Breaking the Heavy-Tailed Noise Barrier in Stochastic Optimization Problems

Nikita Puchkin
HSE University,
IITP RAS

Eduard Gorbunov
MBZUAI

Nikolay Kutuzov
MIPT

Alexander Gasnikov
University Innopolis,
MIPT, ISP RAS

Abstract

We consider stochastic optimization problems with heavy-tailed noise with structured density. For such problems, we show that it is possible to get faster rates of convergence than $\mathcal{O}(K^{-2(\alpha-1)/\alpha})$, when the stochastic gradients have finite moments of order $\alpha \in (1, 2]$. In particular, our analysis allows the noise norm to have an unbounded expectation. To achieve these results, we stabilize stochastic gradients, using smoothed medians of means. We prove that the resulting estimates have negligible bias and controllable variance. This allows us to carefully incorporate them into clipped-SGD and clipped-SSTM and derive new high-probability complexity bounds in the considered setup.

1 INTRODUCTION

Stochastic optimization problems with heavy-tailed noise have been gaining a lot of attention in the machine learning community. This phenomenon can be partially explained due to the growing popularity of large language models (Brown et al., 2020; OpenAI, 2023) where stochastic gradients are often far from being well-concentrated (Zhang et al., 2020). In theoretical studies, such behaviour is reflected in the so-called bounded (central) α-th moment assumption with $\alpha \in (1, 2]$ (Nemirovskij and Yudin, 1983; Zhang et al., 2020), written as

$$\mathbb{E}\big[\|g(x) - \nabla f(x)\|^\alpha\big] \leqslant \sigma^\alpha, \tag{1}$$

where $\nabla f(x)$ is the gradient of the objective function $f(x)$, $g(x)$ is the stochastic gradient, and $\sigma \geqslant 0$. While in the classical literature on stochastic optimization the authors usually require the noise to have bounded variance (see, for instance, (Nemirovski et al., 2009; Ghadimi and Lan, 2013)), many recent results on high-probability/in-expectation rates of convergence were obtained under a strictly weaker condition $1 < \alpha < 2$.

In deep learning and machine learning communities, one of the most popular techniques to deal with heavy-tailed noise is gradient clipping. In (Pascanu et al., 2013), the authors showed that such a simple trick helps to stabilize neural network training via stochastic gradient descent. Their algorithm called clipped-SGD and clipping in general were then studied in a series of papers including (Abadi et al., 2016; Zhang et al., 2019; Chen et al., 2020; Zhang et al., 2020; Mai and Johansson, 2021; Karimireddy et al., 2021). In particular, for strongly convex functions Zhang et al. (2020) showed that the expected error of clipped-SGD (Pascanu et al., 2013) decreases as $\mathcal{O}(K^{-2(\alpha-1)/\alpha})$ when the number of iterations K grows. In (Sadiev et al., 2023), the authors extended this result and proved that a similar bound (up to logarithmic factors) holds with high probability. According to (Zhang et al., 2020, Theorem 5), the rate of convergence $\mathcal{O}(K^{-2(\alpha-1)/\alpha})$ is tight and cannot be improved if no assumptions, except for (1), are made. However, this rate deteriorates when α is close to 1, and, if $\alpha = 1$, the convergence is not even guaranteed. Fortunately, authors usually have to construct quite specific families of discrete distributions to attain the lower bound $\Omega(K^{-2(\alpha-1)/\alpha})$ (see, e. g., (Nemirovskij and Yudin, 1983; Devroye et al., 2016; Zhang et al., 2020; Cherapanamjeri et al., 2022; Vural et al., 2022)). Such an extreme situation is unlikely to hold in practice and we can hope for more optimistic error guarantees. This brings us to a natural question: *is it possible to achieve better rates of convergence in stochastic optimization problems with heavy-tailed noise under refined assumptions on its structure?* In this paper, we give an affirmative answer to this question.

Contribution. We consider a novel stochastic convex optimization setup with smooth (quasi-

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

strongly/strongly) convex objective and structured noise (see Assumption 2.1 below), going beyond the standard bounded α-th moment condition with $\alpha \in (1, 2]$. We provide new high-probability upper bounds on the error of versions of clipped-SGD and its accelerated variant called clipped-SSTM in smooth (quasi-strongly/strongly) problems, properly tailored to our setting. In particular, we do not assume boundedness of α-th moments and get $\widetilde{\mathcal{O}}(K^{-1/2})$ bound, which outperforms $\mathcal{O}(K^{-2(\alpha-1)/\alpha})$ for $\alpha < 4/3$. Moreover, for symmetric noise distributions, we obtain rates of convergence, which match (up to logarithmic factors) the state-of-the-art ones derived under the bounded variance assumption (Nazin et al., 2019; Davis et al., 2021; Gorbunov et al., 2020). In particular, for smooth strongly convex problems, the dominating term in our upper bound decreases as $\widetilde{\mathcal{O}}(K^{-1})$. Our approach relies on new non-asymptotic results on the performance of smoothed median of means.

Paper structure. The rest of the paper is organized as follows. In Section 2, we introduce our notation and formulate problem setup. Section 3 is devoted to an overview of related work. In Sections 4 and 5, we present our main results and illustrate the performance of suggested algorithms in Section 6. Many technical details are deferred to Appendix.

2 SETUP AND NOTATION

Before we formulate our main contributions, we need to introduce the notation and formalize the problem and setup we focus on.

Notation. Throughout the paper, we denote the standard Euclidean norm in $\mathbb{R}^d$ as $\|\cdot\|$. To simplify the bounds in the main text, we use $\widetilde{\mathcal{O}}(\cdot)$ notation that hides constant and polylogarithmic factors. For any $x_1, \ldots, x_n \in \mathbb{R}^d$, we denote $\mathtt{Mean}(x_1, \ldots, x_n) = (x_1 + \ldots + x_n)/n$. For any random vectors $\xi_1, \ldots, \xi_{2m+1}$, $\mathtt{Med}(\xi_1, \ldots, \xi_{2m+1})$ stands for the $(m+1)$-th order statistic (also called the median), taken in the component-wise fashion. For any non-zero $x \in \mathbb{R}^d$ and $\lambda > 0$, $\mathtt{clip}(x, \lambda) = \min\{1, \lambda/\|x\|\}x$ denotes the clipping operator. We also define $\mathtt{clip}(0, \lambda) = 0$ for all $\lambda > 0$. For any $\theta > 0$,

$$\Phi_\theta(t) = \frac{1}{\sqrt{2\pi}\theta} \int_{-\infty}^{t} e^{-u^2/(2\theta^2)}\, du$$

stands for the CDF of a Gaussian random variable with zero mean and variance θ^2. Sometimes, we use the notation $a \wedge b$ and $a \vee b$, instead of $\min\{a, b\}$ and $\max\{a, b\}$, respectively. Along with the standard $\mathcal{O}(\cdot)$ notation, we use the relations $g \lesssim h$ and $h \gtrsim g$, which are equivalent to $g = \mathcal{O}(h)$. Finally, for any functions $g : \mathbb{R}^d \to \mathbb{R}$ and $h : \mathbb{R}^d \to \mathbb{R}$, their convolution is denoted as $g * h(x) = \int_{\mathbb{R}^d} g(x - y)h(y)\, dy$. We also adopt the notation $g^{*k}(x) = \underbrace{g * \ldots * g}_{k \text{ times}}(x)$.

Setup. We consider an unconstrained smooth convex optimization problem

$$\min_{x \in \mathbb{R}^d} f(x), \tag{2}$$

where the function $f : \mathbb{R}^d \to \mathbb{R}$ is accessible through the stochastic first-order oracle $\mathcal{G} : \mathbb{R}^d \to \mathbb{R}^d$ that for given point $x \in \mathbb{R}^d$ returns some estimate of $\nabla f_\xi(x)$ of the true gradient $\nabla f(x)$. We make the following assumption about the distribution of $\nu = \nabla f_\xi(x) - \nabla f(x)$.

Assumption 2.1. *For any $x \in \mathbb{R}^d$ and each $j \in \{1, \ldots, d\}$, the marginal density p_j of the j-th component of the noise $\nu = \nabla f_\xi(x) - \nabla f(x)$ satisfies the following conditions:*

- *there exists $M_j > 0$, such that $\mathsf{r}_j(u) = (\mathsf{p}_j(u) - \mathsf{p}_j(-u))/2$ fulfils*

$$\int_{-\infty}^{+\infty} u\mathsf{r}_j(u)\, du = 0 \quad \text{and} \quad \int_{-\infty}^{+\infty} u^2 |\mathsf{r}_j(u)|\, du \leqslant M_j.$$

- *there are $B_j > 0$ and $\beta_j \geqslant 1$, such that, for any $k \in \mathbb{N}$,*

$$\mathsf{s}_j^{*k}(u) \leqslant \frac{B_j k}{k^{(\beta_j+1)/\beta_j} + |u|^{1+\beta_j}},$$

where $\mathsf{s}_j(x) = (\mathsf{p}_j(x) + \mathsf{p}_j(-x))/2$.

In Assumption 2.1, we split marginal densities of noise components into a sum of symmetric and antisymmetric parts. For each $j \in \{1, \ldots, d\}$, the antisymmetric remainder $\mathsf{r}_j(u)$ is a signed density with a finite second moment. However, the symmetric term $\mathsf{s}_j(u)$ may decay much slower, than $\mathsf{r}_j(u)$. As a result, the density $\mathsf{p}_j(u)$ has finite moments up to order $\alpha < (\beta_j \wedge 2)$. Note that if $\beta_j = 1$ for some $j \in \{1, \ldots, d\}$, then the noise norm $\|\nu\|$ may have no expectation.

We proceed with several examples of $\mathsf{s}_j(u)$, satisfying Assumption 2.1. Obviously, if the j-th component of $\nu(x)$ has a standard Cauchy distribution, that is, $\mathsf{s}_j(u) = 1/\pi \cdot 1/(1 + u^2)$, then, for any $k \in \mathbb{N}$, $\mathsf{s}_j^{*k}(u) = 1/\pi \cdot k/(k^2 + u^2)$, and Assumption 2.1 is fulfilled with $B_j = 1/\pi$ and $\beta_j = 1$. This is a particular example of a symmetric α-stable distribution with parameter $\alpha = 1$. All symmetric α-stable distributions have a characteristic function of the form $\varphi(y) = e^{-|y/\sigma|^\alpha}$,

where > 0 and $2 (0; 2]$ (see, e.g., (Feller, 1971, Chapter XVII, Sections 5-6)). If $1 6 6 2$, they also satisfy Assumption 2.1 with $_j =$ and some$B_j > 0$. In general, if $s_j(u) B=juj^{1+ j}$, then it is known from probability theory that $s_j^k(u) Bk=juj^{1+ j}$ for any $k 2 N$. Assumption 2.1 can be viewed as a non-asymptotic version of this property.

We also make several standard assumptions about function f itself. Similarly to (Gorbunov et al., 2021; Sadiev et al., 2023), it is sufficient for our analysis to make all the assumptions only on some compact subset (ball) of R^d since we show that with high probability the considered methods do not leave this compact. We start with the standard smoothness assumption.

Assumption 2.2. There exists a set$Q R^d$ and constant $L > 0$ such that for all $x; y 2 Q$

$$kr f(x) r f(y)k 6 Lkx yk; \quad (3)$$

$$kr f(x)k^2 6 2L(f(x) f); \quad (4)$$

where$f = \inf_{x2Q} f(x) > 1$.

When $Q = R^d$, (4) follows from (3). However, when $Q 6 R^d$, condition (4) can be derived from(3), if the latter is assumed on a slightly larger set, see (Sadiev et al., 2023, Appendix B) for additional discussion.

We assume convexity or strong convexity off for the results with accelerated rates.

Assumption 2.3. There exists set$Q R^d$ and constant > 0 such that$f$ is -strongly convex, i.e., for all $x; y 2 Q$, it holds that

$$f(y) > f(x) + hr f(x); y xi + \frac{}{2}ky xk^2: \quad (5)$$

Finally, for non-accelerated case, it is sufficient to assume a relaxed condition called quasi-strong convexity.

Assumption 2.4. There exists set$Q R^d$ and constant > 0 such that$f$ is -quasi-strongly convex, that is, for all $x 2 Q$ and $x = \arg\min_{x2R^d} f(x)$

$$f(x) > f(x) + hr f(x); x xi + \frac{}{2}kx x k^2: \quad (6)$$

The above assumption belongs to the class of conditions on structured non-convexity. When > 0 it (together with smoothness) implies linear convergence for Gradient Descent (Necoara et al., 2019).

3 RELATED WORK

High-probability complexity bounds. Under sub-Gaussian noise assumption, optimal (up to logarithmic factors) high-probability complexity bounds[1] are proven by Nemirovski et al. (2009) for (strongly) convex non-smooth problems with bounded sub-gradients, by Ghadimi and Lan (2012) for (strongly) convex smooth problems, and by Li and Orabona (2020) for smooth non-convex problems. These results are achieved for the same methods that are optimal in terms of the in-expectation convergence. However, when the noise has just a nite variance, some algorithmic changes seem to be necessary, e.g., as it is shown in (Sadiev et al., 2023, Section 2), standardSGDhas provably bad (inverse-power instead ofpoly(log(1=))) dependence on the confidence level in this case.

A popular tool for overcoming this issue is gradient clipping, i.e., the application of the clipping operator to the gradient estimator. A version of gradient clipping is used by Nazin et al. (2019) who derive the rst (non-accelerated) high-probability complexity bounds for smooth (strongly) convex problems on compact domains with logarithmic dependence on1= under bounded variance assumption. Accelerated results are obtained by Davis et al. (2021) and Gorbunov et al. (2020) for smooth strongly convex and smooth convex problems respectively. Gorbunov et al. (2021) generalize these results to the case of problems with Hölder continuous gradients.

State-of-the-art high-probability complexity bounds are derived under bounded (central) -th moment assumption (1). The rst work in this direction is (Cutkosky and Mehta, 2021) where the authors derived optimal (up to logarithmic factors) bounds in the smooth non-convex regime with the additional assumption of boundedness of the gradients. Without this assumption, a worse bound is derived by Sadiev et al. (2023), and the optimal one is obtained by Nguyen et al. (2023b). For (strongly) convex problems the results forclipped-SGD and its accelerated versionclipped-SSTM(Gorbunov et al., 2020) are derived by Sadiev et al. (2023). Up to logarithmic factors, these results match the known lower bounds in the strongly convex case (Zhang et al., 2020). Nguyen et al. (2023a) improve the logarithmic factors in the upper bounds from Sadiev et al. (2023); Nguyen et al. (2023b). Recently, the generalization of the results from (Sadiev et al., 2023) to the case of composite and distributed optimization were obtained by Gorbunov et al. (2023).

Other results under heavy-tailed noise. Although in our work we primarily focus on high-probability convergence results, we brie y discuss here

[1] Such results establish upper bounds for the number of oracle calls needed for a method to nd point x such that $f(x) f(x) \varepsilon$ or $kx x k^2 \varepsilon$ or $kr f(x)k^2 \varepsilon$ are less than$"$ with probability at least , where$x$ is a solution of (2).

other existing works devoted to the convergence of stochastic methods under heavy-tailed noise assumption. For convex functions with bounded gradients, Nemirovskij and Yudin (1983) show $O(K^{-(\alpha-1)/\alpha})$ in-expectation convergence rate for Mirror Descent and Vural et al. (2022) propose an extension of this result for uniformly convex functions. For strongly convex functions with bounded gradients, Zhang et al. (2020) show $O(K^{-2(\alpha-1)/\alpha})$ in-expectation rate of convergence. In the smooth non-convex case, $O(K^{-2(\alpha-1)/(3\alpha-2)})$ in-expectation convergence rate is achieved by Zhang et al. (2020) who also derive a matching lower bound.

In the case of the noise with symmetric density function and bounded first moment, Jakovetić et al. (2023) derive $O(K^{-\alpha})$ in-expectation convergence rate for α-strongly convex L-smooth functions and SGD-type methods with general non-linearities. However, parameter α is proportional to $\sigma L^{p-1} d^{\frac{1-p}{2p}}$ in the worst case. Therefore, this rate can be much slower than $O(K^{-2(\alpha-1)/\alpha})$ for ill-conditioned/large-scale problems though Jakovetić et al. (2023) do not assume (1) and consider general class of non-linearities. Our analysis also does not rely on (1), but we additionally allow non-symmetric noise distributions and do not assume the existence of the finite first moment of the noise.

Median estimates. Median, median of means, and smoothed median were extensively used in the problems of robust mean estimation and robust machine learning (see, for instance, (Nemirovskij and Yudin, 1983; Minsker, 2015; Devroye et al., 2016; Lugosi and Mendelson, 2019b, 2020; Lecué and Lerasle, 2020; Cherapanamjeri et al., 2022)). A reader is referred to a comprehensive survey of Lugosi and Mendelson (2019a) on this topic. Usually, the authors use median of means or its modifications to get sub-Gaussian rates of convergence, assuming the existence of only two moments. In Cherapanamjeri et al. (2022), the authors went further and derived a minimax optimal upper bound in the problem of mean estimation when observations have finite moments of order $\alpha \in (1, 2]$. However, the authors faced the same problem as Zhang et al. (2020): the rate of convergence became very slow when α approached 1. This happens, because the family of distributions of interest is extremely large if one assumes the existence of α-th moment only. In our paper, we exploit the special noise structure, described in Section 2. Under Assumption 2.1, we derive new non-asymptotic bounds on the performance of the smoothed median of means, which do not deteriorate even if the underlying density has quite heavy tails.

4 SMOOTHED MEDIAN OF MEANS AND ITS PROPERTIES

In this section, we describe how to get reliable gradient estimates from noisy stochastic gradients given by the first-order oracle. Let us start with a simple example. Fix an arbitrary $x \in \mathbb{R}^d$ and assume that the noise $\xi = \nabla f(x) - \nabla f(x) \in \mathbb{R}^d$ has a symmetric absolutely continuous distribution. For any $j \in \{1, \dots, d\}$, let $p_j(u)$ be the marginal density of ξ_j, the j-th component of ξ. Then the following proposition holds true.

Proposition 4.1. Fix any $j \in \{1, \dots, d\}$ and assume that the marginal density of ξ_j is symmetric, that is, $p_j(u) = p_j(-u)$ for all $u \in \mathbb{R}$. Suppose that there exist positive numbers B_j and β_j, such that

$$p_j(u) \le \frac{B_j}{1 + |u|^{\beta_j + 1}}, \quad \text{for all } u \in \mathbb{R}.$$

Let $\xi_{j,1}, \dots, \xi_{j,(2m+1)}$ be independent copies of ξ_j. If $m > 3/\beta_j$, then $\mathbb{E}\,\mathrm{Med}(\xi_{j,1}, \dots, \xi_{j,(2m+1)}) = 0$ and $\mathbb{E}\,\mathrm{Med}^2(\xi_{j,1}, \dots, \xi_{j,(2m+1)})$ is finite.

The proof of the proposition with an explicit bound on the variance of $\mathrm{Med}(\xi_{j,1}, \dots, \xi_{j,(2m+1)})$ is postponed to Appendix. Proposition 4.1 shows that, despite the heavy tails of the underlying density p, $m > \max\{3/\beta_j : 1 \le j \le d\}$ oracle calls are enough to produce an unbiased estimate $g(x) = \mathrm{Med}(\nabla f_1(x), \dots, \nabla f_m(x))$ of $\nabla f(x)$ with a finite variance. After that, we can use the standard clipping technique to solve the optimization problem (2).

Unfortunately, the symmetry assumption, which played the central role in Proposition 4.1, is rather restrictive. To deal with asymmetric distributions, we use more sophisticated gradient estimates, based on smoothed median of means.

Definition 4.2. Let ξ be a random element in $\mathbb{R}^d$ and let $\lambda > 0$ be an arbitrary number. For any positive integers m and n, the smoothed median of means $\mathrm{SMoM}_{m,n}(\xi, \lambda)$ is defined as follows:

$$\mathrm{SMoM}_{m,n}(\xi, \lambda) = \mathrm{Med}(\zeta_1, \dots, \zeta_{2m+1}),$$

where, for each $j \in \{0, \dots, 2m\}$,

$$\zeta_j = \mathrm{Mean}(\xi_{jn+1}, \dots, \xi_{(j+1)n}) + \lambda \eta_{j+1},$$

$\xi_1, \dots, \xi_{(2m+1)n}$ are i.i.d. copies of ξ, and $\eta_1, \dots, \eta_{2m+1} \sim \mathcal{N}(0, I_d)$ are independent standard Gaussian random vectors.

Let us briefly describe the idea behind our approach. Assuming that $\nabla f(x) - \nabla f(x)$ at a point $x \in \mathbb{R}^d$ has a density $p(u)$, we represent the latter in the following form:

$$p(u) = s(u) + r(u);$$

where $s(u) = (p(u) + p(-u))/2$ is a symmetric part and $r(u) = (p(u) - p(-u))/2$ is an antisymmetric remainder. If the tails of the remainder $r(u)$ are much lighter than the ones of $p(u)$, we can make n oracle calls at the point x and take the average of $\nabla f_1(x), \dots, \nabla f_n(x)$. If n is large enough, then the distribution of $\mathrm{Mean}(\nabla f_1(x), \dots, \nabla f_n(x))$ is almost symmetric. Hence, we can use the same trick as in Proposition 4.1 to get an estimate of $\nabla f(x)$ with a finite variance. We add small Gaussian noise to ensure that the density of our estimate is infinitely differentiable, as we need it for technical purposes. Note that, in general, the expectation of $\mathrm{SMoM}_{m,n}(\nabla f(x) - \nabla f(x); \gamma)$ is not equal to zero, and it is a challenging task to show that it is sufficiently small.

Before we move to the heavy-tailed setup, let us illustrate the efficiency of our approach in the case when the stochastic gradients have a finite second moment.

Lemma 4.3. Assume that stochastic gradient $\nabla f(x)$ at a point $x \in \mathbb{R}^d$ has an absolutely continuous distribution and a finite second moment $\mathbb{E}(\nabla f(x) - \nabla f(x))(\nabla f(x) - \nabla f(x))^\top = \Sigma$. Then, for any positive integer m and n, it holds that

$$\mathbb{E}\left\|\mathrm{SMoM}_{m,n}(\nabla f(x); \gamma) - \nabla f(x)\right\|^2 \le 4(2m+1)\left(\frac{\mathrm{Tr}(\Sigma)}{n} + \gamma^2 d\right).$$

If, in addition, $m > 3$ and $n \ge 8m\ln\frac{1}{\delta}$, then

$$\left\|\mathrm{SMoM}_{m,n}(\nabla f(x); \gamma) - \nabla f(x)\right\| \le \frac{m}{n}\sqrt{\mathrm{Tr}(\Sigma)}.$$

Remark 4.4. Lemma 4.3 also yields that

$$\left\|\mathrm{SMoM}_{m,n}(\nabla f(x); \gamma) - \nabla f(x)\right\| \le \sqrt{m}\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{d}\,\gamma\right).$$

Though this bound is enough for our purposes, nevertheless, we find it useful to prove a dimension-free $O(1/n)$ upper bound on expectation of the smoothed median of means, which follows from the analysis of impact of antisymmetric density part on the expectation (see Lemma A.1 in Appendix).

Lemma 4.3 shows that, if the noise vector has a finite second moment, then the smoothed median of means has a small controllable shift and bounded variance. In this case, it behaves similarly to clipping. However, if we deal with heavy-tailed noise, the standard clipping technique fails, while the smoothed median of means still has a small bias and finite variance. We proceed with the main result of this section.

Lemma 4.5. Assume that the stochastic gradient $\nabla f(x)$ at a point $x \in \mathbb{R}^d$ has an absolutely continuous distribution. Suppose that, for any $j \in \{1, \dots, d\}$ and any $x \in \mathbb{R}^d$, the density of $\eta_j = \nabla f(x) - \nabla f(x)$ meets Assumption 2.1. Then, if $m > 2 + 3\alpha_j$ and $n \ge (2 - \alpha_j)m\ln\frac{1}{\delta}$ for all $j \in \{1, \dots, d\}$, it holds that

$$\mathbb{E}\left\|\mathrm{SMoM}_{m,n}(\nabla f(x); \gamma) - \nabla f(x)\right\|^2 \le \left(m(1+\gamma^2)d + \sum_{j=1}^d \left(\frac{M_j}{n}^{2-\alpha_j} + \left(\frac{2^{\alpha_j}B_j}{n^{\alpha_j-1}}\right)^{2/\alpha_j} + \left(\frac{B_j M_j}{n^{\alpha_j}}\right)^{2/(\alpha_j+1)}\right)\right)$$

and

$$\mathbb{E}\mathrm{SMoM}_{m,n}(\nabla f(x); \gamma) - \nabla f(x) \le \frac{m(1+\gamma^2)}{2n}\sqrt{\sum_{j=1}^d M_j^2} + \frac{m}{2n}\sqrt{\sum_{j=1}^d M_j^2\left(\frac{2^{\alpha_j}B_j}{n^{\alpha_j-1}}\right)^{2/\alpha_j}}.$$

Remark 4.6. The bounds on $\mathbb{E}\|\mathrm{SMoM}_{m,n}(\nabla f(x); \gamma) - \nabla f(x)\|$ in Lemma 4.3 and Lemma 4.5 rely on the fact that convolution with the Gaussian density is infinitely differentiable. However, if $s_1, \dots, s_d$ are also sufficiently smooth, then one can take $\gamma = 0$ and apply a similar technique as in Lemma A.1 and Lemma A.4.

The proof of Lemma 4.5 is moved to Appendix. It shows that the bias of the smoothed median of means decays rapidly with the growth of the batch size, though the noise may have extremely heavy tails. The reason for that is the special noise structure guaranteed by Assumption 2.1. This favourable property allows us to obtain faster rates of convergence in heavy-tailed stochastic convex optimization problems.

5 MAIN RESULTS FOR STOCHASTIC OPTIMIZATION

In view of the results of the previous section, Assumption 2.1 allows constructing an estimator with bounded bias and variance using the smoothed median of means. Since in the analysis of the stochastic first-order methods we only use these two properties, we formulate them as a separate assumption for convenience.

Assumption 5.1. There exists $N \ge 2$, aggregation rule R and (possibly dependent on N) constants $b > 0$, $\gamma > 0$ such that for any $x \in \mathbb{R}^d$ i.i.d. samples $\nabla f_1(x), \dots, \nabla f_N(x)$ from the oracle $G(x)$ satisfy the following relations:

$$\left\|\mathbb{E}[\nabla f(x)] - \nabla f(x)\right\| \le b, \tag{7}$$

$$\mathbb{E}\left\|\nabla f(x) - \mathbb{E}[\nabla f(x)]\right\|^2 \le \gamma^2, \tag{8}$$

where $\nabla f(x) = R(\nabla f_1(x), \dots, \nabla f_N(x))$ and expectations are taken w.r.t. $\nabla f_1(x), \dots, \nabla f_N(x)$.

We emphasize once again that Assumption 5.1 holds whenever Assumption 2.1 is satisfied. Indeed, we can take $\nabla f(x) = \mathrm{SMoM}_{m,n}\nabla f(x)$; with parameters m, n, and δ, satisfying the conditions of Lemma 4.5. In this case, the batch size is equal to $N = (2m+1)n$.

5.1 Convergence of clipped-SGD

We start with clipped-SGD defined as follows:

$$x^{k+1} = x^k - \eta_k \mathrm{clip}(\nabla f_k(x^k), \lambda_k); \qquad (9)$$

where $\nabla f_k(x^k)$ is an estimator satisfying Assumption 5.1 sampled independently from previous iterations. Below we formulate the main convergence result for clipped-SGD in the quasi-convex case.

Theorem 5.2. Let Assumptions 2.2 and 2.4 with $\mu = 0$ hold on $Q = B_{2R}(x^0)$, where $R > \|x^0 - x^*\|$. Suppose that $\nabla f_k(x^k)$ satisfies Assumption 5.1 with parameters b_k, σ_k for $k = 0, 1, \ldots, K$, $K > 0$ and $\eta_k = \eta = \min\{1/L, A, R/(\sigma\sqrt{AK}), R/(bA), R/(b(K+1))\}$, $\lambda_k = \lambda = R/(A\eta)$, where $A = \ln(4(K+1)/\beta)$ and $b = \max_{k=0,1,\ldots,K} b_k$, $\sigma = \max_{k=0,1,\ldots,K} \sigma_k$. Then the iterates produced by clipped-SGD after K iterations with probability at least $1-\beta$ satisfy

$$f(\bar{x}^K) - f(x^*) = \mathcal{O}\left(\max\left\{\frac{LR^2}{K}, \frac{R\sigma}{\sqrt{K}}, bR\right\}\right),$$

where $\bar{x}^K = \frac{1}{K+1}\sum_{k=0}^{K} x^k$.

The rate of convergence in the above result matches (up to logarithmic factors) the best-known one for clipped-SGD under bounded variance assumption (Gorbunov et al., 2021). Due to systematic bias bounded by b the method reaches only $\mathcal{O}(bR)$ error after a sufficiently large number of steps. When the bias is just bounded and cannot be controlled, this situation is standard (Devolder et al., 2014). The proof of the above result follows the ones given in (Gorbunov et al., 2021; Sadiev et al., 2023): using the induction argument, we show that under a proper choice of parameters, the iterates stay in a bounded set with high probability, which allows us to apply standard Bernstein inequality (see Lemma B.1). In particular, this proof technique differs from the standard ones that rely on the boundedness of the noise (Rakhlin et al., 2011) or on the assumption that the noise is sub-Gaussian (Harvey et al., 2019).

However, in our setup, we can control the bias. For example, if the distribution is symmetric and has a bounded moment of the order α for some $\alpha > 0$, then according to Proposition 4.1, it is sufficient to use coordinate-wise median estimator to get $\nabla f(x)$ satisfying Assumption 5.1 with $b = 0$ and $\sigma^2 = d(2m+1)\left(1 - \frac{4B_j}{\delta_j}\right)^{2/\alpha_j}$ (see (18)) using $O(1/\delta)$ samples of $\nabla f(x)$. In this case, we have the following result.

Corollary 5.3 (Symmetric noise). Let the assumptions of Theorem 5.2 hold and for all $x \in \mathbb{R}^d$ the noise $\nabla f(x) - \nabla f(x)$ satisfies the conditions from Proposition 4.1. Then the iterates produced after K iterations of clipped-SGD with $\nabla f_k(x^k)$ being a coordinate-wise median of $2m+1$ samples $\nabla f(x^k)$ with $m > \max\{3/\alpha_j : 1 \le j \le d\}$ and $\eta_k = \eta = \min\{1/L, A, R/(\sigma\sqrt{AK})\}$, $\lambda_k = \lambda = R/A$ for $\sigma^2 = d(2m+1)\left(1 - \frac{4B_j}{\delta_j}\right)^{2/\alpha_j}$ and where $A = \ln(4(K+1)/\beta)$ with probability at least $1-\beta$ satisfy

$$f(\bar{x}^K) - f(x^*) = \mathcal{O}\left(\max\left\{\frac{LR^2}{K}, \frac{R\sigma}{\sqrt{K}}\right\}\right)$$

and the overall number of stochastic oracle calls equals $(2m+1)K = O(K\max\{1/\alpha_j : 1 \le j \le d\})$.

This result implies that as long as the distribution is symmetric, its tails can be even heavier than the ones of Cauchy distribution, i.e., moments of order larger than α for some $\alpha \in (0, 1]$ can be unbounded, but clipped-SGD with coordinate-wise median estimator inside will still converge as in the case when the stochastic gradients are unbiased and have bounded variance. We emphasize that the existing state-of-the-art high probability convergence results (Sadiev et al., 2023; Nguyen et al., 2023b,a) have a slower decreasing main term (of the order $\mathcal{O}(K^{(1-\alpha)/\alpha})$) and are derived for much lighter tails. However, in contrast to Corollary 5.3, the mentioned results do not rely on the symmetry.

Finally, we consider the general case, when the noise satisfies Assumption 2.1, i.e., the noise also has a non-symmetric component. In this case, Lemma 4.5 implies that the smoothed median of means gives an estimator $\nabla f(x)$ satisfying Assumption 5.1 with

$$b = O(C\Delta/n); \quad \sigma^2 = O(d(1+\Delta^2) + D); \qquad (10)$$

$$C = \frac{(1+\Delta)}{2}\sqrt{\sum_{j=1}^{d} M_j^2} + \frac{1}{2}\sqrt{\sum_{j=1}^{d} M_j^2 \left(\frac{2^j B_j}{n^{j-1}}\right)^{2/\alpha_j}}; \qquad (11)$$

$$D = \sum_{j=1}^{d}\left(\frac{M_j}{n}\right)^2 + \left(\frac{2^j B_j}{n^{j-1}}\right)^{2/\alpha_j} + \sum_{j=1}^{d}\left(\frac{B_j M_j}{n^{\alpha_j}}\right)^{2/(\alpha_j+1)} \qquad (12)$$

using $O(n)$ samples $\nabla f(x)$ (when $m = O(1)$). Together with Theorem 5.2 this implies the following result.

Corollary 5.4 (General noise) Let the assumptions of Theorem 5.2 hold and for all $x \in \mathbb{R}^d$ the noise

$= \nabla f(x) - \nabla f(x)$ satisfies Assumption 2.1. Then clipped-SGD with $\nabla f_k(x^k)$ being the smoothed median of means of $O(n)$ samples $\nabla f(x^k)$ and $\gamma_k = \gamma = \left(\min\left\{\frac{1}{4LA}, \frac{R}{\sqrt{KA}}, \frac{R}{bA}, \frac{R}{b(K+1)}\right\}\right)$, $\lambda_k = \lambda = \left(\frac{R}{A}\right)$ for b and λ defined in (10) with probability at least $1-\beta$ after K iterations ensures that $f(\bar{x}^K) - f(x^*)$ equals

$$\mathcal{O}\left(\max\left\{\frac{LR^2}{K}, \frac{\sqrt{(1+\ln\frac{1}{\beta})d + \ln\frac{1}{\beta}}\,DR}{K}, \frac{(1+\ln\frac{1}{\beta})CR}{\varepsilon n}\right\}\right),$$

where $n > \left(\max_{j\in[d]} M_j/\varepsilon\right)$ and C, D are defined in (11)-(12). The overall number of oracle calls equals $O(nK)$.

Due to the heavy-tailedness of the symmetric part of the noise distribution and the presence of bias, the variance term does not necessarily improve with the growth of the number of samples. However, the bias term still can be smaller than any predefined level via increasing n. When the bias is large, i.e., the noise is sufficiently non-symmetric, then one can take $n = \mathcal{O}(\varepsilon^{-1})$ and $K = \mathcal{O}(\varepsilon^{-2})$ to guarantee $f(\bar{x}^K) - f(x^*) \le \varepsilon$ with probability at least $1-\beta$, i.e., the total oracle complexity is $\mathcal{O}(\varepsilon^{-3})$. However, when the bias is small, i.e., $\max_{j\in[d]} M_j$ are sufficiently small [2], then for $\varepsilon = \left(\frac{(1+\ln\frac{1}{\beta})CR}{\varepsilon^2}\right)$ one can achieve even $K = \mathcal{O}(\varepsilon^{-2})$ total oracle complexity that matches (up to logarithmic factors and constants related to the variance) the main term in the optimal complexities under bounded variance assumption (Gorbunov et al., 2020). However, in contrast to the existing results, we do not require the noise to have a finite first moment.

The following theorem gives a convergence rate for clipped-SGD in the quasi-strongly convex case.

Theorem 5.5. Let Assumptions 2.2 and 2.4 with > 0 hold on $Q = B_{2R}(x^*)$, where $R > \|x^0 - x^*\|$. Suppose that $\nabla f_k(x^k)$ satisfies Assumption 5.1 with parameters b_k, λ_k for $k = 0, 1, \ldots, K$, $K > 0$ and $\gamma_k = \gamma = \left(\min\left\{\frac{1}{4LA}, \frac{\min\{\ln B_K, \ln C_K, \ln D\}}{(K+1)}\right\}\right)$, $\lambda_k = \left(\exp\left(\frac{-(1+k/2)}{3}\right)\frac{R}{A}\right)$, where $A = \ln(4(K+1)/\beta)$ and B_K, C_K, D are some parameters dependent on $b = \max_{k=0,1,\ldots,K} b_k$, $\lambda = \max_{k=0,1,\ldots,K} \lambda_k$, γ, R, and K. Then the final iterate produced by clipped-SGD after K iterations with probability at least $1-\beta$ satisfies

$$\|x^K - x^*\|^2 = \mathcal{O}\left(\max\left\{R^2\exp\left(-\frac{K}{L\ln\frac{K}{\beta}}\right), \frac{\sqrt{(1+\ln\frac{1}{\beta})d + \ln\frac{1}{\beta}}\,DR}{K}, \frac{(1+\ln\frac{1}{\beta})CR}{\varepsilon n}\right\}\right).$$

Similarly to the convex case, in the case of symmetric noise with bounded α-th moment for some $\alpha > 0$, the above result matches the best-known one for clipped-SGD under bounded variance and strong convexity (Gorbunov et al., 2021). In particular, for such noise distributions, condition (1) is not necessarily satisfied, and even if it is satisfied, our rate $\mathcal{O}(K^{-1})$ is better than the lower bound $\left(K^{-2(\alpha-1)/\alpha}\right)$ under condition (1) for $\alpha \in (1; 2)$ (Zhang et al., 2020). However, it is worth mentioning that we do rely on the symmetry of the noise distribution to achieve this rate, while the lower bound holds for any distributions satisfying (1).

In the non-symmetric case, we combine Theorem 5.5 with Lemma 4.5 and get the following result.

Corollary 5.6 (General noise). Let the assumptions of Theorem 5.5 hold and for all $x \in \mathbb{R}^d$ the noise $= \nabla f(x) - \nabla f(x)$ satisfies Assumption 2.1. Then the iterates produced after K iterations of clipped-SGD with $\nabla f_k(x^k)$ being the smoothed median of means of $O(n)$ samples $\nabla f(x^k)$ and $\gamma_k = \gamma = \left(\min\left\{\frac{1}{4LA}, \frac{\min\{\ln B_K, \ln C_K, \ln D\}}{(K+1)}\right\}\right)$, $\lambda_k = \left(\exp\left(\frac{-(1+k/2)}{3}\right)\frac{R}{A}\right)$, where $A = \ln(4(K+1)/\beta)$ and B_K, C_K, D are some parameters dependent on b and λ defined in (10), with probability at least $1-\beta$ satisfy

$$\|x^K - x^*\|^2 = \mathcal{O}\left(R^2\exp\left(-\frac{K}{L\ln\frac{K}{\beta}}\right)\right) + \mathcal{O}\left(\max\left\{\frac{(1+\ln\frac{1}{\beta})d + \ln\frac{1}{\beta}}{K}\,DR, \frac{(1+\ln\frac{1}{\beta})CR}{\varepsilon n}\right\}\right),$$

where $n > \left(\max_{j\in[d]} M_j/\varepsilon\right)$ and C, D are defined in (11)-(12). The overall number of oracle calls equals $O(nK)$.

Taking $n = \mathcal{O}(\varepsilon^{-1})$ and $K = \mathcal{O}(\varepsilon^{-1})$, one can guarantee $\|x^K - x^*\|^2 \le \varepsilon$ with probability at least $1-\beta$, i.e., the total oracle complexity is $\mathcal{O}(\varepsilon^{-2})$. This result is worse than the one for the case of symmetric distribution, but it still does not require the existence of the expectation of the noise and is better than $\mathcal{O}(\varepsilon^{-2(\alpha-1)/\alpha})$, which is known to be optimal (up to logarithmic factors) under assumption (1), when $< 4/3$.

5.2 Convergence of clipped-SSTM

Next, we consider an accelerated variant of clipped-SGD called clipped-SSTM (Gorbunov et al., 2020; Gasnikov

[2] In practice, this can happen when a non-symmetric noise is added to the stochastic gradients with symmetric noise, e.g., this can happen in some mechanisms for ensuring differential privacy such as the one from (Guo et al., 2023) (the non-symmetric part of the noise in the resulting vector after averaging over multiple clients can have a small variance when the number of clients is large, which is typical for modern Federated Learning applications (Kairouz et al., 2021)).

and Nesterov, 2016):

$$x^{k+1} = \frac{A_k y^k + \alpha_{k+1} z^k}{A_{k+1}}; \quad (13)$$

$$z^{k+1} = z^k - \alpha_{k+1}\,\mathrm{clip}\left(\nabla f_k(x^{k+1}); \lambda_{k+1}\right); \quad (14)$$

$$y^{k+1} = \frac{A_k y^k + \alpha_{k+1} z^{k+1}}{A_{k+1}}; \quad (15)$$

where $z^0 = y^0 = x^0$, $\lambda_0 = A_0$, $\lambda_{k+1} = \frac{k+2}{2aL}$ for some parameter $a_{k+1} > 0$, $A_{k+1} = A_k + \alpha_{k+1}$, and $\nabla f_k(x^{k+1})$ is an estimator satisfying Assumption 5.1 sampled independently from previous iterations. Below we formulate the main convergence result for clipped-SSTM in the convex case.

Theorem 5.7. Let Assumptions 2.2 and 2.3 with $\mu = 0$ hold on $Q = B_{3R}(x^*)$, where $R > \|x^0 - x^*\|$. Suppose that $\nabla f_k(x^{k+1})$ satisfies Assumption 5.1 with parameters b_k, σ_k for $k = 0, 1, \ldots, K$, $K > 0$ and $a = \max\left\{\min\left\{A^2, \frac{(K+1)^{3/2}\sqrt{A}}{4LR}\right\}, \frac{b(K+2)^2}{4LR}\right\}$, $\lambda_k = \max\left\{\frac{R}{\alpha_{k+1}A}\right\}$, where $A = \ln(4(K+1)/\beta)$ and $b = \max_{k=0,1,\ldots,K} b_k$, $\sigma = \max_{k=0,1,\ldots,K} \sigma_k$. Then the iterates produced by clipped-SSTM after K iterations with probability at least $1 - \beta$ satisfy

$$f(y^K) - f(x^*) = \tilde{\mathcal{O}}\left(\max\left\{\frac{LR^2}{K^2}, \frac{R\sigma}{\sqrt{K}}, bR\right\}\right).$$

As expected for accelerated methods, the above bound has a better $\tilde{\mathcal{O}}(K^{-2})$ deterministic term in contrast to the $\tilde{\mathcal{O}}(K^{-1})$ corresponding term in the upper bound for clipped-SGD. When the noise is symmetric and has bounded κ-th moment for some $\kappa > 0$ (not necessarily larger than 1), then one can construct an estimator with $b = 0$ and finite σ (see Proposition 4.1). In this case, the result matches (up to logarithmic factors) the optimal ones derived under bounded variance (Gorbunov et al., 2020) or sub-Gaussian noise (Ghadimi and Lan, 2012) assumptions. The improvement of the deterministic part can be utilized when parallel independent computations of the estimator are possible with marginal overheads (e.g., on communications/aggregation of the results of parallel computations).

Finally, for non-symmetric noise distributions satisfying Assumption 2.1, Theorem 5.7 with Lemma 4.5 imply the following result.

Corollary 5.8 (General noise). Let the assumptions of Theorem 5.7 hold and for all $x \in \mathbb{R}^d$ the noise $\nabla f(x) - \nabla f(x)$ satisfies Assumption 2.1. Then clipped-SSTM with $\nabla f_k(x^k)$ being the smoothed median of means of $O(n)$ samples $\nabla f(x^k)$ and $a = \max\left\{\min\left\{A^2, \frac{(K+1)^{3/2}\sqrt{A}}{4LR}\right\}, \frac{b(K+2)^2}{4LR}\right\}$, $\lambda_k = \max\left\{\frac{R}{\alpha_{k+1}A}\right\}$, where $A = \ln(4(K+1)/\beta)$ and b and σ defined in (10), with probability at least $1 - \beta$ after K iterations ensures that $f(y^K) - f(x^*)$ equals

$$\tilde{\mathcal{O}}\left(\max\left\{\frac{LR^2}{K^2}, \frac{\sqrt{(1+\ln^2 2)d + \ln\frac{1}{\beta}}\,\sigma R}{\sqrt{K}}, \frac{(1+\ln\frac{1}{\beta})CR}{2^n}\right\}\right),$$

where $n > \log_2\left(\max_{j \in [d]} M_j / \sigma^2\right)$ and C, D are defined in (11)-(12). The overall number of oracle calls equals $O(nK)$.

When the non-symmetric part is large, then the same comments are valid as the ones we make after Corollary 5.4. However, when the non-symmetric part is small, then there are regimes when the effect of acceleration is noticeable (for small enough σ).

For the strongly convex problems, we consider a restarted version of clipped-SSTM. We provide the results for this method in Appendix C.2.

6 NUMERICAL EXPERIMENTS

In this section, we illustrate the performance of clipped-SGD combined with the median and smoothed median of means on a simple quadratic problem:

$$\min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x; \quad (16)$$

where $A \in \mathbb{R}^{d \times d}$ is a randomly generated symmetric positive definite matrix. The code of numerical experiments is available on GitHub[3]. We consider stochastic gradients of the form $\nabla f(x) = Ax + \xi$, where ξ is an artificial noise following one of the following distributions.

Example 1. Cauchy distribution with the density $p_c(x) = \frac{1}{\pi(1+x^2)}$.

Example 2. The mixture of Cauchy and exponential distributions with the density $p(x) = 0.7\, p_c(x) + 0.3\, e^{-(x+1)} \mathbb{1}\{x > -1\}$.

Example 3. The mixture of Cauchy and Pareto distributions with the density $p(x) = 0.7\, p_c(x) + 0.3\, \frac{3}{(x+1.5)^4} \mathbb{1}\{x > -1.5\}$.

The experiments check the ability of median and smoothed median of means to deal with symmetric and asymmetrical heavy-tailed noise. We consider two examples of asymmetrical distributions with rapidly (Example 2) and slowly (Example 3) decaying antisymmetric part to examine its influence on the performance of optimization procedures.

We compare the following baselines:

- clipped-MB-SGD: clipped-SGD, where clipping is taken after mini-batching/averaging;

[3] https://github.com/Kutuz4/AISTATS2024_SMoM

Figure 1: Dependence of the mean error on the oracle calls number with a 95th and 5th percentile bounds.

- MB-clipped-SGD (mini-batched clipped-SGD, where averaging is taken after clipping);
- Med-MB-SGD (mini-batched SGD with median instead of averaging);
- clipped-Med-MB-SGD (mini-batched SGD with median instead of averaging and clipping operation after median);
- SMoM-MB-SGD (mini-batched SGD with the smoothed median of means);
- clipped-SMoM-MB-SGD (mini-batched SGD with clipping of the smoothed median of means).

For all methods, except for SMoM-MB-SGD and clipped-SMoM-MB-SGD, the batch size is 5, while for SMoM-MB-SGD and clipped-SMoM-MB-SGD we took SMoM$_{m,n}$ with $m = n = 2$. We have chosen $x_0 = 8\sqrt{d}\,(1; 1; 1; \ldots; 1)^\top$, where $d = 50$, as a starting point, launched all the methods 50 times and computed the average errors. The results are displayed in Figure 1.

In the case of a symmetric distribution, Med-MB-SGD and clipped-Med-MB-SGD perform better than clipped-SMoM-MB-SGD due to lower oracle calls count for one iteration. However, as we expected, Med-MB-SGD and clipped-Med-MB-SGD cannot achieve high accuracy in the case of asymmetric distributions due to the presence of a bias, while the smoothed median of means successfully adapts to this situation. Suddenly, averaging gradients after clipping has good performance in the asymmetric case, but it still converges slower compared to clipped-SMoM-MB-SGD. In terms of the number of steps, clipped-SMoM-MB-SGD converges much faster than other methods on asymmetric noise because it needs $(2m + 1)\,n = 10$ oracle calls on each iteration. We also see that SMoM-MB-SGD has a similar convergence rate to clipped-SMoM-MB-SGD in the case of distributions with less heavy tails.

7 CONCLUSION

In this work, we show that under some structural assumptions on the noise distribution with heavy tails, one can achieve faster convergence in solving of stochastic optimization problems. The key instrument we use is the smoothed median of means, which provably has a small bias and a finite variance for quite a wide class of distributions. Although our results are given for smooth convex/strongly convex problems, using similar technique, one can derive high-probability convergence results for smooth non-convex problems (Sadiev et al., 2023; Nguyen et al., 2023b), non-smooth convex and strongly convex problem (Gorbunov et al., 2021), variational inequalities under some structured non-monotonicity assumptions (Gorbunov et al., 2022), and composite and distributed optimization problems (Gorbunov et al., 2023). One can also improve the logarithmic factors in our results using the technique from (Nguyen et al., 2023a).

Acknowledgements

The work of A. Gasnikov 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 000000D730321P5Q0002) and the agreement with the Ivannikov Institute for System Programming of the Russian Academy of Sciences dated November 2, 2021 No. 70-2021-00142.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of

the 2016 ACM SIGSAC conference on computer and communications security, pages 308 318.

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33:1877 1901.

Chen, X., Wu, S. Z., and Hong, M. (2020). Understanding gradient clipping in private sgd: A geometric perspective. Advances in Neural Information Processing Systems, 33:13773 13782.

Cherapanamjeri, Y., Tripuraneni, N., Bartlett, P., and Jordan, M. (2022). Optimal mean estimation without a variance. In Conference on Learning Theory, pages 356 357. PMLR.

Cutkosky, A. and Mehta, H. (2021). High-probability bounds for non-convex stochastic optimization with heavy tails. Advances in Neural Information Processing Systems, 34:4883 4895.

Davis, D., Drusvyatskiy, D., Xiao, L., and Zhang, J. (2021). From low probability to high con dence in stochastic convex optimization. The Journal of Machine Learning Research, 22(1):2237 2274.

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

Devroye, L., Lerasle, M., Lugosi, G., and Oliveira, R. I. (2016). Sub-Gaussian mean estimators. The Annals of Statistics, 44(6):2695 2725.

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

Feller, W. (1971). An introduction to probability theory and its applications. Vol. II. Second edition. John Wiley & Sons Inc., New York.

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

Gasnikov, A. and Nesterov, Y. (2016). Universal fast gradient method for stochastic composit optimization problems. arXiv preprint arXiv:1604.05275 .

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 rst- and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization , 23(4):2341 2368.

Gorbunov, E., Danilova, M., Dobre, D., Dvurechenskii, P., Gasnikov, A., and Gidel, G. (2022). Clipped stochastic methods for variational inequalities with heavy-tailed noise. Advances in Neural Information Processing Systems, 35:31319 31332.

Gorbunov, E., Danilova, M., and Gasnikov, A. (2020). Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. Advances in Neural Information Processing Systems, 33:15042 15053.

Gorbunov, E., Danilova, M., Shibaev, I., Dvurechensky, P., and Gasnikov, A. (2021). Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. arXiv preprint arXiv:2106.05958.

Gorbunov, E., Sadiev, A., Danilova, M., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A., and Richtárik, P. (2023). High-probability convergence for composite and distributed stochastic minimization and variational inequalities with heavy-tailed noise. arXiv preprint arXiv:2310.01860 .

Guo, C., Chaudhuri, K., Stock, P., and Rabbat, M. (2023). Privacy-aware compression for federated learning through numerical mechanism design. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pages 11888 11904. PMLR.

Harvey, N. J., Liaw, C., and Randhawa, S. (2019). Simple and optimal high-probability bounds for strongly-convex stochastic gradient descent. arXiv preprint arXiv:1909.00843.

Indritz, J. (1961). An inequality for hermite polynomials. Proceedings of the American Mathematical Society, 12(6):981 983.

Jakoveti¢, D., Bajovi¢, D., Sahu, A. K., Kar, S., Milœsevi¢, N., and Stamenkovi¢, D. (2023). Nonlinear gradient mappings and stochastic optimization: A general framework with applications to heavy-tail noise. SIAM Journal on Optimization , 33(2):394 423.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. Foundations and Trends® in Machine Learning, 14(1 2):1 210.

Karimireddy, S. P., He, L., and Jaggi, M. (2021). Learning from history for byzantine robust optimization. In International Conference on Machine Learning, pages 5311 5319. PMLR.

Lecué, G. and Lerasle, M. (2020). Robust machine learning by median-of-means: Theory and practice. The Annals of Statistics, 48(2):906 931.

Li, X. and Orabona, F. (2020). A high probability analysis of adaptive sgd with momentum. arXiv preprint arXiv:2007.14294.

Lugosi, G. and Mendelson, S. (2019a). Mean estimation and regression under heavy-tailed distributions: A survey. Foundations of Computational Mathematics, 19(5):1145 1190.

Lugosi, G. and Mendelson, S. (2019b). Sub-Gaussian estimators of the mean of a random vector. The Annals of Statistics, 47(2):783 794.

Lugosi, G. and Mendelson, S. (2020). Risk minimization by median-of-means tournaments. Journal of the European Mathematical Society 22(3):925 965.

Mai, V. V. and Johansson, M. (2021). Stability and convergence of stochastic gradient clipping: Beyond lipschitz continuity and smoothness. In International Conference on Machine Learning, pages 7325 7335. PMLR.

Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. Bernoulli, 21(4):2308 2335.

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, 80:1607 1627.

Necoara, I., Nesterov, Y., and Glineur, F. (2019). Linear convergence of first order methods for non-strongly convex optimization. Mathematical Programming, 175:69 107.

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

Nemirovskij, A. S. and Yudin, D. B. (1983). Problem complexity and method efficiency in optimization.

Nguyen, T. D., Ene, A., and Nguyen, H. L. (2023a). Improved convergence in high probability of clipped gradient methods with heavy tails. arXiv preprint arXiv:2304.01119.

Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. L. (2023b). High probability convergence of clipped-sgd under heavy-tailed noise. arXiv preprint arXiv:2302.05437.

OpenAI (2023). GPT-4 technical report.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In International conference on machine learning pages 1310 1318. Pmlr.

Rakhlin, A., Shamir, O., and Sridharan, K. (2011). Making gradient descent optimal for strongly convex stochastic optimization. arXiv preprint arXiv:1109.5647.

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 Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, Proceedings of the 40th International Conference on Machine Learning volume 202 of Proceedings of Machine Learning Research, pages 29563 29648. PMLR.

Vural, N. M., Yu, L., Balasubramanian, K., Volgushev, S., and Erdogdu, M. A. (2022). Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In Conference on Learning Theory, pages 65 102. PMLR.

Zhang, J., He, T., Sra, S., and Jadbabaie, A. (2019). Why gradient clipping accelerates training: A theoretical justification for adaptivity. arXiv preprint arXiv:1905.11881.

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S., Kumar, S., and Sra, S. (2020). Why are adaptive methods good for attention models? Advances in Neural Information Processing Systems, 33:15383 15393.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes, see Sections 2, 4 and 5.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes, see our results in Sections 4 and 5.
   (c) (Optional) Anonymized source code, with speci cation of all dependencies, including external libraries. Yes, we provide the source code with supplementary materials.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes.
   (b) Complete proofs of all theoretical results. Yes, the proofs are collected in Appendix.
   (c) Clear explanations of any assumptions. Yes.

3. For all gures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes, see Section 6 and Appendix.
   (c) A clear de nition of the speci c measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes.
   (d) A description of the computing infrastructure used. Not Applicable. We do not use any computing infrastructure, except for an ordinary laptop.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Not Applicable.
   (b) The license information of the assets, if applicable. Not Applicable.
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable.
   (d) Information about consent from data providers/curators. Not Applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identi able information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

Contents

A SMOOTHED MEDIAN OF MEANS ESTIMATOR

A.1 Proof of Proposition 4.1

Let P_j be the cumulative distribution function of ξ_j:

$$P_j(t) = \int_{-\infty}^{t} p_j(u)du; \quad \text{for all } t \in \mathbb{R}.$$

Then the probability density of the median $\mathrm{Med}(\xi_{j,1}, \ldots, \xi_{j,2m+1})$ is given by

$$(2m+1)\binom{2m}{m} P_j(t)^m \left(1 - P_j(t)\right)^m p_j(t).$$

Let us prove that the variance is bounded. To be more precise, we are going to show that

$$\int_{-\infty}^{\infty} t^2 P_j(t)^m \left(1 - P_j(t)\right)^m p_j(t)\,dt \le \left(1 - \frac{4B_j}{\varkappa_j}^{2/\alpha_j}\right) 4^{-m}.$$

Since, according to the conditions of the proposition, $p_j(u) \le B_j/(1+|u|)^{1+\alpha_j}$ for any $u \in \mathbb{R}$, it holds that

$$1 - P_j(t) \le \int_{t}^{\infty} \frac{B_j}{u^{1+\alpha_j}}du = \frac{B_j}{\alpha_j}\frac{1}{t^{\alpha_j}} \quad \text{for any } t > 1.$$

Similarly, for any $t \le -1$, we have

$$P_j(t) \le \frac{B_j}{\alpha_j}\frac{1}{|t|^{\alpha_j}}.$$

Hence, for any $t \in \mathbb{R}$, $P_j(t)\left(1 - P_j(t)\right)$ satisfies the inequality

$$P_j(t)\left(1 - P_j(t)\right) \le \begin{cases} 4^{-m} \wedge \left(\frac{B_j}{\alpha_j|t|^{\alpha_j}}\right)^m & \text{if } |t| > 1; \\ 4^{-m} & \text{otherwise.} \end{cases} \tag{17}$$

This implies that the integral of $t^2 P_j(t)^m(1 - P_j(t))^m p_j(t)$ over the real line does not exceed

$$\begin{aligned}
\int_{-\infty}^{\infty} t^2 P_j(t)^m \left(1 - P_j(t)\right)^m p_j(t)\,dt &\le \sup_{t\in\mathbb{R}} t^2 P_j(t)^m \left(1 - P_j(t)\right)^m \int_{-\infty}^{\infty} p_j(u)\,du \\
&= \sup_{t\in\mathbb{R}} t^2 P_j(t)^m \left(1 - P_j(t)\right)^m \\
&= \max\left(\sup_{|t|\le 1} t^2 P_j(t)^m \left(1 - P_j(t)\right)^m; \sup_{|t|>1} t^2 P_j(t)^m \left(1 - P_j(t)\right)^m\right) \\
&\le \max\left(4^{-m}; \sup_{|t|>1} t^2 P_j(t)^m \left(1 - P_j(t)\right)^m\right).
\end{aligned}$$

We use (17) to bound the supremum in the right-hand side:

$$\begin{aligned}
\sup_{|t|>1} t^2 P_j(t)^m \left(1 - P_j(t)\right)^m &\le \sup_{|t|>1} t^2 \left(\frac{1}{4} \wedge \frac{B_j}{\alpha_j|t|^{\alpha_j}}\right)^m \\
&\le \max\left(\sup_{1\le|t|\le(4B_j/\alpha_j)^{1/\alpha_j}} t^2 4^{-m}; \sup_{|t|>(4B_j/\alpha_j)^{1/\alpha_j}} t^2 \left(\frac{B_j}{\alpha_j|t|^{\alpha_j}}\right)^m\right).
\end{aligned}$$

If $m_j > 2$, then we have

$$\sup_{|t|>(4B_j/\sigma_j)^{1/\alpha_j}} t^2 \left(\frac{B_j}{\sigma_j |t|^{\alpha_j}}\right)^m = \left(\frac{4B_j}{\sigma_j}\right)^{2-\alpha_j} \le 4^m,$$

and, hence,

$$\sup_{|t|>1} |t^2 P_j(t)^m (1 - P_j(t))^m| \le \sup_{|t|>1} t^2 \left(\frac{1}{4} \wedge \frac{B_j}{\sigma_j |t|^{\alpha_j}}\right)^m \le \left(1 - \frac{4B_j}{\sigma_j}\right)^{2-\alpha_j} \le 4^m.$$

Thus, we obtain that

$$\int_{-1}^{1} |t^2 P_j(t)^m (1 - P_j(t))^m| p_j(t)\, dt \le \sup_{t \in \mathbb{R}} |t^2 P_j(t)^m (1 - P_j(t))^m| \le \left(1 - \frac{4B_j}{\sigma_j}\right)^{2-\alpha_j} \le 4^m.$$

It only remains to note that

$$\binom{2m}{m} = \frac{(2m)!}{m!\, m!} = \prod_{j=1}^{m} \frac{2j}{j} \prod_{j=1}^{m} \frac{2j-1}{j} \le 2^m \cdot 2^m = 4^m$$

to derive the desired bound

$$\mathbb{E}\, \mathrm{Med}(\xi_{j,1}, \ldots, \xi_{j,2m+1})^2 \le (2m+1) \left(1 - \frac{4B_j}{\sigma_j}\right)^{2-\alpha_j}. \tag{18}$$

Concerning the expectation of $\mathrm{Med}(\xi_{j,1}, \ldots, \xi_{j,2m+1})$, we point out that it is finite, because $\mathrm{Med}(\xi_{j,1}, \ldots, \xi_{j,2m+1})$ has a finite second moment. Moreover, due to the symmetry of p_j, we have $P_j(-t) = 1 - P_j(t)$ and, thus,

$$(-t) P_j(-t)^m (1 - P_j(-t))^m p_j(-t) = -t (1 - P_j(t))^m P_j(t)^m p_j(t) \quad \text{for all } t \in \mathbb{R}.$$

Hence, it holds that

$$\mathbb{E}\, \mathrm{Med}(\xi_{j,1}, \ldots, \xi_{j,2m+1}) = \int_{-1}^{1} t P_j(t)^m (1 - P_j(t))^m p_j(t)\, dt = 0.$$

A.2 Proof of Lemma 4.3

Let $\Gamma_{11}, \ldots, \Gamma_{dd}$ be the diagonal elements of Γ. Denote the difference $\nabla f(x) - \nabla f(x)$ by $\Delta = (\Delta_1, \ldots, \Delta_d)^\top$. It is enough to show that

$$|\mathbb{E}\, \mathrm{SMoM}_{m,n}(\Delta_j, \Gamma_{jj})| \le (2m+1) \left(\frac{\Gamma_{jj}}{2n} \left(2 + \frac{k}{n}\right)\right)^{1/2} \left(\frac{4m}{(2m-1)\sqrt{2e}} + \frac{m(2m-1)}{(2m-2)\, e} \frac{k}{2n} + 32 \frac{mk}{2n}\right)^2$$
$$\cdot (2m+1) \left(\frac{\Gamma_{jj}}{n} \left(1 + \frac{k}{2n}\right)\right)^{1/2} \left(1 + \frac{mk}{2n} + \frac{mk}{2n}\right)^2$$

and

$$\mathbb{E}\, \mathrm{SMoM}_{m,n}(\Delta_j, \Gamma_{jj})^2 \le 4(2m+1) \left(\frac{\Gamma_{jj}}{n} + \Delta_j^2\right).$$

for all $j \in \{1, \ldots, d\}$. We start with the upper bound on the second moment. We split the rest of the proof into several steps for convenience.

Step 1: bound on the second moment. For a xed $j \in \{1,\dots,d\}$, let $p_j(u)$ be the marginal density of ξ_j and let

$$F(t) = \int_{-\infty}^{t} p_j^{*n}(u)\,du$$

stand for the cumulative distribution function of $\mathrm{Mean}(\xi_{j,1},\dots,\xi_{j,n}) + \eta_j$, where $\xi_{j,1},\dots,\xi_{j,n}$ are i.i.d. copies of ξ_j. Then the density of $\mathrm{SMoM}_{k,n}(\xi_j,\eta)$ is equal to

$$(2m+1)\binom{2m}{m} F(t)^m\,(1-F(t))^m\,F'(t).$$

If we manage to prove that

$$\sup_{t\in\mathbb{R}} t^2 F(t)^m\,(1-F(t))^m \le \frac{4(\sigma_j^2/n+\varkappa^2)}{4^m};$$

then we immediately obtain

$$\int_{-\infty}^{+\infty} t^2 F(t)^m\,(1-F(t))^m\,F'(t)\,dt \le \sup_{t\in\mathbb{R}} t^2 F(t)^m\,(1-F(t))^m \int_{-\infty}^{+\infty} F'(u)\,du = \sup_{t\in\mathbb{R}} t^2 F(t)^m\,(1-F(t))^m.$$

Let $\xi_{j,1},\dots,\xi_{j,n}$ be independent copies of ξ_j and let $\eta_j \sim \mathcal{N}(0,1)$ be a Gaussian random variable, which is independent of $\xi_{j,1},\dots,\xi_{j,n}$. Then, according to the de nition of $F(t)$, for any $t \in \mathbb{R}$, it holds that

$$1-F(t) = P\left\{\frac{\xi_{j,1}+\dots+\xi_{j,n}}{n} + \eta_j > t\right\}.$$

Since $\xi_{j,1},\dots,\xi_{j,n}$ and η_j have nite variance, we can apply Chebyshev's inequality to derive an upper bound on the right tail of $F(t)$:

$$1-F(t) \le \frac{\mathrm{E}(\xi_{j,1}+\dots+\xi_{j,n})^2/n^2 + \varkappa^2\mathrm{E}\eta_j^2}{t^2} = \frac{\sigma_j^2}{nt^2} + \frac{\varkappa^2}{t^2} \quad \text{for all } t > 0.$$

Similarly, for any $t < 0$, we have

$$F(t) \le \frac{\sigma_j^2}{nt^2} + \frac{\varkappa^2}{t^2}.$$

Combining these bounds with the inequality $F(t)(1-F(t)) \le 1/4$, which holds for any $t \in \mathbb{R}$, we deduce that

$$F(t)\,(1-F(t)) \le \frac{1}{4}\wedge\left(\frac{\sigma_j^2}{nt^2}+\frac{\varkappa^2}{t^2}\right) = \begin{cases} \frac{1}{4}, & \text{if } |t| \le 2\sqrt{\frac{\sigma_j^2}{n}+\varkappa^2}, \\ \frac{\sigma_j^2}{n}+\varkappa^2, & \text{otherwise.}\end{cases} \tag{19}$$

Hence, for any $m > 1$,

$$\sup_{t\in\mathbb{R}} t^2 F(t)^m\,(1-F(t))^m \le \max_{t^2\le 4(\sigma_j^2/n+\varkappa^2)} \frac{t^2}{4^m} \vee \max_{t^2>4(\sigma_j^2/n+\varkappa^2)} t^2\left(\frac{\sigma_j^2}{nt^2}+\frac{\varkappa^2}{t^2}\right)^m = \frac{\sigma_j^2/n+\varkappa^2}{4^{m-1}};$$

as we announced. This implies that

$$\mathrm{E}\,\mathrm{SMoM}_{k,n}(\xi_j,\eta)^2 \le (2m+1)\binom{2m}{m}\int_{-\infty}^{+\infty} t^2 F(t)^m\,(1-F(t))^m\,F'(t)\,dt \le (2m+1)\binom{2m}{m}\frac{\sigma_j^2/n+\varkappa^2}{4^{m-1}}.$$

Similarly to the proof of Proposition 4.1, we use the inequality

$$\binom{2m}{m} = \frac{(2m)!}{m!\,m!} = \prod_{j=1}^{m}\frac{2j}{j}\prod_{j=1}^{m}\frac{2j-1}{j} \le 4^m;$$

which yields that

$$\mathbb{E}\,\mathrm{SMoM}_{m,n}(\Delta_j;\sigma)^2 \le 4(2m+1)\left(\frac{\sigma\sigma}{n}+\Delta^2\right):$$

Step 2: bound on the expectation. The rest of the proof is devoted to an upper bound on the expectation of $\mathrm{SMoM}_{m,n}(\Delta_j;\sigma)$. One could apply the Cauchy-Schwarz inequality to show that $\mathbb{E}\,\mathrm{SMoM}_{m,n}(\Delta_j;\sigma)$ decreases with the growth of n. However, we are going to prove a stronger bound. Our approach is based on decomposition of $p_j(u)$ into the sum of symmetric and antisymmetric part:

$$p_j(u) = s_j(u) + r_j(u), \quad \text{where} \quad s_j(u) = \frac{p_j(u)+p_j(\Delta-u)}{2} \quad \text{and} \quad r_j(u) = \frac{p_j(u)-p_j(\Delta-u)}{2}:$$

If r_j was equal to zero, we could say that $\mathbb{E}\,\mathrm{SMoM}_{m,n}(\Delta_j;\sigma) = 0$ as well. However, in a general situation, has some impact on the mean of $\mathrm{SMoM}_{m,n}(\Delta_j;\sigma)$. To quantify it, we compare the integrals

$$\int_1^{+1} t\,F(t)^m\left(1-F(t)\right)^m F'(t)dt \quad \text{and} \quad \int_1^{+1} t\,G(t)^m\left(1-G(t)\right)^m G'(t)dt;$$

where G is a cumulative distribution function defined as

$$G(t) = \int_1^{+1} t\,\frac{u}{n}\,s_j^{\,n}(u)\,du:$$

In other words, G corresponds to the CDF of $\mathrm{Mean}(\xi_{j,1};\ldots;\xi_{j,n})+\Delta_j$, where $\xi_{j,1};\ldots;\xi_{j,n}$ are i.i.d. copies of ξ_j and $\xi_j \sim \mathcal{N}(0;1)$, in the symmetric case. We are going to show that r_j has minor influence on the expectation of the smoothed median of means if m and n are sufficiently large.

First, let us show that the cumulative distribution functions $F(t)$ and $G(t)$ are close to each other. It is straightforward to check that

$$\sup_{x\in\mathbb{R}}\left|\varphi''(x)\right| = \frac{1}{\sqrt{2\pi}}\,\frac{1}{\sigma^2}\sup_{x\in\mathbb{R}}\left|\frac{x}{\sigma}e^{-x^2/(2\sigma^2)}\right| \le \frac{1}{\sqrt{2\pi}\,\sigma^2}\,ye^{-y^2/2}\Big|_{y=1} = \frac{1}{\sqrt{2\pi e}\,\sigma^2}: \tag{20}$$

Then Lemma A.1 (see Appendix A.4 below) implies that

$$\left|F(t)-G(t)\right| \le \frac{1}{2}\sqrt{\frac{1}{2\pi e}}\,\frac{\sigma\sigma}{\sigma^2 n} \quad \text{for all } t\in\mathbb{R}. \tag{21}$$

In view of (19), for any $m > 3$ it holds that

$$|t|\left|F(t)\right|^{m-1}\left(1-F(t)\right)^m \to 0 \quad \text{and} \quad |t|\left|F(t)\right|^{m-1}\left(1-F(t)\right)^{m-1} \to 0 \quad \text{as } t \to \pm\infty.$$

Then, according to Lemma A.2, we have

$$\begin{aligned}
&\int_1^{+1} t\,F(t)^m\left(1-F(t)\right)^m F'(t)\,dt - \int_1^{+1} t\,G(t)^m\left(1-G(t)\right)^m G'(t)\,dt \\
&\le \sqrt{\frac{1}{2\pi e}}\,\frac{\sigma\sigma}{\sigma^2 n}\int_1^{+1}\left|F(t)^m\left(1-F(t)\right)^m\right|dt + \frac{m}{16\,e}\left(\frac{\sigma\sigma}{\sigma^2 n}\right)^2\int_1^{+1}\left|F(t)^{m-1}\left(1-F(t)\right)^{m-1}\right|dt \\
&\quad + \frac{m}{\sigma^2 n}\sup_{t\in\mathbb{R}}|t|\max_{2[F(t)\wedge G(t);F(t)\vee G(t)]}\left|\xi^{m-2}(1-\xi)^{m-2}\right|:
\end{aligned}$$

Let us remind the reader that the CDF $F(t)$ satisfies the inequalities

$$1-F(t) \le \frac{\sigma\sigma}{nt^2}+\frac{\Delta^2}{t^2} \quad \text{for all } t>0 \text{ and} \quad F(t) \le \frac{\sigma\sigma}{nt^2}+\frac{\Delta^2}{t^2} \quad \text{for all } t<0:$$

Due to the Chebyshev inequality, a similar bound holds for $G(t)$:

$$1 - G(t) \le \frac{\sigma^2}{nt^2} + \frac{\sigma^2}{t^2} \quad \text{for all } t > 0 \text{ and } \quad G(t) \le \frac{\sigma^2}{nt^2} + \frac{\sigma^2}{t^2} \quad \text{for all } t < 0.$$

This yields that

$$\int_1^{+\infty} t\,F(t)^{m-1}\left(1 - F(t)\right)^m F'(t)\,dt \le \int_1^{+\infty} t\,G(t)^{m-1}\left(1 - G(t)\right)^m G'(t)\,dt$$
$$\le \frac{1}{2\sqrt{2e}}\,\frac{\sigma^2}{n}\int_1^{+\infty}\left(\frac{1}{4} \wedge \frac{\sigma^2/n + 2}{t^2}\right)^m dt + \frac{m}{16e}\,\frac{\sigma^2}{n}\int_1^{+\infty}\left(\frac{1}{4} \wedge \frac{\sigma^2/n + 2}{t^2}\right)^{m-1} dt$$
$$+ \frac{m\sigma^2}{2n}\left(\sup_{t \in \mathbb{R}} |t|\left(\frac{1}{4} \wedge \frac{\sigma^2/n + 2}{t^2}\right)^{m-2}\right).$$

Applying Proposition A.3, we obtain that

$$\int_1^{+\infty} t\,F(t)^{m-1}\left(1 - F(t)\right)^m F'(t)\,dt \le \int_1^{+\infty} t\,G(t)^{m-1}\left(1 - G(t)\right)^m G'(t)\,dt$$
$$\le \frac{1}{2\sqrt{2e}}\,\frac{\sigma^2}{n}\,\frac{2m}{(2m-1)\,4^{m-1}}\sqrt{2 + \frac{\sigma^2}{n}}\sqrt{2 + \frac{\sigma^2}{n}}$$
$$+ \frac{m}{16e}\,\frac{\sigma^2}{n}\,\frac{2m-1}{(2m-2)\,4^{m-2}}\left(2 + \frac{\sigma^2}{n}\right)$$
$$+ \frac{m\sigma^2}{2n}\left(\sup_{t \in \mathbb{R}} |t|\left(\frac{1}{4} \wedge \frac{\sigma^2/n + 2}{t^2}\right)^{m-2}\right).$$

Similarly to Step 1, we can prove that

$$\left(\sup_{t \in \mathbb{R}} |t|\left(\frac{1}{4} \wedge \frac{2 + \sigma^2/n}{t^2}\right)^{m-2}\right) = \frac{2^{\sqrt{2 + \sigma^2/n}}}{4^{m-2}};$$

and then it holds that

$$\int_1^{+\infty} t\,F(t)^{m-1}\left(1 - F(t)\right)^m F'(t)\,dt \le \int_1^{+\infty} t\,G(t)^{m-1}\left(1 - G(t)\right)^m G'(t)\,dt$$
$$\le \left(\frac{m}{(2m-1)\,4^{m-1}\sqrt{2e}}\,\frac{\sigma^2}{n} + \frac{m(2m-1)}{(2m-2)\,4^m e}\,\frac{\sigma^2}{n}\right)^2 + \frac{32}{4^m}\left(\frac{m\sigma^2}{2n}\right)^2\sqrt{2 + \frac{\sigma^2}{n}}.$$

Taking into account

$$\binom{2m}{m} = \frac{(2m)!}{m!\,m!} = \prod_{j=1}^{m}\frac{2j}{j}\prod_{j=1}^{m}\frac{2j-1}{j} \le 4^m;$$

we immediately obtain that

$$\mathbb{E}\,\mathrm{SMoM}_{k,n}(\xi_j;\tau) \le (2m+1)\left(\frac{\sigma^2}{n}\sqrt{1 + \frac{\sigma^2}{2n}}\left(\frac{4m}{(2m-1)\sqrt{2e}} + \frac{m(2m-1)}{(2m-2)\,e}\,\frac{\sigma^2}{2n}\right) + 32\left(\frac{m\sigma^2}{2n}\right)^2\right)$$
$$\le (2m+1)\left(\frac{\sigma^2}{n}\sqrt{1 + \frac{k-k}{2n}}\left(\frac{4m}{(2m-1)\sqrt{2e}} + \frac{m(2m-1)}{(2m-2)\,e}\,\frac{k-k}{2n}\right) + 32\left(\frac{mk-k}{2n}\right)^2\right)$$
$$\le (2m+1)\left(\frac{\sigma^2}{n}\sqrt{1 + \frac{k-k}{2n}}\left(1 + \frac{mk-k}{2n}\right) + \left(\frac{mk-k}{2n}\right)^2\right).$$

A.3 Proof of Lemma 4.5

Let $\Delta = (\Delta_1, \ldots, \Delta_d)^\top$ stand for the difference $r f(x) - r f(x)$ and let us show that, for any $j \in \{1, \ldots, d\}$, it holds that

$$\mathbb{E}\,\mathrm{SMoM}_{m,n}(\Delta_j, \tau) \le \frac{m M_j}{2n}\left(1 + \tau + \left(\frac{2^j B_j}{n^{j-1}}\right)^{1/j}\right).$$

and

$$\mathbb{E}\,\mathrm{SMoM}_{m,n}(\Delta_j, \tau)^2 \le m\left(1 + \tau^2 + \left(\frac{M_j}{n}\right)^2 + \left(\frac{2^j B_j}{\tau^j n^{j-1}}\right)^{2/j} + \left(\frac{B_j M_j}{n^j}\right)^{2/(j+1)}\right).$$

From now on, we fix an arbitrary $j \in \{1, \ldots, d\}$. Similarly to the proof of Lemma 4.3, the core idea is to compare the cumulative distribution functions

$$F(t) = \int_1^{+\infty} t^{\frac{u}{n}}\, p_j^n(u)\, du \quad \text{and} \quad G(t) = \int_1^{+\infty} t^{\frac{u}{n}}\, s_j^n(u)\, du.$$

The first one is directly related to the density of $\mathrm{SMoM}_{m,n}(\Delta_j, \tau)$, which is equal to

$$(2m+1)\binom{2m}{m} F(t)^m \left(1 - F(t)\right)^m F'(t).$$

However, the proof of Lemma 4.5 is far more technical. The main obstacle is that we cannot use Chebyshev's inequality to specify the rate of decay of $F(t) - 1 - F(t)$ and of $G(t) - 1 - G(t)$ as t approaches infinity. Instead, we prove the following non-trivial result (see Lemma A.4 below): if Assumption 2.1 holds and $2M_j \le n^{j-1}$, then, for any $t \in \mathbb{R}$ it holds that

$$|F(t) - G(t)| \le \frac{M_j}{nt}\left(1 + \frac{1}{t} + \frac{B_j}{n^{j-1}|t|^j}\right).$$

Combining this result with the bound on the second derivative of (20) and Lemma A.1, we obtain that

$$|F(t) - G(t)| \le \frac{M_j}{2n}\left(1 \wedge \left(\frac{1}{t} + \frac{B_j}{n^{j-1}|t|^{j+1}}\right)\right). \tag{22}$$

Despite the simple statement, the proof of Lemma A.4 is quite technical. A reader can find it in Appendix A.4. With the bound (22) at hand, the proof of Lemma 4.5 is relatively simple. For convenience, we divide it into several steps.

Step 1: a bound on the tails of G. The goal of this step is to specify the rate of decay of $G(t) - 1 - G(t)$ as t tends to infinity. First, consider the case $t > 0$. By the definition of $G(t)$, it holds that

$$\begin{aligned}
1 - G(t) &= \int_1^{+\infty} \left(1 - t^{\frac{u}{n}}\right) s_j^n(u)\, du \\
&= n \int_1^{+\infty} \left(1 - t^{-y}\right) s_j^n(nt + ny)\, dy \\
&= n \int_1^{t/2} \left(1 - t^{-y}\right) s_j^n(nt + ny)\, dy + n \int_{t/2}^{+\infty} \left(1 - t^{-y}\right) s_j^n(nt + ny)\, dy.
\end{aligned} \tag{23}$$

If $y \le t/2$, then

$$1 - t^{-y} \le 1 - t^{-t/2} \le \exp\left(\frac{t^2}{8}\right);$$

and we have

$$n \int_{-1}^{t/2} (1-y)\, s_j^n(nt+ny)\,dy \le \exp\left(-\frac{t^2}{8\sigma^2}\right) \int_{-1}^{t/2} s_j^n(nt+ny)\,n\,dy \le \exp\left(-\frac{t^2}{8\sigma^2}\right). \tag{24}$$

Otherwise, if $y > t/2$, then, due to Assumption 2.1, it holds that

$$n \int_{t/2}^{+1} (1-y)\, s_j^n(nt+ny)\,dy \le \int_{t/2}^{+1} \frac{B_j n^2}{n^{(1+\alpha_j)/\alpha_j} + n^{1+\alpha_j}(t+y)^{1+\alpha_j}}\,dy$$
$$\le \int_0^{+1} \frac{B_j}{n^{\alpha_j-1}(t/2+v)^{1+\alpha_j}}\,dv = \frac{2^{\alpha_j} B_j}{\alpha_j n^{\alpha_j-1} t^{\alpha_j}}. \tag{25}$$

Plugging the inequalities (24), (25) into (23), we obtain that

$$1 - G(t) \le \frac{2^{\alpha_j} B_j}{n^{\alpha_j-1} t^{\alpha_j}} + \exp\left(-\frac{t^2}{8\sigma^2}\right) \quad \text{for all } t > 0.$$

Similarly, we can prove that

$$G(t) \le \frac{2^{\alpha_j} B_j}{\alpha_j n^{\alpha_j-1} |t|^{\alpha_j}} + \exp\left(-\frac{t^2}{8\sigma^2}\right) \quad \text{for all } t < 0.$$

Hence, for any $t \in \mathbb{R}$, we have

$$G(t) - 1 + G(t) \le \min\left\{\frac{1}{4}; \frac{2^{\alpha_j} B_j}{\alpha_j n^{\alpha_j-1}|t|^{\alpha_j}} + \exp\left(-\frac{t^2}{8\sigma^2}\right)\right\}. \tag{26}$$

Step 2: bound on the second moment. The second moment of $\mathrm{SMoM}_{k,n}(\alpha_j; \beta)$ satisfies the inequality

$$\mathbb{E}\,\mathrm{SMoM}_{k,n}(\alpha_j; \beta)^2 = (2m+1)\binom{2m}{m} \int_{-1}^{+1} t^2 F(t)^m \big(1 - F(t)\big)^m F'(t)\,dt$$
$$\le (2m+1)\binom{2m}{m} \sup_{t \in \mathbb{R}} t^2 F(t)^m \big(1 - F(t)\big)^m.$$

On the other hand, (22) and (26) imply that

$$F(t) - 1 + F(t) \le \min\left\{\frac{1}{4}; \frac{2^{\alpha_j} B_j}{\alpha_j n^{\alpha_j-1}|t|^{\alpha_j}} + \exp\left(-\frac{t^2}{8\sigma^2}\right)\right\} + \frac{M_j}{nt}\left(1 + \frac{B_j}{n^{\alpha_j-1}|t|^{\alpha_j}}\right) \quad \text{for all } t \in \mathbb{R}. \tag{27}$$

This yields

$$\sup_{t \in \mathbb{R}} t^2 F(t)^m \big(1 - F(t)\big)^m \le 4^m \left(1 + \sqrt{2} + \frac{M_j}{\sigma}\right)^2 + \left(\frac{2^{\alpha_j} B_j}{\alpha_j n^{\alpha_j-1}}\right)^{2/\alpha_j} + \left(\frac{B_j M_j}{n^{\alpha_j}}\right)^{2/(\alpha_j+1)} \quad \text{if } m \ge \alpha_j > 2.$$

Since

$$\binom{2m}{m} = \frac{(2m)!}{m!\,m!} = \prod_{j=1}^{m} \frac{2j}{j} \prod_{j=1}^{m} \frac{2j-1}{j} \le 4^m,$$

we obtain that

$$\mathbb{E}\,\mathrm{SMoM}_{k,n}(\alpha_j; \beta)^2 \le m\left(1 + \sqrt{2} + \frac{M_j}{\sigma}\right)^2 + \left(\frac{2^{\alpha_j} B_j}{\alpha_j n^{\alpha_j-1}}\right)^{2/\alpha_j} + \left(\frac{B_j M_j}{n^{\alpha_j}}\right)^{2/(\alpha_j+1)}.$$

Step 3: bound on the expectation. It remains to bound the expectation of $\mathrm{SMoM}_{m,n}(\ _j;\ )$. For this purpose, we use Lemma A.2, which yields that

$$
\begin{aligned}
&(2m+1)\binom{2m}{m}^{-1} j\mathbb{E}\mathrm{SMoM}_{m,n}(\ _j;\ )j \\
&= \int_1^1 t\,F(t)^m\ \ 1\ \ F(t)\ ^m F'(t)\,dt \int_1^1 t\,G(t)^m\ \ 1\ \ G(t)\ ^m G'(t)\,dt \\
&\quad 6 \int_1^1 G(t)^m\ \ 1\ \ G(t)\ ^m\ F(t)\ \ G(t)\ dt + \frac{m}{2}\int_1^1 G(t)^{m\ \ 1}\ 1\ \ G(t)\ ^{m\ \ 1}\ F(t)\ \ G(t)\ ^2\,dt \\
&\quad + m^2 \sup_{t2R} jtj\ F(t)\ \ G(t)\ ^2 \max_{2[F(t)\wedge G(t);F(t)\_\ G(t)]}{}^{m\ \ 2}(1\ \ )^{m\ \ 2}\ :
\end{aligned}
$$

Note that the requirement

$$
jtjG(t)^m\ \ 1\ \ G(t)\ ^m\ F(t)\ \ G(t)\ !\ \ 0 \quad \text{and} \quad jtjG(t)^{m\ \ 1}\ 1\ \ G(t)\ ^{m\ \ 1}\ F(t)\ \ G(t)\ ^2\ !\ \ 0 \quad t\ !\ 1
$$

is satis ed, because of the inequalities(22), (26) and the conditions of the lemma. These inequalities also imply that

$$
\begin{aligned}
&\int_1^1 t\,F(t)^m\ \ 1\ \ F(t)\ ^m F'(t)\,dt \int_1^1 t\,G(t)^m\ \ 1\ \ G(t)\ ^m G'(t)\,dt \\
&\cdot \int_1^1 \frac{1}{4}\wedge \frac{2\ ^j B}{_j n^{\ j\ \ 1} jtj^{\ j}} + \exp\left(\frac{t^2}{8\ ^2}\right)^m F(t)\ \ G(t)\ dt \\
&\quad + \frac{m}{2}\int_1^1 \frac{1}{4}\wedge \frac{2\ ^j B_j}{_j n^{\ j\ \ 1} jtj^{\ j}} + \exp\left(\frac{t^2}{8\ ^2}\right)^{m\ \ 1} F(t)\ \ G(t)\ ^2\,dt \\
&\quad + m^2 \sup_{t2R} jtj\ F(t)\ \ G(t)\ ^2 \left(\frac{1}{4}\wedge \frac{2\ ^j B_j}{_j n^{\ j\ \ 1} jtj^{\ j}} + \frac{M_j}{nt}\ 1 + \frac{B_j}{n^{\ j\ \ 1} jtj^{\ j}} + \exp\left(\frac{t^2}{8\ ^2}\right)^{m\ \ 2}\right)\ :
\end{aligned}
$$

Taking into account the bound (22) on the absolute value of the di erence$F(t)\ \ G(t)$, we obtain that

$$
\begin{aligned}
&\int_1^1 t\,F(t)^m\ \ 1\ \ F(t)\ ^m F'(t)\,dt \int_1^1 t\,G(t)^m\ \ 1\ \ G(t)\ ^m G'(t)\,dt \\
&\cdot \frac{M_j}{^2 n}\int_1^1 \frac{1}{4}\wedge \frac{2\ ^j B_j}{_j n^{\ j\ \ 1} jtj^{\ j}} + \exp\left(\frac{t^2}{8\ ^2}\right)^m dt \\
&\quad + \frac{m}{2}\left(\frac{M_j}{^2 n}\right)^2 \int_1^1 \frac{1}{4}\wedge \frac{2\ ^j B_j}{_j n^{\ j\ \ 1} jtj^{\ j}} + \exp\left(\frac{t^2}{8\ ^2}\right)^{m\ \ 1} dt \\
&\quad + m^2\left(\frac{M_j}{^2 n}\right)^2 \sup_{t2R} jtj\ \frac{1}{4}\wedge \frac{2\ ^j B_j}{_j n^{\ j\ \ 1} jtj^{\ j}} + \frac{M_j}{nt}\ 1 + \frac{B_j}{n^{\ j\ \ 1} jtj^{\ j}} + \exp\left(\frac{t^2}{8\ ^2}\right)^{m\ \ 2}\ :
\end{aligned}
$$

The inequality $(a+b)^k \le 2^{k-1}(a^k+b^k)$, which holds for any $k>1$ and positive a and b, implies that

$$\int_1^{+\infty} t\,F(t)^m\left(1-F(t)\right)^m F'(t)\,dt \int_1^{+\infty} t\,G(t)^m\left(1-G(t)\right)^m G'(t)\,dt$$

$$\le \frac{M_j}{2n}\int_1^{+\infty}\left(\frac{1}{4}\wedge\frac{2^{j+1}B_j}{_jn^{j-1}|t|^j}\right)^m dt + \frac{M_j}{2n}\int_1^{+\infty}\left(\frac{1}{4}\wedge 2\exp\left(-\frac{t^2}{8\sigma^2}\right)\right)^m dt$$

$$+ \frac{m}{2}\left(\frac{M_j}{2n}\right)^2\int_1^{+\infty}\left(\frac{1}{4}\wedge\frac{2^{j+1}B_j}{_jn^{j-1}|t|^j}\right)^{m-1} dt + \frac{m}{2}\left(\frac{M_j}{2n}\right)^2\int_1^{+\infty}\left(\frac{1}{4}\wedge 2\exp\left(-\frac{t^2}{8\sigma^2}\right)\right)^{m-1} dt$$

$$+ m^2\left(\frac{M_j}{2n}\right)^2\left(\sup_{t\in\mathbb{R}}|t|^j\left(\frac{1}{4}\wedge\frac{2^{j+1}B_j}{_jn^{j-1}|t|^j}\right)^{m-2}\right) + m^2\left(\frac{M_j}{2n}\right)^2\left(\sup_{t\in\mathbb{R}}|t|^j\left(\frac{1}{4}\wedge\frac{4M_j}{n|t|^j}\right)^{m-2}\right)$$

$$+ m^2\left(\frac{M_j}{2n}\right)^2\left(\sup_{t\in\mathbb{R}}|t|^j\left(\frac{1}{4}\wedge\frac{8M_jB_j}{n^j|t|^{j+1}}\right)^{m-2}\right) + m^2\left(\frac{M_j}{2n}\right)^2\left(\sup_{t\in\mathbb{R}}|t|^j\left(\frac{1}{4}\wedge 8\exp\left(-\frac{t^2}{8\sigma^2}\right)\right)^{m-2}\right).$$

Due to Proposition A.3, it holds that

$$\int_1^{+\infty}\left(\frac{1}{4}\wedge\frac{2^{j+1}B_j}{n^{j-1}|t|^j}\right)^m dt + \int_1^{+\infty}\left(\frac{1}{4}\wedge 2\exp\left(-\frac{t^2}{8\sigma^2}\right)\right)^m dt \le 4^m\left(1+\frac{2^{j+1}B_j}{_jn^{j-1}}\right)^{1/j}.$$

Since

$$\left(\sup_{t\in\mathbb{R}}|t|^j\left(\frac{1}{4}\wedge\frac{2^{j+1}B_j}{_jn^{j-1}|t|^j}\right)^{m-2}\right) + \left(\sup_{t\in\mathbb{R}}|t|^j\left(\frac{1}{4}\wedge\frac{4M_j}{n|t|^j}\right)^{m-2}\right)$$

$$+\left(\sup_{t\in\mathbb{R}}|t|^j\left(\frac{1}{4}\wedge\frac{8M_jB_j}{n^j|t|^{j+1}}\right)^{m-2}\right) + \left(\sup_{t\in\mathbb{R}}|t|^j\left(\frac{1}{4}\wedge 4\exp\left(-\frac{t^2}{8\sigma^2}\right)\right)^{m-2}\right)$$

$$\le 4^m\left(1+\frac{M_j}{n}+\frac{M_jB_j}{n^j}\right)^{1/(j+1)} + \left(\frac{2^{j+1}B_j}{_jn^{j-1}}\right)^{1/j};$$

we conclude that

$$\int_1^{+\infty} t\,F(t)^m\left(1-F(t)\right)^m F'(t)\,dt \int_1^{+\infty} t\,G(t)^m\left(1-G(t)\right)^m G'(t)\,dt$$

$$\le 4^m\left(\frac{M_j}{2n}\right)\left(1+\frac{2^{j+1}B_j}{_jn^{j-1}}\right)^{1/j}$$

$$+ m^2\,4^m\left(\frac{M_j}{2n}\right)^2\left(1+\frac{M_j}{n}+\frac{M_jB_j}{n^j}\right)^{1/(j+1)} + \left(\frac{2^{j+1}B_j}{_jn^{j-1}}\right)^{1/j}.$$

Then it holds that

$$|\mathbb{E}\,\mathrm{SMoM}_{m,n}(\ell_j;\cdot)| \le (2m+1)\binom{2m}{m}4^m\left(\frac{M_j}{2n}\right)\left(1+\frac{2^{j+1}B_j}{_jn^{j-1}}\right)^{1/j}$$

$$+ (2m+1)\binom{2m}{m}m^2 4^m\left(\frac{M_j}{2n}\right)^2\left(1+\frac{M_j}{n}+\frac{M_jB_j}{n^j}\right)^{1/(j+1)} + \left(\frac{2^{j+1}B_j}{_jn^{j-1}}\right)^{1/j}$$

$$\le \frac{mM_j}{2n}\left(1+\frac{2^jB_j}{n^{j-1}}\right)^{1/j}.$$

The proof is finished.

A.4 Technical results

Lemma A.1. Let $h : \mathbb{R} \to \mathbb{R}$ be a twice differentiable function with uniformly bounded second derivative:

$$|h''(x)| \le H \quad \text{for all } x \in \mathbb{R}.$$

Let $q(x)$ be a probability density of a random variable and let

$$\chi(x) = \frac{q(x) + q(-x)}{2} \quad \text{and} \quad \psi(x) = \frac{q(x) - q(-x)}{2}$$

stand for its symmetric and antisymmetric parts, respectively. Assume that there exists $M > 0$ such that the function $\psi(x)$ fulfils

$$\int_{-\infty}^{\infty} x\,\psi(x)dx = 0 \quad \text{and} \quad \int_{-\infty}^{\infty} x^2|\psi(x)|dx \le M.$$

Then, for any positive integer n, it holds that

$$\int h\left(\frac{x}{n}\right) q^{*n}(x)dx - \int h\left(\frac{x}{n}\right) \chi^{*n}(x)dx \le \frac{HM}{2n},$$

provided that the integrals in the left-hand side converge.

Proof. Let $X_1, \ldots, X_n$ be i.i.d. random variables with the density $q(x)$. It is known that $(X_1 + \ldots + X_n) \sim q^{*n}(x)$. Then the integral

$$\int h\left(\frac{x}{n}\right) q^{*n}(x)dx$$

admits a representation

$$\int h\left(\frac{x}{n}\right) q^{*n}(x)dx = \mathbb{E}h\left(\frac{X_1 + \ldots + X_n}{n}\right) = \int h\left(\frac{x_1 + \ldots + x_n}{n}\right) q(x_1)\ldots q(x_n)dx_1\ldots dx_n.$$

Similarly, it holds that

$$\int h\left(\frac{x}{n}\right) \chi^{*n}(x)dx = \int h\left(\frac{x_1 + \ldots + x_n}{n}\right) \chi(x_1)\ldots\chi(x_n)dx_1\ldots dx_n,$$

and thus,

$$\int h\left(\frac{x}{n}\right) q^{*n}(x)dx - \int h\left(\frac{x}{n}\right) \chi^{*n}(x)dx = \int h\left(\frac{x_1 + \ldots + x_n}{n}\right) q(x_1)\ldots q(x_n)dx_1\ldots dx_n - \int h\left(\frac{x_1 + \ldots + x_n}{n}\right) \chi(x_1)\ldots\chi(x_n)dx_1\ldots dx_n.$$

Let us introduce

$$\rho_k(x_1, \ldots, x_n) = \prod_{i=1}^{k} q(x_i) \prod_{i=k+1}^{n} \chi(x_i), \quad \text{where } k \in \{0, \ldots, n\}.$$

Then it holds that

$$\begin{aligned}
&\int h\left(\frac{x_1 + \ldots + x_n}{n}\right) q(x_1)\ldots p(x_n)dx_1\ldots dx_n \\
&\quad - \int h\left(\frac{x_1 + \ldots + x_n}{n}\right) \chi(x_1)\ldots\chi(x_n)dx_1\ldots dx_n \\
&= \int h\left(\frac{x_1 + \ldots + x_n}{n}\right) \rho_n(x_1, \ldots, x_n)dx_1\ldots dx_n \\
&\quad - \int h\left(\frac{x_1 + \ldots + x_n}{n}\right) \rho_0(x_1, \ldots, x_n)dx_1\ldots dx_n \\
&= \sum_{k=1}^{n} \int h\left(\frac{x_1 + \ldots + x_n}{n}\right) \left(\rho_k(x_1, \ldots, x_n) - \rho_{k-1}(x_1, \ldots, x_n)\right) dx_1\ldots dx_n.
\end{aligned}$$

Let us fix any $k \in \{1, \dots, n\}$ and consider

$$\int h\left(\frac{x_1 + \dots + x_n}{n}\right) \varphi_k(x_1, \dots, x_n) \psi_{k-1}(x_1, \dots, x_n)\, dx_1 \dots dx_n.$$

Note that, according to the definition of φ_k, we have

$$\varphi_k(x_1, \dots, x_n) \psi_{k-1}(x_1, \dots, x_n) = \left(\prod_{i=1}^{k-1} q(x_i)\right) \left(\prod_{i=k+1}^{n} \psi(x_i)\right) \psi(x_k).$$

Moreover, due to the Taylor's expansion with the Lagrange remainder term, it holds that

$$h\left(\frac{x_1 + \dots + x_n}{n}\right) \le h\left(\frac{1}{n}\sum_{i \ne k} x_i\right) + h'\left(\frac{1}{n}\sum_{i \ne k} x_i\right)\frac{x_k}{n} + \frac{H x_k^2}{2n^2}.$$

Since, according to the definition of ψ and the conditions of the lemma,

$$\int_{-1}^{1} \psi(x_k) dx_k = 0; \quad \int_{-1}^{1} x_k \psi(x_k) dx_k = 0; \quad \text{and} \quad \int_{-1}^{1} x_k^2 |\psi(x_k)| dx_k \le M^2;$$

we have

$$\int h\left(\frac{1}{n}\sum_{i \ne k} x_i\right) \varphi_k(x_1, \dots, x_n) \psi_{k-1}(x_1, \dots, x_n)\, dx_1 \dots dx_n = 0;$$

$$\int h'\left(\frac{1}{n}\sum_{i \ne k} x_i\right)\frac{x_k}{n} \varphi_k(x_1, \dots, x_n) \psi_{k-1}(x_1, \dots, x_n)\, dx_1 \dots dx_n = 0;$$

and then

$$\int h\left(\frac{x_1 + \dots + x_n}{n}\right) \varphi_k(x_1, \dots, x_n) \psi_{k-1}(x_1, \dots, x_n)\, dx_1 \dots dx_n$$
$$\le \frac{H}{2n^2} \int x_k^2 \left(\prod_{i=1}^{k-1} q(x_i)\right) \left(\prod_{i=k+1}^{n} \psi(x_i)\right) |\psi(x_k)|\, dx_1 \dots dx_n \le \frac{HM}{2n^2}.$$

Finally, applying the triangle inequality, we obtain that

$$\int h\left(\frac{x_1 + \dots + x_n}{n}\right) q(x_1) \dots q(x_n) dx_1 \dots dx_n$$
$$\int h\left(\frac{x_1 + \dots + x_n}{n}\right) \psi(x_1) \dots \psi(x_n) dx_1 \dots dx_n$$
$$\le \sum_{k=1}^{n} \int h\left(\frac{x_1 + \dots + x_n}{n}\right) \varphi_k(x_1, \dots, x_n) \psi_{k-1}(x_1, \dots, x_n)\, dx_1 \dots dx_n$$
$$\le \sum_{k=1}^{n} \frac{HM}{2n^2} = \frac{HM}{2n}.$$

Lemma A.2. Let F and G be any differentiable cumulative distribution functions, such that

$$|t|G(t)^m (1 - G(t))^m F(t)(1 - G(t)) \to 0 \quad \text{and} \quad |t|G(t)^{m-1}(1 - G(t))^{m-1} F(t)(1 - G(t)) \to 0 \quad \text{as } t \to -\infty.$$

Then it holds that

$$\int_{-1}^{1} t\,F(t)^m - \big|F(t)\big|^m F'(t)\,dt \le \int_{-1}^{1} t\,G(t)^m - \big|G(t)\big|^m G'(t)\,dt$$
$$+ 6\int_{-1}^{1} \big|G(t)^m - \big|G(t)\big|^m\big|\,|F(t) - G(t)|\,dt + \frac{m}{2}\int_{-1}^{1} \big|G(t)^{m-1} - \big|G(t)\big|^{m-1}\big|\,|F(t) - G(t)|^2\,dt$$
$$+ m^2 \sup_{t\in\mathbb{R}} |t|\,|F(t) - G(t)|^2 \max_{\in[F(t)\wedge G(t);F(t)\vee G(t)]} |-|^{m-2}(1-)^{m-2}.$$

Proof. The proof is based on integration by parts. Let us define a function $\varphi : [0;1] \to \mathbb{R}$ as follows:

$$\varphi(x) = \frac{1}{m^2}\big(x^m(1-x)^m\big)'' = \frac{1}{m}x^{m-2}(1-x)^{m-2}\big((m-1)(1-2x)^2 - 2x(1-x)\big).$$

Note that, for any $x \in [0;1]$, we have

$$\frac{1}{2m}x^{m-2}(1-x)^{m-2} \le \varphi(x) \le \Big(1 - \frac{1}{m}\Big)x^{m-2}(1-x)^{m-2}. \tag{28}$$

Due to Taylor's expansion with the Lagrange remainder term, for any $t \in \mathbb{R}$, there exists $(t) \in [F(t)\wedge G(t); F(t)\vee G(t)]$, such that

$$F(t)^m - \big|F(t)\big|^m - G(t)^m - \big|G(t)\big|^m = mG(t)^{m-1} - \big|G(t)\big|^{m-1}\big|\,2G(t) - F(t) - G(t)\big|$$
$$+ \frac{m^2}{2}\varphi(-(t))\,|F(t) - G(t)|^2.$$

Then it holds that

$$\int_{-1}^{1} t\,F(t)^m - \big|F(t)\big|^m F'(t)\,dt \le \int_{-1}^{1} t\,G(t)^m - \big|G(t)\big|^m G'(t)\,dt$$
$$+ \int_{-1}^{1} t\,G(t)^m - \big|G(t)\big|^m\big|\,|F'(t) - G'(t)|\,dt$$
$$+ m\int_{-1}^{1} t\,G(t)^{m-1} - \big|G(t)\big|^{m-1}\big|\,|2G(t) - F(t) - G(t)|\,|F'(t) - G'(t)|\,dt \tag{29}$$
$$+ \frac{m^2}{2}\int_{-1}^{1} t\,\varphi(-(t))\,|F(t) - G(t)|^2 F'(t)\,dt.$$

Let us focus on the first term in the right-hand side. Integration by parts yields that

$$\int_{-1}^{1} t\,G(t)^m - \big|G(t)\big|^m\big|\,|F'(t) - G'(t)|\,dt = \int_{-1}^{1} \big|G(t)^m - \big|G(t)\big|^m\big|\,|F(t) - G(t)|\,dt$$
$$m\int_{-1}^{1} t\,G(t)^{m-1} - \big|G(t)\big|^{m-1}\big|\,|2G(t) - F(t) - G(t)|\,|F(t) - G(t)|\,dt.$$

Substituting this equality into (29), we obtain that

$$\int_1^{+\infty} t\,F(t)^{m-1}\left(1-F(t)\right)^{m}F'(t)\,dt-\int_1^{+\infty} t\,G(t)^{m-1}\left(1-G(t)\right)^{m}G'(t)\,dt$$
$$=\int_1^{+\infty} G(t)^{m-1}\left(1-G(t)\right)^{m}\left(F(t)-G(t)\right)dt$$
$$+m\int_1^{+\infty} t\,G(t)^{m-1}\left(1-G(t)\right)^{m-1}\left(1-2G(t)+F(t)G(t)-F'(t)-G'(t)\right)dt \qquad (30)$$
$$+\frac{m^2}{2}\int_1^{+\infty} t\,(\quad(t))\,F(t)-G(t)\Big)^2F'(t)\,dt:$$

Now we can apply integration by parts to the second term in the right-hand side of (30):

$$m\int_1^{+\infty} t\,G(t)^{m-1}\left(1-G(t)\right)^{m-1}\left(1-2G(t)+F(t)G(t)-F'(t)-G'(t)\right)dt$$
$$=\frac{m}{2}\int_1^{+\infty} G(t)^{m-1}\left(1-G(t)\right)^{m-1}\left(1-2G(t)+F(t)G(t)\right)^2dt$$
$$\frac{m^2}{2}\int_1^{+\infty} t\,\left(G(t)-F(t)G(t)\right)^2G'(t)\,dt:$$

Hence, we can rewrite the equality (30) in the following form:

$$\int_1^{+\infty} t\,F(t)^{m-1}\left(1-F(t)\right)^{m}F'(t)\,dt-\int_1^{+\infty} t\,G(t)^{m-1}\left(1-G(t)\right)^{m}G'(t)\,dt$$
$$=\int_1^{+\infty} G(t)^{m-1}\left(1-G(t)\right)^{m}\left(F(t)-G(t)\right)dt$$
$$\frac{m}{2}\int_1^{+\infty} G(t)^{m-1}\left(1-G(t)\right)^{m-1}\left(1-2G(t)+F(t)G(t)\right)^2dt$$
$$+\frac{m^2}{2}\int_1^{+\infty} t\,\left(\quad(t)-F'(t)\right)\left(G(t)-G'(t)\right)\left(F(t)-G(t)\right)^2dt:$$

Then, due to (28) and the triangle inequality, it holds that

$$\left|\int_1^{+\infty} t\,F(t)^{m-1}\left(1-F(t)\right)^{m}F'(t)\,dt-\int_1^{+\infty} t\,G(t)^{m-1}\left(1-G(t)\right)^{m}G'(t)\,dt\right|$$
$$\le 6\int_1^{+\infty} G(t)^{m-1}\left(1-G(t)\right)^{m}\left|F(t)-G(t)\right|dt+\frac{m}{2}\int_1^{+\infty} G(t)^{m-1}\left(1-G(t)\right)^{m-1}\left(1-2G(t)+F(t)G(t)\right)^2dt$$
$$+\frac{m^2}{2}\int_1^{+\infty} jtj\,\left|(t)-F(t)\right|\left(G(t)-F(t)G(t)\right)^2F'(t)\,dt+\frac{m^2}{2}\int_1^{+\infty} jtj\,\left|G(t)-F(t)\right|\left(G(t)-F(t)G(t)\right)^2G'(t)\,dt:$$

Since $1 - 2G(t) \le 1$,

$$\int_{-\infty}^{+\infty} |t| \left| (t) - F(t) - G(t) \right|^2 F'(t)\, dt \le \sup_{t \in \mathbb{R}} \left\{ |t| \left| (t) - F(t) - G(t) \right|^2 \right\};$$

and, similarly,

$$\int_{-\infty}^{+\infty} |t| \left| G(t) - F(t) - G(t) \right|^2 G'(t)\, dt \le \sup_{t \in \mathbb{R}} \left\{ |t| \left| G(t) - F(t) - G(t) \right|^2 \right\};$$

we finally obtain that

$$\begin{aligned}
&\int_{-\infty}^{+\infty} t\, F(t)^m \left(1 - F(t)\right)^m F'(t)\, dt - \int_{-\infty}^{+\infty} t\, G(t)^m \left(1 - G(t)\right)^m G'(t)\, dt \\
&\le \int_{-\infty}^{+\infty} G(t)^m \left(1 - G(t)\right)^m \left(F(t) - G(t)\right) dt + \frac{m}{2} \int_{-\infty}^{+\infty} G(t)^{m-1} \left(1 - G(t)\right)^{m-1} \left(F(t) - G(t)\right)^2 dt \\
&\quad + m^2 \sup_{t \in \mathbb{R}} \left\{ |t| \left| F(t) - G(t) \right|^2 \right\} \max_{\in [F(t) \wedge G(t);\, F(t) \vee G(t)]} {}^{m-2} (1 -)^{m-2} :
\end{aligned}$$

Proposition A.3. For any positive numbers $a, k, $, and , such that $k > + 1$, it holds that

$$\int_{-\infty}^{+\infty} |t| \left(\frac{1}{4} \wedge \frac{a}{|t|} \right)^k dt = \frac{k}{(+ 1)(k - 1)} \cdot \frac{2(4^{1/} a)^{+1}}{4^k}.$$

Proof. The proof follows from simple calculations:

$$\begin{aligned}
\int_{-\infty}^{+\infty} |t| \left(\frac{1}{4} \wedge \frac{a}{|t|} \right)^k dt &= 2 \int_0^{+\infty} t \left(\frac{1}{4} \wedge \frac{a}{t} \right)^k dt \\
&= \frac{2}{4^k} \int_0^{4^{1/} a} t\, dt + 2 a^k \int_{4^{1/} a}^{+\infty} t^{-k}\, dt \\
&= \frac{2(4^{1/} a)^{+1}}{(+ 1)\, 4^k} + \frac{2 a^k \left(4^{1/} a\right)^{k+1}}{(k - 1)} \\
&= \frac{2(4^{1/} a)^{+1}}{4^k} \left(\frac{1}{+ 1} + \frac{1}{k - 1} \right) \\
&= \frac{k}{(+ 1)(k - 1)} \cdot \frac{2(4^{1/} a)^{+1}}{4^k}.
\end{aligned}$$

Lemma A.4. Grant Assumption 2.1 and let

$$F(t) = \int_{-\infty}^{+\infty} t \left(\frac{u}{n} \right) p_j^{\,n}(u)\, du = \int_{\mathbb{R}^n} t \left(\text{Mean}(u_1, \ldots, u_n) \right) p_j(u_1) \ldots p_j(u_n)\, du$$

and

$$G(t) = \int_{-\infty}^{+\infty} t \left(\frac{u}{n} \right) s_j^{\,n}(u)\, du = \int_{\mathbb{R}^n} t \left(\text{Mean}(u_1, \ldots, u_n) \right) s_j(u_1) \ldots s_j(u_n)\, du$$

be two cumulative distribution functions. Assume that $2M_j \le n^2$. Then, for any $t \in \mathbb{R}$ it holds that

$$|F(t) - G(t)| \cdot \frac{M_j}{n^{j-1}|t|^j}\left(1 + |t| + \frac{B_j}{j}\right).$$

Proof. Since $p_j(u) = s_j(u) + r_j(u)$, it holds that

$$p_j^n = \left(s_j(u) + r_j(u)\right)^n = \sum_{k=0}^{n} \binom{n}{k} r_j^k \, s_j^{(n-k)};$$

and then

$$\begin{aligned} F(t) - G(t) &= \int_1^{+\infty} \cos\left(t\,\frac{u}{n}\right) p_j^n(u)\,du - \int_1^{+\infty} \cos\left(t\,\frac{u}{n}\right) s_j^n(u)\,du \\ &= \sum_{k=1}^{n} \binom{n}{k} \int_1^{+\infty} \cos\left(t\,\frac{u}{n}\right) r_j^k \, s_j^{(n-k)}(u)\,du \qquad (31) \\ &= \sum_{k=1}^{n} \binom{n}{k} \int_{\mathbb{R}^n} \cos\left(t\,\mathrm{Mean}(u_1,\ldots,u_n)\right) \left(\prod_{i=1}^{k} r_j(u_i)\right) \left(\prod_{i=k+1}^{n} s_j(u_i)\right) du_1 \ldots du_n. \end{aligned}$$

In the rest of the proof, we bound the summands in the right-hand side one by one. For readability, we split our derivations into several steps.

Step 1: Taylor's expansion. Let us fix an arbitrary $k \in \{1,\ldots,n\}$ and consider

$$\int_{\mathbb{R}^n} \cos\left(t\,\mathrm{Mean}(u_1,\ldots,u_n)\right) \left(\prod_{i=1}^{k} r_j(u_i)\right) \left(\prod_{i=k+1}^{n} s_j(u_i)\right) du_1 \ldots du_n.$$

Using Taylor's expansion with the integral remainder term, we rewrite the expression of interest in the form

$$\begin{aligned} &\int_{\mathbb{R}^n} \cos\left(t\,\mathrm{Mean}(u_1,\ldots,u_n)\right) \left(\prod_{i=1}^{k} r_j(u_i)\right) \left(\prod_{i=k+1}^{n} s_j(u_i)\right) du_1 \ldots du_n \\ &= \int_{\mathbb{R}^n} \left[\cos\left(t\,\mathrm{Mean}(u_1,\ldots,u_n) + \frac{u_1}{n}\right) - \cos\left(t\,\mathrm{Mean}(u_1,\ldots,u_n) + \frac{u_1}{n}\right) - \frac{u_1}{n}\right] \qquad (32) \\ &\quad + \frac{u_1^2}{n^2} \int_0^1 \cos''\left(t\,\mathrm{Mean}(u_1,\ldots,u_n) + \frac{(1-v_1)u_1}{n}\right) v_1\,dv_1 \left(\prod_{i=1}^{k} r_j(u_i)\right) \left(\prod_{i=k+1}^{n} s_j(u_i)\right) du_1 \ldots du_n. \end{aligned}$$

Note that

$$\cos\left(t\,\mathrm{Mean}(u_1,\ldots,u_n) + \frac{u_1}{n}\right) \quad \text{and} \quad \cos\left(t\,\mathrm{Mean}(u_1,\ldots,u_n) + \frac{u_1}{n}\right)$$

do not depend on u_1. Since, according to Assumption 2.1, it holds that

$$\int_1^{+\infty} r_j(u_1)\,du_1 = 0 \quad \text{and} \quad \int_1^{+\infty} u_1 r_j(u_1)\,du_1 = 0;$$

the right-hand side of (32) simplifies to

$$\begin{aligned} &\int_{\mathbb{R}^n} \cos\left(t\,\mathrm{Mean}(u_1,\ldots,u_n)\right) \left(\prod_{i=1}^{k} r_j(u_i)\right) \left(\prod_{i=k+1}^{n} s_j(u_i)\right) du_1 \ldots du_n \\ &= \int_0^1 v_1\,dv_1 \int_{\mathbb{R}^n} du_1 \ldots du_n \cos''\left(t\,\mathrm{Mean}(u_1,\ldots,u_n) + \frac{(1-v_1)u_1}{n}\right) \frac{u_1^2}{n^2} \left(\prod_{i=1}^{k} r_j(u_i)\right) \left(\prod_{i=k+1}^{n} s_j(u_i)\right). \end{aligned}$$

Repeating this trick $(k-1)$ more times, we obtain that

$$\int_{\mathbb{R}^n} t\,\mathrm{Mean}(u_1,\dots,u_n)\prod_{i=1}^{k} r_j(u_i)\prod_{i=k+1}^{n} s_j(u_i)\,du_1\dots du_n$$
$$= \int_0^1 v_1 dv_1 \dots \int_0^1 v_k dv_k \int_{\mathbb{R}^n} du_1\dots du_n$$
$$\cdot\, t^{(2k)}\,\mathrm{Mean}(u_1,\dots,u_n) + \sum_{i=1}^{k}\frac{(1-v_i)u_i}{n}\prod_{i=1}^{k}\frac{u_i^2 r_j(u_i)}{n^2}\prod_{i=k+1}^{n} s_j(u_i);$$

where $t^{(2k)}$ stands for the $(2k)$-th derivative of t. Due to the properties of convolution, the integral in the right-hand side is equal to

$$\int_0^1 v_1 dv_1 \dots \int_0^1 v_k dv_k \int_{\mathbb{R}^n} du_1\dots du_n \; t^{(2k)}\left(\sum_{i=1}^{k}\frac{v_i u_i}{n}+\sum_{i=k+1}^{n}\frac{u_i}{n}\right)\prod_{i=1}^{k}\frac{u_i^2 r_j(u_i)}{n^2}\, s_j^{(n-k)}\left(\sum_{i=k+1}^{n} u_i\right):$$

Substituting $u_{k+1}+\dots+u_n$ by y, we conclude that

$$\int_{\mathbb{R}^n} t\,\mathrm{Mean}(u_1,\dots,u_n)\prod_{i=1}^{k} r_j(u_i)\prod_{i=k+1}^{n} s_j(u_i)\,du_1\dots du_n$$
$$= \int_0^1 v_1 dv_1 \dots \int_0^1 v_k dv_k \int_{\mathbb{R}^k} du_1\dots du_k \int_1^{+\infty} dy \; t^{(2k)}\left(\sum_{i=1}^{k}\frac{v_i u_i}{n}+\frac{y}{n}\right)\prod_{i=1}^{k}\frac{u_i^2 r_j(u_i)}{n^2}\, s_j^{(n-k)}(y). \qquad (33)$$

Step 2: bound on the integral (33). We represent the expression (33) as a sum of two terms:

$$\int_0^1 v_1 dv_1 \dots \int_0^1 v_k dv_k \int_{\mathbb{R}^k} du_1\dots du_k \int_1^{+\infty} dy \; t^{(2k)}\left(\sum_{i=1}^{k}\frac{v_i u_i}{n}+\frac{y}{n}\right)\prod_{i=1}^{k}\frac{u_i^2 r_j(u_i)}{n^2}\, s_j^{(n-k)}(y)$$
$$= \int_0^1 v_1 dv_1 \dots \int_0^1 v_k dv_k \int_{[-\frac{nt}{2k};\frac{nt}{2k}]^k} du_1\dots du_k \int_1^{+\infty} dy \; t^{(2k)}\left(\sum_{i=1}^{k}\frac{v_i u_i}{n}+\frac{y}{n}\right)\prod_{i=1}^{k}\frac{u_i^2 r_j(u_i)}{n^2}\, s_j^{(n-k)}(y)$$
$$+ \int_0^1 v_1 dv_1 \dots \int_0^1 v_k dv_k \int_{\mathbb{R}^k\setminus[-\frac{nt}{2k};\frac{nt}{2k}]^k} du_1\dots du_k \int_1^{+\infty} dy \; t^{(2k)}\left(\sum_{i=1}^{k}\frac{v_i u_i}{n}+\frac{y}{n}\right)\prod_{i=1}^{k}\frac{u_i^2 r_j(u_i)}{n^2}\, s_j^{(n-k)}(y);$$

and bound the former and the latter summands in the right-hand side separately. According to Lemma A.5 and A.6, it holds that

$$\int_0^1 v_1 dv_1 \dots \int_0^1 v_k dv_k \int_{[-\frac{nt}{2k};\frac{nt}{2k}]^k} du_1\dots du_k \int_1^{+\infty} dy \; t^{(2k)}\left(\sum_{i=1}^{k}\frac{v_i u_i}{n}+\frac{y}{n}\right)\prod_{i=1}^{k}\frac{u_i^2 r_j(u_i)}{n^2}\, s_j^{(n-k)}(y)$$
$$\le 6\,\frac{\sqrt{(2k-1)!}}{\sqrt{2}}\left(\frac{2^{3+2\beta_j}B_j}{n^{\beta_j-1}|t|^{1+\beta_j}}+\exp\left(-\frac{t^2}{64\,\sigma^2}\right)\right)\left(\frac{M_j}{2\,\sigma^2 n^2}\right)^k:$$

and

$$\int_0^1 v_1 dv_1 \dots \int_0^1 v_k dv_k \int_{\mathbb{R}^k / [\frac{nt}{2k}; \frac{nt}{2k}]^k} du_1 \dots du_k \int_1^1 dy \; {}^{(2k)} t \left(\sum_{i=1}^k \frac{v_i u_i}{n} - \frac{y}{n} \right)^! \left(\sum_{i=1}^k \frac{u_i^2 r_j(u_i)}{n^2} \right)^! s_j^{(n-k)}(y)$$
$$\leq \sqrt{(2k-2)!} \left(\frac{M_j}{2^2 n^2} \right)^k \left(\frac{4k^2}{t^2} + \frac{1}{\sqrt{2}} \frac{2k}{t} \right).$$

Hence, we obtain that

$$\begin{aligned} &\int_0^1 v_1 dv_1 \dots \int_0^1 v_k dv_k \int_{\mathbb{R}^k} du_1 \dots du_k \int_1^1 dy \; {}^{(2k)} t \left(\sum_{i=1}^k \frac{v_i u_i}{n} - \frac{y}{n} \right)^! \left(\sum_{i=1}^k \frac{u_i^2 r_j(u_i)}{n^2} \right)^! s_j^{(n-k)}(y) \\ &\leq \frac{\sqrt{(2k-1)!}}{\sqrt{2}} \left(\frac{2^{3+2j} B_j}{n^j - 1} t j^{1+j} + \exp\left(\frac{t^2}{64^2} \right) \left(\frac{M_j}{2^2 n^2} \right)^k \right) \\ &\quad + \sqrt{(2k-2)!} \left(\frac{M_j}{2^2 n^2} \right)^k \left(\frac{4k^2}{t^2} + \frac{1}{\sqrt{2}} \frac{2k}{t} \right). \end{aligned} \tag{34}$$

Step 3: final bound. Summing up the equalities (31), (33), and the equality (34), we obtain that

$$\begin{aligned} F(t) - G(t) &\leq \sum_{k=1}^n \binom{n}{k} \left(\frac{\sqrt{(2k-1)!}}{\sqrt{2}} \left(\frac{2^{3+2j} B_j}{n^j - 1} t j^{1+j} + \exp\left(\frac{t^2}{64^2} \right) \left(\frac{M_j}{2^2 n^2} \right)^k \right) \right. \\ &\quad + \sum_{k=1}^n \binom{n}{k} \sqrt{(2k-2)!} \left(\frac{4k^2}{t^2} + \frac{1}{\sqrt{2}} \frac{2k}{t} \right) \left(\frac{M_j}{2^2 n^2} \right)^k \\ &\quad \left. \cdot \sum_{k=1}^n \binom{n}{k} \sqrt{(2k)!} \left(\frac{B_j}{n^j - 1} t j^{1+j} + \exp\left(\frac{t^2}{64^2} + \frac{k^2}{t^2} + \frac{k}{t} \right) \left(\frac{M_j}{2^2 n^2} \right)^k \right). \end{aligned}$$

Finally, Lemma A.7 implies that

$$\begin{aligned} F(t) - G(t) &\leq \sum_{k=1}^n \binom{n}{k} \sqrt{(2k)!} \left(\frac{B_j}{n^j - 1} t j^{1+j} + \exp\left(\frac{t^2}{64^2} + \frac{k^2}{t^2} + \frac{k}{t} \right) \left(\frac{M_j}{2^2 n^2} \right)^k \right. \\ &\leq \left(\frac{B_j}{n^j - 1} t j^{1+j} + \exp\left(\frac{t^2}{64^2} + \frac{2^2}{t^2} + \frac{2}{t} \right) \frac{2M_j}{2n} \right) \\ &\quad \cdot \left(\frac{M_j}{nt} \left(1 + \frac{2}{t} + \frac{B_j}{n^j - 1} t j^{1+j} \right) \right); \end{aligned}$$

whenever $2M_j \leq n^2$. The proof is finished.

Lemma A.5. Under Assumption 2.1, it holds that

$$\begin{aligned} &\int_0^1 v_1 dv_1 \dots \int_0^1 v_k dv_k \int_{[\frac{nt}{2k}; \frac{nt}{2k}]^k} du_1 \dots du_k \int_1^1 dy \; {}^{(2k)} t \left(\sum_{i=1}^k \frac{v_i u_i}{n} - \frac{y}{n} \right)^! \left(\sum_{i=1}^k \frac{u_i^2 r_j(u_i)}{n^2} \right)^! s_j^{(n-k)}(y) \\ &\leq \frac{\sqrt{(2k-1)!}}{\sqrt{2}} \left(\frac{2^{3+2j} B_j}{n^j - 1} t j^{1+j} + \exp\left(\frac{t^2}{64^2} \right) \left(\frac{M_j}{2^2 n^2} \right)^k \right). \end{aligned}$$

Proof. To prove Lemma A.5, it is enough to show that

$$\int_0^1 v_1 dv_1 \int_0^1 \ldots \int_0^1 v_k dv_k \int_{[\frac{nt}{2k};\frac{nt}{2k}]^k} du_1 \ldots du_k \int_1^1 dy \; \varphi^{(2k)}\left(t\left(\sum_{i=1}^k \frac{v_i u_i}{n} - \frac{y}{n}\right)\left(\sum_{i=1}^k \frac{u_i^2 r_j(u_i)}{n^2}\right)\right) s_j^{(n\ k)}(y)$$
$$\leq \frac{\sqrt{(2k-1)!}}{\sqrt{2}} \left(\frac{2^{3+2\ j} B_j}{n^{\ j\ 1} |t|^{1+\ j}} + \exp\left(-\frac{t^2}{64\ ^2}\right)\right) \left(\frac{M_j}{2\ ^2 n^2}\right)^k.$$

Our argument is quite technical, so we divide the proof into several parts.

Step 1: a bound on the 2k-th derivative. First, let us consider the 2k-th derivative of φ. Note that

$$\varphi^{(2k)}(w) = \frac{1}{2k\ 1}\varphi^{(2k)}\left(\frac{w}{1}\right) = \frac{1}{2k\sqrt{2}} H_{2k\ 1}(w/\sqrt{2})\exp\left(-\frac{w^2}{2\ ^2}\right),$$

where $H_{2k\ 1}$ is the $(2k-1)$-th probabilist's Hermite polynomial. We provide a brief information about Hermite polynomials in Appendix D. In particular, we refer to the result of Indritz (1961), which implies that

$$\max_{w \in R} |H_{2k\ 1}(w/\sqrt{2})\exp\left(-\frac{w^2}{4\ ^2}\right)| \leq \sqrt{(2k-1)!} \quad \text{for all } k \in N.$$

Thus, it holds that

$$|\varphi^{(2k)}(w)| \leq \frac{\sqrt{(2k-1)!}}{2k\sqrt{2}} \exp\left(-\frac{w^2}{4\ ^2}\right) \quad \text{for all } w \in R \text{ and all } k, \tag{35}$$

and we obtain the inequality

$$\int_0^1 v_1 dv_1 \int_0^1 \ldots \int_0^1 v_k dv_k \int_{[\frac{nt}{2k};\frac{nt}{2k}]^k} du_1 \ldots du_k \int_1^1 dy \; \varphi^{(2k)}\left(t\left(\sum_{i=1}^k \frac{v_i u_i}{n} - \frac{y}{n}\right)\left(\sum_{i=1}^k \frac{u_i^2 r_j(u_i)}{n^2}\right)\right) s_j^{(n\ k)}(y)$$
$$\leq \frac{\sqrt{(2k-1)!}}{2k\sqrt{2}} \int_0^1 v_1 dv_1 \int_0^1 \ldots \int_0^1 v_k dv_k \int_{[\frac{nt}{2k};\frac{nt}{2k}]^k} du_1 \ldots du_k \int_1^1 dy \exp\left\{\frac{1}{4\ ^2} t\left(\sum_{i=1}^k \frac{v_i u_i}{n} - \frac{y}{n}\right)^2\right\} \left(\sum_{i=1}^k \frac{u_i^2 r_j(u_i)}{n^2}\right) s_j^{(n\ k)}(y). \tag{36}$$

Step 2: a bound on the convolution. Our next goal is to bound the convolution

$$\int_1^1 \exp\left\{\frac{1}{4\ ^2} t\left(\sum_{i=1}^k \frac{v_i u_i}{n} - \frac{y}{n}\right)^2\right\} s_j^{(n\ k)}(y) dy$$

using the properties of $s_j^{(n\ k)}$ from Assumption 2.1. Let us fix an arbitrary $w \in R$ and consider

$$\int_1^1 \exp\left(-\frac{(w - y/n)^2}{4\ ^2}\right) s_j^{(n\ k)}(y) dy.$$

Since $\exp(-(nw - y)^2/(4n^2\ ^2)) \leq 1$, it holds that

$$\int_1^1 \exp\left(-\frac{(w - y/n)^2}{4\ ^2}\right) s_j^{(n\ k)}(y) dy \leq 1. \tag{37}$$

On the other hand, we have

$$\int_{1}^{+\infty} \exp\left(-\frac{(w-y/n)^2}{4\sigma^2}\right) s_j^{(n,k)}(y)\,dy = n\int_{1}^{+\infty} \exp\left(-\frac{y^2}{4\sigma^2}\right) s_j^{(n,k)}(ny+nw)\,dy$$
$$\leq 6\sum_{w=2}^{+\infty} \exp\left(-\frac{y^2}{4\sigma^2}\right) \frac{B_j n(n-k)}{(n-k)^{(j+1)/j} + n^{1+j} j w + y j^{1+j}}\,dy$$
$$+ n\int_{|y|>w/2} \exp\left(-\frac{y^2}{4\sigma^2}\right) s_j^{(n,k)}(ny+nw)\,dy.$$

If $|y| \leq w/2$, then

$$\frac{B_j n(n-k)}{(n-k)^{(j+1)/j} + n^{1+j} j w + y j^{1+j}} \leq \frac{B_j n(n-k)}{(n-k)^{(j+1)/j} + n^{1+j} j w/2 j^{1+j}}$$

and it holds that

$$\sum_{w=2}^{+\infty} \exp\left(-\frac{y^2}{4\sigma^2}\right) \frac{B_j n(n-k)}{(n-k)^{(j+1)/j} + n^{1+j} j w + y j^{1+j}}\,dy$$
$$\leq 6\,\frac{B_j n(n-k)}{(n-k)^{(j+1)/j} + j n w/2 j^{1+j}} \int_{w=2}^{+\infty} \exp\left(-\frac{y^2}{4\sigma^2}\right) dy \tag{38}$$
$$\leq 6\,\frac{B_j n(n-k)}{(n-k)^{(j+1)/j} + j n w/2 j^{1+j}}\,\sqrt{4\pi\sigma^2} \leq 6\,\frac{2^{2+j} B_j \sqrt{\pi}}{n^{j-1} j w j^{1+j}}.$$

Otherwise,

$$n\int_{|y|>w/2} \exp\left(-\frac{y^2}{4\sigma^2}\right) s_j^{(n,k)}(ny+nw)\,dy \leq \exp\left(-\frac{w^2}{16\sigma^2}\right) \int_{1}^{+\infty} s_j^{(n,k)}(ny+nw)\,ndy$$
$$= \exp\left(-\frac{w^2}{16\sigma^2}\right). \tag{39}$$

Taking into account (37), (38), and (39), we obtain that

$$\int_{1}^{+\infty} \exp\left(-\frac{(w-y/n)^2}{4\sigma^2}\right) s_j^{(n,k)}(y)\,dy \leq 1 \wedge \left(\frac{2^{2+j} B_j \sqrt{\pi}}{n^{j-1} j w j^{1+j}} + \exp\left(-\frac{w^2}{16\sigma^2}\right)\right)$$

and, hence,

$$\int_{1}^{+\infty} \exp\left\{-\frac{1}{4\sigma^2}\left(\sum_{i=1}^{k} \frac{v_i u_i}{n} - \frac{y}{n}\right)^2\right\} s_j^{(n,k)}(y)\,dy$$
$$\leq \min\left\{1; \left(\frac{2^{2+j} B_j \sqrt{\pi}}{n^{j-1}} \sum_{i=1}^{k} \frac{v_i u_i}{n}\right)^{1/j} + \exp\left\{\frac{1}{16\sigma^2}\left(\sum_{i=1}^{k} \frac{v_i u_i}{n}\right)^2\right\}\right\}. \tag{40}$$

Step 3: final bound. The inequalities (36) and (40) yield that

$$\int_0^1 v_1 dv_1 \dots \int_0^1 v_k dv_k \int_{[\frac{nt}{2k};\frac{nt}{2k}]^k} du_1 \dots du_k \int_1^1 dy^{(2k)} t^{\sum_{i=1}^k \frac{v_i u_i}{n} - \frac{y}{n}} \prod_{i=1}^k \frac{u_i^2 r_j(u_i)}{n^2} s_j^{(n-k)}(y)$$

$$\le \frac{\sqrt{(2k-1)!}}{2^k \sqrt{2}} \int_0^1 v_1 dv_1 \dots \int_0^1 v_k dv_k \int_{[\frac{nt}{2k};\frac{nt}{2k}]^k} du_1 \dots du_k \prod_{i=1}^k \frac{u_i^2 r_j(u_i)}{n^2}$$

$$\cdot \min\left\{1; \frac{2^{2+j} B_j \sqrt{-}}{n^{j-1}} t^{\sum_{i=1}^k \frac{v_i u_i}{n}}{}^{1-j} + \exp\left\{\frac{1}{16^2} t^{\sum_{i=1}^k \frac{v_i u_i}{n}}\right\}\right\}.$$

On the set $[nt/(2k); nt/(2k)]^k$, we have

$$t^{\sum_{i=1}^k \frac{v_i u_i}{n}} > \frac{t}{2},$$

and, hence,

$$\int_0^1 v_1 dv_1 \dots \int_0^1 v_k dv_k \int_{[\frac{nt}{2k};\frac{nt}{2k}]^k} du_1 \dots du_k \int_1^1 dy^{(2k)} t^{\sum_{i=1}^k \frac{v_i u_i}{n} - \frac{y}{n}} \prod_{i=1}^k \frac{u_i^2 r_j(u_i)}{n^2} s_j^{(n-k)}(y)$$

$$\le \frac{\sqrt{(2k-1)!}}{2^k \sqrt{2}} \frac{2^{3+2j} B_j \sqrt{-}}{n^{j-1} |t|^{1+j}} + \exp\left\{\frac{t^2}{64^2}\right\} \int_0^1 v_1 dv_1 \dots \int_0^1 v_k dv_k \int_{[\frac{nt}{2k};\frac{nt}{2k}]^k} du_1 \dots du_k \prod_{i=1}^k \frac{u_i^2 r_j(u_i)}{n^2}.$$

Due to Assumption 2.1, the right-hand side does not exceed

$$\frac{\sqrt{(2k-1)!}}{2^k \sqrt{2}} \frac{2^{3+2j} B_j \sqrt{-}}{n^{j-1} |t|^{1+j}} + \exp\left\{\frac{t^2}{64^2}\right\} \frac{M_j}{n^2}{}^k \int_0^1 v_1 dv_1 \dots \int_0^1 v_k dv_k$$

$$= \frac{\sqrt{(2k-1)!}}{2^k \sqrt{2}} \frac{2^{3+2j} B_j \sqrt{-}}{n^{j-1} |t|^{1+j}} + \exp\left\{\frac{t^2}{64^2}\right\} \frac{M_j}{2^2 n^2}{}^k.$$

The proof is finished.

Lemma A.6. Let Assumption 2.1 be fulfilled. Then, for any $k > 2$, it holds that

$$\int_0^1 v_1 dv_1 \dots \int_0^1 v_k dv_k \int_{R^k \setminus [\frac{nt}{2k};\frac{nt}{2k}]^k} du \int_1^1 dy^{(2k)} t^{\sum_{i=1}^k \frac{v_i u_i}{n} - \frac{y}{n}} \prod_{i=1}^k \frac{u_i^2 r_j(u_i)}{n^2} s_j^{(n-k)}(y)$$

$$\le \sqrt{(2k-2)!} \left(\frac{M_j}{2^2 n^2}\right)^k \left(\frac{4k^2}{t^2} + \sqrt{\frac{1}{2}} \frac{2k}{t}\right)^k.$$

Proof. Let $\{A_i : 1 \le i \le k\}$ be a collection of sets in R^k, such that

$A_1, \dots, A_k$ form a partition of $R^k \setminus [\frac{nt}{2k}; \frac{nt}{2k}]^k$, that is,

$$R^k \setminus \left[\frac{nt}{2k}; \frac{nt}{2k}\right]^k = \bigcup_{i=1}^k A_i \quad \text{and} \quad A_i \cap A_\ell = \varnothing \quad \text{for all } i \ne \ell;$$

ˆ for all $i \in \{1,\dots,k\}$, it holds that

$$A_i \quad \{u : |u_i| > \frac{nt}{2k}\} \quad :$$

Let us fix any $\ell \in \{1,\dots,k\}$ and consider the integral

$$\int_0^1 v_1 dv_1 \cdots \int_0^1 v_k dv_k \int_{A_\ell} du_1 \cdots du_k \int_1^{+\infty} dy \quad (2k) \quad t \left(\sum_{i=1}^k \frac{v_i u_i}{n} \le \frac{y}{n}\right) \prod_{i=1}^k \frac{u_i^2 r_j(u_i)}{n^2} \quad s_j^{(n\ k)}(y):$$

Applying the Newton-Leibnitz formula, we obtain that

$$\frac{u_\ell^2}{n^2} \int_0^1 (2k) \quad t \left(\sum_{i=1}^k \frac{v_i u_i}{n} \le \frac{y}{n}\right) v_\ell\, dv_\ell = (2k\ \ 2)\, t \left(\sum_{\substack{1 \le i \le k;\\ i \ne \ell}} \frac{v_i u_i}{n} \le \frac{y}{n}\right)$$
$$+ (2k\ \ 1)\, t \left(\sum_{\substack{1 \le i \le k;\\ i \ne \ell}} \frac{v_i u_i}{n} \le \frac{y}{n}\right) \frac{u_\ell}{n}:$$

This implies that

$$\int_0^1 v_1 dv_1 \cdots \int_0^1 v_k dv_k \int_{A_\ell} du_1 \cdots du_k \int_1^{+\infty} dy \quad (2k) \quad t \left(\sum_{i=1}^k \frac{v_i u_i}{n} \le \frac{y}{n}\right) \prod_{i=1}^k \frac{u_i^2 r_j(u_i)}{n^2} \quad s_j^{(n\ k)}(y)$$
$$6 \int_0^1 v_1 dv_1 \cdots \int_0^1 v_k dv_k \int_{A_\ell} du_1 \cdots du_k \int_1^{+\infty} dy \quad (2k\ \ 2)\, t \left(\sum_{\substack{1 \le i \le k;\\ i \ne \ell}} \frac{v_i u_i}{n} \le \frac{y}{n}\right)$$
$$\prod_{\substack{1 \le i \le k\\ i \ne \ell}} \frac{u_i^2 r_j(u_i)}{n^2}\, r_j(u_\ell)\ \ s_j^{(n\ k)}(y)$$
$$+ \int_0^1 v_1 dv_1 \cdots \int_0^1 v_k dv_k \int_{A_\ell} du_1 \cdots du_k \int_1^{+\infty} dy \quad (2k\ \ 1)\, t \left(\sum_{\substack{1 \le i \le k;\\ i \ne \ell}} \frac{v_i u_i}{n} \le \frac{y}{n}\right)$$
$$\prod_{\substack{1 \le i \le k\\ i \ne \ell}} \frac{u_i^2 r_j(u_i)}{n^2}\, \frac{u_\ell r_j(u_\ell)}{n}\ \ s_j^{(n\ k)}(y)$$

Let us apply the inequality (35) we derived in the proof of Lemma A.5 to $^{(2k\ \ 2)}$ and $^{(2k\ \ 1)}$: for all $w \in \mathbb{R}$, it holds that

$$0 \le \ (w) \le 1; \qquad ^{(2k\ \ 2)}(w) \le \frac{\sqrt{(2k-3)!}}{2^{k-2}\sqrt{2}} \exp\left\{\frac{w^2}{4}\right\} \le \frac{\sqrt{(2k-2)!}}{2^{k-2}}; \quad \text{for all } k > 2.$$

and

$$^{(2k\ \ 1)}(w) \le \frac{\sqrt{(2k-2)!}}{2^{k-1}\sqrt{2}} \exp\left\{\frac{w^2}{4}\right\} \le \frac{\sqrt{(2k-2)!}}{2^{k-1}\sqrt{2}}:$$

Then, for any $k \in \mathbb{N}$,

$$\int_0^1 v_1 dv_1 \dots \int_0^{v_k} dv_k \int_{A_\ell} du_1 \dots du_k \int_1^\infty dy \; (2k-2)!! \left(t - \sum_{1 \le i \le k;\, i \in \ell} \frac{v_i - u_i}{n} - \frac{y}{n} \right)_+^{n-k} \left(\prod_{1 \le i \le k,\, i \in \ell} \frac{u_i^2 - r_j(u_i)}{n^2} \right) r_j(u_\ell) \, s_j^{(n-k)}(y)$$

$$= \frac{\sqrt{(2k-2)!}}{(2k-2)!!} \int_0^1 v_1 dv_1 \dots \int_0^{v_k} dv_k \int_{A_\ell} \left(\prod_{1 \le i \le k,\, i \in \ell} \frac{u_i^2 - r_j(u_i)}{n^2} \right) r_j(u_\ell) \, du_1 \dots du_k .$$

Due to Assumption 2.1,

$$\frac{\sqrt{(2k-2)!}}{(2k-2)!!} \int_0^1 v_1 dv_1 \dots \int_0^{v_k} dv_k \int_{A_\ell} \left(\prod_{1 \le i \le k,\, i \in \ell} \frac{u_i^2 - r_j(u_i)}{n^2} \right) r_j(u_\ell) \, du_1 \dots du_k$$

$$\le \sqrt{(2k-2)!} \, \frac{M_j}{2^{k-1} n^2} \int_{\|u_\ell\| \ge t\sqrt{2k}} r_j(u_\ell) \, du_\ell$$

$$\le \sqrt{(2k-2)!} \, \frac{M_j}{2^{k-1} n^2} \int_{\|u_\ell\| \ge t\sqrt{2k}} \frac{2k^2}{n^2 t^2} \, u^2 \, r_j(u_\ell) \, du_\ell$$

$$\le \sqrt{(2k-2)!} \, \frac{M_j}{2^{k-1} n^2} \, \frac{2k^2 M_j}{n^2 t^2} .$$

Similarly, it holds that

$$\int_0^1 v_1 dv_1 \dots \int_0^{v_k} dv_k \int_{A_\ell} du_1 \dots du_k \int_1^\infty dy \; (2k-1)!! \left(t - \sum_{1 \le i \le k;\, i \in \ell} \frac{v_i - u_i}{n} - \frac{y}{n} \right)_+^{n-k} \left(\prod_{1 \le i \le k,\, i \in \ell} \frac{u_i^2 - r_j(u_i)}{n^2} \right) \frac{u_\ell \, r_j(u_\ell)}{n} \, s_j^{(n-k)}(y)$$

$$= \frac{\sqrt{(2k-2)!}}{(2k-1)!! \sqrt{2}} \int_0^1 v_1 dv_1 \dots \int_0^{v_k} dv_k \int_{A_\ell} \left(\prod_{1 \le i \le k,\, i \in \ell} \frac{u_i^2 - r_j(u_i)}{n^2} \right) \frac{u_\ell \, r_j(u_\ell)}{n} \, du_1 \dots du_k$$

$$\le \frac{\sqrt{(2k-2)!}}{\sqrt{2}} \, \frac{M_j}{2^{k-1} n^2} \int_{\|u_\ell\| \ge t\sqrt{2k}} \frac{u_\ell \, r_j(u_\ell)}{n} \, du_\ell$$

$$\le \frac{\sqrt{(2k-2)!}}{\sqrt{2}} \, \frac{M_j}{2^{k-1} n^2} \int_{\|u_\ell\| \ge t\sqrt{2k}} \frac{k}{n^2 t} \, u^2 \, r_j(u_\ell) \, du_\ell$$

$$\le \frac{\sqrt{(2k-2)!}}{\sqrt{2}} \, \frac{M_j}{2^{k-1} n^2} \, \frac{k M_j}{n^2 t} .$$

Thus, we obtain that

$$\int_0^1 v_1 dv_1 \dots \int_0^1 v_k dv_k \int_A \dots \int_1 dy \quad {}^{(2k)} \quad t \sum_{i=1}^k \frac{v_i u_i}{n} \quad \frac{y}{n}^{!} \quad \prod_{i=1}^k \frac{u_i^2 r_j(u_i)}{n^2}^{!} \quad s_j^{(n\ k)}(y)$$
$$6 \quad \frac{p}{(2k\ \ 2)!} \quad \frac{M_j}{2^{\ 2} n^2}^{k} \quad \frac{4k^{2\ 2}}{t^2} + \frac{p\overline{(2k\ \ 2)!}}{p\overline{2}} \quad \frac{M_j}{2^{\ 2} n^2}^{k} \quad \frac{2k}{t}:$$

Hence, due to the triangle inequality,

$$\int_0^1 v_1 dv_1 \dots \int_0^1 v_k dv_k \int_{R^k \setminus [-\frac{nt}{2k}, \frac{nt}{2k}]^k} \dots \int_1 dy \quad {}^{(2k)} \quad t \sum_{i=1}^k \frac{v_i u_i}{n} \quad \frac{y}{n}^{!} \quad \prod_{i=1}^k \frac{u_i^2 r_j(u_i)}{n^2}^{!} \quad s_j^{(n\ k)}(y)$$
$$6 \sum_{\ell=1}^k \int_0^1 v_1 dv_1 \dots \int_0^1 v_k dv_k \int_A \dots \int_1 dy \quad {}^{(2k)} \quad t \sum_{i=1}^k \frac{v_i u_i}{n} \quad \frac{y}{n}^{!} \quad \prod_{i=1}^k \frac{u_i^2 r_j(u_i)}{n^2}^{!} \quad s_j^{(n\ k)}(y)$$
$$6 \quad \frac{p}{(2k\ \ 2)!} \quad \frac{M_j}{2^{\ 2} n^2}^{k} \quad \frac{4k^{2\ 2}}{t^2} + p\frac{1}{2} \quad \frac{2k}{t} \quad :$$

Lemma A.7. Let $n \in \mathbb{N}$ and $a > 0$ be such that $na \leq 1/2$. Then it holds that

$$\sum_{k=1}^n \binom{n}{k} \frac{p}{(2k)!} \quad \frac{a}{2}^{k} \leq 2na \quad \text{and} \quad \sum_{k=1}^n \binom{n}{k} \frac{p}{(2k)!} \quad k \quad \frac{a}{2}^{k} \leq 4na:$$

Proof. First, note that, for any positive integer k, we have

$$(2k)! = \prod_{j=1}^k (2j) \prod_{j=1}^k (2j \quad 1) \leq \prod_{j=1}^k (2j) \prod_{j=1}^k (2j) \leq 4^k \quad k! \quad k!:$$

This implies that

$$\sum_{k=1}^n \binom{n}{k} \frac{p}{(2k)!} \quad \frac{a}{2}^{k} \leq \sum_{k=1}^n \frac{n!}{(n \quad k)!} \quad a^k \leq \sum_{k=1}^n (na)^k \leq \sum_{k=1}^{1} (na)^k = \frac{na}{1 \quad na} \leq 2na:$$

Similarly, it holds that

$$\sum_{k=1}^n \binom{n}{k} \frac{p}{(2k)!} \quad k \quad \frac{a}{2}^{k} \leq \sum_{k=1}^n \frac{n!}{(n \quad k)!} \quad ka^k \leq \sum_{k=1}^n k(na)^k \leq na \sum_{k=0}^{1} k(na)^{k \quad 1} = \frac{na}{(1 \quad na)^2} \leq 4na:$$

B Proofs for clipped-SGD

B.1 Auxiliary Results

Bernstein inequality. The following lemma (known as Bernstein inequality for martingale differences (Bennett, 1962; Dzhaparidze and Van Zanten, 2001; Freedman et al., 1975)) is essential for deriving high-probability upper bounds in our analysis.

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

$$\mathbb{P}\left\{\sum_{i=1}^n X_i > b \text{ and } \sum_{i=1}^n \sigma_i^2 \le G\right\} \le 2\exp\left(-\frac{b^2}{2G + \frac{2cb}{3}}\right). \tag{41}$$

Bias and variance of the clipped stochastic vector. We also rely on the following result from (Gorbunov et al., 2020).

Lemma B.2 (Simplified version of Lemma F.5 from (Gorbunov et al., 2020)). Let X be a random vector in $\mathbb{R}^d$ and $\widehat{X} = \text{clip}(X, \lambda)$. Then,

$$\|\widehat{X} - \mathbb{E}[\widehat{X}]\| \le 2\lambda. \tag{42}$$

Moreover, if for some $\sigma > 0$

$$\mathbb{E}[X] = x \in \mathbb{R}^d, \quad \mathbb{E}[\|X - x\|^2] \le \sigma^2 \tag{43}$$

and $\lambda \le \|x\|/2$, then

$$\|\mathbb{E}[\widehat{X}] - x\| \le \frac{4\sigma^2}{\lambda}; \tag{44}$$

$$\mathbb{E}\left[\|\widehat{X} - \mathbb{E}[\widehat{X}]\|^2\right] \le 18\sigma^2. \tag{45}$$

B.2 Quasi-Convex Case

The analysis of clipped-SGD in the quasi-convex case relies on the following lemma from (Sadiev et al., 2023).

Lemma B.3. Let Assumptions 2.2 and 2.4[4] with $\sigma = 0$ hold on $Q = B_{2R}(x^*)$, where $R > \|x^0 - x^*\|$, and let stepsize γ_k satisfy $\gamma_k \le \frac{1}{L}$. If $x^k \in Q$ for all $k = 0, 1, \ldots, K+1$, $K > 0$, then after K iterations of clipped-SGD we have

$$f(\bar{x}^K) - f(x^*) \le \frac{\|x^0 - x^*\|^2 - \|x^{K+1} - x^*\|^2}{2\sum_{k=0}^K \gamma_k} + \frac{\sum_{k=0}^K \gamma_k \langle x^k - x^*, \nabla f(x^k) - \omega_k \rangle}{\sum_{k=0}^K \gamma_k} + \frac{\sum_{k=0}^K \gamma_k^2 \|\omega_k\|^2}{\sum_{k=0}^K \gamma_k}; \tag{46}$$

$$\bar{x}^K = \frac{\sum_{k=0}^K \gamma_k x^k}{\sum_{k=0}^K \gamma_k}; \tag{47}$$

$$\omega_k \stackrel{\text{def}}{=} \text{clip}\left(\nabla f_{i_k}(x^k), \lambda_k\right) - \nabla f(x^k). \tag{48}$$

Theorem B.4. Let Assumptions 2.2 and 2.4 with $\sigma = 0$ hold on $Q = B_{2R}(x^*)$, where $R > \|x^0 - x^*\|$. Assume

[4] Although Sadiev et al. (2023) claim that they use Assumption 2.3 with $\sigma = 0$, their proof relies on Assumption 2.4 with $\sigma = 0$ instead, which is strictly weaker.

that $\nabla f_{\xi_k}(x^k)$ satisfies Assumption 5.1 with parameters b_k, σ_k for $k = 0, 1, \ldots, K$, $K > 0$ and

$$\gamma_k \leq \min\left\{\frac{1}{160L\ln\frac{4(K+1)}{\beta}}, \frac{1}{208\sigma_k}\sqrt{\frac{R}{K\ln\frac{4(K+1)}{\beta}}}, \frac{R}{160b_k\ln\frac{4(K+1)}{\beta}}, \frac{R}{1600b_k(K+1)}\right\}, \tag{49}$$

$$\gamma_k = \frac{R}{40\ln\frac{4(K+1)}{\beta}}; \tag{50}$$

for some $\beta \in (0, 1]$. Then, after K iterations of clipped-SGD the iterates with probability at least $1-\beta$ satisfy

$$f(x^K) - f(x^*) \leq \frac{2R^2}{(K+1)} \quad \text{and} \quad \{x^k\}_{k=0}^K \subseteq B_{\sqrt{2}R}(x^*). \tag{51}$$

In particular, when

$$\gamma = \min\left\{\frac{1}{160L\ln\frac{4(K+1)}{\beta}}, \frac{1}{208\sigma}\sqrt{\frac{R}{K\ln\frac{4(K+1)}{\beta}}}, \frac{R}{160b\ln\frac{4(K+1)}{\beta}}, \frac{R}{1600b(K+1)}\right\}, \tag{52}$$

$$\text{where} \quad \sigma = \max_{k=0,1,\ldots,K}\sigma_k; \quad b = \max_{k=0,1,\ldots,K} b_k; \tag{53}$$

then the iterates produced by clipped-SGD after K iterations with probability at least $1-\beta$ satisfy

$$f(x^K) - f(x^*) = \mathcal{O}\left(\max\left\{\frac{LR^2\ln\frac{K}{\beta}}{K}, \frac{R\sigma\sqrt{\ln\frac{K}{\beta}}}{\sqrt{K}}, \frac{bR\ln\frac{K}{\beta}}{K}, bR\right\}\right). \tag{54}$$

Proof. Our proof follows similar steps to the one given by Sadiev et al. (2023). The main difference comes due to the presence of the bias in $\nabla f_{\xi_k}(x^k)$. Therefore, for completeness, we provide the full proof here.

Let $R_k = \|x^k - x^*\|$ for all $k > 0$. Our next objective is to establish, by induction, that $R_l \leq 2R$ with a high probability. This will enable us to apply the result from Lemma B.3 and subsequently utilize Bernstein's inequality to estimate the stochastic component of the upper bound. To be more precise, for each $k = 0, \ldots, K+1$, we consider the probability event E_k, defined as follows: inequalities

$$2\sum_{l=0}^{k-1}\langle x^l - x^*, \nabla f(x^l), \eta_l\rangle + 2\sum_{l=0}^{k-1}\|\eta_l\|^2 \leq R^2, \tag{55}$$

$$R_t \leq \sqrt{2}R \tag{56}$$

hold for all $t = 0, 1, \ldots, k$ simultaneously. We aim to demonstrate through induction that $\Pr\{E_k\} \geq 1 - k\beta/(K+1)$ for all $k = 0, 1, \ldots, K+1$. The base case, $k = 0$, is trivial. Assuming that the statement holds for some $k = T - 1 \leq K$, specifically, $\Pr\{E_{T-1}\} \geq 1 - (T-1)\beta/(K+1)$, we need to establish that $\Pr\{E_T\} \geq 1 - T\beta/(K+1)$.

To begin, we observe that the probability event E_{T-1} implies that $x_t \in B_{\sqrt{2}R}(x^*)$ for all $t = 0, 1, \ldots, T-1$. Furthermore, E_{T-1} implies that

$$\|x^T - x^*\| = \|x^{T-1} - x^* - \gamma\nabla f_{\xi_{T-1}}(x^{T-1})\| \leq \|x^{T-1} - x^*\| + \gamma\|\nabla f_{\xi_{T-1}}(x^{T-1})\| \leq \sqrt{2}R + \gamma \overset{(50)}{\leq} 2R,$$

i.e., $x^0, x^1, \ldots, x^T \in B_{2R}(x^*)$. Hence, with E_{T-1} implying $\{x^k\}_{k=0}^T \subseteq Q$, we confirm that the conditions of Lemma B.3 are met, resulting in

$$f(x^t) - f(x^*) \leq \frac{\|x^0 - x^*\|^2 - \|x^t - x^*\|^2}{t} - \frac{2}{t}\sum_{l=0}^{t-1}\langle x^l - x^*, \nabla f(x^l), \eta_l\rangle + \frac{2}{t}\sum_{l=0}^{t-1}\|\eta_l\|^2 \tag{57}$$

for all $t = 1; \dots; T$ simultaneously and for all $t = 1; \dots; T$ 1 this probability event also implies that

$$f(x^{t+1}) - f(x^*) \le \frac{1}{t}\left(R^2 - 2\sum_{l=0}^{t-1}\eta_l\langle x^l - x^* - r f(x^l), \xi_l\rangle + \sum_{l=0}^{t-1}\eta_l^2\|\xi_l\|^2\right) \overset{(55)}{\le} \frac{2R^2}{t}. \tag{58}$$

Considering that $f(x^{T+1}) - f(x^*) > 0$, we can further deduce from (57) that when E_{T+1} holds, the following holds as well:

$$R_T^2 \le R^2 - 2\sum_{l=0}^{t-1}\eta_l\langle x^l - x^* - r f(x^l), \xi_l\rangle + \sum_{l=0}^{t-1}\eta_l^2\|\xi_l\|^2. \tag{59}$$

Next, we define random vectors

$$\xi_t = \begin{cases} x^t - x^* - r f(x^t), & \text{if } \|x^t - x^* - r f(x^t)\| \le 2R, \\ 0, & \text{otherwise;} \end{cases}$$

for all $t = 0; 1; \dots; T-1$. As per their definition, these random vectors are bounded with probability 1

$$\|\xi_t\| \le 2R. \tag{60}$$

Moreover, for $t = 0; \dots; T-1$ event E_{T+1} implies

$$\|r f(x^t)\| \overset{(3)}{\le} L\|x^t - x^*\| \overset{(56)}{\le} \sqrt{2}LR \overset{(49);(50)}{\le} \frac{\sqrt{2}}{2}; \tag{61}$$

$$E_t[r f(x^t)] \le E_t[r f(x^t)] - r f(x^t) + \|r f(x^t)\| \overset{(7);(61)}{\le} b_t + \sqrt{2}LR \overset{(49);(50)}{\le} \frac{\sqrt{2}}{2}; \tag{62}$$

$$\|x^t - x^* - r f(x^t)\| \le \|x^t - x^*\| + \|r f(x^t)\| \overset{(61)}{\le} \sqrt{2}R(1+L) \overset{(49)}{\le} 2R.$$

The latter inequality means that E_{T+1} implies $\xi_t = x^t - x^* - r f(x^t)$ for $t = 0; \dots; T-1$. Next, we define the unbiased part and the bias of ξ_t as ξ_t^u and ξ_t^b, respectively:

$$\xi_t^u = \text{clip}(r f_t(x^t), \lambda_t) - E_t[\text{clip}(r f_t(x^t), \lambda_t)]; \quad \xi_t^b = E_t[\text{clip}(r f_t(x^t), \lambda_t)] - r f(x^t). \tag{63}$$

We notice that $\xi_t = \xi_t^u + \xi_t^b$. Using new notation, we get that E_{T+1} implies

$$\begin{aligned} R_T^2 \le R^2 &- 2\sum_{t=0}^{T-1}\langle \xi_t^u, \xi_t\rangle - 2\sum_{t=0}^{T-1}\langle \xi_t^b, \xi_t\rangle + 2\sum_{t=0}^{T-1}\|\xi_t^u\|^2 - E_t\|\xi_t^u\|^2 \\ &+ 2\sum_{t=0}^{T-1}E_t\|\xi_t^u\|^2 + 2\sum_{t=0}^{T-1}\|\xi_t^b\|^2. \end{aligned} \tag{64}$$

To conclude our inductive proof successfully, we must obtain sufficiently strong upper bounds with high probability for the terms $\alpha; \beta; \gamma; \delta; \varepsilon$. In other words, we need to demonstrate that $\alpha + \beta + \gamma + \delta + \varepsilon \le R^2$ with a high probability. In the subsequent stages of the proof, we will rely on the bounds for the norms and second moments of ξ_t^u and ξ_t^b. First, as per the definition of the clipping operator, we can assert with probability 1 that

$$\|\xi_t^u\| \le 2\lambda_t. \tag{65}$$

Furthermore, given that E_{T+1} implies $\|E_t[r f_t(x^t)]\| \le \frac{\sigma}{\sqrt{2}}$ for $t = 0; 1; \dots; T-1$ (as per (62)), then, according to Lemma B.2, we can deduce that E_{T+1} implies

$$\|\xi_t^b\| \le \|E_t[\text{clip}(r f_t(x^t), \lambda_t)] - E_t[r f_t(x^t)]\| + \|E_t[r f_t(x^t)] - r f(x^t)\| \overset{(44);(7)}{\le} \frac{4\sigma^2}{\lambda_t} + b_t; \tag{66}$$

$$E_t\|\xi_t^u\|^2 \overset{(45)}{\le} 18\sigma^2. \tag{67}$$

Upper bound for $\neg$. By definition of $\overset{u}{t}$, we readily observe that $E_t[\overset{u}{t}] = 0$ and

$$E_t[\ 2\ h\overset{u}{t};\ _t i] = 0:$$

Next, the sum $\neg$ contains only the terms that are bounded with probability 1:

$$j2\ h\overset{u}{t};\ _t ij \ 6\ 2\ k\overset{u}{t}k\ \ k\ _t k \overset{(60);(65)}{6}\ 8\ R\ \overset{(50)}{=}\ \frac{R^2}{5\ln\frac{4(K+1)}{}} \overset{def}{=} c: \tag{68}$$

The conditional variances $\overset{2}{t} \overset{def}{=} E_t[4\ ^2 h\overset{u}{t};\ _t i^2]$ of the summands are bounded:

$$\overset{2}{t} \ 6\ E_t\ \ 4\ ^2 k\overset{u}{t}k^2\ \ k\ _t k^2 \overset{(60)}{6}\ 16\ ^2R^2E_t\ \ k\overset{u}{t}k^2\ : \tag{69}$$

To summarize, we have demonstrated that $f\ \ 2\ h\overset{u}{t};\ _t ig_{t=0}^{T\ 1}$ is a bounded martingale difference sequence with bounded conditional variances $f\ \overset{2}{t}g_{t=0}^{T\ 1}$. Therefore, one can apply Bernstein's inequality (Lemma B.1) with $X_t = \ \ 2\ h\overset{u}{t};\ _t i$, parameter c as in (68), $b = \frac{R^2}{5}$, $G = \frac{R^4}{150\ln\frac{4(K+1)}{}}$ and get

$$P\ \ j\neg j > \frac{R^2}{5}\ \ \text{and}\ \ \overline{X}^{\ 1}_{t=0}\ \overset{2}{t} \ 6\ \frac{R^4}{150\ln\frac{4(K+1)}{}}\ \ \Big) \ \ 6\ 2\exp\ \ \frac{b^2}{2G + \ ^2cb=3}\ \ = \frac{}{2(K\ +1)};$$

which is equivalent to

$$P\mathrm{f}\,E_\neg g > 1\ \ \frac{}{2(K\ +1)};\ \ \text{for}\ \ E_\neg = \ \ \Big(\ \text{either}\ \overline{X}^{\ 1}_{t=0}\ \overset{2}{t} > \frac{R^4}{150\ln\frac{4(K+1)}{}}\ \ \text{or}\ \ j\neg j\ 6\ \frac{R^2}{5}\ \Big) : \tag{70}$$

Additionally, event $E_{T\ \ 1}$ implies that

$$\overline{X}^{\ 1}_{t=0}\ \overset{2}{t} \overset{(69)}{6}\ 16\ ^2R^2\overline{X}^{\ 1}_{t=0}\ E_t\ \ k\overset{u}{t}k^2 \overset{(67)}{6}\ 288\ ^2R^2\ \ ^2T \overset{(49)}{6}\ \frac{R^4}{150\ln\frac{4(K+1)}{}}: \tag{71}$$

Upper bound for $\ -$. From $E_{T\ \ 1}$ it follows that

$$-\ \ = \ \ 2\overline{X}^{\ 1}_{t=0}\ h\overset{b}{t};\ _t i\ 6\ 2\overline{X}^{\ 1}_{t=0}\ k\overset{b}{t}k\ \ k\ _t k \overset{(60);(66)}{6}\ \frac{16\ \ ^2TR}{} + 4\ RbT$$
$$\overset{(50)}{=}\ \ \frac{2\ \ ^2\ \ ^2T\ln\frac{4(K+1)}{}}{5} + 4\ RbT\ \overset{(49)}{6}\ \frac{R^2}{5}: \tag{72}$$

Upper bound for ®. By construction, we have

$$E_t\ \ 2\ \ ^2\ \ k\overset{u}{t}k^2\ \ \ E_t\ \ k\overset{u}{t}k^2\ \ \ = 0:$$

Next, the sum ® contains only the terms that are bounded with probability 1:

$$2\ \ ^2\ \ k\overset{u}{t}k^2\ \ \ E_t\ \ k\overset{u}{t}k^2\ \ \ \ 6\ \ 2\ \ ^2\ \ k\overset{u}{t}k^2 + E_t\ \ k\overset{u}{t}k^2$$
$$\overset{(65)}{6}\ \ 16\ \ ^2\ \ ^2\ \overset{(50)}{=}\ \frac{R^2}{100\ln^2\frac{4(K+1)}{}}\ 6\ \frac{R^2}{5\ln\frac{4(K+1)}{}} \overset{def}{=} c: \tag{73}$$

The conditional variances $e_t^2 \overset{def}{=} E_t\ \ 4\ ^4\ \ k\overset{u}{t}k^2\ \ \ E_t\ \ k\overset{u}{t}k^2\ \ \ ^2$ of the summands are bounded:

$$e_t^2 \ \overset{(73)}{6}\ \ \frac{R^2}{5\ln\frac{4(K+1)}{}}E_t\ \ 2\ \ ^2\ \ k\overset{u}{t}k^2\ \ \ E_t\ \ k\overset{u}{t}k^2\ \ \ \ 6\ \ \frac{4\ ^2R^2}{5\ln\frac{4(K+1)}{}}E_t\ \ k\overset{u}{t}k^2\ : \tag{74}$$

To summarize, we have demonstrated that $\left\{2^{2}\left\|u_t\right\|^2 - \mathbb{E}_t\left\|u_t\right\|^2\right\}_{t=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\left\{e_t^2\right\}_{t=0}^{T-1}$. Therefore, one can apply Bernstein's inequality (Lemma B.1) with $X_t = 2^{2}\left\|u_t\right\|^2 - \mathbb{E}_t\left\|u_t\right\|^2$, parameter c as in (73), $b = \frac{R^2}{5}$, $G = \frac{R^4}{150\ln\frac{4(K+1)}{}}$ and get

$$P\left\{|\cdot| > \frac{R^2}{5} \text{ and } \sum_{t=0}^{T-1} e_t^2 \le \frac{R^4}{150\ln\frac{4(K+1)}{}}\right\} \le 2\exp\left(\frac{b^2}{2G + \frac{2cb}{3}}\right) = \frac{}{2(K+1)};$$

which is equivalent to

$$P\{E_\cdot\} > 1 - \frac{}{2(K+1)}; \quad \text{for } E_\cdot = \left(\text{either } \sum_{t=0}^{T-1} e_t^2 > \frac{R^4}{150\ln\frac{4(K+1)}{}} \text{ or } |\cdot| \le \frac{R^2}{5}\right): \tag{75}$$

Additionally, event E_{T-1} implies that

$$\sum_{t=0}^{T-1} e_t^2 \overset{(74)}{\le} \frac{4^{2}R^2}{5\ln\frac{4(K+1)}{}} \sum_{t=0}^{T-1} \mathbb{E}_t\left\|u_t\right\|^2 \overset{(67)}{\le} \frac{72^{2}R^2\,^{2}T}{5\ln\frac{4(K+1)}{}} \overset{(49)}{\le} \frac{R^4}{150\ln\frac{4(K+1)}{}}. \tag{76}$$

Upper bound for $\bar{}$. From E_{T-1} it follows that

$$\bar{} = 2^{2}\sum_{t=0}^{T-1}\left[\mathbb{E}_t\left\|u_t\right\|^2\right] \overset{(67)}{\le} 36^{2}\,^{2}T \overset{(49)}{\le} \frac{R^2}{5}. \tag{77}$$

Upper bound for $\circ$. From E_{T-1} it follows that

$$\circ = 2^{2}\sum_{t=0}^{T-1} b_t^{2} \overset{(66)}{\le} \frac{32^{4}T^{2}}{^{2}} + 2^{2}b^2T \overset{(50)}{=} 51200\frac{^{4}T^{4}\ln\frac{4(K+1)}{}}{R^2} + 2^{2}b^2T \overset{(49)}{\le} \frac{R^2}{5}. \tag{78}$$

That is, we derived the upper bounds for $\neg; \cdot; \circledR; \bar{}; \circ$. More specically, the probability event E_{T-1} implies:

$$\begin{gathered}
R_T^2 \overset{(64)}{\le} R^2 + \neg + \cdot + \circledR + \bar{} + \circ; \\
\cdot \overset{(72)}{\le} \frac{R^2}{5}; \quad \bar{} \overset{(77)}{\le} \frac{R^2}{5}; \quad \circ \overset{(78)}{\le} \frac{R^2}{5}; \\
\sum_{t=0}^{T-1} b_t^{2} \overset{(71)}{\le} \frac{R^4}{150\ln\frac{4(K+1)}{}}; \quad \sum_{t=0}^{T-1} e_t^2 \overset{(76)}{\le} \frac{R^4}{150\ln\frac{4(K+1)}{}}.
\end{gathered}$$

In addition, we also have (see (70), (75) and our induction assumption)

$$P\{E_{T-1}\} > 1 - \frac{(T-1)}{K+1}; \quad P\{E_\neg\} > 1 - \frac{}{2(K+1)}; \quad P\{E_\circledR\} > 1 - \frac{}{2(K+1)};$$

where

$$\begin{aligned}
E_\neg &= \left(\text{either } \sum_{t=0}^{T-1} b_t^{2} > \frac{R^4}{150\ln\frac{4(K+1)}{}} \text{ or } |\neg| \le \frac{R^2}{5}\right); \\
E_\circledR &= \left(\text{either } \sum_{t=0}^{T-1} e_t^2 > \frac{R^4}{150\ln\frac{4(K+1)}{}} \text{ or } |\circledR| \le \frac{R^2}{5}\right):
\end{aligned}$$

Therefore, probability event $E_{T-1} \setminus E_\neg \setminus E_\circledR$ implies

$$R_T^2 \le R^2 + \frac{R^2}{5} + \frac{R^2}{5} + \frac{R^2}{5} + \frac{R^2}{5} + \frac{R^2}{5} = 2R^2;$$

which is equivalent to (55) and (56) for $t = T$, and

$$\mathbb{P}\{E_T\} > \mathbb{P}\{E_{T-1} \setminus E_\neg \setminus E_®\} = 1 - \mathbb{P}\{\overline{E}_{T-1} \cup \overline{E}_\neg \cup \overline{E}_®\} > 1 - \mathbb{P}\{\overline{E}_{T-1}\} - \mathbb{P}\{\overline{E}_\neg\} - \mathbb{P}\{\overline{E}_®\} > 1 - \frac{T}{K+1}.$$

We have now completed the inductive part of our proof. That is, for all $k = 0, 1, \ldots, K+1$, we have $\mathbb{P}\{E_k\} > 1 - k/(K+1)$. Notably, when $k = K+1$, we can conclude that with a probability of at least $1 - \delta$:

$$f(\bar{x}^K) - f(x^*) \overset{(58)}{\leq} \frac{2R^2}{(K+1)}$$

and $\{x^k\}_{k=0}^K \subseteq Q$, which follows from (56).

Finally, if

$$\gamma = \min\left\{\frac{1}{160L\ln\frac{4(K+1)}{\delta}}, \frac{1}{208}\sqrt{\frac{R}{K\ln\frac{4(K+1)}{\delta}}}, \frac{R}{160b\ln\frac{4(K+1)}{\delta}}, \frac{R}{1600b(K+1)}\right\},$$

then with probability at least $1 - \delta$

$$\begin{aligned} f(\bar{x}^K) - f(x^*) &\leq \frac{2R^2}{(K+1)} \max\left\{\frac{320LR^2\ln\frac{4(K+1)}{\delta}}{K+1}, \frac{416R\sqrt{K\ln\frac{4(K+1)}{\delta}}}{K+1}, \frac{320bR\ln\frac{4(K+1)}{\delta}}{K+1}, 320bR\right\} \\ &= \mathcal{O}\left(\max\left\{\frac{LR^2\ln\frac{K}{\delta}}{K}, \frac{R\sqrt{\ln\frac{K}{\delta}}}{\sqrt{K}}, \frac{bR\ln\frac{K}{\delta}}{K}, bR\right\}\right). \end{aligned}$$

This concludes the proof.

B.3 Quasi-Strongly Convex Case

Lemma B.5. Let Assumptions 2.2 and 2.4 with $\mu = 0$ hold on $Q = B_{2R}(x^*)$, where $R > \|x^0 - x^*\|$, and let stepsize $\gamma_k = \gamma$ satisfy $\gamma \leq \frac{1}{L}$. If $x^k \in Q$ for all $k = 0, 1, \ldots, K$, $K > 0$, then after K iterations of clipped-SGD we have

$$\begin{aligned} \|x^{K+1} - x^*\|^2 \leq \exp(-\gamma\mu(K+1))\|x^0 - x^*\|^2 &- 2\sum_{k=0}^K \exp(-\gamma\mu(K-k))\langle x^k - x^* - \gamma\nabla f(x^k), \gamma\eta_k\rangle \\ &+ \gamma^2\sum_{k=0}^K \exp(-\gamma\mu(K-k))\|\gamma\eta_k\|^2, \end{aligned} \tag{79}$$

where η_k is defined in (48).

Proof. Using the update rule of clipped-SGD, we obtain

$$\begin{aligned}
\|x^{k+1} - x\|^2 &= \|x^k - x\|^2 - 2\langle x^k - x, \operatorname{clip}(\nabla f_{\xi_k}(x^k), \lambda_k)\rangle + \eta^2\|\operatorname{clip}(\nabla f_{\xi_k}(x^k), \lambda_k)\|^2 \\
&= \|x^k - x\|^2 - 2\langle x^k - x, \nabla f(x^k)\rangle - 2\langle x^k - x, \eta_k\rangle \\
&\quad + \eta^2\|\nabla f(x^k)\|^2 + 2\eta\langle \nabla f(x^k), \eta_k\rangle + \eta^2\|\eta_k\|^2 \\
&= \|x^k - x\|^2 - 2\eta\langle x^k - x, \nabla f(x^k)\rangle \\
&\quad - 2\eta\langle x^k - x, \nabla f(x^k)\rangle + \eta^2\|\nabla f(x^k)\|^2 + \eta^2\|\eta_k\|^2 \\
&\overset{(6);(4)}{\leq} (1-\eta\mu)\|x^k - x\|^2 - 2\eta\langle x^k - x, \nabla f(x^k)\rangle \\
&\quad - 2\eta(f(x^k) - f(x)) + 2L\eta^2(f(x^k) - f(x)) + \eta^2\|\eta_k\|^2 \\
&\overset{\eta \leq 1/L}{\leq} (1-\eta\mu)\|x^k - x\|^2 - 2\eta\langle x^k - x, \nabla f(x^k)\rangle + \eta^2\|\eta_k\|^2 \\
&\quad - 2\eta(f(x^k) - f(x)) + 2\eta(f(x^k) - f(x)) + \eta^2\|\eta_k\|^2 \\
&\leq (1-\eta\mu)\|x^k - x\|^2 - 2\eta\langle x^k - x, \nabla f(x^k)\rangle + \eta^2\|\eta_k\|^2 \\
&\leq \exp(-\eta\mu)\|x^k - x\|^2 - 2\eta\langle x^k - x, \nabla f(x^k)\rangle + \eta^2\|\eta_k\|^2.
\end{aligned}$$

Unrolling the recurrence, we obtain (79).

Theorem B.6. Let Assumptions 2.2 and 2.4 with $\mu > 0$ hold on $Q = B_{2R}(x^*)$, where $R > \|x^0 - x^*\|$. Assume that $\nabla f_{\xi_k}(x^k)$ satisfies Assumption 5.1 with parameters b_k, α_k for $k = 0, 1, \ldots, K$, $K > 0$ and

$$0 < \eta \leq \min\left(\frac{1}{400L\ln\frac{4(K+1)}{\beta}}; \frac{\ln(B_K)}{\mu(K+1)}; \frac{\ln(C_K)}{\mu(1+\alpha/2)(K+1)}; \frac{2\ln(D)}{\mu(K+1)}\right); \tag{80}$$

$$B_K = \max\left\{2; \frac{(K+1)\mu^2R^2}{5400\sigma^2\ln\frac{4(K+1)}{\beta}\ln^2(B_K)}\right\} = O\left(\max\left\{2; \frac{K\mu^2R^2}{\sigma^2\ln\frac{K}{\beta}\ln^2\max\left\{2; \frac{K\mu^2R^2}{\sigma^2\ln\frac{K}{\beta}}\right\}}\right\}\right); \tag{81}$$

$$C_K = \max\left\{2; \frac{(\frac{K}{2}+1)R}{480b\ln\frac{4(K+1)}{\beta}\ln(C_K)}\right\} = O\left(\max\left\{2; \frac{KR}{b\ln\frac{K}{\beta}\ln\max\left\{2; \frac{KR}{b\ln\frac{K}{\beta}}\right\}}\right\}\right); \tag{82}$$

$$D = \max\left\{2; \frac{R}{80b\ln(D)}\right\} = O\left(\max\left\{2; \frac{R}{b\ln\max\left\{2; \frac{R}{b}\right\}}\right\}\right); \tag{83}$$

$$\eta_k = \frac{\exp(-\mu\eta(1+\alpha/2))R}{120\ln\frac{4(K+1)}{\beta}}; \tag{84}$$

for some $\beta \in (0; 1]$ and $b = \max_{k=0,1,\ldots,K} b_k$, $\alpha = \max_{k=0,1,\ldots,K} \alpha_k$. Then, after K iterations the iterates produced by clipped-SGD with probability at least $1-\beta$ satisfy

$$\|x^{K+1} - x^*\|^2 \leq 2\exp(-\mu\eta(K+1))R^2. \tag{85}$$

In particular, when η equals the minimum from (80), then the iterates produced by clipped-SGD after K iterations with probability at least $1-\beta$ satisfy

$$\|x^K - x^*\|^2 = O\left(\max\left\{R^2\exp\left(-\frac{\mu K}{L\ln\frac{K}{\beta}}\right); \frac{\sigma^2\ln\frac{K}{\beta}\ln^2(B_K)}{\mu^2K}; \frac{bR\ln\frac{K}{\beta}\ln(C_K)}{K}; \frac{bR\ln(D)}{K}\right\}\right). \tag{86}$$

Proof. Our proof follows similar steps to the one given by Sadiev et al. (2023). The main difference comes due to the presence of the bias in $\nabla f_{\xi_k}(x^k)$. Therefore, for completeness, we provide the full proof here.

Let $R_k = \|x^k - x^*\|$ for all $k > 0$. As in the previous results, the main part of the proof is inductive. More precisely, for each $k = 0, 1, \ldots, K+1$ we consider probability event E_k as follows: inequalities

$$R_t^2 \leq 2\exp(-\mu\eta t)R^2 \tag{87}$$

hold for $t = 0; 1; : : : ; k$ simultaneously. We aim to demonstrate through induction that $\mathbb{P}\{E_k\} > 1 - k=(K+1)$ for all $k = 0; 1; : : : ; K+1$. The base case, $k = 0$, is trivial. Assuming that the statement holds for some $k = T - 1 \leq K$, specifically, $\mathbb{P}\{E_{T-1}\} > 1 - (T-1)=(K+1)$, we need to establish that $\mathbb{P}\{E_T\} > 1 - T=(K+1)$. Since $R_t^2 \leq 2\exp(-t)R^2 \leq 2R^2$, we have $x^t \in B_{2R}(x^*)$, where function f is L-smooth. Thus, E_{T-1} implies

$$\|\nabla f(x^t)\| \overset{(87)}{\leq} L\|x^t - x^*\| \leq \sqrt{2}L\exp(-t/2)R; \tag{88}$$

$$\mathbb{E}_t[\nabla f_t(x^t)] \overset{(80);(82);(84)}{\leq} \mathbb{E}_t[\nabla f_t(x^t)] - \nabla f(x^t) + \|\nabla f(x^t)\| \overset{(7);(88)}{\leq} b + \sqrt{2}L\exp(-t/2)R \tag{89}$$

and

$$\|g_t\|^2 \leq 2\|\nabla f(x^t)\|^2 + 2\|\nabla f(x^t)\|^2 \overset{(88)}{\leq} \frac{5}{2}\lambda_t^2 \overset{(84)}{\leq} \frac{\exp(-t)R^2}{4\lambda^2} \tag{90}$$

for all $t = 0; 1; : : : ; T-1$, where we use that $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ holding for all $a; b \in \mathbb{R}^d$.

Using Lemma B.5, we obtain that E_{T-1} implies

$$R_T^2 \leq \exp(-T)R^2 - 2\sum_{t=0}^{T-1}\exp(-(T-1-t))\langle x^t - x^* - \gamma\nabla f(x^t); \gamma g_t\rangle + \gamma^2\sum_{t=0}^{T-1}\exp(-(T-1-t))\|\gamma g_t\|^2.$$

Next, we define random vectors

$$\delta_t = \begin{cases} x^t - x^* - \gamma\nabla f(x^t); & \text{if } \|x^t - x^* - \gamma\nabla f(x^t)\| \leq \sqrt{2}(1+L)\exp(-t/2)R; \\ 0; & \text{otherwise;} \end{cases} \tag{91}$$

for $t = 0; 1; : : : ; T-1$. As per their definition, these random vectors are bounded with probability 1

$$\|\delta_t\| \leq \sqrt{2}(1+L)\exp(-t/2)R \tag{92}$$

for all $t = 0; 1; : : : ; T-1$. Moreover, for $t = 0; : : : ; T-1$ event E_{T-1} implies $\|\nabla f(x^t)\| \leq \sqrt{2}L\exp(-t/2)R$ (due to (88)) and

$$\|x^t - x^* - \gamma\nabla f(x^t)\| \leq \|x^t - x^*\| + \gamma\|\nabla f(x^t)\| \overset{(88)}{\leq} \sqrt{2}(1+L)\exp(-t/2)R$$

for $t = 0; 1; : : : ; T-1$. The latter inequality means that E_{T-1} implies $\delta_t = x^t - x^* - \gamma\nabla f(x^t)$ for all $t = 0; 1; : : : ; T-1$, meaning that from E_{T-1} it follows that

$$R_T^2 \leq \exp(-T)R^2 - 2\sum_{t=0}^{T-1}\exp(-(T-1-t))\langle\delta_t; \gamma g_t\rangle + \gamma^2\sum_{t=0}^{T-1}\exp(-(T-1-t))\|\gamma g_t\|^2.$$

Next, we define the unbiased part and the bias of γg_t as γg_t^u and γg_t^b, respectively:

$$\gamma g_t^u \overset{\text{def}}{=} \text{clip}(\nabla f_t(x^t); \lambda_t) - \mathbb{E}_t\left[\text{clip}(\nabla f_t(x^t); \lambda_t)\right]; \quad \gamma g_t^b \overset{\text{def}}{=} \mathbb{E}_t\left[\text{clip}(\nabla f_t(x^t); \lambda_t)\right] - \nabla f(x^t); \tag{93}$$

for all $t = 0; \dots; T-1$. We notice that $\nabla_t = \nabla^u_t + \nabla^b_t$. Using new notation, we get that E_{T-1} implies

$$R_T^2 \le \exp(-\gamma(T-1))R^2 - 2\sum_{t=0}^{T-1}\exp(-\gamma(T-1-t))\langle x^t - x^*, \nabla^u_t\rangle$$
$$- 2\sum_{t=0}^{T-1}\exp(-\gamma(T-1-t))\langle x^t - x^*, \nabla^b_t\rangle + 2\eta^2\sum_{t=0}^{T-1}\exp(-\gamma(T-1-t))\mathbb{E}\|\nabla^u_t\|^2$$
$$+ 2\eta^2\sum_{t=0}^{T-1}\exp(-\gamma(T-1-t))\|\nabla^u_t\|^2 - \mathbb{E}\|\nabla^u_t\|^2 + 2\eta^2\sum_{t=0}^{T-1}\exp(-\gamma(T-1-t))\|\nabla^b_t\|^2. \quad (94)$$

where we also apply inequality $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ holding for all $a, b \in \mathbb{R}^d$ to upper bound $\|\nabla_t\|^2$. To conclude our inductive proof successfully, we must obtain sufficiently strong upper bounds with high probability for the terms $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$. In other words, we need to demonstrate that $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 \le \exp(-\gamma T)R^2$ with high probability. In the subsequent stages of the proof, we will rely on the bounds for the norms and second moments of ∇^u_t and ∇^b_t. First, as per the definition of the clipping operator, we can assert with probability 1 that

$$\|\nabla^u_t\| \le 2\lambda_t. \quad (95)$$

Furthermore, given that E_{T-1} implies $\|\mathbb{E}_t[\nabla f_t(x^t)]\| \le \frac{\zeta}{2}$ for $t = 0, 1, \dots, T-1$ (as per (89)), then, according to Lemma B.2, we can deduce that E_{T-1} implies

$$\|\nabla^b_t\| \le \|\mathbb{E}_t\,\mathrm{clip}(\nabla f_t(x^t), \lambda_t) - \mathbb{E}_t\nabla f_t(x^t)\| + \|\mathbb{E}_t\nabla f_t(x^t) - \nabla f(x^t)\|$$
$$\le \frac{4\sigma^2}{\lambda_t} + b, \quad (96)$$
$$\mathbb{E}_t\left[\|\nabla^u_t\|^2\right] \le 18\sigma^2, \quad (97)$$

for all $t = 0, 1, \dots, T-1$.

Upper bound for α_1. By definition of ∇^u_t, we readily observe that $\mathbb{E}_t[\nabla^u_t] = 0$ and

$$\mathbb{E}_t[-2\exp(-\gamma(T-1-t))\langle x^t - x^*, \nabla^u_t\rangle] = 0.$$

Next, the sum α_1 contains only the terms that are bounded with probability 1:

$$|-2\exp(-\gamma(T-1-t))\langle x^t - x^*, \nabla^u_t\rangle| \le 2\exp(-\gamma(T-1-t))\|x^t - x^*\|\,\|\nabla^u_t\|$$
$$\overset{(92),(95)}{\le} 4\sqrt{2}(1+L)\exp(-\gamma(T-1-t))R_t$$
$$\overset{(80),(84)}{\le} \frac{\exp(-\gamma T)R^2}{5\ln\frac{4(K+1)}{\gamma}} \overset{\text{def}}{=} c. \quad (98)$$

The conditional variances $\sigma_t^2 \overset{\text{def}}{=} \mathbb{E}_t 4\eta^2\exp(-2\gamma(T-1-t))\langle x^t - x^*, \nabla^u_t\rangle^2$ of the summands are bounded:

$$\sigma_t^2 \le \mathbb{E}_t 4\eta^2\exp(-(2T-2-2t))\|x^t - x^*\|^2\|\nabla^u_t\|^2$$
$$\overset{(92)}{\le} 8\eta^2(1+L)^2\exp(-(2T-2-t))R^2\mathbb{E}_t\|\nabla^u_t\|^2$$
$$\overset{(80)}{\le} 10\eta^2\exp(-(2T-t))R^2\mathbb{E}_t\|\nabla^u_t\|^2. \quad (99)$$

To summarize, we have demonstrated that $\{2(1-\gamma)^{T-1-t}\langle x^t - x^*, \nabla^u_t\rangle\}_{t=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\sigma_t^2\}_{t=0}^{T-1}$. Therefore, one can apply Bernstein's inequality (Lemma B.1)

with $X_t = 2\ (1\quad)^{T\ 1\ t}h_t;\ {}^{u}_{t}i$, parameter c as in (98), $b = \frac{1}{5}\exp(\quad T\ )R^2$, $G = \frac{\exp(\quad 2\ T\ )R^4}{150\ln\frac{4(K+1)}{}}$ and get

$$P\left(j\neg j > \frac{1}{5}\exp(\quad T\ )R^2 \text{ and } \sum_{t=0}^{\infty}{}^{2}_{t}\ 6\ \frac{\exp(\quad 2\ T\ )R^4}{150\ln\frac{4(K+1)}{}}\right)\ 6\ 2\exp\quad\frac{b^2}{2F+\ 2cb\sqsupset 3}\ =\ \frac{}{2(K\ +1)};$$

which is equivalent to

$$\text{Pf}\,E_{\neg}g > 1\quad\frac{}{2(K\ +1)};\quad \text{for}\quad E_{\neg} = \left(\text{either } \sum_{t=0}^{\infty}{}^{2}_{t} > \frac{\exp(\quad 2\ T\ )R^4}{150\ln\frac{4(K+1)}{}} \text{ or } j\neg j\ 6\ \frac{1}{5}\exp(\quad T\ )R^2\right) : \quad (100)$$

Additionally, event $E_{T\ 1}$ implies that

$$\begin{aligned}
\sum_{t=0}^{\infty}{}^{2}_{t} &\overset{(99)}{6} 10\ {}^{2}\exp(\quad 2\ T\ )R^2\sum_{t=0}^{\infty}\frac{E_{\ t}\ k\ {}^{u}_{t}k^2}{\exp(\quad t\ )}\\
&\overset{(97);T\ 6\ K+1}{6} 180\ {}^{2}\exp(\quad 2\ T\ )R^2\ {}^{2}\sum_{t=0}^{K}\frac{1}{\exp(\quad t\ )}\\
&\overset{(84)}{6} 180\ {}^{2}\exp(\quad 2\ T\ )R^2\ {}^{2}(K+1)\exp(\quad K\ )\\
&\overset{(80);(81)}{6} \frac{\exp(\quad 2\ T\ )R^4}{150\ln\frac{4(K+1)}{}}:
\end{aligned} \quad (101)$$

Upper bound for $\quad$ - . From $E_{T\ 1}$ it follows that

$$\begin{aligned}
-\quad &6\quad 2\quad\exp(\quad\ (T\quad 1))\sum_{t=0}^{\infty}\frac{k\ {}_{t}k\ k\ {}^{b}_{t}k}{\exp(\quad t\ )}\\
&\overset{(92);(96)}{6}\ p\ \overline{2}\ (1+\ L\ )\exp(\quad\ (T\quad 1))R\sum_{t=0}^{\infty}\exp(\ {}^{t}\sqsupset 2)\quad\frac{4\ {}^{2}}{}+\ b\\
&\overset{(80);(84)}{6}\ 3840\ {}^{2}\exp(\quad T\ )\ {}^{2}(K+1)\exp(\ T\ )\ln\frac{4(K+1)}{}\\
&\quad +2\quad\exp(\quad T\ )R(K+1)\exp(\ {}^{T}\sqsupset 2)b\\
&\overset{(80);(81);(83)}{6}\ \frac{1}{5}\exp(\quad T\ )R^2:
\end{aligned} \quad (102)$$

Upper bound for $\quad$ ®. From $E_{T\ 1}$ it follows that

$$\begin{aligned}
® \quad = \quad 2\ {}^{2}\exp(\quad\ (T\quad 1))\sum_{t=0}^{\infty}\frac{E_{\ t}\ k\ {}^{u}_{t}k^2}{\exp(\quad t\ )}\\
\overset{(97)}{6}\quad 144\ {}^{2}\exp(\quad\ (T\quad 1))\ {}^{2}\sum_{t=0}^{\infty}\frac{1}{\exp(\quad t\ )}\\
\overset{(84)}{6}\quad 144\ {}^{2}\exp(\quad\ (T\quad 1))\ {}^{2}(K+1)\exp(\quad K\ )\\
\overset{(80)}{6}\quad\frac{1}{5}\exp(\quad T\ )R^2:
\end{aligned} \quad (103)$$

Upper bound for $\quad\bar{}$. By construction, we have

$$2\ {}^{2}\exp(\quad\ (T\quad 1\quad t))E_{\ t}\ k\ {}^{u}_{t}k^2\quad E_{\ t}\ k\ {}^{u}_{t}k^2\quad = 0:$$

Next, the sum $\bar{\ }$ contains only the terms that are bounded with probability 1:

$$\begin{aligned} 2\ ^2\exp(\quad(T\quad 1\quad t))\ k\ _t^u k^2\quad E_{\ t}\ k\ _t^u k^2 &\overset{(95)}{6} \frac{16\ ^2\exp(\quad T\ )}{\exp(\quad(1+\ t))}\ ^2_l \\ &\overset{(84)}{6} \frac{\exp(\quad T\ )R^2}{5\ln\frac{4(K+1)}{}} \\ &\overset{\text{def}}{=} c: \end{aligned} \tag{104}$$

The conditional variances

$$e_t^2 \overset{\text{def}}{=} E_{\ t}\ \Big[4\ ^4\exp(\ \ 2\quad(T\quad 1\quad t))\ k\ _t^u k^2\quad E_{\ t}\ k\ _t^u k^2\ \ ^2\Big]$$

of the summands are bounded:

$$\begin{aligned} e_t^2\quad &\overset{(104)}{6} \frac{2\ ^2\exp(\ \ 2\ T\ )R^2}{5\exp(\quad(1+\ t))\ln\frac{4(K+1)}{}} E_{\ t}\quad k\ _t^u k^2\quad E_{\ t}\ k\ _t^u k^2 \\ &6 \quad \frac{4\ ^2\exp(\ \ 2\ T\ )R^2}{5\exp(\quad(1+\ t))\ln\frac{4(K+1)}{}} E_{\ t}\ \ k\ _t^u k^2\ : \end{aligned} \tag{105}$$

To summarize, we have demonstrated that $2\ ^2(1\quad\ )^{T\ \ 1\ t}\ k\ _t^u k^2\quad E_{\ t}\ k\ _t^u k^2\quad {}^{T\ \ 1}_{t=0}$ is a bounded martingale di erence sequence with bounded conditional variances $\{e_t^2\}_{t=0}^{T\ \ 1}$. Therefore, one can apply Bernstein's inequality (Lemma B.1) with $X_t = 2\ ^2(1\quad\ )^{T\ \ 1\ t}\ k\ _t^u k^2\quad E_{\ t}\ k\ _t^u k^2$, parameter c as in (104), $b = \frac{1}{5}\exp(\quad T\ )R^2$, $G = \frac{\exp(\ \ 2\ T\ )R^4}{150\ln\frac{4(K+1)}{}}$ and get

$$P\quad\Big(\ \bar{\ }\ j > \frac{1}{5}\exp(\quad T\ )R^2 \text{ and } \bar{\sum_{l=0}^{\ \ 1}} e_t^2\ 6\ \frac{\exp(\ \ 2\ T\ )R^4}{150\ln\frac{4(K+1)}{}}\Big)\quad 6\ 2\exp\quad\frac{b^2}{2G+\ ^{2cb}\!\!=3}\quad =\ \frac{}{2(K\ +1)};$$

which is equivalent to

$$Pf\ E^{-}g > 1\quad\frac{}{2(K\ +1)};\quad \text{for}\quad E^{-}=\quad\Big(\text{either}\quad \bar{\sum_{t=0}^{\ \ 1}} e_t^2 > \frac{\exp(\ \ 2\ T\ )R^4}{150\ln\frac{4(K+1)}{}}\quad\text{or}\quad\ \bar{\ }\ j\ 6\ \frac{1}{5}\exp(\quad T\ )R^2\Big)\ : \tag{106}$$

Additionally, event $E_{T\ \ 1}$ implies that

$$\begin{aligned} \bar{\sum_{l=0}^{\ \ 1}} e_t^2\quad &\overset{(105)}{6} \frac{4\ ^2\exp(\quad(2T\quad 1))R^2}{5\ln\frac{4(K+1)}{}} \bar{\sum_{t=0}^{\ \ 1}} \frac{E_{\ t}\ k\ _l^u k^2}{\exp(\quad t\ )} \\ &\overset{(97);T\ 6\ K+1}{6} \frac{72\ ^2\exp(\quad(2T\quad 1))R^2\ \ ^2}{5\ln\frac{4(K+1)}{}} \sum_{t=0}^{K} \frac{1}{\exp(\quad t\ )} \\ &\overset{(84)}{6} \frac{72\ ^2\exp(\quad(2T\quad 1))R^2\ \ ^2(K\ +1)\exp(\quad K\ )}{5\ln\frac{4(K+1)}{}} \\ &\overset{(80)}{6} \frac{\exp(\ \ 2\ T\ )R^4}{150\ln\frac{4(K+1)}{}}: \end{aligned} \tag{107}$$

Upper bound for $\circ$. From E_{T-1} it follows that

$$\begin{aligned}
\circ &= 2^{2}\sum_{l=0}^{T-1}\exp(-(T-1-t))\,k\,{}_t^{b}k^2 \\
&\overset{(96)}{\leq} 2^{2}\exp(-(T-1))\sum_{t=0}^{T-1}\exp(-t)\left(\frac{64^{4}}{{}_t^{2}}+2b^2\right) \\
&\overset{(84);T\leq K+1}{\leq} \frac{921600^{4}\exp(-(T-1))^{4}\ln^{2}\frac{4(K+1)}{\delta}}{R^2}\sum_{t=0}^{K}\exp\left(2-1+\frac{t}{2}\right)\exp(-t) \\
&\quad +2^{2}\exp(-(T-1))b^2\sum_{t=0}^{K}\exp(-t) \\
&\leq \frac{921600^{4}\exp(-(T-3))^{4}\ln^{2}\frac{4(K+1)}{\delta}(K+1)\exp(2-K)}{R^2} \\
&\quad +2^{2}\exp(-(T-1))b^2\exp(-K)(K+1) \\
&\overset{(80);(81);(83)}{\leq} \frac{1}{5}\exp(-T)R^2. \qquad (108)
\end{aligned}$$

That is, we derived the upper bounds for $\neg;-;\circledR;\bar{}\,;\circ$. More speci cally, the probability event E_{T-1} implies:

$$\begin{gathered}
R_T^2 \overset{(94)}{\leq} 6\exp(-T)R^2+\neg+-+\circledR+\bar{}+\circ; \\
-\overset{(102)}{\leq}6\frac{1}{5}\exp(-T)R^2;\quad \circledR\overset{(103)}{\leq}6\frac{1}{5}\exp(-T)R^2; \\
\circ\overset{(108)}{\leq}6\frac{1}{5}\exp(-T)R^2 \\
\sum_{t=0}^{T-1}{}_t^{2}\overset{(101)}{\leq}6\frac{\exp(-2T)R^4}{150\ln\frac{4(K+1)}{\delta}};\quad \sum_{t=0}^{T-1}e_t^2\overset{(107)}{\leq}6\frac{\exp(-2T)R^4}{150\ln\frac{4(K+1)}{\delta}}.
\end{gathered}$$

Moreover, we also have (see (100), (106) and our induction assumption)

$$\begin{gathered}
\Pr\{E_{T-1}\}>1-\frac{(T-1)}{K+1}; \\
\Pr\{E_\neg\}>1-\frac{}{2(K+1)};\quad \Pr\{E^-\}>1-\frac{}{2(K+1)}.
\end{gathered}$$

where

$$\begin{aligned}
E_\neg &= \left(\text{either } \sum_{t=0}^{T-1}{}_t^{2}>\frac{\exp(-2T)R^4}{150\ln\frac{4(K+1)}{\delta}} \text{ or } |\neg|\leq\frac{1}{5}\exp(-T)R^2\right); \\
E^- &= \left(\text{either } \sum_{t=0}^{T-1}e_t^2>\frac{\exp(-2T)R^4}{150\ln\frac{4(K+1)}{\delta}} \text{ or } |\bar{}|\leq\frac{1}{5}\exp(-T)R^2\right).
\end{aligned}$$

Therefore, probability event $E_{T-1}\setminus E_\neg\setminus E^-$ implies

$$\begin{aligned}
R_T^2 &\overset{(94)}{\leq} 6\exp(-T)R^2+\neg+-+\circledR+\bar{}+\circ \\
&\leq 6\cdot 2\exp(-T)R^2;
\end{aligned}$$

which is equivalent to (87) for $t=T$, and

$$\Pr\{E_T\}>\Pr\{E_{T-1}\setminus E_\neg\setminus E^-\}=1-\Pr\{\bar{E}_{T-1}\cup\bar{E}_\neg\cup\bar{E}^-\}>1-\frac{T}{K+1}.$$

We have now completed the inductive part of our proof. That is, for all $k = 0; 1; : : : ; K + 1$, we have $\mathbb{P}\{E_k\} > 1 - k=(K+1)$. Notably, when $k = K + 1$, we can conclude that with a probability of at least $1 - $:

$$kx^{K+1} - x^* k^2 \leq 2 \exp(- (K+1)) R^2:$$

Finally, if

$$= \min\left(\frac{1}{400L \ln \frac{4(K+1)}{}}; \frac{\ln(B_K)}{(K+1)}; \frac{\ln(C_K)}{(1+K=2)}; \frac{2\ln(D)}{(K+1)} \right);$$

$$B_K = \max\left\{ 2; \frac{(K+1)^2 R^2}{5400^2 \ln \frac{4(K+1)}{} \ln^2(B_K)} \right\} = O\left(\max\left\{ 2; \frac{K^2 R^2}{^2 \ln \frac{K}{} \ln^2 \max\left\{ 2; \frac{K^2 R^2}{^2 \ln\left(\frac{K}{}\right)} \right\}} \right\} \right);$$

$$C_K = \max\left\{ 2; \frac{(\frac{K}{2}+1) R}{480b \ln \frac{4(K+1)}{} \ln(C_K)} \right\} = O\left(\max\left\{ 2; \frac{K R}{b \ln \frac{K}{} \ln \max\left\{ 2; \frac{K R}{b \ln\left(\frac{K}{}\right)} \right\}} \right\} \right);$$

$$D = \max\left\{ 2; \frac{R}{80b \ln(D)} \right\} = O\left(\max\left\{ 2; \frac{R}{b \ln \max\left\{ 2; \frac{R}{b} \right\}} \right\} \right);$$

then with probability at least $1 - $

$$\begin{aligned} kx^{K+1} - x^* k^2 &\leq 2 \exp\left(- \frac{(K+1)}{} \right) R^2 \\ &\leq 2R^2 \max\left\{ \exp\left(- \frac{(K+1)}{400L \ln \frac{4(K+1)}{}} \right); \frac{1}{B_K}; \frac{1}{C_K}; \frac{1}{D} \right\} \\ &= O\left(\max\left\{ R^2 \exp\left(- \frac{K}{L \ln \frac{K}{}} \right); \frac{^2 \ln \frac{K}{} \ln^2(B_K)}{K^2}; \frac{bR \ln \frac{K}{} \ln(C_K)}{K}; \frac{bR \ln(D)}{} \right\} \right): \end{aligned}$$

This concludes the proof.

C Proofs for clipped-SSTM

C.1 Convex Case

The analysis of clipped-SSTM in the convex case relies on the following lemma from (Sadiev et al., 2023).

Lemma C.1 (Lemma F.1 from (Sadiev et al., 2023)) Let Assumptions 2.2 and 2.3 with $\sigma = 0$ hold on $Q = B_{3R}(x^*)$, where $R > \|x^0 - x^*\|$, and let stepsize parameter a satisfy $a > 1$. If $x^k, y^k, z^k \in B_{3R}(x^*)$ for all $k = 0, 1, \ldots, N$, $N > 0$, then after N iterations of clipped-SSTM for all $z \in B_{3R}(x^*)$ we have

$$A_N (f(y^N) - f(z)) \le \frac{1}{2}\|z^0 - z\|^2 - \frac{1}{2}\|z^N - z\|^2 + \sum_{k=0}^{N-1} \langle \tilde{\nabla}_{k+1} - \nabla f(x^{k+1}), z - z^k \rangle + \sum_{k=0}^{N-1} \alpha_{k+1}^2 \|\tilde{\nabla}_{k+1} - \nabla f(x^{k+1})\|^2, \tag{109}$$

$$\tilde{\nabla}_{k+1} \stackrel{\text{def}}{=} \text{clip}\left(\nabla f_{i_k}(x^{k+1}), \lambda_k\right) - \nabla f(x^{k+1}). \tag{110}$$

Next, we also use the following technical result from (Gorbunov et al., 2020).

Lemma C.2 (Lemma E.1 from (Gorbunov et al., 2020)). Let sequences $\{\alpha_k\}_{k>0}$ and $\{A_k\}_{k>0}$ satisfy

$$\alpha_0 = A_0 = 0, \quad A_{k+1} = A_k + \alpha_{k+1}, \quad \alpha_{k+1} = \frac{k+2}{2aL} \quad \forall k > 0, \tag{111}$$

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

$$A_{k+1} = \frac{(k+1)(k+4)}{4aL}, \tag{112}$$

$$A_{k+1} > aL\alpha_{k+1}^2. \tag{113}$$

Theorem C.3. Let Assumptions 2.2 and 2.3 with $\sigma = 0$ hold on $Q = B_{3R}(x^*)$, where $R > \|x^0 - x^*\|$. Assume that $\nabla f_{i_k}(x^{k+1})$ satisfies Assumption 5.1 with parameters b_k, λ_k for $k = 0, 1, \ldots, K$, $K > 0$ and

$$a > \max\left\{97200 \ln^2 \frac{4K}{\beta}; \frac{1800(K+1)}{LR}\sqrt{\frac{p}{K}\ln\frac{4K}{\beta}}; \frac{4b(K+2)^2}{15LR}; \frac{60b(K+2)\ln\frac{4K}{\beta}}{LR}\right\}; \tag{114}$$

$$\lambda_k = \frac{R}{30\alpha_{k+1}\ln\frac{4K}{\beta}}; \tag{115}$$

for some $\beta \in (0, 1]$ and $b = \max_{k=0,1,\ldots,K} b_k$, $\lambda = \max_{k=0,1,\ldots,K} \lambda_k$. Then, after K iterations of clipped-SSTM the iterates with probability at least $1 - \beta$ satisfy

$$f(y^K) - f(x^*) \le \frac{6aLR^2}{K(K+3)} \quad \text{and} \quad \{x^k\}_{k=0}^{K+1}, \{z^k\}_{k=0}^{K}, \{y^k\}_{k=0}^{K} \subset B_{2R}(x^*). \tag{116}$$

In particular, when parameter a equals the maximum from (114), then the iterates produced by clipped-SSTM after K iterations with probability at least $1 - \beta$ satisfy

$$f(y^K) - f(x^*) = O\left(\max\left\{\frac{LR^2\ln^2\frac{K}{\beta}}{K^2}; \frac{R\sqrt{\frac{p}{K}\ln\frac{K}{\beta}}}{K}; \frac{bR\ln\frac{K}{\beta}}{K}; bR\right\}\right). \tag{117}$$

Proof. Our proof follows similar steps to the one given by Sadiev et al. (2023). The main difference comes due to the presence of the bias in $\nabla f_{i_k}(x^k)$. Therefore, for completeness, we provide the full proof here.

Let $R_k = \|z^k - x^*\|$, $\hat{R}_0 = R_0$, and $\hat{R}_{k+1} = \max\{\hat{R}_k, R_{k+1}\}$ for all $k > 0$. We will initially demonstrate through induction that for all $k > 0$, the iterates x^{k+1}, z^k, y^k belong to $B_{\hat{R}_k}(x^*)$. The base of the induction is straightforward because $y^0 = z^0$, $\hat{R}_0 = R_0$, and $x^1 = \frac{A_0 y^0 + \alpha_1 z^0}{A_1} = z^0$. Now, assume that for some $l > 1$,

$x^l; z^{l-1}; y^{l-1} \in B_{R_{l-1}}(x)$. By the definitions of R_l and $\hat{R}_l$, we have $z^l \in B_{R_l}(x) \subseteq B_{\hat{R}_l}(x)$. As y^l is a convex combination of $y^{l-1} \in B_{\hat{R}_{l-1}}(x) \subseteq B_{\hat{R}_l}(x)$, it follows that $z^l \in B_{\hat{R}_l}(x)$ and, given the convex nature of $B_{\hat{R}_l}(x)$, we can conclude that $y^l \in B_{\hat{R}_l}(x)$. Finally, as x^{l+1} is a convex combination of y^l and z^l, it is evident that x^{l+1} also lies in $B_{\hat{R}_l}(x)$.

Our next objective is to establish, by induction, that $\hat{R}_l \le 3R$ with high probability. This will enable us to apply the result from Lemma B.3 and subsequently utilize Bernstein's inequality to estimate the stochastic component of the upper bound. To be more precise, for each $k = 0; \dots; K+1$, we consider the probability event E_k, defined as follows: inequalities

$$\sum_{l=0}^{k-1} \alpha_{l+1}\langle \nabla f(x^{l+1}), z^l - x\rangle + \sum_{l=0}^{k-1} \alpha_{l+1}^2 \|\nabla f(x^{l+1})\|^2 \le R^2; \tag{118}$$

$$R_t \le 2R \tag{119}$$

hold for all $t = 0; 1; \dots; k$ simultaneously. We aim to demonstrate through induction that $\mathbb{P}\{E_k\} \ge 1 - k/(K+1)$ for all $k = 0; 1; \dots; K+1$. The base case, $k = 0$, is trivial: the left-hand side of (118) equals zero and $R > R_0$ by definition. Assuming that the statement holds for some $k = T-1 \le K$, specifically, $\mathbb{P}\{E_{T-1}\} \ge 1 - (T-1)/(K+1)$, we need to establish that $\mathbb{P}\{E_T\} \ge 1 - T/(K+1)$.

To begin, we observe that the probability event E_{T-1} implies that $\hat{R}_t \le 2R$ for all $t = 0; 1; \dots; T-1$. Moreover, it implies that

$$\|z^T - x\| \overset{(14)}{\le} \|z^T - x\| + \alpha_T \|\nabla f(x^T)\| \le 2R + \alpha_T \overset{(115)}{\le} 3R.$$

Hence, with E_{T-1} implying $\{x^k\}_{k=0}^T \subseteq Q$, we confirm that the conditions of Lemma C.1 are met, resulting in

$$A_t\,(f(y^t) - f(x)) \le \frac{1}{2}R_0^2 - \frac{1}{2}R_t^2 + \sum_{l=0}^{t-1} \alpha_{l+1}\langle \nabla f(x^{l+1}), z^l - x\rangle + \sum_{l=0}^{t-1} \alpha_{l+1}^2 \|\nabla f(x^{l+1})\|^2 \tag{120}$$

for all $t = 0; 1; \dots; T$ simultaneously and for all $t = 1; \dots; T-1$ this probability event also implies that

$$f(y^t) - f(x) \overset{(118);(120)}{\le} \frac{\frac{1}{2}R_0^2 - \frac{1}{2}R_t^2 + R^2}{A_t} \le \frac{3R^2}{2A_t} = \frac{6aLR^2}{t(t+3)}. \tag{121}$$

Considering that $f(y^T) - f(x) > 0$, we can further deduce from (120) that when E_{T-1} holds, the following holds as well:

$$\begin{aligned} R_T^2 &\le R_0^2 + 2\underbrace{\sum_{t=0}^{T-1} \alpha_{t+1}\langle \nabla f(x^{t+1}), z^t - x\rangle}_{2B_T} + 2\sum_{t=0}^{T-1} \alpha_{t+1}^2 \|\nabla f(x^{t+1})\|^2 \\ &\le R^2 + 2B_T. \end{aligned} \tag{122}$$

Prior to our estimation of B_T, we need to establish several helpful inequalities. We start with showing that E_{T-1} implies $\|\nabla f(x^{t+1})\| \le 4$ for all $t = 0; 1; \dots; T-1$. For $t = 0$ we have $x^1 = x^0$ and

$$\|\nabla f(x^1)\| = \|\nabla f(x^0)\| \overset{(3)}{\le} L\|x^0 - x^*\| \le \frac{R}{a\alpha_1} = \frac{\alpha_0}{2} \le \frac{60\ln\frac{4K}{a}}{a} \overset{(114)}{\le} \frac{\alpha_0}{4}. \tag{123}$$

Next, for $t = 1;\dots;T-1$ we have $\psi_{t+1}(x^{t+1}-z^t) = A_t(y^t-x^{t+1})$ and event E_{T-1} implies

$$\begin{aligned}
\|\nabla f(x^{t+1})\| &\le \|\nabla f(x^{t+1}) - \nabla f(y^t)\| + \|\nabla f(y^t)\| \\
&\overset{(3);(4)}{\le} L\|x^{t+1}-y^t\| + \sqrt{2L(f(y^t)-f(x^*))} \\
&\overset{(121)}{\le} \frac{L\psi_{t+1}}{A_t}\|x^{t+1}-z^t\| + \sqrt{\frac{12aL^2R^2}{t(t+3)}} \\
&\le \frac{4LR\psi_{t+1}}{A_t} + \sqrt{\frac{12aL^2R^2}{t(t+3)}} \\
&= \frac{R}{60\psi_{t+1}\ln\frac{4K}{\beta}}\left(\frac{240L^2\psi_{t+1}^2\ln\frac{4K}{\beta}}{A_t} + 60\sqrt{\frac{12aL^2\psi_{t+1}^2\ln^2\frac{4K}{\beta}}{t(t+3)}}\right) \\
&\overset{(112);(115)}{\le} \frac{t}{2}\left(\frac{240L^2\frac{(t+2)^2}{2aL}\ln\frac{4K}{\beta}}{\frac{t(t+3)}{4aL}} + 60\sqrt{\frac{12aL^2\frac{(t+2)^2}{2aL}\ln^2\frac{4K}{\beta}}{t(t+3)}}\right) \\
&= \frac{t}{2}\left(\frac{240(t+2)^2\ln\frac{4K}{\beta}}{t(t+3)a} + 60\sqrt{\frac{3(t+2)^2\ln\frac{4K}{\beta}}{t(t+3)a}}\right) \\
&\le \frac{t}{2}\left(\frac{540\ln\frac{4K}{\beta}}{a} + \frac{90\sqrt{3}\ln\frac{4K}{\beta}}{\sqrt{a}}\right) \overset{(114)}{\le} \frac{t}{4};
\end{aligned} \tag{124}$$

where in the last row we use $\frac{(t+2)^2}{t(t+3)} \le \frac{9}{4}$ for all $t > 1$. Therefore, probability event E_{T-1} implies that

$$\mathbb{E}_t[\|\nabla f_t(x^{t+1})\|] \le \mathbb{E}_t[\|\nabla f_t(x^{t+1})\| - \|\nabla f(x^{t+1})\| + \|\nabla f(x^{t+1})\|] \le b + \frac{t}{4} \le \frac{t}{2} \tag{125}$$

and

$$\|x - z^t + \psi_{t+1}\nabla f(x^{t+1})\| \le \|x - z^t\| + \psi_{t+1}\|\nabla f(x^{t+1})\| \overset{(119);(123);(124)}{\le} 2R + \frac{R}{60\ln\frac{4K}{\beta}} \le 3R \tag{126}$$

for all $t = 0;1;\dots;T-1$. Next, we define random vectors

$$\eta_t = \begin{cases} x - z^t + \psi_{t+1}\nabla f(x^{t+1}); & \text{if } \|x - z^t + \psi_{t+1}\nabla f(x^{t+1})\| \le 3R; \\ 0; & \text{otherwise;} \end{cases}$$

for all $t = 0;1;\dots;T-1$. As per their definition, these random vectors are bounded with probability 1

$$\|\eta_t\| \le 3R. \tag{127}$$

This means that E_{T-1} implies $\eta_t = x - z^t + \psi_{t+1}\nabla f(x^{t+1})$ for all $t = 0;1;\dots;T-1$. Then, form E_{T-1} it follows that

$$B_T = \sum_{t=0}^{T-1}\psi_{t+1}\langle\nabla f(x^{t+1}), \eta_t\rangle + \sum_{t=0}^{T-1}\psi_{t+1}^2\|\nabla f(x^{t+1})\|^2.$$

Next, we define the unbiased part and the bias of η_t as η_t^u and η_t^b, respectively:

$$\eta_t^u = \operatorname{clip}(\nabla f_t(x^{t+1}); \lambda_t) - \mathbb{E}_t\operatorname{clip}(\nabla f_t(x^{t+1}); \lambda_t); \quad \eta_t^b = \mathbb{E}_t\operatorname{clip}(\nabla f_t(x^{t+1}); \lambda_t) - \nabla f(x^{t+1}). \tag{128}$$

We notice that $\xi_t = \xi_t^u + \xi_t^b$. Using new notation, we get that E_{T-1} implies

$$B_T = \sum_{t=0}^{T-1} \left\langle \xi_{t+1}, \xi_{t+1}^u + \xi_{t+1}^b \right\rangle + \sum_{t=0}^{T-1} \left\| \xi_{t+1}^u + \xi_{t+1}^b \right\|^2$$
$$\leq \underbrace{\sum_{t=0}^{T-1} \left\langle \xi_{t+1}, \xi_{t+1}^u \right\rangle}_{\neg} + \underbrace{\sum_{t=0}^{T-1} \left\langle \xi_{t+1}, \xi_{t+1}^b \right\rangle}_{-} + 2 \underbrace{\sum_{t=0}^{T-1} \left\| \xi_{t+1}^u \right\|^2}_{®} + 2 \sum_{t=0}^{T-1} \left\| \xi_{t+1}^u \right\|^2 \mathbb{E}_t \left[\left\| \xi_{t+1}^u \right\|^2 \right]$$
$$+ 2 \underbrace{\sum_{t=0}^{T-1} \left\| \xi_{t+1}^u \right\|^2}_{\bar{}} + 2 \underbrace{\sum_{t=0}^{T-1} \left\| \xi_{t+1}^b \right\|^2}_{\circ}. \tag{129}$$

To conclude our inductive proof successfully, we must obtain sufficiently strong upper bounds with high probability for the terms $\neg; -; ®; \bar{}; \circ$. In other words, we need to demonstrate that $\neg + - + ® + \bar{} + \circ \leq R^2$ with a high probability. In the subsequent stages of the proof, we will rely on the bounds for the norms and second moments of ξ_t^u and ξ_t^b. First, as per the definition of the clipping operator, we can assert with probability 1 that

$$\|\xi_{t+1}^u\| \leq 2\lambda_t. \tag{130}$$

Moreover, since E_{T-1} implies that $\|\mathbb{E}_t[\nabla f_t(x^{t+1})]\| \leq \lambda_t/2$ for $t = 0, 1, \ldots, T-1$ (see (125)), then, according to Lemma B.2, we can deduce that E_{T-1} implies

$$\|\xi_{t+1}^b\| \leq \left\| \mathbb{E}_t\left[\text{clip}(\nabla f_t(x^{t+1}), \lambda_t)\right] - \mathbb{E}_t[\nabla f_t(x^{t+1})] \right\| + \left\| \mathbb{E}_t[\nabla f_t(x^{t+1})] - \nabla f(x^{t+1}) \right\|$$
$$\leq \frac{4\sigma^2}{\lambda_t} + b, \tag{131}$$

$$\mathbb{E}_t\left[\|\xi_{t+1}^u\|^2\right] \leq 18\sigma^2. \tag{132}$$

Upper bound for $\neg$. By definition of ξ_t^u, we readily observe that $\mathbb{E}_t[\xi_t^u] = 0$ and

$$\mathbb{E}_t\left[\langle \xi_{t+1}, \xi_{t+1}^u \rangle\right] = 0.$$

Next, the sum $\neg$ contains only the terms that are bounded with probability 1:

$$|\langle \xi_{t+1}, \xi_{t+1}^u \rangle| \leq \|\xi_{t+1}\| \|\xi_{t+1}^u\| \leq \|\xi_t\| \overset{(127),(130)}{\leq} 2\lambda_{t+1}\lambda_t \overset{(115)}{=} \frac{R^2}{5\ln\frac{4K}{\beta}} \overset{\text{def}}{=} c. \tag{133}$$

The conditional variances $\sigma_t^2 \overset{\text{def}}{=} \mathbb{E}_t[\langle \xi_{t+1}, \xi_{t+1}^u \rangle^2]$ of the summands are bounded:

$$\sigma_t^2 \leq \mathbb{E}_t\left[\|\xi_{t+1}\|^2 \|\xi_{t+1}^u\|^2\right] \leq \|\xi_t\|^2 \overset{(127)}{\leq} 9\lambda_{t+1}^2 R^2 \mathbb{E}_t\left[\|\xi_{t+1}^u\|^2\right]. \tag{134}$$

To summarize, we have demonstrated that $\{\langle \xi_{t+1}, \xi_{t+1}^u \rangle\}_{t=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\sigma_t^2\}_{t=0}^{T-1}$. Therefore, one can apply Bernstein's inequality (Lemma B.1) with $X_t = \langle \xi_{t+1}, \xi_{t+1}^u \rangle$, parameter c as in (133), $b = \frac{R^2}{5}$, $G = \frac{R^4}{150\ln\frac{4K}{\beta}}$ and get

$$P\left(|\neg| > \frac{R^2}{5} \text{ and } \sum_{t=0}^{T-1} \sigma_t^2 \leq \frac{R^4}{150\ln\frac{4K}{\beta}}\right) \leq 2\exp\left(-\frac{b^2}{2G + \frac{2cb}{3}}\right) = \frac{\beta}{2K},$$

which is equivalent to

$$P\{E_\neg\} > 1 - \frac{\beta}{2K}, \quad \text{for } E_\neg = \left(\text{either } \sum_{t=0}^{T-1} \sigma_t^2 > \frac{R^4}{150\ln\frac{4K}{\beta}} \text{ or } |\neg| \leq \frac{R^2}{5}\right). \tag{135}$$

In addition, E_{T-1} implies that

$$\sum_{t=0}^{T-1} \alpha_t^2 \overset{(134)}{\leq} 9R^2 \sum_{t=0}^{T-1} \alpha_{t+1}^2 \|\nabla f(x_{t+1}) - u_{t+1}\|^2 \overset{(132)}{\leq} 162\,\alpha^2 R^2 \sum_{t=0}^{T-1} \alpha_{t+1}^2$$
$$= \frac{162\,\alpha^2 R^2}{4a^2L^2} \sum_{t=0}^{T-1} (t+2)^2$$
$$\leq \frac{1}{a^2} \cdot \frac{81\,\alpha^2 R^2 T(T+1)^2}{2a^2L^2} \overset{(114)}{\leq} \frac{R^4}{150 \ln \frac{4K}{\beta}}. \tag{136}$$

Upper bound for Φ. From E_{T-1} it follows that

$$\Phi \leq \sum_{t=0}^{T-1} \|x_{t+1} - x_b\| \|x_{t+1} - x_t\| \overset{(127),(131)}{\leq} 3R \sum_{t=0}^{T-1} \alpha_{t+1} \left(\frac{4}{t} + b\right)$$
$$\overset{(115)}{\leq} 360\,\alpha^2 \ln \frac{4K}{\beta} \sum_{t=0}^{T-1} \alpha_{t+1}^2 + 3bR \sum_{t=0}^{T-1} \alpha_{t+1}$$
$$\leq \frac{360\,\alpha^2 \ln \frac{4K}{\beta}}{4a^2L^2} \sum_{t=0}^{T-1} (t+2)^2 + \frac{3bR}{2aL} \sum_{t=0}^{T-1} (t+2)$$
$$\leq \frac{360\,\alpha^2 \ln \frac{4K}{\beta} T(T+1)^2}{4a^2L^2} + \frac{3bRT(T+1)}{2aL} \overset{(114)}{\leq} \frac{R^2}{5}. \tag{137}$$

Upper bound for Ψ. First, we have

$$E_t\left[2\alpha_{t+1}^2 \langle u_{t+1}, \nabla f(x_{t+1}) - u_{t+1} \rangle\right] = 0.$$

Next, the sum Ψ contains only the terms that are bounded with probability 1:

$$2\alpha_{t+1}^2 \|u_{t+1}\| \|\nabla f(x_{t+1}) - u_{t+1}\| \leq 2\alpha_{t+1}^2 \|\nabla f(x_{t+1}) - u_{t+1}\|^2 \leq 2\alpha_{t+1}^2 \|\nabla f(x_{t+1}) - \nabla f(x_t)\|^2 + 2\alpha_{t+1}^2 \|u_{t+1}\|^2$$
$$\overset{(130)}{\leq} 16\,\alpha_{t+1}^2 \alpha_t^2 \overset{(115)}{\leq} \frac{R^2}{5 \ln \frac{4K}{\beta}} \overset{\text{def}}{=} c. \tag{138}$$

The conditional variances $\sigma_t^2 \overset{\text{def}}{=} E_t\left[4\alpha_{t+1}^4 \|u_{t+1} - \nabla f(x_{t+1})\|^2\right] \leq E_t\left[\|u_{t+1} - \nabla f(x_{t+1})\|^2\right]$ of the summands are bounded:

$$\sigma_t^2 \overset{(138)}{\leq} \frac{R^2}{5 \ln \frac{4K}{\beta}} E_t\left[2\alpha_{t+1}^2 \|u_{t+1} - \nabla f(x_{t+1})\|^2\right] \leq E_k\left[\|u_{t+1} - \nabla f(x_{t+1})\|^2\right] \leq 2\alpha_{t+1}^2 R^2 E_t\left[\|\nabla f(x_{t+1}) - u_{t+1}\|^2\right]; \tag{139}$$

To summarize, we have demonstrated that $\left\{2\alpha_{t+1}^2 \|u_{t+1} - \nabla f(x_{t+1})\|^2 - E_t\left[\|u_{t+1} - \nabla f(x_{t+1})\|^2\right]\right\}_{t=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\sigma_t^2\}_{t=0}^{T-1}$. Therefore, one can apply Bernstein's inequality (Lemma B.1) with $X_t = 2\alpha_{t+1}^2 \|u_{t+1} - \nabla f(x_{t+1})\|^2 - E_t\left[\|u_{t+1} - \nabla f(x_{t+1})\|^2\right]$, parameter c as in (138), $b = \frac{R^2}{5}$, $G = \frac{R^4}{150 \ln \frac{4K}{\beta}}$ and get

$$P\left\{|\Psi| > \frac{R^2}{5} \text{ and } \sum_{t=0}^{T-1} \sigma_t^2 \leq \frac{R^4}{150 \ln \frac{4K}{\beta}}\right\} \leq 2\exp\left(-\frac{b^2}{2G + \frac{2cb}{3}}\right) = \frac{\beta}{2K};$$

which is equivalent to

$$P\{E_\Psi\} \geq 1 - \frac{\beta}{2K}; \quad \text{for } E_\Psi = \left\{\text{either } \sum_{t=0}^{T-1} \sigma_t^2 > \frac{R^4}{150 \ln \frac{4K}{\beta}} \text{ or } |\Psi| \leq \frac{R^2}{5}\right\}. \tag{140}$$

In addition, E_{T-1} implies that

$$\sum_{t=0}^{T-1} \sigma_t^2 \overset{(139)}{\leq} R^2 \sum_{t=0}^{T-1} \alpha_{t+1}^2 \|\nabla f(x_{t+1}) - u_{t+1}\|^2 \leq 9R^2 \sum_{t=0}^{T-1} \alpha_{t+1}^2 \|\nabla f(x_{t+1}) - u_{t+1}\|^2 \overset{(136)}{\leq} \frac{R^4}{150 \ln \frac{4K}{\beta}}. \tag{141}$$

Upper bound for $\bar{}$. From $E_{T\ 1}$ it follows that

$$\bar{} = 2\sum_{t=0}^{} {}^{2}_{t+1}E_{t}\left[{}^{u}_{t+1}{}^{2}\right] 6 \frac{1}{R^2} 9R^2\sum_{t=0}^{} {}^{2}_{t+1}E_{t}\left[{}^{u}_{t+1}{}^{2}\right] \overset{(136)}{6} \frac{R^2}{150\ln\frac{4K}{}} 6 \frac{R^2}{5}: \tag{142}$$

Upper bound for $^{\circ}$. From $E_{T\ 1}$ it follows that

$$\begin{aligned}
^{\circ} &= 2\sum_{t=0}^{} {}^{2}_{t+1}\ {}^{b}_{t+1}{}^{2} 6\ 4\sum_{t=0}^{} {}^{2}_{t+1}\ \frac{16\ {}^{4}}{{}^{2}_{t}} + b^2 \\
&\overset{(115)}{=} \frac{57600\ {}^{4}\ln^2\frac{4K}{}}{R^2}\sum_{t=0}^{} {}^{4}_{t+1} + 4b^2\sum_{t=0}^{} {}^{2}_{t+1} = \frac{57600\ {}^{4}\ln^2\frac{4K}{}}{16a^4L^4R^2}\sum_{t=0}^{}(t+2)^4 + \frac{4b^2}{4a^2L^2}\sum_{t=0}^{}(t+2)^2 \\
&6 \frac{57600\ {}^{4}\ln^2\frac{4K}{}T(T+1)^4}{16a^4L^4R^2} + \frac{4b^2T(T+1)^2}{4a^2L^2} \overset{(114)}{6} \frac{R^2}{5}:
\end{aligned} \tag{143}$$

That is, we derived the upper bounds for $\neg\ ;\ -\ ;\ ®;\ \bar{}\ ;\ ^{\circ}$. More speci cally, the probability event $E_{T\ 1}$ implies:

$$\begin{gathered}
B_T \overset{(129)}{6} R^2 + \neg + - + ® + \bar{} + {}^{\circ}; \\
- \overset{(137)}{6} \frac{R^2}{5}; \quad \bar{} \overset{(142)}{6} \frac{R^2}{5}; \quad {}^{\circ} \overset{(143)}{6} \frac{R^2}{5}; \\
\sum_{t=0}^{} {}^{2}_{t} \overset{(136)}{6} \frac{R^4}{150\ln\frac{4K}{}}; \quad \sum_{t=0}^{} e_t^2 \overset{(141)}{6} \frac{R^4}{150\ln\frac{4K}{}}:
\end{gathered}$$

In addition, we also have (see (135), (140) and our induction assumption)

$$\mathrm{Pf}\,E_{T\ 1}g > 1 \quad \frac{(T\ 1)}{K}; \quad \mathrm{Pf}\,E_{\neg}g > 1 \quad \frac{}{2K}; \quad \mathrm{Pf}\,E_{®}g > 1 \quad \frac{}{2K};$$

where

$$\begin{aligned}
E_{\neg} &= \left(\text{either} \sum_{t=0}^{} {}^{2}_{t} > \frac{R^4}{150\ln\frac{4K}{}} \quad \text{or} \quad j\neg j\ 6\ \frac{R^2}{5} \right); \\
E_{®} &= \left(\text{either} \sum_{t=0}^{} e_t^2 > \frac{R^4}{150\ln\frac{4K}{}} \quad \text{or} \quad j®j\ 6\ \frac{R^2}{5} \right):
\end{aligned}$$

Therefore, probability event $E_{T\ 1} \setminus E_{\neg} \setminus E_{®}$ implies

$$\begin{aligned}
B_T \quad 6 \quad & R^2 + \frac{R^2}{5} + \frac{R^2}{5} + \frac{R^2}{5} + \frac{R^2}{5} + \frac{R^2}{5} = 2R^2; \\
R_T^2 \overset{(122)}{6} \quad & R^2 + 2R^2\ 6\ (2R)^2;
\end{aligned}$$

which is equivalent to (118) and (119) for $t = T$, and

$$\mathrm{Pf}\,E_T g > \mathrm{Pf}\,E_{T\ 1} \setminus E_{\neg} \setminus E_{®}g = 1 \quad P\ \bar{E}_{T\ 1}[\ \bar{E}_{\neg}[\ \bar{E}_{®}\ > 1 \quad \mathrm{Pf}\,\bar{E}_{T\ 1}g \quad \mathrm{Pf}\,\bar{E}_{\neg}g \quad \mathrm{Pf}\,\bar{E}_{®}g > 1 \quad \frac{T}{K}:$$

We have now completed the inductive part of our proof. That is, for all $k = 0; 1; : : : ; K$, we have $\mathrm{Pf}\,E_k g > 1 \quad {}^{k} \ni {}^{K}$. Notably, when $k = K$, we can conclude that with a probability of at least $1 \quad :$

$$f(y^K) \quad f(x^{\ }) \overset{(121)}{6} \frac{6aLR^2}{K(K\ +3)}$$

and $f x^k g_{k=0}^{K+1}; f z^k g_{k=0}^{K}; f y^k g_{k=0}^{K} \quad B_{2R}(x^{\ })$, which follows from (119).

Finally, if

$$a = \max\left\{97200\ln^2\frac{4K}{\delta}, \frac{1800\sigma(K+1)\sqrt{K}\sqrt{\ln\frac{4K}{\delta}}}{LR}, \frac{4b(K+2)^2}{15LR}, \frac{60b(K+2)\ln\frac{4K}{\delta}}{LR}\right\},$$

then with probability at least $1-\delta$

$$\begin{aligned} f(y^K) - f(x^*) &\leqslant \frac{6aLR^2}{K(K+3)} \\ &= \mathcal{O}\left(\max\left\{\frac{LR^2\ln^2\frac{K}{\delta}}{K^2}, \frac{\sigma R\sqrt{\ln\frac{K}{\delta}}}{\sqrt{K}}, \frac{bR\ln\frac{K}{\delta}}{K}, bR\right\}\right). \end{aligned}$$

□

C.2 Strongly Convex Case

In the strongly convex case, we consider a restarted version of SSTM, see Algorithm 1.

Algorithm 1 Restarted clipped-SSTM (R-clipped-SSTM) (Gorbunov et al., 2020)

Input: starting point x^0, number of restarts τ, number of steps of clipped-SSTM between restarts $\{K_t\}_{t=1}^{\tau}$, stepsize parameters $\{a_t\}_{t=1}^{\tau}$, clipping levels $\{\lambda_k^1\}_{k=0}^{K_1-1}$, $\{\lambda_k^2\}_{k=0}^{K_2-1}$, $\ldots$, $\{\lambda_k^\tau\}_{k=0}^{K_\tau-1}$, smoothness constant L.

1: $\hat{x}^0 = x^0$
2: **for** $t = 1, \ldots, \tau$ **do**
3: Run clipped-SSTM for K_t iterations with stepsize parameter a_t, clipping levels $\{\lambda_k^t\}_{k=0}^{K_t-1}$, and starting point $\hat{x}^{t-1}$. Define the output of clipped-SSTM by $\hat{x}^t$.
4: **end for**

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

The main result for R-clipped-SSTM is given below.

Theorem C.4. *Let Assumptions 2.2 and 2.3 with $\mu > 0$ hold on $Q = B_{3R}(x^*)$, where $R \geqslant \|x^0 - x^*\|$ and* R-clipped-SSTM *runs* clipped-SSTM *τ times. Assume that estimator $\nabla f_{i_k,t}(x^{k+1,t})$ used in* clipped-SSTM *at k-th iteration of stage t satisfies Assumption 5.1 with parameters b_k^t and σ_k^t such that*

$$b_k^t \leqslant \frac{15\mu R}{24\cdot 2^{t+1}}. \tag{144}$$

Let

$$K_t = \left\lceil \max\left\{2160\sqrt{\frac{LR_t^2}{\varepsilon_t}}\ln\frac{4320\sqrt{LR_t^2}\tau}{\sqrt{\varepsilon_t}\delta}, 4\left(\frac{5400\sigma^t R_t}{\varepsilon_t}\right)^2\ln\left(\frac{8\tau}{\delta}\left(\frac{5400\sigma^t R_t}{\varepsilon_t}\right)^2\right)\right\}\right\rceil, \tag{145}$$

$$\varepsilon_t = \frac{\mu R_t^2}{4}, \quad R_t = \frac{R}{2^{(t-1)/2}}, \quad \tau = \left\lceil \log_2\frac{\mu R^2}{2\varepsilon}\right\rceil, \tag{146}$$

$$a_t = \max\left\{97200\ln^2\frac{4K_t\tau}{\delta}, \frac{1800\sigma(K_t+1)\sqrt{K_t}\sqrt{\ln\frac{4K_t\tau}{\delta}}}{LR_t}, \frac{4b_t(K_t+2)^2}{15LR_t}, \frac{60b_t(K_t+2)\ln\frac{4K_t}{\delta}}{LR_t}\right\}, \tag{147}$$

$$\lambda_k^t = \frac{R_t}{30\alpha_{k+1}^t\ln\frac{4K_t\tau}{\delta}} \tag{148}$$

for $t = 1, \ldots, \tau$. Then to guarantee $f(\hat{x}^\tau) - f(x^*) \leqslant \varepsilon$ with probability $\geqslant 1-\delta$ R-clipped-SSTM *requires*

$$\mathcal{O}\left(\max\left\{\sqrt{\frac{L}{\mu}}\ln\left(\frac{\mu R^2}{\varepsilon}\right)\ln\left(\frac{\sqrt{L}}{\sqrt{\mu}\delta}\ln\left(\frac{\mu R^2}{\varepsilon}\right)\right), \frac{\sigma^2}{\mu\varepsilon}\ln\left(\frac{\sigma^2}{\mu\varepsilon\delta}\ln\left(\frac{\mu R^2}{\varepsilon}\right)\right)\right\}\right) \tag{149}$$

iterations. Moreover, with probability $\geqslant 1-\delta$ the iterates of R-clipped-SSTM *at stage t stay in the ball $B_{2R_{t-1}}(x^*)$.*

Proof. The proof of this theorem follows the same steps as the one given for Theorem F.3 from (Sadiev et al., 2023). By induction we derive that for any $t = 1, \ldots, \tau$ with probability at least $1 - t\delta/\tau$ inequalities

$$f(\hat{x}^l) - f(x^*) \leqslant \varepsilon_l, \quad \|\hat{x}^l - x^*\|^2 \leqslant R_l^2 = \frac{R^2}{2^l} \tag{150}$$

hold for $l = 1, \ldots, t$ simultaneously. We start with the base of the induction. Theorem C.3 implies that with probability at least $1 - \delta/\tau$

$$\begin{aligned}
f(\hat{x}^1) - f(x^*) &\leqslant \frac{6a_1 L R^2}{K_1(K_1+3)} \\
&\overset{(147)}{=} \max\left\{\frac{583200 L R^2 \ln^2 \frac{4K_1\tau}{\delta}}{K_1(K_1+3)}, \frac{10800\sigma R(K_1+1)\sqrt{K_1 \ln \frac{4K_1\tau}{\delta}}}{K_1(K_1+3)}, \right. \\
&\qquad \left. \frac{24 b_1 R (K_1+2)^2}{15 K_1(K_1+3)}, \frac{360 b_1 R (K_1+2)\ln \frac{4K_1}{\delta}}{K_1(K_1+3)}\right\} \\
&\leqslant \max\left\{\frac{583200 L R^2 \ln^2 \frac{4K_1\tau}{\delta}}{K_1^2}, \frac{10800\sigma R\sqrt{\ln \frac{4K_1\tau}{\delta}}}{\sqrt{K_1}}, \right. \\
&\qquad \left. \frac{24 b_1 R}{15}, \frac{360 b_1 R \ln \frac{4K_1}{\delta}}{K_1}\right\} \\
&\overset{(144),(145)}{\leqslant} \varepsilon_1 = \frac{\mu R^2}{4}
\end{aligned}$$

and, due to the strong convexity,

$$\|\hat{x}^1 - x^*\|^2 \leqslant \frac{2(f(\hat{x}^1) - f(x^*))}{\mu} \leqslant \frac{R^2}{2} = R_1^2.$$

The base of the induction is proven. Now, assume that the statement holds for some $t = T < \tau$, i.e., with probability at least $1 - T\delta/\tau$ inequalities

$$f(\hat{x}^l) - f(x^*) \leqslant \varepsilon_l, \quad \|\hat{x}^l - x^*\|^2 \leqslant R_l^2 = \frac{R^2}{2^l} \tag{151}$$

hold for $l = 1, \ldots, T$ simultaneously. In particular, with probability at least $1 - T\delta/\tau$ we have $\|\hat{x}^T - x^*\|^2 \leqslant R_T^2$. Applying Theorem C.3 and using union bound for probability events, we get that with probability at least $1 - (T+1)\delta/\tau$

$$\begin{aligned}
f(\hat{x}^{T+1}) - f(x^*) &\leqslant \frac{6a_{T+1} L R_T^2}{K_{T+1}(K_{T+1}+3)} \\
&\overset{(147)}{=} \max\left\{\frac{583200 L R_T^2 \ln^2 \frac{4K_{T+1}\tau}{\delta}}{K_{T+1}(K_{T+1}+3)}, \frac{10800\sigma R_T(K_{T+1}+1)\sqrt{K_{T+1} \ln \frac{4K_{T+1}\tau}{\delta}}}{K_{T+1}(K_{T+1}+3)}, \right. \\
&\qquad \left. \frac{24 b_{T+1} R_T (K_{T+1}+2)^2}{15 K_{T+1}(K_{T+1}+3)}, \frac{360 b_{T+1} R_T (K_{T+1}+2)\ln \frac{4K_{T+1}}{\delta}}{K_{T+1}(K_{T+1}+3)}\right\} \\
&\leqslant \max\left\{\frac{583200 L R_T^2 \ln^2 \frac{4K_{T+1}\tau}{\delta}}{K_{T+1}^2}, \frac{10800\sigma R_T\sqrt{\ln \frac{4K_{T+1}\tau}{\delta}}}{\sqrt{K_{T+1}}}, \right. \\
&\qquad \left. \frac{24 b_{T+1} R_T}{15}, \frac{360 b_{T+1} R_T \ln \frac{4K_{T+1}}{\delta}}{K_{T+1}}\right\} \\
&\overset{(144),(145)}{\leqslant} \varepsilon_{T+1} = \frac{\mu R_T^2}{4}
\end{aligned}$$

and, due to the strong convexity,

$$\|\hat{x}^{T+1} - x^*\|^2 \leqslant \frac{2(f(\hat{x}^{T+1}) - f(x^*))}{\mu} \leqslant \frac{R_T^2}{2} = R_{T+1}^2.$$

Thus, we finished the inductive part of the proof. In particular, with probability at least $1-\delta$ inequalities

$$f(\hat{x}^l) - f(x^*) \leqslant \varepsilon_l, \quad \|\hat{x}^l - x^*\|^2 \leqslant R_l^2 = \frac{R^2}{2^l}$$

hold for $l = 1, \ldots, \tau$ simultaneously, which gives for $l = \tau$ that with probability at least $1-\delta$

$$f(\hat{x}^\tau) - f(x^*) \leqslant \varepsilon_\tau = \frac{\mu R_{\tau-1}^2}{4} = \frac{\mu R^2}{2^{\tau+1}} \overset{(146)}{\leqslant} \varepsilon.$$

It remains to calculate the overall number of iterations during all runs of clipped-SSTM. We have

$$\begin{aligned}
\sum_{t=1}^{\tau} K_t &= \mathcal{O}\left(\sum_{t=1}^{\tau} \max\left\{\sqrt{\frac{LR_{t-1}^2}{\varepsilon_t}} \ln\left(\frac{\sqrt{LR_{t-1}^2}\tau}{\sqrt{\varepsilon_t}\delta}\right), \left(\frac{\sigma R_{t-1}}{\varepsilon_t}\right)^2 \ln\left(\frac{\tau}{\delta}\left(\frac{\sigma R_{t-1}}{\varepsilon_t}\right)^2\right)\right\}\right) \\
&= \mathcal{O}\left(\sum_{t=1}^{\tau} \max\left\{\sqrt{\frac{L}{\mu}} \ln\left(\frac{\sqrt{L}\tau}{\sqrt{\mu}\delta}\right), \left(\frac{\sigma}{\mu R_{t-1}}\right)^2 \ln\left(\frac{\tau}{\delta}\left(\frac{\sigma}{\mu R_{t-1}}\right)^2\right)\right\}\right) \\
&= \mathcal{O}\left(\max\left\{\tau\sqrt{\frac{L}{\mu}} \ln\left(\frac{\sqrt{L}\tau}{\sqrt{\mu}\delta}\right), \sum_{t=1}^{\tau}\left(\frac{\sigma \cdot 2^{t/2}}{\mu R}\right)^2 \ln\left(\frac{\tau}{\delta}\left(\frac{\sigma \cdot 2^{t/2}}{\mu R}\right)^2\right)\right\}\right) \\
&= \mathcal{O}\left(\max\left\{\sqrt{\frac{L}{\mu}} \ln\left(\frac{\mu R^2}{\varepsilon}\right) \ln\left(\frac{\sqrt{L}}{\sqrt{\mu}\delta} \ln\left(\frac{\mu R^2}{\varepsilon}\right)\right), \left(\frac{\sigma}{\mu R}\right)^2 \ln\left(\frac{\tau}{\delta}\left(\frac{\sigma \cdot 2^{\tau/2}}{\mu R}\right)^2\right) \sum_{t=1}^{\tau} 2^t\right\}\right) \\
&= \mathcal{O}\left(\max\left\{\sqrt{\frac{L}{\mu}} \ln\left(\frac{\mu R^2}{\varepsilon}\right) \ln\left(\frac{\sqrt{L}}{\sqrt{\mu}\delta} \ln\left(\frac{\mu R^2}{\varepsilon}\right)\right), \left(\frac{\sigma}{\mu R}\right)^2 \ln\left(\frac{\tau}{\delta}\left(\frac{\sigma}{\mu R}\right)^2 \cdot 2\right) 2^\tau\right\}\right) \\
&= \mathcal{O}\left(\max\left\{\sqrt{\frac{L}{\mu}} \ln\left(\frac{\mu R^2}{\varepsilon}\right) \ln\left(\frac{\sqrt{L}}{\sqrt{\mu}\delta} \ln\left(\frac{\mu R^2}{\varepsilon}\right)\right), \left(\frac{\sigma^2}{\mu\varepsilon}\right) \ln\left(\frac{1}{\delta}\left(\frac{\sigma^2}{\mu\varepsilon}\right) \ln\left(\frac{\mu R^2}{\varepsilon}\right)\right)\right\}\right),
\end{aligned}$$

which concludes the proof. □

D PROPERTIES OF HERMITE POLYNOMIALS

This section collects some properties of Hermite polynomials which are used in the proof of Lemma A.4. First, let us recall the definition. There are two versions of Hermite polynomials, which are referred to as "physicist's" and "probabilist's", given by

$$\mathscr{H}_n(x) = (-1)^n e^{x^2} \frac{\mathrm{d}^n}{\mathrm{d}x^n} e^{-x^2} \quad \text{and} \quad \mathcal{H}_n(x) = (-1)^n e^{x^2/2} \frac{\mathrm{d}^n}{\mathrm{d}x^n} e^{-x^2/2} \quad n \in \mathbb{N},$$

respectively. Obviously, for any positive integer n, the following relation holds true:

$$\mathcal{H}_n(x) \equiv 2^{-n/2} \mathscr{H}_n\left(\frac{x}{\sqrt{2}}\right). \tag{152}$$

Hence, "probabilist's" Hermite polynomials inherit all the properties of "physicist's" ones. In (Indritz, 1961), the author proved that

$$\max_{x \in \mathbb{R}} \left| \mathscr{H}_n(x)\, e^{-x^2/2} \right| \leqslant \sqrt{2^n \cdot n!} \quad \text{for all } n \in \mathbb{N}.$$

In view of (152), this implies that

$$\max_{x \in \mathbb{R}} \left| \mathcal{H}_n(x)\, e^{-x^2/4} \right| = 2^{-n/2} \max_{x \in \mathbb{R}} \left| \mathscr{H}_n\left(\frac{x}{\sqrt{2}}\right) e^{-x^2/4} \right| \leqslant \sqrt{n!} \quad \text{for all } n \in \mathbb{N}.$$

E NUMERICAL EXPERIMENTS: ADDITIONAL DETAILS

For every combination of noise distribution and method, we tuned optimal parameters for 70000 steps and ran methods on 95000 steps, where one step is one oracle call.

The optimal values of the learning rate and the clipping parameter were selected via grid search over the sets $\{0.002, 0.004, 0.008, 0.01, 0.02, 0.04\}$ and $\{0.75, 1, 1.5, 2, 4, 8\}$, respectively.

Distribution	Method	Learning Rate	Clipping parameter
Cauchy	clipped-MB-SGD	0.004	4
	Med-MB-SGD	0.002	-
	MB-clipped-SGD	0.01	4
	clipped-Med-MB-SGD	0.002	2
	SMoM-MB-SGD	0.002	-
	clipped-SMoM-MB-SGD	0.008	1
Cauchy + Exponential	clipped-MB-SGD	0.008	1.5
	Med-MB-SGD	0.002	-
	MB-clipped-SGD	0.04	4
	clipped-Med-MB-SGD	0.002	8
	SMoM-MB-SGD	0.002	-
	clipped-SMoM-MB-SGD	0.002	8
Cauchy + Pareto	clipped-MB-SGD	0.008	1.5
	Med-MB-SGD	0.002	-
	MB-clipped-SGD	0.02	0.75
	clipped-Med-MB-SGD	0.002	8
	SMoM-MB-SGD	0.002	-
	clipped-SMoM-MB-SGD	0.008	1

Table 1: Optimal parameters for different distributions and methods.

We also provide plots, reflecting the dependence of the error on the number of iterations, where one iteration is one method's update, see Figure 2 below. As we can see, clipped-SMoM-MB-SGD converges much faster than the competitors due to the larger batch size.

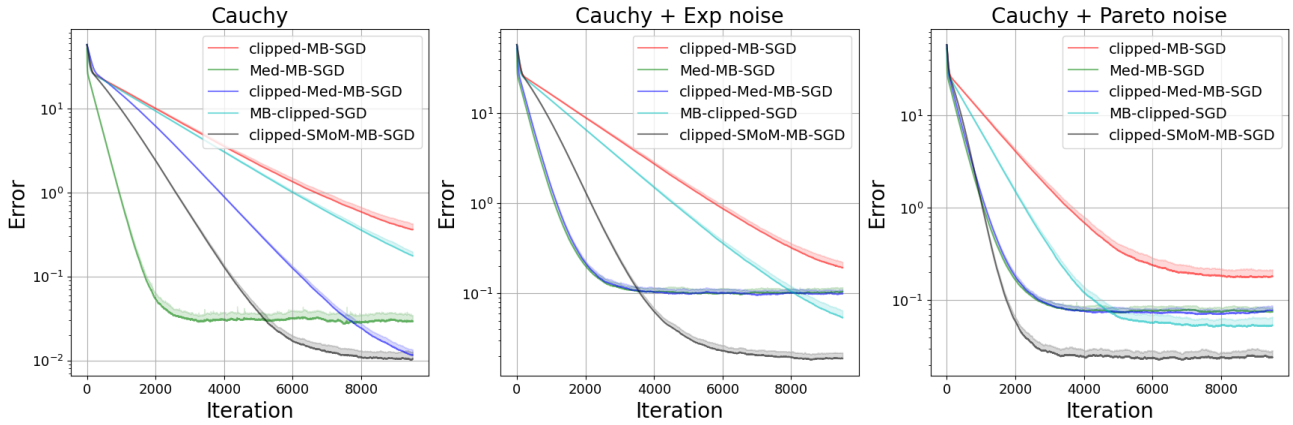

Figure 2: Dependence of the mean error on the number of iterations with a standard deviation upper bound.

Finally, according to our theoretical findings, the error bound for the mini-batched SGD with clipped smoothed median of means grows logarithmically with $1/\delta$. In Figure 3, we plot the dependence of the confidence interval width on the number of iterations to illustrate this point.

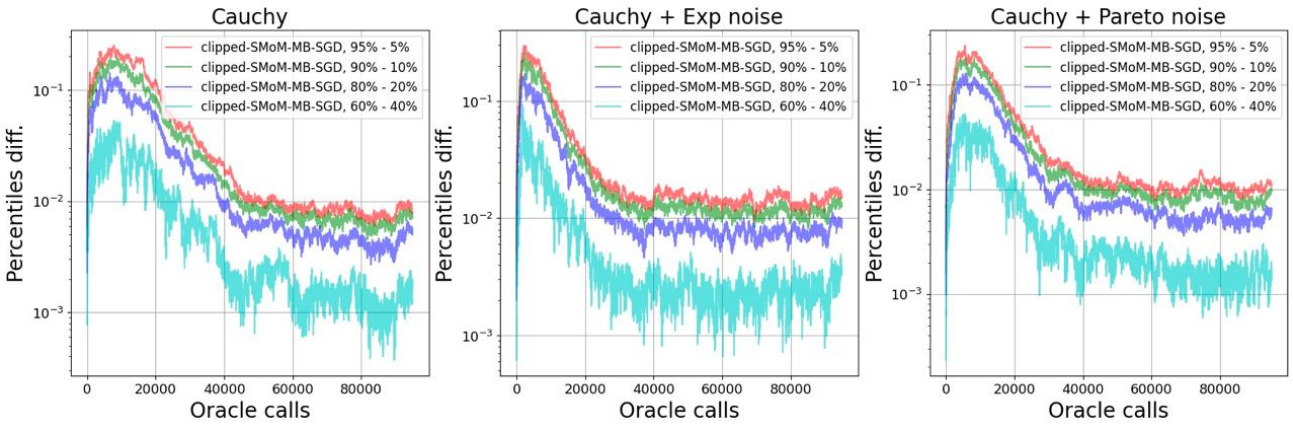

Figure 3: Dependence of the confidence interval width for the error of mini-batched SGD with clipped smoothed median of means on the number of iterations.