Learning Privately from Multiparty Data: Supplementary Material

A. Proofs

A.1. Proof of Theorem 1

Theorem 1: The perturbed output $w_p = w_s + \eta$ from Algorithm 1 with $p(\eta) \propto e^{-\frac{\lambda\epsilon}{2}\|\eta\|}$ is ϵ-differentially private.

Proof. We will compute the sensitivity of the minimizer w_s of the regularized empirical risk with majority-voted labels (8). Suppose $\mathcal{D} = (S^{(1)}, ..., S^{(M)})$ is the ordered set of private training data (5) for M parties, and $\mathcal{D}' = ((S')^{(1)}, ..., S^{(M)})$ is a neighboring set which differs from $\mathcal{D}$ only at party 1's data, without loss of generality. The local classifiers after training with $\mathcal{D}$ and $\mathcal{D}'$ are $H = (h_1, ..., h_M)$ and $H' = (h'_1, ..., h_M)$, respectively, which are again different only for classifier 1. The majority votes $v(x)$ and $v'(x)$ from $\mathcal{D}$ and $\mathcal{D}'$ generates two auxiliary training sets $S = \{(x_i, v(x_i))\}$ and $S' = \{(x_i, v'(x_i)\}$ which have the same features but possibly different labels.

Let $R^\lambda_S(w)$ and $R^\lambda_{S'}(w)$ be the regularized empirical risks for training sets S and S', and let w_s and $w_{s'}$ be the minimizers of the respective risks. From Corollaries 7 and 8 (Chaudhuri et al., 2011), the L_2 difference of w_s and $w_{s'}$ is bounded by

$$\|w_s - w_{s'}\| \le \frac{1}{\lambda} \max_w \|\nabla g(w)\|, \tag{26}$$

where $g(w)$ is the risk difference $R^\lambda_S(w) - R^\lambda_{S'}(w)$, which, in our case, satisfies

$$\begin{aligned}\|\nabla g(w)\| &\le \frac{1}{N}\sum_{i=1}^N \|v(x_i)x_i l'(v(x_i)w^T x_i) \\ &\quad -v'(x_i)x_i l'(v'(x_i)w^T x_i)\|. \\ &\le \frac{1}{N}\sum_{i=1}^N \|x_i\| \times \\ &\quad |l'(w^T x_i) + l'(-w^T x_i)|. \end{aligned} \tag{27}$$

Recall that $\|x\| \le 1$ and $|l'(\cdot)| \le 1$ by assumption. In the worst case, $v(x_i) \ne v'(x_i)$ for all $i = 1, ..., N$, and therefore the RHS of (27) is bounded by 2. Consequently, the L_2 sensitivity of the minimizer w_s is

$$\max_{S,S'} \|w_s - w_{s'}\| \le \frac{2}{\lambda}. \tag{28}$$

ϵ-differential privacy follows from the sensitivity result (3). □

A.2. Proof of Theorem 3

Theorem 3: The perturbed output $w_p = w_s + \eta$ from Algorithm 2 with $p(\eta) \propto e^{-\frac{M\lambda\epsilon}{2}\|\eta\|}$ is ϵ-differentially private.

Proof. The proof parallels the proof of Theorem 1. We again assume $\mathcal{D} = (S^{(1)}, ..., S^{(M)})$ is the ordered set of private training data (5) for M parties, and $\mathcal{D}' = ((S')^{(1)}, ..., S^{(M)})$ is a neighboring set which differs from $\mathcal{D}$ only at party 1's data, without loss of generality. Let $S = \{(x_i, \alpha_i)\}$ and $S' = \{(x_i, \alpha'_i)\}$ be the two resulting datasets which have the same the features but possibly different α's. We first compute the sensitivity of the minimizer of the weighted regularized empirical risk (19). Let $R^\lambda_S(w)$ and $R^\lambda_{S'}(w)$ be the regularized empirical risks for training sets S and S', and let w_s and $w_{s'}$ be the minimizers of the respective risks. Also let $g(w)$ be the difference $R^\lambda_S(w) - R^\lambda_{S'}(w)$ of two risks

$$\begin{aligned} g(w) &= \frac{1}{N}\sum_{i=1}^N [\alpha_i l(w^T x_i) + (1-\alpha_i) l(-w^T x_i) \\ &\quad -\alpha'_i l(w^T x_i) - (1-\alpha'_i) l(-w^T x_i)]. \end{aligned} \tag{29}$$

The gradient of $g(w)$ is bounded by

$$\begin{aligned}\|\nabla g(w)\| &\le \frac{1}{N}\sum_{i=1}^N [|\alpha_i - \alpha'_i|\|x_i\||l'(w^T x_i)| \\ &\quad +|\alpha_i - \alpha'_i|\|x_i\||l'(-w^T x_i)|] \qquad (30) \\ &\le \frac{1}{N}\sum_{i=1}^N 2|\alpha_i - \alpha'_i|. \qquad (31)\end{aligned}$$

In the worst case, $\alpha_i \ne \alpha'_i$ for all $i = 1, ..., N$. Since α_i is the fraction of positive votes, $|\alpha_i - \alpha'_i| \le 1/M$ holds for all $i = 1, ..., N$. Therefore the L_2 sensitivity of the minimizer w_s is at most $\frac{2}{\lambda M}$ and the ϵ-differential privacy follows. □

A.3. Lemma 5

We use the following lemma.

Lemma 5 (Lemma 17 of (Chaudhuri et al., 2011)). *If $X \sim \Gamma(k,\theta)$, where k is an integer, then with probability of at least $1-\delta$,*

$$X \leq k\theta \log(k/\delta).$$

A.4. Lemma 6

Lemma 6. *If w_s is the minimizer of (19) and w_p is the ϵ-differentially private version from Algorithm 2, then with probability of at least $1-\delta_p$ over the privacy mechanism,*

$$R_S^\lambda(w_p) \leq R_S^\lambda(w_s) + \frac{2d^2(c+\lambda)\log^2(d/\delta)}{\lambda^2 M^2 \epsilon^2} \tag{32}$$

Proof. A differentiable function $f:\mathbb{R}^d \to \mathbb{R}$ is called β-smooth, if $\exists \beta > 0$ such that $\|\nabla f(v) - \nabla f(u)\| \leq \beta\|v-u\|$ for all u, v. From the Mean Value Theorem, such a function satisfies

$$f(v) \leq f(u) + \nabla^T f(u)(v-u) + \frac{\beta}{2}\|v-u\|^2,\ \forall u, v.$$

Since $|l'(\cdot)|$ is c-Lipschitz, $R_S^\lambda(w)$ is $(c+\lambda)$-smooth:

$$\begin{aligned}
&\|\nabla R_S^\lambda(v) - \nabla R_S^\lambda(u)\| \\
&\leq \frac{1}{N}\sum_i \big\|\alpha_i x_i l'(v^T x_i) - (1-\alpha_i) x_i l'(-v^T x_i) \\
&-\alpha_i x_i l'(u^T x_i) + (1-\alpha_i) x_i l'(-u^T x_i)\big\| \\
&+\lambda\|v-u\| \\
&\leq \frac{1}{N}\sum_i \big[\alpha_i c\|(v-u)^T x_i\| + \\
&(1-\alpha_i)c\|(u-v)^T x_i\|\big] + \lambda\|v-u\| \\
&\leq (c+\lambda)\|u-v\|.
\end{aligned} \tag{33}$$

By setting $v = w_p$ and $u = w_s$ and using the $(c+\lambda)$-smoothness of $R_S^\lambda(w)$, we have

$$\begin{aligned}
R_S^\lambda(w_p) &\leq R_S^\lambda(w_s) + \nabla^T R_S^\lambda(w_s)(w_p - w_s) \\
&\quad + \frac{(c+\lambda)}{2}\|w_p - w_s^*\|^2 \\
&= R_S^\lambda(w_s) + \frac{(c+\lambda)}{2}\|w_p - w_s\|^2.
\end{aligned} \tag{34}$$

Since

$$P\left(\|w_p - w_s^*\| \leq \frac{2d\log(d/\delta)}{\lambda M \epsilon}\right) \geq 1-\delta_p \tag{35}$$

from Lemma 5 with $k = d$ and $\theta = \frac{2}{\lambda M\epsilon}$, we have the desired result. □

A.5. Proof of Theorem 4

Theorem 4: Let w_0 be any reference hypothesis. Then with probability of at least $1-\delta_p-\delta_s$ over the privacy mechanism (δ_p) and over the choice of samples (δ_s),

$$\begin{aligned}
R(w_p) &\leq R(w_0) + \frac{4d^2(c+\lambda)\log^2(d/\delta_p)}{\lambda^2 M^2\epsilon^2} \\
&\quad + \frac{16(32+\log(1/\delta_s))}{\lambda N} + \frac{\lambda}{2}\|w_0\|^2.
\end{aligned} \tag{36}$$

Proof. Let w_s and w^* be the minimizers of the regularized empirical risk R_S^λ and R^λ, respectively. The risk at w_p relative to a reference classifier w_0 can be written as

$$\begin{aligned}
R(w_p) - R(w_0) &= R^\lambda(w_p) - R^\lambda(w^*) \\
&\quad + R^\lambda(w^*) - R^\lambda(w_0) \\
&\quad + \frac{\lambda}{2}\|w_0\|^2 - \frac{\lambda}{2}\|w_p\|^2 \\
&\leq R^\lambda(w_p) - R^\lambda(w^*) + \frac{\lambda}{2}\|w_0\|^2.
\end{aligned} \tag{37}$$

The inequality above follows from $R^\lambda(w^*) \leq R^\lambda(w_0)$ by definition. Note that since $\|x\| \leq 1$ and $|l'| \leq 1$ by assumption, the weighted loss $\alpha(x)l(w^T x) + (1-\alpha(x))l(w^T x)$ is 1-Lipschitz in w. From Theorem 1 of (Sridharan et al., 2009) with $a = 1$, we can also bound $R^\lambda(w_p) - R^\lambda(w^*)$ as

$$\begin{aligned}
R^\lambda(w_p) - R^\lambda(w^*) &\leq 2(R_S^\lambda(w_p) - R_S^\lambda(w_s^*)) \\
&\quad + \frac{16(32+\log(1/\delta_s))}{\lambda N}
\end{aligned} \tag{38}$$

with probability of $1-\delta_s$ over the choice of samples. By combining this inequality with Lemma 6 using the union bound, we have

$$\begin{aligned}
R^\lambda(w_p) - R^\lambda(w^*) &\leq \frac{4d^2(c+\lambda)\log^2(d/\delta_p)}{\lambda^2 M^2\epsilon^2} \\
&\quad + \frac{16(32+\log(1/\delta_s))}{\lambda N}.
\end{aligned} \tag{39}$$

The theorem follows from (37).

□

B. Differentially private multiclass logistic regression

We extend our methods to multiclass classification problems and provide a sketch of ϵ-differential privacy proofs for multiclass logistic regression loss.

B.1. Standard ERM

Suppose $y \in 1, ..., K$, and let $w = [w_1; \ \ldots \ ; w_K]$ be a stacked $(d\ K) \times 1$ vector. The multiclass logistic loss (i.e. softmax) is

$$l(h(x), y) = -w_y^T x + \log(\sum_l e^{w_l^T x}), \tag{40}$$

and the regularized empirical risk is

$$R_S^\lambda(w) = -\frac{1}{N}\sum_i [w_{y_i}^T x_i - \log(\sum_l e^{w_l^T x_i})] + \frac{\lambda}{2}\|w\|^2. \tag{41}$$

Note that $R_S^\lambda(w)$ is λ-strongly convex in w.

The sensitivity of w_s which minimizes (41) can be computed as follows. Suppose S and S' are two different datasets which are not necessarily neighbors: $S = \{(x_i, y_i))\}$ and $S' = \{(x'_i, y'_i)\}$. Let $g(w)$ be the difference $R_S^\lambda(w) - R_{S'}^\lambda(w)$ of the two risks. Then the partial gradient w.r.t. w_k is

$$\nabla_{w_k} R_S^\lambda(w) = -\frac{1}{N}\sum_i x_i \Delta_k(x_i, y_i, w) + \lambda w_k, \tag{42}$$

where

$$\Delta_k(x_i, y_i, w) = I[y_i = k] - \frac{e^{w_k^T x_i}}{\sum_l e^{w_l^T x_i}} = I[y_i = k] - P_k(x_i). \tag{43}$$

Since $I[y_i = k]$ can be non-zero (i.e. 1) for only one k, and $\sum_k P_k(x_i) = 1$ with $0 \le P_k(x_i) \le 1$, we have

$$\sum_k \Delta_k^2 = \sum_k (I_k - P_k)^2 \le \sum_k (I_k^2 + P_k^2) \le 2, \tag{44}$$

Let $\Delta(x_i, y_i, w) = [\Delta_1(x_i, y_i, w), ..., \Delta_K(x_i, y_i, w)]$ be a $K \times 1$ vector (which depends on x_i, y_i, w.) The gradient of the risk difference $g(w)$ is then

$$\nabla g(w) = -\frac{1}{N}\sum_i \Delta(x_i, y_i, w) \otimes x_i - \Delta(x'_i, y'_i, w) \otimes x'_i, \tag{45}$$

where $\otimes$ is a Kronecker product of two vectors. Note that

$$\|\Delta \otimes x\|^2 = \sum_k \|\Delta_k x\|^2 \le \|x\|^2 \sum_k \Delta_k^2 \le 2\|x\|^2. \tag{46}$$

Without loss of generality, we assume that only (x_1, y_1) and (x'_1, y'_1) are possibly different and $(x_i, y_i) = (x'_i, y'_i)$ for all $i = 2, ..., N$. In this case we have

$$\begin{aligned} \|\nabla g(w)\| &\le \frac{1}{N}\|\Delta(x_1, y_1, w) \otimes x_1\| \\ &\quad + \frac{1}{N}\|\Delta(x'_1, y'_1, w) \otimes x'_1\| \\ &\le \frac{\sqrt{2}}{N}(\|x_1\| + \|x'_1\|) \le \frac{2\sqrt{2}}{N}, \end{aligned} \tag{47}$$

and the therefore the L_2 sensitive of the minimizer of a multiclass logistic regression is

$$\frac{2\sqrt{2}}{N\lambda} \tag{48}$$

from Corollaries 7 and 8 (Chaudhuri et al., 2011). Note that the sensitivity does not depend on the number of classes K.

B.2. Majority-voted ERM

Let $S = \{(x_i, v_i)\}$ and $S' = \{(x_i, v'_i)\}$ be two datasets with the same features but with possibly different labels for all $i = 1, ..., N$. Then the partial gradient of the risk difference $g(w)$ is

$$\begin{aligned} \nabla_{w_k} g(w) &= -\frac{1}{N}\sum_i x_i [I[v_i = k] - I[v'_i = k]] \\ &= -\frac{1}{N}\sum_i x_i a_k(v_i, v'_i), \end{aligned} \tag{49}$$

where $a_k(v_i, v'_i)$ is

$$a_k(v_i, v'_i) = I[v_i = k] - I[v'_i = k] \in \{-1, 0, 1\}. \tag{50}$$

Let $a = [a_1, ..., a_K]$ be a $K \times 1$ vector (which depends on v_i, v'_i.) Note that at most two elements of a can be nonzero (i.e. ± 1.) The gradient can be rewritten using the Kronecker product $\otimes$ as

$$\nabla g(w) = -\frac{1}{N}\sum_i a(v_i, v'_i) \otimes x_i, \tag{51}$$

and its norm is bounded by

$$\|\nabla g(w)\| \le \frac{1}{N}\sum_i \sqrt{2}\|x_i\| \le \sqrt{2}. \tag{52}$$

Therefore the L_2 sensitivity of the minimizer of majority-labeled multiclass logistic regression is

$$\frac{\sqrt{2}}{\lambda}. \tag{53}$$

B.3. Weighted ERM

A natural multiclass extension of the weighted loss (14) is

$$l^\alpha(w) = \sum_k \alpha^k(x) l(w_k^T x), \tag{54}$$

where $\alpha^k(x)$ is the unbiased estimate of the probability $P(v = k|x)$. The corresponding weighted regularized empirical risk is

$$R_S^\lambda(w) = \frac{1}{N}\sum_i \sum_k \alpha^k(x_i) l(w_k^T x) + \frac{\lambda}{2}\|w\|^2$$

$$
\begin{aligned}
&= \frac{1}{N}\sum_i\sum_k \alpha^k(x_i)[\log(\sum_l e^{w_l^T x_i}) - w_k^T x_i] \\
&\quad + \frac{\lambda}{2}\|w\|^2 \\
&= -\frac{1}{N}\sum_i[\sum_k \alpha^k(x_i) w_k^T x_i - \log(\sum_l e^{w_l^T x_i})] \\
&\quad + \frac{\lambda}{2}\|w\|^2, \qquad (55)
\end{aligned}
$$

and its partial gradient is

$$
\nabla_{w_k} R_S^\lambda(w) = -\frac{1}{N}\sum_i x_i \left[\alpha^k(x_i) - \frac{e^{w_k^T x_i}}{\sum_l e^{w_l^T x_i}}\right] + \lambda w_k. \qquad (56)
$$

Let $S = \{(x_i, \alpha_i)\}$ and $S' = \{(x_i, \alpha'_i)\}$ be two datasets with the same features but with possibly different labels for all $i = 1, ..., N$. Then the partial gradient of the risk difference $g(w)$ is

$$
\begin{aligned}
\nabla_{w_k} g(w) &= -\frac{1}{N}\sum_i x_i[\alpha^k(x_i) - (\alpha')^k(x_i)] \\
&= -\frac{1}{N}\sum_i x_i b_k(\alpha_i^k, (\alpha')_i^k), \qquad (57)
\end{aligned}
$$

where $b_k(\alpha_i^k, (\alpha')_i^k) = \alpha^k(x_i) - (\alpha')^k(x_i)$. Let $b = [b_1, ..., b_K]$ be a $K \times 1$ vector (which depends α_i, α'_i.) Note that at most two elements of b can be nonzero (i.e., $\pm 1/M$.) The gradient can then be rewritten as

$$
\nabla g(w) = -\frac{1}{N}\sum_i b(\alpha_i, \alpha'_i) \otimes x_i, \qquad (58)
$$

and its norm is bounded by

$$
\|\nabla g(w)\| \leq \frac{1}{N}\sum_i \frac{\sqrt{2}}{M}\|x_i\| \leq \frac{\sqrt{2}}{M}. \qquad (59)
$$

Therefore the L_2 sensitivity of the minimizer of the weighted multiclass logistic regression is

$$
\frac{\sqrt{2}}{M\lambda}. \qquad (60)
$$

B.4. Parameter averaging

For the purposes of comparison, we also derive the sensitivity of parameter averaging (Pathak et al., 2010) for multiclass logistic regression. Let the two neighboring datasets be $W = (w_1, w_2, ..., w_M)$ and $W' = (w'_1, w'_2, ..., w'_M)$, which are collections of parameters from M parties. The corresponding averages for the two sets are $\bar{w} = \frac{1}{M}\sum_i w_i$ and $\bar{w}' = \frac{1}{M}\sum_i w'_i$. Without loss of generality, we assume the parameters w_1 and w'_1 differ only for party 1 and $w_i = w'_i$ for others $i = 2, ..., M$. Since $\|\bar{w} - \bar{w}'\| = \frac{1}{M}\|w_1 - w'_1\|$, the L_2 sensitivity is $1/M$ times the sensitivity of the minimizer of the minimizer of a single classifier, when all training samples of party 1 are allowed to change. Therefore the L_2 sensitivity of the average parameters for multiclass logistic regression is $\frac{2\sqrt{2}}{M\lambda}$.

References

Chaudhuri, K., Monteleoni, C., and Sarwate, A.D. Differentially private empirical risk minimization. *The Journal of Machine Learning Research*, 12:1069–1109, 2011.

Pathak, M., Rane, S., and Raj, B. Multiparty differential privacy via aggregation of locally trained classifiers. In *Advances in Neural Information Processing Systems*, pp. 1876–1884, 2010.

Sridharan, K., Shalev-Shwartz, S., and Srebro, N. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems*, pp. 1545–1552, 2009.