$$
\|\nabla f(x^{t+1})\|_2 \leq \frac{\lambda_{t+1}}{2} \tag{36}
$$

for all $t = 0, 1, \ldots, T-1$. Having inequality (36) in hand, we show in the rest of the proof that (32) holds for $t = T$ with large enough probability. First of all, we introduce new random variables:

$$\eta_l = \begin{cases} x^* - z^l, & \text{if } \|x^* - z^l\|_2 \leq CR_0, \\ 0, & \text{otherwise,} \end{cases}, \zeta_l = \begin{cases} \nabla f(x^{l+1}), & \text{if } \|\nabla f(x^{l+1})\|_2 \leq \frac{B}{2\alpha_{l+1}}, \\ 0, & \text{otherwise,} \end{cases} \tag{37}$$

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

$$\|\eta_l\|_2 \leq CR_0 \quad \text{and} \quad \|\zeta_l\|_2 \leq \frac{B}{2\alpha_{l+1}}. \tag{38}$$

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

$$\begin{aligned} R_T^2 & \\ \overset{(35),(32),(36),(37)}{\leq} & R_0^2 + 2\sum_{l=0}^{T-1} \alpha_{l+1} \langle \theta_{l+1}, \eta_l \rangle + 2\sum_{l=0}^{T-1} \alpha_{l+1}^2 \|\theta_{l+1}\|_2^2 \\ & + 2\sum_{l=0}^{T-1} \alpha_{l+1}^2 \langle \theta_{l+1}, \zeta_l \rangle + \frac{A_N \varepsilon}{2} \\ = & R_0^2 + \sum_{l=0}^{T-1} \alpha_{l+1} \langle \theta_{l+1}, 2\eta_l + 2\alpha_{l+1}\zeta_l \rangle + 2\sum_{l=0}^{T-1} \alpha_{l+1}^2 \|\theta_{l+1}\|_2^2 + \frac{A_N \varepsilon}{2}. \end{aligned}$$

Finally, we do some preliminaries in order to apply Bernstein's inequality (see Lemma A.2) and obtain that E_{T-1} implies

$$\begin{aligned} R_T^2 \overset{(85)}{\leq} & R_0^2 + \underbrace{\sum_{l=0}^{T-1} \alpha_{l+1} \left\langle \theta_{l+1}^u, 2\eta_l + 2\alpha_{l+1}\zeta_l \right\rangle}_{①} + \underbrace{\sum_{l=0}^{T-1} \alpha_{l+1} \left\langle \theta_{l+1}^b, 2\eta_l + 2\alpha_{l+1}\zeta_l \right\rangle}_{②} \\ & + \underbrace{\sum_{l=0}^{T-1} 4\alpha_{l+1}^2 \left(\|\theta_{l+1}^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l} \left[\|\theta_{l+1}^u\|_2^2 \right] \right)}_{③} + \underbrace{\sum_{l=0}^{T-1} 4\alpha_{l+1}^2 \mathbb{E}_{\boldsymbol{\xi}^l} \left[\|\theta_{l+1}^u\|_2^2 \right]}_{④} \\ & + \underbrace{\sum_{l=0}^{T-1} 4\alpha_{l+1}^2 \|\theta_{l+1}^b\|_2^2}_{⑤} + \frac{A_N \varepsilon}{2} \end{aligned} \tag{39}$$

where we introduce new notations:

$$\theta_{l+1}^u \overset{\text{def}}{=} \widetilde{\nabla} f(x^{l+1}, \boldsymbol{\xi}^l) - \mathbb{E}_{\boldsymbol{\xi}^l} \left[\widetilde{\nabla} f(x^{l+1}, \boldsymbol{\xi}^l) \right], \theta_{l+1}^b \overset{\text{def}}{=} \mathbb{E}_{\boldsymbol{\xi}^l} \left[\widetilde{\nabla} f(x^{l+1}, \boldsymbol{\xi}^l) \right] - \nabla f(x^{l+1}), \tag{40}$$

$$\theta_{l+1} \overset{(13)}{=} \theta_{l+1}^u + \theta_{l+1}^b.$$

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 $\mathbb{E}_{\boldsymbol{\xi}^l}[\theta^u_{l+1}] = 0$ and random variables η_l, ζ_l are independent from $\boldsymbol{\xi}^l$ (which is a collection of i.i.d. samples) summands in ① are conditionally unbiased:

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

Secondly, these summands are bounded with probability 1:

$$\begin{aligned}
&\left|\alpha_{l+1}\left\langle\theta^u_{l+1}, 2\eta_l + 2\alpha_{l+1}\zeta_l\right\rangle\right| \\
&\quad\leq \alpha_{l+1}\|\theta^u_{l+1}\|_2\,\|2\eta_l + 2\alpha_{l+1}\zeta_l\|_2 \\
&\quad\overset{(19),(38)}{\leq} 2\alpha_{l+1}\lambda_{l+1}\left(2CR_0 + B\right) = 2B(2CR_0 + B) \\
&\quad= \left(1 + \frac{1}{32\ln\frac{4N}{\beta}}\right)\frac{C^2R_0^2}{4\ln\frac{4N}{\beta}} \overset{(23)}{\leq} \left(1 + \frac{1}{64}\right)\frac{C^2R_0^2}{4\ln\frac{4N}{\beta}}.
\end{aligned}$$

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

$$\begin{aligned}
\sigma_l^2 &\leq \mathbb{E}_{\boldsymbol{\xi}^l}\left[\alpha^2_{l+1}\left\|\theta^u_{l+1}\right\|_2^2\|2\eta_l + 2\alpha_{l+1}\zeta_l\|_2^2\right] \\
&\overset{(38)}{\leq} \alpha^2_{l+1}\mathbb{E}_{\boldsymbol{\xi}^l}\left[\left\|\theta^u_{l+1}\right\|_2^2\right](2CR_0 + B)^2 \\
&= 4\alpha^2_{l+1}\mathbb{E}_{\boldsymbol{\xi}^l}\left[\left\|\theta^u_{l+1}\right\|_2^2\right]\left(1 + \frac{1}{32\ln\frac{4N}{\beta}}\right)^2 C^2R_0^2 \\
&\overset{(23)}{\leq} 4\alpha^2_{l+1}\mathbb{E}_{\boldsymbol{\xi}^l}\left[\left\|\theta^u_{l+1}\right\|_2^2\right]\left(1 + \frac{1}{64}\right)^2 C^2R_0^2, \qquad (41)
\end{aligned}$$

i.e., σ_l^2 is finite due to finiteness of $\|\theta^u_{l+1}\|_2$ (see Lemma 4.2). 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 a bounded martingale difference sequence with bounded conditional variances $\{\sigma_l^2\}_{l\geq 0}$. Thus, we can apply Bernstein's inequality, i.e. we apply Lemma A.2 with $X_l = \alpha_{l+1}\left\langle\theta^u_{l+1}, 2\eta_l + 2\alpha_{l+1}\zeta_l\right\rangle$, $c = \left(1 + \frac{1}{64}\right)\frac{C^2R_0^2}{4\ln\frac{4N}{\beta}}$ and $F = \frac{c^2\ln\frac{4N}{\beta}}{18}$ (The choice of F will be clarified below.) 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 $1 - 2\exp\left(-\frac{b^2}{2F+2cb/3}\right)$

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

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

$$\begin{aligned} b &= \frac{c\ln\frac{4N}{\beta}}{3}+\sqrt{\frac{c^2\ln^2\frac{4N}{\beta}}{9}+2F\ln\frac{4N}{\beta}}=\frac{c\ln\frac{4N}{\beta}}{3}+\sqrt{\frac{2c^2\ln^2\frac{4N}{\beta}}{9}} \\ &= \frac{1+\sqrt{2}}{3}c\ln\frac{4N}{\beta}\le c\ln\frac{4N}{\beta}=\left(1+\frac{1}{64}\right)\frac{C^2R_0^2}{4}=\left(\frac{1}{4}+\frac{1}{256}\right)C^2R_0^2. \end{aligned}$$

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

$$\underbrace{\text{either } \sum_{l=0}^{T-1}\sigma_l^2>F \quad \text{or} \quad |①|\le\left(\frac{1}{4}+\frac{1}{256}\right)C^2R_0^2}_{\text{probability event } E_①}.$$

Here and below, we notice that the conditions of Lemma 4.2 hold when E_{T-1} holds, since event E_{T-1} implies that $x^0,x^1,\ldots,x^T$ lie in $B_{3R_0}(x^*)$. Therefore, probability event E_{T-1} implies that

$$\begin{aligned} \sum_{l=0}^{T-1}\sigma_l^2 &\overset{(41)}{\le} 4\left(1+\frac{1}{64}\right)^2C^2R_0^2\sum_{l=0}^{T-1}\alpha_{l+1}^2\mathbb{E}_{\boldsymbol{\xi}^l}\left[\left\|\theta_{l+1}^u\right\|_2^2\right] \\ &\overset{(22),(36)}{\le} 72\left(1+\frac{1}{64}\right)^2\sigma^2C^2R_0^2\sum_{l=0}^{T-1}\frac{\alpha_{l+1}^2}{m_l} \\ &\overset{(24)}{\le} \frac{\left(1+\frac{1}{64}\right)^2C^4R_0^4}{288\ln\frac{4N}{\beta}}\sum_{l=0}^{T-1}\frac{1}{N} \\ &\overset{T\le N}{\le} \frac{\left(1+\frac{1}{64}\right)^2C^4R_0^4}{288\ln\frac{4N}{\beta}}=\frac{c^2\ln\frac{4N}{\beta}}{18}=F. \end{aligned}$$

Upper bound for ②. The probability event E_{T-1} implies

$$\begin{aligned} &\alpha_{l+1}\left\langle\theta_{l+1}^b,2\eta_l+2\alpha_{l+1}\zeta_l\right\rangle \\ &\quad\le \alpha_{l+1}\left\|\theta_{l+1}^b\right\|_2\|2\eta_l+2\alpha_{l+1}\zeta_l\|_2 \\ &\quad\overset{(20),(38)}{\le} \alpha_{l+1}\cdot\frac{4\sigma^2}{m_l\lambda_{l+1}}(2CR_0+B) \\ &\quad= \frac{4\sigma^2\alpha_{l+1}^2}{m_l}\left(1+\frac{2CR_0}{B}\right)=\frac{4\sigma^2\alpha_{l+1}^2\left(1+32\ln\frac{4N}{\beta}\right)}{m_l} \\ &\quad\overset{(24)}{\le} \frac{4\left(\frac{1}{\ln\frac{4N}{\beta}}+32\right)C^2R_0^2}{20736N}\overset{(23)}{\le}\frac{11C^2R_0^2}{1728N}. \end{aligned}$$

This implies that

$$② = \sum_{l=0}^{T-1}\alpha_{l+1}\left\langle\theta_{l+1}^b,2\eta_l+2\alpha_{l+1}\zeta_l\right\rangle\overset{T\le N}{\le}\frac{11C^2R_0^2}{1728}.$$

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 unbiased:

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

Secondly, the summands are bounded with probability 1:

$$\begin{aligned}\left|4\alpha_{l+1}^2\left(\|\theta_{l+1}^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right]\right)\right| &\leq 4\alpha_{l+1}^2\left(\|\theta_{l+1}^u\|_2^2 + \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right]\right)\\ &\overset{(19)}{\leq} 4\alpha_{l+1}^2\left(4\lambda_{l+1}^2 + 4\lambda_{l+1}^2\right)\\ &= 32B^2 = \frac{C^2R_0^2}{8\ln^2\frac{4N}{\beta}} \overset{(23)}{\leq} \frac{C^2R_0^2}{16\ln\frac{4N}{\beta}} \overset{\text{def}}{=} c_1. \end{aligned} \tag{42}$$

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:

$$\begin{aligned}\hat{\sigma}_l^2 &\overset{(42)}{\leq} c_1\mathbb{E}_{\boldsymbol{\xi}^l}\left[\left|4\alpha_{l+1}^2\left(\|\theta_{l+1}^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right]\right)\right|\right]\\ &\leq 4c_1\alpha_{l+1}^2\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2 + \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right]\right] = 8c_1\alpha_{l+1}^2\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right], \end{aligned} \tag{43}$$

i.e., $\hat{\sigma}_l^2$ is finite due to finiteness of $\|\theta_{l+1}^u\|_2$ (see Lemma 4.2). In other words, sequence $\left\{4\alpha_{l+1}^2\left(\|\theta_{l+1}^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right]\right.\right.$ is 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 A.2 with $X_l = \hat{X}_l = 4\alpha_{l+1}^2\left(\|\theta_{l+1}^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right]\right)$, $c = c_1 = \frac{C^2R_0^2}{16\ln\frac{4N}{\beta}}$ and $F = F_1 = \frac{c_1^2\ln\frac{4N}{\beta}}{18}$ and get that for all $b > 0$

$$\mathbb{P}\left\{\left|\sum_{l=0}^{T-1}\hat{X}_l\right| > b \text{ and } \sum_{l=0}^{T-1}\hat{\sigma}_l^2 \leq F_1\right\} \leq 2\exp\left(-\frac{b^2}{2F_1 + 2c_1b/3}\right)$$

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.,

$$b = \frac{c_1\ln\frac{4N}{\beta}}{3} + \sqrt{\frac{c_1^2\ln^2\frac{4N}{\beta}}{9} + 2F_1\ln\frac{4N}{\beta}} = \frac{1+\sqrt{2}}{3}c_1\ln\frac{4N}{\beta} \leq \frac{C^2R_0^2}{16}.$$

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

$$\underbrace{\text{either } \sum_{l=0}^{T-1}\hat{\sigma}_l^2 > F_1 \quad \text{or} \quad |③| \leq \frac{C^2R_0^2}{16}}_{\text{probability event } E_③}.$$

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

$$\begin{aligned}\sum_{l=0}^{T-1}\hat{\sigma}_l^2 &\overset{(43)}{\leq} 8c_1\sum_{l=0}^{T-1}\alpha_{l+1}^2\mathbb{E}_{\boldsymbol{\xi}^l}\left[\left\|\theta_{l+1}^u\right\|_2^2\right]\\ &\overset{(22),(36)}{\leq} \frac{9\sigma^2C^2R_0^2}{\ln\frac{4N}{\beta}}\sum_{l=0}^{T-1}\frac{\alpha_{l+1}^2}{m_l} \overset{(24)}{\leq} \frac{C^4R_0^4}{2304\ln^2\frac{4N}{\beta}}\sum_{l=0}^{T-1}\frac{1}{N}\\ &\overset{T\leq N}{\leq} \frac{C^4R_0^4}{2304\ln^2\frac{4N}{\beta}} \overset{(23)}{\leq} \frac{C^4R_0^4}{4608\ln\frac{4N}{\beta}} = \frac{c_1^2\ln\frac{4N}{\beta}}{18} = F_1.\end{aligned}$$

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

$$\begin{aligned}④ &= \sum_{l=0}^{T-1}4\alpha_{l+1}^2\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_{l+1}^u\|_2^2\right] \overset{(22),(36)}{\leq} \sum_{l=0}^{T-1}\frac{72\alpha_{l+1}^2\sigma^2}{m_l} \overset{(24)}{\leq} \sum_{l=0}^{T-1}\frac{C^2R_0^2}{288N\ln\frac{4N}{\beta}}\\ &\overset{T\leq N}{\leq} \frac{C^2R_0^2}{288\ln\frac{4N}{\beta}} \overset{(23)}{\leq} \frac{C^2R_0^2}{576}.\end{aligned}$$

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

$$\begin{aligned}⑤ &= \sum_{l=0}^{T-1}4\alpha_{l+1}^2\|\theta_{l+1}^b\|_2^2 \overset{(20),(36)}{\leq} \sum_{l=0}^{T-1}\frac{64\alpha_{l+1}^2\sigma^4}{m_l^2\lambda_{l+1}^2} = \frac{64\sigma^4}{B^2}\sum_{l=0}^{T-1}\frac{\alpha_{l+1}^4}{m_l^2}\\ &\overset{(24),(25)}{\leq} \frac{256\cdot 64\sigma^4\ln^2\frac{4N}{\beta}}{C^2R_0^2}\sum_{l=0}^{T-1}\frac{C^4R_0^4}{20736^2N^2\sigma^4\ln^2\frac{4N}{\beta}} \overset{T\leq N}{\leq} \frac{C^2R_0^2}{26244}.\end{aligned}$$

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

$$\begin{gathered}R_T^2 \overset{(39)}{\leq} R_0^2 + ① + ② + ③ + ④ + ⑤ + \frac{A_N\varepsilon}{2},\\ \text{where}\quad ② \leq \frac{11C^2R_0^2}{1728},\quad ④ \leq \frac{CR_0^2}{576},\quad ⑤ \leq \frac{C^2R_0^2}{26244},\\ \sum_{l=0}^{T-1}\sigma_l^2 \leq F,\quad \sum_{l=0}^{T-1}\hat{\sigma}_l^2 \leq F_1\end{gathered}$$

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

$$\begin{aligned}E_① &= \left\{\text{either } \sum_{l=0}^{T-1}\sigma_l^2 > F \quad \text{or} \quad |①| \leq \left(\frac{1}{4} + \frac{1}{256}\right)C^2R_0^2\right\},\\ E_③ &= \left\{\text{either } \sum_{l=0}^{T-1}\hat{\sigma}_l^2 > F_1 \quad \text{or} \quad |③| \leq \frac{C^2R_0^2}{16}\right\}.\end{aligned}$$

Moreover, since $N \overset{(28)}{\leq} \frac{2^{\frac{1+\nu}{1+3\nu}} a^{\frac{1+\nu}{1+3\nu}} C^{\frac{2(1+\nu)}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}} M_\nu^{\frac{2}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}} + 1$ and $\varepsilon \overset{(26)}{\leq} \frac{2^{\frac{1+\nu}{2}} a^{\frac{1+\nu}{2}} C^{1+\nu} R_0^{1+\nu} M_\nu}{100^{\frac{1+3\nu}{2}}}$ we have

$$\begin{aligned}\frac{A_N \varepsilon}{2} &\overset{(93)}{\leq} \frac{N^{\frac{1+3\nu}{1+\nu}} \varepsilon^{\frac{2}{1+\nu}}}{4a M_\nu^{\frac{2}{1+\nu}}} \overset{(28)}{\leq} \left(\frac{2^{\frac{1+\nu}{1+3\nu}} a^{\frac{1+\nu}{1+3\nu}} C^{\frac{2(1+\nu)}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}} M_\nu^{\frac{2}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}} + 1\right)^{\frac{1+3\nu}{1+\nu}} \frac{\varepsilon^{\frac{2}{1+\nu}}}{4a M_\nu^{\frac{2}{1+\nu}}} \\ &\overset{(26)}{\leq} \left(\frac{101}{100}\right)^{\frac{1+3\nu}{1+\nu}} \frac{C^2 R_0^2}{2} \leq \frac{10201 C^2 R_0^2}{20000}.\end{aligned}$$

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

$$\begin{aligned}R_T^2 &\leq \left(1 + \left(\frac{1}{4} + \frac{1}{256} + \frac{11}{1728} + \frac{1}{16} + \frac{1}{576} + \frac{1}{26244} + \frac{10201}{20000}\right) C^2\right) R_0^2 \\ &\overset{(30)}{\leq} C^2 R_0^2. \end{aligned} \tag{44}$$

Moreover, using the bound for the union, we derive

$$\mathbb{P}\left\{E_{T-1} \cap E_① \cap E_③\right\} = 1 - \mathbb{P}\left\{\overline{E}_{T-1} \cup \overline{E}_① \cup \overline{E}_③\right\} \geq 1 - \frac{T\beta}{N}. \tag{45}$$

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

$$\mathbb{P}\{E_T\} \overset{(44)}{\geq} \mathbb{P}\left\{E_{T-1} \cap E_① \cap E_③\right\} \overset{(45)}{\geq} 1 - \frac{T\beta}{N},$$

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 $1 - \beta$

$$\begin{aligned}A_N &\left(f(y^N) - f(x^*)\right) \\ &\overset{(33)}{\leq} \frac{1}{2}\|z^0 - x^*\|_2^2 - \frac{1}{2}\|z^N - x^*\|_2^2 + \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \theta_{k+1}, x^* - z^k \right\rangle \\ &\quad + \sum_{k=0}^{N-1} \alpha_{k+1}^2 \|\theta_{k+1}\|_2^2 + \sum_{k=0}^{N-1} \alpha_{k+1}^2 \left\langle \theta_{k+1}, \nabla f(x^{k+1}) \right\rangle + \frac{A_N \varepsilon}{4} \overset{(32)}{\leq} \frac{C^2 R_0^2}{2}.\end{aligned}$$

Since $A_N \overset{(91)}{\geq} \frac{N^{\frac{1+3\nu}{1+\nu}} (\varepsilon/2)^{\frac{1-\nu}{1+\nu}}}{2^{\frac{1+3\nu}{1+\nu}} a M_\nu^{\frac{2}{1+\nu}}}$ we get that with probability at least $1 - \beta$

$$f(y^N) - f(x^*) \leq \frac{4a C^2 R_0^2 M_\nu^{\frac{2}{1+\nu}}}{N^{\frac{1+3\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+\nu}}}.$$

In other words, `clipped-SSTM` with $a = 16384 \ln^2 \frac{4N}{\beta}$ achieves $f(y^N) - f(x^*) \leq \varepsilon$ with probability

at least $1-\beta$ after $\mathcal{O}\left(\frac{M_\nu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}} \ln^{\frac{2(1+\nu)}{1+3\nu}} \frac{M_\nu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}\beta}\right)$ iterations and requires

$$
\begin{aligned}
\sum_{k=0}^{N-1} m_k &\overset{(24)}{=} \sum_{k=0}^{N-1} \mathcal{O}\left(\max\left\{1, \frac{\sigma^2 \alpha_{k+1}^2 N \ln\frac{N}{\beta}}{R_0^2}\right\}\right) \\
&= \mathcal{O}\left(\max\left\{N, \sum_{k=0}^{N-1} \frac{\sigma^2 (k+1)^{\frac{4\nu}{1+\nu}} \varepsilon^{\frac{2(1-\nu)}{1+\nu}} N \ln\frac{N}{\beta}}{M_\nu^{\frac{4}{1+\nu}} R_0^2 a^2}\right\}\right) \\
&\overset{(25)}{=} \mathcal{O}\left(\max\left\{N, \frac{\sigma^2 \varepsilon^{\frac{2(1-\nu)}{1+\nu}} N^{\frac{2(1+3\nu)}{1+\nu}}}{M_\nu^{\frac{4}{1+\nu}} R_0^2 \ln^3\frac{N}{\beta}}\right\}\right) \\
&= \mathcal{O}\left(\max\left\{\frac{M_\nu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}} \ln^{\frac{2(1+\nu)}{1+3\nu}} \frac{M_\nu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}\beta}, \frac{\sigma^2 R_0^2}{\varepsilon^2} \ln \frac{M_\nu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}\beta}\right\}\right).
\end{aligned}
$$

oracle calls, where in the last equality we substituted the number of iterations N from the statement of the theorem. □ □

4.1.3 On the Batchsizes and Numerical Constants

The obtained complexity result is discussed in detail in Section 2. Here, we discuss the choice of the parameters. For convenience, we provide all assumptions from Theorem 4.1 on the parameters below:

$$\ln\frac{4N}{\beta} \geq 2 \tag{46}$$

$$\alpha = \frac{(\varepsilon/2)^{\frac{1-\nu}{1+\nu}}}{2^{\frac{2\nu}{1+\nu}} a M_\nu^{\frac{2}{1+\nu}}}, \quad m_k = \max\left\{1, \frac{20736 N \sigma^2 \alpha_{k+1}^2 \ln\frac{4N}{\beta}}{C^2 R_0^2}\right\}, \tag{47}$$

$$B = \frac{C R_0}{16 \ln\frac{4N}{\beta}}, \quad a \geq 16384 \ln^2\frac{4N}{\beta}, \tag{48}$$

$$\varepsilon^{\frac{1-\nu}{1+\nu}} \leq \frac{a C M_\nu^{\frac{1-\nu}{1+\nu}} R_0^{1-\nu}}{16 \ln\frac{4N}{\beta}}, \quad \varepsilon \leq \frac{2^{\frac{1+\nu}{2}} a^{\frac{1+\nu}{2}} C^{1+\nu} R_0^{1+\nu} M_\nu}{100^{\frac{1+3\nu}{2}}}, \tag{49}$$

$$
\begin{aligned}
\varepsilon^{\frac{1-\nu}{1+3\nu}} \leq \min\Bigg\{ & \frac{a^{\frac{2+3\nu-\nu^2}{2(1+3\nu)}}}{2^{2+4\nu+\frac{3+8\nu-5\nu^2-6\nu^3}{(1+\nu)(1+3\nu)}} \ln\frac{4N}{\beta}}, \\
& \frac{a^{\frac{(1+\nu)^2}{1+3\nu}}}{2^{4+7\nu+\frac{2+7\nu+2\nu^2-3\nu^3}{(1+\nu)(1+3\nu)}} \ln^{1+\nu}\frac{4N}{\beta}} \Bigg\} C^{\frac{1-\nu^2}{1+3\nu}} R_0^{\frac{1-\nu^2}{1+3\nu}} M_\nu^{\frac{1-\nu}{1+3\nu}},
\end{aligned} \tag{50}
$$

$$N = \left\lceil \frac{2^{\frac{1+\nu}{1+3\nu}} a^{\frac{1+\nu}{1+3\nu}} C^{\frac{2(1+\nu)}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}} M_\nu^{\frac{2}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}} \right\rceil + 1, \quad C = \sqrt{7}. \tag{51}$$

We emphasize that (46), (49), and (50) are not restrictive at all since the target accuracy ε and confidence level β are often chosen to be small enough, whereas a can be made large enough.

One can notice that the assumptions on parameter a and batch size m_k contain huge numerical constants (see (47)-(48)) that result in large numerical constants in the expression for the number of iterations N and the total number of oracle calls required to guarantee accuracy ε of the solution. However, for the sake of simplicity of the proofs, we do not try to provide an analysis with better dependence on the numerical constants. Moreover, the main goal of this paper is to derive improved high-probability complexity guarantees in terms of $\mathcal{O}(\cdot)$-notation – such guarantees are insensitive to numerical constants by definition.

Finally, (47) implies that the batch size at iteration k is

$$m_k = \Theta\left(\max\left\{1, \frac{N\sigma^2(k+1)^{\frac{4\nu}{1+\nu}}\varepsilon^{\frac{2(1-\nu)}{1+\nu}}\ln\frac{N}{\beta}}{a^2 M_\nu^{\frac{4}{1+\nu}} R_0^2}\right\}\right)$$

meaning that for $k \sim N$ and $a = \mathcal{O}\left(\ln^2\frac{N}{\beta}\right)$ we have that the second term in the maximum is proportional to $N^{\frac{1+5\nu}{1+\nu}}\varepsilon^{\frac{2(1-\nu)}{1+\nu}}$. When ν is close to 1 and $\sigma^2 \gg 0$, it implies that m_k is huge for big enough k, making the method completely impractical. Fortunately, this issue can be easily solved without sacrificing the oracle complexity of the method: it is sufficient to choose large enough a.

Corollary 4.1. *Let the assumptions of Theorem 4.1 hold and*

$$a = \max\left\{16384\ln^2\frac{4N}{\beta}, \frac{5184^{\frac{1+3\nu}{1+\nu}}\cdot 2^{\frac{2(1+5\nu)(1+2\nu)}{(1+\nu)^2}}\sigma^{\frac{2(1+3\nu)}{1+\nu}}C^{\frac{4\nu}{1+\nu}}R_0^{\frac{4\nu}{1+\nu}}\ln^{\frac{1+3\nu}{1+\nu}}\frac{4N}{\beta}}{M_\nu^{\frac{2}{1+\nu}}\varepsilon^{\frac{6\nu}{1+\nu}}}\right\}. \tag{52}$$

Then for all $k = 0, 1, \ldots, N-1$ we have $m_k = 1$ and to achieve $f(y^N) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$ clipped-SSTM *requires*

$$\mathcal{O}\left(\max\left\{\frac{M_\nu^{\frac{2}{1+3\nu}}R_0^{\frac{2(1+\nu)}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}}\ln^{\frac{2(1+\nu)}{1+3\nu}}\frac{M_\nu^{\frac{2}{1+3\nu}}R_0^{\frac{2(1+\nu)}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}\beta}, \frac{\sigma^2R_0^2}{\varepsilon^2}\ln\frac{\sigma^2R_0^2}{\varepsilon^2\beta}\right\}\right) \tag{53}$$

iterations/oracle calls.

Proof. We start with showing that for the new choice of a we have $m_k = 1$ for all $k = 0, 1, \ldots, N-1$. Indeed, using the assumptions on the parameters from Theorem 4.1, we derive

$$\begin{aligned}
m_k &= \max\left\{1, \frac{20736N\sigma^2\alpha_{k+1}^2\ln\frac{4N}{\beta}}{C^2R_0^2}\right\} \\
&= \max\left\{1, \frac{5184N\sigma^2(k+1)^{\frac{4\nu}{1+\nu}}\varepsilon^{\frac{2(1-\nu)}{1+\nu}}}{a^2M_\nu^{\frac{4}{1+\nu}}C^2R_0^2}\right\} \\
&\overset{k<N}{\leq} \max\left\{1, \frac{5184\sigma^2N^{\frac{1+5\nu}{1+\nu}}\varepsilon^{\frac{2(1-\nu)}{1+\nu}}}{a^2M_\nu^{\frac{4}{1+\nu}}C^2R_0^2}\right\} \\
&\overset{(30)}{\leq} \max\left\{1, \frac{5184\cdot 2^{\frac{2(1+5\nu)(1+2\nu)}{(1+\nu)(1+3\nu)}}\sigma^2C^{\frac{4\nu}{1+3\nu}}R_0^{\frac{4\nu}{1+3\nu}}}{a^{\frac{1+\nu}{1+3\nu}}M_\nu^{\frac{2}{1+3\nu}}\varepsilon^{\frac{6\nu}{1+3\nu}}}\right\} \overset{(52)}{\leq} 1.
\end{aligned}$$

That is, with the choice of the stepsize parameter a as in (52), the method uses unit batch sizes at each iteration. Therefore, iteration and oracle complexities coincide in this case. Next, we consider two possible situations.

1. If $a = 16384 \ln^2 \frac{4N}{\beta}$, then

$$\begin{aligned} N &\overset{(30)}{=} \left\lceil \frac{2^{\frac{1+\nu}{1+3\nu}} a^{\frac{1+\nu}{1+3\nu}} C^{\frac{2(1+\nu)}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}} M_\nu^{\frac{2}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}} \right\rceil + 1 \\ &= \mathcal{O}\left(\frac{M_\nu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}} \ln^{\frac{2(1+\nu)}{1+3\nu}} \frac{N}{\beta} \right) \\ &= \mathcal{O}\left(\frac{M_\nu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}} \ln^{\frac{2(1+\nu)}{1+3\nu}} \frac{M_\nu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}} \beta} \right). \end{aligned}$$

2. If $a = \dfrac{5184^{\frac{1+3\nu}{1+\nu}} \cdot 2^{\frac{2(1+5\nu)(1+2\nu)}{(1+\nu)^2}} \sigma^{\frac{2(1+3\nu)}{1+\nu}} C^{\frac{4\nu}{1+\nu}} R_0^{\frac{4\nu}{1+\nu}} \ln^{\frac{1+3\nu}{1+\nu}} \frac{4N}{\beta}}{M_\nu^{\frac{2}{1+\nu}} \varepsilon^{\frac{6\nu}{1+\nu}}}$, then

$$\begin{aligned} N &\overset{(30)}{=} \left\lceil \frac{2^{\frac{1+\nu}{1+3\nu}} a^{\frac{1+\nu}{1+3\nu}} C^{\frac{2(1+\nu)}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}} M_\nu^{\frac{2}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}} \right\rceil + 1 \\ &= \mathcal{O}\left(\frac{M_\nu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}}}{\varepsilon^{\frac{2}{1+3\nu}}} \cdot \frac{\sigma^2 R_0^{\frac{4\nu}{1+3\nu}} \ln \frac{4N}{\beta}}{M_\nu^{\frac{2}{1+3\nu}} \varepsilon^{\frac{6\nu}{1+3\nu}}} \right) = \mathcal{O}\left(\frac{\sigma^2 R_0^2}{\varepsilon^2} \ln \frac{\sigma^2 R_0^2}{\varepsilon^2 \beta} \right). \end{aligned}$$

Putting all together, we derive (53). □ □

4.2 Convergence in the Strongly Convex Case

In this section, we provide the full proof of Theorem 2.2 together with a complete statement of the result. Note that due to strong convexity, the solution x^* is unique.

Theorem 4.2. *Assume that function f is μ-strongly convex, its stochastic gradient and its gradient satisfy (2) and (3) respectively with $\sigma > 0$, $\nu \in [0,1]$, $M_\nu > 0$ on $Q = B_{3R_0}(x^*)$, where $R_0 \geq \|x^0 - x^*\|_2$. Let $\varepsilon > 0$, $\beta \in (0,1)$ and for $t = 1, \ldots, \tau$*

$$N_t = \left\lceil \frac{a_t^{\frac{1+\nu}{1+3\nu}} C^{\frac{2(1+\nu)}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}} M_\nu^{\frac{2}{1+3\nu}}}{2^{\frac{(1+\nu)(t-2)}{1+3\nu}} \varepsilon_t^{\frac{2}{1+3\nu}}} \right\rceil + 1, \quad \varepsilon_t = \frac{\mu R_0^2}{2^{t+1}}, \tag{54}$$

$$\tau = \left\lceil \log_2 \frac{\mu R_0}{2\varepsilon} \right\rceil - 1, \quad \ln \frac{4N_t \tau}{\beta} \geq 2, \quad C = \sqrt{7}, \tag{55}$$

$$\alpha^t = \frac{\varepsilon_t^{\frac{1-\nu}{1+\nu}}}{2 a_t M_\nu^{\frac{2}{1+\nu}}}, \quad m_k^t = \max\left\{ 1, \frac{20736 \cdot 2^{t-1} N_t \sigma^2 (\alpha_{k+1}^t)^2 \ln \frac{4N_t\tau}{\beta}}{C^2 R_0^2} \right\}, \tag{56}$$

$$\alpha_{k+1}^t = \alpha^t (k+1)^{\frac{2\nu}{1+\nu}}, \quad B_t = \frac{CR_0}{16 \ln \frac{4N_t\tau}{\beta}}, \quad a_t = 16384 \ln^2 \frac{4N_t\tau}{\beta}, \tag{57}$$

$$\varepsilon_t^{\frac{1-\nu}{1+\nu}} \leq \frac{a_t C M_\nu^{\frac{1-\nu}{1+\nu}} R_0^{1-\nu}}{16 \cdot 2^{\frac{(1-\nu)(t-1)}{2}} \ln \frac{4N_t\tau}{\beta}}, \quad \varepsilon_t \leq \frac{a_t^{\frac{1+\nu}{2}} C^{1+\nu} R_0^{1+\nu} M_\nu}{100^{\frac{1+3\nu}{2}} \cdot 2^{\frac{(1+\nu)(t-2)}{2}}}, \tag{58}$$

$$\begin{aligned} \varepsilon_t^{\frac{1-\nu}{1+3\nu}} \quad \leq \quad & \min \left\{ \frac{a_t^{\frac{2+3\nu-\nu^2}{2(1+3\nu)}}}{2^{2+4\nu+\frac{3+8\nu-5\nu^2-6\nu^3}{(1+\nu)(1+3\nu)}} \ln \frac{4N_t\tau}{\beta}}, \right. \\ & \left. \frac{a_t^{\frac{(1+\nu)^2}{1+3\nu}}}{2^{4+7\nu+\frac{2+7\nu+2\nu^2-3\nu^3}{(1+\nu)(1+3\nu)}} \ln^{1+\nu} \frac{4N_t\tau}{\beta}} \right\} \frac{C^{\frac{1-\nu^2}{1+3\nu}} R_0^{\frac{1-\nu^2}{1+3\nu}} M_\nu^{\frac{1-\nu}{1+3\nu}}}{2^{\frac{(1-\nu^2)(t-1)}{2(1+3\nu)}}}. \end{aligned} \tag{59}$$

Then, after τ restarts R-clipped-SSTM *produces $\hat{x}^\tau$ such that with probability at least $1-\beta$*

$$f(\hat{x}^\tau) - f(x^*) \leq \varepsilon. \tag{60}$$

That is, to achieve (60) with probability at least $1-\beta$ the method requires

$$\hat{N} = \mathcal{O}\left(\max\left\{ \left(\frac{M_\nu}{\mu R_0^{1-\nu}}\right)^{\frac{2}{1+3\nu}} \ln \frac{\mu R_0^2}{\varepsilon}, \left(\frac{M_\nu^2}{\mu^{1+\nu}\varepsilon^{1-\nu}}\right)^{\frac{1}{1+3\nu}} \right\} \ln^{\frac{2(1+\nu)}{1+3\nu}} \frac{M_\nu^{\frac{2}{1+3\nu}} \ln \frac{\mu R_0^2}{\varepsilon}}{\mu^{\frac{1+\nu}{1+3\nu}} \varepsilon^{\frac{1-\nu}{1+3\nu}} \beta} \right) \tag{61}$$

iterations of Algorithm 1 and

$$\mathcal{O}\left(\max\left\{ \hat{N}, \frac{\sigma^2}{\mu\varepsilon} \ln \frac{M_\nu^{\frac{2}{1+3\nu}} \ln \frac{\mu R_0^2}{\varepsilon}}{\mu^{\frac{1+\nu}{1+3\nu}} \varepsilon^{\frac{1-\nu}{1+3\nu}} \beta} \right\} \right) \quad \textit{oracle calls.} \tag{62}$$

Proof. Applying the convergence rate result (29) in Theorem 4.1 together with our choice of the parameters of the first restart, we obtain that with probability at least $1 - \frac{\beta}{\tau}$ it holds that $f(\hat{x}^1) - f(x^*) \leq \frac{\mu R_0^2}{4}$. Since f is μ-strongly convex we have $\frac{\mu \|\hat{x}^1 - x^*\|_2^2}{2} \leq f(\hat{x}^1) - f(x^*)$. Therefore, with probability at least $1 - \frac{\beta}{\tau}$

$$f(\hat{x}^1) - f(x^*) \leq \frac{\mu R_0^2}{4}, \quad \|\hat{x}^1 - x^*\|_2^2 \leq \frac{R_0^2}{2}.$$

From mathematical induction and the union bound for probability events, it follows that the inequalities

$$f(\hat{x}^t) - f(x^*) \leq \frac{\mu R_0^2}{2^{t+1}}, \quad \|\hat{x}^t - x^*\|_2^2 \leq \frac{R_0^2}{2^t}$$

hold simultaneously for $t = 1, \ldots, \tau$ with probability at least $1 - \beta$. Thus, it means that after $\tau = \left\lceil \log_2 \frac{\mu R_0^2}{\varepsilon} \right\rceil - 1$ restarts R-clipped-SSTM finds an ε-solution with probability at least $1 - \beta$. The

total number of iterations $\hat{N}$ is

$$
\begin{aligned}
&\sum_{t=1}^{\tau} N_t \\
\overset{(54),(57)}{=} &\mathcal{O}\left(\sum_{t=1}^{\tau} \frac{M_\nu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}}}{2^{\frac{(1+\nu)t}{1+3\nu}} \varepsilon_t^{\frac{2}{1+3\nu}}} \ln^{\frac{2(1+\nu)}{1+3\nu}} \frac{M_\nu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}} \tau}{2^{\frac{(1+\nu)t}{1+3\nu}} \varepsilon_t^{\frac{2}{1+3\nu}} \beta}\right) \\
= &\mathcal{O}\left(\sum_{t=1}^{\tau} \frac{M_\nu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}} 2^{\frac{2t}{1+3\nu}}}{2^{\frac{(1+\nu)t}{1+3\nu}} \mu^{\frac{2}{1+3\nu}} R_0^{\frac{4}{1+3\nu}}} \ln^{\frac{2(1+\nu)}{1+3\nu}} \frac{M_\nu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1+\nu)}{1+3\nu}} 2^{\frac{2t}{1+3\nu}} \tau}{2^{\frac{(1+\nu)t}{1+3\nu}} \mu^{\frac{2}{1+3\nu}} R_0^{\frac{4}{1+3\nu}} \beta}\right) \\
= &\mathcal{O}\left(\sum_{t=1}^{\tau} \frac{M_\nu^{\frac{2}{1+3\nu}} 2^{\frac{(1-\nu)t}{1+3\nu}}}{\mu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1-\nu)}{1+3\nu}}} \ln^{\frac{2(1+\nu)}{1+3\nu}} \frac{M_\nu^{\frac{2}{1+3\nu}} 2^{\frac{(1-\nu)t}{1+3\nu}} \tau}{\mu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1-\nu)}{1+3\nu}} \beta}\right) \\
= &\mathcal{O}\left(\frac{M_\nu^{\frac{2}{1+3\nu}} \max\left\{\tau, 2^{\frac{(1-\nu)\tau}{1+3\nu}}\right\}}{\mu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1-\nu)}{1+3\nu}}} \ln^{\frac{2(1+\nu)}{1+3\nu}} \frac{M_\nu^{\frac{2}{1+3\nu}} 2^{\frac{(1-\nu)\tau}{1+3\nu}} \tau}{\mu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1-\nu)}{1+3\nu}} \beta}\right) \\
= &\mathcal{O}\left(\max\left\{\left(\frac{M_\nu}{\mu R_0^{1-\nu}}\right)^{\frac{2}{1+3\nu}} \ln \frac{\mu R_0^2}{\varepsilon}, \left(\frac{M_\nu^2}{\mu^{1+\nu}\varepsilon^{1-\nu}}\right)^{\frac{1}{1+3\nu}}\right\} \ln^{\frac{2(1+\nu)}{1+3\nu}} \frac{M_\nu^{\frac{2}{1+3\nu}} \ln \frac{\mu R_0^2}{\varepsilon}}{\mu^{\frac{1+\nu}{1+3\nu}} \varepsilon^{\frac{1-\nu}{1+3\nu}} \beta}\right),
\end{aligned}
$$

and the total number of oracle calls equals

$$
\begin{aligned}
\sum_{t=1}^{\tau} \sum_{k=0}^{N_t-1} m_k^t &= \mathcal{O}\left(\max\left\{\sum_{t=1}^{\tau} N_t, \sum_{t=1}^{\tau} \frac{\sigma^2 R_0^2}{2^t \varepsilon_t^2} \ln \frac{M_\nu^{\frac{2}{1+3\nu}} 2^{\frac{(1-\nu)t}{1+3\nu}} \tau}{\mu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1-\nu)}{1+3\nu}} \beta}\right\}\right) \\
&= \mathcal{O}\left(\max\left\{\hat{N}, \sum_{t=1}^{\tau} \frac{\sigma^2 \cdot 2^t}{\mu^2 R_0^2} \ln \frac{M_\nu^{\frac{2}{1+3\nu}} 2^{\frac{(1-\nu)\tau}{1+3\nu}} \tau}{\mu^{\frac{2}{1+3\nu}} R_0^{\frac{2(1-\nu)}{1+3\nu}} \beta}\right\}\right) \\
&= \mathcal{O}\left(\max\left\{\hat{N}, \frac{\sigma^2}{\mu\varepsilon} \ln \frac{M_\nu^{\frac{2}{1+3\nu}} \ln \frac{\mu R_0^2}{\varepsilon}}{\mu^{\frac{1+\nu}{1+3\nu}} \varepsilon^{\frac{1-\nu}{1+3\nu}} \beta}\right\}\right). \qquad \square
\end{aligned}
$$

$\square$

One can also derive a similar result for R-clipped-SSTM when stepsize parameter a is chosen as in Corollary 4.1 for all restarts. In this case, on can choose unit batch sizes: $m_k^t = 1$ for all k and t.

5 SGD with Clipping: Missing Details and Proofs

5.1 Convex Case

In this section, we provide a full statement of Theorem 3.1 together with its proof. The proof is based on a similar idea as the proof of the complexity bounds for clipped-SSTM.

Theorem 5.1. *Assume that the function f is convex, achieves its minimum at a point x^*, and its stochastic gradient and its gradient satisfy (2) and (3) respectively with $\sigma > 0$, $\nu \in [0,1]$, $M_\nu > 0$ on $Q = B_{7R_0}(x^*)$, where $R_0 \geq \|x^0 - x^*\|_2$. Then, for all $\beta \in (0,1)$ and N such that*

$$\ln \frac{4N}{\beta} \geq 2, \tag{63}$$

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

$$\lambda = \frac{R_0}{\gamma \ln \frac{4N}{\beta}}, \quad m \geq \max\left\{1, \frac{81N\sigma^2}{\lambda^2 \ln \frac{4N}{\beta}}\right\} \tag{64}$$

and stepsize

$$\gamma \leq \min\left\{\frac{\varepsilon^{\frac{1-\nu}{1+\nu}}}{8M_\nu^{\frac{2}{1+\nu}}}, \frac{R_0}{\sqrt{2N}\varepsilon^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}}}, \frac{R_0^{1-\nu}}{2C^\nu M_\nu \ln \frac{4N}{\beta}}\right\}, \tag{65}$$

with probability at least $1-\beta$ it holds that

$$f(\bar{x}^N) - f(x^*) \leq \frac{C^2 R_0^2}{\gamma N}, \tag{66}$$

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

$$C = 7. \tag{67}$$

In other words, clipped-SGD *with $\gamma = \min\left\{\frac{\varepsilon^{\frac{1-\nu}{1+\nu}}}{8M_\nu^{\frac{2}{1+\nu}}}, \frac{R_0}{\sqrt{2N}\varepsilon^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}}}, \frac{R_0^{1-\nu}}{2C^\nu M_\nu \ln \frac{4N}{\beta}}\right\}$ achieves $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$ after $\mathcal{O}\left(\max\left\{\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{2}{1+\nu}}}, \frac{M_\nu R_0^{1+\nu}}{\varepsilon} \ln \frac{M_\nu R_0^{1+\nu}}{\varepsilon\beta}\right\}\right)$ iterations and requires*

$$\mathcal{O}\left(\max\left\{\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{2}{1+\nu}}}, \max\left\{\frac{M_\nu R_0^{1+\nu}}{\varepsilon}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\} \ln \frac{M_\nu R_0^{1+\nu}}{\varepsilon\beta}\right\}\right) \tag{68}$$

oracle calls.

Proof. Since $f(x)$ is convex and its gradients satisfy (3), we get the following inequality under assumption that $x^k \in B_{7R_0}(x^*)$:

$$\begin{aligned}
& \|x^{k+1} - x^*\|_2^2 = \|x^k - \gamma \widetilde{\nabla} f(x^k, \boldsymbol{\xi}^k) - x^*\|_2^2 \\
= \;& \|x^k - x^*\|_2^2 + \gamma^2 \|\widetilde{\nabla} f(x^k, \boldsymbol{\xi}^k)\|_2^2 - 2\gamma \left\langle x^k - x^*, \widetilde{\nabla} f(x^k, \boldsymbol{\xi}^k)\right\rangle \\
\overset{(13)}{=} \;& \|x^k - x^*\|_2^2 + \gamma^2 \|\nabla f(x^k) + \theta_k\|_2^2 - 2\gamma \left\langle x^k - x^*, \nabla f(x^k) + \theta_k\right\rangle \\
\overset{(85)}{\leq} \;& \|x^k - x^*\|_2^2 + 2\gamma^2 \|\nabla f(x^k)\|_2^2 + 2\gamma^2 \|\theta_k\|_2^2 - 2\gamma \left\langle x^k - x^*, \nabla f(x^k) + \theta_k\right\rangle \\
\overset{(5)}{\leq} \;& \|x^k - x^*\|_2^2 - 2\gamma \left(1 - 2\gamma \left(\frac{1}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}\right)\left(f(x^k) - f(x^*)\right) + 2\gamma^2 \|\theta_k\|_2^2 \\
& -2\gamma \left\langle x^k - x^*, \theta_k\right\rangle + 2\gamma^2 \varepsilon^{\frac{2\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}},
\end{aligned}$$

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 $R_k \overset{\text{def}}{=} \|x^k - x^*\|_2$, $k > 0$ we derive that for all $k \geq 0$

$$\begin{aligned} R_{k+1}^2 &\leq R_k^2 - 2\gamma\left(1 - 2\gamma\left(\frac{1}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}\right)\left(f(x^k) - f(x^*)\right) + 2\gamma^2\|\theta_k\|_2^2 \\ &\quad -2\gamma\left\langle x^k - x^*, \theta_k\right\rangle + 2\gamma^2\varepsilon^{\frac{2\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} \end{aligned}$$

under assumption that $x^k \in B_{7R_0}(x^*)$. Let us define $A = 2\gamma\left(1 - 2\gamma\left(\frac{1}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}\right) \overset{(65)}{\geq} 2\gamma\left(1 - 2\cdot\frac{\varepsilon^{\frac{1-\nu}{1+\nu}}}{8M_\nu^{\frac{2}{1+\nu}}}\cdot\left(\frac{1}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}}\right.$ $\geq \gamma$ for $\varepsilon = \left(\frac{1}{8M_\nu}\right)^{1+\nu}$ and $\gamma \leq \frac{1}{2}\left(\frac{1}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} M_\nu^{-\frac{2}{1+\nu}} = \frac{1}{16}$. Since $\gamma > 0$, then

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

under assumption that $x^k \in B_{7R_0}(x^*)$. Summing up these inequalities for $k = 0, 1, \ldots, N-1$, we obtain

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

under assumption that $x^k \in B_{7R_0}(x^*)$. 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

$$\begin{aligned} AN\left(f(\bar{x}^N) - f(x^*)\right) &\leq R_0^2 - R_N^2 + 2\gamma^2 N\varepsilon^{\frac{2\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} + 2\gamma^2\sum_{k=0}^{N-1}\|\theta_k\|_2^2 \\ &\quad -2\gamma\sum_{k=0}^{N-1}\left\langle x^k - x^*, \theta_k\right\rangle \end{aligned} \tag{69}$$

under assumption that $x^k \in B_{7R_0}(x^*)$ for $k = 0, 1, \ldots, N-1$. Taking into account that $f(\bar{x}^N) - f(x^*) \geq 0$ and changing the indices we get that for all $k = 0, 1, \ldots, N$

$$R_k^2 \leq R_0^2 + 2\gamma^2 k\varepsilon^{\frac{2\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} + 2\gamma^2\sum_{l=0}^{k-1}\|\theta_l\|_2^2 - 2\gamma\sum_{l=0}^{k-1}\left\langle x^l - x^*, \theta_l\right\rangle. \tag{70}$$

under assumption that $x^l \in B_{7R_0}(x^*)$ for $l = 0, 1, \ldots, k-1$. The remaining part of the proof is based on the analysis of inequality (70). In particular, via induction we prove that for all $k = 0, 1, \ldots, N$ with probability at least $1 - \frac{k\beta}{N}$ we have $\mathbb{P}\{E_k\} = 1 - \frac{k\beta}{N}$ for probability event E_k defined as follows:

Event E_k:

Inequalities

$$R_t^2 \overset{(70)}{\leq} R_0^2 + 2\gamma^2 t\varepsilon^{\frac{2\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} + 2\gamma^2 \sum_{l=0}^{t-1} \|\theta_l\|_2^2 - 2\gamma \sum_{l=0}^{t-1} \left\langle x^l - x^*, \theta_l \right\rangle$$
$$\leq C^2 R_0^2 \tag{71}$$

hold for $t = 0, 1, \ldots, k$ simultaneously where C is defined in (67).

For $t = 0$ inequality (71) 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 $x^t \in B_{7R_0}(x^*)$ for $t = 0, 1, \ldots, T-1$, and, as a consequence, (70) holds for $k = T$. Since $\nabla f(x)$ is (ν, M_ν)-Hölder continuous on $B_{7R_0}(x^*)$, we have that probability event E_{T-1} implies

$$\left\|\nabla f(x^t)\right\|_2 \overset{(3)}{\leq} M_\nu \|x^t - x^0\|^\nu \overset{(71)}{\leq} M_\nu C^\nu R_0^\nu \overset{(65)}{\leq} \frac{\lambda}{2} \tag{72}$$

for $t = 0, 1, \ldots, T-1$. Next, we introduce new random variables:

$$\eta_l = \begin{cases} x^* - x^l, & \text{if } \|x^* - x^l\|_2 \leq CR_0, \\ 0, & \text{otherwise,} \end{cases} \tag{73}$$

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

$$\|\eta_l\|_2 \leq CR_0. \tag{74}$$

Using the introduced notation, we obtain that E_{T-1} implies

$$\begin{aligned} R_T^2 &\overset{(71),(65)}{\leq} R_0^2 + 2\left(\frac{R_0}{\sqrt{2N}\varepsilon^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}}}\right)^2 N\varepsilon^{\frac{2\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} \\ &\quad + 2\gamma^2 \sum_{l=0}^{t-1} \|\theta_l\|_2^2 - 2\gamma \sum_{l=0}^{t-1} \left\langle x^l - x^*, \theta_l \right\rangle \\ &= 2R_0^2 + 2\gamma^2 \sum_{l=0}^{t-1} \|\theta_l\|_2^2 - 2\gamma \sum_{l=0}^{t-1} \left\langle x^l - x^*, \theta_l \right\rangle \\ &\overset{(73),(74)}{=} 2R_0^2 + 2\gamma \sum_{l=0}^{T-1} \langle \theta_l, \eta_l \rangle + 2\gamma^2 \sum_{l=0}^{T-1} \|\theta_l\|_2^2. \end{aligned}$$

Finally, we do some preliminaries in order to apply Bernstein's inequality (see Lemma A.2) and obtain that E_{T-1} implies

$$\begin{aligned} R_T^2 \overset{(85)}{\leq} 2R_0^2 &+ \underbrace{2\gamma \sum_{l=0}^{T-1} \langle \theta_l^u, \eta_l \rangle}_{①} + \underbrace{2\gamma \sum_{l=0}^{T-1} \left\langle \theta_l^b, \eta_l \right\rangle}_{②} + \underbrace{4\gamma^2 \sum_{l=0}^{T-1} \left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right] \right)}_{③} \\ &+ \underbrace{4\gamma^2 \sum_{l=0}^{T-1} \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]}_{④} + \underbrace{4\gamma^2 \sum_{l=0}^{T-1} \|\theta_l^b\|_2^2}_{⑤}, \end{aligned} \tag{75}$$

where we introduce new notations:

$$\theta_l^u \stackrel{\text{def}}{=} \widetilde{\nabla} f(x^l, \boldsymbol{\xi}^l) - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\widetilde{\nabla} f(x^l, \boldsymbol{\xi}^l)\right], \quad \theta_l^b \stackrel{\text{def}}{=} \mathbb{E}_{\boldsymbol{\xi}^l}\left[\widetilde{\nabla} f(x^l, \boldsymbol{\xi}^l)\right] - \nabla f(x^l), \tag{76}$$

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

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: $\left|2\gamma \left\langle \theta_l^u, \eta_l \right\rangle\right| \leq 2\gamma \|\theta_l^u\|_2 \|\eta_l\|_2 \overset{(19),(74)}{\leq} 4\gamma\lambda C R_0$. Finally, one can bound conditional variances $\sigma_l^2 \stackrel{\text{def}}{=} \mathbb{E}_{\boldsymbol{\xi}^l}\left[4\gamma^2 \left\langle \theta_l^u, \eta_l \right\rangle^2\right]$ in the following way: $\sigma_l^2 \leq \mathbb{E}_{\boldsymbol{\xi}^l}\left[4\gamma^2 \|\theta_l^u\|_2^2 \|\eta_l\|_2^2\right] \overset{(74)}{\leq} 4\gamma^2 (CR_0)^2 \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]$, i.e., σ_l^2 is finite due to finiteness of $\|\theta_{l+1}^u\|_2$ (see Lemma 4.2). In other words, sequence $\left\{2\gamma \left\langle \theta_l^u, \eta_l \right\rangle\right\}_{l \geq 0}$ 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 A.2 with $X_l = 2\gamma \left\langle \theta_l^u, \eta_l \right\rangle$, $c = 4\gamma\lambda C R_0$ 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

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

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

$$\underbrace{\text{either } \sum_{l=0}^{T-1} \sigma_l^2 > F \quad \text{or} \quad |①| \leq 4\gamma\lambda C R_0 \ln \frac{4N}{\beta}}_{\text{probability event } E_①}.$$

Here and below, we notice that the conditions of Lemma 4.2 hold when E_{T-1} holds, since event E_{T-1} implies that $x^0, x^1, \ldots, x^T$ lie in $B_{7R_0}(x^*)$. Therefore, probability event E_{T-1} implies that

$$\begin{aligned}\sum_{l=0}^{T-1}\sigma_l^2 &\leq 4\gamma^2(CR_0)^2\sum_{l=0}^{T-1}\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right] \overset{(22)}{\leq} 72\gamma^2(CR_0)^2\sigma^2\frac{T}{m}\\ &\overset{T\leq N}{\leq} 72\gamma^2(CR_0)^2\sigma^2\frac{N}{m} \leq \frac{c^2\ln\frac{4N}{\beta}}{6} = F,\end{aligned}$$

where the last inequality follows from $c = 4\gamma\lambda CR_0$ and simple arithmetic.

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

$$2\gamma\left\langle\theta_l^b, \eta_l\right\rangle \leq 2\gamma\left\|\theta_l^b\right\|_2\|\eta_l\|_2 \overset{(20),(74)}{\leq} 2\gamma\frac{4\sigma^2}{m\lambda}CR_0 = \frac{8\gamma\sigma^2CR_0}{m\lambda}.$$

This implies that

$$② = 2\gamma\sum_{l=0}^{T-1}\left\langle\theta_l^b, \eta_l\right\rangle \overset{T\leq N}{\leq} \frac{8\gamma\sigma^2CR_0N}{m\lambda} \overset{(64)}{\leq} \frac{8}{81}\lambda\gamma CR_0\ln\frac{4N}{\beta}.$$

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 unbiased: $\mathbb{E}_{\boldsymbol{\xi}^l}\left[4\gamma^2\left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right)\right] = 0$. Second, the summands are bounded with probability 1:

$$\begin{aligned}\left|4\gamma^2\left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right)\right| &\leq 4\gamma^2\left(\|\theta_l^u\|_2^2 + \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right) \overset{(19)}{\leq} 4\gamma^2\left(4\lambda^2 + 4\lambda^2\right)\\ &= 32\gamma^2\lambda^2 \overset{\text{def}}{=} c_1.\end{aligned} \tag{77}$$

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:

$$\begin{aligned}\hat{\sigma}_l^2 &\overset{(77)}{\leq} c_1\mathbb{E}_{\boldsymbol{\xi}^l}\left[\left|4\gamma^2\left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right)\right|\right]\\ &\leq 4\gamma^2c_1\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2 + \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right] = 8\gamma^2c_1\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right],\end{aligned} \tag{78}$$

i.e., $\hat{\sigma}_l^2$ is finite due to finiteness of $\|\theta_{l+1}^u\|_2$ (see Lemma 4.2). 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 A.2 with $X_l = \hat{X}_l = 4\gamma^2\left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right]\right)$, $c = c_1 = 32\gamma^2\lambda^2$ and $F = F_1 = \frac{c_1^2\ln\frac{4N}{\beta}}{18}$ and get that for all $b > 0$

$$\mathbb{P}\left\{\left|\sum_{l=0}^{T-1}\hat{X}_l\right| > b \text{ and } \sum_{l=0}^{T-1}\hat{\sigma}_l^2 \leq F_1\right\} \leq 2\exp\left(-\frac{b^2}{2F_1 + 2c_1b/3}\right)$$

or, equivalently, with probability at least $1-2\exp\left(-\frac{b^2}{2F_1+2c_1b/3}\right)$

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

As in our derivations of the upper bound for ① we choose such b that $2\exp\left(-\frac{b^2}{2F_1+2c_1b/3}\right)=\frac{\beta}{2N}$, i.e.,

$$b = \frac{c_1\ln\frac{4N}{\beta}}{3}+\sqrt{\frac{c_1^2\ln^2\frac{4N}{\beta}}{9}+2F_1\ln\frac{4N}{\beta}} \le c_1\ln\frac{4N}{\beta} = 32\gamma^2\lambda^2\ln\frac{4N}{\beta}.$$

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

$$\underbrace{\text{either } \sum_{l=0}^{T-1}\hat{\sigma}_l^2 > F_1 \quad \text{or} \quad |③| \le 32\gamma^2\lambda^2\ln\frac{4N}{\beta}}_{\text{probability event } E_③}.$$

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

$$\begin{aligned}\sum_{l=0}^{T-1}\hat{\sigma}_l^2 &\overset{(78)}{\le} 8\gamma^2c_1\sum_{l=0}^{T-1}\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right] \overset{(22)}{\le} 144\gamma^2c_1\sigma^2\frac{T}{m}\\ &\overset{T\le N}{\le} 144\gamma^2c_1\sigma^2\frac{N}{m}\overset{(64)}{\le}\frac{144\gamma^2\lambda^2c_1\ln\frac{4N}{\beta}}{81} = \frac{c_1^2\ln\frac{4N}{\beta}}{18} = F_1.\end{aligned}$$

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

$$④ = 4\gamma^2\sum_{l=0}^{T-1}\mathbb{E}_{\boldsymbol{\xi}^l}\left[\|\theta_l^u\|_2^2\right] \overset{(22)}{\le} 72\gamma^2\sigma^2\sum_{l=0}^{T-1}\frac{1}{m} \overset{T\le N}{\le} \frac{72\gamma^2N\sigma^2}{m} \overset{(64)}{\le} \frac{8}{9}\lambda^2\gamma^2\ln\frac{4N}{\beta}.$$

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

$$⑤ = 4\gamma^2\sum_{l=0}^{T-1}\|\theta_l^b\|_2^2 \overset{(20)}{\le} 64\gamma^2\sigma^4\frac{T}{m^2\lambda^2} \overset{T\le N}{\le} 64\gamma^2\sigma^4\frac{N}{m^2\lambda^2} \overset{(64)}{\le} \frac{64}{6561}\frac{\lambda^2\gamma^2\ln^2\frac{4N}{\beta}}{N}.$$

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

$$R_T^2 \overset{(75)}{\le} 2R_0^2 + ① + ② + ③ + ④ + ⑤,$$

$$② \le \frac{8}{81}\lambda\gamma CR_0\ln\frac{4N}{\beta}, \quad ④ \le \frac{8}{9}\lambda^2\gamma^2\ln\frac{4N}{\beta}, \quad ⑤ \le \frac{64}{6561}\frac{\lambda^2\gamma^2\ln^2\frac{4N}{\beta}}{N},$$

$$\sum_{l=0}^{T-1}\sigma_l^2 \le F, \quad \sum_{l=0}^{T-1}\hat{\sigma}_l^2 \le F_1$$

and $\mathbb{P}\{E_{T-1}\} \geq 1 - \frac{(T-1)\beta}{N}$, $\mathbb{P}\{E_{①}\} \geq 1 - \frac{\beta}{2N}$, $\mathbb{P}\{E_{③}\} \geq 1 - \frac{\beta}{2N}$,

$$\text{where } E_{①} = \left\{ \text{either } \sum_{l=0}^{T-1} \sigma_l^2 > F \quad \text{or} \quad |①| \leq 4\gamma\lambda C R_0 \ln\frac{4N}{\beta} \right\},$$
$$E_{③} = \left\{ \text{either } \sum_{l=0}^{T-1} \hat{\sigma}_l^2 > F_1 \quad \text{or} \quad |③| \leq 32\gamma^2\lambda^2 \ln\frac{4N}{\beta} \right\}.$$

Taking into account these inequalities and our assumptions on λ and γ (see (64) and (65)) we get that probability event $E_{T-1} \cap E_{①} \cap E_{③}$ implies

$$R_T^2 \leq 2R_0^2 + \left(\frac{4}{7} + \frac{8}{567} + \frac{16}{49} + \frac{4}{441} + \frac{64}{321489}\right) C^2 R_0^2 \overset{(67)}{\leq} C^2 R_0^2. \tag{79}$$

Moreover, using the bound for the union, we derive

$$\mathbb{P}\left\{E_{T-1} \cap E_{①} \cap E_{③}\right\} = 1 - \mathbb{P}\left\{\overline{E}_{T-1} \cup \overline{E}_{①} \cup \overline{E}_{③}\right\} \geq 1 - \frac{T\beta}{N}. \tag{80}$$

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

$$\mathbb{P}\{E_T\} \overset{(79)}{\geq} \mathbb{P}\left\{E_{T-1} \cap E_{①} \cap E_{③}\right\} \overset{(80)}{\geq} 1 - \frac{T\beta}{N},$$

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 $1 - \beta$

$$\begin{aligned} AN\left(f(\bar{x}^N) - f(x^*)\right) &\overset{(69),(65)}{\leq} R_0^2 + 2\left(\frac{R_0}{\sqrt{2N}\varepsilon^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}}}\right)^2 N\varepsilon^{\frac{2\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} \\ &\quad + 2\gamma^2 \sum_{k=0}^{N-1} \|\theta_k\|_2^2 - 2\gamma \sum_{k=0}^{N-1} \left\langle x^k - x^*, \theta_k \right\rangle \\ &\leq 2R_0^2 + 2\gamma^2 \sum_{k=0}^{N-1} \|\theta_k\|_2^2 - 2\gamma \sum_{k=0}^{N-1} \left\langle x^k - x^*, \theta_k \right\rangle \\ &\overset{(71)}{\leq} C^2 R_0^2. \end{aligned}$$

Since $A = 2\gamma\left(1 - 2\gamma\left(\frac{1}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}\right) \overset{(65)}{\geq} \gamma$ we get that with probability at least $1 - \beta$, it holds that $f(\bar{x}^N) - f(x^*) \leq \frac{C^2 R_0^2}{AN} \leq \frac{C^2 R_0^2}{\gamma N}$. When

$$\gamma = \min\left\{\frac{\varepsilon^{\frac{1-\nu}{1+\nu}}}{8M_\nu^{\frac{2}{1+\nu}}}, \frac{R_0}{\sqrt{2N}\varepsilon^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}}}, \frac{R_0^{1-\nu}}{2C^\nu M_\nu \ln\frac{4N}{\beta}}\right\}$$

we have that with probability at least $1 - \beta$

$$\begin{aligned} &f(\bar{x}^N) - f(x^*) \\ &\leq \max\left\{\frac{8C^2 M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{1-\nu}{1+\nu}} N}, \frac{\sqrt{2}C^2 M_\nu^{\frac{1}{1+\nu}} R_0 \varepsilon^{\frac{\nu}{1+\nu}}}{\sqrt{N}}, \frac{2C^{2+\nu} M_\nu R_0^{1+\nu} \ln\frac{4N}{\beta}}{N}\right\}. \end{aligned}$$

Next, we estimate the iteration and oracle complexities of the method and consider 3 possible situations.

1. If $\gamma = \frac{\varepsilon^{\frac{1-\nu}{1+\nu}}}{8M_\nu^{\frac{2}{1+\nu}}}$, then with probability at least $1-\beta$

$$f(\bar{x}^N) - f(x^*) \leq \frac{8C^2 M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{1-\nu}{1+\nu}} N}.$$

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

$$\mathcal{O}\left(\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{2}{1+\nu}}}\right)$$

iterations and requires

$$\begin{aligned} Nm &\overset{(64)}{=} \mathcal{O}\left(\max\left\{N, \frac{N^2\sigma^2\gamma^2 \ln\frac{N}{\beta}}{R_0^2}\right\}\right) \\ &= \mathcal{O}\left(\max\left\{N, \frac{N^2\varepsilon^{\frac{2(1-\nu)}{1+\nu}}\sigma^2 \ln\frac{N}{\beta}}{M_\nu^{\frac{4}{1+\nu}} R_0^2}\right\}\right) \\ &= \mathcal{O}\left(\max\left\{\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{2}{1+\nu}}}, \frac{\sigma^2 R_0^2}{\varepsilon^2} \ln\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{2}{1+\nu}}\beta}\right\}\right) \end{aligned}$$

oracle calls.

2. If $\gamma = \frac{R_0}{\sqrt{2N}\varepsilon^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}}}$, then with probability at least $1-\beta$

$$f(\bar{x}^N) - f(x^*) \leq \frac{\sqrt{2}C^2 M_\nu^{\frac{1}{1+\nu}} R_0 \varepsilon^{\frac{\nu}{1+\nu}}}{\sqrt{N}}.$$

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

$$\mathcal{O}\left(\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{2}{1+\nu}}}\right)$$

iterations and requires

$$\begin{aligned} Nm &\overset{(64)}{=} \mathcal{O}\left(\max\left\{N, \frac{N^2\sigma^2\gamma^2 \ln\frac{N}{\beta}}{R_0^2}\right\}\right) = \mathcal{O}\left(\max\left\{N, \frac{N\sigma^2 \ln\frac{N}{\beta}}{\varepsilon^{\frac{2\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}}\right\}\right) \\ &= \mathcal{O}\left(\max\left\{\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{2}{1+\nu}}}, \frac{\sigma^2 R_0^2}{\varepsilon^2} \ln\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{2}{1+\nu}}\beta}\right\}\right) \end{aligned}$$

oracle calls.

3. If $\gamma = \frac{R_0^{1-\nu}}{2C^\nu M_\nu \ln \frac{4N}{\beta}}$, then with probability at least $1-\beta$

$$f(\bar{x}^N) - f(x^*) \leq \frac{2C^{2+\nu} M_\nu R_0^{1+\nu} \ln \frac{4N}{\beta}}{N}.$$

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

$$\mathcal{O}\left(\frac{M_\nu R_0^{1+\nu} \ln \frac{M_\nu R_0^{1+\nu}}{\varepsilon\beta}}{\varepsilon}\right)$$

iterations and requires

$$\begin{aligned} Nm &\overset{(64)}{=} \mathcal{O}\left(\max\left\{N, \frac{N^2\sigma^2\gamma^2 \ln \frac{N}{\beta}}{R_0^2}\right\}\right) = \mathcal{O}\left(\max\left\{N, \frac{N^2\sigma^2}{M_\nu^2 R_0^{2\nu} \ln \frac{N}{\beta}}\right\}\right) \\ &= \mathcal{O}\left(\max\left\{\frac{M_\nu R_0^{1+\nu}}{\varepsilon}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\} \ln \frac{M_\nu R_0^{1+\nu}}{\varepsilon\beta}\right) \end{aligned}$$

oracle calls.

Putting all together and noticing that $\ln \frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{2}{1+\nu}}\beta} = \mathcal{O}\left(\ln \frac{M_\nu R_0^{1+\nu}}{\varepsilon\beta}\right)$ we get the desired result. □ □

As for clipped-SSTM, it is possible to get rid of using large batch sizes without sacrificing the oracle complexity via a proper choice of γ.

Corollary 5.1. *Let the assumptions of Theorem 5.1 hold and*

$$\gamma = \min\left\{\frac{\varepsilon^{\frac{1-\nu}{1+\nu}}}{8M_\nu^{\frac{2}{1+\nu}}}, \frac{R_0}{\sqrt{2N}\varepsilon^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}}}, \frac{R_0^{1-\nu}}{2C^\nu M_\nu \ln \frac{4N}{\beta}}, \frac{R_0}{9\sigma\sqrt{N \ln \frac{4N}{\beta}}}\right\}. \tag{81}$$

Then for all $k = 0, 1, \ldots, N-1$ one can use $m = 1$ and to achieve $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$ clipped-SGD *requires*

$$\mathcal{O}\left(\max\left\{\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{2}{1+\nu}}}, \max\left\{\frac{M_\nu R_0^{1+\nu}}{\varepsilon}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\} \ln\left(\frac{M_\nu R_0^{1+\nu}}{\varepsilon\beta} + \frac{\sigma^2 R_0^2}{\varepsilon^2\beta}\right)\right\}\right) \tag{82}$$

iterations/oracle calls.

Proof. First of all, we verify that $m = 1$ is a valid choice. The only assumption on m is given in (64): $m \overset{(64)}{\geq} \max\left\{1, \frac{81N\sigma^2}{\lambda^2 \ln \frac{4N}{\beta}}\right\}$. Since $\gamma \leq \frac{R_0}{9\sigma\sqrt{N \ln \frac{4N}{\beta}}}$, we have

$$\max\left\{1, \frac{81N\sigma^2}{\lambda^2 \ln \frac{4N}{\beta}}\right\} \overset{(64)}{=} \max\left\{1, \frac{81\gamma^2\sigma^2 N \ln \frac{4N}{\beta}}{R_0^2}\right\} \leq 1$$

Therefore, for γ given in (81) one can use $m = 1$.

Next, if the minimum in (81) is attained on any of the first three terms, then applying the derivations from the end of the proof of Theorem 5.1, we get that the method requires

$$\mathcal{O}\left(\max\left\{\frac{M_\nu^{\frac{2}{1+\nu}}R_0^2}{\varepsilon^{\frac{2}{1+\nu}}}, \max\left\{\frac{M_\nu R_0^{1+\nu}}{\varepsilon}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\}\ln\frac{M_\nu R_0^{1+\nu}}{\varepsilon\beta}\right\}\right)$$

iterations/oracle calls to achieve $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$. If $\gamma = \frac{R_0}{9\sigma\sqrt{N\ln\frac{4N}{\beta}}}$, then with probability at least $1-\beta$

$$f(\bar{x}^N) - f(x^*) \overset{(66)}{\leq} \frac{9C^2 R_0\sigma\sqrt{\ln\frac{4N}{\beta}}}{\sqrt{N}}.$$

In other words, clipped-SGD achieves $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$ after[7]

$$\mathcal{O}\left(\frac{\sigma^2 R_0^2 \ln\frac{\sigma^2 R_0^2}{\varepsilon^2\beta}}{\varepsilon^2}\right)$$

iterations/oracle calls. Putting all together, we get the desired result. □ □

5.2 Strongly Convex Case

In this section, we provide a full statement of Theorem 3.2 together with its proof. Note that due to strong convexity, the solution x^* is unique.

Theorem 5.2. *Assume that function f is μ-strongly convex, its stochastic gradient and its gradient satisfy (2) and (3) respectively with $\sigma > 0$, $\nu \in [0,1]$, $M_\nu > 0$ on $Q = B_{7R_0}(x^*)$, where $R_0 \geq \|x^0 - x^*\|_2$. Let $\varepsilon > 0$, $\beta \in (0,1)$, and for all $t = 1, \ldots, \tau$, where $\tau = \left\lceil \log_2 \frac{\mu R_0^2}{\varepsilon}\right\rceil - 1$,*

$$N_t = \max\left\{\frac{2C^4 M_\nu^{\frac{2}{1+\nu}} R_0^2}{2^t \varepsilon_t^{\frac{2}{1+\nu}}}, \frac{4C^{2+\nu} M_\nu R_0^{1+\nu} \ln\frac{16C^{2+\nu}M_\nu R_0^{1+\nu}}{2^{\frac{(1+\nu)t}{2}}\varepsilon_t\beta}}{2^{\frac{(1+\nu)t}{2}}\varepsilon_t}\right\}, \quad \varepsilon_t = \frac{\mu R_0^2}{2^{t+1}},$$

$$\lambda_t = \frac{R_0}{2^{\frac{t}{2}}\gamma_t \ln\frac{4N_t\tau}{\beta}}, \quad m_t \geq \max\left\{1, \frac{81 N_t\sigma^2}{\lambda_t^2 \ln\frac{4N_t\tau}{\beta}}\right\}, \quad \ln\frac{4N_t\tau}{\beta} \geq 2,$$

$$\gamma_t = \min\left\{\frac{\varepsilon_t^{\frac{1-\nu}{1+\nu}}}{8M_\nu^{\frac{2}{1+\nu}}}, \frac{R_0}{2^{\frac{t}{2}}\sqrt{2N_t}\varepsilon_t^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}}}, \frac{R_0^{1-\nu}}{2^{1+\frac{(1-\nu)t}{2}} C^\nu M_\nu \ln\frac{4N_t\tau}{\beta}}\right\}.$$

[7] To achieve $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ it is sufficient to take N such that $\frac{9C^2R_0\sigma\sqrt{\ln\frac{4N}{\beta}}}{\sqrt{N}} \leq \varepsilon$. Solving this inequality w.r.t. N, we get that it is sufficient to take N such that $N \geq \frac{81C^4\sigma^2R_0^2\ln\frac{4N}{\beta}}{\varepsilon^2}$, e.g., $N = \left\lceil \frac{162C^4\sigma^2R_0^2\ln\left(\frac{648C^4\sigma^2R_0^2}{\varepsilon^2\beta}\right)}{\varepsilon^2}\right\rceil$ satisfies this inequality.

Then R-clipped-SGD *achieves* $f(\bar{x}^\tau) - f(x^*) \leq \varepsilon$ *with probability at least* $1-\beta$ *after*

$$\mathcal{O}\left(\max\left\{D_1^{\frac{2}{1+\nu}} \ln \frac{\mu R_0^2}{\varepsilon}, D_2^{\frac{2}{1+\nu}}, \max\left\{D_1 \ln \frac{\mu R_0^2}{\varepsilon}, D_2\right\} \ln \frac{D}{\beta}\right\}\right)$$

iterations of Algorithm 3 in total and requires

$$\mathcal{O}\left(\max\left\{D_1^{\frac{2}{1+\nu}} \ln \frac{\mu R_0^2}{\varepsilon}, D_2^{\frac{2}{1+\nu}}, \max\left\{D_1 \ln \frac{\mu R_0^2}{\varepsilon}, D_2, \frac{\sigma^2}{\mu\varepsilon}\right\} \ln \frac{D}{\beta}\right\}\right) \tag{83}$$

oracle calls, where

$$D_1 = \frac{M_\nu}{\mu R_0^{1-\nu}}, \quad D_2 = \frac{M_\nu}{\mu^{\frac{1+\nu}{2}} \varepsilon^{\frac{1-\nu}{2}}}, \quad D = D_2 \ln \frac{\mu R_0^2}{\varepsilon}.$$

Proof. Applying Theorem 5.1, we obtain that with probability at least $1 - \frac{\beta}{\tau}$ it holds that $f(\hat{x}^1) - f(x^*) \leq \frac{\mu R_0^2}{4}$. Since f is μ-strongly convex we have $\frac{\mu\|\hat{x}^1 - x^*\|_2^2}{2} \leq f(\hat{x}^1) - f(x^*)$. Therefore, with probability at least $1 - \frac{\beta}{\tau}$

$$f(\hat{x}^1) - f(x^*) \leq \frac{\mu R_0^2}{4}, \quad \|\hat{x}^1 - x^*\|_2^2 \leq \frac{R_0^2}{2}.$$

From mathematical induction and the union bound for probability events, it follows that inequalities $f(\hat{x}^t) - f(x^*) \leq \frac{\mu R_0^2}{2^{t+1}}, \quad \|\hat{x}^t - x^*\|_2^2 \leq \frac{R_0^2}{2^t}$ hold simultaneously for $t = 1, \ldots, \tau$ with probability at least $1-\beta$. In particular, it means that after $\tau = \left\lceil \log_2 \frac{\mu R_0^2}{\varepsilon} \right\rceil - 1$ restarts R-clipped-SGD finds an ε-solution with probability at least $1-\beta$. The total number of iterations $\hat{N}$ is

$$\begin{aligned}
\sum_{t=1}^{\tau} N_t &= \mathcal{O}\left(\sum_{t=1}^{\tau} \max\left\{\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{2^t \varepsilon_t^{\frac{2}{1+\nu}}}, \frac{M_\nu R_0^{1+\nu}}{2^{\frac{(1+\nu)t}{2}} \varepsilon_t} \ln \frac{M_\nu R_0^{1+\nu} \tau}{2^{\frac{(1+\nu)t}{2}} \varepsilon_t \beta}\right\}\right) \\
&= \mathcal{O}\left(\sum_{t=1}^{\tau} \max\left\{\frac{M_\nu^{\frac{2}{1+\nu}} \cdot 2^{\frac{(1-\nu)t}{1+\nu}}}{\mu^{\frac{2}{1+\nu}} R_0^{\frac{2(1-\nu)}{1+\nu}}}, \frac{M_\nu \cdot 2^{\frac{(1-\nu)t}{2}}}{\mu R_0^{1-\nu}} \ln \frac{M_\nu \cdot 2^{\frac{(1-\nu)\tau}{2}} \tau}{\mu R_0^{1-\nu} \beta}\right\}\right) \\
&= \mathcal{O}\left(\max\left\{\frac{M_\nu^{\frac{2}{1+\nu}}}{\mu^{\frac{2}{1+\nu}} R_0^{\frac{2(1-\nu)}{1+\nu}}}, \frac{M_\nu}{\mu R_0^{1-\nu}} \ln \frac{M_\nu \ln \frac{\mu R_0^2}{\varepsilon}}{\mu^{\frac{1+\nu}{2}} \varepsilon^{\frac{1-\nu}{2}} \beta}\right\} \cdot \left(\frac{\mu R_0^2}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}}\right) \\
&= \mathcal{O}\left(\max\left\{D_1^{\frac{2}{1+\nu}} \ln \frac{\mu R_0^2}{\varepsilon}, D_2^{\frac{2}{1+\nu}}, \max\left\{D_1 \ln \frac{\mu R_0^2}{\varepsilon}, D_2\right\} \ln \frac{D}{\beta}\right\}\right),
\end{aligned}$$

where

$$D_1 = \frac{M_\nu}{\mu R_0^{1-\nu}}, \quad D_2 = \frac{M_\nu}{\mu^{\frac{1+\nu}{2}} \varepsilon^{\frac{1-\nu}{2}}}, \quad D = D_2 \ln \frac{\mu R_0^2}{\varepsilon}.$$

Finally, the total number of oracle calls equals

$$\begin{aligned}
\sum_{t=1}^{\tau} \sum_{k=0}^{N_t - 1} m_k^t &= \mathcal{O}\left(\max\left\{\sum_{t=1}^{\tau} N_t, \sum_{t=1}^{\tau} \frac{\sigma^2 R_0^2}{2^t \varepsilon_t^2} \ln \frac{M_\nu R_0^{1+\nu} \tau}{2^{\frac{(1+\nu)t}{2}} \varepsilon_t \beta}\right\}\right) \\
&= \mathcal{O}\left(\max\left\{\hat{N}, \sum_{t=1}^{\tau} \frac{\sigma^2 \cdot 2^t}{\mu^2 R_0^2} \ln \frac{D}{\beta}\right\}\right) = \mathcal{O}\left(\max\left\{\hat{N}, \frac{\sigma^2}{\mu\varepsilon} \ln \frac{D}{\beta}\right\}\right). \quad \square
\end{aligned}$$

$\square$

One can also derive a similar result for R-clipped-SGD when stepsize γ is chosen as in Corollary 5.1 for all restarts. In this case, one can choose unit batch sizes: $m_t = 1$ for all t.

6 Numerical Experiments

In this section, we present the results of our numerical experiments in synthetic and real-world data. We defer additional details regarding the choice of parameters to Appendix B.

6.1 Experiments on Synthetic Data

First of all, we tested the considered methods on the following problem, which corresponds to the linear regression with the noise having generalized Gaussian distribution (Example 4.4 from (Chaux et al., 2007)):

$$\min_{x\in\mathbb{R}^n}\left\{f_p(x) = \frac{1}{m}\|\mathbf{A}x - y\|_p^p =: \frac{1}{m}\sum_{i=1}^m f_{i,p}(x)\right\}, \quad f_{i,p}(x) = |a_i^\top x - y_i|^p, \tag{84}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $y \in \mathbb{R}^m$, $p \in [1,2]$, and $a_i^\top$ denotes the i-th row of matrix $\mathbf{A}$. One can show that $f_p(x)$ is convex and has (ν, M_ν)-Hölder continuous (sub)gradient[8] with $\nu = p-1$ and $M_\nu = \frac{2^{1-\nu}(1+\nu)}{m}\sum_{i=1}^m \|a_i\|_2^{1+\nu}$. Moreover, to rewrite the considered problem in the form (1), we define ξ as a random index having a uniform distribution on $\{1,\ldots,m\}$. Since the (sub)gradient of f_i is bounded on any compact set and any $i \in \{1,\ldots,m\}$, the variance of the stochastic gradient can be uniformly upper bounded on any compact set as well. That is, problem (84) fits the setup we consider in the theoretical analysis in the previous sections.

We generate matrix $\mathbf{A}$ as follows: 1) assemble the matrix $\mathbf{A}_1 \in \mathbb{R}^{m\times n}$ from mn i.i.d. samples from the standard Gaussian distribution $\mathcal{N}(0,1)$, 2) multiply the rows of matrix $\mathbf{A}_1$ on i.i.d. samples from the Levy distribution with parameters 0 and 0.5 that are smaller than the threshold $t = 10^4$ (we redraw a sample if it is larger than t to avoid numerical instabilities during the experiments), denote the result by $\mathbf{A}_2$, 3) divide the columns of $\mathbf{A}_2$ by their empirical standard deviations (again due to numerical instabilities), denote the result by $\mathbf{A}$, 4) split the dataset equally into the train and test sets and add i.i.d. samples from the Levy distribution with parameters 0 and 0.5 to the train part (we redraw a sample if it is larger than $t\cdot\alpha$ with $\alpha = 10$). Next, we generate $x = x_{\text{true}}$ as a random vector from the uniform distribution on the unit Euclidean sphere and set $y := y_{\text{true}} = \mathbf{A}x_{\text{true}}$. We use $m = 10000$, $n = 200$ (5000 for the train set and 5000 for the test set). The starting point for the methods was generated from the uniform distribution on the Euclidean sphere with radius 10. We also use x_{pred} to denote the output of a method and $y_{\text{pred}} = \mathbf{A}x_{\text{pred}}$ to denote the "answer" of the trained model.

The resulting problem has a heavy-tailed stochastic gradient noise. To illustrate this, for different values of p, we sample a large enough number of batched stochastic gradients $\nabla f_p(x^0,\boldsymbol{\xi}_1),\ldots,\nabla f_p(x^0,\boldsymbol{\xi}_K)$ with the batch size we use to run the methods and plot the histograms for $\|\nabla f_p(x^0,\boldsymbol{\xi}_1) - \nabla f_p(x^0)\|_2,\ldots,\|\nabla f_p(x^0,\boldsymbol{\xi}_K) - \nabla f_p(x^0)\|_2$, see Figure 1.

[8] For $p \in (1,2]$ function $f_{i,p}(x)$ is differentiable and $\nabla f_{i,p}(x) = p|a_i^\top x - y_i|^{p-1}\text{sign}(a_i^\top x - y_i)a_i$ and for $p = 1$ it has subdifferential $\partial f_{i,p}(x) = \begin{cases} a_i, & \text{if } a_i^\top x - y_i > 0, \\ [-a_i, a_i], & \text{if } a_i^\top x - y_i = 0, \\ -a_i, & \text{if } a_i^\top x - y_i < 0. \end{cases}$

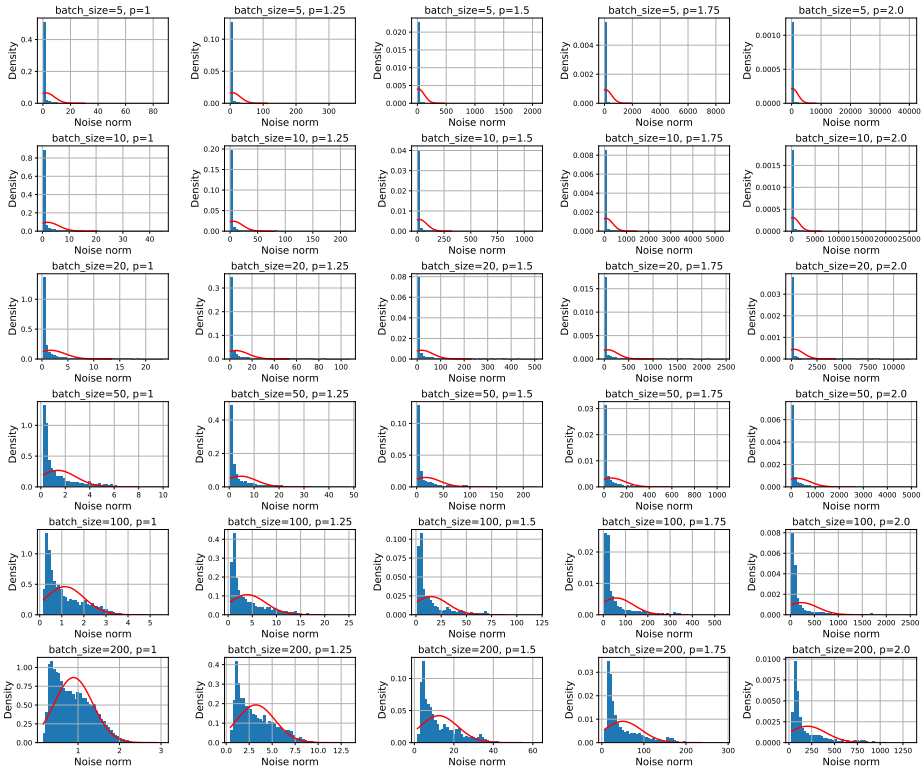

Figure 1: Noise distribution of the stochastic gradients for synthetic dataset, depending on batch size and p of the loss function (84). Red lines: Gaussian probability density functions with means and variances empirically estimated by the samples. The total number of batches for each graph is $5 \cdot 10^5$.

We compared 4 different methods on this problem with different p: Adam, SGD, clipped-SGD, and clipped-SSTM. The results w.r.t. the best relative loss achieved on the training dataset are reported in Figure 2. In all our experiments, clipped-SSTM performs significantly better than other tested methods for all values of p. We also observe that for $p < 1.5$ SGD has a comparable or even faster convergence than clipped-SGD, while for larger values of p SGD is much slower than clipped-SGD. Taking into account the noise distributions reported in Figure 1, this behavior is expected since the stochastic gradient noise in the considered problem has heavier tails due to the specifics of the dataset generation. We also notice that, in this series of experiments, Adam is never faster than clipped-SSTM and, moreover, for $p \geq 1.5$ Adam converges slower than clipped-SGD. Additionally, we compared these methods w.r.t. the number of epochs needed to achieve a 2.0 relative loss on the train, the results are reported in Appendix B.1.2.

6.2 Neural Networks Training

In our experiments with the training of neural networks, we tested the performance of the methods on the following non-convex non-smooth problems[9] (in both tasks, we use standard cross-entropy

[9]We conduct these experiments to illustrate that clipped-SSTM and clipped-SGD might be useful even for the problems that are not theoretically studied in this paper. Since (Gorbunov et al., 2020) does not provide numerical

loss functions):

- `BERT` fine-tuning on `CoLA` dataset (Warstadt et al., 2019). We use pretrained `BERT` from Transformers library (Wolf et al., 2020) (`bert-base-uncased`) and freeze all layers except the last two linear ones.
- `ResNet-18` training on `ImageNet-100` (first 100 classes of `ImageNet` (Russakovsky et al., 2015)).

First, we study the noise distribution for both problems as follows: at the starting point we sample large enough number of batched stochastic gradients $\nabla f(x^0, \boldsymbol{\xi}_1), \ldots, \nabla f(x^0, \boldsymbol{\xi}_K)$ with batch size 32 and plot the histograms for $\|\nabla f(x^0, \boldsymbol{\xi}_1) - \nabla f(x^0)\|_2, \ldots, \|\nabla f(x^0, \boldsymbol{\xi}_K) - \nabla f(x^0)\|_2$, see Figure 3. As one can see, the noise distribution for `BERT` + `CoLA` is substantially non-sub-Gaussian, whereas the distribution for `ResNet-18` + `Imagenet-100` is almost Gaussian. We observe a similar phenomenon for other points along the trajectories of the methods; see Appendix B.2.3.

Next, we compared four different optimizers on these problems: Adam, SGD (with Momentum), clipped-SGD (with Momentum and coordinate-wise[10] clipping) and clipped-SSTM (with norm-clipping and $\nu = 1$). The results are presented in Figure 4. We observed that the noise distributions do not change significantly along the trajectories of the considered methods, see Appendix B. During the hyper-parameters search, we compared different batch sizes emulated via gradient accumulation (thus, we compare methods with different batch sizes by the number of base batches used). The base batch size was 32 for both problems; stepsizes and clipping levels were tuned. One can find additional details regarding our experiments in Appendix B.

Image classification. On `ResNet-18` + `ImageNet-100` task, SGD performs relatively well, and even ties with Adam (with batch size of 4×32) in validation loss. clipped-SSTM (with batch size of 2×32) also ties with Adam and clipped-SGD is not far from them. The results were averaged from 5 different launches (with different starting points/weight initializations). Since the noise distribution is almost Gaussian, even vanilla SGD performs well, i.e., gradient clipping is not required. At the same time, the clipping does not slow down the convergence significantly.

Text classification. On `BERT` + `CoLA` task, when the noise distribution is heavy-tailed, the methods with clipping outperform SGD by a large margin. This result is in good correspondence with the derived high-probability complexity bounds for clipped-SGD, clipped-SSTM, and the best-known ones for SGD. Moreover, clipped-SSTM (with batch size of 8×32) achieves the same loss on validation as Adam, and has better accuracy. These results were averaged from 5 different train-val splits and 20 launches (with different starting points/weight initializations) for each of the splits, 100 launches in total. We provide additional experiments with different NLP tasks in Appendix B.2.4.

experiments with clipped-SSTM on the training of neural networks, our experiments are the first ones showing the behavior of clipped-SSTM on the considered tasks.

[10] Following standard practice in the usage of clipping, we use coordinate-wise clipping in clipped-SGD Zhang et al. (2020b). In the preliminary experiments, we also tried norm-clipping for clipped-SGD, but it showed worse results than the coordinate-wise one. Our analysis can be generalized to the case of coordinate-wise clipping if we assume the boundedness of the coordinate-wise variance σ_c^2 of the stochastic gradients. Then, the result of Lemma 4.2 will hold with $\sigma^2 = n\sigma_c^2$, and the norm of the clipped vector will be bounded by $\sqrt{n}\lambda$. These changes will lead to the explicit dependence on the dimension in the complexity bounds, similarly to Zhang et al. (2020b).

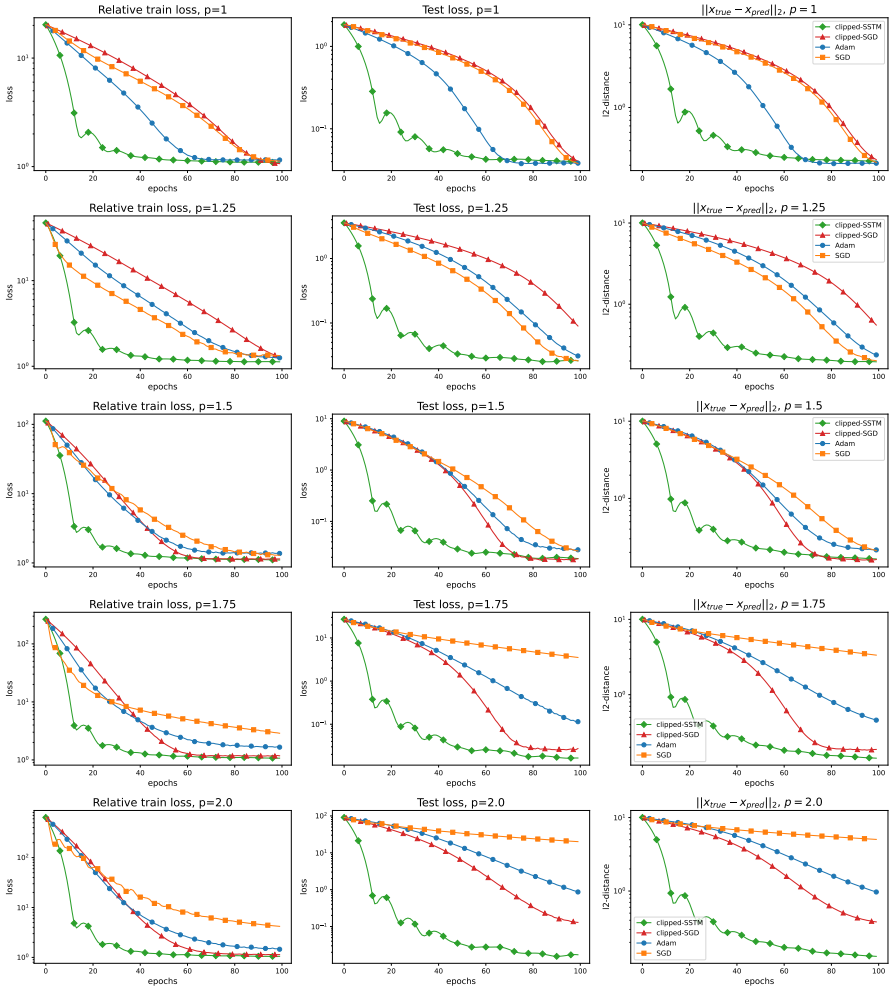

Figure 2: Results obtained for different p by the best relative train loss achieved. To calculate relative loss, we use $f_p(x_{\text{pred}})/f_p(x_{\text{true}})$, where $f_p(x_{true})$ is non-zero because of the noise added to the train part of the dataset.

Acknowledgements

This work was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730324P540002) and the agreement with the Moscow Institute of Physics and Technology dated November 1, 2021 No. 70-2021-00138.

Data availability statement The codes for the conducted numerical experiments are publicly available:
https://github.com/ClippedStochasticMethods/clipped-SSTM.

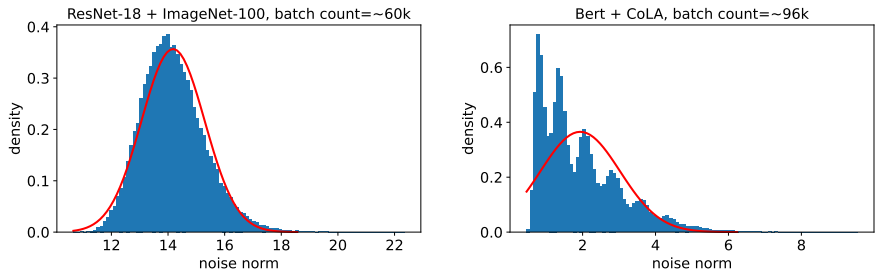

Figure 3: Noise distribution of the stochastic gradients for `ResNet-18` on `ImageNet-100` and `BERT` fine-tuning on the `CoLA` dataset before the training. Red lines: Gaussian probability density functions with means and variances empirically estimated by the samples. Batch count is the total number of samples used to build a histogram.

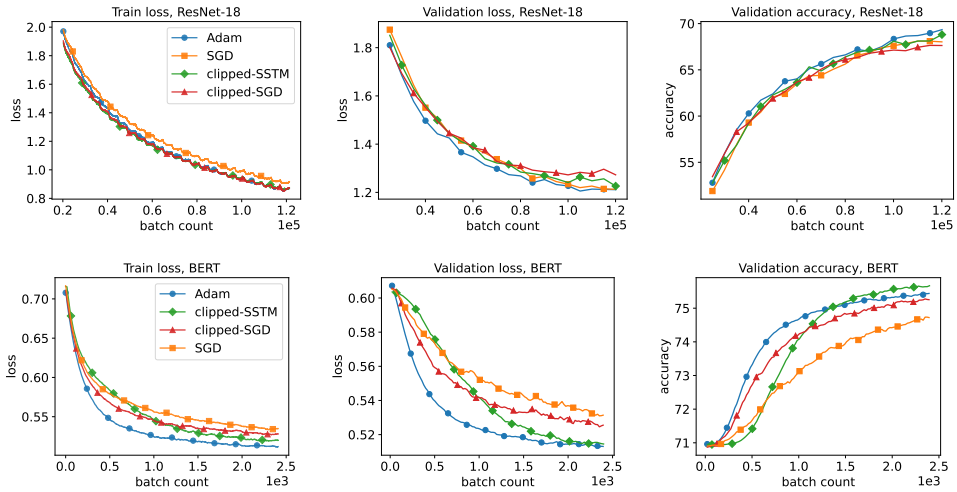

Figure 4: Train and validation loss + accuracy for different optimizers on both problems. Here, "batch count" denotes the total number of used stochastic gradients.

A Basic Facts, Technical Lemmas, and Auxiliary Results

A.1 Useful Inequalities

For all $a, b \in \mathbb{R}^n$

$$\|a+b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2, \tag{85}$$

$$\langle a, b\rangle = \frac{1}{2}\left(\|a+b\|_2^2 - \|a\|_2^2 - \|b\|_2^2\right). \tag{86}$$

A.2 Auxiliary Lemmas

The following lemma is a standard result about functions with (ν, M_ν)-Hölder continuous gradient (Devolder et al., 2014; Nesterov, 2015).

Lemma A.1. *Let f has (ν, M_ν)-Hölder continuous gradient on $Q \subseteq \mathbb{R}^n$. Then for all $x, y \in Q$ and for all $\delta > 0$*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M_\nu}{1+\nu} \|x - y\|_2^{1+\nu}, \tag{87}$$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L(\delta, \nu)}{2} \|x - y\|_2^2 + \frac{\delta}{2}, \quad L(\delta, \nu) = \left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}. \tag{88}$$

The next result is known as Bernstein inequality for martingale differences (Bennett, 1962; Dzhaparidze and Van Zanten, 2001; Freedman et al., 1975).

Lemma A.2. *Let the sequence of random variables $\{X_i\}_{i \geq 1}$ form a martingale difference sequence, i.e. $\mathbb{E}[X_i \mid X_{i-1}, \ldots, X_1] = 0$ for all $i \geq 1$. Assume that conditional variances $\sigma_i^2 \overset{def}{=} \mathbb{E}\left[X_i^2 \mid X_{i-1}, \ldots, X_1\right]$ exist and are bounded and also assume 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 $n \geq 1$*

$$\mathbb{P}\left\{ \left| \sum_{i=1}^n X_i \right| > b \text{ and } \sum_{i=1}^n \sigma_i^2 \leq F \right\} \leq 2 \exp\left(-\frac{b^2}{2F + 2cb/3} \right). \tag{89}$$

A.3 Technical Lemmas

Lemma A.3. *Let sequences $\{\alpha_k\}_{k \geq 0}$ and $\{A_k\}_{k \geq 0}$ satisfy*

$$\alpha_0 = A_0 = 0, \quad \alpha_{k+1} = \frac{(k+1)^{\frac{2\nu}{1+\nu}} (\varepsilon/2)^{\frac{1-\nu}{1+\nu}}}{2^{\frac{2\nu}{1+\nu}} a M_\nu^{\frac{2}{1+\nu}}}, \quad A_{k+1} = A_k + \alpha_{k+1}, \quad a, \varepsilon, M_\nu > 0, \ \nu \in [0, 1] \tag{90}$$

for all $k \geq 0$. Then for all $k \geq 0$ we have

$$A_k \geq a L_k \alpha_k^2, \quad A_k \geq \frac{k^{\frac{1+3\nu}{1+\nu}} (\varepsilon/2)^{\frac{1-\nu}{1+\nu}}}{2^{\frac{1+3\nu}{1+\nu}} a M_\nu^{\frac{2}{1+\nu}}}, \tag{91}$$

where $L_0 = 0$ and for $k > 0$

$$L_k = \left(\frac{2A_k}{\alpha_k \varepsilon} \right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}. \tag{92}$$

Moreover, for all $k \geq 0$

$$A_k \leq \frac{k^{\frac{1+3\nu}{1+\nu}} (\varepsilon/2)^{\frac{1-\nu}{1+\nu}}}{2^{\frac{2\nu}{1+\nu}} a M_\nu^{\frac{2}{1+\nu}}}. \tag{93}$$

Proof. We start with deriving the second inequality from (91). The proof goes by induction. For $k = 0$, the inequality holds. Next, we assume that it holds for all $k \leq K$. Then,

$$A_{K+1} = A_K + \alpha_{K+1} \geq \frac{K^{\frac{1+3\nu}{1+\nu}} (\varepsilon/2)^{\frac{1-\nu}{1+\nu}}}{2^{\frac{1+3\nu}{1+\nu}} a M_\nu^{\frac{2}{1+\nu}}} + \frac{(K+1)^{\frac{2\nu}{1+\nu}} (\varepsilon/2)^{\frac{1-\nu}{1+\nu}}}{2^{\frac{2\nu}{1+\nu}} a M_\nu^{\frac{2}{1+\nu}}}.$$

Let us estimate the right-hand side of the previous inequality. We want to show that

$$\frac{K^{\frac{1+3\nu}{1+\nu}} (\varepsilon/2)^{\frac{1-\nu}{1+\nu}}}{2^{\frac{1+3\nu}{1+\nu}} a M_\nu^{\frac{2}{1+\nu}}} + \frac{(K+1)^{\frac{2\nu}{1+\nu}} (\varepsilon/2)^{\frac{1-\nu}{1+\nu}}}{2^{\frac{2\nu}{1+\nu}} a M_\nu^{\frac{2}{1+\nu}}} \quad \geq \quad \frac{(K+1)^{\frac{1+3\nu}{1+\nu}} (\varepsilon/2)^{\frac{1-\nu}{1+\nu}}}{2^{\frac{1+3\nu}{1+\nu}} a M_\nu^{\frac{2}{1+\nu}}}$$

that is equivalent to the inequality:

$$\frac{K^{\frac{1+3\nu}{1+\nu}}}{2} + (K+1)^{\frac{2\nu}{1+\nu}} \geq \frac{(K+1)^{\frac{1+3\nu}{1+\nu}}}{2} \iff \frac{K^{\frac{1+3\nu}{1+\nu}}}{2} \geq \frac{(K+1)^{\frac{2\nu}{1+\nu}}(K-1)}{2}.$$

If $K = 1$, it trivially holds. If $K > 1$, it is equivalent to

$$\frac{K}{K-1} \geq \left(\frac{K+1}{K}\right)^{2-\frac{2}{1+\nu}}.$$

Since $2 - \frac{2}{1+\nu}$ is monotonically increasing function for $\nu \in [0,1]$ we have that

$$\left(\frac{K+1}{K}\right)^{2-\frac{2}{1+\nu}} \leq \frac{K+1}{K} \leq \frac{K}{K-1}.$$

That is, the second inequality in (91) holds for $k = K+1$, and, as a consequence, it holds for all $k \geq 0$. Next, we derive the first part of (91). For $k = 0$, it trivially holds. For $k > 0$ we consider cases $\nu = 0$ and $\nu > 0$ separately. When $\nu = 0$ the inequality is equivalent to

$$1 \geq \frac{2a\alpha_k M_0^2}{\varepsilon}, \text{ where } \frac{2a\alpha_k M_0^2}{\varepsilon} \overset{(90)}{=} 1,$$

i.e., we have $A_k = aL_k\alpha_k^2$ for all $k \geq 0$. When $\nu > 0$ the first inequality in (91) is equivalent to

$$A_k \geq a^{\frac{1+\nu}{2\nu}} \alpha_k^{\frac{1+3\nu}{2\nu}} (\varepsilon/2)^{-\frac{1-\nu}{2\nu}} M_\nu^{\frac{1}{\nu}} \overset{(90)}{\iff} A_k \geq \frac{k^{\frac{1+3\nu}{1+\nu}} (\varepsilon/2)^{\frac{1-\nu}{1+\nu}}}{2^{\frac{1+3\nu}{1+\nu}} a M_\nu^{\frac{2}{1+\nu}}},$$

where the last inequality coincides with the second inequality from (91) that we derived earlier in the proof.

To finish the proof, it remains to derive (93). Again, the proof goes by induction. For $k = 0$ inequality (93) is trivial. Next, we assume that it holds for all $k \leq K$. Then,

$$A_{K+1} = A_K + \alpha_{K+1} \leq \frac{K^{\frac{1+3\nu}{1+\nu}} (\varepsilon/2)^{\frac{1-\nu}{1+\nu}}}{2^{\frac{2\nu}{1+\nu}} a M_\nu^{\frac{2}{1+\nu}}} + \frac{(K+1)^{\frac{2\nu}{1+\nu}} (\varepsilon/2)^{\frac{1-\nu}{1+\nu}}}{2^{\frac{2\nu}{1+\nu}} a M_\nu^{\frac{2}{1+\nu}}}.$$

Let us estimate the right-hand side of the previous inequality. We want to show that

$$\frac{K^{\frac{1+3\nu}{1+\nu}} (\varepsilon/2)^{\frac{1-\nu}{1+\nu}}}{2^{\frac{2\nu}{1+\nu}} a M_\nu^{\frac{2}{1+\nu}}} + \frac{(K+1)^{\frac{2\nu}{1+\nu}} (\varepsilon/2)^{\frac{1-\nu}{1+\nu}}}{2^{\frac{2\nu}{1+\nu}} a M_\nu^{\frac{2}{1+\nu}}} \leq \frac{(K+1)^{\frac{1+3\nu}{1+\nu}} (\varepsilon/2)^{\frac{1-\nu}{1+\nu}}}{2^{\frac{2\nu}{1+\nu}} a M_\nu^{\frac{2}{1+\nu}}}$$

that is equivalent to the inequality:

$$K^{\frac{1+3\nu}{1+\nu}} + (K+1)^{\frac{2\nu}{1+\nu}} \leq (K+1)^{\frac{1+3\nu}{1+\nu}}.$$

This inequality holds due to

$$K^{\frac{1+3\nu}{1+\nu}} \leq (K+1)^{\frac{2\nu}{1+\nu}} K.$$

That is, (93) holds for $k = K+1$, and, as a consequence, it holds for all $k \geq 0$. □ □

Lemma A.4. *Let f have Hölder continuous gradients on $\mathbb{R}^n$ for some $\nu \in [0,1]$ with constant $M_\nu > 0$, be convex and x^* be some minimum of $f(x)$ on $\mathbb{R}^n$. Then, for all $x \in \mathbb{R}^n$*

$$\|\nabla f(x)\|_2 \leq \left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}} \left(f(x) - f(x^*)\right)^{\frac{\nu}{1+\nu}}, \tag{94}$$

where for $\nu = 0$ we use $\left[\left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}}\right]_{\nu=0} := \lim_{\nu\to 0}\left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}} = 1$.

Proof. For $\nu = 0$ inequality (94) follows from (3) and[11] $\nabla f(x^*) = 0$. When $\nu > 0$ for arbitrary point $x \in \mathbb{R}^n$ we consider the point $y = x - \alpha\nabla f(x)$, where $\alpha = \left(\frac{\|\nabla f(x)\|_2^{1-\nu}}{M_\nu}\right)^{\frac{1}{\nu}}$. For the pair of points x, y we apply (87) and get

$$\begin{aligned} f(y) &\leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{M_\nu}{1+\nu}\|x-y\|_2^{1+\nu} \\ &= f(x) - \alpha\|\nabla f(x)\|^2 + \frac{\alpha^{\nu+1}M_\nu}{1+\nu}\|\nabla f(x)\|_2^{1+\nu} \\ &= f(x) - \frac{\|\nabla f(x)\|_2^{\frac{1+\nu}{\nu}}}{M_\nu^{\frac{1}{\nu}}} + \frac{\|\nabla f(x)\|_2^{\frac{1+\nu}{\nu}}}{(1+\nu)M_\nu^{\frac{1}{\nu}}} = f(x) - \frac{\nu\|\nabla f(x)\|_2^{\frac{1+\nu}{\nu}}}{(1+\nu)M_\nu^{\frac{1}{\nu}}} \end{aligned}$$

implying

$$\|\nabla f(x)\|_2 \leq \left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}} \left(f(x) - f(y)\right)^{\frac{\nu}{1+\nu}} \leq \left(\frac{1+\nu}{\nu}\right)^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}} \left(f(x) - f(x^*)\right)^{\frac{\nu}{1+\nu}}. \quad \square$$

$\square$

Lemma A.5. *Let f have Hölder continuous gradients on $\mathbb{R}^n$ for some $\nu \in [0,1]$ with constant $M_\nu > 0$, be convex and x^* be some minimum of $f(x)$ on $\mathbb{R}^n$. Then, for all $x \in \mathbb{R}^n$ and all $\delta > 0$,*

$$\|\nabla f(x)\|_2^2 \leq 2\left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} \left(f(x) - f(x^*)\right) + \delta^{\frac{2\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}. \tag{95}$$

Proof. For a given $\delta > 0$ we consider an arbitrary point $x \in Q$ and $y = x - \frac{1}{L(\delta,\nu)}\nabla f(x)$, where $L(\delta,\nu) = \left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}$. For the pair of points x, y we apply (88) and get

$$\begin{aligned} f(y) &\leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L(\delta,\nu)}{2}\|x-y\|_2^2 + \frac{\delta}{2} \\ &= f(x) - \frac{1}{2L(\delta,\nu)}\|\nabla f(x)\|_2^2 + \frac{\delta}{2} \end{aligned}$$

implying

$$\begin{aligned} \|\nabla f(x)\|_2^2 &\leq 2L(\delta,\nu)\left(f(x) - f(y)\right) + \delta L(\delta,\nu) \\ &\leq 2\left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} \left(f(x) - f(x^*)\right) + \delta^{\frac{2\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}. \quad \square \end{aligned}$$

$\square$

[11] When f is not differentiable, we use subgradients. In this case, 0 belongs to the subdifferential of f at the point x^*, and we take it as $\nabla f(x^*)$.

B Additional Experimental Details and Results

B.1 Experiments on Synthetic Data

B.1.1 Hyper-Parameters tuning

We grid-searched hyper-parameters for each model. Commonly for all models we considered batch sizes of $\{5, 10, 20, 50, 100, 200\}$ and stepsizes $lr \in [1e{-}1, 1e{-}5]$. As to model-specific parameters:

- for Adam we grid-search over $betas \in (\{0.8, 0.9, 0.95, 0.99\}, \{0.9, 0.99, 0.999\})$,
- for SGD — over $momentum \in \{0.8, 0.9, 0.99, 0.999\}$,
- for clipped-SSTM — over clipping parameter $B \in \{1e{-}0, 1e{-}1, 1e{-}2, 1e{-}3\}$,
- for clipped-SGD — over $momentum \in \{0.8, 0.9, 0.99, 0.999\}$ and clipping parameter $B \in \{1e{-}0, 1e{-}1, 1e{-}2, 1e{-}3\}$.

For clipped-SSTM we additionally use $\nu = 1$ and norm clipping (we did not gridsearch over it extensively; however, in our experiments on real data, these parameters were the best). For clipped-SGD we use coordinate-wise clipping.

For Adam, clipped-SSTM and clipped-SGD the best parameters for each p were approximately the same:

- Adam: $lr = 1e{-}3$, $betas = (0.9, 0.9)$ and batch size of 10
- clipped-SSTM: $lr = 1e{-}3$, $\nu = 1$, $B = 1e{-}2$, norm clipping and a batch size of 5
- clipped-SGD: $lr = 1e{-}3$ and $B = 1e{-}1$ or $lr = 1e{-}2$ and $B = 1e{-}2$, $momentum = 0.8$, coordinate-wise clipping and a batch size of 5

B.1.2 Comparison w.r.t. certain relative train loss level

In Figure 2, we reported the performance of the methods in terms of the best models w.r.t. train loss achieved. However, it is also interesting to compare the methods w.r.t. the speed they achieve a certain (2.0) level of relative loss on train ($f_p(x_{\text{pred}})/f_p(x_{\text{true}})$). This is a valid metric, since $f_p(x_{\text{true}})$ is non-zero, after adding noise to the train part of the dataset, and x_{true} is still a good approximation of the optimal solution. The results are represented in Figure 5. As in the previous set of experiments, one can see that clipped-SSTM outperforms other algorithms and achieves this 2.0 level of relative loss much faster, though later loses to Adam/clipped-SGD.

B.2 Neural Networks Training

B.2.1 Hyper-Parameters

In our experiments with the training of neural networks, we use standard implementations of Adam and SGD from PyTorch Paszke et al. (2019); we write only the parameters we changed from the default.

To conduct these experiments, we used Nvidia RTX 2070s. The longest experiment (evolution of the noise distribution for image classification task) took 53 hours (we iterated several times over the train dataset to build a better histogram; see Appendix B.2.3).

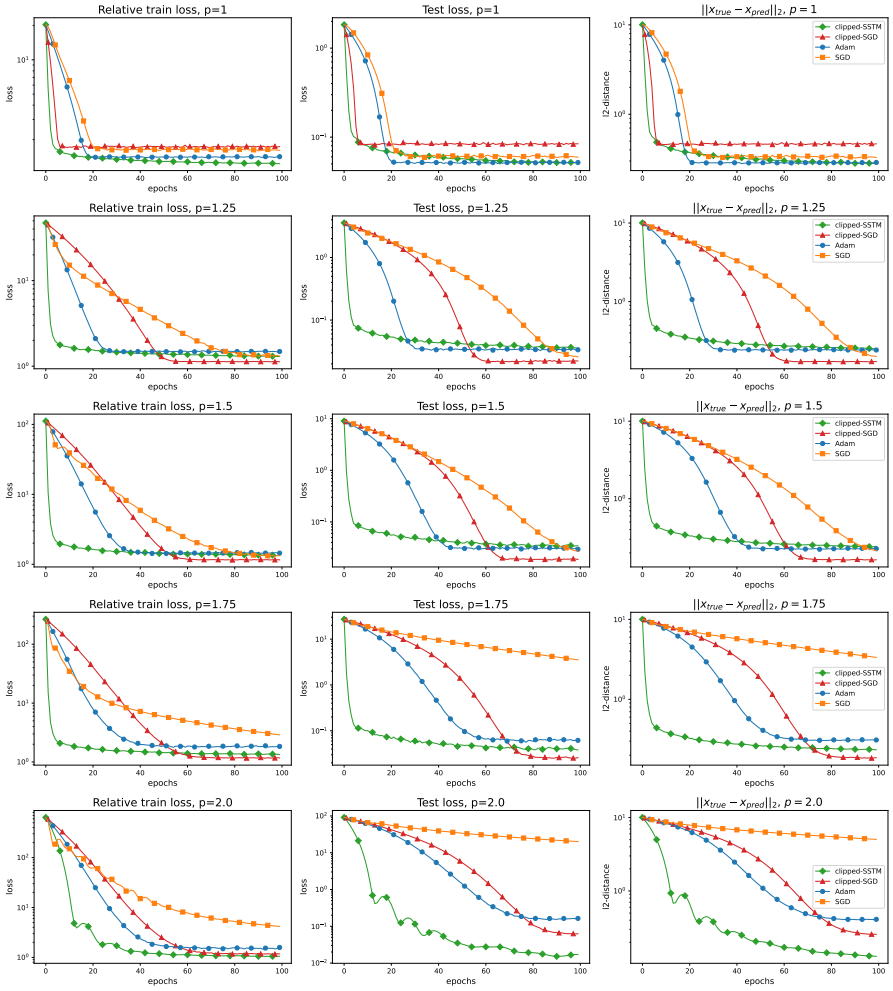

Figure 5: Results obtained for different p by the lowest epoch when model achieved $\times 2$ from loss in x_{true}

Image Classification. For `ResNet-18` + `ImageNet-100` the parameters of the methods were chosen as follows:

- Adam: $lr = 1e-3$ and a batch size of 4×32
- SGD: $lr = 1e-2$, $momentum = 0.9$ and a batch size of 32
- clipped-SGD: $lr = 5e-2$, $momentum = 0.9$, coordinate-wise clipping with clipping parameter $B = 0.1$ and a batch size of 32
- clipped-SSTM: $\nu = 1$, stepsize parameter $\alpha = 1e-3$ (in code we use separately $lr = 1e-2$ and $L = 10$ and $\alpha = \frac{lr}{L}$), norm clipping with clipping parameter $B = 1$ and a batch size of 2×32. We also upper bounded the ratio A_k/A_{k+1} by 0.99 (see $a\_k\_ratio\_upper\_bound$ parameter in code)

The main two parameters that we grid-searched were lr and batch size. For both of them, we used a logarithmic grid (i.e., for lr, we used $1e-5, 2e-5, 5e-5, 1e-4, \ldots, 1e-2, 2e-2, 5e-2$ for Adam). Batchsize was chosen from $32, 2 \cdot 32, 4 \cdot 32$, and $8 \cdot 32$. For SGD, we also tried various momentum parameters.

For clipped-SSTM and clipped-SGD, we used clipping levels of 1 and 0.1, respectively. Too small a choice of the clipping level, e.g. 0.01, slows down the convergence significantly.

Another important parameter for clipped-SSTM here was $a\_k\_ratio\_upper\_bound$ – we used it to upper bound the maximum ratio of A_k/A_{k+1}. Without this modification, the method is too conservative. e.g., after 10^4 steps $A_k/A_{k+1} \approx 0.9999$. Effectively, it can be seen as a momentum parameter of SGD.

Text Classification, CoLA. For `BERT` + `CoLA` the parameters of the methods were chosen as follows:

- Adam: $lr = 5e-5$, $weight\_decay = 5e-4$ and a batch size of 32
- SGD: $lr = 1e-3$, $momentum = 0.9$ and a batch size of 32
- clipped-SSTM: $\nu = 1$, stepsize parameter $\alpha = 8e-3$, norm clipping with clipping parameter $B = 1$ and a batch size of 8×32
- clipped-SGD: $lr = 2e-3$, $momentum = 0.9$, coordinate-wise clipping with clipping parameter $B = 0.1$ and a batch size of 32

There, we used the same grid as in the previous task. The main difference here is that we didn't bound clipped-SSTM A_k/A_{k+1} ratio – there are only ≈ 300 steps of the method (because the batch size is $8 \cdot 32$). Thus, the method is still not too conservative.

B.2.2 On the Relation Between Stepsize Parameter α and Batchsize

In our experiments, we noticed that clipped-SSTM shows similar results when the ratio bs^2/α is kept unchanged, where bs is batch size (see Figure 6). We compare the performance of clipped-SSTM with 4 different choices of α and the batch size.

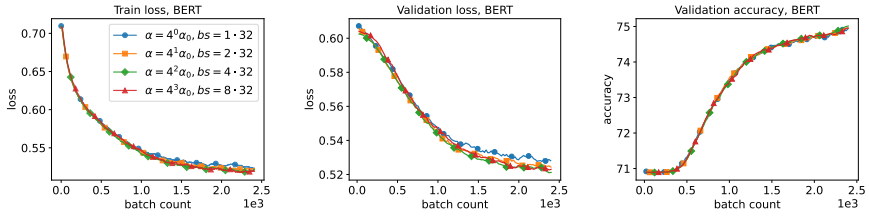

Figure 6: Train and validation loss + accuracy for clipped-SSTM with different parameters. Here $\alpha_0 = 0.000125$, bs means batch size. As we can see from the plots, increasing α 4 times and batch size 2 times almost does not affect the method's behavior.

Theorem 4.1 explains this phenomenon in the convex case. For the case of $\nu = 1$ we have (from

(24) and (30)):

$$\alpha \sim \frac{1}{aM_1}, \quad \alpha_k \sim k\alpha, \quad m_k \sim \frac{Na\sigma^2\alpha_{k+1}^2}{C^2R_0^2 \ln \frac{4N}{\beta}}, \quad N \sim \frac{a^{\frac{1}{2}}CR_0M_1^{\frac{1}{2}}}{\varepsilon^{\frac{1}{2}}} \sim \frac{CR_0}{\alpha^{\frac{1}{2}}\varepsilon^{\frac{1}{2}}},$$

whence

$$m_k \sim \frac{CR_0a\sigma^2\alpha^2(k+1)^2}{\alpha^{\frac{1}{2}}\varepsilon^{\frac{1}{2}}C^2R_0^2 \ln \frac{4N}{\beta}} \sim \frac{\sigma^2\alpha^2(k+1)^2}{\alpha^{\frac{1}{2}}\alpha M_1\varepsilon^{\frac{1}{2}}CR_0 \ln \frac{4N}{\beta}} \sim \alpha^{\frac{1}{2}},$$

where the dependencies on numerical constants and logarithmic factors are omitted. Therefore, the observed empirical relation between batch size (m_k) and α correlates well with the established theoretical results for clipped-SSTM.

B.2.3 Evolution of the Noise Distribution

In this section, we provide our empirical study of the noise distribution evolution along the trajectories of different optimizers. As one can see from the plots in Figures 7 and 8, the noise distribution for `ResNet-18` + `ImageNet-100` task is always close to Gaussian distribution, whereas for `BERT` + `CoLA` task it is significantly heavy-tailed.

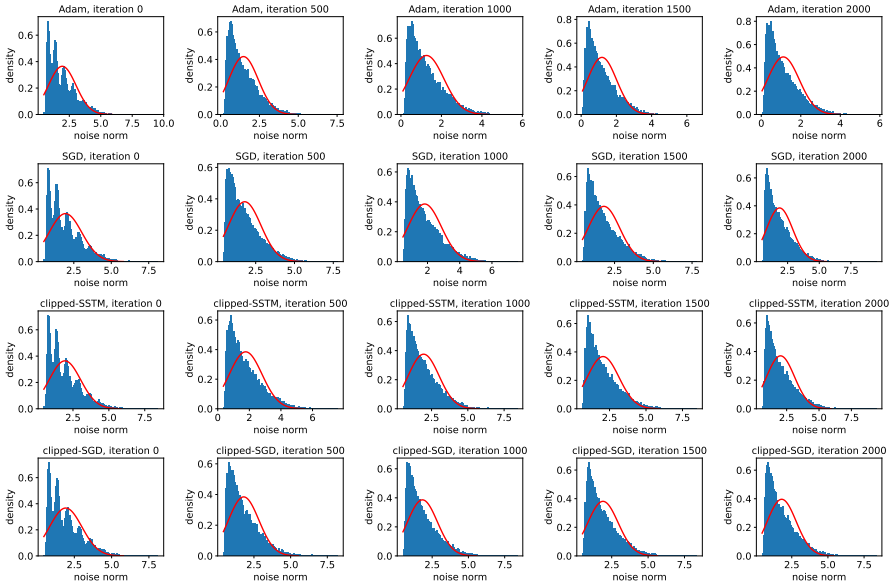

Figure 7: Evolution of the noise distribution for `BERT` + `CoLA` task.

B.2.4 Additional Results on NLP Tasks

In addition to the previous experiments, we tried several different tasks from GLUE benchmark (Wang et al., 2018): binary classification (`SST-2`), paraphrase (`MRPC`) and text similarity (`STS-B`). Noise distributions for these tasks are represented in Figure 9. As for `BERT` + `CoLA`, the noise in the stochastic gradients is heavy-tailed.

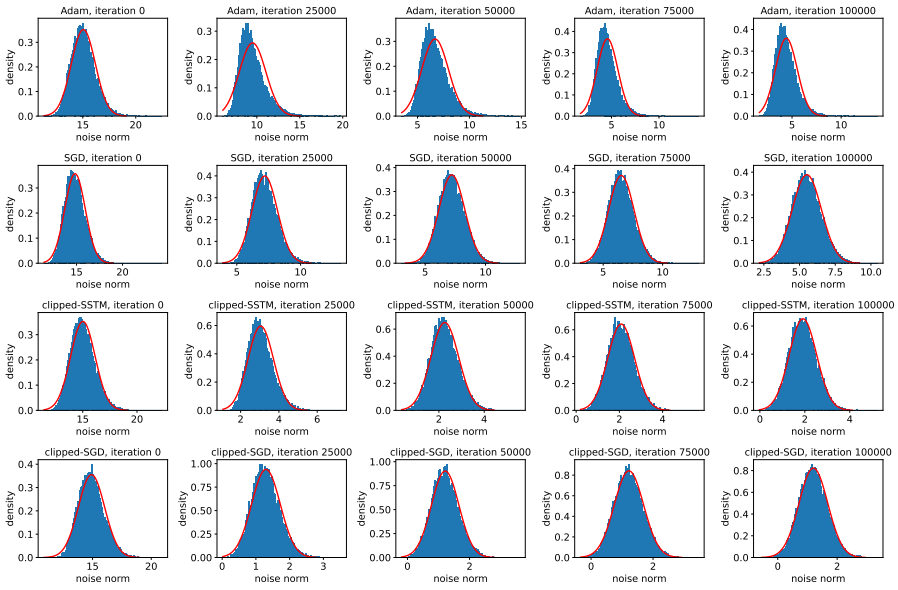

Figure 8: Evolution of the noise distribution for `ResNet-18` + `ImageNet-100` task.

In these additional tests, we also consider clipped-SSTM* – a modification of clipped-SSTM with linearly increasing batch size. This method provides faster convergence at early steps, as well as better results overall in comparison to clipped-SSTM. Parameters for all methods were tuned and are reported below.

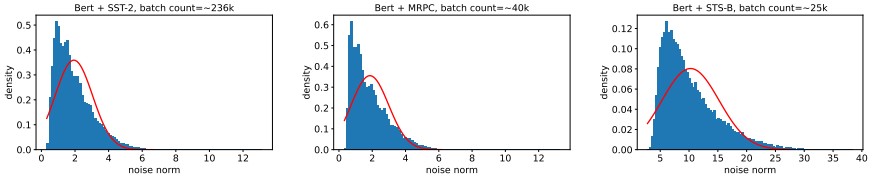

Figure 9: Noise distribution of the stochastic gradients for GLUE benchmark and `BERT` before the training. Red lines: probability density functions with means and variances empirically estimated by the samples. Batch count is the total number of samples used to build a histogram.

Text Classification, SST-2. For `BERT` + `SST-2` the parameters of the methods were chosen as follows:

- Adam: $lr = 5e-4$ and a batch size of 32
- SGD: $lr = 1e-3$, $momentum = 0.9$ and a batch size of 32
- clipped-SGD: $lr = 2e-3$, $momentum = 0.9$, coordinate-wise clipping with clipping parameter $B = 0.1$ and a batch size of 32
- clipped-SSTM: $\nu = 1$, stepsize parameter $\alpha = 1e-3$, norm clipping with clipping parameter $B = 1$ and a batch size of 8×32

- clipped-SSTM*: $\nu = 1$, stepsize parameter $\alpha = 1e-3$, norm clipping with clipping parameter $B = 1$ and a batch size of $k \times 32$ where $k = 1 + \lfloor \frac{batch\ count}{400} \rfloor$

The results are represented in Figure 10.

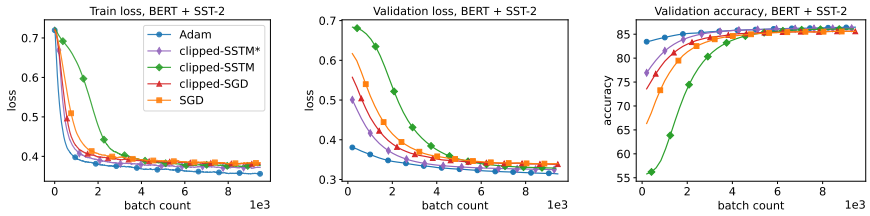

Figure 10: Train and validation loss + accuracy for different optimizers for `BERT` + `SST-2` task.

Paraphrase, MRPC. For `BERT` + `MRPC` the parameters of the methods were chosen as follows:

- Adam: $lr = 5e-4$ and a batch size of 32
- SGD: $lr = 3e-4$, $momentum = 0.99$ and a batch size of 32
- clipped-SGD: $lr = 1e-3$, $momentum = 0.95$, coordinate-wise clipping with clipping parameter $B = 0.1$ and a batch size of 32
- clipped-SSTM: $\nu = 1$, stepsize parameter $\alpha = 3e-3$, norm clipping with clipping parameter $B = 1$ and a batch size of 4×32
- clipped-SSTM*: $\nu = 1$, stepsize parameter $\alpha = 1e-2$, norm clipping with clipping parameter $B = 1$ and a batch size of $k \times 32$ where $k = 1 + \lfloor \frac{batch\ count}{100} \rfloor$

The results are represented in Figure 11.

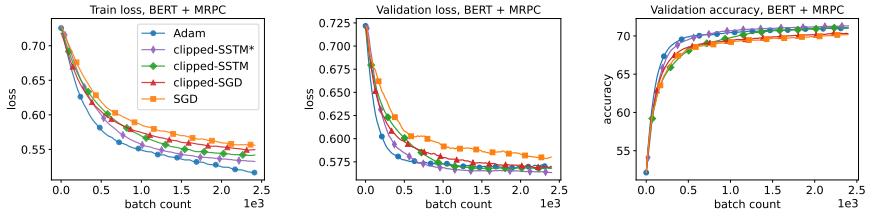

Figure 11: Train and validation loss + accuracy for different optimizers for `BERT` + `MRPC` task.

Text similarity, STS-B. For `BERT` + `STS-B` the parameters of the methods were chosen as follows:

- Adam: $lr = 1e-3$ and a batch size of 32
- SGD: $lr = 1e-4$, $momentum = 0.99$ and a batch size of 32
- clipped-SGD: $lr = 1e-3$, $momentum = 0.995$, coordinate-wise clipping with clipping parameter $B = 0.1$ and a batch size of 32

- clipped-SSTM: $\nu = 1$, stepsize parameter $\alpha = 1e-2$, norm clipping with clipping parameter $B = 1$ and a batch size of 8×32
- clipped-SSTM*: $\nu = 1$, stepsize parameter $\alpha = 3e-3$, norm clipping with clipping parameter $B = 1$ and a batch size of $k \times 32$ where $k = 1 + \lfloor \frac{batch\ count}{200} \rfloor$

The results are represented in Figure 12.

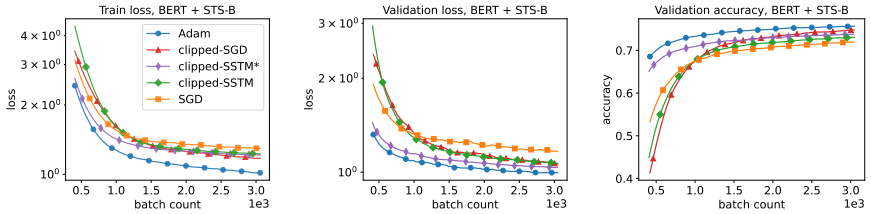

Figure 12: Train and validation loss + accuracy for different optimizers for `BERT` + `STS-B` task.

As for `BERT` + `CoLA`, we see that methods with clipping are superior to SGD. This is expected in view of the histograms reported in Figure 9 and previous empirical studies. We also point out that clipped-SSTM/clipped-SSTM */clipped-SGD achieve either comparable or even better validation accuracy than Adam.

References

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

Chaux, C., Combettes, P. L., Pesquet, J.-C., and Wajs, V. R. (2007). A variational formulation for frame-based inverse problems. *Inverse Problems*, 23(4):1495–1518.

Davis, D., Drusvyatskiy, D., Xiao, L., and Zhang, J. (2021). From low probability to high confidence in stochastic convex optimization. *Journal of Machine Learning Research*, 22(49):1–38.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Devolder, O. (2013). *Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization*. PhD thesis, UCLouvain.

Devolder, O., Glineur, F., and Nesterov, Y. (2014). First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1):37–75.

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

Dzhaparidze, K. and Van Zanten, J. (2001). On Bernstein-type inequalities for martingales. *Stochastic processes and their applications*, 93(1):109–117.

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

Gasnikov, A. V. and Nesterov, Y. E. (2018). Universal method for stochastic composite optimization problems. *Computational Mathematics and Mathematical Physics*, 58:48–64.

Gasnikov, A. V., Nesterov, Y. E., and Spokoiny, V. G. (2015). On the efficiency of a randomized mirror descent algorithm in online optimization problems. *Computational Mathematics and Mathematical Physics*, 55(4):580–596.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org.

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

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

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Gorbunov, E., Danilova, M., and Gasnikov, A. (2020). Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15042–15053. Curran Associates, Inc.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). SGD: General analysis and improved rates. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5200–5209. PMLR.

Guigues, V., Juditsky, A., and Nemirovski, A. (2017). Non-asymptotic confidence bounds for the optimal value of a stochastic program. *Optimization Methods and Software*, 32(5):1033–1058.

Guzmán, C. and Nemirovski, A. (2015). On lower complexity bounds for large-scale smooth convex optimization. *Journal of Complexity*, 31(1):1–14.

Hazan, E., Levy, K., and Shalev-Shwartz, S. (2015). Beyond convexity: Stochastic quasi-convex optimization. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Juditsky, A. and Nemirovski, A. (2011). First order methods for nonsmooth convex large-scale optimization, i: general purpose methods. *Optimization for Machine Learning*, pages 121–148.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

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

Mai, V. V. and Johansson, M. (2021). Stability and Convergence of Stochastic Gradient Clipping: Beyond Lipschitz Continuity and Smoothness. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7325–7335. PMLR.

Menon, A. K., Rawat, A. S., Reddi, S. J., and Kumar, S. (2020). Can gradient clipping mitigate label noise? In *International Conference on Learning Representations*.

Moulines, E. and Bach, F. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 24, pages 451–459. Curran Associates, Inc.

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

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

Nemirovski, A. S. and Yudin, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience publication. Wiley.

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

Nesterov, Y. E. (1983). A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1310–1318, Atlanta, Georgia, USA. PMLR.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Sadiev, A., Danilova, M., Gorbunov, E., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A., and Richtárik, P. (2023). High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *International Conference on Machine Learning*, pages 29563–29648. PMLR.

Şimşekli, U., Gürbüzbalaban, M., Nguyen, T. H., Richard, G., and Sagun, L. (2019). On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint arXiv:1912.00018*.

Simsekli, U., Sagun, L., and Gurbuzbalaban, M. (2019). A tail-index analysis of stochastic gradient noise in deep neural networks. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5827–5837. PMLR.

Spokoiny, V. (2012). Parametric estimation. finite sample theory. *The Annals of Statistics*, 40(6):2877–2909.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Warstadt, A., Singh, A., and Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhang, J., He, T., Sra, S., and Jadbabaie, A. (2020a). Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*.

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S., Kumar, S., and Sra, S. (2020b). Why are adaptive methods good for attention models? In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15383–15393. Curran Associates, Inc.