Unified Differentially Private Stochastic Gradient Descent with Tail-Aware Discriminative Clipping

Haichao Sha[1], Yuncheng Wu[1], Yang Cao[2], Yong Liu[1], Yuhan Liu[1], Ruixuan Liu[3], Hong Chen[1]
[1]Renmin University of China, [2]Tokyo Institute of Technology, [3]Emory University
{sha,wuyuncheng,liuyonggsai,liuyh2019,chong}@ruc.edu.cn, cao@c.titech.ac.jp, ruixuan.liu2@emory.edu

ABSTRACT

Differentially Private Stochastic Gradient Descent (DPSGD) is a *de facto* and principled approach for training deep models with formal privacy guarantees, where per-sample gradient clipping is a key technique that shrinks the L_2 norm of each sample's gradient into a specific threshold to stabilize training. Most prior works typically assume that the stochastic gradient noise (GN) follows a light-tailed distribution (e.g., sub-Gaussian) to determine optimal clipping thresholds. However, recent studies show that GN in deep learning often exhibits heavy-tailed behavior, leading to excessive clipping loss and degraded performance. While several approaches consider heavy-tailed settings, they lack analytical guidance for threshold selection and are limited to specific heavy-tailed distributions.

In this paper, we present a novel clipping mechanism for DPSGD under a generalized sub-Weibull GN assumption. We first establish unified high-probability optimization guarantees that achieve the best-known convergence rates in heavy-tailed settings while retaining optimal rates in light-tailed settings. Building on this, we propose a tail-aware clipping mechanism DC-DPSGD, which privately distinguishes body and tail gradients, and applies discriminative clipping to clip them with different thresholds, thereby balancing clipping loss and DP noise. Further, we theoretically analyze the convergence of DC-DPSGD and provide tighter optimization guarantees. Extensive experiments on eleven real-world datasets demonstrate that our approach outperforms three baselines by up to 3.49%, 5.03%, and 3.70% in terms of accuracy, respectively.

PVLDB Reference Format:
Haichao Sha, Yuncheng Wu, Yang Cao, Yong Liu, Yuhan Liu, Ruixuan Liu, Hong Chen. Unified Differentially Private Stochastic Gradient Descent with Tail-Aware Discriminative Clipping. PVLDB, 19(1): XXX-XXX, 2026.
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/NoNameSha/Discriminative_Clipping_DPSGD.

1 Introduction

Data privacy has become a critical concern in modern deep learning, as models are often trained on sensitive data from diverse sources, ranging from structured records with personally identifiable attributes to unstructured data containing implicit sensitive information. Differentially Private Stochastic Gradient Descent (DPSGD) is a *de facto* and principled approach to training deep models with formal privacy guarantees [1]. It adds calibrated noise to aggregated gradients during training to obscure individual contributions. To ensure stability and control the privacy loss, DPSGD typically employs per-sample gradient clipping, which bounds each sample's gradient by a specified L_2 norm threshold before noise injection.

Recently, a number of works [9, 17, 27, 38, 48, 69, 71, 73] have been proposed to optimize the clipping mechanisms for DPSGD with per-sample gradient clipping (aka. clipped DPSGD). Most existing works assume that the stochastic gradient noise (GN)—defined as the deviation between the randomly sampled gradient and the average gradient estimated on the full training dataset—follows a light-tailed distribution, such as sub-Gaussian. For instance, [9, 69, 73] focus on concentrated gradients near the center of the light-tailed GN distribution and uniformly trim the tail region for efficient clipping, while [3, 27, 38, 71, 78] leverage the boundedness of GN under the light-tailed distribution to determine the clipping thresholds and derive (near) optimal convergence rates.

Nevertheless, these light-tailed GN assumptions do not align with empirical findings, as extensive studies [7, 11, 56, 57, 76] have shown that the GN of SGD in deep learning often exhibits a heavy-tailed distribution. This occurs even when the dataset originates from a light-tailed distribution, the gradient noise still diverges to a heavy-tailed distribution with unbounded variance [31]. This heavy-tailed behavior, reflecting the presence of hard-to-learn samples with large gradient deviations, typically slows down the convergence rate and impairs training performance [30, 42, 43, 49]. To mitigate this problem, several approaches [17, 48] analyze clipped DPSGD in heavy-tailed settings and derive expectation-based convergence bounds under a specific heavy-tailed Lipschitz assumption. However, these approaches fail to give analytical guidance for determining the optimal clipping threshold in non-convex clipped DPSGD, as solving the nonlinear clipping function is intractable. Moreover, their analysis under the heavy-tailed Lipschitz assumption applies only to limited cases and cannot be generalized to more severe heavy-tailed behaviors.

Our goal is to design a theoretically sound clipping mechanism for clipped DPSGD that effectively handles general heavy-tailed GN scenarios and provides analytical guidance for clipping threshold selection. However, there are two main challenges. First, the upper bound of moment generating functions (MGF) [62] is unmeasurable with heavy-tailed GN, which means that the variance of GN in expectation could be dominated by more extreme values, leading to unboundedness. Thus, the expectation tools widely used in prior works are hardly applicable to obtaining analytical solutions for

Proceedings of the VLDB Endowment, Vol. 19, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

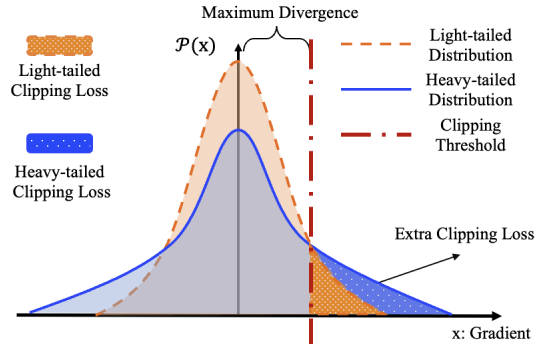

Figure 1: The trade-off comparison between light-tailed and heavy-tailed GN distributions.

the clipping threshold and achieving optimal convergence rates for differentially private optimization. Second, it is challenging to balance the magnitude of random perturbations added to the gradients, i.e., differential privacy (DP) noise, and the clipping loss that is tied to the clipping operation, under the heavy-tail GN setting. Figure 1 shows an example of this trade-off under light-tailed and heavy-tailed GN distributions. On the one hand, as the clipping threshold increases (i.e., when the red dotted line moves to the right), the clipping loss decreases, while the scale of DP noise increases as the maximum divergence (i.e., DP sensitivity) is larger. This could lead to a high-dimensional catastrophe [81] and negatively impact model performance. On the other hand, under the same DP noise magnitude (i.e., when fixing the red dotted line), the slower decay rate of the heavy-tailed distribution (blue line) will introduce extra clipping loss compared to the light-tailed distribution (orange line). It means that aggressive clipping on the tail region can severely distort gradient information under heavy-tailed GN distributions, leading to substantial gradient bias.

In this paper, we present a novel clipping mechanism that achieves the goal above. We address the first challenge by modeling GN using a unified heavy-tailed distribution and employing high-probability analytical tools in place of traditional expectation tools. This allows us to rigorously characterize tail behaviors and bound the convergence rate of clipped DPSGD under heavy-tailed GN. Specifically, we adopt the sub-Weibull distribution, which generalizes both light-tailed and heavy-tailed distributions through a single form parameterized with a tail index θ (Definition 2.2). For instance, $\theta = 1/2$ corresponds to the light-tailed sub-Gaussian distribution, $\theta < 1$ to the heavy-tailed Lipschitz distribution, and $\theta \geq 1$ to even heavier tails. Building on this, we derive unified optimization guarantees for clipped DPSGD that achieve the best-known convergence rates in heavy-tailed GN settings while retaining optimal convergence rates in light-tailed GN settings. To our knowledge, this is the first unified optimization guarantee for clipped DPSGD under both light-tailed and heavy-tailed GN assumptions. Our theoretical analysis further reveals that the convergence performance deteriorates as θ increases, motivating us to establish a principled relationship between the clipping threshold and the tail index.

We address the second challenge by proposing a tail-aware approach, named **Discriminative Clipping (DC)-DPSGD**, which effectively balances the extra clipping loss and required DP noise. We observe that the central body of heavy-tailed GN distributions behaves similarly to light-tailed distributions, while their tails decay much more slowly. Therefore, our key idea is to decompose the gradients into a light body region and a heavy tail region, and to use a larger clipping threshold for the heavy tail region, so as to retain more information from tail gradients while preventing unnecessary noise perturbation of body gradients. Specifically, we design a private subspace identification technique to detect tail gradients and devise a discriminative clipping method with two thresholds—one for the body and a larger one for the tails. We theoretically analyze the choice of these two clipping thresholds and provide tighter optimization guarantees for DC-DPSGD.

We compare DC-DPSGD against three differentially private baselines on eleven real-world datasets. The experimental results show that DC-DPSGD consistently outperforms the baselines. In particular, it achieves up to 3.49%, 5.03%, and 3.70% improvements in accuracy over standard DPSGD, Auto-S, and DP-PSAC, respectively, demonstrating the effectiveness of our approach.

In summary, we make the following contributions in this paper.

- We provide high-probability optimization guarantees for clipped DPSGD under the sub-Weibull GN assumption, achieving the best-known convergence rates in heavy-tailed settings while preserving optimal rates in light-tailed cases. To our knowledge, this is the first unified theoretical guarantee for clipped DPSGD across both light-tailed and heavy-tailed GN regimes.
- We propose a tail-aware clipping mechanism DC-DPSGD with a private subspace identification technique and a discriminative clipping method to optimize clipped DPSGD under generalized heavy-tailed GN settings. We further analyze the convergence of DC-DPSGD and provide tighter optimization guarantees.
- We conduct extensive experiments on eleven real-world datasets, where DC-DPSGD consistently outperforms three differentially private baselines, achieving up to 3.49%, 5.03%, and 3.70% accuracy improvements respectively, which demonstrates the effectiveness of our proposed approach.

The remainder of the paper is structured as follows. Section 2 introduces the preliminaries. Section 3 presents our unified optimization guarantees for clipped DPSGD. Section 4 details the proposed discriminative clipping DPSGD approach. Section 5 presents the experimental evaluation. Section 6 reviews the related works, followed by the conclusion in Section 7.

2 Preliminaries

2.1 Differentially Private Optimization

Let S be a private training dataset, which consists of n training data $S = \{z_1, ..., z_n\}$ with a sample domain Z drawn i.i.d. from an underlying distribution $\mathcal{P}$. We aim to train a private model parameterized with $\mathbf{w} \in W \subseteq \mathbb{R}^d$, where W represents the model parameter space. Since the underlying distribution $\mathcal{P}$ is unknown and inaccessible in practice, we instead minimize the empirical risk in a differentially private manner:

$$L_S(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}, z_i), \tag{1}$$

where the loss function $\ell : W \times Z \to \mathbb{R}$ is typically non-convex. We denote by $\nabla\ell(\mathbf{w}, z_i)$ the gradient of the loss $\ell(\mathbf{w}, z_i)$ with respect to $\mathbf{w}$, evaluated at sample z_i. At every iteration t, we randomly sample a mini-batch $B_t \subseteq S$ by drawing j_t uniformly from the set $\{j : j \in [n]\}$, and define the stochastic gradient as $\nabla\ell(\mathbf{w}, z_{j_t})$, $z_{j_t} \in B_t$ and the average empirical gradient as$\nabla L_S(\mathbf{w}_t) := \frac{1}{n}\sum_{i=1}^{n}\nabla\ell(\mathbf{w}_t, z_i)$.

DPSGD [1] is widely used to preserve training data privacy in deep learning with SGD. In each iteration, it first clips the gradients into a specified threshold c and then adds random perturbation noise proportional to c. As a result, the overall training process ensures differential privacy (DP) based on composition theorems and post-processing properties [5, 6, 10, 23, 24, 50, 66].

*Definition 2.1 (**Differential Privacy**).* A randomized algorithm M is (ε, δ)-differentially private if for any two neighboring datasets S, S' differ in exactly one data point and any event Y, we have

$$\mathbb{P}(M(S) \in Y) \leq \exp(\varepsilon) \cdot \mathbb{P}(M(S') \in Y) + \delta, \tag{2}$$

where ε is the privacy budget and δ is a small probability.

Given Definition 2.1 of DP, we characterize our convergence rates in terms of the following key quantity:

$$\varphi = \frac{\sqrt{d\log(T/\delta)}}{n\varepsilon}, \tag{3}$$

where T is the number of iterations and d is the number of model parameters. We use $\mathbb{O}(\cdot)$ to represent dominant higher-order terms. Throughout the paper, $\|v\|_2$ denotes the L_2 norm for any vector v.

For ease of reference, we summarize the frequently used notations in Appendix A.2 of the full paper [32].

2.2 Sub-Weibull Distribution

We employ the sub-Weibull distribution (see Definition 2.2) to model the gradient noise as it unifies light-tailed and heavy-tailed behaviors within a single analytical form.

Definition 2.2 (Sub-Weibull Distribution [62]). A random variable X is said to follow a *sub-Weibull distribution* with tail index $\theta > 0$ and scale parameter $K > 0$, denoted by $X \sim subW(\theta, K)$, if

$$\mathbb{E}_X\left[\exp\left(|X/K|^{1/\theta}\right)\right] \leq 2. \tag{4}$$

In particular, the sub-Weibull distribution corresponds to light-tailed sub-Gaussian variables when $\theta = 1/2$, and sub-Exponential variables when $\theta = 1$. Larger values with $\theta > 1$ indicate heavier tails. In this paper, we refer to stochastic gradient noise as heavy-tailed when $\theta > 1/2$, and as light-tailed when $\theta = 1/2$.

2.3 Assumptions

We now introduce the assumption of sub-Weibull gradient noise and other assumptions commonly used in DPSGD.

Assumption 2.1 (**Sub-Weibull Gradient Noise** [42]). *Conditioned on the iterative parameter $\mathbf{w}_t$, the gradient noise $\mathcal{G}_t := \nabla\ell(\mathbf{w}_t, z_{j_t}) - \nabla L_S(\mathbf{w}_t)$ satisfies $\mathbb{E}[\nabla\ell(\mathbf{w}_t, z_{j_t}) - \nabla L_S(\mathbf{w}_t)] = 0$ and $\|\nabla\ell(\mathbf{w}_t, z_{j_t}) - \nabla L_S(\mathbf{w}_t)\|_2 \sim subW(\theta, K)$, where $subW(\theta, K)$ denotes a Sub-Weibull distribution with tail index θ and positive scale parameter K, such that $\theta \geq \frac{1}{2}$, and we have*

$$\mathbb{E}_{z_{j_t}\sim S}[\exp((\|\nabla\ell(\mathbf{w}_t, z_{j_t}) - \nabla L_S(\mathbf{w}_t)\|_2/K)^{\frac{1}{\theta}})] \leq 2. \tag{5}$$

Assumption 2.1 is a relaxed version of gradient noise following sub-Gaussian distributions, that is, $\mathbb{E}_t[\exp((\|\nabla\ell(\mathbf{w}_t, z_{j_t}) - \nabla L(\mathbf{w}_t)\|_2/K)^2)] \leq 2$. It implies that finding upper bounds for moment generating function under Assumption 2.1 is impracticable by standard tools [62]. Therefore, we use the truncated tail theory [4] and martingale difference inequality [49] to address this problem in our analysis.

Assumption 2.2 (β-**Smoothness**). *The loss function ℓ is β-smooth, for any sample $z \in Z$ $\mathbf{w}_t, \mathbf{w}'_t \in W$, we have*

$$\|\nabla\ell(\mathbf{w}_t, z) - \nabla\ell(\mathbf{w}'_t, z)\|_2 \leq \beta\|\mathbf{w}_t - \mathbf{w}'_t\|_2. \tag{6}$$

Assumption 2.3 (**G-Bounded**). *For any $\mathbf{w}_t \in \mathbb{R}^d$, there exists a positive real number $G > 0$, and the expectation gradient satisfies*

$$\|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq G. \tag{7}$$

Assumption 2.2 is widely used in optimization literature [28, 42, 81] and is essential for ensuring the convergence of gradients to zero [44]. Assumption 2.3 is milder compared to the bounded stochastic gradient assumption [42, 43, 81], i.e., $\|\nabla\ell(\mathbf{w}_t, z_i)\|_2^2 \leq G$. Note that Assumption 2.3 constrains the L_2 norm of the average gradient under the objective [42, 49], while Assumption 2.1 characterizes the randomness of individual gradients by quantifying their deviation from the empirical average over private training data. These assumptions provide a complementary view of the gradient behavior during optimization. We include other lemmas and theoretical tools in Appendix B.

3 Unified Optimization Guarantees for Clipped DPSGD under Heavy-tailed Gradient Noise

In this section, we first revisit existing optimization guarantees for clipped DPSGD and establish a unified guarantee under heavy-tailed gradient noise. We then discuss the theoretical insights that motivate our approach. Due to space limitations, we provide only the main results here, while the detailed proofs and extended discussions are included in Appendix C of the full paper [32].

3.1 Optimization Guarantees

Most existing analyses of clipped DPSGD focus on light-tailed sub-Gaussian gradient noise, relying on strong symmetry assumptions to handle clipping loss [9, 14], which are unrealistic as GN often exhibits heavy-tailed behaviors. Recent approaches [17, 48] analyze optimization guarantees by assuming that gradients satisfy a specific heavy-tailed Lipschitz condition, that is, the per-sample gradient has unbounded upper constants. Nevertheless, they could not provide analytical guidance on selecting the clipping threshold.

In this paper, we derive unified optimization guarantees under a more general assumption that GN follows a sub-Weibull distribution. This assumption subsumes the light-tailed sub-Gaussian and heavy-tailed Lipschitz assumptions (see Lemma 22 of [49]), which correspond to $\theta = 1/2$ and $\theta < 1$ in Definition 2.2, respectively. Hence, prior results become special cases of ours, while our results extend naturally to even heavier tails with $\theta \geq 1$. We follow the outline of Abadi's clipped DPSGD scheme [1], as summarized in Algorithm 1. The clipping mechanism is defined as $\overline{\mathbf{g}}_t(z_i) = \mathbf{g}_t(z_i)/\max(1, \frac{\|\mathbf{g}_t(z_i)\|_2}{c})$, where $\mathbf{g}_t(z_i) = \nabla\ell(\mathbf{w}_t, z_i)$ and c denotes the clipping threshold, serving as a biased estimate. Our

Algorithm 1 Outline of Clipped DPSGD [1]

Input: sample size n, mini-batch size B, clipping threshold c, learning rate η_t, noise scale σ, the number of iterations T.

Initialize $\mathbf{w}_0$ randomly.
for iteration $t = 1, \ldots, T$ **do**
  Take a random mini-batch B_t with sampling ratio B/n.
  $\widetilde{\mathbf{g}}_t = \text{Clip\_and\_Perturbation}(c, B_t)$.
  Update model parameters: $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \widetilde{\mathbf{g}}_t$.
end for
return $\mathbf{w}_T$

Function Clip_and_Perturbation(c, π):
  for every sample z_i in π **do**
    Compute per-sample gradient: $\mathbf{g}_t(z_i) = \nabla \ell(\mathbf{w}_t, z_i)$.
    Abadi's clipping: $\bar{\mathbf{g}}_t(z_i) = \mathbf{g}_t(z_i) / \max(1, \frac{\|\mathbf{g}_t(z_i)\|_2}{c})$.
  end for
  Compute the weighted average of the gradients: $\bar{\mathbf{g}}_t = \sum_{i=1}^{|\pi|} \bar{\mathbf{g}}_t(z_i)$.
  Add the corresponding noise: $\widetilde{\mathbf{g}}_t = \frac{1}{|\pi|}\left(\bar{\mathbf{g}}_t + \mathbb{N}\left(0, c^2\sigma^2 \mathbb{I}_d\right)\right)$.
return $\widetilde{\mathbf{g}}_t$

Output: trained model parameters $\mathbf{w}_T$.

unified optimization guarantee for clipped DPSGD under the sub-Weibull GN assumption is presented in the following theorem.

Theorem 3.1 (**Convergence Rate of Clipped DPSGD under Sub-Weibull Gradient Noise Assumption**). *Let $\mathbf{w}_t$ be the iterative parameter $\mathbf{w}_t$ produced by Algorithm 1 with learning rate $\eta_t = \frac{1}{\sqrt{T}}$, where $T = \mathbb{O}(\frac{n\varepsilon}{\sqrt{d\log(1/\delta)}})$, $T \geq 1$, and d is the number of model parameters. Under Assumptions 2.1 and 2.2, given that the clipping threshold $c = \max\left(4K\log^{\theta}(\sqrt{T}), 39K\log^{\theta}(2/\delta)\right)$, for any $\delta \in (0,1)$, with probability $1-\delta$, we have:*

$$\frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} \leq \mathbb{O}\left(\log^{\max(1,\theta)}(T/\delta)\log^{2\theta}(\sqrt{T})\varphi^{\frac{1}{2}}\right). \tag{8}$$

Proof Sketch. We give a proof sketch and provide the complete proof in our full paper [32]. In the derivation process, a key quantity is the first-order term $\langle \bar{\mathbf{g}}_t(z_i) - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle$. We split it into two components, i.e.,

$$\langle \bar{\mathbf{g}}_t(z_i) - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle = \tag{9}$$
$$\langle \bar{\mathbf{g}}_t(z_i) - \mathbb{E}_t[\bar{\mathbf{g}}_t(z_i)], \nabla L_S(\mathbf{w}_t)\rangle + \langle \mathbb{E}_t[\bar{\mathbf{g}}_t(z_i)] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle.$$

We first derive high-probability bounds by constructing martingale difference sequences $\sum_{t=1}^{T}(\langle \bar{\mathbf{g}}_t(z_i) - \mathbb{E}_t[\bar{\mathbf{g}}_t(z_i)], \nabla L_S(\mathbf{w}_t)\rangle)$ and applying advanced concentration inequalities to bound the term. Then, we utilize the sub-Weibull properties to obtain an upper bound of the second term $\langle \mathbb{E}_t[\bar{\mathbf{g}}_t(z_i)] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle$. □

From this theorem, we observe that the convergence rate degrades as θ increases, since both $\log^{\max(1,\theta)}(T/\delta)$ and $\log^{2\theta}(\sqrt{T})$ grow with θ. This insight motivates the need to establish a relationship between the clipping threshold c and the heavy tail index θ, which we will elaborate on in Section 3.2. We further compare our result in Theorem 3.1 with the current optimization guarantees of clipped DPSGD under light-tailed and heavy-tailed settings [9, 17, 48, 73], as summarized in Appendix A.1. There are two main observations. First, when $\theta = 1/2$, corresponding to the light-tailed sub-Gaussian GN setting, our optimization guarantees $\mathbb{O}(\log(T/\delta)\log(\sqrt{T})\varphi^{\frac{1}{2}})$ aligns with the current optimal rates without the gradient symmetry assumption [73], except for extra high probability terms. Second, the convergence rates in [17] depend on θ through φ and δ, which deteriorate rapidly as the tail becomes heavier. Their approach also fails to extend to sub-Exponential or heavier-tailed distributions ($\theta \geq 1$), where the rates degenerate to $\mathbb{O}(1)$, meaning that their training algorithm cannot converge in this setting. In contrast, θ is only related to the logarithmic terms in our result, which remains valid with heavier tails when $\theta \geq 1$.

3.2 Theoretical Insights

As discussed in Section 1, extreme samples that deviate significantly from the mean occur more frequently under heavy-tailed distributions, resulting in gradient noise with unbounded L_2 norms and thus more severe clipping loss. To analyze the clipping loss of DPSGD, we need to estimate the deviation between the clipped gradient and the true mean gradient, denoted by $\|\bar{\mathbf{g}}_t - \nabla L_S(\mathbf{w}_t)\|_2$. Unlike the light-tailed case, this deviation cannot be easily bounded by its expectation over t iterations, necessitating alternative theoretical treatment in the heavy-tailed setting. Therefore, we decompose the deviation term to explicitly isolate the gradient noise component $\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2$ that dominates the clipping loss.

$$\begin{aligned}\|\mathbb{E}_t[\bar{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t)\|_2 &= \|\mathbb{E}_t[(\mathbf{g}_t - \nabla L_S(\mathbf{w}_t))b_t]\|_2 \\ &\leq \sqrt{\mathbb{E}_t[\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2^2]\,\mathbb{E}_t[b_t^2]}.\end{aligned} \tag{10}$$

Intuitively, when the stochastic gradient $\mathbf{g}_t$ is close to the true mean gradient $\nabla L_S(\mathbf{w}_t)$, it contributes little to the final error. Thus, the upper bound of the deviation term is mainly determined by cases where the stochastic gradient deviates substantially from the true mean. To capture such cases, we define the indicator $b_t = \mathbb{I}_{\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > \frac{c}{2}}$, which identifies large-deviation gradient noise that dominates the clipping loss. The constant $\frac{1}{2}$ is chosen for analytical tractability and can be replaced by any constant in the interval of $(0,1)$ without loss of generality. Then, according to Assumption 2.1, we have:

$$\mathbb{E}_t[b_t^2] = \mathbb{P}(\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > \frac{c}{2}) \leq 2\exp(-(\frac{c}{4K})^{\frac{1}{\theta}}). \tag{11}$$

From Eq. 11, the optimal clipping threshold increases with the tail index θ, as heavier tails require a larger c for optimality. Thus, directly adopting the same clipping threshold c as used in light-tailed settings tends to exacerbate the extra error $\mathbb{O}(\varphi^{-1/2})$ in $\mathbb{E}_t[b_t^2]$, since the clipping threshold is insufficient to accommodate the heavier-tailed gradients. Compared with the error introduced by DP noise, the effect of clipping is more severe, and the optimization guarantee of clipped DPSGD may become suboptimal or even collapse due to the extra clipping loss.

Optimal clipping threshold under heavy-tailed GN. Our unified optimization guarantee in Theorem 3.1 provides an important insight that the ideal clipping threshold c should scale up as the

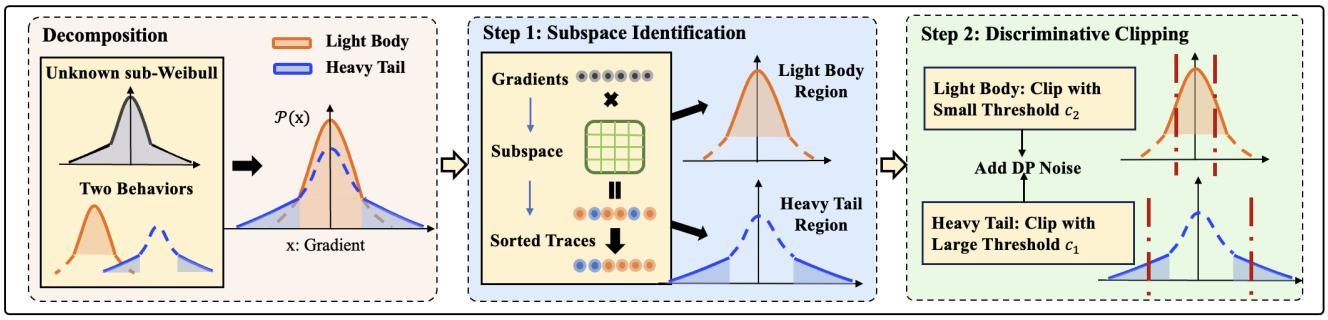

Figure 2: Overview of DC-DPSGD.

tail index θ increases because the theoretical value of c is positively correlated to θ. Specifically, as derived in Theorem 3.1, the optimal clipping threshold is given by $c = \max\left(\mathbb{O}\left(\log^{\theta}(1/\delta), \log^{\theta}(\sqrt{T})\right)\right)$. The two logarithmic terms originate from distinct factors. The term $\mathbb{O}(\log^{\theta}(1/\delta))$ arises from the high-probability concentration under the heavy-tailed assumption, ensuring that the deviation bound holds with high confidence $1-\delta$. The term $\log^{\theta}(\sqrt{T})$ accounts for the cumulative clipping loss across T iterations. When c deviates from its optimal value, the deviation bound deteriorates in opposite directions. A smaller c enlarges the probability term $\mathbb{P}(\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > c/2)$, slowing the exponential decay term $\exp(-(\frac{c}{4K})^{1/\theta})$ and inducing the extra error $\mathbb{O}(\varphi^{-1/2})$ in the upper bound. Conversely, an excessively large c reduces clipping loss but amplifies the variance of the injected DP noise, which scales as c^2 (see line 15 in Algorithm 1), thereby loosening the overall error bound. Hence, the clipping mechanism must strike a balance between clipping loss and injected DP noise to ensure tighter optimization guarantees in the heavy-tailed setting.

4 Discriminative Clipping DPSGD

In this section, we first present our rationale and give an overview of the proposed mechanism DC-DPSGD. Then, we detail the private subspace identification and tail-aware discriminative clipping steps in DC-DPSGD. Finally, we provide a theoretical analysis of the optimization guarantee and privacy guarantee.

4.1 Rationale and Overview

Rationale. According to the optimization guarantee and theoretical insights presented in Section 3, the clipping threshold shall scale with the heavy tail index θ, while balancing the trade-off between clipping loss and DP noise. In practice, most GN values remain small and concentrated around the mean of the distribution; hence, it is desirable to maintain a low DP noise level for these gradients while selectively mitigating clipping loss for those with large magnitudes. To achieve this, our key idea is to classify the sub-Weibull gradient noise into two regions: a *light body* region with small deviations, whose decay resembles a sub-Gaussian distribution, and a *heavy tail* region with large deviations, which exhibits heavier-tailed behavior. Subsequently, we can apply different clipping thresholds to the two regions: a smaller threshold for the light body to reduce DP noise following existing optimal guidance, and a larger threshold for the heavy tail to alleviate excessive clipping loss.

Overview. Figure 2 illustrates an overview of DC-DPSGD, which consists of two steps. In the first step (Section 4.2), we propose a private subspace identification technique to distinguish gradients from light body and heavy tail, where we estimate distributional properties of per-sample gradients by computing and sorting their traces under a specific subspace. To make the trace sorting satisfy differential privacy, we add DP noise with scale σ_{tr} to this step, and provide the corresponding utility guarantees (Theorem 4.1). In the second step (Section 4.3), we present a discriminative clipping method that utilizes different clipping thresholds for the two regions and adds DP noise with scale σ_{dp} for gradient perturbation. To establish the comprehensive theoretical guarantees of our differentially private optimization method, we utilize the sharp sub-Weibull concentration tools [4] to derive high-probability optimization guarantees (Theorem 4.2) for the light body and heavy tail regions, respectively. Moreover, we combine the results of the two steps to obtain a complete optimization guarantee (Theorem 4.3) for DC-DPSGD (Section 4.4). At last, we rigorously analyze the privacy guarantee of DC-DPSGD (Theorem 4.5) and provide the privacy budget allocation for the two steps (Section 4.5). Algorithm 2 presents the detailed steps of DC-DPSGD.

4.2 Private Subspace Identification

We now introduce our subspace identification technique in the first step. We note that numerous studies [46, 81] leverage in-distribution public subspaces to preserve the low-rank private training information. This implies that gradients tend to exhibit more pronounced eigen signals in similar subspaces due to shared eigenvector directions. In light of this, samples from the light-body region are expected to generate stronger responses within subspaces characterized by sub-Gaussian distributions, while heavy-tail samples tend to be more active in subspaces associated with heavier sub-Weibull distributions with $\theta \geq 1$. In addition, due to the high-dimensional nature of gradients in DP stochastic learning, their normalized versions can act as mutually orthogonal directional eigenvectors [63] and provide effective optimization information for model training [9, 73]. Given that normalized gradients have bounded L_2 sensitivity, we can bypass the unbounded norm of heavy-tailed gradients and make DP estimation practicable. Considering the above two points, a higher inner product between the normalized gradients and the subspace following heavy-tailed distributions serves as a

Algorithm 2 Discriminative Clipping DPSGD

Input: tail proportion p; learning rate η_t; number of iterations T; tail index θ; heavy-tailed and light-tailed clipping thresholds c_1, c_2.

1: Initialize $\mathbf{w}_0$ randomly and initialize $V_{t,k}$ to None.
2: **for** each iteration $t = 1 \dots T$ **do**
3: **if** `Strategy_trigger` **then**
4: **Step 1: Private Subspace Identification**
5: Private sorting: $(S^{\text{tail}}, S^{\text{body}}) \leftarrow$ Algorithm 3(σ_{tr}),
6: where $i \in S$, $S^{\text{tail}} = \{\widetilde{z}_i\}_{i=1}^{pn}$, $S^{\text{body}} = \{\widetilde{z}_i\}_{i=1}^{(1-p)n}$.
7: Sample mini-batches with rates q_1 and q_2:
8: $S^{\text{tail}} \leftarrow \{B_1^{\text{tail}}, \dots, B_{q_1 pn}^{\text{tail}}\}$,
9: $S^{\text{body}} \leftarrow \{B_1^{\text{body}}, \dots, B_{q_2(1-p)n}^{\text{body}}\}$.
10: Apply random permutation:
11: $\Pi \leftarrow \text{Perm}\Big(B_1^{\text{body}}, \dots, B_{q_2(1-p)n}^{\text{body}}, B_1^{\text{tail}}, \dots, B_{q_1 pn}^{\text{tail}}\Big)$.
12: **end if**
13: **Step 2: Discriminative Clipping**
14: **for** each batch $\pi \in \Pi$ **do**
15: **if** $\pi \in S^{\text{tail}}$ **then**
16: $\widetilde{\mathbf{g}}_t \leftarrow$ Clip_and_Perturbation(c_1, π)
17: **else if** $\pi \in S^{\text{body}}$ **then**
18: $\widetilde{\mathbf{g}}_t \leftarrow$ Clip_and_Perturbation(c_2, π)
19: **end if**
20: Update parameters: $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \widetilde{\mathbf{g}}_t$
21: **end for**
22: **end for**

Output: model parameters $\mathbf{w}_T$.

measure of linear transformation similarity, indicating closer alignment with the heavy-tail region. Conversely, a lower similarity suggests membership in the light body region.

Private sorting by Gaussian mechanism. To capture these properties, we compute the trace of each per-sample gradient projected onto a specific subspace, using it as a signal to distinguish the sample's behavior. Then, we apply a private sorting mechanism to perturb the trace sequence, enabling the identification of heavy-tailed samples in a privacy-preserving manner. From a theoretical perspective, the trace corresponds to the empirical total variance and serves as a proxy for characterizing distributional properties. Accordingly, we establish utility guarantees showing that empirical traces measured in a randomly sampled projection subspace can reliably approximate the population variance, enabling us to identify heavy-tailed samples with high-probability guarantees.

To be concrete, in the private subspace identification step of each iteration $t \in [1, T]$, we first construct a projection matrix composed of k random orthogonal unit vectors $V_{t,k} = [v_{t,1}, v_{t,2}, ..., v_{t,k}] \in \mathbb{R}^{d \times k}$, which is sampled from the sub-Weibull distributions with different candidate tail indices $\theta = [1/2, 1, 2]$. The candidate tail indices are selected to span a representative spectrum of tail behaviors, ranging from sub-Gaussian ($\theta = 1/2$) to sub-Exponential ($\theta = 1$), and heavier-tailed regimes ($\theta = 2$). This enables the subspace to capture diverse gradient noise patterns. Secondly, we compute the per-sample projected trace by $\lambda_{t,i}^{\text{tr}} = \text{tr}(V_{t,k}^\top \hat{\mathbf{g}}_t(z_i) \hat{\mathbf{g}}_t^\top(z_i) V_{t,k})$, where $\hat{\mathbf{g}}_t(z_i)$ is the normalized version of $\mathbf{g}_t(z_i)$ with bounded directional information and $V_{t,k}^\top$ is the transpose of $V_{t,k}$. Nevertheless,

Algorithm 3 Private Sorting by Gaussian Mechanism

Input: noise multiplier σ_{tr}.

1: Extract orthogonal vectors $[v_1, ..., v_k]$ from a sub-Weibull distribution with θ and construct projection subspace with $V_{t,k} V_{t,k}^\top = \frac{1}{k} \sum_{i=j}^{k} v_j v_j^\top$.
2: **for** per-sample $i \in [1, n]$ **do**
3: Normalize per-sample gradient: $\hat{\mathbf{g}}_t(z_i) = \mathbf{g}_t(z_i) / \|\mathbf{g}_t(z_i)\|$.
4: Calculate the trace of the projected second moment: $\lambda_{t,i}^{\text{tr}} = \text{tr}(V_{t,k}^\top \hat{\mathbf{g}}_t(z_i) \hat{\mathbf{g}}_t^\top(z_i) V_{t,k})$.
5: Perturb traces $\widetilde{\lambda}_{t,i}^{\text{tr}} = \lambda_{t,i}^{\text{tr}} + \mathbb{N}(0, \sigma_{\text{tr}}^2 \mathbb{I})$.
6: **end for**
7: Sort all samples $\{z_i\}_{i=1}^n$ by $\widetilde{\lambda}_{t,i}^{\text{tr}}$ in descending order.
8: Select top-pn as heavy tail set S^{tail} and label them as heavy tail samples $\overline{z}_i, \overline{z}_i \in S^{\text{tail}}$.
9: Assign the remaining $(1-p)n$ samples to S^{body} and label them as light body samples $\widetilde{z}_i, \widetilde{z}_i \in S^{\text{body}}$.

Output: discriminative sample set $(S^{\text{tail}}, S^{\text{body}})$.

despite normalization, the traces still carry characteristic information of the underlying private sample. Thirdly, since directly using unperturbed traces can cause privacy leakage risks, we present a differentially private sorting mechanism (see Algorithm 3) that integrates Report Noisy Argmax techniques [24] that are widely applied in private query systems [20, 47, 59, 65] and data collection [21, 80] with calibrated Gaussian noise, enforcing differential privacy on the trace sequence $\lambda_t^{\text{tr}} = [\lambda_{t,1}^{\text{tr}}, \dots, \lambda_{t,B}^{\text{tr}}]$. In contrast to private gradients on the high-dimensional model parameters, the per-sample trace is a scalar quantity, allowing us to inject one-dimensional Gaussian noise, i.e., $\widetilde{\lambda}_{t,i}^{\text{tr}} = \lambda_{t,i}^{\text{tr}} + \mathbb{N}(0, \sigma_{\text{tr}}^2 \mathbb{I})$. Finally, we introduce a tail proportion parameter $p \in (0, 1)$, according to which the perturbed trace sequence $\widetilde{\lambda}_t^{\text{tr}} = [\widetilde{\lambda}_{t,1}^{\text{tr}}, \dots, \widetilde{\lambda}_{t,B}^{\text{tr}}]$ is sorted in ascending order to determine the boundary between the light-body and heavy-tail regions.

Lines 3-12 in Algorithm 2 summarize the execution of this step. The `strategy_trigger` command (line 3) controls the frequency of private subspace identification. It can be activated only once at the beginning or periodically every few iterations, which effectively balances the stability of subspace and privacy guarantees. Given that a larger trace indicates a higher similarity between the gradient and the sampled subspace, for the top-pn traces in $\widetilde{\lambda}_t^{\text{tr}}$, we classify the corresponding samples into the region associated with the candidate tail index θ. For instance, if $\theta > 1$, they are assigned to the heavy-tail region. The remaining samples with lower trace values are classified into the light-body region.

Accuracy analysis of private subspace identification. We first analyze the utility guarantee of the subspace identification, for which we need to bound the skewing between the empirical and the population second moment, i.e., $\|V_{t,k} V_{t,k}^\top - \mathbb{E}_{V_{t,k} \sim \mathcal{D}}[V_{t,k} V_{t,k}^\top]\|_2$, where $V_{t,k} V_{t,k}^\top = \frac{1}{k} \sum_{i=1}^{k} v_{t,i} v_{t,i}^\top$. We analyze the error caused by the skewing based on Ahlswede-Winter Inequality [63]. In addition, as extra DP noise is injected into private trace sorting (line 8 in Algorithm 3), the accuracy of identification could be affected by the perturbation. Thus, we also need to constrain the error introduced by DP noise. Consequently, we derive the high-probability bound

for private subspace identification in Theorem 4.1. The detailed proof is provided in Appendix D of the full paper [32].

Theorem 4.1 (**Subspace Skewing for Identification**). *Assume that the empirical projection subspace $M = V_{t,k}V_{t,k}^\top \in \mathbb{R}^{d\times d}$ with $V_{t,k}^\top V_{t,k} = \mathbb{I}_k$ approximates the population projection subspace $\hat{M} = \hat{V}_{t,k}\hat{V}_{t,k}^\top = \mathbb{E}_{V_{t,k}\sim\mathcal{P}}[V_{t,k}V_{t,k}^\top]$, $\lambda_{t,i}^{\text{tr}} = \text{tr}(V_{t,k}^\top \hat{\mathbf{g}}_t(z_i)\hat{\mathbf{g}}_t^\top(z_i)V_{t,k})$ and $\hat{\lambda}_{t,i}^{\text{tr}} = \text{tr}(\hat{V}_{t,k}^\top \hat{\mathbf{g}}_t(z_i)\hat{\mathbf{g}}_t^\top(z_i)\hat{V}_{t,k})$, for any gradient $\hat{\mathbf{g}}_t(z_i)$ that satisfies $\|\hat{\mathbf{g}}_t(z_i)\|_2 = 1$, $\zeta_t^{\text{tr}} \sim \mathbb{N}(0, \sigma_{\text{tr}}^2\mathbb{I})$, we have:*

$$|\lambda_{t,i}^{\text{tr}} - \hat{\lambda}_{t,i}^{\text{tr}} + \zeta_t^{\text{tr}}| \leq \frac{4\log(2d/\delta_m)}{k} + \frac{m_2\sqrt{B}\log^{\frac{1}{2}}(1/\delta_{\text{tr}})}{d^{\frac{1}{2}}}, \quad (12)$$

with probability $1-\delta_m-\delta_{\text{tr}}$, where δ_m and δ_{tr} are introduced by subspace concentration and DP noise respectively.

By comparing the magnitudes $\log(2d/\delta_m)/k$ and $\log^{\frac{1}{2}}(1/\delta_{\text{tr}})/d^{\frac{1}{2}}$ in Theorem 4.1, it is evident that the first term dominates since $d \gg k$ (please refer to Appendix D for more discussion). Thus, the error is negligible especially when $k \geq \Omega(\sqrt{d})$, where $\Omega(\cdot)$ denotes an asymptotic lower bound. It means that this bound can be converted into $\mathbb{O}(1/\sqrt{d})$, indicating that the gradients can be correctly identified with high probability $1-\delta_m'$, where $\delta_m' = \delta_{\text{tr}} + \delta_m$.

4.3 Tail-aware Discriminative Clipping

Next, we present our tail-aware discriminative clipping strategy in the second step. In practice, gradients from the heavy tail region often exhibit larger deviations from the mean, resulting in larger L_2 norms. In contrast, the light body gradients tend to be more concentrated and exhibit smaller norms. We observe that under the same clipping threshold, heavy-tailed gradients suffer from more severe clipping loss than light-tailed gradients, which ultimately degrades the utility of the privatized algorithm. This necessitates different clipping strategies for heavy-tailed and light-tailed gradients, respectively. However, existing adaptive approaches [9, 69, 73] can be viewed as an approximated version of Abadi's clipping function (Algorithm 1) under small clipping threshold regimes. They primarily focus on allocating more weights to scale concentrated gradients with relatively small norms, as explained below.

- Auto-S [9] and NSGD [73] employ a normalized clipping strategy with the form $\frac{\mathbf{g}_t(z_i)}{\|\mathbf{g}_t(z_i)\|_2+\gamma}$, where γ is a regularization term and is often set to a small positive value.
- DP-PSAC [69] adopts a conservative clipping strategy to control the amplification using the weight function $\frac{\mathbf{g}_t(z_i)}{\|\mathbf{g}_t(z_i)\|_2+\frac{\gamma}{\|\mathbf{g}_t(z_i)\|_2+\gamma}}$.

These methods overlook the optimization for heavy-tailed gradients and weaken their contributions after clipping, whose large deviations are more susceptible to information loss under uniform clipping. As shown in Figure 3, Auto-S and NSGD achieve intense amplification as the gradient norm decreases, imposing excessive weight on small-norm gradients. DP-PSAC mitigates the amplification of small-norm gradients by employing a non-monotonic adaptive weight function, which estimates the true averaged gradient better. Moreover, these methods rely on the assumption of light-tailed GN, which are inapplicable in heavy-tailed settings.

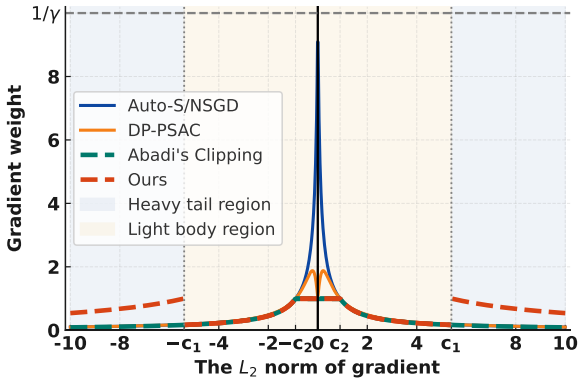

Figure 3: The gradient weight under different per-sample clipping functions, and $\gamma = 0.1$, $c_2 = 1-\gamma$ and $c_1 = 5c_2$.

To tackle this problem, we propose a discriminative clipping mechanism. After identification, the partitioned samples are randomly permuted and divided into batches, where an equivalent level of amplification can be achieved by adopting a smaller batch size (we provide a detailed privacy analysis of this process in Section 4.5). Then, as shown in line 9 of Algorithm 2, we tailor two different clipping thresholds (denoted as c_1 and c_2) for the tail and body gradients classified in the subspace identification step, and perturb the clipped gradients scaled with corresponding clipping thresholds. To make a clear comparison, we set the threshold in Abadi's clipping by $c_2 = 1-\gamma$. Taking $c_1 = 5c_2$ as an example, we define the light body region as gradients with norms less than c_1 and present the gradient weights assigned by our discriminative clipping. As illustrated in Figure 3, we assign more weights to large-norm gradients in the heavy tail region while preserving the original scale of concentrated gradients in the light body region, in contrast to the listed methods. This approach reduces the clipping loss for heavy-tailed gradients with large norms while simultaneously ensuring that body gradients with relatively small norms are not excessively affected by DP perturbation noise.

To characterize the convergence behaviors for the two regions, we assume that the gradients are classified into the correct heavy tail and light body regions. In this way, we conduct separate analyses for the two regions, each accompanied by the respective high-probability optimization guarantee. Next, to establish the optimization guarantee for the two regions, we generalize the sharp heavy-tailed concentration [4] and sub-Weibull Freedman inequality [49] to truncate the heavy-tailed distribution and find the optimal clipping threshold for each region. As a result, we have the following theorem.

Theorem 4.2 (**Convergence Rate of DC-DPSGD**). *Let $\mathbf{w}_t$ be the iterative parameter produced by discriminative clipping with $T = \mathbb{O}(\frac{n\varepsilon}{\sqrt{d\log(1/\delta)}})$, $T \geq 1$ and $\eta_t = \frac{1}{\sqrt{T}}$. Define $\Gamma(x) := \int_0^\infty t^{x-1}e^{-t}dt$, $a = 2$ if $\theta = \frac{1}{2}$, $a = (4\theta)^{2\theta}e^2$ if $\theta \in (\frac{1}{2}, 1]$, and $a = (2^{2\theta+1}+2)\Gamma(2\theta+1) + 2^{3\theta}\Gamma(3\theta+1)/3$ if $\theta > 1$. Under Assumptions 2.1, 2.2 and 2.3, for any $\delta \in (0,1)$, we have:*

(i) **In the heavy tail region**:

suppose that $c_1 = \max\left(4^\theta 2K\log^\theta(\sqrt{T}), 4^\theta 33K\log^\theta(2/\delta)\right)$,

$$C_{\text{tail}}(c_1) := \frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} \leq \mathbb{O}\left(\log^{\max(1,\theta)}(T/\delta)\log^{2\theta}(\sqrt{T})\varphi^{\frac{1}{2}}\right). \tag{13}$$

(ii) **In the light body region**:

suppose that $c_2 = \max\left(2\sqrt{2a}K\log^{\frac{1}{2}}(\sqrt{T}), 33\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)\right)$,

$$C_{\text{body}}(c_2) := \frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} \leq \mathbb{O}\left(\log(T/\delta)\log(\sqrt{T})\varphi^{\frac{1}{2}}\right). \tag{14}$$

Proof Sketch. In Theorem 4.2, the convergence rates for the two regions correspond to the discriminative clipping thresholds c_1 and c_2, denoted by $C_{\text{tail}}(c_1)$ and $C_{\text{body}}(c_2)$ respectively. First, we optimize the theoretical tools by transforming the concentration inequalities for the sum of sub-Weibull random variables X into two-region versions distinguished by the tail probability $\mathbb{P}(|X| > x)$, namely sub-Gaussian tail decay rate $\exp(-x^2)$ and heavy-tailed decay rate $\exp(-x^{1/\theta})$, $\theta > \frac{1}{2}$. Then, we analyze the high-probability bounds for the gradient noise of clipped DPSGD in each region. In the heavy tail region, we make the inequality $\mathbb{P}(\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > c_1) \leq 2\exp(-c_1^{1/\theta})$ hold and derive the dependence of factor $\log^\theta(1/\delta)$ for c_1. In the light body region, we have $\mathbb{P}(\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > c_2) \leq 2\exp(-c_2^2)$, resulting in the factor $\log^{1/2}(1/\delta)$ of c_2. Next, we investigate the high-probability error on the unbounded DP noise using Gaussian distribution properties. Finally, we integrate the results regarding gradient noise and privacy noise to determine the optimal clipping thresholds for both regions and achieve faster convergence rates for the optimization guarantee. We provide the full proof in Appendix E [32]. □

From Theorem 4.2, we can observe that when the gradients fall into the light body region, our rate $C_{\text{body}}(c_2)$ does not contain the heavy tail index θ, implying that the optimization guarantee is not affected by θ and always converges with respect to the light-tailed sub-Gaussian behavior. Moreover, the bound $C_{\text{tail}}(c_1)$ reveals the influence of the tail index θ in the heavy-tail region, which becomes deteriorated as θ increases and leads to a slower convergence rate compared to the light body region. However, since the θ-dependent effect is confined to the heavy-tail region rather than the full gradients, the result that combines the two-behavior rates can yield significantly better performance than that of the classical heavy-tailed clipped DPSGD in Theorem 3.1.

Guidance for clipping threshold selection. Existing adaptive methods implicitly couple the clipping threshold c with the learning rate η_t, forming one single tunable parameter that ultimately guides the gradient clipping. Notably, Abadi's clipping can also be transformed as a form of adaptive clipping Auto-S [9], provided a sufficiently small clipping threshold is used with a large learning rate. The guidance stated in [9] has been widely applied in practice and proven in theory [14, 55]. Prior works [9, 14] have theoretically shown that both Abadi's clipping and the adaptive clipping can achieve the same optimal order of convergence rate. However, their results cannot be extended to heavy-tailed scenarios. Here, we provide a simplified explanation for the transformation: $\eta_t\mathbf{g}_t(z_i)/\max(1, \frac{\|\mathbf{g}_t(z_i)\|_2}{c}) \overset{c\to 0}{\approx} c\eta_t\mathbf{g}_t(z_i)/(\|\mathbf{g}_t(z_i)\|_2 + \gamma)$, where γ is a small constant. Consequently, our discriminative clipping does not conflict with the majority of clipping guidance. Accordingly, for gradients in the light body region, we can follow the existing practice in Abadi's clipping and set c_2 by a sufficiently small threshold to guarantee the proven optimality of Abadi's clipping. Furthermore, to achieve optimality in the heavy tail region, we design a more relaxed threshold based on our theoretical analysis in Theorem 4.2, which shows that the clipping threshold c_1 should be about $\log^{(\theta-1/2)}(1/\delta)$ times greater than c_2. When $\theta = 1/2$, we have $c_1 = c_2$, recovering standard Abadi's clipping.

4.4 Optimization Guarantee for DC-DPSGD

In this subsection, we provide the formal optimization guarantees for DC-DPSGD. Note that the boundary derived in Section 4.3 is based on the assumption of perfectly classifying each sample into its corresponding region. However, in practice, the private subspace identification may incur utility errors by misidentification, which are jointly determined by the subspace skewing and the tail proportion p, as analyzed in Section 4.2. Since the subspace skewing of $\mathbb{O}(1/\sqrt{d})$ is non-dominant compared to the optimization guarantee in Theorem 4.2, we only consider its impact guaranteed by the high-probability term $1-\delta'_m$. Then, the parameter p, representing the probability beyond $x_{\max}$, is theoretically determined by the tail index θ and scale K of the sub-Weibull distribution. Nevertheless, estimating its true value is generally intractable. Therefore, we treat p as a tunable hyperparameter in our method. In practice, we consider values of p typically ranging in the interval refer to $[0.05, 0.2]$ [61].

Suppose that p is the proportion of heavy-tailed gradients in the mini-batch and each gradient is correctly identified into the corresponding region with probability at least $1-\delta'_m$ according to Theorem 4.1. For a mini-batch, the expected fraction of correctly identified heavy-tailed gradients is $p(1-\delta'_m)$. Thus, we can analyze the convergence rate by combining Theorems 4.1 and 4.2 to derive the formal bound for DC-DPSGD, as stated in Corollary 4.3.

Corollary 4.3 (**Optimization Guarantee for DC-DPSGD**). *Let* $\mathbf{w}_t$ *be the parameter produced by DC-DPSGD. Given Assumptions 2.1, 2.2, 2.3, and Theorem 4.2, for any* $\delta' \in (0,1)$*:*

$$\begin{aligned} C_{\text{m}}(c_1, c_2) := \frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} \\ \leq p * \mathbb{O}\left(\log^{\max(1,\theta)}(T/\delta')\log^{2\theta}(\sqrt{T})\varphi^{\frac{1}{2}}\right) \\ + (1-p) * \mathbb{O}\left(\log(T/\delta')\log(\sqrt{T})\varphi^{\frac{1}{2}}\right), \end{aligned} \tag{15}$$

with probability $1-\delta'$*, where* $\delta' = \delta'_m + \delta$*, with* δ'_m *being the error of subspace identification, and* δ *being the convergence probability.*

Proof Sketch. The bound $C_{\text{m}}(c_1, c_2)$ includes three parts: (i) the convergence rate $C_{\text{tail}}(c_1)$ from correctly identified heavy-tailed gradients with proportion p and probability $1-\delta'_m$; (ii) the convergence rate $C_{\text{body}}(c_2)$ from correctly identified light body gradients with proportion $1-p$ and probability $1-\delta'_m$; and (iii) an ignorable misidentification error $\delta'_m\left|C_{\text{tail}}(c_1) - C_{\text{body}}(c_2)\right|$ due to the small

probability δ'_m, reflecting the worst-case gap caused by applying incorrect clipping thresholds to misclassified gradients. The full proof is provided in Appendix F [32]. □

Corollary 4.3 indicates that the optimization guarantee of DC-DPSGD is composed of p-weighted average bounds, where the heavy-tailed convergence rate merely accounts for the portion of p, with the rest made up of the light body rate. Therefore, our bound is tighter than Theorem 3.1 owing to the value of p being always less than 1. Moreover, even for large tail indices θ, the tail proportion p can be restricted with a sufficiently small variance K [62]. Especially, if $p \leq 1/(\frac{C_{\text{tail}}(c_1)}{C_{\text{body}}(c_2)} + 1)$, it enables us to achieve θ-independent rates of $(1-p) * \mathbb{O}\left(\log(T/\delta')\log(\sqrt{T})\varphi^{1/2}\right)$.

4.5 Privacy Analysis

Given partitioned sampling with heterogeneous subsampling rates, the required noise must be rescaled to maintain an equivalent privacy budget. In Algorithm 2, we partition the dataset into a heavy tail region and a light body region with proportions p and $1-p$, and corresponding sampling rates q_1 and q_2. Let $\bar{q} = pq_1 + (1-p)q_2$ denote the average sampling rate. We study the noise multiplier σ_{dp} of the discriminative mechanism with the partitioned sampling.

Theorem 4.4 (Noise scaling under partitioned sampling). *Under the same privacy budget ε, the partitioned mechanism requires a noise multiplier that requires*

$$\sigma_{\text{dp}} \approx \sqrt{\frac{pq_1^2 + (1-p)q_2^2}{\bar{q}^2}}\sigma_{\text{Pois}}. \tag{16}$$

Equality holds if and only if $q_1 = q_2 = \bar{q}$.

Theorem 4.4 formalizes this relation between the noise multiplier σ_{dp} in DC-DPSGD and σ_{Pois} in standard clipped DPSGD. The proof is provided in Appendix G.1.

Finally, we analyze the privacy guarantee of DC-DPSGD. For a fair comparison to existing clipped DPSGD works, the total privacy budget allocated by DC-DPSGD to ε_{tr} and ε_{dp} is equal to the privacy budget ε in DPSGD variants, i.e., $\varepsilon = \varepsilon_{\text{tr}} + \varepsilon_{\text{dp}}$. Theorem 4.5 gives the privacy guarantee of our DC-DPSGD approach.

Theorem 4.5 (**Privacy Guarantee**). *There exist constants m_1 and m_2 such that for any $\varepsilon_{\text{tr}} \leq m_1 T$, $\varepsilon_{\text{dp}} \leq m_1 q^2 T$ and $\delta > 0$, the noise multiplier $\sigma_{\text{tr}}^2 = \frac{m_2 T \ln \frac{1}{\delta}}{\varepsilon_{\text{tr}}^2}$ and $\sigma_{\text{dp}}^2 = \frac{m_2 T \bar{q}^2 \ln \frac{1}{\delta}}{\varepsilon_{\text{dp}}^2}$ over T iterations, where $\bar{q} = pq_1 + (1-p)q_2$, and DC-DPSGD (Algorithm 2) is $(\varepsilon_{\text{tr}} + \varepsilon_{\text{dp}}, \delta)$-differentially private.*

Proof Sketch. Our private subspace identification technique follows the proof of Report Noisy Argmax (RNA) in Claim 3.9 of [24]. For the discriminative clipping mechanism, since it uses two clipping thresholds to separately handle gradients from different parts of the mini-batch, we analyze the gradient perturbation with the Gaussian mechanism [1] and subsampling. The detailed proof of privacy guarantees is provided in Appendix G [32]. □

5 Experiments

In this section, we evaluate the performance of our DC-DPSGD approach and compare it with state-of-the-art clipping mechanisms on a wide range of datasets. We first introduce the experimental setup and then report the evaluation results.

5.1 Experimental Setup

Datasets. We evaluate DC-DPSGD on eleven real-world datasets, including four image datasets, one natural language dataset, and six tabular datasets.

- **MNIST** [40] contains 60,000 training images and 10,000 testing images of handwritten digits from 10 classes.
- **FMNIST** [70] contains 60,000 training images and 10,000 testing images of fashion products from 10 categories.
- **CIFAR10** [39] contains 60,000 color images of size 32×32 from 10 object categories, with 50,000 for training and 10,000 for testing.
- **ImageNette** [18] is a curated subset of ImageNet containing 10 classifiable categories with a total of 13,394 images.
- **E2E** [22] is a natural language generation dataset for end-to-end dialogue systems, containing over 42,000 instances describing restaurant information in natural language.
- **Product** [2] contains 35,311 samples with 7 attributes. The task is to distinguish products from 10 classes.
- **Breast Cancer** [68] contains 569 samples with 30 attributes. The task is to distinguish malignant from benign tumors.
- **Android Malware** [8] contains 4,464 samples extracted from Android applications, labeled as benign or malicious.
- **Adult** [37] contains 48,842 samples. The task is to predict whether an individual belongs to the higher-income group.
- **Bank Marketing** [51] contains 45,211 samples with 16 attributes. The task predicts whether a customer will subscribe.
- **Credit Card** [74] contains 30,000 samples with 23 attributes. The task is to predict whether a customer will default on credit card payments in the following month.

We further construct two heavy-tailed datasets, namely CIFAR10-HT [12] (a heavy-tailed version of CIFAR10) and ImageNette-HT (modified on [54]) to evaluate the performance in heavier-tail settings.

Models. For MNIST and FMNIST, we use a two-layer CNN model. For CIFAR10 and CIFAR10-HT, we fine-tune SimCLRv2 pre-trained by unlabeled ImageNet and ResNeXt-29 pre-trained by CIFAR100 [60] with a linear classifier, respectively. For ImageNette and ImageNette-HT, we adopt the same setting as [9] and ResNet9 without pre-training. For E2E, we utilize a transformer-based large language model (LLM) GPT-2 (163 million parameters) and fine-tune it with the dataset. For tabular tasks, we adopt MLP models equipped with ReLU activations and a two-unit output layer for classification.

We evaluate classification tasks using accuracy that measures the portion of correct predictions, and natural language generation tasks using the BLEU score [53] that measures the quality of generated data with a modified n-gram score.

Baselines. We compare DC-DPSGD with a non-private baseline: non-DP ($\varepsilon = \infty$) and three differentially private baselines:

- **DPSGD** [1], which is the DPSGD with Abadi's clipping mechanism. It clips gradients with norms exceeding the threshold onto the L_2-ball of radius c.

Table 1: Test accuracy (%) comparison between DC-DPSGD and baselines on image datasets.

Algorithm	Privacy	MNIST	FMNIST	CIFAR10	ImageNette	CIFAR10-HT / ImageNette-HT
DPSGD	$\varepsilon = 8, \delta = 1/n^{-1.1}$	97.65±0.09	83.63±0.12	93.31±0.01	66.81±0.42	57.98±0.59 / 34.98±1.47
Auto-S		97.55±0.16	83.38±0.09	93.28±0.06	65.57±0.85	58.30±0.61 / 31.96±2.39
DP-PSAC		97.67±0.06	83.75±0.21	93.30±0.03	65.68±1.71	57.99±0.58 / 34.07±1.55
Ours (DC-DPSGD)		**98.14±0.13**	**84.76±0.34**	**93.80±0.03**	**67.66±0.29**	**61.38±1.00 / 36.72±0.91**
DPSGD	$\varepsilon = 4, \delta = 1/n^{-1.1}$	96.82±0.05	83.32±0.33	93.06±0.09	65.67±0.58	56.81±0.69/31.05±1.67
Auto-S		96.78±0.34	83.08±0.12	93.08±0.06	64.20±0.95	56.63±0.62/30.99±1.69
DP-PSAC		96.35±0.51	83.13±0.20	93.11±0.08	64.15±1.14	56.62±0.63/31.37±2.33
Ours (DC-DPSGD)		**97.92±0.11**	**84.07±0.25**	**93.36±0.14**	**66.09±0.82**	**59.03±0.81/33.58±1.37**
SGD	$\varepsilon = \infty, \delta = 1/n^{-1.1}$	99.10±0.02	89.95±0.32	94.62±0.03	72.98±0.50	71.74±0.65/39.91±1.46
DPSGD		98.16±0.06	84.03±0.08	93.69±0.08	68.61±0.42	63.22±0.68/36.62±1.07
Auto-S		98.20±0.11	83.86±0.10	93.59±0.03	68.90±0.26	62.86±0.92/35.14 ±0.86
DP-PSAC		98.20±0.06	84.18±0.15	93.70±0.01	67.89±0.48	62.87±0.94/36.84±0.84
Ours (DC-DPSGD)		**98.44** ±0.10	**84.92±0.18**	**94.21±0.06**	**70.32±0.48**	**66.57±1.22/38.78±1.04**

Table 2: BLEU (%) comparison between DC-DPSGD and baselines on natural language dataset.

Algorithm	Privacy	E2E Full	E2E LoRA
DPSGD	$\varepsilon = 8$ $\delta = \frac{1}{n^{-1.1}}$	63.189	63.389
Auto-S		63.600	63.518
DP-PSAC		63.627	63.502
Ours (DC-DPSGD)		**64.180**	**63.920**
DPSGD	$\varepsilon = 3$ $\delta = \frac{1}{n^{-1.1}}$	61.519	61.220
Auto-S		61.340	61.220
DP-PSAC		61.340	61.263
Ours (DC-DPSGD)		**61.732**	**61.563**
DPSGD	$\varepsilon = \infty$ $\delta = \frac{1}{n^{-1.1}}$	69.463	69.692
Auto-S		69.463	69.682
DP-PSAC		69.473	69.692
Ours (DC-DPSGD)		**70.328**	**70.455**

Table 3: Test accuracy (%) comparison between DC-DPSGD and baselines on tabular datasets.

Algorithm	Privacy	Product	Malware	Cancer	Adult	Bank	Credit
DPSGD	$\varepsilon = 0.8$ $\delta = \frac{1}{n^{-1.1}}$	83.22	96.86	94.74	85.12	88.62	80.90
Auto-S		82.22	96.86	94.74	85.20	88.51	80.95
DP-PSAC		83.69	96.75	95.09	85.15	88.51	80.92
Ours (DC-DPSGD)		**85.90**	**97.49**	**95.52**	**85.61**	**88.73**	**81.28**
DPSGD	$\varepsilon = 0.5$ $\delta = \frac{1}{n^{-1.1}}$	78.06	93.06	85.09	81.94	86.63	77.95
Auto-S		77.37	93.39	85.96	82.17	86.62	78.45
DP-PSAC		78.45	93.06	84.21	81.86	86.51	77.60
Ours (DC-DPSGD)		**80.03**	**93.58**	**86.32**	**82.30**	**86.72**	**78.63**
DPSGD	$\varepsilon = \infty$ $\delta = \frac{1}{n^{-1.1}}$	95.45	99.55	96.49	85.55	89.94	81.95
Auto-S		95.47	99.33	96.49	85.54	89.72	81.95
DP-PSAC		95.50	99.55	95.61	85.62	89.72	82.08
Ours (DC-DPSGD)		**96.11**	**99.78**	**97.36**	**85.79**	**90.26**	**82.44**

- **Auto-S/NSGD** [9, 73], which is the DPSGD equipped with an automatic clipping mechanism that adaptively normalizes per-sample gradient norms toward unity.
- **DP-PSAC** [69], which extends DPSGD with a controlled clipping mechanism to mitigate the unbounded amplification of small-norm gradients often induced by automatic clipping.

Implementation details. We implement pre-sample clipping by BackPACK [16]. All experiments are conducted on a server with an Intel(R) Xeon(R) E5-2640 v4 CPU at 2.40GHz and a NVIDIA Tesla P40 GPU running on Ubuntu. By default, we uniformly set subspace dimension $k = 200$, $\varepsilon = \varepsilon_{\text{tr}} + \varepsilon_{\text{dp}}$ with $\varepsilon_{\text{tr}}/\varepsilon = 0.05$, $p = 0.1$, and sub-Weibull index $\theta = 2$ for all datasets.

5.2 Effectiveness Comparison with Baselines

We compare the performance of DC-DPSGD with three differentially private baselines and a non-DP baseline on eleven datasets under various DP constraints. Tables 1, 2, and 3 summarize the results on image, natural language, and tabular datasets, respectively.

Results on image datasets. From Table 1, we observe that on normal datasets, DC-DPSGD outperforms DPSGD, Auto-S, and DP-PSAC by up to 1.71%, 2.09%, and 2.43%, respectively. While on heavy-tailed datasets (i.e., CIFAR10-HT and ImageNette-HT), the corresponding improvements are 3.49%, 5.03%, and 3.70%. Specifically, DPSGD with Abadi's clipping performs comparably to adaptive methods under small clipping thresholds, where small-norm gradients receive higher weights. However, these methods degrade notably on complex datasets (e.g., ImageNette) and those with heavy-tailed gradient noise characteristics (e.g., CIFAR10-HT and ImageNette-HT). In contrast, DC-DPSGD consistently improves accuracy on normal datasets while remaining robust under heavy-tailed settings. This robustness arises from our tail-aware discriminative clipping strategy, which assigns a larger clipping threshold to heavy-tailed gradients. Such a design effectively mitigates excessive clipping loss in the tail region, enhancing algorithm stability.

Results on natural language dataset. Table 2 presents the results on text generation tasks. We evaluate DC-DPSGD on the E2E dataset under both full fine-tuning and parameter-efficient fine-tuning (LoRA). The results demonstrate that DC-DPSGD consistently achieves superior BLEU scores compared with baselines, indicating better preservation of linguistic fluency and content fidelity under various privacy constraints. Moreover, these improvements on GPT-2–based text generation suggest that our method remains scalable and effective even in large-scale transformer architectures.

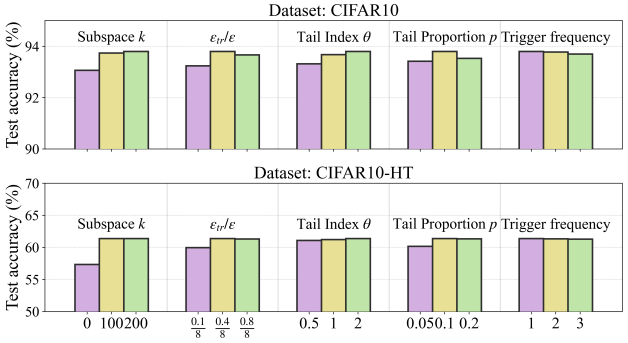

Figure 4: Effects of different parameters on test accuracy.

Results on tabular datasets. As shown in Table 3, under relatively loose privacy constraints (e.g., $\varepsilon = 0.8$), DC-DPSGD achieves superior performance across all tabular datasets, surpassing DPSGD baselines by a clear margin. Even when the privacy budget is reduced to $\varepsilon = 0.5$, the performance degradation of DC-DPSGD remains moderate compared to baselines, indicating its strong robustness to injected DP noise. These results demonstrate that DC-DPSGD maintains competitive accuracy in the low-privacy regime.

5.3 Parameter Evaluation

We next investigate the effects of various parameters in DC-DPSGD on the performance.

Parameter sensitivity analysis for DC-DPSGD. We first analyze the sensitivity of five parameters on model performance, including the dimension of the subspace k, the allocation of privacy budget ε, the tail index θ of the sub-Weibull distribution, the heavy tail proportion p and the identification strategy trigger frequency, with other parameters kept at default. The results on CIFAR10, CIFAR10-HT, and Malware are shown in Figure 4.

- **Subspace k**: We observe that the test accuracy increases with the value of k, which aligns with the theoretical analysis that the trace error is related to $\mathbb{O}(1/k)$ and has a small impact on the results.
- **$\varepsilon_{tr}/\varepsilon$**: For the allocation of privacy budget between ε_{tr} and ε, where $\varepsilon = \varepsilon_{\text{tr}} + \varepsilon_{\text{dp}}$, we empirically find that a moderate privacy budget of around 5% of the total budget allows subspace identification to maintain acceptable performance.
- **Tail index θ**: Since 'HT' datasets are extracted by sub-Exponential distributions, the gradient exhibits a heavier tail phenomenon. Therefore, adopting a heavier-tailed latent distribution with larger θ for the identification step tends to yield higher accuracy.
- **Tail proportion p**: We observe that $p = 0.1$ yields the best performance. When p is too small, it fails to sufficiently mitigate clipping loss; when too large, it introduces excessive noise. The identified proportion of heavy-tailed samples aligns with statistical expectations, while assigning larger clipping thresholds to light-body samples unnecessarily amplifies noise.
- **Trigger frequency**: Varying the trigger frequency has only a moderate effect on model performance, consistent with the empirical observation that the subspace remains stable during training. Therefore, subspace identification can be performed infrequently to reduce extra privacy overhead.

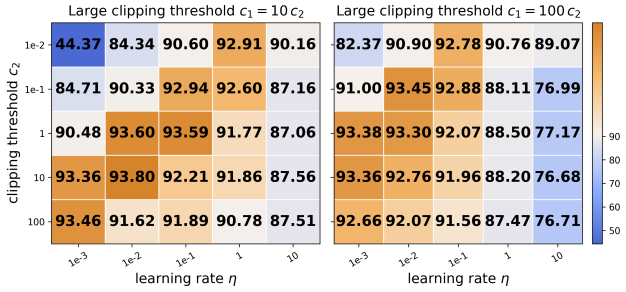

Figure 5: Test accuracy heatmap on CIFAR10 with c_1, c_2 and η.

Table 4: Effect of sampling strategy on test accuracy.

Task	Fixed Parameter	Test Accuracy
CIFAR10	Sampling rate	93.80%
$\varepsilon = 8$	Batch size	93.69%
CIFAR10-HT	Sampling rate	61.38%
$\varepsilon = 8$	Batch size	60.51%
Malware	Sampling rate	97.49%
$\varepsilon = 0.8$	Batch size	97.42%

The results demonstrate that the parameters involved in our method show low sensitivity, implying that our approach remains stable under a wide range of configurations.

Guidance for clipping threshold and learning rate. We now validate our empirical guidance for the clipping threshold in Theorem 4.2. The results in Figure 5 indicate that the optimal ratio is approximately $c_1 \approx 10c_2$. We note that when $c_1 = 100c_2$, the maximum performance declines noticeably, and when $c_1 = c_2$, it corresponds to standard clipped DPSGD. From a theoretical perspective, given the parameters $\delta = 1e^{-5}$, $\eta/B = 0.04$, and $\theta \approx 2$ (following [25, 31, 62]), we can obtain $c_1 = \mathbb{O}(\log^{\theta}(1/\delta))$, which is roughly $\sqrt{125}$ times larger than $c_2 = \mathbb{O}(\log^{1/2}(1/\delta))$. This implies $c_1 = \log^{3/2}(1/\delta)c_2 \approx 10c_2$, confirming that the empirically observed optimal threshold ratio is theoretically consistent.

Effect of sampling strategies on test accuracy. We further investigate how different sampling strategies affect model performance. Table 4 reports test accuracy under two configurations: (1) equal sampling rates $q_1 = q_2 = q$ ($\sigma_{\text{dp}} = \sigma_{\text{Pois}}$), and (2) equal batch sizes $B_1 = B_2 = B$ (e.g., $B = 64$, $p = 0.1$, leading to $\sigma_{\text{dp}} = 1.3\sigma_{\text{Pois}}$ for CIFAR10). The accuracy difference between the two strategies is marginal (within 0.2–0.3%), suggesting that the minor degradation caused by smaller batch sizes is acceptable. Overall, this implies that maintaining consistent sampling rates can preserve privacy amplification in DC-DPSGD without noticeable performance loss.

In addition, we include the training trajectory comparison and performance evaluation on more datasets in the Appendix [32].

5.4 Evaluation of DP Auditing Guarantees

Finally, we conduct state-of-the-art DP auditing methods to examine whether our algorithm maintains formal privacy guarantees in practice. DP guarantees are described using a theoretical privacy budget ε and a failure probability δ, which can limit the ability of adversary $\mathcal{A}$ to distinguish between $M(S)$ and $M(S')$. This inference can be seen as the adversary's membership inference attack

(MIA) attempt, from which both false positive rate (FPR) and false negative rate (FNR) are derived. DP auditing [34] is a tool to verify whether an algorithm satisfies the claimed ε and δ, which leverages MIA [13] to obtain FPR and FNR, thereby deriving the empirical privacy budget ε^*. As demonstrated by [35], the privacy region of any (ε,δ)-DP mechanism is related to μ_{emp}-Gaussian Differential Privacy (GDP) [19] and can be defined as:

$$\mu_{\text{emp}} = \Phi^{-1}(1 - \text{FPR}) - \Phi^{-1}(\text{FNR}), \tag{17}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution. The empirical μ_{emp}-GDP can directly be converted to $(\varepsilon^*, \delta(\varepsilon^*))$-DP by the transformation formula of

$$\delta(\varepsilon^*) = \Phi\left(-\frac{\varepsilon^*}{\mu_{\text{emp}}} + \frac{\mu_{\text{emp}}}{2}\right) - e^{\varepsilon^*}\Phi\left(-\frac{\varepsilon^*}{\mu_{\text{emp}}} - \frac{\mu_{\text{emp}}}{2}\right). \tag{18}$$

This metric provides a convenient bridge between the GDP parameter μ_{emp} and the standard (ε, δ)-DP guarantee. In particular, given an empirical estimate of μ_{emp} obtained from auditing, the above transformation enables one to directly compute the empirical privacy budget ε^* for any target δ. Intuitively, a smaller ε^* indicates stronger empirical privacy preservation, and when ε^* approaches the claimed theoretical ε from below, it implies that the auditing process is stronger in detecting potential privacy leakage. We adopt two auditing methods in this set of experiments. One is worst-case auditing that pre-trains the private model by in-distribution data and manufactures out-of-distribution canaries [52]. These canaries are deliberately inserted and easily identifiable records that allow the auditor to detect whether the model has memorized or exposed private data. The other is one-round auditing, which inserts batches of canaries to detect potential vulnerabilities [58]. We evaluate the empirical ε^* of DC-DPSGD and compare it with that of DPSGD. To further justify the effectiveness of our private subspace identification method, we include DC-DPSGD-NSI with non-private subspace identification for comparison, i.e., ε_{tr} = null .

Table 5 presents the comparison results on CIFAR10 between our method and standard DPSGD [1] under the above two privacy auditing methods, evaluated at the 95% confidence interval. Under worst-case auditing, the empirical privacy budget ε^* of DC-DPSGD remains below the theoretically claimed privacy budget ε, demonstrating that the algorithm strictly adheres to differential privacy guarantees in practice. Under one-round auditing, DC-DPSGD exhibits a privacy detectability level comparable to that of standard DPSGD, indicating a similarly high level of privacy protection. The slightly reduced audibility may result from the insertion of batched canaries, which can dilute the worst-case privacy signal due to averaging effects across batches. Furthermore, in DC-DPSGD-NSI, where subspace identification is performed without privacy constraints, the audited privacy budget exceeds ϵ, suggesting potential leakage through non-private identification. This observation validates the necessity of our proposed private subspace identification mechanism, confirming its effectiveness in enforcing privacy guarantees and mitigating the leakage risks in non-private alternatives.

Table 5: Empirical ε^* of DC-DPSGD with DP auditing.

Method	Worst-case Audit		One-round Audit	
	$\varepsilon = 4$	$\varepsilon = 8$	$\varepsilon = 4$	$\varepsilon = 8$
DPSGD	2.07	4.02	1.80	3.72
DC-DPSGD	2.15	4.31	1.80	3.88
DC-DPSGD-NSI	4.13	7.87	3.35	7.23

6 Related Work

Heavy-tailed gradient noise and high-probability bounds. From the perspective of escaping from stationary points and Langevin dynamics, the gradient noise in neural networks is more inclined to anisotropic and non-Gaussian properties [30, 31, 56, 76], with specific heavy-tailed phenomena discovered and defined in gradient descent in deep neural networks. Recently, several works have focused on heavy-tailed convex optimization in privacy-preserving deep learning [36, 64]. However, the convergence characteristics of heavy-tailed clipped DPSGD in non-convex learning are not addressed. Meanwhile, due to the ability to capture tail behaviors of stochastic gradients, high probability theoretical tools [42, 43, 49] are widely used in non-private learning such as convex and non-convex optimization. Specifically, under bounded α-th moments assumption, [43] provide a high-probability analysis for variants like clipped SGD with momentum and adaptive step sizes by using concentration inequalities for martingales. However, these tools remain under-explored in the context of private learning. Existing works [17, 36, 48] on optimizing clipped DPSGD rely on expectation bounds, which are unsuitable for heavier assumptions.

Gradient clipping. Gradient clipping is a widely adopted technique to ensure the sensitivity of gradients is bounded in both practical implementations and theoretical analysis of DPSGD [14, 29, 38, 67, 71, 75, 79]. Since the tuning parameters in the Abadi's clipping function [1] are complex, various adaptive gradient clipping schemes have been proposed [9, 73]. These schemes scale per-sample gradients based on their norms. In particular, gradients with concentrated norms are amplified infinitely. Building upon this, [69] controls the amplification of gradients with concentrated norms in a finite manner. For the theoretical guidance of clipping thresholds, [17] proposes to set the clipping threshold as a constant strictly smaller than the minimum per-sample gradient norm under the convex heavy-tailed Lipschitz condition. Moreover, Additionally, research on clipping loss has gradually gained importance. [67] and [38] argue for the connection between sampling noise and clipping loss, and mitigate clipping loss through group sampling. However, none of these works can be adapted to gradient clipping under the heavy-tailed GN assumption in DPSGD.

7 Conclusion

In this paper, we present unified high-probability optimization guarantees for clipped DPSGD, achieving the best-known convergence rates under heavy-tailed GN while preserving optimal rates in light-tailed settings. Motivated by these guarantees, we propose a novel tail-aware clipping mechanism DC-DPSGD that applies discriminative clipping thresholds to body and tail gradients, effectively balancing clipping loss and DP noise. We further analyze the convergence of DC-DPSGD and provide tighter optimization guarantees. We conduct extensive experiments on eleven real-world datasets, and the results demonstrate that DC-DPSGD outperforms three state-of-the-art baselines by up to 3.49%, 5.03%, and 3.70% accuracy improvements, respectively.

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *CCS*. 308–318.

[2] Leonidas Akritidis. 2020. Product Classification and Clustering. https://doi.org/10.24432/C5M91Z. UCI Machine Learning Repository.

[3] Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. 2021. Differentially private learning with adaptive clipping. *NeurIPS* 34 (2021), 17455–17466.

[4] Milad Bakhshizadeh, Arian Maleki, and Victor H De La Pena. 2023. Sharp concentration results for heavy-tailed distributions. *Information and Inference: A Journal of the IMA* 12, 3 (2023), 1655–1685.

[5] Ergute Bao, Fei Wei, Yin Yang, Xiaokui Xiao, Tianyu Pang, and Chao Du. 2025. Towards Learning on Vertically Partitioned Data with Distributed Differential Privacy. In *ICDE*. 2121–2134.

[6] Ergute Bao, Yizheng Zhu, Xiaokui Xiao, Yin Yang, Beng Chin Ooi, Benjamin Hong Meng Tan, and Khin Mi Mi Aung. 2022. Skellam Mixture Mechanism: a Novel Approach to Federated Learning with Differential Privacy. *Proc. VLDB Endow.* 15, 11 (2022), 2348–2360.

[7] Melih Barsbey, Milad Sefidgaran, Murat A Erdogdu, Gael Richard, and Umut Simsekli. 2021. Heavy tails in SGD and compressibility of overparametrized neural networks. *NeurIPS* 34 (2021), 29364–29378.

[8] Parthajit Borah and Dhruba K. Bhattacharyya. 2023. TUANDROMD (Tezpur University Android Malware Dataset). UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/855/tuandromd+(tezpur+university+android+malware+dataset) 4465 instances, 241 attributes, binary classification (malware vs goodware).

[9] Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. 2024. Automatic clipping: Differentially private deep learning made easier and stronger. *NeurIPS* 36 (2024).

[10] Kuntai Cai, Xiaokui Xiao, and Graham Cormode. 2023. PrivLava: Synthesizing Relational Data with Foreign Keys under Differential Privacy. *Proc. ACM Manag. Data* 1, 2 (2023), 142:1–142:25.

[11] Alexander Camuto, Xiaoyu Wang, Lingjiong Zhu, Chris Holmes, Mert Gurbuzbalaban, and Umut Simsekli. 2021. Asymmetric heavy tails and implicit bias in gaussian noise injections. In *ICML*. 1249–1260.

[12] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *NeurIPS* 32 (2019).

[13] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*. IEEE, 1897–1914.

[14] Xiangyi Chen, Steven Z Wu, and Mingyi Hong. 2020. Understanding gradient clipping in private sgd: A geometric perspective. *NeurIPS* 33 (2020), 13773–13782.

[15] Ashok Cutkosky and Harsh Mehta. 2020. Momentum improves normalized sgd. In *ICML*. PMLR, 2260–2268.

[16] Felix Dangel, Frederik Kunstner, and Philipp Hennig. 2020. BackPACK: Packing more into Backprop. In *ICLR*. https://openreview.net/forum?id=BJlrF24twB

[17] Rudrajit Das, Satyen Kale, Zheng Xu, Tong Zhang, and Sujay Sanghavi. 2023. Beyond uniform lipschitz condition in differentially private optimization. In *ICML*. PMLR, 7066–7101.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*. Ieee, 248–255.

[19] Jinshuo Dong, Aaron Roth, and Weijie J Su. 2022. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 84, 1 (2022), 3–37.

[20] Wei Dong and Ke Yi. 2021. Residual sensitivity for differentially private multi-way joins. In *Proceedings of the 2021 International Conference on Management of Data*. 432–444.

[21] Jiawei Duan, Qingqing Ye, and Haibo Hu. 2022. Utility analysis and enhancement of LDP mechanisms in high-dimensional space. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE, 407–419.

[22] Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Computer Speech & Language* 59 (2020), 123–156.

[23] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *TCC*. Springer, 265–284.

[24] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.

[25] Khaled Eldowa and Andrea Paudice. 2024. General tail bounds for non-smooth stochastic mirror descent. In *AISTATS*. PMLR, 3205–3213.

[26] Xiequan Fan and Davide Giraudo. 2019. Large deviation inequalities for martingales in Banach spaces. *arXiv preprint arXiv:1909.05584* (2019).

[27] Huang Fang, Xiaoyun Li, Chenglin Fan, and Ping Li. 2022. Improved convergence of differential private sgd with gradient clipping. In *ICLR*.

[28] Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. 2018. Uniform convergence of gradients for non-convex learning and optimization. *NeurIPS* 31 (2018).

[29] Jie Fu, Qingqing Ye, Haibo Hu, Zhili Chen, Lulu Wang, Kuncan Wang, and Xun Ran. 2024. DPSUR: Accelerating Differentially Private Stochastic Gradient Descent Using Selective Update and Release. *Proceedings of the VLDB Endowment* 17, 6 (2024), 1200–1213.

[30] Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. 2020. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *NeurIPS* 33 (2020), 15042–15053.

[31] Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. 2021. The heavy-tail phenomenon in SGD. In *ICML*. PMLR, 3964–3975.

[32] Yong Liu Yang Cao Yuhan Liu Ruixuan Liu Hong Chen Haichao Sha, Yuncheng Wu. 2025. *Tailor for Tails: Differentially Private Stochastic Optimization with Heavy Tails via Discriminative Clipping*. https://github.com/NoNameSha/Discriminative_Clipping_DPSGD/blob/main/full_version_DC_DPSGD.pdf Full Version and Supplementary Materials.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[34] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. 2020. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems* 33 (2020), 22205–22216.

[35] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2015. The composition theorem for differential privacy. In *ICML*. PMLR, 1376–1385.

[36] Gautam Kamath, Xingtu Liu, and Huanyu Zhang. 2022. Improved rates for differentially private stochastic convex optimization with heavy-tailed data. In *ICML*. PMLR, 10633–10660.

[37] Ron Kohavi and Barry Becker. 1996. Adult (Census Income) Data Set. https://archive.ics.uci.edu/ml/datasets/adult. UCI Machine Learning Repository.

[38] Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. 2023. Revisiting Gradient Clipping: Stochastic bias and tight convergence guarantees. In *ICML*. PMLR, 17343–17363.

[39] Alex Krizhevsky. 2009. Learning Multiple Layers of Features from Tiny Images. (2009), 32–33.

[40] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[41] Chris Junchi Li. 2018. A note on concentration inequality for vector-valued martingales with weak exponential-type tails. *arXiv preprint arXiv:1809.02495* (2018).

[42] Shaojie Li and Yong Liu. 2022. High probability guarantees for nonconvex stochastic gradient descent with heavy tails. In *ICML*. PMLR, 12931–12963.

[43] Shaojie Li and Yong Liu. 2023. High Probability Analysis for Non-Convex Stochastic Optimization with Clipping. In *ECAI*. IOS Press.

[44] Xiaoyu Li and Francesco Orabona. 2020. A high probability analysis of adaptive sgd with momentum. *arXiv preprint arXiv:2007.14294* (2020).

[45] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. 2022. Large Language Models Can Be Strong Differentially Private Learners. In *ICLR*.

[46] Junxu Liu, Jian Lou, Li Xiong, Jinfei Liu, and Xiaofeng Meng. 2021. Projected federated averaging with heterogeneous differential privacy. *Proceedings of the VLDB Endowment* 15, 4 (2021), 828–840.

[47] Yuhan Liu, Sheng Wang, Yixuan Liu, Feifei Li, and Hong Chen. 2024. Unleash the Power of Ellipsis: Accuracy-Enhanced Sparse Vector Technique with Exponential Noise. *Proceedings of the VLDB Endowment* 18, 2 (2024), 187–199.

[48] Andrew Lowy and Meisam Razaviyayn. 2023. Private stochastic optimization with large worst-case lipschitz parameter: Optimal rates for (non-smooth) convex losses and extension to non-convex losses. In *ALT*. PMLR, 986–1054.

[49] Liam Madden, Emiliano Dall'Anese, and Stephen Becker. 2024. High probability convergence bounds for non-convex stochastic gradient descent with sub-weibull noise. *Journal of Machine Learning Research* 25, 241 (2024), 1–36.

[50] Ilya Mironov. 2017. Rényi differential privacy. In *CSF*. IEEE, 263–275.

[51] Paulo Moro, Paulo Cortez, and Paulo Rita. 2014. Bank Marketing Data Set. https://archive.ics.uci.edu/dataset/222/bank+marketing. UCI Machine Learning Repository.

[52] Meenatchi Sundaram Muthu Selva Annamalai and Emiliano De Cristofaro. 2024. Nearly tight black-box auditing of differentially private machine learning. *Advances in Neural Information Processing Systems* 37 (2024), 131482–131502.

[53] K Papinesi. 2002. Bleu: A method for automatic evaluation of machine translation. In *ACL*. 311–318.

[54] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. 2021. Influence-balanced loss for imbalanced visual classification. In *ICCV*. 735–744.

[55] Haichao Sha, Ruixuan Liu, Yixuan Liu, and Hong Chen. 2023. PCDP-SGD: Improving the Convergence of Differentially Private SGD via Projection in Advance. *arXiv preprint arXiv:2312.03792* (2023).

[56] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. 2019. A tail-index analysis of stochastic gradient noise in deep neural networks. In *ICML*. PMLR, 5827–5837.

[57] Umut Simsekli, Lingjiong Zhu, Yee Whye Teh, and Mert Gurbuzbalaban. 2020. Fractional underdamped langevin dynamics: Retargeting sgd with momentum under heavy-tailed gradient noise. In *ICML*. PMLR, 8970–8980.

[58] Thomas Steinke, Milad Nasr, and Matthew Jagielski. 2023. Privacy auditing with one (1) training run. *Advances in Neural Information Processing Systems* 36 (2023), 49268–49280.

[59] Dajun Sun, Wei Dong, and Ke Yi. 2023. Confidence Intervals for Private Query Processing. *Proceedings of the VLDB Endowment* 17, 3 (2023), 373–385.

[60] Florian Tramer and Dan Boneh. 2021. Differentially Private Learning Needs Better Features (or Much More Data). In *ICLR*.

[61] Roman Vershynin. 2018. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press.

[62] Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. 2020. Sub-Weibull distributions: Generalizing sub-Gaussian and sub-Exponential properties to heavier tailed distributions. *Stat* 9, 1 (2020), e318.

[63] Martin J Wainwright. 2019. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge university press.

[64] Di Wang, Hanshen Xiao, Srinivas Devadas, and Jinhui Xu. 2020. On differentially private stochastic convex optimization with heavy-tailed data. In *ICML*. PMLR, 10081–10091.

[65] Tianhao Wang, Bolin Ding, Jingren Zhou, Cheng Hong, Zhicong Huang, Ninghui Li, and Somesh Jha. 2019. Answering multi-dimensional analytical queries under local differential privacy. In *Proceedings of the 2019 International Conference on Management of Data*. 159–176.

[66] Fei Wei, Ergute Bao, Xiaokui Xiao, Yin Yang, and Bolin Ding. 2024. AAA: an Adaptive Mechanism for Locally Differential Private Mean Estimation. *Proc. VLDB Endow.* 17, 8 (2024), 1843–1855.

[67] Jianxin Wei, Ergute Bao, Xiaokui Xiao, and Yin Yang. 2022. Dpis: An enhanced mechanism for differentially private sgd with importance sampling. In *CCS*. 2885–2899.

[68] William H. Wolberg, W. Nick Street, and Olvi L. Mangasarian. 1995. Breast Cancer Wisconsin (Diagnostic) Data Set. https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic. UCI Machine Learning Repository.

[69] Tianyu Xia, Shuheng Shen, Su Yao, Xinyi Fu, Ke Xu, Xiaolong Xu, and Xing Fu. 2023. Differentially private learning with per-sample adaptive clipping. In *AAAI*, Vol. 37. 10444–10452.

[70] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *CoRR* (2017).

[71] Hanshen Xiao, Zihang Xiang, Di Wang, and Srinivas Devadas. 2023. A theory to instruct differentially-private learning via clipping bias reduction. In *SP*. IEEE, 2170–2189.

[72] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *CVPR*. 1492–1500.

[73] Xiaodong Yang, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. 2022. Normalized/clipped sgd with perturbation for differentially private non-convex optimization. *arXiv preprint arXiv:2206.13033* (2022).

[74] I-Cheng Yeh and Che-hui Lien. 2009. Default of Credit Card Clients Data Set. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients. UCI Machine Learning Repository.

[75] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. 2020. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *ICLR* (2020).

[76] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. 2020. Why are adaptive methods good for attention models? *NeurIPS* 33 (2020), 15383–15393.

[77] Tong Zhang. 2005. Data dependent concentration bounds for sequential prediction algorithms. In *COLT*. Springer, 173–187.

[78] Xinwei Zhang, Zhiqi Bu, Steven Wu, and Mingyi Hong. 2023. Differentially Private SGD Without Clipping Bias: An Error-Feedback Approach. In *ICLR*.

[79] Xinwei Zhang, Xiangyi Chen, Mingyi Hong, Zhiwei Steven Wu, and Jinfeng Yi. 2022. Understanding clipping for federated learning: Convergence and client-level differential privacy. In *ICML*.

[80] Y Zhang, Q Ye, R Chen, H Hu, Q Han, et al. 2023. Trajectory data collection with local differential privacy. *Proceedings of the VLDB Endowment* 16, 10 (2023), 2591–2604.

[81] Yingxue Zhou, Steven Wu, and Arindam Banerjee. 2021. Bypassing the Ambient Dimension: Private SGD with Gradient Subspace Identification. In *ICLR*.

[82] Yuqing Zhu and Yu-Xiang Wang. 2020. Improving sparse vector technique with renyi differential privacy. *NeurIPS* 33 (2020), 20249–20258.

APPENDIX

A Theoretical Foundations and Notations

A.1 Summary of Theoretical Results

Table 6: Summary of state-of-the-art optimization results, where c is the clipping threshold, θ is the heavy tail index, T is the number of iterations, and δ is a small probability. 'Gradient Symmetry' means the gradient noise $\mathcal{G}_t$ satisfies $\mathbb{P}(\mathcal{G}_t) = \mathbb{P}(-\mathcal{G}_t)$. $G_{\min}$ is the minimum Lipschitz constant, $0 < p < 1$ is the tail proportion and $f_c(\theta) := \max\left(\mathbb{O}\left(\log^{\theta}(1/\delta), \log^{\theta}(\sqrt{T})\right)\right)$.

Method	Upper Bound	Loss Function	Gradient Assumption	Clipping Guidance
NSGD [73]	$\mathbb{O}\left(\varphi^{1/2}\right)$	Non-convex	Light-tailed Gradient Noise	Normalized $c = 1$
Auto-S [9]	$\mathbb{O}\left(\varphi\right)$	Non-convex	Light-tailed Gradient Noise; Gradient Symmetry	Normalized $c = 1$
Clipped DPSGD [17]	$\mathbb{O}\left(\varphi^{1-\theta}\right)$	Convex	Heavy-tailed Lipschitz	$c \leq G_{\min}$
Clipped DPSGD [17]	$\mathbb{O}\left(\delta^{-\frac{2\theta}{2-\theta}}\varphi^{1-\frac{\theta}{2-\theta}}\right)$	Non-convex	Heavy-tailed Lipschitz	$c = \mathbb{O}\left((\delta^2\varphi)^{-\frac{\theta}{2-\theta}}\right)$
Our Clipped DPSGD (Thm 3.1)	$\mathbb{O}\left(\log^{\max(1,\theta)}(T/\delta)\log^{2\theta}(\sqrt{T})\varphi^{1/2}\right)$	Non-convex	Heavy-tailed Gradient Noise	$c = f_c(\theta)$
Our DC-DPSGD (Thm 4.2 Cor 4.3)	$p * \mathbb{O}\left(\log^{\max(1,\theta)}(T/\delta)\log^{2\theta}(\sqrt{T})\varphi^{\frac{1}{2}}\right)$ $+(1-p) * \mathbb{O}\left(\log(T/\delta)\log(\sqrt{T})\varphi^{\frac{1}{2}}\right)$	Non-convex	Heavy-tailed Gradient Noise	Tail: $c_1 = f_c(\theta)$ Body: $c_2 = f_c(\frac{1}{2})$

A.2 Summary of notations

B Preliminaries

A random variable X called a sub-Weibull random variable with tail parameter θ and scale factor K, which is denoted by $X \sim subW(\theta, K)$. We next introduce the equivalent properties and theoretical tools of sub-Weibull distributions.

B.1 Properties

Definition B.1 (Sub-Weibull Equivalent Properties [62]). *Let X be a random variable and $\theta \geq 0$, and there exists some constant K_1, K_2, K_3, K_4 depending on θ. Then the following characterizations are equivalent:*

(1) The tails of X satisfy

$$\exists K_1 > 0 \text{ such that } \mathbb{P}(|X| > t) \leq 2\exp(-(t/K_1)^{\frac{1}{\theta}}), \forall t > 0.$$

(2) The moments of X satisfy

$$\exists K_2 > 0 \text{ such that } \|X\|_p \leq K_2 p^{\theta}, \forall k \geq 1.$$

(3) The moment generating function (MGF) of $|X|^{\frac{1}{\theta}}$ satisfies

$$\exists K_3 > 0 \text{ such that } \mathbb{E}[\exp((\lambda|X|)^{\frac{1}{\theta}})] \leq \exp((\lambda K_3)^{\frac{1}{\theta}}), \forall \lambda \in (0, 1/K_3).$$

(4) The MGF of $|X|^{\frac{1}{\theta}}$ is bounded at some point,

$$\exists K_4 > 0 \text{ such that } \mathbb{E}[\exp((|X|/K_4)^{\frac{1}{\theta}})] \leq 2.$$

Table 7: Summary of notations

Definition of Notations	
$\mathbf{w}$	the model parameter
d	the dimension of model parameters
z	the training sample
n	the training data size
B	the mini-batch size
ℓ	the loss function
$S\ S'$	the neighboring datasets
ε_{dp}	the privacy budget for differential privacy
ε_{tr}	the privacy budget for preserving traces
σ_{dp}	the noise multiplier for differential privacy
σ_{tr}	the noise multiplier for preserving traces
$V_{t,k}$	top-k dimensional the random projection vector
K	the variance-related positive constant
$\nabla L(\mathbf{w}_t)$	the true average gradient for training data
T	the total iterations of training
η_t	the learning rate in t iteration
c	the clipping threshold
c_1	the large clipping threshold for heavy tail
c_2	the small clipping threshold for light body
θ	the heavy tail index
p	the proportion of heavy tail
$\lambda_{t,i}^{\text{tr}}$	the empirical trace of the sample
$\hat{\lambda}_{t,i}^{\text{tr}}$	the population trace of the sample
$\widetilde{\lambda}_{t,i}^{\text{tr}}$	the perturbed trace of the sample
S^{tail}	the set of gradients in the heavy tail region
S^{body}	the set of gradients in the light body region
$\mathcal{C}_{gt}$	the gradient noise

B.2 Theoretical Tools

Based on the properties of sub-Weibull variables, we have the following high probability bounds and concentration inequalities for heavier tails as theoretical tools. Besides, We define l_p norm as $\|\|\|_p$, for any $p \geq 1$.

Lemma B.1. *Let a variable $X \sim subW(\theta, K)$, for any $\delta \in (0,1)$, then with probability $(1-\delta)$ we have*

$$|X| \leq K \log^{\theta}(2/\delta). \tag{1}$$

Proof. Let $K_1 = K$ in Definition B.1, and take $t = K\log^{\theta}(2/\delta)$, then the inequality holds with probability $1-\delta$. □

Lemma B.2 ([49, 62]). *Let $X_1, ..., X_n$ are $subW(\theta, K_i)$ random variables with scale parameters $K_1, ...K_n$. $\forall x \geq 0$, we have*

$$\mathbb{P}(|\sum_{i=1}^{n} X_i| \geq x) \leq 2\exp(-(\frac{x}{g(\theta)\sum_{i=1}^{n} K_i})^{\frac{1}{\theta}}), \tag{2}$$

where $g(\theta) = (4e)^{\theta}$ for $\theta \leq 1$ and $g(\theta) = 2(2e\theta)^{\theta}$ for $\theta \geq 1$.

Lemma B.3 (Sub-Weibull Freedman Inequality [49]). *Let $(\Omega, \mathcal{F}, (\mathcal{F}_i), \mathbb{P})$ be a filtered probability space. Let (ξ_i) and (K_i) be adapted to $(\mathcal{F}_i)$. Let $n \in \mathbb{N}$, then $\forall i \in [n]$, assume $K_{i-1} \geq 0$, $\mathbb{E}[\xi_i|\mathcal{F}_{i-1}] = 0$, and $\mathbb{E}[\exp((|\xi_i|/K_{i-1})^{\frac{1}{\theta}})|\mathcal{F}_{i-1}] \leq 2$ where $\theta \geq 1/2$. If $\theta > 1/2$, assume there exists (m_i) such that $K_{i-1} \leq m_i$.*

if $\theta = 1/2$, let $a = 2$, then $\forall x, \beta \geq 0$, $\alpha > 0$, and $\lambda \in [0, \frac{1}{2\alpha}]$,

$$\mathbb{P}\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k}\xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \alpha\sum_{i=1}^{k}\xi_i + \beta\right\}\right) \leq \exp(-\lambda x + 2\lambda^2\beta), \tag{3}$$

and $\forall x, \beta, \lambda \geq 0$,

$$\mathbb{P}\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k}\xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \beta\right\}\right) \leq \exp(-\lambda x + \frac{\lambda^2}{2}\beta). \tag{4}$$

If $\theta \in (\frac{1}{2}, 1]$, *let* $a = (4\theta)^{2\theta}e^2$ *and* $b = (4\theta)^{\theta}e$. $\forall x, \beta \geq 0$, *and* $\alpha \geq b\max_{i\in[n]} m_i$, *and* $\lambda \in [0, \frac{1}{2\alpha}]$,

$$\mathbb{P}\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k}\xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \alpha\sum_{i=1}^{k}\xi_i + \beta\right\}\right) \leq \exp(-\lambda x + 2\lambda^2\beta), \tag{5}$$

and $\forall x, \beta \geq 0$, *and* $\lambda \in [0, \frac{1}{b\max_{i\in[n]} m_i}]$,

$$\mathbb{P}\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k}\xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \beta\right\}\right) \leq \exp(-\lambda x + \frac{\lambda^2}{2}\beta). \tag{6}$$

If $\theta > 1$, *let* $\delta \in (0,1)$. *Let* $a = (2^{2\theta+1}+2)\Gamma(2\theta+1) + 2^{3\theta}\Gamma(3\theta+1)/3$ *and* $b = 2\log n/\delta^{\theta-1}$, *where* $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. $\forall x, \beta \geq 0$, $\alpha \geq b\max_{i\in[n]} m_i$, *and* $\lambda \in [0, \frac{1}{2\alpha}]$,

$$\mathbb{P}\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k}\xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \alpha\sum_{i=1}^{k}\xi_i + \beta\right\}\right) \leq \exp(-\lambda x + 2\lambda^2\beta) + 2\delta, \tag{7}$$

and $\forall x, \beta \geq 0$, *and* $\lambda \in [0, \frac{1}{b\max_{i\in[n]} m_i}]$,

$$\mathbb{P}\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k}\xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \beta\right\}\right) \leq \exp(-\lambda x + \frac{\lambda^2}{2}\beta) + 2\delta. \tag{8}$$

Lemma B.4 ([77]). *Let* $z_1, ..., z_n$ *be a sequence of randoms variables such that* z_k *may depend the previous variables* $z_1, ..., z_{k-1}$ *for all* $k = 1, ..., n$. *Consider a sequence of functionals* $\xi_k(z_1, ..., z_k)$, $k = 1, ..., n$. *Let* $\sigma_n^2 = \sum_{k=1}^n \mathbb{E}_{z_k}[(\xi_k - \mathbb{E}_{z_k}[\xi_k])^2]$ *be the conditional variance. Assume* $|\xi_k - \mathbb{E}_{z_k}[\xi_k]| \leq b$ *for each* k. *Let* $\rho \in (0,1]$ *and* $\delta \in (0,1)$. *With probability at least* $1-\delta$ *we have*

$$\sum_{k=1}^{n}\xi_k - \sum_{k=1}^{n}\mathbb{E}_{z_k}[\xi_k] \leq \frac{\rho\sigma_n^2}{b} + \frac{b\log\frac{1}{\delta}}{\rho}. \tag{9}$$

Lemma B.5 ([15]). *For any vector* $\mathbf{g} \in \mathbb{R}^d$, $\langle \mathbf{g}/\|\mathbf{g}\|_2, \nabla L_S(\mathbf{w})\rangle \geq \frac{\|\nabla L_S(\mathbf{w})\|_2}{3} - \frac{8\|\mathbf{g}-L_S(\mathbf{w})\|_2}{3}$.

Lemma B.6 ([49]). *If* $X \sim subW(\theta, K)$, *then* $\mathbb{E}[|X^p|] \leq 2\Gamma(p\theta+1)K^p$ $\forall p > 0$. *In particular,* $\mathbb{E}[X^2] \leq 2\Gamma(2\theta+1)K^2$.

Lemma B.7 ([4]). *Suppose* $X_1, ..., X_m \overset{d}{=} X$ *are independent and identically distributed random variables whose right tails are captured by an increasing and continuous function* $I : \mathbb{R} \to \mathbb{R}^{\geq 0}$ *with the property* $I(x) = \mathbb{O}(x)$ *as* $x \to \infty$. *Let* $X^L = X\mathbb{I}(X \leq L)$, $S_m = \sum_{i=1}^m X_i$ *and* $Z^L := X^L - \mathbb{E}[X]$. *Define* $x_{\max} := \sup\{x \geq 0 : x \leq \eta v(mx, \eta)\frac{I(mx)}{mx}\}$, *then*

$$\mathbb{P}(S_m - \mathbb{E}[S_m] > mx) \leq \begin{cases} \exp(-c_x\eta I(mx)) + m\exp(-I(mx)), & \text{if } x \geq x_{\max}, \\ \exp(-\frac{mx^2}{2v(mx_{\max},\eta)}) + m\exp(-\frac{mx_{\max}^2(\eta)}{\eta v(mx_{\max},\eta)}), & \text{if } 0 \leq x \leq x_{\max}, \end{cases} \tag{10}$$

where $c_x = 1 - \frac{\eta v(mx,\eta)I(mx)}{2mx^2}$ *and* $v(L,\eta) = \mathbb{E}\big[(Z^L)^2\mathbb{I}(Z^L \leq 0) + (Z^L)^2\exp(\eta\frac{I(L)}{L}Z^L)\mathbb{I}(Z^L > 0)\big], \forall \beta \in (0,1]$.

Lemma B.8 ([4]). *Consider the same settings as the ones in Lemma B.7. Assume* $\mathbb{E}[X_i] = 0$, *then* $\forall t \geq 0$ *we have*

$$\mathbb{P}(S_m > mt) \leq \exp(-\frac{mt^2}{2v(mt,\eta)}) + \exp(-\eta\max\{c_t, \frac{1}{2}\}I(mt)) + m\exp(-I(mt)). \tag{11}$$

Lemma B.9 (**Ahlswede-Winter Inequality**). *Let Y be a random, symmetric, positive semi-definite dd matrix such that $\|\mathbb{E}[Y]\|_2 \leq 1$. Suppose $\|Y\|_2 \leq R$ for some fixed scalar $R \geq 1$. Let $Y_1, ..., Y_m$ be independent copies of Y (i.e., independently sampled matrix with the same distribution as Y). For any $\mu \in (0,1)$, we have*

$$\mathbb{P}(\|\frac{1}{m}\sum_{i=1}^{m} Y_i - \mathbb{E}[Y_i]\|_2 > \mu) \leq 2d \cdot \exp(-m\mu^2/4R). \tag{12}$$

Lemma B.10 ([26, 41]). *Let $\theta \in (0,\infty)$ be given. Assume that $(\mathbf{X}_i, i = 1, \cdots, N)$ is a sequence of $\mathbb{R}^d$-valued martingale differences with respect to filtration $\mathcal{F}_i$, i.e. $\mathbb{E}[\mathbf{X}_i|\mathcal{F}_{i-1}] = 0$, and it satisfies the following weak exponential-type tail condition: for some $\theta > 0$ and all $i = 1, \cdots, N$ we have for some scalar $0 < K_i$,*

$$\mathbb{E}\left[\exp\left(\left\|\frac{\mathbf{X}_i}{K_i}\right\|^{\frac{1}{\theta}}\right)\right] \leq 2.$$

Assume that $K_i < \infty$ for each $i = 1, \cdots, N$. Then for an arbitrary $N \geq 1$ and $t > 0$,

$$\mathbb{P}\left(\max_{n\leq N}\left\|\sum_{i=1}^{n}\mathbf{X}_i\right\| \geq t\right) \leq 4\left[3 + (3\theta)^{2\theta}\frac{128\sum_{i=1}^{N}K_i^2}{t^2}\right]\exp\left\{-\left(\frac{t^2}{64\sum_{i=1}^{N}K_i^2}\right)^{\frac{1}{2\theta+1}}\right\}. \tag{13}$$

C Convergence of Heavy-tailed Clipped DPSGD

Theorem C.1 (**Convergence of Clipped DPSGD under Heavy-tailed Sub-Weibull Gradient Noise Assumption**). *Under Assumptions 2.1 and 2.2, let* $\mathbf{w}_t$ *be the iterative parameter produced by clipped DPSGD of Algorithm 1 with* $T = \mathbb{O}(\frac{n\varepsilon}{\sqrt{d\log(1/\delta)}})$, $T \geq 1$, *and* $\eta_t = \frac{1}{\sqrt{T}}$. *Define* $\hat{\sigma}_{\text{dp}}^2 := m_2\frac{Tdc^2B^2\log(1/\delta)}{n^2\varepsilon^2}$. *If* $\theta = \frac{1}{2}$ *and* $K \leq \hat{\sigma}_{\text{dp}}$, *then* $c = \max\big(4K\log^\theta(\sqrt{T}), \frac{19K\log^{\frac{1}{2}}(1/\delta)}{12}\big)$. *If* $\theta = \frac{1}{2}$ *and* $K \geq \hat{\sigma}_{\text{dp}}$, *then* $c = \max\big(4K\log^\theta(\sqrt{T}), 39K\log^{\frac{1}{2}}(2/\delta)\big)$. *If* $\theta > \frac{1}{2}$, *then* $c = \max\big(4K\log^\theta(\sqrt{T}), 20K\log^\theta(2/\delta)\big)$. *For any* $\delta \in (0,1)$, *with probability* $1-\delta$, *we have*

$$\frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} \leq \mathbb{O}\Big(\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(T/\delta)\hat{\log}(T/\delta)\log^{2\theta}(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}}\Big),$$

where $\hat{\log}(T/\delta) := \log^{\max(1,\theta)}(T/\delta)$.

Proof. We consider two cases: $\|\nabla L_S(\mathbf{w}_t)\|_2 \leq c/2$ and $\|\nabla L_S(\mathbf{w}_t)\|_2 \geq c/2$. To simplify notation, we omit the subscript of privacy parameters throughout, such as ε_{dp}.

Firstly, we first consider the case $\|\nabla L_S(\mathbf{w}_t)\|_2 \leq c/2$.

$$\begin{aligned}
&L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t) \leq \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \nabla L_S(\mathbf{w}_t)\rangle + \frac{1}{2}\beta\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&\leq -\eta_t\langle \overline{\mathbf{g}}_t + \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle + \frac{1}{2}\beta\eta_t^2\|\overline{\mathbf{g}}_t + \zeta_t\|_2^2 \\
&= -\eta_t\langle \overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t] + \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle \\
&- \eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 + \frac{1}{2}\beta\eta_t^2\|\overline{\mathbf{g}}_t\|_2^2 + \frac{1}{2}\beta\eta_t^2\|\zeta_t\|_2^2 + \beta\eta_t^2\langle \overline{\mathbf{g}}_t, \zeta_t\rangle \\
&= -\eta_t\langle \overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t], \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle \\
&- \eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 + \frac{1}{2}\beta\eta_t^2\|\overline{\mathbf{g}}_t\|_2^2 + \frac{1}{2}\beta\eta_t^2\|\zeta_t\|_2^2 + \beta\eta_t^2\langle \overline{\mathbf{g}}_t, \zeta_t\rangle
\end{aligned} \tag{14}$$

Considering all T iterations, we get

$$\begin{aligned}
&\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S) + \sum_{t=1}^{T}\frac{1}{2}\beta\eta_t^2c^2 + \underbrace{\sum_{t=1}^{T}\frac{1}{2}\beta\eta_t^2\|\zeta_t\|_2^2}_{\text{Eq.1}} + \underbrace{\sum_{t=1}^{T}\beta\eta_t^2\langle \overline{\mathbf{g}}_t, \zeta_t\rangle}_{\text{Eq.2}} \\
&\underbrace{-\sum_{t=1}^{T}\eta_t\langle \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle}_{\text{Eq.3}} \underbrace{-\sum_{t=1}^{T}\eta_t\langle \overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t], \nabla L_S(\mathbf{w}_t)\rangle}_{\text{Eq.4}} \underbrace{-\sum_{t=1}^{T}\eta_t\langle \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle}_{\text{Eq.5}}
\end{aligned} \tag{15}$$

For Eq.1, Eq.2 and Eq.3, since $\zeta_t \sim \mathbb{N}(0, c\sigma_{\text{dp}}\mathbb{I}_d)$, according to sub-Gaussian properties and Lemma B.2, with probability at least $1-\delta$, we have

$$\begin{aligned}
\sum_{t=1}^{T}\frac{1}{2}\beta\eta_t^2\|\zeta_t\|_2^2 &\leq 2\beta K^2e\log(2/\delta)\sum_{t=1}^{T}\eta_t^2 \\
&\leq 2\beta m_2ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2}\sum_{t=1}^{T}\eta_t^2.
\end{aligned} \tag{16}$$

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

$$\begin{aligned}
\sum_{t=1}^{T}\beta\eta_t^2\langle \overline{\mathbf{g}}_t, \zeta_t\rangle &\leq \sum_{t=1}^{T}\beta\eta_t^2\|\overline{\mathbf{g}}_t\|_2\|\zeta_t\|_2 \\
&\leq \sum_{t=1}^{T}2\beta cK\sqrt{e}\log^{\frac{1}{2}}(2/\delta)\eta_t^2 \\
&\leq 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon}\sum_{t=1}^{T}\eta_t^2.
\end{aligned} \tag{17}$$

Due to $\nabla L_S(\mathbf{w}_t) \leq c/2$, for the term $-\sum_{t=1}^{T} \eta_t \langle \zeta_t, \nabla L_S(\mathbf{w}_t) \rangle$, with probability at least $1-\delta$, we have

$$\begin{aligned} -\sum_{t=1}^{T} \eta_t \langle \zeta_t, \nabla L_S(\mathbf{w}_t) \rangle &\leq \sum_{t=1}^{T} \eta_t \|\zeta_t\|_2 \|\nabla L_S(\mathbf{w}_t)\|_2 \\ &\leq \sum_{t=1}^{T} 2cK\sqrt{e} \log^{\frac{1}{2}}(2/\delta)\eta_t \\ &\leq 2\sqrt{em_2 Td} \frac{c^2 B \log(2/\delta)}{n\varepsilon} \sum_{t=1}^{T} \eta_t. \end{aligned} \tag{18}$$

Since $\mathbb{E}_t[-\eta_t \langle \overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t], \nabla L_S(\mathbf{w}_t) \rangle] = 0$, the sequence $(-\eta_t \langle \overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t], \nabla L_S(\mathbf{w}_t) \rangle, t \in \mathbb{N})$ is a martingale difference sequence. Applying Lemma B.4, we define $\xi_t = -\eta_t \langle \overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t], \nabla L_S(\mathbf{w}_t) \rangle$ and have

$$|\xi_t| \leq \eta_t (\|\overline{\mathbf{g}}_t\|_2 + \|\mathbb{E}_t[\overline{\mathbf{g}}_t]\|_2) \|\nabla L_S(\mathbf{w}_t)\|_2 \leq \eta_t c^2. \tag{19}$$

Applying $\mathbb{E}_t[(\xi_t - \mathbb{E}_t\xi_t)^2] \leq \mathbb{E}_t[\xi_t^2]$, we have

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{E}_t[(\xi_t - \mathbb{E}_t\xi_t)^2] &\leq \sum_{t=1}^{T} \eta_t^2 \mathbb{E}_t[\|\overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t]\|_2^2 \|\nabla L_S(\mathbf{w}_t)\|_2^2] \\ &\leq 4c^2 \sum_{t=1}^{T} \eta_t^2 \|\nabla L_S(\mathbf{w}_t)\|_2^2. \end{aligned} \tag{20}$$

Then, with probability $1-\delta$, we obtain

$$\sum_{t=1}^{T} \xi_t \leq \frac{\rho 4c^2 \sum_{t=1}^{T} \eta_t^2 \|\nabla L_S(\mathbf{w}_t)\|_2^2}{\eta_t c^2} + \frac{\eta_t c^2 \log(1/\delta)}{\rho}. \tag{21}$$

Next, to bound term Eq.5, we have

$$\sum_{t=1}^{T} \eta_t \langle \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t) \rangle \leq \frac{1}{2} \sum_{t=1}^{T} \eta_t \|\mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t)\|_2^2 + \frac{1}{2} \sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2^2.$$

Setting $a_t = \mathbb{I}_{\|\mathbf{g}_t\|_2 > c}$ and $b_t = \mathbb{I}_{\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > \frac{c}{2}}$, for term $\|\mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t)\|_2$, we have

$$\begin{aligned} \|\mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t)\|_2 &= \|\mathbb{E}_t[(\overline{\mathbf{g}}_t - \mathbf{g}_t)a_t]\|_2 \\ &= \|\mathbb{E}_t[(\mathbf{g}_t(\frac{c}{\|\mathbf{g}_t\|_2} - 1)a_t]\|_2 \\ &\leq \mathbb{E}_t[\|(\mathbf{g}_t(\frac{c}{\|\mathbf{g}_t\|_2} - 1)a_t\|_2] \\ &\leq \mathbb{E}_t[|\|\mathbf{g}_t\|_2 - c|a_t] \\ &\leq \mathbb{E}_t[|\|\mathbf{g}_t\|_2 - \|\nabla L_S(\mathbf{w}_t)\|_2|a_t] \\ &\leq \mathbb{E}_t[|\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2|a_t] \\ &\leq \mathbb{E}_t[|\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2|b_t] \\ &\leq \sqrt{\mathbb{E}_t[\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2^2]\mathbb{E}_t b_t^2}. \end{aligned} \tag{22}$$

Applying Lemma B.6, we get $\mathbb{E}_t[\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2^2] \leq 2K^2\Gamma(2\theta+1)$. Then, for term $\mathbb{E}_t b_t^2$, with sub-Weibull properties and probability $1-\delta$ we have

$$\mathbb{E}_t b_t^2 = \mathbb{P}(\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > \frac{c}{2}) \leq 2\exp(-(\frac{c}{4K})^{\frac{1}{\theta}}) \tag{23}$$

So, we get formula (22) as

$$\sqrt{\mathbb{E}_t[\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2^2]\mathbb{E}_t b_t^2} \leq 2\sqrt{K^2\Gamma(2\theta+1)\exp(-(\frac{c}{4K})^{\frac{1}{\theta}})}. \tag{24}$$

Thus, for Eq.5, with probability $1 - T\delta$ we finally obtain

$$\sum_{t=1}^{T} \eta_t \langle \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t) \rangle \\ \leq 2K^2\Gamma(2\theta+1) \sum_{t=1}^{T} \eta_t \exp(-(\frac{c}{4K})^{\frac{1}{\theta}}) + \frac{1}{2}\sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2^2. \tag{25}$$

Combining Eq.1-5 with the inequality (10), with probability $1 - 4\delta - T\delta$, we have

$$\sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S) + \sum_{t=1}^{T} \frac{1}{2}\beta\eta_t^2 c^2 + 2\beta m_2 e d \frac{T c^2 B^2 \log^2(2/\delta)}{n^2\varepsilon^2} \sum_{t=1}^{T} \eta_t^2 \\ + 2\beta\sqrt{em_2 T d}\frac{c^2 B \log(2/\delta)}{n\varepsilon} \sum_{t=1}^{T} \eta_t^2 + 2\sqrt{em_2 T d}\frac{c^2 B \log(2/\delta)}{n\varepsilon} \sum_{t=1}^{T} \eta_t + \frac{\eta_t c^2 \log(1/\delta)}{\rho} \\ + \frac{4\rho c^2 \sum_{t=1}^{T} \eta_t^2 \|\nabla L_S(\mathbf{w}_t)\|_2^2}{\eta_t c^2} + 2K^2\Gamma(2\theta+1)\exp(-(\frac{c}{4K})^{\frac{1}{\theta}}) \sum_{t=1}^{T} \eta_t + \frac{1}{2}\sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2^2. \tag{26}$$

Setting $\rho = \frac{1}{16}$, $T = \mathbb{O}(\frac{n\varepsilon}{\sqrt{d\log(1/\delta)}})$ and $\eta_t = \frac{1}{\sqrt{T}}$, we have

$$\frac{1}{4}\sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S) + \frac{1}{2}\beta c^2 + 2\beta m_2 e \frac{d^{\frac{1}{2}} c^2 B^2 \log^{\frac{3}{2}}(2/\delta)}{n\varepsilon} \\ + 2\beta\sqrt{em_2}\frac{d^{\frac{1}{4}} c^2 B \log^{\frac{1}{2}}(2/\delta)}{\sqrt{n\varepsilon}} + 2\sqrt{em_2} c^2 B \log^{\frac{1}{2}}(2/\delta) + \frac{16 d^{\frac{1}{4}} c^2 \log^{\frac{5}{4}}(1/\delta)}{\sqrt{n\varepsilon}} \\ \underbrace{+ 2K^2\Gamma(2\theta+1)\exp(-(\frac{c}{4K})^{\frac{1}{\theta}})\sqrt{T}}_{\text{Eq.6}}. \tag{27}$$

Then, we pay attention to term Eq.6. If $c \to 0$, then $\exp(-(\frac{c}{4K})^{\frac{1}{\theta}}) \to 1$ and $\sqrt{T}$ will dominate term Eq.6. We know that in classical clipped DPSGD, a small c is regarded as the clipping threshold guide, which will cause the variance term Eq.6 to dominate the entire bound. For this, we will provide guidance on the clipping values of DPSGD under the heavy-tailed assumption.

Let $\exp(-(\frac{c}{4K})^{\frac{1}{\theta}}) \leq \frac{1}{\sqrt{T}}$, then we have $c \geq 4K\log^{\theta}(\sqrt{T})$. So, we obtain

$$\sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq 4(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) + 2\beta c^2 + 8\beta m_2 e \frac{d^{\frac{1}{2}} c^2 B^2 \log^{\frac{3}{2}}(2/\delta)}{n\varepsilon} \\ + 8\beta\sqrt{em_2}\frac{d^{\frac{1}{4}} c^2 B \log^{\frac{1}{2}}(2/\delta)}{\sqrt{n\varepsilon}} + 8\sqrt{em_2} c^2 B \log^{\frac{1}{2}}(2/\delta) + \frac{64 d^{\frac{1}{4}} c^2 \log^{\frac{5}{4}}(1/\delta)}{\sqrt{n\varepsilon}} + 8K^2\Gamma(2\theta+1). \tag{28}$$

Multiplying $\frac{1}{\sqrt{T}}$ on both sides, we get

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq \frac{1}{\sqrt{T}}\Bigg(4(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) + 2\beta c^2 + 8\beta m_2 e \frac{d^{\frac{1}{2}} c^2 B^2 \log^{\frac{3}{2}}(2/\delta)}{n\varepsilon} \\ +8\beta\sqrt{em_2}\frac{d^{\frac{1}{4}} c^2 B \log^{\frac{1}{2}}(2/\delta)}{\sqrt{n\varepsilon}} + 8\sqrt{em_2} c^2 B \log^{\frac{1}{2}}(2/\delta) + \frac{64 d^{\frac{1}{4}} c^2 \log^{\frac{5}{4}}(1/\delta)}{\sqrt{n\varepsilon}} + 8K^2\Gamma(2\theta+1)\Bigg). \tag{29}$$

Taking $c = 4K\log^{\theta}(\sqrt{T})$, due to $T \geq 1$, we achieve

$$\begin{aligned}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq\ & \frac{4(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S))}{\sqrt{T}} + \frac{8K^2\Gamma(2\theta+1)}{\sqrt{T}} \\ &+ \frac{16K^2\log^{2\theta}(\sqrt{T})\log(2/\delta)}{\sqrt{T}}\left(2\beta + 8\beta m_2 e\frac{d^{\frac{1}{2}}B^2\log^{\frac{1}{2}}(2/\delta)}{n\varepsilon}\right. \\ &\left.+8\beta\sqrt{em_2}\frac{d^{\frac{1}{4}}B\log^{-\frac{1}{2}}(2/\delta)}{\sqrt{n\varepsilon}} + 8\sqrt{em_2}B\log^{-\frac{1}{2}}(2/\delta) + \frac{64d^{\frac{1}{4}}\log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\varepsilon}}\right) \\ \leq\ & \mathbb{O}\left(\frac{\log^{2\theta}(\sqrt{T})\log(1/\delta)}{\sqrt{T}}\cdot\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\varepsilon}}\right) \\ \leq\ & \mathbb{O}\left(\frac{\log^{2\theta}(\sqrt{T})\log(1/\delta)d^{\frac{1}{4}}\log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\varepsilon}}\right). \end{aligned} \tag{30}$$

Due to $\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq \mathbb{O}\left(\frac{d^{\frac{1}{4}}\log^{2\theta}(\sqrt{T})\log^{\frac{5}{4}}(1/\delta)}{(n\varepsilon)^{\frac{1}{2}}}\right), \tag{31}$$

with probability $1 - T\delta - 4\delta$.

By substitution, with probability $1-\delta$, we get

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq \mathbb{O}\left(\frac{d^{\frac{1}{4}}\log^{2\theta}(\sqrt{T})\log^{\frac{5}{4}}(T/\delta)}{(n\varepsilon)^{\frac{1}{2}}}\right). \tag{32}$$

Secondly, we consider the case $\|\nabla L_S(\mathbf{w}_t)\|_2 \geq c/2$.

$$\begin{aligned} L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t) &\leq \langle \mathbf{w}_{t+1}-\mathbf{w}_t, \nabla L_S(\mathbf{w}_t)\rangle + \frac{1}{2}\beta\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2 \\ &\leq \underbrace{-\eta_t\langle \overline{\mathbf{g}}_t + \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle}_{\text{Eq.7}} + \underbrace{\frac{1}{2}\beta\eta_t^2\|\overline{\mathbf{g}}_t+\zeta_t\|_2^2}_{\text{Eq.8}} \end{aligned} \tag{33}$$

We have discussed term Eq.8 in the above case, so we focus on Eq.7 here. Setting $s_t^+ = \mathbb{I}_{\|\mathbf{g}_t\|_2 \geq c}$ and $s_t^- = \mathbb{I}_{\|\mathbf{g}_t\|_2 \leq c}$.

$$\begin{aligned} &-\eta_t\langle \overline{\mathbf{g}}_t + \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle \\ &= -\eta_t\langle \frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2}s_t^+ + \mathbf{g}_t s_t^-, \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle. \end{aligned} \tag{34}$$

Applying Lemma B.5 to term $-\eta_t\langle \frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2}s_t^+, \nabla L_S(\mathbf{w}_t)\rangle$, we have

$$\begin{aligned} -\eta_t\langle \frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2}s_t^+, \nabla L_S(\mathbf{w}_t)\rangle &\leq -\frac{c\eta_t s_t^+\|\nabla L_S(\mathbf{w}_t)\|_2}{3} + \frac{8c\eta_t\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2}{3} \\ &\leq -\frac{c\eta_t(1-s_t^-)\|\nabla L_S(\mathbf{w}_t)\|_2}{3} + \frac{8c\eta_t\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2}{3}. \end{aligned} \tag{35}$$

For term $-\eta_t\langle \mathbf{g}_t s_t^-, \nabla L_S(\mathbf{w}_t)\rangle$, we obtain

$$\begin{aligned} -\eta_t\langle \mathbf{g}_t s_t^-, \nabla L_S(\mathbf{w}_t)\rangle &= -\eta_t s_t^-(\langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle + \|\nabla L_S(\mathbf{w}_t)\|_2^2) \\ &\leq -\eta_t s_t^-(-\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2\|\nabla L_S(\mathbf{w}_t)\|_2 + \|\nabla L_S(\mathbf{w}_t)\|_2^2) \\ &\leq \eta_t\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2\|\nabla L_S(\mathbf{w}_t)\|_2 - \frac{c}{2}\eta_t s_t^-\|\nabla L_S(\mathbf{w}_t)\|_2 \\ &\leq \eta_t\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2\|\nabla L_S(\mathbf{w}_t)\|_2 - \frac{c}{3}\eta_t s_t^-\|\nabla L_S(\mathbf{w}_t)\|_2. \end{aligned} \tag{36}$$

According to Lemma B.1, with probability at least $1-\delta$, we have

$$\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 \leq K\log^{\theta}(2/\delta), \tag{37}$$

then we get

$$-\eta_t\langle \mathbf{g}_t s_t^-, \nabla L_S(\mathbf{w}_t)\rangle \leq K\log^{\theta}(2/\delta)\|\nabla L_S(\mathbf{w}_t)\|_2 - \frac{c}{3}\eta_t s_t^-\|\nabla L_S(\mathbf{w}_t)\|_2, \tag{38}$$

and

$$-\eta_t\langle\frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2}s_t^+,\nabla L_S(\mathbf{w}_t)\rangle\leq-\frac{c\eta_t(1-s_t^-)\|\nabla L_S(\mathbf{w}_t)\|_2}{3}+\frac{8c\eta_tK\log^\theta(2/\delta)}{3}. \tag{39}$$

Using Lemma B.2 to term $-\sum_{t=1}^T\eta_t\langle\zeta_t,\nabla L_S(\mathbf{w}_t)\rangle$, with probability at least $1-\delta$, we have

$$-\sum_{t=1}^T\eta_t\langle\zeta_t,\nabla L_S(\mathbf{w}_t)\rangle\leq4\sqrt{em_2Td}\frac{cB\log(2/\delta)}{n\varepsilon}\sum_{t=1}^T\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2. \tag{40}$$

So, combining formula (38), formula (39) and formula (40) with term Eq.7, with probability at least $1-2\delta-T\delta$, we obtain

$$\begin{aligned}
&-\sum_{t=1}^T\eta_t\langle\overline{\mathbf{g}}_t+\zeta_t,\nabla L_S(\mathbf{w}_t)\rangle\leq-\sum_{t=1}^T\frac{c\eta_t}{3}\|\nabla L_S(\mathbf{w}_t)\|_2+\sum_{t=1}^T\frac{8c\eta_tK\log^\theta(2/\delta)}{3}\\
&+K\log^\theta(2/\delta)\sum_{t=1}^T\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2+4\sqrt{em_2Td}\frac{cB\log(2/\delta)}{n\varepsilon}\sum_{t=1}^T\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2\\
&\leq-\sum_{t=1}^T\frac{c\eta_t}{3}\|\nabla L_S(\mathbf{w}_t)\|_2+(\frac{19}{3}K\log^\theta(2/\delta)+4\sqrt{em_2Td}\frac{cB\log(2/\delta)}{n\varepsilon})\sum_{t=1}^T\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2.
\end{aligned} \tag{41}$$

Next, considering all T iterations and term Eq.8 with $\hat{\sigma}_{\mathrm{dp}}^2:=dc^2\sigma_{\mathrm{dp}}^2=m_2\frac{Tdc^2B^2\log(1/\delta)}{n^2\varepsilon^2}$ and probability $1-4\delta-T\delta$, we have

$$\begin{aligned}
&(\frac{c}{3}-\frac{19}{3}K\log^\theta(2/\delta)-4\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta))\sum_{t=1}^T\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2\leq L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S)\\
&+(2\beta m_2ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2}+2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon}+\frac{1}{2}\beta c^2)\sum_{t=1}^T\eta_t^2.
\end{aligned} \tag{42}$$

If $\theta=\frac{1}{2}$ and $K\geq\hat{\sigma}_{\mathrm{dp}}$, let $\frac{c}{3}\geq\frac{39}{3}K\log^{\frac{1}{2}}(2/\delta)$, i.e. $c\geq39K\log^{\frac{1}{2}}(2/\delta)$, taking $c=39K\log^{\frac{1}{2}}(2/\delta)$, $T=\mathbb{O}(\frac{n\varepsilon}{\sqrt{d\log(1/\delta)}})$ and $\eta_t=\frac{1}{\sqrt{T}}$, we have

$$\begin{aligned}
&\sum_{t=1}^T\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2\leq\frac{3}{K\log^{\frac{1}{2}}(2/\delta)}(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S))\\
&+\frac{3\sum_{t=1}^T\eta_t^2}{K\log^{\frac{1}{2}}(2/\delta)}\left(2\beta m_2ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2}+2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon}+\frac{1}{2}\beta c^2\right)\\
&\leq\frac{L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S)+2\beta e\hat{\sigma}_{\mathrm{dp}}^2\log(2/\delta)+2\beta c\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(2/\delta)+\frac{39^2}{2}\beta K^2\log(2/\delta)}{\frac{1}{3}K\log^{\frac{1}{2}}(2/\delta)}\\
&\leq\frac{3(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S))}{K\log^{\frac{1}{2}}(2/\delta)}+6\beta eK\log^{\frac{1}{2}}(2/\delta)+6\beta\sqrt{e}\log^{\frac{1}{2}}(2/\delta)+3\beta\frac{(39)^2}{2}K\log^{\frac{1}{2}}(2/\delta).
\end{aligned} \tag{43}$$

Thus, with probability $1-4\delta-T\delta$, we have

$$\frac{1}{T}\sum_{t=1}^T\|\nabla L_S(\mathbf{w}_t)\|_2\leq\frac{1}{\sqrt{T}}\sum_{t=1}^T\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2\leq\mathbb{O}(\frac{\log^{\frac{1}{2}}(1/\delta)}{\sqrt{T}})=\mathbb{O}(\frac{\log^{\frac{1}{2}}(1/\delta)d^{\frac{1}{4}}\log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\varepsilon}}),$$

implying that with probability $1-\delta$, we have

$$\frac{1}{T}\sum_{t=1}^T\|\nabla L_S(\mathbf{w}_t)\|_2\leq\mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(T/\delta)}{\sqrt{n\varepsilon}}). \tag{44}$$

If $\theta = \frac{1}{2}$ and $K \leq \hat{\sigma}_{\text{dp}}$, that is, $c \geq \frac{19 \log^{\frac{1}{2}}(1/\delta) K}{12}$, thus there exists $T = \mathbb{O}(\frac{n\varepsilon}{\sqrt{d \log(1/\delta)}})$, $T \geq 1$ and $\eta_t = \frac{1}{\sqrt{T}}$ that we obtain

$$\begin{aligned}
\sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2 &\leq \frac{1}{\sqrt{e}\hat{\sigma}_{\text{dp}} \log^{\frac{1}{2}}(1/\delta)} (L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) \\
&+ \frac{\sum_{t=1}^{T} \eta_t^2}{\sqrt{e}\hat{\sigma}_{\text{dp}} \log^{\frac{1}{2}}(1/\delta)} \left(2\beta m_2 e d \frac{T c^2 B^2 \log^2(2/\delta)}{n^2 \varepsilon^2} + 2\beta \sqrt{e m_2 T d} \frac{c^2 B \log(2/\delta)}{n\varepsilon} + \frac{1}{2}\beta c^2 \right) \\
&\leq \frac{1}{\sqrt{e}\hat{\sigma}_{\text{dp}} \log^{\frac{1}{2}}(1/\delta)} (L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) \\
&+ \frac{\sum_{t=1}^{T} \eta_t^2}{\sqrt{e}\hat{\sigma}_{\text{dp}} \log^{\frac{1}{2}}(1/\delta)} \left(2\beta e \hat{\sigma}_{\text{dp}}^2 \log(2/\delta) + 2\beta \sqrt{e} \hat{\sigma}_{\text{dp}} \log^{\frac{1}{2}}(2/\delta) + \frac{27^2}{2} \beta e \hat{\sigma}_{\text{dp}}^2 \log(2/\delta) \right) \\
&\leq \frac{L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)}{K \log^{\frac{1}{2}}(2/\delta)} + 2\beta e K \log^{\frac{1}{2}}(2/\delta) + 2\beta \sqrt{e} \log^{\frac{1}{2}}(2/\delta) + \beta \frac{(27)^2}{2} K \log^{\frac{1}{2}}(2/\delta).
\end{aligned} \tag{45}$$

Therefore, with probability $1 - 4\delta - T\delta$, we have

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla L_S(\mathbf{w}_t)\|_2 \leq \mathbb{O}(\frac{\log^{\frac{1}{2}}(1/\delta) d^{\frac{1}{4}} \log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\varepsilon}}),$$

then, with probability $1 - \delta$, we have

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla L_S(\mathbf{w}_t)\|_2 \leq \mathbb{O}(\frac{d^{\frac{1}{4}} \log^{\frac{3}{4}}(T/\delta)}{\sqrt{n\varepsilon}}). \tag{46}$$

If $\theta > \frac{1}{2}$, then term $\log^{\theta}(2/\delta)$ dominates the left-hand inequality, i.e. $\frac{19}{3} K \log^{\theta}(2/\delta) \geq 4\sqrt{e}\hat{\sigma}_{\text{dp}} \log^{\frac{1}{2}}(1/\delta)$. Let $\frac{c}{3} \geq \frac{20}{3} K \log^{\theta}(2/\delta)$, $T = \mathbb{O}(\frac{n\varepsilon}{\sqrt{d \log(1/\delta)}})$ and $\eta_t = \frac{1}{\sqrt{T}}$, we obtain

$$\begin{aligned}
\sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2 &\leq \frac{3}{K \log^{\theta}(2/\delta)} (L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) \\
&+ \frac{3 \sum_{t=1}^{T} \eta_t^2}{K \log^{\theta}(2/\delta)} \left(2\beta m_2 e d \frac{T c^2 B^2 \log^2(2/\delta)}{n^2 \varepsilon^2} + 2\beta \sqrt{e m_2 T d} \frac{c^2 B \log(2/\delta)}{n\varepsilon} + \frac{1}{2}\beta c^2 \right) \\
&\leq \frac{3(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S))}{K \log^{\theta}(2/\delta)} + \frac{19^2}{24} \beta K \log^{\theta}(2/\delta) + 190 \beta K \log^{\theta}(2/\delta) + 3\beta (20)^2 K \log^{\theta}(2/\delta).
\end{aligned} \tag{47}$$

Consequently, with probability $1 - \delta$, we have

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla L_S(\mathbf{w}_t)\|_2 \leq \mathbb{O}(\frac{\log^{\theta}(T/\delta) d^{\frac{1}{4}} \log^{\frac{1}{4}}(T/\delta)}{\sqrt{n\varepsilon}}). \tag{48}$$

Integrating the above results, when $\nabla L_S(\mathbf{w}_t) \geq c/2$ we have

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla L_S(\mathbf{w}_t)\|_2 \leq \mathbb{O}(\frac{d^{\frac{1}{4}} \log^{\theta + \frac{1}{4}}(T/\delta)}{\sqrt{n\varepsilon}}), \tag{49}$$

with probability $1 - \delta$ and $\theta \geq \frac{1}{2}$.

To sum up, covering the two cases, we ultimately come to the conclusion with probability $1 - \delta$, $T = \mathbb{O}(\frac{n\varepsilon}{\sqrt{d \log(1/\delta)}})$, $T \geq 1$, and $\eta_t = \frac{1}{\sqrt{T}}$

$$\begin{aligned}
\frac{1}{T} \sum_{t=1}^{T} \min \left\{ \|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2 \right\} &\leq \mathbb{O}(\frac{d^{\frac{1}{4}} \log^{\theta + \frac{1}{4}}(T/\delta)}{(n\varepsilon)^{\frac{1}{2}}}) + \mathbb{O}(\frac{d^{\frac{1}{4}} \log^{2\theta}(\sqrt{T}) \log^{\frac{5}{4}}(T/\delta)}{(n\varepsilon)^{\frac{1}{2}}}) \\
&\leq \mathbb{O}(\frac{d^{\frac{1}{4}} \log^{\frac{5}{4}}(T/\delta) \left(\log^{\theta - 1}(T/\delta) + \log^{2\theta}(\sqrt{T}) \right)}{(n\varepsilon)^{\frac{1}{2}}}) \\
&\leq \mathbb{O}(\frac{d^{\frac{1}{4}} \log^{\frac{1}{4}}(T/\delta) \hat{\log}(T/\delta) \log^{2\theta}(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}}),
\end{aligned} \tag{50}$$

where $\hat{\log}(T/\delta) = \log^{\max(1,\theta)}(T/\delta)$. If $\theta = \frac{1}{2}$ and $K \leq \hat{\sigma}_{\mathrm{dp}}$, then $c = \max\left(4K\log^{\frac{1}{2}}(\sqrt{T}), \frac{19K\log^{\frac{1}{2}}(1/\delta)}{12}\right)$. If $\theta = \frac{1}{2}$ and $K \geq \hat{\sigma}_{\mathrm{dp}}$, then $c = \max\left(4K\log^{\frac{1}{2}}(\sqrt{T}), 39K\log^{\frac{1}{2}}(2/\delta)\right)$. If $\theta > \frac{1}{2}$, then $c = \max\left(4K\log^{\theta}(\sqrt{T}), 20K\log^{\theta}(2/\delta)\right)$. □

The proof of Theorem 3.1 is completed.

D Subspace Skewing for Identification

Theorem D.1 (**Subspace Skewing for Identification**). *Assume that the empirical projection subspace $M = V_{t,k}V_{t,k}^\top \in \mathbb{R}^{d\times d}$ with $V_{t,k}^\top V_{t,k} = \mathbb{I}_k$ approximates the population projection subspace $\hat{M} = \hat{V}_{t,k}\hat{V}_{t,k}^\top = \mathbb{E}_{V_{t,k}\sim\mathcal{P}}[V_{t,k}V_{t,k}^\top]$, $\lambda_{t,i}^{\text{tr}} = \text{tr}(V_{t,k}^\top \hat{\mathbf{g}}_t(z_i)\hat{\mathbf{g}}_t^\top(z_i)V_{t,k})$ and $\hat{\lambda}_{t,i}^{\text{tr}} = \text{tr}(\hat{V}_{t,k}^\top \hat{\mathbf{g}}_t(z_i)\hat{\mathbf{g}}_t^\top(z_i)\hat{V}_{t,k})$, for any gradient $\hat{\mathbf{g}}_t(z_i)$ that satisfies $\|\hat{\mathbf{g}}_t(z_i)\|_2 = 1$, $\zeta_t^{\text{tr}} \sim \mathbb{N}(0, \sigma_{\text{tr}}^2\mathbb{I})$, with probability $1-\delta_m-\delta_{\text{tr}}$, we have*

$$|\lambda_{t,i}^{\text{tr}} - \hat{\lambda}_{t,i}^{\text{tr}} + \zeta_t^{\text{tr}}| \le \frac{4\log(2d/\delta_m)}{k} + \frac{m_2\sqrt{B}\log^{\frac{1}{2}}(1/\delta_{\text{tr}})}{d^{\frac{1}{2}}}.$$

Proof. For simplicity, we abbreviate $\hat{\mathbf{g}}_t(z_i)$ as $\hat{\mathbf{g}}_t$. Due to the Fact.1, $V_{t,k}^\top V_{t,k} = \mathbb{I}$ and $\hat{V}_{t,k}^\top \hat{V}_{t,k} = \mathbb{I}$, we omit subscripts of expectation and have

$$\begin{aligned}
|\lambda_{t,i}^{\text{tr}} - \hat{\lambda}_{t,i}^{\text{tr}}| &:= |\text{tr}(V_{t,k}^\top \hat{\mathbf{g}}_t\hat{\mathbf{g}}_t^\top V_{t,k}) - \text{tr}(\hat{V}_{t,k}^\top \hat{\mathbf{g}}_t\hat{\mathbf{g}}_t^\top \hat{V}_{t,k})| \\
&= |\|V_{t,k}^\top \hat{\mathbf{g}}_t\|_2^2 - \|\hat{V}_{t,k}^\top \hat{\mathbf{g}}_t\|_2^2| \\
&= |\|V_{t,k}V_{t,k}^\top \hat{\mathbf{g}}_t\|_2^2 - \|\hat{V}_{t,k}\hat{V}_{t,k}^\top \hat{\mathbf{g}}_t\|_2^2| \\
&\le \|V_{t,k}V_{t,k}^\top \hat{\mathbf{g}}_t - \hat{V}_{t,k}\hat{V}_{t,k}^\top \hat{\mathbf{g}}_t\|_2^2 \\
&\le \|V_{t,k}V_{t,k}^\top - \hat{V}_{t,k}\hat{V}_{t,k}^\top\|_2^2\|\hat{\mathbf{g}}_t\|_2^2.
\end{aligned} \tag{51}$$

To bound $\mathbb{E}\|V_{t,k}V_{t,k}^\top - \hat{V}_{t,k}\hat{V}_{t,k}^\top\|_2^2$, we need to bound the gap between the sum of the random positive semidefinite matrix $M := V_{t,k}V_{t,k}^\top = \frac{1}{k}\sum_{i=1}^k v_{t,i}v_{t,i}^\top$ and the expectation $\hat{M} := \hat{V}_{t,k}\hat{V}_{t,k}^\top = \mathbb{E}[V_{t,k}V_{t,k}^\top]$.

Due to $\|v_j\|_2 = 1$, we can easily get

$$\begin{aligned}
\|M\|_2 = \|\frac{1}{k}\sum_{i=1}^k v_{t,i}v_{t,i}^\top\|_2 &\le \frac{1}{k}\sum_{i=1}^k \|v_{t,i}v_{t,i}^\top\|_2 \\
&= \sup_{x:\|x\|_2=1}\frac{1}{k}\sum_{i=1}^k x^T v_{t,i}v_{t,i}^\top x \\
&= \sup_{x:\|x\|_2=1}\frac{1}{k}\sum_{i=1}^k \langle x, v_{t,i}\rangle \\
&\le \frac{1}{k}\sum_{i=1}^k \|x\|_2\|v_{t,i}\|_2 \\
&= 1.
\end{aligned} \tag{52}$$

Thus, $\|M\|_2 \le 1$ and $\|\mathbb{E}M\|_2 = \|M\cdot\mathbb{P}(M)\|_2 \le 1$ because of $\mathbb{P}(M) \le 1$.

Then, according to Ahlswede-Winter Inequality with $R = 1$ and $m = k$, we have for any $\mu \in (0,1)$

$$\mathbb{P}(\|M - \hat{M}\|_2 > \mu) \le 2d\cdot\exp(\frac{-k\mu^2}{4}), \tag{53}$$

where d is dimension of gradients. The inequality shows that the bounded spectral norm of random matrix $\|M\|_2$ concentrates around its expectation with high probability $1 - 2d\cdot\exp(-k\mu^2/4)$.

Since $\|M\|_2 \in [0,1]$ and $\|\mathbb{E}M\|_2 \in [0,1]$, $\|M-\hat{M}\|_2$ is always bounded by 1. Therefore, for $\mu \ge 1$, $\|M-\hat{M}\|_2 > u$ holds with probability 0. So that for any $\mu > 0$, we have

$$\mathbb{P}(\|M - \hat{M}\|_2 > 2\sqrt{\frac{\log 2d}{k}}\mu) \le \exp(-\mu^2). \tag{54}$$

Based on the inequality above, with probability $1-\delta_m$, we have

$$\|M - \hat{M}\|_2 \le 2\frac{\log^{\frac{1}{2}}(2d/\delta_m)}{\sqrt{k}}. \tag{55}$$

Next, considering that we have implicitly normalized the term $\|\hat{\mathbf{g}}_t\|_2^2$ by the threshold 1, the upper bound of $\|\hat{\mathbf{g}}_t\|_2^2$ is 1. As a result, we obtain

$$\begin{aligned}
|\lambda_{t,i}^{\mathrm{tr}} - \hat{\lambda}_{t,i}^{\mathrm{tr}}| &\leq \|V_{t,k}V_{t,k}^\top - \hat{V}_{t,k}\hat{V}_{t,k}^\top\|_2^2 \|\hat{\mathbf{g}}_t\|_2^2 \\
&\leq \|V_{t,k}V_{t,k}^\top - \hat{V}_{t,k}\hat{V}_{t,k}^\top\|_2^2 \\
&\leq \|M - \hat{M}\|_2^2 \\
&\leq \frac{4\log(2d/\delta_m)}{k},
\end{aligned} \tag{56}$$

with probability $1 - \delta_m$.

Due to the shared random subspace of per-sample gradient, the exposed trace may pose potential privacy risks. Thus, we add the noise that satisfies differential privacy to the trace $\lambda_{t,i}^{\mathrm{tr}}$, i.e. $\lambda_{t,i}^{\mathrm{tr}} + \zeta_t^{\mathrm{tr}}$. The upper bound of the trace for per-sample gradient is limited to 1, because we normalize per-sample gradient in advance. So, the sensitivity in differential privacy can be regarded as 1, which in fact means $\zeta_t^{\mathrm{tr}} \sim \mathbb{N}(0, \sigma_{\mathrm{tr}}^2\mathbb{I})$. Then, applying Gaussian properties, with probability $1 - \delta_m - \delta_{\mathrm{tr}}$, we have

$$\begin{aligned}
|\lambda_{t,i}^{\mathrm{tr}} - \hat{\lambda}_{t,i}^{\mathrm{tr}} + \zeta_t^{\mathrm{tr}}| &\leq |\lambda_{t,i}^{\mathrm{tr}} - \hat{\lambda}_{t,i}^{\mathrm{tr}}| + |\zeta_t^{\mathrm{tr}}| \\
&\leq \frac{4\log(2d/\delta_m)}{k} + \sigma_{\mathrm{tr}} \log^{\frac{1}{2}}(2/\delta_{\mathrm{tr}}).
\end{aligned} \tag{57}$$

Regarding to $\sigma_{\mathrm{tr}} = \frac{m_2\sqrt{TB\log(1/\delta)}}{n\varepsilon_{\mathrm{tr}}}$, we take T as $\frac{n\varepsilon_{\mathrm{tr}}}{\sqrt{d\log(1/\delta)}}$ to maintain consistency with the context and have

$$\begin{aligned}
|\lambda_{t,i}^{\mathrm{tr}} - \hat{\lambda}_{t,i}^{\mathrm{tr}} + \zeta_t^{\mathrm{tr}}| &\leq \frac{4\log(2d/\delta_m)}{k} + \frac{m_2\sqrt{B}\log^{\frac{3}{4}}(1/\delta_{\mathrm{tr}})}{d^{\frac{1}{4}}\sqrt{n\varepsilon_{\mathrm{tr}}}} \\
&\leq \frac{4\log(2d/\delta_m)}{k} + \frac{m_2\sqrt{B}\log^{\frac{1}{2}}(1/\delta_{\mathrm{tr}})}{d^{\frac{1}{2}}},
\end{aligned}$$

where the last inequality holds due to $T \geq 1$.

Intuitively, the conclusion tells us that, since $\lambda_{t,i}^{\mathrm{tr}}$ is a constant, the scale $\sigma_{\mathrm{tr}}\mathbb{I}_1$ of noise added is actually small compared to the noise $\sigma_{\mathrm{dp}}\mathbb{I}_d$ added to gradients, where the latter has a tricky dependence on the dimension space d. Concretely, comparing the first term $\frac{4\log(2d/\delta_m)}{k}$, we observe that in the second term $\frac{m_2\sqrt{B}\log^{\frac{1}{2}}(1/\delta_{\mathrm{tr}})}{\sqrt{d}}$, the model parameter $d \gg k$, we concerned in private learning and coupled with noise scale, is in the denominator, which is far better than the factor $\log(d)$ in the numerator of the first term. Therefore the term $\frac{4\log(2d/\delta_m)}{k}$ will dominate the error of subspace skewing, and we can control this part of the error by adopting a larger k.

In conclusion, for the per-sample trace, there is a high probability $1 - \delta_m'$, where $\delta_m' = \delta_m + \delta_{\mathrm{tr}}$, that we can accurately identify heavy-tailed samples within a finite and minor error dependent on the factor $\mathbb{O}(\frac{1}{k})$.

□

The proof of Theorem 4.1 is completed.

E Convergence of Discriminative Clipping

In DC-DPSGD, the convergence bounds for the two regions correspond to c_1 and c_2, respectively. First, we optimize the theoretical tools by transforming the concentration inequalities for the sum of sub-Weibull random variables X into two-region versions distinguished by the tail probability $\mathbb{P}(|X| > x)$, namely sub-Gaussian tail decay rate $\exp(-x^2)$ and heavy-tailed decay rate $\exp(-x^{1/\theta})$, $\theta > \frac{1}{2}$. Then, we analyze the high probability bounds for the gradient noise of clipped DPSGD in each region. In the heavy tail region, we make the inequality $\mathbb{P}(\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > c_1) \leq 2\exp(-c_1^{1/\theta})$ hold and derive the dependence of factor $\log^{\theta}(1/\delta)$ for c_1. In the light body region, we have $\mathbb{P}(\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > c_2) \leq 2\exp(-c_2^2)$, resulting in the factor $\log^{1/2}(1/\delta)$ of c_2. Next, we investigate the high probability error on the unbounded clipped DPSGD privacy noise using Gaussian distribution properties. Finally, we integrate the results regarding gradient noise and privacy noise to determine the optimal clipping thresholds for both regions and achieve faster convergence rates for the optimization performance. To simplify the notation, we emphasize the **heavy tail region** to refer to the impact of $\mathbf{g}_t^{\text{tail}}(z_i)$ on the convergence of the model parameters $\mathbf{w}_t$, and the **light body region** to refer to the impact of $\mathbf{g}_t^{\text{body}}(z_i)$ on the $\mathbf{w}_t$, *i.e.*, splitting $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \cdot \text{DC-DPSGD}\left(\mathbf{g}_t^{\text{tail}}(z_i) + \mathbf{g}_t^{\text{body}}(z_i)\right)$ into two regions, each subject to bound separately. In the proof, we take it as a default that the clipping threshold c corresponds to c_1 for the heavy tail region and to c_2 for the light body region.

Theorem E.1 (**Convergence of Discriminative Clipping**). *Under Assumptions 2.1, 2.2 and 2.3, Let $\mathbf{w}_t$ be the iterative parameter produced by discriminative clipping of Algorithm 2 with with $T = \mathbb{O}(\frac{n\varepsilon}{\sqrt{d\log(1/\delta)}})$, $T \geq 1$ and $\eta_t = \frac{1}{\sqrt{T}}$. Define $\hat{\log}(T/\delta) = \log^{\max(1,\theta)}(T/\delta)$, $\hat{\sigma}_{\text{dp}}^2 = m_2\frac{Tc^2dB^2\log(1/\delta)}{n^2\varepsilon^2}$, $a = 2$ if $\theta = 1/2$, $a = (4\theta)^{2\theta}e^2$ if $\theta \in (1/2, 1]$ and $a = (2^{2\theta+1}+2)\Gamma(2\theta+1) + \frac{2^{3\theta}\Gamma(3\theta+1)}{3}$ if $\theta > 1$, for any $\delta \in (0,1)$, with probability $1-\delta$, then we have:*

*(i). **In the heavy tail region** ($c = c_1$):*

$$\frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} \leq \mathbb{O}\left(\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(T/\delta)\hat{\log}(T/\delta)\log^{2\theta}(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}}\right).$$

(1) *If $\theta = \frac{1}{2}$ and $K \leq \hat{\sigma}_{\text{dp}}$, then $c_1 = \max\left(4K\log^{\frac{1}{2}}(\sqrt{T}), \frac{16aK\log^{\frac{1}{2}}(1/\delta)}{12}\right)$.* (2) *If $\theta = \frac{1}{2}$ and $K \geq \hat{\sigma}_{\text{dp}}$, then $c_1 = \max\left(4K\log^{\frac{1}{2}}(\sqrt{T}), 33\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)\right)$.*
(3) *If $\theta > \frac{1}{2}$, then $c_1 = \max\left(4^{\theta}2K\log^{\theta}(\sqrt{T}), 17K\log^{\theta}(2/\delta)\right)$.*

*(ii). **In the light body region** ($c = c_2$):*

$$\frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} \leq \mathbb{O}\left(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta)\log(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}}\right).$$

(1) *If $K \leq \hat{\sigma}_{\text{dp}}$, then $c_2 = \max\left(2\sqrt{2a}K\log^{\frac{1}{2}}(\sqrt{T}), 27\sqrt{e}\hat{\sigma}_{\text{dp}}\log^{\frac{1}{2}}(1/\delta)\right)$.* (2) *If $K \geq \hat{\sigma}_{\text{dp}}$, then $c_2 = \max\left(2\sqrt{2a}K\log^{\frac{1}{2}}(\sqrt{T}), 33\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)\right)$.*

Proof. We review two cases in Discriminative Clipping DPSGD: $\|\nabla L_S(\mathbf{w}_t)\|_2 \leq c/2$ and $\|\nabla L_S(\mathbf{w}_t)\|_2 \geq c/2$. To simplify notation, we write ε_{dp} as ε, omitting the subscript throughout.

Firstly, in the case $\|\nabla L_S(\mathbf{w}_t)\|_2 \leq c/2$:

$$\begin{aligned}
&L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t) \leq \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \nabla L_S(\mathbf{w}_t)\rangle + \frac{1}{2}\beta\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&\leq -\eta_t\langle \overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t], \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle \\
&- \eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 + \frac{1}{2}\beta\eta_t^2\|\overline{\mathbf{g}}_t\|_2^2 + \frac{1}{2}\beta\eta_t^2\|\zeta_t\|_2^2 + \beta\eta_t^2\langle \overline{\mathbf{g}}_t, \zeta_t\rangle
\end{aligned}$$

Applying the properties of Gaussian tails and Lemma B.2 to ζ_t, Lemma B.4 to term $\sum_{t=1}^{T}\eta_t\langle \overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t], \nabla L_S(\mathbf{w}_t)\rangle$, with probability $1-4\delta$, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S) + \sum_{t=1}^{T}\frac{1}{2}\beta\eta_t^2c^2 + 2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2}\sum_{t=1}^{T}\eta_t^2 \\
&+ 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon}\sum_{t=1}^{T}\eta_t^2 + 2\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon}\sum_{t=1}^{T}\eta_t + \frac{\eta_t c^2\log(1/\delta)}{\rho} \\
&+ \frac{4\rho c^2\sum_{t=1}^{T}\eta_t^2\|\nabla L_S(\mathbf{w}_t)\|_2^2}{\eta_t c^2} - \underbrace{\sum_{t=1}^{T}\eta_t\langle \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle}_{\text{Eq.9}}.
\end{aligned} \tag{58}$$

We will consider a truncated version of term Eq.9 in the following. Similarly,

$$\sum_{t=1}^{T}\eta_t\langle\mathbb{E}_t[\overline{\mathbf{g}}_t]-\nabla L_S(\mathbf{w}_t),\nabla L_S(\mathbf{w}_t)\rangle\leq\frac{1}{2}\sum_{t=1}^{T}\eta_t\|\mathbb{E}_t[\overline{\mathbf{g}}_t]-\nabla L_S(\mathbf{w}_t)\|_2^2+\frac{1}{2}\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2.$$

For term $\|\mathbb{E}_t[\overline{\mathbf{g}}_t]-\nabla L_S(\mathbf{w}_t)\|_2$, we also define $a_t=\mathbb{I}_{\|\mathbf{g}_t\|_2>c}$ and $b_t=\mathbb{I}_{\|\mathbf{g}_t-\nabla L_S(\mathbf{w}_t)\|_2>\frac{c}{2}}$, and have

$$\begin{aligned}\|\mathbb{E}_t[\overline{\mathbf{g}}_t]-\nabla L_S(\mathbf{w}_t)\|_2&=\|\mathbb{E}_t[(\overline{\mathbf{g}}_t-\mathbf{g}_t)a_t]\|_2\\&\leq\mathbb{E}_t[\|(\mathbf{g}_t(\frac{c-\|\mathbf{g}_t\|_2}{\|\mathbf{g}_t\|_2})a_t\|_2]\\&\leq\mathbb{E}_t[|\|\mathbf{g}_t\|_2-\|\nabla L_S(\mathbf{w}_t)\|_2|a_t]\\&\leq\mathbb{E}_t[|\|\mathbf{g}_t-\nabla L_S(\mathbf{w}_t)\|_2|b_t]\\&\leq\sqrt{\mathbb{E}_t[\|\mathbf{g}_t-\nabla L_S(\mathbf{w}_t)\|_2^2]\mathbb{E}_t b_t^2}.\end{aligned}\tag{59}$$

Due to $\mathbb{E}[\mathbf{g}_t-\nabla L_S(\mathbf{w}_t)]=0$, applying Lemma B.7 and B.8 with

$$\begin{aligned}&m=1\\&\sup_{\eta\in(0,1]}\{v(L,\eta)\}=aK^2\\&x_{\max}=\frac{\eta I(x)}{x}aK^2\\&c_t\in[\frac{1}{2},1]\\&\eta=\frac{1}{2}.\end{aligned}$$

In the light body region, i.e. $x\geq x_{\max}$, we have

$$\begin{aligned}\mathbb{P}(\|\mathbf{g}_t-\nabla L_S(\mathbf{w}_t)\|_2>x)&\leq\exp(-c_t\eta I(x))+\exp(-I(x))\\&\leq\exp(-\frac{1}{4}I(x))+\exp(-I(x))\\&\leq 2\exp(-\frac{1}{4}I(x)).\end{aligned}\tag{60}$$

Then, in the heavy tail region, i.e. $0\leq x\leq x_{\max}$, the inequality

$$\begin{aligned}\mathbb{P}(\|\mathbf{g}_t-\nabla L_S(\mathbf{w}_t)\|_2>x)&\leq\exp(-\frac{x^2}{2v(x_{\max},\eta)})+m\exp(-\frac{x_{\max}^2(\eta)}{\eta v(x_{\max},\eta)})\\&\leq 2\exp(-\frac{x^2}{2v(x_{\max},\eta)})\\&\leq 2\exp(-\frac{x^2}{2aK^2})\end{aligned}\tag{61}$$

holds.

Therefore, when $0\leq x\leq x_{\max}$, we have the follow-up truncated conclusions:
If $\theta=\frac{1}{2}$, $\forall\alpha>0$ and $a=2$, we have the following inequality with probability at least $1-\delta$

$$\|\mathbf{g}_t-\nabla L_S(\mathbf{w}_t)\|_2\leq 2K\log^{\frac{1}{2}}(2/\delta).$$

If $\theta\in(\frac{1}{2},1]$, let $a=(4\theta)^{2\theta}e^2$, we have the following inequality with probability at least $1-\delta$

$$\|\mathbf{g}_t-\nabla L_S(\mathbf{w}_t)\|_2\leq\sqrt{2}e(4\theta)^{\theta}K\log^{\frac{1}{2}}(2/\delta).$$

If $\theta>1$, let $a=(2^{2\theta+1}+2)\Gamma(2\theta+1)+\frac{2^{3\theta}\Gamma(3\theta+1)}{3}$, we have the following inequality with probability at least $1-\delta$

$$\|\mathbf{g}_t-\nabla L_S(\mathbf{w}_t)\|_2\leq\sqrt{2(2^{2\theta+1}+2)\Gamma(2\theta+1)+\frac{2^{3\theta}\Gamma(3\theta+1)}{3}}K\log^{\frac{1}{2}}(2/\delta).$$

When $x\geq x_{\max}$, let $I(x)=(x/K)^{\frac{1}{\theta}}$, $\forall\theta\in(\frac{1}{2},1]$, with probability at least $1-\delta$, then we have

$$\|\mathbf{g}_t-\nabla L_S(\mathbf{w}_t)\|_2\leq 4^{\theta}K\log^{\theta}(2/\delta).$$

Apply the truncated corollary above, when $0 \le x \le x_{\max}$, we have

$$\mathbb{E}_t[\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2] \le \sqrt{2a}K \tag{62}$$

and with probability $1-\delta$,

$$\mathbb{E}_t b_t^2 = \mathbb{P}(\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > \frac{c}{2}) \le 2\exp(-(\frac{c}{2\sqrt{2a}K})^2) \tag{63}$$

where $a = 2$ if $\theta = 1/2$, $a = (4\theta)^{2\theta}e^2$ if $\theta \in (1/2, 1]$ and $a = (2^{2\theta+1}+2)\Gamma(2\theta+1) + \frac{2^{3\theta}\Gamma(3\theta+1)}{3}$ if $\theta > 1$.

When $x \ge x_{\max}$, the inequalities

$$\mathbb{E}_t[\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2] \le 4^{\theta}K \tag{64}$$

and

$$\mathbb{E}_t b_t^2 = \mathbb{P}(\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > \frac{c}{2}) \le 2\exp(-\frac{1}{4}(\frac{c}{2K})^{\frac{1}{\theta}}) \tag{65}$$

hold with probability $1-\delta$, where $\theta \ge \frac{1}{2}$.

Thus, with probability $1-T\delta$, we get

$$\sum_{t=1}^{T}\eta_t\langle\mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle \le 2aK^2\sum_{t=1}^{T}\eta_t\exp(-(\frac{c}{2\sqrt{2a}K})^2) + \frac{1}{2}\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2, \tag{66}$$

when $0 \le x \le x_{\max}$.

With probability $1-T\delta$, we obtain

$$\sum_{t=1}^{T}\eta_t\langle\mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle \le 4^{2\theta}K^2\sum_{t=1}^{T}\eta_t\exp(-\frac{1}{4}(\frac{c}{2K})^{\frac{1}{\theta}}) + \frac{1}{2}\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2, \tag{67}$$

when $x \ge x_{\max}$.

By setting $\rho = \frac{1}{16}$, $T = \mathbb{O}(\frac{n\varepsilon}{\sqrt{d\log(1/\delta)}})$ and $\eta_t = \frac{1}{\sqrt{T}}$, with probability $1-4\delta-T\delta$, we have

$$\begin{aligned}
&\frac{1}{4}\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 \le L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S) + \frac{1}{2}\beta c^2 + 2\beta m_2 e\frac{d^{\frac{1}{2}}c^2B^2\log^{\frac{3}{2}}(2/\delta)}{n\varepsilon}\\
&+ 2\beta\sqrt{em_2}\frac{d^{\frac{1}{4}}c^2B\log^{\frac{1}{2}}(2/\delta)}{\sqrt{n\varepsilon}} + 2\sqrt{em_2}c^2B\log^{\frac{1}{2}}(2/\delta) + \frac{16d^{\frac{1}{4}}c^2\log^{\frac{5}{4}}(1/\delta)}{\sqrt{n\varepsilon}}\\
&+ \text{Eq.10}\begin{cases} 2aK^2\sum_{t=1}^{T}\eta_t\exp(-(\frac{c}{2\sqrt{2a}K})^2), & \text{if } 0 \le x \le x_{\max},\\ 4^{2\theta}K^2\sum_{t=1}^{T}\eta_t\exp(-\frac{1}{4}(\frac{c}{2K})^{\frac{1}{\theta}}), & \text{if } x \ge x_{\max}.\end{cases}
\end{aligned} \tag{68}$$

Let the term Eq.10 $\le \frac{1}{\sqrt{T}}$, and we have $c \ge 2\sqrt{2a}K\log^{\frac{1}{2}}(\sqrt{T})$ if $0 \le x \le x_{\max}$ and $c \ge 4^{\theta}2K\log^{\theta}(\sqrt{T})$ if $x \ge x_{\max}$.

In the light body region that $0 \le x \le x_{\max}$, by taking $c_2 = c = 2\sqrt{2a}K\log^{\frac{1}{2}}(\sqrt{T})$ we achieve

$$\begin{aligned}
\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 \le{}& \frac{4(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S))}{\sqrt{T}} + \frac{2aK^2}{\sqrt{T}}\\
&+ \frac{8aK^2\log(\sqrt{T})\log(2/\delta)}{\sqrt{T}}\left(2\beta + 8\beta m_2 eB^2(\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(2/\delta)}{\sqrt{n\varepsilon}})^2\right.\\
&\left.+8\beta\sqrt{em_2}\frac{d^{\frac{1}{4}}B\log^{-\frac{1}{2}}(2/\delta)}{\sqrt{n\varepsilon}} + 8\sqrt{em_2}B\log^{-\frac{1}{2}}(2/\delta) + \frac{64d^{\frac{1}{4}}\log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\varepsilon}}\right)\\
\le{}& \mathbb{O}(\frac{\log(\sqrt{T})\log(1/\delta)}{\sqrt{T}}\cdot\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\varepsilon}})\\
\le{}& \mathbb{O}(\frac{\log(\sqrt{T})d^{\frac{1}{4}}\log^{\frac{5}{4}}(1/\delta)}{\sqrt{n\varepsilon}}).
\end{aligned} \tag{69}$$

In the heavy tail region that $x \geq x_{\max}$, by taking $c_1 = c = 4^{\theta} 2K \log^{\theta}(\sqrt{T})$ we achieve

$$\begin{aligned}
\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq & \frac{4(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S))}{\sqrt{T}} + \frac{2aK^2}{\sqrt{T}} \\
& + \frac{4^{2\theta+1} \log^{2\theta}(\sqrt{T}) \log(2/\delta)}{\sqrt{T}} \left(2\beta + 8\beta m_2 e B^2 (\frac{d^{\frac{1}{4}} \log^{\frac{1}{4}}(2/\delta)}{\sqrt{n\varepsilon}})^2 \right. \\
& \left. + 8\beta\sqrt{em_2} \frac{d^{\frac{1}{4}} B \log^{-\frac{1}{2}}(2/\delta)}{\sqrt{n\varepsilon}} + 8\sqrt{em_2} B \log^{-\frac{1}{2}}(2/\delta) + \frac{64 d^{\frac{1}{4}} \log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\varepsilon}} \right) \\
\leq & \mathbb{O}(\frac{\log^{2\theta}(\sqrt{T}) \log(1/\delta)}{\sqrt{T}} \cdot \frac{d^{\frac{1}{4}} \log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\varepsilon}}) \\
\leq & \mathbb{O}(\frac{\log^{2\theta}(\sqrt{T}) d^{\frac{1}{4}} \log^{\frac{5}{4}}(1/\delta)}{\sqrt{n\varepsilon}}).
\end{aligned} \tag{70}$$

Secondly, we pay extra attention to the bound in the case $\|\nabla L_S(\mathbf{w}_t)\|_2 \geq c/2$.

$$\begin{aligned}
L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t) \leq & \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \nabla L_S(\mathbf{w}_t) \rangle + \frac{1}{2}\beta \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
\leq & \underbrace{-\eta_t \langle \overline{\mathbf{g}}_t + \zeta_t, \nabla L_S(\mathbf{w}_t) \rangle}_{\text{Eq.11}} + \frac{1}{2}\beta \eta_t^2 \|\overline{\mathbf{g}}_t + \zeta_t\|_2^2.
\end{aligned} \tag{71}$$

We revisit term Eq.11 in the case and also set $s_t^+ = \mathbb{I}_{\|\mathbf{g}_t\|_2 \geq c}$ and $s_t^- = \mathbb{I}_{\|\mathbf{g}_t\|_2 \leq c}$.

$$-\eta_t \langle \overline{\mathbf{g}}_t + \zeta_t, \nabla L_S(\mathbf{w}_t) \rangle = -\eta_t \langle \frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2} s_t^+ + \mathbf{g}_t s_t^-, \nabla L_S(\mathbf{w}_t) \rangle - \eta_t \langle \zeta_t, \nabla L_S(\mathbf{w}_t) \rangle. \tag{72}$$

For term $-\sum_{t=1}^{T} \eta_t \langle \mathbf{g}_t s_t^-, \nabla L_S(\mathbf{w}_t) \rangle$, we obtain

$$\begin{aligned}
-\sum_{t=1}^{T} \eta_t \langle \mathbf{g}_t s_t^-, \nabla L_S(\mathbf{w}_t) \rangle = & -\sum_{t=1}^{T} \eta_t s_t^- (\langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t) \rangle + \|\nabla L_S(\mathbf{w}_t)\|_2^2) \\
\leq & -\sum_{t=1}^{T} \eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t) \rangle - \sum_{t=1}^{T} \eta_t s_t^- \|\nabla L_S(\mathbf{w}_t)\|_2^2 \\
\leq & -\sum_{t=1}^{T} \eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t) \rangle - \frac{c}{2} \sum_{t=1}^{T} \eta_t s_t^- \|\nabla L_S(\mathbf{w}_t)\|_2^2 \\
\leq & \underbrace{-\sum_{t=1}^{T} \eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t) \rangle}_{\text{Eq.12}} - \frac{c}{3} \sum_{t=1}^{T} \eta_t s_t^- \|\nabla L_S(\mathbf{w}_t)\|_2^2.
\end{aligned} \tag{73}$$

Let consider the term Eq.12. Since $\mathbb{E}_t[\eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t) \rangle] = 0$, the sequence $(-\eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t) \rangle, t \in \mathbb{N})$ is a martingale difference sequence. In addition, the term $\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)$ is a $subW(\theta, K)$ random variable, thus we apply sub-Weibull Freedman inequality with Lemma B.3 and concentration inequality with Lemma B.7 and B.8 to bound it.

In Lemma B.3, Define

$$v(L, \eta) := \mathbb{E}\big[(X^L - \mathbb{E}[X])^2 \mathbb{I}(X^L \leq \mathbb{E}[X])\big] + \mathbb{E}\big[(X^L - \mathbb{E}[X])^2 \exp\big(\eta(X^L - \mathbb{E}[X])\big) \mathbb{I}(X^L > \mathbb{E}[X])\big],$$

and make $\beta = kv(L, \eta)$, then we have $\sup_{\eta \in (0,1]} \{kv(L, \eta)\} = a \sum_{i=1}^{k} K_i^2$ based on Lemma B.7 and B.8 in [4] and obtain

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{k \in \mathbb{N}} \left\{ \sum_{i=1}^{k} \xi_i \geq kx \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \beta \right\} \right) \leq & \exp(-\lambda kx + \frac{\lambda^2}{2}\beta) \\
= & \exp(-\lambda kx + kv(L, \eta)\frac{\lambda^2}{2}).
\end{aligned} \tag{74}$$

Subsequently, we define the inflection point $x_{\max} := \frac{\eta I(kx)}{kx} a \sum_{i=1}^{k} K_i^2$ and have

(1) In the light body region where $x \geq x_{\max}$, we choose $L = kx$ and $\lambda = \frac{\eta I(kx)}{kx}$, that is $\frac{x}{v(kx,\eta)} \geq \frac{x_{\max}}{v(kx,\eta)} = \frac{\eta I(kx)}{kx}$. Then the inequality achieves

$$\begin{aligned}\mathbb{P}\left(\bigcup_{k\in\mathbb{N}}\left\{\sum_{i=1}^{k}\xi_i \geq kx \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \beta\right\}\right) &\leq \exp(-\eta I(kx) + v(L,\eta)\frac{\eta^2 I^2(kx)}{2kx^2}) \\ &\leq \exp(-\eta I(kx)(1 - v(L,\eta)\frac{\eta I(kx)}{2kx^2})) \\ &\leq \exp(-\eta c_x I(kx)) \\ &\leq \exp(-\frac{1}{2}\eta I(kx)),\end{aligned} \tag{75}$$

where $c_x = 1 - \frac{\eta v(kx,\eta) I(kx)}{2kx^2}$ and the last inequality holds due to $c_x \geq \frac{1}{2}$.

(2) In the heavy tail region where $x \leq x_{\max}$, we choose $L = kx_{\max}$ and $\lambda = \frac{x}{v(L,\eta)} \leq \frac{x_{\max}}{v(L,\eta)} = \frac{\eta I(L)}{L}$. Then, we get

$$\begin{aligned}\mathbb{P}\left(\bigcup_{k\in\mathbb{N}}\left\{\sum_{i=1}^{k}\xi_i \geq kx \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \beta\right\}\right) &\leq \exp(-\frac{kx^2}{v(L,\eta)} + \frac{kx^2}{2v(L,\eta)}) \\ &\leq \exp(-\frac{kx^2}{2v(L,\eta)}).\end{aligned} \tag{76}$$

Implementing the above inferences and propositions with

$$\begin{aligned}&\xi_t = \eta_t \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle \\ &\Lambda := -\sum_{i=1}^{T} \eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle \\ &K_{t-1} = \eta_t K \|\nabla L_S(\mathbf{w}_t)\|_2 \\ &m_t = \eta_t KG \\ &k = T \\ &\eta = 1/2\end{aligned}$$

If $\theta = \frac{1}{2}$, $\forall \alpha > 0$ and $a = 2$, when $x \leq x_{\max}$ we have the following inequality with probability at least $1 - \delta$

$$\begin{aligned}-\sum_{t=1}^{T} \eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle &\leq \sqrt{2Tv(L,\eta)}\log^{\frac{1}{2}}(1/\delta) \\ &\leq \sqrt{2a\sum_{t=1}^{T} K_t^2}\log^{\frac{1}{2}}(1/\delta) \\ &\leq 2\sqrt{\sum_{t=1}^{T}\eta_t^2 K^2 \|\nabla L_S(\mathbf{w}_t)\|_2^2}\log^{\frac{1}{2}}(1/\delta) \\ &\leq 2KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta),\end{aligned} \tag{77}$$

when $x \geq x_{\max}$, with $I(Tx) = (Tx/\sum_{i=1}^{T} K_i)^2$, we have

$$\begin{aligned}-\sum_{t=1}^{T} \eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle &\leq 4^{\frac{1}{2}}\frac{1}{T}\sum_{t=1}^{T} K_t \log^{\frac{1}{2}}(1/\delta) \\ &\leq 2\frac{KG}{T}\sum_{t=1}^{T}\eta_t \log^{\frac{1}{2}}(1/\delta).\end{aligned} \tag{78}$$

If $\theta \in (\frac{1}{2}, 1]$, let $a = (4\theta)^{2\theta}e^2$, when $x \le x_{\max}$ we have the following inequality with probability at least $1-\delta$

$$-\sum_{t=1}^{T}\eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle \le \sqrt{2a\sum_{t=1}^{T}K_t^2}\log^{\frac{1}{2}}(1/\delta)$$
$$\le \sqrt{2}(4\theta)^{\theta}eKG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta), \tag{79}$$

when $x \ge x_{\max}$, let $I(Tx) = (Tx/\sum_{i=1}^{T}K_i)^{\frac{1}{\theta}},\ \forall\theta \in (\frac{1}{2}, 1]$, then we have

$$-\sum_{t=1}^{T}\eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle \le \frac{4^{\theta}}{T}\sum_{t=1}^{T}K_t\log^{\frac{1}{2}}(1/\delta)$$
$$\le \frac{4^{\theta}KG}{T}\sum_{t=1}^{T}\eta_t\log^{\theta}(1/\delta). \tag{80}$$

If $\theta > 1$, let $a = (2^{2\theta+1}+2)\Gamma(2\theta+1) + \frac{2^{3\theta}\Gamma(3\theta+1)}{3}$, when $x \le x_{\max}$ we have the following inequality with probability at least $1-3\delta$

$$-\sum_{t=1}^{T}\eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle \le \sqrt{2a\sum_{t=1}^{T}K_t^2}\log^{\frac{1}{2}}(1/\delta)$$
$$\le \sqrt{2(2^{2\theta+1}+2)\Gamma(2\theta+1) + \frac{2^{3\theta}\Gamma(3\theta+1)}{3}}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta), \tag{81}$$

when $x \ge x_{\max}$, let $I(Tx) = (Tx/\sum_{i=1}^{T}K_i)^{\frac{1}{\theta}},\ \forall\theta > 1$, then we have

$$-\sum_{t=1}^{T}\eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle \le \frac{4^{\theta}}{T}\sum_{t=1}^{T}K_t\log^{\frac{1}{2}}(1/\delta)$$
$$\le \frac{4^{\theta}KG}{T}\sum_{t=1}^{T}\eta_t\log^{\theta}(1/\delta). \tag{82}$$

To continue the proof, employing Lemma B.5 in term $-\eta_t\langle\frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2}s_t^+, \nabla L_S(\mathbf{w}_t)\rangle$ and covering all T iterations, we have

$$-\sum_{t=1}^{T}\eta_t\langle\frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2}s_t^+, \nabla L_S(\mathbf{w}_t)\rangle \le -\frac{c\sum_{t=1}^{T}\eta_t s_t^+\|\nabla L_S(\mathbf{w}_t)\|_2}{3} + \frac{8c\sum_{t=1}^{T}\eta_t\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2}{3}$$
$$\le -\frac{c\sum_{t=1}^{T}\eta_t(1-s_t^-)\|\nabla L_S(\mathbf{w}_t)\|_2}{3}$$
$$+\frac{16\sum_{t=1}^{T}\eta_t\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2\|\nabla L_S(\mathbf{w}_t)\|_2}{3}. \tag{83}$$

With the truncated corollaries above, we have

(1) If $0 \le x \le x_{\max}$, with probability at least $1-3\delta$

$$-\sum_{t=1}^{T}\eta_t\langle\frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2}s_t^+, \nabla L_S(\mathbf{w}_t)\rangle \le -\frac{c\sum_{t=1}^{T}\eta_t(1-s_t^-)\|\nabla L_S(\mathbf{w}_t)\|_2}{3}$$
$$+\frac{16\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2}{3}\begin{cases} 2K\log^{\frac{1}{2}}(2/\delta), & \text{if } \theta = \frac{1}{2}, \\ \sqrt{2}e(4\theta)^{\theta}K\log^{\frac{1}{2}}(2/\delta), & \text{if } \theta \in (\frac{1}{2}, 1], \\ \sqrt{2(2^{2\theta+1}+2)\Gamma(2\theta+1) + \frac{2^{3\theta}\Gamma(3\theta+1)}{3}}K\log^{\frac{1}{2}}(2/\delta) & \text{if } \theta > 1. \end{cases} \tag{84}$$

(2) If $x \geq x_{\max}$ and $\theta \geq \frac{1}{2}$, with probability at least $1-3\delta$

$$
\begin{aligned}
&-\sum_{t=1}^{T}\eta_t\langle\frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2}s_t^+,\nabla L_S(\mathbf{w}_t)\rangle \leq -\frac{c\sum_{t=1}^{T}\eta_t(1-s_t^-)\|\nabla L_S(\mathbf{w}_t)\|_2}{3}\\
&+\frac{16\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2}{3}4^{\theta}K\log^{\theta}(2/\delta).
\end{aligned}
\tag{85}
$$

Then, according to Lemma B.1, combining the truncated results of $-\sum_{t=1}^{T}\eta_t\langle\mathbf{g}_t s_t^-,\nabla L_S(\mathbf{w}_t)\rangle$ and $-\sum_{t=1}^{T}\eta_t\langle\frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2}s_t^+,\nabla L_S(\mathbf{w}_t)\rangle$, we have the inequality:

(1) If $0 \leq x \leq x_{\max}$, with probability at least $1-3\delta-T\delta$

$$
\begin{aligned}
&-\sum_{t=1}^{T}\eta_t\langle\overline{\mathbf{g}}_t,\nabla L_S(\mathbf{w}_t)\rangle \leq -\frac{c\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2}{3}\\
&+\begin{cases}
2KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta), & \text{if } \theta=\frac{1}{2},\\
\sqrt{2}(4\theta)^{\theta}eKG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta), & \text{if } \theta\in(\frac{1}{2},1],\\
\sqrt{2(2^{2\theta+1}+2)\Gamma(2\theta+1)+\frac{2^{3\theta}\Gamma(3\theta+1)}{3}}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta) & \text{if } \theta>1.
\end{cases}\\
&+\frac{16\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2}{3}\begin{cases}
2K\log^{\frac{1}{2}}(2/\delta), & \text{if } \theta=\frac{1}{2},\\
\sqrt{2}e(4\theta)^{\theta}K\log^{\frac{1}{2}}(2/\delta), & \text{if } \theta\in(\frac{1}{2},1],\\
\sqrt{2(2^{2\theta+1}+2)\Gamma(2\theta+1)+\frac{2^{3\theta}\Gamma(3\theta+1)}{3}}K\log^{\frac{1}{2}}(2/\delta) & \text{if } \theta>1.
\end{cases}
\end{aligned}
\tag{86}
$$

(2) If $x \geq x_{\max}$ and $\theta \geq \frac{1}{2}$, with probability at least $1-3\delta-T\delta$

$$
\begin{aligned}
&-\sum_{t=1}^{T}\eta_t\langle\overline{\mathbf{g}}_t,\nabla L_S(\mathbf{w}_t)\rangle \leq -\frac{c\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2}{3}+\frac{4^{\theta}KG}{T}\sum_{t=1}^{T}\eta_t\log^{\theta}(1/\delta)\\
&+\frac{16\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2}{3}4^{\theta}K\log^{\theta}(2/\delta).
\end{aligned}
\tag{87}
$$

Therefore, we refer to formula (14) and formula (15), and apply Lemma B.2 due to $\zeta_t \sim \mathbb{N}(0, c\sigma_{\mathrm{dp}}\mathbb{I}_d)$. Then, to simplify the notation, we define $\hat{\sigma}_{\mathrm{dp}}^2 = dc^2\sigma_{\mathrm{dp}}^2$. With $\hat{\sigma}_{\mathrm{dp}}^2 = m_2\frac{Tc^2dB^2\log(1/\delta)}{n^2\varepsilon^2}$ and probability $1-6\delta-T\delta$, if $0 \leq x \leq x_{\max}$, we have

$$
\begin{aligned}
&(\frac{c}{3}-\frac{16}{3}aK\log^{\frac{1}{2}}(2/\delta)-4\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta))\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 \leq L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S)\\
&+(2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2}+2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon}+\frac{1}{2}\beta c^2)\sum_{t=1}^{T}\eta_t^2\\
&+\sqrt{2a}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta),
\end{aligned}
\tag{88}
$$

if $x \le x_{\max}$, we have

$$
\begin{aligned}
&(\frac{c}{3} - \frac{16}{3}aK\log^{\theta}(2/\delta) - 4\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta))\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 \le L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)\\
&+ (2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon} + \frac{1}{2}\beta c^2)\sum_{t=1}^{T}\eta_t^2\\
&+ \sqrt{2a}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\theta}(1/\delta),
\end{aligned}
\tag{89}
$$

where $a = 2$ if $\theta = 1/2$, $a = (4\theta)^{2\theta}e^2$ if $\theta \in (1/2, 1]$ and $a = (2^{2\theta+1}+2)\Gamma(2\theta+1) + \frac{2^{3\theta}\Gamma(3\theta+1)}{3}$ if $\theta > 1$.

Afterwards,

(1) In case of light body, when $0 \le x \le x_{\max}$ and $\theta \ge \frac{1}{2}$:

If $K \ge \hat{\sigma}_{\mathrm{dp}}$, let $\frac{c}{3} \ge \frac{33}{3}\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)$, $T = \mathbb{O}(\frac{n\varepsilon}{\sqrt{d\log(1/\delta)}})$ and $\eta_t = \frac{1}{\sqrt{T}}$, we obtain

$$
\begin{aligned}
&\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 \le \frac{3}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) + \frac{3\sqrt{2a}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}\\
&+ \frac{3\sum_{t=1}^{T}\eta_t^2}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}\left(2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon} + \frac{1}{2}\beta c^2\right)\\
&\le \frac{3(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S))}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} + \frac{3\sqrt{2a}KG\log^{\frac{1}{2}}(1/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}\\
&+ \frac{6\beta ea^2K^2\log(2/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} + \frac{6\beta\sqrt{e}\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} + \frac{3\beta(33\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta))^2}{2\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}.
\end{aligned}
\tag{90}
$$

Therefore, with probability at least $1 - 6\delta - T\delta$, we have

$$
\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2 \le \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(1/\delta)}{\sqrt{n\varepsilon}}),
$$

then, with probability $1 - \delta$, we have

$$
\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2 \le \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(T/\delta)}{\sqrt{n\varepsilon}}).
\tag{91}
$$

If $K \le \hat{\sigma}_{\mathrm{dp}}$, let $\frac{c}{3} \ge 9\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta)$, that is, $c \ge 27\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta)$, thus there exists $T = \mathbb{O}(\frac{n\varepsilon}{\sqrt{d\log(1/\delta)}})$, $T \ge 1$ and $\eta_t = \frac{1}{\sqrt{T}}$ that we obtain

$$
\begin{aligned}
&\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 \le \frac{1}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta)}(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) + \frac{\sqrt{2a}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta)}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta)}\\
&+ \frac{\sum_{t=1}^{T}\eta_t^2}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta)}\left(2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon} + \frac{1}{2}\beta c^2\right)\\
&\le \frac{L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(2/\delta)} + \frac{\sqrt{2a}KG}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}} + 2\beta eK\log^{\frac{1}{2}}(2/\delta) + 2\beta\sqrt{e}\log^{\frac{1}{2}}(2/\delta) + \beta\frac{(27)^2}{2}K\log^{\frac{1}{2}}(2/\delta).
\end{aligned}
\tag{92}
$$

Therefore, with probability $1 - 6\delta - T\delta$, we have

$$
\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2 \le \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(1/\delta)}{\sqrt{n\varepsilon}}),
$$

then, with probability $1 - \delta$, we have

$$
\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2 \le \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(T/\delta)}{\sqrt{n\varepsilon}}).
\tag{93}
$$

(2) In case of heavy tail, when $x \geq x_{\max}$:
If $\theta = \frac{1}{2}$ and $K \geq \hat{\sigma}_{\mathrm{dp}}$, let $\frac{c}{3} \geq \frac{33}{3}\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)$, $T = \mathbb{O}(\frac{n\varepsilon}{\sqrt{d\log(1/\delta)}})$ and $\eta_t = \frac{1}{\sqrt{T}}$, we obtain

$$\begin{aligned}
&\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \frac{3}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S)) + \frac{3\sqrt{2a}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} \\
&+ \frac{3\sum_{t=1}^{T}\eta_t^2}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}\left(2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon} + \frac{1}{2}\beta c^2\right) \\
&\leq \frac{3(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S))}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} + \frac{3\sqrt{2a}KG\log^{\frac{1}{2}}(1/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} \\
&+ \frac{6\beta ea^2K^2\log(2/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} + \frac{6\beta\sqrt{e}\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} + \frac{3\beta(33\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta))^2}{2\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}.
\end{aligned} \tag{94}$$

Therefore, with probability at least $1-6\delta-T\delta$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(1/\delta)}{\sqrt{n\varepsilon}}),$$

then, with probability $1-\delta$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(T/\delta)}{\sqrt{n\varepsilon}}). \tag{95}$$

If $\theta = \frac{1}{2}$ and $K \leq \hat{\sigma}_{\mathrm{dp}}$, that is, $c \geq \frac{16aK\log^{\frac{1}{2}}(1/\delta)}{12}$, thus there exists $T = \mathbb{O}(\frac{n\varepsilon}{\sqrt{d\log(1/\delta)}})$, $T \geq 1$ and $\eta_t = \frac{1}{\sqrt{T}}$ that we obtain

$$\begin{aligned}
&\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \frac{1}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta)}(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S)) + \frac{\sqrt{2a}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta)}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta)} \\
&+ \frac{\sum_{t=1}^{T}\eta_t^2}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta)}\left(2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon} + \frac{1}{2}\beta c^2\right) \\
&\leq \frac{L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S)}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(2/\delta)} + \frac{\sqrt{2a}KG}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}} + 2\beta eK\log^{\frac{1}{2}}(2/\delta) + 2\beta\sqrt{e}\log^{\frac{1}{2}}(2/\delta) + \beta\frac{(27)^2}{2}K\log^{\frac{1}{2}}(2/\delta).
\end{aligned} \tag{96}$$

Therefore, with probability $1-6\delta-T\delta$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(1/\delta)}{\sqrt{n\varepsilon}}),$$

then, with probability $1-\delta$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(T/\delta)}{\sqrt{n\varepsilon}}). \tag{97}$$

If $\theta > \frac{1}{2}$, then term $\log^{\theta}(2/\delta)$ dominates the inequality. Let $\frac{c}{3} \geq \frac{17}{3}K\log^{\theta}(2/\delta)$, $T = \mathbb{O}(\frac{n\varepsilon}{\sqrt{d\log(1/\delta)}})$ and $\eta_t = \frac{1}{\sqrt{T}}$, we obtain

$$\begin{aligned}
&\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \frac{3}{\sqrt{2a}K\log^{\theta}(2/\delta)}(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S)) + \frac{3\sqrt{2a}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\theta}(1/\delta)}{\sqrt{2a}K\log^{\theta}(2/\delta)} \\
&+ \frac{3\sum_{t=1}^{T}\eta_t^2}{\sqrt{2a}K\log^{\theta}(2/\delta)}\left(2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon} + \frac{1}{2}\beta c^2\right) \\
&\leq \frac{3(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S))}{\sqrt{2a}K\log^{\theta}(2/\delta)} + 3G + \frac{16^2}{24}\beta K\log^{\theta}(2/\delta) + 136\beta K\log^{\theta}(2/\delta) + 3\beta(17)^2K\log^{\theta}(2/\delta).
\end{aligned} \tag{98}$$

As a result, with probability $1-\delta$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \mathbb{O}(\frac{\log^{\theta}(T/\delta)d^{\frac{1}{4}}\log^{\frac{1}{4}}(T/\delta)}{\sqrt{n\varepsilon}}). \tag{99}$$

Consequently, integrate the above results on the condition that $\nabla L_S(\mathbf{w}_t) \geq c/2$.

For light body, we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(T/\delta)}{\sqrt{n\varepsilon}}), \tag{100}$$

For heavy tail, we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\theta+\frac{1}{4}}(T/\delta)}{\sqrt{n\varepsilon}}), \tag{101}$$

with probability $1-\delta$ and $\theta \geq \frac{1}{2}$.

In a word, covering the two cases, we ultimately come to the conclusion with probability $1-\delta$, $T = \mathbb{O}(\frac{n\varepsilon}{\sqrt{d\log(1/\delta)}})$, $T \geq 1$ and $\eta_t = \frac{1}{\sqrt{T}}$:

1. In the heavy tail region:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} &\leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\theta+\frac{1}{4}}(T/\delta)}{(n\varepsilon)^{\frac{1}{2}}}) + \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{2\theta}(\sqrt{T})\log^{\frac{5}{4}}(T/\delta)}{(n\varepsilon)^{\frac{1}{2}}}) \\ &\leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(T/\delta)\left(\log^{\theta}(T/\delta)+\log^{2\theta}(\sqrt{T})\log(T/\delta)\right)}{(n\varepsilon)^{\frac{1}{2}}}) \\ &\leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(T/\delta)\hat{\log}(T/\delta)\log^{2\theta}(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}}),\end{aligned} \tag{102}$$

where $\hat{\log}(T/\delta) = \log^{\max(1,\theta)}(T/\delta)$. If $\theta = \frac{1}{2}$ and $K \leq \hat{\sigma}_{\mathrm{dp}}$, then $c_1 = \max\left(4K\log^{\frac{1}{2}}(\sqrt{T}), \frac{16aK\log^{\frac{1}{2}}(1/\delta)}{12}\right)$. If $\theta = \frac{1}{2}$ and $K \geq \hat{\sigma}_{\mathrm{dp}}$, then $c_1 = \max\left(4K\log^{\frac{1}{2}}(\sqrt{T}), 33\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)\right)$. If $\theta > \frac{1}{2}$, then $c_1 = \max\left(4^{\theta}2K\log^{\theta}(\sqrt{T}), 17K\log^{\theta}(2/\delta)\right)$.

2. In the light body region:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} &\leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(T/\delta)}{(n\varepsilon)^{\frac{1}{2}}}) + \mathbb{O}(\frac{d^{\frac{1}{4}}\log(\sqrt{T})\log^{\frac{5}{4}}(T/\delta)}{(n\varepsilon)^{\frac{1}{2}}}) \\ &\leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(T/\delta)\left(\log^{\frac{1}{2}}(T/\delta)+\log(\sqrt{T})\log(T/\delta)\right)}{(n\varepsilon)^{\frac{1}{2}}}) \\ &\leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta)\log(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}}),\end{aligned} \tag{103}$$

where if $K \leq \hat{\sigma}_{\mathrm{dp}}$, then $c_2 = \max\left(2\sqrt{2a}K\log^{\frac{1}{2}}(\sqrt{T}), 27\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta)\right)$. If $K \geq \hat{\sigma}_{\mathrm{dp}}$, then $c_2 = \max\left(2\sqrt{2a}K\log^{\frac{1}{2}}(\sqrt{T}), 33\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)\right)$. □

The proof of Theorem 4.2 is completed.

F Union Bound (Formal Version) for Discriminative Clipping DPSGD

Corollary F.1 (**Union Bound (Formal Version) for Discriminative Clipping DPSGD**). *Let $\mathbf{w}_t$ be the iterative parameter produced by DC-DPSGD. Under Assumptions 2.1, 2.2 and 2.3, combining Theorem 2 and Theorem 3, for any $\delta' \in (0,1)$, with probability $1-\delta'$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} \leq p * \mathbb{O}\left(\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(T/\delta')\hat{\log}(T/\delta')\log^{2\theta}(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}}\right) + (1-p)*\mathbb{O}\left(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta')\log(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}}\right),$$

where $\delta' = \delta'_m + \delta$, $\hat{\log}(T/\delta') = \log^{\max(1,\theta)}(T/\delta')$ and p is the proportion of heavy-tailed samples.

Proof. We combine the subspace skewing error (Theorem 4.1) with the optimization bound of Discriminative Clipping DPSGD (Theorem 4.2) in this section to align with our algorithm outline. We have already discussed the error of traces in previous chapters and considered the condition of additional noise that satisfies DP, obtaining an upper bound on the error that depends on the factor $\mathbb{O}(\frac{1}{\sqrt{d}})$. This conclusion means that, the divergence between the empirical trace $\lambda_{t,i}^{\text{tr}}$ and the true trace $\hat{\lambda}_{t,i}^{\text{tr}}$ under the high probability guarantee of $1-\delta'_m$, we can accurately identify the trace of the per-sample gradient with minimal error, and classify gradients into the light body and heavy tail based on the metric.

Specifically, based on statistical characteristics, approximately 5% -10% of the data will fall into the tail part. Thus, we select the top p% samples in the trace ranking as the tailed samples, where $p \in [0.05, 0.1]$. Furthermore, based on the relationship between trace and variance, the pn-th of sorted trace $\lambda_t^{\text{tr},p}$ can be seen as the inflection point $x_{\max}$ of distribution defined in truncated theories B.7 and B.8, which corresponds to the empirical sample results with theoretical population variance and the approximation error has bounded in Theorem 4.1. Therefore, in discriminative clipping DPSGD, we can accurately partition the sample into the heavy-tailed convergence bound with a high probability of $(1-\delta'_m)*p$, and exactly induce the sample to the bound of light bodies with a high probability of $(1-\delta'_m)*(1-p)$, while there is a discrimination error with probability δ'_m. Accordingly, we have

$$\begin{aligned}
&\mathcal{C}_{\text{m}}(c_1,c_2) \coloneqq \frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} \\
&= (1-\delta'_m)*p*\mathcal{C}_{\text{tail}}(c_1) + (1-\delta'_m)*(1-p)*\mathcal{C}_{\text{body}}(c_2) + \delta'_m * |\mathcal{C}_{\text{tail}}(c_1) - \mathcal{C}_{\text{body}}(c_2)|. \qquad (104)
\end{aligned}$$

where $\mathcal{C}_{\text{tail}}(c_1)$ means the convergence bound of $\frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\}$ when $\lambda_{t,i}^{\text{tr}} \geq \lambda_t^{\text{tr},p}$, i.e. $\mathbb{O}\left(\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(T/\delta)\hat{\log}(1/\delta)\log^{2\theta}(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}}\right)$, $\mathcal{C}_{\text{body}}(c_2)$ denotes the bound of $\frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\}$ when $0 \leq \lambda_{t,i}^{\text{tr}} \leq \lambda_t^{\text{tr},p}$ i.e. $\mathbb{O}\left(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta)\log(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}}\right)$, with $c_1 = 4^{\theta}2K\log^{\theta}(\sqrt{T})$ and $c_2 = 2\sqrt{2a}K\log^{\frac{1}{2}}(\sqrt{T})$.

If $\theta = \frac{1}{2}$, then $\mathcal{C}_{\text{tail}}(c_1) = \mathcal{C}_{\text{body}}(c_2)$ and $\delta'_m \to 0$, thus we have

$$\mathcal{C}_{\text{m}}(c_1,c_2) = \mathcal{C}_{\text{tail}}(c_1) = \mathbb{O}\left(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta)\log(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}}\right). \qquad (105)$$

If $\theta > \frac{1}{2}$, then $\mathcal{C}_{\text{tail}}(c_1) \geq \mathcal{C}_{\text{body}}(c_2)$, and we need to proof that $\mathcal{C}_{\text{tail}}(c_1) \geq \mathcal{C}_{\text{m}}(c_1,c_2)$, i.e.

$$\begin{aligned}
\mathcal{C}_{\text{tail}}(c_1) &\geq \mathcal{C}_{\text{m}}(c_1,c_2) \\
&\geq (1-\delta'_m)*p*\mathcal{C}_{\text{tail}}(c_1) + (1-\delta'_m)*(1-p)*\mathcal{C}_{\text{body}}(c_2) + \delta'_m*|\mathcal{C}_{\text{tail}}(c_1) - \mathcal{C}_{\text{body}}(c_2)|.
\end{aligned}$$

By transposition, we have

$$(1-\delta'_m)(1-p)*\mathcal{C}_{\text{tail}}(c_1) + \delta'_m*\mathcal{C}_{\text{body}}(c_2) \geq (1-\delta'_m)*(1-p)*\mathcal{C}_{\text{body}}(c_2).$$

Then, we have

$$\mathcal{C}_{\text{tail}}(c_1) \geq \mathcal{C}_{\text{body}}(c_2) - \frac{\delta'_m}{(1-\delta'_m)*(1-p)}\mathcal{C}_{\text{body}}(c_2), \qquad (106)$$

due to $\frac{\delta'_m}{(1-\delta'_m)*(1-p)} \geq 0$, it is proved that $\mathcal{C}_{\text{tail}}(c_1) \geq \mathcal{C}_{\text{m}}(c_1,c_2)$.

From another perspective, for $\mathcal{C}_{\text{m}}(c_1,c_2)$, with probability $1-\delta'_m$, we have

$$\mathcal{C}_{\text{m}}(c_1,c_2) = p*\mathcal{C}_{\text{tail}}(c_1) + *(1-p)*\mathcal{C}_{\text{body}}(c_2). \qquad (107)$$

In other words, for the formula (104), we define $\delta' = \delta'_m + \delta$. Then, with probability $1 - \delta'$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} \leq p * \mathbb{O}\left(\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(T/\delta')\hat{\log}(T/\delta')\log^{2\theta}(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}}\right) + (1-p) * \mathbb{O}\left(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta')\log(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}}\right), \tag{108}$$

where $\hat{\log}(T/\delta') = \log^{\max(1,\theta)}(T/\delta')$.

□

Thus, if $p \leq \frac{1}{\mathbb{O}(\log^{\max(0,\theta-1)}(T/\delta')\log^{2\theta-1}(\sqrt{T}))+1}$ and $p \leq 1$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} \leq \mathbb{O}\left(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta')\log(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}}\right).$$

The proof of Corollary 4.3 is completed.

G Privacy Guarantee

G.1 Privacy Analysis of Sampling Mechanism

Theorem G.1 (Noise scaling under partitioned sampling). *Under the same privacy budget ε, the partitioned mechanism requires a noise multiplier that requires*

$$\sigma_{\mathrm{dp}} \approx \sqrt{\frac{pq_1^2 + (1-p)q_2^2}{\bar{q}^{\,2}}}\,\sigma_{\mathrm{Pois}}. \tag{109}$$

Equality holds if and only if $q_1 = q_2 = \bar{q}$.

Proof. Denote by $\varepsilon_{\mathrm{Pois}}(\alpha, q, \sigma)$ the Rényi Differential Privacy (RDP) cost of a Poisson-subsampled Gaussian mechanism with sampling rate q and noise scale σ at order $\alpha > 1$.

(1) RDP upper bound for partitioned sampling. Consider a partitioned mechanism where the dataset is divided into a *tail subset* (sampling rate q_1) and a *body subset* (sampling rate q_2), with mixing probability p for the tail subset. The total RDP of this mixed mechanism is upper bounded by

$$\varepsilon_{\mathrm{dp}}(\alpha, \sigma) = \frac{1}{\alpha - 1} \log\left(p\, e^{(\alpha-1)\varepsilon_{\mathrm{Pois}}(\alpha, q_1, \sigma)} + (1-p)\, e^{(\alpha-1)\varepsilon_{\mathrm{Pois}}(\alpha, q_2, \sigma)}\right). \tag{110}$$

(2) Convexity in sampling rate. The function $\varepsilon_{\mathrm{Pois}}(\alpha, q, \sigma)$ is monotonically increasing and convex in q. Let $\phi(q) = \exp((\alpha-1)\varepsilon_{\mathrm{Pois}}(\alpha, q, \sigma))$. By Jensen's inequality,

$$p\,\phi(q_1) + (1-p)\,\phi(q_2) \geq \phi(pq_1 + (1-p)q_2) = \phi(\bar{q}), \tag{111}$$

where $\bar{q} = pq_1 + (1-p)q_2$ denotes the average sampling rate.

Substituting (111) into (110), we have

$$\varepsilon_{\mathrm{dp}}(\alpha, \sigma) \geq \varepsilon_{\mathrm{Pois}}(\alpha, \bar{q}, \sigma). \tag{112}$$

Hence, under the same noise scale σ, the per-step RDP of the partitioned mechanism is almost the same as that of Poisson sampling with the same average rate $\bar{q}$, which shares an approximately equivalent level of privacy amplification with uniform sampling without replacement.

Consequently, to achieve an identical target privacy loss ε, the required noise scale must satisfy

$$\sigma_{\mathrm{dp}} \geq \sigma_{\mathrm{Pois}}, \qquad \text{with equality iff } q_1 = q_2 = \bar{q}. \tag{113}$$

(3) Closed-form ratio under small sampling rate approximation. For small sampling rate $q \ll 1$, the RDP of the Poisson-subsampled Gaussian mechanism can be approximated by

$$\varepsilon_{\mathrm{Pois}}(\alpha, q, \sigma) \approx \frac{\alpha}{2\sigma^2} q^2. \tag{114}$$

Substituting into (110), we obtain

$$\varepsilon_{\mathrm{dp}}(\alpha, \sigma) \approx \frac{\alpha}{2\sigma^2}\left(pq_1^2 + (1-p)q_2^2\right), \quad \varepsilon_{\mathrm{Pois}}(\alpha, \sigma) \approx \frac{\alpha}{2\sigma^2}\bar{q}^{\,2}. \tag{115}$$

Equating their privacy losses $\varepsilon_{\mathrm{dp}}(\alpha, \sigma_{\mathrm{dp}}) = \varepsilon_{\mathrm{Pois}}(\alpha, \sigma_{\mathrm{Pois}})$, we have

$$\sigma_{\mathrm{dp}} \approx \sigma_{\mathrm{Pois}} \sqrt{\frac{pq_1^2 + (1-p)q_2^2}{\bar{q}^{\,2}}} \geq \sigma_{\mathrm{Pois}}. \tag{116}$$

The inequality follows from Jensen's inequality, $pq_1^2 + (1-p)q_2^2 \geq (pq_1 + (1-p)q_2)^2 = \bar{q}^{\,2}$.

(4) Conclusion. Therefore, to maintain the same privacy level ε, the partitioned mechanism must employ a noise scale at least as large as that of Poisson sampling:

$$\sigma_{\mathrm{dp}} \approx \sigma_{\mathrm{Pois}} \sqrt{\frac{pq_1^2 + (1-p)q_2^2}{\bar{q}^{\,2}}}.$$

□

G.2 Privacy Guarantee of DC-DPSGD

Next, we provide the complete privacy guarantee proof of Theorem 4.5 for our differential private mechanism M': Subsample∘TraceSorting (TS)∘ GradientPerturbation (GP). The specific proof process is as follows, and our proof comprehensively encompasses mechanism M':

- **TraceSorting**: We prove that TraceSorting is $(\varepsilon_{\mathrm{tr}}, \delta_{\mathrm{tr}})$-DP.
- **TraceSorting∘GradientPerturbation**: We prove that based on the results of TraceSorting, with two different clipping threshold, the unified composition of TraceSorting and GradientPerturbation is $(\varepsilon_{\mathrm{tr}} + \varepsilon_{\mathrm{dp}}, \delta)$-DP, where $\delta = \delta_{\mathrm{tr}} + \delta_{\mathrm{dp}}$.
- **Subsample∘TraceSorting∘GradientPerturbation**: We prove that, under the premise of subsampling, the privacy amplification effect remains valid for our composition mechanism.

Proof. **(1) *Firstly***, we show the TraceSorting with Gaussian noise here is $(\varepsilon_{\text{tr}}, \delta_{\text{tr}})$-DP and follow the proof of Report Noisy Argmax (RNA) in Claim 3.9 [24] to clarify that. Our trace sorting is to choose traces ranked from 1 to pn. To prove that this process satisfies differential privacy (DP), we need to demonstrate that the method of Report i-th Noisy Argmax for any $i \in \mathbb{Z}^+$ and $i \in (0, m]$ is $(\varepsilon_{\text{tr}}, \delta_{\text{tr}})$-DP, where m is sample size. Fix the neighboring datasets $S = S' \cup \{a\}$. Let λ, respectively λ', denote the vector of traces when the dataset is S, respectively S'. We have discussed the default L_2 sensitivity is 1 and use two properties:

(1) **Monotonicity of Traces.** For all $j \in [m]$, $\lambda_j \geq \lambda_j^{'}$;

(2) **Lipschitz Property.** For all $j \in [m]$, $1 + \lambda_j^{'} \geq \lambda_j$.

Fix any $i \in [m]$. We will bound from above and below the ratio of the probabilities that i is selected with S and with S'. Fix r_{-i}^{+}, a set from $\text{Gauss}(1/\varepsilon_{\text{tr}})^{m-i}$ used for all the noisy traces greater than the i-th trace. Defines r_{-i}^{-}, a set from $\text{Gauss}(1/\varepsilon_{\text{tr}})^{i-1}$ used for all the noisy traces less than the i-th trace. We will argue for each $r_{-i} = r_{-i}^{+} \cup r_{-i}^{-}$ independently. We use the notation $\mathbb{P}[i \mid \xi]$ to mean the probability that the output of the Report Noisy Max algorithm is i, conditioned on ξ.

We first argue that $\mathbb{P}[i \mid S, r_{-i}^{-}] \leq e_{\text{tr}}^{\varepsilon}\mathbb{P}[i \mid S', r_{-i}^{-}] + \delta_{\text{tr}}$. Define

$$r^* = \min_{r_i} : \lambda_i + r_i > \lambda_j + r_j \quad \forall j \in \arg(r_{-i}^{-}).$$

Note that, having fixed r_{-i}^{-}, i will be the output (the i-th argmax noisy trace) when the dataset is S if and only if $r_i \geq r^*$. We have, for all $j \in \arg(r_{-i}^{-})$:

$$\begin{aligned} &\lambda_i + r^* > \lambda_j + r_j \\ \Rightarrow\ &(1 + \lambda_i^{'}) + r^* \geq \lambda_i + r^* > \lambda_j + r_j \geq \lambda_j^{'} + r_j \\ \Rightarrow\ &\lambda_i^{'} + (r^* + 1) > \lambda_j^{'} + r_j. \end{aligned}$$

Thus, if $r_i \geq r^* + 1$, then the i-th trace will be the i-th maximum on one side when the dataset is S' and the noise vector is (r_i, r_{-i}^{-}). The probabilities below are over the choice of $r_i \sim \text{Gauss}(1/\varepsilon_{\text{tr}})$, then with probability $1 - \delta_{\text{tr}}$:

$$\begin{aligned} &\mathbb{P}[r_i \geq 1 + r^*] \geq e^{-\varepsilon_{\text{tr}}}\mathbb{P}[r_i \geq r^*] = e^{-\varepsilon_{\text{tr}}}\mathbb{P}[i \mid S, r_{-i}^{-}] \\ \Rightarrow\ &\mathbb{P}[i \mid S', r_{-i}^{-}] \geq \mathbb{P}[r_i \geq 1 + r^*] \geq e^{-\varepsilon_{\text{tr}}}\mathbb{P}[r_i \geq r^*] = e^{-\varepsilon_{\text{tr}}}\mathbb{P}[i \mid S, r_{-i}^{-}], \end{aligned}$$

which, after multiplying through by $e_{\text{tr}}^{\varepsilon}$ and adding probability δ for $\mathbb{P}[r^* - r_i \geq 1] \leq \delta_{\text{tr}}$, yields what we wanted to show:

$$\mathbb{P}[i \mid S, r_{-i}^{-}] \leq e_{\text{tr}}^{\varepsilon}\mathbb{P}[i \mid S', r_{-i}^{-}] + \delta_{\text{tr}}.$$

Then, we argue that $\mathbb{P}[i \mid S, r_{-i}^{+}] \leq e_{\text{tr}}^{\varepsilon}\mathbb{P}[i \mid S', r_{-i}^{+}] + \delta_{\text{tr}}$. Define

$$r^* = \max_{r_i} : \lambda_i + r_i < \lambda_j + r_j \quad \forall j \in \arg(r_{-i}^{+}).$$

Note that, having fixed r_{-i}^{+}, i will be the output (the i-th argmax noisy trace) when the dataset is S if and only if $r_i \leq r^*$. We have, for all $j \in \arg(r_{-i}^{+})$:

$$\begin{aligned} &\lambda_i + r^* < \lambda_j + r_j \\ \Rightarrow\ &\lambda_i^{'} + r^* \leq \lambda_i + r^* < \lambda_j + r_j \leq (\lambda_j^{'} + 1) + r_j \\ \Rightarrow\ &\lambda_i^{'} + (r^* - 1) < \lambda_j^{'} + r_j. \end{aligned}$$

Thus, if $r_i \leq r^* - 1$, then the i-th trace will be the i-th maximum on the other side when the dataset is S' and the noise vector is (r_i, r_{-i}^{+}). The probabilities below are over the choice of $r_i \sim \text{Gauss}(1/\varepsilon_{\text{tr}})$, with probability $1 - \delta_{\text{tr}}$, and we have:

$$\begin{aligned} &\mathbb{P}[r_i \leq r^* - 1] \geq e^{-\varepsilon_{\text{tr}}}\mathbb{P}[r_i \leq r^*] = e^{-\varepsilon_{\text{tr}}}\mathbb{P}[i \mid S, r_{-i}^{+}] \\ \Rightarrow\ &\mathbb{P}[i \mid S', r_{-i}^{+}] \geq \mathbb{P}[r_i \leq r^* - 1] \geq e^{-\varepsilon_{\text{tr}}}\mathbb{P}[r_i \leq r^*] = e^{-\varepsilon_{\text{tr}}}\mathbb{P}[i \mid S, r_{-i}^{+}]. \end{aligned}$$

After multiplying through by $e_{\text{tr}}^{\varepsilon}$ and adding probability δ_{tr} for $\mathbb{P}[r_i - r^* \geq -1] \leq \delta$, we get:

$$\mathbb{P}[i \mid S, r_{-i}^{+}] \leq e_{\text{tr}}^{\varepsilon}\mathbb{P}[i \mid S', r_{-i}^{+}] + \delta_{\text{tr}}.$$

Overall, combing the both cases with $\delta_{\text{tr}} = 2\delta_{\text{tr}}$, we have

$$\begin{aligned} e^{\varepsilon_{\text{tr}}}(\mathbb{P}[i \mid S', r_{-i}^{+}] + \mathbb{P}[i \mid S', r_{-i}^{-}]) + \delta_{\text{tr}} &\geq \mathbb{P}[i \mid S, r_{-i}^{+}] + \mathbb{P}[i \mid S, r_{-i}^{-}] \\ e^{\varepsilon_{\text{tr}}}\mathbb{P}[i \mid S', r_{-i}] + \delta_{\text{tr}} &\geq \mathbb{P}[i \mid S, r_{-i}], \end{aligned}$$

more precisely, we can explicitly bound δ_{tr} to $\mathbb{O}(\frac{1}{pn})$ by refering to [82].

Using the same approach, we can prove that

$$e^{\varepsilon_{\text{tr}}}\mathbb{P}[i \mid S, r_{-i}] + +\delta_{\text{tr}} \geq \mathbb{P}[i \mid S', r_{-i}].$$

□

Thus, TraceSorting with Gaussian noise satisfies $(\varepsilon_{\text{tr}}, \delta_{\text{tr}})$-DP.

(2) *Secondly*, we prove the unified composition of TraceSorting∘GradientPerturbation is $(\varepsilon_{\text{tr}} + \varepsilon_{\text{dp}}, \delta)$-DP. Based on the results of TraceSorting, we employ two different clipping thresholds for GradientPerturbation.

Proof. We define the clipping threshold vector c for per-sample gradient by TraceSorting, for example, with $B = 3$ and $p = 1/3$, if heavy tailed indicator $\lambda = [1, 0, 0]$ then $c = [c_1, c_2, c_2]$.

$$\begin{aligned}
\mathbb{P}[M(S) = Y] &= \mathbb{P}[\text{TraceSorting=index } i \text{ AND GP}|S] \\
&= \int_{-\infty}^{\infty} \mathbb{P}[i|S, r_{-i}] \cdot \mathbb{P}[\text{GP with heavy tailed samples } i] dr \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{P}[i|S, r_{-i}] \cdot \mathbb{P}[\frac{1}{B}(\sum_{j}^{B\in S} g_j + c_j\zeta_j) = Y|c] dr d\zeta \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{P}[i|S, r_{-i}] \cdot \mathbb{P}[f(S) = Y|c] \cdot \mathbb{P}[\zeta = c_j\zeta_j/B] dr d\zeta = *,
\end{aligned}$$

where $r \sim \text{Gauss}(1/\varepsilon_{\text{tr}})$ and $\zeta \sim \text{Gauss}(1/\varepsilon_{\text{dp}})$. We define $f(\cdot) = \text{GradientDescent}$ and $\Delta f = \|f(S) - f(S')\|_2 = c_1$ if $S \in S^{\text{tail}}$ else c_2. With $1 - (\delta_{\text{tr}} + \delta_{\text{dp}})$, we have

$$\begin{aligned}
* &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp(\varepsilon_{\text{tr}}) \mathbb{P}[i|S', r_{-i}] \cdot \mathbb{P}[\frac{1}{B}(\sum_{j}^{B\in S'} g_j + c_j\zeta_j) = Y|c] dr d\zeta \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp(\varepsilon_{\text{tr}}) \mathbb{P}[i|S', r_{-i}] \cdot \mathbb{P}[f(S') + c_j\zeta_j/B = Y + \Delta f|c] dr d\zeta \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp(\varepsilon_{\text{tr}}) \mathbb{P}[i|S', r_{-i}] \cdot \mathbb{I}[f(S') = Y] \cdot \mathbb{P}[\zeta = c_j\zeta_j/B - \Delta f|c] dr d\zeta \\
&\le \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp(\varepsilon_{\text{tr}}) \mathbb{P}[i|S', r_{-i}] \cdot \mathbb{I}[f(S') = Y] \cdot \exp(\varepsilon_{\text{dp}}) \mathbb{P}[\zeta = c_j\zeta_j/B|c] dr d\zeta \\
&\le \exp(\varepsilon_{\text{tr}} + \varepsilon_{\text{dp}}) \mathbb{P}[M(S') = Y],
\end{aligned}$$

where we have taken into account the randomness of c through r with λ, then the first inequality comes from TraceSorting satisfying DP, and the penultimate inequality is derived from the basic Gaussian-based DP mechanism. Thus, define $\delta = \delta_{\text{tr}} + \delta_{\text{dp}}$, TraceSorting∘GradientPerturbation is $(\varepsilon_{\text{tr}} + \varepsilon_{\text{dp}}, \delta)$-DP. □

(3) *Thirdly*, we provide the proof that privacy amplification with subsampling still holds with the mechanism M: TraceSorting∘GradientPerturbation.

Proof. Let S and $S' = S \cup \{i\}$ be two adjacent datasets. In each iteration, the algorithm partitions the samples into a tail subset and a body subset with probability p and $1 - p$, respectively. Each subset is then subsampled independently with sampling rates q_1 and q_2, leading to an effective average sampling rate

$$\bar{q} = pq_1 + (1 - p)q_2.$$

Let M' denote the composed mechanism including the private sorting step and the discriminative clipping step, which together satisfy $(\varepsilon_{\text{tr}} + \varepsilon_{\text{dp}}, \delta)$-DP on the full dataset.

To show $(\bar{q}(e^{\varepsilon_{\text{tr}}+\varepsilon_{\text{dp}}} - 1), \bar{q}\delta)$-DP, we have to bound the ratio with $S' = S \cup i$:

$$\frac{\mathbb{P}[M'(S) = Y] - \bar{q}\delta}{\mathbb{P}[M'(S') = Y]} = \frac{\bar{q}\mathbb{P}[M(S_B) = Y \mid i \in B] + (1 - \bar{q})\mathbb{P}[M(S_B) = Y \mid i \notin B] - \bar{q}\delta}{\bar{q}\mathbb{P}[M(S'_B) = Y \mid i \in B] + (1 - \bar{q})\mathbb{P}[M(S'_B) = Y \mid i \notin B]}$$

To prove that M' satisfies $(\bar{q}(e^{\varepsilon_{\text{tr}}+\varepsilon_{\text{dp}}} - 1), \bar{q}\delta)$-DP, we follow the standard subsampling argument. Let $B \subseteq [n]$ denote the indices of the subsampled data. The probability that i is included in B equals $\bar{q}$, composed of two disjoint events: $(i \in \text{tail}) \wedge (i \in B)$ and $(i \in \text{body}) \wedge (i \in B)$.

For convenience, define the following quantities:

$$\begin{aligned}
C_{\text{tail}} &= \Pr[M(S_B) = Y \mid i \in B, \text{tail}], \\
C_{\text{body}} &= \Pr[M(S_B) = Y \mid i \in B, \text{body}], \\
C' &= \Pr[M(S'_B) = Y \mid i \in B], \\
E &= \Pr[M(S_B) = Y \mid i \notin B] = \Pr[M(S'_B) = Y \mid i \notin B].
\end{aligned}$$

Then the overall probabilities can be expressed as

$$\begin{aligned}
\Pr[M'(S) = Y] &= pq_1C_{\text{tail}} + (1 - p)q_2C_{\text{body}} + (1 - \bar{q})E, \\
\Pr[M'(S') = Y] &= \bar{q}C' + (1 - \bar{q})E.
\end{aligned}$$

Since both tail and body mechanisms satisfy $(\varepsilon_{\text{tr}} + \varepsilon_{\text{dp}}, \delta)$-DP, we have

$$C_{\text{tail}} \le e^{\varepsilon_{\text{tr}}+\varepsilon_{\text{dp}}} \min\{C', E\} + \delta, \qquad C_{\text{body}} \le e^{\varepsilon_{\text{tr}}+\varepsilon_{\text{dp}}} \min\{C', E\} + \delta.$$

Substituting the above inequalities, we obtain

$$\Pr[M'(S) = Y] \le \bar{q}\left(e^{\varepsilon_{\text{tr}}+\varepsilon_{\text{dp}}} \min\{C', E\} + \delta\right) + (1 - \bar{q})E.$$

Dividing both sides by $\Pr[M'(S') = Y] = \bar{q}C' + (1 - \bar{q})E$ and applying the same algebraic manipulation as in the standard subsampling lemma, we get

$$\frac{\Pr[M'(S) = Y] - \bar{q}\delta}{\Pr[M'(S') = Y]} \le e^{\bar{q}(\varepsilon_{\text{tr}}+\varepsilon_{\text{dp}})}.$$

Hence, M' satisfies

$$(\bar{q}(e^{\varepsilon_{\text{tr}}+\varepsilon_{\text{dp}}} - 1),\ \bar{q}\delta)\text{-DP}.$$

□

To sum up, Theorem 4.5 is proven.

H Supplemental Experiments

H.1 Implementation Details

All experiments are conducted on a server with an Intel(R) Xeon(R) E5-2640 v4 CPU at 2.40GHz and a NVIDIA Tesla P40 GPU running on Ubuntu. By default, we uniformly set subspace dimension $k = 200$, $\varepsilon = \varepsilon_{\text{tr}} + \varepsilon_{\text{dp}}$ with $\varepsilon_{\text{tr}} = \varepsilon_{\text{dp}}$, $p = 0.1$, and sub-Weibull index $\theta = 2$ for all datasets. In particular, we use the LDAM [12] loss function for heavy-tailed tasks. Besides, we set $c_2 = 0.1$, $B = 128$, and $\eta = 0.1$ for MNIST and FMNIST. For CIFAR10, we set $c_2 = 0.1$, $B = 256$, and $\eta = 1$. For ImageNette, we set $c_2 = 0.15$, $\eta = 0.0001$ and $B = 1000$. For E2E, we adopt the DPAdam optimizer and use the same settings as [45], where $c_2 = 0.1$. By default, we set $c_1 = 10 * c_2$, and the heavy-tailed proportion p is 0.1. We implement pre-sample clipping by BackPACK [16]. Specially, we list the implementation details by categorizing the dataset below.

- **MNIST**: MNIST has ten classes, 60,000 training samples and 10,000 testing samples. We construct a two-layer CNN network and replace the BatchNorm of the convolutional layer with GroupNorm. We set 40 epochs, 128 batchsize, 0.1 small clipping threshold, 1 large clipping threshold, and 1 learning rate.
- **FMNIST**: FMNIST has ten classes, 60,000 training samples and 10,000 testing samples. we use the same two-layer CNN architecture, and the other hyperparameters are the same as MNIST.
- **CIFAR10**: CIFAR10 has 50,000 training samples and 10,000 testing. We set 50 epoch, 256 batchsize, 0.1 small clipping threshold and 1 large clipping threshold with model SimCLRv2 [60] pre-trained by unlabeled ImageNet. We refer the code for pre-trained SimCLRv2 to https://github.com/ftramer/Handcrafted-DP.
- **CIFAR10-HT**: CIFAR10-HT contains 32×32 pixel 12,406 training samples and 10,000 testing samples, and the proportion of 10 classes in training samples is as follows: [0:5000, 1:2997, 2:1796, 3:1077, 4:645, 5:387, 6:232, 7:139, 8:83, 9:50]. We train CIFAR10-HT on model ResNeXt-29 [72] pre-trained by CIFAR100 with the same parameters as CIFAR10. We can see pre-trained ResNeXt in https://github.com/ftramer/Handcrafted-DP and CIFAR10-HT with LDAM-DRW loss function in https://github.com/kaidic/LDAM-DRW.
- **ImageNette**: ImageNette is a 10-subclass set of ImageNet and contains 9469 training samples and 3925 testing samples. We train on model ResNet-9 [33] without pre-train and set 1000 batchsize, 0.15 small clipping threshold, 1.5 large clipping threshold and 0.0001 learning rate with 50 runs.
- **ImageNette-HT**: We construct the heavy-tailed version of ImageNette by the method in [12]. ImageNette-HT contains 2345 trainging samples and 3925 testing samples, which is difficult to train, and proportion of 10 classes in training data follows: [0:946, 1:567, 2:340, 3:204, 4:122, 5:73, 6:43, 7:26, 8:15, 9:9]. The other settings are the same as ImageNette. Our ResNet-9 refers to https://github.com/cbenitez81/Resnet9/ with 2.5M network parameters.
- **E2E**: We have conducted experiments on transform-based NLP tasks for the dataset E2E with BLEU metric and GPT-2 model, which generates natural language from tabular data in the catering industry. We adopt the DPAdam optimizer and use the same settings as [45], where small clipping threshold $c_2 = 0.1$ and large clipping threshold $c_1 = 10 * c_2$.
- **Tabular Dataset**: We evaluate our method on six representative tabular datasets, including Product, Breast Cancer, Android Malware, Adult (Census Income), Bank Marketing, and Credit Card Default (Taiwan). The Product Classification and Clustering dataset contains 24,794 training samples and 6199 test samples with 7 textual attributes, where the 10-class classification task distinguishes products from different categories collected from 306 merchants on the PriceRunner platform. The Breast Cancer dataset contains 569 samples with 30 continuous attributes, where the binary classification task distinguishes malignant from benign tumors. The Android Malware dataset includes 4,464 samples extracted from Android applications, labeled as benign or malicious. The Adult (Census Income) dataset comprises 48,842 samples and aims to predict whether an individual belongs to the higher-income group. The Bank Marketing dataset contains 4,521 samples with 16 client and campaign-related features, where the task is to predict whether a customer will subscribe to a term deposit. Finally, the Credit Card Default (Taiwan) dataset includes 30,000 samples with 23 attributes and predicts whether a customer will default on credit card payments in the following month. All categorical features are one-hot encoded, and continuous features are normalized. Each dataset is randomly split into 80% training and 20% testing sets. We apply the DPSGD configuration with clipping threshold $c_2 = 0.1$, $c_1 = 1$, batch size 64, learning rate 0.1-0.5.

Moreover, we open our source code and simplified version for discriminative clipping on the following link:
Our source code is available at: https://github.com/NoNameSha/Discriminative_Clipping_DPSGD.

H.2 Training Trajectory of DC-DPSGD

To provide intuitive evidence of the optimization performance during the training, we further demonstrate the trajectories of training accuracy using MNIST with $\epsilon = 8$ in Figure 6, which clearly reveal the evolutionary pattern of model learning across epochs and highlight how different clipping strategies affect convergence behavior. These trajectories serve as an important diagnostic tool for understanding the stability and efficiency of private optimization, showing that DC-DPSGD achieves faster convergence, smoother training dynamics, and consistently higher accuracy compared to existing clipping mechanisms.

H.3 Ablation Experiment

We have included the remaining parameter ablation experiments in the appendix. For MNIST, FMNIST, ImageNette and ImageNette-HT, we evaluate the effects of four parameters on test accuracy in Table 8, Table 9, and Table 10, including the subspace-k, the allocation of privacy

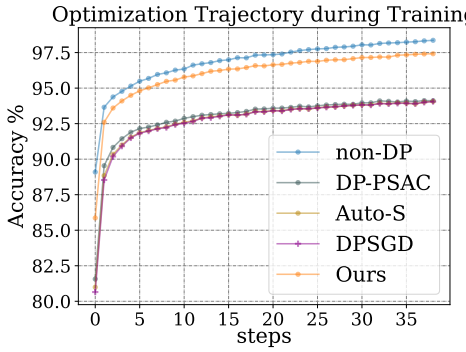

Figure 6: Optimization performance of DC-DPSGD.

budget ε, the heavy tail index sub-Weibull-θ, and the heavy tail proportion p, with other parameters kept at default. The experimental results are consistent with our discussion on CIFAR10 in main text. To investigate the effect of p, we have added a set of new experiments by varying $p \in [0.01, 0.2]$. The results are presented in Table 10. We observe that the test accuracy is minimally affected when p is less than 0.1, but shows a negative impact at around 0.2. We believe that the proportion of heavy-tailed samples aligns with statistical expectations. Assigning larger clipping thresholds to more light-body samples introduces more noise, while conservatively estimating heavy-tails does not fully exploit the algorithm's potential. Additionally, we acknowledge that since ImageNette-HT has only 2,345 training data, which is one-fifth of ImageNette, it is difficult to support the convergence of the model. In the future, we will improve this aspect in our work.

Table 8: Effects of parameters on test accuracy with MNIST and FMNIST with $\varepsilon = 8$.

Dataset	Subspace-k				ε_{tr} / ε			sub-Weibull-θ		
	None	100	150	200	0.2/8	0.4/8	0.8/8	1/2	1	2
MNIST	97.33	97.86	97.95	98.14	97.97	98.14	98.02	97.90	98.06	98.14
FMNIST	83.56	84.62	84.70	84.76	84.62	84.76	84.72	84.05	84.41	84.76

Table 9: Effects of parameters on test accuracy with ImageNette and ImageNette-HT with $\varepsilon = 8$.

Dataset	Subspace-k				ε_{tr} / ε			sub-Weibull-θ		
	None	100	150	200	0.2/8	0.4/8	0.8/8	1/2	1	2
ImageNette	64.98	65.34	66.52	67.66	65.23	67.66	66.12	65.91	66.28	67.66
ImageNette-HT	31.33	35.44	36.17	36.72	35.65	36.72	36.11	35.75	36.37	36.72

Table 10: Effects of parameter on p with ImageNette and $\varepsilon = 8$.

Dataset	Heavy tail Proportion-p				
	0.2	0.1	0.05	0.02	0.01
ImageNette	66.82	67.66	66.02	66.14	65.89

H.4 Simulation Experiment

To support our assumption, we conduct the simulation experiment to characterize the empirical distribution of gradient noise norms $\|\nabla \ell(\mathbf{w}_t, z_{j_t}) - \nabla L_S(\mathbf{w}_t)\|_2$ on the realistic dataset. Note that these quantities represent the L_2 norm of gradient noise rather than raw gradients, and are therefore non-negative. Consequently, the resulting empirical distribution is one-sided and its upper tail reflects the extent of gradient deviation. In the experiment, we model the per-sample gradient noise norms fitted by a sub-Weibull distribution, which is estimated on the upper tail via quantile-initialized maximum likelihood estimation, and compare it against the sub-Gaussian fit. The estimated shape parameter θ and scale parameter K quantify the degree of tail heaviness. As shown in Figure 7, the empirical gradient noise

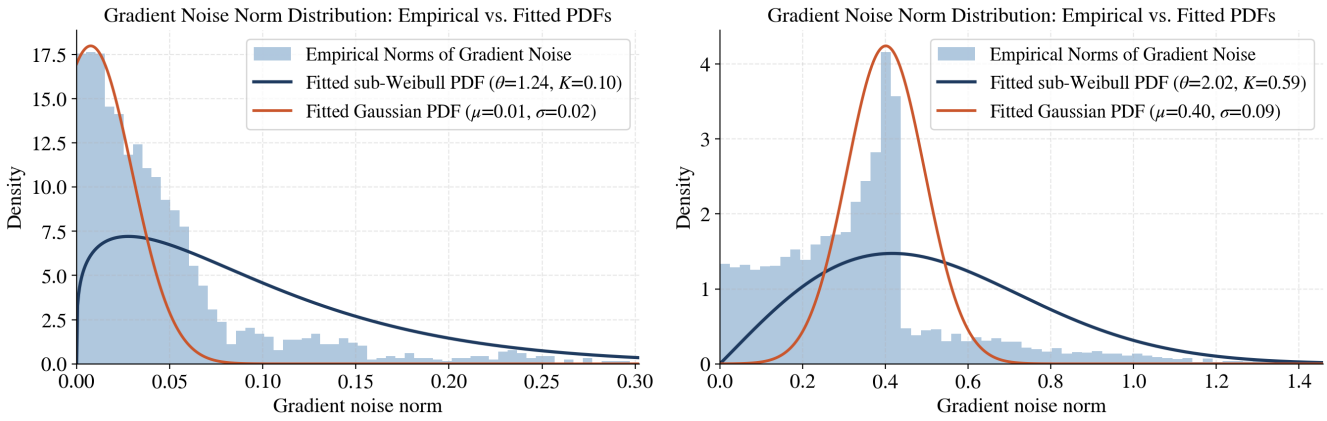

Figure 7: Sub-Weibull simulation of gradient noise norms on Imagenette dataset with ResNet-9. Left: early epoch of training; Right: late epoch of training.

norms exhibit the transition from light-tailed to heavy-tailed behavior. Specifically, the central concentrated region aligns closely with a sub-Gaussian fit, while the fit gradually deviates from the tail trend that is better captured by a sub-Weibull fit with $\theta > 1$. This observation demonstrates the practical validity of our sub-Weibull assumption. Since the sub-Weibull family naturally encompasses the sub-Gaussian distribution, it enables us to optimize clipped DP-SGD within a unified theoretical framework.