A 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.

A.1 PROPERTIES

Definition A.1 (Sub-Weibull Equivalent Properties Vladimirova et al. (2020)). *Let X be a random variable and $\theta \geq 0$, and there exists some constant K_1, K_2, K_3, K_4 depending on θ. Then the following characterizations are equivalent:*

1. *The tails of X satisfy*
$$\exists K_1 > 0 \text{ such that } \mathbb{P}(|X| > t) \leq 2\exp(-(t/K_1)^{\frac{1}{\theta}}), \forall t > 0.$$
2. *The moments of X satisfy*
$$\exists K_2 > 0 \text{ such that } \|X\|_p \leq K_2 p^{\theta}, \forall k \geq 1.$$
3. *The moment generating function (MGF) of $|X|^{\frac{1}{\theta}}$ satisfies*
$$\exists K_3 > 0 \text{ such that } \mathbb{E}[\exp((\lambda|X|)^{\frac{1}{\theta}})] \leq \exp((\lambda K_3)^{\frac{1}{\theta}}), \forall \lambda \in (0, 1/K_3).$$
4. *The MGF of $|X|^{\frac{1}{\theta}}$ is bounded at some point,*
$$\exists K_4 > 0 \text{ such that } \mathbb{E}[\exp((|X|/K_4)^{\frac{1}{\theta}})] \leq 2.$$

A.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 A.1. *Let a variable $X \sim subW(\theta, K)$, for any $\delta \in (0, 1)$, then with probability $(1 - \delta)$ we have*
$$|X| \leq K \log^{\theta}(2/\delta).$$

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

Lemma A.2 (Vladimirova et al. (2020); Madden et al. (2020)). *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*
$$\mathbb{P}(|\sum_{i=1}^{n} X_i| \geq x) \leq 2\exp(-(\frac{x}{g(\theta)\sum_{i=1}^{n} K_i})^{\frac{1}{\theta}})$$
where $g(\theta) = (4e)^{\theta}$ for $\theta \leq 1$ and $g(\theta) = 2(2e\theta)^{\theta}$ for $\theta \geq 1$.

Lemma A.3 (Sub-Weibull Freedman Inequality Madden et al. (2020)). *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 $\forall x, \beta, \lambda \geq 0$,

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

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* $\lambda \in [0, \frac{1}{2\alpha}]$,

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

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

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

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* $\lambda \in [0, \frac{1}{2\alpha}]$,

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

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

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

Lemma A.4 (Zhang (2005)). *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 A.5 (Cutkosky & Mehta (2020)). *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 A.6 (Madden et al. (2020)). *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 A.7 (Bakhshizadeh et al. (2023)). *Suppose* $X_1, ..., X_m \stackrel{d}{=} X$ *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 A.8 (Bakhshizadeh et al. (2023)). *Consider the same settings as the ones in Lemma A.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 A.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*

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

A.3 Notations

Table 4: Summary of notations

Definition of Notations	
$\mathbf{w}$	the model parameter
d	the dimension of model parameters
z	the training sample
n	the sample size
B	the batch sample size
ℓ	the loss function
D D'	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_k	k-dimensional the random projection vector
K	the variance-related positive constant
$\nabla L(\mathbf{w}_t)$	k-dimensional the random projection vector
T	the iterations of training
η_t	the learning rate in t iteration
θ	the heavy tail index
p	the ratio of heavy tail
$\lambda_{t,i}^{\text{tr}}$	the empirical trace of the sample
$\hat{\lambda}_t^{\text{tr}}$	the population trace for dividing heavy tail or light body

B Convergence of Heavy-tailed DPSGD

Theorem B.1 (Convergence of Heavy-tailed DPSGD). *Under Assumptions 3.1 and 3.2, let* $\mathbf{w}_t$ *be the iterate produced by Algorithm DPSGD with* $T = \mathbb{O}(\frac{n\epsilon}{\sqrt{d\log(1/\delta)}})$, $T \geq 1$, *and* $\eta_t = \frac{1}{\sqrt{T}}$. *Define* $\hat{\sigma}_{\mathrm{dp}}^2 := m_2 \frac{Tdc^2B^2\log(1/\delta)}{n^2\epsilon^2}$. *If* $\theta = \frac{1}{2}$ *and* $K \leq \hat{\sigma}_{\mathrm{dp}}$, *then* $c = \max\left(4K\log^{\theta}(\sqrt{T}), \frac{19K\log^{\frac{1}{2}}(1/\delta)}{12}\right)$. *If* $\theta = \frac{1}{2}$ *and* $K \geq \hat{\sigma}_{\mathrm{dp}}$, *then* $c = \max\left(4K\log^{\theta}(\sqrt{T}), 39K\log^{\frac{1}{2}}(2/\delta)\right)$. *If* $\theta > \frac{1}{2}$, *then* $c = \max\left(4K\log^{\theta}(\sqrt{T}), 20K\log^{\theta}(2/\delta)\right)$. *For any* $\delta \in (0,1)$, *with probability* $1-\delta$, *we have*

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

where $\hat{\log}(T/\delta) := \log^{\max(0,\theta-1)}(T/\delta)$.

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

We first consider the case $\nabla L_S(\mathbf{w}_t) \leq c/2$.

$$\begin{aligned}
L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t) &\leq \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \nabla L_S(\mathbf{w}_t)\rangle + \frac{1}{2}\beta\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\
&\leq -\eta_t\langle \overline{\mathbf{g}}_t + \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle + \frac{1}{2}\beta\eta_t^2\|\overline{\mathbf{g}}_t + \zeta_t\|^2 \\
&= -\eta_t\langle \overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t] + \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle \\
&\quad - \eta_t\|\nabla L_S(\mathbf{w}_t)\|^2 + \frac{1}{2}\beta\eta_t^2\|\overline{\mathbf{g}}_t\|^2 + \frac{1}{2}\beta\eta_t^2\|\zeta_t\|^2 + \beta\eta_t^2\langle \overline{\mathbf{g}}_t, \zeta_t\rangle \\
&= -\eta_t\langle \overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t], \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle \\
&\quad - \eta_t\|\nabla L_S(\mathbf{w}_t)\|^2 + \frac{1}{2}\beta\eta_t^2\|\overline{\mathbf{g}}_t\|^2 + \frac{1}{2}\beta\eta_t^2\|\zeta_t\|^2 + \beta\eta_t^2\langle \overline{\mathbf{g}}_t, \zeta_t\rangle
\end{aligned} \tag{12}$$

Considering all T iterations, we get

$$\begin{aligned}
\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|^2 &\leq L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S) + \sum_{t=1}^{T}\frac{1}{2}\beta\eta_t^2c^2 + \underbrace{\sum_{t=1}^{T}\frac{1}{2}\beta\eta_t^2\|\zeta_t\|^2}_{\text{Eq.1}} + \underbrace{\sum_{t=1}^{T}\beta\eta_t^2\langle \overline{\mathbf{g}}_t, \zeta_t\rangle}_{\text{Eq.2}} \\
&\quad - \underbrace{\sum_{t=1}^{T}\eta_t\langle \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle}_{\text{Eq.3}} - \underbrace{\sum_{t=1}^{T}\eta_t\langle \overline{\mathbf{g}}_t - \mathbb{E}_t[\overline{\mathbf{g}}_t], \nabla L_S(\mathbf{w}_t)\rangle}_{\text{Eq.4}} - \underbrace{\sum_{t=1}^{T}\eta_t\langle \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle}_{\text{Eq.5}}
\end{aligned} \tag{13}$$

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 A.2, with probability at least $1-\delta$, we have

$$\begin{aligned}
\sum_{t=1}^{T}\frac{1}{2}\beta\eta_t^2\|\zeta_t\|^2 &\leq 2\beta K^2 e\log(2/\delta)\sum_{t=1}^{T}\eta_t^2 \\
&\leq 2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\epsilon^2}\sum_{t=1}^{T}\eta_t^2.
\end{aligned} \tag{14}$$

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

$$\begin{aligned}\sum_{t=1}^{T}\beta\eta_t^2\langle\overline{\mathbf{g}}_t,\zeta_t\rangle &\leq \sum_{t=1}^{T}\beta\eta_t^2\|\overline{\mathbf{g}}_t\|\|\zeta_t\| \\ &\leq \sum_{t=1}^{T}2\beta cK\sqrt{e}\log^{\frac{1}{2}}(2/\delta)\eta_t^2 \\ &\leq 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\epsilon}\sum_{t=1}^{T}\eta_t^2. \end{aligned}\tag{15}$$

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

$$\begin{aligned}-\sum_{t=1}^{T}\eta_t\langle\zeta_t,\nabla L_S(\mathbf{w}_t)\rangle &\leq \sum_{t=1}^{T}\eta_t\|\zeta_t\|\|\nabla L_S(\mathbf{w}_t)\| \\ &\leq \sum_{t=1}^{T}2cK\sqrt{e}\log^{\frac{1}{2}}(2/\delta)\eta_t \\ &\leq 2\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\epsilon}\sum_{t=1}^{T}\eta_t. \end{aligned}\tag{16}$$

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 A.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

$$|\xi_t|\leq\eta_t(\|\overline{\mathbf{g}}_t\|_2+\|\mathbb{E}_t[\overline{\mathbf{g}}_t]\|_2)\|\nabla L_S(\mathbf{w}_t)\|_2\leq\eta_tc^2.\tag{17}$$

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

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}_t[(\xi_t-\mathbb{E}_t\xi_t)^2] &\leq \sum_{t=1}^{T}\eta_t^2\mathbb{E}_t[\|\overline{\mathbf{g}}_t-\mathbb{E}_t[\overline{\mathbf{g}}_t]\|_2^2\|\nabla L_S(\mathbf{w}_t)\|_2^2] \\ &\leq 4c^2\sum_{t=1}^{T}\eta_t^2\|\nabla L_S(\mathbf{w}_t)\|_2^2. \end{aligned}\tag{18}$$

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

$$\sum_{t=1}^{T}\xi_t\leq\frac{\rho4c^2\sum_{t=1}^{T}\eta_t^2\|\nabla L_S(\mathbf{w}_t)\|_2^2}{\eta_tc^2}+\frac{\eta_tc^2\log(1/\delta)}{\rho}.\tag{19}$$

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

$$\begin{aligned}\|\mathbb{E}_t[\overline{\mathbf{g}}_t]-\nabla L_S(\mathbf{w}_t)\|_2 &= \|\mathbb{E}_t[(\overline{\mathbf{g}}_t-\mathbf{g}_t)a_t]\|_2 \\ &= \|\mathbb{E}_t[(\mathbf{g}_t(\frac{c}{\|\mathbf{g}_t\|_2}-1)a_t]\|_2 \\ &\leq \mathbb{E}_t[\|(\mathbf{g}_t(\frac{c}{\|\mathbf{g}_t\|_2}-1)a_t\|_2] \\ &\leq \mathbb{E}_t[|\|\mathbf{g}_t\|_2-c|a_t] \\ &\leq \mathbb{E}_t[|\|\mathbf{g}_t\|_2-\|\nabla L_S(\mathbf{w}_t)\|_2|a_t] \\ &\leq \mathbb{E}_t[|\|\mathbf{g}_t-\nabla L_S(\mathbf{w}_t)\|_2|a_t] \\ &\leq \mathbb{E}_t[|\|\mathbf{g}_t-\nabla L_S(\mathbf{w}_t)\|_2|b_t] \\ &\leq \sqrt{\mathbb{E}_t[\|\mathbf{g}_t-\nabla L_S(\mathbf{w}_t)\|_2^2]\mathbb{E}_tb_t^2}. \end{aligned}\tag{20}$$

Applying Lemma A.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

$$\mathbb{E}_t b_t^2 = \mathbb{P}(\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > \frac{c}{2}) \leq 2\exp(-(\frac{c}{4K})^{\frac{1}{\theta}}) \tag{21}$$

So, we get formula.(20) as

$$\sqrt{\mathbb{E}_t[\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2^2]\mathbb{E}_t b_t^2} \leq 2\sqrt{K^2\Gamma(2\theta+1)\exp(-(\frac{c}{4K})^{\frac{1}{\theta}})}. \tag{22}$$

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

$$\begin{aligned}
&\sum_{t=1}^{T}\eta_t\langle\mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle \\
&\leq 2K^2\Gamma(2\theta+1)\sum_{t=1}^{T}\eta_t\exp(-(\frac{c}{4K})^{\frac{1}{\theta}}) + \frac{1}{2}\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2.
\end{aligned} \tag{23}$$

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

$$\begin{aligned}
&\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S) + \sum_{t=1}^{T}\frac{1}{2}\beta\eta_t^2c^2 + 2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\epsilon^2}\sum_{t=1}^{T}\eta_t^2 \\
&+ 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\epsilon}\sum_{t=1}^{T}\eta_t^2 + 2\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\epsilon}\sum_{t=1}^{T}\eta_t + \frac{\eta_t c^2\log(1/\delta)}{\rho} \\
&+ \frac{4\rho c^2\sum_{t=1}^{T}\eta_t^2\|\nabla L_S(\mathbf{w}_t)\|_2^2}{\eta_t c^2} + 2K^2\Gamma(2\theta+1)\exp(-(\frac{c}{4K})^{\frac{1}{\theta}})\sum_{t=1}^{T}\eta_t + \frac{1}{2}\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2.
\end{aligned} \tag{24}$$

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

$$\begin{aligned}
&\frac{1}{4}\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S) + \frac{1}{2}\beta c^2 + 2\beta m_2 e\frac{d^{\frac{1}{2}}c^2B^2\log^{\frac{3}{2}}(2/\delta)}{n\epsilon} \\
&+ 2\beta\sqrt{em_2}\frac{d^{\frac{1}{4}}c^2B\log^{\frac{1}{2}}(2/\delta)}{\sqrt{n\epsilon}} + 2\sqrt{em_2}c^2B\log^{\frac{1}{2}}(2/\delta) + \frac{16d^{\frac{1}{4}}c^2\log^{\frac{5}{4}}(1/\delta)}{\sqrt{n\epsilon}} \\
&+ \underbrace{2K^2\Gamma(2\theta+1)\exp(-(\frac{c}{4K})^{\frac{1}{\theta}})\sqrt{T}}_{\text{Eq.6}}.
\end{aligned} \tag{25}$$

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 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

$$\begin{aligned}
&\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq 4(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) + 2\beta c^2 + 8\beta m_2 e\frac{d^{\frac{1}{2}}c^2B^2\log^{\frac{3}{2}}(2/\delta)}{n\epsilon} \\
&+ 8\beta\sqrt{em_2}\frac{d^{\frac{1}{4}}c^2B\log^{\frac{1}{2}}(2/\delta)}{\sqrt{n\epsilon}} + 8\sqrt{em_2}c^2B\log^{\frac{1}{2}}(2/\delta) + \frac{64d^{\frac{1}{4}}c^2\log^{\frac{5}{4}}(1/\delta)}{\sqrt{n\epsilon}} + 8K^2\Gamma(2\theta+1).
\end{aligned} \tag{26}$$

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

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq \frac{1}{\sqrt{T}}\Bigg(4(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S))+2\beta c^2+8\beta m_2 e\frac{d^{\frac{1}{2}}c^2B^2\log^{\frac{3}{2}}(2/\delta)}{n\epsilon}$$
$$+8\beta\sqrt{em_2}\frac{d^{\frac{1}{4}}c^2B\log^{\frac{1}{2}}(2/\delta)}{\sqrt{n\epsilon}}+8\sqrt{em_2}c^2B\log^{\frac{1}{2}}(2/\delta)+\frac{64d^{\frac{1}{4}}c^2\log^{\frac{5}{4}}(1/\delta)}{\sqrt{n\epsilon}}+8K^2\Gamma(2\theta+1)\Bigg). \tag{27}$$

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

$$\begin{aligned}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 &\leq \frac{4(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S))}{\sqrt{T}}+\frac{8K^2\Gamma(2\theta+1)}{\sqrt{T}}\\ &+\frac{16K^2\log^{2\theta}(\sqrt{T})\log(2/\delta)}{\sqrt{T}}\Bigg(2\beta+8\beta m_2 e\frac{d^{\frac{1}{2}}B^2\log^{\frac{1}{2}}(2/\delta)}{n\epsilon}\\ &+8\beta\sqrt{em_2}\frac{d^{\frac{1}{4}}B\log^{-\frac{1}{2}}(2/\delta)}{\sqrt{n\epsilon}}+8\sqrt{em_2}B\log^{-\frac{1}{2}}(2/\delta)+\frac{64d^{\frac{1}{4}}\log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\epsilon}}\Bigg)\\ &\leq \mathbb{O}(\frac{\log^{2\theta}(\sqrt{T})\log(1/\delta)}{\sqrt{T}}\cdot\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\epsilon}})\\ &\leq \mathbb{O}(\frac{\log^{2\theta}(\sqrt{T})\log(1/\delta)d^{\frac{1}{4}}\log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\epsilon}}).\end{aligned} \tag{28}$$

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

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{2\theta}(\sqrt{T})\log^{\frac{5}{4}}(1/\delta)}{(n\epsilon)^{\frac{1}{2}}}), \tag{29}$$

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

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

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{2\theta}(\sqrt{T})\log^{\frac{5}{4}}(T/\delta)}{(n\epsilon)^{\frac{1}{2}}}). \tag{30}$$

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

$$\begin{aligned}L_S(\mathbf{w}_{t+1})-L_S(\mathbf{w}_t) &\leq \langle \mathbf{w}_{t+1}-\mathbf{w}_t,\nabla L_S(\mathbf{w}_t)\rangle+\frac{1}{2}\beta\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2\\ &\leq \underbrace{-\eta_t\langle\overline{\mathbf{g}}_t+\zeta_t,\nabla L_S(\mathbf{w}_t)\rangle}_{\text{Eq.7}}+\underbrace{\frac{1}{2}\beta\eta_t^2\|\overline{\mathbf{g}}_t+\zeta_t\|_2^2}_{\text{Eq.8}}\end{aligned} \tag{31}$$

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

$$\begin{aligned}&-\eta_t\langle\overline{\mathbf{g}}_t+\zeta_t,\nabla L_S(\mathbf{w}_t)\rangle\\ &=-\eta_t\langle\frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2}s_t^++\mathbf{g}_ts_t^-,\nabla L_S(\mathbf{w}_t)\rangle-\eta_t\langle\zeta_t,\nabla L_S(\mathbf{w}_t)\rangle.\end{aligned} \tag{32}$$

Applying Lemma A.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

$$\begin{aligned}-\eta_t\langle\frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2}s_t^+,\nabla L_S(\mathbf{w}_t)\rangle &\leq -\frac{c\eta_t s_t^+\|\nabla L_S(\mathbf{w}_t)\|_2}{3}+\frac{8c\eta_t\|\mathbf{g}_t-\nabla L_S(\mathbf{w}_t)\|_2}{3}\\ &\leq -\frac{c\eta_t(1-s_t^-)\|\nabla L_S(\mathbf{w}_t)\|_2}{3}+\frac{8c\eta_t\|\mathbf{g}_t-\nabla L_S(\mathbf{w}_t)\|_2}{3}.\end{aligned} \tag{33}$$

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

$$\begin{aligned}-\eta_t\langle \mathbf{g}_t s_t^-, \nabla L_S(\mathbf{w}_t)\rangle &= -\eta_t s_t^-(\langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle + \|\nabla L_S(\mathbf{w}_t)\|_2^2) \\ &\leq -\eta_t s_t^-(-\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2\|\nabla L_S(\mathbf{w}_t)\|_2 + \|\nabla L_S(\mathbf{w}_t)\|_2^2) \\ &\leq \eta_t\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2\|\nabla L_S(\mathbf{w}_t)\|_2 - \frac{c}{2}\eta_t s_t^-\|\nabla L_S(\mathbf{w}_t)\|_2 \\ &\leq \eta_t\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2\|\nabla L_S(\mathbf{w}_t)\|_2 - \frac{c}{3}\eta_t s_t^-\|\nabla L_S(\mathbf{w}_t)\|_2. \end{aligned} \tag{34}$$

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

$$\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 \leq K\log^\theta(2/\delta), \tag{35}$$

then we get

$$-\eta_t\langle \mathbf{g}_t s_t^-, \nabla L_S(\mathbf{w}_t)\rangle \leq K\log^\theta(2/\delta)\|\nabla L_S(\mathbf{w}_t)\|_2 - \frac{c}{3}\eta_t s_t^-\|\nabla L_S(\mathbf{w}_t)\|_2, \tag{36}$$

and

$$-\eta_t\langle \frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2} s_t^+, \nabla L_S(\mathbf{w}_t)\rangle \leq -\frac{c\eta_t(1-s_t^-)\|\nabla L_S(\mathbf{w}_t)\|_2}{3} + \frac{8c\eta_t K\log^\theta(2/\delta)}{3}. \tag{37}$$

Using Lemma A.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

$$-\sum_{t=1}^T \eta_t\langle \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle \leq 4\sqrt{em_2Td}\frac{cB\log(2/\delta)}{n\epsilon}\sum_{t=1}^T \eta_t\|\nabla L_S(\mathbf{w}_t)\|_2. \tag{38}$$

So, combining formula.(35), formula.(37) and formula.(38) with term Eq.7, with probability at least $1-2\delta-T\delta$, we obtain

$$\begin{aligned}&-\sum_{t=1}^T \eta_t\langle \overline{\mathbf{g}}_t + \zeta_t, \nabla L_S(\mathbf{w}_t)\rangle \leq -\sum_{t=1}^T \frac{c\eta_t}{3}\|\nabla L_S(\mathbf{w}_t)\|_2 + \sum_{t=1}^T \frac{8c\eta_t K\log^\theta(2/\delta)}{3} \\ &+ K\log^\theta(2/\delta)\sum_{t=1}^T \eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 + 4\sqrt{em_2Td}\frac{cB\log(2/\delta)}{n\epsilon}\sum_{t=1}^T \eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 \\ &\leq -\sum_{t=1}^T \frac{c\eta_t}{3}\|\nabla L_S(\mathbf{w}_t)\|_2 + (\frac{19}{3}K\log^\theta(2/\delta) + 4\sqrt{em_2Td}\frac{cB\log(2/\delta)}{n\epsilon})\sum_{t=1}^T \eta_t\|\nabla L_S(\mathbf{w}_t)\|_2. \end{aligned} \tag{39}$$

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

$$\begin{aligned}&(\frac{c}{3} - \frac{19}{3}K\log^\theta(2/\delta) - 4\sqrt{e}\hat{\sigma}_{\text{dp}}\log^{\frac{1}{2}}(1/\delta))\sum_{t=1}^T \eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 \leq L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S) \\ &+ (2\beta m_2ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\epsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\epsilon} + \frac{1}{2}\beta c^2)\sum_{t=1}^T \eta_t^2. \end{aligned} \tag{40}$$

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

$$\begin{aligned}
\sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2 &\leq \frac{3}{K \log^{\frac{1}{2}}(2/\delta)} (L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) \\
&+ \frac{3 \sum_{t=1}^{T} \eta_t^2}{K \log^{\frac{1}{2}}(2/\delta)} \left(2\beta m_2 e d \frac{T c^2 B^2 \log^2(2/\delta)}{n^2 \epsilon^2} + 2\beta \sqrt{e m_2 T d} \frac{c^2 B \log(2/\delta)}{n\epsilon} + \frac{1}{2} \beta c^2 \right) \\
&\leq \frac{L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S) + 2\beta e \hat{\sigma}_{\text{dp}}^2 \log(2/\delta) + 2\beta c \sqrt{e} \hat{\sigma}_{\text{dp}} \log^{\frac{1}{2}}(2/\delta) + \frac{39^2}{2} \beta K^2 \log(2/\delta)}{\frac{1}{3} K \log^{\frac{1}{2}}(2/\delta)} \\
&\leq \frac{3(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S))}{K \log^{\frac{1}{2}}(2/\delta)} + 6\beta e K \log^{\frac{1}{2}}(2/\delta) + 6\beta \sqrt{e} \log^{\frac{1}{2}}(2/\delta) + 3\beta \frac{(39)^2}{2} K \log^{\frac{1}{2}}(2/\delta).
\end{aligned} \tag{41}$$

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

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

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

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla L_S(\mathbf{w}_t)\|_2 \leq \mathbb{O}(\frac{d^{\frac{1}{4}} \log^{\frac{3}{4}}(T/\delta)}{\sqrt{n\epsilon}}). \tag{42}$$

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 $T = \mathbb{O}(\frac{n\epsilon}{\sqrt{d \log(1/\delta)}})$, $T \geq 1$ and $\eta_t = \frac{1}{\sqrt{T}}$ that we obtain

$$\begin{aligned}
\sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2 &\leq \frac{1}{\sqrt{e} \hat{\sigma}_{\text{dp}} \log^{\frac{1}{2}}(1/\delta)} (L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) \\
&+ \frac{\sum_{t=1}^{T} \eta_t^2}{\sqrt{e} \hat{\sigma}_{\text{dp}} \log^{\frac{1}{2}}(1/\delta)} \left(2\beta m_2 e d \frac{T c^2 B^2 \log^2(2/\delta)}{n^2 \epsilon^2} + 2\beta \sqrt{e m_2 T d} \frac{c^2 B \log(2/\delta)}{n\epsilon} + \frac{1}{2} \beta c^2 \right) \\
&\leq \frac{1}{\sqrt{e} \hat{\sigma}_{\text{dp}} \log^{\frac{1}{2}}(1/\delta)} (L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) \\
&+ \frac{\sum_{t=1}^{T} \eta_t^2}{\sqrt{e} \hat{\sigma}_{\text{dp}} \log^{\frac{1}{2}}(1/\delta)} \left(2\beta e \hat{\sigma}_{\text{dp}}^2 \log(2/\delta) + 2\beta \sqrt{e} \hat{\sigma}_{\text{dp}} \log^{\frac{1}{2}}(2/\delta) + \frac{27^2}{2} \beta e \hat{\sigma}_{\text{dp}}^2 \log(2/\delta) \right) \\
&\leq \frac{L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)}{K \log^{\frac{1}{2}}(2/\delta)} + 2\beta e K \log^{\frac{1}{2}}(2/\delta) + 2\beta \sqrt{e} \log^{\frac{1}{2}}(2/\delta) + \beta \frac{(27)^2}{2} K \log^{\frac{1}{2}}(2/\delta).
\end{aligned} \tag{43}$$

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

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

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

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

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

$$\begin{aligned}
&\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \frac{3}{K\log^{\theta}(2/\delta)}(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) \\
&+ \frac{3\sum_{t=1}^{T}\eta_t^2}{K\log^{\theta}(2/\delta)}\left(2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\epsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\epsilon} + \frac{1}{2}\beta c^2\right) \\
&\leq \frac{3(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S))}{K\log^{\theta}(2/\delta)} + \frac{19^2}{24}\beta K\log^{\theta}(2/\delta) + 190\beta K\log^{\theta}(2/\delta) + 3\beta(20)^2K\log^{\theta}(2/\delta).
\end{aligned} \tag{45}$$

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

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

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

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

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 $1-\delta$, $T = \mathbb{O}(\frac{n\epsilon}{\sqrt{d\log(1/\delta)}})$, $T \geq 1$, and $\eta_t = \frac{1}{\sqrt{T}}$

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} &\leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\theta+\frac{1}{4}}(T/\delta)}{(n\epsilon)^{\frac{1}{2}}}) + \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{2\theta}(\sqrt{T})\log^{\frac{5}{4}}(T/\delta)}{(n\epsilon)^{\frac{1}{2}}}) \\
&\leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta)\big(\log^{\theta-1}(T/\delta) + \log^{2\theta}(\sqrt{T})\big)}{(n\epsilon)^{\frac{1}{2}}}) \\
&\leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta)\hat{\log}(T/\delta)\log^{2\theta}(\sqrt{T})}{(n\epsilon)^{\frac{1}{2}}}),
\end{aligned} \tag{48}$$

where $\hat{\log}(T/\delta) = \log^{\max(0,\theta-1)}(T/\delta)$. If $\theta = \frac{1}{2}$ and $K \leq \hat{\sigma}_{\rm dp}$, then $c = \max\left(4K\log^{\theta}(\sqrt{T}), \frac{19K\log^{\frac{1}{2}}(1/\delta)}{12}\right)$. If $\theta = \frac{1}{2}$ and $K \geq \hat{\sigma}_{\rm dp}$, then $c = \max\left(4K\log^{\theta}(\sqrt{T}), 39K\log^{\frac{1}{2}}(2/\delta)\right)$. If $\theta > \frac{1}{2}$, then $c = \max\left(4K\log^{\theta}(\sqrt{T}), 20K\log^{\theta}(2/\delta)\right)$. $\square$

The proof of Theorem 4.1 is completed.

C Privacy Guarantee

We provide the complete privacy guarantee proof of Theorem 5.1 for our differential private mechanism M': Subsample$\circ$TraceSorting (TS)$\circ$GradientPerturbation (GP). The specific proof process is as follows, and our proof comprehensively encompasses mechanism M':

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

(1) Firstly, we show the TS with Gaussian noise here is $(\epsilon_{\text{tr}}, \delta_{\text{tr}})$-DP and follow the proof of Report Noisy Argmax (RNA) in Claim 3.9 Dwork et al. (2014) to clarify that.

Proof. Our trace sorting is to choose traces ranked from 1 to pB. 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 $(\epsilon_{\text{tr}}, \delta_{\text{tr}})$-DP, where m is sample size. Fix the neighboring datasets $D = D' \cup \{a\}$. Let λ, respectively λ', denote the vector of traces when the dataset is D, respectively D'. 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 D and with D'. Fix r_{-i}^{+}, a set from $\text{Gauss}(1/\epsilon_{\text{tr}})^{m-i}$ used for all the noisy traces greater than the i-th trace. Defines r_{-i}^{-}, a set from $\text{Gauss}(1/\epsilon_{\text{tr}})^{i-1}$ 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 $\mathbb{P}[i \mid D, r_{-i}^{-}] \leq e_{\text{tr}}^{\epsilon}\mathbb{P}[i \mid D', r_{-i}^{-}] + \delta_{\text{tr}}$. Define

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

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

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

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 D' and the noise vector is (r_i, r_{-i}^{-}). The probabilities below are over the choice of $r_i \sim \text{Gauss}(1/\epsilon_{\text{tr}})$, then with probability $1 - \delta_{\text{tr}}$:

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

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

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

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

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

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

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

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 D' and the noise vector is (r_i, r^+_{-i}). The probabilities below are over the choice of $r_i \sim \text{Gauss}(1/\epsilon_{\text{tr}})$, with probability $1 - \delta_{\text{tr}}$, and we have:

$$\mathbb{P}[r_i \leq r^* - 1] \geq e^{-\epsilon_{\text{tr}}}\mathbb{P}[r_i \leq r^*] = e^{-\epsilon_{\text{tr}}}\mathbb{P}[i \mid D, r^+_{-i}]$$
$$\Rightarrow \mathbb{P}[i \mid D', r^+_{-i}] \geq \mathbb{P}[r_i \leq r^* - 1] \geq e^{-\epsilon_{\text{tr}}}\mathbb{P}[r_i \leq r^*] = e^{-\epsilon_{\text{tr}}}\mathbb{P}[i \mid D, r^+_{-i}].$$

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

$$\mathbb{P}[i \mid D, r^+_{-i}] \leq e^{\epsilon}_{\text{tr}}\mathbb{P}[i \mid D', r^+_{-i}] + \delta_{\text{tr}}.$$

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

$$e^{\epsilon_{\text{tr}}}(\mathbb{P}[i \mid D', r^+_{-i}] + \mathbb{P}[i \mid D', r^-_{-i}]) + \delta_{\text{tr}} \geq \mathbb{P}[i \mid D, r^+_{-i}] + \mathbb{P}[i \mid D, r^-_{-i}]$$
$$e^{\epsilon_{\text{tr}}}\mathbb{P}[i \mid D', r_{-i}] + \delta_{\text{tr}} \geq \mathbb{P}[i \mid D, r_{-i}],$$

more precisely, we can explicitly bound δ_{tr} to $\mathbb{O}(\frac{1}{pB})$ by refering to Zhu & Wang (2020).

Using the same approach, we can prove that

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

□

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

(2) Secondly, we prove the unified composition of TraceSorting∘GradientPerturbation is $(\epsilon_{\text{tr}} + \epsilon_{\text{dp}}, \delta)$-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 $c = [c_1, c_2, c_2]$.

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

where $r \sim \text{Gauss}(1/\epsilon_{\text{tr}})$ and $\zeta \sim \text{Gauss}(1/\epsilon_{\text{dp}})$. We define $f(\cdot) = \text{GradientDiscent}$ and $\Delta f = \|f(D) - f(D')\|_2 = \frac{1}{B}(pBc_1 + (1-p)Bc_2) = pc_1 + (1-p)c_2$. With $1 - (\delta_{\text{tr}} + \delta_{\text{dp}})$, we have

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

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 $(\epsilon_{\text{tr}} + \epsilon_{\text{dp}}, \delta)$-DP. □

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

Proof. We use $B \subseteq \{1, ..., n\}$ to denote the identities of the B-subsampled samples from $D = \{z_1, \dots, z_n\}$. Note that the randomness of M' includes both the randomness of the random sample B and the random coins of M. Let D_B (or D'_B) be a subsample from D (or D'). Let Y be an arbitrary output range. For convenience, define $q = B/n$.

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

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

by $e^{q(e^{\epsilon_{\text{tr}}+\epsilon_{\text{dp}}}-1)}$. For convenience, define the quantities:

$$\begin{aligned} C &= \mathbb{P}[M(D_B) = Y \mid i \in B] \\ C' &= \mathbb{P}[M(D'_B) = Y \mid i \in B] \\ E &= \mathbb{P}[M(D_B) = Y \mid i \notin B] = \mathbb{P}[M(D'_B) = Y \mid i \notin B] \end{aligned}$$

We can rewrite the ratio as:

$$\frac{\mathbb{P}[M'(D) = Y] - q\delta}{\mathbb{P}[M'(D') = Y]} = \frac{qC + (1-q)E - q\delta}{qC' + (1-q)E}$$

Now we use the fact that, by $(\epsilon_{\text{tr}}+\epsilon_{\text{dp}}, \delta)$-DP, $C \le e^{\epsilon_{\text{tr}}+\epsilon_{\text{dp}}} \min\{C', E\} + \delta$. The rest is a calculation:

$$\begin{aligned} qC + (1-q)E - q\delta &\le q(e^{\epsilon_{\text{tr}}+\epsilon_{\text{dp}}} \min\{C', E\} + \delta) + (1-q)E - q\delta \\ &= q(\min\{C', E\} + (e^{\epsilon_{\text{tr}}+\epsilon_{\text{dp}}} - 1)\min\{C', E\} + \delta) + (1-q)E - q\delta \\ &\le q(\min\{C', E\} + (e^{\epsilon_{\text{tr}}+\epsilon_{\text{dp}}} - 1)\min\{C', E\} + \delta) + (1-q)E - q\delta \\ &\le q(C' + (e^{\epsilon_{\text{tr}}+\epsilon_{\text{dp}}} - 1)(qC' + (1-q)E) + \delta) + (1-q)E - q\delta \\ &\le q(C' + (e^{\epsilon_{\text{tr}}+\epsilon_{\text{dp}}} - 1)(qC' + (1-q)E) + \delta) + (1-q)E \\ &\le (1 + q(e^{\epsilon_{\text{tr}}+\epsilon_{\text{dp}}} - 1))(qC' + (1-q)E). \end{aligned}$$

Thus, we have:

$$\frac{\mathbb{P}[M'(D) = Y] - q\delta}{\mathbb{P}[M'(D') = Y]} \le q(e^{\epsilon_{\text{tr}}+\epsilon_{\text{dp}}} - 1) \cdot \frac{\mathbb{P}[M(D) = Y]}{\mathbb{P}[M(D') = Y]},$$

and we can derive the simpler conclusion $(\mathbb{O}(q\epsilon_{\text{tr}} + q\epsilon_{\text{dp}}), \mathbb{O}(q\delta))$-DP for mechanism M', i.e Subsample∘TraceSorting∘GradientPerturbation is $(\mathbb{O}(q\epsilon_{\text{tr}} + q\epsilon_{\text{dp}}), \mathbb{O}(\delta))$-DP. Furthermore, according to RenyiDP Mironov (2017) and tCDP Bun et al. (2018), we can calculate the corresponding noise multiplier $\sigma_{\text{tr,dp}} = \mathbb{O}(\frac{q\sqrt{T\log(1/\delta)}}{\epsilon})$ with $\epsilon = \epsilon_{\text{tr}}, \epsilon_{\text{dp}}$ for the composition of iterations in model training. □

To sum up, Theorem 5.1 is proven.

D SUBSPACE SKEWING FOR IDENTIFICATION

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

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

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

$$\begin{aligned}
|\lambda_{t,i}^{\text{tr}} - \hat{\lambda}_t^{\text{tr}}| &:= |\text{tr}(V_k^T \hat{\mathbf{g}}_t \hat{\mathbf{g}}_t^T V_k) - \text{tr}(\hat{V}_k^T \hat{\mathbf{g}}_t \hat{\mathbf{g}}_t^T \hat{V}_k)| \\
&= |\|V_k^T \hat{\mathbf{g}}_t\|_2^2 - \|\hat{V}_k^T \hat{\mathbf{g}}_t\|_2^2| \\
&= |\|V_k V_k^T \hat{\mathbf{g}}_t\|_2^2 - \|\hat{V}_k \hat{V}_k^T \hat{\mathbf{g}}_t\|_2^2| \\
&\leq \|V_k V_k^T \hat{\mathbf{g}}_t - \hat{V}_k \hat{V}_k^T \hat{\mathbf{g}}_t\|_2^2 \\
&\leq \|V_k V_k^T - \hat{V}_k \hat{V}_k^T\|_2^2 \|\hat{\mathbf{g}}_t\|_2^2
\end{aligned} \tag{49}$$

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

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

$$\begin{aligned}
\|M\|_2 = \|\frac{1}{k}\sum_{i=1}^k v_i v_i^T\|_2 &\leq \frac{1}{k}\sum_{i=1}^k \|v_i v_i^T\|_2 \\
&= \sup_{x:\|x\|_2=1} \frac{1}{k}\sum_{i=1}^k x^T v_i v_i^T x \\
&= \sup_{x:\|x\|_2=1} \frac{1}{k}\sum_{i=1}^k \langle x, v_i\rangle \\
&\leq \frac{1}{k}\sum_{i=1}^k \|x\|_2 \|v_i\|_2 \\
&= 1
\end{aligned} \tag{50}$$

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

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

$$\mathbb{P}(\|M - \hat{M}\|_2 > \mu) \leq 2d \cdot \exp(\frac{-k\mu^2}{4}), \tag{51}$$

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 $\mu \geq 1$, $\|M - \hat{M}\|_2 > u$ holds with probability 0. So that for any $\mu > 0$, we have

$$\mathbb{P}(\|M - \hat{M}\|_2 > 2\sqrt{\frac{\log 2d}{k}}\mu) \leq \exp(-\mu^2). \tag{52}$$

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

$$\|M - \hat{M}\|_2 \leq 2\frac{\log^{\frac{1}{2}}(2d/\delta_m)}{\sqrt{k}}. \tag{53}$$

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

$$\begin{aligned}
|\lambda_{t,i}^{\text{tr}} - \hat{\lambda}_t^{\text{tr}}| &\leq \|V_k V_k^T - \hat{V}_k \hat{V}_k^T\|_2^2 \|\hat{\mathbf{g}}_t\|_2^2 \\
&\leq \|V_k V_k^T - \hat{V}_k \hat{V}_k^T\|_2^2 \\
&\leq \|M - \hat{M}\|_2^2 \\
&\leq \frac{4\log{(2d/\delta_m)}}{k},
\end{aligned} \tag{54}$$

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 $\lambda_{t,i}^{\text{tr}}$, i.e. $\lambda_{t,i}^{\text{tr}} + \zeta_t^{\text{tr}}$. 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 $\zeta_t^{\text{tr}} \sim \mathbb{N}(0, \sigma_{\text{tr}}^2 \mathbb{I}_1)$. Then, applying Gaussian properties, with probability $1 - \delta_m - \delta_{\text{tr}}$, we have

$$\begin{aligned}
|\lambda_{t,i}^{\text{tr}} - \hat{\lambda}_t^{\text{tr}} + \zeta_t^{\text{tr}}| &\leq |\lambda_{t,i}^{\text{tr}} - \hat{\lambda}_t^{\text{tr}}| + |\zeta_t^{\text{tr}}| \\
&\leq \frac{4\log{(2d/\delta_m)}}{k} + \sigma_{\text{tr}} \log^{\frac{1}{2}}(2/\delta_{\text{tr}}).
\end{aligned} \tag{55}$$

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

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

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

Intuitively, the conclusion tells us that, since $\lambda_{t,i}^{\text{tr}}$ is a constant, the scale $\sigma_{\text{tr}}\mathbb{I}_1$ of noise added is actually small compared to the noise $\sigma_{\text{dp}}\mathbb{I}_d$ 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 $\frac{m_2\sqrt{B}\log^{\frac{1}{2}}(1/\delta_{\text{tr}})}{\sqrt{d}}$, 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 $\delta_m' = \delta_m + \delta_{\text{tr}}$, 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 5.2 is completed.

E Convergence of Discriminative Clipping

Theorem E.1 (Convergence of Discriminative Clipping). *Under Assumptions 3.1, 3.2 and 3.3, let $\mathbf{w}_t$ be the iterate produced by Algorithm Discriminative Clipping DPSGD with $T = \mathbb{O}(\frac{n\epsilon}{\sqrt{d\log(1/\delta)}})$, $T \geq 1$ and $\eta_t = \frac{1}{\sqrt{T}}$. Define $\hat{\log}(T/\delta) = \log^{\max(0,\theta-1)}(T/\delta)$, $\hat{\sigma}_{\text{dp}}^2 = m_2 \frac{Tc^2 dB^2 \log(1/\delta)}{n^2\epsilon^2}$, $a = 2$ 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** ($c = c_1$):*

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

(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** ($c = c_2$):*

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

(1) *If $K \leq \hat{\sigma}_{\text{dp}}$, then $c_2 = \max\left(2\sqrt{2a}K\log^{\frac{1}{2}}(\sqrt{T}), \frac{16aK\log^{\frac{1}{2}}(1/\delta)}{12}\right)$.* (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: $\nabla L_S(\mathbf{w}_t) \leq c/2$ and $\nabla L_S(\mathbf{w}_t) \geq c/2$. To simplify notation, we write ϵ_{dp} as ϵ, omitting the subscript throughout.

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

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

Applying the properties of Gaussian tails and Lemma A.2 to ζ_t, Lemma A.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

$$\begin{aligned}
&\sum_{t=1}^{T} \eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 \leq L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S) + \sum_{t=1}^{T}\frac{1}{2}\beta\eta_t^2 c^2 + 2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\epsilon^2}\sum_{t=1}^{T}\eta_t^2 \\
&+ 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\epsilon}\sum_{t=1}^{T}\eta_t^2 + 2\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\epsilon}\sum_{t=1}^{T}\eta_t + \frac{\eta_t c^2\log(1/\delta)}{\rho} \\
&+ \frac{4\rho c^2\sum_{t=1}^T \eta_t^2\|\nabla L_S(\mathbf{w}_t)\|_2^2}{\eta_t c^2} - \underbrace{\sum_{t=1}^{T}\eta_t\langle \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle}_{\text{Eq.9}}.
\end{aligned} \tag{56}$$

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

$$\begin{aligned}
\|\mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t)\|_2 &= \|\mathbb{E}_t[(\overline{\mathbf{g}}_t - \mathbf{g}_t)a_t]\|_2 \\
&\leq \mathbb{E}_t[\|(\mathbf{g}_t(\frac{c - \|\mathbf{g}_t\|_2}{\|\mathbf{g}_t\|_2})a_t\|_2] \\
&\leq \mathbb{E}_t[|\|\mathbf{g}_t\|_2 - \|\nabla L_S(\mathbf{w}_t)\|_2|a_t] \\
&\leq \mathbb{E}_t[|\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2|b_t] \\
&\leq \sqrt{\mathbb{E}_t[\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2^2]\mathbb{E}_t b_t^2}.
\end{aligned} \tag{57}$$

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

$$\begin{aligned}
\mathbb{P}(\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > x) &\leq \exp(-c_t \eta I(x)) + \exp(-I(x)) \\
&\leq \exp(-\frac{1}{4} I(x)) + \exp(-I(x)) \\
&\leq 2\exp(-\frac{1}{4} I(x)).
\end{aligned} \tag{58}$$

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

$$\begin{aligned}
\mathbb{P}(\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > x) &\leq \exp(-\frac{x^2}{2v(x_{\max}, \eta)}) + m\exp(-\frac{x_{\max}^2(\eta)}{\eta v(x_{\max}, \eta)}) \\
&\leq 2\exp(-\frac{x^2}{2v(x_{\max}, \eta)}) \\
&\leq 2\exp(-\frac{x^2}{2aK^2})
\end{aligned} \tag{59}$$

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 $0 \leq x \leq x_{\max}$, we have

$$\mathbb{E}_t[\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2] \leq \sqrt{2a}K \tag{60}$$

and with probability $1-\delta$,

$$\mathbb{E}_t b_t^2 = \mathbb{P}(\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > \frac{c}{2}) \leq 2\exp(-(\frac{c}{2\sqrt{2a}K})^2) \tag{61}$$

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 $x \geq x_{\max}$, the inequalities

$$\mathbb{E}_t[\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2] \leq 4^\theta K \tag{62}$$

and

$$\mathbb{E}_t b_t^2 = \mathbb{P}(\|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 > \frac{c}{2}) \leq 2\exp(-\frac{1}{4}(\frac{c}{2K})^{\frac{1}{\theta}}) \tag{63}$$

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

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

$$\sum_{t=1}^{T} \eta_t \langle \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle \leq 2aK^2 \sum_{t=1}^{T} \eta_t \exp(-(\frac{c}{2\sqrt{2a}K})^2) + \frac{1}{2}\sum_{t=1}^{T}\eta_t \|\nabla L_S(\mathbf{w}_t)\|_2^2, \tag{64}$$

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

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

$$\sum_{t=1}^{T} \eta_t \langle \mathbb{E}_t[\overline{\mathbf{g}}_t] - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle \leq 4^{2\theta} K^2 \sum_{t=1}^{T} \eta_t \exp(-\frac{1}{4}(\frac{c}{2K})^{\frac{1}{\theta}}) + \frac{1}{2}\sum_{t=1}^{T}\eta_t \|\nabla L_S(\mathbf{w}_t)\|_2^2, \tag{65}$$

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

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

$$\begin{aligned}\frac{1}{4}\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 &\leq L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S) + \frac{1}{2}\beta c^2 + 2\beta m_2 e \frac{d^{\frac{1}{2}} c^2 B^2 \log^{\frac{3}{2}}(2/\delta)}{n\epsilon} \\ &+ 2\beta\sqrt{em_2}\frac{d^{\frac{1}{4}} c^2 B \log^{\frac{1}{2}}(2/\delta)}{\sqrt{n\epsilon}} + 2\sqrt{em_2}c^2 B \log^{\frac{1}{2}}(2/\delta) + \frac{16 d^{\frac{1}{4}} c^2 \log^{\frac{5}{4}}(1/\delta)}{\sqrt{n\epsilon}} \\ &+ \text{Eq.10} \begin{cases} 2aK^2 \sum_{t=1}^{T} \eta_t \exp(-(\frac{c}{2\sqrt{2a}K})^2), & \text{if } 0 \leq x \leq x_{\max}, \\ 4^{2\theta} K^2 \sum_{t=1}^{T} \eta_t \exp(-\frac{1}{4}(\frac{c}{2K})^{\frac{1}{\theta}}), & \text{if } x \geq x_{\max}. \end{cases}\end{aligned} \tag{66}$$

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

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

$$\begin{aligned}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 &\leq \frac{4(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S))}{\sqrt{T}} + \frac{2aK^2}{\sqrt{T}} \\ &+ \frac{8aK^2\log(\sqrt{T})\log(2/\delta)}{\sqrt{T}}\Bigg(2\beta + 8\beta m_2 eB^2(\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(2/\delta)}{\sqrt{n\epsilon}})^2 \\ &+8\beta\sqrt{em_2}\frac{d^{\frac{1}{4}}B\log^{-\frac{1}{2}}(2/\delta)}{\sqrt{n\epsilon}} + 8\sqrt{em_2}B\log^{-\frac{1}{2}}(2/\delta) + \frac{64d^{\frac{1}{4}}\log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\epsilon}}\Bigg) \\ &\leq \mathbb{O}(\frac{\log(\sqrt{T})\log(1/\delta)}{\sqrt{T}}\cdot\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\epsilon}}) \\ &\leq \mathbb{O}(\frac{\log(\sqrt{T})d^{\frac{1}{4}}\log^{\frac{5}{4}}(1/\delta)}{\sqrt{n\epsilon}}). \end{aligned} \tag{67}$$

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

$$\begin{aligned}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2^2 &\leq \frac{4(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S))}{\sqrt{T}} + \frac{2aK^2}{\sqrt{T}} \\ &+ \frac{4^{2\theta+1}\log^{2\theta}(\sqrt{T})\log(2/\delta)}{\sqrt{T}}\Bigg(2\beta + 8\beta m_2 eB^2(\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(2/\delta)}{\sqrt{n\epsilon}})^2 \\ &+8\beta\sqrt{em_2}\frac{d^{\frac{1}{4}}B\log^{-\frac{1}{2}}(2/\delta)}{\sqrt{n\epsilon}} + 8\sqrt{em_2}B\log^{-\frac{1}{2}}(2/\delta) + \frac{64d^{\frac{1}{4}}\log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\epsilon}}\Bigg) \\ &\leq \mathbb{O}(\frac{\log^{2\theta}(\sqrt{T})\log(1/\delta)}{\sqrt{T}}\cdot\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(1/\delta)}{\sqrt{n\epsilon}}) \\ &\leq \mathbb{O}(\frac{\log^{2\theta}(\sqrt{T})d^{\frac{1}{4}}\log^{\frac{5}{4}}(1/\delta)}{\sqrt{n\epsilon}}). \end{aligned} \tag{68}$$

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

$$\begin{aligned}L_S(\mathbf{w}_{t+1}) - L_S(\mathbf{w}_t) &\leq \langle \mathbf{w}_{t+1}-\mathbf{w}_t, \nabla L_S(\mathbf{w}_t)\rangle + \frac{1}{2}\beta\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2 \\ &\leq \underbrace{-\eta_t\langle\overline{\mathbf{g}}_t+\zeta_t,\nabla L_S(\mathbf{w}_t)\rangle}_{\text{Eq.11}} + \frac{1}{2}\beta\eta_t^2\|\overline{\mathbf{g}}_t+\zeta_t\|_2^2. \end{aligned} \tag{69}$$

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

$$-\eta_t\langle\overline{\mathbf{g}}_t+\zeta_t,\nabla L_S(\mathbf{w}_t)\rangle = -\eta_t\langle\frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2}s_t^+ + \mathbf{g}_t s_t^-, \nabla L_S(\mathbf{w}_t)\rangle - \eta_t\langle\zeta_t,\nabla L_S(\mathbf{w}_t)\rangle. \tag{70}$$

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

$$
\begin{aligned}
-\sum_{t=1}^{T}\eta_t\langle \mathbf{g}_t s_t^-, \nabla L_S(\mathbf{w}_t)\rangle &= -\sum_{t=1}^{T}\eta_t s_t^-(\langle \mathbf{g}_t-\nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle + \|\nabla L_S(\mathbf{w}_t)\|_2^2) \\
&\le -\sum_{t=1}^{T}\eta_t s_t^-\langle \mathbf{g}_t-\nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle - \sum_{t=1}^{T}\eta_t s_t^-\|\nabla L_S(\mathbf{w}_t)\|_2^2 \\
&\le -\sum_{t=1}^{T}\eta_t s_t^-\langle \mathbf{g}_t-\nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle - \frac{c}{2}\sum_{t=1}^{T}\eta_t s_t^-\|\nabla L_S(\mathbf{w}_t)\|_2^2 \\
&\le \underbrace{-\sum_{t=1}^{T}\eta_t s_t^-\langle \mathbf{g}_t-\nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t)\rangle}_{\text{Eq.12}} - \frac{c}{3}\sum_{t=1}^{T}\eta_t s_t^-\|\nabla L_S(\mathbf{w}_t)\|_2^2.
\end{aligned}
\tag{71}
$$

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 A.3 and concentration inequality with Lemma A.7 and A.8 to bound it.

In Lemma A.3, Define

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

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 A.7 and A.8 in Bakhshizadeh et al. (2023) and obtain

$$
\begin{aligned}
\mathbb{P}\left(\bigcup_{k\in\mathbb{N}}\left\{\sum_{i=1}^{k}\xi_i\ge kx \text{ and } \sum_{i=1}^{k}aK_{i-1}^2\le\beta\right\}\right) &\le \exp(-\lambda kx+\frac{\lambda^2}{2}\beta) \\
&= \exp(-\lambda kx + kv(L,\eta)\frac{\lambda^2}{2}).
\end{aligned}
\tag{72}
$$

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 $x\ge x_{\max}$, we choose $L = kx$ and $\lambda = \frac{\eta I(kx)}{kx}$, that is $\frac{x}{v(kx,\eta)}\ge\frac{x_{\max}}{v(kx,\eta)} = \frac{\eta I(kx)}{kx}$. Then the inequality achieves

$$
\begin{aligned}
\mathbb{P}\left(\bigcup_{k\in\mathbb{N}}\left\{\sum_{i=1}^{k}\xi_i\ge kx \text{ and } \sum_{i=1}^{k}aK_{i-1}^2\le\beta\right\}\right) &\le \exp(-\eta I(kx)+v(L,\eta)\frac{\eta^2I^2(kx)}{2kx^2}) \\
&\le \exp(-\eta I(kx)(1-v(L,\eta)\frac{\eta I(kx)}{2kx^2})) \\
&\le \exp(-\eta c_x I(kx)) \\
&\le \exp(-\frac{1}{2}\eta I(kx)),
\end{aligned}
\tag{73}
$$

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

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

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

Implementing the above inferences and propositions with

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

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 $1 - \delta$

$$
\begin{aligned}
-\sum_{t=1}^{T} \eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t) \rangle &\leq \sqrt{2Tv(L,\eta)} \log^{\frac{1}{2}}(1/\delta) \\
&\leq \sqrt{2a \sum_{t=1}^{T} K_t^2 \log^{\frac{1}{2}}(1/\delta)} \\
&\leq 2\sqrt{\sum_{t=1}^{T} \eta_t^2 K^2 \|\nabla L_S(\mathbf{w}_t)\|_2^2 \log^{\frac{1}{2}}(1/\delta)} \\
&\leq 2KG\sqrt{\sum_{t=1}^{T} \eta_t^2 \log^{\frac{1}{2}}(1/\delta)}, \qquad (75)
\end{aligned}
$$

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

$$
\begin{aligned}
-\sum_{t=1}^{T} \eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t) \rangle &\leq 4^{\frac{1}{2}} \frac{1}{T} \sum_{t=1}^{T} K_t \log^{\frac{1}{2}}(1/\delta) \\
&\leq 2\frac{KG}{T} \sum_{t=1}^{T} \eta_t \log^{\frac{1}{2}}(1/\delta). \qquad (76)
\end{aligned}
$$

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

$$
\begin{aligned}
-\sum_{t=1}^{T} \eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t) \rangle &\leq \sqrt{2a \sum_{t=1}^{T} K_t^2 \log^{\frac{1}{2}}(1/\delta)} \\
&\leq \sqrt{2}(4\theta)^{\theta} eKG \sqrt{\sum_{t=1}^{T} \eta_t^2 \log^{\frac{1}{2}}(1/\delta)}, \qquad (77)
\end{aligned}
$$

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

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

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 $1 - 3\delta$

$$
\begin{aligned}
-\sum_{t=1}^{T} \eta_t s_t^- \langle \mathbf{g}_t - \nabla L_S(\mathbf{w}_t), \nabla L_S(\mathbf{w}_t) \rangle &\le \sqrt{2a \sum_{t=1}^{T} K_t^2 \log^{\frac{1}{2}}(1/\delta)} \\
&\le \sqrt{2(2^{2\theta+1} + 2)\Gamma(2\theta + 1) + \frac{2^{3\theta}\Gamma(3\theta + 1)}{3}} KG \sqrt{\sum_{t=1}^{T} \eta_t^2 \log^{\frac{1}{2}}(1/\delta)}, \qquad (79)
\end{aligned}
$$

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

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

To continue the proof, employing Lemma A.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

$$
\begin{aligned}
-\sum_{t=1}^{T} \eta_t \langle \frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2} s_t^+, \nabla L_S(\mathbf{w}_t) \rangle &\le -\frac{c \sum_{t=1}^{T} \eta_t s_t^+ \|\nabla L_S(\mathbf{w}_t)\|_2}{3} + \frac{8c \sum_{t=1}^{T} \eta_t \|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2}{3} \\
&\le -\frac{c \sum_{t=1}^{T} \eta_t (1 - s_t^-) \|\nabla L_S(\mathbf{w}_t)\|_2}{3} \\
&\quad + \frac{16 \sum_{t=1}^{T} \eta_t \|\mathbf{g}_t - \nabla L_S(\mathbf{w}_t)\|_2 \|\nabla L_S(\mathbf{w}_t)\|_2}{3}. \qquad (81)
\end{aligned}
$$

With the truncated corollaries above, we have

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

$$
\begin{aligned}
-\sum_{t=1}^{T} \eta_t \langle \frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2} s_t^+, \nabla L_S(\mathbf{w}_t) \rangle &\le -\frac{c \sum_{t=1}^{T} \eta_t (1 - s_t^-) \|\nabla L_S(\mathbf{w}_t)\|_2}{3} \\
&+ \frac{16 \sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2}{3}
\begin{cases}
2K \log^{\frac{1}{2}}(2/\delta), & \text{if } \theta = \frac{1}{2}, \\
\sqrt{2} e (4\theta)^\theta K \log^{\frac{1}{2}}(2/\delta), & \text{if } \theta \in (\frac{1}{2}, 1], \\
\sqrt{2(2^{2\theta+1} + 2)\Gamma(2\theta + 1) + \frac{2^{3\theta}\Gamma(3\theta + 1)}{3}} K \log^{\frac{1}{2}}(2/\delta) & \text{if } \theta > 1.
\end{cases} \qquad (82)
\end{aligned}
$$

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

$$
\begin{aligned}
-\sum_{t=1}^{T} \eta_t \langle \frac{c\mathbf{g}_t}{\|\mathbf{g}_t\|_2} s_t^+, \nabla L_S(\mathbf{w}_t) \rangle &\le -\frac{c \sum_{t=1}^{T} \eta_t (1 - s_t^-) \|\nabla L_S(\mathbf{w}_t)\|_2}{3} \\
&+ \frac{16 \sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2}{3} 4^\theta K \log^{\theta}(2/\delta). \qquad (83)
\end{aligned}
$$

Then, according to Lemma A.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 $0 \le x \le x_{\max}$, with probability at least $1 - 3\delta - T\delta$

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

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

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

Therefore, we refer to formula.(12) and formula.(13), and apply Lemma A.2 due to $\zeta_t \sim \mathbb{N}(0, c\sigma_{\text{dp}}\mathbb{I}_d)$. Then, to simplify the notation, we define $\hat{\sigma}_{\text{dp}}^2 = dc^2\sigma_{\text{dp}}^2$. With $\hat{\sigma}_{\text{dp}}^2 = m_2\frac{Tc^2dB^2\log(1/\delta)}{n^2\epsilon^2}$ and probability $1 - 6\delta - T\delta$, if $0 \le x \le x_{\max}$, we have

$$
\begin{aligned}
&(\frac{c}{3} - \frac{16}{3}aK\log^{\frac{1}{2}}(2/\delta) - 4\sqrt{e}\hat{\sigma}_{\text{dp}}\log^{\frac{1}{2}}(1/\delta))\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 \le L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S) \\
&+ (2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\epsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\epsilon} + \frac{1}{2}\beta c^2)\sum_{t=1}^{T}\eta_t^2 \\
&+ \sqrt{2}aKG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta),
\end{aligned} \tag{86}
$$

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

$$
\begin{aligned}
&(\frac{c}{3} - \frac{16}{3}aK\log^{\theta}(2/\delta) - 4\sqrt{e}\hat{\sigma}_{\text{dp}}\log^{\frac{1}{2}}(1/\delta))\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 \le L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S) \\
&+ (2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\epsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\epsilon} + \frac{1}{2}\beta c^2)\sum_{t=1}^{T}\eta_t^2 \\
&+ \sqrt{2}aKG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\theta}(1/\delta),
\end{aligned} \tag{87}
$$

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 $0 \leq x \leq x_{\max}$ and $\theta \geq \frac{1}{2}$:

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

$$\begin{aligned}
\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 &\leq \frac{3}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S)) + \frac{3\sqrt{2a}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}\\
&+ \frac{3\sum_{t=1}^{T}\eta_t^2}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}\left(2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\epsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\epsilon} + \frac{1}{2}\beta c^2\right)\\
&\leq \frac{3(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S))}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} + \frac{3\sqrt{2a}KG\log^{\frac{1}{2}}(1/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}\\
&+ \frac{6\beta ea^2K^2\log(2/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} + \frac{6\beta\sqrt{e}\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} + \frac{3\beta(33\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta))^2}{2\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}. \qquad (88)
\end{aligned}$$

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

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

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

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \mathbb{O}(\frac{d^{\frac{3}{4}}\log^{\frac{3}{4}}(T/\delta)}{\sqrt{n\epsilon}}). \qquad (89)$$

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

$$\begin{aligned}
\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 &\leq \frac{1}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta)}(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S)) + \frac{\sqrt{2a}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta)}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta)}\\
&+ \frac{\sum_{t=1}^{T}\eta_t^2}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(1/\delta)}\left(2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\epsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\epsilon} + \frac{1}{2}\beta c^2\right)\\
&\leq \frac{L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S)}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}\log^{\frac{1}{2}}(2/\delta)} + \frac{\sqrt{2a}KG}{\sqrt{e}\hat{\sigma}_{\mathrm{dp}}} + 2\beta eK\log^{\frac{1}{2}}(2/\delta) + 2\beta\sqrt{e}\log^{\frac{1}{2}}(2/\delta) + \beta\frac{(27)^2}{2}K\log^{\frac{1}{2}}(2/\delta).
\end{aligned} \qquad (90)$$

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

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

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

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

2. In case of heavy tail, when $x \geq x_{\max}$:

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

$$
\begin{aligned}
&\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \frac{3}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S)) + \frac{3\sqrt{2a}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} \\
&+ \frac{3\sum_{t=1}^{T}\eta_t^2}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}\left(2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\epsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\epsilon} + \frac{1}{2}\beta c^2\right) \\
&\leq \frac{3(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S))}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} + \frac{3\sqrt{2a}KG\log^{\frac{1}{2}}(1/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} \\
&+ \frac{6\beta ea^2K^2\log(2/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} + \frac{6\beta\sqrt{e}\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}{\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)} + \frac{3\beta(33\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta))^2}{2\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)}.
\end{aligned} \tag{92}
$$

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

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

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

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

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 $T = \mathbb{O}(\frac{n\epsilon}{\sqrt{d\log(1/\delta)}})$, $T \geq 1$ and $\eta_t = \frac{1}{\sqrt{T}}$ that we obtain

$$
\begin{aligned}
&\sum_{t=1}^{T}\eta_t\|\nabla L_S(\mathbf{w}_t)\|_2 \leq \frac{1}{\sqrt{e}\hat{\sigma}_{\text{dp}}\log^{\frac{1}{2}}(1/\delta)}(L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S)) + \frac{\sqrt{2a}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\frac{1}{2}}(1/\delta)}{\sqrt{e}\hat{\sigma}_{\text{dp}}\log^{\frac{1}{2}}(1/\delta)} \\
&+ \frac{\sum_{t=1}^{T}\eta_t^2}{\sqrt{e}\hat{\sigma}_{\text{dp}}\log^{\frac{1}{2}}(1/\delta)}\left(2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\epsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\epsilon} + \frac{1}{2}\beta c^2\right) \\
&\leq \frac{L_S(\mathbf{w}_1)-L_S(\mathbf{w}_S)}{\sqrt{e}\hat{\sigma}_{\text{dp}}\log^{\frac{1}{2}}(2/\delta)} + \frac{\sqrt{2a}KG}{\sqrt{e}\hat{\sigma}_{\text{dp}}} + 2\beta eK\log^{\frac{1}{2}}(2/\delta) + 2\beta\sqrt{e}\log^{\frac{1}{2}}(2/\delta) + \beta\frac{(27)^2}{2}K\log^{\frac{1}{2}}(2/\delta).
\end{aligned} \tag{94}
$$

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

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

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

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

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

$$\begin{aligned}
&\sum_{t=1}^{T} \eta_t \|\nabla L_S(\mathbf{w}_t)\|_2 \leq \frac{3}{\sqrt{2a}K\log^{\theta}(2/\delta)}(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S)) + \frac{3\sqrt{2a}KG\sqrt{\sum_{t=1}^{T}\eta_t^2}\log^{\theta}(1/\delta)}{\sqrt{2a}K\log^{\theta}(2/\delta)} \\
&+ \frac{3\sum_{t=1}^{T}\eta_t^2}{\sqrt{2a}K\log^{\theta}(2/\delta)}\left(2\beta m_2 ed\frac{Tc^2B^2\log^2(2/\delta)}{n^2\epsilon^2} + 2\beta\sqrt{em_2Td}\frac{c^2B\log(2/\delta)}{n\epsilon} + \frac{1}{2}\beta c^2\right) \\
&\leq \frac{3(L_S(\mathbf{w}_1) - L_S(\mathbf{w}_S))}{\sqrt{2a}K\log^{\theta}(2/\delta)} + 3G + \frac{16^2}{24}\beta K\log^{\theta}(2/\delta) + 136\beta K\log^{\theta}(2/\delta) + 3\beta(17)^2 K\log^{\theta}(2/\delta).
\end{aligned} \tag{96}$$

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

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

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

For light body, we have

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

For heavy tail, we have

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

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 $1-\delta$, $T = \mathbb{O}(\frac{n\epsilon}{\sqrt{d\log(1/\delta)}})$, $T \geq 1$ and $\eta_t = \frac{1}{\sqrt{T}}$:

1. In the heavy tail region:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} &\leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\theta+\frac{1}{4}}(T/\delta)}{(n\epsilon)^{\frac{1}{2}}}) + \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{2\theta}(\sqrt{T})\log^{\frac{5}{4}}(T/\delta)}{(n\epsilon)^{\frac{1}{2}}}) \\
&\leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(T/\delta)\big(\log^{\theta}(T/\delta) + \log^{2\theta}(\sqrt{T})\log(T/\delta)\big)}{(n\epsilon)^{\frac{1}{2}}}) \\
&\leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta)\hat{\log}(T/\delta)\log^{2\theta}(\sqrt{T})}{(n\epsilon)^{\frac{1}{2}}}),
\end{aligned} \tag{100}$$

where $\hat{\log}(T/\delta) = \log^{\max(0,\theta-1)}(T/\delta)$. If $\theta = \frac{1}{2}$ and $K \leq \hat{\sigma}_{\mathrm{dp}}$, then $c_1 = \max\big(4^{\theta}2K\log^{\theta}(\sqrt{T}), \frac{16aK\log^{\frac{1}{2}}(1/\delta)}{12}\big)$. If $\theta = \frac{1}{2}$ and $K \geq \hat{\sigma}_{\mathrm{dp}}$, then $c_1 = \max\big(4^{\theta}2K\log^{\theta}(\sqrt{T}), 33\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)\big)$. If $\theta > \frac{1}{2}$, then $c_1 = \max\big(4^{\theta}2K\log^{\theta}(\sqrt{T}), 17K\log^{\theta}(2/\delta)\big)$.

2. In the light body region:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} &\leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{3}{4}}(T/\delta)}{(n\epsilon)^{\frac{1}{2}}}) + \mathbb{O}(\frac{d^{\frac{1}{4}}\log(\sqrt{T})\log^{\frac{5}{4}}(T/\delta)}{(n\epsilon)^{\frac{1}{2}}}) \\
&\leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{1}{4}}(T/\delta)\big(\log^{\frac{1}{2}}(T/\delta) + \log(\sqrt{T})\log(T/\delta)\big)}{(n\epsilon)^{\frac{1}{2}}}) \\
&\leq \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta)\log(\sqrt{T})}{(n\epsilon)^{\frac{1}{2}}}),
\end{aligned} \tag{101}$$

where $\hat{\log}(T/\delta) = \log^{\max(0,\theta-1)}(T/\delta)$. If $\theta = \frac{1}{2}$ and $K \leq \hat{\sigma}_{\mathrm{dp}}$, then $c_2 = \max\left(2\sqrt{2a}K\log^{\theta}(\sqrt{T}), \frac{16aK\log^{\frac{1}{2}}(1/\delta)}{12}\right)$. If $\theta = \frac{1}{2}$ and $K \geq \hat{\sigma}_{\mathrm{dp}}$, then $c_2 = \max\left(2\sqrt{2a}K\log^{\theta}(\sqrt{T}), 33\sqrt{2a}K\log^{\frac{1}{2}}(2/\delta)\right)$. If $\theta > \frac{1}{2}$, then $c_2 = \max\left(2\sqrt{2a}K\log^{\theta}(\sqrt{T}), 17K\log^{\theta}(2/\delta)\right)$.

□

The proof of Theorem 5.3 is completed.

F UNIFORM BOUND FOR DISCRIMINATIVE CLIPPING DPSGD

Theorem F.1 (**Uniform Bound for Discriminative Clipping DPSGD).** *Under Assumptions 3.1, 3.2 and 3.3, combining Theorem 2 and Theorem 3, for any $\delta' \in (0,1)$, with probability $1-\delta'$, we have*

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

where $\delta' = \delta'_m + \delta$, $\hat{\log}(T/\delta) = \log^{\max(0,\theta-1)}(T/\delta)$ and p is ratio of heavy-tailed samples.

Proof. We combine the subspace skewing error (Theorem 5.2) with the optimization bound of Discriminative Clipping DPSGD (Theorem 5.3) 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 $\mathbb{O}(\frac{1}{k})$. This conclusion means that, 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 $p \in [5\%, 10\%]$. Although a subsampling strategy is used, uniform sampling does not change the proportion of tail samples in the batch. Furthermore, based on the relationship between trace and variance, the pB-th of sorted trace $\lambda_t^{\text{tr},p}$ can be seen as the inflection point $x_{\max}$ of distribution defined in truncated theories A.7 and A.8, which corresponds to the empirical sample results with theoretical population variance and the approximation error has bounded in Theorem 5.2. 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

$$\begin{aligned}
&\mathcal{C}_{\text{u}}(c_1, c_2) := \frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} \\
&= (1-\delta'_m) * p * \mathcal{C}_{\text{tail}}(c_1) + (1-\delta'_m) * (1-p) * \mathcal{C}_{\text{body}}(c_2) + \delta'_m * |\mathcal{C}_{\text{tail}}(c_1) - \mathcal{C}_{\text{body}}(c_2)|.
\end{aligned} \tag{102}$$

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 $\lambda_t^{\text{tr},i} \geq \lambda_t^{\text{tr},p}$, i.e. $\mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta)\hat{\log}(1/\delta)\log^{2\theta}(\sqrt{T})}{(n\epsilon)^{\frac{1}{2}}})$, $\mathcal{C}_{\text{body}}(c_2)$ 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 $0 \leq \lambda_t^{\text{tr},i} \leq \lambda_t^{\text{tr},p}$ i.e. $\mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta)\log(\sqrt{T})}{(n\epsilon)^{\frac{1}{2}}})$, 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

$$\mathcal{C}_{\text{u}}(c_1, c_2) = \mathcal{C}_{\text{tail}}(c_1) = \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta)\log(\sqrt{T})}{(n\epsilon)^{\frac{1}{2}}}). \tag{103}$$

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 $\mathcal{C}_{\text{tail}}(c_1) \geq \mathcal{C}_{u}(c_1, c_2)$, i.e.

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

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

$$\mathcal{C}_{\text{tail}}(c_1) \geq \mathcal{C}_{\text{body}}(c_2) - \frac{\delta'_m}{(1-\delta'_m)*(1-p)}\mathcal{C}_{\text{body}}(c_2), \tag{104}$$

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

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

$$\mathcal{C}_{\text{u}}(c_1, c_2) = p * \mathcal{C}_{\text{tail}}(c_1) + *(1-p) * \mathcal{C}_{\text{body}}(c_2). \tag{105}$$

In other words, for the formula.(102), we define $\delta' = \delta'_m + \delta$. Then, with probability $1-\delta'$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla L_S(\mathbf{w}_t)\|_2, \|\nabla L_S(\mathbf{w}_t)\|_2^2\right\} \leq p * \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta)\hat{\log}(T/\delta)\log^{2\theta}(\sqrt{T})}{(n\epsilon)^{\frac{1}{2}}})$$
$$+ (1-p) * \mathbb{O}(\frac{d^{\frac{1}{4}}\log^{\frac{5}{4}}(T/\delta)\log(\sqrt{T})}{(n\epsilon)^{\frac{1}{2}}}) \tag{106}$$

where $\hat{\log}(T/\delta) = \log^{\max(0,\theta-1)}(T/\delta)$.

□

The proof of Theorem 5.4 is completed.

G Supplemental Experiments

G.1 Implementation Details and Codebase

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 $k = 200$, $\epsilon = \epsilon_{\text{tr}} + \epsilon_{\text{dp}}$ with $\epsilon_{\text{tr}} = \epsilon_{\text{dp}}$, $p = 10\%$ and sub-Weibull index $\theta = 2$ for any datasets. In particular, we use the LDAM Cao et al. (2019) loss function for heavy-tailed tasks.

1. **MNIST**: MNIST has ten categories, 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.
2. **FMNIST**: FMNIST has ten categories, 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.
3. **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 Tramer & Boneh (2021) pre-trained by unlabeled ImageNet. We refer the code for pre-trained SimCLRv2 to `https://github.com/ftramer/Handcrafted-DP`.
4. **CIFAR10-HT**: CIFAR10-HT contains 32×32 pixel 12,406 training data and 10,000 testing data, and the proportion of 10 classes in training data 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 Xie et al. (2017) 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`.
5. **ImageNette**: ImageNette is a 10-subclass set of ImageNet and contains 9469 training examples and 3925 testing examples. We train on model ResNet-9 He et al. (2016) 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.
6. **ImageNette-HT**: We construct the heavy-tailed version of ImageNette by the method in Cao et al. (2019). ImageNette-HT contains 2345 trainging data and 3925 testing data, 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.
7. **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 **?**, where small clipping threshold $c_2 = 0.1$ and large clipping threshold $c_1 = 10 * c_2$.

Moreover, we open our source code and implementation details for discriminative clipping on the following link: `https://anonymous.4open.science/r/DC-DPSGD-N-25C9/`.

G.2 Effects of Parameters on Test Accuracy

Due to space limitations, we place the remaining ablation study on MNIST, FMNIST, ImageNette and ImageNette-HT in Table 5 and Table 6. 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 5: Effects of parameters on test accuracy with MNIST and FMNIST.

Dataset	Subspace-k				$\epsilon_{\rm tr}$ + $\epsilon_{\rm dp}$			sub-Weibull-θ		
	None	100	150	200	2+6	4+4	6+2	1/2	1	2
MNIST	98.16	98.48	98.66	98.72	98.78	98.72	98.42	98.61	98.69	98.72
FMNIST	85.78	87.61	87.71	87.80	87.70	87.80	87.26	87.40	87.55	87.80

Table 6: Effects of parameters on test accuracy with ImageNette and ImageNette-HT.

Dataset	Subspace-k				$\epsilon_{\rm tr}$ + $\epsilon_{\rm dp}$			sub-Weibull-θ		
	None	100	150	200	2+6	4+4	6+2	1/2	1	2
ImageNette	66.08	68.34	69.00	69.29	68.54	69.29	68.12	67.91	68.87	69.29
ImageNette-HT	29.33	31.44	33.17	33.70	34.25	33.70	31.13	33.05	33.37	33.70

To investigate the effect of p, we have added a set of new experiments by varying $p \in [1\%, 20\%]$. The results are presented in Table 7. We observe that the test accuracy is minimally affected when p is less than 10%, but shows a negative impact at around 20%. 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.

Table 7: Effects of parameter on p.

Dataset	Heavy tail ratio-p				
	20%	10%	5%	2%	1%
ImageNette	66.82	69.29	68.44	68.45	68.75