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

Anonymous Author(s)
Affiliation
Address
email

Abstract

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

1 Introduction

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

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

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

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

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

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

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

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

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

2 Technical Preliminaries

The optimization problem considered in this work has the following form

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

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

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

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

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

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

This type of convergence is generally considered superior to in-expectation guarantees (e.g., $\mathbb{E}[\mathcal{P}(x^K)] \le \epsilon$), 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.

$$\|\nabla f(x) - \nabla f(y)\| \le L\,\|x-y\|\,. \tag{3}$$

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 $x, y \in Q$

$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle. \tag{4}$$

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 $x \in Q$

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

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

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 $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ satisfies (ε, δ)-Differential Privacy, if for any adjacent $D, D' \in \mathcal{D}$ and for any $\mathcal{S} \subseteq \mathcal{R}$

$$\mathbb{P}\left(\mathcal{M}(\mathcal{D}) \in \mathcal{S}\right) \le e^{\varepsilon}\,\mathbb{P}\left(\mathcal{M}(\mathcal{D}') \in \mathcal{S}\right) + \delta, \tag{7}$$

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).

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

3 Related Work

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

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

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

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 stepsizes, while Chezhegov et al. (2024) obtained analogous bounds for various versions of Clipped-AdaGrad and Clipped-Adam with both scalar and coordinate-wise stepsizes. 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 this order, 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

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

Algorithm 1 DP-Clipped-SGD

Input: starting point x^0, number of iterations K, stepsize $\gamma > 0$, clipping level λ.

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

DP noise is included in the convex case, while the result for DP case in the non-convex setup is deffered 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 $Q = B_{2R}(x^\star)$, $R \geq \|x^0 - x^\star\|$. Set $\zeta_\lambda := \max\left\{0, 2LR - \frac{\lambda}{2}\right\}$, and further assume that the step-size γ is selected to satisfy*

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

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

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

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 stepsize 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 stepsize 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*

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

where

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

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

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

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

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

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

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

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

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

where

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

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

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

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

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

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

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

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

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.

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

where

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

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

5 Conclusion

In this paper, we present the first high-probability convergence analysis of DP-Clipped-SGD for both convex and non-convex smooth optimization problems under heavy-tailed noise. Our results demonstrate that DP-Clipped-SGD converges to a certain neighborhood of the optimal solution at a rate of $\mathcal{O}(1/\sqrt{K})$. In future work, it would be valuable to extend these results to the Federated 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, and 35)

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, and 6)

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 22)

Freedman, D. A. (1975). On tail probabilities for martingales. *the Annals of Probability*, pages 100–118. (Cited on page 22)

Gasnikov, A. and Nesterov, Y. (2016). Universal fast gradient method for stochastic composit 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)

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

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

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

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

Juditsky, A. and Nemirovski, A. S. (2008). Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*. (Cited on page 22)

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

Koloskova, A., Hendrikx, H., and Stich, S. U. (2023). Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, pages 17343–17363. PMLR. (Cited on pages 30 and 43)

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

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

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pages 1302–1338. (Cited on page 23)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Polyanskiy, Y. and Wu, Y. (2025). *Information theory: From coding to learning*. Cambridge university press. (Cited on page 23)

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

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

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

Sadiev, A., Danilova, M., Gorbunov, E., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A., and Richtárik, P. (2023). High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *International Conference on Machine Learning*, pages 29563–29648. PMLR. (Cited on pages 1, 3, 4, 5, 6, 9, 20, and 31)

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

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

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

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

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

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

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

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

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

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, hyperparameters, 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**

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

Answer: [NA]

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

Guidelines:

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