Supplementary Material for

"Differentially Private Regularized Stochastic Convex Optimization with Heavy-Tailed Data"

Haihan Xie[1] Matthew Pietrosanu[1] Yi Liu[1]

Wei Tu[2] Bei Jiang[1] Linglong Kong[1]

[1] *Department of Mathematical and Statistical Sciences, University of Alberta*

[2] *Department of Public Health Sciences, Queen's University*

S1 Additional Numerical Results

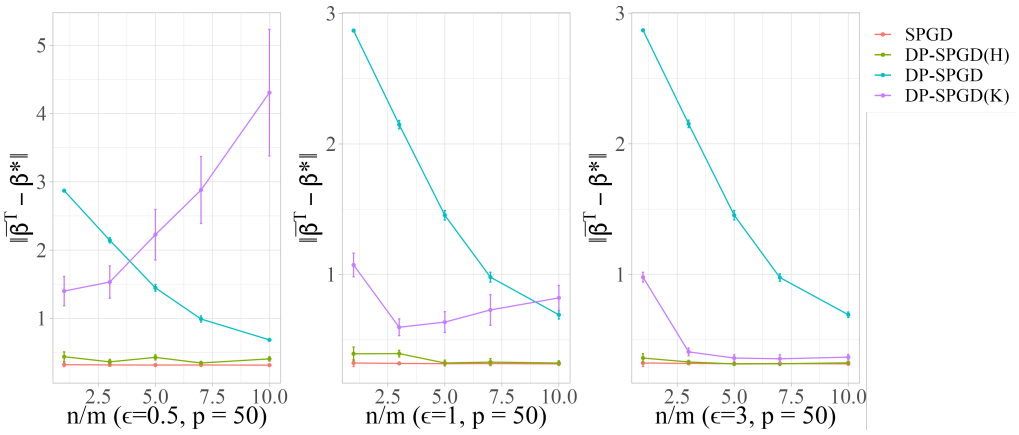

Figure S1: Simulation study results for the lasso model: accuracy vs. batch size n/m under $p = 50$ and different privacy budgets ϵ.

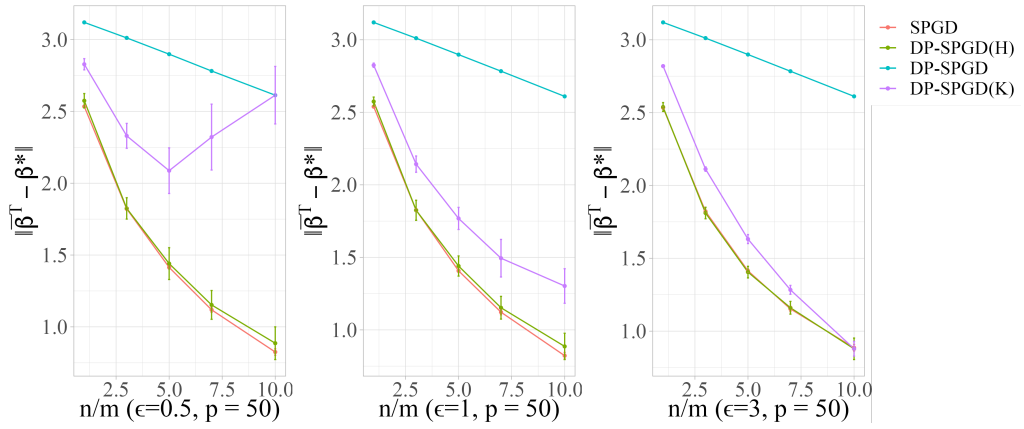

Figure S2: Simulation study results for the ℓ_1-regularized logistic regression model: accuracy vs. batch size n/m under $p = 50$ and different privacy budgets ϵ.

Table S1: Results for the blog feedback analysis: comparison of RMSE under different privacy budget ϵ. Averages (and in parentheses, standard errors) over 20 training–test splits are presented.

Methods	$\epsilon = 0.5$	$\epsilon = 2$
SPGD	$0.867_{(0.041)}$	$0.857_{(0.042)}$
DP-SPGD	$0.893_{(0.059)}$	$0.892_{(0.054)}$
DP-SPGD(K)	$0.954_{(0.051)}$	$0.905_{(0.046)}$
DP-SPGD(H)	$0.949_{(0.044)}$	$0.859_{(0.038)}$

S2 Technical Details

S2.1 Proofs for Algorithm 1

Proof of Proposition 1. The proof follows from Abadi et al. (2016), which provides a tight privacy bound for minibatch SGD under the assumption of

Table S2: Results for the crop-mapping analysis: comparison of classification accuracy under different privacy budget ϵ. Averages (and in parentheses, standard errors) over 20 training–test splits are presented.

Methods	$\epsilon = 0.5$	$\epsilon = 2$
SPGD	$0.935_{(0.004)}$	$0.935_{(0.003)}$
DP-SPGD	$0.916_{(0.004)}$	$0.916_{(0.003)}$
DP-SPGD(K)	$0.886_{(0.010)}$	$0.911_{(0.004)}$
DP-SPGD(H)	$0.909_{(0.007)}$	$0.929_{(0.002)}$

a bounded gradient. In Algorithms 1, the bound of the minibatch gradients is C/m after clipping. Consequently, the result follows from Theorem 1 in Abadi et al. (2016). □

Lemma 1 (Lemma 8 in Atchadé et al. (2017))**.** *Assume that* $\gamma \in (0, K_1^{-1}]$. *Then under Assumption (A1), for all* $\beta, \beta', \omega \in \mathcal{B}$,

$$-2\gamma\big(F_{\mathcal{D}}\left(\mathrm{prox}_{\gamma,g}(\beta)\right) - F_{\mathcal{D}}\left(\beta'\right)\big) \geq \left\|\mathrm{prox}_{\gamma,g}(\beta) - \beta'\right\|^2 + 2\left\langle \mathrm{prox}_{\gamma,g}(\beta) - \beta', \omega - \gamma\nabla L_{\mathcal{D}}(\omega) - \beta\right\rangle - \left\|\beta' - \omega\right\|^2.$$

Proof of Theorem 1. By Jensen's inequality,

$$\mathbb{E}[F_{\mathcal{D}}(\bar{\beta}^T) - F_{\mathcal{D}}(\beta^*)] \leq \mathbb{E}\left[T^{-1}\sum_{t=1}^{T}\{F_{\mathcal{D}}(\beta_t) - F_{\mathcal{D}}(\beta^*)\}\right].$$

Firstly, we consider a single step by taking $\beta = \beta_{t-1} - \gamma_{t-1}\nabla\widehat{L}_{\mathcal{D}}(\beta_{t-1})$,

$\omega = \beta_{t-1}$, $\beta' = \beta^*$, and $\gamma = \gamma_{t-1}$ in Lemma 1, which yields that

$$
\begin{aligned}
&F_{\mathcal{D}}(\beta_t) - F_{\mathcal{D}}(\beta^*) \\
&\leq \frac{1}{2\gamma_{t-1}}(\|\beta_{t-1} - \beta^*\|^2 - \|\beta_t - \beta^*\|^2) + \langle \beta_t - \beta^*, \nabla L_{\mathcal{D}}(\beta_{t-1}) - \nabla \widehat{L}_{\mathcal{D}}(\beta_{t-1}) \rangle \\
&\leq \frac{\Delta^2}{2\gamma_{t-1}} + \langle \beta_t - \beta^*, \nabla L_{\mathcal{D}}(\beta_{t-1}) - \nabla \widehat{L}_{\mathcal{D}}(\beta_{t-1}) \rangle,
\end{aligned}
$$

where the second inequality holds by Assumption (A2).

Summing for $t = 1, \ldots, T$, we have

$$
\sum_{t=1}^{T} \{F_{\mathcal{D}}(\beta_t) - F_{\mathcal{D}}(\beta^*)\} \leq \frac{\Delta^2}{2\gamma_{T-1}} + \sum_{t=1}^{T} \langle \beta_t - \beta^*, \nabla L_{\mathcal{D}}(\beta_{t-1}) - \nabla \widehat{L}_{\mathcal{D}}(\beta_{t-1}) \rangle.
$$

Then,

$$
\begin{aligned}
&\mathbb{E}[F_{\mathcal{D}}(\bar{\beta}^T) - F_{\mathcal{D}}(\beta^*)] \\
&\quad \leq \frac{\Delta^2}{2T\gamma_{T-1}} + \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\langle \beta_t - \beta^*, \nabla L_{\mathcal{D}}(\beta_{t-1}) - \nabla \widehat{L}_{\mathcal{D}}(\beta_{t-1}) \rangle] \\
&\quad \leq \frac{\Delta^2}{2T\gamma_{T-1}} + \frac{\Delta}{T} \sum_{t=1}^{T} \mathrm{Bias}_{\|\cdot\|}(\nabla \widehat{L}_{t-1}),
\end{aligned}
$$

where the second inequality comes from the Hölder's inequality. □

S2.2 Proofs for Algorithm 2

Lemma 2 (Lemma 13 in Kamath et al. (2020))**.** *Let $\mathcal{X}$ be a distribution over $\mathbb{R}$ with mean μ. Suppose that $x \sim \mathcal{X}$ and that $\mathbb{E}[|x - \mu|^k] \leq 1$ for some*

$k \geq 2$. *Furthermore, for any* $c \in \mathbb{R}$, *define*

$$Z = \begin{cases} c - \varrho/2, & x < c - \varrho/2 \\ c + \varrho/2, & x > c + \varrho/2 \\ x, & \textit{else.} \end{cases}$$

If $|\mu - c| \leq \varrho/4$, *then* $|\mu - \mathbb{E}[Z]| \leq 8(4/\varrho)^{k-1}$.

Proof of Theorem 2. For any $j \in [p]$ and $x \in Z_j^i$ (defined in Algorithm 3), by Hölder's inequality,

$$\mathbb{E}[x - \mathbb{E}[Z_j]]^2 \leq (\mathbb{E}[|x - \mathbb{E}[Z_j]|^k])^{2/k} \leq 1$$

for any fixed $i \in [q]$. The second inequality holds under the assumption that $\mathbb{E}[|\langle X - \mu, e_j \rangle|^k] \leq 1$ for $X \sim \mathcal{P}$.

Now, since

$$\begin{aligned} \mathbb{E}[\hat{\mu}_j^i - \mathbb{E}[Z_j]]^2 &= \mathbb{E}\left[\frac{q}{n}\sum_{x \in Z_j^i} x - \mathbb{E}[Z_j]\right]^2 \\ &= \frac{q^2}{n^2}\mathbb{E}\left[\sum_{x \in Z_j^i} x - \frac{n^2}{q^2}\mathbb{E}[Z_j]\right]^2 \\ &= \frac{q^2}{n^2}\mathbb{E}\left[\sum_{x \in Z_j^i} (x - \mathbb{E}[Z_j])^2\right] \\ &\leq \frac{q}{n}, \end{aligned}$$

Chebyshev's inequality implies that

$$\mathbb{P}\left[|\hat{\mu}_j^i - \mathbb{E}[Z_j]| \leq 10\sqrt{\frac{q}{n}}\right] \geq 0.9.$$

Recall that $\hat{\mu}_j = \text{Median}\{\hat{\mu}_j^1, \ldots, \hat{\mu}_j^q\}$. Let $Y_i = \mathbb{1}(|\hat{\mu}_j^i - \mathbb{E}[Z_j]| \geq 10\sqrt{q/n})$. The Y_is are i.i.d. Bernoulli random variables with a success probability of at most $1/10$. It follows that

$$\mathbb{P}\left[|\hat{\mu}_j - \mathbb{E}[Z_j]| \geq 10\sqrt{q/n}\right] = \mathbb{P}\left[\sum_{i=1}^{q} Y_i \geq \frac{q}{2}\right] = \mathbb{P}\left[\sum_{i=1}^{q}(Y_i - \frac{1}{10}) \geq \frac{2q}{5}\right].$$

By Hoeffding's inequality,

$$\mathbb{P}\left[|\hat{\mu}_j - \mathbb{E}[Z_j]| \geq 10\sqrt{q/n}\right] \leq \exp(-q/3).$$

Recall that $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_p)^\top + \mathcal{N}(0, \sigma^2 I_p)$, where $\sigma = \varrho q\sqrt{p\log(\delta^{-1})}/(n\epsilon)$. By Lemma 2 and the tail properties of the chi-squared distribution, with probability no less than $1 - \zeta$,

$$\|\hat{\mu} - \mu\| \leq O\Bigg[\frac{\varrho\log(p/\zeta)\sqrt{p\log(\delta^{-1})}}{n\epsilon}\left\{\sqrt{p} + \sqrt{\log(\zeta^{-1})}\right\} + \sqrt{p}\left\{\sqrt{\frac{\log(p/\zeta)}{n}} + \left(\frac{4}{\varrho}\right)^{k-1}\right\}\Bigg],$$

if we take $q = 3\log(2p/\zeta)$ and $\varrho \geq 4\|\mu\|_\infty$. □

Lemma 3. *For Algorithm 2, under Assumptions (A2)–(A5), with probability at least* $1 - \zeta$*,*

$$\|\nabla\widehat{L}_{\mathcal{D}}(\beta) - \nabla L_{\mathcal{D}}(\beta)\| \leq \tilde{O}\left[p^{\frac{3}{2}}\log(\zeta^{-1})\left\{\frac{\sqrt{p\log(\delta^{-1})}}{m\epsilon}\right\}^{\frac{k-1}{k}}\right]$$

for any $\beta \in \mathcal{B}$*.*

Proof of Lemma 3. Since $\mathcal{B}$ is closed and bounded, it is compact. The number of balls of radius ξ required to cover $\mathcal{B}$ is bounded above as $N_\xi \leq$

$(3\Delta/2\xi)^p$ (Kolmogorov and Tikhomirov, 1959). Let $\tilde{\beta} \in \{\tilde{\beta}_1, \ldots, \tilde{\beta}_{N_\xi}\}$ be any center of this ξ-net. It follows that

$$\|\nabla \widehat{L}_{\mathcal{D}}(\beta) - \nabla L_{\mathcal{D}}(\beta)\| \leq \|\nabla \widehat{L}_{\mathcal{D}}(\beta) - \nabla \widehat{L}_{\mathcal{D}}(\tilde{\beta})\| + \|\nabla \widehat{L}_{\mathcal{D}}(\tilde{\beta}) - \nabla L_{\mathcal{D}}(\tilde{\beta})\| + \|\nabla L_{\mathcal{D}}(\tilde{\beta}) - \nabla L_{\mathcal{D}}(\beta)\|.$$

Then we handle each component separately.

For the first term, let $N \sim \mathcal{N}(0, \sigma^2 I_p)$, where $\sigma = \varrho q \sqrt{p \log(\delta^{-1})}/(m\epsilon)$, we have

$$\begin{aligned}
\|\nabla \widehat{L}_{\mathcal{D}}(\beta) - \nabla \widehat{L}_{\mathcal{D}}(\tilde{\beta})\| &\leq \sqrt{\sum_{j=1}^{p} \left\{ \hat{\mu}_j(\beta) - \hat{\mu}_j(\tilde{\beta}) \right\}^2} + 2\|N\| \\
&= \sqrt{\sum_{j=1}^{p} \left\{ \frac{q}{m} \sum_{k=1}^{m/q} |\nabla \ell_{kj}(\beta) - \nabla \ell_{kj}(\tilde{\beta})| \right\}^2} + 2\|N\| \\
&\leq \sqrt{p} K_2 \|\beta - \tilde{\beta}\| + 2\|N\|.
\end{aligned}$$

Since

$$\mathbb{P}\left[\|N\| \leq 2\sigma \{\sqrt{p} + \sqrt{\log(1/2\zeta)}\} \right] > 1 - \zeta,$$

with probability at least $1 - \zeta$,

$$\begin{aligned}
\|\nabla \widehat{L}_{\mathcal{D}}(\beta) - \nabla \widehat{L}_{\mathcal{D}}(\tilde{\beta})\| &\leq O\left\{ \sqrt{p} \|\beta - \tilde{\beta}\| + \frac{\varrho q p \log^{1/2}(\delta^{-1}) \log(\zeta^{-1})}{m\epsilon} \right\} \\
&\leq \tilde{O}\left[\sqrt{p} \|\beta - \tilde{\beta}\| + \sqrt{p} \log(\zeta^{-1}) \left\{ \frac{\sqrt{p \log(\delta^{-1})}}{m\epsilon} \right\}^{\frac{k-1}{k}} \right],
\end{aligned}$$

where the notation $\tilde{O}$ omits some logarithmic terms.

Next is to bound the second term. From Theorem 2, we know for any center $\tilde{\beta}(\beta)$,

$$\|\nabla \widehat{L}_{\mathcal{D}}(\tilde{\beta}) - \nabla L_{\mathcal{D}}(\tilde{\beta})\| \leq O\Bigg[\frac{\varrho \log\left(\frac{p}{\zeta}\right)\sqrt{p\log(\delta^{-1})}}{m\epsilon}\Big\{\sqrt{p} + \sqrt{\log(\zeta^{-1})}\Big\} + \sqrt{p}\Bigg\{\sqrt{\frac{\log\left(\frac{p}{\zeta}\right)}{m}} + \Big(\frac{4}{\varrho}\Big)^{k-1}\Bigg\}\Bigg]$$

with probability at least $1-\zeta$. The ξ-net for $\mathcal{B}$ circumvents the need to take a supremum over $\beta \in \mathcal{B}$ as

$$\begin{aligned}
&\sup_{\beta\in\mathcal{B}}\|\nabla \widehat{L}_{\mathcal{D}}(\tilde{\beta}(\beta)) - \nabla L_{\mathcal{D}}(\tilde{\beta}(\beta))\| \\
&\quad= \max_{k\in N_\xi}\|\nabla \widehat{L}_{\mathcal{D}}(\tilde{\beta}_k) - \nabla L_{\mathcal{D}}(\tilde{\beta}_k)\| \\
&\quad\leq O\Bigg[\frac{\varrho \log\left(pN_\xi\zeta^{-1}\right)\sqrt{p\log(\delta^{-1})}}{m\epsilon}\Big\{\sqrt{p} + \sqrt{\log(N_\xi\zeta^{-1})}\Big\} + \sqrt{p}\Bigg\{\sqrt{\frac{\log\left(pN_\xi\zeta^{-1}\right)}{m}} + \Big(\frac{4}{\varrho}\Big)^{k-1}\Bigg\}\Bigg] \\
&\quad\leq \tilde{O}\Bigg[p^{\frac{3}{2}}\log(\zeta^{-1})\Big\{\frac{\sqrt{p\log(\delta^{-1})}}{m\epsilon}\Big\}^{\frac{k-1}{k}}\Bigg]
\end{aligned}$$

According to Assumption (A3), the third term $\|\nabla L_{\mathcal{D}}(\tilde{\beta}) - \nabla L_{\mathcal{D}}(\beta)\| \leq K_1\|\tilde{\beta}-\beta\|$.

To sum up,

$$\begin{aligned}
&\sup_{\beta\in\mathcal{B}}\|\nabla \widehat{L}_{\mathcal{D}}(\beta) - \nabla L_{\mathcal{D}}(\beta)\| \\
&\qquad\qquad\leq \tilde{O}\Bigg[\sqrt{p}\|\beta-\tilde{\beta}\| + p^{\frac{3}{2}}\log(\zeta^{-1})\Big\{\frac{\sqrt{p\log(\delta^{-1})}}{m\epsilon}\Big\}^{\frac{k-1}{k}}\Bigg]
\end{aligned}$$

with probability $1-\zeta$. Setting $\xi = m^{\frac{1-k}{k}}$ directly yields Lemma 3. $\square$

Proof of Corollary 1. We will begin by demonstrating that Algorithm 2 satisfies the DP guarantee. This can be directly inferred from the proof of Proposition 1, with minibatch gradients bounded by $\rho q\sqrt{p}/m$. Subsequently, we will derive an upper bound on the excess population risk with high probability:

$$
\begin{aligned}
&F_{\mathcal{D}}(\bar{\beta}^T) - F_{\mathcal{D}}(\beta^*) \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{1}{2\gamma_{t-1}}(\|\beta_{t-1}-\beta^*\|^2 - \|\beta_t-\beta^*\|^2) + \langle \beta_t - \beta^*, \nabla L_{\mathcal{D}}(\beta_{t-1}) - \nabla\widehat{L}_{\mathcal{D}}(\beta_{t-1})\rangle\right\} \\
&\leq \frac{\Delta^2}{2T\gamma_{T-1}} + T^{-1}\sum_{t=1}^{T}\langle \beta_t - \beta^*, \nabla L_{\mathcal{D}}(\beta_{t-1}) - \nabla\widehat{L}_{\mathcal{D}}(\beta_{t-1})\rangle \\
&\leq \frac{\Delta^2}{2T\gamma_{T-1}} + \tilde{O}\left[p^{\frac{3}{2}}\log(\zeta^{-1})\left\{\frac{\sqrt{p\log(\delta^{-1})}}{m\epsilon}\right\}^{\frac{k-1}{k}}\right]
\end{aligned}
$$

with probability at least $1-\zeta$. The first inequality holds by Lemma 1 while the second holds by Assumption (A2). The last inequality follows from the Hölder's inequality and Lemma 3. □

S2.3 Proofs for Algorithm 4

Lemma 4 (Theorem 2 in Wang et al. (2020))**.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples from some distribution Θ. Assume that that $\mathbb{E}_{X\sim\Theta}[X^2] \leq \tau$ for some known τ. Let $\nu = \sqrt{\log(\zeta^{-1})}$ and $s = \sqrt{n\epsilon\tau}/\{\log(\zeta^{-1})\log^{1/4}(\delta^{-1})\}$*

for a given failure probability ζ. Then with probability at least $1-\zeta$,

$$|\mathcal{A}(D) - \mathbb{E}[X]| \leq O\left\{\sqrt{\frac{\tau \log^{1/2}(\delta^{-1}) \log(\zeta^{-1})}{n\epsilon}}\right\}.$$

Lemma 5 (Lemma 7 in Holland (2019))**.** *The estimator $\hat{x}$ defined in* (5.3) *as a function of the data $(x_1, \ldots, x_n)^\top \in \mathbb{R}^n$ satisfies*

$$|\widehat{x}(\boldsymbol{x}) - \widehat{x}(\boldsymbol{x}')| \leq \frac{c_\chi}{n} \|\boldsymbol{x} - \boldsymbol{x}'\|_1,$$

where the factor c_χ takes the form

$$c_\chi = 1 - 2\Phi(-\sqrt{\nu}) + \sqrt{\frac{2}{\nu\pi}} \exp\left(-\frac{\nu}{2}\right)$$

and where Φ is the cumulative distribution function for the standard normal distribution.

Proof of Theorem 3. As the proof of Lemma 3, we apply the standard strategy of covering to obtain the uniform upper bound of the error term $\|\nabla \widehat{L}_{\mathcal{D}}(\beta) - \nabla L_{\mathcal{D}}(\beta)\|$. By triangle inequality, we decompose it into three components:

$$\begin{aligned}\|\nabla \widehat{L}_{\mathcal{D}}(\beta) - \nabla L_{\mathcal{D}}(\beta)\| \leq \|\nabla \widehat{L}_{\mathcal{D}}(\beta) - \nabla \widehat{L}_{\mathcal{D}}(\tilde{\beta})\| + \|\nabla \widehat{L}_{\mathcal{D}}(\tilde{\beta}) - \nabla L_{\mathcal{D}}(\tilde{\beta})\| \\ + \|\nabla L_{\mathcal{D}}(\tilde{\beta}) - \nabla L_{\mathcal{D}}(\beta)\|,\end{aligned}$$

where $\tilde{\beta} \in \{\tilde{\beta}_1, \ldots, \tilde{\beta}_{N_\xi}\}$ is any center of a ξ-net. Then we handle each component separately.

We first deal with the last term. By Assumption (A3),

$$\|\nabla L_{\mathcal{D}}(\tilde{\beta}) - \nabla L_{\mathcal{D}}(\beta)\| \leq K_1 \|\tilde{\beta} - \beta\|.$$

Regarding the first term, let $N \sim \mathcal{N}(0, \sigma^2 I_p)$, where $\sigma = 4s\sqrt{2p\log(\delta^{-1})}/(3m\epsilon)$, we have $\nabla \widehat{L}_{\mathcal{D}}(\beta) = \tilde{\ell}(\beta) + N$ in Algorithm 4 (line 6), so

$$\|\nabla \widehat{L}_{\mathcal{D}}(\beta) - \nabla \widehat{L}_{\mathcal{D}}(\tilde{\beta})\| \leq \|\tilde{\ell}(\beta) - \tilde{\ell}(\tilde{\beta})\| + 2\|N\|.$$

In accordance with the tail property of chi-squared distribution,

$$\mathbb{P}\left[\|N\| \leq 2\sigma\left\{\sqrt{p} + \sqrt{\log\left(\frac{1}{2\zeta}\right)}\right\}\right] > 1 - \zeta.$$

In addition, Lemma 5 yields that

$$\begin{aligned}
\|\tilde{\ell}(\beta) - \tilde{\ell}(\tilde{\beta})\|^2 &\leq \sum_{j=1}^{p}\left\{\frac{c_\chi}{m}\|\nabla \ell_j(\beta) - \nabla \ell_j(\tilde{\beta})\|_1\right\}^2 \\
&\leq \sum_{j=1}^{p}\left\{\frac{c_\chi}{m}\sum_{i=1}^{m}|\nabla \ell_{ij}(\beta) - \nabla \ell_{ij}(\tilde{\beta})|\right\}^2 \\
&\leq p c_\chi^2 K_2^2 \|\beta - \tilde{\beta}\|^2,
\end{aligned}$$

and so

$$\|\tilde{\ell}(\beta) - \tilde{\ell}(\tilde{\beta})\| \leq \sqrt{p} c_\chi K_2 \|\beta - \tilde{\beta}\|.$$

Therefore,

$$\|\nabla \widehat{L}_{\mathcal{D}}(\beta) - \nabla \widehat{L}_{\mathcal{D}}(\tilde{\beta})\| \leq \sqrt{p} c_\chi K_2 \|\beta - \tilde{\beta}\| + \frac{16s\sqrt{2p\log(\delta^{-1})}}{3m\epsilon}\left\{\sqrt{p} + \sqrt{\log\left(\frac{1}{2\zeta}\right)}\right\}.$$

To bound the second term, we apply Lemma 4. It follows that

$$|\nabla \widehat{L}_j(\beta) - \nabla L_j(\beta)| \leq \sqrt{\frac{\tau \log^{1/2}(\delta^{-1})\log(\zeta^{-1})}{m\epsilon}}$$

for $j \in [p]$ and any fixed β as long as $\mathbb{E}[\{\nabla_j \ell(\beta; x, y)\}^2] < \infty$. Hence,

$$\begin{aligned}
&\mathbb{P}\left[\|\nabla \widehat{L}_{\mathcal{D}}(\beta) - \nabla L_{\mathcal{D}}(\beta)\| > \sqrt{\frac{p\tau \log^{1/2}(\delta^{-1}) \log(\zeta^{-1})}{m\epsilon}}\right] \\
&\qquad \leq \sum_{j=1}^{p} \mathbb{P}\left[|\nabla \widehat{L}_j(\beta) - \nabla L_j(\beta)| \leq \sqrt{\frac{\tau \log^{1/2}(\delta^{-1}) \log(\zeta^{-1})}{m\epsilon}}\right] \\
&\qquad \leq p\zeta.
\end{aligned}$$

Let $\tilde{\beta} = \tilde{\beta}(\beta)$ denote the closest center to any fixed β. Thus,

$$\|\nabla \widehat{L}_{\mathcal{D}}(\tilde{\beta}) - \nabla L_{\mathcal{D}}(\tilde{\beta})\| > \sqrt{\frac{p\tau \log^{1/2}(\delta^{-1}) \log(p\zeta^{-1})}{m\epsilon}}$$

occurs with probability no greater than ζ. Then

$$\begin{aligned}
\sup_{\beta \in \mathcal{B}} \|\nabla \widehat{L}_{\mathcal{D}}(\tilde{\beta}(\beta)) - \nabla L_{\mathcal{D}}(\tilde{\beta}(\beta))\| &= \max_{k \in N_\xi} \|\nabla \widehat{L}_{\mathcal{D}}(\tilde{\beta}_k) - \nabla L_{\mathcal{D}}(\tilde{\beta}_k)\| \\
&\leq \sqrt{\frac{p\tau \log^{1/2}(\delta^{-1}) \log(pN_\xi\zeta^{-1})}{m\epsilon}}.
\end{aligned}$$

Therefore, if we let

$$\mathcal{V} = \frac{16s\sqrt{2p\log(\delta^{-1})}}{3m\epsilon}\left\{\sqrt{p} + \sqrt{\log\left(\frac{1}{2\zeta}\right)}\right\},$$

a uniform upper bound on $\|\nabla \widehat{L}_{\mathcal{D}}(\beta) - \nabla L_{\mathcal{D}}(\beta)\|$ is

$$\begin{aligned}
&\sup_{\beta \in \mathcal{B}} \|\nabla \widehat{L}_{\mathcal{D}}(\beta) - \nabla L_{\mathcal{D}}(\beta)\| \\
&\leq \sup_{\beta \in \mathcal{B}} \left(\sqrt{p} c_\chi K_2 \|\beta - \tilde{\beta}\| + K_1 \|\beta - \tilde{\beta}\|\right) + \sqrt{\frac{p\tau \log^{1/2}(\delta^{-1}) \log(pN_\xi\zeta^{-1})}{m\epsilon}} + \mathcal{V} \\
&\leq \max\{K_1, K_2\}\xi(\sqrt{p}c_\chi + 1) + \sqrt{\frac{p\tau \log^{1/2}(\delta^{-1}) \log(pN_\xi\zeta^{-1})}{m\epsilon}} + \mathcal{V}
\end{aligned}$$

with probability at least $1-\zeta$. Setting $s = \sqrt{m\epsilon\tau}/\{\log(\zeta^{-1})\log^{1/4}(\delta^{-1})\}$ and $\xi = m^{-1/2}$ directly yield Theorem 3. □

Proof of Corollary 2. Once the truncation step is performed, the robust gradients in Algorithm 4 are bounded by $4s\sqrt{2p}/(3m)$. As a result, it is easy to demonstrate that the privacy guarantee is met. The proof for the utility guarantee is identical to that of Corollary 1, but with the application of the uniform bound from Theorem 3. □

S2.4 Explicit form of $C(a,b)$ in Catoni and Giulini (2017)

The correction term $C(a,b)$ can be computed as follows. First, define

$$V_- = \frac{\sqrt{2}-a}{b},\ V_+ = \frac{\sqrt{2}+a}{b},\ F_- = \Phi\left(-V_-\right),$$
$$F_+ = \Phi\left(-V_+\right),\ E_- = \exp\left(-\frac{V_-^2}{2}\right),\ \text{and}\ E_+ = \exp\left(-\frac{V_+^2}{2}\right),$$

where Φ denotes the cumulative distribution function for the standard normal distribution. With these elements, we can break the final quantity into

five terms:

$$
\begin{aligned}
T_1 &= \frac{2\sqrt{2}}{3}\left(F_- - F_+\right), \\
T_2 &= -\left(a - \frac{a^3}{6}\right)\left(F_- + F_+\right), \\
T_3 &= \frac{b}{\sqrt{2\pi}}\left(1 - \frac{a^2}{2}\right)\left(E_+ - E_-\right), \\
T_4 &= \frac{ab^2}{2}\left\{F_+ + F_- + \frac{1}{\sqrt{2\pi}}\left(V_+E_+ + V_-E_-\right)\right\}, \\
T_5 &= \frac{b^3}{6\sqrt{2\pi}}\left\{\left(2 + V_-^2\right)E_- - \left(2 + V_+^2\right)E_+\right\}.
\end{aligned}
$$

At last, we explicitly define

$$C(a, b) = T_1 + T_2 + T_3 + T_4 + T_5.$$

Bibliography

Abadi, M., A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318.

Atchadé, Y. F., G. Fort, and E. Moulines (2017). On perturbed proximal gradient algorithms. *The Journal of Machine Learning Research 18*(1), 310–342.

Catoni, O. and I. Giulini (2017). Dimension-free PAC-Bayesian bounds

for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*.

Holland, M. J. (2019). Robust descent using smoothed multiplicative noise. In *The 22nd International Conference on Artificial Intelligence and Statistics*, Volume 89, pp. 703–711. PMLR.

Kamath, G., V. Singhal, and J. Ullman (2020). Private mean estimation of heavy-tailed distributions. In *Conference on Learning Theory*, Volume 125, pp. 2204–2235. PMLR.

Kolmogorov, A. and V. Tikhomirov (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Mat. Nauk 14*, 3–86.

Wang, D., J. Ding, L. Hu, Z. Xie, M. Pan, and J. Xu (2020). Differentially private (gradient) expectation maximization algorithm with statistical guarantees. *arXiv preprint arXiv:2010.13520*.