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

Anonymous Author(s)

Affiliation

Address

email

Abstract

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

16 1 Introduction

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

25 In particular, gradient clipping is a key component in managing heavy-tailed noise, which commonly
26 arises in the training of language models on textual data ([Zhang et al., 2020](#)), in the training of
27 GANs ([Goodfellow et al., 2014](#); [Gorbunov et al., 2022](#)), and even in simpler tasks such as image
28 classification ([Şimşekli et al., 2019](#)). This approach is primarily analyzed through the lens of high-
29 probability convergence, as such guarantees provide a more accurate reflection of the actual behavior
30 of optimization methods compared to their more conventional in-expectation counterparts ([Gorbunov](#)
31 [et al., 2020](#)). Moreover, as demonstrated by [Sadiev et al. \(2023\)](#) for [SGD](#) and by [Chezhegov et al.](#)
32 [\(2024\)](#) for [AdaGrad](#) and [Adam](#), methods without clipping may fail to exhibit high-probability
33 convergence with logarithmic dependence on the failure probability. In contrast, several recent works
34 ([Gorbunov et al., 2020](#); [Cutkosky and Mehta, 2021](#); [Sadiev et al., 2023](#); [Nguyen et al., 2023](#); [Gorbunov](#)
35 [et al., 2024b](#); [Chezhegov et al., 2024](#); [Parletta et al., 2024](#)) have established that various stochastic

36 first-order methods attain significantly better high-probability convergence under heavy-tailed noise
37 assumptions across different settings.

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

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

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

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

57 However, this improvement is achieved by relaxing the requirement for exact convergence and instead
58 demonstrating convergence to a neighborhood whose size depends non-trivially on the clipping level,
59 noise scale, and other problem-dependent parameters. Importantly, the size of this neighborhood,
60 introduced due to the inherent bias in clipped stochastic gradients, can be carefully balanced with
61 the neighborhood induced by the DP noise, allowing for more flexible control over the trade-off
62 between convergence accuracy and privacy. Additionally, we extend our results to the non-convex
63 case, illustrating the broader applicability of our analysis.

64 2 Technical Preliminaries

65 The optimization problem considered in this work has the following form

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

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

69 We assume that at each iteration, we have access to an oracle that provides a stochastic gradi-
70 ent $\nabla f_\xi(x)$, as well as a d -dimensional random vector ω sampled from a Gaussian distribution
71 $\mathcal{N}(0, \sigma_\omega^2 \mathbf{I}_d)$, where \mathbf{I}_d is the $d \times d$ identity matrix. More precisely, the random variables ξ and ω are
72 defined on the probability space $(\Omega_d \times \mathbb{R}^d, \mathcal{B}(\Omega_d) \otimes \mathcal{B}(\mathbb{R}^d), \mathcal{F}^t, \mathbb{P})$, where Ω_d represents the data
73 sample space, and $\mathcal{B}(\mathcal{X})$ denotes the Borel σ -algebra generated by the set \mathcal{X} . This probability space
74 is also equipped with the natural filtration $\mathcal{F}^t = \sigma\left([\nabla f_{\xi^0}(x^0), \omega_0]^T, \dots, [\nabla f_{\xi^t}(x^t), \omega_t]^T\right)$, which
75 captures the history of the stochastic process up to time t . The probability measure \mathbb{P} is defined as the
76 product measure on this space, given by

$$\mathbb{P}\{B_d \times B_\omega\} = (\mu \times \nu)(B_d \times B_\omega) = \mu(B_d) \nu(B_\omega), \quad \forall B_d \in \mathcal{B}(\Omega_d), \forall B_\omega \in \mathcal{B}(\mathbb{R}^d), \quad (2)$$

77 where μ is a probability measure on Ω_d , and ν is the Gaussian measure on \mathbb{R}^d with mean zero and
78 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 $\mathbb{P}\{\mathcal{P}(x^K) \leq \epsilon\} \geq 1 - \beta$, where $\mathcal{P}(x)$ is a performance metric that measures the quality of the solution¹. 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 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., $f^* := \inf_{x \in Q} f(x) > -\infty$.

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.

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|. \quad (3)$$

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 $x, y \in Q$

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

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 $x \in Q$

$$\mathbb{E}_{\xi \sim D} [\nabla f_{\xi}(x) \mid x] = \nabla f(x), \quad (5)$$

$$\mathbb{E}_{\xi \sim D} [\|\nabla f_{\xi}(x) - \nabla f(x)\|^{\alpha} \mid x] \leq \sigma^{\alpha}. \quad (6)$$

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

$$\mathbb{P}(\mathcal{M}(D) \in S) \leq e^{\epsilon} \mathbb{P}(\mathcal{M}(D') \in S) + \delta, \quad (7)$$

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 (Su, 2024).

¹Examples 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 $x^* \in \arg \min_{x \in \mathbb{R}^d} f(x)$.

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 $\sigma_\omega = \Theta\left(\frac{q\lambda}{\epsilon}\sqrt{K \ln \frac{1}{\delta}}\right)$ 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), 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 work that avoids these assumptions is Islamov et al. (2025), where the authors proposed a distributed optimization method based on clipping, error feedback (Seide et al., 2014; Richtárik et al., 2021), and heavy-ball momentum (Polyak, 1964). However, their high-probability convergence analysis critically relies on the assumption that the noise in the stochastic gradients has sub-Gaussian tails. By 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 uses DP mean estimation with a sufficiently large number of samples. However, this approach can be prohibitively expensive in practice, particularly in the training of 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

²Li 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.

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 sub-optimal) 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.

The following two theorems present our newly derived step-size bounds and the corresponding performance guarantees for both convex and non-convex settings. Following each theorem, we provide a table that further simplify the performance bounds under the assumption that the clipping level falls within specific intervals. In these tables, we assume that no DP noise is present, focusing purely on the impact of the clipping bias. The final corollary extend these results to the case where

³In 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 .

```

1: for  $k = 0, \dots, K$  do
2:   Compute  $\hat{g}_k = \text{clip}(\nabla f_{\xi^k}(x^k), \lambda)$  using a fresh sample  $\xi^k \sim \mathcal{D}$ 
3:    $\omega_k \sim \mathcal{N}(0, \sigma_\omega^2 I_d)$ 
4:    $\tilde{g}_k = \hat{g}_k + \omega_k$ 
5:    $x^{k+1} = x^k - \gamma \tilde{g}_k$ 
6: end for

```

DP noise is included in the convex case, while the result for DP case in the non-convex setup is deferred to the supplementary materials due to space limitation.

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 $Q = B_{2R}(x^*)$, $R \geq \|x^0 - x^*\|$. Set $\zeta_\lambda := \max\{0, 2LR - \frac{\lambda}{2}\}$, and further assume that the step-size γ is selected to satisfy*

$$\gamma \leq \mathcal{O} \left(\min \left\{ \frac{1}{L}, \frac{R}{\lambda^{1-\alpha/2} \sqrt{K \ln \left(\frac{K}{\beta} \right) (\sigma^\alpha + \zeta_\lambda^\alpha)}}, \frac{R\lambda^{\alpha-1}}{K(\sigma^\alpha + \zeta_\lambda^\alpha) \left(\frac{LR}{\lambda} + \frac{\lambda^{\alpha-1}\zeta_\lambda}{\sigma^\alpha + \zeta_\lambda^\alpha} + (\sigma^\alpha + \zeta_\lambda^\alpha)^{\frac{-1}{\alpha}} \right)}, \frac{R}{\sigma_\omega \sqrt{dK \ln \left(\frac{K}{\beta} \right)}} \right\} \right). \quad (8)$$

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

$$\min_{t \in [0, K]} f(x^t) - f(x^*) \leq \frac{4R^2}{\gamma(K+1)} + \frac{64LR^4}{\lambda^2 \gamma^2 (K+1)^2}. \quad (9)$$

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 (8).

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 (8), 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 4.2 (Convergence of 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 (8) then with probability at least $1 - \beta$ the error converges to a neighborhood of the global optimum of size*

$$\min_{t \in [0, K]} f(x^t) - f(x^*) \leq \mathcal{O}(\max\{(11), (12), (13), (14)\}). \quad (10)$$

248 where

$$\frac{LR^2}{K} + \frac{L^3 R^4}{\lambda^2 K^2} \quad (11)$$

$$R\lambda^{1-\alpha/2}\sigma^{\alpha/2}\sqrt{\frac{\ln K/\beta}{K} + \frac{LR^2\sigma^\alpha \ln K/\beta}{K}} \quad (12)$$

$$\frac{R(\sigma^\alpha + \zeta_\lambda^\alpha) \left(\frac{LR}{\lambda} + \frac{\lambda^{\alpha-1}\zeta_\lambda}{\sigma^\alpha + \zeta_\lambda^\alpha} + (\sigma^\alpha + \zeta_\lambda^\alpha)^{-\frac{1}{\alpha}} \right)}{\lambda^{\alpha-1}} + \frac{R^2 L(\sigma^\alpha + \zeta_\lambda^\alpha)^2 \left(\frac{LR}{\lambda} + \frac{\lambda^{\alpha-1}\zeta_\lambda}{\sigma^\alpha + \zeta_\lambda^\alpha} + (\sigma^\alpha + \zeta_\lambda^\alpha)^{-\frac{1}{\alpha}} \right)^2}{\lambda^{2\alpha}} \quad (13)$$

$$\frac{R\lambda}{\varepsilon} \sqrt{d \ln \left(\frac{K}{\beta} \right) \ln \left(\frac{K}{\delta} \right) \ln \left(\frac{1}{\delta} \right)} + \frac{LR^2 d \ln \left(\frac{K}{\beta} \right) \ln \left(\frac{K}{\delta} \right) \ln \left(\frac{1}{\delta} \right)}{\varepsilon^2}. \quad (14)$$

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

252 **Corollary 4.3** (Convergence of **DP-Clipped-SGD** for the convex objective). *Let the assump-*
 253 *tions of Theorem 4.1 hold, K is sufficiently large, γ is chosen as the minimum of (8), $\sigma_\omega =$
 254 $\Theta \left(\frac{\lambda}{\varepsilon} \sqrt{K \ln \left(\frac{K}{\delta} \right) \ln \left(\frac{1}{\delta} \right)} \right)$, and $\lambda > 4LR$. Then the optimal value for λ is*

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

255 With this value, the iterates produced by the algorithm with probability of at least $1 - \beta$ satisfy

$$\min_{k \in [0, K]} f(x^t) - f(x^*) = \mathcal{O}(\max \{ (15), (16), (17), (18) \}),$$

256 where

$$\max \left\{ \sqrt{\frac{R^{4-\alpha} L^{2-\alpha} \sigma^\alpha \ln \left(\frac{K}{\beta} \right)}{K}}, R \left(\frac{\varepsilon \sigma^\alpha}{\sqrt{d \ln \left(\frac{K}{\delta} \right) \ln \left(\frac{1}{\delta} \right)}} \right)^{\frac{1}{\alpha}} \sqrt{\frac{\ln^{\frac{3\alpha-2}{2\alpha}} \left(\frac{K}{\beta} \right)}{K}} \right\} \quad (15)$$

$$\min \left\{ \frac{R^{2-\alpha} \sigma^\alpha}{L^{\alpha-1}}, R \sigma \left(\frac{\sqrt{d \ln \left(\frac{K}{\delta} \right) \ln \left(\frac{1}{\delta} \right)}}{\varepsilon} \right)^{\frac{\alpha-1}{\alpha}} \right\} \quad (16)$$

$$\min \left\{ \frac{LR^2}{K^2}, \frac{L^3 R^4 \left(d \ln \left(\frac{K}{\delta} \right) \ln \left(\frac{1}{\delta} \right) \ln \left(\frac{K}{\beta} \right) \right)^{\frac{1}{\alpha}}}{(\varepsilon)^{\frac{1}{\alpha}} \sigma K^2} \right\} + \frac{LR^2}{K} \quad (17)$$

$$\max \left\{ \frac{LR^2}{\varepsilon} \sqrt{d \ln \left(\frac{K}{\delta} \right) \ln \left(\frac{1}{\delta} \right) \ln \left(\frac{K}{\beta} \right)}, \frac{R \sigma \left(d \ln \left(\frac{K}{\delta} \right) \ln \left(\frac{1}{\delta} \right) \ln \left(\frac{K}{\beta} \right) \right)^{\frac{\alpha+2}{2\alpha}}}{\varepsilon^{\frac{\alpha-1}{\alpha}}} \right\} \\ + \frac{LR^2 d}{\varepsilon^2} \ln \left(\frac{K}{\delta} \right) \ln \left(\frac{1}{\delta} \right) \ln \left(\frac{K}{\beta} \right). \quad (18)$$

257 Also, for small λ regime ($\lambda \leq \frac{4}{3}LR$), the optimal value for λ is

$$\lambda = \min \left\{ \frac{4}{3}LR, \frac{2\varepsilon LR}{\left(d \ln \left(\frac{K}{\delta} \right) \ln \left(\frac{1}{\delta} \right) \ln \frac{K}{\beta} \right)^{\frac{1}{2\alpha+2}} + 1} \right\}. \quad (19)$$

258 With this value, the iterates produced by the algorithm with probability of at least $1 - \beta$ satisfy

$$\min_{t \in [0, K]} f(x^t) - f(x^*) = \mathcal{O}(\max \{ (20), (21), (22), (23) \}),$$

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 (9) and the minimized RHS value itself, respectively.

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

259 where

$$\min \left\{ \sqrt{\frac{R^{4-\alpha}L^{2-\alpha}\sigma^\alpha \ln\left(\frac{K}{\beta}\right)}{K}}, \sqrt{\frac{R^{4-\alpha}(\varepsilon L)^{2-\alpha} \ln^{\frac{3\alpha}{4\alpha+4}}\left(\frac{K}{\beta}\right)}{(d \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{1}{\delta}\right))^{\frac{2-\alpha}{4\alpha+4}} K}} \right\} \quad (20)$$

$$\max \left\{ \frac{R^{2-\alpha}\sigma^\alpha}{L^{\alpha-1}}, \frac{R^{2-\alpha}\sigma^\alpha}{\varepsilon} \left(d \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{K}{\beta}\right) \right)^{\frac{\alpha-1}{2\alpha+2}} \right\} \quad (21)$$

$$\max \left\{ \frac{LR^2}{K^2}, \frac{LR^2}{\varepsilon^2 K^2} \left(d \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{K}{\beta}\right) \right)^{\frac{2}{2\alpha+2}} \right\} + \frac{LR^2}{K} \quad (22)$$

$$\min \left\{ \frac{LR^2}{\varepsilon} \sqrt{d \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{K}{\beta}\right)}, \frac{LR^2}{\left(d \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{K}{\beta}\right) \right)^{\frac{1}{2\alpha+2}}} \right\} \\ + \frac{LR^2 d}{\varepsilon^2} \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{K}{\beta}\right). \quad (23)$$

260 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
261 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
262 is applied to each stochastic gradient, and $\varepsilon = \mathcal{O}(q^2 K)$, allowing to have smaller ε and δ for given
263 σ_ω and λ . We note that our analysis holds for the finite-sum case without changes as long as the
264 assumptions of the theorem are satisfied and the mini-batch size equals 1.

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

266 **Theorem 4.4** (Convergence of **DP-Clipped-SGD** for the non-convex objective). *Let the integer*
267 *$K \geq 0$ and $\beta \in (0, 1]$ be given. Let the assumptions 2.1, 2.2, 2.4, hold for the set Q defined*
268 *as $Q = \{x \in \mathbb{R} \mid \exists y \in \mathbb{R}^d : f(y) \leq f^* + 2\Delta \text{ and } \|x - y\| \leq \sqrt{\Delta}/20\sqrt{L}\}$, where $\Delta \geq f(x^0) - f^*$,*

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 (25) and the minimized RHS value itself, respectively.

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

269 $\zeta_\lambda := \max\left\{0, 2\sqrt{L\Delta} - \frac{\lambda}{2}\right\}$, and γ is selected according to

$$\begin{aligned}
 \gamma \leq \mathcal{O} \left(\min \left\{ \frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\lambda^{1-\alpha/2} \sqrt{K \ln \left(\frac{K}{\beta} \right) (\sigma^\alpha + \zeta_\lambda^\alpha)}} \right. \right. \\
 \left. \left. \frac{\sqrt{\frac{\Delta}{L}} \lambda^{\alpha-1}}{K(\sigma^\alpha + \zeta_\lambda^\alpha) \left(\frac{\sqrt{L\Delta}}{\lambda} + \frac{\lambda^{\alpha-1} \zeta_\lambda}{\sigma^\alpha + \zeta_\lambda^\alpha} + (\sigma^\alpha + \zeta_\lambda^\alpha)^{\frac{-1}{\alpha}} \right)}, \frac{\sqrt{\frac{\Delta}{L}}}{\sigma_\omega \sqrt{dK \ln \left(\frac{K}{\beta} \right)}} \right\} \right). \quad (24)
 \end{aligned}$$

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

$$\min_{t \in [0, K]} \|\nabla f(x^t)\|^2 \leq \frac{8\Delta}{\gamma(K+1)} + \frac{128\Delta^2}{\lambda^2 \gamma^2 (K+1)^2} \quad (25)$$

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

274 *Proof sketch.* The proof of Theorems 4.1 and 4.4 is heavily inspired by (Sadiev et al., 2023). Yet,
 275 there is a crucial difference in defining the clipping level parameter. In contrast to (Sadiev et al.,
 276 2023), we treat λ as given rather than calculating it based on other problem parameters. By doing so,
 277 the fundamental assumption regarding the magnitude of λ in comparison to the norm of the gradient
 278 in bias-variance of the clipped vector (Lemma 5.1) of (Sadiev et al., 2023) becomes invalid. Thus, we
 279 develop a general bias-variance lemma (Lemma B.1) to study the statistical properties of the clipped
 280 vector.

281 5 Conclusion

282 In this paper, we present the first high-probability convergence analysis of **DP-Clipped-SGD** for
 283 both convex and non-convex smooth optimization problems under heavy-tailed noise. Our results
 284 demonstrate that **DP-Clipped-SGD** converges to a certain neighborhood of the optimal solution
 285 at a rate of $\mathcal{O}(1/\sqrt{K})$. In future work, it would be valuable to extend these results to the Federated
 286 Learning setting and to investigate the tightness and optimality of the derived 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. (Cited on pages 2, 4, 8, 37, and 53)
- 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. (Cited on page 4)
- 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. (Cited on page 4)
- 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*. (Cited on page 5)
- 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*. (Cited on page 5)
- 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*. (Cited on pages 1, 3, 4, and 5)
- 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. (Cited on pages 1, 4, and 5)
- 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. (Cited on page 5)
- Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7). (Cited on page 1)
- 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. (Cited on page 4)
- 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. (Cited on pages 2, 3, 4, 6, 37, and 53)
- Dzhaparidze, K. and Van Zanten, J. (2001). On bernstein-type inequalities for martingales. *Stochastic processes and their applications*, 93(1):109–117. (Cited on page 20)
- Freedman, D. A. (1975). On tail probabilities for martingales. *the Annals of Probability*, pages 100–118. (Cited on page 20)
- Gasnikov, A. and Nesterov, Y. (2016). Universal fast gradient method for stochastic composite optimization problems. *arXiv preprint arXiv:1604.05275*. (Cited on page 4)
- 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. (Cited on page 4)
- 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. (Cited on page 1)
- 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. (Cited on pages 1 and 4)
- 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. (Cited on pages 1, 4, and 5)
- 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. (Cited on page 4)

340 Gorbunov, E., Sadiev, A., Danilova, M., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A.,
341 and Richtárik, P. (2024b). High-probability convergence for composite and distributed stochastic
342 minimization and variational inequalities with heavy-tailed noise. In Salakhutdinov, R., Kolter, Z.,
343 Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st*
344 *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning*
345 *Research*, pages 15951–16070. PMLR. (Cited on pages 1 and 5)

346 Harvey, N. J., Liaw, C., and Randhawa, S. (2019). Simple and optimal high-probability bounds for
347 strongly-convex stochastic gradient descent. *arXiv preprint arXiv:1909.00843*. (Cited on page 4)

348 Hübler, F., Fatkhullin, I., and He, N. (2024). From gradient clipping to normalization for heavy tailed
349 sgd. *arXiv preprint arXiv:2410.13849*. (Cited on page 5)

350 Islamov, R., Horvath, S., Lucchi, A., Richtarik, P., and Gorbunov, E. (2025). Double momentum and
351 error feedback for clipping with fast rates and differential privacy. *arXiv preprint arXiv:2502.11682*.
352 (Cited on page 4)

353 Juditsky, A. and Nemirovski, A. S. (2008). Large deviations of vector-valued martingales in 2-smooth
354 normed spaces. *arXiv preprint arXiv:0809.0813*. (Cited on page 20)

355 Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint*
356 *arXiv:1412.6980*. (Cited on page 1)

357 Koloskova, A., Hendrikx, H., and Stich, S. U. (2023). Revisiting gradient clipping: Stochastic
358 bias and tight convergence guarantees. In *International Conference on Machine Learning*, pages
359 17343–17363. PMLR. (Cited on pages 30 and 46)

360 Kornilov, N., Shamir, O., Lobanov, A., Dvinskikh, D., Gasnikov, A., Shibaev, I., Gorbunov, E.,
361 and Horváth, S. (2023). Accelerated zeroth-order method for non-smooth stochastic convex
362 optimization problem with infinite variance. *Advances in Neural Information Processing Systems*,
363 36:64083–64102. (Cited on page 5)

364 Kornilov, N., Zmushko, P., Semenov, A., Gasnikov, A., and Beznosikov, A. (2025). Sign operator for
365 coping with heavy-tailed noise: High probability convergence bounds with extensions to distributed
366 optimization and comparison oracle. *arXiv preprint arXiv:2502.07923*. (Cited on page 5)

367 Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection.
368 *Annals of statistics*, pages 1302–1338. (Cited on page 21)

369 Li, B. and Chi, Y. (2025). Convergence and privacy of decentralized nonconvex optimization with
370 gradient clipping and communication compression. *IEEE Journal of Selected Topics in Signal*
371 *Processing*. (Cited on page 4)

372 Li, S. and Liu, Y. (2023). High probability analysis for non-convex stochastic optimization with
373 clipping. In *ECAI 2023*, pages 1406–1413. IOS Press. (Cited on page 5)

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

376 Liu, M., Zhuang, Z., Lei, Y., and Liao, C. (2022). A communication-efficient distributed gradient
377 clipping algorithm for training deep neural networks. *Advances in Neural Information Processing*
378 *Systems*, 35:26204–26217. (Cited on page 4)

379 Liu, Z., Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. (2023). High probability convergence
380 of stochastic gradient methods. In *International Conference on Machine Learning*, pages 21884–
381 21914. PMLR. (Cited on page 4)

382 Madden, L., Dall’Anese, E., and Becker, S. (2024). High probability convergence bounds for non-
383 convex stochastic gradient descent with sub-weibull noise. *Journal of Machine Learning Research*,
384 25(241):1–36. (Cited on page 4)

385 Nazin, A. V., Nemirovsky, A. S., Tsybakov, A. B., and Juditsky, A. B. (2019). Algorithms of
386 robust stochastic optimization based on mirror descent method. *Automation and Remote Control*,
387 80:1607–1627. (Cited on page 4)

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

391 Nemirovskij, A. S. and Yudin, D. B. (1983). Problem complexity and method efficiency in optimiza-
392 tion. (Cited on page 4)

393 Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. (2023). Improved convergence in high
394 probability of clipped gradient methods with heavy tailed noise. (Cited on pages 1 and 5)

395 Noble, M., Bellet, A., and Dieuleveut, A. (2022). Differentially private federated learning on
396 heterogeneous data. In *International conference on artificial intelligence and statistics*, pages
397 10110–10145. PMLR. (Cited on page 4)

398 Parletta, D. A., Paudice, A., Pontil, M., and Salzo, S. (2024). High probability bounds for stochastic
399 subgradient schemes with heavy tailed noise. *SIAM Journal on Mathematics of Data Science*,
400 6(4):953–977. (Cited on pages 1 and 4)

401 Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural
402 networks. In *International conference on machine learning*, pages 1310–1318. Pmlr. (Cited on
403 page 1)

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

406 Polyanskiy, Y. and Wu, Y. (2025). *Information theory: From coding to learning*. Cambridge university
407 press. (Cited on page 21)

408 Puchkin, N., Gorbunov, E., Kutuzov, N., and Gasnikov, A. (2024). Breaking the heavy-tailed noise
409 barrier in stochastic optimization problems. In *International Conference on Artificial Intelligence*
410 *and Statistics*, pages 856–864. PMLR. (Cited on page 5)

411 Richtárik, P., Sokolov, I., and Fatkhullin, I. (2021). EF21: A new, simpler, theoretically better, and
412 practically faster error feedback. *Advances in Neural Information Processing Systems*, 34:4384–
413 4396. (Cited on page 4)

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

416 Sadiev, A., Danilova, M., Gorbunov, E., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A.,
417 and Richtárik, P. (2023). High-probability bounds for stochastic optimization and variational
418 inequalities: the case of unbounded variance. In *International Conference on Machine Learning*,
419 pages 29563–29648. PMLR. (Cited on pages 1, 3, 4, 5, 6, 9, 22, 32, and 48)

420 Seide, F., Fu, H., Droppo, J., Li, G., and Yu, D. (2014). 1-bit stochastic gradient descent and its
421 application to data-parallel distributed training of speech dnns. In *Interspeech*, volume 2014, pages
422 1058–1062. Singapore. (Cited on page 4)

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

425 Şimşekli, U., Gürbüzbalaban, M., Nguyen, T. H., Richard, G., and Sagun, L. (2019). On the
426 heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint*
427 *arXiv:1912.00018*. (Cited on page 1)

428 Streeter, M. and McMahan, H. B. (2010). Less regret via online conditioning. *arXiv preprint*
429 *arXiv:1002.4862*. (Cited on page 1)

430 Su, W. J. (2024). A statistical viewpoint on differential privacy: Hypothesis testing, representation,
431 and blackwell’s theorem. *Annual Review of Statistics and Its Application*, 12. (Cited on page 3)

432 Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., Jin, S., Quek, T. Q., and Poor, H. V.
433 (2020). Federated learning with differential privacy: Algorithms and performance analysis. *IEEE*
434 *transactions on information forensics and security*, 15:3454–3469. (Cited on page 4)

435 Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S., Kumar, S., and Sra, S. (2020). Why are
436 adaptive methods good for attention models? *Advances in Neural Information Processing Systems*,
437 33:15383–15393. (Cited on page 1)

438 Zhang, X., Chen, X., Hong, M., Wu, Z. S., and Yi, J. (2022). Understanding clipping for federated
439 learning: Convergence and client-level differential privacy. In *International Conference on Machine*
440 *Learning, ICML 2022*. (Cited on page 4)

441 Zhao, P., Wu, J., Liu, Z., Wang, C., Fan, R., and Li, Q. (2025). Differential private stochastic
442 optimization with heavy-tailed data: towards optimal rates. In *Proceedings of the AAAI Conference*
443 *on Artificial Intelligence*, volume 39, pages 22795–22803. (Cited on page 4)

444 Zhivotovskiy, N. (2024). Dimension-free bounds for sums of independent matrices and simple tensors
445 via the variational principle. *Electronic Journal of Probability*, 29:1–28. (Cited on page 21)

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: As mentioned in the abstract, this work provides the first high-probability analysis for Clipped SGD with heavy-tailed noise on the gradient and an arbitrary clipping level with added DP noise. This is the main contribution of the paper and it appears in the abstract.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We have explained the limitations of our analysis in Section 4.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Main assumptions are stated in Section 2. Complete correct proofs are provided in the appendices. A proof sketch is provided in the main text in Section 4.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: Only rigorous mathematical analysis is provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: Only rigorous mathematical analysis is provided.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: Only rigorous mathematical analysis is provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: Only rigorous mathematical analysis is provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer “Yes” if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: Only rigorous mathematical analysis is provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The work completely conforms to the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The work investigates the incorporation of differential privacy guarantees in stochastic optimization. Hence, it offers a positive societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No component with potential detrimental effects is released.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: There is no code, data, or models that require licenses.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

707 • If this information is not available online, the authors are encouraged to reach out to
708 the asset’s creators.

709 **13. New assets**

710 Question: Are new assets introduced in the paper well documented and is the documentation
711 provided alongside the assets?

712 Answer: [NA]

713 Justification: No new asset is released.

714 Guidelines:

- 715 • The answer NA means that the paper does not release new assets.
- 716 • Researchers should communicate the details of the dataset/code/model as part of their
717 submissions via structured templates. This includes details about training, license,
718 limitations, etc.
- 719 • The paper should discuss whether and how consent was obtained from people whose
720 asset is used.
- 721 • At submission time, remember to anonymize your assets (if applicable). You can either
722 create an anonymized URL or include an anonymized zip file.

723 **14. Crowdsourcing and research with human subjects**

724 Question: For crowdsourcing experiments and research with human subjects, does the paper
725 include the full text of instructions given to participants and screenshots, if applicable, as
726 well as details about compensation (if any)?

727 Answer: [NA]

728 Justification: There was no crowd-sourcing experiments or research with human subjects.

729 Guidelines:

- 730 • The answer NA means that the paper does not involve crowdsourcing nor research with
731 human subjects.
- 732 • Including this information in the supplemental material is fine, but if the main contribu-
733 tion of the paper involves human subjects, then as much detail as possible should be
734 included in the main paper.
- 735 • According to the NeurIPS Code of Ethics, workers involved in data collection, curation,
736 or other labor should be paid at least the minimum wage in the country of the data
737 collector.

738 **15. Institutional review board (IRB) approvals or equivalent for research with human
739 subjects**

740 Question: Does the paper describe potential risks incurred by study participants, whether
741 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)
742 approvals (or an equivalent approval/review based on the requirements of your country or
743 institution) were obtained?

744 Answer: [NA]

745 Justification: There was no crowd-sourcing experiments or research with human subjects.

746 Guidelines:

- 747 • The answer NA means that the paper does not involve crowdsourcing nor research with
748 human subjects.
- 749 • Depending on the country in which research is conducted, IRB approval (or equivalent)
750 may be required for any human subjects research. If you obtained IRB approval, you
751 should clearly state this in the paper.
- 752 • We recognize that the procedures for this may vary significantly between institutions
753 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the
754 guidelines for their institution.
- 755 • For initial submissions, do not include any information that would break anonymity (if
756 applicable), such as the institution conducting the review.

757 **16. Declaration of LLM usage**

758 Question: Does the paper describe the usage of LLMs if it is an important, original, or
759 non-standard component of the core methods in this research? Note that if the LLM is used
760 only for writing, editing, or formatting purposes and does not impact the core methodology,
761 scientific rigorousness, or originality of the research, declaration is not required.

762 Answer: [NA]

763 Justification: The paper does not use LLMs as an important, original, or non-standard
764 component of the core methods in this research.

765 Guidelines:

- 766 • The answer NA means that the core method development in this research does not
767 involve LLMs as any important, original, or non-standard components.
- 768 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
769 for what should or should not be described.

770 A Notation Table and Auxiliary Facts

771 To facilitate the readability of the proofs, we provide a notation table below⁴.

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 $\ x^0 - x^*\ $ for convex functions
Δ	Upper bound on $f(x^0) - f^*$ for non-convex functions

772

773 **Auxiliary facts.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A sequence $\{\mathcal{F}_i\}_{i \geq 1}$ of nested sigma algebras
 774 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
 775 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
 776 X_i is \mathcal{F}_i -measurable. Furthermore, if $\mathbb{E}[X_i | \mathcal{F}_{i-1}] = X_{i-1} \forall i$, then $\{X_i\}_{i \geq 1}$ is called a martingale.
 777 On the other hand, if $\mathbb{E}[X_i | \mathcal{F}_{i-1}] = 0 \forall i$, then $\{X_i\}_{i \geq 1}$ is called a martingale difference sequence.

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

781 **Lemma A.1.** Let the sequence of random variables $\{X_i\}_{i \geq 1}$ form a martingale difference sequence
 782 on the filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_i\}_{i \geq 1}, \mathbb{P})$. Assume that conditional variances $\sigma_i^2 :=$
 783 $\mathbb{E}[X_i^2 | \mathcal{F}_{i-1}]$ exist and are bounded. Furthermore, there exists a deterministic constant $c \geq 0$ such
 784 that $|X_i| \leq c$ almost surely for all $i \geq 0$. Then for all $b > 0$, $G > 0$ and $n \geq 1$

$$\mathbb{P} \left\{ \left| \sum_{i=1}^n X_i \right| > b \text{ and } \sum_{i=1}^n \sigma_i^2 \leq G \right\} \leq 2 \exp \left(-\frac{b^2}{2G + 2bc/3} \right). \quad (26)$$

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

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

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

$$\mathbb{E} \left[\exp \left(\frac{\|\xi_k\|^2}{\sigma_k^2} \right) \middle| \mathcal{F}_{k-1} \right] \leq \exp(1) \text{ almost surely, } k = 1, \dots, N \quad (27)$$

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

$$\mathbb{P} \left\{ \|S_N\|_2 \geq (\sqrt{2} + \sqrt{2}\phi) \sqrt{\sum_{k=1}^N \sigma_k^2} \right\} \leq \exp \left(-\frac{\phi^2}{3} \right). \quad (28)$$

⁴We fixed minor typos in Table 2 from the main part of the paper. Changes are highlighted using red color.

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

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

791 Let

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

792 Then the following inequalities hold for any positive t :

$$\mathbb{P} \left\{ X \geq 2\|a\|_2 \sqrt{t} + 2\|a\|_\infty t \right\} \leq \exp(-t), \quad (29)$$

$$\mathbb{P} \left\{ X \leq -2\|a\|_2 \sqrt{t} \right\} \leq \exp(-t). \quad (30)$$

793 **Lemma A.4** (Remark 2.8 from (Zhivotovskiy, 2024); see also example 4.3 from (Polyanskiy and Wu,
 794 2025)). Let X be a zero-mean sub-Gaussian random vector in \mathbb{R}^d with covariance matrix Σ . Then
 795 the norm of this vector can be bounded in probability as below

$$\mathbb{P} \left\{ \|X\|_2 > \sqrt{\text{tr}(\Sigma)} + \sqrt{2\|\Sigma\|_2 \ln \frac{1}{\delta}} \right\} \leq \delta. \quad (31)$$

796 B Bound for the Bias and Variance of Clipped Estimator

797 **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
 798 an arbitrary clipping level $\lambda > 0$. Let us assume

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

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

$$\begin{aligned} \|\mathbb{E}[\hat{X}] - \hat{x}\| &\leq \frac{2^{2\alpha-1} \sigma (\sigma^\alpha + (\max\{0, \|x\| - \lambda/2\})^\alpha)^{\frac{\alpha-1}{\alpha}}}{\lambda^{\alpha-1}} \\ &\quad + \max\{\|x\|, \lambda/2\} \frac{2^{2\alpha-1} (\sigma^\alpha + (\max\{0, \|x\| - \lambda/2\})^\alpha)}{\lambda^\alpha} \\ &\quad + \max\{0, \|x\| - \lambda/2\}, \end{aligned} \quad (32)$$

$$\mathbb{E} \|\hat{X} - \mathbb{E}\hat{X}\|^2 \leq \frac{9(2^{2\alpha-1} + 1)\lambda^{2-\alpha}\sigma^\alpha}{4} + \frac{9(2^{2\alpha-1} + 1)\lambda^{2-\alpha}(\max\{0, \|x\| - \lambda/2\})^\alpha}{4}. \quad (33)$$

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

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

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

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

804 **Proof of (32).** For the bias term, we obtain

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

805 where in (i), we used the fact that $\chi \in \{0, 1\}$ and when $\chi = 1$ we have $\left| \frac{\lambda}{\|X\|} - 1 \right| = 1 - \frac{\lambda}{\|X\|} \leq 1$.

806 Then, we continue the derivation as follows:

$$\begin{aligned} \|\mathbb{E}\hat{X} - \hat{x}\| &\leq \mathbb{E}[\chi \|X\|] + \max\left\{0, \|x\| - \frac{\lambda}{2}\right\} \\ &\stackrel{\chi \leq \eta}{\leq} \mathbb{E}[\eta \|X\|] + \max\left\{0, \|x\| - \frac{\lambda}{2}\right\} \\ &\leq \mathbb{E}[\eta \|X - x\|] + \mathbb{E}[\eta \|x\|] + \max\left\{0, \|x\| - \frac{\lambda}{2}\right\} \\ &\stackrel{(i)}{\leq} (\mathbb{E}\|X - x\|^\alpha)^{1/\alpha} (\mathbb{E}[\eta^{\alpha/\alpha-1}])^{(\alpha-1)/\alpha} + \mathbb{E}\eta \|x\| + \max\{0, \|x\| - \lambda/2\}, \end{aligned} \quad (34)$$

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

$$\mathbb{E}[\eta^{\alpha/\alpha-1}] = \mathbb{E}\eta = \mathbb{P}\{\|X - \hat{x}\| > \lambda/2\} = \mathbb{P}\{\|X - \hat{x}\|^\alpha > (\lambda/2)^\alpha\} \leq \frac{2^\alpha \mathbb{E}\|X - \hat{x}\|^\alpha}{\lambda^\alpha}. \quad (35)$$

808 Then, the expected value from the right-hand side (RHS) of (35) can be decomposed as follows

$$\begin{aligned}\mathbb{E}\|X - \hat{x}\|^\alpha &= \mathbb{E}\|X - x + x - \hat{x}\|^\alpha \leq 2^{\alpha-1}(\mathbb{E}\|X - x\|^\alpha + \max\{0, \|x\| - \lambda/2\}^\alpha) \\ &\leq 2^{\alpha-1}(\sigma^\alpha + \max\{0, \|x\| - \lambda/2\}^\alpha),\end{aligned}\quad (36)$$

809 where we use the Jensen's inequality for the convex function $\|x\|^\alpha$. After substitution of (36) into
810 (35), we get

$$\mathbb{E}[\eta^{\alpha/\alpha-1}] = \mathbb{E}\eta \leq \frac{2^{2\alpha-1}(\sigma^\alpha + \max\{0, \|x\| - \lambda/2\}^\alpha)}{\lambda^\alpha}. \quad (37)$$

811 Plugging the above bound in (34), we derive

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

812 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,
813 i.e., bound (32).

814 **Proof of (33).** First, we use the following standard inequality:

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

815 Then, we bound the RHS as

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

816 Applying upper bounds (36) and (37) from the previous part of the proof, we obtain

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

817 which concludes the proof. \square

818 C Missing Proofs

819 In this section, we provide the details of all the missing proofs for the main theorems and also the
820 derivation of the results provided in Tables 1 and 2.

821 C.1 Convex Functions

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

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

$$\begin{aligned} \frac{\gamma}{T+1} \sum_{t=0}^T c_t(f(x^t) - f^*) &\leq \frac{\|x^0 - x^*\|^2 - \|x^{T+1} - x^*\|^2}{T+1} - \frac{2\gamma}{T+1} \sum_{t=0}^T \langle x^t - x^*, \theta_t \rangle \\ &\quad - \frac{2\gamma}{T+1} \sum_{t=0}^T \langle x^t - x^*, \omega_t \rangle + \frac{2\gamma^2}{T+1} \sum_{t=0}^T \|\theta_t\|^2 \\ &\quad + \frac{4\gamma^2}{T+1} \sum_{t=0}^T \|\omega_t\|^2, \end{aligned}$$

827 where we have defined

$$c_t := \min \left\{ 1, \frac{\lambda}{2\|\nabla f(x^t)\|} \right\}, \quad (38)$$

$$\theta_t := \hat{g}_t - c_t \nabla f(x^t). \quad (39)$$

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

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

829 First, we rearrange the terms, and utilize the inequalities $\gamma \leq 1/8L$ and $c_t^2 \leq c_t$. Upon summing over
830 $t = 0, 1, \dots, T$, we obtain the following inequality

$$\begin{aligned} \frac{\gamma}{T+1} \sum_{t=0}^T c_t(f(x^t) - f^*) &\leq \frac{\|x^0 - x^*\|^2 - \|x^{T+1} - x^*\|^2}{T+1} - \frac{2\gamma}{T+1} \sum_{t=0}^T \langle x^t - x^*, \theta_t \rangle \\ &\quad - \frac{2\gamma}{T+1} \sum_{t=0}^T \langle x^t - x^*, \omega_t \rangle + \frac{2\gamma^2}{T+1} \sum_{t=0}^T \|\theta_t\|^2 + \frac{4\gamma^2}{T+1} \sum_{t=0}^T \|\omega_t\|^2, \end{aligned}$$

831 which concludes the proof. \square

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

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

$$\gamma_1 := \frac{R}{42(2^{2\alpha-1} + 1)^{1/2} \sigma^\alpha / 2 \lambda^{1-\alpha/2} \sqrt{6(K+1) \ln \frac{8(K+1)}{\beta} \left(1 + \frac{\zeta_\lambda}{\sigma^\alpha}\right)}}, \quad (40)$$

$$\gamma_2 := \frac{R \lambda^{\alpha-1}}{28(K+1) 2^{2\alpha-1} \sigma^\alpha \left(1 + \frac{\zeta_\lambda}{\sigma^\alpha}\right) \left(\frac{\zeta_\lambda}{\lambda} + \frac{1}{2} + \frac{\lambda^{\alpha-1} \zeta_\lambda}{2^{2\alpha-1} (\sigma^\alpha + \zeta_\lambda)} + \left(1 + \frac{\zeta_\lambda}{\sigma^\alpha}\right)^{-1/\alpha}\right)}, \quad (41)$$

$$\gamma_3 := \frac{R}{56 \sigma_\omega \sqrt{d(K+1)} (\sqrt{2} + \sqrt{2}\phi)}, \quad (42)$$

$$\gamma_4 := \frac{(2 - \sqrt{2})R}{\lambda + \sigma_\omega \left(\sqrt{d} + \sqrt{2 \ln \left(\frac{K+1}{\beta}\right)}\right)}, \quad (43)$$

$$\gamma_5 := \frac{R}{56 \lambda \ln \frac{8(K+1)}{\beta}}, \quad (44)$$

$$\gamma_6 := \frac{R}{2 \sigma_\omega \sqrt{7 \left[(K+1)d + 2 \sqrt{(K+1)d \ln \frac{4(K+1)}{\beta}} + 2 \ln \frac{4(K+1)}{\beta} \right]}}. \quad (45)$$

with $\phi := \sqrt{3 \ln \frac{4(K+1)}{\beta}}$ 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

$$\min_{k \in [0, K]} f(x^k) - f(x^*) \leq \frac{4R^2}{\gamma(K+1)} + \frac{64LR^4}{\lambda^2 \gamma^2 (K+1)^2} \quad \text{and} \quad \{x^k\}_{k=0}^K \subseteq B_{\sqrt{2}R}(x^*). \quad (46)$$

Proof. Let $R_k := \|x^k - x^*\|$ for all $k \geq 0$. Next, our goal is to show by induction that $R_k \leq 2R$ for all $k = 0, 1, \dots, K$ 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 $k = 0, \dots, K+1$ we consider probability event E_k defined as follows: inequalities

$$-2\gamma \sum_{l=0}^{t-1} \langle x^l - x^*, \theta_l \rangle - 2\gamma \sum_{l=0}^{t-1} \langle x^l - x^*, \omega_l \rangle + 2\gamma^2 \sum_{l=0}^{t-1} \|\theta_l\|^2 + 4\gamma^2 \sum_{l=0}^{t-1} \|\omega_l\|^2 \leq R^2, \quad (47)$$

$$R_t \leq \sqrt{2}R, \quad (48)$$

$$\|\omega_t\| \leq \sigma_\omega \left(\sqrt{d} + \sqrt{2 \ln \left(\frac{K+1}{(t+1)\beta} \right)} \right), \quad (49)$$

hold for all $t = 0, 1, \dots, k$ simultaneously. We want to prove via induction that $\mathbb{P}\{E_k\} \geq 1 - (k+1)^\beta / (K+1)^\beta$ for all $k = 0, 1, \dots, K$. For $k = 0$ the statements (47) and (48) trivially hold. Given Lemma A.4, statement (49) will also hold. Assume that the statement is true for some $k = T-1 \leq K$: $\mathbb{P}\{E_{T-1}\} \geq 1 - T^\beta / (K+1)^\beta$. One needs to prove that $\mathbb{P}\{E_T\} \geq 1 - (T+1)^\beta / (K+1)^\beta$. First, we notice that probability event E_{T-1} implies that $x_t \in B_{\sqrt{2}R}(x^*)$ for all $t = 0, 1, \dots, T-1$. For x^T , we can obtain the following inequalities

$$\begin{aligned} \|x^T - x^*\| &= \|x^{T-1} - x^* - \gamma \tilde{g}_{T-1}\| \leq \|x^{T-1} - x^*\| + \gamma \|\hat{g}_{T-1}\| + \gamma \|\omega_{T-1}\| \\ &\leq \sqrt{2}R + \gamma \lambda + \gamma \sigma_\omega \left(\sqrt{d} + \sqrt{2 \ln \left(\frac{K+1}{T\beta} \right)} \right) \stackrel{(43)}{\leq} 2R. \end{aligned} \quad (50)$$

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

$$\begin{aligned} \frac{\gamma}{t} \sum_{l=0}^{t-1} c_l (f(x^l) - f(x^*)) &\leq \frac{\|x^0 - x^*\|^2 - \|x^t - x^*\|^2}{t} + \frac{4\gamma^2}{t} \sum_{l=0}^{t-1} \|\omega_l\|^2 \\ &\quad - \frac{2\gamma}{t} \sum_{l=0}^{t-1} \langle x^l - x^*, \theta_l + \omega_l \rangle + \frac{2\gamma^2}{t} \sum_{l=0}^{t-1} \|\theta_l\|^2, \end{aligned} \quad (51)$$

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

$$\begin{aligned} \gamma \sum_{l=0}^{t-1} c_l (f(x^l) - f(x^*)) &\leq R^2 - 2\gamma \sum_{l=0}^{t-1} \langle x^l - x^*, \theta_l \rangle - 2\gamma \sum_{l=0}^{t-1} \langle x^l - x^*, \omega_l \rangle + 2\gamma^2 \sum_{l=0}^{t-1} \|\theta_l\|^2 \\ &\quad + 4\gamma^2 \sum_{l=0}^{t-1} \|\omega_l\|^2 \\ &\leq 2R^2, \end{aligned} \quad (52)$$

850 where we have used (47) for E_{T-1} . Taking into account that $\sum_{l=0}^{t-1} c_l (f(x^l) - f(x^*)) \geq 0$, (51)
 851 implies

$$R_T^2 \leq R^2 - 2\gamma \sum_{t=0}^{T-1} \langle x^t - x^*, \theta_t \rangle - 2\gamma \sum_{t=0}^{T-1} \langle x^t - x^*, \omega_t \rangle + 2\gamma^2 \sum_{t=0}^{T-1} \|\theta_t\|^2 + 4\gamma^2 \sum_{t=0}^{T-1} \|\omega_t\|^2. \quad (53)$$

852 Next, we define random vectors

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

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

$$\|\eta_t\| \leq 2R. \quad (54)$$

854 Next, we introduce the following vectors

$$\theta_t^u := \hat{g}_t - \mathbb{E}[\hat{g}_t | \mathcal{F}^{t-1}], \quad \theta_t^b := \mathbb{E}[\hat{g}_t | \mathcal{F}^{t-1}] - c_t \nabla f(x^t) \quad (55)$$

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

$$\begin{aligned} R_T^2 &\leq \underbrace{R^2 - 2\gamma \sum_{t=0}^{T-1} \langle \theta_t^u, \eta_t \rangle}_{\textcircled{1}} - \underbrace{2\gamma \sum_{t=0}^{T-1} \langle \theta_t^b, \eta_t \rangle}_{\textcircled{2}} - \underbrace{2\gamma \sum_{t=0}^{T-1} \langle \omega_t, \eta_t \rangle}_{\textcircled{3}} + \underbrace{4\gamma^2 \sum_{t=0}^{T-1} \mathbb{E}[\|\theta_t^u\|^2 | \mathcal{F}^{t-1}]}_{\textcircled{4}} \\ &\quad + \underbrace{4\gamma^2 \sum_{t=0}^{T-1} (\|\theta_t^u\|^2 - \mathbb{E}[\|\theta_t^u\|^2 | \mathcal{F}^{t-1}])}_{\textcircled{5}} + \underbrace{4\gamma^2 \sum_{t=0}^{T-1} \|\theta_t^b\|^2}_{\textcircled{6}} + \underbrace{4\gamma^2 \sum_{t=0}^{T-1} \|\omega_t\|^2}_{\textcircled{7}}. \end{aligned} \quad (56)$$

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

$$\|\theta_t^u\| \leq 2\lambda, \quad (57)$$

859 and

$$\begin{aligned} \|\theta_t^b\| &\leq \frac{2^{2\alpha-1} \sigma (\sigma^\alpha + (\max\{0, \|\nabla f(x^t)\| - \lambda/2\})^\alpha)^{\frac{\alpha-1}{\alpha}}}{\lambda^{\alpha-1}} \\ &\quad + \max\{\|\nabla f(x^t)\|, \lambda/2\} \frac{2^{2\alpha-1} (\sigma^\alpha + (\max\{0, \|\nabla f(x^t)\| - \lambda/2\})^\alpha)}{\lambda^\alpha} \\ &\quad + \max\{0, \|\nabla f(x^t)\| - \lambda/2\}, \end{aligned} \quad (58)$$

860

$$\mathbb{E} \left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1} \right] \leq \frac{9(2^{2\alpha-1} + 1)\lambda^{2-\alpha}\sigma^\alpha}{4} + \frac{9(2^{2\alpha-1} + 1)\lambda^{2-\alpha}(\max\{0, \|\nabla f(x^t)\| - \lambda/2\})^\alpha}{4}. \quad (59)$$

861 As can be seen, these bounds are iteration-dependent due to the presence of $\|\nabla f(x^t)\|$. As a remedy,
 862 we bound $\|\nabla f(x^t)\|$ by $2LR$ inside event E_{T-1} . This bound can be obtained from a combination
 863 of Assumption 2.2, E_{T-1} , and (50). Next, we introduce a new variable $\zeta_\lambda := \max\{0, 2LR - \frac{\lambda}{2}\}$.
 864 Thus, we get the following bounds for the bias and variance of θ_t : E_{T-1} implies

$$\|\theta_t^b\| \leq \frac{2^{2\alpha-1}\sigma(\sigma^\alpha + \zeta_\lambda^\alpha)^{\frac{\alpha-1}{\alpha}}}{\lambda^{\alpha-1}} + \left(\zeta_\lambda + \frac{\lambda}{2}\right) \frac{2^{2\alpha-1}(\sigma^\alpha + \zeta_\lambda^\alpha)}{\lambda^\alpha} + \zeta_\lambda, \quad (60)$$

865

$$\mathbb{E} \left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1} \right] \leq \frac{9(2^{2\alpha-1} + 1)\lambda^{2-\alpha}\sigma^\alpha}{4} + \frac{9(2^{2\alpha-1} + 1)\lambda^{2-\alpha}\zeta_\lambda^\alpha}{4} \quad (61)$$

866 for $t = 0, 1, \dots, T-1$.867 **Upper bound for ①.** By definition of θ_t^u , we have $\mathbb{E}[\theta_t^u \mid \mathcal{F}^{t-1}] = 0$ and

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

868 Furthermore, ① is bounded with probability 1 as

$$|2\gamma\langle\theta_t^u, \eta_t\rangle| \leq 2\gamma\|\theta_t^u\| \cdot \|\eta_t\| \stackrel{(57),(54)}{\leq} 8\gamma\lambda R \stackrel{(44)}{\leq} \frac{R^2}{7 \ln \frac{8(K+1)}{\beta}} := c. \quad (62)$$

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

$$\sigma_t^2 \leq \mathbb{E}[4\gamma^2\|\theta_t^u\|^2 \cdot \|\eta_t\|^2 \mid \mathcal{F}^{t-1}] \stackrel{(54)}{\leq} 16\gamma^2 R^2 \mathbb{E}[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}]. \quad (63)$$

870 In other words, we showed that $\{-2\gamma\langle\theta_t^u, \eta_t\rangle\}_{t=0}^{T-1}$ is a bounded martingale difference sequence with
 871 bounded conditional variances $\{\sigma_t^2\}_{t=0}^{T-1}$. Next, we apply Bernstein's inequality (Lemma A.1) with

872 $X_t = -2\gamma\langle\theta_t^u, \eta_t\rangle$, parameter c as in (62), $b = \frac{R^2}{7}$, $G = \frac{R^4}{294 \ln \frac{8(K+1)}{\beta}}$ to obtain

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

873 Equivalently, we have

$$\mathbb{P}\{E_{\textcircled{1}}\} \geq 1 - \frac{\beta}{4(K+1)}, \quad \text{for} \quad E_{\textcircled{1}} = \left\{ \text{either} \quad \sum_{t=0}^{T-1} \sigma_t^2 > \frac{R^4}{294 \ln \frac{8(K+1)}{\beta}} \quad \text{or} \quad |\textcircled{1}| \leq \frac{R^2}{7} \right\}. \quad (64)$$

874 In addition, E_{T-1} implies

$$\begin{aligned} \sum_{t=0}^{T-1} \sigma_t^2 &\leq 16\gamma^2 R^2 \sum_{t=0}^{T-1} \mathbb{E}[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}] \\ &\stackrel{(61)}{\leq} 4R^2\gamma^2 T (9(2^{2\alpha-1} + 1)\lambda^{2-\alpha}\sigma^\alpha + 9(2^{2\alpha-1} + 1)\lambda^{2-\alpha}\zeta_\lambda^\alpha) \\ &\stackrel{(40)}{\leq} \frac{R^4}{294 \ln \frac{8(K+1)}{\beta}}. \end{aligned} \quad (65)$$

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

$$\begin{aligned} \textcircled{2} &= -2\gamma \sum_{t=0}^{T-1} \langle\theta_t^b, \eta_t\rangle \leq 2\gamma \sum_{t=0}^{T-1} \|\theta_t^b\| \cdot \|\eta_t\| \\ &\stackrel{(60),(54)}{\leq} 4\gamma RT \left(\frac{2^{2\alpha-1}\sigma(\sigma^\alpha + \zeta_\lambda^\alpha)^{\frac{\alpha-1}{\alpha}}}{\lambda^{\alpha-1}} + (\zeta_\lambda + \lambda/2) \frac{2^{2\alpha-1}(\sigma^\alpha + \zeta_\lambda^\alpha)}{\lambda^\alpha} + \zeta_\lambda \right) \\ &\stackrel{T < K+1}{\leq} 4\gamma R(K+1) \frac{2^{2\alpha-1}}{\lambda^{\alpha-1}} (\sigma^\alpha + \zeta_\lambda^\alpha) \left(\left(1 + \frac{\zeta_\lambda^\alpha}{\sigma^\alpha}\right)^{-1/\alpha} + \frac{\zeta_\lambda}{\lambda} + \frac{1}{2} + \frac{\lambda^{\alpha-1}\zeta_\lambda}{2^{2\alpha-1}(\sigma^\alpha + \zeta_\lambda^\alpha)} \right) \\ &\stackrel{(41)}{\leq} \frac{R^2}{7}. \end{aligned} \quad (66)$$

876 **Upper bound for ③.** We have

$$|\textcircled{3}| = \left| -2\gamma \sum_{t=0}^{T-1} \langle \eta_t, \omega_t \rangle \right| = \left| \sum_{t=0}^{T-1} \sum_{i=1}^d 2\gamma \eta_{t,i} \omega_{t,i} \right| \quad (67)$$

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

878 Each summand is the product of a zero-mean Gaussian random variable and a bounded random
879 variable, resulting in the product being a zero-mean sub-Gaussian random variable with parameter
880 $\sigma_{t,i}^2 = 64R^2\gamma^2\sigma_\omega^2$. To prove this, consider

$$\begin{aligned} \mathbb{E} \left[\exp \left(\frac{4\gamma^2}{\sigma_{t,i}^2} |\eta_{t,i}^2 \omega_{t,i}^2| \right) \mid \mathcal{F}^{t-1} \right] &\stackrel{(54)}{\leq} \mathbb{E} \left[\exp \left(\frac{16R^2\gamma^2}{64\gamma^2 R^2 \sigma_\omega^2} |\omega_{t,i}|^2 \right) \right] \\ &\leq \mathbb{E} \left[\exp \left(\frac{|\omega_{t,i}|^2}{4\sigma_\omega^2} \right) \right] \stackrel{(ii)}{\leq} \exp(1) \end{aligned} \quad (68)$$

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

$$\begin{aligned} \mathbb{P} \left\{ \left| \sum_{t=0}^{T-1} \sum_{i=1}^d 2\gamma \eta_{t,i} \omega_{t,i} \right| > (\sqrt{2} + \sqrt{2}\phi) \sqrt{\sum_{t=0}^{T-1} \sum_{i=1}^d 64\gamma^2 R^2 \sigma_\omega^2} \right\} &\leq \exp \left(\frac{-\phi^2}{3} \right) \\ &= \frac{\beta}{4(K+1)}. \end{aligned} \quad (69)$$

883 The choice of $\gamma \leq \gamma_3$ for γ_3 defined (42) implies

$$(\sqrt{2} + \sqrt{2}\phi) \sqrt{\sum_{t=0}^{T-1} \sum_{i=1}^d 64\gamma^2 R^2 \sigma_\omega^2} \leq (\sqrt{2} + \sqrt{2}\phi) \sqrt{64\gamma^2 R^2 (K+1) d \sigma_\omega^2} \stackrel{(42)}{\leq} \frac{R^2}{7},$$

884 and

$$\mathbb{P}\{E_{\textcircled{3}}\} \geq 1 - \frac{\beta}{4(K+1)} \quad \text{for } E_{\textcircled{3}} = \left\{ |\textcircled{3}| \leq \frac{R^2}{7} \right\}. \quad (70)$$

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

$$\begin{aligned} \textcircled{4} &= 4\gamma^2 \sum_{t=0}^{T-1} \mathbb{E} \left[\|\theta_t^u\|^2 \mid \mathcal{F}_\xi^{t-1} \right] \\ &\stackrel{(61)}{\leq} 4\gamma^2 T \left(\frac{9(2^{2\alpha-1} + 1)\lambda^{2-\alpha}\sigma^\alpha}{4} + \frac{9(2^{2\alpha-1} + 1)\lambda^{2-\alpha}\zeta_\lambda^\alpha}{4} \right) \stackrel{(40)}{\leq} \frac{R^2}{7}. \end{aligned} \quad (71)$$

886 **Upper bound for ⑤.** First, we have

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

887 Next, sum ⑤ has bounded with probability 1 terms:

$$\begin{aligned} \left| 4\gamma^2 \left(\|\theta_t^u\|^2 - \mathbb{E} \left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1} \right] \right) \right| &\leq 4\gamma^2 \left(\|\theta_t^u\|^2 + \mathbb{E} \left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1} \right] \right) \\ &\leq 32\gamma^2 \lambda^2 \stackrel{(44)}{\leq} \frac{R^2}{7 \ln \frac{8(K+1)}{\beta}} := c. \end{aligned} \quad (72)$$

888 The summands also have bounded conditional variances

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

889

$$\tilde{\sigma}_t^2 \stackrel{(72)}{\leq} \frac{R^2}{7 \ln \frac{8(K+1)}{\beta}} \mathbb{E} \left[4\gamma^2 \left| \|\theta_t^u\|^2 - \mathbb{E} \left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1} \right] \right| \mid \mathcal{F}^{t-1} \right] \quad (73)$$

$$\leq \frac{8\gamma^2 R^2}{7 \ln \frac{8(K+1)}{\beta}} \mathbb{E} \left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1} \right]. \quad (74)$$

890 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 martin-
 891 gale difference sequence with bounded conditional variances $\{\tilde{\sigma}_t^2\}_{t=0}^{T-1}$. Next, we apply Bernstein's
 892 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 (72),
 893 $b = \frac{R^2}{7}$, $G = \frac{R^4}{294 \ln \frac{8(K+1)}{\beta}}$:

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

894 Equivalently, we have

$$\mathbb{P} \{E_{\mathfrak{E}}\} \geq 1 - \frac{\beta}{4(K+1)}, \quad \text{for} \quad E_{\mathfrak{E}} = \left\{ \text{either} \quad \sum_{t=0}^{T-1} \tilde{\sigma}_t^2 > \frac{R^4}{294 \ln \frac{8(K+1)}{\beta}} \quad \text{or} \quad |\mathfrak{E}| \leq \frac{R^2}{7} \right\}. \quad (75)$$

895 In addition, E_{T-1} implies that

$$\sum_{t=0}^{T-1} \tilde{\sigma}_t^2 \stackrel{(74)}{\leq} \frac{8\gamma^2 R^2 (K+1)}{7 \ln \frac{8(K+1)}{\beta}} \mathbb{E} \left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1} \right] \stackrel{(61),(40)}{\leq} \frac{R^4}{294 \ln \frac{8(K+1)}{\beta}}. \quad (76)$$

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

$$\begin{aligned} \textcircled{6} &= 4\gamma^2 \sum_{t=0}^{T-1} \|\theta_t^b\|^2 \\ &\leq 4\gamma^2 T \left(\frac{2^{2\alpha-1} \sigma (\sigma^\alpha + \zeta_\lambda^\alpha)^{\frac{\alpha-1}{\alpha}}}{\lambda^{\alpha-1}} + (\zeta_\lambda + \lambda/2) \frac{2^{2\alpha-1} (\sigma^\alpha + \zeta_\lambda^\alpha)}{\lambda^\alpha} + \zeta_\lambda \right)^2 \\ &\stackrel{(41)}{\leq} \frac{R^2}{7}. \end{aligned} \quad (77)$$

897 **Upper bound for ⑦.** We have

$$4\gamma^2 \sum_{t=0}^{T-1} \|\omega_t\|^2 = 4\gamma^2 \sigma_\omega^2 \sum_{t=0}^{T-1} \sum_{i=1}^d z_{t,i}^2, \quad (78)$$

898 where $z_{t,i} := \omega_{t,i} / \sigma_\omega$. Using Lemma A.3, we get

$$\mathbb{P} \left\{ \sum_{t=0}^{T-1} \sum_{i=1}^d z_{t,i}^2 > Td + 2\sqrt{Td \ln \frac{4(K+1)}{\beta}} + 2 \ln \frac{4(K+1)}{\beta} \right\} \leq \frac{\beta}{4(K+1)}. \quad (79)$$

899 Since $\gamma \leq \gamma_6$ for γ_6 defined in (45), we obtain

$$\mathbb{P} \left\{ \mathfrak{F} > \frac{R^2}{7} \right\} \leq \frac{\beta}{4(K+1)}, \quad (80)$$

900 which is equivalent to

$$\mathbb{P} \{E_{\mathfrak{F}}\} \geq 1 - \frac{\beta}{4(K+1)} \quad \text{for} \quad E_{\mathfrak{F}} = \left\{ |\mathfrak{F}| \leq \frac{R^2}{7} \right\}. \quad (81)$$

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

$$\begin{aligned} R_T^2 &\leq R^2 - 2\gamma \sum_{l=0}^{t-1} \langle x^l - x^*, \theta_l \rangle - 2\gamma \sum_{l=0}^{t-1} \langle x^l - x^*, \omega_l \rangle + 2\gamma^2 \sum_{l=0}^{t-1} \|\theta_l\|^2 + 4\gamma^2 \sum_{l=0}^{t-1} \|\omega_l\|^2 \\ &\leq R^2 + \textcircled{1} + \textcircled{2} + \textcircled{3} + \textcircled{4} + \textcircled{5} + \textcircled{6} + \textcircled{7} \\ &\leq R^2 + \frac{R^2}{7} + \frac{R^2}{7} + \frac{R^2}{7} + \frac{R^2}{7} + \frac{R^2}{7} + \frac{R^2}{7} + \frac{R^2}{7} = 2R^2, \end{aligned}$$

which is equivalent to (47) and (48) for $t = T$, and

$$\begin{aligned} \mathbb{P}\{E_T\} &\geq \mathbb{P}\{E_{T-1} \cap E_{\textcircled{1}} \cap E_{\textcircled{3}} \cap E_{\textcircled{5}} \cap E_{\textcircled{7}}\} \\ &= 1 - \mathbb{P}\{\overline{E}_{T-1} \cup \overline{E}_{\textcircled{1}} \cup \overline{E}_{\textcircled{3}} \cup \overline{E}_{\textcircled{5}} \cup \overline{E}_{\textcircled{7}}\} \\ &\geq 1 - \mathbb{P}\{\overline{E}_{T-1}\} - \mathbb{P}\{\overline{E}_{\textcircled{1}}\} - \mathbb{P}\{\overline{E}_{\textcircled{3}}\} - \mathbb{P}\{\overline{E}_{\textcircled{5}}\} - \mathbb{P}\{\overline{E}_{\textcircled{7}}\} \\ &\geq 1 - \frac{(T+1)\beta}{K+1}. \end{aligned} \quad (82)$$

This finishes the inductive part of our proof, i.e., for all $k = 0, 1, \dots, K$ 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 $1 - \beta$

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

and $\{x^k\}_{k=0}^K \subseteq Q$, which follows from (48). Now, we have to deal with c_t . To do so, we consider two possible cases for each $t = 0, 1, \dots, K$: either $c_t = 1$ or $c_t = \frac{\lambda}{2\|\nabla f(x^t)\|}$. We define the corresponding sets of indices: $\mathcal{T}_1 := \{t \in \{0, 1, \dots, K\} \mid c_t = 1\}$ and $\mathcal{T}_2 := \{t \in \{0, 1, \dots, K\} \mid c_t = \frac{\lambda}{2\|\nabla f(x^t)\|}\}$. Then, the above inequality can be rewritten as

$$\frac{1}{(K+1)} \sum_{t \in \mathcal{T}_1} (f(x^t) - f(x^*)) + \frac{1}{(K+1)} \sum_{t \in \mathcal{T}_2} \frac{\lambda(f(x^t) - f(x^*))}{2\|\nabla f(x^t)\|} \leq \frac{2R^2}{\gamma(K+1)}, \quad (83)$$

implying

$$\frac{1}{(K+1)} \sum_{t \in \mathcal{T}_1} (f(x^t) - f(x^*)) \leq \frac{2R^2}{\gamma(K+1)} \quad (84)$$

and

$$\frac{1}{(K+1)} \sum_{t \in \mathcal{T}_2} \frac{\lambda(f(x^t) - f(x^*))}{2\|\nabla f(x^t)\|} \leq \frac{2R^2}{\gamma(K+1)}. \quad (85)$$

Using the corollary of smoothness assumption, i.e., $\|\nabla f(x^t)\| \leq \sqrt{2L(f(x^t) - f(x^*))}$, we get from (85) that

$$\frac{1}{K+1} \sum_{t \in \mathcal{T}_2} \sqrt{f(x^t) - f(x^*)} \leq \frac{4\sqrt{2LR^2}}{\lambda\gamma(K+1)}. \quad (86)$$

For inequality (84), 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 $x^2 = f(x^t) - f(x^*)$, we get

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

implying

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

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

$$\frac{1}{K+1} \sum_{t \in \mathcal{T}_1} \sqrt{f(x^t) - f(x^*)} \leq \sqrt{\frac{2R^2}{\gamma(K+1)}}. \quad (87)$$

918 Combining inequalities (86) and (87), we get

$$\frac{1}{K+1} \sum_{t=0}^K \sqrt{f(x^t) - f(x^*)} \leq \sqrt{\frac{2R^2}{\gamma(K+1)}} + \frac{4\sqrt{2}LR^2}{\lambda\gamma(K+1)}, \quad (88)$$

919 which implies

$$\min_{t \in [0, K]} (f(x^t) - f(x^*)) \leq \frac{4R^2}{\gamma(K+1)} + \frac{64LR^4}{\lambda^2\gamma^2(K+1)^2}, \quad (89)$$

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

921 Theorem C.2 states 7 values for step-size, from which the smallest should be selected. To simplify
 922 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
 923 three step-sizes are redundant and can be omitted.

924
 925

926 **Corollary C.3.** *Let all conditions of Theorem C.2 hold. Furthermore, assume that K is large and
 927 one selects $\lambda \leq \mathcal{O}\left(\left(\frac{K}{\ln K}\right)^{1/\alpha}\right)$, then conclusions of Theorem C.2 are valid as long as γ is selected
 928 to satisfy $\gamma \leq \min\{1/8L, \gamma_1, \gamma_2, \gamma_3\}$ where we have*

$$\begin{aligned} \gamma_1 &:= \frac{R}{42(2^{2\alpha-1} + 1)^{1/2} \sigma^{\alpha/2} \lambda^{1-\alpha/2} \sqrt{6(K+1) \ln \frac{8(K+1)}{\beta} \left(1 + \frac{\zeta_\lambda^\alpha}{\sigma^\alpha}\right)}}, \\ \gamma_2 &:= \frac{R\lambda^{\alpha-1}}{28(K+1)2^{2\alpha-1} \sigma^\alpha \left(1 + \frac{\zeta_\lambda^\alpha}{\sigma^\alpha}\right) \left(\frac{\zeta_\lambda}{\lambda} + \frac{1}{2} + \frac{\lambda^{\alpha-1}\zeta_\lambda}{2^{2\alpha-1}(\sigma^\alpha + \zeta_\lambda^\alpha)} + \left(1 + \frac{\zeta_\lambda^\alpha}{\sigma^\alpha}\right)^{-1/\alpha}\right)}, \\ \gamma_3 &:= \frac{R}{56\sigma_\omega \sqrt{d(K+1)}(\sqrt{2} + \sqrt{2}\phi)}. \end{aligned}$$

929 *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 (45)
 930 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. Further-
 931 more, γ_5 in (44) 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
 932 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,
 933 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
 934 in (43), 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
 935 $\mathcal{O}\left(\sigma_\omega \epsilon / \sqrt{K \ln(K/\delta)}\right)$. Therefore, γ_4 decreases by the order $\mathcal{O}\left(\sigma_\omega \epsilon \sqrt{K \ln(K/\delta)}\right)$, which is the same
 936 order as γ_3 . Hence, γ_4 can be omitted, and the proof is complete. \square

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:

$$\gamma \leq \mathcal{O} \left(\min \left\{ \frac{1}{L}, \frac{R}{\sigma^{\alpha/2} \lambda^{1-\alpha/2} \sqrt{K \ln \frac{K}{\beta}}}, \frac{R\lambda^{\alpha-1}}{K\sigma^\alpha} \right\} \right). \quad (90)$$

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

$$\min_{t \in [0, K]} f(x^t) - f(x^*) = \mathcal{O}(\max \{ (92), (93), (94) \}), \quad (91)$$

where

$$R\lambda^{1-\alpha/2} \sigma^{\alpha/2} \sqrt{\frac{\ln K/\beta}{K}} + \frac{LR^2 \sigma^\alpha \ln K/\beta}{\lambda^\alpha K}, \quad (92)$$

$$\frac{R\sigma^\alpha}{\lambda^{\alpha-1}} + \frac{LR^2 \sigma^{2\alpha}}{\lambda^{2\alpha}}, \quad (93)$$

$$\frac{LR^2}{K} + \frac{L^3 R^4}{\lambda^2 K^2}. \quad (94)$$

We clearly see that the dominant term in (92) is an increasing function of λ , and the dominant term in (93) is a decreasing function. Solving for optimal λ as the equilibrium of the dominant terms in (92)

and (93), 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 $1 - \beta$:

$$\min_{t \in [0, K]} f(x^t) - f(x^*) = \mathcal{O}(\max \{ (96), (97) \}), \quad (95)$$

where

$$R\sigma \left(\frac{\ln \frac{K}{\beta}}{K} \right)^{\frac{\alpha-1}{\alpha}} + \frac{LR^2 \ln^2 K/\beta}{K^2}. \quad (96)$$

$$\frac{LR^2}{K} + \frac{L^3 R^4 \left(\ln \frac{K}{\beta} \right)^{\frac{2}{\alpha}}}{\sigma^2 K^{\frac{2\alpha+2}{\alpha}}}. \quad (97)$$

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 (92), (93), 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 (40), (41), 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: $\frac{4}{3}LR < \lambda \leq 4LR$, $\zeta_\lambda < \lambda < \sigma$. In this case, the step-size conditions reduce to the following:

$$\gamma \leq \mathcal{O} \left(\min \left\{ \frac{1}{L}, \frac{R}{\sigma^{\alpha/2} \lambda^{1-\alpha/2} \sqrt{K \ln \frac{K}{\beta}}}, \frac{R\lambda^{\alpha-1}}{K\sigma^\alpha} \right\} \right). \quad (98)$$

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 $1 - \beta$

$$\min_{t \in [0, K]} f(x^t) - f(x^*) = \mathcal{O}(\max\{(\text{100}), (\text{101}), (\text{102})\}), \quad (99)$$

where

$$\sqrt{R^{4-\alpha} L^{2-\alpha} \sigma^\alpha \frac{\ln K/\beta}{K}} + \frac{R^{2-\alpha} \sigma^\alpha \ln K/\beta}{L^{\alpha-1} K}, \quad (100)$$

$$\frac{R^{2-\alpha} \sigma^\alpha}{L^{\alpha-1}} + \frac{\sigma^{2\alpha}}{L^{2\alpha-1} R^{2\alpha-2}}, \quad (101)$$

$$\frac{LR^2}{K} + \frac{LR^2}{K^2}. \quad (102)$$

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

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{R}{\sigma^{\alpha/2} \lambda^{1-\alpha/2} \sqrt{K \ln \frac{K}{\beta}}}, \frac{R\lambda^{\alpha-1}}{K \max\{\sigma^\alpha, \lambda^{\alpha-1} \zeta_\lambda\}}\right\}\right). \quad (103)$$

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

$$R\lambda^{1-\alpha/2} \sigma^{\alpha/2} \sqrt{\frac{\ln K/\beta}{K}} + \frac{LR^2 \sigma^\alpha \ln K/\beta}{\lambda^\alpha K}, \quad (104)$$

$$R\zeta_\lambda + \frac{LR^2 \zeta_\lambda^2}{\lambda^2}, \quad (105)$$

$$\frac{LR^2}{K} + \frac{L^3 R^4}{\lambda^2 K^2}. \quad (106)$$

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 (109) is decreasing in λ , and $\lambda = 4LR$ is not feasible. With this choice of λ , we get with probability at least $1 - \beta$:

$$\min_{t \in [0, K]} f(x^t) - f(x^*) = \mathcal{O}(\max\{(\text{108}), (\text{109}), (\text{110})\}), \quad (107)$$

where

$$R\sqrt{(4LR - \eta)^{2-\alpha} \sigma^\alpha \frac{\ln K/\beta}{K}} + \frac{LR^2 \sigma^\alpha \ln K/\beta}{(LR - \eta)^\alpha K}, \quad (108)$$

$$\frac{R\eta}{2} + \frac{LR^2 \eta^2}{(4LR - \eta)^2}, \quad (109)$$

$$\frac{L\Delta}{K} + \frac{L^2 \Delta^2}{(4\sqrt{L\Delta} - \eta)^2 K^2}. \quad (110)$$

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

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{R}{\zeta_\lambda^{\alpha/2} \lambda^{1-\alpha/2} \sqrt{K \ln \frac{K}{\beta}}}, \frac{R\lambda^{\alpha-1}}{K(\lambda^{\alpha-1} \zeta_\lambda)}\right\}\right), \quad (111)$$

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

$$R\lambda^{1-\alpha/2}\zeta_\lambda^{\alpha/2}\sqrt{\frac{\ln K/\beta}{K}} + \frac{LR^2\zeta_\lambda^\alpha \ln K/\beta}{\lambda^\alpha K}, \quad (112)$$

$$R\zeta_\lambda + \frac{LR^2\zeta_\lambda^2}{\lambda^2}, \quad (113)$$

$$\frac{LR^2}{K} + \frac{L^3R^4}{\lambda^2K^2}. \quad (114)$$

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

$$\min_{t \in [0, K]} f(x^t) - f(x^*) = \mathcal{O}(\max\{(116), (117), (118)\}), \quad (115)$$

982 where

$$R\sqrt{(4LR - 2\sigma)^{2-\alpha}\sigma^\alpha \frac{\ln K/\beta}{K}} + \frac{LR^2\sigma^\alpha \ln K/\beta}{(4LR - 2\sigma)^\alpha K}, \quad (116)$$

$$R\sigma + \frac{LR^2\sigma^2}{(4LR - 2\sigma)^2}, \quad (117)$$

$$\frac{LR^2}{K} + \frac{L^3R^4}{(4LR - 2\sigma)^2K^2}. \quad (118)$$

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

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{R}{\sigma^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K \ln \frac{K}{\beta}}}, \frac{R\lambda^\alpha}{K(\sigma^\alpha\zeta_\lambda)}\right\}\right). \quad (119)$$

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

$$R\lambda^{1-\alpha/2}\sigma^{\alpha/2}\sqrt{\frac{\ln K/\beta}{K}} + \frac{LR^2\sigma^\alpha \ln K/\beta}{\lambda^\alpha K}, \quad (120)$$

$$R\frac{\sigma^\alpha\zeta_\lambda}{\lambda^\alpha} + \frac{LR^2\sigma^{2\alpha}\zeta_\lambda^2}{\lambda^{2\alpha+2}}, \quad (121)$$

$$\frac{LR^2}{K} + \frac{L^3R^4}{\lambda^2K^2}. \quad (122)$$

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

$$\min_{t \in [0, K]} f(x^t) - f(x^*) = \mathcal{O}(\max\{(124), (125), (126)\}), \quad (123)$$

987 where

$$\sqrt{R^{4-\alpha}L^{2-\alpha}\sigma^\alpha \frac{\ln K/\beta}{K}} + \frac{R^{2-\alpha}\sigma^\alpha \ln K/\beta}{L^{\alpha-1}K}, \quad (124)$$

$$\frac{R^{2-\alpha}\sigma^\alpha}{L^{\alpha-1}} + \frac{\sigma^{2\alpha}}{L^{2\alpha-1}R^{2\alpha-2}}, \quad (125)$$

$$\frac{LR^2}{K} + \frac{LR^2}{K^2}. \quad (126)$$

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

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{R}{\zeta_\lambda^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K \ln \frac{K}{\beta}}}, \frac{R\lambda^\alpha}{K(\zeta_\lambda^{\alpha+1})}\right\}\right). \quad (127)$$

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

$$R\lambda^{1-\alpha/2}\zeta_\lambda^{\alpha/2}\sqrt{\frac{\ln K/\beta}{K}} + \frac{LR^2\zeta_\lambda^\alpha \ln K/\beta}{\lambda^\alpha K}, \quad (128)$$

$$\frac{R\zeta_\lambda^{\alpha+1}}{\lambda^\alpha} + \frac{LR^2\zeta_\lambda^{2\alpha}}{\lambda^{2\alpha+2}}, \quad (129)$$

$$\frac{LR^2}{K} + \frac{L^3R^4}{\lambda^2K^2}. \quad (130)$$

991 Next, we find the optimal λ via equalizing the leading terms (the first ones) in (128) and (129). This
 992 results in $\lambda = \frac{4LR}{2C+1}$, where $C = \left(\frac{\ln K/\beta}{K}\right)^{\frac{1}{\alpha+2}}$, which is infeasible. Thus, in this regime, the optimal
 993 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
 994 probability at least $1 - \beta$

$$\min_{t \in [0, K]} f(x^t) - f(x^*) = \mathcal{O}(\max\{(132), (133), (134)\}), \quad (131)$$

995 where

$$R(LR - \eta)^{1-\alpha/2}(LR + \eta)^{\alpha/2}\sqrt{\frac{\ln K/\beta}{K}} + \frac{LR^2(LR + \eta)^{2\alpha} \ln K/\beta}{(LR - \eta)^{2\alpha+2}K}, \quad (132)$$

$$\frac{R(LR + \eta)^{\alpha+1}}{(LR - \eta)^\alpha} + \frac{LR^2(LR + \eta)^{2\alpha}}{(LR - \eta)^{2\alpha+2}}, \quad (133)$$

$$\frac{LR^2}{K} + \frac{L^3R^4}{(LR - \eta)^2K^2}. \quad (134)$$

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

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{R}{\zeta_\lambda^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K \ln \frac{K}{\beta}}}, \frac{R\lambda^{\alpha-1}}{K \max\left\{\frac{\zeta_\lambda^{\alpha+1}}{\lambda}, \zeta_\lambda^{\alpha-1}\sigma\right\}}\right\}\right). \quad (135)$$

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

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{R}{\zeta_\lambda^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{K \ln \frac{K}{\beta}}}, \frac{R\lambda^\alpha}{K(\zeta_\lambda^{\alpha+1})}\right\}\right), \quad (136)$$

999 and $\min_{t \in [0, K]} f(x^t) - f(x^*)$ is bounded with probability at least $1 - \beta$ by the maximum of the
 1000 following terms:

$$R\lambda^{1-\alpha/2}\zeta_\lambda^{\alpha/2}\sqrt{\frac{\ln K/\beta}{K}} + \frac{LR^2\zeta_\lambda^\alpha \ln K/\beta}{\lambda^\alpha K}, \quad (137)$$

$$\frac{R\zeta_\lambda^{\alpha+1}}{\lambda^\alpha} + \frac{LR^2\zeta_\lambda^{2\alpha}}{\lambda^{2\alpha+2}}, \quad (138)$$

$$\frac{LR^2}{K} + \frac{L^3R^4}{\lambda^2K^2}. \quad (139)$$

1001 The optimal λ is $\frac{4}{3}LR$, since the both leading terms in (137) and (138) are decreasing in λ . With this
 1002 choice, we get with probability at least $1 - \beta$

$$\min_{t \in [0, K]} f(x^t) - f(x^*) = \mathcal{O}(\max\{(141), (142), (143)\}), \quad (140)$$

1003 where

$$LR^2\sqrt{\frac{\ln K/\beta}{K}} + \frac{LR^2 \ln K/\beta}{K}, \quad (141)$$

$$R\sigma + \frac{\sigma^2}{L}, \quad (142)$$

$$\frac{LR^2}{K} + \frac{LR^2}{K^2}. \quad (143)$$

1004 Now that we have covered all regions, it's time to consider the DP noise as well.

1005 E Rate and Neighborhood for DP-Clipped-SGD: Convex Case

1006 To ensure the output of the algorithm is (ε, δ) -differentially private in this setting, expectation
 1007 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
 1008 advanced composition theorem of [Dwork et al. \(2014\)](#). In the finite sum case, one can reduce the
 1009 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
 1010 brevity, in the DP case, we only consider two cases: large λ and relatively small λ regimes. The other
 1011 cases can be derived with a similar analysis.

1012 **Case 1:** $\lambda > 4LR$. In this case, $\zeta_\lambda = 0$, and the step-size conditions reduce to the following:

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{R}{\sigma^{\alpha/2} \lambda^{1-\alpha/2} \sqrt{K \ln \frac{K}{\beta}}}, \frac{R \lambda^{\alpha-1}}{K \sigma^\alpha}, \frac{R}{\sigma_\omega \sqrt{d K \ln \frac{K}{\beta}}}\right\}\right). \quad (144)$$

1013 In particular, when γ equals the minimum from step-size condition, then the iterates produced by
 1014 [DP-Clipped-SGD](#) after K iterations with probability at least $1 - \beta$ satisfy

$$\min_{k \in [0, K]} f(x^k) - f(x^*) = \mathcal{O}(\max\{(\text{146}), (\text{147}), (\text{148}), (\text{149})\}), \quad (145)$$

1015 where

$$R \lambda^{1-\alpha/2} \sigma^{\alpha/2} \sqrt{\frac{\ln K/\beta}{K}} + \frac{LR^2 \sigma^\alpha \ln K/\beta}{\lambda^\alpha K}, \quad (146)$$

$$\frac{R \sigma^\alpha}{\lambda^{\alpha-1}} + \frac{LR^2 \sigma^{2\alpha}}{\lambda^{2\alpha}}, \quad (147)$$

$$\frac{LR^2}{K} + \frac{L^3 R^4}{\lambda^2 K^2}, \quad (148)$$

$$R \sigma_\omega \sqrt{\frac{d \ln \frac{K}{\beta}}{K}} + \frac{LR^2 \sigma_\omega^2 d \ln \frac{K}{\beta}}{\lambda^2 K}. \quad (149)$$

1016 Here, (147) accounts for the bias caused by clipping, and (149) accounts for the accumula-
 1017 tion of DP noise. These terms are decreasing and increasing in λ respectively, if we use
 1018 $\sigma_\omega = \Theta\left(\frac{\lambda}{\varepsilon} \sqrt{K \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{1}{\delta}\right)}\right)$. To find the optimal λ , we find the equilibrium of these two
 1019 terms. Solving the equilibrium equation, we get $\lambda = \mathcal{O}\left(\left(\frac{\varepsilon \sigma^\alpha}{d \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{K}{\beta}\right)}\right)^{\frac{1}{\alpha}}\right)$. Unless $\varepsilon \sigma^\alpha$ is
 1020 large enough, this value violates the constraint that $\lambda > 4LR$, and it's not feasible. Thus, we have the
 1021 following formula for the optimal λ :

$$\lambda = \max\left\{4LR, \left(\frac{\varepsilon \sigma^\alpha}{d \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{K}{\beta}\right)}\right)^{\frac{1}{\alpha}}\right\}. \quad (150)$$

1022 For this choice of λ , we get that with probability at least $1 - \beta$

$$\min_{k \in [0, K]} f(x^k) - f(x^*) = \mathcal{O}(\max\{(\text{152}), (\text{153}), (\text{154}), (\text{155})\}), \quad (151)$$

1023 with

$$\max \left\{ \sqrt{R^{4-\alpha} L^{2-\alpha} \sigma^\alpha \frac{\ln K/\beta}{K}}, R \left(\frac{\varepsilon \sigma^\alpha}{\sqrt{d \ln \left(\frac{1}{\delta}\right) \ln \left(\frac{K}{\delta}\right) \ln \left(\frac{K}{\beta}\right)}} \right)^{\frac{1}{\alpha}} \sqrt{\frac{\ln \frac{3\alpha-2}{2\alpha} \frac{K}{\beta}}{K}} \right\}, \quad (152)$$

$$\min \left\{ \frac{R^{2-\alpha} \sigma^\alpha}{L^{\alpha-1}}, R \sigma \left(\frac{d \ln \left(\frac{1}{\delta}\right) \ln \left(\frac{K}{\delta}\right)}{\varepsilon} \right)^{\frac{\alpha-1}{\alpha}} \right\}, \quad (153)$$

$$\min \left\{ \frac{LR^2}{K^2}, \frac{L^3 R^4 \left(d \ln \left(\frac{1}{\delta}\right) \ln \left(\frac{K}{\delta}\right) \right)^{\frac{1}{\alpha}} \ln^{\frac{1}{\alpha}} \frac{K}{\beta}}{(\varepsilon)^{\frac{1}{\alpha}} \sigma} \frac{1}{K^2} \right\} + \frac{LR^2}{K}, \quad (154)$$

$$\max \left\{ \frac{LR^2}{\varepsilon} \sqrt{d \ln \left(\frac{1}{\delta}\right) \ln \left(\frac{K}{\delta}\right) \ln \left(\frac{K}{\beta}\right)}, \frac{R \sigma \left(d \ln \left(\frac{1}{\delta}\right) \ln \left(\frac{K}{\delta}\right) \ln \left(\frac{K}{\beta}\right) \right)^{\frac{\alpha+2}{2\alpha}}}{\varepsilon^{\frac{\alpha-1}{\alpha}}} \right\} \\ + \frac{LR^2}{\varepsilon^2} d \ln \left(\frac{1}{\delta}\right) \ln \left(\frac{K}{\delta}\right) \ln \left(\frac{K}{\beta}\right), \quad (155)$$

1024 where, for the sake of brevity, we only report the dominant terms.

1025 **Case 2:** $\lambda \leq \frac{4}{3}LR$ $\lambda < \sigma < \zeta_\lambda$. In this case, the step-size conditions reduce to

$$\gamma \leq \mathcal{O} \left(\min \left\{ \frac{1}{L}, \frac{R}{\zeta_\lambda^{\alpha/2} \lambda^{1-\alpha/2} \sqrt{K \ln \frac{K}{\beta}}}, \frac{R \lambda^\alpha}{K(\zeta_\lambda^{\alpha+1})}, \frac{R}{\sigma_\omega \sqrt{d K \ln \frac{K}{\beta}}} \right\} \right), \quad (156)$$

1026 Taking γ equal to the right-hand side, we get that with probability at least $1 - \beta$

$$\min_{t \in [0, K]} f(x^t) - f(x^*) = \mathcal{O}(\{(\text{158}), (\text{159}), (\text{160}), (\text{161})\}), \quad (157)$$

1027 with

$$R \lambda^{1-\alpha/2} \sigma^{\alpha/2} \sqrt{\frac{\ln K/\beta}{K}} + \frac{LR^2 \sigma^\alpha \ln K/\beta}{\lambda^\alpha K}, \quad (158)$$

$$\frac{R \zeta_\lambda^{\alpha+1}}{\lambda^\alpha} + \frac{LR^2 \zeta_\lambda^{2\alpha}}{\lambda^{2\alpha+2}}, \quad (159)$$

$$\frac{LR^2}{K} + \frac{L^3 R^4}{\lambda^2 K^2}, \quad (160)$$

$$R \sigma_\omega \sqrt{\frac{d \ln \frac{K}{\beta}}{K}} + \frac{LR^2 \sigma_\omega^2 d \ln \frac{K}{\beta}}{\lambda^2 K}. \quad (161)$$

1028 Similarly to the previous case, we find the optimal λ as the equilibrium of the leading terms in (159)
1029 and (161). By doing so, we get the optimal λ :

$$\lambda = \min \left\{ \frac{4}{3}LR, \frac{2\varepsilon LR}{\left(d \ln \left(\frac{1}{\delta}\right) \ln \left(\frac{K}{\delta}\right) \ln \left(\frac{K}{\beta}\right) \right)^{\frac{1}{2\alpha+2}} + 1} \right\}. \quad (162)$$

1030 For this choice of λ , we get that with probability at least $1 - \beta$

$$\min_{k \in [0, K]} f(x^t) - f(x^*) = \mathcal{O}(\max\{(\text{164}), (\text{165}), (\text{166}), (\text{167})\}), \quad (163)$$

1031 with

$$\min \left\{ \sqrt{R^{4-\alpha} L^{2-\alpha} \sigma^\alpha \frac{\ln K/\beta}{K}}, \sqrt{\frac{R^{4-\alpha} (\varepsilon L)^{2-\alpha} \ln^{\frac{3\alpha}{4\alpha+4}} \frac{K}{\beta}}{(d \ln(\frac{1}{\delta}) \ln(\frac{K}{\delta}))^{\frac{2-\alpha}{4\alpha+4}} K}} \right\}, \quad (164)$$

$$\max \left\{ \frac{R^{2-\alpha} \sigma^\alpha}{L^{\alpha-1}}, \frac{R^{2-\alpha} \sigma^\alpha}{\varepsilon} \left(d \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{K}{\beta}\right) \right)^{\frac{\alpha-1}{2\alpha+2}} \right\}, \quad (165)$$

$$\max \left\{ \frac{LR^2}{K^2}, \frac{LR^2}{\varepsilon^2 K^2} \left(\left(d \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{K}{\beta}\right) \right)^{\frac{1}{2\alpha+2}} + 1 \right)^2 \right\} + \frac{LR^2}{K}, \quad (166)$$

$$\min \left\{ \frac{LR^2}{\varepsilon} \sqrt{d \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{K}{\beta}\right)}, \frac{LR^2 \sqrt{\ln \frac{K}{\beta}}}{\left(d \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{K}{\beta}\right) \right)^{\frac{1}{2\alpha+2}} + 1} \right\} \\ + \frac{LR^2 d}{\varepsilon^2} \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{K}{\beta}\right), \quad (167)$$

1032 where, for the sake of brevity, we only report the dominant terms.

1033 F Non-Convex Functions

1034 Now, we focus on the case of non-convex functions. We start with the following lemma.

1035 **Lemma F.1.** *Let Assumptions 2.1, 2.2 hold on the set $Q =$*
 1036 *$\{x \in \mathbb{R} \mid \exists y \in \mathbb{R}^d : f(y) \leq f^* + 2\Delta \text{ and } \|x - y\| \leq \sqrt{\Delta}/20\sqrt{L}\}$, where $\Delta \geq \Delta_0 = f(x^0) - f^*$ and*
 1037 *let $0 < \gamma \leq 1/4L$. If $x^k \in Q$ for all $k = 0, 1, \dots, K$ for some $K \geq 0$, then the iterates produced by*
 1038 *DP-Clipped-SGD satisfy*

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

1039 for all $T = 0, 1, \dots, K$, and θ_t, c_t are defined in (39), (38) respectively.

1040 *Proof.* The smoothness of f implies

$$\begin{aligned} f(x^{t+1}) &\leq f(x^t) + \langle \nabla f(x^t), x^{t+1} - x^t \rangle + \frac{L}{2} \|x^{t+1} - x^t\|^2 \\ &= f(x^t) - \gamma \langle \nabla f(x^t), \hat{g}_t + \omega_t + c_t \nabla f(x^t) - c_t \nabla f(x^t) \rangle \\ &\quad + \frac{L\gamma^2}{2} \|\hat{g}_t + \omega_t + c_t \nabla f(x^t) - c_t \nabla f(x^t)\|^2 \\ &\leq f(x^t) - \gamma c_t \|\nabla f(x^t)\|^2 - \gamma \langle \nabla f(x^t), \theta_t \rangle - \gamma \langle \nabla f(x^t), \omega_t \rangle + L\gamma^2 \|\omega_t\|^2 \\ &\quad + 2L\gamma^2 \|\theta_t\|^2 + 2L\gamma^2 c_t^2 \|\nabla f(x^t)\|^2 \\ &= f(x^t) - (\gamma c_t - 2\gamma^2 L c_t^2) \|\nabla f(x^t)\|^2 - \gamma \langle \nabla f(x^t), \theta_t \rangle - \gamma \langle \nabla f(x^t), \omega_t \rangle \\ &\quad + L\gamma^2 \|\omega_t\|^2 + 2L\gamma^2 \|\theta_t\|^2. \end{aligned} \tag{168}$$

1041 Rearranging the terms, utilizing $\gamma \leq 1/4L$, and $c_t^2 \leq c_t$, we sum over t to obtain

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

1042 which concludes the proof. □

1043 The above lemma is utilized to prove the main convergence result for DP-Clipped-SGD.

1044 **Theorem F.2.** *Let Assumptions 2.1, 2.2, and 2.4 hold for the following set $Q =$*
 1045 *$\{x \in \mathbb{R} \mid \exists y \in \mathbb{R}^d : f(y) \leq f^* + 2\Delta \text{ and } \|x - y\| \leq \sqrt{\Delta}/20\sqrt{L}\}$, where $\Delta \geq \Delta_0 = f(x^0) - f^*$,*

1046 $\zeta_\lambda = \max\{0, 2\sqrt{L\Delta} - \frac{\lambda}{2}\}$, and $\gamma = \min\{1/4L, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6\}$,

$$\gamma_1 := \frac{\sqrt{\Delta}}{21\sqrt{L}(2^{2\alpha-1} + 1)^{1/2}\sigma^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{6(K+1)\ln\frac{8(K+1)}{\beta}\left(1 + \frac{\zeta_\lambda^\alpha}{\sigma^\alpha}\right)}}, \quad (169)$$

$$\gamma_2 := \frac{\sqrt{\Delta}\lambda^{\alpha-1}}{14\sqrt{L}(K+1)2^{2\alpha-1}(\sigma^\alpha + \zeta_\lambda^\alpha)\left(\frac{\zeta_\lambda}{\lambda} + \frac{1}{2} + \frac{\lambda^{\alpha-1}\zeta_\lambda}{2^{2\alpha-1}(\sigma^\alpha + \zeta_\lambda^\alpha)} + \left(1 + \frac{\zeta_\lambda^\alpha}{\sigma^\alpha}\right)^{-1/\alpha}\right)}, \quad (170)$$

$$\gamma_3 := \frac{\sqrt{\Delta}}{14\sqrt{L}\sigma_\omega\sqrt{d(K+1)}(\sqrt{2} + \sqrt{2\phi})}, \quad (171)$$

$$\gamma_4 := \frac{\sqrt{\Delta}}{20\sqrt{L}\left(\lambda + \sigma_\omega\left(\sqrt{d} + \sqrt{2\ln\left(\frac{K+1}{\beta}\right)}\right)\right)}, \quad (172)$$

$$\gamma_5 := \frac{\sqrt{\Delta}}{28\lambda\sqrt{L}\ln\frac{8(K+1)}{\beta}}, \quad (173)$$

$$\gamma_6 := \frac{\sqrt{\Delta}}{\sqrt{L}\sigma_\omega\sqrt{7\left((K+1)d + 2\sqrt{(K+1)d\ln\frac{4(K+1)}{\beta}} + 2\ln\frac{4(K+1)}{\beta}\right)}}. \quad (174)$$

1047 for some $K > 0$ and $\beta \in (0, 1]$. Then, after K iterations of **DP-Clipped-SGD** the iterates with
1048 probability at least $1 - \beta$ satisfy

$$\min_{t \in [0, K]} \|\nabla f(x^t)\|^2 \leq \frac{8\Delta}{\gamma(K+1)} + \frac{128\Delta^2}{\lambda^2\gamma^2(K+1)^2}. \quad (175)$$

1049 *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
1050 high probability. This fact will allow us to apply Lemma F.1 and then use Bernstein's inequality to
1051 evaluate the stochastic part of the upper-bound. More precisely, for each $k = 0, \dots, K$ we define the
1052 probability event E_k as follows. The inequalities

$$-\gamma \sum_{t=0}^T \langle \nabla f(x^t), \omega_t + \theta_t \rangle + L\gamma^2 \sum_{t=0}^T (2\|\theta_t\|^2 + \|\omega_t\|^2) \leq \Delta, \quad (176)$$

$$\Delta_t \leq 2\Delta, \quad (177)$$

$$\|\omega_t\| \leq \sigma_\omega \left(\sqrt{d} + \sqrt{2\ln\left(\frac{K+1}{(t+1)\beta}\right)} \right), \quad (178)$$

1053 hold for all $t = 0, 1, \dots, k$ simultaneously. We want to prove via induction that $\mathbb{P}\{E_k\} \geq 1 -$
1054 $(k+1)^\beta/(K+1)$ for all $k = 0, 1, \dots, K$. For $k = 0$ the statement is trivial. Assume that the statement
1055 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
1056 $\mathbb{P}\{E_T\} \geq 1 - (T+1)^\beta/(K+1)$. First, we notice that the probability event E_{T-1} implies $\Delta_t \leq 2\Delta$ for
1057 all $t = 0, 1, \dots, T-1$, i.e., $x^t \in \{y \in \mathbb{R}^d \mid f(y) \leq f^* + 2\Delta\}$ for $t = 0, 1, \dots, T-1$. Moreover,
1058 due to the choice of clipping level λ , we have

$$\|x^T - x^{T-1}\| = \gamma\|\hat{g}_{T-1}\| + \gamma\|\omega_{T-1}\| \leq \gamma\lambda + \gamma\sigma_\omega \left(\sqrt{d} + \sqrt{2\ln\left(\frac{K+1}{T\beta}\right)} \right) \stackrel{(172)}{\leq} \frac{\sqrt{\Delta}}{20\sqrt{L}}.$$

1059 Therefore, E_{T-1} implies $\{x^k\}_{k=0}^T \in Q$, meaning that the assumptions of Lemma F.1 are satisfied
1060 and we have

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

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

$$\begin{aligned} \frac{\gamma}{2} \sum_{l=0}^{t-1} c_l \|\nabla f(x^l)\|^2 &\leq \Delta - \gamma \sum_{k=0}^{t-1} \langle \nabla f(x^l), \theta_l \rangle - \gamma \sum_{k=0}^{t-1} \langle \nabla f(x^l), \omega_l \rangle + 2L\gamma^2 \sum_{l=0}^{t-1} \|\theta_l\|^2 \\ &\quad + L\gamma^2 \sum_{l=0}^{t-1} \|\omega_l\|^2 \\ &\leq 2\Delta. \end{aligned} \quad (179)$$

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

$$\eta_t = \begin{cases} \nabla f(x^t), & \text{if } \|\nabla f(x^t)\| \leq 2\sqrt{L\Delta}, \\ 0, & \text{otherwise,} \end{cases} \quad (180)$$

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

$$\|\eta_t\| \leq 2\sqrt{L\Delta}. \quad (181)$$

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

$$\|\nabla f(x^l)\| \stackrel{(180)}{\leq} \sqrt{2L(f(x^l) - f^*)} = \sqrt{2L\Delta_l} \leq 2\sqrt{L\Delta}, \quad (182)$$

meaning that E_{T-1} implies that $\eta_t = \nabla f(x^t)$ for all $t = 0, 1, \dots, T-1$. We notice that $\theta_t = \theta_t^u + \theta_t^b$, where θ_t^u and θ_t^b are defined in (55). Using new notation, we get that E_{T-1} implies

$$\begin{aligned} \Delta_T &\leq \underbrace{\Delta - \gamma \sum_{t=0}^{T-1} \langle \theta_t^u, \eta_t \rangle}_{\textcircled{1}} - \underbrace{\gamma \sum_{t=0}^{T-1} \langle \theta_t^b, \eta_t \rangle}_{\textcircled{2}} - \underbrace{\gamma \sum_{t=0}^{T-1} \langle \omega_t, \eta_t \rangle}_{\textcircled{3}} + \underbrace{4L\gamma^2 \sum_{t=0}^{T-1} \mathbb{E} [\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}]}_{\textcircled{4}} \\ &\quad + \underbrace{4L\gamma^2 \sum_{t=0}^{T-1} (\|\theta_t^u\|^2 - \mathbb{E} [\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}])}_{\textcircled{5}} + \underbrace{4L\gamma^2 \sum_{t=0}^{T-1} \|\theta_t^b\|^2}_{\textcircled{6}} + \underbrace{L\gamma^2 \sum_{t=0}^{T-1} \|\omega_t\|^2}_{\textcircled{7}}. \end{aligned} \quad (183)$$

It remains to derive good enough high-probability upper bounds for the terms $\textcircled{1}, \textcircled{2}, \textcircled{3}, \textcircled{4}, \textcircled{5}, \textcircled{6}, \textcircled{7}$. This amounts to proving $\textcircled{1} + \textcircled{2} + \textcircled{3} + \textcircled{4} + \textcircled{5} + \textcircled{6} + \textcircled{7} \leq \Delta$ 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

$$\|\theta_t^u\| \leq 2\lambda, \quad (184)$$

and from Lemma B.1 we also have

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

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

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

$$\|\theta_t^b\| \leq \frac{2^{2\alpha-1} \sigma (\sigma^\alpha + \zeta_\lambda^\alpha)^{\frac{\alpha-1}{\alpha}}}{\lambda^{\alpha-1}} + \left(\zeta_\lambda + \frac{\lambda}{2} \right) \frac{2^{2\alpha-1} (\sigma^\alpha + \zeta_\lambda^\alpha)}{\lambda^\alpha} + \zeta_\lambda, \quad (185)$$

$$\mathbb{E} [\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}] \leq \frac{9(2^{2\alpha-1} + 1)\lambda^{2-\alpha}\sigma^\alpha}{4} + \frac{9(2^{2\alpha-1} + 1)\lambda^{2-\alpha}\zeta_\lambda^\alpha}{4}. \quad (186)$$

1077 **Upper bound for ①.** By definition of θ_t^u , we have $\mathbb{E}[\theta_t^u \mid \mathcal{F}^{t-1}] = 0$ and

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

1078 Next, sum ① has bounded with probability 1 terms:

$$|\gamma \langle \theta_t^u, \eta_t \rangle| \leq \gamma \|\theta_t^u\| \cdot \|\eta_t\| \stackrel{(180)}{\leq} 4\gamma\lambda\sqrt{L\Delta} \stackrel{(173)}{\leq} \frac{\Delta}{7 \ln \frac{8(K+1)}{\beta}} := c. \quad (187)$$

1079 The summands also have bounded conditional variances $\sigma_t^2 := \mathbb{E}[\gamma^2 \langle \theta_t^u, \eta_t \rangle^2 \mid \mathcal{F}^{t-1}]$:

$$\sigma_t^2 \leq \mathbb{E}[\gamma^2 \|\theta_t^u\|^2 \cdot \|\eta_t\|^2 \mid \mathcal{F}^{t-1}] \leq 4\gamma^2 L\Delta \mathbb{E}[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}]. \quad (188)$$

1080 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
1081 bounded conditional variances $\{\sigma_t^2\}_{t=0}^{T-1}$. Next, we apply Bernstein's inequality (Lemma A.1) with

1082 $X_t = -\gamma \langle \theta_t^u, \eta_t \rangle$, parameter c as in (187), $b = \frac{\Delta}{7}$, $G = \frac{\Delta^2}{294 \ln \frac{8(K+1)}{\beta}}$:

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

1083 Equivalently, we have

$$\mathbb{P}\{E_{\textcircled{1}}\} \geq 1 - \frac{\beta}{4(K+1)}, \quad \text{for } E_{\textcircled{1}} = \left\{ \text{either } \sum_{t=0}^{T-1} \sigma_t^2 > \frac{\Delta^2}{294 \ln \frac{8(K+1)}{\beta}} \quad \text{or} \quad |\textcircled{1}| \leq \frac{\Delta}{7} \right\}. \quad (189)$$

1084 In addition, E_{T-1} implies that

$$\begin{aligned} \sum_{t=0}^{T-1} \sigma_t^2 &\leq 4\gamma^2 L\Delta \sum_{t=0}^{T-1} \mathbb{E}[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}] \\ &\stackrel{(186)}{\leq} 9\gamma^2 L\Delta T ((2^{2\alpha-1} + 1) \lambda^{2-\alpha} \sigma^\alpha + (2^{2\alpha-1} + 1) \lambda^{2-\alpha} \zeta_\lambda) \\ &\stackrel{(169)}{\leq} \frac{\Delta^2}{294 \ln \frac{8(K+1)}{\beta}}. \end{aligned} \quad (190)$$

1085 **Upper bound for ②.** From E_{T-1} it follows that

$$\begin{aligned} \textcircled{2} &= -\gamma \sum_{t=0}^{T-1} \langle \theta_t^b, \eta_t \rangle \leq \gamma \sum_{t=0}^{T-1} \|\theta_t^b\| \cdot \|\eta_t\| \\ &\stackrel{(185)}{\leq} 2\gamma\sqrt{L\Delta}T \left(\frac{2^{2\alpha-1} \sigma (\sigma^\alpha + \zeta_\lambda^\alpha)^{\frac{\alpha-1}{\alpha}}}{\lambda^{\alpha-1}} + (\zeta_\lambda + \lambda/2) \frac{2^{2\alpha-1} (\sigma^\alpha + \zeta_\lambda^\alpha)}{\lambda^\alpha} + \zeta_\lambda \right) \\ &\stackrel{(170)}{\leq} \frac{\Delta}{7}. \end{aligned} \quad (191)$$

1086 **Upper bound for ③.** We have

$$|\textcircled{3}| = \left| -\gamma \sum_{t=0}^{T-1} \langle \omega_t, \eta_t \rangle \right| = \left| \sum_{t=0}^{T-1} \sum_{i=1}^d \gamma \omega_{t,i} \eta_{t,i} \right|, \quad (192)$$

1087 where $\eta_{t,i} := [\eta_t]_i$ and $\omega_{t,i} := [\omega_t]_i$ denote the i -th components of η_t and ω_t respectively.

1088 Each summand is the product of a zero-mean Gaussian random variable and a bounded random
1089 variable, resulting in the product being a zero-mean light-tailed random variable with parameter
1090 $\sigma_{t,i}^2 = 16\gamma^2 L\Delta \sigma_\omega^2$. To prove this, consider

$$\begin{aligned} \mathbb{E} \left[\exp \left(\frac{\gamma^2}{\sigma_{t,i}^2} |\eta_{t,i}^2 \omega_{t,i}^2| \right) \mid \mathcal{F}^{t-1} \right] &\stackrel{(181)}{\leq} \mathbb{E} \left[\exp \left(\frac{4L\Delta\gamma^2}{16\gamma^2 L\Delta \sigma_\omega^2} |\omega_{t,i}|^2 \right) \right] \\ &\leq \exp \left(\frac{|\omega_{t,i}|^2}{4\sigma_\omega^2} \right) \stackrel{(ii)}{\leq} \exp(1), \end{aligned} \quad (193)$$

1091 where (ii) uses the fact that $\omega_{t,i}^2$ is a sub-Gaussian random variable with parameter σ_ω^2 . Now that we
 1092 have established the light-tailedness of summands, we can use the Lemma A.2 to obtain

$$\mathbb{P} \left\{ \left| \sum_{t=0}^{T-1} \sum_{i=1}^d \gamma \eta_{t,i} \omega_{t,i} \right| > (\sqrt{2} + \sqrt{2}\phi) \sqrt{\sum_{t=0}^{T-1} \sum_{i=1}^d 4\gamma^2 L \Delta \sigma_\omega^2} \right\} \leq \exp \left(\frac{-\phi^2}{3} \right) \quad (194)$$

$$= \frac{\beta}{4(K+1)}. \quad (195)$$

1093 The choice of $\gamma \leq \gamma_3$ for γ_3 defined in (171) implies

$$(\sqrt{2} + \sqrt{2}\phi) \sqrt{\sum_{t=0}^{T-1} \sum_{i=1}^d 4\gamma^2 L \Delta \sigma_\omega^2} \leq (\sqrt{2} + \sqrt{2}\phi) \sqrt{4\gamma^2 L \Delta (K+1) d \sigma_\omega^2} \stackrel{(171)}{\leq} \frac{\Delta}{7},$$

1094 and

$$\mathbb{P}\{E_{\textcircled{3}}\} \geq 1 - \frac{\beta}{4(K+1)} \quad \text{for } E_{\textcircled{3}} = \left\{ |\textcircled{3}| > \frac{\Delta}{7} \right\}. \quad (196)$$

1095

1096

1097 **Upper bound for ④.** From E_{T-1} and the conditions on the step-size, it follows that

$$\begin{aligned} \textcircled{4} &= 2L\gamma^2 \sum_{t=0}^{T-1} \mathbb{E} \left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1} \right] \\ &\stackrel{(186)}{\leq} 2LT\gamma^2 \left(\frac{9(2^{2\alpha-1} + 1)\lambda^{2-\alpha}\sigma^\alpha}{4} + \frac{9(2^{2\alpha-1} + 1)\lambda^{2-\alpha}\zeta_\lambda^\alpha}{4} \right) \\ &\stackrel{(169)}{\leq} \frac{\Delta}{7}. \end{aligned} \quad (197)$$

1098 **Upper bound for ⑤.** First, we have

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

1099 Next, sum ⑤ has bounded with probability 1 terms:

$$\begin{aligned} \left| 2L\gamma^2 \left(\|\theta_t^u\|^2 - \mathbb{E} \left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1} \right] \mid \mathcal{F}^{t-1} \right) \right| &\leq 2L\gamma^2 \left(\|\theta_t^u\|^2 + \mathbb{E} \left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1} \right] \right) \\ &\leq 16L\gamma^2 \lambda^2 \stackrel{(173)}{\leq} \frac{\Delta}{7 \ln \frac{8(K+1)}{\beta}} := c. \end{aligned} \quad (198)$$

1100 The summands also have bounded conditional variances as shown below:

$$\tilde{\sigma}_t^2 := \mathbb{E} \left[4L^2\gamma^4 \left(\|\theta_t^u\|^2 - \mathbb{E} \left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1} \right] \right)^2 \mid \mathcal{F}^{t-1} \right] \quad (199)$$

$$\begin{aligned} \tilde{\sigma}_t^2 &\stackrel{(198)}{\leq} \frac{\Delta}{7 \ln \frac{8(K+1)}{\beta}} \mathbb{E} \left[2L\gamma^2 \left| \|\theta_t^u\|^2 - \mathbb{E} \left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1} \right] \right| \mid \mathcal{F}^{t-1} \right] \\ &\leq \frac{4L\gamma^2 \Delta}{7 \ln \frac{8(K+1)}{\beta}} \mathbb{E} \left[\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1} \right], \end{aligned} \quad (200)$$

1101 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

1102 bounded martingale difference sequence with bounded conditional variances $\{\tilde{\sigma}_t^2\}_{t=0}^{T-1}$. Next, we ap-

1103 ply 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

1104 c as in (198), $b = \frac{\Delta}{7}$, $G = \frac{\Delta^2}{294 \ln \frac{8(K+1)}{\beta}}$:

$$\mathbb{P} \left\{ |\textcircled{5}| > \frac{\Delta}{7} \quad \text{and} \quad \sum_{t=0}^{T-1} \tilde{\sigma}_t^2 \leq \frac{\Delta^2}{294 \ln \frac{8(K+1)}{\beta}} \right\} \leq 2 \exp \left(-\frac{b^2}{2G + 2cb/3} \right) = \frac{\beta}{4(K+1)}.$$

1105 Equivalently, we have

$$\mathbb{P}\{E_{\textcircled{5}}\} \geq 1 - \frac{\beta}{4(K+1)}, \quad \text{for } E_{\textcircled{4}} = \left\{ \text{either } \sum_{t=0}^{T-1} \tilde{\sigma}_t^2 > \frac{\Delta^2}{294 \ln \frac{8(K+1)}{\beta}} \quad \text{or} \quad |\textcircled{5}| \leq \frac{\Delta}{7} \right\}. \quad (201)$$

1106 In addition, E_{T-1} implies that

$$\sum_{t=0}^{T-1} \tilde{\sigma}_t^2 \leq \frac{4L\gamma^2\Delta}{7 \ln \frac{8(K+1)}{\beta}} \sum_{t=0}^{T-1} \mathbb{E} [\|\theta_t^u\|^2 \mid \mathcal{F}^{t-1}] \stackrel{(186),(169)}{\leq} \frac{\Delta^2}{294 \ln \frac{8(K+1)}{\beta}}. \quad (202)$$

1107 **Upper bound for ⑥.** From E_{T-1} , and the conditions on the step-size it follows that

$$\textcircled{6} = L\gamma^2 \sum_{t=0}^{T-1} \|\theta_t^b\|^2 \quad (203)$$

$$\begin{aligned} &\leq L\gamma^2 \left(\frac{2^{2\alpha-1} \sigma (\sigma^\alpha + \zeta_\lambda^\alpha)^{\frac{\alpha-1}{\alpha}}}{\lambda^{\alpha-1}} + (\zeta_\lambda + \lambda/2) \frac{2^{2\alpha-1} (\sigma^\alpha + \zeta_\lambda^\alpha)}{\lambda^\alpha} + \zeta_\lambda \right)^2 \\ &\stackrel{(170)}{\leq} \frac{\Delta}{7}. \end{aligned} \quad (204)$$

1108 **Upper bound for ⑦.** We have

$$\textcircled{7} = L\gamma^2 \sum_{t=0}^{T-1} \|\omega_t\|^2 = L\gamma^2 \sigma_\omega^2 \sum_{t=0}^{T-1} \sum_{i=1}^d z_{t,i}^2, \quad (205)$$

1109 where $z_{t,i} := \omega_{t,i}/\sigma_\omega$. Using Lemma A.3, we get

$$\mathbb{P} \left\{ \sum_{t=0}^{T-1} \sum_{i=1}^d z_{t,i}^2 > Td + 2\sqrt{Td \ln \frac{4(K+1)}{\beta}} + 2 \ln \frac{4(K+1)}{\beta} \right\} \leq \frac{\beta}{4(K+1)}. \quad (206)$$

1110 Since $\gamma \leq \gamma_6$, for γ_6 defined in (174)

$$\mathbb{P} \left\{ \textcircled{7} > \frac{\Delta}{7} \right\} \leq \frac{\beta}{4(K+1)}. \quad (207)$$

1111 Equivalently, we have

$$\mathbb{P}\{E_{\textcircled{7}}\} \geq 1 - \frac{\beta}{4(K+1)} \quad \text{for } E_{\textcircled{7}} = \left\{ |\textcircled{7}| \leq \frac{\Delta}{7} \right\}. \quad (208)$$

1112 Now, we have the upper bounds for ①, ②, ③, ④, ⑤, ⑥, ⑦. Thus, probability event $E_{T-1} \cap E_{\textcircled{1}} \cap$
1113 $E_{\textcircled{3}} \cap E_{\textcircled{4}} \cap E_{\textcircled{7}}$ implies

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

1114 which is equivalent to (176) and (177) for $t = T$, and

$$\begin{aligned} \mathbb{P}\{E_T\} &\geq \mathbb{P}\{E_{T-1} \cap E_{\textcircled{1}} \cap E_{\textcircled{3}} \cap E_{\textcircled{4}} \cap E_{\textcircled{7}}\} = 1 - \mathbb{P}\{\bar{E}_{T-1} \cup \bar{E}_{\textcircled{1}} \cup \bar{E}_{\textcircled{3}} \cup \bar{E}_{\textcircled{4}} \cup \bar{E}_{\textcircled{7}}\} \\ &\geq 1 - \mathbb{P}\{\bar{E}_{T-1}\} - \mathbb{P}\{\bar{E}_{\textcircled{1}}\} - \mathbb{P}\{\bar{E}_{\textcircled{3}}\} - \mathbb{P}\{\bar{E}_{\textcircled{4}}\} - \mathbb{P}\{\bar{E}_{\textcircled{7}}\} \geq 1 - \frac{(T+1)\beta}{K+1}. \end{aligned} \quad (209)$$

1116 This finishes the inductive part of our proof, i.e., for all $k = 0, 1, \dots, K$ we have $\mathbb{P}\{E_k\} \geq$
1117 $1 - (k+1)\beta/(K+1)$. In particular, for $k = K$ and with probability at least $1 - \beta$, we have

$$\frac{1}{K+1} \sum_{t=0}^K c_t \|\nabla f(x^t)\|^2 \stackrel{(179)}{\leq} \frac{4\Delta}{\gamma(K+1)},$$

1118 and $\{x^t\}_{t=0}^K \in Q$, which follows from (177). Now we have to deal with c_t . To do so, we consider
1119 two possible cases for each $t = 0, 1, \dots, K$. We either have $c_t = 1$ or $c_t = \frac{\lambda}{2\|\nabla f(x^t)\|}$. We define the

1120 corresponding sets of indices: $\mathcal{T}_1 := \{t \in \{0, 1, \dots, K\} \mid c_t = 1\}$ and $\mathcal{T}_2 := \{t \in \{0, 1, \dots, K\} \mid$
 1121 $c_t = \frac{\lambda}{2\|\nabla f(x^t)\|}\}$. Then, the above inequality can be written as

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

1122 implying

$$\frac{1}{(K+1)} \sum_{t \in \mathcal{T}_1} \|\nabla f(x^t)\|^2 \leq \frac{4\Delta}{\gamma(K+1)}, \quad (210)$$

1123 and

$$\frac{1}{K+1} \sum_{t \in \mathcal{T}_2} \|\nabla f(x^t)\| \leq \frac{8\Delta}{\lambda\gamma(K+1)}, \quad (211)$$

1124 For inequality (210), we follow the technique from (Koloskova et al., 2023) and apply inequality
 1125 $x^2 \geq 2\epsilon x - \epsilon^2$, holding for any $\epsilon, x > 0$. Taking $x = \|\nabla f(x^t)\|^2$, we get

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

1126 implying

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

1127 Upon selecting $\epsilon = \frac{2\sqrt{\Delta}}{\sqrt{\gamma(K+1)}}$, we obtain

$$\frac{1}{K+1} \sum_{t \in \mathcal{T}_1} \|\nabla f(x^t)\| \leq \sqrt{\frac{4\Delta}{\gamma(K+1)}}. \quad (212)$$

1128 Combining inequalities (210) and (211) we get:

$$\frac{1}{K+1} \sum_{t=0}^K \|\nabla f(x^t)\| \leq \sqrt{\frac{4\Delta}{\gamma(K+1)}} + \frac{8\Delta}{\lambda\gamma(K+1)}. \quad (213)$$

1129 Upon considering the best iterate, we have the following bound

$$\min_{t \in [0, K]} \|\nabla f(x^t)\|^2 \leq \frac{8\Delta}{\gamma(K+1)} + \frac{128\Delta^2}{\lambda^2\gamma^2(K+1)^2}. \quad (214)$$

1130 □

1131 Theorem F.2 states 7 values for the step-size, from which the smallest should be selected. To simplify
 1132 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
 1133 three step-sizes are redundant and can be omitted.

1134 **Corollary F.3.** *Let all conditions of Theorem F.2 hold. Furthermore, assume that K is large and*
 1135 *one selects $\lambda \leq \mathcal{O}\left(\left(\frac{K}{\ln K}\right)^{1/\alpha}\right)$, then conclusions of Theorem F.2 are valid as long as γ is selected to*
 1136 *satisfy $\gamma \leq \min\{1/4L, \gamma_1, \gamma_2, \gamma_3\}$ where we have*

$$\begin{aligned} \gamma_1 &:= \frac{\sqrt{\Delta}}{21\sqrt{L}(2^{2\alpha-1} + 1)^{1/2}\sigma^{\alpha/2}\lambda^{1-\alpha/2}\sqrt{6(K+1)\ln\frac{8(K+1)}{\beta}\left(1 + \frac{\zeta_\lambda^\alpha}{\sigma^\alpha}\right)}}, \\ \gamma_2 &:= \frac{\sqrt{\Delta}\lambda^{\alpha-1}}{14\sqrt{L}(K+1)2^{2\alpha-1}(\sigma^\alpha + \zeta_\lambda^\alpha)\left(\frac{\zeta_\lambda}{\lambda} + \frac{1}{2} + \frac{\lambda^{\alpha-1}\zeta_\lambda}{2^{2\alpha-1}(\sigma^\alpha + \zeta_\lambda^\alpha)} + \left(1 + \frac{\zeta_\lambda^\alpha}{\sigma^\alpha}\right)^{-1/\alpha}\right)}, \\ \gamma_3 &:= \frac{\sqrt{\Delta}}{14\sqrt{L}\sigma_\omega\sqrt{d(K+1)}(\sqrt{2} + \sqrt{2}\phi)}. \end{aligned}$$

1137 *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 (174)
 1138 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. Further-
 1139 more, γ_5 in (173) 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
 1140 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,
 1141 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
 1142 in (172), 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
 1143 $\mathcal{O}\left(\sigma_\omega \epsilon / \sqrt{K \ln(K/\delta)}\right)$. Therefore, γ_4 decreases by the order $\mathcal{O}\left(\sigma_\omega \epsilon \sqrt{K \ln(K/\delta)}\right)$, which is the same
 1144 order as γ_3 . Hence, γ_4 can be omitted, and the proof is complete. \square

1145 G Rate and Neighborhood for Clipped-SGD: Non-Convex Case

1146 Now that we have established the convergence properties of **DP-Clipped-SGD** for non-convex
 1147 problems, we turn to evaluating its convergence rate. This rate depends critically on the choice
 1148 of the step-size γ , and in general, the resulting expressions can be quite complex. To obtain more
 1149 interpretable bounds, we consider simplified rate expressions by analyzing separate cases based on
 1150 different ranges of λ . Since we focus on the asymptotic behavior, numerical constants are omitted for
 1151 clarity.

1152 In this section, we consider the cases without the DP noise ($\sigma_\omega = 0$) and investigate all possible
 1153 clipping levels.

1154 **Case 1:** $\lambda > 4\sqrt{L\Delta}$. In this case, $\zeta_\lambda = 0$, and the step-size conditions reduce to the following:

$$\gamma \leq \mathcal{O} \left(\min \left\{ \frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\sigma^{\alpha/2} \lambda^{1-\alpha/2} \sqrt{K \ln \frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}} \lambda^{\alpha-1}}{K \sigma^\alpha} \right\} \right). \quad (215)$$

1155 In particular, when γ equals the minimum from the above condition, the iterates produced by
 1156 **Clipped-SGD** after K iterations with probability at least $1 - \beta$ satisfy

$$\min_{t \in [0, K]} \|\nabla f(x^t)\|^2 = \mathcal{O}(\max\{(217), (218), (219)\}), \quad (216)$$

1157 where

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

$$\frac{\sqrt{L\Delta} \sigma^\alpha}{\lambda^{\alpha-1}} + \frac{L\Delta \sigma^{2\alpha}}{\lambda^{2\alpha}}, \quad (218)$$

$$\frac{L\Delta}{K} + \frac{L^2 \Delta^2}{\lambda^2 K^2}. \quad (219)$$

1158 We clearly see that the dominant term (217) is an increasing function of λ , and the dominant term
 1159 in (218) is a decreasing function. Solving for the optimal λ where the leading terms in (217) and
 1160 (218) 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
 1161 probability at least $1 - \beta$

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

1162 where

$$\sqrt{L\Delta} \sigma \left(\frac{\ln \frac{K}{\beta}}{K} \right)^{\frac{\alpha-1}{\alpha}} + \frac{L\Delta \ln^2 K/\beta}{K^2}, \quad (221)$$

$$\frac{L\Delta}{K} + \frac{L^2 \Delta^2 \left(\ln \frac{K}{\beta} \right)^{\frac{2}{\alpha}}}{\sigma^2 K^{\frac{2\alpha+2}{\alpha}}}. \quad (222)$$

1163 Note in this case, we converge to the exact optimum, and the dominant term matches (Sadiev et al.,
 1164 2023). As it can be seen from (217), (218), when the clipping level is not that large, we converge to a
 1165 neighborhood of the solution, but with a faster rate.

1166 When $\lambda \leq 4\sqrt{L\Delta}$, we have $\zeta_\lambda = \frac{4\sqrt{L\Delta}-\lambda}{2}$. As observed from (169), (170), we also have to consider
 1167 the relation between λ and σ in these cases. Thus, we split the $\lambda \leq 4\sqrt{L\Delta}$ case into 6 different
 1168 regimes to cover all possible cases.

1169 **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
 1170 following:

$$\gamma \leq \mathcal{O} \left(\min \left\{ \frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\sigma^{\alpha/2} \lambda^{1-\alpha/2} \sqrt{K \ln \frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}} \lambda^{\alpha-1}}{K \sigma^\alpha} \right\} \right). \quad (223)$$

1171 As it can be seen, the bounds on step-size are similar to Case 1. However, the optimal λ derived in
 1172 the previous section violates the constraint that $\lambda \leq 4\sqrt{L\Delta}$. Subsequently, the optimal λ becomes
 1173 $\lambda = 4\sqrt{L\Delta}$. For this choice of λ , we have that with probability at least $1 - \beta$

$$\min_{t \in [0, K]} \|\nabla f(x^t)\|^2 = \mathcal{O}(\max\{(225), (226), (227)\}), \quad (224)$$

1174 where

$$\sqrt{(L\Delta)^{\frac{4-\alpha}{2}} \sigma^\alpha \frac{\ln K/\beta}{K}} + \frac{(L\Delta)^{\frac{2-\alpha}{2}} \sigma^\alpha \ln K/\beta}{K}, \quad (225)$$

$$\frac{\sigma^\alpha}{(\sqrt{L\Delta})^{\alpha-2}} + \frac{\sigma^{2\alpha}}{(L\Delta)^{\alpha-1}}, \quad (226)$$

$$\frac{L\Delta}{K} + \frac{L\Delta}{K^2}. \quad (227)$$

1175 **Case 3:** $\frac{4}{3}\sqrt{L\Delta} < \lambda \leq 4\sqrt{L\Delta}$, $\zeta_\lambda < \sigma < \lambda$. In this case, the step-size conditions reduce to

$$\gamma \leq \mathcal{O} \left(\min \left\{ \frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\sigma^{\alpha/2} \lambda^{1-\alpha/2} \sqrt{K \ln \frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}} \lambda^{\alpha-1}}{K \max\{\sigma^\alpha, \lambda^{\alpha-1} \zeta_\lambda\}} \right\} \right). \quad (228)$$

1176 If $\max\{\sigma^\alpha, \lambda^{\alpha-1} \zeta_\lambda\} = \sigma^\alpha$, then the resulting bounds are similar to the previous case. If
 1177 $\max\{\sigma^\alpha, \lambda^{\alpha-1} \zeta_\lambda\} = \lambda^{\alpha-1} \zeta_\lambda$ is satisfied, $\min_{t \in [0, K]} \|\nabla f(x^t)\|^2$ is bounded with probability at
 1178 least $1 - \beta$ by the maximum of the following terms:

$$\sqrt{L\Delta} \lambda^{1-\alpha/2} \sigma^{\alpha/2} \sqrt{\frac{\ln K/\beta}{K}} + \frac{L\Delta \sigma^\alpha \ln K/\beta}{\lambda^\alpha K}, \quad (229)$$

$$\sqrt{L\Delta} \zeta_\lambda + \frac{L\Delta \zeta_\lambda^2}{\lambda^2}, \quad (230)$$

$$\frac{L\Delta}{K} + \frac{L^2 \Delta^2}{\lambda^2 K^2}. \quad (231)$$

1179 In the latter case (i.e., maximum occurring in the second argument), the optimal λ is $4\sqrt{L\Delta} -$
 1180 η , where η is a sufficiently small number such that $\lambda^{\alpha-1} \zeta_\lambda \geq \sigma^\alpha$, i.e., λ satisfies $\zeta_\lambda =$
 1181 $\max \left\{ \frac{\sigma^\alpha}{\lambda^{\alpha-1}}, \lambda^{1-\alpha/2} \sigma^{\alpha/2} \sqrt{\frac{\ln K/\beta}{K}} \right\}$. Note that the (230) is decreasing in λ , and $\lambda = 4\sqrt{L\Delta}$ is
 1182 not feasible. With this choice of λ , we get:

$$\min_{t \in [0, K]} \|\nabla f(x^t)\|^2 = \mathcal{O}(\max\{(233), (234), (235)\}), \quad (232)$$

1183 where

$$\sqrt{L\Delta(4\sqrt{L\Delta} - \eta)^{2-\alpha} \sigma^\alpha \frac{\ln K/\beta}{K}} + \frac{L\Delta \sigma^\alpha \ln K/\beta}{(\sqrt{L\Delta} - \eta)^\alpha K}, \quad (233)$$

$$\frac{\sqrt{L\Delta} \eta}{2} + \frac{L\Delta \eta^2}{(4\sqrt{L\Delta} - \eta)^2}, \quad (234)$$

$$\frac{L\Delta}{K} + \frac{L^2 \Delta^2}{(4\sqrt{L\Delta} - \eta)^2 K^2}. \quad (235)$$

1184 **Case 4:** $\frac{4}{3}\sqrt{L\Delta} < \lambda \leq 4\sqrt{L\Delta}$, $\sigma < \zeta_\lambda < \lambda$. For this case, step-size conditions reduce to

$$\gamma \leq \mathcal{O} \left(\min \left\{ \frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\zeta_\lambda^{\alpha/2} \lambda^{1-\alpha/2} \sqrt{K \ln \frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}} \lambda^{\alpha-1}}{K(\lambda^{\alpha-1} \zeta_\lambda)} \right\} \right), \quad (236)$$

1185 and $\min_{t \in [0, K]} \|\nabla f(x^t)\|^2$ is bounded with probability at least $1 - \beta$ by the maximum of the following
1186 terms

$$\sqrt{L\Delta} \lambda^{1-\alpha/2} \zeta_\lambda^{\alpha/2} \sqrt{\frac{\ln K/\beta}{K}} + \frac{L\Delta \zeta_\lambda^\alpha \ln K/\beta}{\lambda^\alpha K}, \quad (237)$$

$$\sqrt{L\Delta} \zeta_\lambda + \frac{L\Delta \zeta_\lambda^2}{\lambda^2}, \quad (238)$$

$$\frac{L\Delta}{K} + \frac{L^2 \Delta^2}{\lambda^2 K^2}. \quad (239)$$

1187 The optimal λ in this case is $\lambda = 4\sqrt{L\Delta} - 2\sigma$, and we have that with probability at least $1 - \beta$

$$\min_{t \in [0, K]} \|\nabla f(x^t)\|^2 = \mathcal{O}(\max\{(241), (242), (243)\}), \quad (240)$$

1188 where

$$\sqrt{L\Delta(4\sqrt{L\Delta} - 2\sigma)^{2-\alpha} \sigma^\alpha \frac{\ln K/\beta}{K}} + \frac{L\Delta \sigma^\alpha \ln K/\beta}{(4\sqrt{L\Delta} - 2\sigma)^\alpha K}, \quad (241)$$

$$\sqrt{L\Delta} \sigma + \frac{L\Delta \sigma^2}{(4\sqrt{L\Delta} - 2\sigma)^2}, \quad (242)$$

$$\frac{L\Delta}{K} + \frac{L^2 \Delta^2}{(4\sqrt{L\Delta} - 2\sigma)^2 K^2}. \quad (243)$$

1189 **Case 5:** $\lambda \leq \frac{4}{3}\sqrt{L\Delta}$, $\lambda < \zeta_\lambda < \sigma$. In this case, the step-size conditions reduce to

$$\gamma \leq \mathcal{O} \left(\min \left\{ \frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\sigma^{\alpha/2} \lambda^{1-\alpha/2} \sqrt{K \ln \frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}} \lambda^\alpha}{K(\sigma^\alpha \zeta_\lambda)} \right\} \right), \quad (244)$$

1190 and $\min_{t \in [0, K]} \|\nabla f(x^t)\|^2$ is bounded with probability at least $1 - \beta$ by the maximum of the following
1191 terms

$$\sqrt{L\Delta} \lambda^{1-\alpha/2} \sigma^{\alpha/2} \sqrt{\frac{\ln K/\beta}{K}} + \frac{L\Delta \sigma^\alpha \ln K/\beta}{\lambda^\alpha K}, \quad (245)$$

$$\sqrt{L\Delta} \frac{\sigma^\alpha \zeta_\lambda}{\lambda^\alpha} + \frac{L\Delta \sigma^{2\alpha} \zeta_\lambda^2}{\lambda^{2\alpha+2}}, \quad (246)$$

$$\frac{L\Delta}{K} + \frac{L^2 \Delta^2}{\lambda^2 K^2}. \quad (247)$$

1192 In this regime, the optimal $\lambda = \frac{4}{3}\sqrt{L\Delta}$. With this choice of λ , we get with probability at least $1 - \beta$

$$\min_{t \in [0, K]} \|\nabla f(x^t)\|^2 = \mathcal{O}(\max\{(249), (250), (251)\}), \quad (248)$$

1193 where

$$\sqrt{(L\Delta)^{\frac{4-\alpha}{2}} \sigma^\alpha \frac{\ln K/\beta}{K}} + \frac{(L\Delta)^{\frac{2-\alpha}{2}} \sigma^\alpha \ln K/\beta}{K}, \quad (249)$$

$$\frac{\sigma^\alpha}{(\sqrt{L\Delta})^{\alpha-2}} + \frac{\sigma^{2\alpha}}{(L\Delta)^{\alpha-1}}, \quad (250)$$

$$\frac{L\Delta}{K} + \frac{L\Delta}{K^2}. \quad (251)$$

1194 **Case 6:** $\lambda \leq \frac{4}{3}\sqrt{L\Delta}$, $\lambda < \sigma < \zeta_\lambda$. In this case, the step-size conditions reduce to

$$\gamma \leq \mathcal{O} \left(\min \left\{ \frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\zeta_\lambda^{\alpha/2} \lambda^{1-\alpha/2} \sqrt{K \ln \frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}} \lambda^\alpha}{K(\zeta_\lambda^{\alpha+1})} \right\} \right), \quad (252)$$

1195 and $\min_{t \in [0, K]} \|\nabla f(x^t)\|^2$ is bounded with probability at least $1 - \beta$ by the maximum of the following
1196 terms

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

$$\frac{\sqrt{L\Delta} \zeta_\lambda^{\alpha+1}}{\lambda^\alpha} + \frac{L\Delta \zeta_\lambda^{2\alpha}}{\lambda^{2\alpha+2}}, \quad (254)$$

$$\frac{L\Delta}{K} + \frac{L^2 \Delta^2}{\lambda^2 K^2}. \quad (255)$$

1197 Next, we find the optimal λ via equalizing the leading terms (the first ones) in (253) and (254). This

1198 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

1199 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
1200 probability at least $1 - \beta$

$$\min_{t \in [0, K]} \|\nabla f(x^t)\|^2 = \mathcal{O}(\max\{(257), (258), (259)\}), \quad (256)$$

1201 where

$$(\sqrt{L\Delta} - \eta)^{1-\alpha/2} (\sqrt{L\Delta} + \eta)^{\alpha/2} \sqrt{L\Delta \frac{\ln K/\beta}{K}} + \frac{L\Delta (\sqrt{L\Delta} + \eta)^\alpha \ln K/\beta}{(\sqrt{L\Delta} - \eta)^\alpha K}, \quad (257)$$

$$\frac{\sqrt{L\Delta} (\sqrt{L\Delta} + \eta)^{\alpha+1}}{(\sqrt{L\Delta} - \eta)^\alpha} + \frac{L\Delta (\sqrt{L\Delta} + \eta)^{2\alpha}}{(\sqrt{L\Delta} - \eta)^{2\alpha+2}}, \quad (258)$$

$$\frac{L\Delta}{K} + \frac{L^2 \Delta^2}{(\sqrt{L\Delta} - \eta)^2 K^2}. \quad (259)$$

1202 **Case 7:** $\lambda \leq \frac{4}{3}\sqrt{L\Delta}$, $\sigma < \lambda < \zeta_\lambda$. In this case, the step-size conditions reduce to

$$\gamma \leq \mathcal{O} \left(\min \left\{ \frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\zeta_\lambda^{\alpha/2} \lambda^{1-\alpha/2} \sqrt{K \ln \frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}} \lambda^{\alpha-1}}{K \max\left\{ \frac{\zeta_\lambda^{\alpha+1}}{\lambda}, \zeta_\lambda^{\alpha-1} \sigma \right\}} \right\} \right). \quad (260)$$

1203 We note that $\max\left\{ \frac{\zeta_\lambda^{\alpha+1}}{\lambda}, \zeta_\lambda^{\alpha-1} \sigma \right\} = \zeta_\lambda^\alpha \max\left\{ \frac{\zeta_\lambda}{\lambda}, \frac{\sigma}{\lambda} \right\} = \frac{\zeta_\lambda^{\alpha+1}}{\lambda}$ since $\sigma < \lambda < \zeta_\lambda$. Therefore,
1204 similarly to the previous case, we have

$$\gamma \leq \mathcal{O} \left(\min \left\{ \frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\zeta_\lambda^{\alpha/2} \lambda^{1-\alpha/2} \sqrt{K \ln \frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}} \lambda^\alpha}{K \zeta_\lambda^{\alpha+1}} \right\} \right), \quad (261)$$

1205 and $\min_{t \in [0, K]} \|\nabla f(x^t)\|^2$ is bounded with probability at least $1 - \beta$ by the maximum of the following
1206 terms

$$\sqrt{L\Delta} \lambda^{1-\alpha/2} \zeta_\lambda^{\alpha/2} \sqrt{\frac{\ln K/\beta}{K}} + \frac{L\Delta \zeta_\lambda^\alpha \ln K/\beta}{\lambda^\alpha K}, \quad (262)$$

$$\frac{\sqrt{L\Delta} \zeta_\lambda^{\alpha+1}}{\lambda^\alpha} + \frac{L\Delta \zeta_\lambda^{2\alpha}}{\lambda^{2\alpha+2}}, \quad (263)$$

$$\frac{L\Delta}{K} + \frac{L^2 \Delta^2}{\lambda^2 K^2}. \quad (264)$$

1207 The optimal λ equals $\frac{4}{3}\sqrt{L\Delta}$. This happens because both leading terms in (262) and (263) are
 1208 decreasing in λ . With this choice, we get with probability at least $1 - \beta$

$$\min_{t \in [0, K]} \|\nabla f(x^t)\|^2 = \mathcal{O}(\max\{(266), (267), (268)\}), \quad (265)$$

1209 where

$$\sqrt{L\Delta \frac{\ln K/\beta}{K}} + \frac{L\Delta \ln K/\beta}{K}, \quad (266)$$

$$\sqrt{L\Delta}\sigma + \frac{\sigma^2}{L\Delta}, \quad (267)$$

$$\frac{L\Delta}{K} + \frac{L\Delta}{K^2}. \quad (268)$$

1210 Now that we have covered all possible regions, it's time to consider the DP noise as well.

1211 H Rate and Neighborhood for DP-Clipped-SGD: Non-Convex Case

1212 To ensure the output of the algorithm is (ε, δ) -differentially private in this setting, expectation
 1213 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
 1214 advanced composition theorem of [Dwork et al. \(2014\)](#). In the finite sum case, one can reduce the
 1215 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
 1216 brevity, in the DP case, we only consider two cases: large λ and relatively small λ regimes. The other
 1217 cases can be derived with a similar analysis.

1218 **Case 1:** $\lambda > 4\sqrt{L\Delta}$. In this case, $\zeta_\lambda = 0$, and the step-size conditions reduce to the following:

$$\gamma \leq \mathcal{O}\left(\min\left\{\frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\sigma^{\alpha/2} \lambda^{1-\alpha/2} \sqrt{K \ln \frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}} \lambda^{\alpha-1}}{K \sigma^\alpha}, \frac{\sqrt{\frac{\Delta}{L}}}{\sigma_\omega \sqrt{dK \ln \frac{K}{\beta}}}\right\}\right) \quad (269)$$

1219 In particular, when γ equals the minimum from the step-size condition, then the iterates produced by
 1220 [DP-Clipped-SGD](#) after K iterations with probability at least $1 - \beta$ satisfy

$$\min_{k \in [0, K]} \|\nabla f(x^k)\|^2 = \mathcal{O}(\max\{(271), (272), (273), (274)\}) \quad (270)$$

1221 where

$$\sqrt{L\Delta} \lambda^{1-\alpha/2} \sigma^{\alpha/2} \sqrt{\frac{\ln K/\beta}{K}} + \frac{L\Delta \sigma^\alpha \ln K/\beta}{\lambda^\alpha K} \quad (271)$$

$$\frac{\sqrt{L\Delta} \sigma^\alpha}{\lambda^{\alpha-1}} + \frac{L\Delta \sigma^{2\alpha}}{\lambda^{2\alpha}} \quad (272)$$

$$\frac{L\Delta}{K} + \frac{L^2 \Delta^2}{\lambda^2 K^2} \quad (273)$$

$$\sqrt{L\Delta} \sigma_\omega \sqrt{\frac{d \ln \frac{K}{\beta}}{K}} + \frac{L\Delta \sigma_\omega^2 d \ln \frac{K}{\beta}}{\lambda^2 K}. \quad (274)$$

1222 Here, (272) accounts for the bias caused by clipping, and (274) accounts for the accumula-
 1223 tion of DP noise. These terms are decreasing and increasing in λ respectively, if we use

1224 $\sigma_\omega = \Theta\left(\frac{\lambda}{\varepsilon} \sqrt{K \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{1}{\delta}\right)}\right)$. To find the optimal λ , we find the equilibrium of these two

1225 terms. Solving the equilibrium equation, we get $\lambda = \mathcal{O}\left(\left(\frac{\varepsilon \sigma^\alpha}{d \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{K}{\beta}\right)}\right)^{\frac{1}{\alpha}}\right)$. Unless $\varepsilon \sigma^\alpha$ is

1226 large enough, this value violates the constraint that $\lambda > 4\sqrt{L\Delta}$, and it is not feasible. Thus, we have
 1227 the following formula for the optimal λ :

$$\lambda = \max\left\{4\sqrt{L\Delta}, \left(\frac{\varepsilon \sigma^\alpha}{d \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{K}{\beta}\right)}\right)^{\frac{1}{\alpha}}\right\}. \quad (275)$$

1228 For this choice of λ , we get that with probability at least $1 - \beta$

$$\min_{k \in [0, K]} \|\nabla f(x^k)\|^2 = \mathcal{O}(\max\{(277), (278), (279), (280)\}) \quad (276)$$

1229 with

$$\max \left\{ \sqrt{(L\Delta)^{\frac{4-\alpha}{2}} \sigma^\alpha \frac{\ln K/\beta}{K}}, \sqrt{L\Delta} \left(\frac{\varepsilon \sigma^\alpha}{\sqrt{d \ln \left(\frac{1}{\delta}\right) \ln \left(\frac{K}{\delta}\right)}} \right)^{\frac{1}{\alpha}} \sqrt{\frac{\ln^{\frac{3\alpha-2}{2\alpha}} \frac{K}{\beta}}{K}} \right\} \quad (277)$$

$$\min \left\{ \frac{\sigma^\alpha}{(\sqrt{L\Delta})^{\alpha-2}}, \sqrt{L\Delta} \sigma \left(\frac{\sqrt{d \ln \left(\frac{1}{\delta}\right) \ln \left(\frac{K}{\delta}\right) \ln \left(\frac{K}{\beta}\right)}}{\varepsilon} \right)^{\frac{\alpha-1}{\alpha}} \right\} \quad (278)$$

$$\min \left\{ \frac{L\Delta}{K^2}, \frac{L^2 \Delta^2 (d \ln \left(\frac{1}{\delta}\right) \ln \left(\frac{K}{\delta}\right))^{\frac{1}{\alpha}} \ln^{\frac{1}{\alpha}} \frac{K}{\beta}}{(\varepsilon)^{\frac{1}{\alpha}} \sigma} \frac{1}{K^2} \right\} + \frac{L\Delta}{K} \quad (279)$$

$$\max \left\{ \frac{L\Delta}{\varepsilon} \sqrt{d \ln \left(\frac{1}{\delta}\right) \ln \left(\frac{K}{\delta}\right) \ln \left(\frac{K}{\beta}\right)}, \frac{R\sigma \left(d \ln \left(\frac{1}{\delta}\right) \ln \left(\frac{K}{\delta}\right) \ln \left(\frac{K}{\beta}\right) \right)^{\frac{\alpha+2}{2\alpha}}}{\varepsilon^{\frac{\alpha-1}{\alpha}}} \right\} \\ + \frac{L\Delta}{\varepsilon^2} d \ln \left(\frac{1}{\delta}\right) \ln \left(\frac{K}{\delta}\right) \ln \left(\frac{K}{\beta}\right), \quad (280)$$

1230 where, for the sake of brevity, we only report the dominant terms.

1231 **Case 2:** $\lambda \leq \frac{4}{3}\sqrt{L\Delta}$ $\lambda < \sigma < \zeta_\lambda$. In this case, the step-size conditions reduce to the following:

$$\gamma \leq \mathcal{O} \left(\min \left\{ \frac{1}{L}, \frac{\sqrt{\frac{\Delta}{L}}}{\zeta_\lambda^{\alpha/2} \lambda^{1-\alpha/2} \sqrt{K \ln \frac{K}{\beta}}}, \frac{\sqrt{\frac{\Delta}{L}} \lambda^\alpha}{K(\zeta_\lambda^{\alpha+1})}, \frac{\sqrt{\frac{\Delta}{L}}}{\sigma_\omega \sqrt{dK \ln \frac{K}{\beta}}} \right\} \right). \quad (281)$$

1232 Taking γ equal to the right-hand side, we get that with probability at least $1 - \beta$

$$\min_{t \in [0, K]} \|\nabla f(x^t)\|^2 = \mathcal{O}(\{(283), (284), (285), (286)\}) \quad (282)$$

1233 with

$$\sqrt{L\Delta} \lambda^{1-\alpha/2} \sigma^{\alpha/2} \sqrt{\frac{\ln K/\beta}{K}} + \frac{L\Delta \sigma^\alpha \ln K/\beta}{\lambda^\alpha K} \quad (283)$$

$$\frac{\sqrt{L\Delta} \zeta_\lambda^{\alpha+1}}{\lambda^\alpha} + \frac{L\Delta \zeta_\lambda^{2\alpha}}{\lambda^{2\alpha+2}} \quad (284)$$

$$\frac{L\Delta}{K} + \frac{L^2 \Delta^2}{\lambda^2 K^2} \quad (285)$$

$$\sqrt{L\Delta} \sigma_\omega \sqrt{\frac{d \ln \frac{K}{\beta}}{K}} + \frac{L\Delta \sigma_\omega^2 d \ln \frac{K}{\beta}}{\lambda^2 K}. \quad (286)$$

1234 Similarly to the previous case, we find the optimal λ as the equilibrium of the leading terms in (284)
1235 and (286). By doing so, we get the following optimal λ :

$$\lambda = \min \left\{ \frac{4}{3} \sqrt{L\Delta}, \frac{2\varepsilon \sqrt{L\Delta}}{\left(d \ln \left(\frac{1}{\delta}\right) \ln \left(\frac{K}{\delta}\right) \ln \left(\frac{K}{\beta}\right) \right)^{\frac{1}{2\alpha+2}} + 1} \right\} \quad (287)$$

1236 For this choice of λ , we get that with probability at least $1 - \beta$

$$\min_{k \in [0, K]} \|\nabla f(x^t)\|^2 = \mathcal{O}(\max \{(289), (290), (291), (292)\}) \quad (288)$$

1237 with

$$\min \left\{ \sqrt{(L\Delta)^{\frac{4-\alpha}{2}} \sigma^\alpha \frac{\ln K/\beta}{K}}, \sqrt{\frac{(L\Delta)^{\frac{4-\alpha}{2}} \varepsilon^{2-\alpha} \ln^{\frac{3\alpha}{4\alpha+4}} \frac{K}{\beta}}{(d \ln(\frac{1}{\delta}) \ln(\frac{K}{\delta}))^{\frac{2-\alpha}{4\alpha+4}} K}} \right\} \quad (289)$$

$$\max \left\{ \frac{\sigma^\alpha}{\sqrt{L\Delta}^{\alpha-2}}, \frac{(\sqrt{L\Delta})^{2-\alpha} \sigma^\alpha}{\varepsilon} \left(d \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{K}{\beta}\right) \right)^{\frac{\alpha-1}{2\alpha+2}} \right\} \quad (290)$$

$$\max \left\{ \frac{L\Delta}{K^2}, \frac{L\Delta}{\varepsilon^2 K^2} \left(\left(d \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{K}{\beta}\right) \right)^{\frac{1}{2\alpha+2}} + 1 \right)^2 \right\} + \frac{L\Delta}{K} \quad (291)$$

$$\min \left\{ \frac{L\Delta}{\varepsilon} \sqrt{d \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{K}{\beta}\right)}, \frac{L\Delta \sqrt{\ln \frac{K}{\beta}}}{\left(d \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{K}{\beta}\right) \right)^{\frac{1}{2\alpha+2}} + 1} \right\} \\ + \frac{L\Delta d}{\varepsilon^2} d \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{K}{\delta}\right) \ln\left(\frac{K}{\beta}\right), \quad (292)$$

1238 where, for the sake of brevity, we only report the dominant terms.

1239 NeurIPS Paper Checklist

1240 1. Claims

1241 Question: Do the main claims made in the abstract and introduction accurately reflect the
1242 paper’s contributions and scope?

1243 Answer: [\[Yes\]](#)

1244 Justification: As mentioned in the abstract, this work provides the first high-probability
1245 analysis for Clipped SGD with heavy-tailed noise on the gradient and an arbitrary clipping
1246 level with added DP noise. This is the main contribution of the paper and it appears in the
1247 abstract.

1248 Guidelines:

- 1249 • The answer NA means that the abstract and introduction do not include the claims
1250 made in the paper.
- 1251 • The abstract and/or introduction should clearly state the claims made, including the
1252 contributions made in the paper and important assumptions and limitations. A No or
1253 NA answer to this question will not be perceived well by the reviewers.
- 1254 • The claims made should match theoretical and experimental results, and reflect how
1255 much the results can be expected to generalize to other settings.
- 1256 • It is fine to include aspirational goals as motivation as long as it is clear that these goals
1257 are not attained by the paper.

1258 2. Limitations

1259 Question: Does the paper discuss the limitations of the work performed by the authors?

1260 Answer: [\[Yes\]](#)

1261 Justification: We have explained the limitations of our analysis in Section 4.

1262 Guidelines:

- 1263 • The answer NA means that the paper has no limitation while the answer No means that
1264 the paper has limitations, but those are not discussed in the paper.
- 1265 • The authors are encouraged to create a separate "Limitations" section in their paper.
- 1266 • The paper should point out any strong assumptions and how robust the results are to
1267 violations of these assumptions (e.g., independence assumptions, noiseless settings,
1268 model well-specification, asymptotic approximations only holding locally). The authors
1269 should reflect on how these assumptions might be violated in practice and what the
1270 implications would be.
- 1271 • The authors should reflect on the scope of the claims made, e.g., if the approach was
1272 only tested on a few datasets or with a few runs. In general, empirical results often
1273 depend on implicit assumptions, which should be articulated.
- 1274 • The authors should reflect on the factors that influence the performance of the approach.
1275 For example, a facial recognition algorithm may perform poorly when image resolution
1276 is low or images are taken in low lighting. Or a speech-to-text system might not be
1277 used reliably to provide closed captions for online lectures because it fails to handle
1278 technical jargon.
- 1279 • The authors should discuss the computational efficiency of the proposed algorithms
1280 and how they scale with dataset size.
- 1281 • If applicable, the authors should discuss possible limitations of their approach to
1282 address problems of privacy and fairness.
- 1283 • While the authors might fear that complete honesty about limitations might be used by
1284 reviewers as grounds for rejection, a worse outcome might be that reviewers discover
1285 limitations that aren’t acknowledged in the paper. The authors should use their best
1286 judgment and recognize that individual actions in favor of transparency play an impor-
1287 tant role in developing norms that preserve the integrity of the community. Reviewers
1288 will be specifically instructed to not penalize honesty concerning limitations.

1289 3. Theory assumptions and proofs

1290 Question: For each theoretical result, does the paper provide the full set of assumptions and
1291 a complete (and correct) proof?

Answer: [Yes]

Justification: Main assumptions are stated in Section 2. Complete correct proofs are provided in the appendices. A proof sketch is provided in the main text in Section 4.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: Only rigorous mathematical analysis is provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: Only rigorous mathematical analysis is provided.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: Only rigorous mathematical analysis is provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: Only rigorous mathematical analysis is provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer “Yes” if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: Only rigorous mathematical analysis is provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The work completely conforms to the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The work investigates the incorporation of differential privacy guarantees in stochastic optimization. Hence, it offers a positive societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No component with potential detrimental effects is released.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: There is no code, data, or models that require licenses.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new asset is released.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: There was no crowd-sourcing experiments or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: There was no crowd-sourcing experiments or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

1551 Question: Does the paper describe the usage of LLMs if it is an important, original, or
1552 non-standard component of the core methods in this research? Note that if the LLM is used
1553 only for writing, editing, or formatting purposes and does not impact the core methodology,
1554 scientific rigorousness, or originality of the research, declaration is not required.

1555 Answer: [NA]

1556 Justification: The paper does not use LLMs as an important, original, or non-standard
1557 component of the core methods in this research.

1558 Guidelines:

- 1559 • The answer NA means that the core method development in this research does not
1560 involve LLMs as any important, original, or non-standard components.
- 1561 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
1562 for what should or should not be described.