Tailor for Tails: Differentially Private Stochastic Optimization with Heavy Tails via Discriminative Clipping

Haichao Sha[1], Yuncheng Wu[1], Yong Liu[1], Yang Cao[2], 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,liuyuhan,chong}@ruc.edu.cn, cao@c.titech.ac.jp, ruixuan.liu2@emory.edu

ABSTRACT

Per-sample gradient clipping is a key technique in differentially private stochastic gradient descent (DPSGD), which shrinks the L_2 norm of an individual sample's gradient into a specific threshold. Most prior works rely on the assumption that the stochastic gradient noise (GN) follows a light-tailed distribution (e.g., sub-Gaussian). However, recent studies have shown that GN often exhibits a heavy-tailed distribution, rendering excessive clipping loss with existing mechanisms. In this paper, we design a novel clipping mechanism for DPSGD under the more generalized and heavy-tailed GN assumption (i.e., sub-Weibull distribution). We first present high probability guarantees with best-known rates for the optimization performance of DPSGD with gradient clipping under this assumption. Then, inspired by the guarantees, we propose a tail-aware clipping mechanism DC-DPSGD, which privately classifies gradients into body and tail with a subspace identification technique and clips them separately using two different clipping thresholds with a discriminative clipping method. Further, we theoretically analyze the convergence of DC-DPSGD and provide tighter optimization guarantees. Extensive experiments on ten real-world datasets demonstrate that our approach outperforms three baselines by up to 5.03% in terms of accuracy.

PVLDB Reference Format:
Haichao Sha[1], Yuncheng Wu[1], Yong Liu[1], Yang Cao[2], 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,liuyuhan,chong}@ruc.edu.cn, cao@c.titech.ac.jp, ruixuan.liu2@emory.edu . Tailor for Tails: Differentially Private Stochastic Optimization with Heavy Tails via Discriminative Clipping. PVLDB, 14(1): XXX-XXX, 2020.
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at URL_TO_YOUR_ARTIFACTS.

1 Introduction

Data privacy has become a paramount concern in modern deep learning. Models are increasingly trained on sensitive information derived from diverse sources, including structured tabular data with personally identifiable attributes and unstructured data that carries implicit privacy risks. DPSGD has become a cornerstone technique for verifiable and theoretically grounded data-driven privacy preservation. To ensure rigorous privacy guarantees, per-sample gradient clipping underpins the stable functioning of DPSGD [1]. It is designed to shrink the L_2 norm of an individual sample's gradient whenever it exceeds a specified threshold, allowing the addition of calibrated random perturbations.

Recently, a number of works [7, 12, 15, 23, 34, 40, 41, 59, 60, 62] have been proposed to optimize the clipping mechanism for DPSGD with per-sample gradient clipping (aka. clipped DPSGD). Most of these works rely on the assumption that the stochastic gradient noise (GN) in model optimization, referring to the deviation between the stochastic gradient randomly sampled and the average gradient estimated over the full training dataset, follows a light-tailed distribution (e.g., sub-Gaussian distribution). In particular, [7, 59, 62] focus on concentrated-norm gradients near the center of the light-tailed GN distribution and amplify the weight of gradients with relatively small norms. [3, 23, 34, 60, 68] utilize the boundedness of GN under the light-tailed distribution to provide theoretical guidance for the selection of clipping thresholds, which subsequently results in (near) optimal convergence rates.

Nevertheless, numerous empirical findings [5, 8, 49, 50, 66] have shown that the GN of SGD in deep learning often exhibits a heavy-tailed distribution instead of a light-tailed distribution, which may slow down the convergence rate and impair training performance [26, 37, 38, 42]. 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 [27]. Although [15, 41] provide the expectation convergence bounds for clipped DPSGD under a specific heavy-tailed Lipschitz assumption, that is, the per-sample gradient has unbounded upper constants, these results are not applicable to sub-Exponential tails and heavier ones. Besides, they fail to provide guidance on clipping strategies in non-convex DPSGD learning, making it difficult to determine the clipping threshold.

On this basis, designing a theoretically sound and effective clipping mechanism for clipped DPSGD under the heavy-tailed GN assumption faces two challenges. First, the upper bound of their moment generating functions (MGF) [54] is unmeasurable with heavy-tailed GN, which means that the variance of GN in expectation could be dominated by extreme values, leading to unboundedness. Thus, the expectation tools widely used in prior works are hardly applicable to obtaining analytical solutions for the clipping threshold and achieving optimal convergence guarantees 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. Figure 1 shows an example of this trade-off

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

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 (aka. sensitivity) is larger. This could lead to a high-dimensional catastrophe [70] 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 result in a significant loss of gradient information, especially under heavy-tailed GN distributions, which may lead to significant stochastic gradient bias.

In this paper, we present a novel clipping mechanism that addresses the above challenges under the heavy-tailed GN assumption. We address the first challenge by employing high probability tools to characterize the extreme probabilities of tails and bound the convergence of heavy-tailed clipped DPSGD. We then present guarantees that achieve the best-known rates for clipped DPSGD under the heavy-tailed GN assumption while retaining the optimal rates under the light-tailed GN assumption. We address the second challenge by proposing a tail-aware approach inspired by our guarantees, named **Discriminative Clipping (DC)-DPSGD**, which effectively balances the trade-off between clipping loss and required DP noise. The key idea is to decompose the gradients following a heavy-tailed distribution into body gradients and tail gradients, and utilize different clipping thresholds for them respectively, so as to keep more information from tail gradients that can withstand more severe DP noise. More specifically, we design a subspace identification technique to privately identify tail gradients with high probability guarantees. Note that the body of heavy-tailed distributions exhibits characteristics similar to those of light-tailed distributions, and the main difference lies in the decay rate at the tails. Therefore, we extract orthogonal random vectors from heavy-tailed distributions (e.g., sub-Weibull distribution) to construct a random projection subspace, and compute the trace of the second moment between gradients and this subspace to distinguish tail gradients. After that, we devise a discriminative clipping method, which applies a large clipping threshold for the identified tail gradients and a smaller one for the remaining body gradients. We theoretically analyze the choice of these two clipping thresholds and provide tighter optimization guarantees for DC-DPSGD. Further, we improve the guarantee with sharper rates by considering a gradient dominance curvature PL condition [32], which mitigates the dependence of convergence rates on the heavy tail degree. We summarize our main contributions below and the main theoretical results in Table 4 of Appendix A.1 (please see the full version).

- We present high probability guarantees with the best-known rates for the optimization performance of clipped DPSGD under the heavy-tailed sub-Weibull GN assumption, while retaining the optimal rates under the light-tailed GN assumption.
- We propose a tail-aware clipping mechanism DC-DPSGD with a subspace identification technique and a discriminative clipping method to optimize clipped DPSGD under the more generalized GN assumption. We theoretically 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 baselines with up to 5.03% accuracy improvements, demonstrating the effectiveness of our proposed approach.

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) with composition theorems and post-processing properties [19, 20, 43].

*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.

Under Definition 2.1, we characterize the convergence results 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.

Projection Subspace. Next, we introduce several notations regarding the projection subspace, which is used in our approach. Let $V_{t,k} = [v_{t,1}, v_{t,2}, \ldots, v_{t,k}] \in \mathbb{R}^{d\times k}$ denote a random projection matrix at iteration t, constructed by independently sampling k column vectors $v_{t,i} \in \mathbb{R}^{d\times 1}$, each drawn from a heavy-tailed sub-Weibull distribution, for $i \in [1, k]$. Then, the notation $V_{t,k}^\top$ means the transpose of $V_{t,k}$, and the product $V_{t,k}V_{t,k}^\top$ represents the corresponding random projection subspace. Given a normalized per-sample gradient $\nabla\hat{\ell}(\mathbf{w}_t, z_i) = \frac{\nabla\ell(\mathbf{w}_t, z_i)}{\|\nabla\ell(\mathbf{w}_t, z_i)\|_2}$, the empirical second moment in the projected space is expressed as $V_{t,k}^\top \nabla\hat{\ell}(\mathbf{w}_t, z_i)\nabla\hat{\ell}^\top(\mathbf{w}_t, z_i)V_{t,k} \in \mathbb{R}^{k\times k}$. For the population (true) projection subspace following the underlying distribution $\mathcal{P}$, we define it as $\hat{V}_{t,k}\hat{V}_{t,k}^\top$, where $\hat{V}_{t,k}$ =

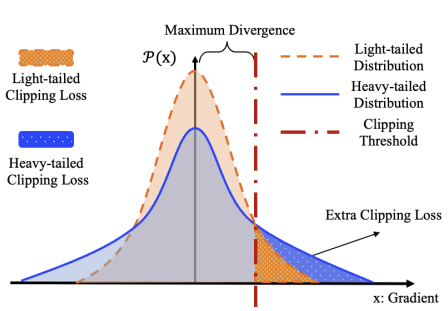

Figure 1: The challenged trade-off induced by different tail decay rates.

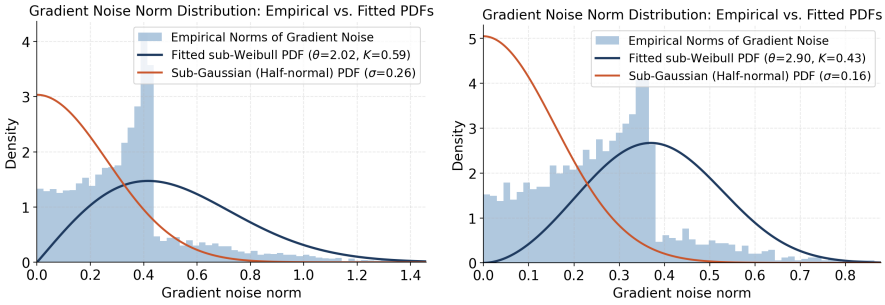

Figure 2: Sub-Weibull simulation of gradient noise norms on Imagenette dataset with ResNet-9 (epoch 49). Left: label-balanced sampling; Right: unbalanced sampling.

$\mathbb{E}_{V_{t,k}\sim\mathcal{P}}[V_{t,k}]$. Further, the estimated total variance in the empirical projection subspace can be measured by the trace, which is denoted as $\lambda_{t,i}^{\text{tr}} = \text{tr}\left(V_{t,k}^\top \nabla\hat{\ell}(\mathbf{w}_t, z_i)\nabla\hat{\ell}^\top(\mathbf{w}_t, z_i)V_{t,k}\right) \in \mathbb{R}$. Then, the true total variance of every per-sample gradient can be obtained by the trace $\hat{\lambda}_{t,i}^{\text{tr}} = \text{tr}\left(\hat{V}_{t,k}^\top \nabla\hat{\ell}(\mathbf{w}_t, z_i)\nabla\hat{\ell}^\top(\mathbf{w}_t, z_i)\hat{V}_{t,k}\right)$. Throughout, for any vector v, $\|v\|_2$ denotes L_2 norm.

For ease of reference, we summarize the frequently used notations in Appendix A.2.

2.2 Sub-Weibull Distributions

We next introduce the detailed characterization of sub-Weibull random variables, and make a comparison with light-tailed distributions using their moment generating functions (MGF) [53].

Definition 2.2 (Sub-Weibull Distribution [54]). 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

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

Sub-Weibull distributions generalize the sub-Gaussian and sub-Exponential families to potentially heavier-tailed distributions. Specifically, sub-Gaussian variables correspond to $\theta = 1/2$, and sub-Exponential variables to $\theta = 1$. Larger values of θ indicate heavier tails. In contrast, light-tailed distributions (often refer to sub-Gaussian) are defined by $\theta = 1/2$. Throughout this paper, when referring to stochastic gradient noise with heavy tails, we mean the gradient noise that satisfies Definition 2.2 with $\theta > 1/2$. Nevertheless, our bounds also cover the light-tailed case with $\theta = 1/2$.

A key problem with sub-Weibull distributions is that the MGF may not exist when $\theta > 1$, which corresponds to distributions heavier than sub-Exponential. In such cases, the standard moment-based techniques would fail, and the theoretical guarantees of clipped DPSGD can be invalid [7, 34, 40, 59, 60, 62], due to the fact that the MGF of X is not well-defined [4]. It is worth noting that in Definition 2.2, the MGF is defined over the variable $|X|^{1/\theta}$. As a result, classical techniques that rely on bounding the MGF are no longer applicable when dealing with heavy-tailed sub-Weibull distributions. Thus, it is difficult to establish the concentration of measure inequalities for sub-Weibull variables.

Fortunately, building on recent advances in martingale concentration for heavy-tailed settings [4, 37, 42], we leverage these tools to establish rigorous optimization guarantees for non-convex differentially private stochastic gradient descent under heavy-tailed gradient noise. In particular, the general truncation-based concentration method [4] enables the decomposition of heavy-tailed distributions into two distinct decay behaviors: sub-Gaussian light body ($\theta = 1/2$) and heavy tail ($\theta > 1/2$). This decomposition allows for effectively balancing the trade-off between gradient clipping and DP noise in DPSGD, which plays a pivotal role in our method.

2.3 Assumptions

We first formalize the assumption of heavy-tailed sub-Weibull gradient noise based on the definition described in Section 2.2, and then introduce other assumptions used in this paper below.

Assumption 2.1 (**Sub-Weibull Gradient Noise** [37]). *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, $\mathbb{E}_t[\exp((\|\nabla\ell(\mathbf{w}_t, z_{j_t}) - \nabla L(\mathbf{w}_t)\|_2/K)^2)] \leq 2$. Note that it implies that finding upper bounds for moment generating function under Assumption 2.1 is impracticable by standard tools [54]. Therefore, we use the truncated tail theory [4] and martingale difference inequality [42] in our analysis.

To support our assumption, we conduct a toy 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 Imagenette dataset with non-DP settings, as shown in Figure 2. We model per-sample gradient noise norms using a sub-Weibull distribution, fitted on the upper tail via quantile-initialized maximum likelihood estimation. The estimated shape parameter θ and scale parameter K quantify the degree of tail heaviness, while a half-normal fit serves as a sub-Gaussian proxy to confirm the presence of heavier tails during training.

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*

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

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.
  $\widetilde{\mathbf{g}}_t$ =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: $\overline{\mathbf{g}}_t(z_i) = \mathbf{g}_t(z_i) / \max(1, \frac{\|\mathbf{g}_t(z_i)\|_2}{c})$.
  end for
  Compute the weighted average of the gradients: $\overline{\mathbf{g}}_t = \sum_{i=1}^{|\pi|} \overline{\mathbf{g}}_t(z_i)$.
  Add the corresponding noise: $\widetilde{\mathbf{g}}_t = \frac{1}{|\pi|}\left(\overline{\mathbf{g}}_t + \mathbb{N}\left(0, c^2\sigma^2 \mathbb{I}_d\right)\right)$.
return $\widetilde{\mathbf{g}}_t$

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

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 [24, 37, 70] and is essential for ensuring the convergence of gradients to zero [39]. Assumption 2.3 is milder compared to the bounded stochastic gradient assumption [37, 38, 70], i.e., $\|\nabla \ell(\mathbf{w}_t, z_i)\|_2^2 \leq G$, making our results more applicable. It is worth noting that Assumption 2.3 constrains the L_2 norm of the average gradient under the objective, which is also commonly used in the stochastic optimization literature [37, 42], while Assumption 2.1 characterizes the randomness of individual sample gradients by quantifying their deviation from the empirical average over private training data.

3 Motivation and Rationale

In this section, we first provide the theoretical optimization guarantees of clipped DPSGD [1] under the heavy-tailed gradient noise (GN) assumption with high probability bounds. Guided by this analysis, we explain our motivation and idea of optimizing clipped DPSGD under heavy tails. For brevity, we only present the key results and analysis in the main paper, and provide the full versions and proofs in Appendix C.

3.1 Revisiting Guarantees of Clipped DPSGD with Heavy Tails

Clipped DPSGD is a golden standard for privacy-preserving deep learning, and has been widely studied in the field of private optimization [7, 15, 59, 60, 62]. However, existing works mainly analyze the optimization performance of clipped DPSGD under the light-tailed GN assumption. Due to the fact that clipped DPSGD serves as a biased optimizer for stochastic gradient descent with the per-sample clipping operation, even under light-tailed assumptions, addressing the bias induced by the clipping threshold to derive the boundary remains non-trivial.

To circumvent this difficulty, prior studies [7, 12] obtain convergence guarantees of clipped DPSGD by using the strong assumed symmetry of light-tailed gradient noise. However, their assumptions, though associated with finite variance, is overly restrictive and unrealistic in practice. Moreover, [15, 41] analyze the optimization guarantee of clipped DPSGD under the heavy-tailed Lipschitz assumption that describes the high-order moment constraint on gradients. Compared to this assumption, our heavy-tailed assumption captures higher-order tail behavior of gradient noise, which holds for the situations with heavier tails. As established in Lemma 22 of [42], the sub-Weibull gradient noise assumption subsumes the heavy-tailed Lipschitz assumption, which means our assumption offers a more general framework. Our analysis follows the outline of Abadi's clipped DPSGD [1], which is detailed 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. In the derivation process, a key quantity is the first-order term $-\langle \overline{\mathbf{g}}_t(z_i) - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle$, which we split into two components, i.e.,

$$\begin{aligned} &-\langle \overline{\mathbf{g}}_t(z_i) - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle = \\ &-\langle \overline{\mathbf{g}}_t(z_i) - \mathbb{E}_t[\overline{\mathbf{g}}_t(z_i)], \nabla L_S(\mathbf{w}_t)\rangle - \langle \mathbb{E}_t[\overline{\mathbf{g}}_t(z_i)] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle. \end{aligned} \tag{8}$$

Next, we present our results for clipped DPSGD under the assumption of heavy-tailed gradient noise.

Theorem 3.1 (**Convergence of Clipped DPSGD under Heavy-tailed 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 $c = \max(4K\log^{\theta}(\sqrt{T}), 39K\log^{\theta}(2/\delta))$, for any $\delta \in (0,1)$, with probability $1-\delta$, we have:*

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^{T} \min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} \\ &\qquad \leq \mathbb{O}\left(\log^{\max(1,\theta)}(T/\delta)\log^{2\theta}(\sqrt{T})\varphi^{\frac{1}{2}}\right). \end{aligned} \tag{9}$$

Proof Sketch. In this proof, firstly, unlike expectation-based analysis, the crucial steps in deriving high probability bounds involve constructing martingale difference sequences $\sum_{t=1}^{T}(-\langle \overline{\mathbf{g}}_t(z_i) - \mathbb{E}_t[\overline{\mathbf{g}}_t(z_i)], \nabla L_S(\mathbf{w}_t)\rangle)$, and applying advanced concentration inequalities to bound the term. Secondly, to handle the term $\langle \mathbb{E}_t[\overline{\mathbf{g}}_t(z_i)] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle$, we utilize the sub-Weibull properties to obtain an upper bound. Due to space limitations, we provide the full proof in Appendix C. □

In this theorem, we provide guidance for establishing the relationship between the clipping threshold c and the heavy tail index θ. We can observe that the optimization performance gradually deteriorates as θ ascends, because both the $\log^{\max(1,\theta)}(T/\delta)$ and $\log^{2\theta}(\sqrt{T})$ terms increase. We compare our result in Theorem 3.1 to the current optimal optimization error of clipped DPSGD [7, 62] and the representative results under the heavy-tailed Lipschitz assumption, as summarized in Table 4 of Appendix A.1. When

$\theta = 1/2$ (i.e., light-tailed sub-Gaussian GN scenarios), our convergence bound becomes $\mathbb{O}(\log(T/\delta)\log(\sqrt{T})\varphi^{\frac{1}{2}})$. It aligns with the current optimal bounds of clipped DPSGD without the gradient symmetry assumption [62], except for extra high probability terms. For results based on the assumption of heavy-tailed Lipschitz [15], their analysis derives bounds with θ-dependent on φ and δ, which deteriorate significantly as the tail heaviness increases. While in our result, θ is only related to the logarithmic terms. Moreover, their approach cannot be extended to sub-exponential or heavier-tailed distributions ($\theta \geq 1$), where the bounds degenerate to $\mathbb{O}(1)$. In contrast, our results can accommodate scenarios with heavier tails. To our knowledge, we are the first to analyze the optimization performance of clipped DPSGD under heavier sub-Weibull GN with high probability bounds.

3.2 Motivation: Better Balance Clipping Error and DP Noise

Clipping error is worse under heavy tails. Gradient noise can introduce more severe clipping bias under heavy-tailed distributions. Due to extreme samples that deviate significantly from the concentrated center occurring more frequently during optimization, such outliers effectively relax the ℓ_2-norm constraint. From a theoretical perspective, the alignment term of the clipped and true gradients $\|\tilde{\mathbf{g}}_t - \nabla L_S(\mathbf{w}_t)\|_2$ cannot be controlled by its expectation over t iterations as in the sub-Gaussian case; as shown below, it is instead bounded by the deviation term $\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2^2$, whose behavior is governed by heavy tails.

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

This bound follows from the Cauchy–Schwarz inequality, where $b_t = \mathbb{I}_{\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > \frac{c}{2}}$ acts as an indicator of large deviation. Then, according to Assumption 2.1, we have

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

Considering Eq. 11, the optimal clipping threshold increases with the tail parameter θ, as heavier tails require a larger c for optimality. Thus, directly adopting the same clipping threshold c as used under light-tailed conditions will cause this error term to be aggravated, resulting in an enlarged error bound of the form $\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2^2$ under heavy tails. Compared with the error introduced by DP noise, this effect is equally serious and may even hinder model convergence. Empirically, small clipping thresholds sometimes perform well. To explain this phenomenon, such small c values are usually paired with large learning rates, and c together with the learning rate can be regarded as coupled parameters that jointly determine the effective optimization scale. However, under heavy-tailed conditions, this coupled parameter should be increased to counteract stronger gradient deviations.

Balance the trade-off between clipping bias and DP noise. Bounding gradient noise in this setting is nontrivial. Our result in Theorem 3.1 also provides an important insight that the ideal clipping threshold c should scale up as the heavy tail index θ increases because the theoretical value of c is positively correlated to θ. Specifically, the optimal clipping threshold is given by $c = \max\left(\mathbb{O}\left(\log^{\theta}(1/\delta), \log^{\theta}(\sqrt{T})\right)\right)$. When c deviates from its optimal value, the deviation bound deteriorates in opposite directions. A smaller c enlarges the probability term $\Pr(\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > c/2)$, 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 bias but amplifies the variance of the injected DP noise, which scales as c^2, thereby loosening the overall error bound. Hence, the clipping mechanism must strike a judicious balance between bias and variance, ensuring a tighter theoretical convergence guarantee under heavy tails.

Insight into tail-aware clipping mechanism. However, a one-size-fits-all increase of the clipping threshold is not ideal under heavy-tailed conditions. In practice, most gradient norms remain relatively small and concentrated, so it is desirable to keep the DP noise level as low as possible while selectively optimizing the clipping loss for large-norm gradients. To this end, we take advantage of the sub-Weibull framework, which can capture both Gaussian-like concentration and tail heaviness simultaneously. This allows a dual optimization perspective, where small clipping thresholds are preserved for the light-tailed portion of gradients to maintain reliability, while larger thresholds are assigned to tail samples to effectively mitigate excessive clipping bias and improve performance under heavy-tailed conditions.

4 Discriminative Clipping DPSGD

4.1 Idea: Tail-Aware Gradient Clipping

Modeling gradient noise with sub-Weibull framework. As shown in Figure 2, the simulation of gradient noise conforms to the sub-Weibull assumption, and the tail becomes more heavier when sampling is unbalanced. Next, we provide more details about our ideas. First, according to the analysis in Theorem 3.1, it is necessary for clipped DPSGD to achieve better optimization performance by selecting a relatively larger clipping threshold related to tail index θ. This occurs when the heavy-tailed gradient noise exhibits a larger deviation from the mean, particularly for large-norm unconcentrated gradients. Second, the sub-Weibull gradient noise can be decomposed into two regions [4]: the region of small deviations, where the decay behaves similarly to a sub-Gaussian distribution, and the region of large deviations, which displays heavier-tailed behavior. Specifically, for the sub-Weibull variables $X \sim subW(\theta, K)$, the tail probability $\mathbb{P}(|X| > x) = \exp(-I(x))\ \forall x > 0$ exhibits two different behaviors:

(1) **Light body**: for small and concentrated x values at the center of the distribution, the tail rate capturing function $I(x)$ decays like a light-tailed sub-Gaussian tail;
(2) **Heavy tail**: for x greater than the normal convergence region, i.e., $x \geq x_{\max}$ is a large deviation region, its decay is slower than that of the normal distribution, where $x_{\max}$ is a mathematical inflection point related to the population variance of underlying distributions [4].

Motivated by the insights, our objective is to identify which private samples belong to the *light body* or *heavy tail* regions, and apply tailored clipping mechanisms to maximize the optimization benefit of each regime accordingly. However, there still exist two challenges. First, obtaining prior knowledge of the true tail index

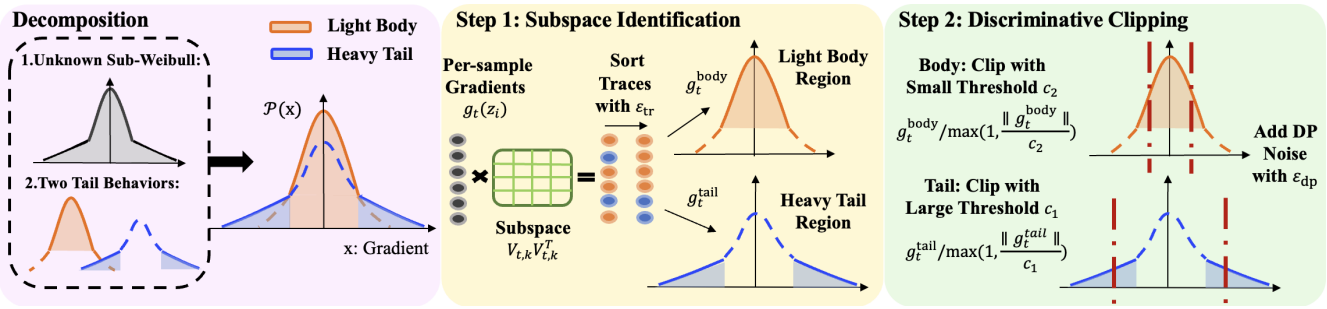

Figure 3: Overview of DC-DPSGD.

and the population variance is often impractical in private optimization. Another way is to estimate distributional properties from private samples, but it typically incurs an expensive privacy cost, especially when DP estimation is involved, which uses gradient norms with unbounded sensitivity. Second, decomposing private optimization into two separate regions makes it significantly more challenging to obtain both privacy analysis and convergence guarantees with existing theoretical tools. For this purpose, we next formally present our proposed solution.

4.2 Overview

In this section, we first give an overview of our novel tail-aware per-sample gradient clipping mechanism, and then give the detailed constructions. To begin with, we introduce the discriminative clipping DPSGD approach, named DC-DPSGD, as outlined in Algorithm 2. The key idea is to classify all mini-batch gradients following a heavy-tailed sub-Weibull distribution into two parts: light body and heavy tail, and employ different clipping thresholds for the two parts respectively. In particular, a small clipping threshold is applied for the light body to achieve optimality under existing light-tailed guidance, and a larger clipping threshold is applied for the heavy tail to mitigate the extra clipping loss.

Figure 3 illustrates an overview of DC-DPSGD, which consists of two steps. In the first step (Section 4.3), we propose a subspace identification technique to distinguish gradients from light body and heavy tail privately, 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 utility guarantees for this step (Theorem 4.1). In the second step (Section 4.4), we present a discriminative clipping method that utilizes different clipping thresholds for the two parts and adds DP noise with scale σ_{dp} for gradient perturbation. To establish the comprehensive theoretical guarantees of our proposed private optimization method, we utilize the sharp sub-Weibull concentration tools [4] to derive high-probability convergence bounds (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 convergence bound (Theorem 4.3) for DC-DPSGD and sharp the convergence rates (Theorem 4.4) under the PL condition (Section 4.5). At last, we rigorously analyze the privacy guarantee of DC-DPSGD (Theorem 4.6) and provide the privacy budget allocation for the two steps (Section 4.6). Algorithm 2 presents the detailed steps of DC-DPSGD.

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.

Initialize $\mathbf{w}_0$ randomly and initialize $V_{t,k}$ to None.
for each iteration $t = 1 \dots T$ **do**
  if `Strategy_trigger` **then**
    Step 1: Private Subspace Identification
    Private sorting: $(S^{\text{tail}}, S^{\text{body}}) \leftarrow$ Algorithm 3(σ_{tr}),
      where $i \in S$, $S^{\text{tail}} = \{\widetilde{z}_i\}_{i=1}^{pn}$, $S^{\text{body}} = \{\widetilde{z}_i\}_{i=1}^{(1-p)n}$.
    Sample mini-batches with rates q_1 and q_2:
      $S^{\text{tail}} \leftarrow \{B_1^{\text{tail}}, \dots, B_{q_1 pn}^{\text{tail}}\}$,
      $S^{\text{body}} \leftarrow \{B_1^{\text{body}}, \dots, B_{q_2(1-p)n}^{\text{body}}\}$.
    Apply random permutation:
      $\Pi \leftarrow \text{Permutation}\left(B_1^{\text{body}}, \dots, B_{q_2(1-p)n}^{\text{body}}, B_1^{\text{tail}}, \dots, B_{q_1 pn}^{\text{tail}}\right)$.
  end if
  Step 2: Discriminative Clipping for DPSGD
  for each batch $\pi \in \Pi$ **do**
    if $\pi \in S^{\text{tail}}$ **then**
      $\widetilde{\mathbf{g}}_t \leftarrow \text{Clip\_and\_Perturbation}(c_1, \pi)$
    else if $\pi \in S^{\text{body}}$ **then**
      $\widetilde{\mathbf{g}}_t \leftarrow \text{Clip\_and\_Perturbation}(c_2, \pi)$
    end if
    Update parameters: $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \widetilde{\mathbf{g}}_t$
  end for
end for

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

4.3 Subspace Identification

We now introduce our subspace identification technique in the first step. We note that numerous studies [25, 64, 70] 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 by 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 [55] and provide effective optimization information for model training [7, 62]. Given that normalized gradients have

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 $(S^{\text{tail}}, S^{\text{body}})$.

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

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 the empirical traces, that are measured in sampled random projection subspace, can reliably approximate the population variance, which enables us to identify the heavy-tailed samples with high-probability guarantees.

To be concrete, in the subspace identification step of each iteration $t \in [1, T]$, we first construct a projection matrix composed of k random orthogonal unit vectors $V_{t,k} = [v_{t,1}, v_{t,2}, ..., v_{t,k}]$, 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 ($\theta = 2$), enabling the subspace to capture diverse gradient noise patterns. Secondly, we compute the per-sample projected trace by $\lambda_{t,i}^{\text{tr}} = V_{t,k}^\top \hat{\mathbf{g}}_t(z_i)\hat{\mathbf{g}}_t^\top(z_i)V_{t,k}$, where $\hat{\mathbf{g}}_t(z_i)$ is the normalized version of $\mathbf{g}_t(z_i)$ with bounded directional information. Nevertheless, despite normalization, the traces still carry characteristic information of the underlying private sample. Thirdly, directly using unperturbed traces may increase the risk of privacy leakage. Therefore, we adopt a differentially private sorting mechanism that integrates Report Noisy Argmax techniques [20] with calibrated Gaussian noise, enforcing differential privacy on the trace sequence $\lambda_t^{\text{tr}} = [\lambda_{t,1}^{\text{tr}}, \ldots, \lambda_{t,B}^{\text{tr}}]$. 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 $\widetilde{\lambda}_t^{\text{tr}} = [\widetilde{\lambda}_{t,1}^{\text{tr}}, \ldots, \widetilde{\lambda}_{t,B}^{\text{tr}}]$ is sorted in ascending order to determine the boundary between the light-body and heavy-tail regions.

The `strategy_trigger` command controls the frequency of 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.

Subspace identification accuracy analysis. 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., $\|V_{t,k}V_{t,k}^\top - \mathbb{E}_{V_{t,k}\sim\mathcal{P}}[V_{t,k}V_{t,k}^\top]\|_2$, where $V_{t,k}V_{t,k}^\top = \frac{1}{k}\sum_{i=1}^{k} v_{t,i}v_{t,i}^\top$. According to Ahlswede-Winter Inequality [55], we analyze the error caused by the skewing. In addition, it is worth noting that in line 8 of Algorithm 2, as extra DP noise is injected into private trace sorting, the accuracy of identification could be affected by the perturbation. Thus, we also need to constrain the bias introduced by DP noise. Building upon the results, we derive the high probability bound for subspace identification in Theorem 4.1 and provide the detailed proof in Appendix D.

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 $\delta_m' = \delta_{\text{tr}} + \delta_m$.

4.4 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

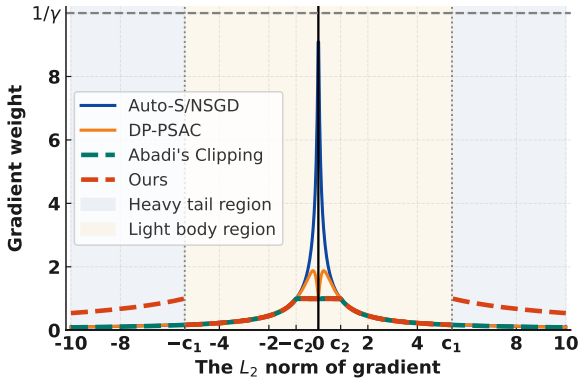

Figure 4: The gradient weight under different per-sample clipping functions, and $\gamma = 0.1$, $c_2 = 1 - \gamma$ and $c_1 = 5c_2$.

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 compared to 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 [7, 59, 62] can be interpreted as an approximated version of Abadi's clipping function defined in Algorithm 1 under small clipping threshold regimes, These adaptive approaches which predominantly focus on allocating more weights to scale concentrated gradients that have relatively small norms. We provide a detailed explanation of their impact on gradient norms below.

(1) Auto-S [7] and NSGD [62] 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.
(2) DP-PSAC [59] adopts a conservative clipping strategy to control the amplification using the weight function $\frac{\mathbf{g}_t(z_i)}{\|\mathbf{g}_t(z_i)\|_2+\frac{\gamma}{\|\mathbf{g}_t(z_i)\|_2+\gamma}}$.

Nevertheless, 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 4, 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, most optimal theoretical guarantees for clipped DPSGD are grounded in the assumption of light-tailed GN [7, 48, 59, 60, 62], leaving a gap in optimization theory under heavy-tailed settings.

To tackle this problem, we propose a discriminative clipping mechanism in this paper. After identification, the partitioned samples are divided into batches and randomly permuted, where an equivalent level of amplification can be achieved by adopting a smaller batch size. A detailed privacy analysis of this process is provided in Section 4.6. 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 that are classified in the subspace identification of Section 4.3, 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 4, 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 light body and heavy tail regions, each accompanied by the respective high-probability convergence guarantee. Next, in the construction of the convergence boundary for the two regions, we generalize the sharp heavy-tailed concentration [4] and sub-Weibull Freedman inequality [42] to truncate the assumed heavy-tailed distribution and find the optimal clipping threshold for each region. As a result, we have the following theorem.

Theorem 4.2 (**Two-behavior Convergence 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 bounds 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 performance. We provide the full proof in Appendix E. □

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 performance 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 slower convergence 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 [7]), under the condition of using a sufficiently small clipping threshold paired with a large learning rate. The guidance stated as a diagonal pattern in [7] has been widely applied in practice and proven in theory [12, 48]. Prior works [7, 12] have theoretically shown that both Abadi's clipping and the adaptive clipping can achieve the same optimal order of convergence rate, but their results cannot be extended to heavy-tailed scenarios. Here, we offer 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.

4.5 Tighter Guarantees for DC-DPSGD

In this subsection, we provide the formal and sharper optimization guarantees for DC-DPSGD. Note that the boundary derived in Section 4.4 is based on the assumption of perfectly classifying each sample into its corresponding region. However, in practice, the subspace identification incurs utility errors by misidentification, which are jointly determined by the subspace skewing and the tail proportion p, as analyzed in Section 4.3. Since the subspace skewing of $\mathbb{O}(1/\sqrt{d})$ is non-dominant compared to the convergence bound 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$[0.05, 0.2]$ [53].

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 heavy-tailed gradients correctly identified is $p(1-\delta'_m)$. Thus, we can analyze the convergence 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 (**Formal Version 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 $\delta' \in (0,1)$, with probability $1-\delta'$:*

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

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 $C_{\text{tail}}(c_1)$ from correctly identified heavy-tailed gradients with the proportion p and the high probability $1-\delta'_m$; (ii) the convergence $C_{\text{body}}(c_2)$ from correctly identified light body gradients with the proportion $1-p$ and the 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. □

Corollary 4.3 indicates that the optimization performance 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 true value of p is always less than 1. Moreover, as even for large tail indices θ, the tail proportion p can be restricted with a sufficiently small variance K [54]. Especially, if $p \le 1/(\frac{C_{\text{tail}}(c_1)}{C_{\text{body}}(c_2)} + 1)$, it enables us to achieve θ-independent rates of $(1-p) * \mathbb{O}\left(\log(T/\delta')\log(\sqrt{T})\varphi^{1/2}\right)$.

Further, instead of measuring by gradient norms, we can derive optimization guarantees of function values under the PL condition, which satisfies a weaker curvature property.

Theorem 4.4 (**Sharper Guarantee for DC-DPSGD**). *Under Assumptions 2.1, 2.2 and B.1, assuming that $\eta_t \le \frac{1}{2\beta}$, and c_1, c_2 as defined in Theorem 4.2, for any $\delta' \in (0,1)$, we have:*

$$\begin{aligned} L_S(\mathbf{w}_{T+1}) - L_S(\mathbf{w}_S) &\le p * \mathbb{O}\left(\log^{2\theta}(T)\log^{\theta+\frac{1}{2}}(T/\delta')\varphi\right) \\ &\quad + (1-p) * \mathbb{O}\left(\log(T)\log(T/\delta')\varphi\right). \end{aligned} \tag{16}$$

Proof. We provide the full proof in Appendix G. □

By substituting Assumption 2.3, we use PL condition to obtain sharper guarantees of an order of φ for DC-DPSGD under heavy-tailed GN assumptions, which is close to the optimal rate of DPSGD (without clipping analysis) and the rate of clipped DPSGD with gradient symmetry under light-tailed GN assumptions. In general, the result is faster than the order of $\varphi^{1/2}$ when $n\varepsilon = \mathbb{O}(\sqrt{d})$. Similar to Corollary 4.3, Theorem 4.4 confirms that when p satisfies the condition of $p \leq \frac{1}{\log^{2\theta-1}(T)\log^{\theta-\frac{1}{2}}(T/\delta')+1}$, the optimization guarantees of $(1-p) * \mathbb{O}\left(\log(T)\log(T/\delta')\varphi\right)$ are also θ-independent.

4.6 Privacy Analysis

Under 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 tail subset and a body subset 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 4.5 (Noise scaling under partitioned sampling). *Under the same privacy budget ε, the partitioned mechanism requires a noise multiplier that requires*

$$\sigma_{\text{dp}} \approx \sqrt{\frac{pq_1^2 + (1-p)q_2^2}{\bar{q}^2}}\sigma_{\text{Pois}}. \tag{17}$$

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

Theorem H.1 formalizes this relation between the noise multiplier σ_{dp} in DC-DPSGD and σ_{Pois} in standard clipped DPSGD.

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 4.6 gives the privacy guarantee of our DC-DPSGD approach.

Theorem 4.6 (**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 [20]. For the discriminative clipping mechanism, since it uses two clipping thresholds to separately handle gradients from different parts of the mini-batch, we re-analyze the gradient perturbation with Gaussian mechanism [1] and subsampling. The detailed proof of privacy guarantees is provided in Appendix H. □

5 Experiments

5.1 Experimental Setup

Datasets and models. We evaluate DC-DPSGD on eleven real-world datasets, including MNIST, FMNIST, CIFAR10, and ImageNette [16] for image classification, E2E [18] for natural language generation, and six tabular datasets—Product [2], Breast Cancer [58], Android Malware [6], Adult [33], Bank Marketing [44], and Credit Card

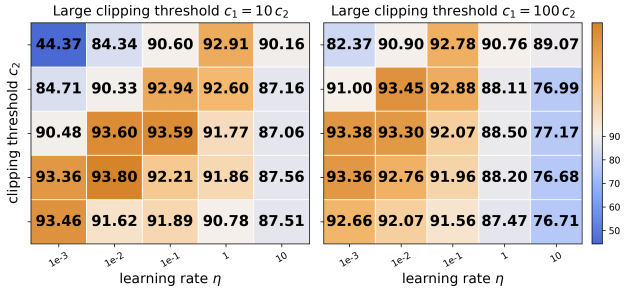

Figure 5: Test accuracy heatmap on CIFAR10 with c_1, c_2 and η.

Default [63]. All tabular tasks are binary classification with standardized categorical and continuous features. Moreover, we use two heavy-tailed versions: namely CIFAR10-HT [9] (a heavy-tailed version of CIFAR10) and ImageNette-HT (modified on [47]) to evaluate the utility under heavy tail assumption.

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 [52] with a linear classifier, respectively. For ImageNette and ImageNette-HT, we adopt the same setting as [7] and ResNet9 without pre-training. For E2E, we use a transformer-based GPT-2 model (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 image classification tasks using accuracy that measures the portion of correct predictions, and natural language generation tasks using the BLEU score [46] that measures the quality of generated data with a modified n-gram score.

Baselines. We compare DC-DPSGD with three differentially private baselines based on the light-tailed gradient noise assumption: DPSGD with Abadi's clipping [1], Auto-S/NSGD [7, 62], DP-PSAC [59], and a non-private baseline: non-DP ($\varepsilon = \infty$).

5.2 Effectiveness Evaluation

We summarize the performance of various methods in Tables 1 and 2, including non-DP baselines for comparison.

DC-DPSGD outperforms Clipped DPSGD variants. In 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, the corresponding improvements are 3.49%, 5.03%, and 3.70%. Specifically, DPSGD with Abadi's clipping demonstrates comparable performance to adaptive methods under small clipping thresholds, where relatively small-norm gradients are assigned higher weights. However, such methods exhibit degraded performance on more complex (e.g., ImageNette) and heavy-tailed datasets (e.g., CIFAR10-HT and ImageNette-HT). In contrast, our method not only improves accuracy on normal datasets but also retains strong performance on heavy-tailed datasets. The reason is that our approach places a larger clipping threshold for heavy-tailed gradients, thereby alleviating the clipping bias of heavy-tailed gradients and preserving more information about them. In addition, for the language dataset E2E, we show that DC-DPSGD is scalable and improved on the GPT-2 model in both fine-tuned methods.

Tabular tasks in low privacy regimes. As shown in Table 2, under relatively loose privacy constraints (e.g., $\varepsilon = 0.8$), DC-DPSGD

Table 1: Test accuracy (%) comparison between DC-DPSGD and baselines under different privacy constraints.

Privacy constraint	Algorithm	MNIST	FMNIST	CIFAR10	ImageNette	CIFAR10-HT / ImageNette-HT
$\varepsilon = 8, \delta = 1/n^{-1.1}$	DPSGD	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**
$\varepsilon = 4, \delta = 1/n^{-1.1}$	DPSGD	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**
$\varepsilon = \infty, \delta = 1/n^{-1.1}$	SGD	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 (%) and accuracy (%) comparison between DC-DPSGD and baselines under different privacy constraints.

Algorithm	Privacy constraint	E2E Full	E2E LoRA	Privacy constraint	Product	Malware	Cancer	Adult	Bank	Credit
DPSGD	$\varepsilon = 8, \delta = 1/n^{-1.1}$	63.189	63.389	$\varepsilon = 0.8, \delta = 1/n^{-1.1}$	83.22	96.86	94.74	85.12	88.62	80.90
Auto-S		63.600	63.518		82.22	96.86	94.74	85.20	88.51	80.95
DP-PSAC		63.627	63.502		83.69	96.75	95.09	85.15	88.51	80.92
Ours (DC-DPSGD)		**64.180**	**63.920**		**85.90**	**97.49**	**95.52**	**85.61**	**88.73**	**81.28**
DPSGD	$\varepsilon = 3, \delta = 1/n^{-1.1}$	61.519	61.220	$\varepsilon = 0.4, \delta = 1/n^{-1.1}$	78.06	93.06	85.09	81.94	86.63	77.95
Auto-S		61.340	61.220		77.37	93.39	85.96	82.17	86.62	78.45
DP-PSAC		61.340	61.263		78.45	93.06	84.21	81.86	86.51	77.60
Ours (DC-DPSGD)		**61.732**	**61.563**		**80.03**	**93.58**	**86.32**	**82.30**	**86.72**	**78.63**
DPSGD	$\varepsilon = \infty, \delta = 1/n^{-1.1}$	69.463	69.692	$\varepsilon = \infty, \delta = 1/n^{-1.1}$	95.45	99.55	96.49	85.55	89.94	81.95
Auto-S		69.463	69.682		95.47	99.33	96.49	85.54	89.72	81.95
DP-PSAC		69.473	69.692		95.50	99.55	95.61	85.62	89.72	82.08
Ours (DC-DPSGD)		**70.328**	**70.455**		**96.11**	**99.78**	**97.36**	**85.79**	**90.26**	**82.44**

Table 3: Effects of different sampling ways 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%

achieves consistently superior performance across all tabular datasets, surpassing standard DPSGD variants by a clear margin. Even when the privacy budget is reduced to $\varepsilon = 0.4$, 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.

Effects of parameters on test accuracy. Next, we evaluate the effects of four parameters on test accuracy, including the dimension of the subspace k, the allocation of privacy budget ε, the heavy tail index θ of the sub-Weibull distribution, the heavy tail proportion p and identification strategy trigger frequency, with other parameters kept at default. The results on CIFAR10, CIFAR10-HT and Android Malware are shown in Tables 3. **1) *Subspace k***: We can 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. **2) $\varepsilon_{tr}/\varepsilon$**: 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. **3) *Tail index θ***: Since the 'HT' dataset is 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

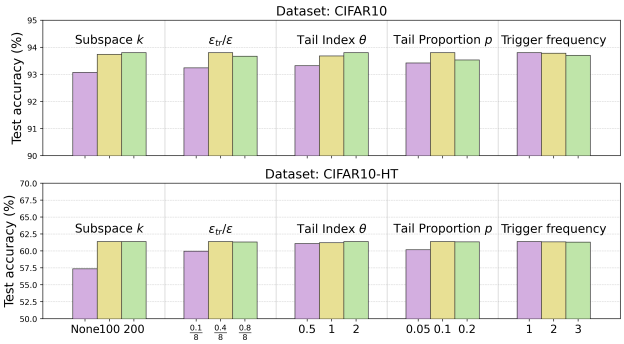

Figure 6: Ablation experiments under different parameters.

accuracy. **4)** ***Tail proportion p***: For the tail proportion, $p = 0.1$ achieves better results. If p is too low, it fails to mitigate clipping loss, while if p is too large, it could introduce extra noise. The proportion of heavy-tailed samples aligns with statistical expectations. Assigning larger clipping thresholds to more light-body samples introduces more noise. **5)** ***Trigger frequency***: Changing the trigger frequency has a moderate effect on model performance, which aligns with the empirical observation that the subspace remains stable during training. Hence, it is advisable to perform subspace identification infrequently to reduce extra privacy overhead.

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 classical clipped DPSGD. From a theoretical perspective, given the parameters $\delta = 1e^{-5}$, $\eta/B = 0.04$, and $\theta \approx 2$ (following [21, 27, 54]), we can obtain $c_1 = \mathbb{O}(\log^{\theta}(1/\delta))$, which is $\sqrt{125}$ times larger than $c_2 = \mathbb{O}(\log^{1/2}(1/\delta))$, that is, $c_1 = \log^{3/2}(1/\delta)c_2$, i.e., $c_1 \approx 10c_2$. Therefore, the optimal clipping threshold aligns with our empirical guidance.

Effects of sampling strategies on test accuracy. Table 6 compares the impact of different sampling strategies on test accuracy across several datasets. One strategy sets $q_1 = q_2 = q$ ($\sigma_{\text{dp}} = \sigma_{\text{Pois}}$), while the other fixes $B_1 = B_2 = B$ (for example, in CIFAR10, $B = 64$, $p = 0.1$, and $\sigma_{\text{dp}} = 1.3\sigma_{\text{Pois}}$). The results show that the accuracy gap between the two settings is marginal (within 0.2–0.3%), indicating that although sacrificing a small portion of batch size generally tends to yield slightly poorer performance, it is acceptable to preserve the same level of privacy amplification in DC-DPSGD.

To give a more complete and transparent view of our method's implementation and empirical performance, we include additional analyses in the Appendix of the full version, covering implementation details, training trajectories, ablation studies, and privacy auditing results. In particular, the privacy auditing section presents the empirical verification of the privacy guarantees of DC-DPSGD using DP auditing methods. Specifically, the auditing process evaluates the empirical privacy budget by measuring the false positive rate and false negative rate of membership inference attacks, which are further transformed to the auditing results by Gaussian DP [17]. The results show that DC-DPSGD maintains comparable empirical privacy leakage to standard DPSGD under both worst-case and one-round auditing paradigms.

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 [26, 27, 49, 66], 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 [31, 56]. Building upon [56], [31] relax the assumption of Lipschitz condition and sub-Exponential distribution to a more general α-th moment bounded condition. 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 [37, 38, 42] are widely used in non-private learning such as convex and non-convex optimization. Specifically, under bounded α-th moments assumption, [38] 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 [15, 31, 41] 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 [3, 12, 34, 48, 57, 60, 65, 69]. Since the tuning parameters in the Abadi's clipping function [1] are complex, various adaptive gradient clipping schemes have been proposed [7, 62]. These schemes scale per-sample gradients based on their norms. In particular, gradients with concentrated norms are amplified infinitely. Building upon this, [59] controls the amplification of gradients with concentrated norms in a finite manner. For the theoretical guidance of clipping thresholds, [15] 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, [34, 68] establish precise convergence guarantees for arbitrary clipping thresholds under the light-tailed bounded GN assumption. Additionally, research on clipping bias has gradually gained importance. [57] and [34] argue for the connection between sampling noise and clipping bias, and mitigate clipping bias 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 high probability guarantees with the best-known rates for the clipped DPSGD optimization under the heavy-tailed gradient noise assumption. Based on the guarantees, we propose a novel tail-aware clipping mechanism DC-DPSGD that employs different clipping thresholds for body and tail gradients. We rigorously analyze the high probability bound of DC-DPSGD under non-convex settings and obtain optimal results with heavier conditions. Furthermore, we sharpen the optimization performance of DC-DPSGD to near tail-independent rates. We conduct extensive experiments on eleven real-world datasets, and the results demonstrate that DC-DPSGD outperforms three state-of-the-art gradient clipping mechanisms by up to 5.03% accuracy improvement.

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] 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.

[6] 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).

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

[8] 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.

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

[10] 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.

[11] Zachary Charles and Dimitris Papailiopoulos. 2018. Stability and generalization of learning algorithms that converge to global optima. In *ICML*. PMLR, 745–754.

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

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

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

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

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

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

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

[25] Aditya Golatkar, Alessandro Achille, Yu-Xiang Wang, Aaron Roth, Michael Kearns, and Stefano Soatto. 2022. Mixed differential privacy in computer vision. In *CVPR*. 8376–8386.

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

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

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

[29] 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.

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

[31] 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.

[32] Hamed Karimi, Julie Nutini, and Mark Schmidt. 2016. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *ECML PKDD*. Springer, 795–811.

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

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

[35] Yunwen Lei and Yiming Ying. 2021. Sharper generalization bounds for learning with gradient-dominated objective functions. In *ICLR*.

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

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

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

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

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

[41] 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.

[42] 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.

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

[44] 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.

[45] 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.

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

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

[48] 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).

[49] 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.

[50] 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.

[51] 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.

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

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

[54] 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.

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

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

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

[62] 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).

[63] 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.

[64] Da Yu, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. 2021. Do not Let Privacy Overbill Utility: Gradient Embedding Perturbation for Private Learning. In *ICLR*.

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

[66] 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.

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

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

[69] 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*.

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

[71] 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 4: 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 [62]	$\mathbb{O}\left(\varphi^{1/2}\right)$	Non-convex	Light-tailed Gradient Noise	Normalized $c = 1$
Auto-S [7]	$\mathbb{O}\left(\varphi\right)$	Non-convex	Light-tailed Gradient Noise; Gradient Symmetry	Normalized $c = 1$
Clipped DPSGD [15]	$\mathbb{O}\left(\varphi^{1-\theta}\right)$	Convex	Heavy-tailed Lipschitz	$c \le G_{\min}$
Clipped DPSGD [15]	$\mathbb{O}\left(\delta^{-\frac{2\theta}{2-\theta}}\varphi^{1-\frac{\theta}{2-\theta}}\right)$	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: $c_2 = f_c(\frac{1}{2})$
Our DC-DPSGD (Thm 4.4)	$p * \mathbb{O}\left(\log^{2\theta}(T)\log^{\theta+\frac{1}{2}}(T/\delta')\varphi\right)$ $+(1-p) * \mathbb{O}\left(\log(T)\log(T/\delta')\varphi\right)$	Non-convex PL Condition		

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 [54]). *Let X be a random variable and $\theta \ge 0$, 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

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

(2) The moments of X satisfy

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

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

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

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

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

Table 5: 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
$\mathbf{g}_t^{\text{tail}}(\cdot)$	the gradients in the heavy tail region
$\mathbf{g}_t^{\text{body}}(\cdot)$	the gradients in the light body region
$\mathcal{C}_{\mathcal{G}^t}$	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 ([42, 54]). *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 [42]). *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 ([67]). *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 ([13]). *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 ([42]). *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 ([22, 36]). *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 $t > 0$,

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

B.3 Assumptions

Assumption B.1 (**PL Condition** [24]). *Assume that $\forall \mathbf{w} \in W$ in a sample space S, there exists an $\mu_S > 0$ such that*

$$L_S(\mathbf{w}) - L_S(\mathbf{w}_S) \leq (4\mu_S)^{-1}\|\nabla L_S(\mathbf{w})\|_2. \tag{14}$$

PL condition means that gradient grows faster than a quadratic function as moving away from the global optima of function value, which is one of the weakest curvature conditions and is widely employed in non-convex learning [11, 32, 35], such as ResNets with linear activations. Together, these assumptions provide a complementary view of the gradient behavior during optimization. We include other lemmas and theoretical tools in Appendix B.

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* $\hat{\sigma}_{\mathrm{dp}}^2 := m_2\frac{Tdc^2B^2\log(1/\delta)}{n^2\varepsilon^2}$. *If* $\theta = \frac{1}{2}$ *and* $K \leq \hat{\sigma}_{\mathrm{dp}}$, *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* $K \geq \hat{\sigma}_{\mathrm{dp}}$, *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 $\varepsilon_{\mathrm{dp}}$.

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

$$\begin{aligned}
&L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t) \leq \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \nabla L_S(\mathbf{w}_t)\rangle + \frac{1}{2}\beta\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&\leq -\eta_t\langle \overline{\mathbf{g}}_t + \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle + \frac{1}{2}\beta\eta_t^2\|\overline{\mathbf{g}}_t + \zeta_t\|_2^2 \\
&= -\eta_t\langle \overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t] + \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle \\
&- \eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 + \frac{1}{2}\beta\eta_t^2\|\overline{\mathbf{g}}_t\|_2^2 + \frac{1}{2}\beta\eta_t^2\|\zeta_t\|_2^2 + \beta\eta_t^2\langle \overline{\mathbf{g}}_t, \zeta_t\rangle \\
&= -\eta_t\langle \overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t], \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle \\
&- \eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 + \frac{1}{2}\beta\eta_t^2\|\overline{\mathbf{g}}_t\|_2^2 + \frac{1}{2}\beta\eta_t^2\|\zeta_t\|_2^2 + \beta\eta_t^2\langle \overline{\mathbf{g}}_t, \zeta_t\rangle
\end{aligned} \tag{15}$$

Considering all T iterations, we get

$$\begin{aligned}
\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S) + \sum_{t=1}^{T}\frac{1}{2}\beta\eta_t^2c^2 + \underbrace{\sum_{t=1}^{T}\frac{1}{2}\beta\eta_t^2\|\zeta_t\|_2^2}_{\text{Eq.1}} + \underbrace{\sum_{t=1}^{T}\beta\eta_t^2\langle \overline{\mathbf{g}}_t, \zeta_t\rangle}_{\text{Eq.2}} \\
\underbrace{-\sum_{t=1}^{T}\eta_t\langle \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle}_{\text{Eq.3}} \underbrace{-\sum_{t=1}^{T}\eta_t\langle \overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t], \nabla L_S(\mathbf{w}_t)\rangle}_{\text{Eq.4}} \underbrace{-\sum_{t=1}^{T}\eta_t\langle \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle}_{\text{Eq.5}}
\end{aligned} \tag{16}$$

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

$$\begin{aligned}
\sum_{t=1}^{T}\frac{1}{2}\beta\eta_t^2\|\zeta_t\|_2^2 &\leq 2\beta K^2e\log(2/\delta)\sum_{t=1}^{T}\eta_t^2 \\
&\leq 2\beta m_2ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2}\sum_{t=1}^{T}\eta_t^2.
\end{aligned} \tag{17}$$

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

$$\begin{aligned}
\sum_{t=1}^{T}\beta\eta_t^2\langle \overline{\mathbf{g}}_t, \zeta_t\rangle &\leq \sum_{t=1}^{T}\beta\eta_t^2\|\overline{\mathbf{g}}_t\|_2\|\zeta_t\|_2 \\
&\leq \sum_{t=1}^{T}2\beta cK\sqrt{e}\log^{\frac{1}{2}}(2/\delta)\eta_t^2 \\
&\leq 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon}\sum_{t=1}^{T}\eta_t^2.
\end{aligned} \tag{18}$$

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

$$\begin{aligned} -\sum_{t=1}^{T} \eta_t \langle \zeta_t, \nabla L_S(\mathbf{w}_t) \rangle &\leq \sum_{t=1}^{T} \eta_t \|\zeta_t\|_2 \|\nabla L_S(\mathbf{w}_t)\|_2 \\ &\leq \sum_{t=1}^{T} 2cK\sqrt{e} \log^{\frac{1}{2}}(2/\delta)\eta_t \\ &\leq 2\sqrt{em_2 Td} \frac{c^2 B \log(2/\delta)}{n\varepsilon} \sum_{t=1}^{T} \eta_t. \end{aligned} \tag{19}$$

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

$$|\xi_t| \leq \eta_t (\|\overline{\mathbf{g}}_t\|_2 + \|\mathbb{E}_t[\overline{\mathbf{g}}_t]\|_2) \|\nabla L_S(\mathbf{w}_t)\|_2 \leq \eta_t c^2. \tag{20}$$

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

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{E}_t[(\xi_t - \mathbb{E}_t\xi_t)^2] &\leq \sum_{t=1}^{T} \eta_t^2 \mathbb{E}_t[\|\overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t]\|_2^2 \|\nabla L_S(\mathbf{w}_t)\|_2^2] \\ &\leq 4c^2 \sum_{t=1}^{T} \eta_t^2 \|\nabla L_S(\mathbf{w}_t)\|_2^2. \end{aligned} \tag{21}$$

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

$$\sum_{t=1}^{T} \xi_t \leq \frac{\rho 4c^2 \sum_{t=1}^{T} \eta_t^2 \|\nabla L_S(\mathbf{w}_t)\|_2^2}{\eta_t c^2} + \frac{\eta_t c^2 \log(1/\delta)}{\rho}. \tag{22}$$

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

$$\begin{aligned} \|\mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t)\|_2 &= \|\mathbb{E}_t[(\overline{\mathbf{g}}_t - \mathbf{g}_t)a_t]\|_2 \\ &= \|\mathbb{E}_t[(\mathbf{g}_t(\frac{c}{\|\mathbf{g}_t\|_2} - 1)a_t]\|_2 \\ &\leq \mathbb{E}_t[\|(\mathbf{g}_t(\frac{c}{\|\mathbf{g}_t\|_2} - 1)a_t\|_2] \\ &\leq \mathbb{E}_t[|\|\mathbf{g}_t\|_2 - c|a_t] \\ &\leq \mathbb{E}_t[|\|\mathbf{g}_t\|_2 - \|\nabla L_S(\mathbf{w}_t)\|_2|a_t] \\ &\leq \mathbb{E}_t[|\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2|a_t] \\ &\leq \mathbb{E}_t[|\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2|b_t] \\ &\leq \sqrt{\mathbb{E}_t[\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2^2]\mathbb{E}_t b_t^2}. \end{aligned} \tag{23}$$

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

$$\mathbb{E}_t b_t^2 = \mathbb{P}(\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > \frac{c}{2}) \leq 2\exp(-(\frac{c}{4K})^{\frac{1}{\theta}}) \tag{24}$$

So, we get formula (22) as

$$\sqrt{\mathbb{E}_t[\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2^2]\mathbb{E}_t b_t^2} \leq 2\sqrt{K^2\Gamma(2\theta+1)\exp(-(\frac{c}{4K})^{\frac{1}{\theta}})}. \tag{25}$$

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

$$\sum_{t=1}^{T} \eta_t \langle \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t) \rangle \\ \leq 2K^2\Gamma(2\theta+1) \sum_{t=1}^{T} \eta_t \exp(-(\frac{c}{4K})^{\frac{1}{\theta}}) + \frac{1}{2}\sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2^2. \tag{26}$$

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

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

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

$$\frac{1}{4}\sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S) + \frac{1}{2}\beta c^2 + 2\beta m_2 e \frac{d^{\frac{1}{2}} c^2 B^2 \log^{\frac{3}{2}}(2/\delta)}{n\varepsilon} \\ + 2\beta\sqrt{em_2}\frac{d^{\frac{1}{4}} c^2 B \log^{\frac{1}{2}}(2/\delta)}{\sqrt{n\varepsilon}} + 2\sqrt{em_2} c^2 B \log^{\frac{1}{2}}(2/\delta) + \frac{16 d^{\frac{1}{4}} c^2 \log^{\frac{5}{4}}(1/\delta)}{\sqrt{n\varepsilon}} \\ \underbrace{+ 2K^2\Gamma(2\theta+1)\exp(-(\frac{c}{4K})^{\frac{1}{\theta}})\sqrt{T}}_{\text{Eq.6}}. \tag{28}$$

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

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

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

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq \frac{1}{\sqrt{T}}\Bigg(4(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) + 2\beta c^2 + 8\beta m_2 e \frac{d^{\frac{1}{2}} c^2 B^2 \log^{\frac{3}{2}}(2/\delta)}{n\varepsilon} \\ + 8\beta\sqrt{em_2}\frac{d^{\frac{1}{4}} c^2 B \log^{\frac{1}{2}}(2/\delta)}{\sqrt{n\varepsilon}} + 8\sqrt{em_2} c^2 B \log^{\frac{1}{2}}(2/\delta) + \frac{64 d^{\frac{1}{4}} c^2 \log^{\frac{5}{4}}(1/\delta)}{\sqrt{n\varepsilon}} + 8K^2\Gamma(2\theta+1)\Bigg). \tag{30}$$

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

$$\begin{aligned}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq & \frac{4(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S))}{\sqrt{T}} + \frac{8K^2\Gamma(2\theta+1)}{\sqrt{T}} \\ & + \frac{16K^2\log^{2\theta}(\sqrt{T})\log(2/\delta)}{\sqrt{T}}\left(2\beta + 8\beta m_2 e\frac{d^{\frac{1}{2}}B^2\log^{\frac{1}{2}}(2/\delta)}{n\varepsilon}\right. \\ & \left. +8\beta\sqrt{em_2}\frac{d^{\frac{1}{4}}B\log^{-\frac{1}{2}}(2/\delta)}{\sqrt{n\varepsilon}} + 8\sqrt{em_2}B\log^{-\frac{1}{2}}(2/\delta) + \frac{64d^{\frac{1}{4}}\log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\varepsilon}}\right) \\ & \leq \mathbb{O}\left(\frac{\log^{2\theta}(\sqrt{T})\log(1/\delta)}{\sqrt{T}}\cdot\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\varepsilon}}\right) \\ & \leq \mathbb{O}\left(\frac{\log^{2\theta}(\sqrt{T})\log(1/\delta)d^{\frac{1}{4}}\log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\varepsilon}}\right). \end{aligned} \tag{31}$$

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

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

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

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

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq \mathbb{O}\left(\frac{d^{\frac{1}{4}}\log^{2\theta}(\sqrt{T})\log^{\frac{5}{4}}(T/\delta)}{(n\varepsilon)^{\frac{1}{2}}}\right). \tag{33}$$

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

$$\begin{aligned} L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t) &\leq \langle \mathbf{w}_{t+1}-\mathbf{w}_t, \nabla L_S(\mathbf{w}_t)\rangle + \frac{1}{2}\beta\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2 \\ &\leq \underbrace{-\eta_t\langle \overline{\mathbf{g}}_t + \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle}_{\text{Eq.7}} + \underbrace{\frac{1}{2}\beta\eta_t^2\|\overline{\mathbf{g}}_t+\zeta_t\|_2^2}_{\text{Eq.8}} \end{aligned} \tag{34}$$

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 $s_t^- = \mathbb{I}_{\|\mathbf{g}_t\|_2 \leq c}$.

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

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

$$\begin{aligned} -\eta_t\langle \frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2}s_t^+, \nabla L_S(\mathbf{w}_t)\rangle &\leq -\frac{c\eta_t s_t^+\|\nabla L_S(\mathbf{w}_t)\|_2}{3} + \frac{8c\eta_t\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2}{3} \\ &\leq -\frac{c\eta_t(1-s_t^-)\|\nabla L_S(\mathbf{w}_t)\|_2}{3} + \frac{8c\eta_t\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2}{3}. \end{aligned} \tag{36}$$

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

$$\begin{aligned} -\eta_t\langle \mathbf{g}_t s_t^-, \nabla L_S(\mathbf{w}_t)\rangle &= -\eta_t s_t^-(\langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle + \|\nabla L_S(\mathbf{w}_t)\|_2^2) \\ &\leq -\eta_t s_t^-(-\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2\|\nabla L_S(\mathbf{w}_t)\|_2 + \|\nabla L_S(\mathbf{w}_t)\|_2^2) \\ &\leq \eta_t\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2\|\nabla L_S(\mathbf{w}_t)\|_2 - \frac{c}{2}\eta_t s_t^-\|\nabla L_S(\mathbf{w}_t)\|_2 \\ &\leq \eta_t\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2\|\nabla L_S(\mathbf{w}_t)\|_2 - \frac{c}{3}\eta_t s_t^-\|\nabla L_S(\mathbf{w}_t)\|_2. \end{aligned} \tag{37}$$

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

$$\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 \leq K\log^{\theta}(2/\delta), \tag{38}$$

then we get

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

and

$$-\eta_t\langle\frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2}s_t^+,\nabla L_S(\mathbf{w}_t)\rangle\leq-\frac{c\eta_t(1-s_t^-)\|\nabla L_S(\mathbf{w}_t)\|_2}{3}+\frac{8c\eta_tK\log^\theta(2/\delta)}{3}. \tag{40}$$

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

$$-\sum_{t=1}^T\eta_t\langle\zeta_t,\nabla L_S(\mathbf{w}_t)\rangle\leq4\sqrt{em_2Td}\frac{cB\log(2/\delta)}{n\varepsilon}\sum_{t=1}^T\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2. \tag{41}$$

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

$$\begin{aligned}
&-\sum_{t=1}^T\eta_t\langle\overline{\mathbf{g}}_t+\zeta_t,\nabla L_S(\mathbf{w}_t)\rangle\leq-\sum_{t=1}^T\frac{c\eta_t}{3}\|\nabla L_S(\mathbf{w}_t)\|_2+\sum_{t=1}^T\frac{8c\eta_tK\log^\theta(2/\delta)}{3}\\
&+K\log^\theta(2/\delta)\sum_{t=1}^T\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2+4\sqrt{em_2Td}\frac{cB\log(2/\delta)}{n\varepsilon}\sum_{t=1}^T\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2\\
&\leq-\sum_{t=1}^T\frac{c\eta_t}{3}\|\nabla L_S(\mathbf{w}_t)\|_2+(\frac{19}{3}K\log^\theta(2/\delta)+4\sqrt{em_2Td}\frac{cB\log(2/\delta)}{n\varepsilon})\sum_{t=1}^T\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2.
\end{aligned} \tag{42}$$

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

$$\begin{aligned}
&(\frac{c}{3}-\frac{19}{3}K\log^\theta(2/\delta)-4\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta))\sum_{t=1}^T\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2\leq L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S)\\
&+(2\beta m_2ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2}+2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon}+\frac{1}{2}\beta c^2)\sum_{t=1}^T\eta_t^2.
\end{aligned} \tag{43}$$

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

$$\begin{aligned}
&\sum_{t=1}^T\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2\leq\frac{3}{K\log^{\frac{1}{2}}(2/\delta)}(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S))\\
&+\frac{3\sum_{t=1}^T\eta_t^2}{K\log^{\frac{1}{2}}(2/\delta)}\left(2\beta m_2ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2}+2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon}+\frac{1}{2}\beta c^2\right)\\
&\leq\frac{L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S)+2\beta e\hat{\sigma}_{\mathrm{dp}}^2\log(2/\delta)+2\beta c\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(2/\delta)+\frac{39^2}{2}\beta K^2\log(2/\delta)}{\frac{1}{3}K\log^{\frac{1}{2}}(2/\delta)}\\
&\leq\frac{3(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S))}{K\log^{\frac{1}{2}}(2/\delta)}+6\beta eK\log^{\frac{1}{2}}(2/\delta)+6\beta\sqrt{e}\log^{\frac{1}{2}}(2/\delta)+3\beta\frac{(39)^2}{2}K\log^{\frac{1}{2}}(2/\delta).
\end{aligned} \tag{44}$$

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

$$\frac{1}{T}\sum_{t=1}^T\|\nabla L_S(\mathbf{w}_t)\|_2\leq\mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(T/\delta)}{\sqrt{n\varepsilon}}). \tag{45}$$

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

$$\begin{aligned}
\sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2 &\leq \frac{1}{\sqrt{e}\hat{\sigma}_{\text{dp}} \log^{\frac{1}{2}}(1/\delta)} (L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) \\
&+ \frac{\sum_{t=1}^{T} \eta_t^2}{\sqrt{e}\hat{\sigma}_{\text{dp}} \log^{\frac{1}{2}}(1/\delta)} \left(2\beta m_2 ed \frac{Tc^2 B^2 \log^2(2/\delta)}{n^2\varepsilon^2} + 2\beta\sqrt{em_2 Td} \frac{c^2 B \log(2/\delta)}{n\varepsilon} + \frac{1}{2}\beta c^2 \right) \\
&\leq \frac{1}{\sqrt{e}\hat{\sigma}_{\text{dp}} \log^{\frac{1}{2}}(1/\delta)} (L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) \\
&+ \frac{\sum_{t=1}^{T} \eta_t^2}{\sqrt{e}\hat{\sigma}_{\text{dp}} \log^{\frac{1}{2}}(1/\delta)} \left(2\beta e \hat{\sigma}_{\text{dp}}^2 \log(2/\delta) + 2\beta\sqrt{e}\hat{\sigma}_{\text{dp}} \log^{\frac{1}{2}}(2/\delta) + \frac{27^2}{2} \beta e \hat{\sigma}_{\text{dp}}^2 \log(2/\delta) \right) \\
&\leq \frac{L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)}{K \log^{\frac{1}{2}}(2/\delta)} + 2\beta e K \log^{\frac{1}{2}}(2/\delta) + 2\beta\sqrt{e} \log^{\frac{1}{2}}(2/\delta) + \beta \frac{(27)^2}{2} K \log^{\frac{1}{2}}(2/\delta).
\end{aligned} \tag{46}$$

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

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla L_S(\mathbf{w}_t)\|_2 \leq \mathbb{O}(\frac{d^{\frac{1}{4}} \log^{\frac{3}{4}}(T/\delta)}{\sqrt{n\varepsilon}}). \tag{47}$$

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

$$\begin{aligned}
\sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2 &\leq \frac{3}{K \log^{\theta}(2/\delta)} (L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) \\
&+ \frac{3\sum_{t=1}^{T} \eta_t^2}{K \log^{\theta}(2/\delta)} \left(2\beta m_2 ed \frac{Tc^2 B^2 \log^2(2/\delta)}{n^2\varepsilon^2} + 2\beta\sqrt{em_2 Td} \frac{c^2 B \log(2/\delta)}{n\varepsilon} + \frac{1}{2}\beta c^2 \right) \\
&\leq \frac{3(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S))}{K \log^{\theta}(2/\delta)} + \frac{19^2}{24} \beta K \log^{\theta}(2/\delta) + 190\beta K \log^{\theta}(2/\delta) + 3\beta(20)^2 K \log^{\theta}(2/\delta).
\end{aligned} \tag{48}$$

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

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

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

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla L_S(\mathbf{w}_t)\|_2 \leq \mathbb{O}(\frac{d^{\frac{1}{4}} \log^{\theta+\frac{1}{4}}(T/\delta)}{\sqrt{n\varepsilon}}), \tag{50}$$

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 $\eta_t = \frac{1}{\sqrt{T}}$

$$\begin{aligned}
\frac{1}{T} \sum_{t=1}^{T} \min\left\{ \|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2 \right\} &\leq \mathbb{O}(\frac{d^{\frac{1}{4}} \log^{\theta+\frac{1}{4}}(T/\delta)}{(n\varepsilon)^{\frac{1}{2}}}) + \mathbb{O}(\frac{d^{\frac{1}{4}} \log^{2\theta}(\sqrt{T}) \log^{\frac{5}{4}}(T/\delta)}{(n\varepsilon)^{\frac{1}{2}}}) \\
&\leq \mathbb{O}(\frac{d^{\frac{1}{4}} \log^{\frac{5}{4}}(T/\delta) \left(\log^{\theta-1}(T/\delta) + \log^{2\theta}(\sqrt{T}) \right)}{(n\varepsilon)^{\frac{1}{2}}}) \\
&\leq \mathbb{O}(\frac{d^{\frac{1}{4}} \log^{\frac{1}{4}}(T/\delta) \hat{\log}(T/\delta) \log^{2\theta}(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}}),
\end{aligned} \tag{51}$$

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

$$\begin{aligned}
|\lambda_{t,i}^{\text{tr}} - \hat{\lambda}_{t,i}^{\text{tr}}| &:= |\text{tr}(V_{t,k}^\top \hat{\mathbf{g}}_t\hat{\mathbf{g}}_t^\top V_{t,k}) - \text{tr}(\hat{V}_{t,k}^\top \hat{\mathbf{g}}_t\hat{\mathbf{g}}_t^\top \hat{V}_{t,k})| \\
&= |\|V_{t,k}^\top \hat{\mathbf{g}}_t\|_2^2 - \|\hat{V}_{t,k}^\top \hat{\mathbf{g}}_t\|_2^2| \\
&= |\|V_{t,k}V_{t,k}^\top \hat{\mathbf{g}}_t\|_2^2 - \|\hat{V}_{t,k}\hat{V}_{t,k}^\top \hat{\mathbf{g}}_t\|_2^2| \\
&\le \|V_{t,k}V_{t,k}^\top \hat{\mathbf{g}}_t - \hat{V}_{t,k}\hat{V}_{t,k}^\top \hat{\mathbf{g}}_t\|_2^2 \\
&\le \|V_{t,k}V_{t,k}^\top - \hat{V}_{t,k}\hat{V}_{t,k}^\top\|_2^2\|\hat{\mathbf{g}}_t\|_2^2.
\end{aligned} \tag{52}$$

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

$$\begin{aligned}
\|M\|_2 = \|\frac{1}{k}\sum_{i=1}^k v_{t,i}v_{t,i}^\top\|_2 &\le \frac{1}{k}\sum_{i=1}^k \|v_{t,i}v_{t,i}^\top\|_2 \\
&= \sup_{x:\|x\|_2=1}\frac{1}{k}\sum_{i=1}^k x^T v_{t,i}v_{t,i}^\top x \\
&= \sup_{x:\|x\|_2=1}\frac{1}{k}\sum_{i=1}^k \langle x, v_{t,i}\rangle \\
&\le \frac{1}{k}\sum_{i=1}^k \|x\|_2\|v_{t,i}\|_2 \\
&= 1.
\end{aligned} \tag{53}$$

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 $\mu \in (0,1)$

$$\mathbb{P}(\|M - \hat{M}\|_2 > \mu) \le 2d\cdot\exp(\frac{-k\mu^2}{4}), \tag{54}$$

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

$$\mathbb{P}(\|M - \hat{M}\|_2 > 2\sqrt{\frac{\log 2d}{k}}\mu) \le \exp(-\mu^2). \tag{55}$$

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

$$\|M - \hat{M}\|_2 \le 2\frac{\log^{\frac{1}{2}}(2d/\delta_m)}{\sqrt{k}}. \tag{56}$$

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

$$\begin{aligned}
|\lambda_{t,i}^{\mathrm{tr}} - \hat{\lambda}_{t,i}^{\mathrm{tr}}| &\leq \|V_{t,k}V_{t,k}^\top - \hat{V}_{t,k}\hat{V}_{t,k}^\top\|_2^2 \|\hat{\mathbf{g}}_t\|_2^2 \\
&\leq \|V_{t,k}V_{t,k}^\top - \hat{V}_{t,k}\hat{V}_{t,k}^\top\|_2^2 \\
&\leq \|M - \hat{M}\|_2^2 \\
&\leq \frac{4\log(2d/\delta_m)}{k},
\end{aligned} \tag{57}$$

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

$$\begin{aligned}
|\lambda_{t,i}^{\mathrm{tr}} - \hat{\lambda}_{t,i}^{\mathrm{tr}} + \zeta_t^{\mathrm{tr}}| &\leq |\lambda_{t,i}^{\mathrm{tr}} - \hat{\lambda}_{t,i}^{\mathrm{tr}}| + |\zeta_t^{\mathrm{tr}}| \\
&\leq \frac{4\log(2d/\delta_m)}{k} + \sigma_{\mathrm{tr}} \log^{\frac{1}{2}}(2/\delta_{\mathrm{tr}}).
\end{aligned} \tag{58}$$

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

$$\begin{aligned}
&\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S) + \sum_{t=1}^{T}\frac{1}{2}\beta\eta_t^2c^2 + 2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2}\sum_{t=1}^{T}\eta_t^2 \\
&+ 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon}\sum_{t=1}^{T}\eta_t^2 + 2\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon}\sum_{t=1}^{T}\eta_t + \frac{\eta_t c^2\log(1/\delta)}{\rho} \\
&+ \frac{4\rho c^2\sum_{t=1}^{T}\eta_t^2\|\nabla L_S(\mathbf{w}_t)\|_2^2}{\eta_t c^2} - \underbrace{\sum_{t=1}^{T}\eta_t\langle \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle}_{\text{Eq.9}}.
\end{aligned} \tag{59}$$

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

$$\begin{aligned}\|\mathbb{E}_t[\overline{\mathbf{g}}_t]-\nabla L_S(\mathbf{w}_t)\|_2&=\|\mathbb{E}_t[(\overline{\mathbf{g}}_t-\mathbf{g}_t)a_t]\|_2\\&\leq\mathbb{E}_t[\|(\mathbf{g}_t(\frac{c-\|\mathbf{g}_t\|_2}{\|\mathbf{g}_t\|_2})a_t\|_2]\\&\leq\mathbb{E}_t[|\|\mathbf{g}_t\|_2-\|\nabla L_S(\mathbf{w}_t)\|_2|a_t]\\&\leq\mathbb{E}_t[|\|\mathbf{g}_t-\nabla L_S(\mathbf{w}_t)\|_2|b_t]\\&\leq\sqrt{\mathbb{E}_t[\|\mathbf{g}_t-\nabla L_S(\mathbf{w}_t)\|_2^2]\mathbb{E}_t b_t^2}.\end{aligned}\tag{60}$$

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

$$\begin{aligned}\mathbb{P}(\|\mathbf{g}_t-\nabla L_S(\mathbf{w}_t)\|_2>x)&\leq\exp(-c_t\eta I(x))+\exp(-I(x))\\&\leq\exp(-\frac{1}{4}I(x))+\exp(-I(x))\\&\leq 2\exp(-\frac{1}{4}I(x)).\end{aligned}\tag{61}$$

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

$$\begin{aligned}\mathbb{P}(\|\mathbf{g}_t-\nabla L_S(\mathbf{w}_t)\|_2>x)&\leq\exp(-\frac{x^2}{2v(x_{\max},\eta)})+m\exp(-\frac{x_{\max}^2(\eta)}{\eta v(x_{\max},\eta)})\\&\leq 2\exp(-\frac{x^2}{2v(x_{\max},\eta)})\\&\leq 2\exp(-\frac{x^2}{2aK^2})\end{aligned}\tag{62}$$

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

$$\mathbb{E}_t[\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2] \le \sqrt{2a}K \tag{63}$$

and with probability $1-\delta$,

$$\mathbb{E}_t b_t^2 = \mathbb{P}(\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > \frac{c}{2}) \le 2\exp(-(\frac{c}{2\sqrt{2a}K})^2) \tag{64}$$

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

$$\mathbb{E}_t[\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2] \le 4^{\theta}K \tag{65}$$

and

$$\mathbb{E}_t b_t^2 = \mathbb{P}(\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > \frac{c}{2}) \le 2\exp(-\frac{1}{4}(\frac{c}{2K})^{\frac{1}{\theta}}) \tag{66}$$

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

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

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

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

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

$$\sum_{t=1}^{T}\eta_t\langle\mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle \le 4^{2\theta}K^2\sum_{t=1}^{T}\eta_t\exp(-\frac{1}{4}(\frac{c}{2K})^{\frac{1}{\theta}}) + \frac{1}{2}\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2, \tag{68}$$

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

$$\begin{aligned}
&\frac{1}{4}\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 \le L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S) + \frac{1}{2}\beta c^2 + 2\beta m_2 e\frac{d^{\frac{1}{2}}c^2B^2\log^{\frac{3}{2}}(2/\delta)}{n\varepsilon}\\
&+ 2\beta\sqrt{em_2}\frac{d^{\frac{1}{4}}c^2B\log^{\frac{1}{2}}(2/\delta)}{\sqrt{n\varepsilon}} + 2\sqrt{em_2}c^2B\log^{\frac{1}{2}}(2/\delta) + \frac{16d^{\frac{1}{4}}c^2\log^{\frac{5}{4}}(1/\delta)}{\sqrt{n\varepsilon}}\\
&+ \text{Eq.10}\begin{cases} 2aK^2\sum_{t=1}^{T}\eta_t\exp(-(\frac{c}{2\sqrt{2a}K})^2), & \text{if } 0 \le x \le x_{\max},\\ 4^{2\theta}K^2\sum_{t=1}^{T}\eta_t\exp(-\frac{1}{4}(\frac{c}{2K})^{\frac{1}{\theta}}), & \text{if } x \ge x_{\max}.\end{cases}
\end{aligned} \tag{69}$$

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

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

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

$$\begin{aligned}
\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq & \frac{4(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S))}{\sqrt{T}} + \frac{2aK^2}{\sqrt{T}} \\
& + \frac{4^{2\theta+1} \log^{2\theta}(\sqrt{T}) \log(2/\delta)}{\sqrt{T}} \left(2\beta + 8\beta m_2 e B^2 (\frac{d^{\frac{1}{4}} \log^{\frac{1}{4}}(2/\delta)}{\sqrt{n\varepsilon}})^2 \right. \\
& \left. + 8\beta\sqrt{em_2} \frac{d^{\frac{1}{4}} B \log^{-\frac{1}{2}}(2/\delta)}{\sqrt{n\varepsilon}} + 8\sqrt{em_2} B \log^{-\frac{1}{2}}(2/\delta) + \frac{64 d^{\frac{1}{4}} \log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\varepsilon}} \right) \\
\leq & \mathbb{O}(\frac{\log^{2\theta}(\sqrt{T}) \log(1/\delta)}{\sqrt{T}} \cdot \frac{d^{\frac{1}{4}} \log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\varepsilon}}) \\
\leq & \mathbb{O}(\frac{\log^{2\theta}(\sqrt{T}) d^{\frac{1}{4}} \log^{\frac{5}{4}}(1/\delta)}{\sqrt{n\varepsilon}}).
\end{aligned} \tag{71}$$

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

$$\begin{aligned}
L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t) \leq & \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \nabla L_S(\mathbf{w}_t) \rangle + \frac{1}{2}\beta \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
\leq & \underbrace{-\eta_t \langle \overline{\mathbf{g}}_t + \zeta_t, \nabla L_S(\mathbf{w}_t) \rangle}_{\text{Eq.11}} + \frac{1}{2}\beta \eta_t^2 \|\overline{\mathbf{g}}_t + \zeta_t\|_2^2.
\end{aligned} \tag{72}$$

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

$$-\eta_t \langle \overline{\mathbf{g}}_t + \zeta_t, \nabla L_S(\mathbf{w}_t) \rangle = -\eta_t \langle \frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2} s_t^+ + \mathbf{g}_t s_t^-, \nabla L_S(\mathbf{w}_t) \rangle - \eta_t \langle \zeta_t, \nabla L_S(\mathbf{w}_t) \rangle. \tag{73}$$

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

$$\begin{aligned}
-\sum_{t=1}^{T} \eta_t \langle \mathbf{g}_t s_t^-, \nabla L_S(\mathbf{w}_t) \rangle = & -\sum_{t=1}^{T} \eta_t s_t^- (\langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t) \rangle + \|\nabla L_S(\mathbf{w}_t)\|_2^2) \\
\leq & -\sum_{t=1}^{T} \eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t) \rangle - \sum_{t=1}^{T} \eta_t s_t^- \|\nabla L_S(\mathbf{w}_t)\|_2^2 \\
\leq & -\sum_{t=1}^{T} \eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t) \rangle - \frac{c}{2} \sum_{t=1}^{T} \eta_t s_t^- \|\nabla L_S(\mathbf{w}_t)\|_2^2 \\
\leq & \underbrace{-\sum_{t=1}^{T} \eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t) \rangle}_{\text{Eq.12}} - \frac{c}{3} \sum_{t=1}^{T} \eta_t s_t^- \|\nabla L_S(\mathbf{w}_t)\|_2^2.
\end{aligned} \tag{74}$$

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

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

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

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

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

$$\begin{aligned}\mathbb{P}\left(\bigcup_{k\in\mathbb{N}}\left\{\sum_{i=1}^{k}\xi_i \geq kx \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \beta\right\}\right) &\leq \exp(-\frac{kx^2}{v(L,\eta)} + \frac{kx^2}{2v(L,\eta)}) \\ &\leq \exp(-\frac{kx^2}{2v(L,\eta)}).\end{aligned} \tag{77}$$

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 $1 - \delta$

$$\begin{aligned}-\sum_{t=1}^{T} \eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle &\leq \sqrt{2Tv(L,\eta)}\log^{\frac{1}{2}}(1/\delta) \\ &\leq \sqrt{2a\sum_{t=1}^{T} K_t^2}\log^{\frac{1}{2}}(1/\delta) \\ &\leq 2\sqrt{\sum_{t=1}^{T}\eta_t^2 K^2 \|\nabla L_S(\mathbf{w}_t)\|_2^2}\log^{\frac{1}{2}}(1/\delta) \\ &\leq 2KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta),\end{aligned} \tag{78}$$

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

$$\begin{aligned}-\sum_{t=1}^{T} \eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle &\leq 4^{\frac{1}{2}}\frac{1}{T}\sum_{t=1}^{T} K_t \log^{\frac{1}{2}}(1/\delta) \\ &\leq 2\frac{KG}{T}\sum_{t=1}^{T}\eta_t \log^{\frac{1}{2}}(1/\delta).\end{aligned} \tag{79}$$

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 $1-\delta$

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

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

$$-\sum_{t=1}^{T}\eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle \le \frac{4^{\theta}}{T}\sum_{t=1}^{T}K_t\log^{\frac{1}{2}}(1/\delta)$$
$$\le \frac{4^{\theta}KG}{T}\sum_{t=1}^{T}\eta_t\log^{\theta}(1/\delta). \tag{81}$$

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 $1-3\delta$

$$-\sum_{t=1}^{T}\eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle \le \sqrt{2a\sum_{t=1}^{T}K_t^2}\log^{\frac{1}{2}}(1/\delta)$$
$$\le \sqrt{2(2^{2\theta+1}+2)\Gamma(2\theta+1) + \frac{2^{3\theta}\Gamma(3\theta+1)}{3}}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta), \tag{82}$$

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

$$-\sum_{t=1}^{T}\eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle \le \frac{4^{\theta}}{T}\sum_{t=1}^{T}K_t\log^{\frac{1}{2}}(1/\delta)$$
$$\le \frac{4^{\theta}KG}{T}\sum_{t=1}^{T}\eta_t\log^{\theta}(1/\delta). \tag{83}$$

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

$$-\sum_{t=1}^{T}\eta_t\langle \frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2}s_t^+, \nabla L_S(\mathbf{w}_t)\rangle \le -\frac{c\sum_{t=1}^{T}\eta_t s_t^+\|\nabla L_S(\mathbf{w}_t)\|_2}{3} + \frac{8c\sum_{t=1}^{T}\eta_t\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2}{3}$$
$$\le -\frac{c\sum_{t=1}^{T}\eta_t(1-s_t^-)\|\nabla L_S(\mathbf{w}_t)\|_2}{3}$$
$$+\frac{16\sum_{t=1}^{T}\eta_t\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2\|\nabla L_S(\mathbf{w}_t)\|_2}{3}. \tag{84}$$

With the truncated corollaries above, we have

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

$$-\sum_{t=1}^{T}\eta_t\langle \frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2}s_t^+, \nabla L_S(\mathbf{w}_t)\rangle \le -\frac{c\sum_{t=1}^{T}\eta_t(1-s_t^-)\|\nabla L_S(\mathbf{w}_t)\|_2}{3}$$
$$+\frac{16\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2}{3}\begin{cases} 2K\log^{\frac{1}{2}}(2/\delta), & \text{if } \theta = \frac{1}{2}, \\ \sqrt{2}e(4\theta)^{\theta}K\log^{\frac{1}{2}}(2/\delta), & \text{if } \theta \in (\frac{1}{2}, 1], \\ \sqrt{2(2^{2\theta+1}+2)\Gamma(2\theta+1) + \frac{2^{3\theta}\Gamma(3\theta+1)}{3}}K\log^{\frac{1}{2}}(2/\delta) & \text{if } \theta > 1. \end{cases} \tag{85}$$

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

$$
\begin{aligned}
&-\sum_{t=1}^{T}\eta_t\langle\frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2}s_t^+,\nabla L_S(\mathbf{w}_t)\rangle \leq -\frac{c\sum_{t=1}^{T}\eta_t(1-s_t^-)\|\nabla L_S(\mathbf{w}_t)\|_2}{3}\\
&+\frac{16\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2}{3}4^{\theta}K\log^{\theta}(2/\delta).
\end{aligned}
\tag{86}
$$

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 $1-3\delta-T\delta$

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

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

$$
\begin{aligned}
&-\sum_{t=1}^{T}\eta_t\langle\overline{\mathbf{g}}_t,\nabla L_S(\mathbf{w}_t)\rangle \leq -\frac{c\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2}{3}+\frac{4^{\theta}KG}{T}\sum_{t=1}^{T}\eta_t\log^{\theta}(1/\delta)\\
&+\frac{16\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2}{3}4^{\theta}K\log^{\theta}(2/\delta).
\end{aligned}
\tag{88}
$$

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

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

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

$$
\begin{aligned}
&(\frac{c}{3} - \frac{16}{3}aK\log^{\theta}(2/\delta) - 4\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta))\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 \le L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)\\
&+ (2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon} + \frac{1}{2}\beta c^2)\sum_{t=1}^{T}\eta_t^2\\
&+ \sqrt{2a}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\theta}(1/\delta),
\end{aligned}
\tag{90}
$$

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

$$
\begin{aligned}
\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 &\le \frac{3}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) + \frac{3\sqrt{2a}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}\\
&+ \frac{3\sum_{t=1}^{T}\eta_t^2}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}\left(2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon} + \frac{1}{2}\beta c^2\right)\\
&\le \frac{3(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S))}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} + \frac{3\sqrt{2a}KG\log^{\frac{1}{2}}(1/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}\\
&+ \frac{6\beta ea^2K^2\log(2/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} + \frac{6\beta\sqrt{e}\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} + \frac{3\beta(33\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta))^2}{2\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}.
\end{aligned}
\tag{91}
$$

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

$$
\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2 \le \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(T/\delta)}{\sqrt{n\varepsilon}}).
\tag{92}
$$

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

$$
\begin{aligned}
\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 &\le \frac{1}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta)}(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) + \frac{\sqrt{2a}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta)}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta)}\\
&+ \frac{\sum_{t=1}^{T}\eta_t^2}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta)}\left(2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon} + \frac{1}{2}\beta c^2\right)\\
&\le \frac{L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(2/\delta)} + \frac{\sqrt{2a}KG}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}} + 2\beta eK\log^{\frac{1}{2}}(2/\delta) + 2\beta\sqrt{e}\log^{\frac{1}{2}}(2/\delta) + \beta\frac{(27)^2}{2}K\log^{\frac{1}{2}}(2/\delta).
\end{aligned}
\tag{93}
$$

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

$$
\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2 \le \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(T/\delta)}{\sqrt{n\varepsilon}}).
\tag{94}
$$

(2) In case of heavy tail, when $x \geq x_{\max}$:
If $\theta = \frac{1}{2}$ and $K \geq \hat{\sigma}_{\text{dp}}$, 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

$$\begin{aligned}
&\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \frac{3}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) + \frac{3\sqrt{2a}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} \\
&+ \frac{3\sum_{t=1}^{T}\eta_t^2}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}\left(2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon} + \frac{1}{2}\beta c^2\right) \\
&\leq \frac{3(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S))}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} + \frac{3\sqrt{2a}KG\log^{\frac{1}{2}}(1/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} \\
&+ \frac{6\beta ea^2K^2\log(2/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} + \frac{6\beta\sqrt{e}\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} + \frac{3\beta(33\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta))^2}{2\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}.
\end{aligned} \tag{95}$$

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

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(T/\delta)}{\sqrt{n\varepsilon}}). \tag{96}$$

If $\theta = \frac{1}{2}$ and $K \leq \hat{\sigma}_{\text{dp}}$, 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

$$\begin{aligned}
&\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \frac{1}{\sqrt{e}\hat{\sigma}_{\text{dp}}\log^{\frac{1}{2}}(1/\delta)}(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) + \frac{\sqrt{2a}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta)}{\sqrt{e}\hat{\sigma}_{\text{dp}}\log^{\frac{1}{2}}(1/\delta)} \\
&+ \frac{\sum_{t=1}^{T}\eta_t^2}{\sqrt{e}\hat{\sigma}_{\text{dp}}\log^{\frac{1}{2}}(1/\delta)}\left(2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon} + \frac{1}{2}\beta c^2\right) \\
&\leq \frac{L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)}{\sqrt{e}\hat{\sigma}_{\text{dp}}\log^{\frac{1}{2}}(2/\delta)} + \frac{\sqrt{2a}KG}{\sqrt{e}\hat{\sigma}_{\text{dp}}} + 2\beta eK\log^{\frac{1}{2}}(2/\delta) + 2\beta\sqrt{e}\log^{\frac{1}{2}}(2/\delta) + \beta\frac{(27)^2}{2}K\log^{\frac{1}{2}}(2/\delta).
\end{aligned} \tag{97}$$

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

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(T/\delta)}{\sqrt{n\varepsilon}}). \tag{98}$$

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

$$\begin{aligned}
&\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \frac{3}{\sqrt{2a}K\log^{\theta}(2/\delta)}(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) + \frac{3\sqrt{2a}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\theta}(1/\delta)}{\sqrt{2a}K\log^{\theta}(2/\delta)} \\
&+ \frac{3\sum_{t=1}^{T}\eta_t^2}{\sqrt{2a}K\log^{\theta}(2/\delta)}\left(2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\varepsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\varepsilon} + \frac{1}{2}\beta c^2\right) \\
&\leq \frac{3(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S))}{\sqrt{2a}K\log^{\theta}(2/\delta)} + 3G + \frac{16^2}{24}\beta K\log^{\theta}(2/\delta) + 136\beta K\log^{\theta}(2/\delta) + 3\beta(17)^2K\log^{\theta}(2/\delta).
\end{aligned} \tag{99}$$

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

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

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

For light body, we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(T/\delta)}{\sqrt{n\varepsilon}}), \tag{101}$$

For heavy tail, we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\theta+\frac{1}{4}}(T/\delta)}{\sqrt{n\varepsilon}}), \tag{102}$$

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:

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

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:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} &\leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(T/\delta)}{(n\varepsilon)^{\frac{1}{2}}}) + \mathbb{O}(\frac{d^{\frac{1}{4}}\log(\sqrt{T})\log^{\frac{5}{4}}(T/\delta)}{(n\varepsilon)^{\frac{1}{2}}}) \\ &\leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(T/\delta)\left(\log^{\frac{1}{2}}(T/\delta)+\log(\sqrt{T})\log(T/\delta)\right)}{(n\varepsilon)^{\frac{1}{2}}}) \\ &\leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta)\log(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}}),\end{aligned} \tag{104}$$

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

$$\begin{aligned}
&\mathcal{C}_{\text{m}}(c_1, c_2) \coloneqq \frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} \\
&= (1-\delta'_m) * p * \mathcal{C}_{\text{tail}}(c_1) + (1-\delta'_m) * (1-p) * \mathcal{C}_{\text{body}}(c_2) + \delta'_m * |\mathcal{C}_{\text{tail}}(c_1) - \mathcal{C}_{\text{body}}(c_2)|. \qquad (105)
\end{aligned}$$

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. $\mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(T/\delta)\hat{\log}(1/\delta)\log^{2\theta}(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}})$, $\mathcal{C}_{\text{body}}(c_2)$ 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. $\mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta)\log(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}})$, 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

$$\mathcal{C}_{\text{m}}(c_1, c_2) = \mathcal{C}_{\text{tail}}(c_1) = \mathbb{O}\left(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta)\log(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}}\right). \qquad (106)$$

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

$$\mathcal{C}_{\text{tail}}(c_1) \geq \mathcal{C}_{\text{body}}(c_2) - \frac{\delta'_m}{(1-\delta'_m) * (1-p)}\mathcal{C}_{\text{body}}(c_2), \qquad (107)$$

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

$$\mathcal{C}_{\text{m}}(c_1, c_2) = p * \mathcal{C}_{\text{tail}}(c_1) + *(1-p) * \mathcal{C}_{\text{body}}(c_2). \qquad (108)$$

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

$$\frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} \leq p * \mathbb{O}\left(\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(T/\delta')\hat{\log}(T/\delta')\log^{2\theta}(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}}\right) + (1-p) * \mathbb{O}\left(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta')\log(\sqrt{T})}{(n\varepsilon)^{\frac{1}{2}}}\right), \tag{109}$$

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 Non-convex Results with PL Condition with Sharper Rates

THEOREM G.1. *Let $\mathbf{w}_t$ be the iterative parameter produced by DC-DPSGD. Under Assumptions 2.1, 2.2 and B.1, assuming that $\eta_t \leq \frac{1}{2\beta}$, $c_1 = \max\left(\mathbb{O}(\log^{\theta}(\sqrt{T})), \mathbb{O}(\log^{\theta}(1/\delta'))\right)$ and $c_2 = \max\left(\mathbb{O}(\log^{1/2}(\sqrt{T})), \mathbb{O}(\log^{1/2}(1/\delta'))\right)$, for any $\delta' \in (0,1)$, and we achieve:*

$$L_S(\mathbf{w}_{T+1}) - L_S(\mathbf{w}_S) \leq p * \mathbb{O}\left(\frac{\log^{2\theta}(T)\log^{\theta+\frac{1}{2}}(T/\delta')\sqrt{d\log(T/\delta')}}{n\varepsilon}\right) + (1-p) * \mathbb{O}\left(\frac{\log(\sqrt{T})\log(T/\delta')\sqrt{d\log(T/\delta')}}{n\varepsilon}\right),$$

where the probability δ' contains the broken probability of subspace identification and the convergence of DC-DPSGD under PL condition.

PROOF. For the iterative process $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t(\widetilde{\mathbf{g}}_t - \nabla L_S(\mathbf{w}_t) + \nabla L_S(\mathbf{w}_t))$, by using $\mathbf{w}_0 = 0$, we summate the iterations and get

$$\mathbf{w}_{t+1} = \sum_{i=1}^{t} -\eta_i(\widetilde{\mathbf{g}}_i - \nabla L_S(\mathbf{w}_i)) - \sum_{i=1}^{t}\eta_i \nabla L_S(\mathbf{w}_i).$$

Firstly, in the case $\|\nabla L_S(\mathbf{w}_i)\|_2 \leq c/2$. This implies that

$$\begin{aligned}\|\mathbf{w}_{t+1}\|_2 &= \|\sum_{i=1}^{t}\eta_i(\widetilde{\mathbf{g}}_i - \nabla L_S(\mathbf{w}_i))\|_2 + \|\sum_{i=1}^{t}\eta_i\nabla L_S(\mathbf{w}_i)\|_2 \\ &= \|\sum_{i=1}^{t}\eta_i(\overline{\mathbf{g}}_i - [\overline{\mathbf{g}}_i])\|_2 + \|\sum_{i=1}^{t}\eta_i(\mathbb{E}_i[\overline{\mathbf{g}}_i] - \nabla L_S(\mathbf{w}_i))\|_2 + \|\sum_{i=1}^{t}\eta_i\nabla L_S(\mathbf{w}_i)\|_2 + \|\sum_{i=1}^{t}\eta_i\zeta_i\|_2. \end{aligned} \tag{110}$$

In the first term, by setting $\xi_i = \eta_i(\overline{\mathbf{g}}_i - \mathbb{E}_i[\overline{\mathbf{g}}_i])$, since $|\xi_i| \leq 2\eta_i c$ and $\mathbb{E}_i[(\xi_i - \mathbb{E}_i\xi_i)^2] \leq \mathbb{E}_i[\xi_i^2] \leq 4\eta_i^2c^2$, according to Lemma B.4, with probability $1-\delta$ and $\rho > 0$ then we have

$$\sum_{i=1}^{t}\xi_i \leq \frac{4\rho c^2\sum_{i=1}^{t}\eta_i^2}{\eta_i c^2} + \frac{\eta_i c^2\log(1/\delta)}{\rho}.$$

Thus, we get

$$\|\sum_{i=1}^{t}\eta_i(\overline{\mathbf{g}}_i - \mathbb{E}_i[\overline{\mathbf{g}}_i])\|_2 \leq \frac{\rho 4c^2\sum_{i=1}^{t}\eta_i^2}{\eta_i c^2} + \frac{\eta_i c^2\log(1/\delta)}{\rho}. \tag{111}$$

From the formula (22), we can derive that the second term $\|\sum_{i=1}^{t}\eta_i(\mathbb{E}_i[\overline{\mathbf{g}}_i] - \nabla L_S(\mathbf{w}_i))\|_2 \leq \sqrt{\mathbb{E}_i[\|\sum_{i=1}^{t}\eta_i(\mathbf{g}_i - \nabla L_S(\mathbf{w}_i))\|_2^2]\mathbb{E}_i b_i^2}$, where $b_i = \mathbb{I}_{\|\mathbf{g}_i - \nabla L_S(\mathbf{w}_i)\|_2 > \frac{c}{2}}$. Due to the term $\|\mathbf{g}_i - \nabla L_S(\mathbf{w}_i)\| \sim subW(\theta, K)$ and $(\eta_i(\mathbf{g}_i - \nabla L_S(\mathbf{w}_i)), i \in \mathbb{N})$ is a martingale difference sequence, we have $\|\eta_i(\mathbf{g}_i - \nabla L_S(\mathbf{w}_i))\| \sim subW(\eta_i\theta, K)$. Then, we apply Lemma B.10 to derive the result

$$\begin{aligned}&\mathbb{P}\left(\max_{1\leq t\leq T}\left\|\sum_{i=1}^{t}\eta_i\left(\mathbf{g}_i - \nabla L_S(\mathbf{w}_i)\right)\right\|_2 \geq x\right) \\ &\leq 4\left[3 + (3\theta)^{2\theta}\frac{128K^2\sum_{i=1}^{T}\eta_i^2}{x^2}\right]\exp\left\{-\left(\frac{x^2}{64K^2\sum_{i=1}^{T}\eta_i^2}\right)^{\frac{1}{2\theta+1}}\right\}. \end{aligned} \tag{112}$$

Taking the dominated term $4\exp\left\{-\left(\frac{x^2}{64K^2\sum_{i=1}^{T}\eta_i^2}\right)^{\frac{1}{2\theta+1}}\right\}$ as δ, we have $x = 8\log^{\theta+\frac{1}{2}}(4/\delta)K(\sum_{i=1}^{T}\eta_i^2)^{\frac{1}{2}}$. Thus, with high probability $1 - 3\delta - \frac{8(3\theta)^{2\theta}}{\log^{2\theta+1}(4/\delta)}\delta$, we achieve

$$\max_{1\leq t\leq T}\left\|\sum_{i=1}^{t}\eta_i\left(\mathbf{g}_i - \nabla L_S(\mathbf{w}_i)\right)\right\|_2 \leq 8\log^{\theta+\frac{1}{2}}(4/\delta)K(\sum_{i=1}^{T}\eta_i^2)^{\frac{1}{2}}.$$

Considering that $\theta \geq 1/2$ and $\delta \in (0,1)$, the term $\log^{2\theta+1}(4/\delta) > 1$. Thus, with probability at least $1 - 3\delta - 8(3\theta)^{2\theta}\delta$, we have

$$\max_{1\leq t\leq T}\left\|\sum_{i=1}^{t}\eta_i\left(\mathbf{g}_i - \nabla L_S(\mathbf{w}_i)\right)\right\|_2 \leq 8\log^{\theta+\frac{1}{2}}(4/\delta)K(\sum_{i=1}^{T}\eta_i^2)^{\frac{1}{2}}. \tag{113}$$

This means that with probability $1-\delta$, we have

$$\max_{1\le t\le T}\left\|\sum_{i=1}^{t}\eta_i\left(\mathbf{g}_i-\nabla L_S(\mathbf{w}_i)\right)\right\|_2 \le 8\log^{\theta+\frac{1}{2}}\left(\frac{4(3+8(3\theta)^{2\theta})}{\delta}\right)K\left(\sum_{i=1}^{T}\eta_i^2\right)^{\frac{1}{2}}. \tag{114}$$

Combining the formula (63) and formula (65) to bound the $\mathbb{E}_i b_t^2$, by dividing into the light body region and heavy tail region, we have

(1) In the light body region, due to $\|\mathbf{g}_i-\nabla L_S(\mathbf{w}_i)\|_2 \sim subW(1/2,K)$, we can formalize the formula (113) with $\theta=1/2$ and have

$$\begin{aligned}\|\sum_{i=1}^{t}\eta_i(\mathbb{E}_i[\overline{\mathbf{g}}_i]-\nabla L_S(\mathbf{w}_i))\|_2 &\le \sqrt{\mathbb{E}_i[\|\sum_{i=1}^{t}\eta_i(\mathbf{g}_i-\nabla L_S(\mathbf{w}_i))\|_2^2]\mathbb{E}_i b_i^2}\\ &\le 16\log\left(\frac{60}{\delta}\right)K\left(\sum_{i=1}^{T}\eta_i^2\right)^{\frac{1}{2}}\exp\left(-\left(\frac{c}{2\sqrt{2aK}}\right)^2\right).\end{aligned}$$

Taking $c_2=c=2\sqrt{2aK}\log^{\frac{1}{2}}(\sqrt{T})$, we derive

$$\|\sum_{i=1}^{t}\eta_i(\mathbb{E}_i[\overline{\mathbf{g}}_i]-\nabla L_S(\mathbf{w}_i))\|_2 \le 16\log\left(\frac{60}{\delta}\right)K\left(\sum_{i=1}^{T}\eta_i^2\right)^{\frac{1}{2}}\frac{1}{\sqrt{T}}. \tag{115}$$

(2) In the heavy tail region, due to $\|\mathbf{g}_i-\nabla L_S(\mathbf{w}_i)\|_2 \sim subW(\theta,K)$, where $\theta>1/2$, we get

$$\|\sum_{i=1}^{t}\eta_i(\mathbb{E}_i[\overline{\mathbf{g}}_i]-\nabla L_S(\mathbf{w}_i))\|_2 \le 16\log^{\theta+\frac{1}{2}}\left(\frac{4(3+8(3\theta)^{2\theta})}{\delta}\right)K\left(\sum_{i=1}^{T}\eta_i^2\right)^{\frac{1}{2}}\exp\left(-\frac{1}{4}\left(\frac{c}{2K}\right)^{\frac{1}{\theta}}\right).$$

Setting $c_1=c=4^{\theta}2K\log^{\theta}(\sqrt{T})$, we have

$$\|\sum_{i=1}^{t}\eta_i(\mathbb{E}_i[\overline{\mathbf{g}}_i]-\nabla L_S(\mathbf{w}_i))\|_2 \le 16\log^{\theta+\frac{1}{2}}\left(\frac{4(3+8(3\theta)^{2\theta})}{\delta}\right)K\left(\sum_{i=1}^{T}\eta_i^2\right)^{\frac{1}{2}}\frac{1}{\sqrt{T}}. \tag{116}$$

For the third term $\|\sum_{i=1}^{t}\eta_i\nabla L_S(\mathbf{w}_i)\|_2$, with Schwartz's inequality and high probability $1-\delta$ uniformly over $t=1,...,T$, we have

$$\begin{aligned}\|\sum_{i=1}^{t}\eta_i\nabla L_S(\mathbf{w}_i)\|_2^2 &\le \left(\sum_{i=1}^{t}\eta_i\right)\left(\sum_{i=1}^{t}\eta_i\|\nabla L_S(\mathbf{w}_i)\|_2^2\right)\\ &\le \left(\sum_{i=1}^{t}\eta_i\right)\mathbb{O}\left(\frac{\log^{2\theta}(\sqrt{T})d^{\frac{1}{4}}\log^{\frac{5}{4}}(1/\delta)\sqrt{T}}{\sqrt{n\varepsilon}}\right),\end{aligned} \tag{117}$$

where the second inequality follows the formulas.(67) and (68), and $\theta=1/2$ in the light body region as well. □

Regard to the fourth term $\|\sum_{i=1}^{t}\eta_i\zeta_i\|_2$, we can use sub-Gaussian properties and Lemma B.2 to bound it and have

$$\begin{aligned}\|\sum_{i=1}^{t}\eta_i\zeta_i\|_2 &\le \sum_{i=1}^{t}\eta_i\|\zeta_i\|_2\\ &\le \frac{\sqrt{m_2edT}cB\log(1/\delta)}{n\varepsilon}\sum_{i=1}^{t}\eta_i.\end{aligned} \tag{118}$$

To sum up all terms in (107), we have

(1) In the light body region,

$$\begin{aligned}\|\mathbf{w}_{t+1}\|_2 \le{}& \mathbb{O}\left(\frac{\sum_{i=1}^{t}\eta_i^2}{\eta_i}+c^2\eta_i\log(1/\delta)\right)+\mathbb{O}\left(\frac{\log(1/\delta)}{\sqrt{T}}\left(\sum_{i=1}^{T}\eta_i^2\right)^{\frac{1}{2}}\right)\\ &+\mathbb{O}\left(\sum_{i=1}^{t}\eta_i\frac{\log(\sqrt{T})d^{\frac{1}{4}}\log^{\frac{5}{4}}(1/\delta)\sqrt{T}}{\sqrt{n\varepsilon}}\right)+\mathbb{O}\left(\sum_{i=1}^{t}\eta_i\frac{\sqrt{d}Tc\log(1/\delta)}{n\varepsilon}\right).\end{aligned} \tag{119}$$

(2) In the heavy tail region,

$$\|\mathbf{w}_{t+1}\|_2 \leq \mathbb{O}\left(\frac{\sum_{i=1}^t \eta_i^2}{\eta_i} + c^2\eta_i \log(1/\delta)\right) + \mathbb{O}\left(\frac{\log^{\theta+\frac{1}{2}}(1/\delta)}{\sqrt{T}}(\sum_{i=1}^T \eta_i^2)^{\frac{1}{2}}\right) + \mathbb{O}\left(\sum_{i=1}^t \eta_i \frac{\log^{2\theta}(\sqrt{T}) d^{\frac{1}{4}} \log^{\frac{5}{4}}(1/\delta)\sqrt{T}}{\sqrt{n\varepsilon}}\right) + \mathbb{O}\left(\sum_{i=1}^t \eta_i \frac{\sqrt{d}Tc\log(1/\delta)}{n\varepsilon}\right). \tag{120}$$

Secondly, in the case $\|\nabla L_S(\mathbf{w}_i)\|_2 > c/2$. By setting $s_i^+ = \mathbb{I}_{\|\mathbf{g}_i\|_2 \geq c}$ and $s_i^- = \mathbb{I}_{\|\mathbf{g}_i\|_2 \leq c}$, it implies that

$$\begin{aligned}
\|\mathbf{w}_{t+1}\|_2 &= \|\sum_{i=1}^t \eta_i(\widetilde{\mathbf{g}}_i - \nabla L_S(\mathbf{w}_i))\|_2 + \|\sum_{i=1}^t \eta_i \nabla L_S(\mathbf{w}_i)\|_2 \\
&\leq \|\sum_{i=1}^t \eta_i(\frac{c\mathbf{g}_i}{\|\mathbf{g}_i\|_2}s_i^+ + \mathbf{g}_i s_i^- - (\frac{cs_i^+}{\|\mathbf{g}_i\|_2} + s_i^-)\nabla L_S(\mathbf{w}_i))\|_2 + \|\sum_{i=1}^t \eta_i \nabla L_S(\mathbf{w}_i)\|_2 + \|\sum_{i=1}^t \eta_i \zeta_i\|_2 \\
&\leq \|\sum_{i=1}^t \eta_i s_i^+(\frac{c\mathbf{g}_i}{\|\mathbf{g}_i\|_2} - \frac{c}{\|\mathbf{g}_i\|_2}\nabla L_S(\mathbf{w}_i))\|_2 + \|\sum_{i=1}^t \eta_i s_i^-(\mathbf{g}_i - \nabla L_S(\mathbf{w}_i))\|_2 \\
&+ \|\sum_{i=1}^t \eta_i \nabla L_S(\mathbf{w}_i)\|_2 + \|\sum_{i=1}^t \eta_i \zeta_i\|_2.
\end{aligned} \tag{121}$$

For the term $\|\sum_{i=1}^t \eta_i s_i^+(\frac{c\mathbf{g}_i}{\|\mathbf{g}_i\|_2} - \frac{c}{\|\mathbf{g}_i\|_2}\nabla L_S(\mathbf{w}_i))\|_2$, we have

$$\|\sum_{i=1}^t \eta_i s_i^+(\frac{c\mathbf{g}_i}{\|\mathbf{g}_i\|_2} - \frac{c}{\|\mathbf{g}_i\|_2}\nabla L_S(\mathbf{w}_i))\|_2 \leq \|\sum_{i=1}^t \eta_i(\mathbf{g}_i - \nabla L_S(\mathbf{w}_i))\|_2.$$

Thus, by using the formula (113), we can also bound the term with probability $1-\delta$

$$\max_{1\leq t\leq T}\left\|\sum_{i=1}^t \eta_i\left(\mathbf{g}_i - \nabla L_S(\mathbf{w}_i)\right)\right\| \leq 8\log^{\theta+\frac{1}{2}}(\frac{4(3+8(3\theta)^{2\theta})}{\delta})K(\sum_{i=1}^T \eta_i^2)^{\frac{1}{2}}. \tag{122}$$

Similarly, the bound of (117) for the term $\|\sum_{i=1}^t \eta_i s_i^-(\mathbf{g}_i - \nabla L_S(\mathbf{w}_i))\|_2$ also holds true.

Since the remaining two terms can already be covered by the above conclusions (the formulas.(88-99), (112) and (113)), we can achieve

(1) In the light body region,

$$\|\mathbf{w}_{t+1}\|_2 \leq \mathbb{O}\left(\log(\frac{1}{\delta})(\sum_{i=1}^T \eta_i^2)^{\frac{1}{2}}\right) + \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(T/\delta)\sqrt{T}}{\sqrt{n\varepsilon}}) + \mathbb{O}\left(\sum_{i=1}^t \eta_i \frac{\sqrt{d}Tc\log(1/\delta)}{n\varepsilon}\right). \tag{123}$$

(2) In the heavy tail region,

$$\|\mathbf{w}_{t+1}\|_2 \leq \mathbb{O}\left(\log^{\theta+\frac{1}{2}}(\frac{1}{\delta})(\sum_{i=1}^T \eta_i^2)^{\frac{1}{2}}\right) + \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\theta+\frac{1}{4}}(T/\delta)\sqrt{T}}{\sqrt{n\varepsilon}}) + \mathbb{O}\left(\sum_{i=1}^t \eta_i \frac{\sqrt{d}Tc\log(1/\delta)}{n\varepsilon}\right). \tag{124}$$

Next, we revisit the convergence rate under the PL condition (Assumption B.1) with the prepared $\mathbf{w}_{t+1}$. In the same way, we divide it into cases: $\|\nabla L_S(\mathbf{w}_i)\|_2 \leq c/2$ and $\|\nabla L_S(\mathbf{w}_i)\|_2 > c/2$.

In the first case of $\|\nabla L_S(\mathbf{w}_i)\|_2 \leq c/2$, we know that,

$$\begin{aligned}
L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t) &\leq \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \nabla L_S(\mathbf{w}_t)\rangle + \frac{1}{2}\beta\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&\leq -\eta_t\langle \overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t], \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle \\
&- \eta_t\langle \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 + \frac{1}{2}\beta\eta_t^2\|\overline{\mathbf{g}}_t\|_2^2 + \frac{1}{2}\beta\eta_t^2\|\zeta_t\|_2^2 + \beta\eta_t^2\langle \overline{\mathbf{g}}_t, \zeta_t\rangle,
\end{aligned} \tag{125}$$

from which we get

$$\begin{aligned}
L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t) &\leq -\eta_t\langle \overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t], \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle \\
&- \eta_t\langle \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle - \frac{3\eta_t}{4}\|\nabla L_S(\mathbf{w}_t)\|_2^2 - \eta_t\mu_S(L_S(\mathbf{w}_t) - L_S(\mathbf{w}_S)) \\
&+ \frac{1}{2}\beta\eta_t^2\|\overline{\mathbf{g}}_t\|_2^2 + \frac{1}{2}\beta\eta_t^2\|\zeta_t\|_2^2 + \beta\eta_t^2\langle \overline{\mathbf{g}}_t, \zeta_t\rangle,
\end{aligned}$$

then

$$L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_S) + \frac{3\eta_t}{4}\|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq -\eta_t\langle\overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t], \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle\mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle$$
$$- \eta_t\langle\zeta_t, \nabla L_S(\mathbf{w}_t)\rangle + (1 - \eta_t\mu_S)(L_S(\mathbf{w}_t) - L_S(\mathbf{w}_S))$$
$$+ \frac{1}{2}\beta\eta_t^2\|\overline{\mathbf{g}}_t\|_2^2 + \frac{1}{2}\beta\eta_t^2\|\zeta_t\|_2^2 + \beta\eta_t^2\langle\overline{\mathbf{g}}_t, \zeta_t\rangle.$$

Setting $\mu_S = \frac{2}{\eta_t(t+t_0)}$ and multiplying both sides by $(t+t_0)(t+t_0-1)$, we have

$$(t+t_0)(t+t_0-1)(L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_S)) + \frac{3(t+t_0-1)}{2\mu_S}\|\nabla L_S(\mathbf{w}_t)\|_2^2$$
$$\leq -(t+t_0)(t+t_0-1)(\eta_t\langle\overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t], \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle\mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle\zeta_t, \nabla L_S(\mathbf{w}_t)\rangle)$$
$$+ (t+t_0-1)(t+t_0-2)(L_S(\mathbf{w}_t) - L_S(\mathbf{w}_S)) + 2\frac{\beta}{\mu_S^2}\|\overline{\mathbf{g}}_t\|_2^2 + 2\frac{\beta}{\mu_S^2}\|\zeta_t\|_2^2 + \frac{4\beta}{\mu_S^2}\langle\overline{\mathbf{g}}_t, \zeta_t\rangle.$$

Taking a summation from $t=1$ to $t=T$, we can achieve

$$(T+t_0)(T+t_0-1)(L_S(\mathbf{w}_{T+1}) - L_S(\mathbf{w}_S)) + \sum_{t=1}^{T}\frac{3(t+t_0-1)}{2\mu_S}\|\nabla L_S(\mathbf{w}_t)\|_2^2$$
$$\leq -\sum_{t=1}^{T}(t+t_0)(t+t_0-1)(\eta_t\langle\overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t], \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle\mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle\zeta_t, \nabla L_S(\mathbf{w}_t)\rangle)$$
$$+ t_0(t_0-1)(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) + 2\sum_{t=1}^{T}\frac{\beta}{\mu_S^2}(\|\overline{\mathbf{g}}_t\|_2^2 + \|\zeta_t\|_2^2 + 2\langle\overline{\mathbf{g}}_t, \zeta_t\rangle).$$

Following the results of the formulas.(64-66), we can use Lemma B.4 to bound the term $\eta_t\langle\overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t], \nabla L_S(\mathbf{w}_t)\rangle$, Lemma B.7 and Lemma B.8 to bound the term $\eta_t\langle\mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle$, and Lemma B.2 to bound the sub-Gaussian DP noise ζ_t. Then, due to $\|\overline{\mathbf{g}}_t\|_2 \leq c$ and $\rho = \frac{1}{24}$, with high probability $1-\delta$, we have

$$(T+t_0)(T+t_0-1)(L_S(\mathbf{w}_{T+1}) - L_S(\mathbf{w}_S)) + \sum_{t=1}^{T}\left(\frac{t+t_0-1}{6\mu_S}\right)\|\nabla L_S(\mathbf{w}_t)\|_2^2$$
$$\leq \sum_{t=1}^{T}(t+t_0)(t+t_0-1)\eta_t\mathbb{O}\left(\sqrt{Td}\frac{c^2\log(1/\delta)}{n\varepsilon} + c^2\log(1/\delta)\right)$$
$$+ \begin{cases} \sum_{t=1}^{T}(t+t_0)(t+t_0-1)\eta_t 2aK^2\exp(-(\frac{c}{2\sqrt{2a}K})^2), & \text{if located in the light body region,} \\ \sum_{t=1}^{T}(t+t_0)(t+t_0-1)\eta_t 4^{2\theta}K^2\exp(-\frac{1}{4}(\frac{c}{2K})^{\frac{1}{\theta}}), & \text{if located in the heavy tail region.} \end{cases}$$
$$+ t_0(t_0-1)(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) + \frac{1}{2}\sum_{t=1}^{T}\frac{\beta}{\mu_S^2}\mathbb{O}\left(c^2 + d\frac{Tc^2\log^2(1/\delta)}{n^2\varepsilon^2} + \sqrt{Td}\frac{c^2\log(1/\delta)}{n\varepsilon}\right). \tag{126}$$

When $t_0 \geq \frac{4\beta}{\mu_S}$ and $\eta_t = \frac{2}{\mu_S(t+t_0)}$, we have $\eta_t \leq \frac{1}{2\beta}$ and

$$
\begin{aligned}
&(T+t_0)(T+t_0-1)(L_S(\mathbf{w}_{T+1}) - L_S(\mathbf{w}_S)) + \sum_{t=1}^{T}\left(\frac{t+t_0-1}{6\mu_S}\right)\|\nabla L_S(\mathbf{w}_t)\|_2^2 \\
&\leq \sum_{t=1}^{T} 2(t+t_0-1)\mu_S^{-1}\mathbb{O}\left(\sqrt{Td}\frac{c^2\log(1/\delta)}{n\varepsilon} + c^2\log(1/\delta)\right) \\
&+ \begin{cases} \sum_{t=1}^{T}(t+t_0-1)\mu_S^{-1}4aK^2\exp(-(\frac{c}{2\sqrt{2a}K})^2), & \text{if located in the light body region,} \\ \sum_{t=1}^{T} 2(t+t_0-1)\mu_S^{-1}4^{2\theta}K^2\exp(-\frac{1}{4}(\frac{c}{2K})^{\frac{1}{\theta}}), & \text{if located in the heavy tail region.} \end{cases} \\
&+ t_0(t_0-1)(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) + \frac{1}{2}\sum_{t=1}^{T}\frac{\beta}{\mu_S^2}\mathbb{O}\left(c^2 + d\frac{Tc^2\log^2(1/\delta)}{n^2\varepsilon^2} + \sqrt{Td}\frac{c^2\log(1/\delta)}{n\varepsilon}\right).
\end{aligned} \tag{127}
$$

Thus, from (122), considering that $T = \mathbb{O}\left(\frac{n\varepsilon}{\sqrt{d\log(1/\delta)}}\right)$, we just focus on privacy parameters, probability δ and iterations T, and have

(1) In the light body region, setting $c_2 = c = 2\sqrt{2a}K\log^{\frac{1}{2}}(\sqrt{T})$, the term $\sum_{t=1}^{T} 2(t+t_0-1)\mu_S^{-1}\mathbb{O}\left(\sqrt{Td}\frac{c^2\log(1/\delta)}{n\varepsilon} + c^2\log(1/\delta)\right)$ is dominated and

$$
L_S(\mathbf{w}_{T+1}) - L_S(\mathbf{w}_S) \leq \mathbb{O}\left(\frac{\log(\sqrt{T})\log(1/\delta)\sqrt{d\log(1/\delta)}}{n\varepsilon}\right). \tag{128}
$$

(2) In the heavy tail region, setting $c_1 = c = 4^{\theta}2K\log^{\theta}(\sqrt{T})$, we have

$$
L_S(\mathbf{w}_{T+1}) - L_S(\mathbf{w}_S) \leq \mathbb{O}\left(\frac{\log^{2\theta}(\sqrt{T})\log(1/\delta)\sqrt{d\log(1/\delta)}}{n\varepsilon}\right). \tag{129}
$$

In the second case of $\|\nabla L_S(\mathbf{w}_i)\|_2 > c/2$, we know that,

$$
\begin{aligned}
L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t) &\leq \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \nabla L_S(\mathbf{w}_t)\rangle + \frac{1}{2}\beta\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&\leq -\eta_t\langle \frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2}s_t^+ + \mathbf{g}_t s_t^-, \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle + \frac{1}{2}\beta\eta_t^2\|\overline{\mathbf{g}}_t + \zeta_t\|_2^2.
\end{aligned} \tag{130}
$$

Similarly, referring to Lemma B.3, B.7 and B.8, and based on the truncation corollaries applied in formulas.(72-85), we replace Assumption 2.3 with the upper bound of $\|\nabla L_S(\mathbf{w}_t)\|_2$ obtained using the following conclusion. According to Assumption 2.2, there exist $b' > 0$ and $\mu_S = \frac{3}{c\eta_t(t+t_0)}$, such that

$$
K_{t-1} = (t+t_0)(t+t_0-1)\eta_t K\|\nabla L_S(\mathbf{w}_t)\|_2 = 3(t+t_0-1)c^{-1}\mu_S^{-1}\|\nabla L_S(\mathbf{w}_t)\|_2,
$$

and

$$
\|\nabla L_S(\mathbf{w}_t)\|_2 \leq (\beta\|\mathbf{w}_t\|_2 + \|\nabla L_S(0)\|_2) \leq (\beta\|\mathbf{w}_t\|_2 + b').
$$

From formulas.(118) and (119), since $t_0 \geq \frac{6\beta}{c\mu_S}$ and $\eta_t = \frac{3}{c\mu_S(t+t_0)}$, by setting $T = \mathbb{O}\left(\frac{n\varepsilon}{\sqrt{d\log(1/\delta)}}\right)$, we have

$$
\text{light body: } \beta\|\mathbf{w}_t\|_2 + b' \leq \mathbb{O}\left(\log(1/\delta)\log(T)\right), \tag{131}
$$

$$
\text{heavy tail: } \beta\|\mathbf{w}_t\|_2 + b' \leq \mathbb{O}\left(\log^{(\theta+\frac{1}{2})}(1/\delta)\log(T)\right), \tag{132}
$$

where the inequalities hold true due to $\sum_{t=1}^{T}\frac{1}{t+t_0} \leq \log(T+1)$.

Thus, according to formulas.(84-87) and Assumption B.1, by multiplying both sides by $(t+t_0)(t+t_0-1)$, substituting G with the upper bound of K_{t-1} with $(\beta\|\mathbf{w}_t\|_2 + b')$ and covering all iterations, we focus on the privacy parameters and T, and with probability $1 - 6\delta - T\delta$ can obtain a series of new bounds:

(1) In the light body region,

$$
\begin{aligned}
&(T+t_0)(T+t_0-1)(L_S(\mathbf{w}_{T+1})-L_S(\mathbf{w}_S))\\
&+\sum_{t=1}^{T}\frac{3(t+t_0-1)}{c\mu_S}(\frac{c}{6}-\frac{16}{3}aK\log^{\frac{1}{2}}(2/\delta)-4\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta))\|\nabla L_S(\mathbf{w}_t)\|_2\\
&\leq t_0(t_0-1)(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S))+\sum_{t=1}^{T}\frac{\sqrt{2a}(t+t_0-1)}{c\mu_S}\mathbb{O}\left(\log^{\frac{1}{2}}(1/\delta)\log(1/\delta)\log(T)\right)\\
&+\sum_{t=1}^{T}\frac{9}{c^2\mu_S^2}\left(\mathbb{O}\left(\frac{dTc^2\log^2(1/\delta)}{n^2\varepsilon^2}\right)+\mathbb{O}\left(\frac{\sqrt{dT}c^2\log(1/\delta)}{n\varepsilon}\right)+\mathbb{O}\left(c^2\right)\right),
\end{aligned}
\tag{133}
$$

(2) and in the heavy tail region,

$$
\begin{aligned}
&(T+t_0)(T+t_0-1)(L_S(\mathbf{w}_{T+1})-L_S(\mathbf{w}_S))\\
&+\sum_{t=1}^{T}\frac{3(t+t_0-1)}{c\mu_S}(\frac{c}{6}-\frac{16}{3}aK\log^{\theta}(2/\delta)-4\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta))\|\nabla L_S(\mathbf{w}_t)\|_2\\
&\leq t_0(t_0-1)(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S))+\sum_{t=1}^{T}\frac{\sqrt{2a}(t+t_0-1)}{c\mu_S}\mathbb{O}\left(\log^{\theta}(1/\delta)\log^{\theta+\frac{1}{2}}(1/\delta)\log(T)\right)\\
&+\sum_{t=1}^{T}\frac{9}{c^2\mu_S^2}\left(\mathbb{O}\left(\frac{dTc^2\log^2(1/\delta)}{n^2\varepsilon^2}\right)+\mathbb{O}\left(\frac{\sqrt{dT}c^2\log(1/\delta)}{n\varepsilon}\right)+\mathbb{O}\left(c^2\right)\right),
\end{aligned}
\tag{134}
$$

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 $\theta>1$.

Therefore, if

$$
\frac{c}{6}-\frac{16}{3}aK\log^{\frac{1}{2}}(2/\delta)-4\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta)\geq 0,
$$

and

$$
\frac{c}{6}-\frac{16}{3}aK\log^{\theta}(2/\delta)-4\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta)\geq 0,
$$

respectively, it means that $c_2=c\geq\mathbb{O}\left(\log^{\frac{1}{2}}(1/\delta)\right)$ and $c_1=c\geq\mathbb{O}\left(\log^{\theta}(1/\delta)\right)$.

Then, with probability $1-\delta$, the function value satisfies the situation:

(1) In the light body region,

$$
L_S(\mathbf{w}_{T+1})-L_S(\mathbf{w}_S)\leq\mathbb{O}\left(\log(T)\log(1/\delta)\frac{\sqrt{d\log(T/\delta)}}{n\varepsilon}\right).
\tag{135}
$$

(2) In the heavy tail region,

$$
L_S(\mathbf{w}_{T+1})-L_S(\mathbf{w}_S)\leq\mathbb{O}\left(\log(T)\log^{\theta+\frac{1}{2}}(T/\delta)\frac{\sqrt{d\log(T/\delta)}}{n\varepsilon}\right).
\tag{136}
$$

Finally, to sum up two cases with formulas.(123-124) and (130-131), we achieve

(1) In the light body region,

$$
\begin{aligned}
L_S(\mathbf{w}_{T+1})-L_S(\mathbf{w}_S)&\leq\mathbb{O}\left(\frac{\log(\sqrt{T})\log(1/\delta)\sqrt{d\log(1/\delta)}}{n\varepsilon}\right)+\mathbb{O}\left(\frac{\log(T)\log(T/\delta)\sqrt{d\log(T/\delta)}}{n\varepsilon}\right)\\
&\leq\mathbb{O}\left(\frac{\log(T)\log(T/\delta)\sqrt{d\log(T/\delta)}}{n\varepsilon}\right).
\end{aligned}
\tag{137}
$$

(2) In the heavy tail region,

$$
\begin{aligned}
L_S(\mathbf{w}_{T+1})-L_S(\mathbf{w}_S)&\leq\mathbb{O}\left(\frac{\log^{2\theta}(\sqrt{T})\log(1/\delta)\sqrt{d\log(1/\delta)}}{n\varepsilon}\right)+\mathbb{O}\left(\frac{\log(T)\log^{\theta+\frac{1}{2}}(T/\delta)\sqrt{d\log(T/\delta)}}{n\varepsilon}\right)\\
&\leq\mathbb{O}\left(\frac{\log^{2\theta}(T)\log^{\theta+\frac{1}{2}}(T/\delta)\sqrt{d\log(T/\delta)}}{n\varepsilon}\right).
\end{aligned}
\tag{138}
$$

Assuming that the prior of the tail proportion is p, we can combine the results of subspace identification following the procedure of Corollary 4.3 and, with probability $1-\delta'$, obtain the union bound of

$$L_S(\mathbf{w}_{T+1}) - L_S(\mathbf{w}_S) \leq p * \mathbb{O}\left(\frac{\log^{2\theta}(T)\log^{\theta+\frac{1}{2}}(T/\delta')\sqrt{d\log(T/\delta')}}{n\varepsilon}\right) + (1-p) * \mathbb{O}\left(\frac{\log(T)\log(T/\delta')\sqrt{d\log(T/\delta')}}{n\varepsilon}\right), \tag{139}$$

where the probability δ' contains the broken probability of subspace identification and the convergence of DC-DPSGD under PL condition.

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

$$L_S(\mathbf{w}_{T+1}) - L_S(\mathbf{w}_S) \leq \mathbb{O}\left(\frac{\log(T)\log(T/\delta')\sqrt{d\log(T/\delta')}}{n\varepsilon}\right). \tag{140}$$

The proof of Theorem 4.4 is completed.

H Privacy Guarantee

H.1 Privacy Analysis of Sampling Mechanism

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

$$\sigma_{\text{dp}} \approx \sqrt{\frac{pq_1^2 + (1-p)q_2^2}{\bar{q}^{\,2}}}\,\sigma_{\text{Pois}}. \tag{141}$$

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

Proof. Denote by $\varepsilon_{\text{Pois}}(\alpha, q, \sigma)$ 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

$$\varepsilon_{\text{dp}}(\alpha, \sigma) = \frac{1}{\alpha - 1} \log\left(p\, e^{(\alpha-1)\varepsilon_{\text{Pois}}(\alpha, q_1, \sigma)} + (1-p)\, e^{(\alpha-1)\varepsilon_{\text{Pois}}(\alpha, q_2, \sigma)}\right). \tag{142}$$

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

$$p\,\phi(q_1) + (1-p)\,\phi(q_2) \geq \phi(pq_1 + (1-p)q_2) = \phi(\bar{q}), \tag{143}$$

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

Substituting (143) into (142), we have

$$\varepsilon_{\text{dp}}(\alpha, \sigma) \geq \varepsilon_{\text{Pois}}(\alpha, \bar{q}, \sigma). \tag{144}$$

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

$$\sigma_{\text{dp}} \geq \sigma_{\text{Pois}}, \qquad \text{with equality iff } q_1 = q_2 = \bar{q}. \tag{145}$$

(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

$$\varepsilon_{\text{Pois}}(\alpha, q, \sigma) \approx \frac{\alpha}{2\sigma^2} q^2. \tag{146}$$

Substituting into (142), we obtain

$$\varepsilon_{\text{dp}}(\alpha, \sigma) \approx \frac{\alpha}{2\sigma^2}\left(pq_1^2 + (1-p)q_2^2\right), \quad \varepsilon_{\text{Pois}}(\alpha, \sigma) \approx \frac{\alpha}{2\sigma^2}\bar{q}^{\,2}. \tag{147}$$

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

$$\sigma_{\text{dp}} \approx \sigma_{\text{Pois}} \sqrt{\frac{pq_1^2 + (1-p)q_2^2}{\bar{q}^{\,2}}} \geq \sigma_{\text{Pois}}. \tag{148}$$

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:

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

□

H.2 Privacy Guarantee of DC-DPSGD

Next, we provide the complete privacy guarantee proof of Theorem 4.6 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 $(\varepsilon_{\text{tr}}, \delta_{\text{tr}})$-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 $(\varepsilon_{\text{tr}} + \varepsilon_{\text{dp}}, \delta)$-DP, where $\delta = \delta_{\text{tr}} + \delta_{\text{dp}}$.
- **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 [20] 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 [71].

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 4.6 is proven.

I Supplemental Experiments

I.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 [9] 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 [40], 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 [14]. 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 [52] 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 [61] 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 [28] 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 [9]. 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 [40], 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.

Moreover, we open our source code and simplified version for discriminative clipping on the following link:

Our source code is anonymously available at: https://anonymous.4open.science/r/DC-DPSGD-N-25C9/.

I.2 Train Trajectories

To provide intuitive evidence of the optimization dynamics, we further demonstrate the trajectories of training accuracy in Figure 7, 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.

I.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 6, Table 7, and Table 8, including the subspace-k, the allocation of privacy

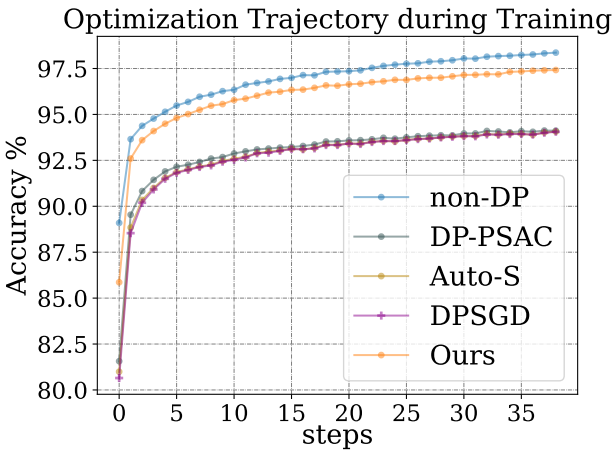

Figure 7: Optimization performance during MNIST Training.

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 8. 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 6: Effects of parameters on test accuracy with MNIST and FMNIST with $\varepsilon = 8$.

Dataset	Subspace-k				ε_{tr} / $\varepsilon_{\text{total}}$			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 7: Effects of parameters on test accuracy with ImageNette and ImageNette-HT with $\varepsilon = 8$.

Dataset	Subspace-k				ε_{tr} / $\varepsilon_{\text{total}}$			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 8: 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

I.4 Evaluation of DP Auditing

We further utilize DP auditing methods to validate the privacy guarantee of DC-DPSGD. DP auditing [29] is a tool to verify whether an algorithm satisfies the claimed privacy budget ε and a failure probability δ, which leverages membership inference attacks (MIAs) [10] to obtain false positive rate (FPR) and false negative rate (FNR), thereby deriving the empirical privacy budget ε^*. As demonstrated by [30], the

privacy region of any (ε,δ)-DP mechanism is constructed by FPR and FNR (also known as Type I α and Type II β errors), which is related to μ_{emp}-Gaussian Differential Privacy (GDP) [17] and is defined as:

$$\mu_{\text{emp}} = \Phi^{-1}(1 - \text{FPR}) - \Phi^{-1}(\text{FNR}), \tag{149}$$

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

$$\delta(\varepsilon^*) = \Phi\left(-\frac{\varepsilon^*}{\mu_{\text{emp}}} + \frac{\mu_{\text{emp}}}{2}\right) - e^{\varepsilon^*}\Phi\left(-\frac{\varepsilon^*}{\mu_{\text{emp}}} - \frac{\mu_{\text{emp}}}{2}\right). \tag{150}$$

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 [45]. The other is an efficient method that needs only one round for auditing by inserting batches of canaries to detect potential vulnerabilities [51], in which we construct the canaries following [45]. Compared to the strongest baseline DPSGD, we evaluate the empirical ε^* of DC-DPSGD and DC-DPSGD-NSI which means we set ε_{tr} = null with non-private subspace identification. Our auditing runs on CIFAR10 with the same implementation in Section I.1.

Table 9 compares our method with standard DPSGD [1] at 95% confidence interval under mainstream privacy auditing paradigms. First, under worst-case auditing, the empirical privacy budget ε^* of DC-DPSGD is lower than the theoretically claimed privacy budget ε, indicating that the algorithm adheres to differential privacy. Second, under one-round auditing, DC-DPSGD exhibits similar privacy detection to DPSGD, suggesting a high level of privacy preservation. The reduced audibility might stem from the insertion of batched canaries, which could dilute the worst-case signal. Furthermore, in the DC-DPSGD-NSI setting, the audited privacy budget exceeds ε, indicating that non-private subspace identification would leak privacy. This outcome validates that our proposed subspace identification method indeed enforces privacy guarantees.

Table 9: Auditing empirical ε^* for DC-DPSGD.

Method	Worst-case Audit [45]		One-round Audit [51]	
	$\varepsilon = 4$	$\varepsilon = 8$	$\varepsilon = 4$	$\varepsilon = 8$
DPSGD	2.07	4.02	1.80	3.72
DC-DPSGD	2.21	4.57	1.80	3.88
DC-DPSGD-NSI	4.13	7.87	3.35	7.23