
Clipping Improves Adam-Norm and AdaGrad-Norm when the Noise Is Heavy-Tailed

Savelii Chezhegov^{1 2 3} Yaroslav Klyukin¹ Andrei Semenov⁴ Aleksandr Beznosikov^{1 2 5 6}
Alexander Gasnikov^{5 1 7} Samuel Horváth⁸ Martin Takáč⁸ Eduard Gorbunov⁸

Abstract

Methods with adaptive stepsizes, such as **AdaGrad** and **Adam**, are essential for training modern Deep Learning models, especially Large Language Models. Typically, the noise in the stochastic gradients is heavy-tailed for the later ones. Gradient clipping provably helps to achieve good high-probability convergence for such noises. However, despite the similarity between **AdaGrad/Adam** and **Clip-SGD**, the current understanding of the high-probability convergence of **AdaGrad/Adam**-type methods is limited in this case. In this work, we prove that **AdaGrad/Adam** (and their delayed version) can have provably bad high-probability convergence if the noise is heavy-tailed. We also show that gradient clipping fixes this issue, i.e., we derive new high-probability convergence bounds with polylogarithmic dependence on the confidence level for **AdaGrad-Norm** and **Adam-Norm** with clipping and with/without delay for smooth convex/non-convex stochastic optimization with heavy-tailed noise. We extend our results to the case of **Clip-AdaGrad/Clip-Adam** with delayed stepsizes. Our empirical evaluations highlight the superiority of clipped versions of **AdaGrad/Adam** in handling the heavy-tailed noise.

1. Introduction

Stochastic first-order optimization methods such as Stochastic Gradient Descent (**SGD**) (Robbins & Monro, 1951) are the methods of choice in training modern Machine Learning (ML) and Deep Learning (DL) models (Shalev-Shwartz & Ben-David, 2014; Goodfellow et al., 2016). There are multiple reasons for that, including but not limited to their simplicity, computation cost, memory usage, and generalization. However, standard **SGD** is rarely used due to its sensitivity to the choice of stepsize. Therefore, methods such as **AdaGrad** (Streeter & McMahan, 2010; Duchi et al., 2011) and **Adam** (Kingma & Ba, 2014), which use adaptive¹ stepsizes, are much more popular in the DL community (Vaswani et al., 2017; You et al., 2019; Nikishina et al., 2022; Li et al., 2022; Abdulkhakimov et al., 2024; 2023; Li et al., 2024; Schaipp et al., 2023; Loizou et al., 2021; Moskvoretskii et al., 2024a;b; Shi et al., 2023). In particular, **Adam**-type methods are not just easier to tune but they also achieve better results in terms of the model performance than **SGD** in the training of Large Language Models (LLMs) (Devlin et al., 2019; Zhang et al., 2020).

In the attempt to explain the later phenomenon, Zhang et al. (2020) consider the noise distribution in the stochastic gradients appearing in the pre-training of the BERT model (Devlin et al., 2019) and show that (i) the gradient noise is heavy-tailed in this case, (ii) **Adam** significantly outperforms **SGD** (with momentum), (iii) **Clip-SGD** (Pascanu et al., 2013) also converges better than **SGD** for such problems, and (iv) **Clip-SGD** is provably convergent (in-expectation) when the noise has bounded α -th moment for some $\alpha \in (1, 2]$ while **SGD** can diverge for $\alpha < 2$. Moreover, gradient clipping also plays a central role in the recent advances on the *high-probability convergence* of stochastic methods under the heavy-tailed noise (Gorbunov et al., 2020; Cutkosky & Mehta, 2021; Sadiev et al., 2023; Nguyen et al., 2023). Taking into account the similarities between **Adam**

¹Moscow Institute of Physics and Technology, Russia

²Ivannikov Institute for System Programming RAS, Russia ³Sber

AI Lab, Russia ⁴Machine Learning and Optimization Laboratory

(MLO), EPFL, Lausanne, Switzerland ⁵Innopolis University, Rus-

sia ⁶The Russian Presidential Academy of National Economy and

Public Administration, Russia ⁷Skolkovo Institute of Science and

Technology, Russia ⁸Mohamed bin Zayed University of Artificial

Intelligence, UAE. Correspondence to: Eduard Gorbunov <eduard.gorbunov@mbzuai.ac.ae>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

¹Throughout the paper, we use the word “adaptivity” in its general meaning: stepsizes are adaptive if they depend on the (stochastic) gradients or function values. We emphasize that, in this sense, an adaptive method can still have parameters affecting its convergence.

and Clip-SGD (the former one can be seen as Clip-SGD with momentum and iteration-dependent clipping level), one can conjecture that Adam enjoys good theoretical high-probability convergence when the gradient noise is heavy-tailed. If this was true, it would be perfectly aligned with the observations from (Zhang et al., 2020) about the connection between the noise in the gradients and Adam’s performance. Moreover, some recent works show that AdaGrad/Adam have provable convergence under generalized smoothness assumptions (Faw et al., 2023; Wang et al., 2023; Li et al., 2023; Wang et al., 2024). Since Clip-SGD has similar convergence properties and since some authors explicitly mention that in this regard Adam and Clip-SGD are similar², it is natural to conjecture that clipping is not needed in Adam/AdaGrad.

However, there are no theoretical results showing the high-probability convergence with *polylogarithmic dependence on the confidence level* of Adam under the heavy-tailed noise and even in the case of the bounded variance. Even for simpler “twin”³ such as AdaGrad there exists a similar gap in the literature. Moreover, Mosbach et al. (2020) apply gradient clipping even for Adam in the fine-tuning of BERT and ALBERT (Lan et al., 2019) models. However, Mosbach et al. (2020) do not report the results that can be achieved by Adam without clipping. Therefore, it remains unclear whether and when the gradient clipping is needed for AdaGrad/Adam and whether AdaGrad/Adam enjoy desirable high-probability convergence under the heavy-tailed noise.

In this work, we address this gap in the literature, i.e., we consider the following questions:

Does the high-probability complexity of Adam/AdaGrad without clipping have polylogarithmic dependence on the confidence level under the heavy-tailed noise?

Does clipping improve the convergence of AdaGrad/Adam under the heavy-tailed noise?

We provide a negative answer to the first question and a positive answer to the second one.

1.1. Our Contributions

The main contributions of this work are summarized below.

- **Negative results for Adam and AdaGrad.** We show that the high-probability complexities of Adam and AdaGrad

²Pan & Li (2023) write in the abstract: “We conclude that the sharpness reduction effect of adaptive coordinate-wise scaling is the reason for Adam’s success in practice.” In addition, Zhou et al. (2020) mention in the discussion of the related work: “...adaptation in ADAM provides a clipping effect.”

³The existing convergence results for Adam often require the choice of parameters that make Adam very similar to AdaGrad with momentum (Défossez et al., 2022); see also Section 1.3.

and their variants with delay by Li & Orabona (2020) do not have polylogarithmic dependence on the confidence level in the worst case when the noise is heavy-tailed. In particular, we design an example of a convex stochastic optimization problem such that the noise is heavy-tailed and the high-probability convergence complexity of Adam/AdaGrad has the inverse-power dependence on the target accuracy and confidence level.

- **Clipping fixes Adam-Norm and AdaGrad-Norm.** We prove that the above issue can be addressed via gradient clipping. That is, we derive high-probability complexity results for Clip-Adam-Norm and Clip-AdaGrad-Norm (with and without momentum) in the case of smooth convex (for the methods with delay) and non-convex (for the methods with and without delay) optimization with the heavy-tailed noise having bounded α -th moment with $\alpha \in (1, 2]$. The obtained results have the desired polylogarithmic dependence on the confidence level. Moreover, in the non-convex case, the derived complexities are optimal up to logarithmic factors, and match the complexity of Clip-SGD in the convex case up to logarithmic factors. We derive similar results for the modifications pf Clip-Adam and Clip-AdaGrad with delay in the non-convex case, showing that *our analysis is applicable to the case of the methods with coordinate-wise stepsizes*.
- **Numerical experiments.** We conducted numerical experiments for synthetic and real-world problems. More precisely, we illustrate the superiority of different versions of Adam/AdaGrad with clipping to the non-clipped versions of Adam/AdaGrad on a simple quadratic problem with additive heavy-tailed noise in the gradients. Next, we also test Adam with and without clipping on the fine-tuning of ALBERT Base model (Lan et al., 2019) on CoLa and RTE datasets (Wang et al., 2018) and observe that Adam with clipping significantly outperforms Adam without clipping when the noise is heavy-tailed. We also obtain similar results for the fine-tuning of RoBERTa Large model (Liu et al., 2019).

1.2. Preliminaries

In this section, we formalize the setup. We focus on unconstrained minimization problems

$$\min_{x \in \mathbb{R}^d} f(x), \quad (1)$$

where the differentiable function $f(x)$ is accessible through the calls of stochastic first-order oracle returning an approximation $\nabla f_\xi(x)$ of $\nabla f(x)$. Here ξ is a random variable following some distribution \mathcal{D} that may be dependent on x and time. In the simplest case, $f_\xi(x)$ is a loss function on the data sample ξ and $f(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[f_\xi(x)]$ is a population risk (Shalev-Shwartz & Ben-David, 2014).

Notation. The notation is quite standard in this work. We use $\mathbb{E}_\xi[\cdot]$ to denote an expectation w.r.t. random variable ξ . All norms are standard Euclidean ones: $\|x\| = \sqrt{\langle x, x \rangle}$. The ball centered at x with a radius R is defined as $B_R(x) := \{y \in \mathbb{R}^d \mid \|y - x\| \leq R\}$. We also use x^* to denote (any) solution of (1) and $f_* := \inf_{x \in \mathbb{R}^d} f(x)$. Clipping operator with clipping level $\lambda > 0$ is defined as $\text{clip}(x, \lambda) := \min\{1, \lambda/\|x\|\}x$ for $x \neq 0$ and $\text{clip}(x, \lambda) := 0$ for $x = 0$.

Assumptions. We start with the assumption⁴ on the noise.

Assumption 1.1. There exists set $Q \subseteq \mathbb{R}^d$ and $\sigma \geq 0, \alpha \in (1, 2]$ such that for all $k \geq 0$ the oracle satisfies $\mathbb{E}[\nabla f_{\xi_k}(x) \mid x] = \nabla f(x)$ and

$$\mathbb{E}[\|\nabla f_{\xi_k}(x) - \nabla f(x)\|^\alpha \mid x] \leq \sigma^\alpha, \quad \forall x \in Q. \quad (2)$$

The sequence $\{\xi_k\}_{k \geq 0}$ is the sequence of independent random variables.

The above assumption is used in many recent works (Zhang et al., 2020; Cutkosky & Mehta, 2021; Sadiev et al., 2023; Nguyen et al., 2023). When $\alpha < 2$, it allows the stochastic gradients to have unbounded variance, e.g., Lévy α -stable noise. Such distributions are usually called heavy-tailed. When $\alpha = 2$, it reduces to the standard bounded variance assumption (Nemirovski et al., 2009; Ghadimi & Lan, 2012; 2013; Takáč et al., 2013).

We also emphasize that the above assumption allows for the time-dependent noise, which we actively use in our negative results from Section 2. Although not often explicitly stated, many existing results in stochastic optimization (Ghadimi & Lan, 2012; Harvey et al., 2019; Sadiev et al., 2023; Zhang et al., 2020; Cutkosky & Mehta, 2021; Nguyen et al., 2023) hold in the case of the time-dependent noise as long as certain moment bounds (e.g., (2)) hold.

Next, we make a standard assumption about the smoothness of the objective function.

Assumption 1.2. There exists set $Q \subseteq \mathbb{R}^d$ and $L > 0$ such that for all $x, y \in Q$

$$\begin{aligned} \|\nabla f(y) - \nabla f(x)\| &\leq L \|y - x\|, \\ \|\nabla f(x)\|^2 &\leq 2L(f(x) - f_*). \end{aligned} \quad (3)$$

We emphasize that the second part of (3) follows from the first part if $Q = \mathbb{R}^d$. However, in more general situations,

⁴Similarly to (Sadiev et al., 2023), for our results, it is sufficient to make all the assumptions only on some set Q . This set is typically bounded and depends on some metric of sub-optimality of the starting point, e.g., the distance from the starting point to the optimum. We emphasize that our assumptions are strictly weaker than corresponding ones for $Q = \mathbb{R}^d$. To achieve this kind of generality, we prove that the proposed method does not leave some set Q with high probability.

this is not always the case; see (Sadiev et al., 2023, Appendix B) for further details. Interestingly, when Q is a compact set, function f can have non-Lipschitz gradients (e.g., polynomially growing with x) on \mathbb{R}^d , see also (Patel et al., 2022; Patel & Berahas, 2022).

In addition, for some of our results, we assume that the objective is convex.

Assumption 1.3 (Optional). There exists set $Q \subseteq \mathbb{R}^d$ such that for all $x, y \in Q$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle. \quad (4)$$

Finally, for the methods without the delay, we assume that function f is bounded.

Assumption 1.4 (Optional). There exists constant $M > 0$ such that for all $x \in \mathbb{R}^d$

$$f(x) - f_* \leq M. \quad (5)$$

A stronger version of the above assumption (boundedness of the empirical risk) is used in (Li & Liu, 2023), which is the only existing work analyzing **AdaGrad** with clipping.

Why high-probability convergence? The vast majority of the existing literature on stochastic optimization focuses on the in-expectation convergence guarantees only. In particular, for some metric $\mathcal{P}(x)$ quantifying the output's quality, e.g., $\mathcal{P}(x) = f(x) - f(x^*)$, $\|\nabla f(x)\|^2$, $\|x - x^*\|^2$, such guarantees provide upper bounds on the number of iterations/oracle calls required for a method to find x such that $\mathbb{E}[\mathcal{P}(x)] \leq \varepsilon$. However, during recent years, *high-probability convergence* guarantees have been gaining a lot of attention as well. Such guarantees give upper bounds on the number of iterations/oracle calls required for a method to find x such that $\mathbb{P}\{\mathcal{P}(x) \leq \varepsilon\} \geq 1 - \delta$, where δ is usually called confidence level or failure probability. One can argue that using Markov's inequality, one can easily deduce a high-probability guarantee from an in-expectation one: if $\mathbb{E}[\mathcal{P}(x_{K(\varepsilon\delta)})] \leq \varepsilon\delta$, where $x_{K(\varepsilon\delta)}$ is an output of the method after $K(\varepsilon\delta)$ iterations/oracle calls, then $\mathbb{P}\{\mathcal{P}(x_{K(\varepsilon\delta)}) > \varepsilon\} < \mathbb{E}[\mathcal{P}(x_{K(\varepsilon\delta)})]/\varepsilon \leq \delta$. Unfortunately, for many methods such as **SGD** (Ghadimi & Lan, 2013) $K(\varepsilon)$ has inverse-power dependence on ε implying that $K(\varepsilon\delta)$ has inverse-power dependence on $\varepsilon\delta$, leading to a noticeable deterioration when δ is small. Therefore, deriving high-probability complexities with *polylogarithmic dependence on δ* requires a separate and thorough consideration and analysis. Moreover, such bounds more accurately reflect the methods' behavior (Gorunov et al., 2020).

1.3. Related Work

High-probability convergence. The first results showing the high-probability convergence of **SGD** and its variants

are derived under the sub-Gaussian noise assumption for convex and strongly convex problems by Nemirovski et al. (2009); Ghadimi & Lan (2012); Harvey et al. (2019) for non-convex problems by Li & Orabona (2020). Although the distribution of the noise is near-sub-Gaussian in some cases, like in the training of ResNet50 (He et al., 2016) on ImageNet (Russakovsky et al., 2015) as shown by Zhang et al. (2020), this assumption does not cover even the distributions with bounded variance. To relax the sub-Gaussian noise assumption, Nazin et al. (2019) consider a truncated version of Stochastic Mirror Descent, which is closely related to Clip-SGD, and prove its high-probability complexity with polylogarithmic dependence on δ under bounded variance assumption for convex smooth problems on the bounded domain. In the strongly convex case, Davis et al. (2021) propose a general approach for obtaining high-probability convergence based on the robust distance estimation and show accelerated high-probability rates in the strongly convex case. Next, for the unconstrained problems, Gorbunov et al. (2020) prove the first high-probability convergence results for Clip-SGD and the first accelerated high-probability rates in the convex case for a version of Clip-SGD with Nesterov's momentum (Nesterov, 1983). This result is generalized to the problems with Hölder-continuous gradients by Gorbunov et al. (2021). Cutkosky & Mehta (2021) derive the first high-probability convergence results under Assumption 1.1 with $\alpha < 2$ for a version of Clip-SGD with normalization and Polyak's momentum (Polyak, 1964) in the case of non-convex problems with bounded gradient. Sadiev et al. (2023) remove the bounded gradient assumption in the non-convex case and also prove the first high-probability convergence results under Assumption 1.1 for Clip-SGD and its accelerated version in the convex and strongly convex cases. Nguyen et al. (2023) provide improved results in the non-convex case under Assumption 1.1 and also improved the dependency on the logarithmic factors in the convergence bounds. The generalization to the composite and distributed optimization problems is developed by Gorbunov et al. (2024). It is also worth mentioning (Jakovetić et al., 2023; Puchkin et al., 2024) who consider potentially heavier noise than in Assumption 1.1 through utilizing the additional structure of the noise such as (near-)symmetry. This direction is further explored by Kornilov et al. (2024) and adjusted to the case of the zeroth-order stochastic oracle.

AdaGrad and Adam. AdaGrad (Streeter & McMahan, 2010; Duchi et al., 2011) has the following update-rule

$$x_{t+1} = x_t - \frac{\gamma}{b_t} \nabla f_{\xi_t}(x_t), \quad (\text{AdaGrad})$$

where $b_t = \sqrt{b_{t-1}^2 + (\nabla f_{\xi_t}(x_t))^2}$,

where all operations (taking a square and taking a square root of a vector, division by a vector) are performed

coordinate-wise. The method is analyzed in many works, including (Streeter & McMahan, 2010; Duchi et al., 2011; Zou et al., 2018; Chen et al., 2018; Ward et al., 2020; Défossez et al., 2022; Faw et al., 2022) to name a few. However, the high-probability convergence of AdaGrad is studied under restrictive assumptions such as almost surely sub-Gaussian noise (Li & Orabona, 2020; Liu et al., 2023) or without such an assumption but with inverse-power dependence on the confidence level δ (Wang et al., 2023) or boundedness of the empirical risk and (non-central) α -th moment (Li & Liu, 2023), which in the worst case implies boundedness of the stochastic gradient (see the discussion after Theorem 3.3). In contrast, our results for Clip-Adam(D)/Clip-M-AdaGrad(D)(-Norm) hold under Assumption 1.1 (and under additional Assumption 1.4 for the methods without delay) and have polylogarithmic dependence on δ .

Adam (Kingma & Ba, 2014) can be seen as a modification of AdaGrad with an exponential moving average b_t^2 of the squared stochastic gradients and with Polyak's momentum (Polyak, 1964):

$$\begin{aligned} x_{t+1} &= x_t - \frac{\gamma}{b_t} m_t, \\ m_t &= \beta_1 m_{t-1} + (1 - \beta_1) \nabla f_{\xi_t}(x_t), \\ b_t &= \sqrt{\beta_2 b_{t-1}^2 + (1 - \beta_2) (\nabla f_{\xi_t}(x_t))^2}, \end{aligned} \quad (6)$$

where all operations (taking a square and taking a square root of a vector, division by a vector) are performed coordinate-wise. Although the original proof by Kingma & Ba (2014) has a flaw spotted by Reddi et al. (2019), one can still show the convergence of Adam when β_2 goes to 1 (Défossez et al., 2022; Zhang et al., 2022; Wang et al., 2024). Moreover, for any fixed β_1 and β_2 such that $\beta_1 < \sqrt{\beta_2}$, e.g., for the default values $\beta_1 = 0.9$ and $\beta_2 = 0.999$, Adam is not guaranteed to converge (Reddi et al., 2019, Theorem 3). Therefore, the standard choice of β_2 in theory is $\beta_2 = 1 - 1/K$, where K is the total number of steps, and that is why, as noticed by Défossez et al. (2022), AdaGrad and Adam are “twins”. Indeed, taking $\beta_1 = 0$ (no momentum) and $\beta_2 = 1 - 1/K$ in (6) we get that $b_t^2 = (1 - 1/K)^{t+1} b_{-1}^2 + \frac{1}{K} \sum_{k=0}^t (1 - 1/K)^{t-k} (\nabla f_{\xi_k}(x_k))^2 = \Theta(b_{-1}^2 + \frac{1}{K} \sum_{k=0}^t (\nabla f_{\xi_k}(x_k))^2)$ since $1/4 = (1 - 1/2)^2 \leq (1 - 1/K)^{t-k} \leq 1$ for $0 \leq k \leq t \leq K$. Thus, up to the rescaling of γ and b_{-1}^2 the effective stepsize of Adam-CW is $\Theta(\cdot)$ of the effective stepsize of AdaGrad-CW (though the points where the gradients are calculated can be quite different for these two methods). This aspect explains why AdaGrad and Adam have similar proofs and convergence guarantees. The high-probability convergence of Adam is studied by Li et al. (2023) under bounded noise and sub-Gaussian noise assumptions, while our results for Clip-Adam(D) do not require such assumptions.

2. Failure of Adam/AdamD and AdaGrad/AdaGradD with Momentum

In this section, we present the negative result on the convergence of Adam, AdaGrad with Momentum (**M-AdaGrad**), and their delayed versions – AdamD/M-AdaGradD (Li & Orabona, 2020).

Theorem 2.1. *For any $\sigma > 0$ and sufficiently small $\varepsilon, \delta \in (0, 1)$, there exist problems (1) such that Assumptions 1.1, 1.2, 1.3, hold with $L = 1$, $\alpha = 2$, and the iterates produced by Adam(D)/M-AdaGrad(D) with x_0 such that $\|x_0 - x^*\| \gg \gamma L$ and with $\beta_2 = 1 - 1/T$ for Adam(D) satisfy: if $\mathbb{P}\{f(x_T) - f(x^*) \geq \varepsilon\} \leq \delta$, then*

$$T = \Omega\left(\text{poly}(\varepsilon^{-1/2}, \delta^{-1/2}, \delta^{-1/3})\right), \quad (7)$$

i.e., the complexity of Adam(D)/M-AdaGrad(D) has inverse-power dependence on δ .

Sketch of the proof. To construct our example, we consider the Huber loss function (Huber, 1992)

$$f(x) = \begin{cases} \frac{1}{2}x^2, & \text{if } |x| \leq \nu, \\ \nu(|x| - \frac{1}{2}\nu), & \text{otherwise,} \end{cases} \quad (8)$$

and design two specific sequences of noises (one for Adam/M-AdaGrad and the second one for AdamD/M-AdaGradD). For Adam/M-AdaGrad, we consider a discrete additive noise for the first step such that Markov's inequality holds as equality, and for the remaining steps, noise equals zero. Then, with high probability, b_t becomes large after the first step, which slowdowns the method. As for AdamD/M-AdaGradD, similarly to Sadiev et al. (2023), we add the noise only to the last step: since b_t is constructed using the norm of the previous stochastic gradient, the noise is independent of the stepsize and can spoil the last iterate. See the complete proofs and details in Appendix B. \square

Interestingly, in the above example, it is sufficient to consider the noise with bounded variance to show that the high-probability convergence rates of Adam(D)/M-AdaGrad(D) depend polynomially on ε^{-1} and $\delta^{-1/2}$. Moreover, following a similar argument to (Zhang et al., 2020, Remark 1), one can show the non-convergence of AdamD/M-AdaGradD when $\alpha < 2$. We also conjecture that for $\alpha < 2$ one can show even worse dependence on ε and δ for Adam/AdaGrad (or even non-convergence) since b_t will grow with high probability even faster in this case. Moreover, we also emphasize that the negative result for Adam(D) is established only for $\beta_2 = 1 - 1/T$, which is a standard assumption to ensure convergence of Adam-type methods. Nevertheless, the negative result of Theorem 2.1 provides necessary evidence that

Adam(D)/M-AdaGrad(D) do not achieve desired high-probability convergence rates and motivates us to apply clipping to Adam(D)/M-AdaGrad(D).

Time-dependent noise. We also emphasize that the noise structure is time-dependent in the provided example, which is used to simplify the proof. Moreover, as discussed in Section 1.2, many existing upper bounds hold for time-dependent noise as well.

Initial condition. The provided negative examples rely on the assumption that $|x_0|$ is sufficiently large, i.e., the method is initialized not too close to the optimum. In particular, for M-AdaGrad, we require $x_0 > \sqrt{2\varepsilon} + 3\gamma$. Since typically $\varepsilon, \gamma \ll 1$, the condition is relatively mild. Moreover, this assumption simplifies the proof. Although we do not provide negative results for an arbitrary choice of x_0 and γ , we conjecture that similar negative results can be obtained in the case of more general choice of γ .

Generalization under Assumption 1.4. The provided example does not satisfy Assumption 1.4 that is used in the next section in the analysis of methods without delay (Theorem 3.3). To address this issue, one can replace function (8) with the following one:

$$f(x) = \begin{cases} \frac{1}{2}x^2, & \text{if } |x| \leq \nu, \\ \nu(|x| - \frac{1}{2}\nu), & \text{if } \nu < |x| \leq D, \\ \nu(D - \frac{1}{2}\nu), & \text{if } |x| > D, \end{cases} \quad (9)$$

where D is such that $D > |x_0|$. Then, the modified function satisfies Assumption 1.4 and the proofs remain the same.

3. New Upper Bounds

Methods. To address the issue indicated in Theorem 2.1, we consider Clip-Adam(D)/Clip-M-AdaGrad(D)-Norm (see Algorithm 2). In contrast to the existing practice (Pan & Li, 2023), we use clipping of the stochastic gradient not only in the update rule for momentum buffer m_t (Line 3 in Algorithm 2), but also in the computation of the scaling factor b_t (Lines 5 and 7 in Algorithm 2). The role of clipping in m_t is similar to the role of clipping in Clip-SGD-type methods: it prevents the method from too large steps that may occur due to the presence of the heavy-tailed noise in the gradients. In this regard, it is important to select clipping level in such a way that bias and variance of the estimator are balanced. However, the role of clipping in b_t is different: clipping prevents b_t from growing too quickly since such a growth can lead to poor high-probability guarantees (see the proof's sketch of Theorem 2.1). We note that clipping is also used in Clip-AdaGrad-Norm (without momentum, i.e., with $\beta_1 = 0$) for both m_t and b_t computation by Li &

Algorithm 1 Adam-norm/AdamD-norm and M-AdaGrad-norm/M-AdaGradD-norm

Input: Stepsize $\gamma > 0$, starting point $x_0 \in \mathbb{R}^d$, initial constant $b_{-1} > 0$ (for Adam-norm and M-AdaGrad-norm) or $b_0 > 0$ (for AdamD-norm and M-AdaGradD-norm), momentum parameters $\beta_1, \beta_2 \in [0, 1]$

- 1: Set $m_{-1} = 0$
- 2: **for** $t = 0, 1, \dots$ **do**
- 3: $m_t = \beta_1 m_{t-1} + (1 - \beta_1) \nabla f_{\xi_t}(x_t)$
- 4: **if** no delay **then**
- 5: $b_t = \begin{cases} \sqrt{\beta_2 b_{t-1}^2 + (1 - \beta_2) \|\nabla f_{\xi_t}(x_t)\|^2} & \text{for Adam-norm} \\ \sqrt{b_{t-1}^2 + \|\nabla f_{\xi_t}(x_t)\|^2} & \text{for M-AdaGrad-norm} \end{cases}$
- 6: **else**
- 7: $b_{t+1} = \begin{cases} \sqrt{\beta_2 b_t^2 + (1 - \beta_2) \|\nabla f_{\xi_t}(x_t)\|^2} & \text{for AdamD-norm} \\ \sqrt{b_t^2 + \|\nabla f_{\xi_t}(x_t)\|^2} & \text{for M-AdaGradD-norm} \end{cases}$
- 8: **end if**
- 9: $x_{t+1} = x_t - \frac{\gamma}{b_t} m_t$
- 10: **end for**

Algorithm 2 Clip-Adam-norm/Clip-AdamD-Norm and Clip-M-AdaGrad-norm/Clip-M-AdaGradD-Norm

Input: Stepsize $\gamma > 0$, starting point $x_0 \in \mathbb{R}^d$, initial constant $b_{-1} > 0$ (for Clip-Adam-norm and Clip-M-AdaGrad-norm) or $b_0 > 0$ (for Clip-AdamD-norm and Clip-M-AdaGradD-norm), momentum parameters $\beta_1, \beta_2 \in [0, 1]$, level of clipping $\lambda > 0$

- 1: Set $m_{-1} = 0$
- 2: **for** $t = 0, 1, \dots$ **do**
- 3: $m_t = \beta_1 m_{t-1} + (1 - \beta_1) \text{clip}(\nabla f_{\xi_t}(x_t), \lambda)$
- 4: **if** no delay **then**
- 5: $b_t = \begin{cases} \sqrt{\beta_2 b_{t-1}^2 + (1 - \beta_2) \|\text{clip}(\nabla f_{\xi_t}(x_t), \lambda)\|^2} & \text{for Clip-Adam-Norm} \\ \sqrt{b_{t-1}^2 + \|\text{clip}(\nabla f_{\xi_t}(x_t), \lambda)\|^2} & \text{for Clip-M-AdaGrad-Norm} \end{cases}$
- 6: **else**
- 7: $b_{t+1} = \begin{cases} \sqrt{\beta_2 b_t^2 + (1 - \beta_2) \|\text{clip}(\nabla f_{\xi_t}(x_t), \lambda)\|^2} & \text{for Clip-AdamD-Norm} \\ \sqrt{b_t^2 + \|\text{clip}(\nabla f_{\xi_t}(x_t), \lambda)\|^2} & \text{for Clip-M-AdaGradD-Norm} \end{cases}$
- 8: **end if**
- 9: $x_{t+1} = x_t - \frac{\gamma}{b_t} m_t$
- 10: **end for**

Liu (2023) but the authors do not comment about the role of clipping in b_t and use restrictive assumptions as we explain later in this section.

Convergence results. We derive new high-probability convergence bounds for the generalized method formalized as Algorithm 2 in the convex and non-convex cases. The following theorem gives the main result for Clip-AdamD/Clip-AdaGradD-Norm in the convex case.

Theorem 3.1 (Convex Case). *Let $K > 0$ and $\delta \in (0, 1]$ and Assumptions 1.1, 1.2, and 1.3 hold for $Q = B_{2R}(x^*)$ for some $R \geq \|x_0 - x^*\|$. Assume that $\beta_1 \in [0, 1]$, $\beta_2 = \frac{K}{K+1}$ (for Clip-AdamD-Norm)*

$$\gamma = \Theta \left(\min \left\{ \frac{(1-\beta_1)^2 b_0}{LA}, \frac{\sqrt{1-\beta_1} R b_0}{\sigma(K+1)^{\frac{1}{\alpha}} A^{\frac{\alpha-1}{\alpha}}} \right\} \right) \text{ and } \lambda =$$

$\Theta \left(\frac{\sqrt{1-\beta_1} b_0 R}{\gamma A} \right)$, where $A = \ln \left(\frac{4(K+1)}{\delta} \right)$. Then, to guarantee $f(\bar{x}_K) - f(x^*) \leq \varepsilon$ with probability at least $1 - \delta$ for $\bar{x}_K = \frac{1}{K+1} \sum_{t=0}^K x_t$ Clip-AdamD/Clip-M-AdaGradD-Norm requires :

$$\tilde{O} \left(\max \left\{ \frac{LR^2}{(1-\beta_1)^3 \varepsilon}, \left(\frac{\sigma R}{(1-\beta_1)^{\frac{3}{2}} \varepsilon} \right)^{\frac{\alpha}{\alpha-1}} \right\} \right) \quad (10)$$

iterations/oracle calls. Moreover, with probability at least $1 - \delta$, all iterates $\{x_t\}_{t=0}^K$ stay in Q .

Next, we present our main results for Clip-AdamD/Clip-M-AdaGradD-Norm and Clip-Adam/Clip-M-AdaGrad-Norm in the non-convex case.

Theorem 3.2 (Non-Convex Case: Methods with Delay). *Let $K > 0$ and $\delta \in (0, 1]$ and Assumptions 1.1 and 1.2 hold for*

$Q = \left\{ x \in \mathbb{R}^d \mid \exists y \in \mathcal{L}_f(2\Delta) : \|x - y\| \leq \frac{\sqrt{\Delta}}{20\sqrt{L}} \right\}$ with $\mathcal{L}_f(2\Delta) := \{y \in \mathbb{R}^d \mid f(y) \leq f_* + 2\Delta\}$ for some $\Delta \geq f(x_0) - f_*$. Assume that $\beta_1 \in [0, 1]$, $\beta_2 = \frac{K}{K+1}$ (for Clip-AdamD-Norm) and

$$\gamma = \Theta \left(\min \left\{ \frac{(1 - \beta_1)^2 b_0}{L(K+1)^{\frac{\alpha-1}{3\alpha-2}} A}, \frac{\sqrt{1 - \beta_1} b_0 \sqrt{\Delta}}{\sqrt{L} \sigma (K+1)^{\frac{\alpha}{3\alpha-2}} A^{\frac{\alpha-1}{\alpha}}}, \frac{(1 - \beta_1)^{\frac{\alpha-1}{2\alpha-1}} b_0 \Delta^{\frac{\alpha}{2\alpha-1}}}{\sigma^{\frac{2\alpha}{2\alpha-1}} L^{\frac{\alpha-1}{2\alpha-1}} (K+1)^{\frac{\alpha}{3\alpha-2}} A^{\frac{2\alpha-2}{2\alpha-1}}} \right\} \right),$$

$$\lambda = \Theta \left(\frac{\sqrt{1 - \beta_1} b_0 \sqrt{\Delta}}{\sqrt{L} \gamma A (K+1)^{\frac{\alpha-1}{3\alpha-2}}} \right), \text{ where } A = \ln \left(\frac{4(K+1)}{\delta} \right).$$

Then, to guarantee $\frac{1}{K+1} \sum_{t=0}^K \|\nabla f(x_t)\|^2 \leq \varepsilon$ with probability at least $1 - \delta$ Clip-AdamD/Clip-M-AdaGradD-Norm requires the following number of iterations/oracle calls:

$$\tilde{O} \left(\max \left\{ \left(\frac{L\Delta}{(1 - \beta_1)^3 \varepsilon} \right)^{\frac{3\alpha-2}{2\alpha-1}}, \left(\frac{\sigma \sqrt{L\Delta}}{(1 - \beta_1)^{\frac{3}{2}} \varepsilon} \right)^{\frac{3\alpha-2}{2\alpha-2}}, \left(\frac{\sigma^{\frac{2\alpha}{2\alpha-1}} (L\Delta)^{\frac{\alpha-1}{2\alpha-1}}}{(1 - \beta_1)^{\frac{3\alpha-2}{2\alpha-1}} \varepsilon} \right)^{\frac{3\alpha-2}{2\alpha-2}} \right\} \right). \quad (11)$$

Moreover, with probability at least $1 - \delta$, all iterates $\{x_t\}_{t=0}^K$ stay in Q .

Theorem 3.3 (Non-Convex Case: Methods without Delay). Let $K > 0$ and $\delta \in (0, 1]$ and Assumptions 1.1, 1.2, 1.4 hold for $Q = \mathbb{R}^d$. Assume that $\beta_1 \in [0, 1]$, $\beta_2 = 1 - \frac{1}{K}$ (for Clip-Adam-Norm) and

$$\gamma = \Theta \left(\min \left\{ \frac{b_{-1}}{L(K+1)^{\frac{\alpha-1}{3\alpha-2}} A}, \frac{b_{-1} \sqrt{M}}{\sqrt{L} \sigma (K+1)^{\frac{\alpha}{3\alpha-2}} A^{\frac{\alpha-1}{\alpha}}}, \frac{b_{-1} M^{\frac{\alpha}{2\alpha-1}}}{\sigma^{\frac{2\alpha}{2\alpha-1}} L^{\frac{\alpha-1}{2\alpha-1}} (K+1)^{\frac{\alpha}{3\alpha-2}} A^{\frac{2\alpha-2}{2\alpha-1}}} \right\} \right),$$

$\lambda = \Theta \left(\frac{b_{-1} \sqrt{M}}{\sqrt{L} \gamma A (K+1)^{\frac{\alpha-1}{3\alpha-2}}} \right)$, where $A = \ln(\frac{4}{\delta})$. Then, to guarantee $\frac{1}{K+1} \sum_{t=0}^K \|\nabla f(x_t)\|^2 \leq \varepsilon$ with probability at least $1 - \delta$ Clip-Adam/Clip-M-AdaGrad-Norm requires the following number of iterations/oracle calls:

$$\tilde{O} \left(\frac{1}{(1 - \beta_1)^{\frac{3}{2}}} \max \left\{ \left(\frac{LM}{\varepsilon} \right)^{\frac{3\alpha-2}{2\alpha-1}}, \left(\frac{\sigma \sqrt{LM}}{\varepsilon} \right)^{\frac{3\alpha-2}{2\alpha-2}}, \left(\frac{\sigma^{\frac{2\alpha}{2\alpha-1}} (LM)^{\frac{\alpha-1}{2\alpha-1}}}{\varepsilon} \right)^{\frac{3\alpha-2}{2\alpha-2}} \right\} \right). \quad (12)$$

Discussion of the results. Theorems 3.1, 3.2, and 3.3 provide high-probability complexities for Clip-Adam(D)Clip-M-AdaGrad(D)-Norm with polylogarithmic dependence

on the confidence level δ . Up to the differences in logarithmic factors, these complexities coincide with the best-known ones for Clip-SGD (Sadiev et al., 2023; Nguyen et al., 2023). Moreover, the leading terms in (11) and (12) are optimal up to logarithmic factors (Zhang et al., 2020), though the first terms in (11) and (12) can be improved (Arjevani et al., 2023). In the convex case, the first term in (10) is not optimal (Nemirovskij & Yudin, 1983) and can be improved (Gorbunov et al., 2020; Sadiev et al., 2023). The optimality of the second term in (10) is still an open question.

It is also worth mentioning that the existing high-probability complexities for Adam/AdaGrad-type (without clipping) methods either have inverse power dependence on δ (Wang et al., 2023) or have polylogarithmic dependence on δ but rely on the assumption that the noise is sub-Gaussian/bounded (Li & Orabona, 2020; Liu et al., 2023; Li et al., 2023), which is stronger than bounded variance assumption. Under the additional assumption that the empirical risk is bounded and the (non-central) α -th moment of the stochastic gradient are bounded and the empirical risk is smooth, which are stronger than Assumptions 1.4, 1.1 and 1.2 respectively, Li & Liu (2023) derive a similar bound to (12) for Clip-AdaGrad-Norm. We emphasize that boundedness and smoothness of the empirical risk imply the boundedness and smoothness of all $f_\xi(x)$ in the worst case (e.g., when the distribution \mathcal{D} is discrete). Therefore, in the worst case, these assumptions imply the boundedness of $\nabla f_\xi(x)$ (in view of the second part of (3) for function f_ξ), meaning that the noise is bounded and, thus, sub-Gaussian. In this case, clipping is not needed for AdaGrad to achieve good high-probability convergence guarantees as shown by Li & Orabona (2020); Liu et al. (2023). Our Theorem 3.3 extends this result to the momentum version of Clip-AdaGrad-Norm under less restrictive assumptions (not implying sub-Gaussianity of the noise) and gives the first high-probability convergence bounds for Clip-Adam with polylogarithmic dependence on δ .

Moreover, to the best of our knowledge, Theorems 3.1 and 3.2 are the first results showing high-probability convergence of Adam/AdaGrad-type methods with polylogarithmic dependence on the confidence level in the case of the heavy-tailed noise without extra assumptions such as Assumption 1.4. We also show that the iterates of Clip-AdamD/Clip-M-AdaGradD do not leave set Q with high probability, where $Q = B_{2R}(x^*)$ in the convex case and $Q = \left\{ x \in \mathbb{R}^d \mid \exists y \in \mathcal{L}_f(2\Delta) : \|x - y\| \leq \frac{\sqrt{\Delta}}{20\sqrt{L}} \right\}$ with $\mathcal{L}_f(2\Delta) := \{y \in \mathbb{R}^d \mid f(y) \leq f_* + 2\Delta\}$ in the non-convex case. Further details and proofs are deferred to Appendix C.

Assumption 1.4 in Theorem 3.3. As we explain above, Assumption 1.4 is weaker than the one used in Li & Liu

(2023). It is worth mentioning that Assumption 1.4 is relatively restrictive. Nevertheless, we need this assumption in our proof to overcome the difficulty of analyzing stochastic methods with correlated stepsizes, i.e., to handle the fact that g_t and b_t are dependent. The existing approaches typically use boundedness of the variance and the norm of the gradient (see Lemma 5.1 in Défossez et al. (2022)) or assume that the noise is sub-Gaussian (Li & Liu, 2023) to tackle this issue. In the heavy-tailed noise regime, these assumptions do not hold. Therefore, we use Assumption 1.4. More precisely, to decouple b_t and g_t in the analysis, we multiply inequality (68) by b_t . However, it eventually leads to the non-trivial weighted sum of function values $\sum_{t=1}^{T-1} \left(\frac{b_t}{p_t} - \frac{b_{t-1}}{p_{t-1}} \right) (f(x_t) - f_*)$ in inequality (72). To estimate this sum, we apply Assumption 1.4. We are not aware of the alternative ways of analyzing versions of AdaGrad/Adam or closely related methods in the heavy-tailed noise regime.

Analysis of coordinate-wise methods. In Appendix C.5, we derive new results for Clip-AdamD and Clip-M-AdaGradD in the non-convex case, i.e., for the methods with coordinate-wise scaling. The analysis and the derived bounds are similar to the ones from Theorem 3.2. The new bounds explicitly depend on the dimension of the problem, which is standard for the methods with coordinate-wise scaling. In particular, under the coordinate-wise version of Assumption 1.1 (see Assumption C.11), we show that there exists a proper choice of γ and λ ensuring that $\frac{1}{K+1} \sum_{k=0}^K \|\nabla f(x_k)\|^2 \leq \varepsilon$ with probability at least $1 - \delta$ after the following number of iterations/oracle calls of Clip-AdamD/Clip-M-AdaGradD:

$$\tilde{\mathcal{O}} \left(\max \left\{ \left(\frac{\sqrt{d}L\Delta}{(1-\beta_1)^3\varepsilon} \right)^{\frac{3\alpha-2}{2\alpha-1}}, \left(\frac{\sqrt{L\Delta} \left(\sum_{i=1}^d \sigma_i^\alpha \right)^{\frac{1}{\alpha}}}{(1-\beta_1)^{\frac{3}{2}} \sqrt{d}^{\frac{2-\alpha}{\alpha}} \varepsilon} \right)^{\frac{3\alpha-2}{2\alpha-2}}, \left(\frac{d^{\frac{\alpha-1}{2\alpha-1}} \left(\sum_{i=1}^d \sigma_i^{2\alpha} \right)^{\frac{1}{2\alpha-1}} (L\Delta)^{\frac{\alpha-1}{2\alpha-1}}}{(1-\beta_1)^{\frac{\alpha-1}{2\alpha-1}} \varepsilon} \right)^{\frac{3\alpha-2}{2\alpha-2}} \right\} \right).$$

Discussion of Kunstner et al. (2023). Kunstner et al. (2023) show that Adam outperforms SGD even in full-batch settings, hinting that the success of Adam is not only related to its better interaction with the heavy-tailed noise. Our focus is on high-probability convergence of AdaGrad/Adam-based methods, showing that without clipping, they have poor high-probability complexities, similar to SGD. Thus, our results complement Kunstner et al. (2023) by highlight-

ing the necessity of gradient clipping in AdaGrad/Adam for high-probability convergence.

4. Numerical Experiments

In this section, we illustrate numerically that clipping indeed helps AdaGrad and Adam to achieve better high-probability convergence. Our code is available online: <https://github.com/yaroslavkliukin/Clipped-AdaGrad-and-Adam>.

Quadratic problem. In the first experiment, we test the performance of different versions of AdaGrad with and without clipping on the 1-dimensional quadratic objective with additive heavy-tailed noise: $f(x) = x^2/2$, $\nabla f_\xi(x) = x + \xi$, where the noise ξ has probability density function $p(t) = \frac{3}{4(1+|t|)^{2.5}}$. In this case, Assumption 1.1 is satisfied with any $\alpha \in (1, 1.5)$ and the α -th moment is unbounded for $\alpha \geq 1.5$. Moreover, the function is strongly convex and L -smooth with $L = 1$. We choose $x_0 = 2$, $b_0 = 3$ (for the versions of AdaGrad with delay), $b_{-1} = 3$ (for other cases), $\lambda = 1/2$ for the methods with clipping, and choose γ from $\{1, 1/16, 1/128\}$. Each method was run 100 times with different seeds.

The results are given in Figure 1, where for each method, we show its trajectory in terms of the squared distance to the solution for $\gamma = 1$ and $\gamma = 1/16$ (the results for $\gamma = 1/128$ are given in Appendix D.1). Solid lines correspond to the median value of the squared distances, and the error bands cover the areas from the 10-th to 90-th percentiles of $(x_t - x^*)^2$. These results show that clipped versions of AdaGrad (with and without delay) achieve better convergence with higher probability than their non-clipped counterparts. Moreover, versions with clipping exhibit similar behavior to each other. That is, the error bands for Clip-AdaGrad(D) are lower than for AdaGrad(D) (note that the vertical axis is shown in the logarithmic scale making the error bands for Clip-AdaGrad(D) look wider than for AdaGrad(D), while they are not). In general, the observed results for AdaGrad-type methods are perfectly aligned with the theory developed in this paper. We provide the results for Adam with and without clipping/delay in Appendix D.1.

ALBERT Base v2 fine-tuning. In the second part of our experiments, we consider fine-tuning the pre-trained ALBERT Base v2 model (Lan et al., 2019) on CoLa and RTE datasets (Wang et al., 2018). Since Adam-based algorithms are the methods of choice for NLP tasks, in the main part of the paper, we focus on Adam and its clipped versions – Clip-Adam and Clip-AdamD – and provide additional experiments with AdaGrad-based methods in Appendix D.2. We took a pre-trained model from the Hugging Face library. Then, the model was fine-tuned following the methodology

Clipping Improves Adam-Norm and AdaGrad-Norm

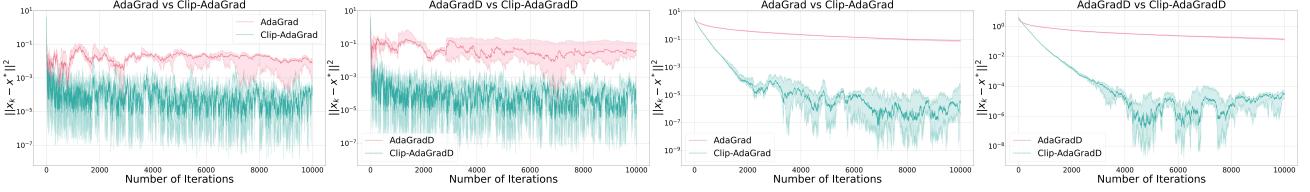


Figure 1: Performance of different versions of **AdaGrad** (with and without clipping/delay) with stepsizes $\gamma = 1$ (two left plots) and $\gamma = 1/16$ (two right plots) on the quadratic problem.

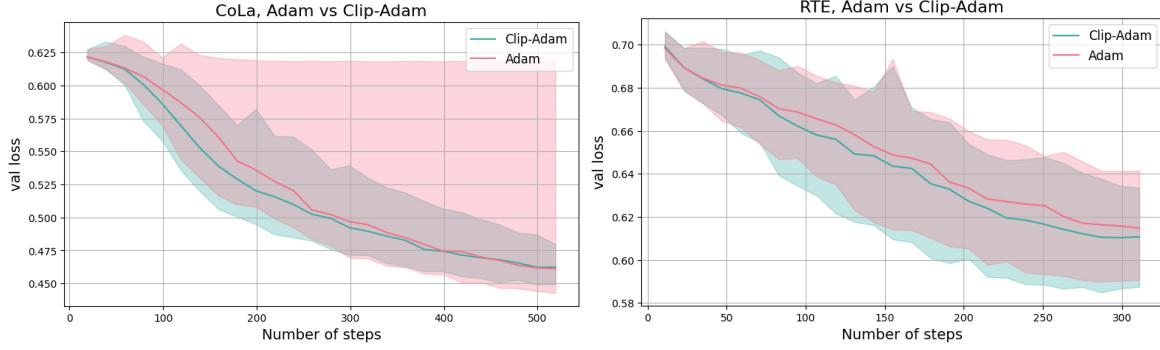


Figure 2: Validation loss for ALBERT Base v2 fine-tuning task on the CoLa and RTE datasets.

suggested by [Mosbach et al. \(2020\)](#). For the methods with clipping, we used the same batchsize and stepsize as for **Adam** and tuned the clipping level for the two types of clipping⁵. In the main text, we show the results with layer-wise clipping. Further details and additional results are deferred to Appendix D.2.

Before comparing the methods, we ran **Adam** and checked how heavy-tailed the noise in the stochastic gradients is along the trajectory. In particular, for both tasks, we selected 4 iterates corresponding to the starting point, points generated after $\approx 1/3$ and $\approx 2/3$ of all steps, and the last iterate. Then, for each of these points, we sampled size-16 (for CoLa) and size-8 (for RTE) mini-batched estimator $\nabla f_\xi(x)$ of the gradient 1000 times, saved the resulting norms of the differences $\|\nabla f_\xi(x) - \nabla f(x)\|$, and plotted their histogram, i.e., we plotted the histograms of the noise norm. Moreover, we also measure the heavy-tailedness of the noise following the approach from ([Gorbunov et al., 2022](#)): we compute two metrics $p_{mR} = F_{1.5}(\|\nabla f_\xi(x) - \nabla f(x)\|)$, which quantifies ‘‘mild’’ heavy tails, and $p_{eR} = F_3(\|\nabla f_\xi(x) - \nabla f(x)\|)$ introduced by [Jordanova & Petkova \(2017\)](#), which quantifies ‘‘extreme’’ heavy tails, where $F_a(\|\nabla f_\xi(x) - \nabla f(x)\|) = \mathbb{P}\{\|\nabla f_\xi(x) - \nabla f(x)\| > Q_3 + a(Q_3 - Q_1)\}$ and Q_i is the i -th quartile of $\|\nabla f_\xi(x) - \nabla f(x)\|$. To illustrate the heavy-tailedness clearly, we divide these metrics to the ones

⁵We did not consider the global/norm clipping (the considered in theory), since typically coordinate-wise or layer-wise clipping work better in fine-tuning, see ([Lv et al., 2024](#)).

computed for the standard normal distribution (p_{mRN} and p_{eRN}) and show $\rho_{mR} = p_{mR}/p_{mRN}$ and $\rho_{eR} = p_{eR}/p_{eRN}$ on the plots. The histograms are provided in Figure 6 (see Appendix D.2). They show that the noise distribution has much heavier tails for CoLa than for RTE.

Then, similarly to the experiments with the quadratic problem, we ran the methods 100 times, and for each step, we computed the median value of the validation loss and its 5-th and 95-th percentiles. The results are presented in Figure 2, where the solid lines correspond to the medians and the error bands cover the areas between 5-th and 95-th percentiles. As expected, **Adam** exhibits poor high-probability convergence on the CoLa datasets where the noise is significantly heavy-tailed, and **Clip-Adam** shows much better performance: the area between 5-th and 95-th percentiles is relatively narrow for **Clip-Adam**. In contrast, for the RTE dataset, **Clip-Adam** performs similarly to **Adam**. This is also expected since the noise is much less heavy for RTE, as Figure 6 shows. Taking into account the negative results from Section 2, and the upper bounds from Section 3, we conclude that these numerical results are well-aligned with the theory developed in the paper.

In Appendix D.3, we also provide additional experiments with the fine-tuning the 355M parameter RoBERTa Large model ([Liu et al., 2019](#)) on the two GLUE ([Wang et al., 2018](#)) datasets: QNLI (116k question-answer pairs) and CoLa (10.7k linguistic acceptability examples). Similarly to the previous results, the clipped variants consistently outperform their unclipped counterparts for the larger model.

Acknowledgements

The work of Savelii Chezhegov and Aleksandr Beznosikov on the final version of this paper was supported by the Ministry of Economic Development of the Russian Federation (agreement with MIPT No. 139-15-2025-013, dated June 20, 2025, IGK 000000C313925P4B0002).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

- Abdukhakimov, F., Xiang, C., Kamzolov, D., Gower, R., and Takáč, M. Sania: Polyak-type optimization framework leads to scale invariant stochastic algorithms. *arXiv preprint arXiv:2312.17369*, 2023.
- Abdukhakimov, F., Xiang, C., Kamzolov, D., and Takáč, M. Stochastic gradient descent with preconditioned polyak step-size. *Computational Mathematics and Mathematical Physics*, 64(4):621–634, 2024.
- Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):165–214, 2023.
- Bennett, G. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.
- Chen, X., Liu, S., Sun, R., and Hong, M. On the convergence of a class of adam-type algorithms for non-convex optimization. *arXiv preprint arXiv:1808.02941*, 2018.
- Cutkosky, A. and Mehta, H. High-probability bounds for non-convex stochastic optimization with heavy tails. *Advances in Neural Information Processing Systems*, 34:4883–4895, 2021.
- Davis, D., Drusvyatskiy, D., Xiao, L., and Zhang, J. From low probability to high confidence in stochastic convex optimization. *The Journal of Machine Learning Research*, 22(1):2237–2274, 2021.
- Défossez, A., Bottou, L., Bach, F., and Usunier, N. A simple convergence proof of adam and adagrad. *Transactions on Machine Learning Research*, 2022.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.
- Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.
- Dzhaparidze, K. and Van Zanten, J. On bernstein-type inequalities for martingales. *Stochastic processes and their applications*, 93(1):109–117, 2001.
- Faw, M., Tziotis, I., Caramanis, C., Mokhtari, A., Shakkottai, S., and Ward, R. The power of adaptivity in sgd: Self-tuning step sizes with unbounded gradients and affine variance. In *Conference on Learning Theory*, pp. 313–355. PMLR, 2022.
- Faw, M., Rout, L., Caramanis, C., and Shakkottai, S. Beyond uniform smoothness: A stopped analysis of adaptive sgd. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 89–160. PMLR, 2023.
- Freedman, D. A. et al. On tail probabilities for martingales. *the Annals of Probability*, 3(1):100–118, 1975.
- Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.
- Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.
- Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.
- Gorbunov, E., Danilova, M., and Gasnikov, A. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020.
- Gorbunov, E., Danilova, M., Shibaev, I., Dvurechensky, P., and Gasnikov, A. Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. *arXiv preprint arXiv:2106.05958*, 2021.
- Gorbunov, E., Danilova, M., Dobre, D., Dvurechenskii, P., Gasnikov, A., and Gidel, G. Clipped stochastic methods for variational inequalities with heavy-tailed noise. *Advances in Neural Information Processing Systems*, 35:31319–31332, 2022.
- Gorbunov, E., Sadiev, A., Danilova, M., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A., and Richtárik, P. High-probability convergence for composite and

- distributed stochastic minimization and variational inequalities with heavy-tailed noise. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15951–16070. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/gorbunov24a.html>.
- Harvey, N. J., Liaw, C., and Randhawa, S. Simple and optimal high-probability bounds for strongly-convex stochastic gradient descent. *arXiv preprint arXiv:1909.00843*, 2019.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Huber, P. J. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.
- Jakovetić, D., Bajović, D., Sahu, A. K., Kar, S., Milošević, N., and Stamenković, D. Nonlinear gradient mappings and stochastic optimization: A general framework with applications to heavy-tail noise. *SIAM Journal on Optimization*, 33(2):394–423, 2023.
- Jordanova, P. K. and Petkova, M. P. Measuring heavy-tailedness of distributions. In *AIP Conference Proceedings*, volume 1910. AIP Publishing, 2017.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Kornilov, N., Dorn, Y., Lobanov, A., Kutuzov, N., Shibaev, I., Gorbunov, E., Gasnikov, A., and Nazin, A. Zeroth-order median clipping for non-smooth convex optimization problems with heavy-tailed symmetric noise. *arXiv preprint arXiv:2402.02461*, 2024.
- Kunstner, F., Chen, J., Lavington, J. W., and Schmidt, M. Noise is not the main factor behind the gap between sgd and adam on transformers, but sign descent might be. In *The Eleventh International Conference on Learning Representations*, 2023.
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.
- Li, H., Rakhlin, A., and Jadbabaie, A. Convergence of adam under relaxed assumptions. *Advances in Neural Information Processing Systems*, 36, 2023.
- Li, S. and Liu, Y. High probability analysis for non-convex stochastic optimization with clipping. In *ECAI 2023*, pp. 1406–1413. IOS Press, 2023.
- Li, S., Swartworth, W. J., Takáč, M., Needell, D., and Gower, R. M. Sp2: A second order stochastic polyak method. *ICLR 2023*, 2022.
- Li, X. and Orabona, F. A high probability analysis of adaptive sgd with momentum. *arXiv preprint arXiv:2007.14294*, 2020.
- Li, Y., Yuan, R., Fan, C., Schmidt, M., Horváth, S., Gower, R. M., and Takáč, M. Enhancing policy gradient with the polyak step-size adaption. *arXiv preprint arXiv:2404.07525*, 2024.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach, 2019. URL <https://arxiv.org/abs/1907.11692>.
- Liu, Z., Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. High probability convergence of stochastic gradient methods. In *International Conference on Machine Learning*, pp. 21884–21914. PMLR, 2023.
- Loizou, N., Vaswani, S., Laradji, I. H., and Lacoste-Julien, S. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pp. 1306–1314. PMLR, 2021.
- Lv, K., Yang, Y., Liu, T., Gao, Q., Guo, Q., and Qiu, X. Full parameter fine-tuning for large language models with limited resources, 2024. URL <https://arxiv.org/abs/2306.09782>.
- Mosbach, M., Andriushchenko, M., and Klakow, D. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*, 2020.
- Moskvoretskii, V., Panchenko, A., and Nikishina, I. Are large language models good at lexical semantics? a case of taxonomy learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 1498–1510, 2024a.
- Moskvoretskii, V., Tupitsa, N., Biemann, C., Horváth, S., Gorbunov, E., and Nikishina, I. Low-resource machine translation through the lens of personalized federated learning. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8806–8825, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.

514. URL <https://aclanthology.org/2024.findings-emnlp.514/>.
- Nazin, A. V., Nemirovsky, A. S., Tsybakov, A. B., and Juditsky, A. B. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80:1607–1627, 2019.
- Nemirovski, A. S., Juditsky, A. B., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.
- Nemirovskij, A. S. and Yudin, D. B. Problem complexity and method efficiency in optimization. 1983.
- Nesterov, Y. E. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In *Doklady Akademii Nauk*, volume 269, pp. 543–547. Russian Academy of Sciences, 1983.
- Nguyen, T. D., Ene, A., and Nguyen, H. L. Improved convergence in high probability of clipped gradient methods with heavy tails. *arXiv preprint arXiv:2304.01119*, 2023.
- Nikishina, I., Vakhitova, A., Tutubalina, E., and Panchenko, A. Cross-modal contextualized hidden state projection method for expanding of taxonomic graphs. In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, pp. 11–24, 2022.
- Pan, Y. and Li, Y. Toward understanding why adam converges faster than sgd for transformers. *arXiv preprint arXiv:2306.00204*, 2023.
- Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pp. 1310–1318. Pmlr, 2013.
- Patel, V. and Berahas, A. S. Gradient descent in the absence of global lipschitz continuity of the gradients. *arXiv preprint arXiv:2210.02418*, 2022.
- Patel, V., Zhang, S., and Tian, B. Global convergence and stability of stochastic gradient descent. *Advances in Neural Information Processing Systems*, 35:36014–36025, 2022.
- Polyak, B. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.
- Puchkin, N., Gorbunov, E., Kutuzov, N., and Gasnikov, A. Breaking the heavy-tailed noise barrier in stochastic optimization problems. In *International Conference on Artificial Intelligence and Statistics*, pp. 856–864. PMLR, 2024.
- Reddi, S. J., Kale, S., and Kumar, S. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.
- Robbins, H. and Monro, S. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.
- Sadiev, A., Danilova, M., Gorbunov, E., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A., and Richtárik, P. 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. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29563–29648. PMLR, 23–29 Jul 2023.
- Schaipp, F., Ohana, R., Eickenberg, M., Defazio, A., and Gower, R. M. Momo: Momentum models for adaptive learning rates. *arXiv preprint arXiv:2305.07583*, 2023.
- Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- Shi, Z., Sadiev, A., Loizou, N., Richtárik, P., and Takáč, M. Ai-sarah: Adaptive and implicit stochastic recursive gradient methods. *Transactions on Machine Learning Research*, 2023.
- Streeter, M. and McMahan, H. B. Less regret via online conditioning. *arXiv preprint arXiv:1002.4862*, 2010.
- Takáč, M., Bijral, A., Richtárik, P., and Srebro, N. Mini-batch primal and dual methods for svms. In *In 30th International Conference on Machine Learning, ICML 2013*, 2013.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.
- Wang, B., Zhang, H., Ma, Z., and Chen, W. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In *The Thirty Sixth Annual*

Conference on Learning Theory, pp. 161–190. PMLR, 2023.

Wang, B., Zhang, Y., Zhang, H., Meng, Q., Sun, R., Ma, Z.-M., Liu, T.-Y., Luo, Z.-Q., and Chen, W. Provable adaptivity of adam under non-uniform smoothness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2960–2969, 2024.

Ward, R., Wu, X., and Bottou, L. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21(219):1–30, 2020.

Yang, C. and Ma, X. Improving stability of fine-tuning pretrained language models via component-wise gradient norm clipping, 2022. URL <https://arxiv.org/abs/2210.10325>.

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., and Hsieh, C.-J. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

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

Zhang, Y., Chen, C., Shi, N., Sun, R., and Luo, Z.-Q. Adam can converge without any modification on update rules. *Advances in neural information processing systems*, 35: 28386–28399, 2022.

Zhou, P., Feng, J., Ma, C., Xiong, C., Hoi, S. C. H., et al. Towards theoretically understanding why sgd generalizes better than adam in deep learning. *Advances in Neural Information Processing Systems*, 33:21285–21296, 2020.

Zou, F., Shen, L., Jie, Z., Sun, J., and Liu, W. Weighted adagrad with unified momentum. *arXiv preprint arXiv:1808.03408*, 2, 2018.

A. Technical Details and Auxiliary Results

Additional notation. For the ease of exposition, we introduce the following notation for the proofs:

$$\begin{aligned} g_t &= \text{clip}(\nabla f_{\xi_t}(x_t), \lambda), \\ \theta_t &= g_t - \nabla f(x_t), \\ \theta_t^u &= g_t - \mathbb{E}_{\xi_t}[g_t], \\ \theta_t^b &= \mathbb{E}_{\xi_t}[g_t] - \nabla f(x_t), \\ R_t &= \|x_t - x^*\|, \\ \Delta_t &= f(x_t) - f_*. \end{aligned}$$

Auxiliary results. We also use the following standard results.

Proposition A.1 (Young's inequality.). For any $x, y \in \mathbb{R}^d$ and $p > 0$ the following inequality holds:

$$\|x + y\|^2 \leq (1 + p) \|x\|^2 + \left(1 + \frac{1}{p}\right) \|y\|^2.$$

In particular, for $p = 1$

$$\|x + y\|^2 \leq 2\|x\|^2 + 2\|y\|^2.$$

Lemma A.2 (Lemma B.2 from (Défossez et al., 2022)). Let $0 \leq a \leq b$ be some non-negative integers and $0 \leq q < 1$. Then,

$$\sum_{k=a}^b q^k k \leq \frac{q}{(1-q)^2}.$$

Lemma A.3 (Lemma 1 from (Streeter & McMahan, 2010)). Let $\{a_i\}_{i=1}^n$ and c be non-negative reals. Then,

$$\sum_{k=1}^n \frac{a_k}{\sqrt{c + \sum_{i=1}^k a_i}} \leq 2 \sqrt{c + \sum_{k=1}^n a_k}$$

The following lemma by Sadiev et al. (2023) helps to estimate bias and variance of the clipped stochastic gradient satisfying Assumption 1.1.

Lemma A.4 (Lemma 5.1 from (Sadiev et al., 2023)). Let X be a random vector from \mathbb{R}^d and $\widehat{X} = \text{clip}(X, \lambda)$. Then, $\|\widehat{X} - \mathbb{E}[\widehat{X}]\| \leq 2\lambda$. Moreover, if for some $\sigma \geq 0$ and $\alpha \in (1, 2]$ we have $\mathbb{E}[X] = x \in \mathbb{R}^d$, $\mathbb{E}[\|X - x\|^\alpha] \leq \sigma^\alpha$, and $\|x\| \leq \frac{\lambda}{2}$, then

$$\begin{aligned} \|\mathbb{E}[\widehat{X}] - x\| &\leq \frac{2^\alpha \sigma^\alpha}{\lambda^{\alpha-1}}, \\ \mathbb{E}\left[\|\widehat{X} - x\|^2\right] &\leq 18\lambda^{2-\alpha}\sigma^\alpha, \\ \mathbb{E}\left[\|\widehat{X} - \mathbb{E}[\widehat{X}]\|^2\right] &\leq 18\lambda^{2-\alpha}\sigma^\alpha. \end{aligned}$$

Finally, in the analysis of Clip-RAdaGradD, we face the sums of martingale-difference sequences. One of the tools that we use to handle them is Bernstein's inequality (Bennett, 1962; Dzhaparidze & Van Zanten, 2001; Freedman et al., 1975).

Lemma A.5 (Bernstein's inequality). Let the sequence of random variables $\{X_i\}_{i \geq 1}$ form a martingale difference sequence, i.e., $\mathbb{E}[X_i | X_{i-1}, \dots, X_1] = 0$ for all $i \geq 1$. Assume that conditional variances $\sigma_i^2 = \mathbb{E}[X_i^2 | X_{i-1}, \dots, X_1]$ exist and are bounded and also assume that there exists deterministic constant $c > 0$ such that $|X_i| \leq c$ almost surely for all $i \geq 1$. Then for all $b > 0$, $G > 0$ and $n \geq 1$

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

B. Missing Proofs from Section 2

In this section, we provide further details regarding Theorem 2.1 giving a negative result about high-probability convergence of **Adam/M-AdaGrad** and **AdamD/M-AdaGradD**. For all methods, we use the 1-dimensional Huber loss function:

$$f(x) = \begin{cases} \frac{1}{2}x^2, & \text{if } |x| \leq \nu, \\ \nu(|x| - \frac{1}{2}\nu), & \text{otherwise.} \end{cases}$$

This function is convex and L -smooth with $L = 1$. However, the construction of noises and proofs are different for **Adam**, **M-AdaGrad**, **AdamD**, and **M-AdaGradD**. Therefore, we provide the negative results for these methods separately in the following subsections. We emphasize that the constructed examples are 1-dimensional, meaning that they hold for coordinate-wise and norm-versions of the considered methods.

B.1. Failure of M-AdaGrad

We start with the following lemma giving a closed-form expression for the iterates of deterministic **M-AdaGrad** applied to (8).

Lemma B.1. *Suppose that the starting point x_0 is such that $x_0 > 0$. If after T iterations of deterministic **M-AdaGrad** with initial momentum⁶ $m_{-1} \geq 0$ we have $|x_t| > \nu$ and $x_t > 0$ for all $t = \overline{1, T-1}$, then*

$$x_T = x_0 - \gamma\nu \sum_{t=0}^{T-1} \frac{1 - \beta_1^{t+1} + \beta_1^{t+1} \frac{m_{-1}}{\nu}}{\sqrt{b_{-1}^2 + (t+1)\nu^2}}.$$

Proof. Since $|x_t| > \nu$ and x_t is positive, the gradient at x_t is equal to ν . Hence, by substituting the gradient into the algorithm, we get the final result. \square

The above lemma relies on the condition that $|x_t| > \nu$ and $x_t > 0$ for all $t = \overline{1, T-1}$. For any γ, b_{-1} and T this condition can be achieved if we choose sufficiently small ν .

Next, we estimate the interval where x_T lies.

Lemma B.2. *Let the conditions of Lemma B.1 hold. Then, we have*

$$\begin{aligned} x_T &\geq x_0 - \gamma \left(\frac{1}{\sqrt{1+a_0}} + 2\sqrt{a_0+T} - 2\sqrt{a_0+1} + \frac{m_{-1}\beta_1}{\nu(1-\beta_1)} \right), \\ x_T &\leq x_0 - \gamma(1-\beta_1) \left(2\sqrt{a_0+T+1} - 2\sqrt{a_0+1} \right), \end{aligned}$$

where $a_0 = \frac{b_{-1}^2}{\nu^2}$.

Proof. From Lemma B.1 we have:

$$x_T = x_0 - \gamma \sum_{t=0}^{T-1} \frac{1 - \beta_1^{t+1} + \beta_1^{t+1} \frac{m_{-1}}{\nu}}{\sqrt{a_0 + (t+1)}},$$

where $a_0 = \frac{b_{-1}^2}{\nu^2}$. Next, we bound the second term in the following way:

$$\sum_{t=0}^{T-1} \frac{1 - \beta_1^{t+1} + \beta_1^{t+1} \frac{m_{-1}}{\nu}}{\sqrt{a_0 + (t+1)}} \geq (1 - \beta_1) \int_{a_0}^{a_0+T} \frac{1}{\sqrt{1+x}} dx = (1 - \beta_1) (2\sqrt{a_0+T+1} - 2\sqrt{a_0+1}), \quad (13)$$

⁶By default, $m_{-1} = 0$ but the result of this lemma holds for non-zero $m_{-1} \geq 0$ as well. If $m_{-1} \leq 0$, then one can select $x_0 < 0$.

$$\begin{aligned}
 \sum_{t=0}^{T-1} \frac{1 - \beta_1^{t+1} + \beta_1^{t+1} \frac{m_{-1}}{\nu}}{\sqrt{a_0 + (t+1)}} &= \sum_{t=0}^{T-1} \frac{1 - \beta_1^{t+1}}{\sqrt{a_0 + (t+1)}} + \sum_{t=0}^{T-1} \frac{\beta_1^{t+1} \frac{m_{-1}}{\nu}}{\sqrt{a_0 + (t+1)}} \\
 &\leq \frac{1}{\sqrt{1+a_0}} + \int_{a_0}^{a_0+T-1} \frac{1}{\sqrt{1+x}} dx + \sum_{t=0}^{T-1} \beta_1^{t+1} \frac{m_{-1}}{\nu} \\
 &\leq \frac{1}{\sqrt{1+a_0}} + 2\sqrt{a_0+T} - 2\sqrt{a_0+1} + \frac{m_{-1}\beta_1}{\nu(1-\beta_1)}. \tag{14}
 \end{aligned}$$

Combining (13) and (14), we get the final result. \square

Corollary B.3. If $x_0 - \gamma - \nu - \frac{\gamma m_{-1} \beta_1}{\nu(1-\beta_1)} > 0$ and

$$T < \frac{\left(x_0 - \nu - \gamma - \frac{\gamma m_{-1} \beta_1}{\nu(1-\beta_1)}\right)^2 + 4\gamma \left(x_0 - \nu - \gamma - \frac{\gamma m_{-1} \beta_1}{\nu(1-\beta_1)}\right) \sqrt{a_0+1}}{4\gamma^2} + 1,$$

then $x_T > \nu$ for deterministic M-AdaGrad. Alternatively, $|x_T| \leq \nu$ implies that

$$T \geq \frac{\left(x_0 - \nu - \gamma - \frac{\gamma m_{-1} \beta_1}{\nu(1-\beta_1)}\right)^2 + 4\gamma \left(x_0 - \nu - \gamma - \frac{\gamma m_{-1} \beta_1}{\nu(1-\beta_1)}\right) \sqrt{a_0+1}}{4\gamma^2} + 1.$$

Proof. First, let us show that

$$\nu < x_0 - \gamma \left(1 + 2\sqrt{a_0+T} - 2\sqrt{a_0+1} + \frac{m_{-1}\beta_1}{\nu(1-\beta_1)}\right) \tag{15}$$

is equivalent to

$$T < \frac{\left(x_0 - \nu - \gamma - \frac{\gamma m_{-1} \beta_1}{\nu(1-\beta_1)}\right)^2 + 4\gamma \left(x_0 - \nu - \gamma - \frac{\gamma m_{-1} \beta_1}{\nu(1-\beta_1)}\right) \sqrt{a_0+1}}{4\gamma^2} + 1.$$

Rewriting the (15), one can obtain

$$2\gamma\sqrt{a_0+T} < x_0 - \nu - \gamma - \frac{\gamma m_{-1} \beta_1}{\nu(1-\beta_1)} + 2\gamma\sqrt{a_0+1}.$$

Squaring both parts of the inequality above and expressing T , we get the alternative equivalent formula. Noticing that $1 \geq \frac{1}{\sqrt{1+a_0}}$ and applying Lemma B.2, we get the final result. The second part of the corollary is just a negation of the implication stated in the first part of the corollary. \square

Theorem B.4. For any $\varepsilon, \delta \in (0, 1)$, $\sigma > 0$ such that $\sigma/\sqrt{\varepsilon\delta} \geq 8$, there exists convex L -smooth minimization problem (8) and stochastic gradient oracle such that Assumption 1.1 holds with $\alpha = 2$ and the iterates produced by M-AdaGrad after K steps with stepsize γ and starting point x_0 such that $R := x_0 - \sqrt{2\varepsilon} - 3\gamma \geq 3\gamma\beta_1 + \sqrt{9\gamma^2\beta_1^2 + 4\gamma^2\beta_1 K}$ satisfy the following implication:

$$\mathbb{P}\{f(x_K) - f(x^*) \geq \varepsilon\} \leq \delta \implies K = \Omega\left(\frac{b_{-1}R}{\sqrt{\varepsilon\gamma}} + \frac{\sigma R}{\gamma\sqrt{\varepsilon\delta}}\right), \tag{16}$$

i.e., the high-probability complexity of M-AdaGrad has inverse-power dependence on δ .

Proof. Before we delve into the technical details, we provide an intuition behind the proof. We want to use the lower bound from Corollary B.3 and estimate the bound for the number of iterations required to achieve the desired optimization error ε with probability at least $1 - \delta$. Moreover, we need to set ν depending on the accuracy ε (ν is analytically clarified later).

We denote the output of deterministic **M-AdaGrad** after t iterations as \hat{x}_t . Then, we introduce the noise in the stochastic gradient in the following way

$$g_k = \nabla f(x_k) - \sigma \xi_k,$$

where

$$\xi_k = \begin{cases} 0, & \text{for } k > 0, \\ \begin{cases} -A, & \text{with probability } \frac{1}{2A^2} \\ 0, & \text{with probability } 1 - \frac{1}{A^2} \\ A, & \text{with probability } \frac{1}{2A^2} \end{cases} & \text{otherwise,} \end{cases} \quad (17)$$

where the formula for A is given later. The noise construction (17) implies that stochasticity appears only at the first iteration of **M-AdaGrad**, and then it only affects the stepsizes. Therefore,

$$x_1 = x_0 - \frac{\gamma}{b_0} m_0,$$

where $b_0 = \sqrt{b_{-1}^2 + (\nu - \sigma \xi_0)^2}$ and $m_0 = (1 - \beta_1)(\nu - \sigma \xi_0)$. Moreover, x_1 can be bounded in the following way

$$x_0 + \gamma > x_1 > x_0 - \gamma.$$

Also, let us define K_0 as the number of iterations required to achieve at least ε -accuracy. According to the stochasticity construction, we get that for $\xi_0 \neq A$ the momentum m_0 is non-negative:

$$\begin{aligned} m_0 &= \begin{cases} (1 - \beta_1)(\nu + \sigma A), & \text{if } \xi_0 = -A \\ (1 - \beta_1)\nu, & \text{if } \xi_0 = 0 \end{cases} \\ &\geq 0. \end{aligned}$$

Therefore, choosing x_0 in such a way that $x_0 - 2\gamma - \nu - \frac{\gamma m_0 \beta_1}{\nu(1-\beta_1)} \geq 0$, we can apply Corollary B.3 and obtain that K_0 can be chosen as

$$K_0 = \frac{\left(x_0 - \nu - \gamma - \frac{\gamma m_0 \beta_1}{\nu(1-\beta_1)} \right) \sqrt{a_1}}{\gamma}$$

with $a_1 = \frac{b_0^2}{\nu^2}$ and $\varepsilon = \frac{\nu^2}{2}$. Let us specify that this estimate depends on the stochasticity at the first iteration, i.e., the bound on the number of iterations is random. Consequently, if **M-AdaGrad** achieves ε -solution after K steps, we should have $K \geq K_0$. Therefore, $\mathbb{P}\{K \geq K_0\} \geq \mathbb{P}\{f(x_K) - f(x^*) \leq \varepsilon\}$ and we want to estimate K such that

$$\mathbb{P}\{K_0 \leq K\} \geq 1 - \delta.$$

Bounding the left-hand side,

$$\begin{aligned} \mathbb{P}\{K_0 \leq K\} &= \mathbb{P}\{K_0 \leq K | \xi_0 = A\} \mathbb{P}\{\xi_0 = A\} + \mathbb{P}\{K_0 \leq K | \xi_0 \neq A\} \mathbb{P}\{\xi_0 \neq A\} \\ &\leq \mathbb{P}\left\{ \frac{\left(x_0 - \nu - \gamma - \frac{\gamma m_0 \beta_1}{\nu(1-\beta_1)} \right) \sqrt{a_1}}{\gamma} \leq K \middle| \xi_0 \neq A \right\} \mathbb{P}\{\xi_0 \neq A\} + \mathbb{P}\{\xi_0 = A\} \\ &\leq \mathbb{P}\left\{ \frac{\left(x_0 - \nu - 2\gamma - \frac{\gamma m_0 \beta_1}{\nu(1-\beta_1)} \right) \sqrt{a_1}}{\gamma} \leq K \middle| \xi_0 \neq A \right\} \mathbb{P}\{\xi_0 \neq A\} + \mathbb{P}\{\xi_0 = A\}. \end{aligned}$$

Denoting $R = x_0 - \nu - 2\gamma$ and assuming that for $\xi_0 \neq A$ we have $R \geq \frac{2\gamma m_0 \beta_1}{\nu(1-\beta_1)}$, which implies $R - \frac{\gamma m_0 \beta_1}{\nu(1-\beta_1)} \geq \frac{R}{2}$, we derive

$$\begin{aligned} \mathbb{P}\{K_0 \leq K\} &\leq \mathbb{P}\left\{\frac{\left(x_0 - \nu - 2\gamma - \frac{\gamma m_0 \beta_1}{\nu(1-\beta_1)}\right) \sqrt{a_1}}{\gamma} \leq K \middle| \xi_0 \neq A\right\} \mathbb{P}\{\xi_0 \neq A\} + \mathbb{P}\{\xi_0 = A\} \\ &\leq \mathbb{P}\left\{\frac{R}{2\gamma} \sqrt{a_1} \leq K \middle| \xi_0 \neq A\right\} \mathbb{P}\{\xi_0 \neq A\} + \mathbb{P}\{\xi_0 = A\} \\ &= \mathbb{P}\left\{b_{-1}^2 + (\nu - \sigma \xi_0)^2 \leq \frac{4K^2 \nu^2 \gamma^2}{R^2} \middle| \xi_0 \neq A\right\} \mathbb{P}\{\xi_0 \neq A\} + \mathbb{P}\{\xi_0 = A\} \\ &\leq \mathbb{P}\left\{(\nu - \sigma \xi_0)^2 \leq \frac{4K^2 \nu^2 \gamma^2}{R^2} \middle| \xi_0 \neq A\right\} \mathbb{P}\{\xi_0 \neq A\} + \mathbb{P}\{\xi_0 = A\}, \end{aligned}$$

where in the third row, we substitute the analytical form of a_1 , and in the fourth row, we used $K \geq \frac{b_{-1} R}{4\nu\gamma}$. Therefore, we get

$$\begin{aligned} \mathbb{P}\{K_0 \leq K\} &\leq \mathbb{P}\left\{|\sigma \xi_0 - \nu| \leq \frac{2K\nu\gamma}{R} \middle| \xi_0 \neq A\right\} \mathbb{P}\{\xi_0 \neq A\} + \mathbb{P}\{\xi_0 = A\} \\ &\leq \mathbb{P}\left\{\sigma |\xi_0| \leq \frac{2K\nu\gamma}{R} + \nu \middle| \xi_0 \neq A\right\} \mathbb{P}\{\xi_0 \neq A\} + \mathbb{P}\{\xi_0 = A\}. \end{aligned}$$

As a result, $\mathbb{P}\{K_0 \leq K\} \geq 1 - \delta$ implies

$$\mathbb{P}\left\{\sigma |\xi_0| \leq \frac{2K\nu\gamma}{R} + \nu \middle| \xi_0 \neq A\right\} \mathbb{P}\{\xi_0 \neq A\} + \mathbb{P}\{\xi_0 = A\} \geq 1 - \delta.$$

Consequently, choosing $A = \frac{\frac{2K\nu\gamma}{R} + 2\nu}{\sigma}$, the first probability in the inequality above is equal to $1 - \frac{1}{A^2}$, since the only $\xi_0 = 0$ satisfies the condition on random variable. Hence, we have

$$1 - \frac{1}{2A^2} \geq 1 - \delta.$$

Consequently,

$$\frac{1}{A} = \frac{\sigma}{\frac{2K\nu\gamma}{R} + 2\nu} \leq \sqrt{2\delta}.$$

Therefore,

$$K \geq \frac{R}{\gamma} \left(\frac{\sigma}{2\nu\sqrt{2\delta}} - 1 \right) \geq \frac{R\sigma}{8\gamma\sqrt{\varepsilon\delta}},$$

since $\sigma/\sqrt{\varepsilon\delta} \geq 8$ and $\nu = \sqrt{2\varepsilon}$. It remains to find the conditions on x_0 and γ ensuring that $R \geq \frac{2\gamma m_0 \beta_1}{\nu(1-\beta_1)}$ for $\xi_0 \neq A$. It is sufficient to choose R in the following way:

$$R \geq \frac{2\gamma(\nu + \sigma A)\beta_1}{\nu} = 2\gamma \left(3 + \frac{2K\gamma}{R} \right) \beta_1.$$

Solving the quadratic inequality in R , we get $R \geq 3\gamma\beta_1 + \sqrt{9\gamma^2\beta_1^2 + 4\gamma^2\beta_1 K}$. Choosing x_0 such that the previous inequality is satisfied, we conclude the proof. \square

B.2. Failure of M-AdaGradD

Similarly to the case of M-AdaGrad, we start by obtaining the analytic form of iterations of the deterministic M-AdaGradD in the following lemma.

Lemma B.5. Suppose that starting point x_0 is such that $x_0 > 0$. If after T iterations of deterministic M-AdaGradD we have $|x_t| > \nu$ and $x_t > 0$ for all $t = \overline{1, T-1}$ with , then

$$x_T = x_0 - \gamma \nu \sum_{t=0}^{T-1} \frac{1 - \beta_1^{t+1}}{\sqrt{b_0^2 + t\nu^2}}.$$

Proof. The proof is similar to the proof of Lemma B.1. Since $x_t > \nu$, the gradient at point x_t is equal to ν . Substituting that into the iteration of M-AdaGradD for each t , we finish the proof. \square

Now, let us estimate the interval where x_T lies.

Lemma B.6. Let the conditions of Lemma B.5 hold. Then, we have

$$x_0 - \gamma \left(\frac{1}{\sqrt{a_0}} + 2\sqrt{a_0 + T - 1} - 2\sqrt{a_0} \right) \leq x_T \leq x_0 - \gamma(1 - \beta_1) \left(2\sqrt{a_0 + T} - 2\sqrt{a_0} \right),$$

where $a_0 = \frac{b_0^2}{\nu^2}$.

Proof. Let us start with Lemma B.5:

$$x_T = x_0 - \gamma \sum_{t=0}^{T-1} \frac{1 - \beta_1^{t+1}}{\sqrt{a_0 + t}},$$

where $a_0 = \frac{b_0^2}{\nu^2}$. Next, we bound the second term in the following way:

$$\sum_{t=0}^{T-1} \frac{1 - \beta_1^{t+1}}{\sqrt{a_0 + t}} \geq (1 - \beta_1) \int_{a_0}^{a_0 + T} \frac{1}{\sqrt{x}} dx = (1 - \beta_1)(2\sqrt{a_0 + T} - 2\sqrt{a_0}), \quad (18)$$

$$\sum_{t=0}^{T-1} \frac{1 - \beta_1^{t+1}}{\sqrt{a_0 + t}} \leq \frac{1}{\sqrt{a_0}} + \int_{a_0}^{a_0 + T-1} \frac{1}{\sqrt{x}} dx = \frac{1}{\sqrt{a_0}} + 2\sqrt{a_0 + T - 1} - 2\sqrt{a_0}. \quad (19)$$

Combining (18) and (19), we have the final result. \square

Corollary B.7. If $x_0 - \gamma > \nu > 0$, $b_0 \geq \nu$ and

$$T < \frac{(x_0 - \nu - \gamma)^2 + 4\gamma(x_0 - \nu - \gamma)\sqrt{a_0}}{4\gamma^2} + 2,$$

then $x_T > \nu$ for deterministic M-AdaGradD. Conversely, the case $|x_T| \leq \nu$ implies that

$$T \geq \frac{(x_0 - \nu - \gamma)^2 + 4\gamma(x_0 - \nu - \gamma)\sqrt{a_0}}{4\gamma^2} + 2.$$

Proof. The proof is the same as for Corollary B.3. \square

Theorem B.8. For any $\varepsilon, \delta \in (0, 1)$, $\sigma > 0$, there exists convex L -smooth minimization problem (8) and stochastic gradient oracle such that Assumption 1.1 holds with $\alpha = 2$ and the iterates produced by M-AdaGradD after K steps with stepsize γ and starting point x_0 such that $R := x_0 - \sqrt{2\varepsilon} - \gamma > 0$, $b_0 > \nu$ and $(1 - \beta_1)\sigma R/\varepsilon\sqrt{\delta} \geq 16b_0^2$ satisfy the following implication

$$\mathbb{P}\{f(x_K) - f(x^*) \geq \varepsilon\} \leq \delta \implies K = \Omega\left(\frac{\sigma R}{\varepsilon\sqrt{\delta}}\right), \quad (20)$$

i.e., the high-probability complexity of M-AdaGradD has inverse-power dependence on δ .

Proof. The overall idea of the proof resembles the one for Theorem B.4 – we combine the lower bound for the number of iterations from Corollary B.7 with the specific choice of stochasticity. Nevertheless, to prove this theorem, we construct the adversarial noise in another way. More precisely, we consider the following stochastic gradient

$$g_k = \nabla f(x_k) - \sigma \xi_k,$$

where

$$\xi_k = \begin{cases} 0, & \text{if } k < K-1 \text{ or } |\hat{x}_K| > \nu, \\ -A_k, & \text{with probability } \frac{1}{2A_k^2} \\ 0, & \text{with probability } 1 - \frac{1}{A_k^2} \\ A_k, & \text{otherwise,} \end{cases} \quad (21)$$

where \hat{x}_K is the result of deterministic **M-AdaGradD** after K iterations and $A_k = \max \left\{ 1, \frac{2\nu b_k}{(1-\beta_1)\gamma\sigma} \right\}$. What is more, $\mathbb{E}[\xi_k] = 0$ and $\mathbb{E}[\xi_k^2] \leq 1$ by the construction. Therefore, the stochastic gradient satisfies the Assumption 1.1 with $\alpha = 2$.

We want to prove that $\mathbb{P}\{f(x_K) - f(x^*) > \varepsilon\} \leq \delta$. For $\delta < 1$, this implies that $|\hat{x}_K| \leq \nu$ with $\varepsilon = \frac{\nu^2}{2}$. Indeed, assuming the contrary, the noise is equal to 0 for each iteration by the construction, meaning that

$$\mathbb{P}\{f(x_K) - f(x^*) > \varepsilon\} = \mathbb{P}\{f(\hat{x}_K) - f(x^*) > \varepsilon\} = \mathbb{P}\{|\hat{x}_K| > \nu\} = 1 > \delta.$$

As a result, $|\hat{x}_K| \leq \nu$ and, applying Corollary B.7, we obtain

$$K \geq \frac{(x_0 - \nu - \gamma)^2 + 4\gamma(x_0 - \nu - \gamma)\sqrt{a_0}}{4\gamma^2} + 2.$$

What is more, x_K can be written as

$$x_K = \hat{x}_{K-1} - \frac{\gamma}{b_{K-1}} m_{K-1} = \hat{x}_K + \frac{(1-\beta_1)\gamma\sigma\xi_{K-1}}{b_{K-1}}.$$

Hence,

$$\begin{aligned} \mathbb{P}\{f(x_K) - f(x^*) \geq \varepsilon\} &= \mathbb{P}\{|x_K| \geq \nu\} = \mathbb{P}\left\{\left|\hat{x}_K + \frac{(1-\beta_1)\gamma\sigma\xi_{K-1}}{b_{K-1}}\right| \geq \nu\right\} \\ &\geq \mathbb{P}\left\{\left|\frac{(1-\beta_1)\gamma\sigma\xi_{K-1}}{b_{K-1}}\right| \geq \nu + \hat{x}_K\right\} \geq \mathbb{P}\left\{\left|\frac{(1-\beta_1)\gamma\sigma\xi_{K-1}}{b_{K-1}}\right| \geq 2\nu\right\} \\ &= \mathbb{P}\left\{|\xi_{K-1}| \geq \frac{2\nu b_{K-1}}{(1-\beta_1)\gamma\sigma}\right\}. \end{aligned}$$

If $\max \left\{ 1, \frac{2\nu b_{K-1}}{(1-\beta_1)\gamma\sigma} \right\} = 1$, then

$$\delta \geq \mathbb{P}\{f(x_K) - f(x^*) \geq \varepsilon\} \geq \mathbb{P}\left\{|\xi_{K-1}| \geq \frac{2\nu b_{K-1}}{(1-\beta_1)\gamma\sigma}\right\} = 1,$$

which leads us to the contradiction. Therefore $\max \left\{ 1, \frac{2\nu b_{K-1}}{(1-\beta_1)\gamma\sigma} \right\} = \frac{2\nu b_{K-1}}{(1-\beta_1)\gamma\sigma}$, and

$$\delta \geq \mathbb{P}\{f(x_K) - f(x^*) \geq \varepsilon\} \geq \mathbb{P}\left\{|\xi_{K-1}| \geq \frac{2\nu b_{K-1}}{(1-\beta_1)\gamma\sigma}\right\} = \frac{1}{A_{K-1}^2} = \frac{(1-\beta_1)^2\gamma^2\sigma^2}{4\nu^2 b_{K-1}^2},$$

where we used that $A_{K-1} = \max \left\{ 1, \frac{2\nu b_{K-1}}{(1-\beta_1)\gamma\sigma} \right\}$ and the noise structure. Consequently, $\gamma \leq \frac{2\nu b_{K-1}\sqrt{\delta}}{(1-\beta_1)\sigma}$. What is more, b_{K-1} can be bounded as

$$b_{K-1} \leq \sqrt{b_0^2 + K\nu^2}$$

since the gradient of f is uniformly bounded by ν . Hence, we obtain

$$\begin{aligned} K &\geq \frac{(x_0 - \nu - \gamma)^2}{4\gamma^2} + \frac{4(x_0 - \nu - \gamma)\sqrt{a_0}}{4\gamma} \geq \frac{(x_0 - \nu - \gamma)^2}{4\gamma^2} \\ &\geq \frac{(1 - \beta_1)^2(x_0 - \nu - \gamma)^2\sigma^2}{16\nu^2(b_0^2 + K\nu^2)\delta}. \end{aligned}$$

Multiplying both sides by $\nu^2(b_0^2 + K\nu^2)$, we get

$$(b_0^2 + K\nu^2)^2 \geq \nu^2 K(b_0^2 + K\nu^2) \geq \frac{(1 - \beta_1)^2(x_0 - \nu - \gamma)^2\sigma^2}{16\delta},$$

implying that

$$K \geq \frac{(1 - \beta_1)\sigma R}{4\nu^2\sqrt{\delta}} - b_0^2 = \frac{(1 - \beta_1)\sigma R}{8\varepsilon\sqrt{\delta}} - b_0^2 \geq \frac{(1 - \beta_1)\sigma R}{16\varepsilon\sqrt{\delta}},$$

which finishes the proof. \square

B.3. Failure of Adam

Similarly to the case of **M-AdaGrad**, we start by obtaining the analytical form of iterations of the deterministic **Adam** in the following lemma.

Lemma B.9. Suppose that the starting point x_0 is such that $x_0 > 0$. If after T iterations of deterministic **Adam** with initial momentum⁷ $m_{-1} \geq 0$ we have $|x_t| > \nu$ and $x_t > 0$ for all $t = \overline{1, T-1}$, then

$$x_T = x_0 - \gamma \sum_{t=0}^{T-1} \frac{\beta_1^{t+1}m_{-1} + (1 - \beta_1^{t+1})\nu}{\sqrt{\beta_2^{t+1}b_{-1}^2 + (1 - \beta_2^{t+1})\nu^2}}.$$

Proof. Since $|x_t| > \nu$ and x_t is positive, the gradient at x_t is equal to ν . Hence, by substituting the gradient into the algorithm, we get the final result. \square

The above lemma relies on the condition that $|x_t| > \nu$ and $x_t > 0$ for all $t = \overline{1, T-1}$. For any γ, b_{-1} and T this condition can be achieved if we choose sufficiently small ν .

Next, we estimate the interval where x_T lies.

Lemma B.10. Let the conditions of Lemma B.9 hold. Then, if $\beta_2 = 1 - 1/K$, where K is the total number of iterations of deterministic **Adam**, we have

$$x_0 - \gamma \left(\frac{2\beta_1 m_{-1}}{(1 - \beta_1)b_{-1}} + \frac{2T\nu}{b_{-1}} \right) \leq x_T \leq x_0 - \frac{(1 - \beta_1)\gamma\nu T}{\sqrt{b_{-1}^2 + \nu^2}}.$$

Proof. From Lemma B.9 we have:

$$x_T = x_0 - \gamma \sum_{t=0}^{T-1} \frac{\beta_1^{t+1}m_{-1} + (1 - \beta_1^{t+1})\nu}{\sqrt{\beta_2^{t+1}b_{-1}^2 + (1 - \beta_2^{t+1})\nu^2}}.$$

Next, we bound the second term in the inequality above in the following way:

$$\sum_{t=0}^{T-1} \frac{\beta_1^{t+1}m_{-1} + (1 - \beta_1^{t+1})\nu}{\sqrt{\beta_2^{t+1}b_{-1}^2 + (1 - \beta_2^{t+1})\nu^2}} \leq \sum_{t=0}^{T-1} \frac{\beta_1^{t+1}m_{-1}}{\sqrt{\beta_2^{t+1}b_{-1}^2 + (1 - \beta_2^{t+1})\nu^2}} + \frac{2T\nu}{b_{-1}} \leq \frac{2\beta_1 m_{-1}}{(1 - \beta_1)b_{-1}} + \frac{2T\nu}{b_{-1}}, \quad (22)$$

⁷By default, $m_{-1} = 0$ but the result of this lemma holds for non-zero $m_{-1} \geq 0$ as well. If $m_{-1} \leq 0$, then one can select $x_0 < 0$.

$$\sum_{t=0}^{T-1} \frac{\beta_1^{t+1} m_{-1} + (1 - \beta_1^{t+1}) \nu}{\sqrt{\beta_2^{t+1} b_{-1}^2 + (1 - \beta_2^{t+1}) \nu^2}} \geq \frac{(1 - \beta_1) \nu T}{\sqrt{b_{-1}^2 + \nu^2}}, \quad (23)$$

where we use the fact that with $K \geq 2$ next inequalities hold

$$1 \geq \beta_2^k = (1 - 1/K)^k \geq (1 - 1/K)^K \geq 1/4,$$

$$0 \leq 1 - \beta_2^k \leq 3/4 \leq 1.$$

Combining (22) and (23), we get the final result. \square

Corollary B.11. If $x_0 - \nu - \frac{2\gamma\beta_1 m_{-1}}{(1-\beta_1)b_{-1}} > 0$ and

$$T < \frac{\left(x_0 - \nu - \frac{2\gamma\beta_1 m_{-1}}{(1-\beta_1)b_{-1}}\right) b_{-1}}{2\gamma\nu},$$

then $x_T > \nu$ for deterministic Adam with $\beta_2 = 1 - 1/K$. Alternatively, $|x_T| \leq \nu$ implies that

$$T \geq \frac{\left(x_0 - \nu - \frac{2\gamma\beta_1 m_{-1}}{(1-\beta_1)b_{-1}}\right) b_{-1}}{2\gamma\nu}.$$

Proof. Let us note that

$$\nu < x_0 - \gamma \left(\frac{2\beta_1 m_{-1}}{(1-\beta_1)b_{-1}} + \frac{2T\nu}{b_{-1}} \right)$$

is equivalent to

$$T < \frac{\left(x_0 - \nu - \frac{2\gamma\beta_1 m_{-1}}{(1-\beta_1)b_{-1}}\right) b_{-1}}{2\gamma\nu}.$$

The second part of the corollary is just a negation of the implication stated in the first part of the corollary. \square

Theorem B.12. For any $\varepsilon, \delta \in (0, 1)$, $\sigma > 0$, there exists convex L -smooth minimization problem (8) and stochastic gradient oracle such that Assumption 1.1 holds with $\alpha = 2$ and the iterates produced by Adam after K steps with $\beta_2 = 1 - 1/K$ and stepsize γ and starting point x_0 such that $R := x_0 - \nu \geq 2\gamma(1 + \beta_1)\sqrt{K}$ satisfy the following implication:

$$\mathbb{P}\{f(x_K) - f(x^*) \geq \varepsilon\} \leq \delta \implies K = \Omega\left(\frac{b_{-1}R}{\sqrt{\varepsilon}\gamma} + \left(\frac{\sigma R}{\gamma\sqrt{\varepsilon\delta}}\right)^{2/3}\right), \quad (24)$$

i.e., the high-probability complexity of Adam has inverse-power dependence on δ .

Proof. The main idea is quite similar to the proof of Theorem B.4. We introduce the noise in the stochastic gradient in the following way

$$g_k = \nabla f(x_k) - \sigma\xi_k,$$

where

$$\xi_k = \begin{cases} 0, & \text{for } k > 0, \\ \begin{cases} -A, & \text{with probability } \frac{1}{2A^2} \\ 0, & \text{with probability } 1 - \frac{1}{A^2} \\ A, & \text{with probability } \frac{1}{2A^2} \end{cases} & \text{otherwise,} \end{cases} \quad (25)$$

where the formula for A is given later. The noise construction (25) implies that stochasticity appears only at the first iteration of **Adam**, and then it only affects the stepsizes. Therefore,

$$x_1 = x_0 - \frac{\gamma}{b_0} m_0,$$

where $b_0 = \sqrt{\beta_2 b_{-1}^2 + (1 - \beta_2)(\nu - \sigma \xi_0)^2}$ and $m_0 = (1 - \beta_1)(\nu - \sigma \xi_0)$. Denoting K_0 as the number of iterations required to achieve at least ε -accuracy, choosing x_0 in such a way that $x_0 - \frac{\gamma m_0}{b_0} - \nu - \frac{2\gamma\beta_1 m_0}{(1-\beta_1)b_0} > 0$ and considering the case $\xi_0 \neq A$ to guarantee $m_0 \geq 0$, we apply Corollary B.11 and get that the algorithm needs to make at least

$$K_0 = \frac{\left(x_1 - \nu - \frac{2\gamma\beta_1 m_0}{(1-\beta_1)b_0}\right) b_0}{2\gamma\nu}$$

iterations to reach ε -accuracy, where $\varepsilon = \frac{\nu^2}{2}$. Let us specify that this estimate depends on the stochasticity at the first iteration, i.e., the bound on the number of iterations is random. Consequently, if **Adam** achieves ε -solution after K steps, we should have $K \geq K_0$. Therefore, $\mathbb{P}\{K \geq K_0\} \geq \mathbb{P}\{f(x_K) - f(x^*) \leq \varepsilon\}$ and we want to estimate K such that

$$\mathbb{P}\{K_0 \leq K\} \geq 1 - \delta.$$

Bounding the left-hand side,

$$\begin{aligned} \mathbb{P}\{K_0 \leq K\} &= \mathbb{P}\{K_0 \leq K | \xi_0 = A\} \mathbb{P}\{\xi_0 = A\} + \mathbb{P}\{K_0 \leq K | \xi_0 \neq A\} \mathbb{P}\{\xi_0 \neq A\} \\ &\leq \mathbb{P}\left\{\frac{\left(x_1 - \nu - \frac{2\gamma\beta_1 m_0}{(1-\beta_1)b_0}\right) b_0}{2\gamma\nu} \leq K \mid \xi_0 \neq A\right\} \mathbb{P}\{\xi_0 \neq A\} + \mathbb{P}\{\xi_0 = A\} \\ &= \mathbb{P}\left\{\frac{\left(x_0 - \gamma \frac{m_0}{b_0} - \nu - \frac{2\gamma\beta_1 m_0}{(1-\beta_1)b_0}\right) b_0}{2\gamma\nu} \leq K \mid \xi_0 \neq A\right\} \mathbb{P}\{\xi_0 \neq A\} + \mathbb{P}\{\xi_0 = A\}. \end{aligned}$$

Similar to the proof of Theorem B.4, denoting $R = x_0 - \nu$ and assuming that for $\xi_0 \neq A$ we have $R \geq \frac{4\gamma m_0 \beta_1}{b_0(1-\beta_1)} + \frac{2\gamma m_0}{b_0}$, which implies $R - \frac{2\gamma m_0 \beta_1}{b_0(1-\beta_1)} - \frac{\gamma m_0}{b_0} \geq \frac{2\gamma m_0 \beta_1}{b_0(1-\beta_1)} + \frac{\gamma m_0}{b_0}$, we derive

$$\begin{aligned} \mathbb{P}\{K_0 \leq K\} &\leq \mathbb{P}\left\{\frac{\left(x_0 - \gamma \frac{m_0}{b_0} - \nu - \frac{2\gamma\beta_1 m_0}{(1-\beta_1)b_0}\right) b_0}{2\gamma\nu} \leq K \mid \xi_0 \neq A\right\} \mathbb{P}\{\xi_0 \neq A\} + \mathbb{P}\{\xi_0 = A\} \\ &\leq \mathbb{P}\left\{\frac{Rb_0}{4\gamma\nu} \leq K \mid \xi_0 \neq A\right\} \mathbb{P}\{\xi_0 \neq A\} + \mathbb{P}\{\xi_0 = A\} \\ &= \mathbb{P}\left\{\sqrt{\beta_2 b_{-1}^2 + (1 - \beta_2)(\nu - \sigma \xi_0)^2} \leq \frac{4\gamma\nu K}{R} \mid \xi_0 \neq A\right\} \mathbb{P}\{\xi_0 \neq A\} + \mathbb{P}\{\xi_0 = A\} \\ &\leq \mathbb{P}\left\{\sqrt{1 - \beta_2} |\nu - \sigma \xi_0| \leq \frac{4\gamma\nu K}{R} \mid \xi_0 \neq A\right\} \mathbb{P}\{\xi_0 \neq A\} + \mathbb{P}\{\xi_0 = A\} \\ &\leq \mathbb{P}\left\{\sigma |\xi_0| \leq \frac{4\gamma\nu K}{R\sqrt{1 - \beta_2}} + \nu \mid \xi_0 \neq A\right\} \mathbb{P}\{\xi_0 \neq A\} + \mathbb{P}\{\xi_0 = A\}, \end{aligned}$$

where in the fourth row we used $K \geq \frac{b_{-1} R}{4\gamma\nu}$. Therefore, $\mathbb{P}\{K_0 \leq K\} \geq 1 - \delta$ implies

$$\mathbb{P}\left\{\sigma |\xi_0| \leq \frac{4\gamma\nu K}{R\sqrt{1 - \beta_2}} + \nu \mid \xi_0 \neq A\right\} \mathbb{P}\{\xi_0 \neq A\} + \mathbb{P}\{\xi_0 = A\} \geq 1 - \delta.$$

Consequently, if we choose $A = \frac{\frac{4\gamma\nu K}{R\sqrt{1 - \beta_2}} + 2\nu}{\sigma}$, then the only realization of the random variable ξ_0 for which the inequality in the first probability is satisfied is $\xi_0 = 0$. Hence, we have

$$1 - \frac{1}{2A^2} \geq 1 - \delta.$$

As a result, we get

$$\frac{1}{A} = \frac{\sigma}{\frac{4\gamma\nu K}{R\sqrt{1-\beta_2}} + 2\nu} \leq \sqrt{2\delta}.$$

Therefore,

$$K/\sqrt{1-\beta_2} = K^{\frac{3}{2}} \geq \frac{R}{2\gamma} \left(\frac{\sigma}{2\nu\sqrt{2\delta}} - 1 \right) \geq \frac{R\sigma}{16\gamma\sqrt{\varepsilon\delta}} \implies K \geq \left(\frac{R\sigma}{16\gamma\sqrt{\varepsilon\delta}} \right)^{\frac{2}{3}},$$

where we use $1 - \beta_2 = 1/K$, $\sigma/\sqrt{\varepsilon\delta} \geq 8$ and $\nu = \sqrt{2\varepsilon}$. It remains to find the conditions on x_0 and γ ensuring that for $\xi_0 \neq A$

$$R \geq \frac{4\gamma m_0 \beta_1}{b_0(1 - \beta_1)} + \frac{2\gamma m_0}{b_0} = \frac{2\gamma m_0}{b_0} \left(\frac{2\beta_1}{1 - \beta_1} + 1 \right).$$

Therefore, the above is equivalent to

$$R \geq 2\gamma \left(\frac{2\beta_1}{1 - \beta_1} + 1 \right) \max \left\{ \frac{(1 - \beta_1)\nu}{\sqrt{\beta_2 b_{-1}^2 + (1 - \beta_2)\nu^2}}, \frac{(1 - \beta_1)(\nu + \sigma A)}{\sqrt{\beta_2 b_{-1}^2 + (1 - \beta_2)(\nu + \sigma A)^2}} \right\}.$$

To simplify the condition on R , we derive an upper bound for the maximum in the right-hand side:

$$\begin{aligned} \max \left\{ \frac{(1 - \beta_1)\nu}{\sqrt{\beta_2 b_{-1}^2 + (1 - \beta_2)\nu^2}}, \frac{(1 - \beta_1)(\nu + \sigma A)}{\sqrt{\beta_2 b_{-1}^2 + (1 - \beta_2)(\nu + \sigma A)^2}} \right\} &\leq \max \left\{ \frac{(1 - \beta_1)\nu}{\sqrt{(1 - \beta_2)\nu^2}}, \frac{(1 - \beta_1)(\nu + \sigma A)}{\sqrt{(1 - \beta_2)(\nu + \sigma A)^2}} \right\} \\ &= \frac{1 - \beta_1}{\sqrt{1 - \beta_2}} = (1 - \beta_1)\sqrt{K}. \end{aligned}$$

Therefore, it is sufficient to choose R satisfying

$$R \geq 2\gamma(1 - \beta_1)\sqrt{K} \left(\frac{2\beta_1}{1 - \beta_1} + 1 \right) = 2\gamma(1 + \beta_1)\sqrt{K}.$$

This concludes the proof. \square

B.4. Failure of AdamD

We follow the idea for previous proofs and start by obtaining the analytical form of iterations of the deterministic AdamD in the following lemma.

Lemma B.13. *Suppose that the starting point x_0 is such that $x_0 > 0$. If after T iterations of deterministic AdamD we have $|x_t| > \nu$ and $x_t > 0$ for all $t = \overline{1, T-1}$, then*

$$x_T = x_0 - \gamma\nu \sum_{t=0}^{T-1} \frac{1 - \beta_1^{t+1}}{\sqrt{\beta_2^t b_0^2 + (1 - \beta_2^t)\nu^2}}.$$

Proof. Since $|x_t| > \nu$ and x_t is positive, the gradient at x_t is equal to ν . Hence, by substituting the gradient into the algorithm, we get the final result. \square

The above lemma relies on the condition that $|x_t| > \nu$ and $x_t > 0$ for all $t = \overline{1, T-1}$. For any γ, b_0 and T this condition can be achieved if we choose sufficiently small ν .

Next, we estimate the interval where x_T lies.

Lemma B.14. *Let the conditions of Lemma B.13 hold. Then, if $\beta_2 = 1 - 1/K$, where K is the total number of iterations of deterministic AdamD, we have*

$$x_0 - \frac{2\gamma\nu T}{b_0} \leq x_T \leq x_0 - \frac{\gamma\nu(1 - \beta_1)T}{\sqrt{b_0^2 + \nu^2}}.$$

Proof. From Lemma B.13 we have:

$$x_T = x_0 - \gamma\nu \sum_{t=0}^{T-1} \frac{1 - \beta_1^{t+1}}{\sqrt{\beta_2^t b_0^2 + (1 - \beta_2^t) \nu^2}}.$$

Next, we bound the second term in the inequality above in the following way:

$$\sum_{t=0}^{T-1} \frac{1 - \beta_1^{t+1}}{\sqrt{\beta_2^t b_0^2 + (1 - \beta_2^t) \nu^2}} \leq \frac{2T}{b_0}, \quad (26)$$

$$\sum_{t=0}^{T-1} \frac{1 - \beta_1^{t+1}}{\sqrt{\beta_2^t b_0^2 + (1 - \beta_2^t) \nu^2}} \geq \frac{(1 - \beta_1)T}{\sqrt{b_0^2 + \nu^2}}, \quad (27)$$

where we use the fact that with $K \geq 2$ next inequalities hold

$$1 \geq \beta_2^k = (1 - 1/K)^k \geq (1 - 1/K)^K \geq 1/4,$$

$$0 \leq 1 - \beta_2^k \leq 3/4 \leq 1.$$

Combining (26) and (27), we get the final result. \square

Corollary B.15. *If $x_0 > \nu > 0$ and*

$$T < \frac{(x_0 - \nu)b_0}{2\gamma\nu},$$

then $x_T > \nu$ for deterministic AdamD. Alternatively, $|x_T| \leq \nu$ implies that

$$T \geq \frac{(x_0 - \nu)b_0}{2\gamma\nu}.$$

Proof. The proof is the same as for Corollary B.11. \square

Theorem B.16. *For any $\varepsilon, \delta \in (0, 1)$, $\sigma > 0$, there exists convex L -smooth minimization problem (8) and stochastic gradient oracle such that Assumption 1.1 holds with $\alpha = 2$ and the iterates produced by AdamD after K steps with stepsize γ and starting point x_0 such that $R := x_0 - \nu > 0$, $b_0 > \nu$ and $\sigma R/\varepsilon\sqrt{\delta} \geq 16b_0^2$ satisfy the following implication*

$$\mathbb{P}\{f(x_K) - f(x^*) \geq \varepsilon\} \leq \delta \implies K = \Omega\left(\frac{\sigma R}{\varepsilon\sqrt{\delta}}\right), \quad (28)$$

i.e., the high-probability complexity of AdamD has inverse-power dependence on δ .

Proof. The overall idea of the proof resembles the one for Theorem B.12 – we combine the lower bound for the number of iterations from Corollary B.15 with the specific choice of stochasticity. Nevertheless, to prove this theorem, we construct the adversarial noise in another way. More precisely, we consider the following stochastic gradient

$$g_k = \nabla f(x_k) - \sigma\xi_k,$$

where

$$\xi_k = \begin{cases} 0, & \text{if } k < K - 1 \text{ or } |\hat{x}_K| > \nu, \\ \begin{cases} -A_k, & \text{with probability } \frac{1}{2A_k^2} \\ 0, & \text{with probability } 1 - \frac{1}{A_k^2} \\ A_k, & \text{otherwise,} \end{cases} & \text{with probability } \frac{1}{2A_k^2} \\ \begin{cases} 0, & \text{with probability } 1 - \frac{1}{A_k^2} \\ A_k, & \text{otherwise,} \end{cases} & \text{with probability } \frac{1}{2A_k^2} \end{cases} \quad (29)$$

where \hat{x}_K is the result of deterministic **AdamD** after K iterations and $A_k = \max \left\{ 1, \frac{2\nu b_k}{(1-\beta_1)\gamma\sigma} \right\}$. What is more, $\mathbb{E}[\xi_k] = 0$ and $\mathbb{E}[\xi_k^2] \leq 1$ by the construction. Therefore, the stochastic gradient satisfies the Assumption 1.1 with $\alpha = 2$.

We want to prove that $\mathbb{P}\{f(x_K) - f(x^*) > \varepsilon\} \leq \delta$. For $\delta < 1$, this implies that $|\hat{x}_K| \leq \nu$ with $\varepsilon = \frac{\nu^2}{2}$. Indeed, assuming the contrary, the noise is equal to 0 for each iteration by the construction, meaning that

$$\mathbb{P}\{f(x_K) - f(x^*) > \varepsilon\} = \mathbb{P}\{f(\hat{x}_K) - f(x^*) > \varepsilon\} = \mathbb{P}\{|\hat{x}_K| > \nu\} = 1 > \delta.$$

As a result, $|\hat{x}_K| \leq \nu$ and, applying Corollary B.15, we obtain

$$K \geq \frac{(x_0 - \nu)b_0}{2\gamma\nu}.$$

What is more, x_K can be written as

$$x_K = \hat{x}_{K-1} - \frac{\gamma}{b_{K-1}}m_{K-1} = \hat{x}_K + \frac{(1-\beta_1)\gamma\sigma\xi_{K-1}}{b_{K-1}}.$$

Hence,

$$\begin{aligned} \mathbb{P}\{f(x_K) - f(x^*) \geq \varepsilon\} &= \mathbb{P}\{|x_K| \geq \nu\} = \mathbb{P}\left\{\left|\hat{x}_K + \frac{(1-\beta_1)\gamma\sigma\xi_{K-1}}{b_{K-1}}\right| \geq \nu\right\} \\ &\geq \mathbb{P}\left\{\left|\frac{(1-\beta_1)\gamma\sigma\xi_{K-1}}{b_{K-1}}\right| \geq \nu + \hat{x}_K\right\} \geq \mathbb{P}\left\{\left|\frac{(1-\beta_1)\gamma\sigma\xi_{K-1}}{b_{K-1}}\right| \geq 2\nu\right\} \\ &= \mathbb{P}\left\{|\xi_{K-1}| \geq \frac{2\nu b_{K-1}}{(1-\beta_1)\gamma\sigma}\right\}. \end{aligned}$$

If $\max \left\{ 1, \frac{2\nu b_{K-1}}{(1-\beta_1)\gamma\sigma} \right\} = 1$, then

$$\delta \geq \mathbb{P}\{f(x_K) - f(x^*) \geq \varepsilon\} \geq \mathbb{P}\left\{|\xi_{K-1}| \geq \frac{2\nu b_{K-1}}{(1-\beta_1)\gamma\sigma}\right\} = 1,$$

which leads us to the contradiction. Therefore $\max \left\{ 1, \frac{2\nu b_{K-1}}{\gamma\sigma} \right\} = \frac{2\nu b_{K-1}}{(1-\beta_1)\gamma\sigma}$, and

$$\delta \geq \mathbb{P}\{f(x_K) - f(x^*) \geq \varepsilon\} \geq \mathbb{P}\left\{|\xi_{K-1}| \geq \frac{2\nu b_{K-1}}{(1-\beta_1)\gamma\sigma}\right\} = \frac{1}{A_{K-1}^2} = \frac{(1-\beta_1)^2\gamma^2\sigma^2}{4\nu^2 b_{K-1}^2},$$

where we used that $A_{K-1} = \max \left\{ 1, \frac{2\nu b_{K-1}}{(1-\beta_1)\gamma\sigma} \right\}$ and the noise structure. Consequently, $\gamma \leq \frac{2\nu b_{K-1}\sqrt{\delta}}{(1-\beta_1)\sigma}$. What is more, b_{K-1} can be bounded as

$$b_{K-1} \leq \sqrt{b_0^2 + \nu^2}$$

since the gradient of f is uniformly bounded by ν . Hence, we obtain with $b_0 \geq \nu$

$$K \geq \frac{(x_0 - \nu)b_0}{2\gamma\nu} \geq \frac{(1-\beta_1)(x_0 - \nu)\sigma b_0}{4\sqrt{b_0^2 + \nu^2}\nu^2\sqrt{\delta}} \geq \frac{(1-\beta_1)(x_0 - \nu)\sigma}{8\nu^2\sqrt{\delta}} = \frac{(1-\beta_1)R\sigma}{16\varepsilon\sqrt{\delta}},$$

which finishes the proof. \square

C. Missing Proofs from Section 3

In this section, we provide missing proofs for Algorithm 2 in the convex and non-convex cases. For each case, the proof consists of two parts – descent lemma and main theorem. Moreover, for convenience of the proofs, we consider a reweighted version of Algorithm 2 summarized in Algorithm 3, which has an additional parameter $\eta > 0$ appearing in the update rule for b_t . However, Algorithms 2 and 3 are equivalent: if we divide b_t and γ in Algorithm 3 by $\sqrt{\eta}$, the method reduces to Algorithm 2 but produces exactly the same points as before (given the same initialization and source of stochasticity, i.e., seed), since γ/b_t remains unchanged.

Algorithm 3 Reweighted Clip-Adam/Clip-AdamD-Norm and Clip-M-AdaGrad/Clip-M-AdaGradD-Norm

Input: Step size $\gamma > 0$, starting point $x_0 \in \mathbb{R}^d$, initial constant $b_{-1} > 0$ (for Adam and M-AdaGrad) or $b_0 > 0$ (for AdamD and M-AdaGradD), momentum parameters $\beta_1, \beta_2 \in [0, 1]$, level of clipping $\lambda > 0$, reweighting parameter $\eta > 0$

- 1: Set $m_{-1} = 0$
- 2: **for** $t = 0, 1, \dots$ **do**
- 3: $m_t = \beta_1 m_{t-1} + (1 - \beta_1) \text{clip}(\nabla f_{\xi_t}(x_t), \lambda)$
- 4: **if** no delay **then**
- 5: $b_t = \begin{cases} \sqrt{\beta_2 b_{t-1}^2 + \eta(1 - \beta_2) \|\text{clip}(\nabla f_{\xi_t}(x_t), \lambda)\|^2} & \text{for Clip-Adam-Norm} \\ \sqrt{b_{t-1}^2 + \eta \|\text{clip}(\nabla f_{\xi_t}(x_t), \lambda)\|^2} & \text{for Clip-M-AdaGrad-Norm} \end{cases}$
- 6: **else**
- 7: $b_{t+1} = \begin{cases} \sqrt{\beta_2 b_t^2 + \eta(1 - \beta_2) \|\text{clip}(\nabla f_{\xi_t}(x_t), \lambda)\|^2} & \text{for Clip-AdamD-Norm} \\ \sqrt{b_t^2 + \eta \|\text{clip}(\nabla f_{\xi_t}(x_t), \lambda)\|^2} & \text{for Clip-M-AdaGradD-Norm} \end{cases}$
- 8: **end if**
- 9: $x_{t+1} = x_t - \frac{\gamma}{b_t} m_t$
- 10: **end for**

C.1. Technical Lemmas

Here we introduce technical lemmas for the future proofs.

Lemma C.1. *Let the sequence $\{b_t\}_{t=0}$ is generated by Algorithm 3 in K iterations. Then, for every $t, r: t \geq r$ we get*

$$b_t \geq c_m b_r,$$

where the constant c_m depends on the update rule for b_t . To be more precise, $c_m = 1$ for the Clip-M-AdaGrad/Clip-M-AdaGradD-Norm, and $c_m = 1/2$ for Clip-Adam/Clip-AdamD-Norm.

Proof. The case of Clip-M-AdaGrad/Clip-M-AdaGradD-Norm is obvious since the sequence $\{b_t\}_{t=0}$ is non-decreasing. For the Clip-Adam/Clip-AdamD-Norm we obtain that

$$b_t^2 \geq \beta_2^{t-r} b_r^2 = \left(1 - \frac{1}{K}\right)^{t-r} b_r^2 \geq \left(1 - \frac{1}{K}\right)^K b_r^2 \geq \frac{1}{4} b_r^2,$$

where we, without loss of generality, assume that $K \geq 2$ and apply the analytical form of β_2 with fact that $g(K) = (1 - \frac{1}{K})^K$ is increasing function. Taking the square root from both parts, we conclude the proof. \square

Lemma C.2. *Let the sequence $\{m_t\}_{t=0}$ is generated by Algorithm 3 in K iterations. Then, for every $0 \leq t \leq K-1$ it holds that*

$$m_t = \sum_{k=0}^t \beta_1^{t-k} (1 - \beta_1) g_k.$$

Moreover, $\|m_t\|^2$ can be bounded in the following way:

$$\|m_t\|^2 \leq (1 - \beta_1^{t+1}) \sum_{k=0}^t \beta_1^{t-k} (1 - \beta_1) \|g_k\|^2.$$

Proof. The first part of the lemma is the direct consequence of update rule of momentum m_t . For the second part we need to apply the Jensen's inequality as follows:

$$\left\| \sum_{k=0}^t \frac{\beta_1^{t-k} (1 - \beta_1)}{1 - \beta_1^{t+1}} g_k \right\|^2 \leq \sum_{k=0}^t \frac{\beta_1^{t-k} (1 - \beta_1)}{1 - \beta_1^{t+1}} \|g_k\|^2,$$

where we use the convexity of $\|\cdot\|^2$ and $\sum_{k=0}^t \beta_1^{t-k} (1 - \beta_1) = 1 - \beta_1^{t+1}$. Multiplying both sides by $(1 - \beta_1^{t+1})^2$, we get the final result. \square

C.2. Non-Convex Case: Methods with Delay

Lemma C.3 (Descent lemma). *Let Assumption 1.2 hold on $Q = \left\{ x \in \mathbb{R}^d \mid \exists y \in \mathbb{R}^d : f(y) \leq f_* + 2\Delta \text{ and } \|x - y\| \leq \frac{\sqrt{\Delta}}{20\sqrt{L}} \right\}$, where $f(x_0) - f_* = \Delta_0 \leq \Delta$. Then, after T iterations of Clip-M-AdaGradD/Clip-AdamD-Norm with $b_0 \geq 2\gamma L / (1 - \beta_1)^2 c_m^2$, if $x_t \in Q \forall t = \overline{0, T}$, we have*

$$\sum_{t=0}^{T-1} \frac{\gamma C_t}{2} \|\nabla f(x_t)\|^2 \leq \Delta_0 - \Delta_T - \sum_{t=0}^{T-1} (\gamma C_t - 2A_t) \langle \nabla f(x_t), \theta_t^u \rangle + \sum_{t=0}^{T-1} \gamma C_t \|\theta_t^b\|^2 + \sum_{t=0}^{T-1} 2A_t \|\theta_t^u\|^2,$$

where $C_t = \sum_{k=t}^{T-1} \frac{1 - \beta_1}{b_k} \beta_1^{k-t}$, $A_t = \sum_{k=t}^{T-1} \frac{L\gamma^2(1 - \beta_1)}{c_m b_k b_0} (k - t + 1) \beta_1^{k-t}$ and c_m is taken from Lemma C.1.

Proof. We start with the L -smoothness of f :

$$f(x_{t+1}) - f(x_t) \leq \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 = -\frac{\gamma}{b_t} \langle \nabla f(x_t), m_t \rangle + \frac{L\gamma^2}{2b_t^2} \|m_t\|^2. \quad (30)$$

Using the update rule of Algorithm 3, we can obtain

$$\begin{aligned} -\langle \nabla f(x_t), m_t \rangle &= -\beta_1 \langle \nabla f(x_t), m_{t-1} \rangle - (1 - \beta_1) \langle \nabla f(x_t), g_t \rangle \\ &= -\beta_1 \langle \nabla f(x_t) - \nabla f(x_{t-1}), m_{t-1} \rangle - \beta_1 \langle \nabla f(x_{t-1}), m_{t-1} \rangle \\ &\quad - (1 - \beta_1) \langle \nabla f(x_t), g_t \rangle \\ &\leq -\beta_1 \langle \nabla f(x_{t-1}), m_{t-1} \rangle + \beta_1 \|\nabla f(x_t) - \nabla f(x_{t-1})\| \|m_{t-1}\| \\ &\quad - (1 - \beta_1) \langle \nabla f(x_t), g_t \rangle \\ &\leq -\beta_1 \langle \nabla f(x_{t-1}), m_{t-1} \rangle + \beta_1 L \|x_t - x_{t-1}\| \|m_{t-1}\| \\ &\quad - (1 - \beta_1) \langle \nabla f(x_t), g_t \rangle \\ &= -\beta_1 \langle \nabla f(x_{t-1}), m_{t-1} \rangle + \frac{\gamma \beta_1 L}{b_{t-1}} \|m_{t-1}\|^2 \\ &\quad - (1 - \beta_1) \langle \nabla f(x_t), g_t \rangle, \end{aligned}$$

where we use the Cauchy-Schwarz inequality and L -smoothness of f . Applying the same idea for the $t-1, t-2, \dots, 0$ and noting that $m_{-1} = 0$, we get

$$-\langle \nabla f(x_t), m_t \rangle \leq -(1 - \beta_1) \sum_{k=0}^t \beta_1^{t-k} \langle \nabla f(x_k), g_k \rangle + L\gamma \sum_{k=0}^{t-1} \frac{\beta_1^{t-k}}{b_k} \|m_k\|^2. \quad (31)$$

Therefore, substituting (31) into (30), we have

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq -\frac{(1-\beta_1)\gamma}{b_t} \sum_{k=0}^t \beta_1^{t-k} \langle \nabla f(x_k), g_k \rangle + \frac{L\gamma^2}{b_t} \sum_{k=0}^{t-1} \frac{\beta_1^{t-k}}{b_k} \|m_k\|^2 + \frac{L\gamma^2}{2b_t^2} \|m_t\|^2 \\ &\leq -\frac{(1-\beta_1)\gamma}{b_t} \sum_{k=0}^t \beta_1^{t-k} \langle \nabla f(x_k), g_k \rangle + \frac{L\gamma^2}{b_t} \sum_{k=0}^t \frac{\beta_1^{t-k}}{b_k} \|m_k\|^2. \end{aligned}$$

Applying Lemma C.2 with $1 - \beta_1^{k+1} \leq 1$, we can rewrite the inequality above as follows:

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq -\frac{(1-\beta_1)\gamma}{b_t} \sum_{k=0}^t \beta_1^{t-k} \langle \nabla f(x_k), g_k \rangle + \frac{L\gamma^2}{b_t} \sum_{k=0}^t \frac{\beta_1^{t-k}}{b_k} \sum_{j=0}^k \beta_1^{k-j} (1-\beta_1) \|g_j\|^2 \\ &= -\frac{(1-\beta_1)\gamma}{b_t} \sum_{k=0}^t \beta_1^{t-k} \langle \nabla f(x_k), g_k \rangle + \frac{L\gamma^2}{b_t} \sum_{j=0}^t \sum_{k=j}^t \frac{\beta_1^{t-k}}{b_k} \beta_1^{k-j} (1-\beta_1) \|g_j\|^2, \end{aligned} \quad (32)$$

where we change the limits of summation. Now let us bound the second term. Applying Lemma C.1, we obtain that $b_k \geq c_m b_0$ (the constant c_m is taken from Lemma C.1). Consequently,

$$\begin{aligned} \frac{L\gamma^2}{b_t} \sum_{j=0}^t \sum_{k=j}^t \frac{\beta_1^{t-k}}{b_k} \beta_1^{k-j} (1-\beta_1) \|g_j\|^2 &\leq \frac{L\gamma^2(1-\beta_1)}{c_m b_t b_0} \sum_{j=0}^t \sum_{k=j}^t \beta_1^{t-k} \beta_1^{k-j} \|g_j\|^2 \\ &= \frac{L\gamma^2(1-\beta_1)}{c_m b_t b_0} \sum_{j=0}^t \beta_1^{t-j} (t-j+1) \|g_j\|^2. \end{aligned} \quad (33)$$

Thus, substituting (33) into (32), we get

$$f(x_{t+1}) - f(x_t) \leq -\frac{(1-\beta_1)\gamma}{b_t} \sum_{k=0}^t \beta_1^{t-k} \langle \nabla f(x_k), g_k \rangle + \frac{L\gamma^2(1-\beta_1)}{c_m b_t b_0} \sum_{k=0}^t \beta_1^{t-k} (t-k+1) \|g_k\|^2.$$

After summing over $t = 0, \dots, T-1$,

$$f(x_T) - f(x_0) \leq -\sum_{t=0}^{T-1} \frac{(1-\beta_1)\gamma}{b_t} \sum_{k=0}^t \beta_1^{t-k} \langle \nabla f(x_k), g_k \rangle + \sum_{t=0}^{T-1} \frac{L\gamma^2(1-\beta_1)}{c_m b_t b_0} \sum_{k=0}^t \beta_1^{t-k} (t-k+1) \|g_k\|^2.$$

The main idea is to estimate the coefficients corresponding to $\langle \nabla f(x_r), g_r \rangle$ and $\|g_r\|^2$. These multiplicative factors can be estimated as

$$-\sum_{t=r}^{T-1} \frac{\gamma(1-\beta_1)}{b_t} \beta_1^{t-r} \quad (34)$$

for the scalar product and

$$\sum_{t=r}^{T-1} \frac{L\gamma^2(1-\beta_1)}{c_m b_t b_0} (t-r+1) \beta_1^{t-r} \quad (35)$$

for the squared norm, respectively. For (35) we can apply Lemma C.1 in the following way:

$$\sum_{t=r}^{T-1} \frac{L\gamma^2(1-\beta_1)}{c_m b_t b_0} (t-r+1) \beta_1^{t-r} \leq \sum_{t=r}^{T-1} \frac{L\gamma^2(1-\beta_1)}{c_m^2 b_r b_0} (t-r+1) \beta_1^{t-r} = \frac{L\gamma^2(1-\beta_1)}{c_m^2 b_r b_0} \sum_{t=r}^{T-1} (t-r+1) \beta_1^{t-r}.$$

Applying Lemma A.2, and using that $\sum_{t=r}^{T-1} \beta_1^{t-r} \leq \frac{1}{1-\beta_1}$, we get

$$A_r = \sum_{t=r}^{T-1} \frac{L\gamma^2(1-\beta_1)}{c_m b_t b_0} (t-r+1) \beta_1^{t-r} \leq \frac{L\gamma^2}{c_m^2 b_r b_0 (1-\beta_1)} \quad (36)$$

for each $k = 0, \dots, r$. Moreover, let us denote the factor corresponding to the scalar product (34) as $-\gamma C_r$. C_r can be bounded as follows:

$$\frac{(1 - \beta_1)}{b_r} \leq \sum_{t=r}^{T-1} \frac{(1 - \beta_1)}{b_t} \beta_1^{t-r} \leq \sum_{t=r}^{T-1} \frac{(1 - \beta_1)}{c_m b_0} \beta_1^{t-r} \leq \frac{1}{c_m b_0},$$

where we apply Lemma C.1. Therefore, the descent lemma can be formulated as

$$f(x_T) - f(x_0) \leq - \sum_{t=0}^{T-1} \gamma C_t \langle \nabla f(x_t), g_t \rangle + \sum_{t=0}^{T-1} A_t \|g_t\|^2.$$

Substituting the analytical form of g_t , we have

$$\begin{aligned} f(x_T) - f(x_0) &\leq - \sum_{t=0}^{T-1} \gamma C_t \langle \nabla f(x_t), g_t \rangle + \sum_{t=0}^{T-1} A_t \|g_t\|^2 \\ &= - \sum_{t=0}^{T-1} \gamma C_t \left(\langle \nabla f(x_t), \theta_t \rangle + \|\nabla f(x_t)\|^2 \right) \\ &\quad + \sum_{t=0}^{T-1} A_t \left(\|\theta_t\|^2 + 2 \langle \nabla f(x_t), \theta_t \rangle + \|\nabla f(x_t)\|^2 \right) \\ &= - \sum_{t=0}^{T-1} (\gamma C_t - A_t) \|\nabla f(x_t)\|^2 - \sum_{t=0}^{T-1} (\gamma C_t - 2A_t) \langle \nabla f(x_t), \theta_t \rangle \\ &\quad + \sum_{t=0}^{T-1} A_t \|\theta_t\|^2. \end{aligned}$$

Choosing $\gamma \leq \frac{(1 - \beta_1)^2 c_m^2 b_0}{2L}$, we get that $\gamma C_t - 2A_t \geq 0$ since the boundary $C_t \geq \frac{1 - \beta_1}{b_t}$ and (36) hold with $k = t$. Therefore, using that $\theta_t = \theta_t^u + \theta_t^b$, one can obtain

$$\begin{aligned} f(x_T) - f(x_0) &\leq - \sum_{t=0}^{T-1} (\gamma C_t - A_t) \|\nabla f(x_t)\|^2 - \sum_{t=0}^{T-1} (\gamma C_t - 2A_t) \langle \nabla f(x_t), \theta_t \rangle + \sum_{t=0}^{T-1} A_t \|\theta_t\|^2 \\ &\leq - \sum_{t=0}^{T-1} (\gamma C_t - A_t) \|\nabla f(x_t)\|^2 - \sum_{t=0}^{T-1} (\gamma C_t - 2A_t) \langle \nabla f(x_t), \theta_t^u \rangle \\ &\quad + \sum_{t=0}^{T-1} 2A_t \left(\|\theta_t^u\|^2 + \|\theta_t^b\|^2 \right) + \sum_{t=0}^{T-1} \left(\frac{\gamma C_t}{2} - A_t \right) \|\nabla f(x_t)\|^2 + \sum_{t=0}^{T-1} \left(\frac{\gamma C_t}{2} - A_t \right) \|\theta_t^b\|^2 \\ &= - \sum_{t=0}^{T-1} \frac{\gamma C_t}{2} \|\nabla f(x_t)\|^2 - \sum_{t=0}^{T-1} (\gamma C_t - 2A_t) \langle \nabla f(x_t), \theta_t^u \rangle + \sum_{t=0}^{T-1} 2A_t \|\theta_t^u\|^2 + \sum_{t=0}^{T-1} \left(\frac{\gamma C_t}{2} + A_t \right) \|\theta_t^b\|^2. \end{aligned}$$

Using that $\frac{\gamma C_t}{2} \geq A_t$, and rearranging terms with $\Delta_t = f(x_t) - f_*$, we get the final result. \square

Remark C.4. It is important to note that Q can be any non-empty subset of \mathbb{R}^d as long as the iterates belong to it. In this sense, the form of Q is not that important for the proof (a similar observation holds for Lemma C.6 in the convex case). Nevertheless, Q plays a key role in the next part of the proof.

Theorem C.5. Let Assumptions 1.1 and 1.2 hold on $Q = \{x \in \mathbb{R}^d \mid \exists y \in \mathbb{R}^d : f(y) \leq f_* + 2\Delta \text{ and } \|x - y\| \leq \frac{\sqrt{\Delta}}{20\sqrt{L}}\}$ with $f(x_0) - f_* = \Delta_0 \leq \Delta$. Then, after $K + 1$ iterations of Clip-M-AdaGradD/Clip-AdamD-Norm with

$$\gamma \leq \min \left\{ \frac{(1 - \beta_1)^2 c_m^2 b_0 (K + 1)^{\frac{1-\alpha}{3\alpha-2}}}{80L \ln \frac{4(K+1)}{\delta}}, \frac{c_m \sqrt{1 - \beta_1} 35^{\frac{1}{\alpha}} b_0 \sqrt{\Delta}}{432^{\frac{1}{\alpha}} \cdot 20\sqrt{L} \sigma (K + 1)^{\frac{\alpha-1}{3\alpha-2}} \ln^{\frac{\alpha-1}{\alpha}} \frac{4(K+1)}{\delta}}, \right. \\ \left. \frac{c_m (1 - \beta_1)^{\frac{\alpha-1}{2\alpha-1}} b_0 \Delta^{\frac{\alpha}{2\alpha-1}}}{4^{\frac{\alpha+1}{2\alpha-1}} \cdot 20^{\frac{2\alpha-2}{2\alpha-1}} \sigma^{\frac{2\alpha}{2\alpha-1}} L^{\frac{\alpha-1}{2\alpha-1}} (K + 1)^{\frac{\alpha}{3\alpha-2}} \ln^{\frac{2\alpha-2}{2\alpha-1}} \left(\frac{4(K+1)}{\delta} \right)} \right\}, \quad \eta = \frac{L\gamma^2(1 - \beta_1)^2}{\Delta}, \quad (37)$$

and

$$\lambda = \frac{c_m \sqrt{1 - \beta_1} b_0 \sqrt{\Delta} (K + 1)^{\frac{1-\alpha}{3\alpha-2}}}{20\sqrt{L}\gamma \ln\left(\frac{4(K+1)}{\delta}\right)} \quad (38)$$

the bound

$$\sum_{k=0}^K \frac{\gamma C_k}{2} \|\nabla f(x_k)\|^2 \leq 2\Delta$$

holds with probability at least $1 - \delta$. In particular, when γ equals the minimum from (37), the iterates produced by Clip-M-AdaGradD/Clip-AdamD-Norm satisfy

$$\begin{aligned} & \frac{1}{K+1} \sum_{k=0}^K \|\nabla f(x_k)\|^2 \\ &= \mathcal{O}\left(\max\left\{\frac{L\Delta \ln \frac{K+1}{\delta}}{(1-\beta_1)^3(K+1)^{\frac{2\alpha-1}{3\alpha-2}}}, \frac{\sqrt{L\Delta}\sigma \ln \frac{\alpha-1}{\alpha} \frac{K+1}{\delta}}{(1-\beta_1)^{\frac{3}{2}}(K+1)^{\frac{2\alpha-2}{3\alpha-2}}}, \frac{\sigma^{\frac{2\alpha}{2\alpha-1}}(L\Delta)^{\frac{\alpha-1}{2\alpha-1}} \ln \frac{2\alpha-2}{2\alpha-1} \frac{K+1}{\delta}}{(1-\beta_1)^{\frac{3\alpha-2}{2\alpha-1}}(K+1)^{\frac{2\alpha-2}{3\alpha-2}}}\right\}\right) \end{aligned}$$

with probability at least $1 - \delta$.

Proof. Our proof is induction-based (similarly to the one for Clip-SGD by Sadiev et al. (2023)). We introduce probability event E_k as follows: inequalities

$$\begin{aligned} & - \sum_{l=0}^{t-1} (\gamma C_l - 2A_l) \langle \nabla f(x_l), \theta_l^u \rangle + \sum_{l=0}^{t-1} \gamma C_l \|\theta_l^b\|^2 + \sum_{l=0}^{t-1} 2A_l \|\theta_l^u\|^2 \leq \Delta, \\ & \Delta_t \leq 2\Delta \end{aligned}$$

hold simultaneously $\forall t = 0, 1, \dots, k$. We want to show that $\mathbb{P}\{E_k\} \geq 1 - \frac{k\delta}{K+1} \forall k = 0, 1, \dots, K+1$. The case when $k = 0$ is obvious. Now let us make an induction step: let the statement hold for some $k = T-1 \leq K$: $\mathbb{P}\{E_{T-1}\} \geq 1 - \frac{(T-1)\delta}{K+1}$. It remains to prove that $\mathbb{P}\{E_T\} \geq 1 - \frac{T\delta}{K+1}$. The event E_{T-1} implies that $x_t \in \{y \in \mathbb{R}^d : f(y) \leq f_* + 2\Delta\} \forall t = 0, \dots, T-1$ and

$$\|x_T - x_{T-1}\| = \frac{\gamma}{b_t} \|m_{T-1}\| \leq \frac{\gamma\lambda}{c_m b_0} \leq \frac{\sqrt{\Delta}}{20\sqrt{L} \ln \frac{4(K+1)}{\delta}} \leq \frac{\sqrt{\Delta}}{20\sqrt{L}}.$$

Hence, event E_{T-1} implies $\{x_t\}_{t=0}^T \subseteq Q$ and we can apply Lemma C.3:

$$\sum_{l=0}^{t-1} \frac{\gamma C_l}{2} \|\nabla f(x_l)\|^2 \leq \Delta_0 - \Delta_t - \sum_{l=0}^{t-1} (\gamma C_l - 2A_l) \langle \nabla f(x_l), \theta_l^u \rangle + \sum_{l=0}^{t-1} \gamma C_l \|\theta_l^b\|^2 + \sum_{l=0}^{t-1} 2A_l \|\theta_l^u\|^2$$

$\forall t = 1, \dots, T$ and $\forall t = 1, \dots, T-1$ it implies that

$$\sum_{l=0}^{t-1} \frac{\gamma C_l}{2} \|\nabla f(x_l)\|^2 \leq \Delta_0 - \Delta_t - \sum_{l=0}^{t-1} (\gamma C_l - 2A_l) \langle \nabla f(x_l), \theta_l^u \rangle + \sum_{l=0}^{t-1} \gamma C_l \|\theta_l^b\|^2 + \sum_{l=0}^{t-1} 2A_l \|\theta_l^u\|^2 \leq 2\Delta.$$

Taking into account that $\sum_{l=0}^{t-1} \frac{\gamma C_l}{2} \|\nabla f(x_l)\|^2 \geq 0$ for all t , we get that E_{T-1} implies

$$\begin{aligned} \Delta_T &\leq \Delta_0 - \sum_{t=0}^{T-1} (\gamma C_t - 2A_t) \langle \nabla f(x_t), \theta_t^u \rangle + \sum_{t=0}^{T-1} \gamma C_t \|\theta_t^b\|^2 + \sum_{t=0}^{T-1} 2A_t \|\theta_t^u\|^2 \\ &= \Delta_0 - \sum_{t=0}^{T-1} (\gamma C_t - 2A_t) \langle \nabla f(x_t), \theta_t^u \rangle + \sum_{t=0}^{T-1} \gamma C_t \|\theta_t^b\|^2 \\ &\quad + \sum_{t=0}^{T-1} 2A_t \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right) + \sum_{t=0}^{T-1} 2A_t \mathbb{E}_{\xi_t} \|\theta_t^u\|^2. \end{aligned}$$

Next, for vectors

$$\eta_t = \begin{cases} \nabla f(x_t), & \|\nabla f(x_t)\| \leq 2\sqrt{L\Delta} \\ 0, & \text{otherwise} \end{cases}$$

for all $t = 0, 1, \dots, T-1$, we have that that with probability 1

$$\|\eta_t\| \leq 2\sqrt{L\Delta}. \quad (39)$$

What is more, for all $t = 0, \dots, T-1$ E_{T-1} implies

$$\|\nabla f(x_t)\| \leq \sqrt{2L\Delta_t} \leq 2\sqrt{L\Delta} \stackrel{(38)}{\leq} \frac{\lambda}{2}.$$

Thus, E_{T-1} implies $\eta_t = \nabla f(x_t)$ for $t = 0, 1, \dots, T-1$ and

$$\Delta_T \leq \Delta_0 - \underbrace{\sum_{t=0}^{T-1} (\gamma C_t - 2A_t) \langle \eta_t, \theta_t^u \rangle}_{\textcircled{1}} + \underbrace{\sum_{t=0}^{T-1} \gamma C_t \|\theta_t^b\|^2}_{\textcircled{2}} + \underbrace{\sum_{t=0}^{T-1} 2A_t (\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2)}_{\textcircled{3}} + \underbrace{\sum_{t=0}^{T-1} 2A_t \mathbb{E}_{\xi_t} \|\theta_t^u\|^2}_{\textcircled{4}}. \quad (40)$$

It remains to bound each term in (40) separately with high probability. Before we move on, we also note that event E_{T-1} implies $\|\nabla f(x_t)\| \leq \frac{\lambda}{2}$. Therefore, one can apply Lemma A.4 and get

$$\|\theta_t^u\| \leq 2\lambda, \quad (41)$$

$$\|\theta_t^b\| \leq \frac{2^\alpha \sigma^\alpha}{\lambda^{\alpha-1}}, \quad (42)$$

$$\mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \leq 18\lambda^{2-\alpha} \sigma^\alpha. \quad (43)$$

Bound for ①. The definition of θ_t^u implies

$$\mathbb{E}_{\xi_t} [-(\gamma C_t - 2A_t) \langle \eta_t, \theta_t^u \rangle] = 0.$$

What is more, since $C_t \leq \frac{1}{c_m b_0}$, we get

$$|(\gamma C_t - 2A_t) \langle \eta_t, \theta_t^u \rangle| \leq \gamma C_t \|\eta_t\| \|\theta_t^u\| \stackrel{(39),(41)}{\leq} \frac{4\gamma\lambda\sqrt{L\Delta}}{c_m b_0} \leq \frac{\Delta}{5 \ln \left(\frac{4(K+1)}{\delta} \right)} = c.$$

Let us define $\sigma_t^2 = \mathbb{E}_{\xi_t} [(\gamma C_t - 2A_t)^2 \langle \eta_t, \theta_t^u \rangle^2]$. Hence,

$$\sigma_t^2 \stackrel{(39)}{\leq} (\gamma C_t - 2A_t)^2 \cdot 4L\Delta \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \leq \frac{4\gamma^2 L\Delta}{c_m^2 b_0^2} \mathbb{E}_{\xi_t} \|\theta_t^u\|^2. \quad (44)$$

Therefore, we can apply Bernstein's inequality (Lemma A.5) with $G = \frac{7\Delta^2}{480 \ln \frac{4(K+1)}{\delta}}$:

$$\mathbb{P} \left\{ \left| - \sum_{t=0}^{T-1} (\gamma C_t - 2A_t) \langle \nabla f(x_t), \theta_t^u \rangle \right| > \frac{\Delta}{4} \text{ and } \sum_{t=0}^{T-1} \sigma_t^2 \leq G \right\} \leq 2 \exp \left(- \frac{\Delta^2}{16(2G + \frac{\Delta c}{6})} \right) = \frac{\delta}{2(K+1)}.$$

Thus, we get

$$\mathbb{P} \left\{ \text{either } \left| - \sum_{t=0}^{T-1} (\gamma C_t - 2A_t) \langle \nabla f(x_t), \theta_t^u \rangle \right| \leq \frac{\Delta}{4} \text{ or } \sum_{t=0}^{T-1} \sigma_t^2 > G \right\} \geq 1 - \frac{\delta}{2(K+1)}.$$

Moreover, event E_{T-1} implies

$$\begin{aligned} \sum_{t=0}^{T-1} \sigma_t^2 &\stackrel{(43)}{\leq} \frac{72\gamma^2\lambda^{2-\alpha}\sigma^\alpha L\Delta T}{c_m^2 b_0^2} \stackrel{(38)}{=} \frac{72c_m^{2-\alpha}(1-\beta_1)^{1-\frac{\alpha}{2}}\gamma^\alpha b_0^{2-\alpha}\sqrt{\Delta}^{2-\alpha}(K+1)^{\frac{\alpha^2-3\alpha+2}{3\alpha-2}}\sigma^\alpha L\Delta T}{c_m^2 20^{2-\alpha}\sqrt{L}^{2-\alpha}b_0^2 \ln^{2-\alpha}\frac{4(K+1)}{\delta}} \\ &\stackrel{(37)}{\leq} \frac{7\Delta^2}{480 \ln \frac{4(K+1)}{\delta}}. \end{aligned}$$

Bound for ②. For the second term, we get that E_{T-1} implies

$$\begin{aligned} \sum_{t=0}^{T-1} \gamma C_t \|\theta_t^b\|^2 &\leq \sum_{t=0}^{T-1} \frac{\gamma}{c_m b_0} \|\theta_t^b\|^2 \stackrel{(42)}{\leq} \frac{4^\alpha \sigma^{2\alpha} \gamma T}{c_m \lambda^{2\alpha-2} b_0} \\ &\stackrel{(38)}{\leq} \frac{4^\alpha \sigma^{2\alpha} \gamma (K+1)}{c_m b_0} \cdot \frac{20^{2\alpha-2} L^{\alpha-1} \gamma^{2\alpha-2} (K+1)^{\frac{(\alpha-1)(2\alpha-2)}{3\alpha-2}} \ln^{2\alpha-2} \left(\frac{4(K+1)}{\delta} \right)}{c_m^{2\alpha-2} (1-\beta_1)^{\alpha-1} b_0^{2\alpha-2} \Delta^{\alpha-1}} \\ &= \frac{4^\alpha \cdot 20^{2\alpha-2} \sigma^{2\alpha} L^{\alpha-1} (K+1)^{\frac{\alpha(2\alpha-1)}{3\alpha-2}} \ln^{2\alpha-2} \left(\frac{4(K+1)}{\delta} \right)}{c_m^{2\alpha-1} (1-\beta_1)^{\alpha-1} b_0^{2\alpha-1} \Delta^{\alpha-1}} \cdot \gamma^{2\alpha-1} \\ &\stackrel{(37)}{\leq} \frac{\Delta}{4}, \end{aligned}$$

where in the last step, we apply the third condition on γ from (37).

Bound for ③. Similarly to ①, we have unbiased and bounded terms in the sum:

$$\mathbb{E}_{\xi_t} \left[2A_t \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right) \right] = 0$$

and, since (36) from Lemma C.3 hold with $k=0$,

$$\left| 2A_t \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right) \right| \stackrel{(41)}{\leq} \frac{16L\lambda^2\gamma^2}{c_m^2 b_0^2 (1-\beta_1)} \leq \frac{\Delta}{25 \ln \frac{4(K+1)}{\delta}} \leq \frac{15\Delta}{47 \ln \frac{4(K+1)}{\delta}} = c. \quad (45)$$

Next, we define $\hat{\sigma}_t^2 = \mathbb{E}_{\xi_t} \left[4A_t^2 \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right)^2 \right]$. For the introduced quantities, we have

$$\hat{\sigma}_t^2 \stackrel{(45)}{\leq} c \mathbb{E}_{\xi_t} \left[2A_t \left| \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right) \right| \right] \leq \frac{4L\gamma^2 c}{c_m^2 b_0^2 (1-\beta_1)} \mathbb{E}_{\xi_t} \|\theta_t^u\|^2. \quad (46)$$

Therefore, we can apply Bernstein's inequality (Lemma A.5) with $G = \frac{7\Delta^2}{1504 \ln \frac{4(K+1)}{\delta}}$:

$$\mathbb{P} \left\{ \left| \sum_{t=0}^{T-1} 2A_t \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right) \right| > \frac{\Delta}{4} \text{ and } \sum_{t=0}^{T-1} \hat{\sigma}_t^2 \leq G \right\} \leq 2 \exp \left(-\frac{\Delta^2}{16(2G + \frac{\Delta c}{6})} \right) = \frac{\delta}{2(K+1)}.$$

Thus, we get

$$\mathbb{P} \left\{ \text{either } \left| \sum_{t=0}^{T-1} 2A_t \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right) \right| \leq \frac{\Delta}{4} \text{ or } \sum_{t=0}^{T-1} \hat{\sigma}_t^2 > G \right\} \geq 1 - \frac{\delta}{2(K+1)}.$$

Moreover, event E_{T-1} implies

$$\begin{aligned} \sum_{t=0}^{T-1} \hat{\sigma}_t^2 &\stackrel{(46),(41)}{\leq} \frac{72L\gamma^2 c \lambda^{2-\alpha} \sigma^\alpha T}{c_m^2 b_0^2 (1-\beta_1)} \stackrel{(38)}{\leq} \frac{72c \gamma^\alpha b_0^{2-\alpha} \sqrt{\Delta}^{2-\alpha} (K+1)^{\frac{\alpha^2-3\alpha+2}{3\alpha-2}} \sigma^\alpha L T}{20^{2-\alpha} c_m^\alpha (1-\beta_1)^{\frac{\alpha}{2}} \sqrt{L}^{2-\alpha} b_0^2 \ln^{2-\alpha} \frac{4(K+1)}{\delta}} \\ &\stackrel{(37)}{\leq} \frac{7\Delta c}{480} \leq \frac{7\Delta^2}{1504 \ln \frac{4(K+1)}{\delta}}. \end{aligned}$$

Bound for ④. For the last term, we have that E_{T-1} implies

$$\begin{aligned} \sum_{t=0}^{T-1} 2A_t \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 &\leq \sum_{t=0}^{T-1} \frac{2L\gamma^2}{c_m^2 b_0^2(1-\beta_1)} \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \\ &\stackrel{(41)}{\leq} \frac{36L\gamma^2 \lambda^{2-\alpha} \sigma^\alpha T}{c_m^2 b_0^2(1-\beta_1)} \stackrel{(38)}{\leq} \frac{36\gamma^\alpha b_0^{2-\alpha} \sqrt{\Delta}^{2-\alpha} (K+1)^{\frac{\alpha^2-3\alpha+2}{3\alpha-2}} \sigma^\alpha L T}{20^{2-\alpha} (1-\beta_1)^{\frac{\alpha}{2}} \sqrt{L}^{2-\alpha} b_0^2 \ln^{2-\alpha} \frac{4(K+1)}{\delta}} \\ &\stackrel{(37)}{\leq} \frac{7\Delta}{960 \ln \frac{4(K+1)}{\delta}} \leq \frac{\Delta}{4}. \end{aligned}$$

Thus, taking into account the bounds above, the probability event $E_{T-1} \cap E_1 \cap E_2$ implies that

$$\Delta_T \leq \Delta + 4 \frac{\Delta}{4} = 2\Delta,$$

where

$$\begin{aligned} E_1 &= \left\{ \text{either } \left| - \sum_{t=0}^{T-1} \left(\frac{\gamma}{b_t} - \frac{L\gamma^2}{b_t^2} \right) \langle \nabla f(x_t), \theta_t^u \rangle \right| \leq \frac{\Delta}{4} \text{ or } \sum_{t=0}^{T-1} \sigma_t^2 > \frac{7\Delta^2}{480 \ln \frac{4(K+1)}{\delta}} \right\}, \\ E_2 &= \left\{ \text{either } \left| \sum_{t=0}^{T-1} \frac{L\gamma^2}{b_t^2} (\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2) \right| \leq \frac{\Delta}{4} \text{ or } \sum_{t=0}^{T-1} \hat{\sigma}_t^2 > \frac{7\Delta^2}{1504 \ln \frac{4(K+1)}{\delta}} \right\}. \end{aligned}$$

Therefore,

$$\mathbb{P}\{E_T\} \geq \mathbb{P}\{E_{T-1} \cap E_1 \cap E_2\} = 1 - \mathbb{P}\{\bar{E}_{T-1} \cup \bar{E}_1 \cup \bar{E}_2\} \geq 1 - \mathbb{P}\{\bar{E}_{T-1}\} - \mathbb{P}\{\bar{E}_1\} - \mathbb{P}\{\bar{E}_2\} \geq 1 - \frac{T\delta}{K+1}.$$

Hence, for all $k = 0, \dots, K+1$ we get $\mathbb{P}(E_k) \geq 1 - \frac{k\delta}{K+1}$. As revision result, event E_{K+1} implies that

$$\sum_{k=0}^K \frac{\gamma C_k}{2} \|\nabla f(x_k)\|^2 \leq 2\Delta \tag{47}$$

holds with probability at least $1 - \delta$.

Therefore, we get that with probability at least $1 - \delta$

$$\sum_{k=0}^K \|\nabla f(x_k)\|^2 \leq \frac{4\Delta}{\gamma} \max_{k \in [0, K]} \frac{1}{C_k}.$$

and, since $C_k \geq \frac{1-\beta_1}{b_k}$, we obtain

$$\sum_{k=0}^K \|\nabla f(x_k)\|^2 \leq \frac{4\Delta}{\gamma(1-\beta_1)} \max_{k \in [0, K]} b_k. \tag{48}$$

Moreover,

$$b_k^2 \leq b_0^2 + \eta \sum_{k=0}^K \left(3\|\nabla f(x_k)\|^2 + 3\|\theta_k^u\|^2 + 3\|\theta_k^b\|^2 \right) \tag{49}$$

for the **Clip-AdaGradD** of b_k and

$$b_k^2 \leq b_0^2 + \frac{\eta}{K+1} \sum_{k=0}^K \left(3\|\nabla f(x_k)\|^2 + 3\|\theta_k^u\|^2 + 3\|\theta_k^b\|^2 \right) \tag{50}$$

for the Clip-AdamD, respectively. Next, we use that the event E_{K+1} implies

$$\sum_{k=0}^K \frac{\gamma}{c_m b_0} \|\theta_k^b\|^2 \leq \frac{\Delta}{4}; \quad \sum_{k=0}^K \frac{2L\gamma^2}{c_m^2 b_0^2 (1 - \beta_1)} \|\theta_k^u\|^2 \leq \frac{\Delta}{2}$$

because we could replace $b_t \rightarrow c_m b_0$ into C_t and A_t , and all steps in ②, ③ and ④ will be the same. Therefore, with applying Lemma C.1, next bounds

$$\begin{aligned} \sum_{k=0}^K \|\nabla f(x_k)\|^2 &\leq \frac{4\Delta}{\gamma(1 - \beta_1)} \sqrt{b_0^2 + 3\eta \sum_{k=0}^K \|\nabla f(x_k)\|^2 + \frac{3\eta b_0 \Delta}{4\gamma} + \frac{3\eta b_0^2 (1 - \beta_1) \Delta}{4L\gamma^2}}; \\ \sum_{k=0}^K \|\nabla f(x_k)\|^2 &\leq \frac{4\Delta}{\gamma(1 - \beta_1)} \sqrt{b_0^2 + \frac{3\eta}{K+1} \sum_{k=0}^K \|\nabla f(x_k)\|^2 + \frac{3\eta b_0 \Delta}{8\gamma(K+1)} + \frac{3\eta b_0^2 (1 - \beta_1) \Delta}{16L\gamma^2(K+1)}} \end{aligned}$$

hold with probability at least $1 - \delta$, where we substitute different c_m from Lemma C.1 and (49), (50) for Clip-M-AdaGradD-Norm and Clip-AdamD-Norm, respectively. Next, solving quadratic inequalities above with respect to $\sum_{k=0}^K \|\nabla f(x_k)\|^2$, we obtain

$$\begin{aligned} \sum_{k=0}^K \|\nabla f(x_k)\|^2 &\leq \frac{\frac{48\eta\Delta^2}{\gamma^2(1-\beta_1)^2} + \sqrt{\frac{9\cdot4^4\eta^2\Delta^4}{\gamma^4(1-\beta_1)^4} + \frac{16\Delta^2}{\gamma^2(1-\beta_1)^2} \left(\frac{3\eta b_0 \Delta}{4\gamma} + \frac{3\eta b_0^2 (1 - \beta_1) \Delta}{4L\gamma^2} + b_0^2 \right)}}{2} \\ &= \frac{24\eta\Delta^2}{\gamma^2(1-\beta_1)^2} + \sqrt{\frac{576\eta^2\Delta^4}{\gamma^4(1-\beta_1)^4} + \left(\frac{3\eta b_0 \Delta^3}{\gamma^3(1-\beta_1)^2} + \frac{3\eta b_0^2 \Delta^3}{L\gamma^4(1-\beta_1)} + \frac{4b_0^2 \Delta^2}{\gamma^2(1-\beta_1)^2} \right)} \\ &= \frac{\Delta}{\gamma^2} \left(\frac{24\eta\Delta}{(1-\beta_1)^2} + \sqrt{\frac{576\eta^2\Delta^2}{(1-\beta_1)^4} + \left(\frac{3\eta b_0 \gamma \Delta}{(1-\beta_1)^2} + \frac{3\eta b_0^2 \Delta}{L(1-\beta_1)} + \frac{4b_0^2 \gamma^2}{(1-\beta_1)^2} \right)} \right) \end{aligned}$$

for Clip-M-AdaGradD-Norm and

$$\begin{aligned} \sum_{k=0}^K \|\nabla f(x_k)\|^2 &\leq \frac{24\eta\Delta^2}{\gamma^2(1-\beta_1)^2(K+1)} \\ &+ \sqrt{\frac{9 \cdot 4^3 \eta^2 \Delta^4}{\gamma^4(1-\beta_1)^4(K+1)^2} + \frac{4\Delta^2}{\gamma^2(1-\beta_1)^2} \left(\frac{3\eta b_0 \Delta}{8\gamma(K+1)} + \frac{3\eta b_0^2 (1 - \beta_1) \Delta}{16L\gamma^2(K+1)} + b_0^2 \right)} \\ &= \frac{24\eta\Delta^2}{\gamma^2(1-\beta_1)^2(K+1)} \\ &+ \sqrt{\frac{576\eta^2\Delta^4}{\gamma^4(1-\beta_1)^4(K+1)^2} + \left(\frac{3\eta b_0 \Delta^3}{2\gamma^3(1-\beta_1)^2(K+1)} + \frac{3\eta b_0^2 \Delta^3}{4L\gamma^4(1-\beta_1)(K+1)} + \frac{4b_0^2 \Delta^2}{\gamma^2(1-\beta_1)^2} \right)} \\ &= \frac{\Delta}{\gamma^2} \left(\frac{24\eta\Delta}{(1-\beta_1)^2(K+1)} \right. \\ &\left. + \sqrt{\frac{576\eta^2\Delta^2}{(1-\beta_1)^4(K+1)^2} + \left(\frac{3\eta b_0 \gamma \Delta}{2(1-\beta_1)^2(K+1)} + \frac{3\eta b_0^2 \Delta}{4L(1-\beta_1)(K+1)} + \frac{4b_0^2 \gamma^2}{(1-\beta_1)^2} \right)} \right) \end{aligned}$$

for the Clip-AdamD-Norm. Substituting $\eta = \frac{L\gamma^2(1-\beta_1)^2}{\Delta}$ and applying $\sqrt{a^2 + b^2 + c^2 + d^2} \leq a+b+c+d$ for non-negative

numbers, one can obtain the bound for Clip-M-AdaGradD-Norm:

$$\begin{aligned}
 \frac{1}{K+1} \sum_{k=0}^K \|\nabla f(x_k)\|^2 &\leq \frac{\Delta}{(K+1)\gamma^2} \left(48L\gamma^2 + \sqrt{3L\gamma^3 b_0} + \sqrt{3\gamma^2 b_0^2(1-\beta_1)} + \frac{2\gamma b_0}{1-\beta_1} \right) \\
 &\leq \frac{\Delta}{(K+1)\gamma^2} \left(49L\gamma^2 + 3\sqrt{\gamma^2 b_0^2(1-\beta_1)} + \frac{2\gamma b_0}{1-\beta_1} \right) \\
 &\leq \frac{\Delta}{(K+1)\gamma^2} \left(49L\gamma^2 + 3\gamma b_0 + \frac{2\gamma b_0}{1-\beta_1} \right) \\
 &\leq \frac{2\Delta}{(K+1)\gamma^2} \max \left\{ 49L\gamma^2, \frac{5\gamma b_0}{1-\beta_1} \right\} \\
 &= \max \left\{ \frac{98L\Delta}{K+1}, \frac{10\Delta b_0}{\gamma(K+1)(1-\beta_1)} \right\}
 \end{aligned} \tag{51}$$

and for Clip-AdamD-Norm:

$$\begin{aligned}
 \frac{1}{K+1} \sum_{k=0}^K \|\nabla f(x_k)\|^2 &\leq \frac{\Delta}{(K+1)\gamma^2} \left(\frac{48L\gamma^2}{K+1} + \sqrt{\frac{3L\gamma^3 b_0}{2(K+1)}} + \sqrt{\frac{3\gamma^2 b_0^2(1-\beta_1)}{4(K+1)}} + \frac{2\gamma b_0}{1-\beta_1} \right) \\
 &\leq \frac{\Delta}{(K+1)\gamma^2} \left(\frac{48L\gamma^2}{K+1} + 2\sqrt{\frac{L\gamma^3 b_0}{(K+1)}} + \gamma b_0 + \frac{2\gamma b_0}{1-\beta_1} \right) \\
 &\leq \frac{\Delta}{(K+1)\gamma^2} \left(\frac{49L\gamma^2}{K+1} + \frac{4\gamma b_0}{1-\beta_1} \right) \\
 &\leq \frac{2\Delta}{(K+1)\gamma^2} \max \left\{ \frac{49L\gamma^2}{K+1}, \frac{4\gamma b_0}{1-\beta_1} \right\} \\
 &= \max \left\{ \frac{98L\Delta}{(K+1)^2}, \frac{8\Delta b_0}{\gamma(K+1)(1-\beta_1)} \right\},
 \end{aligned} \tag{52}$$

where we use that $2\sqrt{ab} \leq a + b$. Consequently, after substitution of (37) into (51), (52), we get final bounds for Clip-M-AdaGradD/Clip-AdamD-Norm:

$$\begin{aligned}
 &\frac{1}{K+1} \sum_{k=0}^K \|\nabla f(x_k)\|^2 \\
 &= \mathcal{O} \left(\max \left\{ \frac{L\Delta \ln \frac{K+1}{\delta}}{(1-\beta_1)^3(K+1)^{\frac{2\alpha-1}{3\alpha-2}}}, \frac{\sqrt{L\Delta}\sigma \ln^{\frac{\alpha-1}{\alpha}} \frac{K+1}{\delta}}{(1-\beta_1)^{\frac{3}{2}}(K+1)^{\frac{2\alpha-2}{3\alpha-2}}}, \frac{\sigma^{\frac{2\alpha}{2\alpha-1}} (L\Delta)^{\frac{\alpha-1}{2\alpha-1}} \ln^{\frac{2\alpha-2}{2\alpha-1}} \frac{K+1}{\delta}}{(1-\beta_1)^{\frac{3\alpha-2}{2\alpha-1}}(K+1)^{\frac{2\alpha-2}{3\alpha-2}}} \right\} \right)
 \end{aligned}$$

holds with probability at least $1 - \delta$. \square

C.3. Convex Case: Methods with Delay

Lemma C.6 (Descent lemma). *Let Assumptions 1.2 and 1.3 hold on $Q = B_{2R}(x^*)$, where $\|x_0 - x^*\| \leq R$. Assume that $x_t \in Q \forall t = 0, T$. Then, after T iterations of Clip-M-AdaGradD-Norm/Clip-AdamD-Norm with $b_0 \geq \frac{8\gamma L}{(1-\beta_1)^2 c_m^2}$, we have*

$$\sum_{t=0}^{T-1} \gamma C_t (f(x_t) - f_*) \leq R_0^2 - R_t^2 - \sum_{t=0}^{T-1} 2\gamma C_t \langle x_t - x^*, \theta_t \rangle + \sum_{t=0}^{T-1} 2A_t \|\theta_t\|^2,$$

where $C_t = \sum_{i=t}^{T-1} \frac{1-\beta_1}{b_i} \beta_1^{i-t}$ and $A_t = \sum_{i=t}^{T-1} \frac{2\gamma^2(1-\beta_1)}{c_m b_i b_0} \beta_1^{i-t} (i-t+1)$.

Proof. According to the update rule of Algorithm 3, we have

$$\|x_{t+1} - x^*\|^2 = \|x_t - x^*\|^2 - \frac{2\gamma}{b_t} \langle x_t - x^*, m_t \rangle + \frac{\gamma^2}{b_t^2} \|m_t\|^2.$$

To bound the scalar product, we substitute the update rule for m_t :

$$\begin{aligned} -\langle x_t - x^*, m_t \rangle &= -\beta_1 \langle x_t - x^*, m_{t-1} \rangle - (1 - \beta_1) \langle x_t - x^*, g_t \rangle \\ &= -\beta_1 \langle x_t - x_{t-1}, m_{t-1} \rangle - \beta_1 \langle x_{t-1} - x^*, m_{t-1} \rangle \\ &\quad - (1 - \beta_1) \langle x_t - x^*, g_t \rangle \\ &\leq -\beta_1 \langle x_{t-1} - x^*, m_{t-1} \rangle - (1 - \beta_1) \langle x_t - x^*, g_t \rangle \\ &\quad + \beta_1 \|x_t - x_{t-1}\| \|m_{t-1}\| \\ &= -\beta_1 \langle x_{t-1} - x^*, m_{t-1} \rangle - (1 - \beta_1) \langle x_t - x^*, g_t \rangle \\ &\quad + \frac{\gamma\beta_1}{b_{t-1}} \|m_{t-1}\|^2. \end{aligned}$$

Applying the same idea for $t-1, t-2, \dots, 0$ and using that $m_{-1} = 0$, one can obtain

$$-\langle x_t - x^*, m_t \rangle \leq -\sum_{k=0}^t (1 - \beta_1) \beta_1^{t-k} \langle x_k - x^*, g_k \rangle + \sum_{k=0}^{t-1} \frac{\gamma\beta_1^{t-k}}{b_k} \|m_k\|^2.$$

Therefore, we get

$$\|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 - \frac{2\gamma}{b_t} \sum_{k=0}^t (1 - \beta_1) \beta_1^{t-k} \langle x_k - x^*, g_k \rangle + \frac{2\gamma^2}{b_t} \sum_{k=0}^t \frac{\beta_1^{t-k}}{b_k} \|m_k\|^2.$$

Substituting the bound for $\|m_k\|^2$ from Lemma C.2 with $1 - \beta_1^{k+1} \leq 1$, we have

$$\begin{aligned} \|x_{t+1} - x^*\|^2 &\leq \|x_t - x^*\|^2 - \frac{2\gamma}{b_t} \sum_{k=0}^t (1 - \beta_1) \beta_1^{t-k} \langle x_k - x^*, g_k \rangle + \frac{2\gamma^2}{b_t} \sum_{k=0}^t \frac{\beta_1^{t-k}}{b_k} \sum_{j=0}^k \beta_1^{k-j} (1 - \beta_1) \|g_j\|^2 \\ &= \|x_t - x^*\|^2 - \frac{2\gamma}{b_t} \sum_{k=0}^t (1 - \beta_1) \beta_1^{t-k} \langle x_k - x^*, g_k \rangle + \frac{2\gamma^2}{b_t} \sum_{k=0}^t \sum_{j=0}^k \frac{\beta_1^{t-j}}{b_k} (1 - \beta_1) \|g_j\|^2. \end{aligned}$$

Applying the same technique as in Lemma C.3 (see (33)), one can obtain

$$\|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 - \frac{2\gamma(1 - \beta_1)}{b_t} \sum_{k=0}^t \beta_1^{t-k} \langle x_k - x^*, g_k \rangle + \frac{2\gamma^2(1 - \beta_1)}{c_m b_t b_0} \sum_{j=0}^t \beta_1^{t-j} (t - j + 1) \|g_j\|^2.$$

After summing over t :

$$\begin{aligned} \|x_T - x^*\|^2 &\leq \|x_0 - x^*\|^2 - \sum_{t=0}^{T-1} \frac{2\gamma(1 - \beta_1)}{b_t} \sum_{k=0}^t \beta_1^{t-k} \langle x_k - x^*, g_k \rangle \\ &\quad + \sum_{t=0}^{T-1} \frac{2\gamma^2(1 - \beta_1)}{c_m b_t b_0} \sum_{j=0}^t \beta_1^{t-j} (t - j + 1) \|g_j\|^2. \end{aligned} \tag{53}$$

Therefore, multiplicative factors for $\langle x_r - x^*, g_r \rangle$ and $\|g_r\|^2$ are equal to

$$-\sum_{t=r}^{T-1} \frac{2\gamma(1 - \beta_1)}{b_t} \beta_1^{t-r} \quad \text{and} \quad \sum_{t=r}^{T-1} \frac{2\gamma^2(1 - \beta_1)}{c_m b_t b_0} \beta_1^{t-r} (t - r + 1),$$

respectively. Let us denote them as $-2\gamma C_r$ and A_r . Using the same idea as in Lemma C.3, we get

$$\frac{(1 - \beta_1)}{b_r} \leq C_r \leq \frac{1}{c_m b_p}$$

and

$$A_r \leq \frac{2\gamma^2}{c_m^2 b_p b_0 (1 - \beta_1)}$$

for all $p = 0, \dots, r$ because of Lemma C.1. Rewriting (53) in terms of C_r, A_r ,

$$\|x_T - x^*\|^2 \leq \|x_0 - x^*\|^2 - \sum_{t=0}^{T-1} 2\gamma C_t \langle x_t - x^*, g_t \rangle + \sum_{t=0}^{T-1} A_t \|g_t\|^2.$$

Consequently,

$$\begin{aligned} \|x_T - x^*\|^2 - \|x_0 - x^*\|^2 &\leq - \sum_{t=0}^{T-1} 2\gamma C_t \langle x_t - x^*, g_t \rangle + \sum_{t=0}^{T-1} A_t \|g_t\|^2 \\ &= - \sum_{t=0}^{T-1} 2\gamma C_t \langle x_t - x^*, \nabla f(x_t) + \theta_t \rangle + \sum_{t=0}^{T-1} A_t \|\nabla f(x_t) + \theta_t\|^2 \\ &\leq - \sum_{t=0}^{T-1} 2\gamma C_t \langle x_t - x^*, \nabla f(x_t) \rangle - \sum_{t=0}^{T-1} 2\gamma C_t \langle x_t - x^*, \theta_t \rangle \\ &\quad + \sum_{t=0}^{T-1} 2A_t \|\nabla f(x_t)\|^2 + \sum_{t=0}^{T-1} 2A_t \|\theta_t\|^2. \end{aligned}$$

Using Assumptions 1.2 and 1.3, one can obtain

$$\begin{aligned} \sum_{t=0}^{T-1} (2\gamma C_t - 4LA_t) (f(x_t) - f_*) &\leq \sum_{t=0}^{T-1} \left(2\gamma C_t \langle x_t - x^*, \nabla f(x_t) \rangle - 2A_t \|\nabla f(x_t)\|^2 \right) \\ &\leq \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 - \sum_{t=0}^{T-1} 2\gamma C_t \langle x_t - x^*, \theta_t \rangle + \sum_{t=0}^{T-1} 2A_t \|\theta_t\|^2. \end{aligned}$$

If we choose $\gamma \leq \frac{(1-\beta_1)^2 c_m^2 b_0}{8L}$, then $2\gamma C_t - 4LA_t \geq \gamma C_t$ because of lower bound on C_t and upper bound for A_t . This finishes the proof. \square

Theorem C.7. *Let Assumptions 1.1, 1.2, and 1.3 hold on $Q = B_{2R}(x^*)$ with $\|x_0 - x^*\| \leq R$. Then, after $K + 1$ iterations of Clip-M-AdaGradD-Norm/Clip-AdamD-Norm with*

$$\gamma \leq \min \left\{ \frac{(1 - \beta_1)^2 c_m^2 b_0}{160L \ln \left(\frac{4(K+1)}{\delta} \right)}, \frac{\sqrt{1 - \beta_1} c_m R b_0}{40 \cdot 9^{\frac{1}{\alpha}} \sigma (K+1)^{\frac{1}{\alpha}} \ln^{\frac{\alpha-1}{\alpha}} \left(\frac{4(K+1)}{\delta} \right)} \right\}, \quad \eta = \frac{\gamma^2 (1 - \beta_1)^2}{R^2}, \quad (54)$$

and

$$\lambda = \frac{\sqrt{1 - \beta_1} c_m b_0 R}{40\gamma \ln \left(\frac{4(K+1)}{\delta} \right)} \quad (55)$$

the bound

$$\sum_{k=0}^K \gamma C_k (f(x_k) - f_*) \leq 2R^2$$

holds with probability at least $1 - \delta$. In particular, when γ equals the minimum from (54), the iterates produced by Clip-M-AdaGradD-Norm/Clip-AdamD-Norm satisfy

$$f(\bar{x}_K) - f(x^*) = \mathcal{O} \left(\max \left\{ \frac{LR^2 \ln \frac{K+1}{\delta}}{(1-\beta_1)^3(K+1)}, \frac{\sigma R \ln^{\frac{\alpha-1}{\alpha}} \frac{K+1}{\delta}}{(1-\beta_1)^{\frac{3}{2}}(K+1)^{\frac{\alpha-1}{\alpha}}} \right\} \right)$$

with probability at least $1 - \delta$, where $\bar{x}_K = \frac{1}{K+1} \sum_{k=0}^K x_k$.

Proof. Our proof is induction-based (similarly to the one for Clip-SGD by Sadiev et al. (2023)). We introduce probability event E_k as follows: inequalities

$$\begin{aligned} - \sum_{l=0}^{t-1} 2\gamma C_l \langle x_l - x^*, \theta_l \rangle + \sum_{l=0}^{t-1} 2A_l \|\theta_l\|^2 &\leq R^2, \\ R_t &\leq \sqrt{2}R \end{aligned}$$

hold simultaneously $\forall t = 0, 1, \dots, k$. We want to show that $\mathbb{P}\{E_k\} \geq 1 - \frac{k\delta}{K+1} \forall k = 0, 1, \dots, K+1$. The case when $k = 0$ is obvious. Now let us make an induction step: let the statement hold for some $k = T-1 \leq K$: $\mathbb{P}\{E_{T-1}\} \geq 1 - \frac{(T-1)\delta}{K+1}$. It remains to prove that $\mathbb{P}\{E_T\} \geq 1 - \frac{T\delta}{K+1}$. The event E_{T-1} implies $x_t \in B_{\sqrt{2}R}(x^*) \forall t = 0, \dots, T-1$. Hence, E_{T-1} also implies

$$\|x_T - x^*\| \leq \|x_{T-1} - x^*\| + \frac{\gamma}{b_{T-1}} \|m_{T-1}\| \leq \sqrt{2}R + \frac{\gamma\lambda}{b_{T-1}} \leq \sqrt{2}R + \frac{\gamma\lambda}{c_m b_0} \leq 2R.$$

Therefore, E_{T-1} implies $\{x_t\}_{t=0}^T \subseteq B_{2R}(x^*)$ and we can apply Lemma C.6:

$$\sum_{l=0}^{t-1} \gamma C_l (f(x_l) - f_*) \leq R_0^2 - R_t^2 - \sum_{l=0}^{t-1} 2\gamma C_l \langle x_l - x^*, \theta_l \rangle + \sum_{l=0}^{t-1} 2A_l \|\theta_l\|^2$$

$\forall t = 1, \dots, T$ and $\forall t = 1, \dots, T-1$ it implies that

$$\sum_{l=0}^{t-1} \gamma C_l (f(x_l) - f_*) \leq R_0^2 - \sum_{l=0}^{t-1} 2\gamma C_l \langle x_l - x^*, \theta_l \rangle + \sum_{l=0}^{t-1} 2A_l \|\theta_l\|^2 \leq 2R^2.$$

Taking into account that $\sum_{l=0}^{t-1} \gamma C_l (f(x_l) - f_*) \geq 0$, we get that E_{T-1} implies

$$R_T^2 \leq R_0^2 - \sum_{t=0}^{T-1} 2\gamma C_t \langle x_t - x^*, \theta_t \rangle + \sum_{t=0}^{T-1} 2A_t \|\theta_t\|^2. \quad (56)$$

Next, for vectors

$$\eta_t = \begin{cases} x_t - x^*, & \|x_t - x^*\| \leq \sqrt{2}R \\ 0, & \text{otherwise} \end{cases}$$

for all $t = 0, 1, \dots, T-1$, we have that with probability 1

$$\|\eta_t\| \leq \sqrt{2}R. \quad (57)$$

Then, E_{T-1} implies that $\eta_t = x_t - x^*$ for all $t = 0, \dots, T-1$. What is more, for all $t = 0, \dots, T-1$ E_{T-1} implies

$$\|\nabla f(x_t)\| \leq L \|x_t - x^*\| \stackrel{(55)}{\leq} \sqrt{2}LR \leq \frac{\lambda}{2}.$$

Hence, using the notation from Appendix A, we have that E_{T-1} implies

$$\begin{aligned}
 R_T^2 &\leq R_0^2 - \underbrace{\sum_{t=0}^{T-1} 2\gamma C_t \langle x_t - x^*, \theta_t^u \rangle}_{\textcircled{1}} - \underbrace{\sum_{t=0}^{T-1} 2\gamma C_t \langle x_t - x^*, \theta_t^b \rangle}_{\textcircled{2}} + \underbrace{\sum_{t=0}^{T-1} 4A_t (\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2)}_{\textcircled{3}} \\
 &\quad + \underbrace{\sum_{t=0}^{T-1} 4A_t \mathbb{E}_{\xi_t} \|\theta_t^u\|^2}_{\textcircled{4}} + \underbrace{\sum_{t=0}^{T-1} 4A_t \|\theta_t^b\|^2}_{\textcircled{5}}.
 \end{aligned} \tag{58}$$

Next, we bound each term separately with high probability. Before we move on, we also note that event E_{T-1} implies $\|\nabla f(x_t)\| \leq \frac{\lambda}{2}$. Therefore, one can apply Lemma A.4 and get

$$\|\theta_t^u\| \leq 2\lambda, \tag{59}$$

$$\|\theta_t^b\| \leq \frac{2^\alpha \sigma^\alpha}{\lambda^{\alpha-1}}, \tag{60}$$

$$\mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \leq 18\lambda^{2-\alpha} \sigma^\alpha. \tag{61}$$

Bound for ①. The definition of θ_t^u implies

$$\mathbb{E}_{\xi_t} [-2\gamma C_t \langle \eta_t, \theta_t^u \rangle] = 0.$$

Moreover, applying the bound on C_t : $C_t \leq \frac{1}{c_m b_0}$ from Lemma C.6,

$$|-2\gamma C_t \langle \eta_t, \theta_t^u \rangle| \leq 2\gamma C_t \|\eta_t\| \|\theta_t^u\| \stackrel{(57), (59)}{\leq} \frac{6\gamma\lambda R}{c_m b_0} \stackrel{(55)}{\leq} \frac{3R^2}{20 \ln\left(\frac{4(K+1)}{\delta}\right)} = c.$$

For $\sigma_t^2 = \mathbb{E}_{\xi_t} [4\gamma^2 C_t^2 \langle \eta_t, \theta_t^u \rangle^2]$ we also derive

$$\sigma_t^2 \leq 4\gamma^2 C_t^2 \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \|\eta_t\|^2 \leq \frac{8\gamma^2 R^2}{c_m^2 b_0^2} \mathbb{E}_{\xi_t} \|\theta_t^u\|^2. \tag{62}$$

Hence, we can apply Bernstein's inequality (Lemma A.5) with c defined above and $G = \frac{R^4}{100 \ln\left(\frac{4(K+1)}{\delta}\right)}$:

$$\begin{aligned}
 \mathbb{P} \left\{ -\sum_{t=0}^{T-1} \frac{2\gamma}{b_t} \langle x_t - x^*, \theta_t^u \rangle > \frac{R^2}{5} \text{ and } \sum_{t=0}^{T-1} \sigma_t^2 \leq G \right\} &\leq 2 \exp \left(-\frac{R^4}{25 \left(2G + \frac{2cR^2}{15} \right)} \right) \\
 &= \frac{\delta}{2(K+1)}.
 \end{aligned}$$

Therefore,

$$\mathbb{P} \left\{ \text{either } -\sum_{t=0}^{T-1} \frac{2\gamma}{b_t} \langle x_t - x^*, \theta_t^u \rangle \leq \frac{R^2}{5} \text{ or } \sum_{t=0}^{T-1} \sigma_t^2 > G \right\} \geq 1 - \frac{\delta}{2(K+1)}.$$

In addition, event E_{T-1} implies that (due to (62) and (61))

$$\begin{aligned}
 \sum_{t=0}^{T-1} \sigma_t^2 &\leq \frac{144\gamma^2 \lambda^{2-\alpha} \sigma^\alpha R^2 T}{c_m^2 b_0^2} \stackrel{(55)}{\leq} \frac{144(1-\beta_1)^{1-\frac{\alpha}{2}} \gamma^\alpha b_0^{2-\alpha} \sigma^\alpha R^{4-\alpha} T}{40^{2-\alpha} c_m^\alpha b_0^2 \ln^{2-\alpha} \left(\frac{4(K+1)}{\delta} \right)} \\
 &\stackrel{(54)}{\leq} \frac{144(1-\beta_1) R^4 T}{9 \cdot 40^2 (K+1) \ln \left(\frac{4(K+1)}{\delta} \right)} \leq \frac{R^4}{100 \ln \left(\frac{4(K+1)}{\delta} \right)}.
 \end{aligned}$$

Bound for ②. For the second term, one can obtain from (54), (55) and $\alpha \leq 2$ that E_{T-1} implies

$$\begin{aligned} -\sum_{t=0}^{T-1} 2\gamma C_t \langle x_t - x^*, \theta_t^b \rangle &\leq \sum_{t=0}^{T-1} \frac{2\gamma}{c_m b_0} \|\eta_t\| \|\theta_t^b\| \stackrel{(57),(60)}{\leq} \frac{2\sqrt{2} \cdot 2^\alpha \sigma^\alpha \gamma T R}{c_m b_0 \lambda^{\alpha-1}} \\ &\stackrel{(55)}{=} \frac{4 \cdot 2^\alpha 40^\alpha \sigma^\alpha \gamma^\alpha T R^{2-\alpha}}{40(1-\beta_1)^{\frac{\alpha}{2}-1} c_m^\alpha b_0^\alpha \ln^{1-\alpha} \left(\frac{4(K+1)}{\delta}\right)} \stackrel{(54)}{\leq} \frac{4 \cdot 2^\alpha (1-\beta_1) T R^2}{360 \cdot (K+1)} \\ &\leq \frac{2R^2}{45} \leq \frac{R^2}{5}. \end{aligned}$$

Bound for ③. For the third part, we have

$$\mathbb{E}_{\xi_t} \left[4A_t \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right) \right] = 0.$$

What is more,

$$\begin{aligned} \left| 4A_t \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right) \right| &\leq 4A_t \left(\|\theta_t^u\|^2 + \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right) \stackrel{(59)}{\leq} \frac{64\gamma^2 \lambda^2}{c_m^2 b_0^2 (1-\beta_1)} \stackrel{(55)}{=} \frac{R^2}{25 \ln^2 \left(\frac{4(K+1)}{\delta}\right)} \\ &\leq \frac{3R^2}{20 \ln \left(\frac{4(K+1)}{\delta}\right)} = c. \end{aligned} \tag{63}$$

We also define

$$\hat{\sigma}_t^2 = \mathbb{E}_{\xi_t} \left[16A_t^2 \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right)^2 \right].$$

Hence,

$$\begin{aligned} \hat{\sigma}_t^2 &\stackrel{(63)}{\leq} \frac{3R^2}{20 \ln \left(\frac{4(K+1)}{\delta}\right)} \mathbb{E}_{\xi_t} \left[\left| 4A_t \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right) \right| \right] \\ &\leq \frac{12\gamma^2 R^2}{5c_m^2 b_0^2 (1-\beta_1) \ln \left(\frac{4(K+1)}{\delta}\right)} \mathbb{E}_{\xi_t} \|\theta_t^u\|^2. \end{aligned}$$

Therefore, we can apply Bernstein's inequality (Lemma A.5) with c defined above and $G = \frac{R^4}{100 \ln \left(\frac{4(K+1)}{\delta}\right)}$:

$$\begin{aligned} \mathbb{P} \left\{ \sum_{t=0}^{T-1} 4A_t \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right) > \frac{R^2}{5} \text{ and } \sum_{t=0}^{T-1} \hat{\sigma}_t^2 \leq G \right\} &\leq 2 \exp \left(-\frac{R^4}{25(2G + \frac{2cR^2}{15})} \right) \\ &= \frac{\delta}{2(K+1)}. \end{aligned}$$

Consequently,

$$\mathbb{P} \left\{ \text{either } \sum_{t=0}^{T-1} 4A_t \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right) \leq \frac{R^2}{5} \text{ or } \sum_{t=0}^{T-1} \hat{\sigma}_t^2 > G \right\} \geq 1 - \frac{\delta}{2(K+1)}.$$

Moreover, event E_{T-1} implies that

$$\begin{aligned} \sum_{t=0}^{T-1} \hat{\sigma}_t^2 &\leq \sum_{t=0}^{T-1} \frac{12\gamma^2 R^2}{5c_m^2 b_0^2 (1-\beta_1) \ln \left(\frac{4(K+1)}{\delta}\right)} \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \stackrel{(61)}{\leq} \frac{18 \cdot 12\gamma^2 \lambda^{2-\alpha} \sigma^\alpha R^2 T}{5c_m^2 b_0^2 (1-\beta_1) \ln \left(\frac{4(K+1)}{\delta}\right)} \\ &\stackrel{(55)}{=} \frac{18 \cdot 12 \cdot 40^\alpha \gamma^\alpha \sigma^\alpha R^{4-\alpha} T}{5 \cdot 40^2 c_m^\alpha (1-\beta_1)^{\frac{\alpha}{2}} b_0^\alpha \ln^{3-\alpha} \left(\frac{4(K+1)}{\delta}\right)} \stackrel{(54)}{\leq} \frac{18 \cdot 12 R^4 T}{9 \cdot 5 \cdot 40^2 (K+1) \ln^2 \left(\frac{4(K+1)}{\delta}\right)} \\ &\leq \frac{R^4}{100 \ln \left(\frac{4(K+1)}{\delta}\right)}. \end{aligned}$$

Bound for ④. For the fourth part, we get that E_{T-1} implies

$$\begin{aligned} \sum_{t=0}^{T-1} 4A_t E_{\xi_t} \|\theta_t^u\|^2 &\leq \sum_{t=0}^{T-1} \frac{8\gamma^2}{c_m^2 b_0^2(1-\beta_1)} E_{\xi_t} \|\theta_t^u\|^2 \stackrel{(61)}{\leq} \frac{144\gamma^2 \lambda^{2-\alpha} \sigma^\alpha T}{c_m^2 b_0^2(1-\beta_1)} \\ &\stackrel{(54)}{=} \frac{144\gamma^\alpha 40^\alpha R^{2-\alpha} \sigma^\alpha T}{40^2 c_m^\alpha b_0^\alpha (1-\beta_1)^{\frac{\alpha}{2}} \ln^{2-\alpha} \left(\frac{4(K+1)}{\delta}\right)} \stackrel{(54)}{\leq} \frac{144R^2 T}{9 \cdot 40^2 (K+1) \ln \left(\frac{4(K+1)}{\delta}\right)} \\ &\leq \frac{R^2}{100} \leq \frac{R^2}{5}. \end{aligned}$$

Bound for ⑤. For the last term, E_{T-1} implies

$$\begin{aligned} \sum_{t=0}^{T-1} 4A_t \|\theta_t^b\|^2 &\leq \sum_{t=0}^{T-1} \frac{8\gamma^2}{c_m^2 b_0^2(1-\beta_1)} \|\theta_t^b\|^2 \stackrel{(60)}{\leq} \frac{8 \cdot 4^\alpha \sigma^{2\alpha} \gamma^2 T}{c_m^2 b_0^2(1-\beta_1) \lambda^{2(\alpha-1)}} \\ &\stackrel{(55)}{=} \frac{8 \cdot 4^\alpha 40^{2\alpha} \sigma^{2\alpha} \gamma^{2\alpha} T \ln^{2(\alpha-1)} \left(\frac{4(K+1)}{\delta}\right)}{40^2 c_m^{2\alpha} b_0^{2\alpha} (1-\beta_1)^\alpha R^{2(\alpha-1)}} \\ &\stackrel{(54)}{\leq} \frac{8 \cdot 4^\alpha R^2 T}{360^2 (K+1)^2} \leq \frac{8R^2}{45^2} \leq \frac{R^2}{5}. \end{aligned}$$

Thus, taking into account the bounds above, the probability event $E_{T-1} \cap E_1 \cap E_2$ implies that

$$R_T^2 \leq R^2 + 5 \frac{R^2}{5} = 2R^2,$$

where

$$\begin{aligned} E_1 &= \left\{ \text{either } - \sum_{t=0}^{T-1} \frac{2\gamma}{b_t} \langle x_t - x^*, \theta_t^u \rangle \leq \frac{R^2}{5} \text{ or } \sum_{t=0}^{T-1} \sigma_t^2 > \frac{R^4}{100 \ln \left(\frac{4(K+1)}{\delta}\right)} \right\}, \\ E_2 &= \left\{ \text{either } \sum_{t=0}^{T-1} \frac{4\gamma^2}{b_t^2} \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right) \leq \frac{R^2}{5} \text{ or } \sum_{t=0}^{T-1} \hat{\sigma}_t^2 > \frac{R^4}{100 \ln \left(\frac{4(K+1)}{\delta}\right)} \right\}. \end{aligned}$$

Therefore,

$$\mathbb{P}\{E_T\} \geq \mathbb{P}\{E_{T-1} \cap E_1 \cap E_2\} = 1 - \mathbb{P}\{\overline{E}_{T-1} \cup \overline{E}_1 \cup \overline{E}_2\} \geq 1 - \mathbb{P}\{\overline{E}_{T-1}\} - \mathbb{P}\{\overline{E}_1\} - \mathbb{P}\{\overline{E}_2\} \geq 1 - \frac{T\delta}{K+1}.$$

Hence, for all $k = 0, \dots, K+1$ we get $\mathbb{P}\{E_k\} \geq 1 - \frac{k\delta}{K+1}$. As the result, event E_{K+1} implies that

$$\sum_{k=0}^K \gamma C_k (f(x_k) - f_*) \leq 2R^2 \tag{64}$$

with probability at least $1 - \delta$. Next, from (64) we get that with probability at least $1 - \delta$

$$\sum_{k=0}^K (f(x_k) - f_*) \leq \frac{2R^2}{\gamma} \max_{k \in [0, K]} \frac{1}{C_k}.$$

Moreover, $\frac{1}{C_k}$ can be bounded in the following way (from Lemma C.6):

$$\frac{1}{C_k} \leq \frac{b_k}{(1-\beta_1)}.$$

Hence, we get

$$\sum_{k=0}^K (f(x_k) - f_*) \leq \frac{2R^2}{\gamma(1-\beta_1)} \max_{k \in [0, K]} b_k. \quad (65)$$

Also we can bound b_k for **Clip-M-AdaGradD-Norm** using that $g_k = \nabla f(x_k) + \theta_k$ and Assumption 1.2:

$$b_k^2 \leq b_0^2 + \eta \sum_{k=0}^K \left(4L(f(x_k) - f_*) + 2\|\theta_k\|^2 \right)$$

and for **Clip-AdamD-Norm**, respectively

$$b_k^2 \leq b_0^2 + \frac{\eta}{K+1} \sum_{k=0}^K \left(4L(f(x_k) - f_*) + 2\|\theta_k\|^2 \right).$$

Therefore, due to the fact that the event E_{K+1} implies (see the bounds for ③, ④ and ⑤)

$$\sum_{k=0}^K \frac{4\gamma^2}{c_m^2 b_0^2 (1-\beta_1)} \|\theta_k\|^2 \leq \frac{3R^2}{5},$$

we get

$$b_k^2 \leq b_0^2 + \eta \sum_{k=0}^K 4L((f(x_k) - f_*)) + \frac{3\eta(1-\beta_1)b_0^2 R^2}{10\gamma^2}$$

for **Clip-M-AdaGradD-Norm** scheme and

$$b_k^2 \leq b_0^2 + \frac{\eta}{K+1} \sum_{k=0}^K 4L((f(x_k) - f_*)) + \frac{3\eta(1-\beta_1)b_0^2 R^2}{40\gamma^2(K+1)}$$

for **Clip-AdamD-Norm**, where we substitute the constant c_m from Lemma C.1. Consequently, substituting bounds above in (65), we get

$$\left(\sum_{k=0}^K (f(x_k) - f_*) \right)^2 \leq \frac{4R^4}{\gamma^2(1-\beta_1)^2} \left(b_0^2 + \eta \sum_{k=0}^K (4L(f(x_k) - f_*)) + \frac{3\eta(1-\beta_1)R^2b_0^2}{10\gamma^2} \right)$$

for **Clip-M-AdaGradD-Norm** and

$$\left(\sum_{k=0}^K (f(x_k) - f_*) \right)^2 \leq \frac{4R^4}{\gamma^2(1-\beta_1)^2} \left(b_0^2 + \frac{\eta}{K+1} \sum_{k=0}^K (4L(f(x_k) - f_*)) + \frac{3\eta(1-\beta_1)R^2b_0^2}{40\gamma^2(K+1)} \right)$$

for **Clip-AdamD-Norm**, respectively. Solving these quadratic inequalities, we have that E_{K+1} implies

$$\begin{aligned} \sum_{k=0}^K (f(x_k) - f_*) &\leq \frac{2R^2}{\gamma^2} \left(\frac{4L\eta R^2}{(1-\beta_1)^2} + \sqrt{\frac{16L^2\eta^2 R^4}{(1-\beta_1)^4} + b_0^2 \left(\frac{\gamma^2}{(1-\beta_1)^2} + \frac{3\eta R^2}{10(1-\beta_1)} \right)} \right) \\ &\leq \frac{6R^2}{\gamma^2} \max \left\{ \frac{8L\eta R^2}{(1-\beta_1)^2}, \frac{b_0\gamma}{1-\beta_1}, b_0 R \sqrt{\frac{\eta}{1-\beta_1}} \right\} \end{aligned}$$

and

$$\begin{aligned} \sum_{k=0}^K (f(x_k) - f_*) &\leq \frac{2R^2}{\gamma^2} \left(\frac{4L\eta R^2}{(1-\beta_1)^2(K+1)} \right. \\ &\quad \left. + \sqrt{\frac{16L^2\eta^2 R^4}{(1-\beta_1)^4(K+1)^2} + b_0^2 \left(\frac{\gamma^2}{(1-\beta_1)^2} + \frac{3\eta R^2}{40(1-\beta_1)(K+1)} \right)} \right) \\ &\leq \frac{6R^2}{\gamma^2} \max \left\{ \frac{8L\eta R^2}{(1-\beta_1)^2(K+1)}, \frac{b_0\gamma}{1-\beta_1}, b_0 R \sqrt{\frac{\eta}{(1-\beta_1)(K+1)}} \right\}. \end{aligned}$$

with probability at least $1 - \delta$. Choosing $\eta = \frac{\gamma^2(1-\beta_1)^2}{R^2}$, γ equal to the minimum from (54) and using that $2\sqrt{ab} \leq a + b$, we obtain the bound for Clip-M-AdaGradD/Clip-AdamD-Norm for the convex case:

$$\frac{1}{K+1} \sum_{k=0}^K (f(x_k) - f_*) = \mathcal{O} \left(\max \left\{ \frac{LR^2 \ln \frac{K+1}{\delta}}{(1-\beta_1)^3(K+1)}, \frac{\sigma R \ln \frac{\alpha-1}{\alpha} \frac{K+1}{\delta}}{(1-\beta_1)^{\frac{3}{2}}(K+1)^{\frac{\alpha-1}{\alpha}}} \right\} \right)$$

with probability at least $1 - \delta$. To get the final result, it remains to apply Jensen's inequality. \square

C.4. Non-Convex Case: Methods without Delay

Lemma C.8 (Descent lemma). *Let Assumptions 1.2 and 1.4 hold. Then, after T iterations of Clip-M-AdaGrad/Clip-Adam, we have*

$$\sum_{t=0}^{T-1} \frac{\gamma C_t}{2} \|\nabla f(x_t)\|^2 \leq \left(2M + \frac{2L\gamma^2}{\eta(1-\beta_1)} \right) \sqrt{b_{-1}^2 + \eta \sum_{t=0}^{T-1} \|g_t\|^2} - \sum_{t=0}^{T-1} \gamma C_t \langle \nabla f(x_t), \theta_t^u \rangle + \sum_{t=0}^{T-1} \frac{\gamma C_t}{2} \|\theta_t^b\|^2$$

for Clip-M-AdaGrad-Norm, where $C_t = \sum_{k=t}^{T-1} (1-\beta_1) \beta_1^{k-t}$, and

$$\sum_{t=0}^{T-1} \frac{\gamma C_t}{2} \|\nabla f(x_t)\|^2 \leq \left(3M + \frac{16KL\gamma^2}{\eta(1-\beta_1)} \right) \sqrt{b_{-1}^2 + \frac{\eta}{K} \sum_{t=0}^{T-1} \|g_t\|^2} - \sum_{t=0}^{T-1} \gamma C_t \langle \nabla f(x_t), \theta_t^u \rangle + \sum_{t=0}^{T-1} \frac{\gamma C_t}{2} \|\theta_t^b\|^2$$

for Clip-Adam-Norm, where $C_t = \sum_{k=t}^{T-1} (1-\beta_1) \beta_1^{k-t} / (\sqrt{\beta_2})^k$.

Proof. The first part of the proof is similar to the Lemma C.3. We start with the L -smoothness of f :

$$f(x_{t+1}) - f(x_t) \leq \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 = -\frac{\gamma}{b_t} \langle \nabla f(x_t), m_t \rangle + \frac{L\gamma^2}{2b_t^2} \|m_t\|^2. \quad (66)$$

Using the update rule of Algorithm 3, we can obtain

$$\begin{aligned} -\langle \nabla f(x_t), m_t \rangle &= -\beta_1 \langle \nabla f(x_t), m_{t-1} \rangle - (1-\beta_1) \langle \nabla f(x_t), g_t \rangle \\ &= -\beta_1 \langle \nabla f(x_t) - \nabla f(x_{t-1}), m_{t-1} \rangle - \beta_1 \langle \nabla f(x_{t-1}), m_{t-1} \rangle \\ &\quad - (1-\beta_1) \langle \nabla f(x_t), g_t \rangle \\ &\leq -\beta_1 \langle \nabla f(x_{t-1}), m_{t-1} \rangle + \beta_1 \|\nabla f(x_t) - \nabla f(x_{t-1})\| \|m_{t-1}\| \\ &\quad - (1-\beta_1) \langle \nabla f(x_t), g_t \rangle \\ &\leq -\beta_1 \langle \nabla f(x_{t-1}), m_{t-1} \rangle + \beta_1 L \|x_t - x_{t-1}\| \|m_{t-1}\| \\ &\quad - (1-\beta_1) \langle \nabla f(x_t), g_t \rangle \\ &= -\beta_1 \langle \nabla f(x_{t-1}), m_{t-1} \rangle + \frac{\gamma \beta_1 L}{b_{t-1}} \|m_{t-1}\|^2 \\ &\quad - (1-\beta_1) \langle \nabla f(x_t), g_t \rangle, \end{aligned}$$

where we use the Cauchy-Schwarz inequality and L -smoothness of f . Applying the same idea for the $t-1, t-2, \dots, 0$ and noting that $m_{-1} = 0$, we get

$$-\langle \nabla f(x_t), m_t \rangle \leq -(1-\beta_1) \sum_{k=0}^t \beta_1^{t-k} \langle \nabla f(x_k), g_k \rangle + L\gamma \sum_{k=0}^{t-1} \frac{\beta_1^{t-k}}{b_k} \|m_k\|^2. \quad (67)$$

Therefore, substituting (67) into (66), we have

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq -\frac{(1-\beta_1)\gamma}{b_t} \sum_{k=0}^t \beta_1^{t-k} \langle \nabla f(x_k), g_k \rangle + \frac{L\gamma^2}{b_t} \sum_{k=0}^{t-1} \frac{\beta_1^{t-k}}{b_k} \|m_k\|^2 + \frac{L\gamma^2}{2b_t^2} \|m_t\|^2 \\ &\leq -\frac{(1-\beta_1)\gamma}{b_t} \sum_{k=0}^t \beta_1^{t-k} \langle \nabla f(x_k), g_k \rangle + \frac{L\gamma^2}{b_t} \sum_{k=0}^t \frac{\beta_1^{t-k}}{b_k} \|m_k\|^2. \end{aligned}$$

Applying Lemma C.2 with $1 - \beta_1^{k+1} \leq 1$, we can rewrite the inequality above as follows:

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq -\frac{(1 - \beta_1)\gamma}{b_t} \sum_{k=0}^t \beta_1^{t-k} \langle \nabla f(x_k), g_k \rangle + \frac{L\gamma^2}{b_t} \sum_{k=0}^t \frac{\beta_1^{t-k}}{b_k} \sum_{j=0}^k \beta_1^{k-j} (1 - \beta_1) \|g_j\|^2 \\ &= -\frac{(1 - \beta_1)\gamma}{b_t} \sum_{k=0}^t \beta_1^{t-k} \langle \nabla f(x_k), g_k \rangle + \frac{L\gamma^2}{b_t} \sum_{j=0}^t \sum_{k=j}^t \frac{\beta_1^{t-k}}{b_k} \beta_1^{k-j} (1 - \beta_1) \|g_j\|^2, \end{aligned} \quad (68)$$

where we change the limits of summation. Multiplying both sides of the inequality above by $\frac{b_t}{p_t}$, where

$$p_t = \begin{cases} 1, & \text{for Clip-M-AdaGrad-Norm} \\ (\sqrt{\beta_2})^t, & \text{for Clip-Adam-Norm} \end{cases} \quad (69)$$

and using that $b_k \geq c_m b_j$ (see Lemma C.1), one can obtain

$$\frac{b_t}{p_t} (f(x_{t+1}) - f(x_t)) \leq -\frac{(1 - \beta_1)\gamma}{p_t} \sum_{k=0}^t \beta_1^{t-k} \langle \nabla f(x_k), g_k \rangle + \frac{L\gamma^2}{p_t} \sum_{j=0}^t \frac{\beta_1^{t-j}}{c_m b_j} (1 - \beta_1)(t - j + 1) \|g_j\|^2.$$

After summing over t ,

$$\sum_{t=0}^{T-1} \frac{b_t}{p_t} (f(x_{t+1}) - f(x_t)) \leq -(1 - \beta_1)\gamma \sum_{t=0}^{T-1} \sum_{k=0}^t \frac{\beta_1^{t-k}}{p_t} \langle \nabla f(x_k), g_k \rangle + L\gamma^2 \sum_{t=0}^{T-1} \sum_{j=0}^t \frac{\beta_1^{t-j}}{c_m b_j p_t} (1 - \beta_1)(t - j + 1) \|g_j\|^2.$$

Next, applying the same idea as in Lemma C.3, we get that multiplicative factors are equal to

$$-\gamma C_r = -\sum_{t=r}^{T-1} \frac{\gamma(1 - \beta_1)\beta_1^{t-r}}{p_t} \quad (70)$$

for the scalar product $\langle \nabla f(x_r), g_r \rangle$ and

$$A_r = \sum_{t=r}^{T-1} \frac{L\gamma^2(1 - \beta_1)}{c_m b_r p_t} (t - r + 1) \beta_1^{t-r} \quad (71)$$

for the squared norm $\|g_r\|^2$, respectively. Moreover, it can be shown that $p_t \geq c_m$ for corresponding update rule of b_t . Hence, for (71) we apply Lemma A.2 to obtain the next bound:

$$A_r \leq \frac{L\gamma^2}{c_m^2 b_r (1 - \beta_1)}.$$

Therefore, rewriting the descent lemma in terms of (70) and (71), we have

$$\sum_{t=0}^{T-1} \frac{b_t}{p_t} (f(x_{t+1}) - f(x_t)) \leq -\sum_{t=0}^{T-1} \gamma C_t \langle \nabla f(x_t), g_t \rangle + \frac{L\gamma^2}{c_m^2 (1 - \beta_1)} \sum_{t=0}^{T-1} \frac{\|g_t\|^2}{b_t}.$$

Using that $g_t = \nabla f(x_t) + \theta_t$, we get

$$\begin{aligned}
 \sum_{t=0}^{T-1} \gamma C_t \|\nabla f(x_t)\|^2 &\leq \sum_{t=0}^{T-1} \frac{b_t}{p_t} (f(x_t) - f(x_{t+1})) - \sum_{t=0}^{T-1} \gamma C_t \langle \nabla f(x_t), \theta_t \rangle + \frac{L\gamma^2}{c_m^2(1-\beta_1)} \sum_{t=0}^{T-1} \frac{\|g_t\|^2}{b_t} \\
 &= \sum_{t=0}^{T-1} \frac{b_t}{p_t} (f(x_t) - f_* - (f(x_{t+1}) - f_*)) - \sum_{t=0}^{T-1} \gamma C_t \langle \nabla f(x_t), \theta_t \rangle \\
 &\quad + \frac{L\gamma^2}{c_m^2(1-\beta_1)} \sum_{t=0}^{T-1} \frac{\|g_t\|^2}{b_t} \\
 &\leq \frac{b_0}{p_0} (f(x_0) - f_*) + \sum_{t=1}^{T-1} \left(\frac{b_t}{p_t} - \frac{b_{t-1}}{p_{t-1}} \right) (f(x_t) - f_*) - \sum_{t=0}^{T-1} \gamma C_t \langle \nabla f(x_t), \theta_t \rangle \\
 &\quad + \frac{L\gamma^2}{c_m^2(1-\beta_1)} \sum_{t=0}^{T-1} \frac{\|g_t\|^2}{b_t}.
 \end{aligned} \tag{72}$$

Since $p_t = 1$ for **Clip-M-AdaGrad-Norm**, we can use that $b_t \geq b_{t-1}$, and for **Clip-Adam-Norm** we get $b_t \geq \sqrt{\beta_2} b_{t-1}$, what is equal to $\frac{b_t}{p_t} \geq \frac{b_{t-1}}{p_{t-1}}$ with $p_t = (\sqrt{\beta_2})^t$. Therefore, applying Assumption 1.4, we obtain

$$\sum_{t=0}^{T-1} \gamma C_t \|\nabla f(x_t)\|^2 \leq \frac{b_0 M}{p_0} + \frac{b_{T-1} M}{p_{T-1}} - \sum_{t=0}^{T-1} \gamma C_t \langle \nabla f(x_t), \theta_t \rangle + \frac{L\gamma^2}{c_m^2(1-\beta_1)} \sum_{t=0}^{T-1} \frac{\|g_t\|^2}{b_t}.$$

Now we construct descent lemmas for each considering update separately. For **Clip-M-AdaGrad-Norm** we directly apply Lemma A.3 to bound the last term:

$$\begin{aligned}
 \sum_{t=0}^{T-1} \gamma C_t \|\nabla f(x_t)\|^2 &\leq 2M b_{T-1} - \sum_{t=0}^{T-1} \gamma C_t \langle \nabla f(x_t), \theta_t \rangle + \frac{L\gamma^2}{\eta(1-\beta_1)} b_{T-1} \\
 &= \left(2M + \frac{2L\gamma^2}{\eta(1-\beta_1)} \right) b_{T-1} - \sum_{t=0}^{T-1} \gamma C_t \langle \nabla f(x_t), \theta_t \rangle \\
 &\leq \left(2M + \frac{2L\gamma^2}{\eta(1-\beta_1)} \right) b_{T-1} - \sum_{t=0}^{T-1} \gamma C_t \langle \nabla f(x_t), \theta_t^u \rangle \\
 &\quad + \sum_{t=0}^{T-1} \frac{\gamma C_t}{2} \|\nabla f(x_t)\|^2 + \sum_{t=0}^{T-1} \frac{\gamma C_t}{2} \|\theta_t^b\|^2,
 \end{aligned} \tag{73}$$

where we use that $c_m = 1$ and $p_t = 1$ for **Clip-M-AdaGrad-Norm**. For the **Clip-Adam-Norm**, we get

$$\begin{aligned}
 \sum_{t=0}^{T-1} \frac{\|g_t\|^2}{b_t} &= \frac{1}{\eta} \sum_{t=0}^{T-1} \frac{\eta \|g_t\|^2}{\sqrt{\beta_2^{t+1} b_{-1}^2 + (1-\beta_2)\eta \sum_{k=0}^t \beta_2^{t-k} \|g_k\|^2}} \\
 &\leq \frac{K}{\eta} \sum_{t=0}^{T-1} \frac{2 \frac{\eta}{K} \|g_t\|^2}{\sqrt{b_{-1}^2 + \frac{\eta}{K} \sum_{k=0}^t \|g_k\|^2}} \\
 &\leq \frac{4K}{\eta} \sqrt{b_{-1}^2 + \frac{\eta}{K} \sum_{t=0}^{T-1} \|g_t\|^2},
 \end{aligned}$$

where we use that $\beta_2^k \geq 1/4$ for all $k = 0, \dots, K$. Consequently, with upper bound on b_t and $c_m = 1/2$, for **Clip-Adam-Norm**

one can obtain

$$\begin{aligned}
 \sum_{t=0}^{T-1} \gamma C_t \|\nabla f(x_t)\|^2 &\leq b_0 M + \frac{b_{T-1} M}{(\sqrt{\beta_2})^{T-1}} - \sum_{t=0}^{T-1} \gamma C_t \langle \nabla f(x_t), \theta_t \rangle + \frac{16KL\gamma^2}{\eta(1-\beta_1)} \sqrt{b_{-1}^2 + \frac{\eta}{K} \sum_{k=0}^t \|g_k\|^2} \\
 &\leq \left(3M + \frac{16KL\gamma^2}{\eta(1-\beta_1)}\right) \sqrt{b_{-1}^2 + \frac{\eta}{K} \sum_{t=0}^{T-1} \|g_t\|^2} - \sum_{t=0}^{T-1} \gamma C_t \langle \nabla f(x_t), \theta_t \rangle \\
 &\leq \left(3M + \frac{16KL\gamma^2}{\eta(1-\beta_1)}\right) \sqrt{b_{-1}^2 + \frac{\eta}{K} \sum_{t=0}^{T-1} \|g_t\|^2} - \sum_{t=0}^{T-1} \gamma C_t \langle \nabla f(x_t), \theta_t^u \rangle \\
 &\quad + \sum_{t=0}^{T-1} \frac{\gamma C_t}{2} \|\nabla f(x_t)\|^2 + \sum_{t=0}^{T-1} \frac{\gamma C_t}{2} \|\theta_t^b\|^2.
 \end{aligned}$$

After substitution of the analytical form of b_{T-1} in (73) and different options of p_t , we claim the final result. \square

Theorem C.9. Let Assumptions 1.1, 1.2 and 1.4 hold. Then, after K iterations of Clip-M-AdaGrad-Norm/Clip-Adam-Norm with

$$\gamma \leq \min \left\{ \frac{b_{-1} K^{\frac{1-\alpha}{3\alpha-2}}}{48L \ln(\frac{4}{\delta})}, \frac{b_{-1} \sqrt{M}}{4^{\frac{1}{\alpha}} \cdot 12\sqrt{L} \sigma(K+1)^{\frac{\alpha}{3\alpha-2}} \ln^{\frac{\alpha-1}{\alpha}}(\frac{4}{\delta})}, \right. \\
 \left. \frac{b_{-1} M^{\frac{\alpha}{2\alpha-1}}}{4^{\frac{\alpha}{2\alpha-1}} \cdot 12^{\frac{2\alpha-2}{2\alpha-1}} \sigma^{\frac{2\alpha}{2\alpha-1}} L^{\frac{\alpha-1}{2\alpha-1}} (K+1)^{\frac{\alpha}{3\alpha-2}} \ln^{\frac{2\alpha-2}{2\alpha-1}}(\frac{4}{\delta})} \right\}, \quad \eta = \frac{L\gamma^2}{M(1-\beta_1)}, \quad (74)$$

and

$$\lambda = \frac{b_{-1} \sqrt{M} (K+1)^{\frac{1-\alpha}{3\alpha-2}}}{12\sqrt{L}\gamma \ln(\frac{4}{\delta})} \quad (75)$$

the bound

$$\frac{1}{K} \sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2 = \mathcal{O} \left(\frac{1}{(1-\beta_1)^{\frac{3}{2}}} \max \left\{ \frac{LM \ln(\frac{4}{\delta})}{K^{\frac{2\alpha-1}{3\alpha-2}}}, \frac{\sqrt{LM} \sigma \ln^{\frac{\alpha-1}{\alpha}}(\frac{4}{\delta})}{K^{\frac{2\alpha-2}{3\alpha-2}}}, \frac{\sigma^{\frac{2\alpha}{2\alpha-1}} (LM)^{\frac{\alpha-1}{2\alpha-1}} \ln^{\frac{2\alpha-2}{2\alpha-1}}(\frac{4}{\delta})}{K^{\frac{2\alpha-2}{3\alpha-2}}} \right\} \right)$$

holds with probability at least $1 - \delta$.

Proof. The main idea of the proof is similar to the proof of Theorem C.5, but we do not need to introduce any probabilistic events since according to Assumption 1.4 the norm of gradient is always bounded:

$$\|\nabla f(x_t)\| \leq \sqrt{2L(f(x_t) - f_*)} \leq \sqrt{2LM} \stackrel{(75)}{\leq} \frac{\lambda}{2}.$$

Therefore, one can apply Lemma A.4 and get

$$\|\theta_t^u\| \leq 2\lambda, \quad (76)$$

$$\|\theta_t^b\| \leq \frac{2^\alpha \sigma^\alpha}{\lambda^{\alpha-1}}, \quad (77)$$

$$\mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \leq 18\lambda^{2-\alpha} \sigma^\alpha. \quad (78)$$

According to the Lemma C.8, we get

$$\sum_{t=0}^{T-1} \frac{\gamma C_t}{2} \|\nabla f(x_t)\|^2 \leq \left(2M + \frac{2L\gamma^2}{\eta(1-\beta_1)}\right) \sqrt{b_{-1}^2 + \eta \sum_{t=0}^{T-1} \|g_t\|^2} - \sum_{t=0}^{T-1} \gamma C_t \langle \nabla f(x_t), \theta_t^u \rangle + \sum_{t=0}^{T-1} \frac{\gamma C_t}{2} \|\theta_t^b\|^2$$

with $C_t = \sum_{k=t}^{T-1} (1 - \beta_1) \beta_1^{k-t}$ for **Clip-M-AdaGrad-Norm** and

$$\sum_{t=0}^{T-1} \frac{\gamma C_t}{2} \|\nabla f(x_t)\|^2 \leq \left(3M + \frac{16KL\gamma^2}{\eta(1-\beta_1)} \right) \sqrt{b_{-1}^2 + \frac{\eta}{K} \sum_{t=0}^{T-1} \|g_t\|^2} - \sum_{t=0}^{T-1} \gamma C_t \langle \nabla f(x_t), \theta_t^u \rangle + \sum_{t=0}^{T-1} \frac{\gamma C_t}{2} \|\theta_t^b\|^2$$

with $C_t = \sum_{k=t}^{T-1} (1 - \beta_1) \beta_1^{k-t} / (\sqrt{\beta_2})^k$ for **Clip-Adam-Norm**. Let us bound C_t regardless of the method. In can be shown that

$$1 - \beta_1 \leq C_t(\text{Clip-M-AdaGrad-Norm}) \leq \sum_{k=0}^{\infty} (1 - \beta_1) \beta_1^k = 1$$

and

$$1 - \beta_1 \leq C_t(\text{Clip-Adam-Norm}) \leq 2 \sum_{k=0}^{\infty} (1 - \beta_1) \beta_1^k = 2,$$

since $(\sqrt{\beta_2})^{T-1} \geq 1/2$. Therefore, descent lemmas for **Clip-M-AdaGrad-Norm** and **Clip-Adam-Norm** can be rewritten in the following way:

$$\begin{aligned} \frac{\gamma(1-\beta_1)}{2} \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 &\leq \left(2M + \frac{2L\gamma^2}{\eta(1-\beta_1)} \right) \sqrt{b_{-1}^2 + \eta \sum_{t=0}^{T-1} \|g_t\|^2} \\ &\quad - \sum_{t=0}^{T-1} \gamma C_t \langle \nabla f(x_t), \theta_t^u \rangle + \sum_{t=0}^{T-1} \gamma \|\theta_t^b\|^2 \end{aligned} \tag{79}$$

for **Clip-M-AdaGrad-Norm** and

$$\begin{aligned} \frac{\gamma(1-\beta_1)}{2} \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 &\leq \left(3M + \frac{16KL\gamma^2}{\eta(1-\beta_1)} \right) \sqrt{b_{-1}^2 + \frac{\eta}{K} \sum_{t=0}^{T-1} \|g_t\|^2} \\ &\quad - \sum_{t=0}^{T-1} \gamma C_t \langle \nabla f(x_t), \theta_t^u \rangle + \sum_{t=0}^{T-1} \gamma \|\theta_t^b\|^2 \end{aligned} \tag{80}$$

for **Clip-Adam-Norm**. Moreover, $\sum_{t=0}^{T-1} \|g_t\|^2$ can be bounded as follows:

$$\sum_{t=0}^{T-1} \|g_t\|^2 \leq 3 \sum_{t=0}^{T-1} \left(\|\nabla f(x_t)\|^2 + \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right) + \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 + \|\theta_t^b\|^2 \right). \tag{81}$$

The main idea is to give upper bounds for the next terms for all $T \leq K$:

$$\underbrace{\sum_{t=0}^{T-1} \frac{L\gamma^2}{b_{-1}^2} \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right)}_{\textcircled{1}}, \underbrace{\sum_{t=0}^{T-1} \frac{L\gamma^2}{b_{-1}^2} \mathbb{E}_{\xi_t} \|\theta_t^u\|^2}_{\textcircled{2}}, \underbrace{\sum_{t=0}^{T-1} \frac{\gamma}{b_{-1}} \|\theta_t^b\|^2}_{\textcircled{3}}, \underbrace{- \sum_{t=0}^{T-1} \frac{\gamma}{b_{-1}} C_t \langle \nabla f(x_t), \theta_t^u \rangle}_{\textcircled{4}}$$

In cases of $\textcircled{1}$, $\textcircled{2}$ and $\textcircled{3}$ we multiply sums from (81) to the factors to move to the corresponding type of sums from Theorem C.5.

Bound for $\textcircled{1}$. We have bounded and unbiased terms in the sum:

$$\mathbb{E}_{\xi_t} \left[\frac{L\gamma^2}{b_{-1}^2} \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right) \right] = 0$$

and

$$\left| \frac{L\gamma^2}{b_{-1}^2} \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right) \right| \stackrel{(76)}{\leq} \frac{8L\gamma^2\lambda^2}{b_{-1}^2} \leq \frac{24M}{19 \ln \frac{4}{\delta}} = c.$$

Next, we define $\hat{\sigma}_t^2 = \mathbb{E}_{\xi_t} \left[\frac{L^2\gamma^4}{b_{-1}^4} \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right)^2 \right]$. For the introduced quantities, we have

$$\hat{\sigma}_t^2 \leq \frac{cL\gamma^2}{b_{-1}^2} \mathbb{E}_{\xi_t} \left| \|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right| \leq \frac{2cL\gamma^2}{b_{-1}^2} \mathbb{E}_{\xi_t} \|\theta_t^u\|^2.$$

Therefore, we can apply Bernstein's inequality (Lemma A.5) with $G = \frac{3M^2}{38 \ln(\frac{4}{\delta})}$:

$$\mathbb{P} \left\{ \left| \sum_{t=0}^{T-1} \frac{L\gamma^2}{b_{-1}^2} \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right) \right| > M \text{ and } \sum_{t=0}^{T-1} \hat{\sigma}_t^2 \leq G \right\} \leq 2 \exp \left(-\frac{M^2}{2G + \frac{2cM}{3}} \right) = \frac{\delta}{2}.$$

Thus, we get

$$\mathbb{P} \left\{ \text{either } \left| \sum_{t=0}^{T-1} \frac{L\gamma^2}{b_{-1}^2} \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \right) \right| \leq M \text{ or } \sum_{t=0}^{T-1} \hat{\sigma}_t^2 > G \right\} \geq 1 - \frac{\delta}{2}.$$

Moreover,

$$\sum_{t=0}^{T-1} \hat{\sigma}_t^2 \stackrel{(78)}{\leq} \frac{36cTL\gamma^2\lambda^{2-\alpha}\sigma^\alpha}{b_{-1}^2} \stackrel{(75)}{\leq} \frac{36cTL\gamma^\alpha\sqrt{M}^{2-\alpha}K^{\frac{(1-\alpha)(2-\alpha)}{3\alpha-2}}}{12^{2-\alpha}b_{-1}^\alpha\sqrt{L}^{2-\alpha}\ln^{2-\alpha}(\frac{4}{\delta})} \stackrel{(74)}{\leq} \frac{3M^2}{38 \ln(\frac{4}{\delta})}.$$

Bound for ②. For the second term, we get

$$\sum_{t=0}^{T-1} \frac{L\gamma^2}{b_{-1}^2} \mathbb{E}_{\xi_t} \|\theta_t^u\|^2 \stackrel{(78)}{\leq} \frac{18TL\gamma^2\lambda^{2-\alpha}\sigma^\alpha}{b_{-1}^2} \stackrel{(75)}{\leq} \frac{18TL\gamma^\alpha\sqrt{M}^{2-\alpha}K^{\frac{(1-\alpha)(2-\alpha)}{3\alpha-2}}}{12^{2-\alpha}b_{-1}^\alpha\sqrt{L}^{2-\alpha}\ln^{2-\alpha}(\frac{4}{\delta})} \stackrel{(74)}{\leq} \frac{M}{32} \leq M.$$

Bound for ③. For the third sum, we obtain

$$\sum_{t=0}^{T-1} \frac{\gamma}{b_{-1}} \|\theta_t^b\|^2 \stackrel{(77)}{\leq} \frac{4^\alpha \sigma^{2\alpha} \gamma T}{b_{-1} \lambda^{2\alpha-2}} \stackrel{(75)}{=} \frac{4^\alpha 12^{2\alpha-2} \sigma^{2\alpha} \gamma^{2\alpha-1} T L^{\alpha-1} \ln^{2\alpha-2}(\frac{4}{\delta})}{b_{-1}^{2\alpha-1} M^{\alpha-1} K^{\frac{(1-\alpha)(2\alpha-2)}{3\alpha-2}}} \stackrel{(74)}{\leq} M,$$

where we choose the third option for γ .

Bound for ④. Similarly to ①, we have unbiased and bounded terms in sum:

$$\mathbb{E}_{\xi_t} \left[-\frac{\gamma C_t}{b_{-1}} \langle \nabla f(x_t), \theta_t^u \rangle \right] = 0$$

and

$$\left| -\frac{\gamma C_t}{b_{-1}} \langle \nabla f(x_t), \theta_t^u \rangle \right| \leq \frac{2\gamma}{b_{-1}} \|\nabla f(x_t)\| \|\theta_t^u\| \stackrel{(76)}{\leq} \frac{4\gamma\lambda\sqrt{2LM}}{b_{-1}} \leq \frac{3M}{4 \ln(\frac{4}{\delta})} = c.$$

Let us define $\sigma_t^2 = \mathbb{E}_{\xi_t} \left[\frac{\gamma^2 C_t^2}{b_{-1}^2} \langle \nabla f(x_t), \theta_t^u \rangle^2 \right]$. Hence,

$$\sigma_t^2 \leq \frac{8\gamma^2 LM}{b_{-1}^2} \mathbb{E}_{\xi_t} \|\theta_t^u\|^2.$$

Therefore, we can apply Bernstein's inequality (Lemma A.5) with $G = \frac{M^2}{4 \ln(\frac{4}{\delta})}$:

$$\mathbb{P} \left\{ \left| - \sum_{t=0}^{T-1} \frac{\gamma C_t}{b_{-1}} \langle \nabla f(x_t), \theta_t^u \rangle \right| > M \text{ and } \sum_{t=0}^{T-1} \sigma_t^2 \leq G \right\} \leq 2 \exp \left(-\frac{M^2}{2G + \frac{2cM}{3}} \right) = \frac{\delta}{2}.$$

Thus, we get

$$\mathbb{P} \left\{ \text{either } \left| - \sum_{t=0}^{T-1} \frac{\gamma C_t}{b_{-1}} \langle \nabla f(x_t), \theta_t^u \rangle \right| \leq M \text{ or } \sum_{t=0}^{T-1} \sigma_t^2 > G \right\} \geq 1 - \frac{\delta}{2}.$$

Moreover,

$$\sum_{t=0}^{T-1} \sigma_t^2 \stackrel{(78)}{\leq} \frac{144\gamma^2 LMT\lambda^{2-\alpha}\sigma^\alpha}{b_{-1}^2} \stackrel{(75)}{=} \frac{144\sqrt{M}^{2-\alpha} K^{\frac{(1-\alpha)(2-\alpha)}{3\alpha-2}} \gamma^\alpha LMT\sigma^\alpha}{12^{2-\alpha} b_{-1}^\alpha \sqrt{L}^{2-\alpha} \ln^{2-\alpha}(\frac{4}{\delta})} \stackrel{(74)}{\leq} \frac{M^2}{4\ln(\frac{4}{\delta})}.$$

Consequently, next inequality holds with probability at least $1 - \delta$ for all $T \leq K$:

$$\sum_{t=0}^{T-1} \|g_t\|^2 \leq 3 \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 + \frac{6Mb_{-1}^2}{L\gamma^2} + \frac{3Mb_{-1}}{\gamma}.$$

Let us specify η for each method. This parameter can be chosen as follows:

$$\eta = \begin{cases} \frac{L\gamma^2}{M(1-\beta_1)}, & \text{for Clip-M-AdaGrad-Norm} \\ \frac{KL\gamma^2}{M(1-\beta_1)}, & \text{for Clip-Adam-Norm} \end{cases}$$

Therefore, (79) and (80) can be rewritten in an unified form with $T = K$ and ①, ②, ③ and ④:

$$\frac{\gamma(1-\beta_1)}{2} \sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2 \leq 19M \sqrt{b_{-1}^2 + \frac{3L\gamma^2}{M(1-\beta_1)} \sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2 + \frac{6b_{-1}^2}{1-\beta_1} + \frac{3L\gamma b_{-1}}{1-\beta_1} + 2Mb_{-1}}$$

holds with probability at least $1 - \delta$ for both algorithms. Denoting $\sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2$ as S_K and squaring the inequality above, we get

$$\begin{aligned} \frac{\gamma^2(1-\beta_1)^2}{4} S_K^2 &\leq \left(19M \sqrt{b_{-1}^2 + \frac{3L\gamma^2}{M(1-\beta_1)} S_K + \frac{6b_{-1}^2}{1-\beta_1} + \frac{3L\gamma b_{-1}}{1-\beta_1} + 2M} \right)^2 \\ &\leq 762M^2 \left(b_{-1}^2 + \frac{3L\gamma^2}{M(1-\beta_1)} S_K + \frac{6b_{-1}^2}{1-\beta_1} + \frac{3L\gamma b_{-1}}{1-\beta_1} \right) + 8M^2 b_{-1}^2, \end{aligned}$$

where we use the fact that $(a+b)^2 \leq 2a^2 + 2b^2$. Rearranging the terms, we have

$$S_K^2 - \frac{6 \cdot 38^2 LM}{(1-\beta_1)^3} S_K - \frac{2 \cdot 38^2 M^2}{\gamma^2(1-\beta_1)^2} \left(b_{-1}^2 + \frac{8b_{-1}^2}{762} + \frac{6b_{-1}}{1-\beta_1} + \frac{3L\gamma b_{-1}}{1-\beta_1} \right) \leq 0.$$

Solving the quadratic inequality and using that $\sqrt{a^2 + b^2} \leq a + b$, one can obtain

$$\begin{aligned} S_K &\leq \frac{6 \cdot 38^2 LM}{(1-\beta_1)^3} + \frac{38\sqrt{2}M}{\gamma(1-\beta_1)} \sqrt{b_{-1}^2 + \frac{8b_{-1}^2}{762} + \frac{6b_{-1}^2}{1-\beta_1} + \frac{3L\gamma b_{-1}}{1-\beta_1}} \\ &\leq \frac{6 \cdot 38^2 LM}{(1-\beta_1)^3} + \frac{38\sqrt{2}M}{\gamma(1-\beta_1)} \left(\frac{21b_{-1}}{19} + \frac{3b_{-1}}{\sqrt{1-\beta_1}} \right), \end{aligned}$$

because $L\gamma \leq \frac{b_{-1}}{48}$. Therefore, after division of both sides by K and substitution of γ from (74), we get the final bound for Clip-M-AdaGrad-Norm/Clip-Adam-Norm:

$$\frac{1}{K} \sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2 = \mathcal{O} \left(\frac{1}{(1-\beta_1)^{\frac{3}{2}}} \max \left\{ \frac{LM \ln(\frac{4}{\delta})}{K^{\frac{2\alpha-1}{3\alpha-2}}}, \frac{\sqrt{LM} \sigma \ln^{\frac{\alpha-1}{\alpha}}(\frac{4}{\delta})}{K^{\frac{2\alpha-2}{3\alpha-2}}}, \frac{\sigma^{\frac{2\alpha}{2\alpha-1}} (LM)^{\frac{\alpha-1}{2\alpha-1}} \ln^{\frac{2\alpha-2}{2\alpha-1}}(\frac{4}{\delta})}{K^{\frac{2\alpha-2}{3\alpha-2}}} \right\} \right)$$

with probability at least $1 - \delta$. \square

C.5. Non-Convex Case: Coordinate-wise Methods with Delay

Similarly to the methods with scalar stepsizes, for convenience, we consider an reweighted forms of the coordinate-wise methods, which are equivalent to non-reweighted ones.

Algorithm 4 Clip-Adam/Clip-AdamD and Clip-M-AdaGrad/Clip-M-AdaGradD

Input: Stepsize $\gamma > 0$, starting point $x_0 \in \mathbb{R}^d$, initial constant $b_{-1} > 0$ (for Adam and M-AdaGrad) or $b_0 > 0$ (for AdamD and M-AdaGradD), momentum parameters $\beta_1, \beta_2 \in [0, 1]$, level of clipping $\lambda_i > 0$ for each coordinate, reweighting parameter $\eta > 0$

- 1: Set $m_{-1} = 0$
- 2: **for** $t = 0, 1, \dots$ **do**
- 3: $m_{t,i} = \beta_1 m_{t-1,i} + (1 - \beta_1) \text{clip}([\nabla f_{\xi_t}(x_t)]_i, \lambda_i)$
- 4: **if** no delay **then**
- 5: $b_{t,i} = \begin{cases} \sqrt{\beta_2 b_{t-1,i}^2 + \eta(1 - \beta_2) |\text{clip}([\nabla f_{\xi_t}(x_t)]_i, \lambda_i)|^2} & \text{for Clip-Adam} \\ \sqrt{b_{t-1,i}^2 + \eta |\text{clip}([\nabla f_{\xi_t}(x_t)]_i, \lambda_i)|^2} & \text{for Clip-M-AdaGrad} \end{cases}$
- 6: **else**
- 7: $b_{t+1,i} = \begin{cases} \sqrt{\beta_2 b_{t,i}^2 + \eta(1 - \beta_2) |\text{clip}([\nabla f_{\xi_t}(x_t)]_i, \lambda_i)|^2} & \text{for Clip-AdamD} \\ \sqrt{b_{t,i}^2 + \eta |\text{clip}([\nabla f_{\xi_t}(x_t)]_i, \lambda_i)|^2} & \text{for Clip-M-AdaGradD} \end{cases}$
- 8: **end if**
- 9: $x_{t+1,i} = x_{t,i} - \frac{\gamma}{b_{t,i}} m_{t,i}$
- 10: **end for**

To improve the readability of the proof of coordinate-wise case, we introduce the following notation for this subsection:

- $\nabla_{t,i} := [\nabla f(x_t)]_i$,
- $g_{t,i} := [g_t]_i$,
- $\theta_{t,i} := g_{t,i} - \nabla_{t,i}$,
- $\theta_{t,i}^u := g_{t,i} - \mathbb{E}_{\xi_t}[g_{t,i}]$,
- $\theta_{t,i}^b := \mathbb{E}_{\xi_t}[g_{t,i}] - \nabla_{t,i}$.

Lemma C.10 (Descent lemma). *Let Assumption 1.2 hold on $Q = \left\{ x \in \mathbb{R}^d \mid \exists y \in \mathbb{R}^d : f(y) \leq f_* + 2\Delta \text{ and } \|x - y\| \leq \frac{\sqrt{\Delta}}{20\sqrt{L}} \right\}$, where $f(x_0) - f_* = \Delta_0 \leq \Delta$. Then, after T iterations of Clip-M-AdaGradD/Clip-AdamD with $b_0 \geq \gamma L / c_m$, if $x_t \in Q \forall t = \overline{0, T}$, we have*

$$\begin{aligned} \sum_{t=0}^{T-1} \sum_{i=1}^d \frac{\gamma C_{t,i}}{2} \nabla_{t,i}^2 &\leq (f(x_0) - f^*) - (f(x_T) - f^*) - \sum_{t=0}^{T-1} \sum_{i=1}^d (\gamma C_{t,i} - 2A_{t,i}) \nabla_{t,i} \theta_{t,i}^u \\ &\quad + \sum_{t=0}^{T-1} \sum_{i=1}^d 2A_{t,i} (\theta_{t,i}^u)^2 + \sum_{t=0}^{T-1} \sum_{i=1}^d \gamma C_{t,i} (\theta_{t,i}^b)^2, \end{aligned}$$

where $C_{t,i} = (1 - \beta_1) \sum_{k=t}^{T-1} \frac{\beta_1^{k-t}}{b_{k,i}}$ and $A_{t,i} = \frac{L\gamma^2(1-\beta_1)}{c_m} \sum_{k=t}^{T-1} \sum_{p=t}^k \frac{\beta_1^{k-p}}{b_{p,i}^2}$.

Proof. Starting with L -smoothness, we derive

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 = -\gamma \sum_{i=1}^d \frac{\nabla_{t,i} m_{t,i}}{b_{t,i}} + \frac{L\gamma^2}{2} \sum_{i=1}^d \frac{m_{t,i}^2}{b_{t,i}^2} \\ &= \sum_{i=1}^d \left[-\gamma \frac{\nabla_{t,i} m_{t,i}}{b_{t,i}} + \frac{L\gamma^2}{2} \frac{m_{t,i}^2}{b_{t,i}^2} \right]. \end{aligned} \quad (82)$$

Similarly to the proof of Lemma C.3, we get

$$\begin{aligned} -\nabla_{t,i} m_{t,i} &= -\beta_1 \nabla_{t,i} m_{t-1,i} - (1 - \beta_1) \nabla_{t,i} g_{t,i} \\ &= -\beta_1 (\nabla_{t,i} - \nabla_{t-1,i}) m_{t-1,i} - \beta_1 \nabla_{t-1,i} m_{t-1,i} - (1 - \beta_1) \nabla_{t,i} g_{t,i} \\ &\leq \beta_1 |\nabla_{t,i} - \nabla_{t-1,i}| |m_{t-1,i}| - \beta_1 \nabla_{t-1,i} m_{t-1,i} - (1 - \beta_1) \nabla_{t,i} g_{t,i}. \end{aligned}$$

Using that $m_{-1,i} = 0$ with the above recursive inequality, we get

$$-\nabla_{t,i} m_{t,i} \leq -(1 - \beta_1) \sum_{k=0}^t \beta_1^{t-k} \nabla_{k,i} g_{k,i} + \sum_{k=0}^{t-1} \beta_1^{t-k} |\nabla_{k+1,i} - \nabla_{k,i}| |m_{k,i}|.$$

Dividing both sides by $b_{t,i}$ and summing over i , we have

$$-\sum_{i=1}^d \frac{\nabla_{t,i} m_{t,i}}{b_{t,i}} \leq -(1 - \beta_1) \sum_{k=0}^t \beta_1^{t-k} \sum_{i=1}^d \frac{\nabla_{k,i} g_{k,i}}{b_{t,i}} + \sum_{k=0}^{t-1} \beta_1^{t-k} \sum_{i=1}^d \frac{|\nabla_{k+1,i} - \nabla_{k,i}| |m_{k,i}|}{b_{t,i}}. \quad (83)$$

Then, we apply Cauchy-Schwarz inequality to the last term in the right-hand side:

$$\begin{aligned} \sum_{k=0}^{t-1} \beta_1^{t-k} \sum_{i=1}^d \frac{|\nabla_{k+1,i} - \nabla_{k,i}| |m_{k,i}|}{b_{t,i}} &\leq \sum_{k=0}^{t-1} \beta_1^{t-k} \left(\sqrt{\sum_{i=1}^d (\nabla_{k+1,i} - \nabla_{k,i})^2} \right) \left(\sqrt{\sum_{i=1}^d \frac{m_{k,i}^2}{b_{t,i}^2}} \right) \\ &= \sum_{k=0}^{t-1} \beta_1^{t-k} \|\nabla f(x_{k+1}) - \nabla f(x_k)\| \left(\sqrt{\sum_{i=1}^d \frac{m_{k,i}^2}{b_{t,i}^2}} \right) \\ &\leq \sum_{k=0}^{t-1} \beta_1^{t-k} L \|x_{k+1} - x_k\| \left(\sqrt{\sum_{i=1}^d \frac{m_{k,i}^2}{b_{t,i}^2}} \right) \\ &= L \gamma \sum_{k=0}^{t-1} \beta_1^{t-k} \left(\sqrt{\sum_{i=1}^d \frac{m_{k,i}^2}{b_{k,i}^2}} \right) \left(\sqrt{\sum_{i=1}^d \frac{m_{k,i}^2}{b_{t,i}^2}} \right) \\ &\leq \frac{L\gamma}{c_m} \sum_{k=0}^{t-1} \beta_1^{t-k} \sum_{i=1}^d \frac{m_{k,i}^2}{b_{k,i}^2}, \end{aligned}$$

where in the second inequality we apply L -smoothness, and in the last inequality we apply Lemma C.1. Therefore, substituting the inequality above into (83) and combining it with (82), we derive

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq -(1 - \beta_1) \gamma \sum_{k=0}^t \beta_1^{t-k} \sum_{i=1}^d \frac{\nabla_{k,i} g_{k,i}}{b_{t,i}} + \frac{L\gamma^2}{c_m} \sum_{k=0}^{t-1} \beta_1^{t-k} \sum_{i=1}^d \frac{m_{k,i}^2}{b_{k,i}^2} + \frac{L\gamma^2}{2} \sum_{i=1}^d \frac{m_{t,i}^2}{b_{t,i}^2} \\ &\leq -(1 - \beta_1) \gamma \sum_{k=0}^t \beta_1^{t-k} \sum_{i=1}^d \frac{\nabla_{k,i} g_{k,i}}{b_{t,i}} + \frac{L\gamma^2}{c_m} \sum_{k=0}^t \beta_1^{t-k} \sum_{i=1}^d \frac{m_{k,i}^2}{b_{k,i}^2}. \end{aligned}$$

Applying Lemma C.2 to the last term, we get

$$f(x_{t+1}) - f(x_t) \leq -(1 - \beta_1) \gamma \sum_{k=0}^t \beta_1^{t-k} \sum_{i=1}^d \frac{\nabla_{k,i} g_{k,i}}{b_{t,i}} + \frac{L\gamma^2(1 - \beta_1)}{c_m} \sum_{k=0}^t \beta_1^{t-k} \sum_{i=1}^d \sum_{j=0}^k \beta_1^{k-j} \frac{g_{j,i}^2}{b_{k,i}^2}.$$

Summing over t , one can obtain

$$(f(x_T) - f^*) - (f(x_0) - f^*) \leq -(1 - \beta_1)\gamma \sum_{t=0}^{T-1} \sum_{k=0}^t \beta_1^{t-k} \sum_{i=1}^d \frac{\nabla_{k,i} g_{k,i}}{b_{t,i}} + \frac{L\gamma^2(1 - \beta_1)}{c_m} \sum_{t=0}^{T-1} \sum_{k=0}^t \beta_1^{t-k} \sum_{i=1}^d \sum_{j=0}^k \beta_1^{k-j} \frac{g_{j,i}^2}{b_{k,i}^2}. \quad (84)$$

The rest of the proof follows Lemma C.3. Let us denote $-\gamma C_{r,i}$ and $A_{r,i}$ as the coefficients in front of $\nabla_{r,i} g_{r,i}$ and $g_{r,i}^2$ in (84), respectively. These coefficients equal to

$$-\gamma C_{r,i} = -(1 - \beta_1)\gamma \sum_{t=r}^{T-1} \frac{\beta_1^{t-r}}{b_{t,i}}, \quad (85)$$

$$A_{r,i} = \frac{L\gamma^2(1 - \beta_1)}{c_m} \sum_{t=r}^{T-1} \sum_{k=r}^t \frac{\beta_1^{t-r}}{b_{k,i}^2}. \quad (86)$$

Following the same steps as in the proof of Lemma C.3, we get useful bounds (will be used later)

$$\frac{(1 - \beta_1)}{b_{r,i}} \leq C_{r,i} \leq \frac{1}{c_m b_0}, \quad A_{r,i} \leq \frac{L\gamma^2}{c_m^3 b_{r,i} b_0 (1 - \beta_1)} \quad (87)$$

due to Lemma A.2 and Lemma C.1. Rewriting (84) with (85) and (86), we have

$$(f(x_T) - f^*) - (f(x_0) - f^*) \leq - \sum_{t=0}^{T-1} \sum_{i=1}^d \gamma C_{t,i} \nabla_{t,i} g_{t,i} + \sum_{t=0}^{T-1} \sum_{i=1}^d A_{t,i} g_{t,i}^2.$$

Following the same steps as in the proof of Lemma C.3, we get

$$\begin{aligned} (f(x_T) - f^*) - (f(x_0) - f^*) &\leq - \sum_{t=0}^{T-1} \sum_{i=1}^d \gamma C_{t,i} \nabla_{t,i} g_{t,i} + \sum_{t=0}^{T-1} \sum_{i=1}^d A_{t,i} g_{t,i}^2 \\ &\leq - \sum_{t=0}^{T-1} \sum_{i=1}^d (\gamma C_{t,i} - 2A_{t,i}) \nabla_{t,i} \theta_{t,i} - \sum_{t=0}^{T-1} \sum_{i=1}^d (\gamma C_{t,i} - A_{t,i}) \nabla_{t,i}^2 \\ &\quad + \sum_{t=0}^{T-1} \sum_{i=1}^d A_{t,i} \theta_{t,i}^2. \end{aligned}$$

Using the notation of $\theta_{t,i}^u$ and $\theta_{t,i}^b$, with $\gamma \leq \frac{(1-\beta_1)^2 c_m^3 b_0}{2L}$, which implies $\gamma C_{t,i} - 2A_{t,i} \geq 0$ due to (87),

$$\begin{aligned}
 (f(x_T) - f^*) - (f(x_0) - f^*) &\leq -\sum_{t=0}^{T-1} \sum_{i=1}^d (\gamma C_{t,i} - 2A_{t,i}) \nabla_{t,i} \theta_{t,i} - \sum_{t=0}^{T-1} \sum_{i=1}^d (\gamma C_{t,i} - A_{t,i}) \nabla_{t,i}^2 \\
 &\quad + \sum_{t=0}^{T-1} \sum_{i=1}^d A_{t,i} \theta_{t,i}^2 \\
 &\leq -\sum_{t=0}^{T-1} \sum_{i=1}^d (\gamma C_{t,i} - 2A_{t,i}) \nabla_{t,i} \theta_{t,i}^u - \sum_{t=0}^{T-1} \sum_{i=1}^d (\gamma C_{t,i} - A_{t,i}) \nabla_{t,i}^2 \\
 &\quad + \sum_{t=0}^{T-1} \sum_{i=1}^d 2A_{t,i} ((\theta_{t,i}^u)^2 + (\theta_{t,i}^b)^2) + \sum_{t=0}^{T-1} \sum_{i=1}^d \left(\frac{\gamma C_{t,i}}{2} - A_{t,i} \right) (\theta_{t,i}^b)^2 \\
 &\quad + \sum_{t=0}^{T-1} \sum_{i=1}^d \left(\frac{\gamma C_{t,i}}{2} - A_{t,i} \right) \nabla_{t,i}^2 \\
 &= -\sum_{t=0}^{T-1} \sum_{i=1}^d \frac{\gamma C_{t,i}}{2} \nabla_{t,i}^2 - \sum_{t=0}^{T-1} \sum_{i=1}^d (\gamma C_{t,i} - 2A_{t,i}) \nabla_{t,i} \theta_{t,i}^u \\
 &\quad + \sum_{t=0}^{T-1} \sum_{i=1}^d 2A_{t,i} (\theta_{t,i}^u)^2 + \sum_{t=0}^{T-1} \sum_{i=1}^d \gamma C_{t,i} (\theta_{t,i}^b)^2,
 \end{aligned}$$

where we use the fact that $\frac{\gamma C_{t,i}}{2} \geq A_{t,i}$. Rearranging the terms, we conclude the proof. \square

Next, to reflect tighter dependencies on the variance of different coordinates of the stochastic gradient, we make the following assumption.

Assumption C.11. There exists set $Q \subseteq \mathbb{R}^d$ and $\sigma \geq 0, \alpha \in (1, 2]$ such that the oracle satisfies $\mathbb{E}[\nabla f_\xi(x)] = \nabla f(x)$ and

$$\mathbb{E}[\|\nabla_i f_\xi(x) - \nabla_i f(x)\|^\alpha] \leq \sigma_i^\alpha, \quad \forall i \in [d] \text{ and } \forall x \in Q. \quad (88)$$

Moreover, we assume that $\{\nabla f_\xi(x)\}_{i=1}^d$ are independent.

Theorem C.12. Let Assumptions C.11 and I.2 hold on $Q = \left\{ x \in \mathbb{R}^d \mid \exists y \in \mathbb{R}^d : f(y) \leq f_* + 2\Delta \text{ and } \|x - y\| \leq \frac{\sqrt{\Delta}}{20\sqrt{L}} \right\}$ with $f(x_0) - f_* = \Delta_0 \leq \Delta$. Then, after $K + 1$ iterations of Clip-M-AdaGradD/Clip-AdamD with

$$\begin{aligned}
 \gamma &\leq \min \left\{ \frac{(1-\beta_1)^2 c_m b_0 (K+1)^{\frac{1-\alpha}{3\alpha-2}}}{40L\sqrt{d} \ln \frac{4(K+1)}{\delta}}, \frac{35^{\frac{1}{\alpha}} c_m \sqrt{1-\beta_1} b_0 \sqrt{\Delta} \sqrt{d}^{\frac{2-\alpha}{\alpha}}}{432^{\frac{1}{\alpha}} \cdot 20\sqrt{L} \left(\sum_{i=1}^d \sigma_i^\alpha \right)^{\frac{1}{\alpha}} (K+1)^{\frac{\alpha-1}{3\alpha-2}} \ln^{\frac{\alpha-1}{\alpha}} \frac{4(K+1)}{\delta}}, \right. \\
 &\quad \left. \frac{c_m (1-\beta_1)^{\frac{\alpha-1}{2\alpha-1}} b_0 \Delta^{\frac{\alpha}{2\alpha-1}}}{4^{\frac{\alpha+1}{2\alpha-1}} \cdot 20^{\frac{2\alpha-2}{2\alpha-1}} \left(\sum_{i=1}^d \sigma_i^{2\alpha} \right)^{\frac{1}{2\alpha-1}} d^{\frac{\alpha-1}{2\alpha-1}} L^{\frac{\alpha-1}{2\alpha-1}} (K+1)^{\frac{\alpha}{3\alpha-2}} \ln^{\frac{2\alpha-2}{2\alpha-1}} \left(\frac{4(K+1)}{\delta} \right)} \right\}, \quad \eta = \frac{L\gamma^2(1-\beta_1)^2}{\Delta},
 \end{aligned} \quad (89)$$

and

$$\lambda_i \equiv \lambda = \frac{c_m \sqrt{1-\beta_1} b_0 \sqrt{\Delta} (K+1)^{\frac{1-\alpha}{3\alpha-2}}}{20\sqrt{d}\sqrt{L}\gamma \ln \left(\frac{4(K+1)}{\delta} \right)} \quad (90)$$

the bound

$$\sum_{k=0}^K \sum_{i=1}^d \frac{\gamma C_{k,i}}{2} \nabla_{k,i}^2 \leq 2\Delta$$

holds with probability at least $1 - \delta$. In particular, when γ equals the minimum from (89), the iterates produced by Clip-M-AdaGradD/Clip-AdamD satisfy

$$\begin{aligned} & \frac{1}{K+1} \sum_{k=0}^K \|\nabla f(x_k)\|^2 \\ &= \mathcal{O} \left(\max \left\{ \frac{\sqrt{d} L \Delta \ln \frac{K+1}{\delta}}{(1-\beta_1)^3 (K+1)^{\frac{2\alpha-1}{3\alpha-2}}}, \frac{\sqrt{L\Delta} \left(\sum_{i=1}^d \sigma_i^\alpha \right)^{\frac{1}{\alpha}} \ln^{\frac{\alpha-1}{\alpha}} \frac{K+1}{\delta}}{(1-\beta_1)^{\frac{3}{2}} \sqrt{d}^{\frac{2-\alpha}{\alpha}} (K+1)^{\frac{2\alpha-2}{3\alpha-2}}}, \right. \right. \\ & \quad \left. \left. \frac{d^{\frac{\alpha-1}{2\alpha-1}} \left(\sum_{i=1}^d \sigma_i^{2\alpha} \right)^{\frac{1}{2\alpha-1}} (L\Delta)^{\frac{\alpha-1}{2\alpha-1}} \ln^{\frac{2\alpha-2}{2\alpha-1}} \frac{K+1}{\delta}}{(1-\beta_1)^{\frac{\alpha-1}{2\alpha-1}} (K+1)^{\frac{2\alpha-2}{3\alpha-2}}} \right\} \right) \end{aligned}$$

with probability at least $1 - \delta$.

Proof. We construct the proof in a similar way as the proof of Theorem C.5. The probability event E_k is defined as follows: inequalities

$$\begin{aligned} \sum_{l=0}^{t-1} \sum_{i=1}^d \left[-(\gamma C_{l,i} - 2A_{l,i}) \nabla_{l,i} \theta_{l,i}^u + 2A_{l,i} (\theta_{l,i}^u)^2 + \gamma C_{l,i} (\theta_{l,i}^b)^2 \right] &\leq \Delta, \\ \Delta_t &\leq 2\Delta. \end{aligned}$$

hold simultaneously $\forall t = \overline{0, k}$. The main idea is to show that $\mathbb{P}\{E_k\} \geq 1 - \frac{k\delta}{K+1} \forall k = \overline{0, K+1}$. The case $k = 0$ is obvious. According to an induction step, we assume that this statement holds for some $k = T-1 \leq K : \mathbb{P}\{E_{T-1}\} \geq 1 - \frac{(T-1)\delta}{K+1}$. We need to prove that $\mathbb{P}\{E_T\} \geq 1 - \frac{T\delta}{K+1}$. The event E_{T-1} implies that $x_t \in \{y \in \mathbb{R}^d : f(y) \leq f^* + 2\Delta\} \forall t = 0, \dots, T-1$ and

$$\|x_T - x_{T-1}\| = \gamma \sqrt{\sum_{i=1}^d \frac{m_{t,i}^2}{b_{t,i}^2}} \leq \gamma \sqrt{\sum_{i=1}^d \frac{m_{t,i}^2}{c_m b_0^2}} \leq \frac{\gamma}{c_m b_0} \sqrt{\sum_{i=1}^d \lambda_i^2} \leq \frac{\sqrt{\Delta}}{20\sqrt{L}}.$$

Hence, event E_{T-1} implies $\{x_t\}_{t=0}^{T-1} \subseteq Q$ and according to Lemma C.10,

$$\sum_{l=0}^{t-1} \sum_{i=1}^d \gamma C_{l,i} \nabla_{l,i}^2 \leq \Delta_0 - \Delta_t + \sum_{l=0}^{t-1} \sum_{i=1}^d \left[-(\gamma C_{l,i} - 2A_{l,i}) \nabla_{l,i} \theta_{l,i}^u + 2A_{l,i} (\theta_{l,i}^u)^2 + \gamma C_{l,i} (\theta_{l,i}^b)^2 \right]$$

$\forall t = \overline{1, T}$ and $\forall t = \overline{1, T-1}$ it implies that

$$\begin{aligned} \sum_{l=0}^{t-1} \sum_{i=1}^d \gamma C_{l,i} \nabla_{l,i}^2 &\leq \Delta_0 - \Delta_t + \sum_{l=0}^{t-1} \sum_{i=1}^d \left[-(\gamma C_{l,i} - 2A_{l,i}) \nabla_{l,i} \theta_{l,i}^u + 2A_{l,i} (\theta_{l,i}^u)^2 + \gamma C_{l,i} (\theta_{l,i}^b)^2 \right] \\ &\leq 2\Delta. \end{aligned}$$

Taking into account that left-hand side is greater than zero (since every term is nonnegative), E_{T-1} implies

$$\Delta_T \leq \Delta_0 + \sum_{l=0}^{t-1} \sum_{i=1}^d \left[-(\gamma C_{l,i} - 2A_{l,i}) \nabla_{l,i} \theta_{l,i}^u + 2A_{l,i} (\theta_{l,i}^u)^2 + \gamma C_{l,i} (\theta_{l,i}^b)^2 \right].$$

Next, let us denote

$$\eta_{t,i} = \begin{cases} \nabla_{t,i}, & \|\nabla f(x_t)\| \leq 2\sqrt{L\Delta} \\ 0, & \text{otherwise} \end{cases} \quad (91)$$

$\forall t = 0, \dots, T-1$. Therefore, the event E_{T-1} implies $\eta_{t,i} = \nabla_{t,i}$ since

$$|\nabla_{t,i}| \leq \|\nabla f(x_t)\| \leq \sqrt{2L\Delta_t} \leq 2\sqrt{L\Delta} \stackrel{(90)}{\leq} \frac{\lambda_i}{2}.$$

Thus, we obtain that E_{T-1} implies

$$\begin{aligned} \Delta_t &\leq \Delta_0 + \sum_{t=0}^{T-1} \sum_{i=1}^d \left[-(\gamma C_{t,i} - 2A_{t,i}) \eta_{t,i} \theta_{t,i}^u \right] \\ &+ \sum_{t=0}^{T-1} \sum_{i=1}^d \left[2A_{t,i} (\theta_{t,i}^u)^2 \right] + \sum_{t=0}^{T-1} \sum_{i=1}^d \left[\gamma C_{t,i} (\theta_{t,i}^b)^2 \right] \\ &= \underbrace{\sum_{t=0}^{T-1} \sum_{i=1}^d \left[-(\gamma C_{t,i} - 2A_{t,i}) \eta_{t,i} \theta_{t,i}^u \right]}_{\textcircled{1}} + \underbrace{\sum_{t=0}^{T-1} \sum_{i=1}^d \left[2A_{t,i} \left[(\theta_{t,i}^u)^2 - \mathbb{E}_{\xi_t} [(\theta_{t,i}^u)^2] \right] \right]}_{\textcircled{2}} \\ &+ \underbrace{\sum_{t=0}^{T-1} \sum_{i=1}^d \left[2A_{t,i} \mathbb{E}_{\xi_t} (\theta_{t,i}^u)^2 \right]}_{\textcircled{3}} + \underbrace{\sum_{t=0}^{T-1} \sum_{i=1}^d \left[\gamma C_{t,i} (\theta_{t,i}^b)^2 \right]}_{\textcircled{4}}. \end{aligned} \quad (92)$$

It remains to give upper bounds for each term in (92). Moreover, due to (90) we get $|\nabla_{t,i}| \leq \frac{\lambda_i}{2}$. Therefore, we can apply Lemma A.4 and get

$$|\theta_{t,i}^u| \leq 2\lambda_i, \quad (93)$$

$$|\theta_{t,i}^b| \leq \frac{2^\alpha \sigma_i^\alpha}{\lambda_i^{\alpha-1}}, \quad (94)$$

$$\mathbb{E}_{\xi_t} [(\theta_{t,i}^u)^2] \leq 18\lambda_i^{2-\alpha} \sigma_i^\alpha. \quad (95)$$

Bound for ①. The definition of $\theta_{t,i}^u$ implies

$$\mathbb{E}_{\xi_t} \left[\sum_{i=1}^d \left[-(\gamma C_{t,i} - 2A_{t,i}) \eta_{t,i} \theta_{t,i}^u \right] \right] = 0.$$

What is more, we have

$$\begin{aligned} \left| \sum_{i=1}^d \left[-(\gamma C_{t,i} - 2A_{t,i}) \eta_{t,i} \theta_{t,i}^u \right] \right| &\leq \sqrt{\sum_{i=1}^d \eta_{t,i}^2} \sqrt{\sum_{i=1}^d \gamma^2 C_{t,i}^2 (\theta_{t,i}^u)^2} \stackrel{(91), (93)}{\leq} \frac{4\sqrt{d}\gamma\lambda\sqrt{L\Delta}}{c_m b_0} \\ &\leq \frac{\Delta}{5 \ln \left(\frac{4(K+1)}{\delta} \right)} = c, \end{aligned}$$

where we use that $(\gamma C_{t,i} - 2A_{t,i}) \geq 0$ due to the choice of γ . Also let us define $\sigma_t^2 = \mathbb{E}_{\xi_t} \left[\left(\sum_{i=1}^d \left[-(\gamma C_{t,i} - 2A_{t,i}) \eta_{t,i} \theta_{t,i}^u \right] \right)^2 \right]$. Therefore, using (87),

$$\sigma_t^2 \leq \frac{\gamma^2}{c_m^2 b_0^2} \left(\sum_{i=1}^d \eta_{t,i}^2 \right) \left(\sum_{i=1}^d \mathbb{E}_{\xi_t} [(\theta_{t,i}^u)^2] \right) \stackrel{(91)}{\leq} \frac{4\gamma^2 L\Delta}{c_m^2 b_0^2} \sum_{i=1}^d \mathbb{E} [(\theta_{t,i}^u)^2]. \quad (96)$$

Hence, one can apply Bernstein's inequality (Lemma A.5) with $G = \frac{7\Delta^2}{480 \ln \frac{4(K+1)}{\delta}}$:

$$\begin{aligned} \mathbb{P} \left\{ \left| - \sum_{t=0}^{T-1} \sum_{i=1}^d \left[-(\gamma C_{t,i} - 2A_{t,i}) \eta_{t,i} \theta_{t,i}^u \right] \right| > \frac{\Delta}{4} \text{ and } \sum_{t=0}^{T-1} \sigma_t^2 \leq G \right\} &\leq 2 \exp \left(- \frac{\Delta^2}{16(2G + \frac{\Delta c}{6})} \right) \\ &= \frac{\delta}{2(K+1)}. \end{aligned}$$

Thus, we get

$$\mathbb{P} \left\{ \text{either } \left| \sum_{t=0}^{T-1} \sum_{i=1}^d \left[-(\gamma C_{t,i} - 2A_{t,i}) \eta_{t,i} \theta_{t,i}^u \right] \right| \leq \frac{\Delta}{4} \text{ or } \sum_{t=0}^{T-1} \sigma_t^2 > G \right\} \geq 1 - \frac{\delta}{2(K+1)}.$$

Moreover, event E_{T-1} implies

$$\begin{aligned} \sum_{t=0}^{T-1} \sigma_t^2 &\leq \sum_{t=0}^{T-1} \frac{4\gamma^2 L \Delta}{c_m^2 b_0^2} \sum_{i=1}^d \mathbb{E} \left[(\theta_{t,i}^u)^2 \right] \stackrel{(95)}{\leq} \frac{72\gamma^2 \lambda^{2-\alpha} L \Delta T}{c_m^2 b_0^2} \sum_{i=1}^d \sigma_i^\alpha \\ &\stackrel{(90)}{=} \frac{72(1-\beta_1)^{1-\frac{\alpha}{2}} c_m^{2-\alpha} \gamma^\alpha b_0^{2-\alpha} \sqrt{\Delta}^{2-\alpha} (K+1)^{\frac{\alpha^2-3\alpha+2}{3\alpha-2}} L \Delta T}{20^{2-\alpha} c_m^2 b_0^2 \sqrt{L}^{2-\alpha} \sqrt{d}^{2-\alpha} \ln^{2-\alpha} \frac{4(K+1)}{\delta}} \sum_{i=1}^d \sigma_i^\alpha \stackrel{(89)}{\leq} \frac{7\Delta^2}{480 \ln \frac{4(K+1)}{\delta}}. \end{aligned}$$

Bound for ②. Here we also apply Bernstein's inequality. One can check that terms into ② are unbiased and bounded:

$$\mathbb{E}_{\xi_t} \left[\sum_{i=1}^d \left[2A_{t,i} \left[(\theta_{t,i}^u)^2 - \mathbb{E}_{\xi_t} [(\theta_{t,i}^u)^2] \right] \right] \right] = 0$$

because of the definition of $\theta_{t,i}^u$ and

$$\left| \sum_{i=1}^d \left[2A_{t,i} \left[(\theta_{t,i}^u)^2 - \mathbb{E}_{\xi_t} (\theta_{t,i}^u)^2 \right] \right] \right| \leq \frac{4dL\gamma^2\lambda^2}{c_m^2 b_0^2 (1-\beta_1)} \leq \frac{\Delta}{100 \ln \left(\frac{4(K+1)}{\delta} \right)} \leq \frac{15\Delta}{47 \ln \left(\frac{4(K+1)}{\delta} \right)} = c.$$

What is more, let us define $\hat{\sigma}_t = \mathbb{E}_{\xi_t} \left[\left(\sum_{i=1}^d \left[2A_{t,i} \left[(\theta_{t,i}^u)^2 - \mathbb{E}_{\xi_t} (\theta_{t,i}^u)^2 \right] \right] \right)^2 \right]$. Hence, we get

$$\hat{\sigma}_t \leq c \mathbb{E}_{\xi_t} \left[\sum_{i=1}^d \left[2A_{t,i} \left| (\theta_{t,i}^u)^2 - \mathbb{E}_{\xi_t} (\theta_{t,i}^u)^2 \right| \right] \right] \leq \frac{4L\gamma^2 c}{c_m^2 b_0^2 (1-\beta_1)} \sum_{i=1}^d \mathbb{E}_{\xi_t} \left[(\theta_{t,i}^u)^2 \right].$$

Therefore, one can apply Bernstein's inequality (see Lemma A.5) with $G = \frac{7\Delta^2}{1504 \ln \left(\frac{4(K+1)}{\delta} \right)}$:

$$\begin{aligned} \mathbb{P} \left\{ \left| \sum_{t=0}^{T-1} \sum_{i=1}^d \left[2A_{t,i} \left[(\theta_{t,i}^u)^2 - \mathbb{E}_{\xi_t} [(\theta_{t,i}^u)^2] \right] \right] \right| > \frac{\Delta}{4} \text{ and } \sum_{t=0}^{T-1} \hat{\sigma}_t^2 \leq G \right\} &\leq 2 \exp \left(-\frac{\Delta^2}{16 (2G + \frac{\Delta c}{6})} \right) \\ &= \frac{\delta}{2(K+1)}. \end{aligned}$$

Thus, we get

$$\mathbb{P} \left\{ \text{either } \left| \sum_{t=0}^{T-1} \sum_{i=1}^d \left[2A_{t,i} \left[(\theta_{t,i}^u)^2 - \mathbb{E}_{\xi_t} [(\theta_{t,i}^u)^2] \right] \right] \right| \leq \frac{\Delta}{4} \text{ or } \sum_{t=0}^{T-1} \hat{\sigma}_t^2 > G \right\} \geq 1 - \frac{\delta}{2(K+1)}.$$

Moreover, event E_{T-1} implies

$$\begin{aligned} \sum_{t=0}^{T-1} \hat{\sigma}_t^2 &\stackrel{(95)}{\leq} \frac{72L\gamma^2 c \lambda^{2-\alpha} T}{c_m^2 b_0^2 (1-\beta_1)} \sum_{i=1}^d \sigma_i^\alpha \stackrel{(90)}{\leq} \frac{72c\gamma^\alpha \sqrt{\Delta}^{2-\alpha} (K+1)^{\frac{\alpha^2-3\alpha+2}{3\alpha-2}} LT}{20^{2-\alpha} (1-\beta_1)^{\frac{\alpha}{2}} c_m^\alpha b_0^\alpha \sqrt{d}^{2-\alpha} \sqrt{L}^{2-\alpha} \ln^{2-\alpha} \frac{4(K+1)}{\delta}} \sum_{i=1}^d \sigma_i^\alpha \\ &\stackrel{(89)}{\leq} \frac{7\Delta c}{480} = \frac{7\Delta^2}{1504 \ln \left(\frac{4(K+1)}{\delta} \right)}. \end{aligned}$$

Bound for ③. For the third term, we have that the event E_{T-1} implies

$$\begin{aligned} \sum_{t=0}^{T-1} \sum_{i=1}^d \left[2A_{t,i} \mathbb{E}_{\xi_t} (\theta_{t,i}^u)^2 \right] &\stackrel{(95)}{\leq} \frac{36L\gamma^2\lambda^{2-\alpha}T}{c_m^2 b_0^2(1-\beta_1)} \sum_{i=1}^d \sigma_i^\alpha \\ &\stackrel{(90)}{=} \frac{36L\gamma^\alpha \sqrt{\Delta}^{2-\alpha} (K+1)^{\frac{\alpha^2-3\alpha+2}{3\alpha-2}} T}{20^{2-\alpha}(1-\beta_1)^{\frac{\alpha}{2}} c_m^\alpha b_0^\alpha \sqrt{d}^{2-\alpha} \sqrt{L}^{2-\alpha} \ln^{2-\alpha} \left(\frac{4(K+1)}{\delta} \right)} \sum_{i=1}^d \sigma_i^\alpha \\ &\stackrel{(89)}{\leq} \frac{\Delta}{4}. \end{aligned}$$

Bound for ④. Similarly to the previous bound, E_{T-1} implies

$$\begin{aligned} \sum_{t=0}^{T-1} \sum_{i=1}^d \left[\gamma C_{t,i} (\theta_{t,i}^b)^2 \right] &\stackrel{(94)}{\leq} \frac{4^\alpha \gamma T}{c_m b_0 \lambda^{2\alpha-2}} \sum_{i=1}^d \sigma_i^{2\alpha} \\ &\stackrel{(90)}{\leq} \frac{4^\alpha \gamma (K+1)}{c_m b_0} \sum_{i=1}^d \sigma_i^{2\alpha} \cdot \frac{20^{2\alpha-2} \sqrt{d}^{2\alpha-2} \sqrt{L}^{2\alpha-2} \gamma^{2\alpha-2} \ln^{2\alpha-2} \left(\frac{4(K+1)}{\delta} \right)}{(1-\beta_1)^{\alpha-1} c_m^{2\alpha-2} b_0^{2\alpha-2} \sqrt{\Delta}^{2\alpha-2} (K+1)^{\frac{(1-\alpha)(2\alpha-2)}{3\alpha-2}}} \\ &= \frac{4^\alpha (K+1)}{c_m b_0} \sum_{i=1}^d \sigma_i^{2\alpha} \cdot \frac{20^{2\alpha-2} \sqrt{d}^{2\alpha-2} \sqrt{L}^{2\alpha-2} \ln^{2\alpha-2} \left(\frac{4(K+1)}{\delta} \right)}{(1-\beta_1)^{\alpha-1} c_m^{2\alpha-2} b_0^{2\alpha-2} \sqrt{\Delta}^{2\alpha-2} (K+1)^{\frac{(1-\alpha)(2\alpha-2)}{3\alpha-2}}} \cdot \gamma^{2\alpha-1} \\ &\stackrel{(89)}{\leq} \frac{\Delta}{4}. \end{aligned}$$

Therefore, the event $E_{T-1} \cap E_1 \cap E_2$ implies

$$\Delta_T \leq \Delta + \Delta = 2\Delta,$$

where

$$\begin{aligned} E_1 &= \left\{ \text{either } \left| \sum_{t=0}^{T-1} \sum_{i=1}^d \left[-(\gamma C_{t,i} - 2A_{t,i}) \eta_{t,i} \theta_{t,i}^u \right] \right| \leq \frac{\Delta}{4} \text{ or } \sum_{t=0}^{T-1} \sigma_t^2 > \frac{7\Delta^2}{480 \ln \frac{4(K+1)}{\delta}} \right\} \\ E_2 &= \left\{ \text{either } \left| \sum_{t=0}^{T-1} \sum_{i=1}^d \left[2A_{t,i} \left[(\theta_{t,i}^u)^2 - \mathbb{E}_{\xi_t} \left[(\theta_{t,i}^u)^2 \right] \right] \right] \right| \leq \frac{\Delta}{4} \text{ or } \sum_{t=0}^{T-1} \hat{\sigma}_t^2 > \frac{7\Delta^2}{1504 \ln \left(\frac{4(K+1)}{\delta} \right)} \right\}. \end{aligned}$$

Similarly to Theorem C.5, one can obtain

$$\mathbb{P}\{E_k\} \geq 1 - \frac{k\delta}{K+1}$$

for all $k = 0, \dots, K+1$. Consequently, E_{K+1} implies

$$\sum_{k=0}^K \sum_{i=1}^d \frac{\gamma C_{k,i}}{2} \nabla_{k,i}^2 \leq 2\Delta$$

with probability at least $1 - \delta$. Hence, with (87) we get

$$\sum_{k=0}^K \sum_{i=1}^d \nabla_{k,i}^2 = \sum_{k=0}^K \|\nabla f(x_k)\|^2 \leq \frac{4\Delta}{\gamma(1-\beta_1)} \max_{k \in \{0, \dots, K\}, i \in \{1, \dots, d\}} b_{k,i}. \quad (97)$$

What is more,

$$\begin{aligned} \max_{k \in \{0, \dots, K\}, i \in \{1, \dots, d\}} b_{k,i}^2 &\leq b_0^2 + \eta_m \sum_{k=0}^K \left(3 \sum_{i=1}^d \nabla_{k,i}^2 + 3 \sum_{i=1}^d \left((\theta_{k,i}^u)^2 - \mathbb{E}_{\xi_t} \left[(\theta_{k,i}^u)^2 \right] \right) \right. \\ &\quad \left. + 3 \sum_{i=1}^d \mathbb{E}_{\xi_t} \left[(\theta_{k,i}^u)^2 \right] + 3 \sum_{i=1}^d (\theta_{k,i}^b)^2 \right), \end{aligned}$$

where $\eta_m = \eta$ for Clip-M-AdaGradD and $\eta_m = \frac{\eta}{K+1}$ for Clip-AdamD. Also the event E_{K+1} implies

$$\begin{aligned} \sum_{k=0}^K \sum_{i=1}^d \left((\theta_{k,i}^u)^2 - \mathbb{E}_{\xi_t} \left[(\theta_{k,i}^u)^2 \right] + \mathbb{E}_{\xi_t} \left[(\theta_{k,i}^u)^2 \right] \right) &\leq \frac{\Delta c_m^2 b_0^2 (1 - \beta_1)}{4L\gamma^2}, \\ \sum_{k=0}^K \sum_{i=1}^d (\theta_{k,i}^b)^2 &\leq \frac{\Delta c_m b_0}{4\gamma} \end{aligned}$$

due to the bounds ②, ③ and ④ (exchanging $b_{t,i}$ to $c_m b_0$ in these terms allows to obtain the *same* bounds for them). Therefore,

$$\max_{k \in \overline{0, K}, i \in \overline{1, d}} b_{k,i}^2 \leq b_0^2 + 3\eta_m \sum_{k=0}^K \|\nabla f(x_k)\|^2 + \frac{3\eta_m c_m b_0 \Delta}{4\gamma} + \frac{3\eta_m c_m^2 b_0^2 (1 - \beta_1)}{4L\gamma^2}. \quad (98)$$

Denoting the left-hand side of (97) as S_K , squaring both sides and substituting (98) gives the quadratic inequality on S_K . Solving it, we have

$$\begin{aligned} S_K &\leq \frac{24\Delta^2 \eta_m}{\gamma^2 (1 - \beta_1)^2} + \sqrt{\frac{24^2 \Delta^4 \eta_m^2}{\gamma^4 (1 - \beta_1)^4} + \frac{16b_0^2 \Delta^2}{\gamma^2 (1 - \beta_1)^2} + 16 \left(\frac{3\eta_m c_m b_0 \Delta^3}{4\gamma^3 (1 - \beta_1)^2} + \frac{3\eta_m c_m^2 b_0^2 \Delta^3}{4L\gamma^4 (1 - \beta_1)} \right)} \\ &\leq 4 \max \left\{ \frac{48\Delta^2 \eta_m}{\gamma^2 (1 - \beta_1)^2}, \frac{4b_0 \Delta}{\gamma (1 - \beta_1)}, 4\sqrt{\frac{\eta_m c_m b_0 \Delta^3}{\gamma^3 (1 - \beta_1)^2}}, 4\sqrt{\frac{\eta_m c_m^2 b_0^2 \Delta^3}{L\gamma^4 (1 - \beta_1)}} \right\} \\ &\leq \frac{16\Delta}{\gamma} \max \left\{ \frac{12\Delta \eta_m}{\gamma}, \frac{b_0}{1 - \beta_1}, \sqrt{\frac{\eta_m c_m b_0 \Delta}{\gamma (1 - \beta_1)^2}}, \sqrt{\frac{\eta_m c_m^2 b_0^2 \Delta}{L\gamma^2 (1 - \beta_1)}} \right\}. \end{aligned}$$

□

After substitution of η_m and division by $K + 1$, similar to the Theorem C.5, we conclude the final result:

$$\begin{aligned} &\frac{1}{K+1} \sum_{k=0}^K \|\nabla f(x_k)\|^2 \\ &= \mathcal{O} \left(\max \left\{ \frac{\sqrt{d} L \Delta \ln \frac{K+1}{\delta}}{(1 - \beta_1)^3 (K+1)^{\frac{2\alpha-1}{3\alpha-2}}}, \frac{\sqrt{L\Delta} \left(\sum_{i=1}^d \sigma_i^\alpha \right)^{\frac{1}{\alpha}} \ln^{\frac{\alpha-1}{\alpha}} \frac{K+1}{\delta}}{(1 - \beta_1)^{\frac{3}{2}} \sqrt{d}^{\frac{2-\alpha}{\alpha}} (K+1)^{\frac{2\alpha-2}{3\alpha-2}}}, \right. \right. \\ &\quad \left. \left. \frac{d^{\frac{\alpha-1}{2\alpha-1}} \left(\sum_{i=1}^d \sigma_i^{2\alpha} \right)^{\frac{1}{2\alpha-1}} (L\Delta)^{\frac{\alpha-1}{2\alpha-1}} \ln^{\frac{2\alpha-2}{2\alpha-1}} \frac{K+1}{\delta}}{(1 - \beta_1)^{\frac{\alpha-1}{2\alpha-1}} (K+1)^{\frac{2\alpha-2}{3\alpha-2}}} \right\} \right) \end{aligned}$$

with probability at least $1 - \delta$.

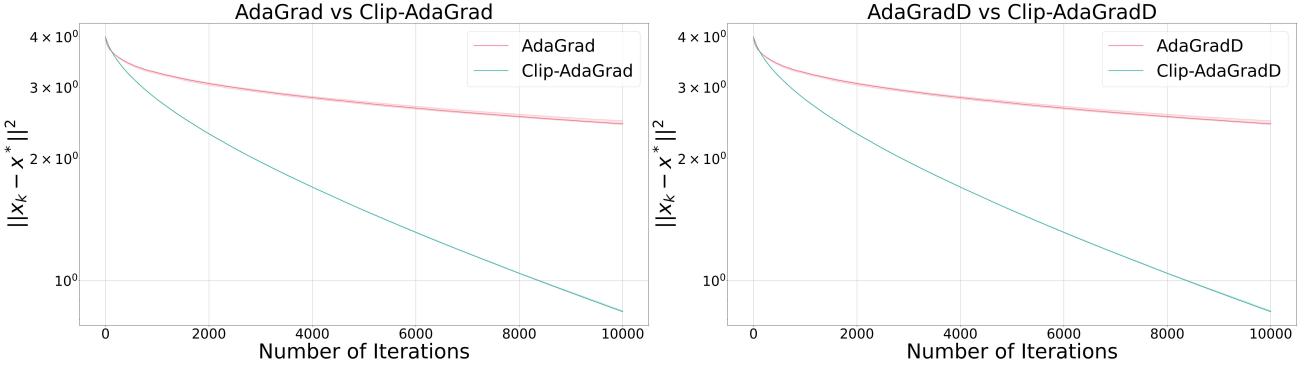


Figure 3: Performance of different versions of **AdaGrad** (with and without clipping/delay) with stepsize $\gamma = 1/128$ on the quadratic problem.

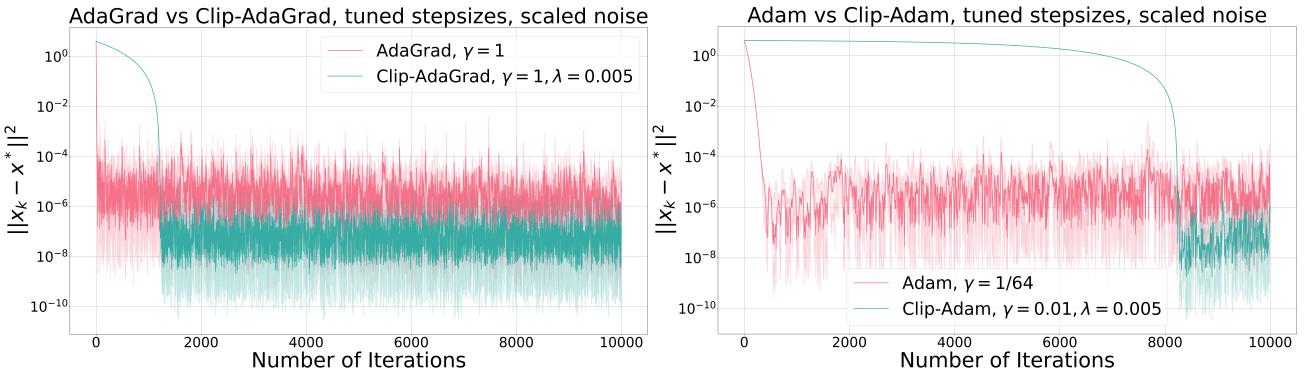


Figure 4: Performance of different versions of **AdaGrad** and **Adam** (with and without clipping) on the quadratic problem with scaled noise.

D. Numerical Experiments: Additional Details and Results

D.1. Quadratic Problem

In addition to the results provided in the main text, we compare the performance of different versions of **AdaGrad** with $\gamma = 1/128$. The results are given in Figure 3. One can notice that methods with clipping consistently outperform the methods without clipping for this stepsize as well.

Moreover, we provide the results of similar experiments for **Adam** with and without clipping/delay in Figure 5 (for $\beta_1 = 0.9$ and $\beta_2 = 0.999$). In general, the observed results for **Adam**-based methods are very similar to the ones obtained for **AdaGrad**: clipped versions of **Adam** show better high-probability convergence than non-clipped ones.

We also run **AdaGrad** and **Adam** and their clipped analogs for the same problem with scaled noise, that is, instead of ξ defined in Section 4, we use $\xi/100$. The clipping level is chosen 100 times smaller as well, i.e., $\lambda = 0.005$. Stepsizes were tuned for each method: for **AdaGrad**, we tried $\gamma = 1, \frac{1}{16}, \frac{1}{20}, \frac{1}{64}, \frac{1}{100}$; for **Clip-AdaGrad**, we tried $\gamma = 1, \frac{1}{16}, \frac{1}{64}$; for **Adam**, we tried $\gamma = \frac{1}{16}, \frac{1}{20}, \frac{1}{64}, \frac{1}{100}$; for **Clip-Adam**, we tried $\gamma = \frac{1}{16}, \frac{1}{20}, \frac{1}{64}, \frac{1}{100}$. The results are reported in Figure 4. As in the original experiments, the methods with clipping achieve a better optimization error with high probability.

D.2. ALBERT Base v2 Fine-tuning

Further details. In our experiments with finetuning of the ALBERT Base v2 model on CoLa and RTE datasets, we follow a standard practice of usage **Adam**, we apply bias correction to **Adam** and **Clip-Adam**. For the delayed version – **Clip-AdamD** – we do not apply bias correction and tune b_0 instead.

We used linear warmup with warmup ratio being 0.1, and hyperparameters were $\beta_1 = 0.9, \beta_2 = 0.999, b = \epsilon\mathbf{1}$, where

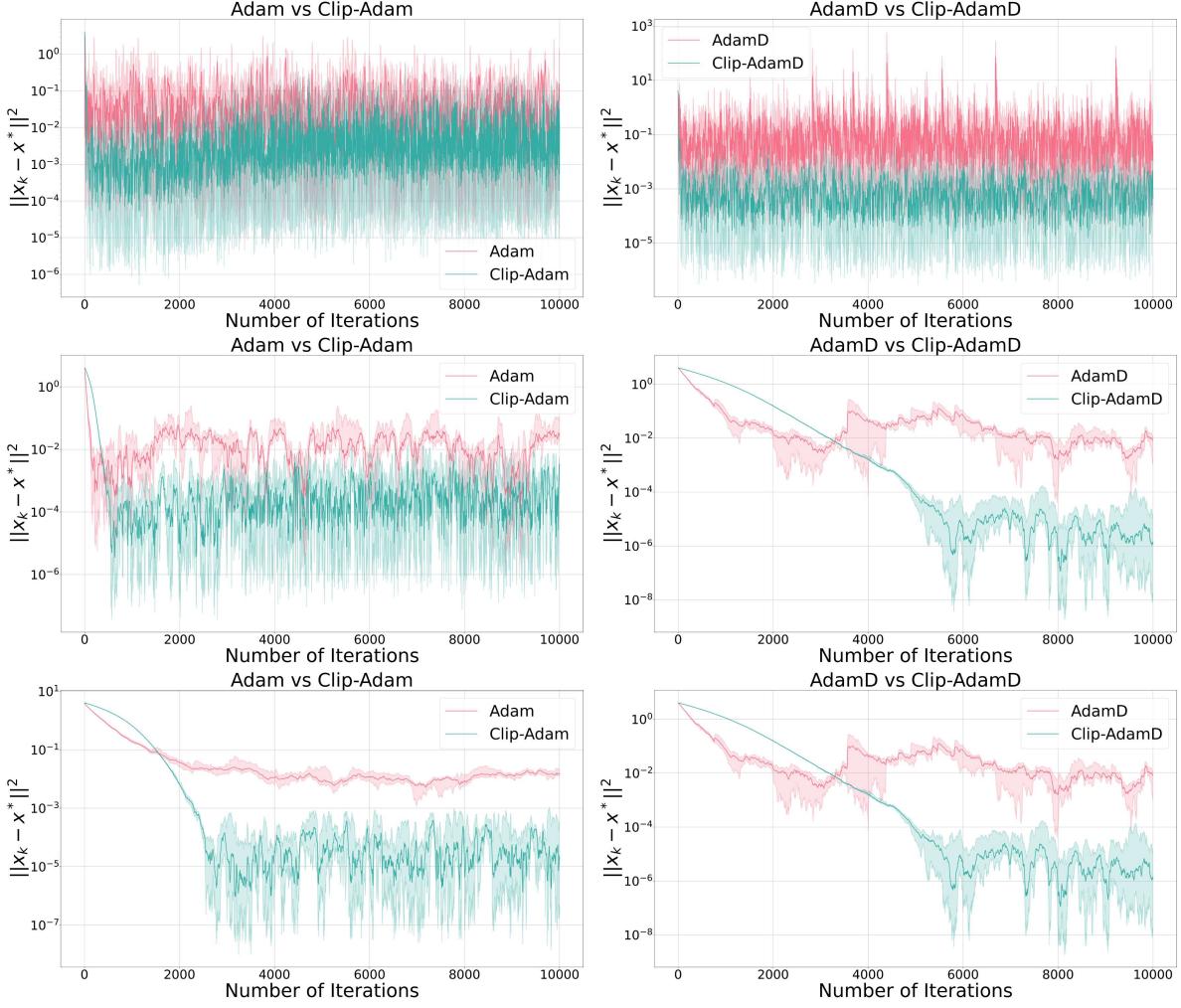


Figure 5: Performance of different versions of Adam (with and without clipping/delay) under the standard setting ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with stepsizes $\gamma = 1$ (first row) and $\gamma = \frac{1}{16}$ (second row) on the quadratic problem.

$\mathbf{1} = (1, 1, \dots, 1)^\top \in \mathbb{R}^d$. We tuned batchsize and stepsize γ for Adam and selected best values from $\{4, 8, 16, 32\}$ for the batchsize and from $\{10^{-6}, 3 \cdot 10^{-6}, 10^{-5}, 3 \cdot 10^{-5}, 10^{-4}\}$ for γ . For the CoLa dataset, the best batchsize was 16 and $\gamma = 10^{-5}$, and for the RTE dataset, the best batchsize was 8 and $\gamma = 10^{-5}$. We tested coordinate-wise clipping with $\lambda \in \{0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1\}$ and layer-wise clipping with $\lambda \in \{0.1, 0.2, 0.5, 1, 2, 5, 10\}$. For the CoLa dataset, the best results were achieved with $\lambda = 1$ for layer-wise clipping and $\lambda = 0.02$ for coordinate-wise clipping, and for the RTE dataset, the best results were achieved with $\lambda = 2$ for layer-wise clipping and $\lambda = 0.005$ for coordinate-wise clipping.

Noise histograms. The histograms are provided in Figure 6, where we additionally estimate the mean and standard deviation and plot the density of the normal distribution with these parameters (black curve). For the CoLa dataset, the noise distribution changes significantly after the start of the training, and its mean drifts to the right. However, the standard deviation does not change significantly, and, more importantly, metrics ρ_{mR} and ρ_{eR} remain quite large, showing that the distribution is significantly heavy-tailed. In contrast, for the RTE dataset, the noise distribution does not drift significantly, and, interestingly, ρ_{eR} decreases towards the end of training and becomes zero, while ρ_{mR} stays in the interval $[5, 10]$. Therefore, the noise distribution has much heavier tails for CoLa than for RTE.

Additional results. In the main part of our work, we present the results for Clip-Adam with layer-wise clipping. In Figure 7, we provide the results in the case of coordinate-wise clipping. In general, they are quite similar to the ones given

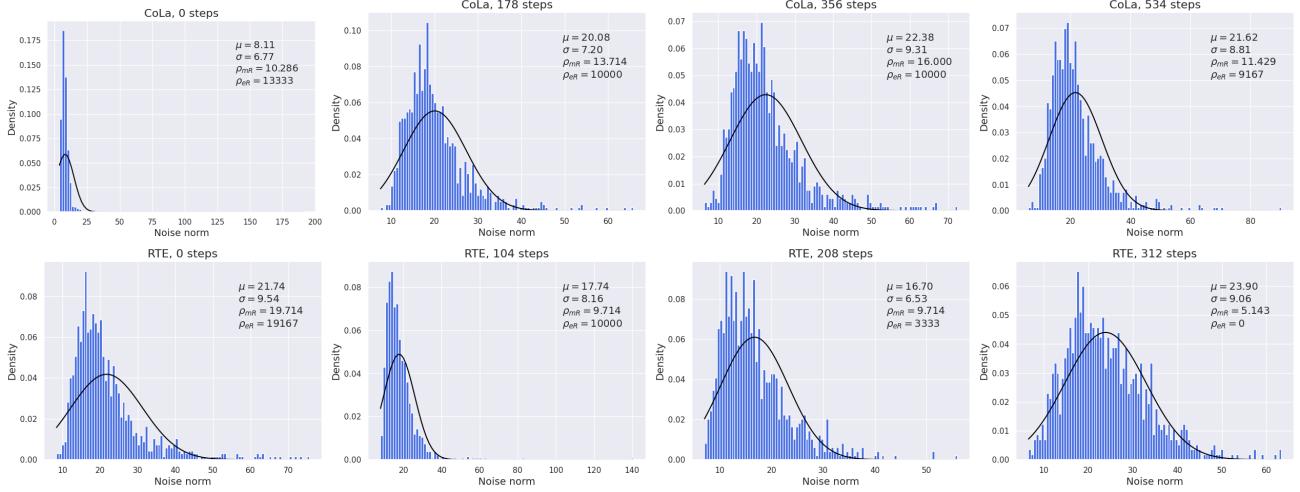


Figure 6: Gradient noise evolution for **Adam** on CoLa (the first row) and RTE (the second row) datasets. Histograms were evaluated after 0 steps, after $\approx 1/3$ and $\approx 2/3$ of all steps, and in the end.

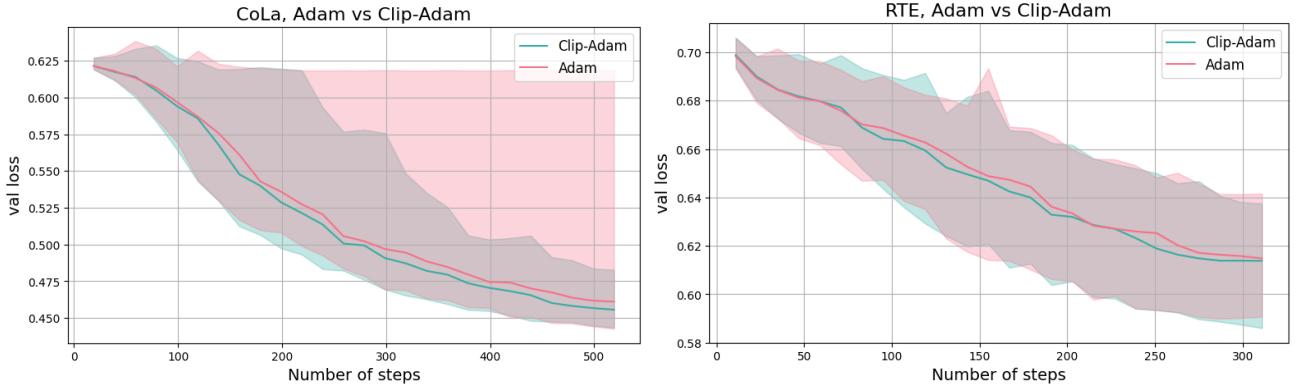


Figure 7: Validation loss for ALBERT Base v2 fine-tuning task on the CoLa and RTE datasets. **Clip-Adam** is used with coordinate-wise clipping ($\lambda = 0.02$ for CoLa and $\lambda = 0.005$ for RTE).

in Figure 2, indicating that both clipping strategies can be useful in practice and improve the high-probability convergence of **Adam**.

We also conducted experiments with **Clip-AdamD** and compared its performance with **Clip-Adam**. We tuned parameter ϵ defining $b = \epsilon \mathbf{1}$, where $\mathbf{1} = (1, 1, \dots, 1)^\top \in \mathbb{R}^d$. Tuning was performed in two phases: during the first phase, we selected the best values of ϵ from $\{10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$, and then for every selected $\hat{\epsilon}$ we tried $\epsilon \in \{0.2\hat{\epsilon}, 0.5\hat{\epsilon}, 0.8\hat{\epsilon}, 2\hat{\epsilon}, 5\hat{\epsilon}, 8\hat{\epsilon}\}$. In the case of CoLa dataset, the best ϵ was $2 \cdot 10^{-6}$, and in the case of RTE dataset, the best ϵ was $2 \cdot 10^{-6}$.

The results are presented⁸ in Figure 8 and show that **Clip-AdamD** performs worse than **Clip-Adam**, especially on CoLa dataset. However, it is worth mentioning that the clipping level was selected the same for both **Clip-Adam** and **Clip-AdamD**. Moreover, we have not tried to use bias correction for **Clip-AdamD** that could also improve its performance. Finally, the tuning of ϵ parameter over multiple runs can also improve the result of **Clip-AdamD**.

Finally, we also conducted similar experiments with **AdaGrad**-based methods with and without clipping/delay. Parameter γ and batchsize were tuned across the same values as in the case of **Adam**. Moreover, similarly to the experiments with **Adam**, we used standard layer-wise clipping for **AdaGrad**-based methods since it gave better results. The final parameters are (i)

⁸In the plots, we use the name **Clip-RAdamD**, which is equivalent to **Clip-AdamD** as explained at the beginning of Appendix C.

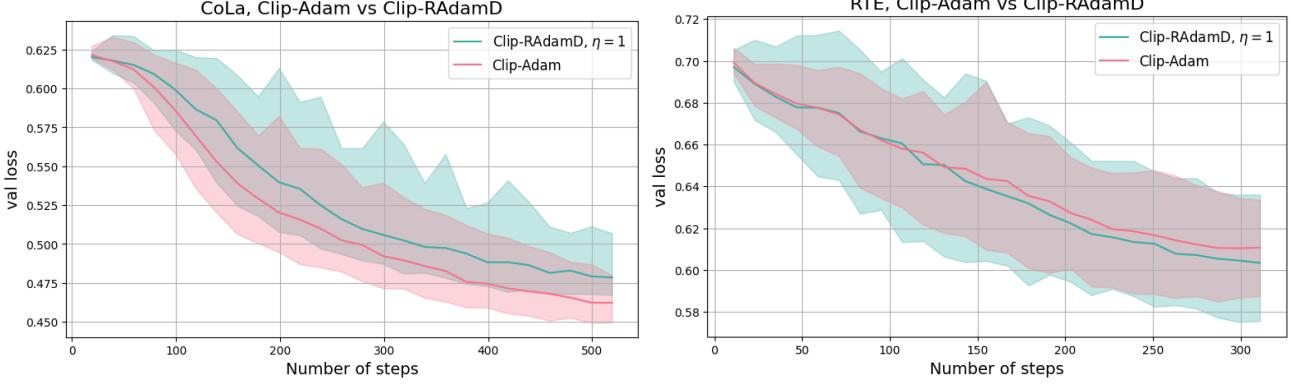


Figure 8: Validation loss for ALBERT Base v2 fine-tuning task on the CoLa and RTE datasets.

$\gamma = 10^{-4}$, batchsize 4, $\lambda = 5$ for **(Clip-)AdaGrad** on CoLa dataset, (ii) $\gamma = 10^{-4}$, batchsize 16, $\lambda = 1$ for **(Clip-)AdaGrad** on RTE dataset, (iii) $\gamma = 10^{-4}$, batchsize 4, $\lambda = 5$ for **(Clip-)AdaGradD** on CoLa dataset, and (iv) $\gamma = 10^{-4}$, batchsize 16, $\lambda = 0.1$ for **(Clip-)AdaGradD** on RTE dataset. The results are presented in Figure 9. For this particular case, there is no big difference between versions of **AdaGrad** with and without clipping, and only for CoLa dataset we see that **Clip-AdaGrad** has much smaller error band than **AdaGrad**.

D.3. Scaling Up: Fine-Tuning of 355M Model

Setup. We replicate the setup from Appendix D.2 for fine-tuning the 355M parameter RoBERTa Large model (Liu et al., 2019) on the two GLUE (Wang et al., 2018) datasets: QNLI (116k question-answer pairs) and CoLa (10.7k linguistic acceptability examples). Keeping identical hyperparameters, including the learning rate scheduler, warmup ratio, and optimizer parameters (β_1, β_2), we employ a moderate batch size of 16 for both datasets to ensure comparability with the previous finding. Through extensive testing of layer-wise clipping with $\lambda \in \{0.1, 0.2, 0.5, 1, 2, 5, 10\}$, we identified $\lambda = 1$ as the best value for both tasks, aligning with prior work specialized in fine-tuning of large language models (Yang & Ma, 2022). The best values that we have selected in this part are used to build noise histograms and compare algorithms with and without clipping.

Noise Histograms. We also build the noise histograms (see Figure 10) for the RoBERTa Large model. In the histogram, we quantitatively present the measure the heavy-tailedness of the noise following the recipe from (Gorbunov et al., 2022). After fine-tuning the model with checkpointing on the full dataset (QNLI or CoLa), we compute the true gradient $\nabla f(x)$ through many gradient accumulation steps. From identical initial checkpoints, we sample 1000 mini-batched gradient estimators for CoLa and 100 for QNLI (all of them are with the batch size of 16), calculating the norm differences $|\nabla f_\xi(x) - \nabla f(x)|$ for histogram construction. As in Section 4 and Appendix D.2, we assess the heavy-tailedness of the noise using the metrics p_{mR} and p_{eR} . Both metrics remain consistently large throughout training, indicating that the noise exhibits significant heavy-tailed behavior. Furthermore, the histogram profiles closely resemble the Lévy α -stable distribution, as shown in Figure 10.

Comparison of Methods With and Without Clipping. We replicate our earlier setup with one key modification: the number of random seeds is reduced from 100 to 5, due to computational constraints when scaling to a 355M-parameter model—particularly given QNLI’s substantially larger size compared to CoLa or RTE. At each step, we compute the median validation loss, along with the 5-th and 95-th percentiles.

The comparison between **Adam** and **Clip-Adam** is shown in Figure 11, while the comparison between **AdaGrad** and **Clip-AdaGrad** is presented in Figure 12. Notably, for the new model and across both datasets, the clipped variants consistently outperform their unclipped counterparts.

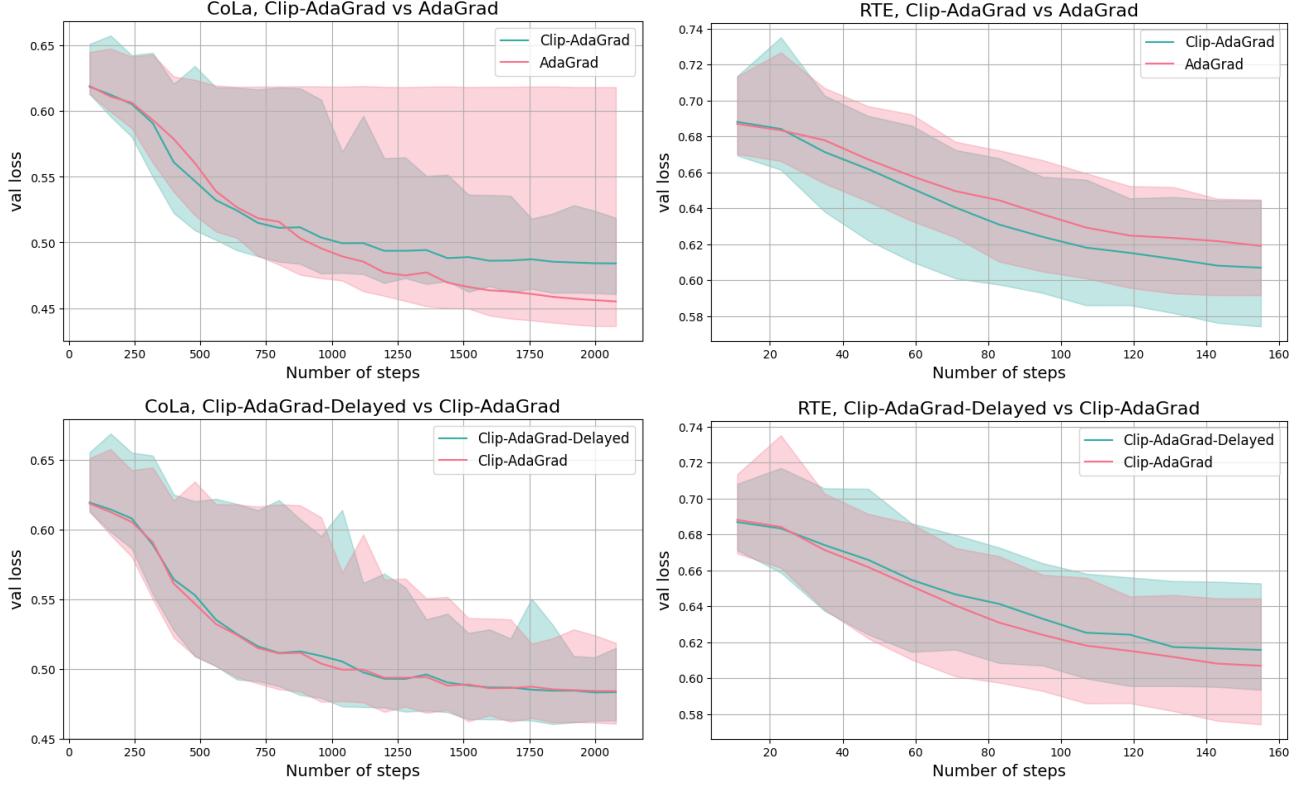


Figure 9: Validation loss for ALBERT Base v2 fine-tuning task on the CoLa and RTE datasets.

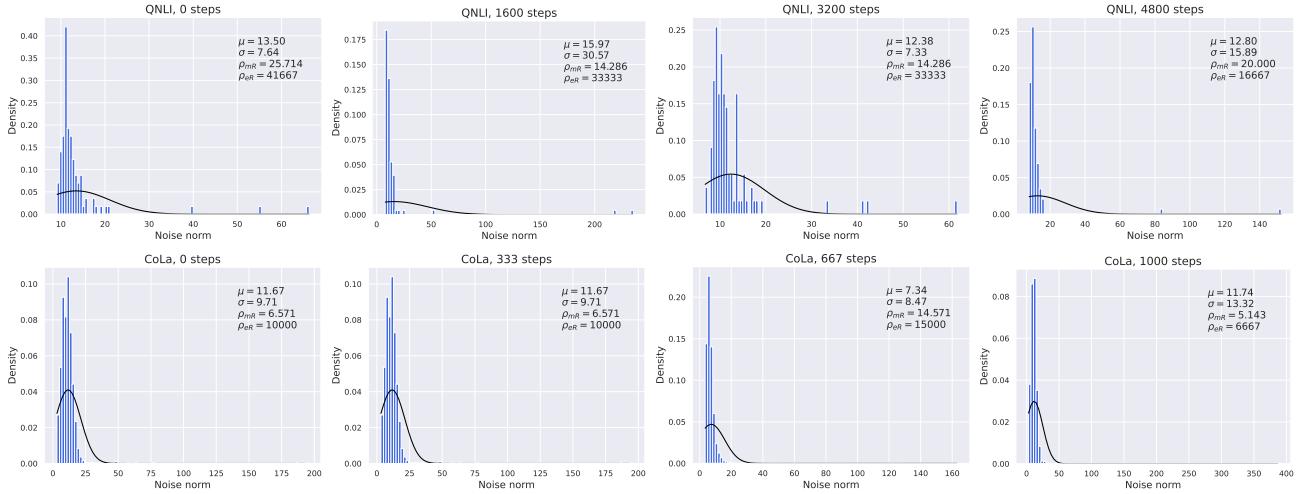


Figure 10: Gradient noise evolution for Adam on QNLI (the first row) and CoLa (the second row) datasets during RoBERTa Large fine-tuning. Histograms were evaluated after 0 steps, after $\approx 1/3$ and $\approx 2/3$ of all steps, and in the end.

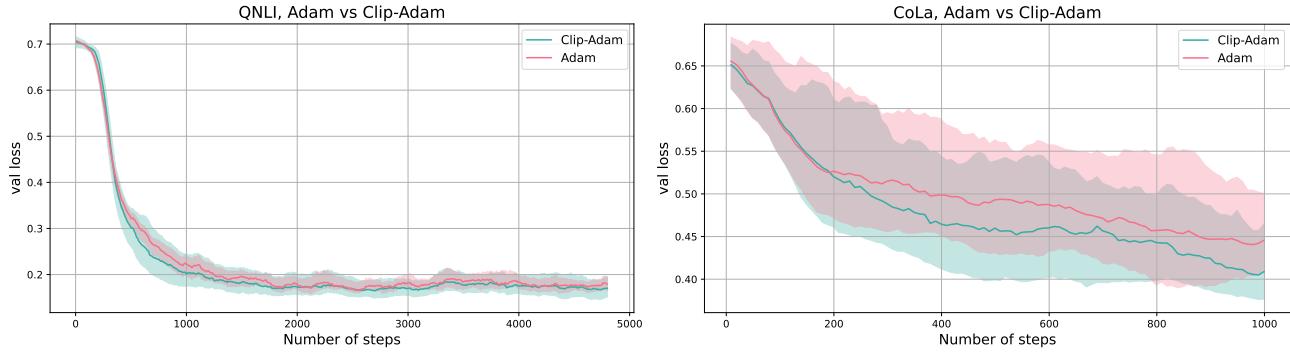


Figure 11: Validation loss for RoBERTa Large fine-tuning task on the QNLI and CoLa datasets. Clip-Adam is used with layer-wise clipping.

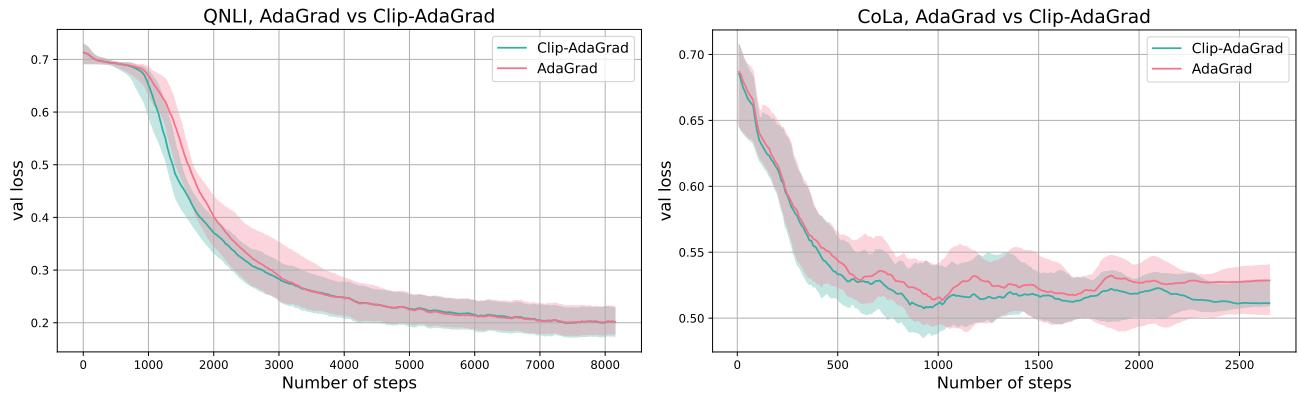


Figure 12: Validation loss for RoBERTa Large fine-tuning task on the QNLI and CoLa datasets. Clip-AdaGrad is used with layer-wise clipping.