Supplementary Material (AISTATS 2019): Rényi Differentially Private ERM for Smooth Objectives

A Proofs

Lemma 2. *Let B and B' be mini-batches that differ on the value of one record. Define the operator $\mathcal{T}_B(\cdot) = \mathrm{Id}(\cdot) - \eta\nabla f_B(\cdot)$ (and similarly for B'). Let $\mathbf{w}$ and $\mathbf{w}'$ be any two vectors in Θ. Let $\rho = \max\{|1-\eta\mu|, |1-\eta L|\}$ (where μ is the strong convexity parameter and L is the smoothness parameter). Then:*

$$||\mathcal{T}_B(\mathbf{w}) - \mathcal{T}_B(\mathbf{w}')|| \le \rho||\mathbf{w}-\mathbf{w}'|| \quad \textit{(same batch B)}$$
$$||\mathcal{T}_B(\mathbf{w}) - \mathcal{T}_{B'}(\mathbf{w}')|| \le \rho||\mathbf{w}-\mathbf{w}'|| + \frac{2\eta R}{|B|}$$

where the first equation shows the effect of using the same operator $\mathcal{T}_B$ and the second equation shows the effect of using $\mathcal{T}_B$ to update $\mathbf{w}$ and a different operator $\mathcal{T}_{B'}$ to update $\mathbf{w}'$.

Proof. We first consider the case where the same operator $\mathcal{T}_B$ is applied to both $\mathbf{w}$ and $\mathbf{w}'$, i.e., $B = B'$.

$$\begin{aligned}
\|\mathcal{T}_B(\mathbf{w}) - \mathcal{T}_B(\mathbf{w}')\|_2 &= \|\mathbf{w} - \eta\nabla f_B(\mathbf{w}) - (\mathbf{w}' - \eta\nabla f_B(\mathbf{w}'))\|_2 \\
&= \|\mathbf{w} - \mathbf{w}' - \eta(\nabla f_B(\mathbf{w}) - \nabla f_B(\mathbf{w}'))\|_2 \\
&= \left\|\int_0^1 \{\mathbf{I} - \eta\nabla^2 f_B(\mathbf{w}' + s(\mathbf{w}-\mathbf{w}'))\}(\mathbf{w}-\mathbf{w}')\,\mathrm{d}s\right\|_2 \\
&\le \int_0^1 \left\|\{\mathbf{I} - \eta\nabla^2 f_B(\mathbf{w}' + s(\mathbf{w}-\mathbf{w}'))\}(\mathbf{w}-\mathbf{w}')\right\|_2 \mathrm{d}s \\
&\le \int_0^1 \left\|\mathbf{I} - \eta\nabla^2 f_B(\mathbf{w}' + s(\mathbf{w}_t-\mathbf{w}_t'))\right\|_2 \|\mathbf{w}-\mathbf{w}'\|_2\,\mathrm{d}s \\
&\le \int_0^1 \sup_{\mathbf{z}} \|\mathbf{I} - \eta\nabla^2 f_B(\mathbf{z})\|_2 \|\mathbf{w}-\mathbf{w}'\|_2\,\mathrm{d}s \\
&\le \sup_{\mathbf{z}} \|\mathbf{I} - \eta_t\nabla^2 f(\mathbf{z})\|_2 \|\mathbf{w}-\mathbf{w}'\|_2 \\
&\le \max\{|1-\eta\mu|,\, |1-\eta L|\}\|\mathbf{w}-\mathbf{w}'\|_2 \\
&= \rho\|\mathbf{w}-\mathbf{w}'\|_2\,,
\end{aligned}$$

where $\mathbf{z} = \mathbf{w}' + s^*(\mathbf{w}-\mathbf{w}')$, $s^* \in [0,1]$ is a point on the line segment joining $\mathbf{w}$ and $\mathbf{w}'$.

Now we consider the case where B and B' differ by one record. Let ξ denote the index of record at which D and D' differ, i.e., $d_i = d_i'$ for all $i \ne \xi$ and $d_\xi \ne d_\xi'$. We introduce the following equality.

$$\begin{aligned}
\nabla f_B(\mathbf{w}) - \nabla f_B(\mathbf{w}') &= \frac{1}{|B|}\left\{\sum_{i\in B}\nabla f(\mathbf{w}, d_i) - \sum_{i\in B'}\nabla f(\mathbf{w}', d_i')\right\} \\
&= \frac{1}{|B|}\Big\{\nabla f(\mathbf{w}, d_\xi) - \nabla f(\mathbf{w}', d_\xi) + \nabla f(\mathbf{w}', d_\xi) - \nabla f(\mathbf{w}', d_\xi') + \sum_{i\in B, i\ne\xi}\nabla f(\mathbf{w}, d_i) - \nabla f(\mathbf{w}', d_i)\Big\} \\
&= \frac{1}{|B|}\left\{\big(\nabla f(\mathbf{w}', d_\xi) - \nabla f(\mathbf{w}', d_\xi')\big) + \sum_{i\in B}\nabla f(\mathbf{w}, d_i) - \nabla f(\mathbf{w}', d_i)\right\} \\
&= \nabla f_B(\mathbf{w}) - \nabla f_B(\mathbf{w}') + \frac{1}{|B|}\big(\nabla f(\mathbf{w}', d_\xi) - \nabla f(\mathbf{w}', d_\xi')\big) \qquad (6)
\end{aligned}$$

Using Equation (6), we get

$$\begin{aligned}
\|\mathcal{T}_B(\mathbf{w}) - \mathcal{T}_{B'}(\mathbf{w}')\|_2 &= \|\mathbf{w} - \eta\nabla f_B(\mathbf{w}) - (\mathbf{w}' - \eta\nabla f_{B'}(\mathbf{w}'))\|_2 \\
&= \|\mathbf{w} - \mathbf{w}' - \eta(\nabla f_B(\mathbf{w}) - \nabla f_{B'}(\mathbf{w}'))\|_2 \\
&= \left\|\mathbf{w} - \mathbf{w} - \eta(\nabla f_B(\mathbf{w}) - \nabla f_B(\mathbf{w}')) + \frac{\eta}{|B|}\left(\nabla f(\mathbf{w}', d_\xi) - \nabla f(\mathbf{w}', d'_\xi)\right)\right\|_2 \\
&\leq \|\mathbf{w} - \mathbf{w}' - \eta(\nabla f_B(\mathbf{w}_t) - \nabla f_B(\mathbf{w}'))\|_2 + \frac{\eta}{|B|}\|\nabla f(\mathbf{w}', d_\xi) - \nabla f(\mathbf{w}', d'_\xi)\|_2 \\
&\leq \|\mathbf{w} - \mathbf{w}' - \eta(\nabla f_B(\mathbf{w}) - \nabla f_B(\mathbf{w}'))\|_2 + \frac{2\eta R}{|B|} \\
&= \|\mathcal{T}_B(\mathbf{w}) - \mathcal{T}_B(\mathbf{w}')\|_2 + \frac{2\eta R}{|B|} \\
&\leq \rho\|\mathbf{w} - \mathbf{w}'\|_2 + \frac{2\eta R}{|B|},
\end{aligned}$$

where the second to last inequality is due to our requirement on the boundedness of gradient. □

Lemma 3. *Define $H_\alpha(P_1; P_2) = e^{(\alpha-1)\,\mathrm{D}_\alpha(P_1 \,\|\, P_2)}$. Let $\mathcal{M}_1, \ldots, \mathcal{M}_m$ be mechanisms and $q = [q_1, \ldots, q_m]$ be a probability vector over $1, \ldots, m$. Let $\mathcal{M}$, on input D, sample $i \sim q$ and return $\mathcal{M}_i(D)$. Then $H_\alpha(\mathcal{M}(D_1); \mathcal{M}(D_2)) \leq \sum_{j=1}^{m} q_j H_\alpha(\mathcal{M}_j(D_1); \mathcal{M}_j(D_2))$.*

Proof. For each j, let P_1^j and P_2^j be the distributions of $\mathcal{M}_j(D_1)$ and $\mathcal{M}_j(D_2)$, respectively. Let P_1 be the distribution of $\mathcal{M}(D_1)$ and let P_2 be the distribution of $\mathcal{M}(D_2)$.

$$\begin{aligned}
&H_\alpha(\mathcal{M}(D_1); \mathcal{M}(D_2)) \\
&= \mathbb{E}_{x\sim P_2}\left[P_1(x)^\alpha P_2(x)^{-\alpha}\right] \\
&= \mathbb{E}_{x\sim P_2}\left[\left(\frac{\sum_{j=1}^m q_j P_1^j(x)}{\sum_{j=1}^m q_j P_2^j(x)}\right)^\alpha\right] \\
&= \mathbb{E}_{x\sim P_2}\left[\left(\sum_{j=1}^m \frac{q_j P_2^j(x)}{\sum_{j'=1}^m q_{j'} P_2^{j'}(x)} \frac{P_1^j(x)}{P_2^j(x)}\right)^\alpha\right] \\
&= \mathbb{E}_{x\sim P_2}\left[\left(\sum_{j=1}^m \frac{q_j P_2^j(x)}{P_2(x)} \frac{P_1^j(x)}{P_2^j(x)}\right)^\alpha\right] \\
&\leq \mathbb{E}_{x\sim P_2}\left[\sum_{j=1}^m \frac{q_j P_2^j(x)}{P_2(x)} \left(\frac{P_1^j(x)}{P_2^j(x)}\right)^\alpha\right] \\
&= \sum_{j=1}^m q_j\, \mathbb{E}_{x\sim P_2^j}\left[\left(\frac{P_1^j(x)}{P_2^j(x)}\right)^\alpha\right] \\
&= \sum_{j=1}^m q_j H_\alpha(\mathcal{M}_j(D_1); \mathcal{M}_j(D_2)),
\end{aligned}$$

where the inequality comes from Jensen's inequality (since the function $z \mapsto z^\alpha$ is convex for $\alpha > 1$) and the second-to-last equality comes from using the definition of expected value. □

Proposition 2. *If we run Algorithm 1 for arbitrary number of epochs with a fixed step size η, its sensitivity Δ satisfies*

$$\Delta \leq \frac{2\eta R}{|B|(1-\rho^m)},$$

where $\rho = \max\{|1-\eta\mu|, |1-\eta L|\}$. In particular, when $m = 1$ and $\eta = \frac{2}{L+\mu}$, $\Delta \leq \frac{2R}{n\mu}$.

Proof. Let D and D' be any two databases that differ on one record. Given a fixed randomness in data permutation, let $B_0, \ldots, B_{m-1}$ and $B'_0, \ldots, B'_{m-1}$ denote m disjoint mini-batches for D and D', respectively. Then there exists an index j such that $B_j \neq B'_j$ and $B_i = B'_i$ for all $i \neq j$.

Algorithm 1 on input D generates a sequence of solutions $\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, \ldots$, using the rule $\mathbf{w}_i = \mathcal{T}_{B_{i-1 \bmod m}}(\mathbf{w}_{i-1})$ (and similarly on input D' using $\mathcal{T}_{B'}$). Define $\Delta^{(k)}$ as the difference between $\mathbf{w}_i$ and $\mathbf{w}'_i$ at the end of k^{th} epoch. Provided that the algorithm for input D and D' starts with the same initial solution, i.e., $\mathbf{w}_0 = \mathbf{w}'_0$, Lemma 2 says that the first $j-1$ updates in an epoch will be contractions, the j^{th} update will be an expansion, and the remaining $m-j$ updates will be contractions. Therefore, at the end of the first epoch, we have $\Delta^{(1)} \leq \rho^{m-j}\frac{2\eta R}{|B|}$. In the second epoch, there will be again $j-1$ contractions, one expansion, and $m-j$ contractions. Hence, we have

$$\begin{aligned}\Delta^{(2)} &\leq \rho^{m-j}\left(\rho \cdot (\rho^{j-1}\Delta^{(1)}) + \frac{2\eta R}{|B|}\right)\\ &= \rho^m \Delta^{(1)} + \rho^{m-j}\frac{2\eta R}{|B|}\\ &\leq \rho^m \cdot \rho^{m-j}\frac{2\eta R}{|B|} + \rho^{m-j}\frac{2\eta R}{|B|}\,.\end{aligned}$$

Likewise, at the end of the k^{th} epoch,

$$\Delta^{(k)} \leq \rho^{m-j}\frac{2\eta R}{|B|}\left(\rho^{(k-1)m} + \rho^{(k-2)m} + \cdots + \rho^m + 1\right)\,.$$

Therefore,

$$\lim_{k\to\infty}\Delta^{(k)} = \frac{\rho^{m-j}2\eta R}{|B|(1-\rho^m)} \leq \frac{2\eta R}{|B|(1-\rho^m)} \tag{7}$$

since $0 < \rho < 1$. Recall that $\rho = \max\{|1-\eta\mu|, |1-\eta L|\}$. We see that ρ is a function of step size η, and the value of η can be optimized to minimize ρ (i.e., to obtain the maximum contraction). It can be seen that ρ has the minimum value of $\frac{L-\mu}{L+\mu}$ when $\eta = \frac{2}{L+\mu}$, which is when $|1-\eta\mu| = |1-\eta L|$. Plugging $\rho = \frac{L-\mu}{L+\mu}$ and $m = 1$ into (7), we obtain the second claim. □

Proposition 3. *Algorithm 3 with averaging satisfies (α, ϵ)-RDP, where* $\epsilon = \frac{1}{\alpha-1}\log\left(\frac{1}{m}\sum_{j=1}^m e^{\frac{\alpha(\alpha-1)(\Delta[j])^2}{2\sigma^2}}\right)$.

Proof. Let D and D' be neighboring databases. Let $\mathcal{M}_j$ be a mechanism with associated sensitivity $\Delta[j]$. Given the randomly permuted input dataset, Algorithm 3, denoted by $\mathcal{M}$, chooses $\mathcal{M}_j$ with probability $q[j] = 1/m$ and releases the output using the Gaussian mechanism with noise scale parameter σ. We show that the Rényi divergence between the output distributions of $\mathcal{M}$ is bounded by ϵ.

$$\begin{aligned}\mathrm{D}_\alpha(\mathcal{M}(D) \,\|\, \mathcal{M}(D')) &= \frac{1}{\alpha-1}\log H_\alpha(\mathcal{M}(D); \mathcal{M}(D'))\\ &\leq \frac{1}{\alpha-1}\log\left(\sum_{j=1}^m q[j] H_\alpha(\mathcal{M}_j(D); \mathcal{M}_j(D'))\right)\\ &= \frac{1}{\alpha-1}\log\left(\frac{1}{m}\sum_{j=1}^m e^{(\alpha-1)\,\mathrm{D}_\alpha(\mathcal{M}_j(D)\,\|\,\mathcal{M}_j(D'))}\right)\\ &\leq \frac{1}{\alpha-1}\log\left(\frac{1}{m}\sum_{j=1}^m e^{\alpha(\alpha-1)\Delta[j]^2/2\sigma^2}\right)\,,\end{aligned}$$

where the first and second inequalities are due to Lemmas 3 and 1, respectively. □

B KDDCup99 Dataset

To demonstrate the performance on a large dataset, we evaluate the proposed algorithm on KDDCup99 dataset. Figure 4 shows the performance for LR and SVM. For LR, output perturbation methods perform better when ϵ is small while gradient perturbation methods outperform when ϵ is large. While OutPert-GD perform very poorly on other 4 datasets, it shows a comparable performance on the large dataset. This is because its sensitivity is inversely proportional to the dataset size.

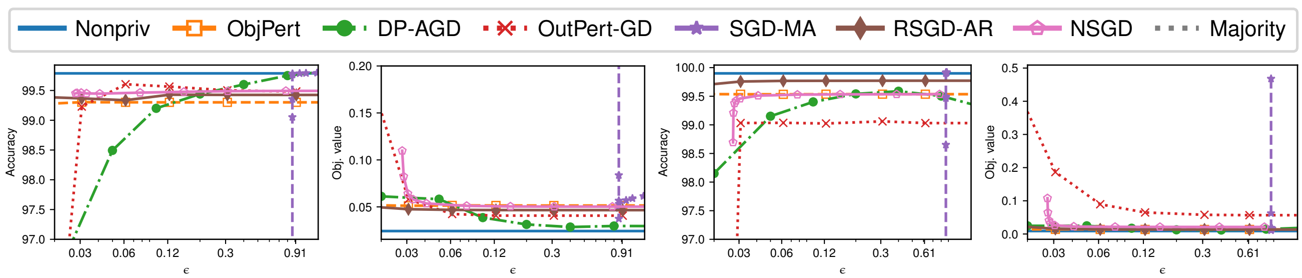

Figure 4: Performance on KDDCup99 dataset (Left: LR, Right: SVM)