High-Probability Bounds for Stochastic Optimization and Variational Inequalities: the Case of Unbounded Variance

Abdurakhmon Sadiev [1] **Marina Danilova** [2] **Eduard Gorbunov** [3] **Samuel Horváth** [3] **Gauthier Gidel** [4 5]
Pavel Dvurechensky [6] **Alexander Gasnikov** [2 7 8] **Peter Richtárik** [1]

Abstract

During recent years the interest of optimization and machine learning communities in high-probability convergence of stochastic optimization methods has been growing. One of the main reasons for this is that high-probability complexity bounds are more accurate and less studied than in-expectation ones. However, SOTA high-probability non-asymptotic convergence results are derived under strong assumptions such as the boundedness of the gradient noise variance or of the objective's gradient itself. In this paper, we propose several algorithms with high-probability convergence results under less restrictive assumptions. In particular, we derive new high-probability convergence results under the assumption that the gradient/operator noise has bounded central α-th moment for $\alpha \in (1, 2]$ in the following setups: (i) smooth non-convex / Polyak-Łojasiewicz / convex / strongly convex / quasi-strongly convex minimization problems, (ii) Lipschitz / star-cocoercive and monotone / quasi-strongly monotone variational inequalities. These results justify the usage of the considered methods for solving problems that do not fit standard functional classes studied in stochastic optimization.

[1]King Abdullah University of Science and Technology, KSA [2]Moscow Institute of Physics and Technology, Russia [3]Mohamed bin Zayed University of Artificial Intelligence, UAE [4]Université de Montréal and Mila, Canada [5]Canada CIFAR AI Chair [6]Weierstrass Institute for Applied Analysis and Stochastics, Germany [7]Skolkovo Institute of Science and Technology, Russia [8]Institute for Information Transmission Problems RAS, Russia. Correspondence to: Eduard Gorbunov <eduard.gorbunov@mbzuai.ac.ae>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

1. Introduction

Training of machine learning models is usually performed via stochastic first-order optimization methods, e.g., Stochastic Gradient Descent (SGD) (Robbins & Monro, 1951)

$$x^{k+1} = x^k - \gamma \nabla f_{\xi^k}(x^k), \tag{1}$$

where $\nabla f_{\xi^k}(x^k)$ represents the stochastic gradient of the objective/loss function f at point x^k. Despite numerous empirical studies and observations validating the good performance of such methods, it is also important for the field to understand their theoretical convergence properties, e.g., under what assumptions a method converges and what the rate is. However, since the methods of interest are stochastic, one needs to specify what type of convergence is considered before moving on to further questions.

Typically, the convergence of the stochastic methods is studied only in expectation, i.e., for some performance metric[1] $\mathcal{P}(x)$, upper bounds are derived for the number of iterations K needed to achieve $\mathbb{E}[\mathcal{P}(x^K)] \le \varepsilon$, where x^K is the output of the method after K steps, ε is an optimization error, and $\mathbb{E}[\cdot]$ is the full expectation. These bounds can be "blind" to some important properties like light-/heavy-tailedness of the noise distribution and, as a result, such guarantees do not accurately describe the methods' convergence in practice (Gorbunov et al., 2020). In contrast, high-probability convergence guarantees are more sensitive to the noise distribution and thus are more accurate. Such results provide upper bounds for the number of iterations K needed to achieve $\mathbb{P}\{\mathcal{P}(x^K) \le \varepsilon\} \ge 1 - \beta$ for some confidence level $\beta \in (0, 1]$, where $\mathbb{P}\{\cdot\}$ denotes some probability measure determined by a setup.

With the ultimate goal of bridging the theory and practice of stochastic methods, recent works on high-probability convergence guarantees (Nazin et al., 2019; Davis et al., 2021; Gorbunov et al., 2020; 2021; 2022a; Cutkosky & Mehta, 2021) focus on an important direction of the relaxing the assumptions under which these guarantees are derived.

[1]Examples of performance metrics for minimization of function f: $\mathcal{P}(x) = f(x) - f(x^*)$, $\mathcal{P}(x) = \|\nabla f(x)\|^2$, $\mathcal{P}(x) = \|x - x^*\|^2$, where $x^* \in \arg\min_{x \in \mathbb{R}^d} f(x)$.

Our paper further extensively complements this line of works in two main aspects: for a plethora of settings, we derive new high-probability results allowing the variance of the noise and the gradient of the objective to be unbounded.

1.1. Technical Preliminaries

Before we move on to the main part of the paper, we introduce the problems considered in the work and all necessary preliminaries. In particular, we consider stochastic unconstrained optimization problems

$$\min_{x\in\mathbb{R}^d} \left\{ f(x) = \mathbb{E}_{\xi\sim\mathcal{D}}\left[f_\xi(x)\right]\right\}, \tag{2}$$

where ξ is a random variable with distribution $\mathcal{D}$. Such problems often arise in machine learning, where $f_\xi(x)$ represents the loss function on the data sample ξ (Shalev-Shwartz & Ben-David, 2014).

Another class of problems that we consider this work is unconstrained variational inequality problems (VIP), i.e., non-linear equations (Harker & Pang, 1990; Ryu & Yin, 2021):

$$\text{find } x^* \in \mathbb{R}^d \text{ such that } F(x^*) = 0, \tag{3}$$

where $F(x) = \mathbb{E}_{\xi\sim\mathcal{D}}[F_\xi(x)]$. These problems arise in adversarial/game formulations of machine learning tasks (Goodfellow et al., 2014; Gidel et al., 2019).

Notation. We use standard notation: $\|x\| = \sqrt{\langle x, x\rangle}$ denotes the standard Euclidean norm in $\mathbb{R}^d$, $\mathbb{E}_\xi[\cdot]$ denotes an expectation w.r.t. the randomness coming from random variable ξ, $B_R(x) = \{y\in\mathbb{R}^d \mid \|y-x\|\leq R\}$ is a ball with center at x and radius R. We define restricted gap-function as $\texttt{Gap}_R(x) = \max_{y\in B_R(x^*)}\langle F(y), x-y\rangle$ – a standard convergence criterion for monotone VIP (Nesterov, 2007). Finally, $\mathcal{O}(\cdot)$ hides numerical factors and $\widetilde{\mathcal{O}}(\cdot)$ hides polylogarithmic and numerical factors.

Assumptions on a subset. Although we consider unconstrained problems, our analysis does not require any assumptions to hold on the whole space. For our purposes, it is sufficient to introduce all assumptions only on some subset of $\mathbb{R}^d$, since we prove that the considered methods do not leave some ball around the solution or some level-set of the objective function with high probability. This allows us to consider quite large classes of problems.

Stochastic oracle. We assume that at given point x we have an access to the unbiased stochastic oracle returning $\nabla f_\xi(x)$ or $F_\xi(x)$ that satisfy the following conditions.

Assumption 1.1. We assume that there exist some set $Q\subseteq\mathbb{R}^d$ and values $\sigma\geq 0$, $\alpha\in(1,2]$ such that for all $x\in Q$

(i) for problem (2) $\mathbb{E}_{\xi\sim\mathcal{D}}[\nabla f_\xi(x)] = \nabla f(x)$ and

$$\mathbb{E}_{\xi\sim\mathcal{D}}[\|\nabla f_\xi(x) - \nabla f(x)\|^\alpha] \leq \sigma^\alpha, \tag{4}$$

(ii) for problem (3) $\mathbb{E}_{\xi\sim\mathcal{D}}[F_\xi(x)] = F(x)$ and

$$\mathbb{E}_{\xi\sim\mathcal{D}}[\|F_\xi(x) - F(x)\|^\alpha] \leq \sigma^\alpha. \tag{5}$$

When $\alpha = 2$, the above assumption recovers the standard uniformly bounded variance assumption (Nemirovski et al., 2009; Ghadimi & Lan, 2012; 2013). However, Assumption 1.1 allows the variance of the estimator to be *unbounded* when $\alpha\in(1,2)$, i.e., the noise can follow some heavy-tailed distribution. For example, the distribution of the gradient noise in the training of large attention models resembles Lévy α-stable distribution with $\alpha < 2$ (Zhang et al., 2020b). There exist also other versions of Assumption 1.1, see (Patel et al., 2022).

Assumptions on f. We start with a very mild assumption since without it, problem (2) does not make sense.

Assumption 1.2. We assume that there exist some set $Q\subseteq\mathbb{R}^d$ such that f is uniformly lower-bounded on Q: $f_* = \inf_{x\in Q} f(x) > -\infty$.

Moreover, when working with minimization problems (2), we always assume smoothness of f.

Assumption 1.3. We assume that there exist some set $Q\subseteq\mathbb{R}^d$ and constant $L>0$ such that for all $x,y\in Q$

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|, \tag{6}$$

$$\|\nabla f(x)\|^2 \leq 2L\left(f(x) - f_*\right), \tag{7}$$

where $f_* = \inf_{x\in Q} f(x) > -\infty$.

We notice here that (7) follows from (6) for $Q=\mathbb{R}^d$, but in the general case, the implication is slightly more involved (see the details in Appendix B). When Q is a compact set, the function f is allowed to be non-L-smooth on the whole $\mathbb{R}^d$, which is related to local-Lipschitzness of the gradients (Patel et al., 2022; Patel & Berahas, 2022).

In each particular special case, we also make *one of the following assumptions* about the structured non-convexity of the objective function. The previous two assumptions hold for a very broad class of functions. The next assumption – Polyak-Łojasiewicz condition (Polyak, 1963; Lojasiewicz, 1963) – narrows the class of non-convex functions.

Assumption 1.4. We assume that there exist some set $Q\subseteq\mathbb{R}^d$ and constant $\mu>0$ such that f satisfies Polyak-Łojasiewicz (PŁ) condition/inequality on Q, i.e., for all $x\in Q$ and $x^* = \arg\min_{x\in\mathbb{R}^d} f(x)$

$$\|\nabla f(x)\|^2 \geq 2\mu\left(f(x) - f(x^*)\right). \tag{8}$$

When function f is μ-strongly convex, it satisfies PŁ condition. However, PŁ inequality can hold even for non-convex functions. Some analogs of this assumption have been observed for over-parameterized models (Liu et al., 2022).

We also consider another relaxation of convexity.

Assumption 1.5. We assume that there exist some set $Q \subseteq \mathbb{R}^d$ and constant $\mu \geq 0$ such that f is μ-quasi-strongly convex, i.e., for all $x \in Q$ and $x^* = \arg\min_{x\in\mathbb{R}^d} f(x)$

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

As PŁ condition, this assumption holds for any μ-strongly convex function but does not imply convexity. Nevertheless, for the above two assumptions, some standard deterministic methods such as Gradient Descent (GD) converge linearly; see more details and examples in (Necoara et al., 2019).

In the analysis of the accelerated method, we also need standard (strong) convexity.

Assumption 1.6. We assume that there exist some set $Q \subseteq \mathbb{R}^d$ and constant $\mu \geq 0$ such that f is μ-strongly convex, i.e., for all $x, y \in Q$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2. \quad (10)$$

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

Assumptions on F. In the context of solving (3), we assume Lipschitzness of F – a standard assumption for VIP.

Assumption 1.7. We assume that there exist some set $Q \subseteq \mathbb{R}^d$ and constant $L > 0$ such that for all $x, y \in Q$

$$\|F(x) - F(y)\| \leq L\|x - y\|, \quad (11)$$

Similarly to the case of minimization problems, we make *one or two of the following assumptions* about the structured non-monotonicity of the operator F. The first assumption we consider is the standard monotonicity.

Assumption 1.8. We assume that there exist some set $Q \subseteq \mathbb{R}^d$ such that F is monotone on Q, i.e., for all $x, y \in Q$

$$\langle F(x) - F(y), x - y\rangle \geq 0. \quad (12)$$

Monotonicity can be seen as an analog of convexity for VIP. When (12) holds with $\mu\|x - y\|^2$ in the r.h.s. instead of just 0, operator F is called μ-strongly monotone.

Next, we consider quasi-strong monotonicity (Mertikopoulos & Zhou, 2019; Song et al., 2020; Loizou et al., 2021) – a relaxation of strong monotonicity. There exist examples of non-monotone problems such that the assumption below holds (Loizou et al., 2021, Appendix A.6).

Assumption 1.9. We assume that there exist some set $Q \subseteq \mathbb{R}^d$ and constant $\mu > 0$ such that F is μ-quasi strongly monotone on Q, i.e., for all $x \in Q$ and x^* such that $F(x^*) = 0$ we have

$$\langle F(x), x - x^*\rangle \geq \mu\|x - x^*\|^2. \quad (13)$$

Another structured non-monotonicity assumption that we consider in this paper is star-cocoercivity.

Assumption 1.10. We assume that there exist some set $Q \subseteq \mathbb{R}^d$ and constant $\ell > 0$ such that F is star-cocoercive on Q, i.e., for all $x \in Q$ and x^* such that $F(x^*) = 0$

$$\|F(x)\|^2 \leq \ell\langle F(x), x - x^*\rangle. \quad (14)$$

This assumption can be seen as a relaxation of the standard cocoercivity: $\|F(x) - F(y)\|^2 \leq \ell\langle F(x) - F(y), x - y\rangle$. However, unlike cocoercivity, star-cocoercivity implies neither monotonicity nor Lipschitzness of operator F (Loizou et al., 2021, Appendix A.6).

1.2. Closely Related Works and Our Contributions

In this subsection, we overview closely related works and describe the contributions of our work. Additional related works are discussed in Appendix A.

Convex optimization and monotone VIPs. Classical high-probability results for (strongly) convex minimization (Nemirovski et al., 2009; Ghadimi & Lan, 2012) and monotone VIP (Juditsky et al., 2011) are derived under the so-called light-tails assumption, meaning that the noise in the stochastic gradients/operators is assumed to be sub-Gaussian: $\mathbb{E}_{\xi\sim\mathcal{D}}[\exp(\|\nabla f_\xi(x) - \nabla f(x)\|^2/\sigma^2)] \leq \exp(1)$ or $\mathbb{E}_{\xi\sim\mathcal{D}}[\exp(\|F_\xi(x) - F(x)\|^2/\sigma^2)] \leq \exp(1)$. In these settings, optimal (up to logarithmic factors) rates of convergence are derived in the mentioned papers.

The first high-probability results with logarithmic dependence[2] on $1/\beta$ under just bounded variance assumption are given by Nazin et al. (2019), where the authors show non-accelerated rates of convergence for a version of Mirror Descent with a special truncation operator for smooth convex and strongly convex problems defined on the bounded sets. Then, Davis et al. (2021) derive accelerated rates in the strongly convex case using robust distance estimation techniques. Gorbunov et al. (2020; 2021) propose an accelerated method with clipping for unconstrained (strongly) convex problems with Lipschitz / Hölder continuous gradients and derive the first high-probability results for clipped-SGD. In the context of VIP, Gorbunov et al. (2022a) derive the first high-probability results for the stochastic methods for solving VIP under bounded variance assumption and different assumptions on structured non-monotonicity.

[2]Note that from in-expectation convergence guarantee, one can always get a high-probability one using Markov's inequality. For example, under bounded variance, smoothness, and strong convexity assumptions SGD achieves $\mathbb{E}\|x^k - x^*\|^2 \leq \varepsilon$ after $k = \widetilde{\mathcal{O}}(\max\{L/\mu, \sigma^2/\mu\varepsilon\})$ iterations. Therefore, taking k such that $\mathbb{E}\|x^k - x^*\|^2 \leq \varepsilon\beta$ we get from Markov's inequality that $\mathbb{P}\{\|x^k - x^*\|^2 \leq \varepsilon\} \leq \beta$. However, in this case, we get bound $k = \widetilde{\mathcal{O}}(\max\{L/\mu, \sigma^2/\mu\varepsilon\beta\})$, having undesirable inverse-power dependence on β.

Table 1: Summary of known and new high-probability complexity results for solving smooth problem (2). Column "Setup" indicates the assumptions made in addition to Assumptions 1.1 and 1.3. All assumptions are made only on some ball around the solution with radius $\sim R \geq \|x^0 - x^*\|$ (unless the opposite is indicated). By the complexity we mean the number of stochastic oracle calls needed for a method to guarantee that $\mathbb{P}\{\text{Metric} \leq \varepsilon\} \geq 1 - \beta$ for some $\varepsilon > 0$, $\beta \in (0, 1]$ and "Metric" is taken from the corresponding column. For simplicity, we omit numerical and logarithmic factors in the complexity bounds. Column "α" shows the allowed values of α, "UD?" shows whether the analysis works on unbounded domains, and "UG?" indicates whether the analysis works without assuming boundedness of the gradient. Notation: L = Lipschitz constant; D = diameter of the domain (for the result from (Nazin et al., 2019)); σ = parameter from Assumption 1.1; R = any upper bound on $\|x^0 - x^*\|$; μ = (quasi-)strong convexity/Polyak-Łojasiewicz parameter; Δ = any upper bound on $f(x^0) - f_*$; G = parameter such that $\mathbb{E}_{\xi\sim\mathcal{D}}\|\nabla f_\xi(x)\|^\alpha \leq G^\alpha$ (for the result from (Cutkosky & Mehta, 2021)). The results of this paper are highlighted in blue.

Setup	Method	Citation	Metric	Complexity	α	UD?	UG?
As. 1.6 ($\mu = 0$)	RSMD	(Nazin et al., 2019)[(1)]	$f(\overline{x}^K) - f(x^*)$	$\max\left\{\frac{LD^2}{\varepsilon}, \frac{\sigma^2 D^2}{\varepsilon^2}\right\}$	2	✗	✓
	clipped-SGD	(Gorbunov et al., 2020) (Gorbunov et al., 2021)	$f(\overline{x}^K) - f(x^*)$	$\max\left\{\frac{LR^2}{\varepsilon}, \frac{\sigma^2 R^2}{\varepsilon^2}\right\}$	2	✓	✓
	clipped-SSTM	(Gorbunov et al., 2020) (Gorbunov et al., 2021)	$f(y^K) - f(x^*)$	$\max\left\{\sqrt{\frac{LR^2}{\varepsilon}}, \frac{\sigma^2 R^2}{\varepsilon^2}\right\}$	2	✓	✓
	clipped-SGD	Theorems 3.1 & E.6	$f(\overline{x}^K) - f(x^*)$	$\max\left\{\frac{LR^2}{\varepsilon}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}$	$(1, 2]$	✓	✓
	clipped-SSTM	Theorems 3.2 & F.2	$f(y^K) - f(x^*)$	$\max\left\{\sqrt{\frac{LR^2}{\varepsilon}}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}$	$(1, 2]$	✓	✓
As. 1.6 ($\mu > 0$)	restarted-RSMD	(Nazin et al., 2019)[(1)]	$f(\overline{x}^K) - f(x^*)$	$\max\left\{\frac{L}{\mu}, \frac{\sigma^2}{\mu\varepsilon}\right\}$	2	✗	✓
	proxBoost	(Davis et al., 2021)[(1)]	$f(\overline{x}^K) - f(x^*)$	$\max\left\{\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{\mu\varepsilon}\right\}$[(2)]	2	✓	✓
	R-clipped-SGD	(Gorbunov et al., 2020) (Gorbunov et al., 2021)	$f(\overline{x}^K) - f(x^*)$	$\max\left\{\frac{L}{\mu}, \frac{\sigma^2}{\mu\varepsilon}\right\}$	2	✓	✓
	R-clipped-SSTM	(Gorbunov et al., 2020) (Gorbunov et al., 2021)	$f(y^K) - f(x^*)$	$\max\left\{\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{\mu\varepsilon}\right\}$	2	✓	✓
	R-clipped-SSTM	Theorems 3.2 & F.3	$f(y^K) - f(x^*)$	$\max\left\{\sqrt{\frac{L}{\mu}}, \left(\frac{\sigma^2}{\mu\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right\}$	$(1, 2]$	✓	✓
As. 1.5 ($\mu > 0$)	clipped-SGD	Theorems 3.1 & E.8	$\|x^K - x^*\|^2$	$\max\left\{\frac{L}{\mu}, \left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right\}$	$(1, 2]$	✓	✓
As. 1.2	MSGD	(Li & Orabona, 2020)[(1)]	$\frac{1}{K+1}\sum\limits_{k=0}^{K}\|\nabla f(x^k)\|^2$	$\max\left\{\frac{L^2\Delta^2}{\varepsilon}, \frac{\sigma^4}{\varepsilon^2}\right\}$	✗[(3)]	✓	✓
	clipped-NMSGD	(Cutkosky & Mehta, 2021)[(1)]	$\left(\frac{1}{K+1}\sum\limits_{k=0}^{K}\|\nabla f(x^k)\|\right)^2$ [(4)]	$\left(\frac{G^2}{\varepsilon}\right)^{\frac{3\alpha-2}{2\alpha-2}}$	$(1, 2]$	✓	✗
	clipped-SGD	Theorems 3.1 & E.2 [(5)]	$\frac{1}{K+1}\sum\limits_{k=0}^{K}\|\nabla f(x^k)\|^2$	$\max\left\{\frac{L\Delta}{\varepsilon}, \left(\frac{\sqrt{L\Delta}\sigma}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}$	$(1, 2]$	✓	✓
As. 1.4	clipped-SGD	Theorems 3.1 & E.4 [(5)]	$f(x^K) - f(x^*)$	$\max\left\{\frac{L}{\mu}, \left(\frac{L\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right\}$	$(1, 2]$	✓	✓

[(1)] All assumptions are made on the whole domain.
[(2)] Complexity has extra logarithmic factor of $\ln(L/\mu)$.
[(3)] Li & Orabona (2020) assume that the noise is sub-Gaussian: $\mathbb{E}\left[\exp\left(\|\nabla f_\xi(x) - \nabla f(x)\|^2/\sigma^2\right)\right] \leq \exp(1)$ for all x from the domain.
[(4)] We notice that $\left(\frac{1}{K+1}\sum_{k=0}^{K}\|\nabla f(x^k)\|\right)^2 \leq \frac{1}{K+1}\sum_{k=0}^{K}\|\nabla f(x^k)\|^2$ and in the worst case the left-hand side is $K+1$ times smaller than the right-hand side.
[(5)] All assumptions are made on the level set $Q = \{x \in \mathbb{R}^d \mid \exists y \in \mathbb{R}^d:\ f(y) \leq f_* + 2\Delta \text{ and } \|x - y\| \leq \sqrt{\Delta}/20\sqrt{L}\}$.

However, there are no high-probability results (with logarithmic dependence on the confidence level) for smooth (strongly) convex minimization problems and Lipschitz VIP without imposing bounded variance assumption. Only recently, Zhang & Cutkosky (2022) derived optimal regret-bounds under Assumption 1.1 in the convex case with bounded gradients on $\mathbb{R}^d$. However, the bounded gradients assumption is quite restrictive when assumed on the whole space. Thus, a noticeable gap in the stochastic optimization literature remains.

Contribution. We obtain new high-probability convergence results under Assumption 1.1 for smooth convex minimization problems and Lipschitz VIP; see the summary in Tables 1 and 2. In particular, for Clipped Stochastic Similar Triangles Method (clipped-SSTM) (Gorbunov et al., 2020) and its restarted version, we derive high-probability convergence results for smooth convex and strongly convex problems. The high-probability complexity in the strongly convex case matches (up to logarithmic factors) the known in-expectation lower bound (Zhang et al., 2020b) and deterministic lower bound (Nemirovskij & Yudin, 1983). In other words, we derive the first optimal high-probability complexity results for smooth strongly convex optimization. Noticeably, the derived results have clear separation between accelerated part and stochastic part that emphasizes a potential of clipped-SSTM for efficient parallelization. Next, we derive high-probability results for clipped-SGD for smooth star-convex and quasi-strongly convex objectives under Assumption 1.1. Finally, under the same assumption, we prove the high-probability convergence of Clipped Stochastic Extragradient (clipped-SEG) (Korpelevich, 1976; Juditsky et al., 2011; Gorbunov et al., 2022a) for Lipschitz monotone and quasi-strongly monotone VIP and also obtain high-probability results for Clipped Stochastic Gradient Descent-Ascent (clipped-SGDA) for star-cocoercive and monotone / quasi-strongly monotone VIP. In the special case of $\alpha = 2$, our analysis recovers SOTA high-probability results under bounded variance assumption.

Non-convex optimization. Under the light-tails and smoothness assumption Li & Orabona (2020) derive high-

Table 2: Summary of known and new high-probability complexity results for solving (3). Column "Setup" indicates the assumptions made in addition to Assumption 1.1. All assumptions are made only on some ball around the solution with radius $\sim R \geq \|x^0 - x^*\|$ (unless the opposite is indicated). By the complexity we mean the number of stochastic oracle calls needed for a method to guarantee that $\mathbb{P}\{\text{Metric} \leq \varepsilon\} \geq 1 - \beta$ for some $\varepsilon > 0, \beta \in (0, 1]$ and "Metric" is taken from the corresponding column. For simplicity, we omit numerical and logarithmic factors in the complexity bounds. Column "α" shows the allowed values of α, "UD?" shows whether the analysis works on unbounded domains, and "UG?" indicates whether the analysis works without assuming boundedness of the gradient. Notation: $\widetilde{x}^K_{\text{avg}} = \frac{1}{K+1}\sum_{k=0}^K \widetilde{x}^k$ (for clipped-SEG), $x^K_{\text{avg}} = \frac{1}{K+1}\sum_{k=0}^K x^k$ (for clipped-SGDA); L = Lipschitz constant; D = diameter of the domain (used in (Juditsky et al., 2011)); $\texttt{Gap}_D(x) = \max_{y\in\mathcal{X}}\langle F(y), x - y\rangle$, where $\mathcal{X}$ is a bounded domain with diameter D where the problem is defined (used in (Juditsky et al., 2011)); D = diameter of the domain (for the result from (Juditsky et al., 2011)); σ = parameter from Assumption 1.1; R = any upper bound on $\|x^0 - x^*\|$; μ = quasi-strong monotonicity parameter; ℓ = star-cocoercivity parameter. The results of this paper are highlighted in blue.

Setup	Method	Citation	Metric	Complexity	α	UD?	UG?
As. 1.7 & 1.8	Mirror-Prox	(Juditsky et al., 2011)[(1)]	$\texttt{Gap}_D(\widetilde{x}^K_{\text{avg}})$	$\max\left\{\frac{LD^2}{\varepsilon}, \frac{\sigma^2 D^2}{\varepsilon^2}\right\}$	✗[(2)]	✗	✓
	clipped-SEG	(Gorbunov et al., 2022a)	$\texttt{Gap}_R(\widetilde{x}^K_{\text{avg}})$	$\max\left\{\frac{LR^2}{\varepsilon}, \frac{\sigma^2 R^2}{\varepsilon^2}\right\}$	2	✓	✓
	clipped-SEG	Theorems 4.1 & G.2	$\texttt{Gap}_R(\widetilde{x}^K_{\text{avg}})$	$\max\left\{\frac{LR^2}{\varepsilon}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}$	$(1, 2]$	✓	✓
As. 1.7 & 1.9	clipped-SEG	(Gorbunov et al., 2022a)	$\|x^k - x^*\|^2$	$\max\left\{\frac{L}{\mu}, \frac{\sigma^2}{\mu^2\varepsilon}\right\}$	2	✓	✓
	clipped-SEG	Theorems 4.1 & G.4	$\|x^k - x^*\|^2$	$\max\left\{\frac{L}{\mu}, \left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right\}$	$(1, 2]$	✓	✓
As. 1.8 & 1.10	clipped-SGDA	(Gorbunov et al., 2022a)	$\texttt{Gap}_R(x^K_{\text{avg}})$	$\max\left\{\frac{\ell R^2}{\varepsilon}, \frac{\sigma^2 R^2}{\varepsilon^2}\right\}$	2	✓	✓
	clipped-SGDA	Theorems 4.2 & H.3	$\texttt{Gap}_R(x^K_{\text{avg}})$	$\max\left\{\frac{\ell R^2}{\varepsilon}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}$	$(1, 2]$	✓	✓
As. 1.10	clipped-SGDA	(Gorbunov et al., 2022a)	$\frac{1}{K+1}\sum_{k=0}^K \|F(x^k)\|^2$	$\max\left\{\frac{\ell^2 R^2}{\varepsilon}, \frac{\ell^2\sigma^2 R^2}{\varepsilon^2}\right\}$	2	✓	✓
	clipped-SGDA	Theorems 4.2 & H.4	$\frac{1}{K+1}\sum_{k=0}^K \|F(x^k)\|^2$	$\max\left\{\frac{\ell^2 R^2}{\varepsilon}, \left(\frac{\ell\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}$	$(1, 2]$	✓	✓
As. 1.9 & 1.10	clipped-SGDA	(Gorbunov et al., 2022a)	$\|x^K - x^*\|^2$	$\max\left\{\frac{\ell}{\mu}, \frac{\sigma^2}{\mu^2\varepsilon}\right\}$	2	✓	✓
	clipped-SGDA	Theorems 4.2 & H.6	$\|x^K - x^*\|^2$	$\max\left\{\frac{\ell}{\mu}, \left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right\}$	$(1, 2]$	✓	✓

[(1)] All assumptions are made on the whole domain.

[(2)] Juditsky et al. (2011) assume that the noise is sub-Gaussian: $\mathbb{E}\left[\exp\left(\|F_\xi(x) - F(x)\|^2/\sigma^2\right)\right] \leq \exp(1)$ for all x from the domain.

probability convergence rates to the first-order stationary point for SGD. These rates match the known in-expectation guarantees for SGD and are optimal up to logarithmic factors (Arjevani et al., 2022). Recently, Cutkosky & Mehta (2021) derived the first high-probability results for non-convex optimization under Assumption 1.1 for a version of SGD with gradient clipping and normalization of the momentum. The results are obtained for the non-standard metric – $\frac{1}{K+1}\sum_{k=0}^K \|\nabla f(x^k)\|$ – and match in-expectation lower bound for the expected (non-squared) norm of the gradient from (Zhang et al., 2020b). However, Cutkosky & Mehta (2021) make an additional assumption that the norm of the gradient is bounded[3] on $\mathbb{R}^d$, which is quite restrictive.

Contribution. We derive the first high-probability result with logarithmic dependence on the confidence level for finding first-order stationary points of smooth (possibly, non-convex) functions without bounded gradients assumption. The result is derived for simple clipped-SGD. Moreover, we extend the analysis to the functions satisfying Polyak-Łojasiewicz condition; see Table 1 for the summary.

Gradient clipping received a lot of attention in the machine learning community due to its successful empirical applications in the training of deep neural networks (Pascanu et al., 2013; Goodfellow et al., 2016). The clipping operator is defined as $\texttt{clip}(x, \lambda) = \min\{1, \lambda/\|x\|\}\, x$ ($\texttt{clip}(x, \lambda) = 0$, when $x = 0$). From the theoretical perspective, gradient clipping is used for multiple different purposes: to handle structured non-smoothness in the objective function (Zhang et al., 2020a), to robustify aggregation (Karimireddy et al., 2021) and to provide privacy guarantees (Abadi et al., 2016) in the distributed training. Moreover, as we already mentioned before, gradient clipping is used to handle heavy-tailed noise (satisfying Assumption 1.1) in the stochastic gradients (Zhang et al., 2020b) and, in particular, to derive better high-probability guarantees under bounded variance assumption (Nazin et al., 2019; Gorbunov et al., 2020). However, there are no results showing the necessity of modifying standard methods like SGD and its accelerated variants to achieve high-probability convergence with logarithmic dependence on the confidence level under bounded variance assumption.

Contribution. We construct an example of a strongly convex smooth problem and stochastic oracle with bounded variance such that to achieve $\mathbb{P}\{\|x^k - x^*\|^2 > \varepsilon\} \leq \beta$ SGD requires $\Omega\left(\sigma^2/\mu\sqrt{\varepsilon\beta}\right)$ iterations, i.e., the algorithm has inverse-power dependence on the confidence level. This justifies the importance of using some non-linearity such as gradient clipping to achieve logarithmic dependence on the confidence level even in the bounded variance case.

[3] More precisely, instead of Assumption 1.1, Cutkosky & Mehta (2021) assume $\mathbb{E}_{\xi\sim}\|\nabla f_\xi(x)\|^\alpha \leq G^\alpha$ for some $G > 0$. This assumption implies Assumption 1.1 and boundedness of $\|\nabla f(x)\|$.

2. Failure of Standard SGD

It is known that SGD $x^{k+1} = x^k - \gamma \nabla f_{\xi^k}(x^k)$ can diverge in expectation, when Assumption 1.1 is satisfied with $\alpha < 2$ (Zhang et al., 2020b, Remark 1). However, it does converge in expectation when $\alpha = 2$, i.e., when the variance is bounded. In contrast, there are no high-probability convergence results for SGD having logarithmic dependence on $1/\beta$. The next theorem establishes the impossibility of deriving such high-probability results.

Theorem 2.1. *For any $\varepsilon > 0$ and sufficiently small $\beta \in (0,1)$ there exist problem (2) such that Assumptions 1.1, 1.3, and 1.6 hold with $Q = \mathbb{R}^d$, $\alpha = 2$, $0 < \mu \le L$ and for the iterates produced by SGD with any stepsize $\gamma > 0$*

$$\mathbb{P}\left\{\|x^k - x^*\|^2 \ge \varepsilon\right\} \le \beta \implies k = \Omega\left(\frac{\sigma}{\mu\sqrt{\varepsilon\beta}}\right).$$

The proof is deferred to Appendix D. We believe that similar examples can be constructed for any stochastic first-order methods having linear dependence on the stochastic gradients in their update rules. Thus, Theorem 2.1 motivates the use of non-linear operators such as gradient clipping in stochastic methods to achieve logarithmic dependence on the confidence level in the high-probability bounds.

3. Main Results for Minimization Problems

3.1. SGD with Clipping

We start with clipped-SGD:

$$x^{k+1} = x^k - \gamma \cdot \texttt{clip}\left(\nabla f_{\xi^k}(x^k), \lambda_k\right), \tag{15}$$

where ξ^k is sampled from $\mathcal{D}_k$ independently from previous steps. We emphasize here and below that distribution of the noise is allowed to be dependent on k: we require just independence of ξ^k from the the previous steps. Our main convergence results for clipped-SGD are summarized in the following theorem.

Theorem 3.1 (Convergence of clipped-SGD)**.** *Let $k \ge 0$ and $\beta \in (0,1]$ are such that $A = \ln\frac{4(K+1)}{\beta} \ge 1$.*
Case 1. *Let Assumptions 1.1, 1.2, 1.3 hold for $Q = \{x \in \mathbb{R}^d \mid \exists y \in \mathbb{R}^d : f(y) \le f_* + 2\Delta$ and $\|x - y\| \le \sqrt{\Delta}/20\sqrt{L}\}$, $\Delta \ge f(x^0) - f_*$ and $0 < \gamma \le \mathcal{O}\left(\min\{1/LA, \sqrt{\Delta}/\sigma\sqrt{L}K^{1/\alpha}A^{(\alpha-1)/\alpha}\}\right)$, $\lambda_k = \lambda = \Theta(\sqrt{\Delta}/\sqrt{L}\gamma A)$.*
Case 2. *Let Assumptions 1.1, 1.3, 1.4 hold for $Q = \{x \in \mathbb{R}^d \mid \exists y \in \mathbb{R}^d : f(y) \le f_* + 2\Delta$ and $\|x - y\| \le \sqrt{\Delta}/20\sqrt{L}\}$, $\Delta \ge f(x^0) - f_*$ and $0 < \gamma = \mathcal{O}\left(\min\{1/LA, \ln(B_K)/\mu(K+1)\}\right)$, $B_K = \Theta\left(\max\{2, (K+1)^{2(\alpha-1)/\alpha}\mu^2\Delta/L\sigma^2A^{2(\alpha-1)/\alpha}\ln^2(B_K)\}\right)$, $\lambda_k = \Theta(\exp(-\gamma\mu(1+k/2))\sqrt{\Delta}/\sqrt{L}\gamma A)$.*
Case 3. *Let Assumptions 1.1, 1.3, 1.6 with $\mu = 0$ hold for $Q = B_{3R}(x^*)$, $R \ge \|x^0 - x^*\|$ and $0 < \gamma \le \mathcal{O}(\min\{1/LA, R/\sigma K^{1/\alpha}A^{(\alpha-1)/\alpha}\})$, $\lambda_k = \lambda = \Theta(R/\gamma A)$.*
Case 4. *Let Assumptions 1.1, 1.3, 1.5 with $\mu > 0$ hold for $Q = B_{3R}(x^*)$, $R \ge \|x^0 - x^*\|$ and $0 < \gamma = \mathcal{O}\left(\min\{1/LA, \ln(B_K)/\mu(K+1)\}\right)$, $B_K = \Theta\left(\max\{2, (K+1)^{2(\alpha-1)/\alpha}\mu^2R^2/\sigma^2A^{2(\alpha-1)/\alpha}\ln^2(B_K)\}\right)$, $\lambda_k = \Theta(\exp(-\gamma\mu(1+k/2))R/\gamma A)$.*
*Then to guarantee $\frac{1}{K+1}\sum_{k=0}^{k}\|\nabla f(x^k)\|^2 \le \varepsilon$ in **Case 1**, $f(x^K) - f(x^*) \le \varepsilon$ in **Case 2**, $f(\bar{x}^K) - f(x^*) \le \varepsilon$ in **Case 3** with $\bar{x}^K = \frac{1}{K+1}\sum_{k=0}^{K}x^k$, $\|x^K - x^*\|^2 \le \varepsilon$ in **Case 4** with probability $\ge 1 - \beta$ clipped-SGD requires*

$$\textit{Case 1:}\quad \widetilde{\mathcal{O}}\left(\max\left\{\frac{L\Delta}{\varepsilon}, \left(\frac{\sqrt{L\Delta}\sigma}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right) \tag{16}$$

$$\textit{Case 2:}\quad \widetilde{\mathcal{O}}\left(\max\left\{\frac{L}{\mu}, \left(\frac{L\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right\}\right) \tag{17}$$

$$\textit{Case 3:}\quad \widetilde{\mathcal{O}}\left(\max\left\{\frac{LR^2}{\varepsilon}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right) \tag{18}$$

$$\textit{Case 4:}\quad \widetilde{\mathcal{O}}\left(\max\left\{\frac{L}{\mu}, \left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right\}\right) \tag{19}$$

oracle calls.

The complete formulation of the result and full proofs are deferred to Appendix E. As one can see from Table 1, for $\alpha = 2$ the derived complexity bounds match the best-known ones for clipped-SGD in the setups where it was analyzed. Next, we emphasize that the second term under the maximum in (19) (quasi-strongly convex functions) is optimal up to logarithmic factors (Zhang et al., 2020b). In the convex case, there are no lower bounds, but we conjecture that the second term in (18) is optimal (up to logarithms) in this case as well.

Next, in the case of PŁ-functions, we are not aware of any high-probability convergence results in the literature. In the special case of $\alpha = 2$, the derived complexity bound (17) matches the best-known in-expectation complexity bound for SGD (Karimi et al., 2016; Khaled & Richtárik, 2020) and the first term coincides (up to logarithms) with the lower bound for deterministic first-order methods in this setup (Yue et al., 2022).

Finally, in the non-convex case, bound (16) is the first high-probability result under Assumption 1.1 without the additional assumption of the boundedness of the gradients. For $\alpha = 2$ it matches (up to logarithms) in-expectation lower bound (Arjevani et al., 2022). However, when $\alpha < 2$, bound (16) is inferior to the existing one $\widetilde{\mathcal{O}}\left((G^2/\varepsilon)^{(3\alpha-2)/2(\alpha-1)}\right)$ by Cutkosky & Mehta

(2021), which relies on the stronger assumption that $\mathbb{E}_{\xi\sim\mathcal{D}}\|\nabla f_\xi(x)\|^\alpha \leq G^\alpha$ for some $G > 0$ and all $x \in \mathbb{R}^d$, and also do not match the lower bound by Zhang et al. (2020b) derived for $\mathbb{E}\|\nabla f(x^k)\|$, where x^k is the output of the stochastic first-order method. It is also worth mentioning that Cutkosky & Mehta (2021) use a different performance metric: $\hat{\mathcal{P}}_K = \left(\frac{1}{K+1}\sum_{k=0}^K \|\nabla f(x^k)\|\right)^2$. This metric is always smaller than $\mathcal{P}_K = \frac{1}{K+1}\sum_{k=0}^K \|\nabla f(x^k)\|^2$, which we use in our result. In the worst case, $\mathcal{P}_K$ can be $K+1$ times larger than $\hat{\mathcal{P}}_K$. Moreover, the lower bound from (Zhang et al., 2020b) is derived for $\mathbb{E}\|\nabla f(x^k)\|$ that is also always smaller than the standard quantity of interest $\mathbb{E}\|\nabla f(x^k)\|^2$. Therefore, the question of optimality of the bound (16) remains open for $\alpha < 2$. Moreover, it will also be interesting to modify our analysis in this case to derive a better bound for metric $\hat{\mathcal{P}}_K$ than (16).

3.2. Acceleration

Next, we focus on the accelerated version of clipped-SGD called Clipped Stochastic Similar Triangles Method clipped-SSTM (Gorbunov et al., 2020). The method constructs three sequences of points $\{x^k\}_{k\geq 0}$, $\{y^k\}_{k\geq 0}$, $\{z^k\}_{k\geq 0}$ satisfying the following update rules: $x^0 = y^0 = z^0$ and

$$x^{k+1} = \frac{A_k y^k + \alpha_{k+1} z^k}{A_{k+1}}, \tag{20}$$

$$z^{k+1} = z^k - \alpha_{k+1} \cdot \texttt{clip}\left(\nabla f_{\xi^k}(x^{k+1}), \lambda_k\right), \tag{21}$$

$$y^{k+1} = \frac{A_k y^k + \alpha_{k+1} z^{k+1}}{A_{k+1}}, \tag{22}$$

where $A_0 = \alpha_0 = 0$, $\alpha_{k+1} = \frac{k+2}{2aL}$, $A_{k+1} = A_k + \alpha_{k+1}$, and ξ^k is sampled from $\mathcal{D}_k$ independently from previous steps. Our main convergence result for clipped-SSTM is given in the following theorem.

Theorem 3.2 (Convergence of clipped-SSTM). *Let Assumptions 1.1, 1.3, 1.6 with $\mu = 0$ hold for $Q = B_{3R}(x^*)$, $R \geq \|x^0 - x^*\|^2$ and $a = \Theta(\max\{A^2, \sigma K^{(\alpha+1)/\alpha} A^{(\alpha-1)/\alpha}/LR\})$, $\lambda_k = \Theta(R/(\alpha_{k+1}A))$, where $A = \ln\frac{4K}{\beta}$, $\beta \in (0,1]$ are such that $A \geq 1$. Then to guarantee $f(y^K) - f(x^*) \leq \varepsilon$ with probability $\geq 1-\beta$ clipped-SSTM requires*

$$\widetilde{\mathcal{O}}\left(\max\left\{\sqrt{\frac{LR^2}{\varepsilon}}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right) \quad \textit{oracle calls.} \tag{23}$$

Moreover, with probability $\geq 1-\beta$ the iterates of clipped-SSTM stay in the ball $B_{2R}(x^*)$: $\{x^k\}_{k=0}^{K+1}, \{y^k\}_{k=0}^K, \{z^k\}_{k=0}^K \subseteq B_{2R}(x^*)$.

The derived high-probability bound matches (see the proof in Appendix F.1) the best-known one in the case of $\alpha = 2$. For $\alpha < 2$ there are no lower bounds in the convex case. However, the first term in (23) is optimal and matches the deterministic lower bound in the convex case (Nemirovskij & Yudin, 1983). The second term is the same as in the bound for clipped-SGD (18) and we conjecture that it cannot be improved.

In the strongly convex case, we consider clipped-SSTM with restarts (R-clipped-SSTM). This method consists of τ stages. On the t-th stage R-clipped-SSTM runs clipped-SSTM for K_0 iterations from the starting point $\hat{x}^t$, which is the output from the previous stage ($\hat{x}^t = x^0$), and defines the obtained point as $\hat{x}^{t+1}$; see Algorithm 3 in Appendix F.2. For this procedure we have the following result.

Theorem 3.3 (Convergence of R-clipped-SSTM). *Let Assumptions 1.1, 1.3, 1.6 with $\mu > 0$ hold for $Q = B_{3R}(x^*)$, $R \geq \|x^0 - x^*\|^2$ and R-clipped-SSTM runs clipped-SSTM $\tau = \lceil\log_2(\mu R^2/2\varepsilon)\rceil$ times. Let $K_t = \widetilde{\Theta}(\max\{\sqrt{LR_{t-1}^2/\varepsilon_t}, (\sigma R_{t-1}/\varepsilon_t)^{\alpha/(\alpha-1)}\})$, $a_t = \widetilde{\Theta}(\max\{1, \sigma K_t^{\alpha+1/\alpha}/LR_t\})$, $\lambda_k^t = \widetilde{\Theta}(R/\alpha_{k+1}^t)$ be the parameters for the stage t of R-clipped-SSTM, where $R_{t-1} = 2^{-(t-1)/2}R$, $\varepsilon_t = \mu R_{t-1}^2/4$, $\ln\frac{4\tau K_t}{\beta} \geq 1$ for all $t = 1,\ldots,\tau$ and some $\beta \in (0,1]$. Then to guarantee $f(\hat{x}^\tau) - f(x^*) \leq \varepsilon$ with probability $\geq 1-\beta$ R-clipped-SSTM requires*

$$\widetilde{\mathcal{O}}\left(\max\left\{\sqrt{\frac{L}{\mu}}, \left(\frac{\sigma^2}{\mu\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right\}\right) \quad \textit{oracle calls.} \tag{24}$$

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

The obtained complexity bound (see the proof in Appendix F.2) is the first optimal (up to logarithms) high-probability complexity bound under Assumption 1.1 for the smooth strongly convex problems. Indeed, the first term cannot be improved in view of the deterministic lower bound by Nemirovskij & Yudin (1983), and the second term is optimal due to Zhang et al. (2020b).

4. Main Results for Variational Inequalities

4.1. Clipped Stochastic Extragradient

For (quasi strongly) monotone VIPs we consider Clipped Stochastic Extragradient method (clipped-SEG):

$$\widetilde{x}^k = x^k - \gamma \cdot \texttt{clip}(F_{\xi_1^k}(x^k), \lambda_k), \tag{25}$$

$$x^{k+1} = x^k - \gamma \cdot \texttt{clip}(F_{\xi_2^k}(\widetilde{x}^k), \lambda_k), \tag{26}$$

where ξ_1^k, ξ_2^k are sampled from $\mathcal{D}_k$ independently from previous steps. Our main convergence results for clipped-SEG are summarized below.

Theorem 4.1 (Convergence of clipped-SEG).
Case 1. *Let Assumptions 1.1, 1.3, 1.8 hold for $Q = B_{4R}(x^*)$ and $0 < \gamma = \mathcal{O}\left(\min\{1/LA, R/K^{1/\alpha}\sigma A^{(\alpha-1)/\alpha}\}\right)$,*

$\lambda_k = \lambda = \Theta(R/\gamma A)$, *where* $A = \ln\frac{6(K+1)}{\beta} \geq 1$, $\beta \in (0,1]$.
Case 2. *Let Assumptions 1.1, 1.3, 1.9 with* $\mu > 0$ *hold for* $Q = B_{3R}(x^*)$ *and* $0 < \gamma = \mathcal{O}\left(\min\{1/LA, \ln(B_K)/\mu(K+1)\}\right)$, $B_K = \Theta\left(\max\{2, (K+1)^{2(\alpha-1)/\alpha}\mu^2R^2/\sigma^2A^{2(\alpha-1)/\alpha}\ln^2(B_K)\}\right)$, $\lambda_k = \Theta(\exp(-\gamma\mu(1+k/2))R/\gamma A)$, *where* $A = \ln\frac{6(K+1)}{\beta}$, $\beta \in (0,1]$ *are such that* $A \geq 1$.
Then to guarantee $\texttt{Gap}_R(\widetilde{x}^K_{avg}) \leq \varepsilon$ *in **Case 1** with* $\widetilde{x}^K_{avg} = \frac{1}{K+1}\sum_{k=0}^K \widetilde{x}^k$, $\|x^K - x^*\|^2 \leq \varepsilon$ *in **Case 2** with probability* $\geq 1-\beta$ clipped-SEG *requires*

$$\textit{Case 1:}\quad \widetilde{\mathcal{O}}\left(\max\left\{\frac{LR^2}{\varepsilon}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right) \tag{27}$$

$$\textit{Case 2:}\quad \widetilde{\mathcal{O}}\left(\max\left\{\frac{L}{\mu}, \left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right\}\right) \tag{28}$$

oracle calls.

The proofs are deferred to Appendix G. When $\alpha = 2$, the above bounds recover SOTA high-probability bounds for monotone and quasi-strongly monotone Lipschitz VIP (Gorbunov et al., 2022a). For the case of $\alpha < 2$ (27) and (28) are the first high-probability results for the mentioned classes. Next, the first terms in these complexity bounds are optimal (up to logarithms) due to the lower bounds for the deterministic methods (Ouyang & Xu, 2021; Zhang et al., 2022). The second term in (28) is also optimal (up to logarithms) due to the lower bounds for stochastic strongly convex minimization (Zhang et al., 2020b). Similarly to the convex case in minimization, we conjecture that the second term in (27) cannot be improved in the monotone case as well.

4.2. Clipped Stochastic Gradient Descent-Ascent

In the star-cocoercive case, we focus on Clipped Stochastic Gradient Descent-Ascent (clipped-SGDA):

$$x^{k+1} = x^k - \gamma\cdot\texttt{clip}(F_{\xi^k}(x^k), \lambda_k), \tag{29}$$

where ξ^k is sampled from $\mathcal{D}_k$ independently from previous steps. For this method we derive the following convergence guarantees.

Theorem 4.2 (Convergence of clipped-SGDA).
Case 1. *Let Assumptions 1.1, 1.10, 1.8 hold for* $Q = B_{2R}(x^*)$ *and* $0 < \gamma = \mathcal{O}\left(\min\{1/\ell A, R/K^{1/\alpha}\sigma A^{(\alpha-1)/\alpha}\}\right)$, $\lambda_k = \lambda = \Theta(R/\gamma A)$, $\beta \in (0,1]$ *are such that* $A \geq 1$.
Case 2. *Let Assumptions 1.1, 1.10 hold for* $Q = B_{2R}(x^*)$ *and* $0 < \gamma = \mathcal{O}\left(\min\{1/\ell A, R/K^{1/\alpha}\sigma A^{(\alpha-1)/\alpha}\}\right)$, $\lambda_k = \lambda = \Theta(R/\gamma A)$, *where* $A = \ln\frac{4(K+1)}{\beta}$, $\beta \in (0,1]$ *are such that* $A \geq 1$.
Case 3. *Let Assumptions 1.1, 1.10, 1.9 with* $\mu > 0$ *hold for* $Q = B_{2R}(x^*)$ *and* $0 < \gamma = \mathcal{O}\left(\min\{1/\ell A, \ln(B_K)/\mu(K+1)\}\right)$, $B_K = \Theta\left(\max\{2, (K+1)^{2(\alpha-1)/\alpha}\mu^2R^2/\sigma^2A^{2(\alpha-1)/\alpha}\ln^2(B_K)\}\right)$, $\lambda_k = \Theta(\exp(-\gamma\mu(1+k/2))R/\gamma A)$, *where* $A = \ln\frac{4(K+1)}{\beta}$, $\beta \in (0,1]$ *are such that* $A \geq 1$.
Then to guarantee $\texttt{Gap}_R(\widetilde{x}^K_{avg}) \leq \varepsilon$ *in **Case 1** with* $\widetilde{x}^K_{avg} = \frac{1}{K+1}\sum_{k=0}^K \widetilde{x}^k$, $\frac{1}{K+1}\sum_{k=0}^k \|F(x^k)\|^2 \leq \ell\varepsilon$ *in **Case 2**,* $\|x^K - x^*\|^2 \leq \varepsilon$ *in **Case 3** with probability* $\geq 1-\beta$ clipped-SGDA *requires*

$$\textit{Case 1 and 2:}\quad \widetilde{\mathcal{O}}\left(\max\left\{\frac{\ell R^2}{\varepsilon}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right) \tag{30}$$

$$\textit{Case 2:}\quad \widetilde{\mathcal{O}}\left(\max\left\{\frac{\ell}{\mu}, \left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right\}\right) \tag{31}$$

oracle calls.

One can find the proofs in Appendix H. The derived high-probability results generalize the existing SOTA results from the case of $\alpha = 2$ (Gorbunov et al., 2022a) to the case of $\alpha < 2$.

5. Key Lemma and Intuition Behind the Proofs

The proofs of all results in this paper follow a very similar pattern. To illustrate the main idea, we consider the analysis of clipped-SGD in the non-convex case. Mimicking the proof of deterministic GD we derive the following inequality:

$$\gamma\sum_{k=0}^K \|\nabla f(x^k)\|^2 \lessapprox \Delta_0 - \Delta_{K+1} - \gamma\sum_{k=0}^K\langle\nabla f(x^k), \theta_k\rangle + L\gamma^2\sum_{k=0}^{K-1}\|\theta_k\|^2, \tag{32}$$

where $\Delta_k = f(x^k) - f_*$ and $\theta_k = \widetilde{\nabla} f_{\xi^k}(x^k) - \nabla f(x^k)$. In other words, we separate the deterministic part of the method from its stochastic part. To obtain the result of Theorem 3.1 (Case 1) it remains to upper bound with high-probability the sums from the second line of the formula above. We do it with the help of Bernstein's inequality (Lemma B.2). However, it requires several preliminary steps. In particular, Bernstein's inequality needs the random variables to be bounded. The magnitudes of summands depend on $\nabla f(x^k)$ that can be arbitrarily large due to the stochasticity in x^k. However, (32) allows to bound Δ_{K+1} inductively and, using smoothness, to bound $\|\nabla f(x^{K+1})\|$. Secondly, Bernstein's inequality requires knowing the bounds on the bias and variance of the clipped stochastic estimator. For such purposes, we derive the following result, which is a generalization of Lemma F.5 from (Gorbunov et al., 2020); see also Lemma 10 from (Zhang et al., 2020b).

Lemma 5.1. *Let X be a random vector in $\mathbb{R}^d$ and $\widetilde{X} = \texttt{clip}(X, \lambda)$. Then, $\|\widetilde{X} - \mathbb{E}[\widetilde{X}]\| \leq 2\lambda$. 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 $\|x\| \leq \lambda/2$, then*

$$\left\|\mathbb{E}[\widetilde{X}] - x\right\| \leq \frac{2^\alpha \sigma^\alpha}{\lambda^{\alpha-1}}, \tag{33}$$

$$\mathbb{E}\left[\left\|\widetilde{X} - x\right\|^2\right] \leq 18\lambda^{2-\alpha}\sigma^\alpha, \tag{34}$$

$$\mathbb{E}\left[\left\|\widetilde{X} - \mathbb{E}[\widetilde{X}]\right\|^2\right] \leq 18\lambda^{2-\alpha}\sigma^\alpha. \tag{35}$$

This lemma can be useful on its own for analyses involving clipping operators. Moreover, our high-probability analysis does not rely on the choice of clipping explicitly: in the proofs, we use only $\|\widetilde{X}\| \leq \lambda$ and inequalities (33)-(35). Therefore, our results hold for the methods considered in this work with any other non-linearity $\phi_\lambda(x)$ (not necessary clipping), if it satisfies the conditions from the above lemma for $\widetilde{X} = \phi_\lambda(X)$.

6. Discussion

In this work, we contributed to the stochastic optimization literature via deriving new high-probability results under Assumption 1.1. Our results can be extended to the minimization of functions with Hölder continuous gradients using similar ideas to (Gorbunov et al., 2021). Another prominent direction is in obtaining new high-probability results for other types of non-linearities, e.g., like in (Polyak & Tsypkin, 1980; Jakovetic et al., 2022).

Acknowledgements

This work was partially supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Moscow Institute of Physics and Technology dated November 1, 2021 No. 70-2021-00138.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016. (Cited on page 5)

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, pp. 1–50, 2022. (Cited on pages 5 and 6)

Bennett, G. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962. (Cited on page 14)

Cutkosky, A. and Mehta, H. High-probability bounds for non-convex stochastic optimization with heavy tails. *Advances in Neural Information Processing Systems*, 34, 2021. (Cited on pages 1, 4, 5, 6, and 7)

Davis, D., Drusvyatskiy, D., Xiao, L., and Zhang, J. From low probability to high confidence in stochastic convex optimization. *Journal of Machine Learning Research*, 22 (49):1–38, 2021. (Cited on pages 1, 3, and 4)

Dvurechenskii, P., Dvinskikh, D., Gasnikov, A., Uribe, C., and Nedich, A. Decentralize and randomize: Faster algorithm for wasserstein barycenters. *Advances in Neural Information Processing Systems*, 31, 2018. (Cited on page 40)

Dzhaparidze, K. and Van Zanten, J. On bernstein-type inequalities for martingales. *Stochastic processes and their applications*, 93(1):109–117, 2001. (Cited on page 14)

Freedman, D. A. et al. On tail probabilities for martingales. *the Annals of Probability*, 3(1):100–118, 1975. (Cited on page 14)

Gasnikov, A. and Nesterov, Y. Universal fast gradient method for stochastic composit optimization problems. *arXiv preprint arXiv:1604.05275*, 2016. (Cited on page 40)

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. (Cited on pages 2 and 3)

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. (Cited on page 2)

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial networks. *International Conference on Learning Representations*, 2019. (Cited on page 2)

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. (Cited on page 2)

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016. (Cited on page 5)

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. (Cited on pages 1, 3, 4, 5, 7, 8, 14, 40, and 48)

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. (Cited on pages 1, 3, 4, 9, 40, and 43)

Gorbunov, E., Danilova, M., Dobre, D., Dvurechensky, P., Gasnikov, A., and Gidel, G. Clipped stochastic methods for variational inequalities with heavy-tailed noise. *arXiv preprint arXiv:2206.01095*, 2022a. (Cited on pages 1, 3, 4, 5, 8, 51, 52, 59, 60, 69, 70, 77, 79, and 80)

Gorbunov, E., Loizou, N., and Gidel, G. Extragradient method: $\mathcal{O}(1/K)$ last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In *International Conference on Artificial Intelligence and Statistics*, pp. 366–402. PMLR, 2022b. (Cited on pages 38, 51, and 69)

Harker, P. T. and Pang, J.-S. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical programming*, 48(1):161–220, 1990. (Cited on page 2)

Jakovetic, D., Bajovic, D., Sahu, A. K., Kar, S., Milosevic, N., and Stamenkovic, D. Nonlinear gradient mappings and stochastic optimization: A general framework with applications to heavy-tail noise. *arXiv preprint arXiv:2204.02593*, 2022. (Cited on page 9)

Juditsky, A., Nemirovski, A., and Tauvel, C. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011. (Cited on pages 3, 4, and 5)

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 795–811. Springer, 2016. (Cited on page 6)

Karimireddy, S. P., He, L., and Jaggi, M. Learning from history for byzantine robust optimization. In *International Conference on Machine Learning*, pp. 5311–5319. PMLR, 2021. (Cited on page 5)

Khaled, A. and Richtárik, P. Better theory for sgd in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020. (Cited on page 6)

Korpelevich, G. M. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976. (Cited on page 4)

Li, X. and Orabona, F. A high probability analysis of adaptive sgd with momentum. *arXiv preprint arXiv:2007.14294*, 2020. (Cited on page 4)

Liu, C., Zhu, L., and Belkin, M. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022. (Cited on page 2)

Loizou, N., Berard, H., Gidel, G., Mitliagkas, I., and Lacoste-Julien, S. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. *Advances in Neural Information Processing Systems*, 34, 2021. (Cited on page 3)

Lojasiewicz, S. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963. (Cited on page 2)

Mertikopoulos, P. and Zhou, Z. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1):465–507, 2019. (Cited on page 3)

Nazin, A. V., Nemirovsky, A., Tsybakov, A. B., and Juditsky, A. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80(9):1607–1627, 2019. (Cited on pages 1, 3, 4, and 5)

Necoara, I., Nesterov, Y., and Glineur, F. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1):69–107, 2019. (Cited on page 3)

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009. (Cited on pages 2 and 3)

Nemirovskij, A. S. and Yudin, D. B. Problem complexity and method efficiency in optimization. 1983. (Cited on pages 4, 7, and 13)

Nesterov, Y. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007. (Cited on page 2)

Nesterov, Y. et al. *Lectures on convex optimization*, volume 137. Springer, 2018. (Cited on page 14)

Ouyang, Y. and Xu, Y. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1):1–35, 2021. (Cited on page 8)

Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pp. 1310–1318, 2013. (Cited on pages 5 and 19)

Patel, V. and Berahas, A. S. Gradient descent in the absence of global lipschitz continuity of the gradients: Convergence, divergence and limitations of its continuous approximation. *arXiv preprint arXiv:2210.02418*, 2022. (Cited on page 2)

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. (Cited on page 2)

Polyak, B. T. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963. (Cited on page 2)

Polyak, B. T. and Tsypkin, Y. Z. Optimal pseudogradient adaptation algorithms. *Avtomatika i Telemekhanika*, (8): 74–84, 1980. (Cited on page 9)

Robbins, H. and Monro, S. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951. (Cited on page 1)

Ryu, E. K. and Yin, W. Large-scale convex optimization via monotone operators, 2021. (Cited on page 2)

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014. (Cited on page 2)

Song, C., Zhou, Z., Zhou, Y., Jiang, Y., and Ma, Y. Optimistic dual extrapolation for coherent non-monotone variational inequalities. *Advances in Neural Information Processing Systems*, 33:14303–14314, 2020. (Cited on page 3)

Vural, N. M., Yu, L., Balasubramanian, K., Volgushev, S., and Erdogdu, M. A. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In *Conference on Learning Theory*, pp. 65–102. PMLR, 2022. (Cited on page 13)

Yue, P., Fang, C., and Lin, Z. On the lower bound of minimizing Polyak-Lojasiewicz functions. *arXiv preprint arXiv:2212.13551*, 2022. (Cited on page 6)

Zhang, J. and Cutkosky, A. Parameter-free regret in high probability with heavy tails. *arXiv preprint arXiv:2210.14355*, 2022. (Cited on page 4)

Zhang, J., He, T., Sra, S., and Jadbabaie, A. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=BJgnXpVYwS. (Cited on page 5)

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S. J., Kumar, S., and Sra, S. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33, 2020b. (Cited on pages 2, 4, 5, 6, 7, 8, and 13)

Zhang, J., Hong, M., and Zhang, S. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming*, 194(1-2):901–935, 2022. (Cited on page 8)

Contents

A. Additional Related Work

In this section, we provide an overview of the existing in-expectation convergence results under Assumption 1.1.

Convex minimization. The first in-expectation result under Assumption 1.1 is given by Nemirovskij & Yudin (1983), who derive[4] $\mathcal{O}(\varepsilon^{-\alpha/(\alpha-1)})$ complexity for Mirror Descent applied to the minimization of convex functions with bounded gradients. This result was recently extended by Vural et al. (2022) to the uniformly convex functions, and matching lower bounds were derived. In the strongly convex case, Zhang et al. (2020b) prove $\mathcal{O}(\varepsilon^{-\alpha/2(\alpha-1)})$ complexity for clipped-SGD. However, all these results rely on the boundedness of the gradient. To the best of our knowledge, there are no results for smooth convex problems under Assumption 1.1 without assuming that the gradient is bounded even in terms of expectation.

Non-convex minimization. In the non-convex smooth case, Zhang et al. (2020b) prove $\mathcal{O}(\varepsilon^{-(3\alpha-2)/(\alpha-1)})$ complexity for clipped-SGD to produce a point x such that $\mathbb{E}\|\nabla f(x)\| \leq \varepsilon$. In the same work, the authors derive the matching lower bound. However, both upper and lower bounds are derived for $\mathbb{E}\|\nabla f(x)\|$ which is smaller than $\sqrt{\mathbb{E}\|\nabla f(x)\|^2}$. The later one is stronger and is more standard performance metric for stochastic non-convex optimization. Therefore, the question of deriving lower and matching upper bounds for the standard metric remains open.

[4]In this section, we hide in $\mathcal{O}(\cdot)$ all dependencies except the dependency on ε.

B. Useful Facts

Smoothness. If f is L-smooth on convex set $Q \subseteq \mathbb{R}^d$, then for all $x, y \in Q$ (Nesterov et al., 2018)

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2. \tag{36}$$

In particular, if x and $y = x - \frac{1}{L}\nabla f(x)$ lie in Q, then the above inequality gives

$$f(y) \le f(x) - \frac{1}{L}\|\nabla f(x)\|^2 + \frac{1}{2L}\|\nabla f(x)\|^2 = f(x) - \frac{1}{2L}\|\nabla f(x)\|^2$$

and

$$\|\nabla f(x)\|^2 \le 2L\left(f(x) - f(y)\right) \le 2L\left(f(x) - f_*\right)$$

under the assumption that $f_* = \inf_{x\in Q} f(x) > -\infty$. In other words, (7) holds for any $x \in Q$ such that $(x - \frac{1}{L}\nabla f(x)) \in Q$. For example, if x^* is an optimum of f, then L-smoothness on $B_{2R}(x^*)$ implies that (7) holds on $B_R(x^*)$: indeed, for any $x \in B_R(x^*)$ we have

$$\left\|x - \frac{1}{L}\nabla f(x) - x^*\right\| \le \|x - x^*\| + \frac{1}{L}\|\nabla f(x)\| \overset{(6)}{\le} 2\|x - x^*\| \le 2R.$$

This derivation means that, in the worst case, to have (7) on a set Q we need to assume smoothness on a slightly larger set.

Parameters in clipped-SSTM. To analyze clipped-SSTM we use the following lemma about its parameters α_k and A_k.

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

$$\alpha_0 = A_0 = 0, \quad A_{k+1} = A_k + \alpha_{k+1}, \quad \alpha_{k+1} = \frac{k+2}{2aL} \quad \forall k \ge 0, \tag{37}$$

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

$$A_{k+1} = \frac{(k+1)(k+4)}{4aL}, \tag{38}$$

$$A_{k+1} \ge aL\alpha_{k+1}^2. \tag{39}$$

Bernstein inequality. One of the final steps in our proofs is in the proper application of the following lemma known as *Bernstein inequality for martingale differences* (Bennett, 1962; Dzhaparidze & Van Zanten, 2001; Freedman et al., 1975).

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

$$\mathbb{P}\left\{\Big|\sum_{i=1}^n X_i\Big| > b \text{ and } \sum_{i=1}^n \sigma_i^2 \le G\right\} \le 2\exp\left(-\frac{b^2}{2G + {}^{2cb}/_{3}}\right). \tag{40}$$

C. Proof of Lemma 5.1

Lemma C.1 (Lemma 5.1)**.** *Let X be a random vector in $\mathbb{R}^d$ and $\widetilde{X} = \texttt{clip}(X,\lambda)$. Then,*

$$\left\|\widetilde{X} - \mathbb{E}[\widetilde{X}]\right\| \leq 2\lambda. \tag{41}$$

Moreover, if for some $\sigma \geq 0$ and $\alpha \in [1,2)$

$$\mathbb{E}[X] = x \in \mathbb{R}^d, \quad \mathbb{E}[\|X - x\|^\alpha] \leq \sigma^\alpha \tag{42}$$

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

$$\left\|\mathbb{E}[\widetilde{X}] - x\right\| \leq \frac{2^\alpha \sigma^\alpha}{\lambda^{\alpha-1}}, \tag{43}$$

$$\mathbb{E}\left[\left\|\widetilde{X} - x\right\|^2\right] \leq 18\lambda^{2-\alpha}\sigma^\alpha, \tag{44}$$

$$\mathbb{E}\left[\left\|\widetilde{X} - \mathbb{E}[\widetilde{X}]\right\|^2\right] \leq 18\lambda^{2-\alpha}\sigma^\alpha. \tag{45}$$

Proof. **Proof of (41):** by definition of a clipping operator, we have

$$\begin{aligned}
\left\|\widetilde{X} - \mathbb{E}\left[\widetilde{X}\right]\right\| &\leq \left\|\widetilde{X}\right\| + \left\|\mathbb{E}\left[\widetilde{X}\right]\right\| \\
&= \|\texttt{clip}(X,\lambda)\| + \|\mathbb{E}\left[\texttt{clip}(X,\lambda)\right]\| \\
&\leq \left\|\min\left\{1, \frac{\lambda}{\|X\|}\right\} X\right\| + \mathbb{E}\left[\left\|\min\left\{1, \frac{\lambda}{\|X\|}\right\} X\right\|\right] \\
&= \min\left\{\|X\|, \lambda\right\} + \mathbb{E}\left[\min\left\{\|X\|, \lambda\right\}\right] \\
&\leq \lambda + \lambda = 2\lambda.
\end{aligned}$$

Proof of (43): To start the proof, we introduce two indicator random variables. Let

$$\chi = \mathbb{I}_{\{X : \|X\| > \lambda\}} = \begin{cases} 1, & \text{if } \|X\| > \lambda, \\ 0, & \text{otherwise} \end{cases}, \quad \eta = \mathbb{I}_{\left\{X : \|X - x\| > \frac{\lambda}{2}\right\}} = \begin{cases} 1, & \text{if } \|X - x\| > \frac{\lambda}{2}, \\ 0, & \text{otherwise} \end{cases}. \tag{46}$$

Moreover, since $\|X\| \leq \|x\| + \|X - x\| \overset{\|x\| \leq \lambda/2}{\leq} \frac{\lambda}{2} + \|X - x\|$, we have $\chi \leq \eta$. We are now in a position to show (43). Using that

$$\widetilde{X} = \min\left\{1, \frac{\lambda}{\|X\|}\right\} X = \chi \frac{\lambda}{\|X\|} X + (1 - \chi) X,$$

we obtain

$$\begin{aligned}
\left\|\mathbb{E}\left[\widetilde{X}\right] - x\right\| &= \left\|\mathbb{E}\left[X + \chi\left(\frac{\lambda}{\|X\|} - 1\right) X\right] - x\right\| \\
&= \left\|\mathbb{E}\left[\chi\left(\frac{\lambda}{\|X\|} - 1\right) X\right]\right\| \\
&\leq \mathbb{E}\left[\left|\chi\left(\frac{\lambda}{\|X\|} - 1\right)\right| \|X\|\right] \\
&= \mathbb{E}\left[\chi\left(1 - \frac{\lambda}{\|X\|}\right) \|X\|\right].
\end{aligned}$$

Since $1 - \lambda/\|X\| \in (0,1)$ when $\chi \neq 0$, we derive

$$\begin{aligned}\left\|\mathbb{E}\left[\widetilde{X}\right] - x\right\| &\leq \mathbb{E}\left[\chi\|X\|\right] \\ &\overset{\chi\leq\eta}{\leq} \mathbb{E}\left[\eta\|X\|\right] \\ &\leq \mathbb{E}\left[\eta\|X-x\| + \eta\|x\|\right] \\ &\overset{(*)}{\leq} \left(\mathbb{E}\left[\|X-x\|^\alpha\right]\right)^{1/\alpha}\left(\mathbb{E}\left[\eta^{\frac{\alpha}{\alpha-1}}\right]\right)^{\frac{\alpha-1}{\alpha}} + \|x\|\mathbb{E}\left[\eta\right] \\ &\overset{(42)}{\leq} \sigma\left(\mathbb{E}\left[\eta^{\frac{\alpha}{1-\alpha}}\right]\right)^{\frac{1-\alpha}{\alpha}} + \frac{\lambda}{2}\mathbb{E}\left[\eta\right],\end{aligned}$$

where in $(*)$ we applied Hölder's inequality. By Markov's inequality,

$$\begin{aligned}\mathbb{E}\left[\eta^{\frac{\alpha}{1-\alpha}}\right] &= \mathbb{E}\left[\eta\right] = \mathbb{P}\left\{\|X-x\| > \frac{\lambda}{2}\right\} = \mathbb{P}\left\{\|X-x\|^\alpha > \frac{\lambda^\alpha}{2^\alpha}\right\} \\ &\leq \frac{2^\alpha}{\lambda^\alpha}\mathbb{E}\left[\|X-x\|^\alpha\right] \\ &\leq \left(\frac{2\sigma}{\lambda}\right)^\alpha. \end{aligned} \tag{47}$$

Thus, in combination with the previous chain of inequalities, we finally have

$$\left\|\mathbb{E}\left[\widetilde{X}\right] - x\right\| \leq \sigma\left(\frac{2\sigma}{\lambda}\right)^{\alpha-1} + \frac{\lambda}{2}\left(\frac{2\sigma}{\lambda}\right)^\alpha = \frac{2^\alpha\sigma^\alpha}{\lambda^{\alpha-1}}.$$

Proof of (44): Using $\|\widetilde{X} - x\| \leq \|\widetilde{X}\| + \|x\| \leq \lambda + \frac{\lambda}{2} = \frac{3\lambda}{2}$, we have

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{X}-x\|^2\right] &= \mathbb{E}\left[\|\widetilde{X}-x\|^\alpha\|\widetilde{X}-x\|^{2-\alpha}\right] \\ &\leq \left(\frac{3\lambda}{2}\right)^{2-\alpha}\mathbb{E}\left[\|\widetilde{X}-x\|^\alpha\chi + \|\widetilde{X}-x\|^\alpha(1-\chi)\right] \\ &= \left(\frac{3\lambda}{2}\right)^{2-\alpha}\mathbb{E}\left[\chi\left\|\frac{\lambda}{\|X\|}X - x\right\|^\alpha + \|X-x\|^\alpha(1-\chi)\right] \\ &\leq \left(\frac{3\lambda}{2}\right)^{2-\alpha}\mathbb{E}\left[\chi\left(\left\|\frac{\lambda}{\|X\|}X\right\| + \|x\|\right)^\alpha + \|X-x\|^\alpha(1-\chi)\right] \\ &\overset{\|x\|\leq\frac{\lambda}{2}}{\leq} \left(\frac{3\lambda}{2}\right)^{2-\alpha}\left(\mathbb{E}\left[\chi\left(\frac{3\lambda}{2}\right)^\alpha + \sigma^\alpha\right]\right),\end{aligned}$$

where in the last inequality we applied (42) and $1-\chi \leq 1$. By (47) and $\chi \leq \eta$, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{X}-x\|^2\right] &\leq \frac{9\lambda^2}{4}\left(\frac{2\sigma}{\lambda}\right)^\alpha + \left(\frac{3\lambda}{2}\right)^{2-\alpha}\sigma^\alpha \\ &\leq \frac{9\lambda^2}{4}\left(2^\alpha + \frac{2^\alpha}{3^\alpha}\right)\frac{\sigma^\alpha}{\lambda^\alpha} \\ &\leq 18\lambda^{2-\alpha}\sigma^\alpha.\end{aligned}$$

Proof of (45): Using variance decomposition and (44), we have

$$\mathbb{E}\left[\left\|\widetilde{X} - \mathbb{E}[\widetilde{X}]\right\|^2\right] \leq \mathbb{E}\left[\|\widetilde{X}-x\|^2\right] \leq 18\lambda^{2-\alpha}\sigma^\alpha.$$

□

D. Proof of Theorem 2.1

In this section, we give an example of the problem for which SGD without clipping leads to a weak high-probability convergence guarantee even under the strong assumption of bounded variance. Theorem below formally states our result, showing that, in the worst-case, the bound for SGD scales worse than that of clipped-SGD in terms of the probability β.

Theorem D.1. *For any $\varepsilon > 0$, $\beta \in (0,1)$, and* SGD *parameterized by the number of steps K and stepsize γ, there exists problem (2) such that Assumptions 1.1, 1.3, and 1.6 hold with $\alpha = 2$, $0 < \mu \leq L$ and for the iterates produced by* SGD *with any stepsize $0 < \gamma \leq 1/\mu$*

$$\mathbb{P}\left\{\|x^K - x^*\|^2 \geq \varepsilon\right\} \leq \beta \implies K = \Omega\left(\frac{\sigma}{\mu\sqrt{\beta\varepsilon}}\right). \tag{48}$$

Proof. To prove the above theorem, we consider the simple one-dimensional problem $f(x) = \mu x^2/2$. It is easy to see that the considered problem is μ-strongly convex, μ-smooth, and has optimum at $x^* = 0$. We construct the noise in an adversarial way with respect to the parameters of the SGD. Concretely, the noise depends on the number of iterates N, stepsize γ, target precision ε, the starting point x^0, and bound on the variance σ^2 such that

$$\nabla f_{\xi_k}(x^k) = \mu x^k - \sigma z_k,$$

where

$$z_k = \begin{cases} 0, & \text{if } k < K-1 \text{ or } (1-\gamma\mu)^K|x^0| > \sqrt{\varepsilon}, \\ \begin{cases} -A, & \text{with probability } \frac{1}{2A^2}, \\ 0 & \text{with probability } 1 - \frac{1}{A^2}, \\ A, & \text{with probability } \frac{1}{2A^2}, \end{cases} & \text{otherwise} \end{cases} \quad \forall k \in \{0,1,\ldots,K-1\}, \tag{49}$$

where $A = \max\left\{\frac{2\sqrt{\varepsilon}}{\gamma\sigma}, 1\right\}$. We note that $\mathbb{E}\left[z^k\right] = 0$. Therefore, $\mathbb{E}\left[\nabla f_{\xi_k}(x^k)\right] = \mu x^k = \nabla f(x^k)$. Furthermore,

$$\mathrm{Var}\left[z^k\right] = \mathbb{E}\left[(z^k)^2\right] \leq \frac{1}{2A^2}A^2 + \frac{1}{2A^2}A^2 = 1,$$

which implies that Assumption 1.1 holds for $\alpha = 2$. We note that our construction depends on the parameters of the algorithm and the target value ε. However, our analysis of the methods with clipping works in such generality.

Let us now analyze the properties of the introduced problem. We are interested in the situation when

$$\mathbb{P}\left\{\|x^K - x^*\|^2 \geq \varepsilon\right\} \leq \beta$$

for $\beta \in (0,1)$. We first prove that this implies that $(1-\gamma\mu)^K|x^0| \leq \sqrt{\varepsilon}$. To do that we proceed by contradiction and assume that

$$(1-\gamma\mu)^K|x^0| > \sqrt{\varepsilon}. \tag{50}$$

By construction, this implies that $z_k = 0,\ \forall k \in \{0,1,\ldots,K\}$. This, in turn, implies that $x^K = (1-\gamma\mu)^K x^0$, and, further, by (50) and since $x^* = 0$, that

$$\mathbb{P}\left\{\|x^K - x^*\|^2 \geq \varepsilon\right\} = \mathbb{P}\left\{\|x^K\|^2 \geq \varepsilon\right\} = 1.$$

Thus, the contradiction shows that $(1-\gamma\mu)^K|x^0| \leq \sqrt{\varepsilon}$, which yields $K \geq \frac{\ln\frac{\sqrt{\varepsilon}}{|x^0|}}{\ln(1-\gamma\mu)} \geq \frac{\ln\frac{\sqrt{\varepsilon}}{|x^0|}}{-\gamma\mu} \geq \frac{1}{\gamma\mu} = \ln\frac{|x^0|}{\sqrt{\varepsilon}}$. Using (49) with $K \geq \frac{1}{\gamma\mu}\log\frac{|x^0|}{\sqrt{\varepsilon}}$, we obtain

$$\|x^K - x^*\|^2 = ((1-\gamma\mu)^K x^0 + \gamma\sigma z_K)^2.$$

Furthermore,

$$\begin{aligned}
\mathbb{P}\left\{\|x^K - x^*\|^2 \geq \varepsilon\right\} &= \mathbb{P}\left\{\left|(1-\gamma\mu)^K x^0 + \gamma\sigma z_K\right| \geq \sqrt{\varepsilon}\right\} \\
&= \mathbb{P}\left\{\gamma\sigma z_K \geq \sqrt{\varepsilon} - (1-\gamma\mu)^K x^0\right\} + \mathbb{P}\left\{\gamma\sigma z_K \leq -\sqrt{\varepsilon} - (1-\gamma\mu)^K x^0\right\} \\
&\geq \mathbb{P}\left\{\gamma\sigma z_K \geq \sqrt{\varepsilon} + (1-\gamma\mu)^K x^0\right\} + \mathbb{P}\left\{\gamma\sigma z_K \leq -\sqrt{\varepsilon} - (1-\gamma\mu)^K x^0\right\} \\
&= \mathbb{P}\left\{|\gamma\sigma z_K| \geq \sqrt{\varepsilon} + (1-\gamma\mu)^K x^0\right\} \\
&\geq \mathbb{P}\left\{|\gamma\sigma z_K| \geq 2\sqrt{\varepsilon}\right\} = \mathbb{P}\left\{|z_K| \geq \frac{2\sqrt{\varepsilon}}{\gamma\sigma}\right\}.
\end{aligned}$$

Now if $\frac{2\sqrt{\varepsilon}}{\gamma\sigma} < 1$ then $A = 1$. Therefore,

$$1 = \mathbb{P}\left\{|z_K| \geq \frac{2\sqrt{\varepsilon}}{\gamma\sigma}\right\} \leq \mathbb{P}\left\{\|x^K - x^*\|^2 > \varepsilon\right\} < \beta,$$

yielding contradiction, which implies that if $\mathbb{P}\left\{\|x^K - x^*\|^2 > \varepsilon\right\} < \beta$ for our constructed problem, then $\frac{2\sqrt{\varepsilon}}{\gamma\sigma} \geq 1$, i.e., $\gamma \leq \frac{2\sqrt{\varepsilon}}{\sigma}$. For $\gamma \leq \frac{2\sqrt{\varepsilon}}{\sigma}$, we have

$$\beta \geq \mathbb{P}\left\{\|x^K - x^*\|^2 \geq \varepsilon\right\} \geq \mathbb{P}\left\{|z_K| \geq \frac{2\sqrt{\varepsilon}}{\gamma\sigma}\right\} = \frac{1}{A^2} = \frac{\gamma^2\sigma^2}{4\varepsilon}.$$

This implies that $\gamma \leq \frac{2\sqrt{\beta\varepsilon}}{\sigma}$. Combining this inequality with $K \geq \frac{1}{\gamma\mu}\log\frac{|x^0|}{\sqrt{\varepsilon}}$ yields

$$K \geq \frac{\sigma}{2\mu\sqrt{\beta\varepsilon}}\log\frac{|x^0|}{\sqrt{\varepsilon}}$$

and concludes the proof. □

E. Missing Proofs for clipped-SGD

In this section, we provide all the missing details and proofs of the results for clipped-SGD. For brevity, we will use the following notation: $\widetilde{\nabla} f_{\xi^k}(x^k) = \texttt{clip}(\nabla f_{\xi^k}(x^k), \lambda_k)$.

Algorithm 1 Clipped Stochastic Gradient Descent (clipped-SGD) (Pascanu et al., 2013)

Input: starting point x^0, number of iterations K, stepsize $\gamma > 0$, clipping levels $\{\lambda_k\}_{k=0}^{K-1}$.
1: **for** $k = 0, \ldots, K-1$ **do**
2: Compute $\widetilde{\nabla} f_{\xi^k}(x^k) = \texttt{clip}\left(\nabla f_{\xi^k}(x^k), \lambda_k\right)$ using a fresh sample $\xi^k \sim \mathcal{D}_k$
3: $x^{k+1} = x^k - \gamma \widetilde{\nabla} f_{\xi^k}(x^k)$
4: **end for**
Output: x^K

E.1. Non-Convex Functions

We start the analysis of clipped-SGD in the non-convex case with the following lemma that follows the proof of deterministic GD and separates the stochasticity from the determinisitc part of the method.

Lemma E.1. *Let Assumptions 1.2 and 1.3 hold on $Q = \{x \in \mathbb{R}^d \mid \exists y \in \mathbb{R}^d : \ f(y) \leq f_* + 2\Delta$ and $\|x - y\| \leq \sqrt{\Delta}/20\sqrt{L}\}$, where $\Delta \geq \Delta_0 = f(x^0) - f_*$, and let stepsize γ satisfy $\gamma < \frac{2}{L}$. If $x^k \in Q$ for all $k = 0, 1, \ldots, K$, $K \geq 0$, then after K iterations of* clipped-SGD *we have*

$$\gamma\left(1 - \frac{L\gamma}{2}\right)\sum_{k=0}^{K-1}\|\nabla f(x^k)\|^2 \leq (f(x^0) - f_*) - (f(x^K) - f_*) - \gamma(1 - L\gamma)\sum_{k=0}^{K-1}\langle \nabla f(x^k), \theta_k\rangle + \frac{L\gamma^2}{2}\sum_{k=0}^{K-1}\|\theta_k\|^2, \tag{51}$$

$$\theta_k \overset{def}{=} \widetilde{\nabla} f_{\xi^k}(x^k) - \nabla f(x^k). \tag{52}$$

Proof. Using $x^{k+1} = x^k - \gamma\widetilde{\nabla} f_{\xi^k}(x^k)$ and smoothness of f (1.3) we get that for all $k = 0, 1, \ldots, K-1$

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\ &= f(x^k) - \gamma\langle \nabla f(x^k), \widetilde{\nabla} f_{\xi^k}(x^k)\rangle + \frac{L\gamma^2}{2}\|\widetilde{\nabla} f_{\xi^k}(x^k)\|^2 \\ &\overset{(52)}{=} f(x^k) - \gamma\|\nabla f(x^k)\|^2 - \gamma\langle \nabla f(x^k), \theta_k\rangle + \frac{L\gamma^2}{2}\|\theta_k\|^2 \\ &\quad + \frac{L\gamma^2}{2}\|\nabla f(x^k)\|^2 + L\gamma^2\langle \nabla f(x^k), \theta_k\rangle \\ &= f(x^k) - \gamma\left(1 - \frac{L\gamma}{2}\right)\|\nabla f(x^k)\|^2 - \gamma(1 - L\gamma)\langle \nabla f(x^k), \theta_k\rangle + \frac{L\gamma^2}{2}\|\theta_k\|^2. \end{aligned}$$

We rearrange the terms and get

$$\gamma\left(1 - \frac{L\gamma}{2}\right)\|\nabla f(x^k)\|^2 \leq f(x^k) - f(x^{k+1}) - \gamma(1 - L\gamma)\langle \nabla f(x^k), \theta_k\rangle + \frac{L\gamma^2}{2}\|\theta_k\|^2.$$

Finally, summing up these inequalities for $k = 0, \ldots, K-1$, we get

$$\begin{aligned}\gamma\left(1-\frac{L\gamma}{2}\right)\sum_{k=0}^{K-1}\|\nabla f(x^k)\|^2 &\leq \sum_{k=0}^{K-1}\left(f(x^k)-f(x^{k+1})\right)-\gamma(1-L\gamma)\sum_{k=0}^{K-1}\langle\nabla f(x^k),\theta_k\rangle\\ &\quad +\frac{L\gamma^2}{2}\sum_{k=0}^{K-1}\|\theta_k\|^2\\ &= (f(x^0)-f_*)-(f(x^K)-f_*)-\gamma(1-L\gamma)\sum_{k=0}^{K-1}\langle\nabla f(x^k),\theta_k\rangle\\ &\quad +\frac{L\gamma^2}{2}\sum_{k=0}^{K-1}\|\theta_k\|^2,\end{aligned}$$

which concludes the proof. □

Using this lemma, we prove the main convergence result for clipped-SGD in the non-convex case.

Theorem E.2. *Let Assumptions 1.1, 1.2, 1.3 hold on $Q = \{x \in \mathbb{R}^d \mid \exists y \in \mathbb{R}^d: \ f(y) \leq f_* + 2\Delta$ and $\|x-y\| \leq \sqrt{\Delta}/20\sqrt{L}\}$, where $\Delta \geq \Delta_0 = f(x^0) - f_*$, stepsize*

$$\gamma \leq \min\left\{\frac{1}{80L\ln\frac{4(K+1)}{\beta}}, \frac{\sqrt{\Delta}}{27^{\frac{1}{\alpha}}20\sigma\sqrt{L}K^{\frac{1}{\alpha}}\left(\ln\frac{4(K+1)}{\beta}\right)^{\frac{\alpha-1}{\alpha}}}\right\}, \tag{53}$$

and clipping level

$$\lambda_k = \lambda = \frac{\sqrt{\Delta}}{20\sqrt{L}\gamma\ln\frac{4(K+1)}{\beta}}, \tag{54}$$

for some $K \geq 0$ and $\beta \in (0,1]$ such that $\ln\frac{4(K+1)}{\beta} \geq 1$. Then, after K iterations of clipped-SGD *the iterates with probability at least $1-\beta$ satisfy*

$$\frac{1}{K+1}\sum_{k=0}^{K}\|\nabla f(x^k)\|^2 \leq \frac{2\Delta}{\gamma\left(1-\frac{L\gamma}{2}\right)(K+1)}. \tag{55}$$

In particular, when γ equals the minimum from (53), then the iterates produced by clipped-SGD *after K iterations with probability at least $1-\beta$ satisfy*

$$\frac{1}{K+1}\sum_{k=0}^{K}\|\nabla f(x^k)\|^2 = \mathcal{O}\left(\max\left\{\frac{L\Delta\ln\frac{K}{\beta}}{K}, \frac{\sqrt{L\Delta}\sigma\ln^{\frac{\alpha-1}{\alpha}}\frac{K}{\beta}}{K^{\frac{\alpha-1}{\alpha}}}\right\}\right), \tag{56}$$

meaning that to achieve $\frac{1}{K+1}\sum\limits_{k=0}^{K}\|\nabla f(x^k)\|^2 \leq \varepsilon$ with probability at least $1-\beta$ clipped-SGD *requires*

$$K = \mathcal{O}\left(\max\left\{\frac{L\Delta}{\varepsilon}\ln\frac{L\Delta}{\beta\varepsilon}, \left(\frac{\sqrt{L\Delta}\sigma}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\ln\left(\frac{1}{\beta}\left(\frac{\sqrt{L\Delta}\sigma}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right\}\right) \quad \text{iterations/oracle calls.} \tag{57}$$

Proof. Let $\Delta_k = f(x^k) - f_*$ for all $k \geq 0$. Next, our goal is to show by induction that $\Delta_l \leq 2\Delta$ with high probability, which allows to apply the result of Lemma E.1 and then use Bernstein's inequality to estimate the stochastic part of the upper-bound. More precisely, for each $k = 0, \ldots, K+1$ we consider probability event E_k defined as follows: inequalities

$$\frac{L\gamma^2}{2}\sum_{l=0}^{t-1}\|\theta_l\|^2 - \gamma(1-L\gamma)\sum_{l=1}^{t-1}\langle\nabla f(x^l),\theta_l\rangle \leq \Delta, \tag{58}$$

$$\Delta_t \leq 2\Delta \tag{59}$$

hold for all $t = 0, 1, \ldots, k$ simultaneously. We want to prove via induction that $\mathbb{P}\{E_k\} \geq 1 - k\beta/(K+1)$ for all $k = 0, 1, \ldots, K+1$. For $k = 0$ the statement is trivial. Assume that the statement is true for some $k = T - 1 \leq K$: $\mathbb{P}\{E_{T-1}\} \geq 1 - (T-1)\beta/(K+1)$. One needs to prove that $\mathbb{P}\{E_T\} \geq 1 - T\beta/(K+1)$. First, we notice that probability event E_{T-1} implies that $\Delta_t \leq 2\Delta$ for all $t = 0, 1, \ldots, T-1$, i.e., $x^t \in \{y \in \mathbb{R}^d \mid f(y) \leq f_* + 2\Delta\}$ for $t = 0, 1, \ldots, T-1$. Moreover, due to the choice of clipping level λ we have

$$\|x^T - x^{T-1}\| = \gamma\|\widetilde{\nabla} f_{\xi^{T-1}}(x^{T-1})\| \leq \gamma\lambda \overset{(54)}{=} \frac{\sqrt{\Delta}}{20\sqrt{L}\ln\frac{4K}{\beta}} \leq \frac{\sqrt{\Delta}}{20\sqrt{L}}.$$

Therefore, E_{T-1} implies $\{x^k\}_{k=0}^T \subseteq Q$, meaning that the assumptions of Lemma E.1 are satisfied and we have

$$A\sum_{l=0}^{t-1}\|\nabla f(x^l)\|^2 \quad\leq\quad \Delta_0 - \Delta_t - \gamma(1 - L\gamma)\sum_{k=0}^{t-1}\langle\nabla f(x^l), \theta_l\rangle + \frac{L\gamma^2}{2}\sum_{l=0}^{t-1}\|\theta_l\|^2, \tag{60}$$

$$A \quad\overset{\text{def}}{=}\quad \gamma\left(1 - \frac{L\gamma}{2}\right) \overset{(53)}{\geq} 0 \tag{61}$$

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

$$\sum_{l=0}^{t-1}\|\nabla f(x^l)\|^2 \overset{(60)}{\leq} \frac{1}{A}\left(\Delta - \gamma(1 - L\gamma)\sum_{k=0}^{t-1}\langle\nabla f(x^l), \theta_l\rangle + \frac{L\gamma^2}{2}\sum_{l=0}^{t-1}\|\theta_l\|^2\right) \overset{(58),}{\leq} \frac{2\Delta}{A}. \tag{62}$$

Taking into account that $A\sum\limits_{t=0}^{T-1}\|\nabla f(x^t)\|^2 \geq 0$, we also derive that E_{T-1} implies

$$\Delta_T \leq \Delta + \frac{L\gamma^2}{2}\sum_{t=0}^{T-1}\|\theta_t\|^2 - \gamma(1 - L\gamma)\sum_{t=0}^{T-1}\langle\nabla f(x^t), \theta_t\rangle. \tag{63}$$

Next, we define random vectors

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

for all $t = 0, 1, \ldots, T-1$. By definition these random vectors are bounded with probability 1

$$\|\eta_t\| \leq 2\sqrt{L\Delta}. \tag{64}$$

Moreover, for $t = 1, \ldots, T-1$ event E_{T-1} implies

$$\|\nabla f(x^l)\| \overset{(7)}{\leq} \sqrt{2L(f(x^l) - f_*)} = \sqrt{2L\Delta_l} \leq 2\sqrt{L\Delta} \overset{(53),(54)}{\leq} \frac{\lambda}{2}, \tag{65}$$

meaning that E_{T-1} implies that $\eta_t = \nabla f(x^t)$ for all $t = 0, 1, \ldots, T-1$. Next, we define the unbiased part and the bias of θ_t as θ_t^u and θ_t^b, respectively:

$$\theta_t^u = \widetilde{\nabla} f_{\xi^t}(x^t) - \mathbb{E}_{\xi^t}\left[\widetilde{\nabla} f_{\xi^t}(x^t)\right], \quad \theta_t^b = \mathbb{E}_{\xi^t}\left[\widetilde{\nabla} f_{\xi^t}(x^t)\right] - \nabla f(x^t). \tag{66}$$

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

$$\begin{aligned}\Delta_T \quad\leq\quad & \underbrace{\Delta - \gamma(1 - L\gamma)\sum_{t=0}^{T-1}\langle\theta_t^u, \eta_t\rangle}_{①} \underbrace{- \gamma(1 - L\gamma)\sum_{t=0}^{T-1}\langle\theta_t^b, \eta_t\rangle}_{②} + \underbrace{L\gamma^2\sum_{t=0}^{T-1}\left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]\right)}_{③}\\ & + \underbrace{L\gamma^2\sum_{t=0}^{T-1}\mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]}_{④} + \underbrace{L\gamma^2\sum_{t=0}^{T-1}\left\|\theta_t^b\right\|^2}_{⑤}.\end{aligned} \tag{67}$$

It remains to derive good enough high-probability upper-bounds for the terms ①, ②, ③, ④, ⑤, i.e., to finish our inductive proof we need to show that ① + ② + ③ + ④ + ⑤ $\leq \Delta$ with high probability. In the subsequent parts of the proof, we will need to use many times the bounds for the norm and second moments of θ_t^u and θ_t^b. First, by definition of clipping operator, we have with probability 1 that

$$\|\theta_t^u\| \leq 2\lambda. \tag{68}$$

Moreover, since E_{T-1} implies that $\|\nabla f(x^t)\| \leq \lambda/2$ for $t = 0, 1, \ldots, T-1$ (see (65)), then, in view of Lemma 5.1, we have that E_{T-1} implies

$$\|\theta_t^b\| \leq \frac{2^\alpha \sigma^\alpha}{\lambda^{\alpha-1}}, \tag{69}$$

$$\mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right] \leq 18\lambda^{2-\alpha}\sigma^\alpha. \tag{70}$$

Upper bound for ①. By definition of θ_t^u, we have $\mathbb{E}_{\xi^t}[\theta_t^u] = 0$ and

$$\mathbb{E}_{\xi^t}\left[-\gamma(1-L\gamma)\langle\theta_t^u, \eta_t\rangle\right] = 0.$$

Next, sum ① has bounded with probability 1 terms:

$$|\gamma(1-L\gamma)\langle\theta_t^u, \eta_t\rangle| \overset{(53)}{\leq} \gamma\|\theta_t^u\| \cdot \|\eta_t\| \overset{(64),(68)}{\leq} 4\gamma\lambda\sqrt{L\Delta} \overset{(54)}{=} \frac{\Delta}{5\ln\frac{4(K+1)}{\beta}} \overset{\text{def}}{=} c. \tag{71}$$

The summands also have bounded conditional variances $\sigma_t^2 \overset{\text{def}}{=} \mathbb{E}_{\xi^t}[\gamma^2(1-L\gamma)^2\langle\theta_t^u, \eta_t\rangle^2]$:

$$\sigma_t^2 \leq \mathbb{E}_{\xi^t}\left[\gamma^2(1-L\gamma)^2\|\theta_t^u\|^2 \cdot \|\eta_t\|^2\right] \overset{(64)}{\leq} 4\gamma^2(1-L\gamma)^2 L\Delta\mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right] \overset{(53)}{\leq} 4\gamma^2 L\Delta\mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]. \tag{72}$$

In other words, we showed that $\{-\gamma(1-L\gamma)\langle\theta_t^u, \eta_t\rangle\}_{t=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\sigma_t^2\}_{t=0}^{T-1}$. Next, we apply Bernstein's inequality (Lemma B.2) with $X_t = -\gamma(1-L\gamma)\langle\theta_t^u, \eta_t\rangle$, parameter c as in (71), $b = \frac{\Delta}{5}$, $G = \frac{\Delta^2}{150\ln\frac{4(K+1)}{\beta}}$:

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

Equivalently, we have

$$\mathbb{P}\{E_①\} \geq 1 - \frac{\beta}{2(K+1)}, \quad \text{for} \quad E_① = \left\{\text{either} \quad \sum_{t=0}^{T-1}\sigma_t^2 > \frac{\Delta^2}{150\ln\frac{4(K+1)}{\beta}} \quad \text{or} \quad |①| \leq \frac{\Delta}{5}\right\}. \tag{73}$$

In addition, E_{T-1} implies that

$$\begin{aligned}\sum_{t=0}^{T-1}\sigma_t^2 &\overset{(72)}{\leq} 4\gamma^2 L\Delta\sum_{t=0}^{T-1}\mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right] \overset{(70)}{\leq} 72\gamma^2 L\Delta\sigma^\alpha T\lambda^{2-\alpha} \\ &\overset{(54)}{=} \frac{9\cdot 20^\alpha\sqrt{\Delta}^{4-\alpha}\sigma^\alpha T\sqrt{L}^\alpha\gamma^\alpha}{50\ln^{2-\alpha}\frac{4(K+1)}{\beta}} \overset{(53)}{\leq} \frac{\Delta^2}{150\ln\frac{4(K+1)}{\beta}}.\end{aligned} \tag{74}$$

Upper bound for ②. From E_{T-1} it follows that

$$\begin{aligned}② &= -\gamma(1-L\gamma)\sum_{t=0}^{T-1}\langle\theta_t^b, \eta_t\rangle \overset{(53)}{\leq} \gamma\sum_{t=0}^{T-1}\|\theta_t^b\| \cdot \|\eta_t\| \overset{(64),(69)}{\leq} \frac{2\cdot 2^\alpha\gamma\sigma^\alpha T\sqrt{L\Delta}}{\lambda^{\alpha-1}} \\ &\overset{(54)}{=} \frac{40^\alpha}{10}\cdot\frac{\sigma^\alpha T\sqrt{\Delta}^{2-\alpha}\sqrt{L}^\alpha\gamma^\alpha}{\ln^{1-\alpha}\frac{4(K+1)}{\beta}} \overset{(53)}{\leq} \frac{\Delta}{5}.\end{aligned} \tag{75}$$

Upper bound for ③. First, we have

$$\mathbb{E}_{\xi^t}\left[L\gamma^2\left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]\right)\right] = 0.$$

Next, sum ③ has bounded with probability 1 terms:

$$\begin{aligned}\left|L\gamma^2\left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]\right)\right| &\leq L\gamma^2\left(\|\theta_t^u\|^2 + \mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]\right) \\ &\overset{(68)}{\leq} 8L\gamma^2\lambda^2 \overset{(54)}{=} \frac{\Delta}{50\ln^2\frac{4(K+1)}{\beta}} \leq \frac{\Delta}{5\ln\frac{4(K+1)}{\beta}} \overset{\text{def}}{=} c. \end{aligned} \tag{76}$$

The summands also have bounded conditional variances $\widetilde{\sigma}_t^2 \overset{\text{def}}{=} \mathbb{E}_{\xi^t}\left[L^2\gamma^4\left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]\right)^2\right]$:

$$\widetilde{\sigma}_t^2 \overset{(76)}{\leq} \frac{\Delta}{5\ln\frac{4(K+1)}{\beta}}\mathbb{E}_{\xi^t}\left[L\gamma^2\left|\|\theta_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]\right|\right] \leq \frac{2L\gamma^2\Delta}{5\ln\frac{4(K+1)}{\beta}}\mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right], \tag{77}$$

since $\ln\frac{4K}{\beta} \geq 1$. In other words, we showed that $\left\{L\gamma^2\left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]\right)\right\}_{t=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\widetilde{\sigma}_t^2\}_{t=0}^{T-1}$. Next, we apply Bernstein's inequality (Lemma B.2) with $X_t = L\gamma^2\left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]\right)$, parameter c as in (76), $b = \frac{\Delta}{5}$, $G = \frac{\Delta^2}{150\ln\frac{4(K+1)}{\beta}}$:

$$\mathbb{P}\left\{|③| > \frac{\Delta}{5} \quad \text{and} \quad \sum_{t=0}^{T-1}\widetilde{\sigma}_t^2 \leq \frac{\Delta^2}{150\ln\frac{4(K+1)}{\beta}}\right\} \leq 2\exp\left(-\frac{b^2}{2G + 2cb/3}\right) = \frac{\beta}{2(K+1)}.$$

Equivalently, we have

$$\mathbb{P}\left\{E_③\right\} \geq 1 - \frac{\beta}{2(K+1)}, \quad \text{for} \quad E_③ = \left\{\text{either} \quad \sum_{t=0}^{T-1}\widetilde{\sigma}_t^2 > \frac{\Delta^2}{150\ln\frac{4(K+1)}{\beta}} \quad \text{or} \quad |③| \leq \frac{\Delta}{5}\right\}. \tag{78}$$

In addition, E_{T-1} implies that

$$\begin{aligned}\sum_{t=0}^{T-1}\widetilde{\sigma}_t^2 &\overset{(77)}{\leq} \frac{2L\gamma^2\Delta}{5\ln\frac{4(K+1)}{\beta}}\sum_{t=0}^{T-1}\mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right] \overset{(70)}{\leq} \frac{36L\gamma^2\Delta\lambda^{2-\alpha}\sigma^\alpha T}{5\ln\frac{4(K+1)}{\beta}} \\ &\overset{(54)}{=} \frac{9\cdot 20^\alpha}{500}\cdot\frac{\sigma^\alpha T\sqrt{\Delta}^{4-\alpha}\sqrt{L}^\alpha\gamma^\alpha}{\ln^{3-\alpha}\frac{4(K+1)}{\beta}} \overset{(53)}{\leq} \frac{\Delta^2}{150\ln\frac{4(K+1)}{\beta}}. \end{aligned} \tag{79}$$

Upper bound for ④. From E_{T-1} it follows that

$$④ = L\gamma^2\sum_{t=0}^{T-1}\mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right] \overset{(70)}{\leq} 18L\gamma^2\lambda^{2-\alpha}\sigma^\alpha T \overset{(54)}{=} \frac{9\cdot 20^\alpha}{200}\cdot\frac{\sqrt{L}^\alpha\gamma^\alpha\sigma^\alpha T\sqrt{\Delta}^{2-\alpha}}{\ln^{2-\alpha}\frac{4(K+1)}{\beta}} \overset{(53)}{\leq} \frac{\Delta}{5}. \tag{80}$$

Upper bound for ⑤. From E_{T-1} it follows that

$$⑤ = L\gamma^2\sum_{t=0}^{T-1}\|\theta_t^b\|^2 \overset{(69)}{\leq} \frac{4^\alpha\sigma^{2\alpha}TL\gamma^2}{\lambda^{2(\alpha-1)}} \overset{(54)}{=} \frac{1600^\alpha}{400}\cdot\frac{\sigma^{2\alpha}TL^\alpha\gamma^{2\alpha}\Delta^{1-\alpha}}{\ln^{2(1-\alpha)}\frac{4(K+1)}{\beta}} \overset{(53)}{\leq} \frac{\Delta}{5}. \tag{81}$$

Now, we have the upper bounds for ①, ②, ③, ④, ⑤. In particular, probability event E_{T-1} implies

$$\Delta_T \overset{(67)}{\leq} \Delta + ① + ② + ③ + ④ + ⑤,$$

$$② \overset{(75)}{\leq} \frac{\Delta}{5}, \quad ④ \overset{(80)}{\leq} \frac{\Delta}{5}, \quad ⑤ \overset{(81)}{\leq} \frac{\Delta}{5},$$

$$\sum_{t=0}^{T-1}\sigma_t^2 \overset{(74)}{\leq} \frac{\Delta^2}{150\ln\frac{4(K+1)}{\beta}}, \quad \sum_{t=0}^{T-1}\widetilde{\sigma}_t^2 \overset{(79)}{\leq} \frac{\Delta^2}{150\ln\frac{4(K+1)}{\beta}}.$$

Moreover, we also have (see (73), (78) and our induction assumption)

$$\mathbb{P}\{E_{T-1}\} \geq 1 - \frac{(T-1)\beta}{K+1}, \quad \mathbb{P}\{E_{①}\} \geq 1 - \frac{\beta}{2(K+1)}, \quad \mathbb{P}\{E_{③}\} \geq 1 - \frac{\beta}{2(K+1)},$$

where

$$\begin{aligned} E_{①} &= \left\{ \text{either} \quad \sum_{t=0}^{T-1} \sigma_t^2 > \frac{\Delta^2}{150 \ln \frac{4(K+1)}{\beta}} \quad \text{or} \quad |①| \leq \frac{\Delta}{5} \right\}, \\ E_{③} &= \left\{ \text{either} \quad \sum_{t=0}^{T-1} \widetilde{\sigma}_t^2 > \frac{\Delta^2}{150 \ln \frac{4(K+1)}{\beta}} \quad \text{or} \quad |③| \leq \frac{\Delta}{5} \right\}. \end{aligned}$$

Thus, probability event $E_{T-1} \cap E_{①} \cap E_{③}$ implies

$$\Delta_T \quad \leq \quad \Delta + \frac{\Delta}{5} + \frac{\Delta}{5} + \frac{\Delta}{5} + \frac{\Delta}{5} + \frac{\Delta}{5} = 2\Delta,$$

which is equivalent to (58) and (59) for $t = T$, and

$$\mathbb{P}\{E_T\} \geq \mathbb{P}\left\{E_{T-1} \cap E_{①} \cap E_{③}\right\} = 1 - \mathbb{P}\left\{\overline{E}_{T-1} \cup \overline{E}_{①} \cup \overline{E}_{③}\right\} \geq 1 - \mathbb{P}\{\overline{E}_{T-1}\} - \mathbb{P}\{\overline{E}_{①}\} - \mathbb{P}\{\overline{E}_{③}\} \geq 1 - \frac{T\beta}{K+1}.$$

This finishes the inductive part of our proof, i.e., for all $k = 0, 1, \ldots, K+1$ we have $\mathbb{P}\{E_k\} \geq 1 - k\beta/(K+1)$. In particular, for $k = K+1$ we have that with probability at least $1 - \beta$

$$\frac{1}{K+1} \sum_{k=0}^{K} \|\nabla f(x^k)\|^2 \overset{(62)}{\leq} \frac{2\Delta}{A(K+1)} \overset{(53)}{=} \frac{2\Delta}{\gamma\left(1 - \frac{L\gamma}{2}\right)(K+1)}$$

and $\{x^k\}_{k=0}^K \subseteq Q$, which follows from (59).

Finally, if

$$\gamma \leq \min\left\{ \frac{1}{80L \ln \frac{4(K+1)}{\beta}}, \frac{\sqrt{\Delta}}{27^{\frac{1}{\alpha}} 20 \sigma \sqrt{L} K^{\frac{1}{\alpha}} \left(\ln \frac{4(K+1)}{\beta}\right)^{\frac{\alpha-1}{\alpha}}} \right\},$$

then with probability at least $1 - \beta$

$$\begin{aligned} \frac{1}{K+1} \sum_{k=0}^{K} \|\nabla f(x^k)\|^2 &\leq \frac{2\Delta}{\gamma\left(1 - \frac{L\gamma}{2}\right)(K+1)} \leq \frac{4\Delta}{\gamma(K+1)} \\ &= \max\left\{ \frac{320\Delta L \ln \frac{4(K+1)}{\beta}}{K+1}, \frac{80\sqrt{\Delta} 27^{\frac{1}{\alpha}} \sigma \sqrt{L} K^{\frac{1}{\alpha}} \left(\ln \frac{4(K+1)}{\beta}\right)^{\frac{\alpha-1}{\alpha}}}{K+1} \right\} \\ &= \mathcal{O}\left(\max\left\{ \frac{L\Delta \ln \frac{K}{\beta}}{K}, \frac{\sqrt{L\Delta}\sigma \ln^{\frac{\alpha-1}{\alpha}} \frac{K}{\beta}}{K^{\frac{\alpha-1}{\alpha}}} \right\} \right). \end{aligned}$$

To get $\frac{1}{K+1} \sum\limits_{k=0}^{K} \|\nabla f(x^k)\|^2 \leq \varepsilon$ with probability at least $1 - \beta$ it is sufficient to choose K such that both terms in the maximum above are $\mathcal{O}(\varepsilon)$. This leads to

$$K = \mathcal{O}\left(\max\left\{ \frac{L\Delta}{\varepsilon} \ln \frac{L\Delta}{\varepsilon\beta}, \left(\frac{\sqrt{L\Delta}\sigma}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}} \ln\left(\frac{1}{\beta} \left(\frac{\sqrt{L\Delta}\sigma}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}} \right) \right\} \right),$$

which concludes the proof. □

E.2. Polyak-Łojasiewicz Functions

In this subsection, we provide a high-probability analysis of clipped-SGD in the case of Polyak-Łojasiewicz functions. As in the non-convex case, we start with the lemma that handles optimization part of the algorithm and separates it from the stochastic one.

Lemma E.3. *Let Assumptions 1.3 and 1.4 hold on $Q = \{x \in \mathbb{R}^d \mid \exists y \in \mathbb{R}^d: \ f(y) \leq f_* + 2\Delta$ and $\|x - y\| \leq \sqrt{\Delta}/20\sqrt{L}\}$, where $\Delta = f(x^0) - f_*$, and let stepsize γ satisfy $\gamma \leq \frac{1}{L}$. If $x^k \in Q$ for all $k = 0, 1, \ldots, K+1$, $K \geq 0$, then after K iterations of* clipped-SGD *for all $x \in Q$ we have*

$$
\begin{aligned}
f(x^{K+1}) - f_* \quad &\leq \quad (1-\gamma\mu)^{K+1}(f(x^0) - f_*) - \gamma(1 - L\gamma)\sum_{k=0}^{K}(1-\gamma\mu)^{K-k}\langle \nabla f(x^k), \theta_k\rangle \\
&\quad + \frac{L\gamma^2}{2}\sum_{k=0}^{K}(1-\gamma\mu)^{K-k}\|\theta_k\|^2, \qquad (82)
\end{aligned}
$$

where θ_k is defined in (52).

Proof. Using $x^{k+1} = x^k - \gamma\widetilde{\nabla} f_{\xi^k}(x^k)$ and smoothness of f (1.3) we get that for all $k = 0, 1, \ldots, K$

$$
\begin{aligned}
f(x^{k+1}) \quad &\leq \quad f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\
&\leq \quad f(x^k) - \gamma\langle \nabla f(x^k), \widetilde{\nabla} f_{\xi^k}(x^k)\rangle + \frac{L\gamma^2}{2}\|\widetilde{\nabla} f_{\xi^k}(x^k)\|^2 \\
&\overset{(52)}{=} \quad f(x^k) - \gamma\left(1 - \frac{L\gamma}{2}\right)\|\nabla f(x^k)\|^2 - \gamma(1 - L\gamma)\langle \nabla f(x^k), \theta_k\rangle + \frac{L\gamma^2}{2}\|\theta_k\|^2 \\
&\overset{\gamma \leq \frac{1}{L}}{\leq} \quad f(x^k) - \frac{\gamma}{2}\|\nabla f(x^k)\|^2 - \gamma(1 - L\gamma)\langle \nabla f(x^k), \theta_k\rangle + \frac{L\gamma^2}{2}\|\theta_k\|^2 \\
&\overset{(8)}{\leq} \quad f(x^k) - \gamma\mu(f(x^k) - f_*) - \gamma(1 - L\gamma)\langle \nabla f(x^k), \theta_k\rangle + \frac{L\gamma^2}{2}\|\theta_k\|^2.
\end{aligned}
$$

By rearranging the terms and subtracting f_*, we obtain

$$
f(x^{k+1}) - f_* \quad \leq \quad (1-\gamma\mu)(f(x^k) - f_*) - \gamma(1 - L\gamma)\langle \nabla f(x^k), \theta_k\rangle + \frac{L\gamma^2}{2}\|\theta_k\|^2.
$$

Unrolling the recurrence, we obtain (82). □

Theorem E.4. *Let Assumptions 1.1, 1.3, 1.4 hold on $Q = \{x \in \mathbb{R}^d \mid \exists y \in \mathbb{R}^d: \ f(y) \leq f_* + 2\Delta$ and $\|x - y\| \leq \sqrt{\Delta}/20\sqrt{L}\}$, where $\Delta \geq \Delta_0 = f(x^0) - f_*$, stepsize*

$$
0 < \gamma \quad \leq \quad \min\left\{\frac{1}{250L\ln\frac{4(K+1)}{\beta}}, \frac{\ln(B_K)}{\mu(K+1)}\right\}, \qquad (83)
$$

$$
B_K \quad = \quad \max\left\{2, \frac{(K+1)^{\frac{2(\alpha-1)}{\alpha}}\mu^2\Delta}{264600^{\frac{2}{\alpha}}L\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{4(K+1)}{\beta}\right)\ln^2(B_K)}\right\} \qquad (84)
$$

$$
= \quad \Theta\left(\max\left\{1, \frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2\Delta}{L\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)\ln^2\left(\max\left\{2, \frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2\Delta}{L\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)}\right\}\right)}\right\}\right), \qquad (85)
$$

and clipping level

$$
\lambda_k = \frac{\exp(-\gamma\mu(1 + k/2))\sqrt{\Delta}}{120\sqrt{L}\gamma\ln\frac{4(K+1)}{\beta}}, \qquad (86)
$$

for some $K > 0$ and $\beta \in (0,1]$ such that $\ln \frac{4(K+1)}{\beta} \geq 1$. Then, after K iterations of clipped-SGD *the iterates with probability at least $1-\beta$ satisfy*

$$f\left(x^{K+1}\right) - f_* \leq 2\exp(-\gamma\mu(K+1))\Delta. \tag{87}$$

In particular, when γ equals the minimum from (83), then the iterates produced by clipped-SGD *after K iterations with probability at least $1-\beta$ satisfy*

$$f\left(x^K\right) - f_* = \mathcal{O}\left(\max\left\{\Delta\exp\left(-\frac{\mu K}{L\ln\frac{K}{\beta}}\right), \frac{L\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)\ln^2\left(\max\left\{2, \frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2\Delta}{L\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)}\right\}\right)}{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2}\right\}\right), \tag{88}$$

meaning that to achieve $f\left(x^K\right) - f_* \leq \varepsilon$ with probability at least $1-\beta$ clipped-SGD *requires*

$$K = \mathcal{O}\left(\frac{L}{\mu}\ln\left(\frac{\Delta}{\varepsilon}\right)\ln\left(\frac{L}{\mu\beta}\ln\frac{\Delta}{\varepsilon}\right), \left(\frac{L\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\ln\left(\frac{1}{\beta}\left(\frac{L\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right)\ln^{\frac{\alpha}{\alpha-1}}\left(B_\varepsilon\right)\right), \tag{89}$$

iterations/oracle calls, where

$$B_\varepsilon = \max\left\{2, \frac{\Delta}{\varepsilon\ln\left(\frac{1}{\beta}\left(\frac{L\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right)}\right\}.$$

Proof. As in the previous results, the proof is based on the induction argument and shows that the iterates do not leave some set with high probability. More precisely, for each $k = 0, 1, \ldots, K+1$ we consider probability event E_k as follows: inequalities

$$\Delta_t \leq 2\exp(-\gamma\mu t)\Delta \tag{90}$$

hold for $t = 0, 1, \ldots, k$ simultaneously, where $\Delta_t = f(x^t) - f_*$. We want to prove $\mathbb{P}\{E_k\} \geq 1 - {k\beta}/{(K+1)}$ for all $k = 0, 1, \ldots, K+1$ by induction. The base of the induction is trivial: for $k = 0$ we have $\Delta_0 \leq \Delta < 2\Delta$ by definition. Next, assume that for $k = T-1 \leq K$ the statement holds: $\mathbb{P}\{E_{T-1}\} \geq 1 - {(T-1)\beta}/{(K+1)}$. Given this, we need to prove $\mathbb{P}\{E_T\} \geq 1 - {T\beta}/{(K+1)}$. Since $\Delta_t \leq 2\exp(-\gamma\mu t)\Delta \leq 2\Delta$, we have $x^t \in \{y \in \mathbb{R}^d \mid f(y) \leq f_* + 2\Delta\}$ for $t = 0, 1, \ldots, T-1$, where function f is L-smooth. Thus, E_{T-1} implies

$$\|\nabla f(x^t)\| \overset{(7)}{\leq} \sqrt{2L(f(x^t) - f_*)} \overset{(90)}{\leq} 2\sqrt{L\exp(-\gamma\mu t)\Delta} \overset{(83),(86)}{\leq} \frac{\lambda_t}{2} \tag{91}$$

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

$$\|x^T - x^{T-1}\| = \gamma\|\widetilde{\nabla} f_{\xi^{T-1}}(x^{T-1})\| \leq \gamma\lambda_{T-1} \overset{(86)}{\leq} \frac{\sqrt{\Delta}}{20\sqrt{L}},$$

meaning that E_{T-1} implies $x^T \in \{x \in \mathbb{R}^d \mid \exists y \in \mathbb{R}^d:\ f(y) \leq f_* + 2\Delta \text{ and } \|x - y\| \leq {\sqrt{\Delta}}/{20\sqrt{L}}\}$. Using Lemma E.3 and $(1-\gamma\mu)^T \leq \exp(-\gamma\mu T)$, we obtain that E_{T-1} implies

$$\begin{aligned}\Delta_T \quad &\leq \quad \exp(-\gamma\mu T)\Delta - \gamma(1 - L\gamma)\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\langle\nabla f(x^l), \theta_l\rangle \\ &\quad + \frac{L\gamma^2}{2}\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\|\theta_l\|^2.\end{aligned}$$

To handle the sums above, we introduce a new notation:

$$\eta_t = \begin{cases}\nabla f(x^t), & \text{if } \|\nabla f(x^t)\| \leq 2\sqrt{L}\exp({-\gamma\mu t}/{2})\sqrt{\Delta}, \\ 0, & \text{otherwise,}\end{cases} \tag{92}$$

for $t = 0, 1, \ldots, T-1$. These vectors are bounded almost surely:

$$\|\eta_t\| \leq 2\sqrt{L}\exp(-\gamma\mu t/2)\sqrt{\Delta} \tag{93}$$

for all $t = 0, 1, \ldots, T-1$. In other words, E_{T-1} implies $\eta_t = \nabla f(x^t)$ for all $t = 0, 1, \ldots, T-1$, meaning that from E_{T-1} it follows that

$$\begin{aligned}\Delta_T \quad &\leq \quad \exp(-\gamma\mu T)\Delta - \gamma(1 - L\gamma)\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\langle \eta_l, \theta_l\rangle \\ &\quad + \frac{L\gamma^2}{2}\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\|\theta_l\|^2.\end{aligned}$$

To handle the sums appeared on the right-hand side of the previous inequality we consider unbiased and biased parts of θ_l:

$$\theta_t^u = \widetilde{\nabla} f_{\xi^t}(x^t) - \mathbb{E}_{\xi^t}\left[\widetilde{\nabla} f_{\xi^t}(x^t)\right], \quad \theta_t^b = \mathbb{E}_{\xi^t}\left[\widetilde{\nabla} f_{\xi^t}(x^t)\right] - \nabla f(x^t). \tag{94}$$

for all $l = 0, \ldots, T-1$. By definition we have $\theta_l = \theta_l^u + \theta_l^b$ for all $l = 0, \ldots, T-1$. Therefore, E_{T-1} implies

$$\begin{aligned}\Delta_T \quad &\leq \quad \exp(-\gamma\mu T)\Delta \underbrace{-\gamma(1-L\gamma)\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\langle \eta_l, \theta_l^u\rangle}_{①} \underbrace{-\gamma(1-L\gamma)\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\langle \eta_l, \theta_l^b\rangle}_{②} \\ &\quad \underbrace{+L\gamma^2\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\mathbb{E}_{\xi^l}\left[\|\theta_l^u\|^2\right]}_{③} \underbrace{+L\gamma^2\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\left(\|\theta_l^u\|^2 - \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|^2\right]\right)}_{④} \\ &\quad \underbrace{+L\gamma^2\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\|\theta_l^b\|^2}_{⑤}.\end{aligned} \tag{95}$$

where we also apply inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ holding for all $a, b \in \mathbb{R}^d$ to upper bound $\|\theta_l\|^2$. It remains to derive good enough high-probability upper-bounds for the terms ①, ②, ③, ④, ⑤, i.e., to finish our inductive proof we need to show that ① + ② + ③ + ④ + ⑤ $\leq \exp(-\gamma\mu T)\Delta$ with high probability. In the subsequent parts of the proof, we will need to use many times the bounds for the norm and second moments of θ_t^u and θ_t^b. First, by definition of clipping operator, we have with probability 1 that

$$\|\theta_l^u\| \leq 2\lambda_l. \tag{96}$$

Moreover, since E_{T-1} implies that $\|\nabla f(x^l)\|^2 \leq \lambda_l/2$ for all $l = 0, 1, \ldots, T-1$ (see (91)), from Lemma 5.1 we also have that E_{T-1} implies

$$\|\theta_l^b\| \leq \frac{2^\alpha\sigma^\alpha}{\lambda_l^{\alpha-1}}, \tag{97}$$

$$\mathbb{E}_{\xi^l}\left[\|\theta_l^u\|^2\right] \leq 18\lambda_l^{2-\alpha}\sigma^\alpha, \tag{98}$$

for all $l = 0, 1, \ldots, T-1$.

Upper bound for ①. By definition of θ_l^u, we have $\mathbb{E}_{\xi^l}[\theta_l^u] = 0$ and

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

Next, sum ① has bounded with probability 1 terms:

$$\begin{aligned}|-\gamma(1-L\gamma)(1-\gamma\mu)^{T-1-l}\langle \eta_l, \theta_l^u\rangle| \quad &\overset{(83)}{\leq} \quad \gamma\exp(-\gamma\mu(T-1-l))\|\eta_l\| \cdot \|\theta_l^u\| \\ &\overset{(93),(96)}{\leq} \quad 4\sqrt{L\Delta}\gamma\exp(-\gamma\mu(T-1-l/2))\lambda_l \\ &\overset{(83),(86)}{\leq} \quad \frac{\exp(-\gamma\mu T)\Delta}{5\ln\frac{4(K+1)}{\beta}} \overset{\text{def}}{=} c.\end{aligned} \tag{99}$$

The summands also have bounded conditional variances $\sigma_l^2 \stackrel{\text{def}}{=} \mathbb{E}_{\xi^l}\left[\gamma^2(1-L\gamma)^2(1-\gamma\mu)^{2T-2-2l}\langle\eta_l,\theta_l^u\rangle^2\right]$:

$$\begin{aligned}\sigma_l^2 &\leq \mathbb{E}_{\xi^l}\left[\gamma^2(1-L\gamma)^2\exp(-\gamma\mu(2T-2-2l))\|\eta_l\|^2\cdot\|\theta_l^u\|^2\right]\\ &\stackrel{(93),(83)}{\leq} 4\gamma^2 L\Delta\exp(-\gamma\mu(2T-2-l))\mathbb{E}_{\xi^l}\left[\|\theta_l^u\|^2\right]\\ &\stackrel{(83)}{\leq} 10\gamma^2 L\Delta\exp(-\gamma\mu(2T-l))R^2\mathbb{E}_{\xi^l}\left[\|\theta_l^u\|^2\right]. \end{aligned} \tag{100}$$

In other words, we showed that $\{-\gamma(1-L\gamma)(1-\gamma\mu)^{T-1-l}\langle\eta_l,\theta_l^u\rangle\}_{l=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\sigma_l^2\}_{l=0}^{T-1}$. Next, we apply Bernstein's inequality (Lemma B.2) with $X_l = -\gamma(1-L\gamma)(1-\gamma\mu)^{T-1-l}\langle\eta_l,\theta_l^u\rangle$, parameter c as in (99), $b = \frac{1}{5}\exp(-\gamma\mu T)\Delta$, $G = \frac{\exp(-2\gamma\mu T)\Delta^2}{150\ln\frac{4(K+1)}{\beta}}$:

$$\mathbb{P}\left\{|①| > \frac{1}{5}\exp(-\gamma\mu T)\Delta \text{ and } \sum_{l=0}^{T-1}\sigma_l^2 \leq \frac{\exp(-2\gamma\mu T)\Delta^2}{150\ln\frac{4(K+1)}{\beta}}\right\} \leq 2\exp\left(-\frac{b^2}{2G + {}^{2cb}/_{3}}\right) = \frac{\beta}{2(K+1)}.$$

Equivalently, we have

$$\mathbb{P}\{E_①\} \geq 1 - \frac{\beta}{2(K+1)}, \quad \text{for} \quad E_① = \left\{\text{either} \quad \sum_{l=0}^{T-1}\sigma_l^2 > \frac{\exp(-2\gamma\mu T)\Delta^2}{150\ln\frac{4(K+1)}{\beta}} \quad \text{or} \quad |①| \leq \frac{1}{5}\exp(-\gamma\mu T)\Delta\right\}. \tag{101}$$

In addition, E_{T-1} implies that

$$\begin{aligned}\sum_{l=0}^{T-1}\sigma_l^2 &\stackrel{(100)}{\leq} 10\gamma^2 L\Delta\exp(-2\gamma\mu T)\sum_{l=0}^{T-1}\frac{\mathbb{E}_{\xi^l}\left[\|\theta_l^u\|^2\right]}{\exp(-\gamma\mu l)}\\ &\stackrel{(98),T\leq K+1}{\leq} 180\gamma^2 L\Delta\exp(-2\gamma\mu T)\sigma^\alpha\sum_{l=0}^{K}\frac{\lambda_l^{2-\alpha}}{\exp(-\gamma\mu l)}\\ &\stackrel{(86)}{=} \frac{180\gamma^\alpha\sqrt{L}^\alpha\sqrt{\Delta}^{4-\alpha}\exp(-2\gamma\mu T)\sigma^\alpha}{120^{2-\alpha}\ln^{2-\alpha}\frac{4(K+1)}{\beta}}\sum_{l=0}^{K}\frac{1}{\exp(-\gamma\mu l)}\cdot(\exp(-\gamma\mu(1+{}^{l}/_{2})))^{2-\alpha}\\ &= \frac{180\gamma^\alpha\sqrt{L}^\alpha\sqrt{\Delta}^{4-\alpha}\exp(-2\gamma\mu T)\sigma^\alpha}{120^{2-\alpha}\ln^{2-\alpha}\frac{4(K+1)}{\beta}}\sum_{l=0}^{K}\exp(\gamma\mu(\alpha-2))\cdot\exp\left(\frac{\gamma\mu\alpha l}{2}\right)\\ &\leq \frac{180\gamma^\alpha\sqrt{L}^\alpha\sqrt{\Delta}^{4-\alpha}\exp(-2\gamma\mu T)\sigma^\alpha(K+1)\exp(\frac{\gamma\mu\alpha K}{2})}{120^{2-\alpha}\ln^{2-\alpha}\frac{4(K+1)}{\beta}}\\ &\stackrel{(83)}{\leq} \frac{\exp(-2\gamma\mu T)\Delta^2}{150\ln\frac{4(K+1)}{\beta}}, \end{aligned} \tag{102}$$

where we also show that E_{T-1} implies

$$\gamma^2 L\Delta\sum_{l=0}^{K}\frac{\lambda_l^{2-\alpha}}{\exp(-\gamma\mu l)} \leq \frac{\gamma^\alpha\sqrt{L}^\alpha\sqrt{\Delta}^{4-\alpha}(K+1)\exp(\frac{\gamma\mu\alpha K}{2})}{120^{2-\alpha}\ln^{2-\alpha}\frac{4(K+1)}{\beta}}. \tag{103}$$

Upper bound for ②. From E_{T-1} it follows that

$$\begin{aligned}
② &\overset{(83)}{\leq} \gamma \exp(-\gamma\mu(T-1)) \sum_{l=0}^{T-1} \frac{\|\eta_l\| \cdot \|\theta_l^b\|}{\exp(-\gamma\mu l)} \\
&\overset{(93),(97)}{\leq} 2^{1+\alpha}\gamma \exp(-\gamma\mu(T-1))\sqrt{\Delta}\sigma^\alpha \sum_{l=0}^{T-1} \frac{1}{\lambda_l^{\alpha-1}\exp(-\gamma\mu l/2)} \\
&\overset{(86)}{=} 2^{1+\alpha}\cdot 120^{\alpha-1}\sqrt{L}^{1-\alpha}\sqrt{\Delta}^{2-\alpha}\exp(-\gamma\mu(T-1))\gamma^\alpha\sigma^\alpha \ln^{\alpha-1}\tfrac{4(K+1)}{\beta} \sum_{l=0}^{T-1} \frac{\exp(\gamma\mu l/2)}{\exp\left(-\gamma\mu(1+l/2)\right)^{\alpha-1}} \\
&\overset{T\leq K+1}{\leq} 2^{1+\alpha}\cdot 120^{\alpha-1}\sqrt{L}^{1-\alpha}\sqrt{\Delta}^{2-\alpha}\exp(-\gamma\mu(T-1))\gamma^\alpha\sigma^\alpha \ln^{\alpha-1}\tfrac{4(K+1)}{\beta} \sum_{l=0}^{K} \exp\left(\frac{\gamma\mu\alpha l}{2}\right) \\
&\leq 2^{1+\alpha}\cdot 120^{\alpha-1}\sqrt{L}^{1-\alpha}\sqrt{\Delta}^{2-\alpha}\exp(-\gamma\mu(T-1))\gamma^\alpha\sigma^\alpha \ln^{\alpha-1}\tfrac{4(K+1)}{\beta}(K+1)\exp\left(\frac{\gamma\mu\alpha K}{2}\right) \\
&\overset{(83)}{\leq} \frac{1}{5}\exp(-\gamma\mu T)\Delta. \qquad (104)
\end{aligned}$$

Upper bound for ③. From E_{T-1} it follows that

$$\begin{aligned}
③ &= L\gamma^2 \exp(-\gamma\mu(T-1)) \sum_{l=0}^{T-1} \frac{\mathbb{E}_{\xi^l}\left[\|\theta_l^u\|^2\right]}{\exp(-\gamma\mu l)} \\
&\overset{(98)}{\leq} 18L\gamma^2\exp(-\gamma\mu(T-1))\sigma^\alpha \sum_{l=0}^{T-1} \frac{\lambda_l^{2-\alpha}}{\exp(-\gamma\mu l)} \\
&\overset{(103)}{\leq} \frac{18\gamma^\alpha\sqrt{L}^\alpha\sqrt{\Delta}^{2-\alpha}\exp(-\gamma\mu(T-1))\sigma^\alpha(K+1)\exp(\frac{\gamma\mu\alpha K}{2})}{120^{2-\alpha}\ln^{2-\alpha}\frac{4(K+1)}{\beta}} \\
&\overset{(83)}{\leq} \frac{1}{5}\exp(-\gamma\mu T)\Delta. \qquad (105)
\end{aligned}$$

Upper bound for ④. First, we have

$$L\gamma^2(1-\gamma\mu)^{T-1-l}\mathbb{E}_{\xi^l}\left[\|\theta_l^u\|^2 - \mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right]\right] = 0.$$

Next, sum ④ has bounded with probability 1 terms:

$$\begin{aligned}
L\gamma^2(1-\gamma\mu)^{T-1-l}\left|\|\theta_l^u\|^2 - \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|^2\right]\right| &\overset{(96)}{\leq} \frac{8L\gamma^2\exp(-\gamma\mu T)\lambda_l^2}{\exp(-\gamma\mu(1+l))} \\
&\overset{(86)}{=} \frac{\exp(-\gamma\mu(T+1))\Delta}{1800\ln^2\frac{4(K+1)}{\beta}} \\
&\leq \frac{\exp(-\gamma\mu T)\Delta}{5\ln\frac{4(K+1)}{\beta}} \\
&\overset{\text{def}}{=} c. \qquad (106)
\end{aligned}$$

The summands also have conditional variances

$$\widehat{\sigma}_l^2 \overset{\text{def}}{=} \mathbb{E}_{\xi^l}\left[L^2\gamma^4(1-\gamma\mu)^{2T-2-2l}\left|\|\theta_l^u\|^2 - \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|^2\right]\right|^2\right]$$

that are bounded

$$\widehat{\sigma}_l^2 \overset{(106)}{\leq} \frac{L\gamma^2\exp(-2\gamma\mu T)\Delta}{5\exp(-\gamma\mu(1+l))\ln\frac{4(K+1)}{\beta}}\mathbb{E}_{\xi^l}\left[\left|\|\theta_l^u\|^2 - \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|^2\right]\right|\right]$$
$$\leq \frac{2L\gamma^2\exp(-2\gamma\mu T)\Delta}{5\exp(-\gamma\mu(1+l))\ln\frac{4(K+1)}{\beta}}\mathbb{E}_{\xi^l}\left[\|\theta_l^u\|^2\right]. \quad (107)$$

In other words, we showed that $\left\{L\gamma^2(1-\gamma\mu)^{T-1-l}\left(\|\theta_l^u\|^2 - \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|^2\right]\right)\right\}_{l=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\widehat{\sigma}_l^2\}_{l=0}^{T-1}$. Next, we apply Bernstein's inequality (Lemma B.2) with $X_l = L\gamma^2(1-\gamma\mu)^{T-1-l}\left(\|\theta_l^u\|^2 - \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|^2\right]\right)$, parameter c as in (106), $b = \frac{1}{5}\exp(-\gamma\mu T)\Delta$, $G = \frac{\exp(-2\gamma\mu T)\Delta^2}{150\ln\frac{4(K+1)}{\beta}}$:

$$\mathbb{P}\left\{|④| > \frac{1}{5}\exp(-\gamma\mu T)\Delta \text{ and } \sum_{l=0}^{T-1}\widehat{\sigma}_l^2 \leq \frac{\exp(-2\gamma\mu T)\Delta^2}{150\ln\frac{4(K+1)}{\beta}}\right\} \leq 2\exp\left(-\frac{b^2}{2G + 2cb/3}\right) = \frac{\beta}{2(K+1)}.$$

Equivalently, we have

$$\mathbb{P}\{E_④\} \geq 1 - \frac{\beta}{2(K+1)}, \quad \text{for} \quad E_④ = \left\{\text{either} \quad \sum_{l=0}^{T-1}\widehat{\sigma}_l^2 > \frac{\exp(-2\gamma\mu T)\Delta^2}{150\ln\frac{4(K+1)}{\beta}} \quad \text{or} \quad |④| \leq \frac{1}{5}\exp(-\gamma\mu T)\Delta\right\}. \quad (108)$$

In addition, E_{T-1} implies that

$$\sum_{l=0}^{T-1}\widehat{\sigma}_l^2 \overset{(107)}{\leq} \frac{2L\gamma^2\exp(-\gamma\mu(2T-1))\Delta}{5\ln\frac{4(K+1)}{\beta}}\sum_{l=0}^{T-1}\frac{\mathbb{E}_{\xi^l}\left[\|\theta_l^u\|^2\right]}{\exp(-\gamma\mu l)}$$
$$\overset{(98), T\leq K+1}{\leq} \frac{36L\gamma^2\exp(-\gamma\mu(2T-1))\Delta\sigma^\alpha}{5\ln\frac{4(K+1)}{\beta}}\sum_{l=0}^{K}\frac{\lambda_l^{2-\alpha}}{\exp(-\gamma\mu l)}$$
$$\overset{(103)}{\leq} \frac{36\sqrt{L}^\alpha\gamma^\alpha\exp(-\gamma\mu(2T-1))\sqrt{\Delta}^{4-\alpha}\sigma^\alpha(K+1)\exp(\frac{\gamma\mu\alpha K}{2})}{5\cdot 120^{2-\alpha}\ln^{3-\alpha}\frac{4(K+1)}{\beta}}$$
$$\overset{(53)}{\leq} \frac{\exp(-2\gamma\mu T)\Delta^2}{150\ln\frac{4(K+1)}{\beta}}. \quad (109)$$

Upper bound for ⑤. From E_{T-1} it follows that

$$⑤ = L\gamma^2\sum_{l=0}^{T-1}\exp(-\gamma\mu(T-1-l))\|\theta_l^b\|^2$$
$$\overset{(97)}{\leq} 2^{2\alpha}L\gamma^2\exp(-\gamma\mu(T-1))\sigma^{2\alpha}\sum_{l=0}^{T-1}\frac{1}{\lambda_l^{2\alpha-2}\exp(-\gamma\mu l)}$$
$$\overset{(86), T\leq K+1}{\leq} \frac{2\cdot 2^{2\alpha}\cdot 120^{2\alpha-2}\gamma^{2\alpha}\sqrt{L}^{2\alpha}\exp(-\gamma\mu T)\sigma^{2\alpha}\ln^{2\alpha-2}\frac{4(K+1)}{\beta}}{\sqrt{\Delta}^{2\alpha-2}}\sum_{l=0}^{K}\exp\left(\gamma\mu(2\alpha-2)\left(1+\frac{l}{2}\right)\right)\exp(\gamma\mu l)$$
$$\leq \frac{4\cdot 2^{2\alpha}\cdot 120^{2\alpha-2}\gamma^{2\alpha}\sqrt{L}^{2\alpha}\exp(-\gamma\mu T)\sigma^{2\alpha}\ln^{2\alpha-2}\frac{4(K+1)}{\beta}}{\sqrt{\Delta}^{2\alpha-2}}\sum_{l=0}^{K}\exp(\gamma\mu\alpha l)$$
$$\leq \frac{4\cdot 2^{2\alpha}\cdot 120^{2\alpha-2}\gamma^{2\alpha}\sqrt{L}^{2\alpha}\exp(-\gamma\mu T)\sigma^{2\alpha}\ln^{2\alpha-2}\frac{4(K+1)}{\beta}(K+1)\exp(\gamma\mu\alpha K)}{\sqrt{\Delta}^{2\alpha-2}}$$
$$\overset{(53)}{\leq} \frac{1}{5}\exp(-\gamma\mu T)\Delta. \quad (110)$$

Now, we have the upper bounds for ①, ②, ③, ④, ⑤. In particular, probability event E_{T-1} implies

$$\Delta_T \overset{(95)}{\leq} \exp(-\gamma\mu T)\Delta + ① + ② + ③ + ④ + ⑤,$$
$$② \overset{(104)}{\leq} \frac{1}{5}\exp(-\gamma\mu T)\Delta, \quad ③ \overset{(105)}{\leq} \frac{1}{5}\exp(-\gamma\mu T)\Delta, \quad ⑤ \overset{(110)}{\leq} \frac{1}{5}\exp(-\gamma\mu T)\Delta,$$
$$\sum_{l=0}^{T-1}\sigma_l^2 \overset{(102)}{\leq} \frac{\exp(-2\gamma\mu T)\Delta^2}{150\ln\frac{4(K+1)}{\beta}}, \quad \sum_{l=0}^{T-1}\widehat{\sigma}_l^2 \overset{(109)}{\leq} \frac{\exp(-2\gamma\mu T)\Delta^2}{150\ln\frac{4(K+1)}{\beta}}.$$

Moreover, we also have (see (101), (108) and our induction assumption)

$$\mathbb{P}\{E_{T-1}\} \geq 1 - \frac{(T-1)\beta}{K+1},$$
$$\mathbb{P}\{E_①\} \geq 1 - \frac{\beta}{2(K+1)}, \quad \mathbb{P}\{E_④\} \geq 1 - \frac{\beta}{2(K+1)},$$

where

$$E_① = \left\{\text{either} \quad \sum_{l=0}^{T-1}\sigma_l^2 > \frac{\exp(-2\gamma\mu T)\Delta^2}{150\ln\frac{4(K+1)}{\beta}} \quad \text{or} \quad |①| \leq \frac{1}{5}\exp(-\gamma\mu T)\Delta\right\},$$
$$E_④ = \left\{\text{either} \quad \sum_{l=0}^{T-1}\widehat{\sigma}_l^2 > \frac{\exp(-2\gamma\mu T)\Delta^2}{150\ln\frac{4(K+1)}{\beta}} \quad \text{or} \quad |④| \leq \frac{1}{5}\exp(-\gamma\mu T)\Delta\right\}.$$

Thus, probability event $E_{T-1} \cap E_① \cap E_④$ implies

$$\Delta_T \overset{(95)}{\leq} \exp(-\gamma\mu T)\Delta + ① + ② + ③ + ④ + ⑤ \leq 2\exp(-\gamma\mu T)\Delta,$$

which is equivalent to (90) for $t = T$, and

$$\mathbb{P}\{E_T\} \geq \mathbb{P}\{E_{T-1} \cap E_① \cap E_④\} = 1 - \mathbb{P}\{\overline{E}_{T-1} \cup \overline{E}_① \cup \overline{E}_④\} \geq 1 - \frac{T\beta}{K+1}.$$

This finishes the inductive part of our proof, i.e., for all $k = 0, 1, \ldots, K+1$ we have $\mathbb{P}\{E_k\} \geq 1 - {k\beta}/{(K+1)}$. In particular, for $k = K+1$ we have that with probability at least $1-\beta$

$$f(x^{K+1}) - f_* \leq 2\exp(-\gamma\mu(K+1))\Delta.$$

Finally, if

$$\gamma = \min\left\{\frac{1}{250L\ln\frac{4(K+1)}{\beta}}, \frac{\ln(B_K)}{\mu(K+1)}\right\},$$
$$B_K = \max\left\{2, \frac{(K+1)^{\frac{2(\alpha-1)}{\alpha}}\mu^2\Delta}{264600^{\frac{2}{\alpha}}L\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{6(K+1)}{\beta}\right)\ln^2(B_K)}\right\}$$
$$= \mathcal{O}\left(\max\left\{2, \frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2\Delta}{L\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)\ln^2\left(\max\left\{2, \frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2\Delta}{L\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)}\right\}\right)}\right\}\right)$$

then with probability at least $1-\beta$

$$\begin{aligned} f(x^{K+1}) - f_* &\leq 2\exp(-\gamma\mu(K+1))\Delta \\ &= 2\Delta \max\left\{\exp\left(-\frac{\mu(K+1)}{250L\ln\frac{4(K+1)}{\beta}}\right), \frac{1}{B_K}\right\} \\ &= \mathcal{O}\left(\max\left\{\Delta\exp\left(-\frac{\mu K}{L\ln\frac{K}{\beta}}\right), \frac{L\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)\ln^2\left(\max\left\{2, \frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2\Delta}{L\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)}\right\}\right)}{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2}\right\}\right). \end{aligned}$$

To get $\|x^{K+1}-x^*\|^2 \leq \varepsilon$ with probability at least $1-\beta$ it is sufficient to choose K such that both terms in the maximum above are $\mathcal{O}(\varepsilon)$. This leads to

$$K = \mathcal{O}\left(\frac{L}{\mu}\ln\left(\frac{\Delta}{\varepsilon}\right)\ln\left(\frac{L}{\mu\beta}\ln\frac{\Delta}{\varepsilon}\right), \left(\frac{L\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\ln\left(\frac{1}{\beta}\left(\frac{L\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right)\ln^{\frac{\alpha}{\alpha-1}}(B_\varepsilon)\right),$$

where

$$B_\varepsilon = \max\left\{2, \frac{\Delta}{\varepsilon\ln\left(\frac{1}{\beta}\left(\frac{L\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right)}\right\}.$$

This concludes the proof. □

E.3. Convex Functions

Now, we focus on the case of convex functions. We start with the following lemma.

Lemma E.5. *Let Assumptions 1.3 and 1.6 with $\mu = 0$ hold on $Q = B_{2R}(x^*)$, where $R \geq \|x^0 - x^*\|$, and let stepsize γ satisfy $\gamma \leq \frac{1}{L}$. If $x^k \in Q$ for all $k = 0,1,\ldots,K+1$, $K \geq 0$, then after K iterations of* clipped-SGD *we have*

$$\begin{aligned} \gamma\left(f(\overline{x}^K) - f(x^*)\right) &\leq \frac{\|x^0-x^*\|^2 - \|x^{K+1}-x^*\|^2}{K+1} \\ &\quad -\frac{2\gamma}{K+1}\sum_{k=0}^{K}\langle x^k - x^* - \gamma\nabla f(x^k), \theta_k\rangle + \frac{\gamma^2}{K+1}\sum_{k=0}^{K}\|\theta_k\|^2, \quad (111) \\ \overline{x}^K &= \frac{1}{K+1}\sum_{k=0}^{K}x^k, \quad (112) \end{aligned}$$

where θ_k is defined in (52).

Proof. Using $x^{k+1} = x^k - \gamma\widetilde{\nabla} f_{\xi^k}(x^k)$, we derive for all $k = 0,1,\ldots,K$ that

$$\begin{aligned} \|x^{k+1}-x^*\|^2 &= \|x^k - x^*\|^2 - 2\gamma\langle x^k - x^*, \widetilde{\nabla} f_{\xi^k}(x^k)\rangle + \gamma^2\|\widetilde{\nabla} f_{\xi^k}(x^k)\|^2 \\ &= \|x^k - x^*\|^2 - 2\gamma\langle x^k - x^*, \nabla f(x^k)\rangle - 2\gamma\langle x^k - x^*, \theta_k\rangle + \gamma^2\|\nabla f(x^k) + \theta_k\|^2 \\ &\overset{(10),\mu=0}{\leq} \|x^k - x^*\|^2 - 2\gamma\left(f(x^k) - f(x^*)\right) - 2\gamma\langle x^k - x^* - \gamma\nabla f(x^k), \theta_k\rangle \\ &\qquad +\gamma^2\|\nabla f(x^k)\|^2 + \gamma^2\|\theta_k\|^2 \\ &\overset{(7)}{\leq} \|x^k - x^*\|^2 - 2\gamma(1-\gamma L)\left(f(x^k) - f(x^*)\right) - 2\gamma\langle x^k - x^* - \gamma\nabla f(x^k), \theta_k\rangle \\ &\qquad +\gamma^2\|\theta_k\|^2 \\ &\overset{\gamma\leq 1/L}{\leq} \|x^k - x^*\|^2 - \gamma\left(f(x^k) - f(x^*)\right) - 2\gamma\langle x^k - x^* - \gamma\nabla f(x^k), \theta_k\rangle + \gamma^2\|\theta_k\|^2. \end{aligned}$$

Summing up the above inequalities for $k = 0, 1, \ldots, K$ and rearranging the terms, we get

$$\begin{aligned}
\frac{\gamma}{K+1}\sum_{k=0}^{K}\left(f(x^k) - f(x^*)\right) &\leq \frac{1}{K+1}\sum_{k=0}^{K}\left(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right) - \frac{2\gamma}{K+1}\sum_{k=0}^{K}\langle x^k - x^* - \gamma\nabla f(x^k), \theta_k\rangle \\
&\quad + \frac{\gamma^2}{K+1}\sum_{k=0}^{K}\|\theta_k\|^2 \\
&= \frac{\|x^0 - x^*\|^2 - \|x^{K+1} - x^*\|^2}{K+1} - \frac{2\gamma}{K+1}\sum_{k=0}^{K}\langle x^k - x^* - \gamma\nabla f(x^k), \theta_k\rangle \\
&\quad + \frac{\gamma^2}{K+1}\sum_{k=0}^{K}\|\theta_k\|^2.
\end{aligned}$$

Finally, we use the definition of $\overline{x}^K$ and Jensen's inequality and get the result. □

Using this lemma we prove the main convergence result for clipped-SGD.

Theorem E.6 (Case 3 from Theorem 3.1)**.** *Let Assumptions 1.1, 1.3 and 1.6 with $\mu = 0$ hold on $Q = B_{2R}(x^*)$, where $R \geq \|x^0 - x^*\|$, and*

$$\gamma \leq \min\left\{\frac{1}{80L\ln\frac{4(K+1)}{\beta}}, \frac{R}{108^{\frac{1}{\alpha}}\cdot 20\sigma K^{\frac{1}{\alpha}}\left(\ln\frac{4(K+1)}{\beta}\right)^{\frac{\alpha-1}{\alpha}}}\right\}, \tag{113}$$

$$\lambda_k \equiv \lambda = \frac{R}{40\gamma\ln\frac{4(K+1)}{\beta}}, \tag{114}$$

for some $K > 0$ and $\beta \in (0, 1]$ such that $\ln\frac{4K}{\beta} \geq 1$. Then, after K iterations of clipped-SGD *the iterates with probability at least $1 - \beta$ satisfy*

$$f(\overline{x}^K) - f(x^*) \leq \frac{2R^2}{\gamma(K+1)} \quad \text{and} \quad \{x^k\}_{k=0}^{K} \subseteq B_{\sqrt{2}R}(x^*). \tag{115}$$

In particular, when γ equals the minimum from (53), then the iterates produced by clipped-SGD *after K iterations with probability at least $1 - \beta$ satisfy*

$$f(\overline{x}^K) - f(x^*) = \mathcal{O}\left(\max\left\{\frac{LR^2\ln\frac{K}{\beta}}{K}, \frac{\sigma R\ln^{\frac{\alpha-1}{\alpha}}\frac{K}{\beta}}{K^{\frac{\alpha-1}{\alpha}}}\right\}\right), \tag{116}$$

meaning that to achieve $f(\overline{x}^K) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ clipped-SGD *requires*

$$K = \mathcal{O}\left(\max\left\{\frac{LR^2}{\varepsilon}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\ln\left(\frac{1}{\beta}\left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right\}\right) \quad \textit{iterations/oracle calls.} \tag{117}$$

Proof. Let $R_k = \|x^k - x^*\|$ for all $k \geq 0$. Next, our goal is to show by induction that $R_l \leq 2R$ with high probability, which allows to apply the result of Lemma E.5 and then use Bernstein's inequality to estimate the stochastic part of the upper-bound. More precisely, for each $k = 0, \ldots, K+1$ we consider probability event E_k defined as follows: inequalities

$$-2\gamma\sum_{l=0}^{t-1}\langle x^l - x^* - \gamma\nabla f(x^l), \theta_l\rangle + \gamma^2\sum_{l=0}^{t-1}\|\theta_l\|^2 \leq R^2, \tag{118}$$

$$R_t \leq \sqrt{2}R \tag{119}$$

hold for all $t = 0, 1, \ldots, k$ simultaneously. We want to prove via induction that $\mathbb{P}\{E_k\} \geq 1 - k\beta/(K+1)$ for all $k = 0, 1, \ldots, K+1$. For $k = 0$ the statement is trivial. Assume that the statement is true for some $k = T - 1 \leq K$:

$\mathbb{P}\{E_{T-1}\} \geq 1 - {(T-1)\beta}/{(K+1)}$. One needs to prove that $\mathbb{P}\{E_T\} \geq 1 - {T\beta}/{(K+1)}$. First, we notice that probability event E_{T-1} implies that $x_t \in B_{\sqrt{2}R}(x^*)$ for all $t = 0, 1, \ldots, T-1$. Moreover, E_{T-1} implies

$$\|x^T - x^*\| = \|x^{T-1} - x^* - \gamma\widetilde{\nabla} f_{\xi^{T-1}}(x^{T-1})\| \leq \|x^{T-1} - x^*\| + \gamma\|\widetilde{\nabla} f_{\xi^{T-1}}(x^{T-1})\| \leq \sqrt{2}R + \gamma\lambda \overset{(114)}{\leq} 2R,$$

i.e., $x^0, x^1, \ldots, x^T \in B_{2R}(x^*)$. Therefore, E_{T-1} implies $\{x^k\}_{k=0}^T \subseteq Q$, meaning that the assumptions of Lemma E.5 are satisfied and we have

$$\begin{aligned}\gamma\left(f(\overline{x}^{t-1}) - f(x^*)\right) \quad \leq \quad & \frac{\|x^0 - x^*\|^2 - \|x^t - x^*\|^2}{t} \\ & -\frac{2\gamma}{t}\sum_{l=0}^{t-1}\langle x^l - x^* - \gamma\nabla f(x^l), \theta_l\rangle + \frac{\gamma^2}{t}\sum_{l=0}^{t-1}\|\theta_l\|^2 \end{aligned} \tag{120}$$

for all $t = 1, \ldots, T$ simultaneously and for all $t = 1, \ldots, T-1$ this probability event also implies that

$$f(\overline{x}^{t-1}) - f(x^*) \quad \leq \quad \frac{1}{\gamma t}\left(R^2 - 2\gamma\sum_{l=0}^{t-1}\langle x^l - x^* - \gamma\nabla f(x^l), \theta_l\rangle + \gamma^2\sum_{l=0}^{t-1}\|\theta_l\|^2\right) \overset{(118)}{\leq} \frac{2R^2}{\gamma t}. \tag{121}$$

Taking into account that $f(\overline{x}^{T-1}) - f(x^*) \geq 0$, we also derive from (120) that E_{T-1} implies

$$R_T^2 \leq R^2 - 2\gamma\sum_{l=0}^{t-1}\langle x^l - x^* - \gamma\nabla f(x^l), \theta_l\rangle + \gamma^2\sum_{l=0}^{t-1}\|\theta_l\|^2. \tag{122}$$

Next, we define random vectors

$$\eta_t = \begin{cases} x^t - x^* - \gamma\nabla f(x^t), & \text{if } \|x^t - x^* - \gamma\nabla f(x^t)\|^2 \leq 2R, \\ 0, & \text{otherwise,}\end{cases}$$

for all $t = 0, 1, \ldots, T-1$. By definition, these random vectors are bounded with probability 1

$$\|\eta_t\| \leq 2R. \tag{123}$$

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

$$\begin{aligned}\|\nabla f(x^t)\| \quad & \overset{(6)}{\leq} \quad L\|x^t - x^*\| \overset{(119)}{\leq} \sqrt{2}LR \overset{(113),(114)}{\leq} \frac{\lambda}{2}, \\ \|x^t - x^* - \gamma\nabla f(x^t)\| \quad & \leq \quad \|x^t - x^*\| + \gamma\|\nabla f(x^t)\| \overset{(124)}{\leq} \sqrt{2}R(1 + L\gamma) \overset{(113)}{\leq} 2R.\end{aligned} \tag{124}$$

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

$$\theta_t^u = \widetilde{\nabla} f_{\xi^t}(x^t) - \mathbb{E}_{\xi^t}\left[\widetilde{\nabla} f_{\xi^t}(x^t)\right], \quad \theta_t^b = \mathbb{E}_{\xi^t}\left[\widetilde{\nabla} f_{\xi^t}(x^t)\right] - \nabla f(x^t). \tag{125}$$

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

$$\begin{aligned} R_T^2 \quad \leq \quad & R^2 \underbrace{-2\gamma\sum_{t=0}^{T-1}\langle\theta_t^u, \eta_t\rangle}_{①} \underbrace{-2\gamma\sum_{t=0}^{T-1}\langle\theta_t^b, \eta_t\rangle}_{②} + \underbrace{2\gamma^2\sum_{t=0}^{T-1}\left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]\right)}_{③} \\ & \underbrace{+2\gamma^2\sum_{t=0}^{T-1}\mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]}_{④} \underbrace{+2\gamma^2\sum_{t=0}^{T-1}\left\|\theta_t^b\right\|^2}_{⑤}. \end{aligned} \tag{126}$$

It remains to derive good enough high-probability upper-bounds for the terms ①, ②, ③, ④, ⑤, i.e., to finish our inductive proof we need to show that ① + ② + ③ + ④ + ⑤ $\leq R^2$ with high probability. In the subsequent parts of the proof, we will need to use many times the bounds for the norm and second moments of θ_t^u and θ_t^b. First, by definition of clipping operator, we have with probability 1 that

$$\|\theta_t^u\| \leq 2\lambda. \tag{127}$$

Moreover, since E_{T-1} implies that $\|\nabla f(x^t)\| \leq \lambda/2$ for $t = 0, 1, \ldots, T-1$ (see (124)), then, in view of Lemma 5.1, we have that E_{T-1} implies

$$\|\theta_t^b\| \leq \frac{2^\alpha \sigma^\alpha}{\lambda^{\alpha-1}}, \tag{128}$$

$$\mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right] \leq 18\lambda^{2-\alpha}\sigma^\alpha. \tag{129}$$

Upper bound for ①. By definition of θ_t^u, we have $\mathbb{E}_{\xi^t}[\theta_t^u] = 0$ and

$$\mathbb{E}_{\xi^t}\left[-2\gamma\langle\theta_t^u, \eta_t\rangle\right] = 0.$$

Next, sum ① has bounded with probability 1 terms:

$$|2\gamma\left\langle\theta_t^u, \eta_t\right\rangle| \leq 2\gamma\|\theta_t^u\| \cdot \|\eta_t\| \overset{(123),(127)}{\leq} 8\gamma\lambda R \overset{(114)}{=} \frac{R^2}{5\ln\frac{4(K+1)}{\beta}} \overset{\text{def}}{=} c. \tag{130}$$

The summands also have bounded conditional variances $\sigma_t^2 \overset{\text{def}}{=} \mathbb{E}_{\xi^t}[4\gamma^2\langle\theta_t^u, \eta_t\rangle^2]$:

$$\sigma_t^2 \leq \mathbb{E}_{\xi^t}\left[4\gamma^2\|\theta_t^u\|^2 \cdot \|\eta_t\|^2\right] \overset{(123)}{\leq} 16\gamma^2 R^2 \mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]. \tag{131}$$

In other words, we showed that $\{-2\gamma\left\langle\theta_t^u, \eta_t\right\rangle\}_{t=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\sigma_t^2\}_{t=0}^{T-1}$. Next, we apply Bernstein's inequality (Lemma B.2) with $X_t = -2\gamma\left\langle\theta_t^u, \eta_t\right\rangle$, parameter c as in (130), $b = \frac{R^2}{5}$, $G = \frac{R^4}{150\ln\frac{4(K+1)}{\beta}}$:

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

Equivalently, we have

$$\mathbb{P}\{E_①\} \geq 1 - \frac{\beta}{2(K+1)}, \quad \text{for} \quad E_① = \left\{\text{either} \quad \sum_{t=0}^{T-1}\sigma_t^2 > \frac{R^4}{150\ln\frac{4(K+1)}{\beta}} \quad \text{or} \quad |①| \leq \frac{R^2}{5}\right\}. \tag{132}$$

In addition, E_{T-1} implies that

$$\begin{aligned}\sum_{t=0}^{T-1}\sigma_t^2 &\overset{(131)}{\leq} 16\gamma^2R^2\sum_{t=0}^{T-1}\mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right] \overset{(129)}{\leq} 288\gamma^2R^2\sigma^\alpha T\lambda^{2-\alpha} \\ &\overset{(114)}{=} \frac{9\cdot 40^\alpha R^{4-\alpha}\sigma^\alpha T\gamma^\alpha}{50\ln^{2-\alpha}\frac{4(K+1)}{\beta}} \overset{(113)}{\leq} \frac{R^4}{150\ln\frac{4(K+1)}{\beta}}.\end{aligned} \tag{133}$$

Upper bound for ②. From E_{T-1} it follows that

$$\begin{aligned}② &= -2\gamma\sum_{t=0}^{T-1}\langle\theta_t^b, \eta_t\rangle \leq 2\gamma\sum_{t=0}^{T-1}\|\theta_t^b\| \cdot \|\eta_t\| \overset{(123),(128)}{\leq} \frac{4\cdot 2^\alpha\gamma\sigma^\alpha TR}{\lambda^{\alpha-1}} \\ &\overset{(114)}{=} \frac{80^\alpha}{10}\cdot\frac{\sigma^\alpha TR^{2-\alpha}\gamma^\alpha}{\ln^{1-\alpha}\frac{4(K+1)}{\beta}} \overset{(113)}{\leq} \frac{R^2}{5}.\end{aligned} \tag{134}$$

Upper bound for ③. First, we have

$$\mathbb{E}_{\xi^t}\left[2\gamma^2\left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]\right)\right] = 0.$$

Next, sum ③ has bounded with probability 1 terms:

$$\left|2\gamma^2\left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]\right)\right| \leq 2\gamma^2\left(\|\theta_t^u\|^2 + \mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]\right) \overset{(127)}{\leq} 16\gamma^2\lambda^2 \overset{(114)}{=} \frac{R^2}{100\ln^2\frac{4(K+1)}{\beta}} \leq \frac{R^2}{5\ln\frac{4(K+1)}{\beta}} \overset{\text{def}}{=} c. \tag{135}$$

The summands also have bounded conditional variances $\widetilde{\sigma}_t^2 \overset{\text{def}}{=} \mathbb{E}_{\xi^t}\left[4\gamma^4\left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]\right)^2\right]$:

$$\widetilde{\sigma}_t^2 \overset{(135)}{\leq} \frac{R^2}{5\ln\frac{4(K+1)}{\beta}}\mathbb{E}_{\xi^t}\left[2\gamma^2\left|\|\theta_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]\right|\right] \leq \frac{4\gamma^2R^2}{5\ln\frac{4(K+1)}{\beta}}\mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right], \tag{136}$$

since $\ln\frac{4(K+1)}{\beta} \geq 1$. In other words, we showed that $\left\{2\gamma^2\left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]\right)\right\}_{t=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\widetilde{\sigma}_t^2\}_{t=0}^{T-1}$. Next, we apply Bernstein's inequality (Lemma B.2) with $X_t = 2\gamma^2\left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right]\right)$, parameter c as in (135), $b = \frac{R^2}{5}$, $G = \frac{R^4}{150\ln\frac{4(K+1)}{\beta}}$:

$$\mathbb{P}\left\{|③| > \frac{R^2}{5} \quad \text{and} \quad \sum_{t=0}^{T-1}\widetilde{\sigma}_t^2 \leq \frac{R^4}{150\ln\frac{4(K+1)}{\beta}}\right\} \leq 2\exp\left(-\frac{b^2}{2G + {}^{2cb}/_3}\right) = \frac{\beta}{2(K+1)}.$$

Equivalently, we have

$$\mathbb{P}\{E_③\} \geq 1 - \frac{\beta}{2(K+1)}, \quad \text{for} \quad E_③ = \left\{\text{either} \quad \sum_{t=0}^{T-1}\widetilde{\sigma}_t^2 > \frac{R^4}{150\ln\frac{4(K+1)}{\beta}} \quad \text{or} \quad |③| \leq \frac{R^2}{5}\right\}. \tag{137}$$

In addition, E_{T-1} implies that

$$\sum_{t=0}^{T-1}\widetilde{\sigma}_t^2 \overset{(136)}{\leq} \frac{4\gamma^2R^2}{5\ln\frac{4(K+1)}{\beta}}\sum_{t=0}^{T-1}\mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right] \overset{(129)}{\leq} \frac{72\gamma^2R^2\lambda^{2-\alpha}\sigma^\alpha T}{5\ln\frac{4(K+1)}{\beta}} \overset{(114)}{=} \frac{9\cdot 40^\alpha}{1000}\cdot\frac{\sigma^\alpha TR^{4-\alpha}\gamma^\alpha}{\ln^{3-\alpha}\frac{4(K+1)}{\beta}} \overset{(113)}{\leq} \frac{R^4}{150\ln\frac{4(K+1)}{\beta}}. \tag{138}$$

Upper bound for ④. From E_{T-1} it follows that

$$④ = 2\gamma^2\sum_{t=0}^{T-1}\mathbb{E}_{\xi^t}\left[\|\theta_t^u\|^2\right] \overset{(129)}{\leq} 36\gamma^2\lambda^{2-\alpha}\sigma^\alpha T \overset{(114)}{=} \frac{9\cdot 40^\alpha}{400}\cdot\frac{\gamma^\alpha\sigma^\alpha TR^{2-\alpha}}{\ln^{2-\alpha}\frac{4(K+1)}{\beta}} \overset{(113)}{\leq} \frac{R^2}{5}. \tag{139}$$

Upper bound for ⑤. From E_{T-1} it follows that

$$⑤ = 2\gamma^2\sum_{t=0}^{T-1}\left\|\theta_t^b\right\|^2 \overset{(128)}{\leq} \frac{2\cdot 4^\alpha\sigma^{2\alpha}T\gamma^2}{\lambda^{2(\alpha-1)}} \overset{(114)}{=} \frac{6400^\alpha}{800}\cdot\frac{\sigma^{2\alpha}T\gamma^{2\alpha}R^{2-2\alpha}}{\ln^{2(1-\alpha)}\frac{4(K+1)}{\beta}} \overset{(113)}{\leq} \frac{R^2}{5}. \tag{140}$$

Now, we have the upper bounds for ①, ②, ③, ④, ⑤. In particular, probability event E_{T-1} implies

$$R_T^2 \overset{(126)}{\leq} R^2 + ① + ② + ③ + ④ + ⑤,$$
$$② \overset{(134)}{\leq} \frac{R^2}{5}, \quad ④ \overset{(139)}{\leq} \frac{R^2}{5}, \quad ⑤ \overset{(140)}{\leq} \frac{R^2}{5},$$
$$\sum_{t=0}^{T-1}\sigma_t^2 \overset{(133)}{\leq} \frac{R^4}{150\ln\frac{4(K+1)}{\beta}}, \quad \sum_{t=0}^{T-1}\widetilde{\sigma}_t^2 \overset{(138)}{\leq} \frac{R^4}{150\ln\frac{4(K+1)}{\beta}}.$$

Moreover, we also have (see (132), (137) and our induction assumption)

$$\mathbb{P}\{E_{T-1}\} \geq 1 - \frac{(T-1)\beta}{K+1}, \quad \mathbb{P}\{E_{①}\} \geq 1 - \frac{\beta}{2(K+1)}, \quad \mathbb{P}\{E_{③}\} \geq 1 - \frac{\beta}{2(K+1)},$$

where

$$E_{①} = \left\{ \text{either} \quad \sum_{t=0}^{T-1} \sigma_t^2 > \frac{R^4}{150 \ln \frac{4(K+1)}{\beta}} \quad \text{or} \quad |①| \leq \frac{R^2}{5} \right\},$$
$$E_{③} = \left\{ \text{either} \quad \sum_{t=0}^{T-1} \widetilde{\sigma}_t^2 > \frac{R^4}{150 \ln \frac{4(K+1)}{\beta}} \quad \text{or} \quad |③| \leq \frac{R^2}{5} \right\}.$$

Thus, probability event $E_{T-1} \cap E_{①} \cap E_{③}$ implies

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

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

$$\mathbb{P}\{E_T\} \geq \mathbb{P}\{E_{T-1} \cap E_{①} \cap E_{③}\} = 1 - \mathbb{P}\left\{\overline{E}_{T-1} \cup \overline{E}_{①} \cup \overline{E}_{③}\right\} \geq 1 - \mathbb{P}\{\overline{E}_{T-1}\} - \mathbb{P}\{\overline{E}_{①}\} - \mathbb{P}\{\overline{E}_{③}\} \geq 1 - \frac{T\beta}{K+1}.$$

This finishes the inductive part of our proof, i.e., for all $k = 0, 1, \ldots, K+1$ we have $\mathbb{P}\{E_k\} \geq 1 - {}^{k\beta}/_{(K+1)}$. In particular, for $k = K+1$ we have that with probability at least $1 - \beta$

$$f(\overline{x}^K) - f(x^*) \overset{(121)}{\leq} \frac{2R^2}{\gamma(K+1)}$$

and $\{x^k\}_{k=0}^K \subseteq Q$, which follows from (119).

Finally, if

$$\gamma \leq \min \left\{ \frac{1}{80L \ln \frac{4(K+1)}{\beta}}, \frac{R}{108^{\frac{1}{\alpha}} \cdot 20\sigma K^{\frac{1}{\alpha}} \left(\ln \frac{4(K+1)}{\beta}\right)^{\frac{\alpha-1}{\alpha}}} \right\},$$

then with probability at least $1 - \beta$

$$\begin{aligned} f(\overline{x}^K) - f(x^*) &\leq \frac{2R^2}{\gamma(K+1)} \\ &= \max \left\{ \frac{160LR^2 \ln \frac{4(K+1)}{\beta}}{K+1}, \frac{40 \cdot 108^{\frac{1}{\alpha}} \sigma R K^{\frac{1}{\alpha}} \left(\ln \frac{4(K+1)}{\beta}\right)^{\frac{\alpha-1}{\alpha}}}{K+1} \right\} \\ &= \mathcal{O}\left(\max \left\{ \frac{LR^2 \ln \frac{K}{\beta}}{K}, \frac{\sigma R \ln^{\frac{\alpha-1}{\alpha}} \frac{K}{\beta}}{K^{\frac{\alpha-1}{\alpha}}} \right\} \right). \end{aligned}$$

To get $f(\overline{x}^K) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ it is sufficient to choose K such that both terms in the maximum above are $\mathcal{O}(\varepsilon)$. This leads to

$$K = \mathcal{O}\left(\max \left\{ \frac{LR^2}{\varepsilon} \ln \frac{LR^2}{\varepsilon\beta}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}} \ln \left(\frac{1}{\beta} \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}} \right) \right\} \right),$$

which concludes the proof.

□

E.4. Quasi-Strongly Convex Functions

Finally, we consider clipped-SGD under smoothness and quasi-strong convexity assumptions. As the next lemma shows, the gradient of such function is quasi-strongly monotone and star-cocoercive operator.

Lemma E.7. *Consider differentiable function $f:\mathbb{R}^d\to\mathbb{R}$. If f satisfies Assumption 1.5 on some set Q with parameter μ, then operator $F(x)=\nabla f(x)$ satisfies Assumption 1.9 on Q with parameter $\mu/2$. If f satisfies Assumptions 1.3 and 1.5 with $\mu=0$ on some set Q, then operator $F(x)=\nabla f(x)$ satisfies Assumption 1.10 on Q with $\ell=2L$.*

Proof. We start with the first part. Assumption 1.5 on set Q means that for any $x\in Q$

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

For $F(x)=\nabla f(x)$ it implies that for all $x\in Q$

$$\langle F(x),x-x^*\rangle\geq f(x)-f(x^*)+\frac{\mu}{2}\|x-x^*\|^2\geq\frac{\mu}{2}\|x-x^*\|^2,$$

i.e., Assumption 1.9 holds on Q with parameter $\mu/2$ for operator $F(x)$.

Next, we prove the second part. Assume that f satisfies Assumptions 1.3 and 1.5 with $\mu=0$ on some set Q. Our goal is to show that $F(x)=\nabla f(x)$ satisfies Assumption 1.10 on Q. In view of (Gorbunov et al., 2022b, Lemma C.6), this is equivalent to showing that operator $\mathrm{Id}-\frac{1}{L}F$ is non-expansive around x^*, i.e., we need to show that $\|(\mathrm{Id}-\frac{1}{L}F)(x)-(\mathrm{Id}-\frac{1}{L}F)(x^*)\|\leq\|x-x^*\|$ for any $x\in Q$. We have

$$\begin{aligned}\left\|\left(\mathrm{Id}-\frac{1}{L}F\right)(x)-\left(\mathrm{Id}-\frac{1}{L}F\right)(x^*)\right\|^2 &= \left\|x-x^*-\frac{1}{L}F(x)\right\|^2\\ &= \|x-x^*\|^2-\frac{2}{L}\langle x-x^*,F(x)\rangle+\frac{1}{L^2}\|F(x)\|^2\\ &= \|x-x^*\|^2-\frac{2}{L}\langle x-x^*,\nabla f(x)\rangle+\frac{1}{L^2}\|\nabla f(x)\|^2\\ &\overset{(9),(7)}{\leq} \|x-x^*\|^2-\frac{2}{L}\left(f(x)-f(x^*)\right)+\frac{2}{L}\left(f(x)-f(x^*)\right)\\ &= \|x-x^*\|^2.\end{aligned}$$

This finishes the proof. □

Therefore, using the result of Theorem H.6 with $\ell:=2L$ and $\mu:=\mu/2$, we get the convergence result for clipped-SGD under smoothness and quasi-strong convexity assumptions.

Theorem E.8 (Case 4 in Theorem 3.1)**.** *Let Assumptions 1.1, 1.3, 1.5, hold for $Q=B_{2R}(x^*)=\{x\in\mathbb{R}^d\mid\|x-x^*\|\leq 2R\}$, where $R\geq\|x^0-x^*\|$, and*

$$0<\gamma \leq \min\left\{\frac{1}{800L\ln\frac{4(K+1)}{\beta}},\frac{2\ln(B_K)}{\mu(K+1)}\right\}, \tag{141}$$

$$B_K = \max\left\{2,\frac{(K+1)^{\frac{2(\alpha-1)}{\alpha}}\mu^2R^2}{4\cdot 5400^{\frac{2}{\alpha}}\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{4(K+1)}{\beta}\right)\ln^2(B_K)}\right\} \tag{142}$$

$$= \mathcal{O}\left(\max\left\{2,\frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2R^2}{\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)\ln^2\left(\max\left\{2,\frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2R^2}{\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)}\right\}\right)}\right\}\right), \tag{143}$$

$$\lambda_k = \frac{\exp(-\gamma(\mu/2)(1+k/2))R}{120\gamma\ln\frac{4(K+1)}{\beta}}, \tag{144}$$

for some $K \geq 0$ *and* $\beta \in (0,1]$ *such that* $\ln \frac{4(K+1)}{\beta} \geq 1$. *Then, after* K *iterations the iterates produced by* clipped-SGD *with probability at least* $1-\beta$ *satisfy*

$$\|x^{K+1} - x^*\|^2 \leq 2\exp(-\gamma(\mu/2)(K+1))R^2. \tag{145}$$

In particular, when γ *equals the minimum from* (141), *then the iterates produced by* clipped-SGD *after* K *iterations with probability at least* $1-\beta$ *satisfy*

$$\|x^K - x^*\|^2 = \mathcal{O}\left(\max\left\{R^2\exp\left(-\frac{\mu K}{\ell \ln\frac{K}{\beta}}\right), \frac{\sigma^2 \ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)\ln^2\left(\max\left\{2, \frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2 R^2}{\sigma^2 \ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)}\right\}\right)}{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2}\right\}\right), \tag{146}$$

meaning that to achieve $\|x^K - x^*\|^2 \leq \varepsilon$ *with probability at least* $1-\beta$ clipped-SGD *requires*

$$K = \mathcal{O}\left(\frac{L}{\mu}\ln\left(\frac{R^2}{\varepsilon}\right)\ln\left(\frac{L}{\mu\beta}\ln\frac{R^2}{\varepsilon}\right), \left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\ln\left(\frac{1}{\beta}\left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right)\ln^{\frac{\alpha}{\alpha-1}}\left(B_\varepsilon\right)\right) \tag{147}$$

iterations/oracle calls, where

$$B_\varepsilon = \max\left\{2, \frac{R^2}{\varepsilon \ln\left(\frac{1}{\beta}\left(\frac{4\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right)}\right\}.$$

F. Missing Proofs for clipped-SSTM and R-clipped-SSTM

In this section, we provide the complete formulation of the main results for clipped-SSTM and R-clipped-SSTM and the missing proofs. For brevity, we will use the following notation: $\widetilde{\nabla} f_{\xi^k}(x^{k+1}) = \texttt{clip}\left(\nabla f_{\xi^k}(x^{k+1}), \lambda_k\right)$.

Algorithm 2 Clipped Stochastic Similar Triangles Method (clipped-SSTM) (Gorbunov et al., 2020)

Input: starting point x^0, number of iterations K, stepsize parameter $a > 0$, clipping levels $\{\lambda_k\}_{k=0}^{K-1}$, smoothness constant L.

Set $A_0 = \alpha_0 = 0$, $y^0 = z^0 = x^0$
for $k = 0, \ldots, K-1$ **do**
  Set $\alpha_{k+1} = \frac{k+2}{2aL}$, $A_{k+1} = A_k + \alpha_{k+1}$
  $x^{k+1} = \frac{A_k y^k + \alpha_{k+1} z^k}{A_{k+1}}$
  Compute $\widetilde{\nabla} f_{\xi^k}(x^{k+1}) = \texttt{clip}\left(\nabla f_{\xi^k}(x^{k+1}), \lambda_k\right)$ using a fresh sample $\xi^k \sim \mathcal{D}_k$
  $z^{k+1} = z^k - \alpha_{k+1} \widetilde{\nabla} f_{\xi^k}(x^{k+1})$
  $y^{k+1} = \frac{A_k y^k + \alpha_{k+1} z^{k+1}}{A_{k+1}}$
end for

Output: y^K

F.1. Convex Functions

We start with the following lemma, which is a special case of Lemma 6 from (Gorbunov et al., 2021). This result can be seen the "optimization" part of the analysis of clipped-SSTM: the proof follows the same steps as the analysis of deterministic Similar Triangles Method (Gasnikov & Nesterov, 2016; Dvurechenskii et al., 2018) and separates stochasticity from the deterministic part of the method.

Lemma F.1 (Special case of Lemma 4.1 from (Gorbunov et al., 2021))**.** *Let Assumptions 1.3 and 1.6 with $\mu = 0$ hold on $Q = B_{3R}(x^*)$, where $R \geq \|x^0 - x^*\|$, and let stepsize parameter a satisfy $a \geq 1$. If $x^k, y^k, z^k \in B_{3R}(x^*)$ for all $k = 0, 1, \ldots, N$, $N \geq 0$, then after N iterations of clipped-SSTM for all $z \in B_{3R}(x^*)$ we have*

$$
\begin{aligned}
A_N\left(f(y^N) - f(z)\right) &\leq \frac{1}{2}\|z^0 - z\|^2 - \frac{1}{2}\|z^N - z\|^2 + \sum_{k=0}^{N-1} \alpha_{k+1}\left\langle \theta_{k+1}, z - z^k + \alpha_{k+1}\nabla f(x^{k+1})\right\rangle \\
&\quad + \sum_{k=0}^{N-1} \alpha_{k+1}^2 \left\|\theta_{k+1}\right\|^2, \qquad (148)\\
\theta_{k+1} &\overset{\text{def}}{=} \widetilde{\nabla} f_{\xi^k}(x^{k+1}) - \nabla f(x^{k+1}). \qquad (149)
\end{aligned}
$$

Proof. For completeness, we provide the full proof. Using $z^{k+1} = z^k - \alpha_{k+1}\widetilde{\nabla} f_{\xi^k}(x^{k+1})$ we get that for all $z \in B_{3R}(x^*)$ and $k = 0, 1, \ldots, N-1$

$$
\begin{aligned}
\alpha_{k+1}\left\langle \widetilde{\nabla} f_{\xi^k}(x^{k+1}), z^k - z\right\rangle &= \alpha_{k+1}\left\langle \widetilde{\nabla} f_{\xi^k}(x^{k+1}), z^k - z^{k+1}\right\rangle + \alpha_{k+1}\left\langle \widetilde{\nabla} f_{\xi^k}(x^{k+1}), z^{k+1} - z\right\rangle \\
&= \alpha_{k+1}\left\langle \widetilde{\nabla} f_{\xi^k}(x^{k+1}), z^k - z^{k+1}\right\rangle + \left\langle z^{k+1} - z^k, z - z^{k+1}\right\rangle \\
&= \alpha_{k+1}\left\langle \widetilde{\nabla} f_{\xi^k}(x^{k+1}), z^k - z^{k+1}\right\rangle - \frac{1}{2}\|z^k - z^{k+1}\|^2 \\
&\quad + \frac{1}{2}\|z^k - z\|^2 - \frac{1}{2}\|z^{k+1} - z\|^2, \qquad (150)
\end{aligned}
$$

where in the last step we apply $2\langle a, b\rangle = \|a + b\|^2 - \|a\|^2 - \|b\|^2$ with $a = z^{k+1} - z^k$ and $b = z - z^{k+1}$. The update rules (22) and (20) give the following formula:

$$
y^{k+1} = \frac{A_k y^k + \alpha_{k+1} z^{k+1}}{A_{k+1}} = \frac{A_k y^k + \alpha_{k+1} z^k}{A_{k+1}} + \frac{\alpha_{k+1}}{A_{k+1}}\left(z^{k+1} - z^k\right) = x^{k+1} + \frac{\alpha_{k+1}}{A_{k+1}}\left(z^{k+1} - z^k\right). \qquad (151)
$$

It implies

$$\begin{aligned}\alpha_{k+1}\left\langle \widetilde{\nabla} f_{\xi^k}(x^{k+1}), z^k - z\right\rangle &\overset{(149),(150)}{\leq} \alpha_{k+1}\left\langle \nabla f(x^{k+1}), z^k - z^{k+1}\right\rangle - \frac{1}{2}\|z^k - z^{k+1}\|^2\\ &\quad +\alpha_{k+1}\left\langle \theta_{k+1}, z^k - z^{k+1}\right\rangle + \frac{1}{2}\|z^k - z\|^2 - \frac{1}{2}\|z^{k+1} - z\|^2\\ &\overset{(151)}{=} A_{k+1}\left\langle \nabla f(x^{k+1}), x^{k+1} - y^{k+1}\right\rangle - \frac{1}{2}\|z^k - z^{k+1}\|^2\\ &\quad +\alpha_{k+1}\left\langle \theta_{k+1}, z^k - z^{k+1}\right\rangle + \frac{1}{2}\|z^k - z\|^2 - \frac{1}{2}\|z^{k+1} - z\|^2\\ &\overset{(36)}{\leq} A_{k+1}\left(f(x^{k+1}) - f(y^{k+1})\right) + \frac{A_{k+1}L}{2}\|x^{k+1} - y^{k+1}\|^2 - \frac{1}{2}\|z^k - z^{k+1}\|^2\\ &\quad +\alpha_{k+1}\left\langle \theta_{k+1}, z^k - z^{k+1}\right\rangle + \frac{1}{2}\|z^k - z\|^2 - \frac{1}{2}\|z^{k+1} - z\|^2\\ &\overset{(151)}{=} A_{k+1}\left(f(x^{k+1}) - f(y^{k+1})\right) + \frac{1}{2}\left(\frac{\alpha_{k+1}^2 L}{A_{k+1}} - 1\right)\|z^k - z^{k+1}\|^2\\ &\quad +\alpha_{k+1}\left\langle \theta_{k+1}, z^k - z^{k+1}\right\rangle + \frac{1}{2}\|z^k - z\|^2 - \frac{1}{2}\|z^{k+1} - z\|^2,\end{aligned}$$

where in the third inequality we use $x^{k+1}, y^{k+1} \in B_{3R}(x^*)$. Since $A_{k+1} \geq aL_{k+1}\alpha_{k+1}^2$ (Lemma B.1) and $a \geq 1$ we can continue our derivation as follows:

$$\begin{aligned}\alpha_{k+1}\left\langle \widetilde{\nabla} f_{\xi^k}(x^{k+1}), z^k - z\right\rangle &\leq A_{k+1}\left(f(x^{k+1}) - f(y^{k+1})\right) + \alpha_{k+1}\left\langle \theta_{k+1}, z^k - z^{k+1}\right\rangle\\ &\quad +\frac{1}{2}\|z^k - z\|^2 - \frac{1}{2}\|z^{k+1} - z\|^2. \end{aligned} \tag{152}$$

Convexity of f gives

$$\begin{aligned}\left\langle \widetilde{\nabla} f_{\xi^k}(x^{k+1}), y^k - x^{k+1}\right\rangle &\overset{(149)}{=} \left\langle \nabla f(x^{k+1}), y^k - x^{k+1}\right\rangle + \left\langle \theta_{k+1}, y^k - x^{k+1}\right\rangle\\ &\leq f(y^k) - f(x^{k+1}) + \left\langle \theta_{k+1}, y^k - x^{k+1}\right\rangle. \end{aligned} \tag{153}$$

The definition of x^{k+1} (20) implies

$$\alpha_{k+1}\left(x^{k+1} - z^k\right) = A_k\left(y^k - x^{k+1}\right) \tag{154}$$

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

$$\begin{aligned}\alpha_{k+1}\left\langle \widetilde{\nabla} f_{\xi^k}(x^{k+1}), x^{k+1} - z\right\rangle &= \alpha_{k+1}\left\langle \widetilde{\nabla} f_{\xi^k}(x^{k+1}), x^{k+1} - z^k\right\rangle + \alpha_{k+1}\left\langle \widetilde{\nabla} f_{\xi^k}(x^{k+1}), z^k - z\right\rangle\\ &\overset{(154)}{=} A_k\left\langle \widetilde{\nabla} f_{\xi^k}(x^{k+1}), y^k - x^{k+1}\right\rangle + \alpha_{k+1}\left\langle \widetilde{\nabla} f_{\xi^k}(x^{k+1}), z^k - z\right\rangle\\ &\overset{(153),(152)}{\leq} A_k\left(f(y^k) - f(x^{k+1})\right) + A_k\left\langle \theta_{k+1}, y^k - x^{k+1}\right\rangle\\ &\quad +A_{k+1}\left(f(x^{k+1}) - f(y^{k+1})\right) + \alpha_{k+1}\left\langle \theta_{k+1}, z^k - z^{k+1}\right\rangle\\ &\quad +\frac{1}{2}\|z^k - z\|^2 - \frac{1}{2}\|z^{k+1} - z\|^2\\ &\overset{(154)}{=} A_k f(y^k) - A_{k+1} f(y^{k+1}) + \alpha_{k+1}\left\langle \theta_{k+1}, x^{k+1} - z^k\right\rangle\\ &\quad +\alpha_{k+1} f(x^{k+1}) + \alpha_{k+1}\left\langle \theta_{k+1}, z^k - z^{k+1}\right\rangle\\ &\quad +\frac{1}{2}\|z^k - z\|^2 - \frac{1}{2}\|z^{k+1} - z\|^2\\ &= A_k f(y^k) - A_{k+1} f(y^{k+1}) + \alpha_{k+1} f(x^{k+1})\\ &\quad +\alpha_{k+1}\left\langle \theta_{k+1}, x^{k+1} - z^{k+1}\right\rangle + \frac{1}{2}\|z^k - z\|^2 - \frac{1}{2}\|z^{k+1} - z\|^2.\end{aligned}$$

Rearranging the terms, we get

$$\begin{aligned}
A_{k+1}f(y^{k+1}) - A_k f(y^k) &\leq \alpha_{k+1}\left(f(x^{k+1}) + \left\langle \widetilde{\nabla} f_{\xi^k}(x^{k+1}), z - x^{k+1}\right\rangle\right) + \frac{1}{2}\|z^k - z\|^2 \\
&\quad -\frac{1}{2}\|z^{k+1} - z\|^2 + \alpha_{k+1}\left\langle \theta_{k+1}, x^{k+1} - z^{k+1}\right\rangle \\
&\overset{(149)}{=} \alpha_{k+1}\left(f(x^{k+1}) + \langle \nabla f(x^{k+1}), z - x^{k+1}\rangle\right) \\
&\quad +\alpha_{k+1}\left\langle \theta_{k+1}, z - x^{k+1}\right\rangle + \frac{1}{2}\|z^k - z\|^2 - \frac{1}{2}\|z^{k+1} - z\|^2 \\
&\quad +\alpha_{k+1}\left\langle \theta_{k+1}, x^{k+1} - z^{k+1}\right\rangle \\
&\leq \alpha_{k+1}f(z) + \frac{1}{2}\|z^k - z\|^2 - \frac{1}{2}\|z^{k+1} - z\|^2 + \alpha_{k+1}\left\langle \theta_{k+1}, z - z^{k+1}\right\rangle,
\end{aligned}$$

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

$$\begin{aligned}
A_N f(y^N) &\leq A_N f(z) + \frac{1}{2}\|z^0 - z\|^2 - \frac{1}{2}\|z^N - z\|^2 + \sum_{k=0}^{N-1}\alpha_{k+1}\left\langle \theta_{k+1}, z - z^{k+1}\right\rangle \\
&= A_N f(z) + \frac{1}{2}\|z^0 - z\|^2 - \frac{1}{2}\|z^N - z\|^2 + \sum_{k=0}^{N-1}\alpha_{k+1}\left\langle \theta_{k+1}, z - z^k + \alpha_{k+1}\widetilde{\nabla} f_{\xi^k}(x^{k+1})\right\rangle \\
&\overset{(149)}{=} A_N f(z) + \frac{1}{2}\|z^0 - z\|^2 - \frac{1}{2}\|z^N - z\|^2 + \sum_{k=0}^{N-1}\alpha_{k+1}\left\langle \theta_{k+1}, z - z^k + \alpha_{k+1}\nabla f_{\xi^k}(x^{k+1})\right\rangle \\
&\quad + \sum_{k=0}^{N-1}\alpha_{k+1}^2\left\|\theta_{k+1}\right\|^2,
\end{aligned}$$

which concludes the proof. □

Using this lemma we prove the main convergence result for clipped-SSTM.

Theorem F.2 (Full version of Theorem 3.2)**.** *Let Assumptions 1.1, 1.3 and 1.6 with $\mu = 0$ hold on $Q = B_{3R}(x^*)$, where $R \geq \|x^0 - x^*\|$, and*

$$a \geq \max\left\{48600\ln^2\frac{4K}{\beta}, \frac{900\sigma(K+1)K^{\frac{1}{\alpha}}\ln^{\frac{\alpha-1}{\alpha}}\frac{4K}{\beta}}{LR}\right\}, \tag{155}$$

$$\lambda_k = \frac{R}{30\alpha_{k+1}\ln\frac{4K}{\beta}}, \tag{156}$$

for some $K > 0$ and $\beta \in (0,1]$ such that $\ln\frac{4K}{\beta} \geq 1$. Then, after K iterations of clipped-SSTM *the iterates with probability at least $1-\beta$ satisfy*

$$f(y^K) - f(x^*) \leq \frac{6aLR^2}{K(K+3)} \quad \text{and} \quad \{x^k\}_{k=0}^{K+1}, \{z^k\}_{k=0}^{K}, \{y^k\}_{k=0}^{K} \subseteq B_{2R}(x^*). \tag{157}$$

In particular, when parameter a equals the maximum from (155), then the iterates produced by clipped-SSTM *after K iterations with probability at least $1-\beta$ satisfy*

$$f(y^K) - f(x^*) = \mathcal{O}\left(\max\left\{\frac{LR^2\ln^2\frac{K}{\beta}}{K^2}, \frac{\sigma R\ln^{\frac{\alpha-1}{\alpha}}\frac{K}{\beta}}{K^{\frac{\alpha-1}{\alpha}}}\right\}\right), \tag{158}$$

meaning that to achieve $f(y^K) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$ clipped-SSTM *requires*

$$K = \mathcal{O}\left(\sqrt{\frac{LR^2}{\varepsilon}}\ln\frac{LR^2}{\varepsilon\beta}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\ln\left(\frac{1}{\beta}\left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right) \quad \text{iterations/oracle calls.} \tag{159}$$

Proof. The proof starts similarly to the proof of Theorem 4.1 from (Gorbunov et al., 2021). Let $R_k = \|z^k - x^*\|$, $\widetilde{R}_0 = R_0$, $\widetilde{R}_{k+1} = \max\{\widetilde{R}_k, R_{k+1}\}$ for all $k \geq 0$. We first show by induction that for all $k \geq 0$ the iterates x^{k+1}, z^k, y^k lie in $B_{\widetilde{R}_k}(x^*)$. The induction base is trivial since $y^0 = z^0$, $\widetilde{R}_0 = R_0$, and $x^1 = \frac{A_0 y^0 + \alpha_1 z^0}{A_1} = z^0$. Next, assume that $x^l, z^{l-1}, y^{l-1} \in B_{\widetilde{R}_{l-1}}(x^*)$ for some $l \geq 1$. By definitions of R_l and $\widetilde{R}_l$ we have that $z^l \in B_{R_l}(x^*) \subseteq B_{\widetilde{R}_l}(x^*)$. Since y^l is a convex combination of $y^{l-1} \in B_{\widetilde{R}_{l-1}}(x^*) \subseteq B_{\widetilde{R}_l}(x^*)$, $z^l \in B_{\widetilde{R}_l}(x^*)$ and $B_{\widetilde{R}_l}(x^*)$ is a convex set we conclude that $y^l \in B_{\widetilde{R}_l}(x^*)$. Finally, since x^{l+1} is a convex combination of y^l and z^l we have that x^{l+1} lies in $B_{\widetilde{R}_l}(x^*)$ as well.

Next, our goal is to show by induction that $\widetilde{R}_l \leq 3R$ with high probability, which allows us to apply the result of Lemma F.1 and then use Bernstein's inequality to estimate the stochastic part of the upper-bound. More precisely, for each $k = 0, \ldots, K$ we consider probability event E_k defined as follows: inequalities

$$\sum_{l=0}^{t-1} \alpha_{l+1} \left\langle \theta_{l+1}, x^* - z^l + \alpha_{l+1} \nabla f_{\xi^l}(x^{l+1}) \right\rangle + \sum_{l=0}^{t-1} \alpha_{l+1}^2 \left\|\theta_{l+1}\right\|^2 \leq R^2, \tag{160}$$

$$R_t \leq 2R \tag{161}$$

hold for all $t = 0, 1, \ldots, k$ simultaneously. We want to prove via induction that $\mathbb{P}\{E_k\} \geq 1 - {k\beta}/{K}$ for all $k = 0, 1, \ldots, K$. For $k = 0$ the statement is trivial: the left-hand side of (160) equals zero and $R \geq R_0$ by definition. Assume that the statement is true for some $k = T - 1 \leq K - 1$: $\mathbb{P}\{E_{T-1}\} \geq 1 - {(T-1)\beta}/{K}$. One needs to prove that $\mathbb{P}\{E_T\} \geq 1 - {T\beta}/{K}$. First, we notice that probability event E_{T-1} implies that $\widetilde{R}_t \leq 2R$ for all $t = 0, 1, \ldots, T-1$. Moreover, it implies that

$$\|z^T - x^*\| \overset{(21)}{\leq} \|z^T - x^*\| + \alpha_T \|\widetilde{\nabla} f_{\xi^{T-1}}(x^T)\| \leq 2R + \alpha_T \lambda_{T-1} \overset{(156)}{\leq} 3R.$$

Therefore, E_{T-1} implies $\{x^k\}_{k=0}^T, \{z^k\}_{k=0}^T, \{y^k\}_{k=0}^T \subseteq B_{3R}(x^*)$, meaning that the assumptions of Lemma F.1 are satisfied and we have

$$A_t \left(f(y^t) - f(x^*)\right) \leq \frac{1}{2}R_0^2 - \frac{1}{2}R_t^2 + \sum_{l=0}^{t-1} \alpha_{l+1} \left\langle \theta_{l+1}, x^* - z^l + \alpha_{l+1} \nabla f(x^{l+1}) \right\rangle + \sum_{l=0}^{t-1} \alpha_{l+1}^2 \left\|\theta_{l+1}\right\|^2 \tag{162}$$

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

$$f(y^t) - f(x^*) \overset{(160),(162)}{\leq} \frac{\frac{1}{2}R_0^2 - \frac{1}{2}R_t^2 + R^2}{A_t} \leq \frac{3R^2}{2A_t} = \frac{6aLR^2}{t(t+3)}. \tag{163}$$

Taking into account that $f(y^T) - f(x^*) \geq 0$, we also derive that E_{T-1} implies

$$\begin{aligned} R_T^2 &\leq R_0^2 + \underbrace{2\sum_{t=0}^{T-1} \alpha_{t+1} \left\langle \theta_{t+1}, x^* - z^t + \alpha_{t+1} \nabla f(x^{t+1}) \right\rangle + 2\sum_{t=0}^{T-1} \alpha_{t+1}^2 \left\|\theta_{t+1}\right\|^2}_{2B_T} \\ &\leq R^2 + 2B_T. \end{aligned} \tag{164}$$

Before we estimate B_T, we need to derive a few useful inequalities. We start with showing that E_{T-1} implies $\|\nabla f(x^{t+1})\| \leq {\lambda_t}/{2}$ for all $t = 0, 1, \ldots, T-1$. For $t = 0$ we have $x^1 = x^0$ and

$$\|\nabla f(x^1)\| = \|\nabla f(x^0)\| \overset{(6)}{\leq} L\|x^0 - x^*\| \leq \frac{R}{a\alpha_1} = \frac{\lambda_0}{2} \cdot \frac{60 \ln \frac{4K}{\beta}}{a} \overset{(155)}{\leq} \frac{\lambda_0}{2}. \tag{165}$$

Next, for $t = 1, \ldots, T-1$ event E_{T-1} implies

$$
\begin{aligned}
\|\nabla f(x^{t+1})\| &\leq \|\nabla f(x^{t+1}) - \nabla f(y^t)\| + \|\nabla f(y^t)\| \\
&\overset{(6),(7)}{\leq} L\|x^{t+1} - y^t\| + \sqrt{2L\left(f(y^t) - f(x^*)\right)} \\
&\overset{(154),(163)}{\leq} \frac{L\alpha_{t+1}}{A_t}\|x^{t+1} - z^t\| + \sqrt{\frac{12aL^2R^2}{t(t+3)}} \\
&\leq \frac{4LR\alpha_{t+1}}{A_t} + \sqrt{\frac{12aL^2R^2}{t(t+3)}} \\
&= \frac{R}{60\alpha_{t+1}\ln\frac{4K}{\beta}}\left(\frac{240L\alpha_{t+1}^2\ln\frac{4K}{\beta}}{A_t} + 60\sqrt{\frac{12aL^2\alpha_{t+1}^2\ln^2\frac{4K}{\beta}}{t(t+3)}}\right) \\
&\overset{(38),(156)}{\leq} \frac{\lambda_t}{2}\left(\frac{240L\left(\frac{t+2}{2aL}\right)^2\ln\frac{4K}{\beta}}{\frac{t(t+3)}{4aL}} + 60\sqrt{\frac{12aL^2\left(\frac{t+2}{2aL}\right)^2\ln^2\frac{4K}{\beta}}{t(t+3)}}\right) \\
&= \frac{\lambda_t}{2}\left(\frac{240(t+2)^2\ln\frac{4K}{\beta}}{t(t+3)a} + 60\sqrt{\frac{3(t+2)^2\ln\frac{4K}{\beta}}{t(t+3)a}}\right) \\
&\leq \frac{\lambda_t}{2}\left(\frac{540\ln\frac{4K}{\beta}}{a} + \frac{90\sqrt{3}\ln\frac{4K}{\beta}}{\sqrt{a}}\right) \overset{(155)}{\leq} \frac{\lambda_t}{2},
\end{aligned} \tag{166}
$$

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

$$
\|x^* - z^t + \alpha_{t+1}\nabla f(x^{t+1})\| \leq \|x^* - z^t\| + \alpha_{t+1}\|\nabla f(x^{t+1})\| \overset{(161),(165),(166)}{\leq} 2R + \frac{R}{60\ln\frac{4K}{\beta}} \leq 3R \tag{167}
$$

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

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

for all $t = 0, 1, \ldots, T-1$. By definition these random vectors are bounded with probability 1

$$
\|\eta_t\| \leq 3R \tag{168}
$$

and probability event E_{T-1} implies that $\eta_t = x^* - z^t + \alpha_{t+1}\nabla f(x^{t+1})$ for all $t = 0, 1, \ldots, T-1$. Then, form E_{T-1} it follows that

$$
B_T = \sum_{t=0}^{T-1}\alpha_{t+1}\left\langle\theta_{t+1}, \eta_t\right\rangle + \sum_{t=0}^{T-1}\alpha_{t+1}^2\left\|\theta_{t+1}\right\|^2.
$$

Next, we define the unbiased part and the bias of θ_{t+1} as θ_{t+1}^u and θ_{t+1}^b respectively:

$$
\theta_{t+1}^u = \widetilde{\nabla} f_{\xi^t}(x^{t+1}) - \mathbb{E}_{\xi^t}\left[\widetilde{\nabla} f_{\xi^t}(x^{t+1})\right], \quad \theta_{t+1}^b = \mathbb{E}_{\xi^t}\left[\widetilde{\nabla} f_{\xi^t}(x^{t+1})\right] - \nabla f(x^{t+1}). \tag{169}
$$

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

$$
\begin{aligned}
B_T &= \sum_{t=0}^{T-1} \alpha_{t+1} \left\langle \theta_{t+1}^u + \theta_{t+1}^b, \eta_t \right\rangle + \sum_{t=0}^{T-1} \alpha_{t+1}^2 \left\| \theta_{t+1}^u + \theta_{t+1}^b \right\|^2 \\
&\leq \underbrace{\sum_{t=0}^{T-1} \alpha_{t+1} \left\langle \theta_{t+1}^u, \eta_t \right\rangle}_{①} + \underbrace{\sum_{t=0}^{T-1} \alpha_{t+1} \left\langle \theta_{t+1}^b, \eta_t \right\rangle}_{②} + \underbrace{2\sum_{t=0}^{T-1} \alpha_{t+1}^2 \left(\left\| \theta_{t+1}^u \right\|^2 - \mathbb{E}_{\xi^t}\left[\left\| \theta_{t+1}^u \right\|^2 \right] \right)}_{③} \\
&\quad \underbrace{+2\sum_{t=0}^{T-1} \alpha_{t+1}^2 \mathbb{E}_{\xi^t}\left[\left\| \theta_{t+1}^u \right\|^2 \right]}_{④} \underbrace{+2\sum_{t=0}^{T-1} \alpha_{t+1}^2 \left\| \theta_{t+1}^b \right\|^2}_{⑤}.
\end{aligned} \tag{170}
$$

It remains to derive good enough high-probability upper-bounds for the terms ①, ②, ③, ④, ⑤, i.e., to finish our inductive proof we need to show that ① + ② + ③ + ④ + ⑤ $\leq R^2$ with high probability. In the subsequent parts of the proof, we will need to use many times the bounds for the norm and second moments of θ_{t+1}^u and θ_{t+1}^b. First, by definition of clipping operator, we have with probability 1 that

$$
\|\theta_{t+1}^u\| \leq 2\lambda_t. \tag{171}
$$

Moreover, since E_{T-1} implies that $\|\nabla f(x^{t+1})\| \leq \lambda_t/2$ for $t = 0, 1, \ldots, T-1$ (see (165) and (166)), then, in view of Lemma 5.1, we have that E_{T-1} implies

$$
\|\theta_{t+1}^b\| \leq \frac{2^\alpha \sigma^\alpha}{\lambda_t^{\alpha-1}}, \tag{172}
$$

$$
\mathbb{E}_{\xi^t}\left[\|\theta_{t+1}^u\|^2 \right] \leq 18\lambda_t^{2-\alpha}\sigma^\alpha. \tag{173}
$$

Upper bound for ①. By definition of θ_{t+1}^u, we have $\mathbb{E}_{\xi^t}[\theta_{t+1}^u] = 0$ and

$$
\mathbb{E}_{\xi^t}\left[\alpha_{t+1} \left\langle \theta_{t+1}^u, \eta_t \right\rangle \right] = 0.
$$

Next, sum ① has bounded with probability 1 terms:

$$
|\alpha_{t+1} \left\langle \theta_{t+1}^u, \eta_t \right\rangle| \leq \alpha_{t+1}\|\theta_{t+1}^u\| \cdot \|\eta_t\| \overset{(168),(171)}{\leq} 6\alpha_{t+1}\lambda_t R \overset{(156)}{=} \frac{R^2}{5\ln\frac{4K}{\beta}} \overset{\text{def}}{=} c. \tag{174}
$$

The summands also have bounded conditional variances $\sigma_t^2 \overset{\text{def}}{=} \mathbb{E}_{\xi^t}[\alpha_{t+1}^2 \left\langle \theta_{t+1}^u, \eta_t \right\rangle^2]$:

$$
\sigma_t^2 \leq \mathbb{E}_{\xi^t}\left[\alpha_{t+1}^2 \|\theta_{t+1}^u\|^2 \cdot \|\eta_t\|^2 \right] \overset{(168)}{\leq} 9\alpha_{t+1}^2 R^2 \mathbb{E}_{\xi^t}\left[\|\theta_{t+1}^u\|^2 \right]. \tag{175}
$$

In other words, we showed that $\{\alpha_{t+1} \left\langle \theta_{t+1}^u, \eta_t \right\rangle\}_{t=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\sigma_t^2\}_{t=0}^{T-1}$. Next, we apply Bernstein's inequality (Lemma B.2) with $X_t = \alpha_{t+1} \left\langle \theta_{t+1}^u, \eta_t \right\rangle$, parameter c as in (174), $b = \frac{R^2}{5}$, $G = \frac{R^4}{150\ln\frac{4K}{\beta}}$:

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

Equivalently, we have

$$
\mathbb{P}\{E_①\} \geq 1 - \frac{\beta}{2K}, \quad \text{for} \quad E_① = \left\{ \text{either} \quad \sum_{t=0}^{T-1} \sigma_t^2 > \frac{R^4}{150\ln\frac{4K}{\beta}} \quad \text{or} \quad |①| \leq \frac{R^2}{5} \right\}. \tag{176}
$$

In addition, E_{T-1} implies that

$$\begin{aligned}\sum_{t=0}^{T-1}\sigma_t^2 &\overset{(175)}{\leq} 9R^2\sum_{t=0}^{T-1}\alpha_{t+1}^2\mathbb{E}_{\xi^t}\left[\|\theta_{t+1}^u\|^2\right] \overset{(173)}{\leq} 162\sigma^\alpha R^2\sum_{t=0}^{T-1}\alpha_{t+1}^2\lambda_t^{2-\alpha}\\ &\overset{(156)}{\leq} \frac{162\sigma^\alpha R^{4-\alpha}}{30^{2-\alpha}\ln^{2-\alpha}\frac{4K}{\beta}}\sum_{t=0}^{T-1}\alpha_{t+1}^\alpha = \frac{162\sigma^\alpha R^{4-\alpha}}{30^{2-\alpha}\cdot 2^\alpha a^\alpha L^\alpha\ln^{2-\alpha}\frac{4K}{\beta}}\sum_{t=0}^{T-1}(t+2)^\alpha\\ &\leq \frac{1}{a^\alpha}\cdot\frac{162\sigma^\alpha R^{4-\alpha}T(T+1)^\alpha}{60L^\alpha\ln^{2-\alpha}\frac{4K}{\beta}} \overset{(155)}{\leq} \frac{R^4}{150\ln\frac{4K}{\beta}}. \end{aligned}\tag{177}$$

Upper bound for ②. From E_{T-1} it follows that

$$\begin{aligned}② &\leq \sum_{t=0}^{T-1}\alpha_{t+1}\|\theta_{t+1}^b\|\cdot\|\eta_t\| \overset{(168),(172)}{\leq} 3R\cdot 2^\alpha\sigma^\alpha\sum_{t=0}^{T-1}\frac{\alpha_{t+1}}{\lambda_t^{\alpha-1}} \overset{(156)}{\leq} 12R\sigma^\alpha\cdot\frac{30^{\alpha-1}\ln^{\alpha-1}\frac{4K}{\beta}}{R^{\alpha-1}}\sum_{t=0}^{T-1}\alpha_{t+1}^\alpha\\ &\leq \frac{360\sigma^\alpha R^{2-\alpha}\ln^{\alpha-1}\frac{4K}{\beta}}{2^\alpha a^\alpha L^\alpha}\sum_{t=0}^{T-1}(t+2)^\alpha \leq \frac{1}{a^\alpha}\cdot\frac{180\sigma^\alpha R^{2-\alpha}T(T+1)^\alpha\ln^{\alpha-1}\frac{4K}{\beta}}{L^\alpha} \overset{(155)}{\leq} \frac{R^2}{5}. \end{aligned}\tag{178}$$

Upper bound for ③. First, we have

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

Next, sum ③ has bounded with probability 1 terms:

$$\begin{aligned}\left|2\alpha_{t+1}^2\left(\left\|\theta_{t+1}^u\right\|^2 - \mathbb{E}_{\xi^t}\left[\left\|\theta_{t+1}^u\right\|^2\right]\right)\right| &\leq 2\alpha_{t+1}^2\left(\|\theta_{t+1}^u\|^2 + \mathbb{E}_{\xi^t}\left[\left\|\theta_{t+1}^u\right\|^2\right]\right)\\ &\overset{(171)}{\leq} 16\alpha_{t+1}^2\lambda_t^2 \overset{(156)}{\leq} \frac{R^2}{5\ln\frac{4K}{\beta}} \overset{\text{def}}{=} c. \end{aligned}\tag{179}$$

The summands also have bounded conditional variances $\widetilde{\sigma}_t^2 \overset{\text{def}}{=} \mathbb{E}_{\xi^t}\left[4\alpha_{t+1}^4\left(\left\|\theta_{t+1}^u\right\|^2 - \mathbb{E}_{\xi^t}\left[\left\|\theta_{t+1}^u\right\|^2\right]\right)^2\right]$:

$$\widetilde{\sigma}_t^2 \overset{(179)}{\leq} \frac{R^2}{5\ln\frac{4K}{\beta}}\mathbb{E}_{\xi^t}\left[2\alpha_{t+1}^2\left|\left\|\theta_{t+1}^u\right\|^2 - \mathbb{E}_{\xi^k}\left[\left\|\theta_{t+1}^u\right\|^2\right]\right|\right] \leq \alpha_{t+1}^2R^2\mathbb{E}_{\xi^t}\left[\|\theta_{t+1}^u\|^2\right], \tag{180}$$

since $\ln\frac{4K}{\beta} \geq 1$. In other words, we showed that $\left\{2\alpha_{t+1}^2\left(\left\|\theta_{t+1}^u\right\|^2 - \mathbb{E}_{\xi^t}\left[\left\|\theta_{t+1}^u\right\|^2\right]\right)\right\}_{t=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\widetilde{\sigma}_t^2\}_{t=0}^{T-1}$. Next, we apply Bernstein's inequality (Lemma B.2) with $X_t = 2\alpha_{t+1}^2\left(\left\|\theta_{t+1}^u\right\|^2 - \mathbb{E}_{\xi^t}\left[\left\|\theta_{t+1}^u\right\|^2\right]\right)$, parameter c as in (179), $b = \frac{R^2}{5}$, $G = \frac{R^4}{150\ln\frac{4K}{\beta}}$:

$$\mathbb{P}\left\{|③| > \frac{R^2}{5} \quad\text{and}\quad \sum_{t=0}^{T-1}\widetilde{\sigma}_t^2 \leq \frac{R^4}{150\ln\frac{4K}{\beta}}\right\} \leq 2\exp\left(-\frac{b^2}{2G + 2cb/3}\right) = \frac{\beta}{2K}.$$

Equivalently, we have

$$\mathbb{P}\{E_③\} \geq 1 - \frac{\beta}{2K}, \quad \text{for} \quad E_③ = \left\{\text{either} \quad \sum_{t=0}^{T-1}\widetilde{\sigma}_t^2 > \frac{R^4}{150\ln\frac{4K}{\beta}} \quad\text{or}\quad |③| \leq \frac{R^2}{5}\right\}. \tag{181}$$

In addition, E_{T-1} implies that

$$\sum_{t=0}^{T-1}\widetilde{\sigma}_t^2 \overset{(180)}{\leq} R^2\sum_{t=0}^{T-1}\alpha_{t+1}^2\mathbb{E}_{\xi^t}\left[\|\theta_{t+1}^u\|^2\right] \leq 9R^2\sum_{t=0}^{T-1}\alpha_{t+1}^2\mathbb{E}_{\xi^t}\left[\|\theta_{t+1}^u\|^2\right] \overset{(177)}{\leq} \frac{R^4}{150\ln\frac{4K}{\beta}}. \tag{182}$$

Upper bound for ④. From E_{T-1} it follows that

$$④ = 2\sum_{t=0}^{T-1} \alpha_{t+1}^2 \mathbb{E}_{\xi^t}\left[\left\|\theta_{t+1}^u\right\|^2\right] \leq \frac{1}{R^2} \cdot 9R^2 \sum_{t=0}^{T-1} \alpha_{t+1}^2 \mathbb{E}_{\xi^t}\left[\left\|\theta_{t+1}^u\right\|^2\right] \overset{(177)}{\leq} \frac{R^2}{150\ln\frac{4K}{\beta}} \leq \frac{R^2}{5}. \tag{183}$$

Upper bound for ⑤. From E_{T-1} it follows that

$$\begin{aligned} ⑤ &= 2\sum_{t=0}^{T-1} \alpha_{t+1}^2 \left\|\theta_{t+1}^b\right\|^2 \leq 2^{2\alpha+1}\sigma^{2\alpha} \sum_{t=0}^{T-1} \frac{\alpha_{t+1}^2}{\lambda_t^{2\alpha-2}} \overset{(156)}{=} \frac{2^{2\alpha+1}\cdot 30^{2\alpha-2}\sigma^{2\alpha}\ln^{2\alpha-2}\frac{4K}{\beta}}{R^{2\alpha-2}} \sum_{t=0}^{T-1} \alpha_{t+1}^{2\alpha} \\ &= \frac{2^{2\alpha+1}\cdot 30^{2\alpha-2}\sigma^{2\alpha}\ln^{2\alpha-2}\frac{4K}{\beta}}{2^{2\alpha}a^{2\alpha}L^{2\alpha}R^{2\alpha-2}} \sum_{t=0}^{T-1}(t+2)^{2\alpha} \leq \frac{1}{a^{2\alpha}} \cdot \frac{1800\sigma^{2\alpha}T(T+1)^{2\alpha}\ln^{2\alpha-2}\frac{4K}{\beta}}{L^{2\alpha}R^{2\alpha-2}} \overset{(155)}{\leq} \frac{R^2}{5}. \end{aligned} \tag{184}$$

Now, we have the upper bounds for ①, ②, ③, ④, ⑤. In particular, probability event E_{T-1} implies

$$B_T \overset{(170)}{\leq} R^2 + ① + ② + ③ + ④ + ⑤,$$
$$② \overset{(178)}{\leq} \frac{R^2}{5}, \quad ④ \overset{(183)}{\leq} \frac{R^2}{5}, \quad ⑤ \overset{(184)}{\leq} \frac{R^2}{5},$$
$$\sum_{t=0}^{T-1}\sigma_t^2 \overset{(177)}{\leq} \frac{R^4}{150\ln\frac{4K}{\beta}}, \quad \sum_{t=0}^{T-1}\widetilde{\sigma}_t^2 \overset{(182)}{\leq} \frac{R^4}{150\ln\frac{4K}{\beta}}.$$

Moreover, we also have (see (176), (181) and our induction assumption)

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

where

$$\begin{aligned} E_{①} &= \left\{\text{either} \quad \sum_{t=0}^{T-1}\sigma_t^2 > \frac{R^4}{150\ln\frac{4K}{\beta}} \quad \text{or} \quad |①| \leq \frac{R^2}{5}\right\}, \\ E_{③} &= \left\{\text{either} \quad \sum_{t=0}^{T-1}\widetilde{\sigma}_t^2 > \frac{R^4}{150\ln\frac{4K}{\beta}} \quad \text{or} \quad |③| \leq \frac{R^2}{5}\right\}. \end{aligned}$$

Thus, probability event $E_{T-1} \cap E_{①} \cap E_{③}$ implies

$$\begin{aligned} B_T &\leq R^2 + \frac{R^2}{5} + \frac{R^2}{5} + \frac{R^2}{5} + \frac{R^2}{5} + \frac{R^2}{5} = 2R^2, \\ R_T^2 &\overset{(164)}{\leq} R^2 + 2R^2 \leq (2R)^2, \end{aligned}$$

which is equivalent to (160) and (161) for $t = T$, and

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

This finishes the inductive part of our proof, i.e., for all $k = 0, 1, \ldots, K$ we have $\mathbb{P}\{E_k\} \geq 1 - {}^{k\beta}/_{K}$. In particular, for $k = K$ we have that with probability at least $1 - \beta$

$$f(y^K) - f(x^*) \overset{(163)}{\leq} \frac{6aLR^2}{K(K+3)}$$

and $\{x^k\}_{k=0}^{K+1}, \{z^k\}_{k=0}^{K}, \{y^k\}_{k=0}^{K} \subseteq B_{2R}(x^*)$, which follows from (161).

Finally, if

$$a = \max\left\{48600\ln^2\frac{4K}{\beta}, \frac{900\sigma(K+1)K^{\frac{1}{\alpha}}\ln^{\frac{\alpha-1}{\alpha}}\frac{4K}{\beta}}{LR}\right\},$$

then with probability at least $1-\beta$

$$\begin{aligned} f(y^K) - f(x^*) &\leq \frac{6aLR^2}{K(K+3)} = \max\left\{\frac{291600LR^2\ln^2\frac{4K}{\beta}}{K(K+3)}, \frac{5400\sigma R(K+1)K^{\frac{1}{\alpha}}\ln^{\frac{\alpha-1}{\alpha}}\frac{4K}{\beta}}{K(K+3)}\right\} \\ &= \mathcal{O}\left(\max\left\{\frac{LR^2\ln^2\frac{K}{\beta}}{K^2}, \frac{\sigma R\ln^{\frac{\alpha-1}{\alpha}}\frac{K}{\beta}}{K^{\frac{\alpha-1}{\alpha}}}\right\}\right). \end{aligned}$$

To get $f(y^K) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$ it is sufficient to choose K such that both terms in the maximum above are $\mathcal{O}(\varepsilon)$. This leads to

$$K = \mathcal{O}\left(\sqrt{\frac{LR^2}{\varepsilon}}\ln\frac{LR^2}{\varepsilon\beta}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\ln\left(\frac{1}{\beta}\left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right)$$

that concludes the proof. □

F.2. Strongly Convex Functions

In the strongly convex case, we consider the restarted version of clipped-SSTM (R-clipped-SSTM). The main result is summarized below.

Algorithm 3 Restarted clipped-SSTM (R-clipped-SSTM) (Gorbunov et al., 2020)

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

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

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

Theorem F.3 (Full version of Theorem 3.3). *Let Assumptions 1.1, 1.3, 1.6 with $\mu > 0$ hold for $Q = B_{3R}(x^*)$, where $R \geq \|x^0 - x^*\|^2$ and R-clipped-SSTM runs clipped-SSTM τ times. Let*

$$K_t = \left\lceil \max\left\{1080\sqrt{\frac{LR_{t-1}^2}{\varepsilon_t}}\ln\frac{2160\sqrt{LR_{t-1}^2}\tau}{\sqrt{\varepsilon_t}\beta}, 2\left(\frac{5400\sigma R_{t-1}}{\varepsilon_t}\right)^{\frac{\alpha}{\alpha-1}}\ln\left(\frac{4\tau}{\beta}\left(\frac{5400\sigma R_{t-1}}{\varepsilon_t}\right)^{\frac{\alpha}{\alpha-1}}\right)\right\}\right\rceil, \tag{185}$$

$$\varepsilon_t = \frac{\mu R_{t-1}^2}{4}, \quad R_{t-1} = \frac{R}{2^{(t-1)/2}}, \quad \tau = \left\lceil \log_2\frac{\mu R^2}{2\varepsilon}\right\rceil, \quad \ln\frac{4K_t\tau}{\beta} \geq 1, \tag{186}$$

$$a_t = \max\left\{48600\ln^2\frac{4K_t\tau}{\beta}, \frac{900\sigma(K_t+1)K_t^{\frac{1}{\alpha}}\ln^{\frac{\alpha-1}{\alpha}}\frac{4K_t\tau}{\beta}}{LR_t}\right\}, \tag{187}$$

$$\lambda_k^t = \frac{R_t}{30\alpha_{k+1}^t\ln\frac{4K_t\tau}{\beta}} \tag{188}$$

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

$$\mathcal{O}\left(\max\left\{\sqrt{\frac{L}{\mu}}\ln\left(\frac{\mu R^2}{\varepsilon}\right)\ln\left(\frac{\sqrt{L}}{\sqrt{\mu}\beta}\ln\left(\frac{\mu R^2}{\varepsilon}\right)\right), \left(\frac{\sigma^2}{\mu\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\ln\left(\frac{1}{\beta}\left(\frac{\sigma^2}{\mu\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\ln\left(\frac{\mu R^2}{\varepsilon}\right)\right)\right\}\right) \tag{189}$$

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

Proof. We show by induction that for any $t = 1, \ldots, \tau$ with probability at least $1 - {t\beta}/{\tau}$ inequalities

$$f(\hat{x}^l) - f(x^*) \leq \varepsilon_l, \quad \|\hat{x}^l - x^*\|^2 \leq R_l^2 = \frac{R^2}{2^l} \tag{190}$$

hold for $l = 1, \ldots, t$ simultaneously. First, we prove the base of the induction. Theorem F.2 implies that with probability at least $1 - {\beta}/{\tau}$

$$\begin{aligned} f(\hat{x}^1) - f(x^*) &\leq \frac{6a_1 LR^2}{K_1(K_1+3)} \overset{(187)}{=} \max\left\{\frac{291600LR^2\ln^2\frac{4K_1\tau}{\beta}}{K_1(K_1+3)}, \frac{5400\sigma R(K_1+1)K_1^{\frac{1}{\alpha}}\ln^{\frac{\alpha-1}{\alpha}}\frac{4K_1\tau}{\beta}}{K_1(K_1+3)}\right\} \\ &\leq \max\left\{\frac{291600LR^2\ln^2\frac{4K_1\tau}{\beta}}{K_1^2}, \frac{5400\sigma R\ln^{\frac{\alpha-1}{\alpha}}\frac{4K_1\tau}{\beta}}{K_1^{\frac{\alpha-1}{\alpha}}}\right\} \\ &\overset{(185)}{\leq} \varepsilon_1 = \frac{\mu R^2}{4} \end{aligned}$$

and, due to the strong convexity,

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

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

$$f(\hat{x}^l) - f(x^*) \leq \varepsilon_l, \quad \|\hat{x}^l - x^*\|^2 \leq R_l^2 = \frac{R^2}{2^l} \tag{191}$$

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

$$\begin{aligned} f(\hat{x}^{T+1}) - f(x^*) &\leq \frac{6a_{T+1}LR_T^2}{K_{T+1}(K_{T+1}+3)} \\ &\overset{(187)}{=} \max\left\{\frac{291600LR_T^2\ln^2\frac{4K_{T+1}\tau}{\beta}}{K_{T+1}(K_{T+1}+3)}, \frac{5400\sigma R_T(K_{T+1}+1)K_{T+1}^{\frac{1}{\alpha}}\ln^{\frac{\alpha-1}{\alpha}}\frac{4K_{T+1}\tau}{\beta}}{K_{T+1}(K_{T+1}+3)}\right\} \\ &\leq \max\left\{\frac{291600LR_T^2\ln^2\frac{4K_{T+1}\tau}{\beta}}{K_{T+1}^2}, \frac{5400\sigma R_T\ln^{\frac{\alpha-1}{\alpha}}\frac{4K_{T+1}\tau}{\beta}}{K_{T+1}^{\frac{\alpha-1}{\alpha}}}\right\} \\ &\overset{(185)}{\leq} \varepsilon_{T+1} = \frac{\mu R_T^2}{4} \end{aligned}$$

and, due to the strong convexity,

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

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

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

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

$$f(\hat{x}^\tau) - f(x^*) \leq \varepsilon_\tau = \frac{\mu R_{\tau-1}^2}{4} = \frac{\mu R^2}{2^{\tau+1}} \overset{(186)}{\leq} \varepsilon.$$

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

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

which concludes the proof. □

G. Missing Proofs for clipped-SEG

In this section, we provide the complete formulation of the main results for clipped-SSTM and R-clipped-SSTM and the missing proofs. For brevity, we will use the following notation: $\widetilde{F}_{\xi_1^k}(x^k) = \texttt{clip}\left(F_{\xi_1^k}(x^k), \lambda_k\right)$ and $\widetilde{F}_{\xi_2^k}(\widetilde{x}^k) = \texttt{clip}\left(F_{\xi_2^k}(\widetilde{x}^k), \lambda_k\right)$.

Algorithm 4 Clipped Stochastic Extragradient (clipped-SEG) (Gorbunov et al., 2022a)

Input: starting point x^0, number of iterations K, stepsize $\gamma > 0$, clipping levels $\{\lambda_k\}_{k=0}^{K-1}$.

for $k = 0, \ldots, K$ **do**

Compute $\widetilde{F}_{\xi_1^k}(x^k) = \texttt{clip}\left(F_{\xi_1^k}(x^k), \lambda_k\right)$ using a fresh sample $\xi_1^k \sim \mathcal{D}_k$

$\widetilde{x}^k = x^k - \gamma \widetilde{F}_{\xi_1^k}(x^k)$

Compute $\widetilde{F}_{\xi_2^k}(\widetilde{x}^k) = \texttt{clip}\left(F_{\xi_2^k}(\widetilde{x}^k), \lambda_k\right)$ using a fresh sample $\xi_2^k \sim \mathcal{D}_k$

$x^{k+1} = x^k - \gamma \widetilde{F}_{\xi_2^k}(\widetilde{x}^k)$

end for

Output: x^{K+1} or $\widetilde{x}^K_{\texttt{avg}} = \frac{1}{K+1}\sum\limits_{k=0}^{K} \widetilde{x}^K$

G.1. Monotone Problems

We start with the following lemma derived by Gorbunov et al. (2022b). Since this lemma handles only deterministic part of the algorithm, the proof is the same as in the original work.

Lemma G.1 (Lemma C.1 from (Gorbunov et al., 2022b))**.** *Let Assumptions 1.7 and 1.8 hold for $Q = B_{4R}(x^*)$, where $R \geq \|x^0 - x^*\|$ and $0 < \gamma \leq 1/\sqrt{2}L$. If x^k and $\widetilde{x}^k$ lie in $B_{4R}(x^*)$ for all $k = 0, 1, \ldots, K$ for some $K \geq 0$, then for all $u \in B_{4R}(x^*)$ the iterates produced by clipped-SEG satisfy*

$$\begin{aligned}
\langle F(u), \widetilde{x}^K_{avg} - u\rangle &\leq \frac{\|x^0 - u\|^2 - \|x^{K+1} - u\|^2}{2\gamma(K+1)} + \frac{\gamma}{2(K+1)}\sum_{k=0}^{K}\left(\|\theta_k\|^2 + 2\|\omega_k\|^2\right) \\
&\quad + \frac{1}{K+1}\sum_{k=0}^{K}\langle x^k - u - \gamma F(\widetilde{x}^k), \theta_k\rangle, \qquad (192) \\
\widetilde{x}^K_{avg} &\overset{def}{=} \frac{1}{K+1}\sum_{k=0}^{K}\widetilde{x}^k, \qquad (193) \\
\theta_k &\overset{def}{=} F(\widetilde{x}^k) - \widetilde{F}_{\xi_2^k}(\widetilde{x}^k), \qquad (194) \\
\omega_k &\overset{def}{=} F(x^k) - \widetilde{F}_{\xi_1^k}(x^k). \qquad (195)
\end{aligned}$$

Using this lemma we prove the main convergence result for clipped-SEG in the monotone case.

Theorem G.2 (Case 1 in Theorem 4.1)**.** *Let Assumptions 1.1, 1.7, 1.8 hold for $Q = B_{4R}(x^*)$, where $R \geq \|x^0 - x^*\|$, and*

$$\begin{aligned}
0 < \gamma &\leq \min\left\{\frac{1}{160L\ln\frac{6(K+1)}{\beta}}, \frac{20^{\frac{2-\alpha}{\alpha}}R}{10800^{\frac{1}{\alpha}}(K+1)^{\frac{1}{\alpha}}\sigma\ln^{\frac{\alpha-1}{\alpha}}\frac{6(K+1)}{\beta}}\right\}, \qquad (196) \\
\lambda_k \equiv \lambda &= \frac{R}{20\gamma\ln\frac{6(K+1)}{\beta}}, \qquad (197)
\end{aligned}$$

for some $K \geq 0$ and $\beta \in (0,1]$ such that $\ln\frac{6(K+1)}{\beta} \geq 1$. Then, after K iterations the iterates produced by clipped-SEG with probability at least $1 - \beta$ satisfy

$$\texttt{Gap}_R(\widetilde{x}^K_{avg}) \leq \frac{9R^2}{2\gamma(K+1)} \quad \text{and} \quad \{x^k\}_{k=0}^{K+1} \subseteq B_{3R}(x^*), \{\widetilde{x}^k\}_{k=0}^{K+1} \subseteq B_{4R}(x^*), \qquad (198)$$

where $\widetilde{x}_{avg}^K$ *is defined in* (193). *In particular, when* γ *equals the minimum from* (196)*, then the iterates produced by* clipped-SEG *after* K *iterations with probability at least* $1-\beta$ *satisfy*

$$\texttt{Gap}_R(\widetilde{x}_{avg}^K) = \mathcal{O}\left(\max\left\{\frac{LR^2\ln\frac{K}{\beta}}{K}, \frac{\sigma R \ln^{\frac{\alpha-1}{\alpha}}\frac{K}{\beta}}{K^{\frac{\alpha-1}{\alpha}}}\right\}\right), \tag{199}$$

meaning that to achieve $\texttt{Gap}_R(\widetilde{x}_{avg}^K) \leq \varepsilon$ *with probability at least* $1-\beta$ clipped-SEG *requires*

$$K = \mathcal{O}\left(\frac{LR^2}{\varepsilon}\ln\frac{LR^2}{\varepsilon\beta}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\ln\frac{\sigma R}{\varepsilon\beta}\right) \quad \textit{iterations/oracle calls.} \tag{200}$$

Proof. The proof follows similar steps as the proof of Theorem C.1 from (Gorbunov et al., 2022a). The key difference is related to the application of Bernstein inequality and estimating biases and variances of stochastic terms.

Let $R_k = \|x^k - x^*\|$ for all $k \geq 0$. As in the previous results, the proof is based on the induction argument and showing that the iterates do not leave some ball around the solution with high probability. More precisely, for each $k = 0, 1, \ldots, K+1$ we consider probability event E_k as follows: inequalities

$$\underbrace{\max_{u\in B_R(x^*)}\left\{\|x^0-u\|^2 + 2\gamma\sum_{l=0}^{t-1}\langle x^l - u - \gamma F(\widetilde{x}^l), \theta_l\rangle + \gamma^2\sum_{l=0}^{t-1}\left(\|\theta_l\|^2 + 2\|\omega_l\|^2\right)\right\}}_{A_t} \leq 9R^2, \tag{201}$$

$$\left\|\gamma\sum_{l=0}^{t-1}\theta_l\right\| \leq R \tag{202}$$

hold for $t = 0, 1, \ldots, k$ simultaneously. We want to prove $\mathbb{P}\{E_k\} \geq 1 - k\beta/(K+1)$ for all $k = 0, 1, \ldots, K+1$ by induction. The base of the induction is trivial: for $k = 0$ we have $\|x^0 - u\|^2 \leq 2\|x^0 - x^*\|^2 + 2\|x^* - u\|^2 \leq 4R^2 \leq 9R^2$ and $\|\gamma\sum_{l=0}^{k-1}\theta_l\| = 0$ for any $u \in B_R(x^*)$. Next, assume that for $k = T-1 \leq K$ the statement holds: $\mathbb{P}\{E_{T-1}\} \geq 1 - (T-1)\beta/(K+1)$. Given this, we need to prove $\mathbb{P}\{E_T\} \geq 1 - T\beta/(K+1)$. We start with showing that E_{T-1} implies $R_t \leq 3R$ for all $t = 0, 1, \ldots, T$ (also by induction). For $t = 0$ this is already shown. Now, assume that $R_t \leq 3R$ for all $t = 0, 1, \ldots, t'$ for some $t' < T$. Then for $t = 0, 1, \ldots, t'$

$$\begin{aligned}
\|\widetilde{x}^t - x^*\| &= \|x^t - x^* - \gamma\widetilde{F}_{\xi_1^t}(x^t)\| \leq \|x^t - x^*\| + \gamma\|\widetilde{F}_{\xi_1^t}(x^t)\| \\
&\leq \|x^t - x^*\| + \gamma\lambda \overset{(197)}{\leq} 3R + \frac{R}{20\ln\frac{6(K+1)}{\beta}} \leq 4R.
\end{aligned} \tag{203}$$

Therefore, the conditions of Lemma G.1 are satisfied and we have that E_{T-1} implies

$$\begin{aligned}
\max_{u\in B_R(x^*)}&\left\{2\gamma(t'+1)\langle F(u), \widetilde{x}_{\text{avg}}^{t'} - u\rangle + \|x^{t'+1} - u\|^2\right\} \\
&\leq \max_{u\in B_R(x^*)}\left\{\|x^0-u\|^2 + 2\gamma\sum_{l=0}^{t'}\langle x^l - u - \gamma F(\widetilde{x}^l), \theta_l\rangle\right. \\
&\qquad\left. + \gamma^2\sum_{l=0}^{t'}\left(\|\theta_l\|^2 + 2\|\omega_l\|^2\right)\right\} \\
&\overset{(201)}{\leq} 9R^2,
\end{aligned}$$

meaning that

$$\|x^{t'+1} - x^*\|^2 \leq \max_{u\in B_R(x^*)}\left\{2\gamma(t'+1)\langle F(u), \widetilde{x}_{\text{avg}}^{t'} - u\rangle + \|x^{t'+1} - u\|^2\right\} \leq 9R^2,$$

i.e., $R_{t'+1} \leq 3R$. In other words, we derived that probability event E_{T-1} implies $R_t \leq 3R$ and

$$\max_{u\in B_R(x^*)}\left\{2\gamma(t+1)\langle F(u), \widetilde{x}_{\text{avg}}^t - u\rangle + \|x^{t+1} - u\|^2\right\} \leq 9R^2 \tag{204}$$

for all $t = 0, 1, \ldots, T$. In addition, due to (203) E_{T-1} also implies that $\|\widetilde{x}^t - x^*\| \le 4R$ for all $t = 0, 1, \ldots, T$. Thus, E_{T-1} implies

$$\begin{aligned} \|x^t - x^* - \gamma F(\widetilde{x}^t)\| &\le \|x^t - x^*\| + \gamma\|F(\widetilde{x}^t)\| \overset{(11)}{\le} 3R + \gamma L\|\widetilde{x}^t - x^*\| \\ &\overset{(203)}{\le} 3R + 4R\gamma L \overset{(196)}{\le} 5R, \end{aligned} \tag{205}$$

for all $t = 0, 1, \ldots, T$. Next, we introduce random vectors

$$\eta_t = \begin{cases} x^t - x^* - \gamma F(\widetilde{x}^t), & \text{if } \|x^t - x^* - \gamma F(\widetilde{x}^t)\| \le 5R, \\ 0, & \text{otherwise,} \end{cases}$$

for all $t = 0, 1, \ldots, T$. These vectors are bounded almost surely:

$$\|\eta_t\| \le 5R \tag{206}$$

for all $t = 0, 1, \ldots, T$. Moreover, due to (205), probability event E_{T-1} implies $\eta_t = x^t - x^* - \gamma F(\widetilde{x}^t)$ for all $t = 0, 1, \ldots, T$ and

$$\begin{aligned} A_T &= \max_{u \in B_R(x^*)} \left\{ \|x^0 - u\|^2 + 2\gamma \sum_{l=0}^{T-1} \langle x^* - u, \theta_l\rangle \right\} + 2\gamma \sum_{l=0}^{T-1} \langle x^l - x^* - \gamma F(\widetilde{x}^l), \theta_l\rangle + \gamma^2 \sum_{l=0}^{T-1} \left(\|\theta_l\|^2 + 2\|\omega_l\|^2\right) \\ &\le 4R^2 + 2\gamma \max_{u \in B_R(x^*)} \left\{ \left\langle x^* - u, \sum_{l=0}^{T-1} \theta_l \right\rangle \right\} + 2\gamma \sum_{l=0}^{T-1} \langle \eta_l, \theta_l\rangle + \gamma^2 \sum_{l=0}^{T-1} \left(\|\theta_l\|^2 + 2\|\omega_l\|^2\right) \\ &= 4R^2 + 2\gamma R \left\| \sum_{l=0}^{T-1} \theta_l \right\| + 2\gamma \sum_{l=0}^{T-1} \langle \eta_l, \theta_l\rangle + \gamma^2 \sum_{l=0}^{T-1} \left(\|\theta_l\|^2 + 2\|\omega_l\|^2\right), \end{aligned}$$

where A_T is defined in (201).

To handle the sums appeared in the right-hand side of the previous inequality we consider unbiased and biased parts of θ_l, ω_l:

$$\theta_l^u \overset{\text{def}}{=} \mathbb{E}_{\xi_2^l}\left[\widetilde{F}_{\xi_2^l}(\widetilde{x}^l)\right] - \widetilde{F}_{\xi_2^l}(\widetilde{x}^l), \quad \theta_l^b \overset{\text{def}}{=} F(\widetilde{x}^l) - \mathbb{E}_{\xi_2^l}\left[\widetilde{F}_{\xi_2^l}(\widetilde{x}^l)\right], \tag{207}$$

$$\omega_l^u \overset{\text{def}}{=} \mathbb{E}_{\xi_1^l}\left[\widetilde{F}_{\xi_1^l}(x^l)\right] - \widetilde{F}_{\xi_1^l}(x^l), \quad \omega_l^b \overset{\text{def}}{=} F(x^l) - \mathbb{E}_{\xi_1^l}\left[\widetilde{F}_{\xi_1^l}(x^l)\right], \tag{208}$$

for all $l = 0, \ldots, T-1$. By definition we have $\theta_l = \theta_l^u + \theta_l^b$, $\omega_l = \omega_l^u + \omega_l^b$ for all $l = 0, \ldots, T-1$. Therefore, E_{T-1} implies

$$\begin{aligned} A_T \le\ & 4R^2 + 2\gamma R \left\| \sum_{l=0}^{T-1} \theta_l \right\| + \underbrace{2\gamma \sum_{l=0}^{T-1} \langle \eta_l, \theta_l^u\rangle}_{①} + \underbrace{2\gamma \sum_{l=0}^{T-1} \langle \eta_l, \theta_l^b\rangle}_{②} \\ & + \underbrace{2\gamma^2 \sum_{l=0}^{T-1} \left(\mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] + 2\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right]\right)}_{③} \\ & + \underbrace{2\gamma^2 \sum_{l=0}^{T-1} \left(\|\theta_l^u\|^2 + 2\|\omega_l^u\|^2 - \mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] - 2\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right]\right)}_{④} \\ & + \underbrace{2\gamma^2 \sum_{l=0}^{T-1} \left(\|\theta_l^b\|^2 + 2\|\omega_l^b\|^2\right)}_{⑤}, \end{aligned} \tag{209}$$

where we also apply inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ holding for all $a, b \in \mathbb{R}^d$ to upper bound $\|\theta_l\|^2$ and $\|\omega_l\|^2$. It remains to derive good enough high-probability upper-bounds for the terms $2\gamma R \left\|\sum_{l=0}^{T-1} \theta_l\right\|$, ①, ②, ③, ④, ⑤, i.e., to finish our inductive proof we need to show that $2\gamma R \left\|\sum_{l=0}^{T-1} \theta_l\right\|$ + ① + ② + ③ + ④ + ⑤ $\leq 5R^2$ with high probability. In the subsequent parts of the proof, we will need use many times the bounds for the norm and second moments of θ_{t+1}^u and θ_{t+1}^b. First, by definition of clipping operator we have with probability 1 that

$$\|\theta_l^u\| \leq 2\lambda, \quad \|\omega_l^u\| \leq 2\lambda. \tag{210}$$

Moreover, since E_{T-1} implies that

$$\|F(x^l)\| \overset{(11)}{\leq} L\|x^l - x^*\| \leq 3LR \overset{(196)}{\leq} \frac{R}{40\gamma \ln \frac{6(K+1)}{\beta}} \overset{(197)}{=} \frac{\lambda}{2},$$
$$\|F(\widetilde{x}^l)\| \overset{(11)}{\leq} L\|\widetilde{x}^l - x^*\| \overset{(203)}{\leq} 4LR \overset{(196)}{\leq} \frac{R}{40\gamma \ln \frac{6(K+1)}{\beta}} \overset{(197)}{=} \frac{\lambda}{2}$$

for $t = 0, 1, \ldots, T-1$. Then, in view of Lemma 5.1, we have that E_{T-1} implies

$$\left\|\theta_l^b\right\| \leq \frac{2^\alpha \sigma^\alpha}{\lambda^{\alpha-1}}, \quad \left\|\omega_l^b\right\| \leq \frac{2^\alpha \sigma^\alpha}{\lambda^{\alpha-1}}, \tag{211}$$
$$\mathbb{E}_{\xi_2^l}\left[\|\theta_l\|^2\right] \leq 18\lambda^{2-\alpha}\sigma^\alpha, \quad \mathbb{E}_{\xi_1^l}\left[\|\omega_l\|^2\right] \leq 18\lambda^{2-\alpha}\sigma^\alpha, \tag{212}$$
$$\mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] \leq 18\lambda^{2-\alpha}\sigma^\alpha, \quad \mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right] \leq 18\lambda^{2-\alpha}\sigma^\alpha, \tag{213}$$

for all $l = 0, 1, \ldots, T-1$.

Upper bound for ①. By definition of θ_l^u, we have $\mathbb{E}_{\xi_2^l}[\theta_l^u] = 0$ and

$$\mathbb{E}_{\xi_2^l}\left[2\gamma\langle \eta_l, \theta_l^u\rangle\right] = 0.$$

Next, sum ① has bounded with probability 1 terms:

$$|2\gamma\langle \eta_l, \theta_l^u\rangle| \leq 2\gamma\|\eta_l\| \cdot \|\theta_l^u\| \overset{(206),(210)}{\leq} 20\gamma R\lambda \overset{(197)}{=} \frac{R^2}{\ln \frac{6(K+1)}{\beta}} \overset{\text{def}}{=} c. \tag{214}$$

The summands also have bounded conditional variances $\sigma_l^2 \overset{\text{def}}{=} \mathbb{E}_{\xi_2^l}\left[4\gamma^2\langle \eta_l, \theta_l^u\rangle^2\right]$:

$$\sigma_l^2 \leq \mathbb{E}_{\xi_2^l}\left[4\gamma^2\|\eta_l\|^2 \cdot \|\theta_l^u\|^2\right] \overset{(206)}{\leq} 100\gamma^2 R^2 \mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right]. \tag{215}$$

In other words, we showed that $\{2\gamma\langle \eta_l, \theta_l^u\rangle\}_{l=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\sigma_l^2\}_{l=0}^{T-1}$. Next, we apply Bernstein's inequality (Lemma B.2) with $X_l = 2\gamma\langle \eta_l, \theta_l^u\rangle$, parameter c as in (214), $b = R^2$, $G = \frac{R^4}{6\ln \frac{6(K+1)}{\beta}}$:

$$\mathbb{P}\left\{|①| > R^2 \text{ and } \sum_{l=0}^{T-1} \sigma_l^2 \leq \frac{R^4}{6\ln \frac{6(K+1)}{\beta}}\right\} \leq 2\exp\left(-\frac{b^2}{2G + 2cb/3}\right) = \frac{\beta}{3(K+1)}.$$

Equivalently, we have

$$\mathbb{P}\{E_①\} \geq 1 - \frac{\beta}{3(K+1)}, \quad \text{for} \quad E_① = \left\{\text{either} \quad \sum_{l=0}^{T-1} \sigma_l^2 > \frac{R^4}{6\ln \frac{6(K+1)}{\beta}} \quad \text{or} \quad |①| \leq R^2\right\}. \tag{216}$$

In addition, E_{T-1} implies that

$$\sum_{l=0}^{T-1}\sigma_l^2 \overset{(215)}{\leq} 100\gamma^2R^2\sum_{l=0}^{T-1}\mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] \overset{(213),T\leq K+1}{\leq} 1800(K+1)\gamma^2R^2\lambda^{2-\alpha}\sigma^\alpha$$
$$\overset{(197)}{\leq} \frac{1800(K+1)\gamma^\alpha\sigma^\alpha R^{4-\alpha}}{20^{2-\alpha}\ln^{2-\alpha}\frac{6(K+1)}{\beta}} \overset{(196)}{\leq} \frac{R^4}{6\ln\frac{6(K+1)}{\beta}}. \tag{217}$$

Upper bound for ②. From E_{T-1} it follows that

$$② \leq 2\gamma\sum_{l=0}^{T-1}\|\eta_l\|\cdot\|\theta_l^b\| \overset{(206),(211),T\leq K+1}{\leq} \frac{10\cdot 2^\alpha(K+1)\gamma R\sigma^\alpha}{\lambda^{\alpha-1}}$$
$$\overset{(197)}{=} \frac{10\cdot 2^\alpha\cdot 20^{\alpha-1}(K+1)\gamma^\alpha\sigma^\alpha\ln^{\alpha-1}\frac{6(K+1)}{\beta}}{R^{\alpha-2}} \overset{(196)}{\leq} R^2. \tag{218}$$

Upper bound for ③. From E_{T-1} it follows that

$$2\gamma^2\sum_{l=0}^{T-1}\mathbb{E}_{\xi_2^l}[\|\theta_l^u\|^2] \overset{(212),T\leq K+1}{\leq} 36\gamma^2(K+1)\lambda^{2-\alpha}\sigma^\alpha \overset{(197)}{=} \frac{36\gamma^\alpha(K+1)\sigma^\alpha}{20^{2-\alpha}\ln^{2-\alpha}\frac{6(K+1)}{\beta}} \overset{(196)}{\leq} \frac{1}{12}R^2, \tag{219}$$

$$4\gamma^2\sum_{l=0}^{T-1}\mathbb{E}_{\xi_1^l}[\|\omega_l^u\|^2] \overset{(212),T\leq K+1}{\leq} 72\gamma^2(K+1)\lambda^{2-\alpha}\sigma^\alpha \overset{(197)}{=} \frac{72\gamma^\alpha(K+1)\sigma^\alpha}{20^{2-\alpha}\ln^{2-\alpha}\frac{6(K+1)}{\beta}} \overset{(196)}{\leq} \frac{1}{12}R^2, \tag{220}$$

$$③ \overset{(219),(220)}{\leq} \frac{1}{6}R^2. \tag{221}$$

Upper bound for ④. By the construction we have

$$2\gamma^2\mathbb{E}_{\xi_1^l,\xi_2^l}\left[\|\theta_l^u\|^2 + 2\|\omega_l^u\|^2 - \mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] - 2\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right]\right] = 0.$$

Next, sum ① has bounded with probability 1 terms:

$$2\gamma^2\left|\|\theta_l^u\|^2 + 2\|\omega_l^u\|^2 - \mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] - 2\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right]\right| \leq 2\gamma^2\|\theta_l^u\|^2 + 2\gamma^2\mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right]$$
$$+4\gamma^2\|\omega_l^u\|^2 + 4\gamma^2\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right]$$
$$\overset{(210)}{\leq} 48\gamma^2\lambda^2$$
$$\overset{(197)}{\leq} \frac{R^2}{6\ln\frac{6(K+1)}{\beta}} \overset{\text{def}}{=} c. \tag{222}$$

The summands also have bounded conditional variances $\widetilde{\sigma}_l^2 \overset{\text{def}}{=} 4\gamma^4\mathbb{E}_{\xi_1^l,\xi_2^l}\left[\left|\|\theta_l^u\|^2 + 2\|\omega_l^u\|^2 - \mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] - 2\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right]\right|^2\right]$:

$$\widetilde{\sigma}_l^2 \overset{(222)}{\leq} \frac{\gamma^2R^2}{3\ln\frac{6(K+1)}{\beta}}\mathbb{E}_{\xi_1^l,\xi_2^l}\left[\left|\|\theta_l^u\|^2 + 2\|\omega_l^u\|^2 - \mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] - 2\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right]\right|\right]$$
$$\leq \frac{2\gamma^2R^2}{3\ln\frac{6(K+1)}{\beta}}\mathbb{E}_{\xi_1^l,\xi_2^l}\left[\|\theta_l^u\|^2 + 2\|\omega_l^u\|^2\right]. \tag{223}$$

In other words, we showed that $\left\{2\gamma^2\left(\|\theta_l^u\|^2 + 2\|\omega_l^u\|^2 - \mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] - 2\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right]\right)\right\}_{l=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\sigma_l^2\}_{l=0}^{T-1}$. Next, we apply Bernstein's inequality (Lemma B.2) with

$X_l = 2\gamma^2\left(\|\theta_l^u\|^2 + 2\|\omega_l^u\|^2 - \mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] - 2\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right]\right)$, parameter c as in (222), $b = \frac{1}{6}R^2$, $G = \frac{R^4}{216\ln\frac{6(K+1)}{\beta}}$:

$$\mathbb{P}\left\{|④| > \frac{1}{6}R^2 \text{ and } \sum_{l=0}^{T-1}\widetilde{\sigma}_l^2 \leq \frac{R^4}{216\ln\frac{6(K+1)}{\beta}}\right\} \leq 2\exp\left(-\frac{b^2}{2G + {}^{2cb}/_{3}}\right) = \frac{\beta}{3(K+1)}.$$

Equivalently, we have

$$\mathbb{P}\{E_④\} \geq 1 - \frac{\beta}{3(K+1)}, \quad \text{for} \quad E_④ = \left\{\text{either} \quad \sum_{l=0}^{T-1}\widetilde{\sigma}_l^2 > \frac{R^4}{216\ln\frac{6(K+1)}{\beta}} \quad \text{or} \quad |④| \leq \frac{1}{6}R^2\right\}. \tag{224}$$

In addition, E_{T-1} implies that

$$\begin{aligned}\sum_{l=0}^{T-1}\widetilde{\sigma}_l^2 &\overset{(223)}{\leq} \frac{2\gamma^2R^2}{3\ln\frac{6(K+1)}{\beta}}\sum_{l=0}^{T-1}\mathbb{E}_{\xi_1^l,\xi_2^l}\left[\|\theta_l^u\|^2 + 2\|\omega_l^u\|^2\right] \\ &\overset{(213),T\leq K+1}{\leq} \frac{36(K+1)\gamma^2R^2\lambda^{2-\alpha}\sigma^\alpha}{\ln\frac{6(K+1)}{\beta}} \\ &\overset{(197)}{\leq} \frac{36(K+1)\gamma^\alpha R^{4-\alpha}\sigma^\alpha}{20^{2-\alpha}\ln^{3-\alpha}\frac{6(K+1)}{\beta}} \overset{(196)}{\leq} \frac{R^4}{216\ln\frac{6(K+1)}{\beta}}.\end{aligned} \tag{225}$$

Upper bound for ⑤. From E_{T-1} it follows that

$$\begin{aligned}⑤ &= 2\gamma^2\sum_{l=0}^{T-1}\left(\|\theta_l^b\|^2 + 2\|\omega_l^b\|^2\right) \overset{(211),T\leq K+1}{\leq} \frac{6\cdot 2^{2\alpha}\gamma^2\sigma^{2\alpha}(K+1)}{\lambda^{2\alpha-2}} \\ &\overset{(197)}{=} \frac{6\cdot 2^{2\alpha}\cdot 20^{2\alpha-2}\gamma^{2\alpha}\sigma^{2\alpha}(K+1)\ln^{2\alpha-2}\frac{6(K+1)}{\beta}}{R^{2\alpha-2}} \overset{(196)}{\leq} \frac{1}{6}R^2.\end{aligned} \tag{226}$$

Upper bound for $2\gamma R\left\|\sum_{l=0}^{T-1}\theta_l\right\|$. To upper-bound this sum, we introduce new random vectors:

$$\zeta_l = \begin{cases}\gamma\sum\limits_{r=0}^{l-1}\theta_r, & \text{if } \left\|\gamma\sum\limits_{r=0}^{l-1}\theta_r\right\| \leq R, \\ 0, & \text{otherwise}\end{cases}$$

for $l = 1, 2, \ldots, T-1$. These vectors are bounded with probability 1:

$$\|\zeta_l\| \leq R. \tag{227}$$

Therefore, taking into account (202), we derive that E_{T-1} implies

$$\begin{aligned}2\gamma R\left\|\sum_{l=0}^{T-1}\theta_l\right\| &= 2R\sqrt{\gamma^2\left\|\sum_{l=0}^{T-1}\theta_l\right\|^2} \\ &= 2R\sqrt{\gamma^2\sum_{l=0}^{T-1}\|\theta_l\|^2 + 2\gamma\sum_{l=0}^{T-1}\left\langle\gamma\sum_{r=0}^{l-1}\theta_r, \theta_l\right\rangle} \\ &= 2R\sqrt{\gamma^2\sum_{l=0}^{T-1}\|\theta_l\|^2 + 2\gamma\sum_{l=0}^{T-1}\langle\zeta_l, \theta_l\rangle} \\ &\overset{(207)}{\leq} 2R\sqrt{③ + ④ + ⑤ + \underbrace{2\gamma\sum_{l=0}^{T-1}\langle\zeta_l, \theta_l^u\rangle}_{⑥} + \underbrace{2\gamma\sum_{l=0}^{T-1}\langle\zeta_l, \theta_l^b\rangle}_{⑦}}.\end{aligned} \tag{228}$$

Similarly to the previous parts of the proof, we bound ⑥ and ⑦.

Upper bound for ⑥. By definition of θ_l^u, we have $\mathbb{E}_{\xi_2^l}[\theta_l^u] = 0$ and

$$\mathbb{E}_{\xi_2^l}\left[2\gamma\langle\zeta_l, \theta_l^u\rangle\right] = 0.$$

Next, sum ⑥ has bounded with probability 1 terms:

$$|2\gamma\langle\zeta_l, \theta_l^u\rangle| \leq 2\gamma\|\eta_l\| \cdot \|\theta_l^u\| \overset{(227),(210)}{\leq} 4\gamma R\lambda \overset{(197)}{\leq} \frac{R^2}{4\ln\frac{6(K+1)}{\beta}} \overset{\text{def}}{=} c. \tag{229}$$

The summands also have bounded conditional variances $\hat{\sigma}_l^2 \overset{\text{def}}{=} \mathbb{E}_{\xi_2^l}\left[4\gamma^2\langle\zeta_l, \theta_l^u\rangle^2\right]$:

$$\hat{\sigma}_l^2 \leq \mathbb{E}_{\xi_2^l}\left[4\gamma^2\|\zeta_l\|^2 \cdot \|\theta_l^u\|^2\right] \overset{(227)}{\leq} 4\gamma^2R^2\mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right]. \tag{230}$$

In other words, we showed that $\{2\gamma\langle\zeta_l, \theta_l^u\rangle\}_{l=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\hat{\sigma}_l^2\}_{l=0}^{T-1}$. Next, we apply Bernstein's inequality (Lemma B.2) with $X_l = 2\gamma\langle\zeta_l, \theta_l^u\rangle$, parameter c as in (229), $b = \frac{R^2}{4}$, $G = \frac{R^4}{96\ln\frac{6(K+1)}{\beta}}$:

$$\mathbb{P}\left\{|⑤| > \frac{1}{4}R^2 \text{ and } \sum_{l=0}^{T-1}\hat{\sigma}_l^2 \leq \frac{R^4}{96\ln\frac{4(K+1)}{\beta}}\right\} \leq 2\exp\left(-\frac{b^2}{2G + {}^{2cb}/_3}\right) = \frac{\beta}{3(K+1)}.$$

Equivalently, we have

$$E_{⑥} = \left\{\text{either} \quad \sum_{l=0}^{T-1}\hat{\sigma}_l^2 > \frac{R^4}{96\ln\frac{6(K+1)}{\beta}} \quad \text{or} \quad |⑤| \leq \frac{1}{4}R^2\right\}. \tag{231}$$

In addition, E_{T-1} implies that

$$\begin{aligned}\sum_{l=0}^{T-1}\hat{\sigma}_l^2 &\overset{(230)}{\leq} 4\gamma^2R^2\sum_{l=0}^{T-1}\mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] \overset{(213),T\leq K+1}{\leq} 72(K+1)\gamma^2R^2\lambda^{2-\alpha}\sigma^\alpha \\ &\overset{(197)}{=} \frac{72(K+1)\gamma^\alpha R^{4-\alpha}\sigma^\alpha}{20^{2-\alpha}\ln^{2-\alpha}\frac{6(K+1)}{\beta}} \overset{(196)}{\leq} \frac{R^4}{96\ln\frac{6(K+1)}{\beta}}.\end{aligned} \tag{232}$$

Upper bound for ⑦. From E_{T-1} it follows that

$$\begin{aligned}⑦ &\leq 2\gamma\sum_{l=0}^{T-1}\|\zeta_l\| \cdot \|\theta_l^b\| \overset{(227),(211),T\leq K+1}{\leq} \frac{2^{\alpha+1}(K+1)\gamma R\sigma^\alpha}{\lambda^{\alpha-1}} \\ &\overset{(197)}{=} \frac{2^{\alpha+1}\cdot 20^{\alpha-1}(K+1)\gamma^\alpha\sigma^\alpha\ln^{\alpha-1}\frac{6(K+1)}{\beta}}{R^{\alpha-2}} \overset{(196)}{\leq} \frac{1}{4}R^2.\end{aligned} \tag{233}$$

Now, we have the upper bounds for $2\gamma R\left\|\sum_{l=0}^{T-1}\theta_l\right\|$, ①, ②, ③, ④, ⑤. In particular, probability event E_{T-1} implies

$$A_T \overset{(209)}{\leq} 4R^2 + 2\gamma R\left\|\sum_{l=0}^{T-1}\theta_l\right\| + ① + ② + ③ + ④ + ⑤,$$

$$2\gamma R\left\|\sum_{l=0}^{T-1}\theta_l\right\| \overset{(228)}{\leq} 2R\sqrt{③ + ④ + ⑤ + ⑥ + ⑦},$$

$$② \overset{(218)}{\leq} R^2, \quad ③ \overset{(221)}{\leq} \frac{1}{6}R^2, \quad ⑤ \overset{(226)}{\leq} \frac{1}{6}R^2, \quad ⑦ \overset{(233)}{\leq} \frac{1}{4}R^2,$$

$$\sum_{l=0}^{T-1}\sigma_l^2 \overset{(217)}{\leq} \frac{R^4}{6\ln\frac{6(K+1)}{\beta}}, \quad \sum_{l=0}^{T-1}\widetilde{\sigma}_l^2 \overset{(225)}{\leq} \frac{R^4}{216\ln\frac{6(K+1)}{\beta}}, \quad \sum_{l=0}^{T-1}\hat{\sigma}_l^2 \overset{(232)}{\leq} \frac{R^4}{96\ln\frac{6(K+1)}{\beta}}.$$

Moreover, we also have (see (216), (224), (231) and our induction assumption)

$$\mathbb{P}\{E_{T-1}\} \geq 1 - \frac{(T-1)\beta}{K+1},$$
$$\mathbb{P}\{E_{①}\} \geq 1 - \frac{\beta}{3(K+1)}, \quad \mathbb{P}\{E_{④}\} \geq 1 - \frac{\beta}{3(K+1)}, \quad \mathbb{P}\{E_{⑥}\} \geq 1 - \frac{\beta}{3(K+1)},$$

where

$$\begin{aligned} E_{①} &= \left\{\text{either} \quad \sum_{l=0}^{T-1} \sigma_l^2 > \frac{R^4}{6\ln\frac{6(K+1)}{\beta}} \quad \text{or} \quad |①| \leq R^2\right\}, \\ E_{④} &= \left\{\text{either} \quad \sum_{l=0}^{T-1} \widetilde{\sigma}_l^2 > \frac{R^4}{216\ln\frac{6(K+1)}{\beta}} \quad \text{or} \quad |④| \leq \frac{1}{6}R^2\right\}, \\ E_{⑥} &= \left\{\text{either} \quad \sum_{l=0}^{T-1} \hat{\sigma}_l^2 > \frac{R^4}{96\ln\frac{6(K+1)}{\beta}} \quad \text{or} \quad |⑥| \leq \frac{1}{4}R^2\right\}. \end{aligned}$$

Thus, probability event $E_{T-1} \cap E_{①} \cap E_{④} \cap E_{⑥}$ implies

$$\left\|\gamma \sum_{l=0}^{T-1} \theta_l\right\| \leq \sqrt{\frac{1}{6}R^2 + \frac{1}{6}R^2 + \frac{1}{6}R^2 + \frac{1}{4}R^2 + \frac{1}{4}R^2} = R, \tag{234}$$

$$\begin{aligned} A_T &\leq 4R^2 + 2R\sqrt{\frac{1}{6}R^2 + \frac{1}{6}R^2 + \frac{1}{6}R^2 + \frac{1}{4}R^2 + \frac{1}{4}R^2} \\ &\quad + R^2 + R^2 + \frac{1}{6}R^2 + \frac{1}{6}R^2 + \frac{1}{6}R^2 \\ &\leq 9R^2, \end{aligned} \tag{235}$$

which is equivalent to (201) and (202) for $t = T$, and

$$\mathbb{P}\{E_T\} \geq \mathbb{P}\{E_{T-1} \cap E_{①} \cap E_{④} \cap E_{⑥}\} = 1 - \mathbb{P}\{\overline{E}_{T-1} \cup \overline{E}_{①} \cup \overline{E}_{④} \cup \overline{E}_{⑥}\} \geq 1 - \frac{T\beta}{K+1}.$$

This finishes the inductive part of our proof, i.e., for all $k = 0, 1, \ldots, K+1$ we have $\mathbb{P}\{E_k\} \geq 1 - k\beta/(K+1)$. In particular, for $k = K+1$ we have that with probability at least $1-\beta$

$$\begin{aligned} \texttt{Gap}_R(\widetilde{x}_{\text{avg}}^K) &= \max_{u \in B_R(x^*)} \left\{\langle F(u), \widetilde{x}_{\text{avg}}^K - u\rangle\right\} \\ &\leq \frac{1}{2\gamma(K+1)} \max_{u \in B_R(x^*)} \left\{2\gamma(K+1)\langle F(u), \widetilde{x}_{\text{avg}}^t - u\rangle + \|x^{K+1} - u\|^2\right\} \\ &\overset{(204)}{\leq} \frac{9R^2}{2\gamma(K+1)}. \end{aligned}$$

Finally, if

$$\gamma = \min\left\{\frac{1}{160L\ln\frac{6(K+1)}{\beta}}, \frac{20^{\frac{2-\alpha}{\alpha}}R}{10800^{\frac{1}{\alpha}}(K+1)^{\frac{1}{\alpha}}\sigma\ln^{\frac{\alpha-1}{\alpha}}\frac{6(K+1)}{\beta}}\right\}$$

then with probability at least $1-\beta$

$$\begin{aligned} \texttt{Gap}_R(\widetilde{x}_{\text{avg}}^K) &\leq \frac{9R^2}{2\gamma(K+1)} = \max\left\{\frac{720LR^2\ln\frac{6(K+1)}{\beta}}{K+1}, \frac{9\sigma R\ln^{\frac{\alpha-1}{\alpha}}\frac{6(K+1)}{\beta}}{2\cdot 20^{\frac{2-\alpha}{\alpha}}(K+1)^{\frac{\alpha-1}{\alpha}}}\right\} \\ &= \mathcal{O}\left(\max\left\{\frac{LR^2\ln\frac{K}{\beta}}{K}, \frac{\sigma R\ln^{\frac{\alpha-1}{\alpha}}\frac{K}{\beta}}{K^{\frac{\alpha-1}{\alpha}}}\right\}\right). \end{aligned}$$

To get $\texttt{Gap}_R(\widetilde{x}^K_{\texttt{avg}}) \le \varepsilon$ with probability at least $1-\beta$ it is sufficient to choose K such that both terms in the maximum above are $\mathcal{O}(\varepsilon)$. This leads to

$$K = \mathcal{O}\left(\frac{LR^2}{\varepsilon}\ln\frac{LR^2}{\varepsilon\beta}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\ln\frac{\sigma R}{\varepsilon\beta}\right)$$

that concludes the proof. □

G.2. Quasi-Strongly Monotone Problems

As in the monotone case, we use another lemma from (Gorbunov et al., 2022a) that handles the deterministic part of clipped-SEG in the quasi-strongly monotone case.

Lemma G.3 (Lemma C.3 from (Gorbunov et al., 2022a)). *Let Assumptions 1.7, 1.9 hold for $Q = B_{3R}(x^*) = \{x \in \mathbb{R}^d \mid \|x - x^*\| \le 3R\}$, where $R \ge \|x^0 - x^*\|$, and $0 < \gamma \le 1/2(L+2\mu)$. If x^k and $\widetilde{x}^k$ lie in $B_{3R}(x^*)$ for all $k = 0, 1, \ldots, K$ for some $K \ge 0$, then the iterates produced by* clipped-SEG *satisfy*

$$\begin{aligned}\|x^{K+1} - x^*\|^2 &\le (1-\gamma\mu)^{K+1}\|x^0 - x^*\|^2 - 4\gamma^3\mu\sum_{k=0}^{K}(1-\gamma\mu)^{K-k}\langle F(x^k), \omega_k\rangle \\ &\quad +2\gamma\sum_{k=0}^{K}(1-\gamma\mu)^{K-k}\langle x^k - x^* - \gamma F(\widetilde{x}^k), \theta_k\rangle \\ &\quad +\gamma^2\sum_{k=0}^{K}(1-\gamma\mu)^{K-k}\left(\|\theta_k\|^2 + 4\|\omega_k\|^2\right), \end{aligned} \tag{236}$$

where θ_k, ω_k are defined in (194), (195).

Using this lemma we prove the main convergence result for clipped-SEG in the quasi-strongly monotone case.

Theorem G.4 (Case 2 in Theorem 4.1). *Let Assumptions 1.1, 1.7, 1.9, hold for $Q = B_{3R}(x^*) = \{x \in \mathbb{R}^d \mid \|x - x^*\| \le 3R\}$, where $R \ge \|x^0 - x^*\|$, and*

$$0 < \gamma \le \min\left\{\frac{1}{650L\ln\frac{6(K+1)}{\beta}}, \frac{\ln(B_K)}{\mu(K+1)}\right\}, \tag{237}$$

$$B_K = \max\left\{2, \frac{(K+1)^{\frac{2(\alpha-1)}{\alpha}}\mu^2R^2}{264600^{\frac{2}{\alpha}}\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{6(K+1)}{\beta}\right)\ln^2(B_K)}\right\} \tag{238}$$

$$= \mathcal{O}\left(\max\left\{2, \frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2R^2}{\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)\ln^2\left(\max\left\{2, \frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2R^2}{\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)}\right\}\right)}\right\}\right), \tag{239}$$

$$\lambda_k = \frac{\exp(-\gamma\mu(1+k/2))R}{120\gamma\ln\frac{6(K+1)}{\beta}}, \tag{240}$$

for some $K \ge 0$ and $\beta \in (0,1]$ such that $\ln\frac{6(K+1)}{\beta} \ge 1$. Then, after K iterations the iterates produced by clipped-SEG *with probability at least $1-\beta$ satisfy*

$$\|x^{K+1} - x^*\|^2 \le 2\exp(-\gamma\mu(K+1))R^2. \tag{241}$$

In particular, when γ equals the minimum from (237), then the iterates produced by clipped-SEG *after K iterations with probability at least $1-\beta$ satisfy*

$$\|x^K - x^*\|^2 = \mathcal{O}\left(\max\left\{R^2\exp\left(-\frac{\mu K}{L\ln\frac{K}{\beta}}\right), \frac{\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)\ln^2\left(\max\left\{2, \frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2R^2}{\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)}\right\}\right)}{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2}\right\}\right), \tag{242}$$

meaning that to achieve $\|x^K - x^*\|^2 \leq \varepsilon$ *with probability at least* $1-\beta$ clipped-SEG *requires*

$$K = \mathcal{O}\left(\frac{L}{\mu}\ln\left(\frac{R^2}{\varepsilon}\right)\ln\left(\frac{L}{\mu\beta}\ln\frac{R^2}{\varepsilon}\right), \left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\ln\left(\frac{1}{\beta}\left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right)\ln^{\frac{\alpha}{\alpha-1}}(B_\varepsilon)\right) \tag{243}$$

iterations/oracle calls, where

$$B_\varepsilon = \max\left\{2, \frac{R^2}{\varepsilon\ln\left(\frac{1}{\beta}\left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right)}\right\}.$$

Proof. Again, we will closely follow the proof of Theorem C.3 from (Gorbunov et al., 2022a) and the main difference will be reflected in the application of Bernstein inequality and estimating biases and variances of stochastic terms.

Let $R_k = \|x^k - x^*\|$ for all $k \geq 0$. As in the previous results, the proof is based on the induction argument and showing that the iterates do not leave some ball around the solution with high probability. More precisely, for each $k = 0, 1, \ldots, K+1$ we consider probability event E_k as follows: inequalities

$$R_t^2 \leq 2\exp(-\gamma\mu t)R^2 \tag{244}$$

hold for $t = 0, 1, \ldots, k$ simultaneously. We want to prove $\mathbb{P}\{E_k\} \geq 1 - {k\beta}/{(K+1)}$ for all $k = 0, 1, \ldots, K+1$ by induction. The base of the induction is trivial: for $k=0$ we have $R_0^2 \leq R^2 < 2R^2$ by definition. Next, assume that for $k = T-1 \leq K$ the statement holds: $\mathbb{P}\{E_{T-1}\} \geq 1 - {(T-1)\beta}/{(K+1)}$. Given this, we need to prove $\mathbb{P}\{E_T\} \geq 1 - {T\beta}/{(K+1)}$. Since $R_t^2 \leq 2\exp(-\gamma\mu t)R^2 \leq 9R^2$, we have $x^t \in B_{3R}(x^*)$, where operator F is L-Lipschitz. Thus, E_{T-1} implies

$$\|F(x^t)\| \leq L\|x^t - x^*\| \overset{(244)}{\leq} \sqrt{2}L\exp(-{\gamma\mu t}/{2})R \overset{(237),(240)}{\leq} \frac{\lambda_t}{2} \tag{245}$$

and

$$\|\omega_t\|^2 \leq 2\|\widetilde{F}_{\xi_1}(x^t)\|^2 + 2\|F(x^t)\|^2 \overset{(245)}{\leq} \frac{5}{2}\lambda_t^2 \overset{(240)}{\leq} \frac{\exp(-\gamma\mu t)R^2}{4\gamma^2} \tag{246}$$

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

Next, we need to prove that E_{T-1} implies $\|\widetilde{x}^t - x^*\| \leq 3R$ and show several useful inequalities related to θ_t. Lipschitzness of F probability event E_{T-1} implies

$$\begin{aligned}
\|\widetilde{x}^t - x^*\|^2 &= \|x^t - x^* - \gamma\widetilde{F}_{\xi_1}(x^t)\|^2 \leq 2\|x^t - x^*\|^2 + 2\gamma^2\|\widetilde{F}_{\xi_1}(x^t)\|^2 \\
&\leq 2R_t^2 + 4\gamma^2\|F(x^t)\|^2 + 4\gamma^2\|\omega_t\|^2 \\
&\overset{(11)}{\leq} 2(1+2\gamma^2L^2)R_t^2 + 4\gamma^2\|\omega_t\|^2 \\
&\overset{(237),(246)}{\leq} 7\exp(-\gamma\mu t)R^2 \leq 9R^2
\end{aligned} \tag{247}$$

and

$$\|F(\widetilde{x}^t)\| \leq L\|\widetilde{x}^t - x^*\| \leq \sqrt{7}L\exp(-{\gamma\mu t}/{2})R \overset{(237),(240)}{\leq} \frac{\lambda_t}{2} \tag{248}$$

for all $t = 0, 1, \ldots, T-1$. Therefore, E_{T-1} implies that $x^t, \widetilde{x}^t \in B_{3R}(x^*)$ for all $t = 0, 1, \ldots, T-1$. Using Lemma G.3 and $(1-\gamma\mu)^T \leq \exp(-\gamma\mu T)$, we obtain that E_{T-1} implies

$$\begin{aligned}
R_T^2 \leq\ & \exp(-\gamma\mu T)R^2 - 4\gamma^3\mu\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\langle F(x^l), \omega_l\rangle \\
&+2\gamma\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\langle x^l - x^* - \gamma F(\widetilde{x}^l), \theta_l\rangle \\
&+\gamma^2\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\left(\|\theta_l\|^2 + 4\|\omega_l\|^2\right).
\end{aligned}$$

To handle the sums above, we introduce a new notation:

$$\zeta_t = \begin{cases} F(x^t), & \text{if } \|F(x^t)\| \leq \sqrt{2}L\exp(-\gamma\mu t/2)R, \\ 0, & \text{otherwise,} \end{cases} \tag{249}$$

$$\eta_t = \begin{cases} x^t - x^* - \gamma F(\widetilde{x}^t), & \text{if } \|x^t - x^* - \gamma F(\widetilde{x}^t)\| \leq \sqrt{7}(1+\gamma L)\exp(-\gamma\mu t/2)R, \\ 0, & \text{otherwise,} \end{cases} \tag{250}$$

for $t = 0, 1, \ldots, T-1$. These vectors are bounded almost surely:

$$\|\zeta_t\| \leq \sqrt{2}L\exp(-\gamma\mu t/2)R, \quad \|\eta_t\| \leq \sqrt{7}(1+\gamma L)\exp(-\gamma\mu t/2)R \tag{251}$$

for all $t = 0, 1, \ldots, T-1$. We also notice that E_{T-1} implies $\|F(x^t)\| \leq \sqrt{2}L\exp(-\gamma\mu t/2)R$ (due to (245)) and

$$\begin{aligned} \|x^t - x^* - \gamma F(\widetilde{x}^t)\| &\leq \|x^t - x^*\| + \gamma\|F(\widetilde{x}^t)\| \\ &\overset{(247),(248)}{\leq} \sqrt{7}(1+\gamma L)\exp(-\gamma\mu t/2)R \end{aligned}$$

for $t = 0, 1, \ldots, T-1$. In other words, E_{T-1} implies $\zeta_t = F(x^t)$ and $\eta_t = x^t - x^* - \gamma F(\widetilde{x}^t)$ for all $t = 0, 1, \ldots, T-1$, meaning that from E_{T-1} it follows that

$$\begin{aligned} R_T^2 \leq\; & \exp(-\gamma\mu T)R^2 - 4\gamma^3\mu\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\langle\zeta_l, \omega_l\rangle \\ & + 2\gamma\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\langle\eta_l, \theta_l\rangle + \gamma^2\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\left(\|\theta_l\|^2 + 4\|\omega_l\|^2\right). \end{aligned}$$

To handle the sums appeared on the right-hand side of the previous inequality we consider unbiased and biased parts of θ_l, ω_l:

$$\theta_l^u \overset{\text{def}}{=} \mathbb{E}_{\xi_2^l}\left[\widetilde{F}_{\xi_2^l}(\widetilde{x}^l)\right] - \widetilde{F}_{\xi_2^l}(\widetilde{x}^l), \quad \theta_l^b \overset{\text{def}}{=} F(\widetilde{x}^l) - \mathbb{E}_{\xi_2^l}\left[\widetilde{F}_{\xi_2^l}(\widetilde{x}^l)\right], \tag{252}$$

$$\omega_l^u \overset{\text{def}}{=} \mathbb{E}_{\xi_1^l}\left[\widetilde{F}_{\xi_1^l}(x^l)\right] - \widetilde{F}_{\xi_1^l}(x^l), \quad \omega_l^b \overset{\text{def}}{=} F(x^l) - \mathbb{E}_{\xi_1^l}\left[\widetilde{F}_{\xi_1^l}(x^l)\right], \tag{253}$$

for all $l = 0, \ldots, T-1$. By definition we have $\theta_l = \theta_l^u + \theta_l^b$, $\omega_l = \omega_l^u + \omega_l^b$ for all $l = 0, \ldots, T-1$. Therefore, E_{T-1} implies

$$\begin{aligned} R_T^2 \leq\; & \exp(-\gamma\mu T)R^2 \underbrace{-4\gamma^3\mu\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\langle\zeta_l, \omega_l^u\rangle}_{①} \underbrace{-4\gamma^3\mu\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\langle\zeta_l, \omega_l^b\rangle}_{②} \\ & \underbrace{+2\gamma\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\langle\eta_l, \theta_l^u\rangle}_{③} \underbrace{+2\gamma\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\langle\eta_l, \theta_l^b\rangle}_{④} \\ & \underbrace{+2\gamma^2\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\left(\mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] + 4\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right]\right)}_{⑤} \\ & \underbrace{+2\gamma^2\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\left(\|\theta_l^u\|^2 + 4\|\omega_l^u\|^2 - \mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] - 4\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right]\right)}_{⑥} \\ & \underbrace{+2\gamma^2\sum_{l=0}^{T-1}(1-\gamma\mu)^{T-1-l}\left(\|\theta_l^b\|^2 + 4\|\omega_l^b\|^2\right)}_{⑦}. \end{aligned} \tag{254}$$

where we also apply inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ holding for all $a, b \in \mathbb{R}^d$ to upper bound $\|\theta_l\|^2$ and $\|\omega_l\|^2$. It remains to derive good enough high-probability upper-bounds for the terms ①, ②, ③, ④, ⑤, ⑥, ⑦, i.e., to finish our inductive proof we need to show that ① + ② + ③ + ④ + ⑤ + ⑥ + ⑦ $\leq \exp(-\gamma\mu T)R^2$ with high probability. In the subsequent parts of the proof, we will need to use many times the bounds for the norm and second moments of θ_{t+1}^u and θ_{t+1}^b. First, by definition of clipping operator, we have with probability 1 that

$$\|\theta_l^u\| \leq 2\lambda_l, \quad \|\omega_l^u\| \leq 2\lambda_l. \tag{255}$$

Moreover, since E_{T-1} implies that $\|F(x^l)\| \leq \lambda_l/2$ and $\|F(\widetilde{x}^l)\| \leq \lambda_l/2$ for all $l = 0, 1, \ldots, T-1$ (see (245) and (248)), from Lemma 5.1 we also have that E_{T-1} implies

$$\|\theta_l^b\| \leq \frac{2^\alpha\sigma^\alpha}{\lambda_l^{\alpha-1}}, \quad \|\omega_l^b\| \leq \frac{2^\alpha\sigma^\alpha}{\lambda_l^{\alpha-1}}, \tag{256}$$

$$\mathbb{E}_{\xi_2^l}\left[\|\theta_l\|^2\right] \leq 18\lambda_l^{2-\alpha}\sigma^\alpha, \quad \mathbb{E}_{\xi_1^l}\left[\|\omega_l\|^2\right] \leq 18\lambda_l^{2-\alpha}\sigma^\alpha, \tag{257}$$

$$\mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] \leq 18\lambda_l^{2-\alpha}\sigma^\alpha, \quad \mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right] \leq 18\lambda_l^{2-\alpha}\sigma^\alpha, \tag{258}$$

for all $l = 0, 1, \ldots, T-1$.

Upper bound for ①. By definition of ω_l^u, we have $\mathbb{E}_{\xi_1^l}[\omega_l^u] = 0$ and

$$\mathbb{E}_{\xi_1^l}\left[-4\gamma^3\mu(1-\gamma\mu)^{T-1-l}\langle\zeta_l, \omega_l^u\rangle\right] = 0.$$

Next, sum ① has bounded with probability 1 terms:

$$\begin{aligned}
|-4\gamma^3\mu(1-\gamma\mu)^{T-1-l}\langle\zeta_l, \omega_l^u\rangle| &\leq 4\gamma^3\mu\exp(-\gamma\mu(T-1-l))\|\zeta_l\|\cdot\|\omega_l^u\| \\
&\overset{(251),(255)}{\leq} 8\sqrt{2}\gamma^3\mu L\exp(-\gamma\mu(T-1-l/2))R\lambda_l \\
&\overset{(237),(240)}{\leq} \frac{\exp(-\gamma\mu T)R^2}{7\ln\frac{6(K+1)}{\beta}} \overset{\text{def}}{=} c.
\end{aligned} \tag{259}$$

The summands also have bounded conditional variances $\sigma_l^2 \overset{\text{def}}{=} \mathbb{E}_{\xi_1^l}\left[16\gamma^6\mu^2(1-\gamma\mu)^{2T-2-2l}\langle\zeta_l, \omega_l^u\rangle^2\right]$:

$$\begin{aligned}
\sigma_l^2 &\leq \mathbb{E}_{\xi_1^l}\left[16\gamma^6\mu^2\exp(-\gamma\mu(2T-2-2l))\|\zeta_l\|^2\cdot\|\omega_l^u\|^2\right] \\
&\overset{(251)}{\leq} 36\gamma^6\mu^2L^2\exp(-\gamma\mu(2T-2-l))R^2\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right] \\
&\overset{(237)}{\leq} \frac{4\gamma^2\exp(-\gamma\mu(2T-l))R^2}{2809\ln\frac{6(K+1)}{\beta}}\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right].
\end{aligned} \tag{260}$$

In other words, we showed that $\{-4\gamma^3\mu(1-\gamma\mu)^{T-1-l}\langle\zeta_l, \omega_l^u\rangle\}_{l=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\sigma_l^2\}_{l=0}^{T-1}$. Next, we apply Bernstein's inequality (Lemma B.2) with $X_l = -4\gamma^3\mu(1-\gamma\mu)^{T-1-l}\langle\zeta_l, \omega_l^u\rangle$, parameter c as in (259), $b = \frac{1}{7}\exp(-\gamma\mu T)R^2$, $G = \frac{\exp(-2\gamma\mu T)R^4}{294\ln\frac{6(K+1)}{\beta}}$:

$$\mathbb{P}\left\{|①| > \frac{1}{7}\exp(-\gamma\mu T)R^2 \text{ and } \sum_{l=0}^{T-1}\sigma_l^2 \leq \frac{\exp(-2\gamma\mu T)R^4}{294\ln\frac{6(K+1)}{\beta}}\right\} \leq 2\exp\left(-\frac{b^2}{2G + 2cb/3}\right) = \frac{\beta}{3(K+1)}.$$

Equivalently, we have

$$\mathbb{P}\{E_①\} \geq 1 - \frac{\beta}{3(K+1)}, \quad \text{for} \quad E_① = \left\{\text{either} \quad \sum_{l=0}^{T-1}\sigma_l^2 > \frac{\exp(-2\gamma\mu T)R^4}{294\ln\frac{6(K+1)}{\beta}} \quad \text{or} \quad |①| \leq \frac{1}{7}\exp(-\gamma\mu T)R^2\right\}. \tag{261}$$

In addition, E_{T-1} implies that

$$
\begin{aligned}
\sum_{l=0}^{T-1}\sigma_l^2 &\overset{(260)}{\leq} \frac{4\gamma^2\exp(-2\gamma\mu T)R^2}{2809\ln\frac{6(K+1)}{\beta}}\sum_{l=0}^{T-1}\frac{\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right]}{\exp(-\gamma\mu l)}\\
&\overset{(258),T\leq K+1}{\leq} \frac{72\gamma^2\exp(-2\gamma\mu T)R^2\sigma^\alpha}{2809\ln\frac{6(K+1)}{\beta}}\sum_{l=0}^{K}\frac{\lambda_l^{2-\alpha}}{\exp(-\gamma\mu l)}\\
&\overset{(240)}{\leq} \frac{72\gamma^\alpha\exp(-2\gamma\mu T)R^{4-\alpha}\sigma^\alpha}{2809\cdot 120^{2-\alpha}\ln^{3-\alpha}\frac{6(K+1)}{\beta}}\sum_{l=0}^{K}\frac{1}{\exp(-\gamma\mu l)}\cdot\left(\exp(-\gamma\mu(1+l/2))\right)^{2-\alpha}\\
&\leq \frac{72\gamma^\alpha\exp(-2\gamma\mu T)R^{4-\alpha}\sigma^\alpha}{2809\cdot 120^{2-\alpha}\ln^{3-\alpha}\frac{6(K+1)}{\beta}}\sum_{l=0}^{K}\exp(\gamma\mu(\alpha-2))\cdot\exp\left(\frac{\gamma\mu\alpha l}{2}\right)\\
&\leq \frac{72\gamma^\alpha\exp(-2\gamma\mu T)R^{4-\alpha}\sigma^\alpha(K+1)\exp\left(\frac{\gamma\mu\alpha K}{2}\right)}{2809\cdot 120^{2-\alpha}\ln^{3-\alpha}\frac{6(K+1)}{\beta}}\\
&\overset{(237)}{\leq} \frac{\exp(-2\gamma\mu T)R^4}{294\ln\frac{6(K+1)}{\beta}},
\end{aligned}
\tag{262}
$$

where we also show that E_{T-1} implies

$$
\gamma^2R^2\sum_{l=0}^{K}\frac{\lambda_l^{2-\alpha}}{\exp(-\gamma\mu l)} \leq \frac{\gamma^\alpha R^{4-\alpha}(K+1)\exp(\frac{\gamma\mu\alpha K}{2})}{120^{2-\alpha}\ln^{2-\alpha}\frac{6(K+1)}{\beta}}.
\tag{263}
$$

Upper bound for ②. From E_{T-1} it follows that

$$
\begin{aligned}
② &\leq 4\gamma^3\mu\sum_{l=0}^{T-1}\exp(-\gamma\mu(T-1-l))\|\zeta_l\|\cdot\|\omega_l^b\|\\
&\overset{(251),(256)}{\leq} 2^{2+\alpha}\cdot\sqrt{2}\exp(-\gamma\mu(T-1))\gamma^3\mu LR\sum_{l=0}^{T-1}\frac{\sigma^\alpha}{\lambda_l^{\alpha-1}\exp(-\gamma\mu l/2)}\\
&\overset{(240)}{=} \frac{2^{2+\alpha}\cdot 120^{\alpha-1}\sqrt{2}\exp(-\gamma\mu(T-1))\gamma^{2+\alpha}\mu L\sigma^\alpha\ln^{\alpha-1}\frac{6(K+1)}{\beta}}{R^{\alpha-2}}\sum_{l=0}^{T-1}\frac{1}{\exp\left(-\gamma\mu(1+l/2)\right)^{\alpha-1}\cdot\exp(-\gamma\mu l/2)}\\
&\overset{T\leq K+1}{\leq} \frac{2^{3+\alpha}\cdot 120^{\alpha-1}\sqrt{2}\exp(-\gamma\mu(T-1))\gamma^{2+\alpha}\mu L\sigma^\alpha\ln^{\alpha-1}\frac{6(K+1)}{\beta}}{R^{\alpha-2}}\sum_{l=0}^{K}\exp\left(\frac{\gamma\mu\alpha l}{2}\right)\\
&\leq \frac{2^{3+\alpha}\cdot 120^{\alpha-1}\sqrt{2}\exp(-\gamma\mu(T-1))\gamma^{2+\alpha}\mu L\sigma^\alpha\ln^{\alpha-1}\frac{6(K+1)}{\beta}(K+1)\exp\left(\frac{\gamma\mu\alpha K}{2}\right)}{R^{\alpha-2}}\\
&\overset{(237)}{\leq} \frac{1}{7}\exp(-\gamma\mu T)R^2,
\end{aligned}
\tag{264}
$$

where we also show that E_{T-1} implies

$$
\gamma R\sum_{l=0}^{T-1}\frac{1}{\lambda_l^{\alpha-1}\exp(-\gamma\mu l/2)} \leq \frac{120^{\alpha-1}\gamma^\alpha(K+1)\exp(\frac{\gamma\mu\alpha K}{2})\ln^{\alpha-1}\frac{6(K+1)}{\beta}}{R^{\alpha-2}}.
\tag{265}
$$

Upper bound for ③. By definition of θ_l^u, we have $\mathbb{E}_{\xi_2^l}[\theta_l^u] = 0$ and

$$
\mathbb{E}_{\xi_2^l}\left[2\gamma(1-\gamma\mu)^{T-1-l}\langle\eta_l,\theta_l^u\rangle\right] = 0.
$$

Next, sum ③ has bounded with probability 1 terms:

$$\begin{aligned}|2\gamma(1-\gamma\mu)^{T-1-l}\langle \eta_l, \theta_l^u\rangle| &\leq 2\gamma\exp(-\gamma\mu(T-1-l))\|\eta_l\|\cdot\|\theta_l^u\| \\ &\overset{(251),(255)}{\leq} 4\sqrt{7}\gamma(1+\gamma L)\exp(-\gamma\mu(T-1-l/2))R\lambda_l \\ &\overset{(237),(240)}{\leq} \frac{\exp(-\gamma\mu T)R^2}{7\ln\frac{6(K+1)}{\beta}} \overset{\text{def}}{=} c. \end{aligned} \tag{266}$$

The summands also have bounded conditional variances $\widetilde{\sigma}_l^2 \overset{\text{def}}{=} \mathbb{E}_{\xi_2^l}\left[4\gamma^2(1-\gamma\mu)^{2T-2-2l}\langle \eta_l, \theta_l^u\rangle^2\right]$:

$$\begin{aligned}\widetilde{\sigma}_l^2 &\leq \mathbb{E}_{\xi_2^l}\left[4\gamma^2\exp(-\gamma\mu(2T-2-2l))\|\eta_l\|^2\cdot\|\theta_l^u\|^2\right] \\ &\overset{(251)}{\leq} 49\gamma^2(1+\gamma L)^2\exp(-\gamma\mu(2T-2-l))R^2\mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] \\ &\overset{(237)}{\leq} 50\gamma^2\exp(-\gamma\mu(2T-l))R^2\mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right]. \end{aligned} \tag{267}$$

In other words, we showed that $\{2\gamma(1-\gamma\mu)^{T-1-l}\langle \eta_l, \theta_l^u\rangle\}_{l=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\widetilde{\sigma}_l^2\}_{l=0}^{T-1}$. Next, we apply Bernstein's inequality (Lemma B.2) with $X_l = 2\gamma(1-\gamma\mu)^{T-1-l}\langle \eta_l, \theta_l^u\rangle$, parameter c as in (266), $b = \frac{1}{7}\exp(-\gamma\mu T)R^2$, $G = \frac{\exp(-2\gamma\mu T)R^4}{294\ln\frac{6(K+1)}{\beta}}$:

$$\mathbb{P}\left\{|③| > \frac{1}{7}\exp(-\gamma\mu T)R^2 \text{ and } \sum_{l=0}^{T-1}\widetilde{\sigma}_l^2 \leq \frac{\exp(-2\gamma\mu T)R^4}{294\ln\frac{6(K+1)}{\beta}}\right\} \leq 2\exp\left(-\frac{b^2}{2G + 2cb/3}\right) = \frac{\beta}{3(K+1)}.$$

Equivalently, we have

$$\mathbb{P}\{E_③\} \geq 1 - \frac{\beta}{3(K+1)}, \quad \text{for} \quad E_③ = \left\{\text{either} \quad \sum_{l=0}^{T-1}\widetilde{\sigma}_l^2 > \frac{\exp(-2\gamma\mu T)R^4}{294\ln\frac{6(K+1)}{\beta}} \quad \text{or} \quad |③| \leq \frac{1}{7}\exp(-\gamma\mu T)R^2\right\}. \tag{268}$$

In addition, E_{T-1} implies that

$$\begin{aligned}\sum_{l=0}^{T-1}\widetilde{\sigma}_l^2 &\overset{(267)}{\leq} 50\gamma^2\exp(-2\gamma\mu T)R^2\sum_{l=0}^{T-1}\frac{\mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right]}{\exp(-\gamma\mu l)} \\ &\overset{(258),T\leq K+1}{\leq} 900\gamma^2\exp(-2\gamma\mu T)R^2\sigma^\alpha\sum_{l=0}^{K}\frac{\lambda_l^{2-\alpha}}{\exp(-\gamma\mu l)} \\ &\overset{(263)}{\leq} \frac{900\gamma^\alpha\exp(-2\gamma\mu T)R^{4-\alpha}\sigma^\alpha(K+1)\exp(\frac{\gamma\mu\alpha K}{2})}{120^{2-\alpha}\ln^{2-\alpha}\frac{6(K+1)}{\beta}} \\ &\overset{(237)}{\leq} \frac{\exp(-2\gamma\mu T)R^4}{294\ln\frac{6(K+1)}{\beta}}. \end{aligned} \tag{269}$$

Upper bound for ④. From E_{T-1} it follows that

$$\begin{aligned}④ &\leq 2\gamma\exp(-\gamma\mu(T-1))\sum_{l=0}^{T-1}\frac{\|\eta_l\|\cdot\|\theta_l^b\|}{\exp(-\gamma\mu l)} \\ &\overset{(251),(256)}{\leq} 2^{1+\alpha}\sqrt{7}\gamma(1+\gamma L)\exp(-\gamma\mu(T-1))R\sigma^\alpha\sum_{l=0}^{T-1}\frac{1}{\lambda_l^{\alpha-1}\exp(-\gamma\mu l/2)} \\ &\overset{(265)}{\leq} \frac{2^{3+\alpha}\cdot 120^{\alpha-1}\sqrt{7}\gamma^\alpha(1+\gamma L)\exp(-\gamma\mu T)(K+1)\exp\left(\frac{\gamma\mu\alpha K}{2}\right)\ln^{\alpha-1}\frac{6(K+1)}{\beta}}{R^{\alpha-2}} \\ &\overset{(237)}{\leq} \frac{1}{7}\exp(-\gamma\mu T)R^2. \end{aligned} \tag{270}$$

Upper bound for ⑤. From E_{T-1} it follows that

$$
\begin{aligned}
⑤ &= 2\gamma^2 \exp(-\gamma\mu(T-1)) \sum_{l=0}^{T-1} \frac{\mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] + 4\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right]}{\exp(-\gamma\mu l)} \\
&\overset{(258)}{\leq} 180\gamma^2 \exp(-\gamma\mu(T-1))\sigma^\alpha \sum_{l=0}^{T-1} \frac{\lambda_l^{2-\alpha}}{\exp(-\gamma\mu l)} \\
&\overset{(263)}{\leq} \frac{180\gamma^\alpha R^{2-\alpha}\exp(-\gamma\mu(T-1))\sigma^\alpha (K+1)\exp(\frac{\gamma\mu\alpha K}{2})}{120^{2-\alpha}\ln^{2-\alpha}\frac{6(K+1)}{\beta}} \\
&\overset{(237)}{\leq} \frac{1}{7}\exp(-\gamma\mu T)R^2.
\end{aligned} \tag{271}
$$

Upper bound for ⑥. First, we have

$$
2\gamma^2(1-\gamma\mu)^{T-1-l}\mathbb{E}_{\xi_1^l,\xi_2^l}\left[\|\theta_l^u\|^2 + 4\|\omega_l^u\|^2 - \mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] - 4\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right]\right] = 0.
$$

Next, sum ⑥ has bounded with probability 1 terms:

$$
\begin{aligned}
2\gamma^2(1-\gamma\mu)^{T-1-l}\left|\|\theta_l^u\|^2 + 4\|\omega_l^u\|^2 - \mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] - 4\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right]\right| &\overset{(255)}{\leq} \frac{80\gamma^2\exp(-\gamma\mu T)\lambda_l^2}{\exp(-\gamma\mu(1+l))} \\
&\overset{(240)}{\leq} \frac{\exp(-\gamma\mu T)R^2}{7\ln\frac{6(K+1)}{\beta}} \\
&\overset{\text{def}}{=} c.
\end{aligned} \tag{272}
$$

The summands also have conditional variances

$$
\widehat{\sigma}_l^2 \overset{\text{def}}{=} \mathbb{E}_{\xi_1^l,\xi_2^l}\left[4\gamma^4(1-\gamma\mu)^{2T-2-2l}\left|\|\theta_l^u\|^2 + 4\|\omega_l^u\|^2 - \mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] - 4\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right]\right|^2\right]
$$

that are bounded

$$
\begin{aligned}
\widehat{\sigma}_l^2 &\overset{(272)}{\leq} \frac{2\gamma^2\exp(-2\gamma\mu T)R^2}{7\exp(-\gamma\mu(1+l))\ln\frac{6(K+1)}{\beta}}\mathbb{E}_{\xi_1^l,\xi_2^l}\left[\left|\|\theta_l^u\|^2 + 4\|\omega_l^u\|^2 - \mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] - 4\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right]\right|\right] \\
&\leq \frac{4\gamma^2\exp(-2\gamma\mu T)R^2}{7\exp(-\gamma\mu(1+l))\ln\frac{6(K+1)}{\beta}}\mathbb{E}_{\xi_1^l,\xi_2^l}\left[\|\theta_l^u\|^2 + 4\|\omega_l^u\|^2\right].
\end{aligned} \tag{273}
$$

In other words, we showed that $\left\{2\gamma^2(1-\gamma\mu)^{T-1-l}\left(\|\theta_l^u\|^2 + 4\|\omega_l^u\|^2 - \mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] - 4\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right]\right)\right\}_{l=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\widehat{\sigma}_l^2\}_{l=0}^{T-1}$. Next, we apply Bernstein's inequality (Lemma B.2) with $X_l = 2\gamma^2(1-\gamma\mu)^{T-1-l}\left(\|\theta_l^u\|^2 + 4\|\omega_l^u\|^2 - \mathbb{E}_{\xi_2^l}\left[\|\theta_l^u\|^2\right] - 4\mathbb{E}_{\xi_1^l}\left[\|\omega_l^u\|^2\right]\right)$, parameter c as in (272), $b = \frac{1}{7}\exp(-\gamma\mu T)R^2$, $G = \frac{\exp(-2\gamma\mu T)R^4}{294\ln\frac{6(K+1)}{\beta}}$:

$$
\mathbb{P}\left\{|⑥| > \frac{1}{7}\exp(-\gamma\mu T)R^2 \text{ and } \sum_{l=0}^{T-1}\widehat{\sigma}_l^2 \leq \frac{\exp(-2\gamma\mu T)R^4}{294\ln\frac{6(K+1)}{\beta}}\right\} \leq 2\exp\left(-\frac{b^2}{2G + 2cb/3}\right) = \frac{\beta}{3(K+1)}.
$$

Equivalently, we have

$$
\mathbb{P}\{E_{⑥}\} \geq 1 - \frac{\beta}{3(K+1)}, \quad \text{for} \quad E_{⑥} = \left\{\text{either} \quad \sum_{l=0}^{T-1}\widehat{\sigma}_l^2 > \frac{\exp(-2\gamma\mu T)R^4}{294\ln\frac{6(K+1)}{\beta}} \quad \text{or} \quad |⑥| \leq \frac{1}{7}\exp(-\gamma\mu T)R^2\right\}. \tag{274}
$$

In addition, E_{T-1} implies that

$$\begin{aligned}
\sum_{l=0}^{T-1} \widehat{\sigma}_l^2 &\overset{(273)}{\leq} \frac{4\gamma^2\exp(-\gamma\mu(2T-1))R^2}{7\ln\frac{6(K+1)}{\beta}} \sum_{l=0}^{T-1} \frac{\mathbb{E}_{\xi_1^l,\xi_2^l}\left[\|\theta_l^u\|^2 + 4\|\omega_l^u\|^2\right]}{\exp(-\gamma\mu l)} \\
&\overset{(258),T\leq K+1}{\leq} \frac{360\gamma^2\exp(-\gamma\mu(2T-1))R^2\sigma^\alpha}{7\ln\frac{6(K+1)}{\beta}} \sum_{l=0}^{K} \frac{\lambda_l^{2-\alpha}}{\exp(-\gamma\mu l)} \\
&\overset{(263)}{\leq} \frac{360\gamma^\alpha\exp(-\gamma\mu(2T-1))R^{4-\alpha}\sigma^\alpha(K+1)\exp(\frac{\gamma\mu\alpha K}{2})}{7\cdot 120^{2-\alpha}\ln^{3-\alpha}\frac{6(K+1)}{\beta}} \\
&\overset{(237)}{\leq} \frac{\exp(-2\gamma\mu T)R^4}{294\ln\frac{6(K+1)}{\beta}}.
\end{aligned} \tag{275}$$

Upper bound for ⑦. From E_{T-1} it follows that

$$\begin{aligned}
⑦ &= 2\gamma^2\sum_{l=0}^{T-1}\exp(-\gamma\mu(T-1-l))\left(\|\theta_l^b\|^2 + 4\|\omega_l^b\|^2\right) \\
&\overset{(256)}{\leq} 10\cdot 2^{2\alpha}\gamma^2\exp(-\gamma\mu(T-1))\sigma^{2\alpha}\sum_{l=0}^{T-1}\frac{1}{\lambda_l^{2\alpha-2}\exp(-\gamma\mu l)} \\
&\overset{(240),T\leq K+1}{\leq} \frac{20\cdot 2^{2\alpha}\cdot 120^{2\alpha-2}\gamma^{2\alpha}\exp(-\gamma\mu T)\sigma^{2\alpha}\ln^{2\alpha-2}\frac{6(K+1)}{\beta}}{R^{2\alpha-2}}\sum_{l=0}^{K}\exp\left(\gamma\mu(2\alpha-2)\left(1+\frac{l}{2}\right)\right)\exp(\gamma\mu l) \\
&\leq \frac{40\cdot 2^{2\alpha}\cdot 120^{2\alpha-2}\gamma^{2\alpha}\exp(-\gamma\mu T)\sigma^{2\alpha}\ln^{2\alpha-2}\frac{6(K+1)}{\beta}}{R^{2\alpha-2}}\sum_{l=0}^{K}\exp(\gamma\mu\alpha l) \\
&\leq \frac{40\cdot 2^{2\alpha}\cdot 120^{2\alpha-2}\gamma^{2\alpha}\exp(-\gamma\mu T)\sigma^{2\alpha}\ln^{2\alpha-2}\frac{6(K+1)}{\beta}(K+1)\exp(\gamma\mu\alpha K)}{R^{2\alpha-2}} \\
&\overset{(237)}{\leq} \frac{1}{7}\exp(-\gamma\mu T)R^2.
\end{aligned} \tag{276}$$

Now, we have the upper bounds for ①, ②, ③, ④, ⑤, ⑥, ⑦. In particular, probability event E_{T-1} implies

$$R_T^2 \overset{(254)}{\leq} \exp(-\gamma\mu T)R^2 + ① + ② + ③ + ④ + ⑤ + ⑥ + ⑦,$$
$$② \overset{(264)}{\leq} \frac{1}{7}\exp(-\gamma\mu T)R^2, \quad ④ \overset{(270)}{\leq} \frac{1}{7}\exp(-\gamma\mu T)R^2,$$
$$⑤ \overset{(271)}{\leq} \frac{1}{7}\exp(-\gamma\mu T)R^2, \quad ⑦ \overset{(276)}{\leq} \frac{1}{7}\exp(-\gamma\mu T)R^2,$$
$$\sum_{l=0}^{T-1}\sigma_l^2 \overset{(262)}{\leq} \frac{\exp(-2\gamma\mu T)R^4}{294\ln\frac{6(K+1)}{\beta}}, \quad \sum_{l=0}^{T-1}\widetilde{\sigma}_l^2 \overset{(269)}{\leq} \frac{\exp(-2\gamma\mu T)R^4}{294\ln\frac{6(K+1)}{\beta}}, \quad \sum_{l=0}^{T-1}\widehat{\sigma}_l^2 \overset{(275)}{\leq} \frac{\exp(-2\gamma\mu T)R^4}{294\ln\frac{6(K+1)}{\beta}}.$$

Moreover, we also have (see (261), (268), (274) and our induction assumption)

$$\mathbb{P}\{E_{T-1}\} \geq 1 - \frac{(T-1)\beta}{K+1},$$
$$\mathbb{P}\{E_{①}\} \geq 1 - \frac{\beta}{3(K+1)}, \quad \mathbb{P}\{E_{③}\} \geq 1 - \frac{\beta}{3(K+1)}, \quad \mathbb{P}\{E_{⑥}\} \geq 1 - \frac{\beta}{3(K+1)},$$

where

$$\begin{aligned}
E_{①} &= \left\{\text{either} \quad \sum_{l=0}^{T-1} \sigma_l^2 > \frac{\exp(-2\gamma\mu T)R^4}{294\ln\frac{6(K+1)}{\beta}} \quad \text{or} \quad |①| \leq \frac{1}{7}\exp(-\gamma\mu T)R^2\right\},\\
E_{③} &= \left\{\text{either} \quad \sum_{l=0}^{T-1} \widetilde{\sigma}_l^2 > \frac{\exp(-2\gamma\mu T)R^4}{294\ln\frac{6(K+1)}{\beta}} \quad \text{or} \quad |③| \leq \frac{1}{7}\exp(-\gamma\mu T)R^2\right\},\\
E_{⑥} &= \left\{\text{either} \quad \sum_{l=0}^{T-1} \widehat{\sigma}_l^2 > \frac{\exp(-2\gamma\mu T)R^4}{294\ln\frac{6(K+1)}{\beta}} \quad \text{or} \quad |⑥| \leq \frac{1}{7}\exp(-\gamma\mu T)R^2\right\}.
\end{aligned}$$

Thus, probability event $E_{T-1} \cap E_{①} \cap E_{③} \cap E_{⑥}$ implies

$$\begin{aligned}
R_T^2 &\overset{(254)}{\leq} \exp(-\gamma\mu T)R^2 + ① + ② + ③ + ④ + ⑤ + ⑥ + ⑦\\
&\leq 2\exp(-\gamma\mu T)R^2,
\end{aligned}$$

which is equivalent to (244) for $t = T$, and

$$\mathbb{P}\{E_T\} \geq \mathbb{P}\{E_{T-1} \cap E_{①} \cap E_{③} \cap E_{⑥}\} = 1 - \mathbb{P}\{\overline{E}_{T-1} \cup \overline{E}_{①} \cup \overline{E}_{③} \cup \overline{E}_{⑥}\} \geq 1 - \frac{T\beta}{K+1}.$$

This finishes the inductive part of our proof, i.e., for all $k = 0, 1, \ldots, K+1$ we have $\mathbb{P}\{E_k\} \geq 1 - {k\beta}/{(K+1)}$. In particular, for $k = K+1$ we have that with probability at least $1-\beta$

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

Finally, if

$$\begin{aligned}
\gamma &= \min\left\{\frac{1}{650L\ln\frac{6(K+1)}{\beta}}, \frac{\ln(B_K)}{\mu(K+1)}\right\},\\
B_K &= \max\left\{2, \frac{(K+1)^{\frac{2(\alpha-1)}{\alpha}}\mu^2R^2}{264600^{\frac{2}{\alpha}}\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{6(K+1)}{\beta}\right)\ln^2(B_K)}\right\}\\
&= \mathcal{O}\left(\max\left\{2, \frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2R^2}{\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)\ln^2\left(\max\left\{2, \frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2R^2}{\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)}\right\}\right)}\right\}\right)
\end{aligned}$$

then with probability at least $1-\beta$

$$\begin{aligned}
\|x^{K+1} - x^*\|^2 &\leq 2\exp(-\gamma\mu(K+1))R^2\\
&= 2R^2\max\left\{\exp\left(-\frac{\mu(K+1)}{650L\ln\frac{6(K+1)}{\beta}}\right), \frac{1}{B_K}\right\}\\
&= \mathcal{O}\left(\max\left\{R^2\exp\left(-\frac{\mu K}{L\ln\frac{K}{\beta}}\right), \frac{\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)\ln^2\left(\max\left\{2, \frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2R^2}{\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)}\right\}\right)}{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2}\right\}\right).
\end{aligned}$$

To get $\|x^{K+1} - x^*\|^2 \leq \varepsilon$ with probability at least $1-\beta$ it is sufficient to choose K such that both terms in the maximum above are $\mathcal{O}(\varepsilon)$. This leads to

$$K = \mathcal{O}\left(\frac{L}{\mu}\ln\left(\frac{R^2}{\varepsilon}\right)\ln\left(\frac{L}{\mu\beta}\ln\frac{R^2}{\varepsilon}\right), \left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\ln\left(\frac{1}{\beta}\left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right)\ln^{\frac{\alpha}{\alpha-1}}(B_\varepsilon)\right),$$

where

$$B_\varepsilon = \max\left\{2, \frac{R^2}{\varepsilon \ln\left(\frac{1}{\beta}\left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right)}\right\}.$$

This concludes the proof. □

H. Missing Proofs for clipped-SGDA

In this section, we provide the complete formulation of the main results for clipped-SGDA and the missing proofs. For brevity, we will use the following notation: $\widetilde{F}_{\xi^k}(x^k) = \texttt{clip}\left(F_{\xi^k}(x^k), \lambda_k\right)$.

Algorithm 5 Clipped Stochastic Gradient Descent Ascent (clipped-SGDA) (Gorbunov et al., 2022a)

Input: starting point x^0, number of iterations K, stepsize $\gamma > 0$, clipping levels $\{\lambda_k\}_{k=0}^{K-1}$.
1: **for** $k = 0, \ldots, K$ **do**
2: Compute $\widetilde{F}_{\xi^k}(x^k) = \texttt{clip}\left(F_{\xi^k}(x^k), \lambda_k\right)$ using a fresh sample $\xi^k \sim \mathcal{D}_k$
3: $x^{k+1} = x^k - \gamma \widetilde{F}_{\xi^k}(x^k)$
4: **end for**
Output: x^{K+1} or $x^K_{\texttt{avg}} = \frac{1}{K+1}\sum\limits_{k=0}^{K} x^K$

H.1. Monotone Star-Cocoercive Problems

We start with the following lemma derived by Gorbunov et al. (2022b). Since this lemma handles only deterministic part of the algorithm, the proof is the same as in the original work.

Lemma H.1 (Lemma D.1 from (Gorbunov et al., 2022b)). *Let Assumptions 1.8 and 1.10 hold for $Q = B_{3R}(x^*)$, where $R \geq \|x^0 - x^*\|$ and $0 < \gamma \leq {}^2/\ell$. If x^k lies in $B_{3R}(x^*)$ for all $k = 0, 1, \ldots, K$ for some $K \geq 0$, then for all $u \in B_{3R}(x^*)$ the iterates produced by* clipped-SGDA *satisfy*

$$\langle F(u), x^K_{avg} - u\rangle \leq \frac{\|x^0 - u\|^2 - \|x^{K+1} - u\|^2}{2\gamma(K+1)} + \frac{\gamma}{2(K+1)}\sum_{k=0}^{K}\left(\|F(x^k)\|^2 + \|\omega_k\|^2\right) + \frac{1}{K+1}\sum_{k=0}^{K}\langle x^k - u - \gamma F(x^k), \omega_k\rangle, \tag{277}$$

$$x^K_{avg} \stackrel{def}{=} \frac{1}{K+1}\sum_{k=0}^{K} x^k, \tag{278}$$

$$\omega_k \stackrel{def}{=} F(x^k) - \widetilde{F}_{\xi^k}(x^k). \tag{279}$$

Also we need to use the following lemma to estimate the term $\sum\limits_{k=0}^{K}\|F(x^k)\|^2$ from the right hand side of (277) in the proof of the main theorem.

Lemma H.2 (Lemma D.2 from (Gorbunov et al., 2022b)). *Let Assumption 1.10 hold for $Q = B_{3R}(x^*)$, where $R \geq R_0 \stackrel{def}{=} \|x^0 - x^*\|$ and $0 < \gamma \leq {}^2/\ell$. If x^k lies in $B_{3R}(x^*)$ for all $k = 0, 1, \ldots, K$ for some $K \geq 0$, then the iterates produced by* clipped-SGDA *satisfy*

$$\frac{\gamma}{K+1}\left(\frac{2}{\ell} - \gamma\right)\sum_{k=0}^{K}\|F(x^k)\|^2 \leq \frac{\|x^0 - x^*\|^2 - \|x^{K+1} - x^*\|^2}{K+1} + \frac{2\gamma}{K+1}\sum_{k=0}^{K}\langle x^k - x^* - \gamma F(x^k), \omega_k\rangle + \frac{\gamma^2}{K+1}\sum_{k=0}^{K}\|\omega_k\|^2, \tag{280}$$

where ω_k is defined in (279).

Using those lemmas, we prove the main convergence result for clipped-SGDA in the monotone star-cocoercive case.

Theorem H.3 (Case 1 in Theorem 4.2). *Let Assumptions 1.1, 1.8, 1.10 hold for $Q = B_{3R}(x^*)$, where $R \geq \|x^0 - x^*\|$, and*

$$0 < \gamma \leq \min\left\{\frac{1}{170\ell \ln \frac{6(K+1)}{\beta}}, \frac{R}{97200^{\frac{1}{\alpha}}(K+1)^{\frac{1}{\alpha}}\sigma \ln^{\frac{\alpha-1}{\alpha}}\frac{6(K+1)}{\beta}}\right\}, \tag{281}$$

$$\lambda_k \equiv \lambda = \frac{R}{60\gamma \ln \frac{6(K+1)}{\beta}}, \tag{282}$$

for some $K \geq 0$ and $\beta \in (0,1]$ such that $\ln \frac{6(K+1)}{\beta} \geq 1$. Then, after K iterations the iterates produced by clipped-SGDA *with probability at least $1-\beta$ satisfy*

$$\texttt{Gap}_R(x^K_{avg}) \leq \frac{5R^2}{\gamma(K+1)} \quad \textit{and} \quad \{x^k\}_{k=0}^{K+1} \subseteq B_{3R}(x^*), \tag{283}$$

where x^K_{avg} is defined in (278). In particular, when γ equals the minimum from (281), then the iterates produced by clipped-SGDA *after K iterations with probability at least $1-\beta$ satisfy*

$$\texttt{Gap}_R(\widetilde{x}^K_{avg}) = \mathcal{O}\left(\max\left\{\frac{\ell R^2 \ln \frac{K}{\beta}}{K}, \frac{\sigma R \ln^{\frac{\alpha-1}{\alpha}}\frac{K}{\beta}}{K^{\frac{\alpha-1}{\alpha}}}\right\}\right), \tag{284}$$

meaning that to achieve $\texttt{Gap}_R(\widetilde{x}^K_{avg}) \leq \varepsilon$ with probability at least $1-\beta$ clipped-SGDA *requires*

$$K = \mathcal{O}\left(\frac{\ell R^2}{\varepsilon}\ln\frac{\ell R^2}{\varepsilon\beta}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\ln\left(\frac{1}{\beta}\left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right) \quad \textit{iterations/oracle calls.} \tag{285}$$

Proof. The proof follows similar steps as the proof of Theorem D.1 from (Gorbunov et al., 2022a). The key difference is related to the application of Bernstein inequality and estimating biases and variances of stochastic terms.

Let $R_k = \|x^k - x^*\|$ for all $k \geq 0$. As in the previous results, the proof is based on the induction argument and showing that the iterates do not leave some ball around the solution with high probability. More precisely, for each $k = 0, 1, \ldots, K+1$ we consider probability event E_k as follows: inequalities

$$\|x^t - x^*\|^2 \leq 2R^2 \quad \text{and} \quad \gamma\left\|\sum_{l=0}^{t-1}\omega_l\right\| \leq R \tag{286}$$

hold for $t = 0, 1, \ldots, k$ simultaneously. We want to prove that $\mathbb{P}\{E_k\} \geq 1 - k\beta/(K+1)$ for all $k = 0, 1, \ldots, K+1$ by induction. The base of the induction is trivial: for $k=0$ we have $R_0^2 \leq 2R^2$ by definition and $\sum_{l=0}^{-1}\omega_l = 0$. Next, assume that the statement holds for $k = T \leq K$, i.e., we have $\mathbb{P}\{E_T\} \geq 1 - T\beta/(K+1)$. Given this, we need to prove that $\mathbb{P}\{E_{T+1}\} \geq 1 - (T+1)\beta/(K+1)$. Since probability event E_T implies $R_t^2 \leq 2R^2$, we have $x^t \in B_{2R}(x^*)$ for all $t = 0, 1, \ldots, T$. According to this, the assumptions of Lemma H.2 hold and E_T implies ($\gamma < 1/\ell$)

$$\frac{\gamma}{\ell(T+1)}\sum_{t=0}^{T}\|F(x^t)\|^2 \leq \frac{\|x^0 - x^*\|^2 - \|x^{T+1} - x^*\|^2}{T+1} + \frac{2\gamma}{T+1}\sum_{t=0}^{T}\langle x^t - x^* - \gamma F(x^t), \omega_t\rangle + \frac{\gamma^2}{T+1}\sum_{t=0}^{T}\|\omega_t\|^2 \tag{287}$$

and by ℓ-star-cocoersivity we have

$$\|F(x^t)\| \leq \ell\|x^t - x^*\| \overset{(286)}{\leq} \sqrt{2}\ell R \overset{(281),(282)}{\leq} \frac{\lambda}{2} \tag{288}$$

for all $t = 0, 1, \ldots, T$. Using (287), we obtain

$$R_{T+1}^2 \leq R_0^2 + 2\gamma\sum_{t=0}^{T}\langle x^t - x^* - \gamma F(x^t), \omega_t\rangle + \gamma^2\sum_{t=0}^{T}\|\omega_t\|^2.$$

Due to (288), we have

$$\begin{aligned}\|x^t - x^* - \gamma F(x^t)\| &\leq \|x^t - x^*\| + \gamma\|F(x^t)\| \overset{(14),(286)}{\leq} 2R + \gamma\ell\|x^t - x^*\| \\ &\overset{(286)}{\leq} 2R + 2R\gamma\ell \overset{(281)}{\leq} 3R, \end{aligned} \tag{289}$$

for all $t = 0, 1, \ldots, T$. To handle the sum above, we introduce a new vector

$$\eta_t = \begin{cases} x^t - x^* - \gamma F(x^t), & \text{if } \|x^t - x^* - \gamma F(x^t)\| \leq 3R, \\ 0, & \text{otherwise,} \end{cases}$$

for all $t = 0, 1, \ldots, T$. This vector η_t is bounded with probability 1:

$$\|\eta_t\| \leq 3R \tag{290}$$

for all $t = 0, 1, \ldots, T$. We also notice that probability event E_T implies $\eta_t = x^t - x^* - \gamma F(x^t)$ for all $t = 0, 1, \ldots, T$ Thus, thanks to (289), E_T implies

$$R_{T+1}^2 \leq R^2 + 2\gamma \sum_{t=0}^{T} \langle \eta_t, \omega_t \rangle + \gamma^2 \sum_{t=0}^{T} \|\omega_t\|^2.$$

To handle the sums appeared on the right-hand side of the previous inequality we consider unbiased and biased parts of ω_t:

$$\omega_t^u \overset{\text{def}}{=} \mathbb{E}_{\xi^t}\left[\widetilde{F}_{\xi^t}(x^t)\right] - \widetilde{F}_{\xi^t}(x^t), \quad \omega_t^b \overset{\text{def}}{=} F(x^t) - \mathbb{E}_{\xi^t}\left[\widetilde{F}_{\xi^t}(x^t)\right] \tag{291}$$

for all $t = 0, \ldots, T$. Also, by definition we have $\omega_t = \omega_t^u + \omega_t^b$ for all $t = 0, \ldots, T$. Therefore, E_T implies

$$\begin{aligned} R_{T+1}^2 &\leq R^2 + \underbrace{2\gamma \sum_{t=0}^{T} \langle \eta_t, \omega_t^u \rangle}_{①} + \underbrace{2\gamma \sum_{t=0}^{T} \langle \eta_t, \omega_t^b \rangle}_{②} + \underbrace{2\gamma^2 \sum_{t=0}^{T} \left(\mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]\right)}_{③} \\ &\quad + \underbrace{2\gamma^2 \sum_{t=0}^{T} \left(\|\omega_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]\right)}_{④} + \underbrace{2\gamma^2 \sum_{t=0}^{T} \left(\|\omega_t^b\|^2\right)}_{⑤}. \end{aligned} \tag{292}$$

We notice that the above inequality does not rely on monotonicity of F.

According to the induction assumption, from probability event E_T we have $x^t \in B_{2R}(x^*)$ for all $t = 0, 1, \ldots, T$. Thus, the assumptions of Lemma H.1 hold and probability event E_T implies

$$\begin{aligned} 2\gamma(T+1)\texttt{Gap}_R(x_{\texttt{avg}}^T) &\leq \max_{u \in B_R(x^*)} \left\{ \|x^0 - u\|^2 + 2\gamma \sum_{t=0}^{T} \langle x^t - u - \gamma F(x^t), \omega_t \rangle \right\} \\ &\quad + \gamma^2 \sum_{t=0}^{T} \left(\|F(x^t)\|^2 + \|\omega_t\|^2\right), \\ &= \max_{u \in B_R(x^*)} \left\{ \|x^0 - u\|^2 + 2\gamma \sum_{t=0}^{T} \langle x^* - u, \omega_t \rangle \right\} \\ &\quad + 2\gamma \sum_{t=0}^{T} \langle x^t - x^* - \gamma F(x^t), \omega_t \rangle \\ &\quad + \gamma^2 \sum_{t=0}^{T} \left(\|F(x^t)\|^2 + \|\omega_t\|^2\right). \end{aligned}$$

As we mentioned before, E_T implies $\eta_t = x^t - x^* - \gamma F(x^t)$ for all $t = 0, 1, \ldots, T$ as well as (287) and $\gamma < 1/\ell$. Due to that, probability event E_T implies

$$\begin{aligned}
2\gamma(T+1)\texttt{Gap}_R(x^T_{\texttt{avg}}) &\leq \max_{u \in B_R(x^*)} \left\{\|x^0 - u\|^2\right\} + 2\gamma \max_{u \in B_R(x^*)} \left\{\sum_{t=0}^T \langle x^* - u, \omega_t\rangle\right\} \\
&\quad + 2\gamma \sum_{t=0}^T \langle \eta_t, \omega_t\rangle + \frac{\gamma}{\ell}\sum_{t=0}^T \|F(x^t)\|^2 + \gamma^2 \sum_{t=0}^T \|\omega_t\|^2 \\
&\leq 4R^2 + 2\gamma \max_{u \in B_R(x^*)} \left\{\left\langle x^* - u, \sum_{t=0}^T \omega_t \right\rangle\right\} \\
&\quad + R^2 + 4\gamma \sum_{t=0}^T \langle \eta_t, \omega_t\rangle + 2\gamma^2 \sum_{t=0}^T \|\omega_t\|^2 \\
&\leq 5R^2 + 2\gamma R \left\|\sum_{t=0}^T \omega_t\right\| + 2\cdot(① + ② + ③ + ④ + ⑤), \qquad (293)
\end{aligned}$$

where we also aplly inequality $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ holding for all $a, b \in \mathbb{R}^d$ to upper bound $\|\omega_t\|^2$.

It remains to derive good enough high-probability upper-bounds for the terms ①, ②, ③, ④, ⑤ and $2\gamma R \left\|\sum_{t=0}^T \omega_t\right\|$, i.e., to finish our inductive proof we need to show that ① + ② + ③ + ④ + ⑤ $\leq R^2$ and $2\gamma R \left\|\sum_{t=0}^T \omega_t\right\| \leq 2R^2$ with high probability.In the subsequent parts of the proof, we will need to use many times the bounds for the norm and second moments of ω_t^u, ω_t^b. First, by Lemma C.1, we have with probability 1 that

$$\|\omega_t^u\| \leq 2\lambda \qquad (294)$$

for all $t = 0, 1, \ldots, T$. Moreover, due to Lemma C.1, we also have that E_T implies

$$\left\|\omega_t^b\right\| \leq \frac{2^\alpha \sigma^\alpha}{\lambda^{\alpha-1}}, \qquad (295)$$

$$\mathbb{E}_{\xi^t}\left[\left\|\omega_t^b\right\|^2\right] \leq 18\lambda^{2-\alpha}\sigma^\alpha, \qquad (296)$$

$$\mathbb{E}_{\xi^t}\left[\left\|\omega_t^u\right\|^2\right] \leq 18\lambda^{2-\alpha}\sigma^\alpha \qquad (297)$$

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

Upper bound for ①. By definition of ω_t^u, we have $\mathbb{E}_{\xi^t}[\omega_t^u] = 0$ and

$$\mathbb{E}_{\xi^t}\left[2\gamma\langle \eta_t, \omega_t^u\rangle\right] = 0.$$

Next, the sum ① has bounded with probability 1 term:

$$|2\gamma\langle \eta_t, \omega_t^u\rangle| \leq 2\gamma\|\eta_t\| \cdot \|\omega_t^u\| \overset{(290),(294)}{\leq} 12\gamma R\lambda \overset{(282)}{\leq} \frac{R^2}{5\ln\frac{6(K+1)}{\beta}} \overset{\text{def}}{=} c. \qquad (298)$$

Moreover, these summands also have bounded conditional variances $\sigma_t^2 \overset{\text{def}}{=} \mathbb{E}_{\xi^t}\left[4\gamma^2\langle \eta_t, \omega_t^u\rangle^2\right]$:

$$\sigma_t^2 \leq \mathbb{E}_{\xi^t}\left[4\gamma^2\|\eta_t\|^2 \cdot \|\omega_t^u\|^2\right] \overset{(290)}{\leq} 36\gamma^2 R^2 \mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]. \qquad (299)$$

In other words, we showed that $\{2\gamma\langle \eta_t, \omega_t^u\rangle\}_{t\geq 0}$ is a bounded martingale difference sequence with bounded conditional variances $\{\sigma_t^2\}_{t\geq 0}$. Next, we apply Bernstein's inequality (Lemma B.2) with $X_t = 2\gamma\langle \eta_t, \omega_t^u\rangle$, parameter c as in (298), $b = \frac{R^2}{5}$, $G = \frac{R^4}{150\ln\frac{6(K+1)}{\beta}}$:

$$\mathbb{P}\left\{|①| > \frac{R^2}{5} \text{ and } \sum_{t=0}^T \sigma_t^2 \leq \frac{R^4}{150\ln\frac{6(K+1)}{\beta}}\right\} \leq 2\exp\left(-\frac{b^2}{2G + 2cb/3}\right) = \frac{\beta}{3(K+1)}.$$

Equivalently, we have

$$\mathbb{P}\{E_{①}\} \geq 1 - \tfrac{\beta}{3(K+1)}, \text{ for } E_{①} = \left\{\text{either} \quad \sum_{t=0}^{T} \sigma_t^2 > \frac{R^4}{150 \ln \frac{6(K+1)}{\beta}} \quad \text{or} \quad |①| \leq \frac{R^2}{5}\right\}. \tag{300}$$

In addition, E_T implies that

$$\begin{aligned}
\sum_{t=0}^{T} \sigma_t^2 &\overset{(299)}{\leq} 36\gamma^2 R^2 \sum_{t=0}^{T} \mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right] \\
&\overset{(297), T \leq K+1}{\leq} 648\gamma^2 R^2 \sigma^\alpha (K+1) \lambda^{2-\alpha} \\
&\overset{(282)}{\leq} 648\gamma^\alpha R^{4-\alpha} \sigma^\alpha (K+1) \ln^{\alpha-2} \frac{6(K+1)}{\beta} \\
&\overset{(281)}{\leq} \frac{R^4}{150 \ln \frac{6(K+1)}{\beta}}.
\end{aligned} \tag{301}$$

Upper bound for ②. From E_T it follows that

$$\begin{aligned}
② &\leq 2\gamma \sum_{t=0}^{T} \|\eta_l\| \cdot \|\omega_t^b\| \overset{(290),(295), T \leq K+1}{\leq} 6 \cdot 2^\alpha \gamma R (K+1) \frac{\sigma^\alpha}{\lambda^{\alpha-1}} \\
&\overset{(282)}{=} 12 \cdot 120^{\alpha-1} \gamma^\alpha \sigma^\alpha R^{2-\alpha} (K+1) \ln^{\alpha-1} \frac{6(K+1)}{\beta} \overset{(281)}{\leq} \frac{R^2}{5}.
\end{aligned} \tag{302}$$

Upper bound for ③. From E_T it follows that

$$\begin{aligned}
③ &= 2\gamma^2 \sum_{t=0}^{T} \mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right] \overset{(297), T \leq K+1}{\leq} 36\gamma^2 \lambda^{2-\alpha} \sigma^\alpha (K+1) \\
&\overset{(282)}{\leq} 36\gamma^\alpha R^{2-\alpha} \sigma^\alpha (K+1) \ln^{\alpha-2} \frac{6(K+1)}{\beta} \overset{(281)}{\leq} \frac{R^2}{5}.
\end{aligned} \tag{303}$$

Upper bound for ④. First, we have

$$2\gamma^2 \mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]\right] = 0.$$

Next, the sum ④ has bounded with probability 1 terms:

$$\begin{aligned}
2\gamma^2 \left|\|\omega_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]\right| &\leq 2\gamma^2 \left(\|\omega_t^u\|^2 + \mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]\right) \overset{(294)}{\leq} 16\gamma^2\lambda^2 \\
&\overset{(282)}{\leq} \frac{R^2}{225 \ln \frac{6(K+1)}{\beta}} \leq \frac{R^2}{5 \ln \frac{6(K+1)}{\beta}} \overset{\text{def}}{=} c.
\end{aligned} \tag{304}$$

The summands also have conditional variances $\widetilde{\sigma}_t^2 \overset{\text{def}}{=} 4\gamma^4 \mathbb{E}_{\xi^t}\left[\left(\|\omega_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]\right)^2\right]$ that are bounded

$$\widetilde{\sigma}_t^2 \overset{(304)}{\leq} \frac{2\gamma^2 R^2}{225 \ln \frac{6(K+1)}{\beta}} \mathbb{E}_{\xi^t}\left[\left|\|\omega_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]\right|\right] \leq \frac{4\gamma^2 R^2}{225 \ln \frac{6(K+1)}{\beta}} \mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]. \tag{305}$$

In other words, we showed that $\{\|\omega_t^u\|^2 - \mathbb{E}_{\xi^t}[\|\omega_t^u\|^2]\}_{t\geq 0}$ is a bounded martingale difference sequence with bounded conditional variances $\{\widetilde{\sigma}_t^2\}_{t\geq 0}$.Next, we apply Bernstein's inequality (Lemma B.2) with $X_t = \|\omega_t^u\|^2 - \mathbb{E}_{\xi^t}[\|\omega_t^u\|^2]$, parameter c as in (304), $b = \frac{R^2}{5}$, $G = \frac{R^4}{150 \ln \frac{6(K+1)}{\beta}}$:

$$\mathbb{P}\left\{|④| > \frac{R^2}{5} \text{ and } \sum_{t=0}^{T} \widetilde{\sigma}_t^2 \leq \frac{R^4}{150 \ln \frac{6(K+1)}{\beta}}\right\} \leq 2 \exp\left(-\frac{b^2}{2G + {}^{2cb}/_{3}}\right) = \frac{\beta}{3(K+1)}.$$

Equivalently, we have

$$\mathbb{P}\{E_{④}\} \geq 1 - \frac{\beta}{3(K+1)}, \text{ for } E_{④} = \left\{\text{either} \quad \sum_{t=0}^{T} \widetilde{\sigma}_t^2 > \frac{R^4}{150 \ln \frac{6(K+1)}{\beta}} \quad \text{or} \quad |④| \leq \frac{R^2}{5}\right\}. \tag{306}$$

In addition, E_T implies that

$$\begin{aligned}\sum_{t=0}^{T} \widetilde{\sigma}_t^2 &\overset{(305)}{\leq} \frac{4\gamma^2 R^2}{225 \ln \frac{6(K+1)}{\beta}} \sum_{t=0}^{T} \mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right] \overset{(297), T\leq K+1}{\leq} \frac{8\gamma^2 R^2 (K+1)}{25 \ln \frac{6(K+1)}{\beta}} \lambda^{2-\alpha}\sigma^\alpha \\ &\overset{(282)}{\leq} \frac{8}{25}\gamma^\alpha R^{4-\alpha}(K+1)\sigma^\alpha \ln^{\alpha-3}\frac{6(K+1)}{\beta} \\ &\overset{(281)}{\leq} \frac{R^4}{150 \ln \frac{6(K+1)}{\beta}}.\end{aligned} \tag{307}$$

Upper bound for ⑤. From E_T it follows that

$$\begin{aligned}⑤ &= 2\gamma^2 \sum_{t=0}^{T} \|\omega_t^b\|^2 \overset{(295), T\leq K+1}{\leq} 2^{2\alpha+1}\cdot 60^{2\alpha-2}\gamma^2 (K+1)\frac{\sigma^{2\alpha}}{\lambda^{2\alpha-2}} \\ &\overset{(282)}{=} 2^{2\alpha+1}\cdot 60^{2\alpha-2}\gamma^{2\alpha}(K+1)\frac{\sigma^{2\alpha}}{R^{2\alpha-2}} \ln^{2\alpha-2}\frac{6(K+1)}{\beta} \\ &\overset{(281)}{\leq} \frac{R^2}{5}.\end{aligned} \tag{308}$$

Upper bound for $\gamma \left\|\sum_{t=0}^{T}\omega_t\right\|$. To estimate this term from above, we consider a new vector:

$$\zeta_l = \begin{cases}\gamma \sum\limits_{r=0}^{l-1}\omega_r, & \text{if } \left\|\gamma \sum\limits_{r=0}^{l-1}\omega_r\right\| \leq R, \\ 0, & \text{otherwise}\end{cases}$$

for $l = 1, 2, \ldots, T-1$.This vector is bounded almost surely:

$$\|\zeta_l\| \leq R. \tag{309}$$

Thus, by (286), probability event E_T implies

$$\begin{aligned}\gamma\left\|\sum_{l=0}^{T}\omega_l\right\| &= \sqrt{\gamma^2\left\|\sum_{l=0}^{T}\omega_l\right\|^2} \\ &= \sqrt{\gamma^2\sum_{l=0}^{T}\|\omega_l\|^2 + 2\gamma\sum_{l=0}^{T}\left\langle \gamma\sum_{r=0}^{l-1}\omega_r, \omega_l\right\rangle} \\ &= \sqrt{\gamma^2\sum_{l=0}^{T}\|\omega_l\|^2 + 2\gamma\sum_{l=0}^{T}\langle\zeta_l, \omega_l\rangle} \\ &\overset{(292)}{\leq} \sqrt{③ + ④ + ⑤ + \underbrace{2\gamma\sum_{l=0}^{T}\langle\zeta_l, \omega_l^u\rangle}_{⑥} + \underbrace{2\gamma\sum_{l=0}^{T}\langle\zeta_l, \omega_l^b\rangle}_{⑦}}.\end{aligned} \tag{310}$$

Following similar steps as before, we bound ⑥ and ⑦.

Upper bound for ⑥. By definition of ω_u^t, we have $\mathbb{E}_{\xi^t}[\omega_t^u] = 0$ and

$$\mathbb{E}_{\xi^t}\left[2\gamma\langle\zeta_t, \omega_t^u\rangle\right] = 0.$$

Next, sum ⑥ has bounded with probability 1 terms:

$$|2\gamma\langle\zeta_t, \omega_t^u\rangle| \leq 2\gamma\|\zeta_t\| \cdot \|\omega_t^u\| \overset{(309),(294)}{\leq} 4\gamma R\lambda \overset{(282)}{\leq} \frac{R^2}{5\ln\frac{6(K+1)}{\beta}} \overset{\text{def}}{=} c. \tag{311}$$

The summands also have bounded conditional variances $\widehat{\sigma}_t^2 \overset{\text{def}}{=} \mathbb{E}_{\xi^t}\left[4\gamma^2\langle\zeta_t, \omega_t^u\rangle^2\right]$:

$$\widehat{\sigma}_t^2 \leq \mathbb{E}_{\xi^t}\left[4\gamma^2\|\zeta_t\|^2 \cdot \|\omega_t^u\|^2\right] \overset{(309)}{\leq} 4\gamma^2 R^2 \mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]. \tag{312}$$

In other words, we showed that $\{2\gamma\langle\zeta_t, \omega_t^u\rangle\}_{t\geq 0}$ is a bounded martingale difference sequence with bounded conditional variances $\{\widehat{\sigma}_t^2\}_{t\geq 0}$. Applying Bernstein's inequality (Lemma B.2) with $X_t = 2\gamma\langle\zeta_t, \omega_t^u\rangle$, parameter c as in (311), $b = \frac{R^2}{5}$, $G = \frac{R^4}{150\ln\frac{6(K+1)}{\beta}}$:

$$\mathbb{P}\left\{|⑥| > \frac{R^2}{5} \text{ and } \sum_{t=0}^{T}\widehat{\sigma}_t^2 \leq \frac{R^4}{150\ln\frac{6(K+1)}{\beta}}\right\} \leq 2\exp\left(-\frac{b^2}{2G + \nicefrac{2cb}{3}}\right) = \frac{\beta}{3(K+1)}.$$

Equivalently, we have

$$\mathbb{P}\{E_{⑥}\} \geq 1 - \frac{\beta}{3(K+1)} \text{ for } E_{⑥} = \left\{\text{either} \quad \sum_{t=0}^{T}\widehat{\sigma}_t^2 > \frac{R^4}{150\ln\frac{6(K+1)}{\beta}} \quad \text{or} \quad |⑥| \leq \frac{R^2}{5}\right\}. \tag{313}$$

In addition, E_T implies that

$$\begin{aligned}
\sum_{t=0}^{T}\widehat{\sigma}_t^2 &\overset{(312)}{\leq} 4\gamma^2 R^2 \sum_{t=0}^{T}\mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right] \\
&\overset{(297),T\leq K+1}{\leq} 72\gamma^2 R^2\sigma^\alpha(K+1)\lambda^{2-\alpha} \\
&\overset{(282)}{\leq} 72\gamma^\alpha R^{4-\alpha}\sigma^\alpha(K+1)\ln^{\alpha-2}\frac{6(K+1)}{\beta} \\
&\overset{(281)}{\leq} \frac{R^4}{150\ln\frac{6(K+1)}{\beta}}.
\end{aligned} \tag{314}$$

Upper bound for ⑦. From E_T it follows that

$$\begin{aligned}
⑦ &\leq 2\gamma\sum_{t=0}^{T}\|\zeta_t\| \cdot \|\omega_t^b\| \overset{(309),(295),T\leq K+1}{\leq} 8 \cdot 2^\alpha\gamma R(K+1)\frac{\sigma^\alpha}{\lambda^{\alpha-1}} \\
&\overset{(282)}{=} 16 \cdot 120^{\alpha-1}\gamma^\alpha\sigma^\alpha R^{2-\alpha}(K+1)\ln^{\alpha-1}\frac{6(K+1)}{\beta} \overset{(281)}{\leq} \frac{R^2}{5}.
\end{aligned} \tag{315}$$

Now, we have the upper bounds for ①, ②, ③, ④, ⑤, ⑥, ⑦. In particular, probability event E_{T-1} implies

$$R_{T+1}^2 \overset{(292)}{\leq} R^2 + ① + ② + ③ + ④ + ⑤,$$

$$2\gamma(T+1)\texttt{Gap}_R(x_{\text{avg}}^T) \overset{(293)}{\leq} 5R^2 + 2\gamma R \left\| \sum_{t=0}^{T} \omega_t \right\| + 2\cdot (① + ② + ③ + ④ + ⑤),$$

$$\gamma \left\| \sum_{l=0}^{T} \omega_l \right\| \overset{(310)}{\leq} \sqrt{③ + ④ + ⑤ + ⑥ + ⑦},$$

$$② \overset{(302)}{\leq} \frac{R^2}{5}, \quad ③ \overset{(303)}{\leq} \frac{R^2}{5}, \quad ⑤ \overset{(308)}{\leq} \frac{R^2}{5}, \quad ⑦ \overset{(315)}{\leq} \frac{R^2}{5},$$

$$\sum_{t=0}^{T} \sigma_t^2 \overset{(301)}{\leq} \frac{R^4}{150 \ln \frac{6(K+1)}{\beta}}, \quad \sum_{t=0}^{T} \widetilde{\sigma}_t^2 \overset{(307)}{\leq} \frac{R^4}{150 \ln \frac{6(K+1)}{\beta}}, \quad \sum_{t=0}^{T} \widehat{\sigma}_t^2 \overset{(314)}{\leq} \frac{R^4}{150 \ln \frac{6(K+1)}{\beta}}.$$

Moreover, we also have (see (300), (306), (315) and our induction assumption)

$$\mathbb{P}\{E_T\} \geq 1 - \frac{T\beta}{K+1},$$

$$\mathbb{P}\{E_{①}\} \geq 1 - \frac{\beta}{3(K+1)}, \quad \mathbb{P}\{E_{④}\} \geq 1 - \frac{\beta}{3(K+1)}, \quad \mathbb{P}\{E_{⑥}\} \geq 1 - \frac{\beta}{3(K+1)},$$

where

$$E_{①} = \left\{ \text{either} \quad \sum_{t=0}^{T} \sigma_t^2 > \frac{R^4}{150 \ln \frac{6(K+1)}{\beta}} \quad \text{or} \quad |①| \leq \frac{R^2}{5} \right\},$$

$$E_{④} = \left\{ \text{either} \quad \sum_{t=0}^{T} \widetilde{\sigma}_t^2 > \frac{R^4}{150 \ln \frac{6(K+1)}{\beta}} \quad \text{or} \quad |④| \leq \frac{R^2}{5} \right\},$$

$$E_{⑥} = \left\{ \text{either} \quad \sum_{t=0}^{T} \widehat{\sigma}_t^2 > \frac{R^4}{150 \ln \frac{6(K+1)}{\beta}} \quad \text{or} \quad |⑥| \leq \frac{R^2}{5} \right\}.$$

Thus, probability event $E_T \cap E_{①} \cap E_{④} \cap E_{⑥}$ implies

$$\begin{aligned} R_{T+1}^2 &\leq R^2 + ① + ② + ③ + ④ + ⑤ \leq 2R^2, \\ \gamma \left\| \sum_{l=0}^{T} \omega_l \right\| &\leq \sqrt{③ + ④ + ⑤ + ⑥ + ⑦} \leq R, \\ 2\gamma(T+1)\texttt{Gap}_R(x_{\text{avg}}^T) &\leq 6R^2 + 2\gamma R \left\| \sum_{t=0}^{T} \omega_t \right\| + 2\cdot (① + ② + ③ + ④ + ⑤) \\ &\leq 10R^2, \end{aligned}$$

which gives (286) for $t = T$, and

$$\mathbb{P}\{E_{T+1}\} \geq \mathbb{P}\{E_T \cap E_{①} \cap E_{④} \cap E_{⑥}\} = 1 - \mathbb{P}\{\overline{E}_T \cup \overline{E}_{①} \cup \overline{E}_{④} \cup \overline{E}_{⑥}\} \geq 1 - \frac{T\beta}{K+1}.$$

This finishes the inductive part of our proof, i.e., for all $k = 0, 1, \ldots, K+1$ we have $\mathbb{P}\{E_k\} \geq 1 - {k\beta}/{(K+1)}$. In particular, for $k = K+1$ we have that with probability at least $1-\beta$

$$\texttt{Gap}_R(x_{\text{avg}}^K) \leq \frac{5R^2}{\gamma(K+1)}.$$

Finally, if

$$\gamma = \min \left\{ \frac{1}{170\ell \ln \frac{6(K+1)}{\beta}}, \frac{R}{97200^{\frac{1}{\alpha}} (K+1)^{\frac{1}{\alpha}} \sigma \ln^{\frac{\alpha-1}{\alpha}} \frac{6(K+1)}{\beta}} \right\}$$

then with probability at least $1-\beta$

$$\begin{aligned}\texttt{Gap}_R(\widetilde{x}^K_{\texttt{avg}}) &\leq \frac{5R^2}{\gamma(K+1)} = \max\left\{\frac{800LR^2\ln\frac{6(K+1)}{\beta}}{K+1}, \frac{5\cdot 97200^{\frac{1}{\alpha}}\sigma R\ln^{\frac{\alpha-1}{\alpha}}\frac{6(K+1)}{\beta}}{(K+1)^{\frac{\alpha-1}{\alpha}}}\right\}\\ &= \mathcal{O}\left(\max\left\{\frac{\ell R^2\ln\frac{K}{\beta}}{K}, \frac{\sigma R\ln^{\frac{\alpha-1}{\alpha}}\frac{K}{\beta}}{K^{\frac{\alpha-1}{\alpha}}}\right\}\right).\end{aligned}$$

To get $\texttt{Gap}_R(\widetilde{x}^K_{\texttt{avg}}) \leq \varepsilon$ with probability at least $1-\beta$ it is sufficient to choose K such that both terms in the maximum above are $\mathcal{O}(\varepsilon)$. This leads to

$$K = \mathcal{O}\left(\frac{\ell R^2}{\varepsilon}\ln\frac{\ell R^2}{\varepsilon\beta}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\ln\left(\frac{1}{\beta}\left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right)$$

that concludes the proof. □

H.2. Star-Cocoercive Problems

Theorem H.4 (Case 2 in Theorem 4.2)**.** *Let Assumptions 1.1, 1.10 hold for $Q = B_{2R}(x^*)$, where $R \geq \|x^0 - x^*\|$, and*

$$0 < \gamma \leq \min\left\{\frac{1}{170\ell\ln\frac{4(K+1)}{\beta}}, \frac{R}{97200^{\frac{1}{\alpha}}(K+1)^{\frac{1}{\alpha}}\sigma\ln^{\frac{\alpha-1}{\alpha}}\frac{4(K+1)}{\beta}}\right\}, \tag{316}$$

$$\lambda_k \equiv \lambda = \frac{R}{60\gamma\ln\frac{4(K+1)}{\beta}}, \tag{317}$$

for some $K \geq 0$ and $\beta \in (0,1]$ such that $\ln\frac{4(K+1)}{\beta} \geq 1$. Then, after K iterations the iterates produced by clipped-SGDA *with probability at least $1-\beta$ satisfy*

$$\frac{1}{K+1}\sum_{k=0}^{K}\|F(x^k)\|^2 \leq \frac{2\ell R^2}{\gamma(K+1)}. \tag{318}$$

In particular, when γ equals the minimum from (316), then the iterates produced by clipped-SGDA *after K iterations with probability at least $1-\beta$ satisfy*

$$\frac{1}{K+1}\sum_{k=0}^{K}\|F(x^k)\|^2 = \mathcal{O}\left(\max\left\{\frac{\ell^2R^2\ln\frac{K}{\beta}}{K+1}, \frac{\ell\sigma R\ln^{\frac{\alpha-1}{\alpha}}\frac{K}{\beta}}{K^{\frac{\alpha-1}{\alpha}}}\right\}\right), \tag{319}$$

meaning that to achieve $\frac{1}{K+1}\sum\limits_{k=0}^{K}\|F(x^k)\|^2 \leq \varepsilon$ with probability at least $1-\beta$ clipped-SGDA *requires*

$$K = \mathcal{O}\left(\frac{\ell^2R^2}{\varepsilon}\ln\frac{\ell^2R^2}{\varepsilon\beta}, \left(\frac{\ell\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\ln\left(\frac{1}{\beta}\left(\frac{\ell\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right) \quad \textit{iterations/oracle calls.} \tag{320}$$

Proof. Again, we will closely follow the proof of Theorem D.2 from (Gorbunov et al., 2022a) and the main difference will be reflected in the application of Bernstein inequality and estimating biases and variances of stochastic terms.

Let $R_k = \|x^k - x^*\|$ for all $k \geq 0$. As the previous result, the proof is based on on the induction argument and showing that the iterates do not leave some ball around the solution with high probability. More precisely, for each $k = 0, \ldots, K+1$ we define probability event E_k as follows: inequalities

$$\|x^t - x^*\|^2 \leq 2R^2, \tag{321}$$

hold for $t = 0, 1, \ldots, k$ simultaneously. We want to prove that $\mathbb{P}\{E_k\} \geq 1 - {k\beta}/{(K+1)}$ for all $k = 0, 1, \ldots, K+1$ by induction. One of the important things is that inequalities (287) and (292) are obtained without assuming monotonicity of F. Thus, if we do exactly the same steps as in the proof of Theorem H.3 (up to the replacement of $\ln \frac{6(K+1)}{\beta}$ by $\ln \frac{4(K+1)}{\beta}$), we gain that

$$R_{T+1}^2 \overset{(292)}{\leq} R^2 + ① + ② + ③ + ④ + ⑤,$$

$$② \overset{(302)}{\leq} \frac{R^2}{5}, \quad ③ \overset{(303)}{\leq} \frac{R^2}{5}, \quad ⑤ \overset{(308)}{\leq} \frac{R^2}{5},$$

$$\sum_{t=0}^{T} \sigma_t^2 \overset{(301)}{\leq} \frac{R^4}{150 \ln \frac{4(K+1)}{\beta}}, \quad \sum_{t=0}^{T} \widetilde{\sigma}_t^2 \overset{(307)}{\leq} \frac{R^4}{150 \ln \frac{4(K+1)}{\beta}}.$$

Moreover, we also have (see (300), (306) and our induction assumption)

$$\mathbb{P}\{E_T\} \geq 1 - \frac{T\beta}{K+1},$$

$$\mathbb{P}\{E_{①}\} \geq 1 - \frac{\beta}{2(K+1)}, \quad \mathbb{P}\{E_{④}\} \geq 1 - \frac{\beta}{2(K+1)},$$

where

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

$$E_{④} = \left\{ \text{either} \quad \sum_{t=0}^{T} \widetilde{\sigma}_t^2 > \frac{R^4}{150 \ln \frac{4(K+1)}{\beta}} \quad \text{or} \quad |④| \leq \frac{R^2}{5} \right\}.$$

Thus probability event $E_{T-1} \cap E_{①} \cap E_{④}$ implies

$$R_{T+1}^2 \leq R^2 + ① + ② + ③ + ④ + ⑤ \leq 2R^2,$$

and

$$\mathbb{P}\{E_{T+1}\} \geq \mathbb{P}\{E_T \cap E_{①} \cap E_{④}\} = 1 - \mathbb{P}\{\overline{E}_T \cup \overline{E}_{①} \cup \overline{E}_{④}\} \geq 1 - \frac{T\beta}{K+1}. \tag{322}$$

This finishes the inductive part of our proof, i.e. for all $k = 0, 1, \ldots, K+1$ we have $\mathbb{P}\{E_k\} \geq 1 - {k\beta}/{(K+1)}$. In particular, for $k = K+1$ we have that with probability at least $1 - \beta$

$$\frac{1}{K+1} \sum_{k=0}^{K} \|F(x^k)\|^2 \overset{(287)}{\leq} \frac{\ell(R^2 - R_{K+1}^2)}{\gamma(K+1)} + \frac{\ell(① + ② + ③ + ④ + ⑤)}{\gamma(K+1)} \leq \frac{2\ell R^2}{\gamma(K+1)}.$$

Finally, if

$$\gamma = \min\left\{ \frac{1}{170\ell \ln \frac{4(K+1)}{\beta}}, \frac{R}{97200^{\frac{1}{\alpha}} (K+1)^{\frac{1}{\alpha}} \sigma \ln^{\frac{\alpha-1}{\alpha}} \frac{4(K+1)}{\beta}} \right\}$$

then with probability at least $1 - \beta$

$$\frac{1}{K+1} \sum_{k=0}^{K} \|F(x^k)\|^2 \leq \frac{2\ell R^2}{\gamma(K+1)} = \max\left\{ \frac{340\ell^2 R^2 \ln \frac{4(K+1)}{\beta}}{K+1}, \frac{2 \cdot 97200^{\frac{1}{\alpha}} \ell\sigma R \ln^{\frac{\alpha-1}{\alpha}} \frac{4(K+1)}{\beta}}{(K+1)^{\frac{\alpha-1}{\alpha}}} \right\}$$

$$= \mathcal{O}\left(\max\left\{ \frac{\ell^2 R^2 \ln \frac{K}{\beta}}{K}, \frac{\ell\sigma R \ln^{\frac{\alpha-1}{\alpha}} \frac{K}{\beta}}{K^{\frac{\alpha-1}{\alpha}}} \right\} \right).$$

To get $\frac{1}{K+1}\sum\limits_{k=0}^{K}\|F(x^k)\|^2 \leq \varepsilon$ with probability at least $1-\beta$ it is sufficient to choose K such that both terms in the maximum above are $\mathcal{O}(\varepsilon)$. This leads to

$$K = \mathcal{O}\left(\frac{\ell^2R^2}{\varepsilon}\ln\frac{\ell^2R^2}{\varepsilon\beta}, \left(\frac{\ell\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\ln\left(\frac{1}{\beta}\left(\frac{\ell\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right)$$

that concludes the proof.

□

H.3. Quasi-Strongly Monotone Star-Cocoercive Problems

As in the monotone case, we use another lemma from (Gorbunov et al., 2022a) that handles the deterministic part of clipped-SGDA in the quasi-strongly monotone case.

Lemma H.5 (Lemma D.3 from (Gorbunov et al., 2022a)). *Let Assumptions 1.9, 1.10 hold for $Q = B_{2R}(x^*) = \{x \in \mathbb{R}^d \mid \|x - x^*\| \leq 2R\}$, where $R \geq \|x^0 - x^*\|$, and $0 < \gamma \leq 1/\ell$. If x^k lie in $B_{2R}(x^*)$ for all $k = 0, 1, \ldots, K$ for some $K \geq 0$, then the iterates produced by* clipped-SGDA *satisfy*

$$\|x^{K+1} - x^*\|^2 \leq (1-\gamma\mu)^{K+1}\|x^0 - x^*\|^2 + 2\gamma\sum_{k=0}^{K}(1-\gamma\mu)^{K-k}\langle x^k - x^* - \gamma F(x^k), \omega_k\rangle + \gamma^2\sum_{k=0}^{K}(1-\gamma\mu)^{K-k}\|\omega_k\|^2, \tag{323}$$

where ω_k are defined in (279).

Using this lemma we prove the main convergence result for clipped-SGDA in the quasi-strongly monotone case.

Theorem H.6 (Case 2 in Theorem 4.2). *Let Assumptions 1.1, 1.9, 1.10, hold for $Q = B_{2R}(x^*) = \{x \in \mathbb{R}^d \mid \|x - x^*\| \leq 2R\}$, where $R \geq \|x^0 - x^*\|$, and*

$$0 < \gamma \leq \min\left\{\frac{1}{400\ell\ln\frac{4(K+1)}{\beta}}, \frac{\ln(B_K)}{\mu(K+1)}\right\}, \tag{324}$$

$$B_K = \max\left\{2, \frac{(K+1)^{\frac{2(\alpha-1)}{\alpha}}\mu^2R^2}{5400^{\frac{2}{\alpha}}\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{4(K+1)}{\beta}\right)\ln^2(B_K)}\right\} \tag{325}$$

$$= \mathcal{O}\left(\max\left\{2, \frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2R^2}{\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)\ln^2\left(\max\left\{2, \frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2R^2}{\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)}\right\}\right)}\right\}\right), \tag{326}$$

$$\lambda_k = \frac{\exp(-\gamma\mu(1+k/2))R}{120\gamma\ln\frac{4(K+1)}{\beta}}, \tag{327}$$

for some $K \geq 0$ and $\beta \in (0,1]$ such that $\ln\frac{4(K+1)}{\beta} \geq 1$. Then, after K iterations the iterates produced by clipped-SGDA *with probability at least $1-\beta$ satisfy*

$$\|x^{K+1} - x^*\|^2 \leq 2\exp(-\gamma\mu(K+1))R^2. \tag{328}$$

In particular, when γ equals the minimum from (324), then the iterates produced by clipped-SGDA *after K iterations with probability at least $1-\beta$ satisfy*

$$\|x^K - x^*\|^2 = \mathcal{O}\left(\max\left\{R^2\exp\left(-\frac{\mu K}{\ell\ln\frac{K}{\beta}}\right), \frac{\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)\ln^2\left(\max\left\{2, \frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2R^2}{\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)}\right\}\right)}{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2}\right\}\right), \tag{329}$$

meaning that to achieve $\|x^K - x^*\|^2 \leq \varepsilon$ *with probability at least* $1-\beta$ clipped-SGDA *requires*

$$K = \mathcal{O}\left(\frac{\ell}{\mu}\ln\left(\frac{R^2}{\varepsilon}\right)\ln\left(\frac{\ell}{\mu\beta}\ln\frac{R^2}{\varepsilon}\right), \left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\ln\left(\frac{1}{\beta}\left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right)\ln^{\frac{\alpha}{\alpha-1}}\left(B_\varepsilon\right)\right) \tag{330}$$

iterations/oracle calls, where

$$B_\varepsilon = \max\left\{2, \frac{R^2}{\varepsilon\ln\left(\frac{1}{\beta}\left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right)}\right\}.$$

Proof. Again, we will closely follow the proof of Theorem D.3 from (Gorbunov et al., 2022a) and the main difference will be reflected in the application of Bernstein inequality and estimating biases and variances of stochastic terms.

Let $R_k = \|x^k - x^*\|$ for all $k \geq 0$. As in the previous results, the proof is based on the induction argument and showing that the iterates do not leave some ball around the solution with high probability. More precisely, for each $k = 0, 1, \ldots, K+1$ we consider probability event E_k as follows: inequalities

$$R_t^2 \leq 2\exp(-\gamma\mu t)R^2 \tag{331}$$

hold for $t = 0, 1, \ldots, k$ simultaneously. We want to prove $\mathbb{P}\{E_k\} \geq 1 - {k\beta}/{(K+1)}$ for all $k = 0, 1, \ldots, K+1$ by induction. The base of the induction is trivial: for $k = 0$ we have $R_0^2 \leq R^2 < 2R^2$ by definition. Next, assume that for $k = T-1 \leq K$ the statement holds: $\mathbb{P}\{E_{T-1}\} \geq 1 - {(T-1)\beta}/{(K+1)}$. Given this, we need to prove $\mathbb{P}\{E_T\} \geq 1 - {T\beta}/{(K+1)}$. Since $R_t^2 \leq 2\exp(-\gamma\mu t)R^2 \leq 2R^2$, we have $x^t \in B_{2R}(x^*)$, where operator F is ℓ-star-cocoersive. Thus, E_{T-1} implies

$$\|F(x^t)\| \quad \leq \quad \ell\|x^t - x^*\| \overset{(331)}{\leq} \sqrt{2}\ell\exp(-{\gamma\mu t}/{2})R \overset{(324),(327)}{\leq} \frac{\lambda_t}{2} \tag{332}$$

and

$$\|\omega_t\|^2 \quad \leq \quad 2\|\widetilde{F}_\xi(x^t)\|^2 + 2\|F(x^t)\|^2 \overset{(332)}{\leq} \frac{5}{2}\lambda_t^2 \overset{(327)}{\leq} \frac{\exp(-\gamma\mu t)R^2}{4\gamma^2} \tag{333}$$

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

Using Lemma H.5 and $(1-\gamma\mu)^T \leq \exp(-\gamma\mu T)$, we obtain that E_{T-1} implies

$$\begin{aligned} R_T^2 \quad &\leq \quad \exp(-\gamma\mu T)R^2 + 2\gamma\sum_{t=0}^{T-1}(1-\gamma\mu)^{T-1-t}\langle x^t - x^* - \gamma F(x^t), \omega_t\rangle \\ &\quad + \gamma^2\sum_{t=0}^{T-1}(1-\gamma\mu)^{T-1-t}\|\omega_t\|^2. \end{aligned}$$

To handle the sums above, we introduce a new notation:

$$\eta_t = \begin{cases} x^t - x^* - \gamma F(x^t), & \text{if } \|x^t - x^* - \gamma F(x^t)\| \leq \sqrt{2}(1+\gamma\ell)\exp(-{\gamma\mu t}/{2})R, \\ 0, & \text{otherwise,} \end{cases} \tag{334}$$

for $t = 0, 1, \ldots, T-1$. This vector is bounded almost surely:

$$\|\eta_t\| \leq \sqrt{2}(1+\gamma\ell)\exp(-{\gamma\mu t}/{2})R \tag{335}$$

for all $t = 0, 1, \ldots, T-1$. We also notice that E_{T-1} implies $\|F(x^t)\| \leq \sqrt{2}\ell\exp(-{\gamma\mu t}/{2})R$ (due to (332)) and

$$\begin{aligned} \|x^t - x^* - \gamma F(x^t)\| \quad &\leq \quad \|x^t - x^*\| + \gamma\|F(x^t)\| \\ &\overset{(332)}{\leq} \quad \sqrt{2}(1+\gamma\ell)\exp(-{\gamma\mu t}/{2})R \end{aligned}$$

for $t = 0, 1, \ldots, T-1$. In other words, E_{T-1} implies $\eta_t = x^t - x^* - \gamma F(x^t)$ for all $t = 0, 1, \ldots, T-1$, meaning that from E_{T-1} it follows that

$$R_T^2 \leq \exp(-\gamma\mu T)R^2 + 2\gamma\sum_{t=0}^{T-1}(1-\gamma\mu)^{T-1-t}\langle \eta_t, \omega_t\rangle + \gamma^2\sum_{t=0}^{T-1}(1-\gamma\mu)^{T-1-t}\|\omega_t\|^2.$$

To handle the sums appeared on the right-hand side of the previous inequality we consider unbiased and biased parts of ω_t:

$$\omega_t^u \overset{\text{def}}{=} \mathbb{E}_{\xi^t}\left[F_{\xi^t}(x^t)\right] - \widetilde{F}_{\xi^t}(x^t), \quad \omega_t^b \overset{\text{def}}{=} F(x^t) - \mathbb{E}_{\xi_1^t}\left[F_{\xi^t}(x^t)\right], \tag{336}$$

for all $t = 0, \ldots, T-1$. By definition we have $\omega_t = \omega_t^u + \omega_t^b$ for all $t = 0, \ldots, T-1$. Therefore, E_{T-1} implies

$$\begin{aligned} R_T^2 \leq\ & \exp(-\gamma\mu T)R^2 + \underbrace{2\gamma\sum_{t=0}^{T-1}(1-\gamma\mu)^{T-1-t}\langle \eta_t, \omega_t^u\rangle}_{①} \\ & + \underbrace{2\gamma\sum_{t=0}^{T-1}(1-\gamma\mu)^{T-1-t}\langle \eta_t, \omega_t^b\rangle}_{②} + \underbrace{2\gamma^2\sum_{t=0}^{T-1}(1-\gamma\mu)^{T-1-t}\mathbb{E}_{\xi}\left[\|\omega_t^u\|^2\right]}_{③} \\ & + \underbrace{2\gamma^2\sum_{t=0}^{T-1}(1-\gamma\mu)^{T-1-t}\left(\|\omega_t^u\|^2 - \mathbb{E}_{\xi}\left[\|\omega_t^u\|^2\right]\right)}_{④} + \underbrace{2\gamma^2\sum_{t=0}^{T-1}(1-\gamma\mu)^{T-1-t}\|\omega_t^b\|^2}_{⑤}. \end{aligned} \tag{337}$$

where we also apply inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ holding for all $a, b \in \mathbb{R}^d$ to upper bound $\|\omega_t\|^2$. It remains to derive good enough high-probability upper-bounds for the terms ①, ②, ③, ④, ⑤, i.e., to finish our inductive proof we need to show that ① + ② + ③ + ④ + ⑤ $\leq \exp(-\gamma\mu T)R^2$ with high probability. In the subsequent parts of the proof, we will need to use many times the bounds for the norm and second moments of ω_t^u and ω_t^b. First, by Lemma 5.1, we have with probability 1 that

$$\|\omega_t^u\| \leq 2\lambda_t. \tag{338}$$

Moreover, since E_{T-1} implies that $\|F(x^t)\| \leq {\lambda_t}/{2}$ and $\|F(x^t)\| \leq {\lambda_t}/{2}$ for all $t = 0, 1, \ldots, T-1$ (see (332)), from Lemma 5.1 we also have that E_{T-1} implies

$$\left\|\omega_t^b\right\| \leq \frac{2^\alpha\sigma^\alpha}{\lambda_t^{\alpha-1}}, \tag{339}$$

$$\mathbb{E}_{\xi^t}\left[\left\|\omega_t^b\right\|^2\right] \leq 18\lambda_t^{2-\alpha}\sigma^\alpha, \tag{340}$$

$$\mathbb{E}_{\xi^t}\left[\left\|\omega_t^u\right\|^2\right] \leq 18\lambda_t^{2-\alpha}\sigma^\alpha, \tag{341}$$

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

Upper bound for ①. By definition of ω_t^u, we have $\mathbb{E}_{\xi^t}[\omega_t^u] = 0$ and

$$\mathbb{E}_{\xi^t}\left[2\gamma(1-\gamma\mu)^{T-1-t}\langle \eta_t, \omega_t^u\rangle\right] = 0.$$

Next, sum ① has bounded with probability 1 terms:

$$\begin{aligned} \left|2\gamma(1-\gamma\mu)^{T-1-t}\langle \eta_t, \omega_t^u\rangle\right| &\leq 2\gamma\exp(-\gamma\mu(T-1-t))\|\eta_t\|\cdot\|\omega_t^u\| \\ &\overset{(335),(338)}{\leq} 4\sqrt{2}\gamma(1+\gamma\ell)\exp(-\gamma\mu(T-1-{t}/{2}))R\lambda_t \\ &\overset{(324),(327)}{\leq} \frac{\exp(-\gamma\mu T)R^2}{5\ln\frac{4(K+1)}{\beta}} \overset{\text{def}}{=} c. \end{aligned} \tag{342}$$

The summands also have bounded conditional variances $\sigma_t^2 \stackrel{\text{def}}{=} \mathbb{E}_{\xi^t}\left[4\gamma^2(1-\gamma\mu)^{2T-2-2t}\langle \eta_t, \omega_t^u\rangle^2\right]$:

$$\begin{aligned}
\sigma_t^2 &\leq \mathbb{E}_{\xi^t}\left[4\gamma^2\exp(-\gamma\mu(2T-2-2t))\|\eta_t\|^2\cdot\|\omega_t^u\|^2\right] \\
&\stackrel{(335)}{\leq} 8\gamma^2(1+\gamma\ell)^2\exp(-\gamma\mu(2T-2-t))R^2\mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right] \\
&\stackrel{(324)}{\leq} 10\gamma^2\exp(-\gamma\mu(2T-t))R^2\mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]. \qquad (343)
\end{aligned}$$

In other words, we showed that $\{2\gamma(1-\gamma\mu)^{T-1-t}\langle\eta_t,\omega_t^u\rangle\}_{t=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\sigma_t^2\}_{t=0}^{T-1}$. Next, we apply Bernstein's inequality (Lemma B.2) with $X_t = 2\gamma(1-\gamma\mu)^{T-1-t}\langle\eta_t,\omega_t^u\rangle$, parameter c as in (342), $b = \frac{1}{5}\exp(-\gamma\mu T)R^2$, $F = \frac{\exp(-2\gamma\mu T)R^4}{300\ln\frac{4(K+1)}{\beta}}$:

$$\mathbb{P}\left\{|①| > \frac{1}{5}\exp(-\gamma\mu T)R^2 \text{ and } \sum_{t=0}^{T-1}\sigma_t^2 \leq \frac{\exp(-2\gamma\mu T)R^4}{300\ln\frac{4(K+1)}{\beta}}\right\} \leq 2\exp\left(-\frac{b^2}{2F+{}^{2cb}/_{3}}\right) = \frac{\beta}{2(K+1)}.$$

Equivalently, we have

$$\mathbb{P}\{E_{①}\} \geq 1-\frac{\beta}{2(K+1)}, \quad \text{for} \quad E_{①} = \left\{\text{either} \quad \sum_{t=0}^{T-1}\sigma_t^2 > \frac{\exp(-2\gamma\mu T)R^4}{300\ln\frac{4(K+1)}{\beta}} \quad \text{or} \quad |①| \leq \frac{1}{5}\exp(-\gamma\mu T)R^2\right\}. \qquad (344)$$

In addition, E_{T-1} implies that

$$\begin{aligned}
\sum_{t=0}^{T-1}\sigma_t^2 &\stackrel{(343)}{\leq} 10\gamma^2\exp(-2\gamma\mu T)R^2\sum_{t=0}^{T-1}\frac{\mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]}{\exp(-\gamma\mu t)} \\
&\stackrel{(341),T\leq K+1}{\leq} 180\gamma^2\exp(-2\gamma\mu T)R^2\sigma^\alpha\sum_{t=0}^{K}\frac{\lambda_t^{2-\alpha}}{\exp(-\gamma\mu t)} \\
&\stackrel{(327)}{\leq} \frac{180\gamma^\alpha\exp(-2\gamma\mu T)R^{4-\alpha}\sigma^\alpha(K+1)\exp(\frac{\gamma\mu\alpha K}{2})}{120^{2-\alpha}\ln^{2-\alpha}\frac{4(K+1)}{\beta}} \\
&\stackrel{(324)}{\leq} \frac{\exp(-2\gamma\mu T)R^4}{300\ln\frac{4(K+1)}{\beta}}. \qquad (345)
\end{aligned}$$

Upper bound for ②. From E_{T-1} it follows that

$$\begin{aligned}
② &\leq 2\gamma\exp(-\gamma\mu(T-1))\sum_{t=0}^{T-1}\frac{\|\eta_t\|\cdot\|\omega_t^b\|}{\exp(-\gamma\mu t)} \\
&\stackrel{(335),(339)}{\leq} 2^{1+\alpha}\sqrt{2}\gamma(1+\gamma\ell)\exp(-\gamma\mu(T-1))R\sigma^\alpha\sum_{t=0}^{T-1}\frac{1}{\lambda_t^{\alpha-1}\exp(-{}^{\gamma\mu t}/_{2})} \\
&\stackrel{(327)}{\leq} \frac{2^{3+\alpha}\cdot 120^{\alpha-1}\sqrt{2}\gamma^\alpha(1+\gamma\ell)\exp(-\gamma\mu T)\sigma^\alpha(K+1)\exp\left(\frac{\gamma\mu\alpha T}{2}\right)\ln^{\alpha-1}\frac{4(K+1)}{\beta}}{R^{\alpha-2}} \\
&\stackrel{(324)}{\leq} \frac{1}{5}\exp(-\gamma\mu T)R^2. \qquad (346)
\end{aligned}$$

Upper bound for ③. From E_{T-1} it follows that

$$\begin{aligned} ③ &= 2\gamma^2 \exp(-\gamma\mu(T-1)) \sum_{t=0}^{T-1} \frac{\mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]}{\exp(-\gamma\mu t)} \\ &\overset{(341)}{\leq} 144\gamma^2 \exp(-\gamma\mu(T-1))\sigma^\alpha \sum_{t=0}^{T-1} \frac{\lambda_t^{2-\alpha}}{\exp(-\gamma\mu t)} \\ &\overset{(327)}{\leq} \frac{144\gamma^\alpha R^{2-\alpha} \exp(-\gamma\mu(T-1))\sigma^\alpha (K+1)\exp(\frac{\gamma\mu\alpha K}{2})}{120^{2-\alpha}\ln^{2-\alpha}\frac{4(K+1)}{\beta}} \\ &\overset{(324)}{\leq} \frac{1}{5}\exp(-\gamma\mu T)R^2. \end{aligned} \tag{347}$$

Upper bound for ④. First, we have

$$2\gamma^2(1-\gamma\mu)^{T-1-t}\mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]\right] = 0.$$

Next, sum ④ has bounded with probability 1 terms:

$$\begin{aligned} 2\gamma^2(1-\gamma\mu)^{T-1-t}\left|\|\omega_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]\right| &\overset{(338)}{\leq} \frac{16\gamma^2\exp(-\gamma\mu T)\lambda_l^2}{\exp(-\gamma\mu(1+t))} \\ &\overset{(327)}{\leq} \frac{\exp(-\gamma\mu T)R^2}{5\ln\frac{4(K+1)}{\beta}} \\ &\overset{\text{def}}{=} c. \end{aligned} \tag{348}$$

The summands also have conditional variances

$$\widetilde{\sigma}_t^2 \overset{\text{def}}{=} \mathbb{E}_{\xi^t}\left[4\gamma^4(1-\gamma\mu)^{2T-2-2t}\left|\|\omega_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]\right|^2\right]$$

that are bounded

$$\begin{aligned} \widetilde{\sigma}_t^2 &\overset{(348)}{\leq} \frac{2\gamma^2\exp(-2\gamma\mu T)R^2}{5\exp(-\gamma\mu(1+t))\ln\frac{4(K+1)}{\beta}}\mathbb{E}_{\xi^t}\left[\left|\|\omega_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]\right|\right] \\ &\leq \frac{4\gamma^2\exp(-2\gamma\mu T)R^2}{5\exp(-\gamma\mu(1+t))\ln\frac{4(K+1)}{\beta}}\mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]. \end{aligned} \tag{349}$$

In other words, we showed that $\left\{2\gamma^2(1-\gamma\mu)^{T-1-t}\left(\|\omega_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]\right)\right\}_{t=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\widetilde{\sigma}_t^2\}_{t=0}^{T-1}$. Next, we apply Bernstein's inequality (Lemma B.2) with $X_t = 2\gamma^2(1-\gamma\mu)^{T-1-t}\left(\|\omega_t^u\|^2 - \mathbb{E}_{\xi^t}\left[\|\omega_t^u\|^2\right]\right)$, parameter c as in (348), $b = \frac{1}{5}\exp(-\gamma\mu T)R^2$, $G = \frac{\exp(-2\gamma\mu T)R^4}{300\ln\frac{4(K+1)}{\beta}}$:

$$\mathbb{P}\left\{|④| > \frac{1}{5}\exp(-\gamma\mu T)R^2 \text{ and } \sum_{l=0}^{T-1}\widetilde{\sigma}_t^2 \leq \frac{\exp(-2\gamma\mu T)R^4}{294\ln\frac{4(K+1)}{\beta}}\right\} \leq 2\exp\left(-\frac{b^2}{2G + {}^{2cb}/_3}\right) = \frac{\beta}{2(K+1)}.$$

Equivalently, we have

$$\mathbb{P}\{E_④\} \geq 1 - \frac{\beta}{2(K+1)}, \quad \text{for} \quad E_④ = \left\{\text{either} \quad \sum_{t=0}^{T-1}\widetilde{\sigma}_t^2 > \frac{\exp(-2\gamma\mu T)R^4}{300\ln\frac{4(K+1)}{\beta}} \quad \text{or} \quad |④| \leq \frac{1}{5}\exp(-\gamma\mu T)R^2\right\}. \tag{350}$$

In addition, E_{T-1} implies that

$$\begin{aligned}
\sum_{l=0}^{T-1} \widetilde{\sigma}_t^2 &\overset{(349)}{\leq} \frac{4\gamma^2 \exp(-\gamma\mu(2T-1))R^2}{5\ln\frac{4(K+1)}{\beta}} \sum_{t=0}^{T-1} \frac{\mathbb{E}_{\xi^t}\left[\|\omega_l^u\|^2\right]}{\exp(-\gamma\mu t)} \\
&\overset{(341),T\leq K+1}{\leq} \frac{72\gamma^2 \exp(-\gamma\mu(2T-1))R^2\sigma^\alpha}{5\ln\frac{4(K+1)}{\beta}} \sum_{t=0}^{K} \frac{\lambda_t^{2-\alpha}}{\exp(-\gamma\mu t)} \\
&\overset{(327)}{\leq} \frac{72\gamma^\alpha \exp(-\gamma\mu(2T-1))R^{4-\alpha}\sigma^\alpha (K+1)\exp(\frac{\gamma\mu\alpha K}{2})}{5\cdot 120^{2-\alpha}\ln^{3-\alpha}\frac{4(K+1)}{\beta}} \\
&\overset{(324)}{\leq} \frac{\exp(-2\gamma\mu T)R^4}{300\ln\frac{4(K+1)}{\beta}}. \qquad (351)
\end{aligned}$$

Upper bound for ⑤. From E_{T-1} it follows that

$$\begin{aligned}
⑤ &= 2\gamma^2 \sum_{l=0}^{T-1} \exp(-\gamma\mu(T-1-t))\left(\|\omega_t^b\|^2\right) \\
&\overset{(339)}{\leq} 2\cdot 2^{2\alpha}\gamma^2 \exp(-\gamma\mu(T-1))\sigma^{2\alpha} \sum_{t=0}^{T-1} \frac{1}{\lambda_t^{2\alpha-2}\exp(-\gamma\mu t)} \\
&\overset{(327),T\leq K+1}{\leq} \frac{2\cdot 2^{2\alpha}\cdot 120^{2\alpha-2}\gamma^{2\alpha}\exp(-\gamma\mu(T-1))\sigma^{2\alpha}\ln^{2\alpha-2}\frac{4(K+1)}{\beta}}{R^{2\alpha-2}} \sum_{t=0}^{K} \exp\left(\gamma\mu(2\alpha-2)\left(1+\frac{t}{2}\right)\right)\exp(\gamma\mu t) \\
&\leq \frac{4\cdot 2^{2\alpha}\cdot 120^{2\alpha-2}\gamma^{2\alpha}\exp(-\gamma\mu(T-3))\sigma^{2\alpha}\ln^{2\alpha-2}\frac{4(K+1)}{\beta}}{R^{2\alpha-2}} \sum_{t=0}^{K} \exp(\gamma\mu\alpha t) \\
&\leq \frac{4\cdot 2^{2\alpha}\cdot 120^{2\alpha-2}\gamma^{2\alpha}\exp(-\gamma\mu(T-3))\sigma^{2\alpha}\ln^{2\alpha-2}\frac{4(K+1)}{\beta}(K+1)\exp(\gamma\mu\alpha K)}{R^{2\alpha-2}} \\
&\overset{(324)}{\leq} \frac{1}{5}\exp(-\gamma\mu T)R^2. \qquad (352)
\end{aligned}$$

Now, we have the upper bounds for ①, ②, ③, ④, ⑤. In particular, probability event E_{T-1} implies

$$R_T^2 \overset{(337)}{\leq} \exp(-\gamma\mu T)R^2 + ① + ② + ③ + ④ + ⑤,$$
$$② \overset{(346)}{\leq} \frac{1}{5}\exp(-\gamma\mu T)R^2, \quad ③ \overset{(347)}{\leq} \frac{1}{5}\exp(-\gamma\mu T)R^2,$$
$$⑤ \overset{(352)}{\leq} \frac{1}{5}\exp(-\gamma\mu T)R^2$$
$$\sum_{t=0}^{T-1}\sigma_t^2 \overset{(345)}{\leq} \frac{\exp(-2\gamma\mu T)R^4}{300\ln\frac{4(K+1)}{\beta}}, \quad \sum_{t=0}^{T-1}\widetilde{\sigma}_t^2 \overset{(351)}{\leq} \frac{\exp(-2\gamma\mu T)R^4}{300\ln\frac{4(K+1)}{\beta}}.$$

Moreover, we also have (see (344), (350) and our induction assumption)

$$\mathbb{P}\{E_{T-1}\} \geq 1 - \frac{(T-1)\beta}{K+1},$$
$$\mathbb{P}\{E_{①}\} \geq 1 - \frac{\beta}{2(K+1)}, \quad \mathbb{P}\{E_{④}\} \geq 1 - \frac{\beta}{2(K+1)}.$$

where

$$\begin{aligned}
E_{①} &= \left\{\text{either} \quad \sum_{t=0}^{T-1} \sigma_t^2 > \frac{\exp(-2\gamma\mu T)R^4}{300\ln\frac{4(K+1)}{\beta}} \quad \text{or} \quad |①| \leq \frac{1}{5}\exp(-\gamma\mu T)R^2\right\}, \\
E_{④} &= \left\{\text{either} \quad \sum_{t=0}^{T-1} \widetilde{\sigma}_t^2 > \frac{\exp(-2\gamma\mu T)R^4}{300\ln\frac{4(K+1)}{\beta}} \quad \text{or} \quad |④| \leq \frac{1}{5}\exp(-\gamma\mu T)R^2\right\}.
\end{aligned}$$

Thus, probability event $E_{T-1} \cap E_{①} \cap E_{④}$ implies

$$\begin{aligned}
R_T^2 &\overset{(337)}{\leq} \exp(-\gamma\mu T)R^2 + ① + ② + ③ + ④ + ⑤ \\
&\leq 2\exp(-\gamma\mu T)R^2,
\end{aligned}$$

which is equivalent to (331) for $t = T$, and

$$\mathbb{P}\{E_T\} \geq \mathbb{P}\{E_{T-1} \cap E_{①} \cap E_{④}\} = 1 - \mathbb{P}\{\overline{E}_{T-1} \cup \overline{E}_{①} \cup \overline{E}_{④}\} \geq 1 - \frac{T\beta}{K+1}.$$

This finishes the inductive part of our proof, i.e., for all $k = 0, 1, \ldots, K+1$ we have $\mathbb{P}\{E_k\} \geq 1 - {k\beta}/{(K+1)}$. In particular, for $k = K+1$ we have that with probability at least $1-\beta$

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

Finally, if

$$\begin{aligned}
\gamma &= \min\left\{\frac{1}{400\ell\ln\frac{4(K+1)}{\beta}}, \frac{\ln(B_K)}{\mu(K+1)}\right\}, \\
B_K &= \max\left\{2, \frac{(K+1)^{\frac{2(\alpha-1)}{\alpha}}\mu^2R^2}{5400^{\frac{2}{\alpha}}\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{4(K+1)}{\beta}\right)\ln^2(B_K)}\right\} \\
&= \mathcal{O}\left(\max\left\{2, \frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2R^2}{\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)\ln^2\left(\max\left\{2, \frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2R^2}{\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)}\right\}\right)}\right\}\right)
\end{aligned}$$

then with probability at least $1-\beta$

$$\begin{aligned}
\|x^{K+1} - x^*\|^2 &\leq 2\exp(-\gamma\mu(K+1))R^2 \\
&= 2R^2\max\left\{\exp\left(-\frac{\mu(K+1)}{400\ell\ln\frac{4(K+1)}{\beta}}\right), \frac{1}{B_K}\right\} \\
&= \mathcal{O}\left(\max\left\{R^2\exp\left(-\frac{\mu K}{\ell\ln\frac{K}{\beta}}\right), \frac{\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)\ln^2\left(\max\left\{2, \frac{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2R^2}{\sigma^2\ln^{\frac{2(\alpha-1)}{\alpha}}\left(\frac{K}{\beta}\right)}\right\}\right)}{K^{\frac{2(\alpha-1)}{\alpha}}\mu^2}\right\}\right).
\end{aligned}$$

To get $\|x^{K+1} - x^*\|^2 \leq \varepsilon$ with probability at least $1-\beta$ it is sufficient to choose K such that both terms in the maximum above are $\mathcal{O}(\varepsilon)$. This leads to

$$K = \mathcal{O}\left(\frac{\ell}{\mu}\ln\left(\frac{R^2}{\varepsilon}\right)\ln\left(\frac{\ell}{\mu\beta}\ln\frac{R^2}{\varepsilon}\right), \left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\ln\left(\frac{1}{\beta}\left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right)\ln^{\frac{\alpha}{\alpha-1}}(B_\varepsilon)\right),$$

where

$$B_\varepsilon = \max\left\{2, \frac{R^2}{\varepsilon \ln\left(\frac{1}{\beta}\left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right)}\right\}.$$

This concludes the proof. □