Differentially Private Clipped-SGD: High-Probability Convergence with Arbitrary Clipping Level

Saleh Vatan Khah
IUST

Savelii Chezhegov
MIRAI

Shahrokh Farahmand
IUST

Samuel Horváth
MBZUAI

Eduard Gorbunov
MBZUAI

Abstract

Gradient clipping is a fundamental tool in Deep Learning, improving the high-probability convergence of stochastic first-order methods like SGD, AdaGrad, and Adam under heavy-tailed noise, which is common in training large language models. It is also a crucial component of Differential Privacy (DP) mechanisms. However, existing high-probability convergence analyses typically require the clipping threshold to increase with the number of optimization steps, which is incompatible with standard DP mechanisms like the Gaussian mechanism. In this work, we close this gap by providing the first high-probability convergence analysis for DP-Clipped-SGD with a fixed clipping level, applicable to both convex and non-convex smooth optimization under heavy-tailed noise, characterized by a bounded central α-th moment assumption, $\alpha \in (1, 2]$. Our results show that, with a fixed clipping level, the method converges to *a neighborhood* of the optimal solution with a *faster rate* than the existing ones. The neighborhood can be balanced against the noise introduced by DP, providing a refined trade-off between convergence speed and privacy guarantees.

Proceedings of the 29$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

Stochastic first-order optimization methods, such as Stochastic Gradient Descent (SGD) (Robbins and Monro, 1951), AdaGrad (Streeter and McMahan, 2010; Duchi et al., 2011), and Adam (Kingma and Ba, 2014), are fundamental for training modern Machine Learning (ML) and Deep Learning (DL) models. However, these methods are often enhanced with additional algorithmic techniques that play a critical role in their convergence and practical performance. Among these, gradient clipping (Pascanu et al., 2013) is one of the most widely used and well-studied approaches. In recent years, substantial efforts have been made to theoretically understand the advantages of gradient clipping and its impact on the convergence of stochastic optimization algorithms.

In particular, gradient clipping is a key component in managing heavy-tailed noise, which commonly arises in the training of language models on textual data (Zhang et al., 2020b), in the training of GANs (Goodfellow et al., 2014; Gorbunov et al., 2022), and even in simpler tasks such as image classification (Şimşekli et al., 2019). This approach is primarily analyzed through the lens of high-probability convergence, as such guarantees provide a more accurate reflection of the actual behavior of optimization methods compared to their more conventional in-expectation counterparts (Gorbunov et al., 2020). Moreover, as demonstrated by Sadiev et al. (2023) for SGD and by Chezhegov et al. (2024) for AdaGrad and Adam, methods without clipping may fail to exhibit high-probability convergence with logarithmic dependence on the failure probability. In contrast, several recent works (Gorbunov et al., 2020; Cutkosky and Mehta, 2021; Sadiev et al., 2023; Nguyen et al., 2023; Gorbunov et al., 2024b; Chezhegov et al., 2024; Parletta et al., 2024) have established that various stochastic first-order methods attain significantly better high-probability convergence under heavy-tailed noise assumptions across different settings.

On the other hand, clipping is a cornerstone of Differentially Private (DP) machine learning. The widely used Gaussian mechanism (Dwork et al., 2014) achieves privacy by adding Gaussian noise to the gradients, thereby introducing uncertainty about their true values. However, the DP guarantees provided by this mechanism rely on the assumption that the gradients have bounded norms, a condition typically enforced through gradient clipping (Abadi et al., 2016).

It is therefore tempting to claim that gradient clipping can provably address two distinct challenges simultaneously: mitigating heavy-tailed noise and ensuring differential privacy (DP). However, this is not entirely accurate, as the clipping policies required for these two objectives differ substantially. In the context of heavy-tailed noise, existing convergence guarantees are typically derived assuming that the clipping level increases with the total number of training steps. In contrast, DP mechanisms require a fixed and bounded clipping threshold to ensure robust privacy guarantees. This fundamental mismatch raises a critical question:

How does differentially private version of Clipped-SGD *converge with high probability under the heavy-tailed noise?*

Our contribution. In this paper, we address the above question by providing the first high-probability convergence bounds for the differentially private version of Clipped-SGD (DP-Clipped-SGD) with an *arbitrary fixed clipping level.* Specifically, for convex smooth optimization under heavy-tailed noise, we assume that the stochastic gradient has a bounded central α-th moment for some $\alpha \in (1, 2]$ and establish that DP-Clipped-SGD achieves a high-probability convergence rate of $\widetilde{\mathcal{O}}(K^{-1/2})$ to a neighborhood of the optimal solution. This rate is significantly faster than the previously known bound of $\widetilde{\mathcal{O}}(K^{-(\alpha-1)/\alpha})$ in this setting.

This improvement comes from relaxing exact convergence and instead proving convergence to a neighborhood whose size depends non-trivially on the clipping level, the DP noise scale, and other problem-dependent parameters. Importantly, the neighborhood induced by clipping bias can be balanced against the neighborhood induced by the DP noise, providing finer control over the trade-off between optimization accuracy and privacy. We also extend the analysis to the non-convex case, illustrating the broader applicability of our techniques.

2 TECHNICAL PRELIMINARIES

The optimization problem considered in this work has the following form

$$\min_{x\in\mathbb{R}^d} \left\{ f(x) := \mathbb{E}_{\xi\sim\mathcal{D}}[f_\xi(x)] \right\}. \quad (1)$$

Here, x denotes the model parameters, $f : \mathbb{R}^d \to \mathbb{R}$ is the expected loss function, and $f_\xi : \mathbb{R}^d \to \mathbb{R}$ represents the loss computed for a random sample ξ drawn from an (often unknown) distribution $\mathcal{D}$. Such problems are fundamental in machine learning (Shalev-Shwartz and Ben-David, 2014).

We assume that at each iteration, we have access to an oracle that provides a stochastic gradient $\nabla f_\xi(x)$, as well as a d-dimensional random vector ω sampled from a Gaussian distribution $\mathcal{N}(0, \sigma_\omega^2 \mathbf{I}_d)$, where $\mathbf{I}_d$ is the $d \times d$ identity matrix. More precisely, the random variables ξ and ω are defined on the probability space $\left(\Omega_d \times \mathbb{R}^d, \mathcal{B}(\Omega_d) \otimes \mathcal{B}(\mathbb{R}^d), \mathcal{F}^t, \mathbb{P}\right)$, where Ω_d represents the data sample space, and $\mathcal{B}(\mathcal{X})$ denotes the Borel σ-algebra generated by the set $\mathcal{X}$. This probability space is also equipped with the natural filtration $\mathcal{F}^t = \sigma\left(\left[\nabla f_{\xi^0}(x^0), \omega_0\right]^T, \dots \left[\nabla f_{\xi^t}(x^t), \omega_t\right]^T\right)$, which captures the history of the stochastic process up to time t. The probability measure $\mathbb{P}$ is defined as the product measure on this space, given by

$$\mathbb{P}\{B_d \times B_\omega\} = (\mu \times \nu)(B_d \times B_\omega) = \mu(B_d)\, \nu(B_\omega),$$
$$\forall B_d \in \mathcal{B}(\Omega_d), \forall B_\omega \in \mathcal{B}(\mathbb{R}^d),$$

where μ is a probability measure on Ω_d, and ν is the Gaussian measure on $\mathbb{R}^d$ with mean zero and covariance matrix $\sigma_\omega^2 \mathbf{I}_d$.

Types of convergence bounds. Several types of convergence bounds are commonly used to analyze the behavior of stochastic optimization methods, ranging from in-expectation bounds to almost sure convergence guarantees. High-probability convergence bounds provide guarantees of the form $\mathbb{P}\left\{\mathcal{P}(x^K) \leq \epsilon\right\} \geq 1 - \beta$, where $\mathcal{P}(x)$ is a performance metric that measures the quality of the solution[1]. Here, $\mathbb{P}\{\cdot\}$ denotes the probability measure defined by the problem setup, x^K is the algorithm's output after K iterations, β is the confidence level (or failure probability), and ϵ is the optimization error.

This type of convergence is generally considered superior to in-expectation guarantees (e.g., $\mathbb{E}[\mathcal{P}(x^K)] \leq \epsilon$), as it captures not only the average behavior of the

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

underlying random variables but also their tail behavior, which is particularly important for distributions with heavy tails. However, it is worth noting that the number of iterations K required to achieve such high-probability guarantees can depend inversely on the failure probability β, as seen in analyses for methods like SGD (Sadiev et al., 2023), AdaGrad, and Adam (Chezhegov et al., 2024). Such inverse-power dependencies on β are generally undesirable, as β is typically chosen to be very small. Consequently, a major objective in the high-probability convergence literature is to establish bounds with polylogarithmic dependence on $1/\beta$, which are significantly tighter and more practical.

Assumptions. In the following, we list the assumptions on the structure of the problem at hand. These assumptions are very mild and cover a wide range of problems.

Assumption 2.1. *We assume the function f is uniformly lower-bounded on some subset $Q \subseteq \mathbb{R}^d$, i.e., $f^* \overset{\text{def}}{=} \inf_{x\in Q} f(x) > -\infty$.*

The above assumption is necessary for problem (1) to be feasible. Next, we make a standard assumption about the smoothness of the objective function.

Assumption 2.2. *We assume that there exists a constant $L > 0$ such that for all $x, y \in Q \subseteq \mathbb{R}^d$ the function f satisfies the following.*

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|. \tag{2}$$

In this work, we consider both classes of convex and non-convex functions. The following assumption holds only for convex functions.

Assumption 2.3. *We assume there exists a subset Q of $\mathbb{R}^d$ such that for all $x, y \in Q$*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle. \tag{3}$$

The following assumption is with respect to the stochastic oracle that our algorithm receives at each iteration. We assume that the stochastic gradients have a bounded central α moment for some $\alpha \in (1, 2]$. This assumption is stated explicitly below.

Assumption 2.4. *We assume there exist some subset $Q \subseteq \mathbb{R}^d$, and some constants $\sigma > 0$, $\alpha \in (1, 2]$ such that for all $x \in Q$*

$$\mathbb{E}_{\xi\sim D}\left[\nabla f_\xi(x) \mid x\right] = \nabla f(x), \tag{4}$$

$$\mathbb{E}_{\xi\sim D}\left[\|\nabla f_\xi(x) - \nabla f(x)\|^\alpha \mid x\right] \leq \sigma^\alpha. \tag{5}$$

As it can be seen, in the case $\alpha = 2$, the aforementioned conditions recover the standard uniformly bounded variance assumption widely used for obtaining convergence guarantees for optimization algorithms in the literature. Since the L^p norms of random variable are non-decreasing in p, this assumption allows the stochastic gradients to have infinite variance.

Next, we use the classical definition of (ε, δ)-differential privacy. Intuitively, it provides probabilistic guarantees that an intruder cannot infer the existence of a particular data in the data set that the algorithm used to train the model.

Definition 2.1. *((ϵ, δ)-Differential Privacy (Dwork et al., 2014)). A randomized method $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ satisfies (ε, δ)-Differential Privacy, if for any adjacent $D, D' \in \mathcal{D}$ and for any $\mathcal{S} \subseteq \mathcal{R}$*

$$\mathbb{P}\left(\mathcal{M}(\mathcal{D}) \in \mathcal{S}\right) \leq e^\varepsilon \mathbb{P}\left(\mathcal{M}(\mathcal{D}') \in \mathcal{S}\right) + \delta, \tag{6}$$

Smaller (ε, δ) provides stronger privacy guarantee. This also can be viewed from the perspective of Bayesian hypothesis testing where the null and alternative hypothesis are about the existence of an individual's data in the dataset (Kairouz et al., 2015; Su, 2024).

3 RELATED WORK

Clipping in Differential Private learning. There are several approaches to ensuring DP guarantees in SGD, but the most common method relies on a combination of gradient clipping and noise injection. In the finite-sum setting, Abadi et al. (2016) demonstrated that it is sufficient to add Gaussian noise (the Gaussian mechanism) with standard deviation $\sigma_\omega = \Theta\left(\frac{q\lambda}{\varepsilon}\sqrt{K \ln \frac{1}{\delta}}\right)$ to the clipped gradients, where q is the sampling probability for each individual summand. This approach reduces the variance of the required Gaussian noise by a factor of $\sqrt{\ln K}$ compared to the advanced composition theorem (Dwork et al., 2014), significantly improving the utility of DP training.

This combination of gradient clipping and the Gaussian mechanism has become a standard approach in many DP training algorithms. However, these methods often rely on restrictive assumptions, such as requiring the clipping level to always be larger than the norm of the transmitted vector (Zhang et al., 2022; Noble et al., 2022; Allouah et al., 2023, 2024; Li and Chi, 2025)[2], assuming symmetry of the noise distribution (Liu et al., 2022), light tails (e.g., sub-exponential

[2] Li and Chi (2025) also provide an in-expectation convergence result without the bounded gradient assumption, but with a worse dependence on the variance bound of the stochastic gradients.

ones) (Wang et al., 2020), large batch sizes (Jin et al., 2024), boundedness of the gradient (Das et al., 2023) or requiring that the full gradients be computed (Wei et al., 2020). These conditions can be quite restrictive, particularly in practical large-scale settings.

To the best of our knowledge, the only works that avoid these restrictive assumptions are Koloskova et al. (2023); Islamov et al. (2025). Specifically, Koloskova et al. (2023) analyzed the in-expectation convergence of DP-Clipped-SGD with mini-batching under the bounded variance assumption, for an arbitrary clipping level in the non-convex (L_0, L_1)-smooth regime (Zhang et al., 2020a). However, they leave open the question of high-probability convergence under heavy-tailed noise. Islamov et al. (2025) proposed a distributed optimization method that incorporates clipping, error feedback (Seide et al., 2014; Richtárik et al., 2021), and heavy-ball momentum (Polyak, 1964). Yet, their high-probability convergence analysis crucially relies on the assumption that the noise in the stochastic gradients has sub-Gaussian tails. In contrast, under the more realistic Assumption 2.4 with $\alpha \geq 2$ (which is still more restrictive than the heavy-tailed case with $\alpha < 2$), Zhao et al. (2025) derive in-expectation convergence bounds for a variant of projected SGD that employs DP mean estimation using a sufficiently large number of samples. However, this approach can be prohibitively expensive in practice, especially for training large language models.

High-probability convergence bounds. If the noise in the stochastic gradient has light tails, then classical stochastic first-order methods like SGD and its adaptive and momentum-based variants can achieve desirable high-probability convergence rates, characterized by polylogarithmic dependence on the failure probability β. For instance, under the sub-Gaussian noise assumption, such results exist for SGD (Nemirovski et al., 2009; Harvey et al., 2019), its accelerated variants (Ghadimi and Lan, 2012; Dvurechensky and Gasnikov, 2016), and its momentum and AdaGrad versions (Li and Orabona, 2020; Liu et al., 2023). Additionally, Madden et al. (2024) demonstrate that polylogarithmic high-probability bounds can also be achieved for SGD under the weaker sub-Weibull noise assumption. However, as highlighted by Sadiev et al. (2023) and Chezhegov et al. (2024), methods like SGD, AdaGrad, and Adam can fail to achieve these desired high-probability rates under heavier-tailed noise distributions.

To address the limitations of high-probability convergence for stochastic methods under heavy-tailed noise, several algorithmic modifications have been proposed and rigorously analyzed in recent years. Nazin et al. (2019) introduced a variant of Stochastic Mirror Descent (Nemirovskij and Yudin, 1983) with *truncation* of the stochastic gradient, establishing high-probability complexity bounds for convex and strongly convex smooth optimization over compact sets under the bounded variance assumption (Assumption 2.4 with $\alpha = 2$). Interestingly, the truncation operator used in this work, while not identical, is closely related to the standard *gradient clipping* technique that has since become the foundation of many subsequent studies.

In particular, Gorbunov et al. (2020) derived the first high-probability complexity bounds for Clipped-SGD and also proposed an accelerated version based on the Stochastic Similar Triangles Method (SSTM) (Gasnikov and Nesterov, 2016). These results were later extended to non-smooth problems by Gorbunov et al. (2024a); Parletta et al. (2024), to unconstrained variational inequalities by Gorbunov et al. (2022), and to settings with noise having a bounded α-th moment by Cutkosky and Mehta (2021) (with an additional bounded gradient assumption in the non-convex case). Building on these foundations, Sadiev et al. (2023) extended the results from Gorbunov et al. (2020) and Gorbunov et al. (2022) to the more challenging setting defined by Assumption 2.4 with $\alpha < 2$, removing the bounded gradient assumption for non-convex objectives. This work also introduced new high-probability bounds for Clipped-SGD in the non-convex regime. These non-convex results were further refined by Nguyen et al. (2023), who also obtained tighter logarithmic factors in the convergence rates for both convex and strongly convex settings.

In the context of distributed optimization, Gorbunov et al. (2024b) extended the results of Sadiev et al. (2023) to distributed composite minimization and variational inequalities using the clipping of gradient differences, thereby broadening the applicability to decentralized and federated learning scenarios.

Adaptive methods have also been analyzed through the lens of high-probability convergence. Li and Liu (2023) derived new high-probability bounds for Clipped-AdaGrad with scalar step-sizes, while Chezhegov et al. (2024) obtained analogous bounds for various versions of Clipped-AdaGrad and Clipped-Adam with both scalar and coordinate-wise step-sizes. Additionally, Kornilov et al. (2023) proposed a zeroth-order variant of Clipped-SSTM and analyzed it under Assumption 2.4, extending the clipping framework to derivative-free settings.

However, a critical limitation shared by all of these methods is that the clipping level λ is typically chosen as an increasing function of the total number of steps

K^3. This choice, while theoretically convenient, leads to prohibitively large DP noise variance when aiming to guarantee (ε, δ)-DP, resulting in utility bounds that grow with K and significantly degrade the practical effectiveness of these methods in privacy-preserving applications.

There exist other alternatives to gradient clipping that also ensure high-probability convergence with polylogarithmic dependency on the failure probability. They include robust distance estimation coupled with inexact proximal point steps (Davis et al., 2021), gradient normalization (Cutkosky and Mehta, 2021; Hübler et al., 2024), and sign-based methods (Kornilov et al., 2025). Notably, the approaches from Hübler et al. (2024); Kornilov et al. (2025) enjoy provable (yet suboptimal) high-probability convergence even when α is unknown. In the special case of symmetric distributions, Armacki et al. (2023, 2024) provide new high-probability convergence bounds for a large class of SGD-type methods with non-linear transformations such as standard clipping, coordinate-wise clipping, normalization, and sign-operator, and Puchkin et al. (2024) derive high-probability convergence of SGD with median-based clipping and also extend this result to problems with structured non-symmetry for SGD with smoothed median of means coupled with gradient clipping.

4 MAIN RESULTS

The well-known Clipped-SGD algorithm with the Gaussian DP mechanism (DP-Clipped-SGD) is described in Algorithm 1. If differential privacy (DP) is not required, one can simply set $\sigma_\omega^2 = 0$. As shown by Sadiev et al. (2023), achieving exact convergence to the optimal solution of problem (1) using Clipped-SGD requires the clipping level to be chosen as $\lambda = \mathcal{O}\left(\sigma \left({}^{K}/(\ln \frac{K}{\beta})\right)^{1/\alpha}\right)$. However, this choice of clipping level, which scales with the total number of iterations K, is problematic from a DP perspective. Specifically, larger clipping levels necessitate larger DP noise to maintain privacy, significantly increasing the variance in gradient estimates and leading to a larger convergence neighborhood.

To address this limitation, in this work, we focus on the more general case of arbitrary fixed clipping levels that do not scale with the total number of iterations. This approach is more compatible with practical DP requirements, where clipping levels are typically kept constant. However, our theoretical results can also accommodate clipping levels that scale with K up to the order $\lambda = \mathcal{O}\left(\sigma \left({}^{K}/(\ln \frac{K}{\beta})\right)^{1/\alpha}\right)$, as we discuss in detail in the appendix. This broader analysis introduces a few additional step-size conditions, which we also explore thoroughly in the supplementary material.

[3] In some cases, such as the analysis of Clipped-SSTM (Gorbunov et al., 2020) or Clipped-SGD under strong convexity (Sadiev et al., 2023), the clipping level decreases as a function of the current iteration counter k but still increases overall as a function of K.

Algorithm 1 DP-Clipped-SGD

Input: Starting point x^0, number of iterations K, step-size $\gamma > 0$, clipping level λ.

for $k = 0, \ldots, K$ **do**
  Sample $\xi^k \sim \mathcal{D}$
  Compute $\hat{g}_k = \texttt{clip}\left(\nabla f_{\xi^k}(x^k), \lambda\right)$
  $\omega_k \sim \mathcal{N}(0, \sigma_\omega^2 I_d)$
  $\widetilde{g}_k = \hat{g}_k + \omega_k$
  $x^{k+1} = x^k - \gamma \widetilde{g}_k$
end for

The next two theorems present our step-size conditions and the resulting performance guarantees in the convex and non-convex settings. After each theorem, we provide a table that simplifies the bound in several representative regimes of the clipping level. In these tables, we set the DP noise to zero in order to isolate the effect of clipping bias. The final corollary makes the DP dependence explicit in the convex case, while the corresponding (ε, δ)-specialized non-convex statement is deferred to the supplementary material due to space limitations.

Convex problems. We start with the convex case.

Theorem 4.1 (Convergence of DP-Clipped-SGD for the convex objectives)**.** *Let the integer $K \geq 0$ and $\beta \in (0, 1]$ be given. Furthermore, let Assumptions 2.1, 2.2, 2.3, 2.4, hold for $Q = B_{2R}(x^\star)$, $R \geq \|x^0 - x^\star\|$. Set $\zeta_\lambda \stackrel{\text{def}}{=} \max\left\{0, 2LR - \frac{\lambda}{2}\right\}$, and further assume that the step-size γ is selected to satisfy*

$$\gamma \leq \mathcal{O}\left(\min\left\{①, ②, ③, ④\right\}\right) \tag{7}$$

$$① = \frac{1}{L}, \quad ② = \frac{R}{\lambda^{1-\alpha/2}\sqrt{K \ln\left(\frac{K}{\beta}\right)\left(\sigma^\alpha + \zeta_\lambda^\alpha\right)}}$$

$$③ = \frac{R\lambda^{\alpha-1}}{K(\sigma^\alpha + \zeta_\lambda^\alpha)\left(\frac{LR}{\lambda} + \frac{\lambda^{\alpha-1}\zeta_\lambda}{\sigma^\alpha + \zeta_\lambda^\alpha} + \left(\sigma^\alpha + \zeta_\lambda^\alpha\right)^{\frac{-1}{\alpha}}\right)}$$

$$④ = \frac{R}{\sigma_\omega\sqrt{dK \ln\left(\frac{K}{\beta}\right)}}.$$

Then, after K iterations of DP-Clipped-SGD*, the iterates with probability at least $1-\beta$ satisfy*

$$\min_{t\in[0,K]} f(x^t) - f(x^\star) \le \frac{4R^2}{\gamma(K+1)} + \frac{64LR^4}{\lambda^2\gamma^2(K+1)^2}. \quad (8)$$

The convergence rate and the neighborhood to which the algorithm converges depend on the magnitude of λ in a non-trivial way. Table 1 summarizes these relationships for different values of λ in the absence of DP noise. In the special case where $\lambda = \mathcal{O}\left(\sigma\left({}^{K}/\ln\frac{K}{\beta}\right)^{1/\alpha}\right)$, our theorem provides a convergence rate of $\mathcal{O}\left(\left({}^{(\ln\frac{K}{\beta})}/K\right)^{(\alpha-1)/\alpha} + {}^{(\ln\frac{K}{\beta})}/K\right)$ to the exact solution in the asymptotic regime. This matches the rate previously derived by Sadiev et al. (2023).

In contrast, if λ is chosen as a constant, independent of K, the leading term in the convergence rate simplifies to $\mathcal{O}(\sqrt{(\ln\frac{K}{\beta})/K})$, which is faster than the more conservative bound $\mathcal{O}\left(\left((\ln\frac{K}{\beta})/K\right)^{(\alpha-1)/\alpha}\right)$. However, this faster rate comes at the cost of only guaranteeing convergence to a neighborhood around the optimal solution, determined by the third term in the step-size condition (7).

To ensure (ε,δ)-DP for DP-Clipped-SGD in our setting (i.e., expectation minimization), one can set the noise scale as $\sigma_\omega = \Theta\left(\frac{\lambda}{\varepsilon}\sqrt{K\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{1}{\delta}\right)}\right)$ and apply the advanced composition theorem (Dwork et al., 2014, Theorem 3.22). Given the fourth term in (7), this choice implies that the step-size decreases as $1/K$, resulting in convergence to a certain neighborhood. This observation is formalized in the next corollary.

Corollary 4.1 (Convergence of DP-Clipped-SGD for the convex objective)**.** *Let the assumptions of Theorem 4.1 hold, $\sigma_\omega = \Theta\left(\frac{\lambda}{\varepsilon}\sqrt{K\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{1}{\delta}\right)}\right)$, and γ is chosen as the minimum of (7). Take $A = (\sigma^\alpha + \zeta_\lambda^\alpha)$, $B = d\ln\left(\frac{K}{\beta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{1}{\delta}\right)$, and $C = \left(\frac{LR}{\lambda} + \frac{\lambda^{\alpha-1}\zeta_\lambda}{A} + A^{\frac{-1}{\alpha}}\right)$. Then, with probability at least $1-\beta$*

$$\min_{t\in[0,K]} f(x^t) - f(x^\star) \le \mathcal{O}\left(\max\left\{(10),(11),(12),(13)\right\}\right). \quad (9)$$

where

$$\frac{LR^2}{K} + \frac{L^3R^4}{\lambda^2K^2} \quad (10)$$

$$R\lambda^{1-\alpha/2}\sqrt{\frac{A\ln(K/\beta)}{K}} + \frac{LR^2\lambda^\alpha A\ln(K/\beta)}{K} \quad (11)$$

$$\frac{RAC}{\lambda^{\alpha-1}} + \frac{R^2LA^2C^2}{\lambda^{2\alpha}} \quad (12)$$

$$\frac{R\lambda}{\varepsilon}\sqrt{B} + \frac{LR^2B}{\varepsilon^2} \quad . \quad (13)$$

One may notice that there is a non-trivial trade-off between the convergence rate, clipping level, and the size of the neighborhood. Therefore, we consider two special cases and provide the result with optimally selected λ in the appendix.

In the finite-sum case, i.e., when $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ for some finite n, Abadi et al. (2016) show that it is sufficient to choose $\sigma_\omega = \Theta\left(\frac{q\lambda}{\varepsilon}\sqrt{K\ln\frac{1}{\delta}}\right)$, where $q = b/n$, b is the mini-batch size, clipping is applied to each stochastic gradient, and $\varepsilon = \mathcal{O}(q^2K)$, allowing to have smaller ε and δ for given σ_ω and λ. We note that our analysis holds for the finite-sum case without changes as long as the assumptions of the theorem are satisfied and the mini-batch size equals 1.

Non-convex problems. In the non-convex case, we derive the following result.

Theorem 4.2 (Convergence of DP-Clipped-SGD for the non-convex objective)**.** *Let the integer $K \ge 0$ and $\beta \in (0,1]$ be given. Let the assumptions 2.1, 2.2, 2.4, hold for the set*

$$Q = \left\{x \in \mathbb{R}^d \mid \exists y \in \mathbb{R}^d : f(y) \le f^* + 2\Delta \text{ and } \|x-y\| \le \frac{\sqrt{\Delta}}{20\sqrt{L}}\right\}, \quad (14)$$

where $\Delta \ge f(x^0) - f^*$, $\zeta_\lambda \overset{\text{def}}{=} \max\left\{0, 2\sqrt{L\Delta} - \frac{\lambda}{2}\right\}$, and γ is selected according to

$$\gamma \le \mathcal{O}\left(\min\{①,②,③,④\}\right), \quad (15)$$

$$① = \frac{1}{L}, \quad ② = \frac{\sqrt{\frac{\Delta}{L}}}{\lambda^{1-\alpha/2}\sqrt{K\ln\left(\frac{K}{\beta}\right)(\sigma^\alpha + \zeta_\lambda^\alpha)}},$$

$$③ = \frac{\sqrt{\frac{\Delta}{L}}\lambda^{\alpha-1}}{K(\sigma^\alpha+\zeta_\lambda^\alpha)\left(\frac{\sqrt{L\Delta}}{\lambda} + \frac{\lambda^{\alpha-1}\zeta_\lambda}{\sigma^\alpha+\zeta_\lambda^\alpha} + (\sigma^\alpha+\zeta_\lambda^\alpha)^{\frac{-1}{\alpha}}\right)},$$

$$④ = \frac{\sqrt{\frac{\Delta}{L}}}{\sigma_\omega\sqrt{dK\ln\left(\frac{K}{\beta}\right)}}.$$

Then, after K iterations of DP-Clipped-SGD *and with probability at least $1-\beta$, we have*

$$\min_{t\in[0,K]} \left\|\nabla f(x^t)\right\|^2 \le \frac{8\Delta}{\gamma(K+1)} + \frac{128\Delta^2}{\lambda^2\gamma^2(K+1)^2} \quad (16)$$

Table 1: Rate, neighborhood and optimal λ in different regimes for the convex objective function. Here, λ denotes the clipping level, L denotes the smoothness parameter, $R \geq \|x^0 - x^*\|$ represents the initial error, $\alpha \in (1,2]$ denotes the moment that is bounded and σ^α is that upper bound value. Furthermore, β is the confidence level, $\zeta_\lambda := \max\{0, 2LR - \frac{\lambda}{2}\}$, and η is a small positive constant. By optimal λ and optimal neighborhood, we refer to the λ that minimizes the right hand side (RHS) of (8) and the minimized RHS value itself, respectively.

Regime	Neighborhood	Optimal λ	Convergence rate	Optimal Neighborhood
$\lambda > 4LR$ $(\zeta_\lambda = 0)$	$\mathcal{O}\left(R\frac{\sigma^\alpha}{\lambda^{\alpha-1}} + LR^2\frac{\sigma^{2\alpha}}{\lambda^{2\alpha}}\right)$	$\mathcal{O}\left(\sigma\left(\frac{K}{\ln\frac{K}{\beta}}\right)^{\frac{1}{\alpha}}\right)$	$\mathcal{O}\left(\left(\frac{\ln\frac{K}{\beta}}{K}\right)^{\frac{\alpha-1}{\alpha}} + \frac{\ln^2\frac{K}{\beta}}{K^2}\right)$	-
$\frac{4}{3}LR < \lambda \leq 4LR$ $\zeta_\lambda < \lambda < \sigma$	$\mathcal{O}\left(R\frac{\sigma^\alpha}{\lambda^{\alpha-1}} + LR^2\frac{\sigma^{2\alpha}}{\lambda^{2\alpha}}\right)$	$4LR$	$\mathcal{O}\left(\sqrt{\frac{\ln\frac{K}{\beta}}{K}} + \frac{\ln\frac{K}{\beta}}{K}\right)$	$\mathcal{O}\left(\frac{R^{2-\alpha}\sigma^\alpha}{L^{\alpha-1}} + \frac{\sigma^{2\alpha}}{L^{2\alpha-1}R^{2\alpha-2}}\right)$
$\frac{4}{3}LR < \lambda \leq 4LR$ $\zeta_\lambda < \sigma < \lambda$	$\mathcal{O}\left(R\frac{\sigma^\alpha}{\lambda^{\alpha-1}} + LR^2\frac{\sigma^{2\alpha}}{\lambda^{2\alpha}}\right)$	$4LR$	$\mathcal{O}\left(\sqrt{\frac{\ln\frac{K}{\beta}}{K}} + \frac{\ln\frac{K}{\beta}}{K}\right)$	$\mathcal{O}\left(\frac{R^{2-\alpha}\sigma^\alpha}{L^{\alpha-1}} + \frac{\sigma^{2\alpha}}{L^{2\alpha-1}R^{2\alpha-2}}\right)$
	$\mathcal{O}\left(R\zeta_\lambda + \frac{LR^2\zeta_\lambda^2}{\lambda^2}\right)$	$4LR - \eta$	$\mathcal{O}\left(\sqrt{\frac{\ln\frac{K}{\beta}}{K}} + \frac{\ln\frac{K}{\beta}}{K}\right)$	$\mathcal{O}\left(R\eta + \frac{LR^2\eta^2}{(LR-\eta)^2}\right)$
$\frac{4}{3}LR < \lambda \leq 4LR$ $(\sigma < \zeta_\lambda < \lambda)$	$\mathcal{O}\left(R\zeta_\lambda + \frac{LR^2\zeta_\lambda^2}{\lambda^2}\right)$	$4LR - 2\sigma$	$\mathcal{O}\left(\sqrt{\frac{\ln\frac{K}{\beta}}{K}} + \frac{\ln\frac{K}{\beta}}{K}\right)$	$\mathcal{O}\left(R\sigma + \frac{LR^2\sigma^2}{(LR-\sigma)^2}\right)$
$\lambda \leq \frac{4}{3}LR$ $(\lambda < \zeta_\lambda < \sigma)$	$\mathcal{O}\left(R\frac{\sigma^\alpha\zeta_\lambda}{\lambda^\alpha} + \frac{LR^2\sigma^{2\alpha}\zeta_\lambda^2}{\lambda^{2\alpha+2}}\right)$	$\frac{4}{3}LR$	$\mathcal{O}\left(\sqrt{\frac{\ln\frac{K}{\beta}}{K}} + \frac{\ln\frac{K}{\beta}}{K}\right)$	$\mathcal{O}\left(\frac{R^{2-\alpha}\sigma^\alpha}{L^{\alpha-1}} + \frac{\sigma^{2\alpha}}{L^{2\alpha-1}R^{2\alpha-2}}\right)$
$\lambda \leq \frac{4}{3}LR$ $(\lambda < \sigma < \zeta_\lambda)$	$\mathcal{O}\left(R\frac{\zeta_\lambda^{\alpha+1}}{\lambda^\alpha} + \frac{LR^2\zeta_\lambda^{2\alpha}}{\lambda^{2\alpha+2}}\right)$	$\frac{4}{3}LR - \eta$	$\mathcal{O}\left(\sqrt{\frac{\ln\frac{K}{\beta}}{K}} + \frac{\ln\frac{K}{\beta}}{K}\right)$	$\mathcal{O}\left(\frac{R(LR+\eta)^{\alpha+1}}{(LR-\eta)^\alpha} + \frac{LR^2(LR+\eta)^{2\alpha}}{(LR-\eta)^{2\alpha+2}}\right)$
$\lambda \leq \frac{4}{3}LR$ $(\sigma < \lambda < \zeta_\lambda)$	$\mathcal{O}\left(R\frac{\zeta_\lambda^{\alpha+1}}{\lambda^\alpha} + \frac{LR^2\zeta_\lambda^{2\alpha}}{\lambda^{2\alpha+2}}\right)$	$\frac{4}{3}LR - \eta$	$\mathcal{O}\left(\sqrt{\frac{\ln\frac{K}{\beta}}{K}} + \frac{\ln\frac{K}{\beta}}{K}\right)$	$\mathcal{O}\left(\frac{R(LR+\eta)^{\alpha+1}}{(LR-\eta)^\alpha} + \frac{LR^2(LR+\eta)^{2\alpha}}{(LR-\eta)^{2\alpha+2}}\right)$
	$\mathcal{O}\left(R\frac{\sigma\zeta_\lambda^{\alpha-1}}{\lambda^{\alpha-1}} + \frac{LR^2\sigma^2\zeta_\lambda^{2\alpha-2}}{\lambda^{2\alpha}}\right)$	$\frac{4}{3}LR$	$\mathcal{O}\left(\sqrt{\frac{\ln\frac{K}{\beta}}{K}} + \frac{\ln\frac{K}{\beta}}{K}\right)$	$\mathcal{O}\left(R\sigma + \frac{\sigma^2}{L}\right)$

Similarly to the convex case, the above result establishes the convergence to a certain neighborhood with a faster $\mathcal{O}(1/\sqrt{K})$ rate. We defer the corollaries for the non-convex case to the appendix and describe different special cases for the no-DP regime in Table 2.

Corollary 4.2 (Convergence of DP-Clipped-SGD for the non-convex objective)**.** *Let the assumption of Theorem 4.2 hold, and γ is chosen as the minimum of (15). Take $A = (\sigma^\alpha + \zeta_\lambda^\alpha)$, and $B = \left(\frac{\sqrt{L\Delta}}{\lambda} + \frac{\lambda^{\alpha-1}\zeta_\lambda}{A} + A^{\frac{-1}{\alpha}}\right)$. Then, with probability at least $1-\beta$*

$$\min_{t\in[0,K]} \left\|\nabla f(x^t)\right\|^2 \leq \mathcal{O}\left(\max\left\{(18),(19),(20),(21)\right\}\right), \tag{17}$$

where

$$\frac{L\Delta}{K} + \frac{L^2\Delta^2}{\lambda^2 K^2} \tag{18}$$

$$\sqrt{L\Delta}\lambda^{1-\alpha/2}\sqrt{\frac{A\ln{K/\beta}}{K}} + \frac{L\Delta A\ln(K/\beta)}{\lambda^\alpha K} \tag{19}$$

$$\frac{\sqrt{\Delta L}AB}{\lambda^{\alpha-1}} + \frac{\Delta L A^2 B^2}{\lambda^{2\alpha}} \tag{20}$$

$$\frac{\sigma_\omega\sqrt{dL\Delta\ln(K/\beta)}}{\sqrt{K}} + \frac{\sigma_\omega^2 dL\Delta\ln(K/\beta)}{\lambda^2 K}. \tag{21}$$

Comparison with the results by Koloskova et al. (2023). Koloskova et al. (2023) derive their *in-expectation* convergence result under the (L_0, L_1)-smoothness assumption (Zhang et al., 2020a) and the σ^2-uniformly bounded variance assumption (i.e., Assumption 2.4 with $\alpha = 2$), for DP-Clipped-SGD with mini-batching. For ease of comparison, we consider the special case $L_1 = 0$ and $L_0 = L$, which corresponds to standard L-smoothness. Moreover, for simplicity, we assume a mini-batch size of 1. In this setting, the result from Koloskova et al. (2023, Appendix C.4.2) for DP-Clipped-SGD can be written as follows: if $\gamma \leq 1/9L$, then

$$\min_{t\in[0,K]} \left(\mathbb{E}\left[\|\nabla f(x^t)\|\right]\right)^2 \leq$$
$$\mathcal{O}\left(\frac{\Delta}{\gamma K} + \frac{\Delta^2}{\lambda^2\gamma^2K^2} + \gamma L\sigma^2 + \min\left\{\sigma^2, \frac{\sigma^4}{\lambda^2}\right\} + \gamma L d\sigma_\omega^2 + \frac{\gamma^2L^2d^2\sigma_\omega^4}{\lambda^2}\right)$$

The structure of our bound is quite similar. Specifically, the terms from (18) correspond to the convergence of DP-Clipped-SGD in the noiseless regime ($\sigma = \sigma_\omega = 0$) and match the $\mathcal{O}\left(\frac{\Delta}{\gamma K} + \frac{\Delta^2}{\lambda^2\gamma^2K^2}\right)$ part when $\gamma = \Theta(1/L)$. Next, the terms in (19) serve as analogs of the $\mathcal{O}(\gamma L\sigma^2)$ term. The leading term in (19) matches the K-dependence of $\mathcal{O}(\gamma L\sigma^2)$ for $\gamma = \Theta(1/\sqrt{K})$. However, these terms also depend on the clipping level λ, which arises from our high-probability convergence analysis and the presence of heavy-tailed noise.

The key difference lies in the terms stemming from the inherent bias of Clipped-SGD (Koloskova et al., 2023, Theorems 3.1–3.2) and the DP noise. In our result, these bias terms appear in (20), while the corresponding term in Koloskova et al. (2023) is $\mathcal{O}\left(\min\left\{\sigma^2, \frac{\sigma^4}{\lambda^2}\right\}\right)$. As shown in Table 2, in the spe-

Table 2: Rate, neighborhood and optimal λ in different regimes for the non-convex objective function. Here, λ denotes the clipping level, L denotes the smoothness parameter, $\Delta \geq f(x^0) - f(x^*)$ represents the initial error, $\alpha \in (1, 2]$ denotes the moment that is bounded and σ^α is that upper bound value. Furthermore, β is the confidence level, $\zeta_\lambda := \max\{0, 2\sqrt{L\Delta} - \frac{\lambda}{2}\}$,and η is a small positive constant. By optimal λ and optimal neighborhood, we refer to the λ that minimizes the right hand side (RHS) of (16) and the minimized RHS value itself, respectively.

Regime	Neighborhood	Optimal λ	Convergence rate	Optimal Neighborhood
$\lambda > 4\sqrt{L\Delta}$ $(\zeta_\lambda = 0)$	$\mathcal{O}\left(\sqrt{L\Delta}\frac{\sigma^\alpha}{\lambda^{\alpha-1}} + L\Delta\frac{\sigma^{2\alpha}}{\lambda^{2\alpha}}\right)$	$\mathcal{O}\left(\sigma\left(\frac{K}{\ln\frac{K}{\beta}}\right)^{\frac{1}{\alpha}}\right)$	$\mathcal{O}\left(\left(\frac{\ln\frac{K}{\beta}}{K}\right)^{\frac{\alpha-1}{\alpha}} + \frac{\ln^2\frac{K}{\beta}}{K^2}\right)$	-
$\frac{4}{3}\sqrt{L\Delta} < \lambda \leq 4\sqrt{L\Delta}$ $\zeta_\lambda < \lambda < \sigma$	$\mathcal{O}\left(\sqrt{L\Delta}\frac{\sigma^\alpha}{\lambda^{\alpha-1}} + L\Delta\frac{\sigma^{2\alpha}}{\lambda^{2\alpha}}\right)$	$4\sqrt{L\Delta}$	$\mathcal{O}\left(\sqrt{\frac{\ln\frac{K}{\beta}}{K}} + \frac{\ln\frac{K}{\beta}}{K}\right)$	$\mathcal{O}\left(\frac{\sigma^\alpha}{(\sqrt{L\Delta})^{\alpha-2}} + \frac{\sigma^{2\alpha}}{(\sqrt{L\Delta})^{2\alpha-2}}\right)$
$\frac{4}{3}\sqrt{L\Delta} < \lambda \leq 4\sqrt{L\Delta}$ $\zeta_\lambda < \lambda < \sigma$	$\mathcal{O}\left(\sqrt{L\Delta}\frac{\sigma^\alpha}{\lambda^{\alpha-1}} + L\Delta\frac{\sigma^{2\alpha}}{\lambda^{2\alpha}}\right)$	$4\sqrt{L\Delta}$	$\mathcal{O}\left(\sqrt{\frac{\ln\frac{K}{\beta}}{K}} + \frac{\ln\frac{K}{\beta}}{K}\right)$	$\mathcal{O}\left(\frac{\sigma^\alpha}{(\sqrt{L\Delta})^{\alpha-2}} + \frac{\sigma^{2\alpha}}{(\sqrt{L\Delta})^{2\alpha-2}}\right)$
	$\mathcal{O}\left(\sqrt{L\Delta}\zeta_\lambda + \frac{L\Delta\zeta_\lambda^2}{\lambda^2}\right)$	$4\sqrt{L\Delta} - \eta$	$\mathcal{O}\left(\sqrt{\frac{\ln\frac{K}{\beta}}{K}} + \frac{\ln\frac{K}{\beta}}{K}\right)$	$\mathcal{O}\left(\sqrt{L\Delta}\eta + \frac{L\Delta\eta^2}{(\sqrt{L\Delta}-\eta)^2}\right)$
$\frac{4}{3}\sqrt{L\Delta} < \lambda \leq 4\sqrt{L\Delta}$ $(\sigma < \zeta_\lambda < \lambda)$	$\mathcal{O}\left(\sqrt{L\Delta}\zeta_\lambda + \frac{L\Delta\zeta_\lambda^2}{\lambda^2}\right)$	$4\sqrt{L\Delta} - 2\sigma$	$\mathcal{O}\left(\sqrt{\frac{\ln\frac{K}{\beta}}{K}} + \frac{\ln\frac{K}{\beta}}{K}\right)$	$\mathcal{O}\left(\sqrt{L\Delta}\sigma + \frac{L\Delta\sigma^2}{(\sqrt{L\Delta}-\sigma)^2}\right)$
$\lambda \leq \frac{4}{3}\sqrt{L\Delta}$ $(\lambda < \zeta_\lambda < \sigma)$	$\mathcal{O}\left(\sqrt{L\Delta}\frac{\sigma^\alpha\zeta_\lambda}{\lambda^\alpha} + \frac{L\Delta\sigma^{2\alpha}\zeta_\lambda^2}{\lambda^{2\alpha+2}}\right)$	$\frac{4}{3}\sqrt{L\Delta}$	$\mathcal{O}\left(\sqrt{\frac{\ln\frac{K}{\beta}}{K}} + \frac{\ln\frac{K}{\beta}}{K}\right)$	$\mathcal{O}\left(\frac{\sigma^\alpha}{(\sqrt{L\Delta})^{\alpha-2}} + \frac{\sigma^{2\alpha}}{(\sqrt{L\Delta})^{2\alpha-2}}\right)$
$\lambda \leq \frac{4}{3}\sqrt{L\Delta}$ $(\lambda < \sigma < \zeta_\lambda)$	$\mathcal{O}\left(\sqrt{L\Delta}\frac{\zeta_\lambda^{\alpha+1}}{\lambda^\alpha} + \frac{L\Delta\zeta_\lambda^{2\alpha}}{\lambda^{2\alpha+2}}\right)$	$\frac{4}{3}\sqrt{L\Delta} - \eta$	$\mathcal{O}\left(\sqrt{\frac{\ln\frac{K}{\beta}}{K}} + \frac{\ln\frac{K}{\beta}}{K}\right)$	$\mathcal{O}\left(\frac{\sqrt{L\Delta}(\sqrt{L\Delta}+\eta)^{\alpha+1}}{(\sqrt{L\Delta}-\eta)^\alpha} + \frac{L\Delta(\sqrt{L\Delta}+\eta)^{2\alpha}}{(\sqrt{L\Delta}-\eta)^{2\alpha+2}}\right)$
$\lambda \leq \frac{4}{3}\cdot 4\sqrt{L\Delta}$ $(\sigma < \lambda < \zeta_\lambda)$	$\mathcal{O}\left(\sqrt{L\Delta}\frac{\zeta_\lambda^{\alpha+1}}{\lambda^\alpha} + \frac{L\Delta\zeta_\lambda^{2\alpha+2}}{\lambda^{2\alpha+2}}\right)$	$\frac{4}{3}\sqrt{L\Delta} - \eta$	$\mathcal{O}\left(\sqrt{\frac{\ln\frac{K}{\beta}}{K}} + \frac{\ln\frac{K}{\beta}}{K}\right)$	$\mathcal{O}\left(\frac{\sqrt{L\Delta}(\sqrt{L\Delta}+\eta)^{\alpha+1}}{(\sqrt{L\Delta}-\eta)^\alpha} + \frac{L\Delta(\sqrt{L\Delta}+\eta)^{2\alpha}}{(\sqrt{L\Delta}-\eta)^{2\alpha+2}}\right)$
	$\mathcal{O}\left(\sqrt{L\Delta}\frac{\sigma\zeta_\lambda^{\alpha-1}}{\lambda^{\alpha-1}} + L\Delta\frac{\sigma^2\zeta_\lambda^{2\alpha-2}}{\lambda^{2\alpha}}\right)$	$\frac{4}{3}\sqrt{L\Delta}$	$\mathcal{O}\left(\sqrt{\frac{\ln\frac{K}{\beta}}{K}} + \frac{\ln\frac{K}{\beta}}{K}\right)$	$\mathcal{O}\left(\sqrt{L\Delta}\sigma + \sigma^2\right)$

cial case $\lambda > 4\sqrt{L\Delta}$, the bias terms (i.e., the convergence neighborhood when $\sigma_\omega = 0$) in (20) reduce to $\mathcal{O}\left(\sqrt{L\Delta}\frac{\sigma^\alpha}{\lambda^{\alpha-1}} + L\Delta\frac{\sigma^{2\alpha}}{\lambda^{2\alpha}}\right)$. Assuming $\lambda > \sigma$ for simplicity, the term from Koloskova et al. (2023) becomes $\mathcal{O}\left(\frac{\sigma^4}{\lambda^2}\right)$, which is strictly larger than the second term and strictly smaller than the first term in our bound when $\alpha = 2$. Furthermore, in this regime, both terms in our bound decrease with increasing α, suggesting that the convergence neighborhood grows with the heaviness of the noise. Whether the bound in (20) is tight and whether improvements are possible in other regimes remain open questions.

Finally, ignoring logarithmic factors (introduced by the high-probability analysis), the DP-noise-related terms in our bound (21) are $\tilde{\mathcal{O}}\left(\frac{\sigma_\omega\sqrt{dL\Delta}}{\sqrt{K}} + \frac{\sigma_\omega^2 dL\Delta}{\lambda^2 K}\right)$, while the corresponding terms in Koloskova et al. (2023) are $\mathcal{O}\left(\gamma L d\sigma_\omega^2 + \frac{\gamma^2 L^2 d^2 \sigma_\omega^4}{\lambda^2}\right)$. Setting $\gamma = \sqrt{\Delta/\sigma_\omega^2 LdK}$ yields the latter bound as $\mathcal{O}\left(\frac{\sigma_\omega\sqrt{dL\Delta}}{\sqrt{K}} + \frac{\sigma_\omega^2 dL\Delta}{\lambda^2 K}\right)$, which matches (21) up to logarithmic factors.

Proof sketch of our main results. The proofs of Theorems 4.1 and 4.2 follow the general template of Sadiev et al. (2023), but the main technical obstacle is different. In Sadiev et al. (2023), the clipping level is chosen as a function of the horizon K, which makes it possible to guarantee $\|\nabla f(x^t)\| \leq \frac{\lambda}{2}$ on the relevant high-probability event. Here, λ is fixed a priori, so this relation need not hold, and the standard bias-variance lemma for the clipped gradient is no longer applicable. We therefore prove Lemma B.1, which controls the bias and conditional second moment of $\texttt{clip}(X, \lambda)$ for an arbitrary λ through the slack quantity ζ_λ.

In the convex case, we first derive a one-step descent inequality (Lemma C.1) with the clipped-gradient coefficient $c_t = \min\{1, \lambda/(2\|\nabla f(x^t)\|)\}$ and the residual term $\theta_t = \hat{g}_t - c_t\nabla f(x^t)$. We then prove by induction that the iterates remain in a controlled region around $x^\star$. Inside this event, smoothness bounds the gradient norm, which turns Lemma B.1 into uniform bounds on the bias part θ_t^b and the conditional variance of the centered part θ_t^u. The stochastic sums involving θ_t^u and the Gaussian noise form bounded martingale difference sequences, so Bernstein-type inequalities together with Gaussian norm concentration control their cumulative contribution with probability at least $1-\beta$.

The non-convex argument is analogous, but it starts from a smoothness-based descent inequality for the function values rather than a distance recursion. After summing the descent bound and partitioning the indices according to whether $c_t = 1$ or $c_t < 1$, we convert the resulting weighted average estimate into either a best-iterate function suboptimality bound (convex case) or a best-iterate gradient-norm bound (non-convex case). The step-size conditions are chosen so that each deterministic bias term and each concentrated stochastic term fits within the induction budget, which yields the claimed $\widetilde{\mathcal{O}}(K^{-1/2})$ convergence to a neighborhood. □

5 CONCLUSIONS

In this paper, we present the first high-probability convergence analysis of DP-Clipped-SGD for both convex and non-convex smooth optimization under heavy-tailed noise. Our results show that DP-Clipped-SGD converges to a neighborhood of the optimum at the rate $\mathcal{O}(1/\sqrt{K})$. Promising directions for future work include extending the analysis to federated learning and sharpening the resulting neighborhood bounds.

References

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

Allouah, Y., Guerraoui, R., Gupta, N., Pinot, R., and Stephan, J. (2023). On the privacy-robustness-utility trilemma in distributed learning. In *International Conference on Machine Learning*, pages 569–626. PMLR.

Allouah, Y., Koloskova, A., El Firdoussi, A., Jaggi, M., and Guerraoui, R. (2024). The privacy power of correlated noise in decentralized learning. In *International Conference on Machine Learning*, pages 1115–1143. PMLR.

Armacki, A., Sharma, P., Joshi, G., Bajovic, D., Jakovetic, D., and Kar, S. (2023). High-probability convergence bounds for nonlinear stochastic gradient descent under heavy-tailed noise. *arXiv preprint arXiv:2310.18784*.

Armacki, A., Yu, S., Bajovic, D., Jakovetic, D., and Kar, S. (2024). Large deviations and improved mean-squared error rates of nonlinear sgd: Heavy-tailed noise and power of symmetry. *arXiv preprint arXiv:2410.15637*.

Chezhegov, S., Klyukin, Y., Semenov, A., Beznosikov, A., Gasnikov, A., Horváth, S., Takáč, M., and Gorbunov, E. (2024). Clipping improves adam-norm and adagrad-norm when the noise is heavy-tailed. *arXiv preprint arXiv:2406.04443*.

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

Das, R., Kale, S., Xu, Z., Zhang, T., and Sanghavi, S. (2023). Beyond uniform lipschitz condition in differentially private optimization. In *International Conference on Machine Learning*, pages 7066–7101. PMLR.

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

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

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

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

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

Freedman, D. A. (1975). On tail probabilities for martingales. *the Annals of Probability*, pages 100–118.

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

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

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

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

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

Gorbunov, E., Danilova, M., Shibaev, I., Dvurechensky, P., and Gasnikov, A. (2024a). High-probability complexity bounds for non-smooth stochastic convex optimization with heavy-tailed noise. *Journal of Optimization Theory and Applications*, pages 1–60.

Gorbunov, E., Sadiev, A., Danilova, M., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A., and Richtárik, P. (2024b). High-probability convergence for composite and distributed stochastic minimization and variational inequalities with heavy-tailed noise. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 15951–16070. PMLR.

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

Hübler, F., Fatkhullin, I., and He, N. (2024). From gradient clipping to normalization for heavy tailed sgd. *arXiv preprint arXiv:2410.13849*.

Islamov, R., Horvath, S., Lucchi, A., Richtarik, P., and Gorbunov, E. (2025). Double momentum and error feedback for clipping with fast rates and differential privacy. *arXiv preprint arXiv:2502.11682*.

Jin, C., Zhou, K., Han, B., Cheng, J., and Zeng, T. (2024). Efficient private sco for heavy-tailed data via averaged clipping. *Machine Learning*, 113(11):8487–8532.

Juditsky, A. and Nemirovski, A. S. (2008). Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*.

Kairouz, P., Oh, S., and Viswanath, P. (2015). The composition theorem for differential privacy. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1376–1385, Lille, France. PMLR.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koloskova, A., Hendrikx, H., and Stich, S. U. (2023). Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, pages 17343–17363. PMLR.

Kornilov, N., Shamir, O., Lobanov, A., Dvinskikh, D., Gasnikov, A., Shibaev, I., Gorbunov, E., and Horváth, S. (2023). Accelerated zeroth-order method for non-smooth stochastic convex optimization problem with infinite variance. *Advances in Neural Information Processing Systems*, 36:64083–64102.

Kornilov, N., Zmushko, P., Semenov, A., Gasnikov, A., and Beznosikov, A. (2025). Sign operator for coping with heavy-tailed noise: High probability convergence bounds with extensions to distributed optimization and comparison oracle. *arXiv preprint arXiv:2502.07923*.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pages 1302–1338.

Li, B. and Chi, Y. (2025). Convergence and privacy of decentralized nonconvex optimization with gradient clipping and communication compression. *IEEE Journal of Selected Topics in Signal Processing.*

Li, S. and Liu, Y. (2023). High probability analysis for non-convex stochastic optimization with clipping. In *ECAI 2023*, pages 1406–1413. IOS Press.

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

Liu, M., Zhuang, Z., Lei, Y., and Liao, C. (2022). A communication-efficient distributed gradient clipping algorithm for training deep neural networks. *Advances in Neural Information Processing Systems*, 35:26204–26217.

Liu, Z., Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. (2023). High probability convergence of stochastic gradient methods. In *International Conference on Machine Learning*, pages 21884–21914. PMLR.

Madden, L., Dall'Anese, E., and Becker, S. (2024). High probability convergence bounds for non-convex stochastic gradient descent with sub-weibull noise. *Journal of Machine Learning Research*, 25(241):1–36.

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

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

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

Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. (2023). Improved convergence in high probability of clipped gradient methods with heavy tailed noise.

Noble, M., Bellet, A., and Dieuleveut, A. (2022). Differentially private federated learning on heterogeneous data. In *International conference on artificial intelligence and statistics*, pages 10110–10145. PMLR.

Parletta, D. A., Paudice, A., Pontil, M., and Salzo, S. (2024). High probability bounds for stochastic subgradient schemes with heavy tailed noise. *SIAM Journal on Mathematics of Data Science*, 6(4):953–977.

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

Polyak, B. T. (1964). Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17.

Polyanskiy, Y. and Wu, Y. (2025). *Information theory: From coding to learning*. Cambridge university press.

Puchkin, N., Gorbunov, E., Kutuzov, N., and Gasnikov, A. (2024). Breaking the heavy-tailed noise barrier in stochastic optimization problems. In *International Conference on Artificial Intelligence and Statistics*, pages 856–864. PMLR.

Richtárik, P., Sokolov, I., and Fatkhullin, I. (2021). EF21: A new, simpler, theoretically better, and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34:4384–4396.

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

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

Seide, F., Fu, H., Droppo, J., Li, G., and Yu, D. (2014). 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Interspeech*, volume 2014, pages 1058–1062. Singapore.

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

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

Streeter, M. and McMahan, H. B. (2010). Less regret via online conditioning. *arXiv preprint arXiv:1002.4862*.

Su, W. J. (2024). A statistical viewpoint on differential privacy: Hypothesis testing, representation, and blackwell's theorem. *Annual Review of Statistics and Its Application*, 12.

Wang, D., Xiao, H., Devadas, S., and Xu, J. (2020). On differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, pages 10081–10091. PMLR.

Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., Jin, S., Quek, T. Q., and Poor, H. V. (2020). Federated learning with differential privacy: Algorithms and performance analysis. *IEEE transactions on information forensics and security*, 15:3454–3469.

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

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

Zhang, X., Chen, X., Hong, M., Wu, Z. S., and Yi, J. (2022). Understanding clipping for federated learning: Convergence and client-level differential privacy. In *International Conference on Machine Learning, ICML 2022*.

Zhao, P., Wu, J., Liu, Z., Wang, C., Fan, R., and Li, Q. (2025). Differential private stochastic optimization with heavy-tailed data: towards optimal rates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22795–22803.

Zhivotovskiy, N. (2024). Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Electronic Journal of Probability*, 29:1–28.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

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

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Notation Table and Auxiliary Facts

To facilitate the readability of the proofs, we provide a notation table below.

Table 3: Our notation.

Notation	Explanation
g_t	Stochastic gradient
$\hat{g}_t$	Clipped stochastic gradient
$\tilde{g}_t$	Clipped stochastic gradient after DP noise injection
c_t	$\min\left\{1, \frac{\lambda}{2\|\nabla f(x^t)\|}\right\}$
ω_t	Injected DP noise at iteration t
β	Confidence level/failure probability
ζ_λ	Convex case: $\max\left\{0, 2LR - \frac{\lambda}{2}\right\}$ Non-convex case: $\max\left\{0, 2\sqrt{L\Delta} - \frac{\lambda}{2}\right\}$
$\mathcal{F}^t$	Filtration up to the time t
σ	Gradient noise parameter
σ_ω	DP noise parameter
R	Upper bound on $\left\|x^0 - x^*\right\|$ for convex functions
Δ	Upper bound on $f(x^0) - f^*$ for non-convex functions

Auxiliary facts. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A sequence $\{\mathcal{F}_i\}_{i\geq 1}$ of nested sigma algebras in $\mathcal{F}$ (i.e., $\mathcal{F}_i \subset \mathcal{F}_{i+1} \subset \mathcal{F}$) is called a filtration, in which case $(\Omega, \mathcal{F}, \{\mathcal{F}_i\}_{i\geq 1}, \mathbb{P})$ is called a filtered probability space. A sequence of random variables $\{X_i\}_{i\geq 1}$ is said to be adapted to $\{\mathcal{F}_i\}_{i\geq 1}$ if each X_i is $\mathcal{F}_i$-measurable. Furthermore, if $\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] = X_{i-1}\ \forall i$, then $\{X_i\}_{i\geq 1}$ is called a martingale. On the other hand, if $\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] = 0\ \forall i$, then $\{X_i\}_{i\geq 1}$ is called a martingale difference sequence.

One of the very useful tools in establishing high probability convergence guarantees in this work is the following lemma, which is known as the Bernstein inequality for martingale difference sequences (Freedman, 1975), (Dzhaparidze and Van Zanten, 2001).

Lemma A.1. *Let the sequence of random variables $\{X_i\}_{i\geq 1}$ form a martingale difference sequence on the filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_i\}_{i\geq 1}, \mathbb{P})$. Assume that conditional variances $\sigma_i^2 := \mathbb{E}\left[X_i^2|\mathcal{F}_{i-1}\right]$ exist and are bounded. Furthermore, there exists a deterministic constant $c \geq 0$ such that $|X_i| \leq c$ almost surely for all $i \geq 0$. Then for all $b > 0$, $G > 0$ and $n \geq 1$*

$$\mathbb{P}\left\{\left|\sum_{i=1}^n X_i\right| > b \ \text{ and } \ \sum_{i=1}^n \sigma_i^2 \leq G\right\} \leq 2\exp\left(-\frac{b^2}{2G + {}^{2bc}/_3}\right). \tag{22}$$

Lemma A.2. *(Corollary of Theorem 2.1, item (ii) from (Juditsky and Nemirovski, 2008)) Let $\{\xi_k\}_{k=1}^N$ be a sequence of random vectors in $\mathbb{R}^n$ such that*

$$\mathbb{E}\left[\xi_k|\mathcal{F}_{k-1}\right] = 0 \ \text{ almost surely}, \quad k = 1, ..., N\ .$$

Define $S_N := \sum_{k=1}^N \xi_k$. Assume that the sequence $\{\xi_k\}_{k=1}^N$ satisfies the following light-tail condition

$$\mathbb{E}\left[\exp\left(\frac{\|\xi_k\|^2}{\sigma_k^2}\right) \mid \mathcal{F}_{k-1}\right] \leq \exp(1) \ \text{ almost surely}, \quad k = 1, ..., N \tag{23}$$

where $\sigma_1, ..., \sigma_N$ are some positive numbers. Then for all $\phi \geq 0$, we have

$$\mathbb{P}\left\{\|S_N\|_2 \geq (\sqrt{2} + \sqrt{2}\phi)\sqrt{\sum_{k=1}^N \sigma_k^2}\right\} \leq \exp\left(-\frac{\phi^2}{3}\right). \tag{24}$$

Lemma A.3 (Lemma 1 from (Laurent and Massart, 2000))**.** *Let $\{Y_i\}_{i=1}^n$ be i.i.d. Gaussian variables, with mean 0 and variance 1. Let $\{a_i\}_{i=1}^n$ be nonnegative constants. Define*

$$\|a\|_\infty = \sup_{i=1,\dots n} |a_i|\,, \quad \|a\|_2^2 = \sum_{i=1}^n a_i^2.$$

Let

$$X = \sum_{i=1}^n a_i \left(Y_i^2 - 1\right).$$

Then the following inequalities hold for any positive t:

$$\mathbb{P}\left\{X \geq 2\|a\|_2\sqrt{t} + 2\|a\|_\infty t\right\} \leq \exp(-t), \tag{25}$$

$$\mathbb{P}\left\{X \leq -2\|a\|_2\sqrt{t}\right\} \leq \exp(-t). \tag{26}$$

Lemma A.4 (Remark 2.8 from (Zhivotovskiy, 2024); see also example 4.3 from (Polyanskiy and Wu, 2025))**.** *Let X be a zero-mean sub-Gaussian random vector in $\mathbb{R}^d$ with covariance matrix Σ. Then the norm of this vector can be bounded in probability as below*

$$\mathbb{P}\left\{\|X\|_2 > \sqrt{\operatorname{tr}(\Sigma)} + \sqrt{2\|\Sigma\|_2 \ln\frac{1}{\delta}}\right\} \leq \delta. \tag{27}$$

B Bound for the Bias and Variance of Clipped Estimator

Lemma B.1. *Let X be a random vector from $\mathbb{R}^d$. We define the random vector $\hat{X} := \text{clip}(X, \lambda)$ for an arbitrary clipping level $\lambda > 0$. Let us assume*

$$\mathbb{E}[X] = x, \qquad \mathbb{E}[\|X - x\|^\alpha] \le \sigma^\alpha,$$

where $\sigma > 0$ is bounded, $\alpha \in (1, 2]$, and we also define $\hat{x} := \text{clip}(x, \lambda/2)$. Then, the following inequalities hold:

$$\begin{aligned}\left\|\mathbb{E}[\hat{X}] - \hat{x}\right\| \le & \frac{2^{2\alpha-1}\sigma\left(\sigma^\alpha + (\max\{0, \|x\| - \lambda/2\})^\alpha\right)^{\frac{\alpha-1}{\alpha}}}{\lambda^{\alpha-1}} \\ & + \max\{\|x\|, \lambda/2\}\frac{2^{2\alpha-1}\left(\sigma^\alpha + (\max\{0, \|x\| - \lambda/2\})^\alpha\right)}{\lambda^\alpha} \\ & + \max\{0, \|x\| - \lambda/2\}, \end{aligned} \tag{28}$$

$$\mathbb{E}\left\|\hat{X} - \mathbb{E}\hat{X}\right\|^2 \le \frac{9(2^{2\alpha-1}+1)\lambda^{2-\alpha}\sigma^\alpha}{4} + \frac{9(2^{2\alpha-1}+1)\lambda^{2-\alpha}(\max\{0, \|x\| - \lambda/2\})^\alpha}{4}. \tag{29}$$

Proof. The proof technique is similar to the proof of Lemma 5.1 from (Sadiev et al., 2023). Define random variables χ and η as

$$\chi = \mathbb{I}_{\{\|X\| > \lambda\}}, \qquad \eta = \mathbb{I}_{\{\|X - \hat{x}\| > \lambda/2\}}.$$

Since $\|X\| \le \|\hat{x}\| + \|X - \hat{x}\| \le \frac{\lambda}{2} + \|X - \hat{x}\|$, we get $\chi \le \eta$. Moreover, note that

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

Proof of (28). For the bias term, we obtain

$$\begin{aligned}\left\|\mathbb{E}\hat{X} - \hat{x}\right\| &= \left\|\mathbb{E}\left(X + \chi\left(\frac{\lambda}{\|X\|} - 1\right)X - \min\left\{1, \frac{\lambda}{2\|x\|}\right\}x\right)\right\| \\ &\le \left\|\mathbb{E}[\chi\left(\frac{\lambda}{\|X\|} - 1\right)X]\right\| + \left(1 - \min\left\{1, \frac{\lambda}{2\|x\|}\right\}\right)\|x\| \\ &= \left\|\mathbb{E}[\chi\left(\frac{\lambda}{\|X\|} - 1\right)X]\right\| + \max\left\{0, \|x\| - \frac{\lambda}{2}\right\} \\ &\le \mathbb{E}\left[\left|\chi\left(\frac{\lambda}{\|X\|} - 1\right)\right|\|X\|\right] + \max\left\{0, \|x\| - \frac{\lambda}{2}\right\} \\ &\overset{(i)}{\le} \mathbb{E}\left[\chi\|X\|\right] + \max\left\{0, \|x\| - \frac{\lambda}{2}\right\},\end{aligned}$$

where in (i), we used the fact that $\chi \in \{0, 1\}$ and when $\chi = 1$ we have $\left|\frac{\lambda}{\|X\|} - 1\right| = 1 - \frac{\lambda}{\|X\|} \le 1$. Then, we continue the derivation as follows:

$$\begin{aligned}\left\|\mathbb{E}\hat{X} - \hat{x}\right\| &\le \mathbb{E}\left[\chi\|X\|\right] + \max\left\{0, \|x\| - \frac{\lambda}{2}\right\} \\ &\overset{\chi\le\eta}{\le} \mathbb{E}\left[\eta\|X\|\right] + \max\left\{0, \|x\| - \frac{\lambda}{2}\right\} \\ &\le \mathbb{E}\left[\eta\|X - x\|\right] + \mathbb{E}\left[\eta\|x\|\right] + \max\left\{0, \|x\| - \frac{\lambda}{2}\right\} \\ &\overset{(i)}{\le} \left(\mathbb{E}\|X - x\|^\alpha\right)^{1/\alpha}\left(\mathbb{E}[\eta^{\alpha/\alpha-1}]\right)^{(\alpha-1)/\alpha} + \mathbb{E}\eta\|x\| + \max\{0, \|x\| - \lambda/2\},\end{aligned} \tag{30}$$

where in (i), we used Hölder inequality. Moreover, due to Markov's inequality, we also have

$$\mathbb{E}[\eta^{\alpha/\alpha-1}] = \mathbb{E}\eta = \mathbb{P}\{\|X - \hat{x}\| > \lambda/2\} = \mathbb{P}\left\{\|X - \hat{x}\|^\alpha > (\lambda/2)^\alpha\right\} \le \frac{2^\alpha\mathbb{E}\|X - \hat{x}\|^\alpha}{\lambda^\alpha}. \tag{31}$$

Then, the expected value from the right-hand side (RHS) of (31) can be decomposed as follows

$$\begin{aligned}\mathbb{E}\|X-\hat{x}\|^{\alpha} &= \mathbb{E}\left\|X-x+x-\hat{x}\right\|^{\alpha} \leq 2^{\alpha-1}(\mathbb{E}\|X-x\|^{\alpha}+\max\{0,\|x\|-\lambda/2\}^{\alpha}) \\ &\leq 2^{\alpha-1}(\sigma^{\alpha}+\max\{0,\|x\|-\lambda/2\}^{\alpha}),\end{aligned} \tag{32}$$

where we use the Jensen's inequality for the convex function $\|x\|^{\alpha}$. After substitution of (32) into (31), we get

$$\mathbb{E}[\eta^{\alpha/\alpha-1}] = \mathbb{E}\eta \leq \frac{2^{2\alpha-1}(\sigma^{\alpha}+\max\{0,\|x\|-\lambda/2\}^{\alpha})}{\lambda^{\alpha}}. \tag{33}$$

Plugging the above bound in (30), we derive

$$\begin{aligned}\left\|\mathbb{E}\hat{X}-\hat{x}\right\| \leq{}& \sigma\left(\frac{2^{2\alpha-1}(\sigma^{\alpha}+\max\{0,\|x\|-\lambda/2\}^{\alpha})}{\lambda^{\alpha}}\right)^{\frac{\alpha-1}{\alpha}} + \|x\|\frac{2^{2\alpha-1}(\sigma^{\alpha}+\max\{0,\|x\|-\lambda/2\}^{\alpha})}{\lambda^{\alpha}} \\ &+\max\{0,\|x\|-\lambda/2\}.\end{aligned}$$

Using that $\frac{\alpha-1}{\alpha}\leq 1$ and $\|x\|\leq\max\{\|x\|,\lambda/2\}$, we conclude the proof of the result for the bias term, i.e., bound (28).

Proof of (29)**.** First, we use the following standard inequality:

$$\mathbb{E}\left\|\hat{X}-\mathbb{E}\hat{X}\right\|^{2} \leq \mathbb{E}\left\|\hat{X}-\hat{x}\right\|^{2}.$$

Then, we bound the RHS as

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

Applying upper bounds (32) and (33) from the previous part of the proof, we obtain

$$\begin{aligned}\mathbb{E}\left\|\hat{X}-\hat{x}\right\|^{2} \leq{}& \left(\frac{3\lambda}{2}\right)^{2}\frac{2^{2\alpha-1}(\sigma^{\alpha}+\max\{0,\|x\|-\lambda/2\}^{\alpha})}{\lambda^{\alpha}} \\ &+\left(\frac{3\lambda}{2}\right)^{2-\alpha}2^{\alpha-1}(\sigma^{\alpha}+\max\{0,\|x\|-\lambda/2\}^{\alpha}) \\ ={}& \frac{9\cdot(2^{2\alpha-1}+1)\lambda^{2-\alpha}\sigma^{\alpha}}{4}+\frac{9\cdot(2^{2\alpha-1}+1)\lambda^{2-\alpha}(\max\{0,\|x\|-\lambda/2\})^{\alpha}}{4},\end{aligned}$$

which concludes the proof. □

C Missing Proofs: Convex Case

We start the analysis with the following lemma. This lemma follows the proof of deterministic GD and separates the stochastic part from the deterministic part of Clipped-SGD.

Lemma C.1. *Let Assumptions 2.1, 2.2, and 2.3, and hold for $Q = B_{2R}(x^\star)$, where $R \geq \|x^0 - x^\star\|$ and $0 < \gamma \leq 1/8L$. If $x^k \in Q$ for all $k = 0, 1, \ldots, K$ for some $K \geq 0$, then for any $0 \leq T \leq K$ the iterates produced by* DP-Clipped-SGD *satisfy*

$$\frac{\gamma}{T+1}\sum_{t=0}^{T} c_t(f(x^t) - f^\star) \leq \frac{\left\|x^0 - x^\star\right\|^2 - \left\|x^{T+1} - x^\star\right\|^2}{T+1} - \frac{2\gamma}{T+1}\sum_{t=0}^{T}\langle x^t - x^\star, \theta_t\rangle - \frac{2\gamma}{T+1}\sum_{t=0}^{T}\langle x^t - x^\star, \omega_t\rangle + \frac{2\gamma^2}{T+1}\sum_{t=0}^{T}\|\theta_t\|^2 + \frac{4\gamma^2}{T+1}\sum_{t=0}^{T}\|\omega_t\|^2,$$

where we have defined

$$c_t \stackrel{\text{def}}{=} \min\left\{1, \frac{\lambda}{2\left\|\nabla f(x^t)\right\|}\right\}, \tag{34}$$

$$\theta_t \stackrel{\text{def}}{=} \hat{g}_t - c_t \nabla f(x^t). \tag{35}$$

Proof. Since $x^{t+1} = x^t - \gamma \tilde{g}_t$, the following set of inequalities hold for all $t = 0, 1, \ldots, K$:

$$\begin{aligned}
\left\|x^{t+1} - x^\star\right\|^2 &= \left\|x^t - x^\star\right\|^2 - 2\gamma\langle x^t - x^\star, \tilde{g}_t\rangle + \gamma^2\left\|\tilde{g}_t\right\|^2 \\
&= \left\|x^t - x^\star\right\|^2 - 2\gamma\langle x^t - x^\star, \hat{g}_t + \omega_t\rangle + \gamma^2\left\|\hat{g}_t + \omega_t\right\|^2 \\
&= \left\|x^t - x^\star\right\|^2 - 2\gamma\langle x^t - x^\star, \hat{g}_t + \omega_t + c_t\nabla f(x^t) - c_t\nabla f(x^t)\rangle \\
&\quad + \gamma^2\left\|\hat{g}_t + \omega_t + c_t\nabla f(x^t) - c_t\nabla f(x^t)\right\|^2 \\
&\leq \left\|x^t - x^\star\right\|^2 - 2\gamma\langle x^t - x^\star, \theta_t + \omega_t\rangle - 2\gamma c_t\langle x^t - x^\star, \nabla f(x^t)\rangle + 2\gamma^2\left\|\theta_t\right\|^2 \\
&\quad + 4\gamma^2\left\|\omega_t\right\|^2 + 4\gamma^2 c_t^2\left\|\nabla f(x^t)\right\|^2 \\
&\leq \left\|x^t - x^\star\right\|^2 - 2\gamma\langle x^t - x^\star, \theta_t + \omega_t\rangle - 2\gamma c_t(f(x^t) - f^\star) + 2\gamma^2\left\|\theta_t\right\|^2 \\
&\quad + 4\gamma^2\left\|\omega_t\right\|^2 + 8\gamma^2 c_t^2 L(f(x^t) - f^\star) \\
&= \left\|x^t - x^\star\right\|^2 - 2\gamma\langle x^t - x^\star, \theta_t + \omega_t\rangle - (2\gamma - 8\gamma^2 L)c_t(f(x^t) - f^\star) + 2\gamma^2\left\|\theta_t\right\|^2 + 4\gamma^2\left\|\omega_t\right\|^2.
\end{aligned}$$

First, we rearrange the terms, and utilize the inequalities $\gamma \leq 1/8L$ and $c_t^2 \leq c_t$. Upon summing over $t = 0, 1, \ldots, T$, we obtain the following inequality

$$\frac{\gamma}{T+1}\sum_{t=0}^{T} c_t(f(x^t) - f^\star) \quad \leq \quad \frac{\left\|x^0 - x^\star\right\|^2 - \left\|x^{T+1} - x^\star\right\|^2}{T+1} - \frac{2\gamma}{T+1}\sum_{t=0}^{T}\langle x^t - x^\star, \theta_t\rangle - \frac{2\gamma}{T+1}\sum_{t=0}^{T}\langle x^t - x^\star, \omega_t\rangle + \frac{2\gamma^2}{T+1}\sum_{t=0}^{T}\|\theta_t\|^2 + \frac{4\gamma^2}{T+1}\sum_{t=0}^{T}\|\omega_t\|^2,$$

which concludes the proof. □

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

Theorem C.1. *Let Assumptions 2.1, 2.2, 2.3, and 2.4 hold for $Q = B_{2R}(x^\star)$, where R is such that $R \geq \|x^0 - x^\star\|$. Let $\zeta_\lambda \stackrel{\text{def}}{=} \max\{0, 2LR - \frac{\lambda}{2}\}$, and $\gamma \leq \min\{1/8L, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6\}$, where*

$$\gamma_1 \quad \stackrel{\text{def}}{=} \quad \frac{R}{42(2^{2\alpha-1}+1)^{1/2}\sigma^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{6(K+1)\ln\frac{8(K+1)}{\beta}\left(1 + \frac{\zeta_\lambda^\alpha}{\sigma^\alpha}\right)}}, \tag{36}$$

$$\gamma_2 \overset{\text{def}}{=} \frac{R\lambda^{\alpha-1}}{28(K+1)2^{2\alpha-1}\sigma^\alpha\left(1+\frac{\zeta_\lambda^\alpha}{\sigma^\alpha}\right)\left(\frac{\zeta_\lambda}{\lambda}+\frac{1}{2}+\frac{\lambda^{\alpha-1}\zeta_\lambda}{2^{2\alpha-1}(\sigma^\alpha+\zeta_\lambda^\alpha)}+\left(1+\frac{\zeta_\lambda^\alpha}{\sigma^\alpha}\right)^{-1/\alpha}\right)}, \tag{37}$$

$$\gamma_3 \overset{\text{def}}{=} \frac{R}{56\sigma_\omega\sqrt{d(K+1)}(\sqrt{2}+\sqrt{2}\phi)}, \tag{38}$$

$$\gamma_4 \overset{\text{def}}{=} \frac{(2-\sqrt{2})R}{\lambda+\sigma_\omega\left(\sqrt{d}+\sqrt{2\ln\left(\frac{K+1}{\beta}\right)}\right)}, \tag{39}$$

$$\gamma_5 \overset{\text{def}}{=} \frac{R}{56\lambda\ln\frac{8(K+1)}{\beta}}, \tag{40}$$

$$\gamma_6 \overset{\text{def}}{=} \frac{R}{2\sigma_w\sqrt{7\left[(K+1)d+2\sqrt{(K+1)d\ln\frac{4(K+1)}{\beta}}+2\ln\frac{4(K+1)}{\beta}\right]}}. \tag{41}$$

with $\phi \overset{\text{def}}{=} \sqrt{3\ln\frac{4(K+1)}{\beta}}$ *for some* $K>0$ *and* $\beta\in(0,1]$. *Then, after* K *iterations of* DP-Clipped-SGD*, the iterates with probability at least* $1-\beta$ *satisfy*

$$\min_{k\in[0,K]} f(x^k)-f(x^\star) \le \frac{4R^2}{\gamma(K+1)}+\frac{64LR^4}{\lambda^2\gamma^2(K+1)^2} \quad \textit{and} \quad \{x^k\}_{k=0}^K \subseteq B_{\sqrt{2}R}(x^\star). \tag{42}$$

Proof. Let $R_k \overset{\text{def}}{=} \|x^k-x^\star\|$ for all $k\ge 0$. Next, our goal is to show by induction that $R_k \le 2R$ for all $k=0,1,\ldots,K$ with high probability, which allows us to apply the result of Lemma C.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

$$-2\gamma\sum_{l=0}^{t-1}\langle x^l-x^\star,\theta_l\rangle-2\gamma\sum_{l=0}^{t-1}\langle x^l-x^\star,\omega_l\rangle+2\gamma^2\sum_{l=0}^{t-1}\|\theta_l\|^2+4\gamma^2\sum_{l=0}^{t-1}\|\omega_l\|^2 \le R^2, \tag{43}$$

$$R_t \le \sqrt{2}R, \tag{44}$$

$$\|\omega_t\| \le \sigma_\omega\left(\sqrt{d}+\sqrt{2\ln\left(\frac{K+1}{(t+1)\beta}\right)}\right), \tag{45}$$

hold for all $t=0,1,\ldots,k$ simultaneously. We want to prove via induction that $\mathbb{P}\{E_k\}\ge 1-{}^{(k+1)\beta}/_{(K+1)}$ for all $k=0,1,\ldots,K$. For $k=0$ the statements (43) and (44) trivially hold. Given Lemma A.4, statement (45) will also hold. Assume that the statement is true for some $k=T-1\le K$: $\mathbb{P}\{E_{T-1}\}\ge 1-{}^{T\beta}/_{(K+1)}$. One needs to prove that $\mathbb{P}\{E_T\}\ge 1-{}^{(T+1)\beta}/_{(K+1)}$. First, we notice that probability event E_{T-1} implies that $x_t\in B_{\sqrt{2}R}(x^\star)$ for all $t=0,1,\ldots,T-1$. For x^T, we can obtain the following inequalities

$$\begin{aligned}\|x^T-x^\star\| &= \|x^{T-1}-x^\star-\gamma\widetilde{g}_{T-1}\| \le \|x^{T-1}-x^\star\|+\gamma\|\hat{g}_{T-1}\|+\gamma\|\omega_{T-1}\|\\ &\le \sqrt{2}R+\gamma\lambda+\gamma\sigma_\omega\left(\sqrt{d}+\sqrt{2\ln\left(\frac{K+1}{T\beta}\right)}\right) \overset{(39)}{\le} 2R.\end{aligned} \tag{46}$$

This means that $x^0,x^1,\ldots,x^T\in B_{2R}(x^\star)$. Therefore, E_{T-1} implies $\{x^k\}_{k=0}^T\subseteq Q$, meaning that the assumptions of Lemma C.1 are satisfied. Subsequently, the following inequality holds

$$\begin{aligned}\frac{\gamma}{t}\sum_{l=0}^{t-1}c_l\left(f(x^l)-f(x^\star)\right) &\le \frac{\|x^0-x^\star\|^2-\|x^t-x^\star\|^2}{t}+\frac{4\gamma^2}{t}\sum_{l=0}^{t-1}\|\omega_l\|^2\\ &\quad -\frac{2\gamma}{t}\sum_{l=0}^{t-1}\langle x^l-x^\star,\theta_l+\omega_l\rangle+\frac{2\gamma^2}{t}\sum_{l=0}^{t-1}\|\theta_l\|^2,\end{aligned} \tag{47}$$

for all $t=1,\ldots,T$ simultaneously. For all $t=1,\ldots,T-1$ this event also implies

$$\gamma\sum_{l=0}^{t-1}c_l(f(x^l)-f(x^\star)) \le R^2-2\gamma\sum_{l=0}^{t-1}\langle x^l-x^\star,\theta_l\rangle-2\gamma\sum_{l=0}^{t-1}\langle x^l-x^\star,\omega_l\rangle+2\gamma^2\sum_{l=0}^{t-1}\|\theta_l\|^2$$

$$\begin{aligned} &+ \quad 4\gamma^2 \sum_{l=0}^{t-1} \|\omega_l\|^2 \\ &\leq \quad 2R^2, \end{aligned} \tag{48}$$

where we have used (43) for E_{T-1}. Taking into account that $\sum_{l=0}^{t-1} c_l(f(x^l) - f(x^\star)) \geq 0$, (47) implies

$$R_T^2 \leq R^2 - 2\gamma \sum_{t=0}^{T-1} \langle x^t - x^\star, \theta_t \rangle - 2\gamma \sum_{t=0}^{T-1} \langle x^t - x^\star, \omega_t \rangle + 2\gamma^2 \sum_{t=0}^{T-1} \|\theta_t\|^2 + 4\gamma^2 \sum_{t=0}^{T-1} \|\omega_t\|^2. \tag{49}$$

Next, we define random vectors

$$\eta_t \overset{\text{def}}{=} \begin{cases} x^t - x^\star, & \text{if } \|x^t - x^\star\| \leq 2R, \\ 0, & \text{otherwise,} \end{cases}$$

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

$$\|\eta_t\| \leq 2R. \tag{50}$$

Next, we introduce the following vectors

$$\theta_t^u \overset{\text{def}}{=} \hat{g}_t - \mathbb{E}\left[\hat{g}_t \mid \mathcal{F}^{t-1}\right], \quad \theta_t^b \overset{\text{def}}{=} \mathbb{E}\left[\hat{g}_t \mid \mathcal{F}^{t-1}\right] - c_t \nabla f(x^t) \tag{51}$$

Using the above notation, we notice that $\theta_t = \theta_t^u + \theta_t^b$. Subsequently, E_{T-1} implies

$$\begin{aligned} R_T^2 \quad \leq \quad & R^2 \underbrace{-2\gamma \sum_{t=0}^{T-1} \langle \theta_t^u, \eta_t \rangle}_{①} \underbrace{-2\gamma \sum_{t=0}^{T-1} \langle \theta_t^b, \eta_t \rangle}_{②} \underbrace{-2\gamma \sum_{t=0}^{T-1} \langle \omega_l, \eta_t \rangle}_{③} + \underbrace{4\gamma^2 \sum_{t=0}^{T-1} \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right]}_{④} \\ & \underbrace{+ 4\gamma^2 \sum_{t=0}^{T-1} \left(\|\theta_t^u\|^2 - \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right]\right)}_{⑤} + \underbrace{4\gamma^2 \sum_{t=0}^{T-1} \left\|\theta_t^b\right\|^2}_{⑥} + \underbrace{4\gamma^2 \sum_{t=0}^{T-1} \|\omega_t\|^2}_{⑦}. \end{aligned} \tag{52}$$

To finish our inductive proof we need to show that ① + ② + ③ + ④ + ⑤ + ⑥ + ⑦ $\leq R^2$ with high probability. In the subsequent parts of the proof, we will utilize the bounds for the norm and norm squared moments of θ_t^u and θ_t^b. First, by definition of clipping operator and Lemma B.1 we have

$$\|\theta_t^u\| \leq 2\lambda, \tag{53}$$

and

$$\begin{aligned} \|\theta_t^b\| \qquad \leq \; & \frac{2^{2\alpha-1}\sigma\left(\sigma^\alpha + (\max\{0, \|\nabla f(x^t)\| - \lambda/2\})^\alpha\right)^{\frac{\alpha-1}{\alpha}}}{\lambda^{\alpha-1}} \\ & + \max\{\left\|\nabla f(x^t)\right\|, \lambda/2\} \frac{2^{2\alpha-1}\left(\sigma^\alpha + (\max\{0, \|\nabla f(x^t)\| - \lambda/2\})^\alpha\right)}{\lambda^\alpha} \\ & + \max\{0, \left\|\nabla f(x^t)\right\| - \lambda/2\}, \end{aligned} \tag{54}$$

$$\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right] \leq \frac{9(2^{2\alpha-1}+1)\lambda^{2-\alpha}\sigma^\alpha}{4} + \frac{9(2^{2\alpha-1}+1)\lambda^{2-\alpha}(\max\{0, \|\nabla f(x^t)\| - \lambda/2\})^\alpha}{4}. \tag{55}$$

As can be seen, these bounds are iteration-dependent due to the presence of $\|\nabla f(x^t)\|$. As a remedy, we bound $\|\nabla f(x^t)\|$ by $2LR$ inside event E_{T-1}. This bound can be obtained from a combination of Assumption 2.2, E_{T-1}, and (46). Next, we introduce a new variable $\zeta_\lambda := \max\{0, 2LR - \frac{\lambda}{2}\}$. Thus, we get the following bounds for the bias and variance of θ_t: E_{T-1} implies

$$\|\theta_t^b\| \qquad \leq \frac{2^{2\alpha-1}\sigma\left(\sigma^\alpha + \zeta_\lambda^\alpha\right)^{\frac{\alpha-1}{\alpha}}}{\lambda^{\alpha-1}} + \left(\zeta_\lambda + \frac{\lambda}{2}\right)\frac{2^{2\alpha-1}\left(\sigma^\alpha + \zeta_\lambda^\alpha\right)}{\lambda^\alpha} + \zeta_\lambda, \tag{56}$$

$$\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right] \leq \frac{9(2^{2\alpha-1}+1)\lambda^{2-\alpha}\sigma^\alpha}{4} + \frac{9(2^{2\alpha-1}+1)\lambda^{2-\alpha}\zeta_\lambda^\alpha}{4} \tag{57}$$

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

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

$$\mathbb{E}\left[-2\gamma\langle\theta_t^u, \eta_t\rangle \mid \mathcal{F}^{t-1}\right] = 0.$$

Furthermore, ① is bounded with probability 1 as

$$|2\gamma\left\langle\theta_t^u, \eta_t\right\rangle| \leq 2\gamma\|\theta_t^u\|\cdot\|\eta_t\| \overset{(53),(50)}{\leq} 8\gamma\lambda R \overset{(40)}{\leq} \frac{R^2}{7\ln\frac{8(K+1)}{\beta}} \overset{\text{def}}{=} c. \tag{58}$$

The summands also have bounded conditional variances $\sigma_t^2 \overset{\text{def}}{=} \mathbb{E}[4\gamma^2\langle\theta_t^u, \eta_t\rangle^2 \mid \mathcal{F}^{t-1}]$ as

$$\sigma_t^2 \leq \mathbb{E}\left[4\gamma^2\|\theta_t^u\|^2\cdot\|\eta_t\|^2 \mid \mathcal{F}^{t-1}\right] \overset{(50)}{\leq} 16\gamma^2R^2\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right]. \tag{59}$$

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 A.1) with $X_t = -2\gamma\left\langle\theta_t^u, \eta_t\right\rangle$, parameter c as in (58), $b = \frac{R^2}{7}$, $G = \frac{R^4}{294\ln\frac{8(K+1)}{\beta}}$ to obtain

$$\mathbb{P}\left\{|①| > \frac{R^2}{7} \quad \text{and} \quad \sum_{t=0}^{T-1}\sigma_t^2 \leq \frac{R^4}{294\ln\frac{8(K+1)}{\beta}}\right\} \leq 2\exp\left(-\frac{b^2}{2G + 2cb/3}\right) = \frac{\beta}{4(K+1)}.$$

Equivalently, we have

$$\mathbb{P}\left\{E_{①}\right\} \geq 1 - \frac{\beta}{4(K+1)}, \quad \text{for} \quad E_{①} = \left\{\text{either} \quad \sum_{t=0}^{T-1}\sigma_t^2 > \frac{R^4}{294\ln\frac{8(K+1)}{\beta}} \quad \text{or} \quad |①| \leq \frac{R^2}{7}\right\}. \tag{60}$$

In addition, E_{T-1} implies

$$\begin{aligned}
\sum_{t=0}^{T-1}\sigma_t^2 &\leq 16\gamma^2R^2\sum_{t=0}^{T-1}\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right] \\
&\overset{(57)}{\leq} 4R^2\gamma^2T\left(9(2^{2\alpha-1}+1)\lambda^{2-\alpha}\sigma^\alpha + 9(2^{2\alpha-1}+1)\lambda^{2-\alpha}\zeta_\lambda^\alpha)\right) \\
&\overset{(36)}{\leq} \frac{R^4}{294\ln\frac{8(K+1)}{\beta}}.
\end{aligned} \tag{61}$$

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

$$\begin{aligned}
② &= -2\gamma\sum_{t=0}^{T-1}\langle\theta_t^b, \eta_t\rangle \leq 2\gamma\sum_{t=0}^{T-1}\|\theta_t^b\|\cdot\|\eta_t\| \\
&\overset{(56),(50)}{\leq} 4\gamma RT\left(\frac{2^{2\alpha-1}\sigma\left(\sigma^\alpha + \zeta_\lambda^\alpha\right)^{\frac{\alpha-1}{\alpha}}}{\lambda^{\alpha-1}} + (\zeta_\lambda + \lambda/2)\frac{2^{2\alpha-1}\left(\sigma^\alpha + \zeta_\lambda^\alpha\right)}{\lambda^\alpha} + \zeta_\lambda\right) \\
&\overset{T<K+1}{\leq} 4\gamma R(K+1)\frac{2^{2\alpha-1}}{\lambda^{\alpha-1}}\left(\sigma^\alpha + \zeta_\lambda^\alpha\right)\left(\left(1 + \frac{\zeta_\lambda^\alpha}{\sigma^\alpha}\right)^{-1/\alpha} + \frac{\zeta_\lambda}{\lambda} + \frac{1}{2} + \frac{\lambda^{\alpha-1}\zeta_\lambda}{2^{2\alpha-1}\left(\sigma^\alpha + \zeta_\lambda^\alpha\right)}\right) \\
&\overset{(37)}{\leq} \frac{R^2}{7}.
\end{aligned} \tag{62}$$

Upper bound for ③. We have

$$|③| = \left|-2\gamma\sum_{t=0}^{T-1}\langle\eta_t, \omega_t\rangle\right| = \left|\sum_{t=0}^{T-1}\sum_{i=1}^{d}2\gamma\eta_{t,i}\omega_{t,i}\right| \tag{63}$$

where $\eta_{t,i} \overset{\text{def}}{=} [\eta_t]_i$ and $\omega_{t,i} \overset{\text{def}}{=} [\omega_t]_i$ denote the i-th components of η_t and ω_t respectively.

Each summand is the product of a zero-mean Gaussian random variable and a bounded random variable, resulting in the product being a zero-mean sub-Gaussian random variable with parameter $\sigma_{t,i}^2 = 64R^2\gamma^2\sigma_\omega^2$. To prove this, consider

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{4\gamma^2}{\sigma_{t,i}^2}\left|\eta_{t,i}^2\omega_{t,i}^2\right|\right) \mid \mathcal{F}^{t-1}\right] &\overset{(50)}{\leq} \mathbb{E}\left[\exp\left(\frac{16R^2\gamma^2}{64\gamma^2R^2\sigma_\omega^2}\left|\omega_{t,i}\right|^2\right)\right] \\ &\leq \mathbb{E}\left[\exp\left(\frac{\left|\omega_{t,i}\right|^2}{4\sigma_\omega^2}\right)\right] \overset{(ii)}{\leq} \exp(1) \end{aligned} \tag{64}$$

where (ii) uses the fact that $\omega_{t,i}^2$ is light-tailed random variable with parameter σ_ω^2. Now that we have established the light-tailedness of summands, we can use the Lemma A.2 to obtain

$$\begin{aligned}\mathbb{P}\left\{\left|\sum_{t=0}^{T-1}\sum_{i=1}^{d} 2\gamma\eta_{t,i}\omega_{t,i}\right| > \left(\sqrt{2}+\sqrt{2}\phi\right)\sqrt{\sum_{t=0}^{K}\sum_{i=1}^{d} 64\gamma^2R^2\sigma_\omega^2}\right\} &\leq \exp\left(\frac{-\phi^2}{3}\right) \\ &= \frac{\beta}{4(K+1)}. \end{aligned} \tag{65}$$

The choice of $\gamma \leq \gamma_3$ for γ_3 defined (38) implies

$$\left(\sqrt{2}+\sqrt{2}\phi\right)\sqrt{\sum_{t=0}^{T-1}\sum_{i=1}^{d} 64\gamma^2R^2\sigma_\omega^2} \leq \left(\sqrt{2}+\sqrt{2}\phi\right)\sqrt{64\gamma^2R^2(K+1)d\sigma_\omega^2} \overset{(38)}{\leq} \frac{R^2}{7},$$

and

$$\mathbb{P}\{E_{③}\} \geq 1 - \frac{\beta}{4(K+1)} \quad \text{for} \quad E_{③} = \left\{|③| \leq \frac{R^2}{7}\right\}. \tag{66}$$

Upper bound for ④. From E_{T-1}, and conditions on the step-size it follows that

$$\begin{aligned} ④ &= 4\gamma^2\sum_{t=0}^{T-1}\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_\xi^{t-1}\right] \\ &\overset{(57)}{\leq} 4\gamma^2 T\left(\frac{9(2^{2\alpha-1}+1)\lambda^{2-\alpha}\sigma^\alpha}{4} + \frac{9(2^{2\alpha-1}+1)\lambda^{2-\alpha}\zeta_\lambda^\alpha}{4}\right) \overset{(36)}{\leq} \frac{R^2}{7}. \end{aligned} \tag{67}$$

Upper bound for ⑤. First, we have

$$\mathbb{E}\left[4\gamma^2\left(\|\theta_t^u\|^2 - \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right]\right) \mid \mathcal{F}^{t-1}\right] = 0.$$

Next, sum ⑤ has bounded with probability 1 terms:

$$\begin{aligned}\left|4\gamma^2\left(\|\theta_t^u\|^2 - \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right]\right)\right| &\leq 4\gamma^2\left(\|\theta_t^u\|^2 + \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right]\right) \\ &\leq 32\gamma^2\lambda^2 \overset{(40)}{\leq} \frac{R^2}{7\ln\frac{8(K+1)}{\beta}} \overset{\text{def}}{=} c. \end{aligned} \tag{68}$$

The summands also have bounded conditional variances

$$\widetilde{\sigma}_t^2 \overset{\text{def}}{=} \mathbb{E}\left[16\gamma^4\left(\|\theta_t^u\|^2 - \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right]\right)^2 \mid \mathcal{F}^{t-1}\right],$$

$$\widetilde{\sigma}_t^2 \overset{(68)}{\leq} \frac{R^2}{7\ln\frac{8(K+1)}{\beta}}\mathbb{E}\left[4\gamma^2\left|\|\theta_t^u\|^2 - \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right]\right| \mid \mathcal{F}^{t-1}\right] \tag{69}$$

$$\leq \quad \frac{8\gamma^2 R^2}{7\ln\frac{8(K+1)}{\beta}}\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right]. \tag{70}$$

To summarize, we have shown that $\left\{4\gamma^2\left(\|\theta_t^u\|^2 - \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right]\right)\right\}_{t=0}^{T-1}$ 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 A.1) with $X_t = 4\gamma^2\left(\|\theta_t^u\|^2 - \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right]\right)$, parameter c as in (68), $b = \frac{R^2}{7}$, $G = \frac{R^4}{294\ln\frac{8(K+1)}{\beta}}$:

$$\mathbb{P}\left\{|⑤| > \frac{R^2}{7} \quad \text{and} \quad \sum_{t=0}^{T-1}\widetilde{\sigma}_t^2 \leq \frac{R^4}{294\ln\frac{8(K+1)}{\beta}}\right\} \leq 2\exp\left(-\frac{b^2}{2G + {}^{2cb}/_3}\right) = \frac{\beta}{4(K+1)}.$$

Equivalently, we have

$$\mathbb{P}\{E_⑤\} \geq 1 - \frac{\beta}{4(K+1)}, \quad \text{for} \quad E_⑤ = \left\{\text{either} \quad \sum_{t=0}^{T-1}\widetilde{\sigma}_t^2 > \frac{R^4}{294\ln\frac{8(K+1)}{\beta}} \quad \text{or} \quad |④| \leq \frac{R^2}{7}\right\}. \tag{71}$$

In addition, E_{T-1} implies that

$$\sum_{t=0}^{T-1}\widetilde{\sigma}_t^2 \overset{(70)}{\leq} \frac{8\gamma^2 R^2(K+1)}{7\ln\frac{8(K+1)}{\beta}}\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right] \overset{(57),(36)}{\leq} \frac{R^4}{294\ln\frac{8(K+1)}{\beta}}. \tag{72}$$

Upper bound for ⑥. From E_{T-1}, and conditions on the step-size it follows that

$$\begin{aligned} ⑥ &= 4\gamma^2\sum_{t=0}^{T-1}\left\|\theta_t^b\right\|^2 \\ &\leq 4\gamma^2 T\left(\frac{2^{2\alpha-1}\sigma\left(\sigma^\alpha+\zeta_\lambda^\alpha\right)^{\frac{\alpha-1}{\alpha}}}{\lambda^{\alpha-1}} + (\zeta_\lambda + {}^{\lambda}/_2)\frac{2^{2\alpha-1}\left(\sigma^\alpha+\zeta_\lambda^\alpha\right)}{\lambda^\alpha} + \zeta_\lambda\right)^2 \\ &\overset{(37)}{\leq} \frac{R^2}{7}. \end{aligned} \tag{73}$$

Upper bound for ⑦. We have

$$4\gamma^2\sum_{t=0}^{T-1}\|\omega_t\|^2 = 4\gamma^2\sigma_\omega^2\sum_{t=0}^{T-1}\sum_{i=1}^{d} z_{t,i}^2, \tag{74}$$

where $z_{t,i} \overset{\text{def}}{=} {}^{\omega_{t,i}}/_{\sigma_\omega}$. Using Lemma A.3, we get

$$\mathbb{P}\left\{\sum_{t=0}^{T-1}\sum_{i=1}^{d} z_{t,i}^2 > Td + 2\sqrt{Td\ln\frac{4(K+1)}{\beta}} + 2\ln\frac{4(K+1)}{\beta}\right\} \leq \frac{\beta}{4(K+1)}. \tag{75}$$

Since $\gamma \leq \gamma_6$ for γ_6 defined in (41), we obtain

$$\mathbb{P}\left\{⑦ > \frac{R^2}{7}\right\} \leq \frac{\beta}{4(K+1)}, \tag{76}$$

which is equivalent to

$$\mathbb{P}\{E_⑦\} \geq 1 - \frac{\beta}{4(K+1)} \quad \text{for} \quad E_③ = \left\{|⑦| \leq \frac{R^2}{7}\right\}. \tag{77}$$

Now, we have the upper bounds for ①, ②, ③, ④, ⑤, ⑥, ⑦ . Thus, probability event $E_{T-1} \cap E_① \cap E_③ \cap E_⑤ \cap E_⑦$ implies

$$R_T^2 \quad \leq \quad R^2 - 2\gamma\sum_{l=0}^{t-1}\langle x^l - x^\star, \theta_l\rangle - 2\gamma\sum_{l=0}^{t-1}\langle x^l - x^\star, \omega_l\rangle + 2\gamma^2\sum_{l=0}^{t-1}\|\theta_l\|^2 + 4\gamma^2\sum_{l=0}^{t-1}\|\omega_l\|^2$$

$$\begin{aligned} &\leq R^2 + ① + ② + ③ + ④ + ⑤ + ⑥ + ⑦ \\ &\leq R^2 + \frac{R^2}{7} + \frac{R^2}{7} + \frac{R^2}{7} + \frac{R^2}{7} + \frac{R^2}{7} + \frac{R^2}{7} + \frac{R^2}{7} = 2R^2, \end{aligned}$$

which is equivalent to (43) and (44) for $t = T$, and

$$\begin{aligned} \mathbb{P}\{E_T\} &\geq \mathbb{P}\{E_{T-1} \cap E_{①} \cap E_{③} \cap E_{⑤} \cap E_{⑦}\} \\ &= 1 - \mathbb{P}\left\{\overline{E}_{T-1} \cup \overline{E}_{①} \cup \overline{E}_{③} \cup \overline{E}_{⑤} \cup \overline{E}_{⑦}\right\} \\ &\geq 1 - \mathbb{P}\{\overline{E}_{T-1}\} - \mathbb{P}\{\overline{E}_{①}\} - \mathbb{P}\{\overline{E}_{③}\} - \mathbb{P}\{\overline{E}_{⑤}\} - \mathbb{P}\{\overline{E}_{⑦}\} \\ &\geq 1 - \frac{(T+1)\beta}{K+1}. \end{aligned} \tag{78}$$

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

$$\frac{1}{(K+1)} \sum_{t=0}^{K} c_t (f(x^t) - f(x^\star)) \leq \frac{2R^2}{\gamma(K+1)}$$

and $\{x^k\}_{k=0}^K \subseteq Q$, which follows from (44). Now, we have to deal with c_t. To do so, we consider two possible cases for each $t = 0, 1, \ldots, K$: either $c_t = 1$ or $c_t = \frac{\lambda}{2\|\nabla f(x^t)\|}$. We define the corresponding sets of indices: $\mathcal{T}_1 \overset{\text{def}}{=} \{t \in \{0, 1, \ldots, K\} \mid c_t = 1\}$ and $\mathcal{T}_2 \overset{\text{def}}{=} \{t \in \{0, 1, \ldots, K\} \mid c_t = \frac{\lambda}{2\|\nabla f(x^t)\|}\}$. Then, the above inequality can be rewritten as

$$\frac{1}{(K+1)} \sum_{t \in \mathcal{T}_1} (f(x^t) - f(x^\star)) + \frac{1}{(K+1)} \sum_{t \in \mathcal{T}_2} \frac{\lambda(f(x^t) - f(x^\star))}{2\|\nabla f(x^t)\|} \leq \frac{2R^2}{\gamma(K+1)}, \tag{79}$$

implying

$$\frac{1}{(K+1)} \sum_{t \in \mathcal{T}_1} (f(x^t) - f(x^\star)) \leq \frac{2R^2}{\gamma(K+1)} \tag{80}$$

and

$$\frac{1}{(K+1)} \sum_{t \in \mathcal{T}_2} \frac{\lambda(f(x^t) - f(x^\star))}{2\|\nabla f(x^t)\|} \leq \frac{2R^2}{\gamma(K+1)}. \tag{81}$$

Using the corollary of smoothness assumption, i.e., $\|\nabla f(x^t)\| \leq \sqrt{2L(f(x^t) - f(x^\star))}$, we get from (81) that

$$\frac{1}{K+1} \sum_{t \in \mathcal{T}_2} \sqrt{f(x^t) - f(x^\star)} \leq \frac{4\sqrt{2L}R^2}{\lambda\gamma(K+1)}. \tag{82}$$

For inequality (80), we follow the technique from (Koloskova et al., 2023) and apply inequality $x^2 \geq 2\epsilon x - \epsilon^2$, which holding for any ϵ, x. Setting $x^2 = f(x^t) - f(x^\star)$, we get

$$\frac{1}{K+1} \sum_{t \in \mathcal{T}_1} \left(2\epsilon\sqrt{f(x^t) - f(x^\star)} - \epsilon^2\right) \leq \frac{2R^2}{\gamma(K+1)},$$

implying

$$\frac{1}{K+1} \sum_{t \in \mathcal{T}_1} \sqrt{f(x^t) - f(x^\star)} \leq \frac{R^2}{\gamma(K+1)\epsilon} + \frac{\epsilon}{2}.$$

Choosing $\epsilon = \frac{\sqrt{2}R}{\sqrt{\gamma(K+1)}}$, we obtain

$$\frac{1}{K+1} \sum_{t \in \mathcal{T}_1} \sqrt{f(x^t) - f(x^\star)} \leq \sqrt{\frac{2R^2}{\gamma(K+1)}}. \tag{83}$$

Combining inequalities (82) and (83), we get

$$\frac{1}{K+1}\sum_{t=0}^{K}\sqrt{f(x^t)-f(x^\star)} \leq \sqrt{\frac{2R^2}{\gamma(K+1)}} + \frac{4\sqrt{2L}R^2}{\lambda\gamma(K+1)}, \tag{84}$$

which implies

$$\min_{t\in[0,K]}\left(f(x^t)-f(x^\star)\right) \leq \frac{4R^2}{\gamma(K+1)} + \frac{64LR^4}{\lambda^2\gamma^2(K+1)^2}, \tag{85}$$

where we have utilized the inequality $(a+b)^2 \leq 2a^2+2b^2$. This concludes the proof. □

Theorem C.1 states 7 values for step-size, from which the smallest should be selected. To simplify matters, we demonstrate that if λ is selected equal or smaller than the order of $\mathcal{O}\left(\left(\frac{K}{\ln K}\right)^{1/\alpha}\right)$, then three step-sizes are redundant and can be omitted.

Corollary C.1. *Let all conditions of Theorem C.1 hold. Furthermore, assume that K is large and one selects $\lambda \leq \mathcal{O}\left(\left(\frac{K}{\ln K}\right)^{1/\alpha}\right)$, then conclusions of Theorem C.1 are valid as long as γ is selected to satisfy $\gamma \leq \min\left\{1/8L, \gamma_1, \gamma_2, \gamma_3\right\}$ where we have*

$$\gamma_1 \stackrel{\text{def}}{=} \frac{R}{42(2^{2\alpha-1}+1)^{1/2}\sigma^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{6(K+1)\ln\frac{8(K+1)}{\beta}\left(1+\frac{\zeta_\lambda^\alpha}{\sigma^\alpha}\right)}},$$

$$\gamma_2 \stackrel{\text{def}}{=} \frac{R\lambda^{\alpha-1}}{28(K+1)2^{2\alpha-1}\sigma^\alpha\left(1+\frac{\zeta_\lambda^\alpha}{\sigma^\alpha}\right)\left(\frac{\zeta_\lambda}{\lambda}+\frac{1}{2}+\frac{\lambda^{\alpha-1}\zeta_\lambda}{2^{2\alpha-1}\left(\sigma^\alpha+\zeta_\lambda^\alpha\right)}+\left(1+\frac{\zeta_\lambda^\alpha}{\sigma^\alpha}\right)^{-1/\alpha}\right)},$$

$$\gamma_3 \stackrel{\text{def}}{=} \frac{R}{56\sigma_\omega\sqrt{d(K+1)}(\sqrt{2}+\sqrt{2}\phi)}.$$

Proof. For large K, it is evident that γ_3 decreases at a rate of $\mathcal{O}\left(\sigma_\omega\sqrt{K\ln K}\right)$, while γ_6 in (41) decreases at a rate of $\mathcal{O}\left(\sigma_\omega\sqrt{K}\right)$. Subsequently, γ_3 dominates γ_6 and γ_6 can be omitted. Furthermore, γ_5 in (40) decreases with a rate of $\mathcal{O}\left(K^{1/\alpha}(\ln K)^{1-1/\alpha}\right)$ which is less than the rate of γ_2. It can be deduced that for large λ, γ_2 decreases at the rate $\mathcal{O}(K)$ which is faster than γ_5. If λ is small, γ_2 dominates γ_5 again due to the λ in the numerator of γ_2. Hence, γ_5 can be discarded. As for γ_4 in (39), we know that σ_ω is on the order of $\mathcal{O}\left(\lambda/\epsilon\sqrt{K\ln(K/\delta)}\right)$. Hence, one can replace λ with $\mathcal{O}\left(\sigma_\omega\epsilon/\sqrt{K\ln(K/\delta)}\right)$. Therefore, γ_4 decreases by the order $\mathcal{O}\left(\sigma_\omega\epsilon\sqrt{K\ln(K/\delta)}\right)$, which is the same order as γ_3. Hence, γ_4 can be omitted, and the proof is complete. □

D Rate and Neighborhood for Clipped-SGD: Convex Case

Now that we have established the convergence properties of DP-Clipped-SGD for convex problems, we turn to evaluating its convergence rate. This rate depends critically on the choice of the step-size γ, and in general, the resulting expressions can be quite complex. To obtain more interpretable bounds, we consider simplified rate expressions by analyzing separate cases based on different ranges of λ. Since we focus on the asymptotic behavior, numerical constants are omitted for clarity.

In this section, we consider the cases without the DP noise ($\sigma_\omega = 0$) and investigate all possible clipping levels.

Case 1: $\lambda > 4LR$. In this case, $\zeta_\lambda = 0$, and the step-size conditions reduce to the following:

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{R}{\sigma^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K\ln\frac{K}{\beta}}}, \frac{R\lambda^{\alpha-1}}{K\sigma^\alpha}\right\}\right). \tag{86}$$

In particular, when γ equals the minimum from the above condition, the iterates produced by Clipped-SGD after K iterations with probability at least $1-\beta$ satisfy

$$\min_{t\in[0,K]} f(x^t) - f(x^\star) = \mathcal{O}\left(\max\left\{(88), (89), (90)\right\}\right), \tag{87}$$

where

$$R\lambda^{1-\alpha/2}\sigma^{\alpha/2}\sqrt{\frac{\ln K/\beta}{K}} + \frac{LR^2\sigma^\alpha \ln K/\beta}{\lambda^\alpha K}, \tag{88}$$

$$\frac{R\sigma^\alpha}{\lambda^{\alpha-1}} + \frac{LR^2\sigma^{2\alpha}}{\lambda^{2\alpha}}, \tag{89}$$

$$\frac{LR^2}{K} + \frac{L^3R^4}{\lambda^2K^2}. \tag{90}$$

We clearly see that the dominant term in (88) is an increasing function of λ, and the dominant term in (89) is a decreasing function. Solving for optimal λ as the equilibrium of the dominant terms in (88) and (89), we get $\lambda = \mathcal{O}\left(\sigma\left(\frac{K}{\ln\frac{K}{\beta}}\right)^{\frac{1}{\alpha}}\right)$. Plugging in this λ, we get with probability at least $1-\beta$:

$$\min_{t\in[0,K]} f(x^t) - f(x^\star) = \mathcal{O}\left(\max\left\{(92), (93)\right\}\right), \tag{91}$$

where

$$R\sigma\left(\frac{\ln\frac{K}{\beta}}{K}\right)^{\frac{\alpha-1}{\alpha}} + \frac{LR^2\ln^2 K/\beta}{K^2}. \tag{92}$$

$$\frac{LR^2}{K} + \frac{L^3R^4\left(\ln\frac{K}{\beta}\right)^{\frac{2}{\alpha}}}{\sigma^2K^{\frac{2\alpha+2}{\alpha}}}. \tag{93}$$

In this case, Clipped-SGD converges to the exact optimum asymptotically with high probability, and the dominant term matches the one from Sadiev et al. (2023). As it can be seen from (88), (89), when the clipping level is not that large, we converge to a neighborhood of the solution, but with a faster $\mathcal{O}(1/\sqrt{K})$ rate.

Next, when $\lambda \leq 4LR$, we have $\zeta_\lambda = \frac{4LR-\lambda}{2}$. As it can be seen from (36), (37), in these cases, we also have to consider the relation between λ and σ. Thus, we split $\lambda \leq 4LR$ regime into 6 different regimes to cover all possible cases.

Case 2: $\frac{4}{3}LR < \lambda \leq 4LR, \quad \zeta_\lambda < \lambda < \sigma$. In this case, the step-size conditions reduce to the following:

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{R}{\sigma^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K\ln\frac{K}{\beta}}}, \frac{R\lambda^{\alpha-1}}{K\sigma^\alpha}\right\}\right). \tag{94}$$

As can be seen, the result is the same as in the previous case. The optimal λ derived in the previous section violates the constraint that $\lambda \leq 4LR$; thus, the optimal $\lambda = 4LR$. For this choice of λ, we have with probability at least $1-\beta$

$$\min_{t\in[0,K]} f(x^t) - f(x^\star) = \mathcal{O}\left(\max\left\{(96),(97),(98)\right\}\right), \tag{95}$$

where

$$\sqrt{R^{4-\alpha}L^{2-\alpha}\sigma^\alpha \frac{\ln {K/\beta}}{K}} + \frac{R^{2-\alpha}\sigma^\alpha \ln {K/\beta}}{L^{\alpha-1}K}, \tag{96}$$

$$\frac{R^{2-\alpha}\sigma^\alpha}{L^{\alpha-1}} + \frac{\sigma^{2\alpha}}{L^{2\alpha-1}R^{2\alpha-2}}, \tag{97}$$

$$\frac{LR^2}{K} + \frac{LR^2}{K^2}. \tag{98}$$

Case 3: $\frac{4}{3}LR < \lambda \leq 4LR, \quad \zeta_\lambda < \sigma < \lambda$. In this case, the step-size conditions reduce to the following:

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{R}{\sigma^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K\ln\frac{K}{\beta}}}, \frac{R\lambda^{\alpha-1}}{K\max\{\sigma^\alpha, \lambda^{\alpha-1}\zeta_\lambda\}}\right\}\right). \tag{99}$$

If $\max\{\sigma^\alpha, \lambda^{\alpha-1}\zeta_\lambda\} = \sigma^\alpha$, then the bounds are similar to the previous case. If $\max\{\sigma^\alpha, \lambda^{\alpha-1}\zeta_\lambda\} = \lambda^{\alpha-1}\zeta_\lambda$ is satisfied, $\min_{t\in[0,K]} f(x^t) - f(x^\star)$ is bounded with probability at least $1-\beta$ by the maximum of the following terms:

$$R\lambda^{1-\alpha/2}\sigma^{\alpha/2}\sqrt{\frac{\ln {K/\beta}}{K}} + \frac{LR^2\sigma^\alpha \ln {K/\beta}}{\lambda^\alpha K}, \tag{100}$$

$$R\zeta_\lambda + \frac{LR^2\zeta_\lambda^2}{\lambda^2}, \tag{101}$$

$$\frac{LR^2}{K} + \frac{L^3R^4}{\lambda^2K^2}. \tag{102}$$

In the latter case (i.e., maximum occurring in the second argument), the optimal λ is $4LR - \eta$, where η is a sufficiently small number such that $\lambda^{\alpha-1}\zeta_\lambda \geq \sigma^\alpha$, i.e., λ satisfies $\zeta_\lambda = \max\left\{\frac{\sigma^\alpha}{\lambda^{\alpha-1}}, \lambda^{1-\alpha/2}\sigma^{\alpha/2}\sqrt{\frac{\ln {K/\beta}}{K}}\right\}$. Note that the (105) is decreasing in λ, and $\lambda = 4LR$ is not feasible. With this choice of λ, we get with probability at least $1-\beta$:

$$\min_{t\in[0,K]} f(x^t) - f(x^\star) = \mathcal{O}\left(\max\left\{(104),(105),(106)\right\}\right), \tag{103}$$

where

$$R\sqrt{(4LR-\eta)^{2-\alpha}\sigma^\alpha \frac{\ln {K/\beta}}{K}} + \frac{LR^2\sigma^\alpha \ln {K/\beta}}{(LR-\eta)^\alpha K}, \tag{104}$$

$$\frac{R\eta}{2} + \frac{LR^2\eta^2}{(4LR-\eta)^2}, \tag{105}$$

$$\frac{L\Delta}{K} + \frac{L^2\Delta^2}{(4\sqrt{L\Delta}-\eta)^2K^2}. \tag{106}$$

Case 4: $\frac{4}{3}LR < \lambda \leq 4LR, \quad \sigma < \zeta_\lambda < \lambda$. In this case, the step-size conditions reduce to the following:

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{R}{\zeta_\lambda^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K\ln\frac{K}{\beta}}}, \frac{R\lambda^{\alpha-1}}{K(\lambda^{\alpha-1}\zeta_\lambda)}\right\}\right), \tag{107}$$

and $\min_{t\in[0,K]} f(x^t) - f(x^\star)$ is bounded with probability at least $1-\beta$ by the maximum of the following terms:

$$R\lambda^{1-\alpha/2}\zeta_\lambda^{\alpha/2}\sqrt{\frac{\ln {}^K\!/\!_\beta}{K}} + \frac{LR^2\zeta_\lambda^\alpha \ln {}^K\!/\!_\beta}{\lambda^\alpha K}, \tag{108}$$

$$R\zeta_\lambda + \frac{LR^2\zeta_\lambda^2}{\lambda^2}, \tag{109}$$

$$\frac{LR^2}{K} + \frac{L^3R^4}{\lambda^2K^2}. \tag{110}$$

The optimal in this case is $\lambda = 4LR - 2\sigma$, and the neighborhood of the convergence and the rate are presented below: with probability at least $1-\beta$

$$\min_{t\in[0,K]} f(x^t) - f(x^\star) = \mathcal{O}\left(\max\left\{(112), (113), (114)\right\}\right), \tag{111}$$

where

$$R\sqrt{(4LR-2\sigma)^{2-\alpha}\sigma^\alpha\frac{\ln {}^K\!/\!_\beta}{K}} + \frac{LR^2\sigma^\alpha \ln {}^K\!/\!_\beta}{(4LR-2\sigma)^\alpha K}, \tag{112}$$

$$R\sigma + \frac{LR^2\sigma^2}{(4LR-2\sigma)^2}, \tag{113}$$

$$\frac{LR^2}{K} + \frac{L^3R^4}{(4LR-2\sigma)^2K^2}. \tag{114}$$

Case 5: $\lambda \leq \frac{4}{3}LR, \quad \lambda < \zeta_\lambda < \sigma$**.** In this case, the step-size conditions reduce to the following:

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{R}{\sigma^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K\ln\frac{K}{\beta}}}, \frac{R\lambda^\alpha}{K(\sigma^\alpha\zeta_\lambda)}\right\}\right). \tag{115}$$

Function sub-optimality $\min_{t\in[0,K]} f(x^t) - f(x^\star)$ is bounded with probability at least $1-\beta$ by the maximum of the following terms:

$$R\lambda^{1-\alpha/2}\sigma^{\alpha/2}\sqrt{\frac{\ln {}^K\!/\!_\beta}{K}} + \frac{LR^2\sigma^\alpha \ln {}^K\!/\!_\beta}{\lambda^\alpha K}, \tag{116}$$

$$R\frac{\sigma^\alpha\zeta_\lambda}{\lambda^\alpha} + \frac{LR^2\sigma^{2\alpha}\zeta_\lambda^2}{\lambda^{2\alpha+2}}, \tag{117}$$

$$\frac{LR^2}{K} + \frac{L^3R^4}{\lambda^2K^2}. \tag{118}$$

In this regime, the optimal $\lambda = \frac{4}{3}LR$. With this choice of λ we get: with probability at least $1-\beta$

$$\min_{t\in[0,K]} f(x^t) - f(x^\star) = \mathcal{O}\left(\max\left\{(120), (121), (122)\right\}\right), \tag{119}$$

where

$$\sqrt{R^{4-\alpha}L^{2-\alpha}\sigma^\alpha\frac{\ln {}^K\!/\!_\beta}{K}} + \frac{R^{2-\alpha}\sigma^\alpha \ln {}^K\!/\!_\beta}{L^{\alpha-1}K}, \tag{120}$$

$$\frac{R^{2-\alpha}\sigma^\alpha}{L^{\alpha-1}} + \frac{\sigma^{2\alpha}}{L^{2\alpha-1}R^{2\alpha-2}}, \tag{121}$$

$$\frac{LR^2}{K} + \frac{LR^2}{K^2}. \tag{122}$$

Case 6: $\lambda \leq \frac{4}{3}LR, \quad \lambda < \sigma < \zeta_\lambda$**.** In this case, the step-size conditions reduce to the following:

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{R}{\zeta_\lambda^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K\ln\frac{K}{\beta}}}, \frac{R\lambda^\alpha}{K(\zeta_\lambda^{\alpha+1})}\right\}\right). \tag{123}$$

Function sub-optimality $\min_{t\in[0,K]} f(x^t) - f(x^\star)$ is bounded with probability at least $1-\beta$ by the maximum of the following terms:

$$R\lambda^{1-\alpha/2}\zeta_\lambda^{\alpha/2}\sqrt{\frac{\ln {}^K\!/\!_\beta}{K}} + \frac{LR^2\zeta_\lambda^\alpha \ln {}^K\!/\!_\beta}{\lambda^\alpha K}, \tag{124}$$

$$\frac{R\zeta_\lambda^{\alpha+1}}{\lambda^\alpha} + \frac{LR^2\zeta_\lambda^{2\alpha}}{\lambda^{2\alpha+2}}, \tag{125}$$

$$\frac{LR^2}{K} + \frac{L^3R^4}{\lambda^2K^2}. \tag{126}$$

Next, we find the optimal λ via equalizing the leading terms (the first ones) in (124) and (125). This results in $\lambda = \frac{4LR}{2C+1}$, where $C = \left(\frac{\ln\frac{K}{\beta}}{K}\right)^{\frac{1}{\alpha+2}}$, which is infeasible. Thus, in this regime, the optimal λ is $\frac{4}{3}LR - \eta$, where $\eta \geq 0$ is such that $\lambda < \sigma < \zeta_\lambda$. Given this choice of λ, we obtain with probability at least $1-\beta$

$$\min_{t\in[0,K]} f(x^t) - f(x^\star) = \mathcal{O}\left(\max\left\{(128), (129), (130)\right\}\right), \tag{127}$$

where

$$R(LR-\eta)^{1-\alpha/2}(LR+\eta)^{\alpha/2}\sqrt{\frac{\ln {}^K\!/\!_\beta}{K}} + \frac{LR^2(LR+\eta)^{2\alpha}\ln {}^K\!/\!_\beta}{(LR-\eta)^{2\alpha+2}K}, \tag{128}$$

$$\frac{R(LR+\eta)^{\alpha+1}}{(LR-\eta)^\alpha} + \frac{LR^2(LR+\eta)^{2\alpha}}{(LR-\eta)^{2\alpha+2}}, \tag{129}$$

$$\frac{LR^2}{K} + \frac{L^3R^4}{(LR-\eta)^2K^2}. \tag{130}$$

Case 7: $\lambda \leq \frac{4}{3}LR, \quad \sigma < \lambda < \zeta_\lambda$**.** In this case, the step-size conditions reduce to the following:

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{R}{\zeta_\lambda^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K\ln\frac{K}{\beta}}}, \frac{R\lambda^{\alpha-1}}{K\max\left\{\frac{\zeta_\lambda^{\alpha+1}}{\lambda}, \zeta_\lambda^{\alpha-1}\sigma\right\}}\right\}\right). \tag{131}$$

We note that $\max\left\{\frac{\zeta_\lambda^{\alpha+1}}{\lambda}, \zeta_\lambda^{\alpha-1}\sigma\right\} = \zeta^\alpha \max\left\{\frac{\zeta_\lambda}{\lambda}, \frac{\sigma}{\lambda}\right\} = \frac{\zeta_\lambda^{\alpha+1}}{\lambda}$ since $\sigma < \lambda < \zeta_\lambda$. Therefore, similarly to the previous case, we have

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{R}{\zeta_\lambda^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K\ln\frac{K}{\beta}}}, \frac{R\lambda^\alpha}{K(\zeta_\lambda^{\alpha+1})}\right\}\right), \tag{132}$$

and $\min_{t\in[0,K]} f(x^t) - f(x^\star)$ is bounded with probability at least $1-\beta$ by the maximum of the following terms:

$$R\lambda^{1-\alpha/2}\zeta_\lambda^{\alpha/2}\sqrt{\frac{\ln {}^K\!/\!_\beta}{K}} + \frac{LR^2\zeta_\lambda^\alpha \ln {}^K\!/\!_\beta}{\lambda^\alpha K}, \tag{133}$$

$$\frac{R\zeta_\lambda^{\alpha+1}}{\lambda^\alpha} + \frac{LR^2\zeta_\lambda^{2\alpha}}{\lambda^{2\alpha+2}}, \tag{134}$$

$$\frac{LR^2}{K} + \frac{L^3R^4}{\lambda^2K^2}. \tag{135}$$

The optimal λ is $\frac{4}{3}LR$, since the both leading terms in (133) and (134) are decreasing in λ. With this choice, we get with probability at least $1-\beta$

$$\min_{t\in[0,K]} f(x^t) - f(x^\star) = \mathcal{O}\left(\max\left\{(137), (138), (139)\right\}\right), \tag{136}$$

where

$$LR^2\sqrt{\frac{\ln {}^K\!/\!_\beta}{K}} + \frac{LR^2\ln {}^K\!/\!_\beta}{K}, \tag{137}$$

$$R\sigma + \frac{\sigma^2}{L}, \tag{138}$$

$$\frac{LR^2}{K} + \frac{LR^2}{K^2}. \tag{139}$$

Now that we have covered all regions, it's time to consider the DP noise as well.

E Rate and Neighborhood for DP-Clipped-SGD: Convex Case

To ensure the output of the algorithm is (ε, δ)-differentially private in this setting, expectation minimization, it suffices to set the noise scale as $\sigma_\omega = \Theta\left(\frac{\lambda}{\varepsilon}\sqrt{K\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{1}{\delta}\right)}\right)$ and apply the advanced composition theorem of Dwork et al. (2014). In the finite sum case, one can reduce the amount of noise by a factor of $\sqrt{\ln\left(\frac{K}{\delta}\right)}$ as it was shown by Abadi et al. (2016). For the sake of brevity, in the DP case, we only consider two cases: large λ and relatively small λ regimes. The other cases can be derived with a similar analysis.

Case 1: $\lambda > 4LR$. In this case, $\zeta_\lambda = 0$, and the step-size conditions reduce to the following:

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{R}{\sigma^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K\ln\frac{K}{\beta}}}, \frac{R\lambda^{\alpha-1}}{K\sigma^\alpha}, \frac{R}{\sigma_\omega\sqrt{dK\ln\frac{K}{\beta}}}\right\}\right). \tag{140}$$

In particular, when γ equals the minimum from step-size condition, then the iterates produced by DP-Clipped-SGD after K iterations with probability at least $1-\beta$ satisfy

$$\min_{k\in[0,K]} f(x^t) - f(x^\star) = \mathcal{O}\left(\max\left\{(142), (143), (144), (145)\right\}\right), \tag{141}$$

where

$$R\lambda^{1-\alpha/2}\sigma^{\alpha/2}\sqrt{\frac{\ln K/\beta}{K}} + \frac{LR^2\sigma^\alpha\ln K/\beta}{\lambda^\alpha K}, \tag{142}$$

$$\frac{R\sigma^\alpha}{\lambda^{\alpha-1}} + \frac{LR^2\sigma^{2\alpha}}{\lambda^{2\alpha}}, \tag{143}$$

$$\frac{LR^2}{K} + \frac{L^3R^4}{\lambda^2K^2}, \tag{144}$$

$$R\sigma_\omega\sqrt{\frac{d\ln\frac{K}{\beta}}{K}} + \frac{LR^2\sigma_\omega^2 d\ln\frac{K}{\beta}}{\lambda^2K}. \tag{145}$$

Here, (143) accounts for the bias caused by clipping, and (145) accounts for the accumulation of DP noise. These terms are decreasing and increasing in λ respectively, if we use $\sigma_\omega = \Theta\left(\frac{\lambda}{\varepsilon}\sqrt{K\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{1}{\delta}\right)}\right)$. To find the optimal λ, we find the equilibrium of these two terms. Solving the equilibrium equation, we get $\lambda = \mathcal{O}\left(\frac{\varepsilon\sigma^\alpha}{d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)}\right)^{\frac{1}{\alpha}}$. Unless $\varepsilon\sigma^\alpha$ is large enough, this value violates the constraint that $\lambda > 4LR$, and it's not feasible. Thus, we have the following formula for the optimal λ:

$$\lambda = \max\left\{4LR, \left(\frac{\varepsilon\sigma^\alpha}{d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)}\right)^{\frac{1}{\alpha}}\right\}. \tag{146}$$

For this choice of λ, we get that with probability at least $1-\beta$

$$\min_{k\in[0,K]} f(x^t) - f(x^\star) = \mathcal{O}\left(\max\left\{(148), (149), (150), (151)\right\}\right), \tag{147}$$

with

$$\max\left\{\sqrt{R^{4-\alpha}L^{2-\alpha}\sigma^\alpha\frac{\ln K/\beta}{K}}, R\left(\frac{\varepsilon\sigma^\alpha}{\sqrt{d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)}}\right)^{\frac{1}{\alpha}}\sqrt{\frac{\ln^{\frac{3\alpha-2}{2\alpha}}\frac{K}{\beta}}{K}}\right\}, \tag{148}$$

$$\min\left\{\frac{R^{2-\alpha}\sigma^\alpha}{L^{\alpha-1}}, R\sigma\left(\frac{d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)}{\varepsilon}\right)^{\frac{\alpha-1}{\alpha}}\right\}, \tag{149}$$

$$\min\left\{\frac{LR^2}{K^2}, \frac{L^3R^4\left(d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\right)^{\frac{1}{\alpha}}}{(\varepsilon)^{\frac{1}{\alpha}}\sigma}\frac{\ln^{\frac{1}{\alpha}}\frac{K}{\beta}}{K^2}\right\} + \frac{LR^2}{K}, \tag{150}$$

$$\max\left\{\frac{LR^2}{\varepsilon}\sqrt{d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)}, \frac{R\sigma\left(d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)\right)^{\frac{\alpha+2}{2\alpha}}}{\varepsilon^{\frac{\alpha-1}{\alpha}}}\right\}$$
$$+\frac{LR^2}{\varepsilon^2}d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right), \tag{151}$$

where, for the sake of brevity, we only report the dominant terms.

Case 2: $\lambda \leq \frac{4}{3}LR \quad \lambda < \sigma < \zeta_\lambda$**.** In this case, the step-size conditions reduce to

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{R}{\zeta_\lambda^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K\ln\frac{K}{\beta}}}, \frac{R\lambda^\alpha}{K(\zeta_\lambda^{\alpha+1})}, \frac{R}{\sigma_\omega\sqrt{dK\ln\frac{K}{\beta}}}\right\}\right), \tag{152}$$

Taking γ equal to the right-hand side, we get that with probability at least $1-\beta$

$$\min_{t\in[0,K]} f(x^t) - f(x^\star) = \mathcal{O}\left(\{(154), (155), (156), (157)\}\right), \tag{153}$$

with

$$R\lambda^{1-\alpha/2}\sigma^{\alpha/2}\sqrt{\frac{\ln K/\beta}{K}} + \frac{LR^2\sigma^\alpha\ln K/\beta}{\lambda^\alpha K}, \tag{154}$$

$$\frac{R\zeta_\lambda^{\alpha+1}}{\lambda^\alpha} + \frac{LR^2\zeta_\lambda^{2\alpha}}{\lambda^{2\alpha+2}}, \tag{155}$$

$$\frac{LR^2}{K} + \frac{L^3R^4}{\lambda^2K^2}, \tag{156}$$

$$R\sigma_\omega\sqrt{\frac{d\ln\frac{K}{\beta}}{K}} + \frac{LR^2\sigma_\omega^2 d\ln\frac{K}{\beta}}{\lambda^2K}. \tag{157}$$

Similarly to the previous case, we find the optimal λ as the equilibrium of the leading terms in (155) and (157). By doing so, we get the optimal λ:

$$\lambda = \min\left\{\frac{4}{3}LR, \frac{2\varepsilon LR}{\left(d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)\right)^{\frac{1}{2\alpha+2}}+1}\right\}. \tag{158}$$

For this choice of λ, we get that with probability at least $1-\beta$

$$\min_{k\in[0,K]} f(x^t) - f(x^\star) = \mathcal{O}\left(\max\{(160), (161), (162), (163)\}\right), \tag{159}$$

with

$$\min\left\{\sqrt{R^{4-\alpha}L^{2-\alpha}\sigma^\alpha\frac{\ln K/\beta}{K}}, \sqrt{\frac{R^{4-\alpha}(\varepsilon L)^{2-\alpha}\ln^{\frac{3\alpha}{4\alpha+4}}\frac{K}{\beta}}{\left(d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\right)^{\frac{2-\alpha}{4\alpha+4}}K}}\right\}, \tag{160}$$

$$\max\left\{\frac{R^{2-\alpha}\sigma^\alpha}{L^{\alpha-1}}, \frac{R^{2-\alpha}\sigma^\alpha}{\varepsilon}\left(d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)\right)^{\frac{\alpha-1}{2\alpha+2}}\right\}, \tag{161}$$

$$\max\left\{\frac{LR^2}{K^2}, \frac{LR^2}{\varepsilon^2K^2}\left(\left(d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)\right)^{\frac{1}{2\alpha+2}}+1\right)^2\right\} + \frac{LR^2}{K}, \tag{162}$$

$$\min\left\{\frac{LR^2}{\varepsilon}\sqrt{d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)}, \frac{LR^2\sqrt{\ln\frac{K}{\beta}}}{\left(d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)\right)^{\frac{1}{2\alpha+2}}+1}\right\}$$
$$+\frac{LR^2 d}{\varepsilon^2}\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right), \tag{163}$$

where, for the sake of brevity, we only report the dominant terms.

F Missing Proofs: Non-Convex Case

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

Lemma F.1. *Let Assumptions 2.1, 2.2 hold on the set* $Q = \left\{x \in \mathbb{R}^d | \exists y \in \mathbb{R}^d : f(y) \leq f^* + 2\Delta \text{ and } \|x - y\| \leq \sqrt{\Delta}/20\sqrt{L}\right\}$, *where* $\Delta \geq \Delta_0 = f(x^0) - f^*$ *and let* $0 < \gamma \leq 1/4L$. *If* $x^k \in Q$ *for all* $k = 0, 1, \ldots, K$ *for some* $K \geq 0$, *then the iterates produced by* DP-Clipped-SGD *satisfy*

$$\frac{\gamma}{2(T+1)}\sum_{t=0}^{T} c_t \left\|\nabla f(x^t)\right\|^2 \leq \frac{(f(x^0)-f^*)-(f(x^{T+1})-f^*)}{T+1} - \frac{\gamma}{T+1}\sum_{t=0}^{T}\langle \nabla f(x^t), \theta_t\rangle - \frac{\gamma}{T+1}\sum_{t=0}^{T}\langle \nabla f(x^t), \omega_t\rangle + \frac{2L\gamma^2}{T+1}\sum_{t=0}^{T}\|\theta_t\|^2 + \frac{L\gamma^2}{T+1}\sum_{t=0}^{T}\|\omega_t\|^2,$$

for all $T = 0, 1, \ldots, K$, *and* θ_t, c_t *are defined in* (35), (34) *respectively.*

Proof. The smoothness of f implies

$$\begin{aligned} f(x^{t+1}) &\leq f(x^t) + \langle \nabla f(x^t), x^{t+1} - x^t\rangle + \frac{L}{2}\left\|x^{t+1} - x^t\right\|^2 \\ &= f(x^t) - \gamma\langle \nabla f(x^t), \hat{g}_t + \omega_t + c_t\nabla f(x^t) - c_t \nabla f(x^t)\rangle \\ &\quad + \frac{L\gamma^2}{2}\left\|\hat{g}_t + \omega_t + c_t\nabla f(x^t) - c_t\nabla f(x^t)\right\|^2 \\ &\leq f(x^t) - \gamma c_t\left\|\nabla f(x^t)\right\|^2 - \gamma\langle \nabla f(x^t), \theta_t\rangle - \gamma\langle\nabla f(x^t), \omega_t\rangle + L\gamma^2\|\omega_t\|^2 \\ &\quad + 2L\gamma^2\|\theta_t\|^2 + 2L\gamma^2 c_t^2\left\|\nabla f(x^t)\right\|^2 \\ &= f(x^t) - (\gamma c_t - 2\gamma^2 L c_t^2)\left\|\nabla f(x^t)\right\|^2 - \gamma\langle \nabla f(x^t), \theta_t\rangle - \gamma\langle \nabla f(x^t), \omega_t\rangle \\ &\quad + L\gamma^2\|\omega_t\|^2 + +2L\gamma^2\|\theta_t\|^2. \end{aligned} \tag{164}$$

Rearranging the terms, utilizing $\gamma \leq 1/4L$, and $c_t^2 \leq c_t$, we sum over t to obtain

$$\frac{\gamma}{2(T+1)}\sum_{t=0}^{T} c_t \left\|\nabla f(x^t)\right\|^2 \leq \frac{(f(x^0)-f^*)-(f(x^{T+1})-f^*)}{T+1} - \frac{\gamma}{T+1}\sum_{t=0}^{T}\langle \nabla f(x^t), \theta_t\rangle - \frac{\gamma}{T+1}\sum_{t=0}^{T}\langle \nabla f(x^t), \omega_t\rangle + \frac{2L\gamma^2}{T+1}\sum_{t=0}^{T}\|\theta_t\|^2 + \frac{L\gamma^2}{T+1}\sum_{t=0}^{T}\|\omega_t\|^2,$$

which concludes the proof. □

The above lemma is utilized to prove the main convergence result for DP-Clipped-SGD.

Theorem F.1. *Let Assumptions 2.1, 2.2, and 2.4 hold for the following set* $Q = \left\{x \in \mathbb{R}^d | \exists y \in \mathbb{R}^d : f(y) \leq f^* + 2\Delta \text{ and } \|x - y\| \leq \sqrt{\Delta}/20\sqrt{L}\right\}$, *where* $\Delta \geq \Delta_0 = f(x^0) - f^*$, $\zeta_\lambda = \max\{0, 2\sqrt{L\Delta} - \frac{\lambda}{2}\}$, *and* $\gamma = \min\{1/4L, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6\}$,

$$\gamma_1 \stackrel{\text{def}}{=} \frac{\sqrt{\Delta}}{21\sqrt{L}(2^{2\alpha-1}+1)^{1/2}\sigma^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{6(K+1)\ln\frac{8(K+1)}{\beta}\left(1+\frac{\zeta_\lambda^\alpha}{\sigma^\alpha}\right)}}, \tag{165}$$

$$\gamma_2 \stackrel{\text{def}}{=} \frac{\sqrt{\Delta}\lambda^{\alpha-1}}{14\sqrt{L}(K+1)2^{2\alpha-1}\left(\sigma^\alpha+\zeta_\lambda^\alpha\right)\left(\frac{\zeta_\lambda}{\lambda}+\frac{1}{2}+\frac{\lambda^{\alpha-1}\zeta_\lambda}{2^{2\alpha-1}\left(\sigma^\alpha+\zeta_\lambda^\alpha\right)}+\left(1+\frac{\zeta_\lambda^\alpha}{\sigma^\alpha}\right)^{-1/\alpha}\right)}, \tag{166}$$

$$\gamma_3 \stackrel{\text{def}}{=} \frac{\sqrt{\Delta}}{14\sqrt{L}\sigma_\omega\sqrt{d(K+1)}(\sqrt{2}+\sqrt{2}\phi)}, \tag{167}$$

$$\gamma_4 \stackrel{\text{def}}{=} \frac{\sqrt{\Delta}}{20\sqrt{L}\left(\lambda+\sigma_\omega\left(\sqrt{d}+\sqrt{2\ln\left(\frac{K+1}{\beta}\right)}\right)\right)}, \tag{168}$$

$$\gamma_5 \quad \stackrel{\text{def}}{=} \quad \frac{\sqrt{\Delta}}{28\lambda\sqrt{L}\ln\frac{8(K+1)}{\beta}}, \tag{169}$$

$$\gamma_6 \quad \stackrel{\text{def}}{=} \quad \frac{\sqrt{\Delta}}{\sqrt{L}\sigma_w\sqrt{7\left((K+1)d+2\sqrt{(K+1)d\ln\frac{4(K+1)}{\beta}}+2\ln\frac{4(K+1)}{\beta}\right)}}. \tag{170}$$

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

$$\min_{t\in[0,K]}\left\|\nabla f(x^t)\right\|^2 \leq \frac{8\Delta}{\gamma(K+1)} + \frac{128\Delta^2}{\lambda^2\gamma^2(K+1)^2}. \tag{171}$$

Proof. Let $\Delta_k = f(x^k) - f^*$ for all $k \geq 0$. We aim to show by induction that $\Delta_l \leq 2\Delta$ with high probability. This fact will allow us to apply Lemma F.1 and then use Bernstein's inequality to evaluate the stochastic part of the upper-bound. More precisely, for each $k = 0, \ldots, K$ we define the probability event E_k as follows. The inequalities

$$-\gamma\sum_{t=0}^{T}\langle\nabla f(x^t), \omega_t+\theta_t\rangle + L\gamma^2\sum_{t=0}^{T}\left(2\left\|\theta_t\right\|^2+\left\|\omega_t\right\|^2\right) \leq \Delta, \tag{172}$$

$$\Delta_t \leq 2\Delta, \tag{173}$$

$$\|\omega_t\| \leq \sigma_\omega\left(\sqrt{d}+\sqrt{2\ln\left(\frac{K+1}{(t+1)\beta}\right)}\right), \tag{174}$$

hold for all $t = 0, 1, \ldots, k$ simultaneously. We want to prove via induction that $\mathbb{P}\{E_k\} \geq 1 - {(k+1)\beta}/{(K+1)}$ for all $k = 0, 1, \ldots, K$. For $k = 0$ the statement is trivial. Assume that the statement is true for some $k = T-1 \leq K$ and $\mathbb{P}\{E_{T-1}\} \geq 1 - {T\beta}/{(K+1)}$. One needs to prove that $\mathbb{P}\{E_T\} \geq 1 - {(T+1)\beta}/{(K+1)}$. First, we notice that the probability event E_{T-1} implies $\Delta_t \leq 2\Delta$ 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

$$\|x^T - x^{T-1}\| = \gamma\|\hat{g}_{T-1}\| + \gamma\|\omega_{T-1}\| \leq \gamma\lambda + \gamma\sigma_\omega\left(\sqrt{d}+\sqrt{2\ln\left(\frac{K+1}{T\beta}\right)}\right) \stackrel{(168)}{\leq} \frac{\sqrt{\Delta}}{20\sqrt{L}}.$$

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

$$\begin{aligned}\frac{\gamma}{2}\sum_{l=0}^{t-1}\|\nabla f(x^l)\|^2 \quad &\leq \quad \Delta_0 - \Delta_t - \gamma\sum_{l=0}^{t-1}\langle\nabla f(x^l), \theta_l\rangle - \gamma\sum_{l=0}^{t-1}\langle\nabla f(x^l), \omega_l\rangle + 2L\gamma^2\sum_{l=0}^{t-1}\|\theta_l\|^2 \\ &+ \quad L\gamma^2\sum_{l=0}^{t-1}\|\omega_l\|^2,\end{aligned}$$

for all $t = 0, 1, \ldots, T$ simultaneously. This event also implies

$$\begin{aligned}\frac{\gamma}{2}\sum_{l=0}^{t-1}c_l\|\nabla f(x^l)\|^2 \quad &\leq \quad \Delta - \gamma\sum_{k=0}^{t-1}\langle\nabla f(x^l), \theta_l\rangle - \gamma\sum_{k=0}^{t-1}\langle\nabla f(x^l), \omega_l\rangle + 2L\gamma^2\sum_{l=0}^{t-1}\|\theta_l\|^2 \\ &+ \quad L\gamma^2\sum_{l=0}^{t-1}\|\omega_l\|^2 \\ &\leq \quad 2\Delta.\end{aligned} \tag{175}$$

Taking into account that $\frac{\gamma}{2}\sum\limits_{l=0}^{T-1}c_l\|\nabla f(x^l)\|^2 \geq 0$, E_{T-1} also implies

$$\Delta_T \leq \Delta - \gamma\sum_{l=0}^{T-1}\langle\nabla f(x^l), \theta_l\rangle - \gamma\sum_{l=0}^{T-1}\langle\nabla f(x^l), \omega_l\rangle + 2L\gamma^2\sum_{l=0}^{T-1}\|\theta_l\|^2 + L\gamma^2\sum_{l=0}^{T-1}\|\omega_l\|^2.$$

Next, we define random vectors

$$\eta_t = \begin{cases} \nabla f(x^t), & \text{if } \|\nabla f(x^t)\| \leq 2\sqrt{L\Delta}, \\ 0, & \text{otherwise,} \end{cases} \tag{176}$$

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

$$\|\eta_t\| \leq 2\sqrt{L\Delta}. \tag{177}$$

Moreover, for $t = 1, \ldots, T-1$ event E_{T-1}, and corollary of smoothness imply

$$\|\nabla f(x^l)\| \overset{(176)}{\leq} \sqrt{2L(f(x^l) - f^*)} = \sqrt{2L\Delta_l} \leq 2\sqrt{L\Delta}, \tag{178}$$

meaning that E_{T-1} implies that $\eta_t = \nabla f(x^t)$ for all $t = 0, 1, \ldots, T-1$. We notice that $\theta_t = \theta_t^u + \theta_t^b$, where θ_t^u and θ_t^b are defined in (51). Using new notation, we get that E_{T-1} implies

$$\begin{aligned} \Delta_T \quad \leq \quad & \Delta \underbrace{-\gamma \sum_{t=0}^{T-1} \langle \theta_t^u, \eta_t \rangle}_{①} \underbrace{-\gamma \sum_{t=0}^{T-1} \langle \theta_t^b, \eta_t \rangle}_{②} \underbrace{-\gamma \sum_{t=0}^{T-1} \langle \omega_t, \eta_t \rangle}_{③} + \underbrace{4L\gamma^2 \sum_{t=0}^{T-1} \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right]}_{④} \\ & + \underbrace{4L\gamma^2 \sum_{t=0}^{T-1} \left(\|\theta_t^u\|^2 - \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right]\right)}_{⑤} + \underbrace{4L\gamma^2 \sum_{t=0}^{T-1} \left\|\theta_t^b\right\|^2}_{⑥} + \underbrace{L\gamma^2 \sum_{t=0}^{T-1} \|\omega_t\|^2}_{⑦}. \end{aligned} \tag{179}$$

It remains to derive good enough high-probability upper bounds for the terms ①, ②, ③, ④, ⑤, ⑥, ⑦. This amounts to proving ① + ② + ③ + ④ + ⑤ + ⑥ + ⑦ $\leq \Delta$ with high probability. In the subsequent parts of the proof, we will need to use the bounds for the norm and second moments of θ_t^u and θ_t^b many times. First, by definition of the clipping operator, we have with probability 1 that

$$\|\theta_t^u\| \leq 2\lambda, \tag{180}$$

and from Lemma B.1 we also have

$$\begin{aligned} \|\theta_t^b\| \quad & \leq \frac{2^{2\alpha-1}\sigma\left(\sigma^\alpha + (\max\{0, \|\nabla f(x^t)\| - \lambda/2\})^\alpha\right)^{\frac{\alpha-1}{\alpha}}}{\lambda^{\alpha-1}} \\ & + \max\{\left\|\nabla f(x^t)\right\|, \lambda/2\} \frac{2^{2\alpha-1}\left(\sigma^\alpha + (\max\{0, \|\nabla f(x^t)\| - \lambda/2\})^\alpha\right)}{\lambda^\alpha} \\ & + \max\{0, \left\|\nabla f(x^t)\right\| - \lambda/2\}, \end{aligned}$$

$$\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right] \leq \frac{9(2^{2\alpha-1}+1)\lambda^{2-\alpha}\sigma^\alpha}{4} + \frac{9(2^{2\alpha-1}+1)\lambda^{2-\alpha}(\max\{0, \|\nabla f(x^t)\| - \lambda/2\})^\alpha}{4}.$$

As can be seen, these bounds are iteration-dependent. To overcome this, we bound $\|\nabla f(x^t)\|$ by $2\sqrt{L\Delta}$, which follows from E_{T-1}, i.e., E_{T-1} implies

$$\|\theta_t^b\| \quad \leq \frac{2^{2\alpha-1}\sigma\left(\sigma^\alpha + \zeta_\lambda^\alpha\right)^{\frac{\alpha-1}{\alpha}}}{\lambda^{\alpha-1}} + \left(\zeta_\lambda + \frac{\lambda}{2}\right) \frac{2^{2\alpha-1}\left(\sigma^\alpha + \zeta_\lambda^\alpha\right)}{\lambda^\alpha} + \zeta_\lambda, \tag{181}$$

$$\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right] \leq \frac{9(2^{2\alpha-1}+1)\lambda^{2-\alpha}\sigma^\alpha}{4} + \frac{9(2^{2\alpha-1}+1)\lambda^{2-\alpha}\zeta_\lambda^\alpha}{4}. \tag{182}$$

Upper bound for ①. By definition of θ_t^u, we have $\mathbb{E}\left[\theta_t^u \mid \mathcal{F}^{t-1}\right] = 0$ and

$$\mathbb{E}\left[-\gamma\langle \theta_t^u, \eta_t \rangle \mid \mathcal{F}^{t-1}\right] = 0.$$

Next, sum ① has bounded with probability 1 terms:

$$|\gamma \langle \theta_t^u, \eta_t \rangle| \leq \gamma \|\theta_t^u\| \cdot \|\eta_t\| \overset{(176)}{\leq} 4\gamma\lambda\sqrt{L\Delta} \overset{(169)}{\leq} \frac{\Delta}{7\ln\frac{8(K+1)}{\beta}} \overset{\text{def}}{=} c. \tag{183}$$

The summands also have bounded conditional variances $\sigma_t^2 \overset{\text{def}}{=} \mathbb{E}\left[\gamma^2 \langle \theta_t^u, \eta_t \rangle^2 \mid \mathcal{F}^{t-1}\right]$:

$$\sigma_t^2 \leq \mathbb{E}\left[\gamma^2 \|\theta_t^u\|^2 \cdot \|\eta_t\|^2 \mid \mathcal{F}^{t-1}\right] \leq 4\gamma^2 L\Delta \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right]. \tag{184}$$

In other words, we showed that $\{-\gamma \langle \theta_t^u, \eta_t \rangle\}_{t=0}^{T-1}$ is a bounded martingale difference sequence with bounded conditional variances $\{\sigma_t^2\}_{t=0}^{T-1}$. Next, we apply Bernstein's inequality (Lemma A.1) with $X_t = -\gamma \langle \theta_t^u, \eta_t \rangle$, parameter c as in (183), $b = \frac{\Delta}{7}$, $G = \frac{\Delta^2}{294\ln\frac{8(K+1)}{\beta}}$:

$$\mathbb{P}\left\{|①| > \frac{\Delta}{7} \quad \text{and} \quad \sum_{t=0}^{T-1} \sigma_t^2 \leq \frac{\Delta^2}{294\ln\frac{8(K+1)}{\beta}}\right\} \leq 2\exp\left(-\frac{b^2}{2G + {}^{2cb}/_{3}}\right) = \frac{\beta}{4(K+1)}.$$

Equivalently, we have

$$\mathbb{P}\{E_①\} \geq 1 - \frac{\beta}{4(K+1)}, \quad \text{for} \quad E_① = \left\{\text{either} \quad \sum_{t=0}^{T-1} \sigma_t^2 > \frac{\Delta^2}{294\ln\frac{8(K+1)}{\beta}} \quad \text{or} \quad |①| \leq \frac{\Delta}{7}\right\}. \tag{185}$$

In addition, E_{T-1} implies that

$$\begin{aligned}
\sum_{t=0}^{T-1} \sigma_t^2 \quad & \leq 4\gamma^2 L\Delta \sum_{t=0}^{T-1} \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right] \\
& \overset{(182)}{\leq} 9\gamma^2 L\Delta T\left(\left(2^{2\alpha-1}+1\right)\lambda^{2-\alpha}\sigma^\alpha + (2^{2\alpha-1}+1)\lambda^{2-\alpha}\zeta_\lambda\right) \\
& \overset{(165)}{\leq} \frac{\Delta^2}{294\ln\frac{8(K+1)}{\beta}}.
\end{aligned} \tag{186}$$

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

$$\begin{aligned}
② \quad & = -\gamma \sum_{t=0}^{T-1} \langle \theta_t^b, \eta_t \rangle \leq \gamma \sum_{t=0}^{T-1} \|\theta_t^b\| \cdot \|\eta_t\| \\
& \overset{(181)}{\leq} 2\gamma\sqrt{L\Delta}T\left(\frac{2^{2\alpha-1}\sigma\left(\sigma^\alpha + \zeta_\lambda^\alpha\right)^{\frac{\alpha-1}{\alpha}}}{\lambda^{\alpha-1}} + (\zeta_\lambda + {}^{\lambda}/_{2})\frac{2^{2\alpha-1}\left(\sigma^\alpha + \zeta_\lambda^\alpha\right)}{\lambda^\alpha} + \zeta_\lambda\right) \\
& \overset{(166)}{\leq} \frac{\Delta}{7}.
\end{aligned} \tag{187}$$

Upper bound for ③. We have

$$|③| = \left|-\gamma \sum_{t=0}^{T-1} \langle \omega_t, \eta_t \rangle\right| = \left|\sum_{t=0}^{T-1} \sum_{i=1}^{d} \gamma\omega_{t,i}, \eta_{t,i}\right|, \tag{188}$$

where $\eta_{t,i} \overset{\text{def}}{=} [\eta_t]_i$ and $\omega_{t,i} \overset{\text{def}}{=} [\omega_t]_i$ denote the i-th components of η_t and ω_t respectively.

Each summand is the product of a zero-mean Gaussian random variable and a bounded random variable, resulting in the product being a zero-mean light-tailed random variable with parameter $\sigma_{t,i}^2 = 16\gamma^2 L\Delta\sigma_\omega^2$. To prove this, consider

$$\mathbb{E}\left[\exp\left(\frac{\gamma^2}{\sigma_{t,i}^2}\left|\eta_{t,i}^2\omega_{t,i}^2\right|\right) \mid \mathcal{F}^{t-1}\right] \quad \overset{(177)}{\leq} \quad \mathbb{E}\left[\exp\left(\frac{4L\Delta\gamma^2}{16\gamma^2 L\Delta\sigma_\omega^2}\left|\omega_{t,i}\right|^2\right)\right]$$

$$\leq \quad \exp\left(\frac{|\omega_{t,i}|^2}{4\sigma_\omega^2}\right) \overset{(ii)}{\leq} \exp(1), \tag{189}$$

where (ii) uses the fact that $\omega_{t,i}^2$ is a sub-Gaussian random variable with parameter σ_ω^2. Now that we have established the light-tailedness of summands, we can use the Lemma A.2 to obtain

$$\mathbb{P}\left\{\left|\sum_{t=0}^{T-1}\sum_{i=1}^{d}\gamma\eta_{t,i}\omega_{t,i}\right| > \left(\sqrt{2}+\sqrt{2}\phi\right)\sqrt{\sum_{t=0}^{K}\sum_{i=1}^{d}4\gamma^2 L\Delta\sigma_\omega^2}\right\} \leq \exp\left(\frac{-\phi^2}{3}\right) \tag{190}$$

$$= \frac{\beta}{4(K+1)}. \tag{191}$$

The choice of $\gamma \leq \gamma_3$ for γ_3 defined in (167) implies

$$\left(\sqrt{2}+\sqrt{2}\phi\right)\sqrt{\sum_{t=0}^{T-1}\sum_{i=1}^{d}4\gamma^2 L\Delta\sigma_\omega^2} \leq \left(\sqrt{2}+\sqrt{2}\phi\right)\sqrt{4\gamma^2 L\Delta(K+1)d\sigma_\omega^2} \overset{(167)}{\leq} \frac{\Delta}{7},$$

and

$$\mathbb{P}\{E_{③}\} \geq 1 - \frac{\beta}{4(K+1)} \quad \text{for} \quad E_{③} = \left\{|③| > \frac{\Delta}{7}\right\}. \tag{192}$$

Upper bound for ④. From E_{T-1} and the conditions on the step-size, it follows that

$$\begin{aligned} ④ &= 2L\gamma^2\sum_{t=0}^{T-1}\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right] \\ &\overset{(182)}{\leq} 2LT\gamma^2\left(\frac{9(2^{2\alpha-1}+1)\lambda^{2-\alpha}\sigma^\alpha}{4} + \frac{9(2^{2\alpha-1}+1)\lambda^{2-\alpha}\zeta_\lambda^\alpha}{4}\right) \\ &\overset{(165)}{\leq} \frac{\Delta}{7}. \end{aligned} \tag{193}$$

Upper bound for ⑤. First, we have

$$\mathbb{E}\left[2L\gamma^2\left(\|\theta_t^u\|^2 - \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right]\right) \mid \mathcal{F}^{t-1}\right] = 0.$$

Next, sum ⑤ has bounded with probability 1 terms:

$$\begin{aligned} \left|2L\gamma^2\left(\|\theta_t^u\|^2 - \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right] \mid \mathcal{F}^{t-1}\right)\right| &\leq 2L\gamma^2\left(\|\theta_t^u\|^2 + \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right]\right) \\ &\leq 16L\gamma^2\lambda^2 \overset{(169)}{\leq} \frac{\Delta}{7\ln\frac{8(K+1)}{\beta}} \overset{\text{def}}{=} c. \end{aligned} \tag{194}$$

The summands also have bounded conditional variances as shown below:

$$\widetilde{\sigma}_t^2 \overset{\text{def}}{=} \mathbb{E}\left[4L^2\gamma^4\left(\|\theta_t^u\|^2 - \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right]\right)^2 \mid \mathcal{F}^{t-1}\right] \tag{195}$$

$$\begin{aligned} \widetilde{\sigma}_t^2 &\overset{(194)}{\leq} \frac{\Delta}{7\ln\frac{8(K+1)}{\beta}}\mathbb{E}\left[2L\gamma^2\left|\|\theta_t^u\|^2 - \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right]\right| \mid \mathcal{F}^{t-1}\right] \\ &\leq \frac{4L\gamma^2\Delta}{7\ln\frac{8(K+1)}{\beta}}\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}\right], \end{aligned} \tag{196}$$

since $\ln\frac{8K}{\beta}\geq 1$. In other words, we showed that $\left\{2L\gamma^2\left(\|\theta_t^u\|^2-\mathbb{E}\left[\|\theta_t^u\|^2\mid\mathcal{F}^{t-1}\right]\right)\right\}_{t=0}^{T-1}$ 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 A.1) with $X_t = 2L\gamma^2\left(\|\theta_t^u\|^2-\mathbb{E}\left[\|\theta_t^u\|^2\mid\mathcal{F}^{t-1}\right]\right)$, parameter c as in (194), $b=\frac{\Delta}{7}$, $G=\frac{\Delta^2}{294\ln\frac{8(K+1)}{\beta}}$:

$$\mathbb{P}\left\{|⑤|>\frac{\Delta}{7}\quad\text{and}\quad\sum_{t=0}^{T-1}\widetilde{\sigma}_t^2\leq\frac{\Delta^2}{294\ln\frac{8(K+1)}{\beta}}\right\}\leq 2\exp\left(-\frac{b^2}{2G+{}^{2cb}/_3}\right)=\frac{\beta}{4(K+1)}.$$

Equivalently, we have

$$\mathbb{P}\left\{E_{⑤}\right\}\geq 1-\frac{\beta}{4(K+1)},\quad\text{for}\quad E_{④}=\left\{\text{either}\quad\sum_{t=0}^{T-1}\widetilde{\sigma}_t^2>\frac{\Delta^2}{294\ln\frac{8(K+1)}{\beta}}\quad\text{or}\quad|⑤|\leq\frac{\Delta}{7}\right\}. \tag{197}$$

In addition, E_{T-1} implies that

$$\sum_{t=0}^{T-1}\widetilde{\sigma}_t^2 \;\leq\; \frac{4L\gamma^2\Delta}{7\ln\frac{8(K+1)}{\beta}}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\theta_t^u\|^2\mid\mathcal{F}^{t-1}\right]\overset{(182),(165)}{\leq}\frac{\Delta^2}{294\ln\frac{8(K+1)}{\beta}}. \tag{198}$$

Upper bound for ⑥. From E_{T-1}, and the conditions on the step-size it follows that

$$⑥ \;=\; L\gamma^2\sum_{t=0}^{T-1}\left\|\theta_t^b\right\|^2 \tag{199}$$

$$\leq\; L\gamma^2\left(\frac{2^{2\alpha-1}\sigma\left(\sigma^\alpha+\zeta_\lambda^\alpha\right)^{\frac{\alpha-1}{\alpha}}}{\lambda^{\alpha-1}}+(\zeta_\lambda+{}^{\lambda}/_2)\frac{2^{2\alpha-1}\left(\sigma^\alpha+\zeta_\lambda^\alpha\right)}{\lambda^\alpha}+\zeta_\lambda\right)^2$$

$$\overset{(166)}{\leq}\;\frac{\Delta}{7}. \tag{200}$$

Upper bound for ⑦. We have

$$⑦ = L\gamma^2\sum_{t=0}^{T-1}\|\omega_t\|^2 = L\gamma^2\sigma_\omega^2\sum_{t=0}^{T-1}\sum_{i=1}^{d}z_{t,i}^2, \tag{201}$$

where $z_{t,i}\overset{\text{def}}{=}{}^{\omega_{t,i}}/_{\sigma_\omega}$. Using Lemma A.3, we get

$$\mathbb{P}\left\{\sum_{t=0}^{T-1}\sum_{i=1}^{d}z_{t,i}^2>Td+2\sqrt{Td\ln\frac{4(K+1)}{\beta}}+2\ln\frac{4(K+1)}{\beta}\right\}\leq\frac{\beta}{4(K+1)}. \tag{202}$$

Since $\gamma\leq\gamma_6$, for γ_6 defined in (170)

$$\mathbb{P}\left\{⑦>\frac{\Delta}{7}\right\}\leq\frac{\beta}{4(K+1)}. \tag{203}$$

Equivalently, we have

$$\mathbb{P}\{E_{⑦}\}\geq 1-\frac{\beta}{4(K+1)}\quad\text{for}\quad E_{⑦}=\left\{|⑦|\leq\frac{\Delta}{7}\right\}. \tag{204}$$

Now, we have the upper bounds for ①, ②, ③, ④, ⑤, ⑥, ⑦ . Thus, probability event $E_{T-1}\cap E_{①}\cap E_{③}\cap E_{④}\cap E_{⑦}$ implies

$$\Delta_T \;\leq\; \Delta+\frac{\Delta}{7}+\frac{\Delta}{7}+\frac{\Delta}{7}+\frac{\Delta}{7}+\frac{\Delta}{7}+\frac{\Delta}{7}+\frac{\Delta}{7}=2\Delta,$$

which is equivalent to (172) and (173) for $t = T$, and

$$\mathbb{P}\{E_T\} \geq \mathbb{P}\left\{E_{T-1} \cap E_{①} \cap E_{③} \cap E_{④} \cap E_{⑦}\right\} = 1 - \mathbb{P}\left\{\overline{E}_{T-1} \cup \overline{E}_{①} \cup \overline{E}_{③} \cup \overline{E}_{④} \cup \overline{E}_{⑦}\right\}$$

$$\geq 1 - \mathbb{P}\{\overline{E}_{T-1}\} - \mathbb{P}\{\overline{E}_{①}\} - \mathbb{P}\{\overline{E}_{③}\} - \mathbb{P}\{\overline{E}_{④}\} - \mathbb{P}\{\overline{E}_{⑦}\} \geq 1 - \frac{(T+1)\beta}{K+1}. \tag{205}$$

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

$$\frac{1}{K+1}\sum_{t=0}^{K} c_t \|\nabla f(x^t)\|^2 \overset{(175)}{\leq} \frac{4\Delta}{\gamma(K+1)},$$

and $\{x^t\}_{t=0}^{K} \in Q$, which follows from (173). Now we have to deal with c_t. To do so, we consider two possible cases for each $t = 0, 1, \ldots, K$. We either have $c_t = 1$ or $c_t = \frac{\lambda}{2\|\nabla f(x^t)\|}$. We define the corresponding sets of indices: $\mathcal{T}_1 \overset{\text{def}}{=} \{t \in \{0, 1, \ldots, K\} \mid c_t = 1\}$ and $\mathcal{T}_2 \overset{\text{def}}{=} \{t \in \{0, 1, \ldots, K\} \mid c_t = \frac{\lambda}{2\|\nabla f(x^t)\|}\}$. Then, the above inequality can be written as

$$\frac{1}{(K+1)}\sum_{t \in \mathcal{T}_1} \left\|\nabla f(x^t)\right\|^2 + \frac{1}{(K+1)}\sum_{t \in \mathcal{T}_2} \frac{\lambda \left\|\nabla f(x^t)\right\|^2}{2\left\|\nabla f(x^t)\right\|} \leq \frac{4\Delta}{\gamma(K+1)},$$

implying

$$\frac{1}{(K+1)}\sum_{t \in \mathcal{T}_1} \left\|\nabla f(x^t)\right\|^2 \leq \frac{4\Delta}{\gamma(K+1)}, \tag{206}$$

and

$$\frac{1}{K+1}\sum_{t \in \mathcal{T}_2} \left\|\nabla f(x^t)\right\| \leq \frac{8\Delta}{\lambda\gamma(K+1)}, \tag{207}$$

For inequality (206), we follow the technique from (Koloskova et al., 2023) and apply inequality $x^2 \geq 2\epsilon x - \epsilon^2$, holding for any $\epsilon, x > 0$. Taking $x = \|\nabla f(x^t)\|^2$, we get

$$\frac{1}{K+1}\sum_{t \in \mathcal{T}_1} \left(2\epsilon\left\|\nabla f(x^t)\right\| - \epsilon^2\right) \leq \frac{4\Delta}{\gamma(K+1)},$$

implying

$$\frac{1}{K+1}\sum_{t \in \mathcal{T}_1} \left\|\nabla f(x^t)\right\| \leq \frac{2\Delta}{\gamma(K+1)\epsilon} + \frac{\epsilon}{2}.$$

Upon selecting $\epsilon = \frac{2\sqrt{\Delta}}{\sqrt{\gamma(K+1)}}$, we obtain

$$\frac{1}{K+1}\sum_{t \in \mathcal{T}_1} \left\|\nabla f(x^t)\right\| \leq \sqrt{\frac{4\Delta}{\gamma(K+1)}}. \tag{208}$$

Combining inequalities (206) and (207) we get:

$$\frac{1}{K+1}\sum_{t=0}^{K} \left\|\nabla f(x^t)\right\| \leq \sqrt{\frac{4\Delta}{\gamma(K+1)}} + \frac{8\Delta}{\lambda\gamma(K+1)}. \tag{209}$$

Upon considering the best iterate, we have the following bound

$$\min_{t \in [0,K]} \left\|\nabla f(x^t)\right\|^2 \leq \frac{8\Delta}{\gamma(K+1)} + \frac{128\Delta^2}{\lambda^2\gamma^2(K+1)^2}. \tag{210}$$

□

Theorem F.1 states 7 values for the step-size, from which the smallest should be selected. To simplify matters, we demonstrate that if λ is selected equal or smaller than the order of $\mathcal{O}\left(\left(\frac{K}{\ln K}\right)^{1/\alpha}\right)$, then three step-sizes are redundant and can be omitted.

Corollary F.1. *Let all conditions of Theorem F.1 hold. Furthermore, assume that K is large and one selects $\lambda \leq \mathcal{O}\left(\left(\frac{K}{\ln K}\right)^{1/\alpha}\right)$, then conclusions of Theorem F.1 are valid as long as γ is selected to satisfy $\gamma \leq \min\{1/4L, \gamma_1, \gamma_2, \gamma_3\}$ where we have*

$$\gamma_1 \overset{\text{def}}{=} \frac{\sqrt{\Delta}}{21\sqrt{L}(2^{2\alpha-1}+1)^{1/2}\sigma^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{6(K+1)\ln\frac{8(K+1)}{\beta}\left(1+\frac{\zeta_\lambda^\alpha}{\sigma^\alpha}\right)}},$$

$$\gamma_2 \overset{\text{def}}{=} \frac{\sqrt{\Delta}\lambda^{\alpha-1}}{14\sqrt{L}(K+1)2^{2\alpha-1}\left(\sigma^\alpha+\zeta_\lambda^\alpha\right)\left(\frac{\zeta_\lambda}{\lambda}+\frac{1}{2}+\frac{\lambda^{\alpha-1}\zeta_\lambda}{2^{2\alpha-1}\left(\sigma^\alpha+\zeta_\lambda^\alpha\right)}+\left(1+\frac{\zeta_\lambda^\alpha}{\sigma^\alpha}\right)^{-1/\alpha}\right)},$$

$$\gamma_3 \overset{\text{def}}{=} \frac{\sqrt{\Delta}}{14\sqrt{L}\sigma_\omega\sqrt{d(K+1)}(\sqrt{2}+\sqrt{2}\phi)}.$$

Proof. For large K, it is evident that γ_3 decreases at a rate of $\mathcal{O}\left(\sigma_\omega\sqrt{K\ln K}\right)$, while γ_6 in (170) decreases at a rate of $\mathcal{O}\left(\sigma_\omega\sqrt{K}\right)$. Subsequently, γ_3 dominates γ_6 and γ_6 can be omitted. Furthermore, γ_5 in (169) decreases with a rate of $\mathcal{O}\left(K^{1/\alpha}(\ln K)^{1-1/\alpha}\right)$ which is less than the rate of γ_2. It can be deduced that for large λ, γ_2 decreases at the rate $\mathcal{O}(K)$ which is faster than γ_5. If λ is small, γ_2 dominates γ_5 again due to the λ in the numerator of γ_2. Hence, γ_5 can be discarded. As for γ_4 in (168), we know that σ_ω is on the order of $\mathcal{O}\left(\lambda/\epsilon\sqrt{K\ln(K/\delta)}\right)$. Hence, one can replace λ with $\mathcal{O}\left(\sigma_\omega\epsilon/\sqrt{K\ln(K/\delta)}\right)$. Therefore, γ_4 decreases by the order $\mathcal{O}\left(\sigma_\omega\epsilon\sqrt{K\ln(K/\delta)}\right)$, which is the same order as γ_3. Hence, γ_4 can be omitted, and the proof is complete. □

G Rate and Neighborhood for Clipped-SGD: Non-Convex Case

Now that we have established the convergence properties of DP-Clipped-SGD for non-convex problems, we turn to evaluating its convergence rate. This rate depends critically on the choice of the step-size γ, and in general, the resulting expressions can be quite complex. To obtain more interpretable bounds, we consider simplified rate expressions by analyzing separate cases based on different ranges of λ. Since we focus on the asymptotic behavior, numerical constants are omitted for clarity.

In this section, we consider the cases without the DP noise ($\sigma_\omega = 0$) and investigate all possible clipping levels.

Case 1: $\lambda > 4\sqrt{L\Delta}$. In this case, $\zeta_\lambda = 0$, and the step-size conditions reduce to the following:

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\sigma^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K\ln\frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}}\lambda^{\alpha-1}}{K\sigma^\alpha}\right\}\right). \tag{211}$$

In particular, when γ equals the minimum from the above condition, the iterates produced by Clipped-SGD after K iterations with probability at least $1-\beta$ satisfy

$$\min_{t\in[0,K]}\left\|\nabla f(x^t)\right\|^2 = \mathcal{O}\left(\max\left\{(213), (214), (215)\right\}\right), \tag{212}$$

where

$$\sqrt{L\Delta}\lambda^{1-\alpha/2}\sigma^{\alpha/2}\sqrt{\frac{\ln K/\beta}{K}} + \frac{L\Delta\sigma^\alpha\ln K/\beta}{\lambda^\alpha K}, \tag{213}$$

$$\frac{\sqrt{L\Delta}\sigma^\alpha}{\lambda^{\alpha-1}} + \frac{L\Delta\sigma^{2\alpha}}{\lambda^{2\alpha}}, \tag{214}$$

$$\frac{L\Delta}{K} + \frac{L^2\Delta^2}{\lambda^2K^2}. \tag{215}$$

We clearly see that the dominant term (213) is an increasing function of λ, and the dominant term in (214) is a decreasing function. Solving for the optimal λ where the leading terms in (213) and (214) become equal, we obtain $\lambda = \mathcal{O}\left(\sigma\left(\frac{K}{\ln\frac{K}{\beta}}\right)^{\frac{1}{\alpha}}\right)$. Substituting back this λ, we get that with probability at least $1-\beta$

$$\min_{t\in[0,K]}\left\|\nabla f(x^t)\right\|^2 = \mathcal{O}\left(\max\left\{(217), (218)\right\}\right), \tag{216}$$

where

$$\sqrt{L\Delta}\sigma\left(\frac{\ln\frac{K}{\beta}}{K}\right)^{\frac{\alpha-1}{\alpha}} + \frac{L\Delta\ln^2 K/\beta}{K^2}, \tag{217}$$

$$\frac{L\Delta}{K} + \frac{L^2\Delta^2\left(\ln\frac{K}{\beta}\right)^{\frac{2}{\alpha}}}{\sigma^2K^{\frac{2\alpha+2}{\alpha}}}. \tag{218}$$

Note in this case, we converge to the exact optimum, and the dominant term matches (Sadiev et al., 2023). As it can be seen from (213), (214), when the clipping level is not that large, we converge to a neighborhood of the solution, but with a faster rate.

When $\lambda \leq 4\sqrt{L\Delta}$, we have $\zeta_\lambda = \frac{4\sqrt{L\Delta}-\lambda}{2}$. As observed from (165), (166), we also have to consider the relation between λ and σ in these cases. Thus, we split the $\lambda \leq 4\sqrt{L\Delta}$ case into 6 different regimes to cover all possible cases.

Case 2: $\frac{4}{3}\sqrt{L\Delta} < \lambda \leq 4\sqrt{L\Delta}$ $\zeta_\lambda < \lambda < \sigma$. In this case, the step-size conditions reduce to the following:

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\sigma^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K\ln\frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}}\lambda^{\alpha-1}}{K\sigma^\alpha}\right\}\right). \tag{219}$$

As it can be seen, the bounds on step-size are similar to Case 1. However, the optimal λ derived in the previous section violates the constraint that $\lambda \leq 4\sqrt{L\Delta}$. Subsequently, the optimal λ becomes $\lambda = 4\sqrt{L\Delta}$. For this choice of λ, we have that with probability at least $1-\beta$

$$\min_{t\in[0,K]} \left\|\nabla f(x^t)\right\|^2 = \mathcal{O}\left(\max\left\{(221), (222), (223)\right\}\right), \tag{220}$$

where

$$\sqrt{(L\Delta)^{\frac{4-\alpha}{2}}\sigma^\alpha \frac{\ln K/\beta}{K}} + \frac{(L\Delta)^{\frac{2-\alpha}{2}}\sigma^\alpha \ln K/\beta}{K}, \tag{221}$$

$$\frac{\sigma^\alpha}{(\sqrt{L\Delta})^{\alpha-2}} + \frac{\sigma^{2\alpha}}{(L\Delta)^{\alpha-1}}, \tag{222}$$

$$\frac{L\Delta}{K} + \frac{L\Delta}{K^2}. \tag{223}$$

Case 3: $\frac{4}{3}\sqrt{L\Delta} < \lambda \leq 4\sqrt{L\Delta}, \quad \zeta_\lambda < \sigma < \lambda$. In this case, the step-size conditions reduce to

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\sigma^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K\ln\frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}}\lambda^{\alpha-1}}{K\max\{\sigma^\alpha, \lambda^{\alpha-1}\zeta_\lambda\}}\right\}\right). \tag{224}$$

If $\max\{\sigma^\alpha, \lambda^{\alpha-1}\zeta_\lambda\} = \sigma^\alpha$, then the resulting bounds are similar to the previous case. If $\max\{\sigma^\alpha, \lambda^{\alpha-1}\zeta_\lambda\} = \lambda^{\alpha-1}\zeta_\lambda$ is satisfied, $\min_{t\in[0,K]}\left\|\nabla f(x^t)\right\|^2$ is bounded with probability at least $1-\beta$ by the maximum of the following terms:

$$\sqrt{L\Delta}\lambda^{1-\alpha/2}\sigma^{\alpha/2}\sqrt{\frac{\ln K/\beta}{K}} + \frac{L\Delta\sigma^\alpha \ln K/\beta}{\lambda^\alpha K}, \tag{225}$$

$$\sqrt{L\Delta}\zeta_\lambda + \frac{L\Delta\zeta_\lambda^2}{\lambda^2}, \tag{226}$$

$$\frac{L\Delta}{K} + \frac{L^2\Delta^2}{\lambda^2 K^2}. \tag{227}$$

In the latter case (i.e., maximum occurring in the second argument), the optimal λ is $4\sqrt{L\Delta} - \eta$, where η is a sufficiently small number such that $\lambda^{\alpha-1}\zeta_\lambda \geq \sigma^\alpha$, i.e., λ satisfies $\zeta_\lambda = \max\left\{\frac{\sigma^\alpha}{\lambda^{\alpha-1}}, \lambda^{1-\alpha/2}\sigma^{\alpha/2}\sqrt{\frac{\ln K/\beta}{K}}\right\}$. Note that the (226) is decreasing in λ, and $\lambda = 4\sqrt{L\Delta}$ is not feasible. With this choice of λ, we get:

$$\min_{t\in[0,K]} \left\|\nabla f(x^t)\right\|^2 = \mathcal{O}\left(\max\left\{(229), (230), (231)\right\}\right), \tag{228}$$

where

$$\sqrt{L\Delta(4\sqrt{L\Delta}-\eta)^{2-\alpha}\sigma^\alpha \frac{\ln K/\beta}{K}} + \frac{L\Delta\sigma^\alpha \ln K/\beta}{(\sqrt{L\Delta}-\eta)^\alpha K}, \tag{229}$$

$$\frac{\sqrt{L\Delta}\eta}{2} + \frac{L\Delta\eta^2}{(4\sqrt{L\Delta}-\eta)^2}, \tag{230}$$

$$\frac{L\Delta}{K} + \frac{L^2\Delta^2}{(4\sqrt{L\Delta}-\eta)^2K^2}. \tag{231}$$

Case 4: $\frac{4}{3}\sqrt{L\Delta} < \lambda \leq 4\sqrt{L\Delta}, \quad \sigma < \zeta_\lambda < \lambda$. For this case, step-size conditions reduce to

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\zeta_\lambda^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K\ln\frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}}\lambda^{\alpha-1}}{K(\lambda^{\alpha-1}\zeta_\lambda)}\right\}\right), \tag{232}$$

and $\min_{t\in[0,K]}\left\|\nabla f(x^t)\right\|^2$ is bounded with probability at least $1-\beta$ by the maximum of the following terms

$$\sqrt{L\Delta}\lambda^{1-\alpha/2}\zeta_\lambda^{\alpha/2}\sqrt{\frac{\ln K/\beta}{K}} + \frac{L\Delta\zeta_\lambda^\alpha \ln K/\beta}{\lambda^\alpha K}, \tag{233}$$

$$\sqrt{L\Delta}\zeta_\lambda + \frac{L\Delta\zeta_\lambda^2}{\lambda^2}, \tag{234}$$

$$\frac{L\Delta}{K} + \frac{L^2\Delta^2}{\lambda^2 K^2}. \tag{235}$$

The optimal λ in this case is $\lambda = 4\sqrt{L\Delta} - 2\sigma$, and we have that with probability at least $1-\beta$

$$\min_{t\in[0,K]}\left\|\nabla f(x^t)\right\|^2 = \mathcal{O}\left(\max\left\{(237),(238),(239)\right\}\right), \tag{236}$$

where

$$\sqrt{L\Delta(4\sqrt{L\Delta}-2\sigma)^{2-\alpha}\sigma^\alpha\frac{\ln K/\beta}{K}} + \frac{L\Delta\sigma^\alpha \ln K/\beta}{(4\sqrt{L\Delta}-2\sigma)^\alpha K}, \tag{237}$$

$$\sqrt{L\Delta}\sigma + \frac{L\Delta\sigma^2}{(4\sqrt{L\Delta}-2\sigma)^2}, \tag{238}$$

$$\frac{L\Delta}{K} + \frac{L^2\Delta^2}{(4\sqrt{L\Delta}-2\sigma)^2K^2}. \tag{239}$$

Case 5: $\lambda \leq \frac{4}{3}\sqrt{L\Delta}, \quad \lambda < \zeta_\lambda < \sigma$**.** In this case, the step-size conditions reduce to

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\sigma^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K\ln\frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}}\lambda^\alpha}{K(\sigma^\alpha\zeta_\lambda)}\right\}\right), \tag{240}$$

and $\min_{t\in[0,K]}\left\|\nabla f(x^t)\right\|^2$ is bounded with probability at least $1-\beta$ by the maximum of the following terms

$$\sqrt{L\Delta}\lambda^{1-\alpha/2}\sigma^{\alpha/2}\sqrt{\frac{\ln K/\beta}{K}} + \frac{L\Delta\sigma^\alpha \ln K/\beta}{\lambda^\alpha K}, \tag{241}$$

$$\sqrt{L\Delta}\frac{\sigma^\alpha\zeta_\lambda}{\lambda^\alpha} + \frac{L\Delta\sigma^{2\alpha}\zeta_\lambda^2}{\lambda^{2\alpha+2}}, \tag{242}$$

$$\frac{L\Delta}{K} + \frac{L^2\Delta^2}{\lambda^2 K^2}. \tag{243}$$

In this regime, the optimal $\lambda = \frac{4}{3}\sqrt{L\Delta}$. With this choice of λ, we get with probability at least $1-\beta$

$$\min_{t\in[0,K]}\left\|\nabla f(x^t)\right\|^2 = \mathcal{O}\left(\max\left\{(245),(246),(247)\right\}\right), \tag{244}$$

where

$$\sqrt{(L\Delta)^{\frac{4-\alpha}{2}}\sigma^\alpha\frac{\ln K/\beta}{K}} + \frac{(L\Delta)^{\frac{2-\alpha}{2}}\sigma^\alpha \ln K/\beta}{K}, \tag{245}$$

$$\frac{\sigma^\alpha}{(\sqrt{L\Delta})^{\alpha-2}} + \frac{\sigma^{2\alpha}}{(L\Delta)^{\alpha-1}}, \tag{246}$$

$$\frac{L\Delta}{K} + \frac{L\Delta}{K^2}. \tag{247}$$

Case 6: $\lambda \leq \frac{4}{3}\sqrt{L\Delta}, \quad \lambda < \sigma < \zeta_\lambda$**.** In this case, the step-size conditions reduce to

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\zeta_\lambda^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K\ln\frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}}\lambda^\alpha}{K(\zeta_\lambda^{\alpha+1})}\right\}\right), \tag{248}$$

and $\min_{t\in[0,K]} \left\|\nabla f(x^t)\right\|^2$ is bounded with probability at least $1-\beta$ by the maximum of the following terms

$$\sqrt{L\Delta}\lambda^{1-\alpha/2}\zeta_\lambda^{\alpha/2}\sqrt{\frac{\ln K/\beta}{K}} + \frac{L\Delta\zeta_\lambda^\alpha \ln K/\beta}{\lambda^\alpha K}, \tag{249}$$

$$\frac{\sqrt{L\Delta}\zeta_\lambda^{\alpha+1}}{\lambda^\alpha} + \frac{L\Delta\zeta_\lambda^{2\alpha}}{\lambda^{2\alpha+2}}, \tag{250}$$

$$\frac{L\Delta}{K} + \frac{L^2\Delta^2}{\lambda^2K^2}. \tag{251}$$

Next, we find the optimal λ via equalizing the leading terms (the first ones) in (249) and (250). This yields $\lambda = \frac{4\sqrt{L\Delta}}{2C+1}$, where $C = \left(\frac{\ln\frac{K}{\beta}}{K}\right)^{\frac{1}{\alpha+2}}$, which is infeasible. Thus, the optimal λ in this regime is $\lambda = \frac{4}{3}\sqrt{L\Delta} - \eta$, where $\eta \geq 0$ is such that $\lambda < \sigma < \zeta_\lambda$. Given this choice of λ, we obtain with probability at least $1-\beta$

$$\min_{t\in[0,K]} \left\|\nabla f(x^t)\right\|^2 = \mathcal{O}\left(\max\left\{(253), (254), (255)\right\}\right), \tag{252}$$

where

$$(\sqrt{L\Delta}-\eta)^{1-\alpha/2}(\sqrt{L\Delta}+\eta)^{\alpha/2}\sqrt{L\Delta\frac{\ln K/\beta}{K}} + \frac{L\Delta(\sqrt{L\Delta}+\eta)^\alpha \ln K/\beta}{(\sqrt{L\Delta}-\eta)^\alpha K}, \tag{253}$$

$$\frac{\sqrt{L\Delta}(\sqrt{L\Delta}+\eta)^{\alpha+1}}{(\sqrt{L\Delta}-\eta)^\alpha} + \frac{L\Delta(\sqrt{L\Delta}+\eta)^{2\alpha}}{(\sqrt{L\Delta}-\eta)^{2\alpha+2}}, \tag{254}$$

$$\frac{L\Delta}{K} + \frac{L^2\Delta^2}{(\sqrt{L\Delta}-\eta)^2K^2}. \tag{255}$$

Case 7: $\lambda \leq \frac{4}{3}\sqrt{L\Delta}, \quad \sigma < \lambda < \zeta_\lambda$**.** In this case, the step-size conditions reduce to

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\zeta_\lambda^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K\ln\frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}}\lambda^{\alpha-1}}{K\max\left\{\frac{\zeta_\lambda^{\alpha+1}}{\lambda}, \zeta_\lambda^{\alpha-1}\sigma\right\}}\right\}\right). \tag{256}$$

We note that $\max\left\{\frac{\zeta_\lambda^{\alpha+1}}{\lambda}, \zeta_\lambda^{\alpha-1}\sigma\right\} = \zeta^\alpha \max\left\{\frac{\zeta_\lambda}{\lambda}, \frac{\sigma}{\lambda}\right\} = \frac{\zeta_\lambda^{\alpha+1}}{\lambda}$ since $\sigma < \lambda < \zeta_\lambda$. Therefore, similarly to the previous case, we have

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\zeta_\lambda^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K\ln\frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}}\lambda^\alpha}{K\zeta_\lambda^{\alpha+1}}\right\}\right), \tag{257}$$

and $\min_{t\in[0,K]} \left\|\nabla f(x^t)\right\|^2$ is bounded with probability at least $1-\beta$ by the maximum of the following terms

$$\sqrt{L\Delta}\lambda^{1-\alpha/2}\zeta_\lambda^{\alpha/2}\sqrt{\frac{\ln K/\beta}{K}} + \frac{L\Delta\zeta_\lambda^\alpha \ln K/\beta}{\lambda^\alpha K}, \tag{258}$$

$$\frac{\sqrt{L\Delta}\zeta_\lambda^{\alpha+1}}{\lambda^\alpha} + \frac{L\Delta\zeta_\lambda^{2\alpha}}{\lambda^{2\alpha+2}}, \tag{259}$$

$$\frac{L\Delta}{K} + \frac{L^2\Delta^2}{\lambda^2K^2}. \tag{260}$$

The optimal λ equals $\frac{4}{3}\sqrt{L\Delta}$. This happens because both leading terms in (258) and (259) are decreasing in λ. With this choice, we get with probability at least $1-\beta$

$$\min_{t\in[0,K]} \left\|\nabla f(x^t)\right\|^2 = \mathcal{O}\left(\max\left\{(262), (263), (264)\right\}\right), \tag{261}$$

where

$$\sqrt{L\Delta\frac{\ln {}^{K}\!/\!_{\beta}}{K}} + \frac{L\Delta \ln {}^{K}\!/\!_{\beta}}{K}, \tag{262}$$

$$\sqrt{L\Delta}\sigma + \frac{\sigma^2}{L\Delta}, \tag{263}$$

$$\frac{L\Delta}{K} + \frac{L\Delta}{K^2}. \tag{264}$$

Now that we have covered all possible regions, it's time to consider the DP noise as well.

H Rate and Neighborhood for DP-Clipped-SGD: Non-Convex Case

To ensure the output of the algorithm is (ε, δ)-differentially private in this setting, expectation minimization, it suffices to set the noise scale as $\sigma_\omega = \Theta\left(\frac{\lambda}{\varepsilon}\sqrt{K \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{1}{\delta}\right)}\right)$ and apply the advanced composition theorem of Dwork et al. (2014). In the finite sum case, one can reduce the amount of noise by a factor of $\sqrt{\ln\left(\frac{K}{\delta}\right)}$ as it was shown by Abadi et al. (2016). For the sake of brevity, in the DP case, we only consider two cases: large λ and relatively small λ regimes. The other cases can be derived with a similar analysis.

Case 1: $\lambda > 4\sqrt{L\Delta}$. In this case, $\zeta_\lambda = 0$, and the step-size conditions reduce to the following:

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\sigma^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K \ln \frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}}\lambda^{\alpha-1}}{K\sigma^\alpha}, \frac{\sqrt{\frac{\Delta}{L}}}{\sigma_\omega\sqrt{dK \ln \frac{K}{\beta}}}\right\}\right) \tag{265}$$

In particular, when γ equals the minimum from the step-size condition, then the iterates produced by DP-Clipped-SGD after K iterations with probability at least $1-\beta$ satisfy

$$\min_{k\in[0,K]} \left\|\nabla f(x^t)\right\|^2 = \mathcal{O}\left(\max\left\{(267), (268), (269), (270)\right\}\right) \tag{266}$$

where

$$\sqrt{L\Delta}\lambda^{1-\alpha/2}\sigma^{\alpha/2}\sqrt{\frac{\ln K/\beta}{K}} + \frac{L\Delta\sigma^\alpha \ln K/\beta}{\lambda^\alpha K} \tag{267}$$

$$\frac{\sqrt{L\Delta}\sigma^\alpha}{\lambda^{\alpha-1}} + \frac{L\Delta\sigma^{2\alpha}}{\lambda^{2\alpha}} \tag{268}$$

$$\frac{L\Delta}{K} + \frac{L^2\Delta^2}{\lambda^2K^2} \tag{269}$$

$$\sqrt{L\Delta}\sigma_\omega\sqrt{\frac{d\ln\frac{K}{\beta}}{K}} + \frac{L\Delta\sigma_\omega^2 d\ln\frac{K}{\beta}}{\lambda^2 K}. \tag{270}$$

Here, (268) accounts for the bias caused by clipping, and (270) accounts for the accumulation of DP noise. These terms are decreasing and increasing in λ respectively, if we use $\sigma_\omega = \Theta\left(\frac{\lambda}{\varepsilon}\sqrt{K \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{1}{\delta}\right)}\right)$. To find the optimal λ, we find the equilibrium of these two terms. Solving the equilibrium equation, we get $\lambda = \mathcal{O}\left(\frac{\varepsilon\sigma^\alpha}{d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)}\right)^{\frac{1}{\alpha}}$. Unless $\varepsilon\sigma^\alpha$ is large enough, this value violates the constraint that $\lambda > 4\sqrt{L\Delta}$, and it is not feasible. Thus, we have the following formula for the optimal λ:

$$\lambda = \max\left\{4\sqrt{L\Delta}, \left(\frac{\varepsilon\sigma^\alpha}{d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)}\right)^{\frac{1}{\alpha}}\right\}. \tag{271}$$

For this choice of λ, we get that with probability at least $1-\beta$

$$\min_{k\in[0,K]} \left\|\nabla f(x^t)\right\|^2 = \mathcal{O}\left(\max\left\{(273), (274), (275), (276)\right\}\right) \tag{272}$$

with

$$\max\left\{\sqrt{(L\Delta)^{\frac{4-\alpha}{2}}\sigma^\alpha\frac{\ln K/\beta}{K}}, \sqrt{L\Delta}\left(\frac{\varepsilon\sigma^\alpha}{\sqrt{d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)}}\right)^{\frac{1}{\alpha}}\sqrt{\frac{\ln^{\frac{3\alpha-2}{2\alpha}}\frac{K}{\beta}}{K}}\right\} \tag{273}$$

$$\min\left\{\frac{\sigma^\alpha}{\left(\sqrt{L\Delta}\right)^{\alpha-2}}, \sqrt{L\Delta}\sigma\left(\frac{\sqrt{d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)}}{\varepsilon}\right)^{\frac{\alpha-1}{\alpha}}\right\} \tag{274}$$

$$\min\left\{\frac{L\Delta}{K^2}, \frac{L^2\Delta^2\left(d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\right)^{\frac{1}{\alpha}}}{(\varepsilon)^{\frac{1}{\alpha}}\sigma}\frac{\ln^{\frac{1}{\alpha}}\frac{K}{\beta}}{K^2}\right\} + \frac{L\Delta}{K} \tag{275}$$

$$\max\left\{\frac{L\Delta}{\varepsilon}\sqrt{d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)}, \frac{R\sigma\left(d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)\right)^{\frac{\alpha+2}{2\alpha}}}{\varepsilon^{\frac{\alpha-1}{\alpha}}}\right\}$$
$$+\frac{L\Delta}{\varepsilon^2}d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right), \tag{276}$$

where, for the sake of brevity, we only report the dominant terms.

Case 2: $\lambda \leq \frac{4}{3}\sqrt{L\Delta} \quad \lambda < \sigma < \zeta_\lambda$**.** In this case, the step-size conditions reduce to the following:

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\zeta_\lambda^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K\ln\frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}}\lambda^\alpha}{K(\zeta_\lambda^{\alpha+1})}, \frac{\sqrt{\frac{\Delta}{L}}}{\sigma_\omega\sqrt{dK\ln\frac{K}{\beta}}}\right\}\right). \tag{277}$$

Taking γ equal to the right-hand side, we get that with probability at least $1-\beta$

$$\min_{t\in[0,K]}\left\|\nabla f(x^t)\right\|^2 = \mathcal{O}\left(\{(279), (280), (281), (282)\}\right) \tag{278}$$

with

$$\sqrt{L\Delta}\lambda^{1-\alpha/2}\sigma^{\alpha/2}\sqrt{\frac{\ln K/\beta}{K}} + \frac{L\Delta\sigma^\alpha\ln K/\beta}{\lambda^\alpha K} \tag{279}$$

$$\frac{\sqrt{L\Delta}\zeta_\lambda^{\alpha+1}}{\lambda^\alpha} + \frac{L\Delta\zeta_\lambda^{2\alpha}}{\lambda^{2\alpha+2}} \tag{280}$$

$$\frac{L\Delta}{K} + \frac{L^2\Delta^2}{\lambda^2K^2} \tag{281}$$

$$\sqrt{L\Delta}\sigma_\omega\sqrt{\frac{d\ln\frac{K}{\beta}}{K}} + \frac{L\Delta\sigma_\omega^2 d\ln\frac{K}{\beta}}{\lambda^2K}. \tag{282}$$

Similarly to the previous case, we find the optimal λ as the equilibrium of the leading terms in (280) and (282). By doing so, we get the following optimal λ:

$$\lambda = \min\left\{\frac{4}{3}\sqrt{L\Delta}, \frac{2\varepsilon\sqrt{L\Delta}}{\left(d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)\right)^{\frac{1}{2\alpha+2}}+1}\right\} \tag{283}$$

For this choice of λ, we get that with probability at least $1-\beta$

$$\min_{k\in[0,K]}\left\|\nabla f(x^t)\right\|^2 = \mathcal{O}\left(\max\{(285), (286), (287), (288)\}\right) \tag{284}$$

with

$$\min\left\{\sqrt{(L\Delta)^{\frac{4-\alpha}{2}}\sigma^\alpha\frac{\ln K/\beta}{K}}, \sqrt{\frac{(L\Delta)^{\frac{4-\alpha}{2}}\varepsilon^{2-\alpha}\ln^{\frac{3\alpha}{4\alpha+4}}\frac{K}{\beta}}{\left(d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\right)^{\frac{2-\alpha}{4\alpha+4}}K}}\right\} \tag{285}$$

$$\max\left\{\frac{\sigma^\alpha}{\sqrt{L\Delta}^{\alpha-2}}, \frac{(\sqrt{L\Delta})^{2-\alpha}\sigma^\alpha}{\varepsilon}\left(d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)\right)^{\frac{\alpha-1}{2\alpha+2}}\right\} \tag{286}$$

$$\max\left\{\frac{L\Delta}{K^2}, \frac{L\Delta}{\varepsilon^2K^2}\left(\left(d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)\right)^{\frac{1}{2\alpha+2}}+1\right)^2\right\} + \frac{L\Delta}{K} \tag{287}$$

$$\min\left\{\frac{L\Delta}{\varepsilon}\sqrt{d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)}, \frac{L\Delta\sqrt{\ln\frac{K}{\beta}}}{\left(d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right)\right)^{\frac{1}{2\alpha+2}}+1}\right\}$$
$$+\frac{L\Delta d}{\varepsilon^2}d\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{K}{\delta}\right)\ln\left(\frac{K}{\beta}\right), \tag{288}$$

where, for the sake of brevity, we only report the dominant terms.

I Preliminary Numerical Experiments

We conclude with a simple synthetic experiment that illustrates the clipping-level trade-off predicted by our theory. We apply DP-Clipped-SGD to the one-dimensional Huber-loss problem:

$$\min_{x\in\mathbb{R}} f(x), \quad \text{where} \quad f(x) = \begin{cases} \frac{x^2}{2\tau}, & \text{for } |x| \le \tau, \\ |x| - \frac{\tau}{2}, & \text{for } |x| > \tau, \end{cases} \quad \tau > 0.$$

The Huber loss is convex and L-smooth with $L = 1/\tau$. We use the stochastic gradient model $\nabla f_\xi(x) = \nabla f(x) + \xi$, where $\xi = s \cdot \eta \cdot \nu$, $s > 0$ is a scaling factor, η is a Rademacher random variable, and ν is a Pareto random variable with scale parameter 1 and shape parameter $\hat{\alpha} \in (1, 2]$. This construction satisfies Assumption 2.4 for every $\alpha \in (1, \hat{\alpha}]$. Unless stated otherwise, we set $\sigma_\omega = 0.03\,\lambda\sqrt{K}$ and $\gamma = 0.05/\sqrt{K}$.

For each value of K, we run DP-Clipped-SGD 100 times and report the median together with the 5th and 95th percentiles. Figure 1 shows the final loss after $K \in \{1000, 5000, 10000\}$ steps as a function of the clipping level λ for several values of the Pareto shape parameter $\hat{\alpha}$.

The main qualitative pattern is consistent with our theory: the empirically best clipping level stays in a moderate range and does not increase with K. This is natural in the DP setting because the injected Gaussian noise grows linearly with λ, so overly large clipping levels eventually hurt utility. Moreover, for a fixed horizon K, the curves are relatively stable across the tested values of $\hat{\alpha}$, which is consistent with our bounds in the fixed-clipping regime.

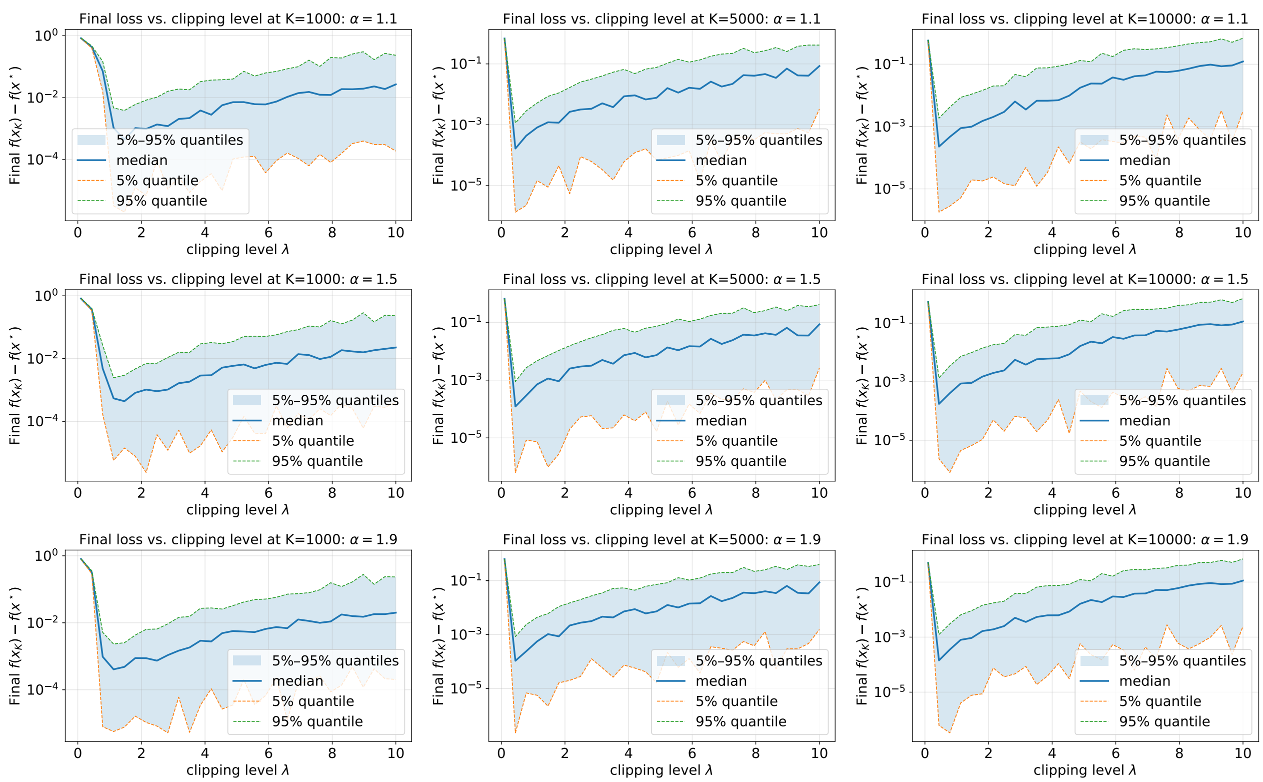

Figure 1: Final Huber loss achieved by DP-Clipped-SGD as a function of the clipping level λ for different iteration budgets K and Pareto shape parameters $\hat{\alpha}$. Solid lines show the median over 100 runs, and the shaded band corresponds to the 5th–95th percentile range.