

A. Proof of Dimension Independence for Output Perturbation (Theorem 1)

First, we prove the following lemma, which bounds the excess loss (empirical risk) due parameter vector θ_{priv} compared to $\hat{\theta}$.

Lemma 1. Let $\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(\langle \theta, \mathbf{x}_i \rangle; y_i)$. We have, $\mathbb{E}_{\mathbf{b}} [\mathcal{L}(\theta_{priv}) - \mathcal{L}(\hat{\theta})] = O\left(\frac{(LR_2)^2 \sqrt{\log(1/\delta) + \epsilon}}{\lambda \epsilon}\right)$.

Proof. Now,

$$\begin{aligned} \mathcal{L}(\theta_{priv}) - \mathcal{L}(\hat{\theta}) &= \frac{1}{n} \sum_{i=1}^n (\ell(\langle \theta_{priv}, \mathbf{x}_i \rangle; y_i) \\ &\quad - \ell(\langle \hat{\theta}, \mathbf{x}_i \rangle; y_i)). \end{aligned}$$

By the Lipschitz property of the loss function ℓ , we have

$$\begin{aligned} \mathcal{L}(\theta_{priv}) - \mathcal{L}(\hat{\theta}) &\leq \frac{1}{n} \sum_{i=1}^n L |\langle \theta_{priv} - \hat{\theta}, \mathbf{x}_i \rangle| \\ &\leq \frac{L}{n} \sum_{i=1}^n |\langle \mathbf{b}, \mathbf{x}_i \rangle|. \end{aligned}$$

Notice that, each inner product $\langle \mathbf{b}, \mathbf{x}_i \rangle$ is distributed as $\mathcal{N}(0, \sigma^2 \|\mathbf{x}_i\|_2)$, where $\sigma = \frac{(LR_2) \sqrt{\log(1/\delta) + \epsilon}}{\lambda \epsilon}$. Therefore,

$$\begin{aligned} \mathbb{E}_{\mathbf{b}} [\mathcal{L}(\theta_{priv}) - \mathcal{L}(\hat{\theta})] &\leq \frac{L}{n} \sum_{i=1}^n \mathbb{E}_{\mathbf{b}} [|\langle \mathbf{b}, \mathbf{x}_i \rangle|] \\ &\leq \frac{L\sigma}{n} \sum_{i=1}^n \|\mathbf{x}_i\|_2 \leq LR_2\sigma. \end{aligned}$$

Hence Proved. \square

Now, let $J(\theta) = \mathbb{E}_{(\mathbf{x}, y) \sim Dist} [\ell(\langle \theta, \mathbf{x} \rangle; y)] + \frac{\lambda}{2n} \|\theta\|_2^2$ and $\tilde{J}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(\langle \theta, \mathbf{x}_i \rangle; y_i) + \frac{\lambda}{2n} \|\theta\|_2^2$. Also, let $\theta^* = \arg \min_{\theta \in \mathbb{R}^p} J(\theta)$ and $\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^p} \tilde{J}(\theta)$. Then, using Lemma 1, we have:

$$\mathbb{E}_{\mathbf{b}} [\tilde{J}(\theta_{priv}) - \tilde{J}(\hat{\theta})] \leq O(LR_2\sigma) + \mathbb{E}_{\mathbf{b}} \left[\frac{\lambda \|\theta_{priv}\|_2^2}{2n} \right]. \quad (13)$$

Now, we use the following excess risk theorem by (Shalev-Shwartz et al., 2009).

Theorem 5 (One sided uniform convergence (Shalev-Shwartz et al., 2009)). Let $J(\theta)$, $\tilde{J}(\theta)$, $\hat{\theta}$, λ and the loss function ℓ be defined as above. Then, the following holds $\forall \theta \in \mathbb{R}^p$ (with probability at least $1 - \gamma$):

$$J(\theta) - J(\theta^*) \leq 2 \left(\tilde{J}(\theta) - \tilde{J}(\hat{\theta}) \right) + O\left(\frac{(LR_2)^2 \log(1/\gamma)}{\lambda}\right),$$

where L is the Lipschitz constant of the loss function ℓ , and R_2 is an upper bound on the L_2 -norm of the feature vectors in the training data set.

Let $F(\theta) = \mathbb{E}_{(\mathbf{x}, y) \sim Dist} [\ell(\langle \theta, \mathbf{x} \rangle; y)]$. From Theorem 5 and (13), we have the following with probability at least $2/3$ over the data generating distribution $Dist$:

$$\begin{aligned} \mathbb{E}_{\mathbf{b}} [J(\theta_{priv}) - J(\theta^*)] &\leq O(LR_2\sigma) + \mathbb{E}_{\mathbf{b}} \left[\frac{\lambda \|\theta_{priv}\|_2^2}{2n} \right] \\ &\quad + O\left(\frac{(LR_2)^2}{\lambda}\right). \end{aligned}$$

That is,

$$\mathbb{E}_{\mathbf{b}} [F(\theta_{priv}) - F(\theta^*)] \leq O\left(LR_2\sigma + \frac{(LR_2)^2}{\lambda}\right) + \frac{\lambda}{2n} \|\theta^*\|_2^2.$$

Theorem now follows by using $\sigma = \frac{(LR_2) \sqrt{\log(1/\delta) + \epsilon}}{\lambda \epsilon}$, by setting $\lambda = \frac{LR_2 \sqrt{n}}{\|\theta^*\|_2}$ in the above given bound and by using Markov's inequality.

B. Proofs for Private ERM over Simplex

B.1. Proof of Privacy Guarantee (Theorem 3)

Proof. We first characterize the optimal non-private $\hat{\theta}$ obtained by solving (8). To this end, we form the Lagrangian of (8):

$$\begin{aligned} \mathcal{L}(\theta, \nu) &= \frac{1}{n} \sum_{i=1}^n \ell(\langle \mathbf{x}_i, \theta \rangle; y_i) + \frac{\lambda}{n} \sum_{j=1}^p \theta_j \log(\theta_j) \\ &\quad + \frac{\nu}{n} \left(\sum_i \theta_i - 1 \right) \quad (14) \end{aligned}$$

Now, using optimality conditions:

$$(\hat{\theta}, \nu^*) = \max_{\nu} \min_{\theta \in \Delta} \mathcal{L}(\theta, \nu).$$

By setting the gradient of the Lagrangian to be zero and by using primal feasibility, we get:

$$\begin{aligned} \hat{\theta}_j &= \exp\left(-\frac{\nu^*}{\lambda} - 1 - \frac{1}{\lambda} \sum_i \ell'(\langle \mathbf{x}_i, \hat{\theta} \rangle; y_i) \mathbf{x}_i^j\right), \\ \exp\left(\frac{\nu^*}{\lambda}\right) &= \sum_{r \in [p]} \exp\left(-1 - \frac{1}{\lambda} \sum_{i \in [n]} \ell'(\langle \mathbf{x}_i, \hat{\theta} \rangle; y_i) \mathbf{x}_i^r\right), \end{aligned}$$

where ℓ' is the derivative of ℓ and \mathbf{x}_i^j denotes the j -th coordinate of \mathbf{x}_i .

That is,

$$\hat{\theta}_j = \frac{\exp\left(-\frac{1}{\lambda} \sum_i \ell'(\langle \mathbf{x}_i, \hat{\theta} \rangle; y_i) \mathbf{x}_i^j\right)}{\sum_r \exp\left(-\frac{1}{\lambda} \sum_i \ell'(\langle \mathbf{x}_i, \hat{\theta} \rangle; y_i) \mathbf{x}_i^r\right)}. \quad (15)$$

Similarly, let $\widehat{\theta}'_j$ be the solution to (8) but by using a different data set \mathcal{D}' that differs from $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ in exactly *one* data point. Without loss of generality, we assume that \mathcal{D} and \mathcal{D}' differs only in the first entry (\mathbf{x}'_1, y'_1) .

Now, consider an index a_s that is sampled from the probability distribution $\widehat{\theta}$. Now, probability of sampling $a_s = j$, given that $\widehat{\theta}$ is learned using data set \mathcal{D} is given by: $Pr(a_s = j|\mathcal{D}) = \widehat{\theta}_j$. Similarly, $Pr(a_s = j|\mathcal{D}') = \widehat{\theta}'_j$. Hence,

$$\begin{aligned} \max_j \frac{Pr(a_s = j|\mathcal{D})}{Pr(a_s = j|\mathcal{D}')} &= \max_j \frac{\widehat{\theta}_j}{\widehat{\theta}'_j} \\ &= \max_j \frac{\exp\left(-\frac{1}{\lambda} \sum_i \ell'(\langle \mathbf{x}_i, \widehat{\theta} \rangle; y_i) \mathbf{x}_i^j\right)}{\sum_r \exp\left(-\frac{1}{\lambda} \sum_i \ell'(\langle \mathbf{x}_i, \widehat{\theta} \rangle; y_i) \mathbf{x}_i^r\right)} \\ &\quad \cdot \frac{\sum_r \exp\left(-\frac{1}{\lambda} \ell'(\langle \mathbf{x}'_1, \widehat{\theta}' \rangle; y'_1) \mathbf{x}'_1{}^r - \frac{1}{\lambda} \sum_{i=2}^n \ell'(\langle \mathbf{x}_i, \widehat{\theta}' \rangle; y_i) \mathbf{x}_i^r\right)}{\exp\left(-\frac{1}{\lambda} \ell'(\langle \mathbf{x}'_1, \widehat{\theta}' \rangle; y'_1) \mathbf{x}'_1{}^j - \frac{1}{\lambda} \sum_{i=2}^n \ell'(\langle \mathbf{x}_i, \widehat{\theta}' \rangle; y_i) \mathbf{x}_i^j\right)}. \end{aligned} \quad (16)$$

Now, first consider the following:

$$\begin{aligned} &\frac{\exp\left(-\frac{1}{\lambda} \sum_i \ell'(\langle \mathbf{x}_i, \widehat{\theta} \rangle; y_i) \mathbf{x}_i^j\right)}{\exp\left(-\frac{1}{\lambda} \ell'(\langle \mathbf{x}'_1, \widehat{\theta}' \rangle; y'_1) \mathbf{x}'_1{}^j - \frac{1}{\lambda} \sum_{i=2}^n \ell'(\langle \mathbf{x}_i, \widehat{\theta}' \rangle; y_i) \mathbf{x}_i^j\right)} \\ &= \exp\left(-\frac{1}{\lambda} \ell'(\langle \mathbf{x}_1, \widehat{\theta} \rangle; y_1) \mathbf{x}_1^j + \frac{1}{\lambda} \ell'(\langle \mathbf{x}'_1, \widehat{\theta}' \rangle; y'_1) \mathbf{x}'_1{}^j\right) \\ &\quad + \frac{1}{\lambda} \sum_{i=2}^n \left(\ell'(\langle \mathbf{x}_i, \widehat{\theta} \rangle; y_i) - \ell'(\langle \mathbf{x}_i, \widehat{\theta}' \rangle; y_i)\right) \mathbf{x}_i^j \\ &\leq \exp\left(\frac{2LR_\infty}{\lambda} + \frac{nR_\infty^2 L_g \|\widehat{\theta} - \widehat{\theta}'\|_1}{\lambda}\right) = A, \end{aligned} \quad (17)$$

where the last inequality follows by: a) using Lipschitz continuity of ℓ , i.e., $\ell'(\cdot; \cdot) \leq L$, b) $\|\mathbf{x}_i\|_\infty \leq R_\infty$, c) by using Lipschitz continuity of ℓ' , and d) by applying Holder's inequality $|\langle \mathbf{x}_i, \widehat{\theta} - \widehat{\theta}' \rangle| \leq \|\mathbf{x}_i\|_\infty \|\widehat{\theta} - \widehat{\theta}'\|_1$.

Now, we bound $\|\widehat{\theta} - \widehat{\theta}'\|_1$ using strong convexity of the entropy regularizer w.r.t. L_1 norm. Let $J(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(\langle \mathbf{x}_i, \theta \rangle; y_i) + \frac{\lambda}{n} \sum_{j=1}^p \theta_j \log(\theta_j)$. As $\widehat{\theta}$ is the minimum of (8):

$$\frac{\lambda}{2n} \|\widehat{\theta} - \widehat{\theta}'\|_1^2 + J(\widehat{\theta}|\mathcal{D}) \leq J(\widehat{\theta}'|\mathcal{D}).$$

Similarly, using optimality of $\widehat{\theta}'$ for (8) with data set \mathcal{D}' :

$$\frac{\lambda}{2n} \|\widehat{\theta} - \widehat{\theta}'\|_1^2 + J(\widehat{\theta}'|\mathcal{D}') \leq J(\widehat{\theta}|\mathcal{D}').$$

Adding the above two equations, using the fact that $\mathcal{D} - \mathcal{D}' = (\mathbf{x}_1, y_1)$, by applying the Lipschitz continuity of ℓ , and by using Holder's inequality, we get:

$$\|\widehat{\theta} - \widehat{\theta}'\|_1 \leq \frac{LR_\infty}{\lambda}.$$

Now plugging the above bound in (17), we get:

$$A \leq \exp\left(\frac{2LR_\infty}{\lambda} + \frac{nLR_\infty^3 L_g}{\lambda^2}\right).$$

Using the above equation with (16), we get:

$$\max_j \frac{Pr(a_s = j|\mathcal{D})}{Pr(a_s = j|\mathcal{D}')} \leq \exp\left(\frac{4LR_\infty}{\lambda} + \frac{2nLR_\infty^3 L_g}{\lambda^2}\right). \quad (18)$$

Note that this ensures, that each ‘‘sample’’ a_s is $\epsilon = \exp\left(\frac{4LR_\infty}{\lambda} + \frac{2nLR_\infty^3 L_g}{\lambda^2}\right)$ differentially private. Hence, ϵ and (ϵ, δ) differential privacy for the computation of the collection of m samples $\{a_1, a_2, \dots, a_m\}$ and consequently θ_{priv} follows by using the *weak* and the *strong composition* theorems of (Dwork et al., 2006b; 2010c) respectively. \square

B.2. Proof Utility Guarantee (Theorem 4)

We first prove in Lemma 2 the excess risk bound of Algorithm (8) for any choice of m and λ . We then set $m = \left(\frac{\epsilon\lambda}{\log(1/\delta)}\right)^2 \left(32 + \frac{16nR_\infty^2 L_g}{\lambda}\right)^{-2}$ and $\lambda = \frac{n^{2/3}}{\epsilon^{1/3} \log^{1/3} p}$ to get the final guarantee.

Lemma 2. *Let L, L_g be as defined in Theorem 3. With probability at least $2/3$ over the randomness of $Dist$ and the randomness of θ_{priv} , the following is true.*

$$\begin{aligned} &\mathbb{E}_{(\mathbf{x}, y) \sim Dist} [\ell(\langle \theta_{priv}, \mathbf{x} \rangle; y) - \ell(\langle \theta^*, \mathbf{x} \rangle; y)] = \\ &O\left(\frac{LR_\infty \log m}{\sqrt{m}} + \frac{\lambda}{n} \log p + \frac{(LR_\infty)^2}{\lambda}\right). \end{aligned}$$

Here $\theta^* = \arg \min_{\theta \in \Delta} \mathbb{E}_{(\mathbf{x}, y) \sim Dist} [\ell(\langle \theta, \mathbf{x} \rangle; y)]$.

Proof. Recall that,

$$\theta_{priv} = \frac{1}{m} \sum_s e_{a_s},$$

where e_{a_s} is the a_s -th canonical basis vector and $a_s \in \{1, 2, \dots, p\}, \forall s \in [m]$ are sampled i.i.d. according to the probability distribution $\widehat{\theta}$.

Now, for any fixed \mathbf{x} : $\langle \mathbf{x}, \theta_{priv} \rangle = \frac{1}{m} \sum_s \langle \mathbf{x}, e_{a_s} \rangle$. Note that, $\mathbb{E}_{a_s} [\langle \mathbf{x}, e_{a_s} \rangle] = \langle \mathbf{x}, \widehat{\theta} \rangle$. Therefore,

$$\mathbb{E}_{\theta_{priv}} [\langle \mathbf{x}, \theta_{priv} \rangle] = \mathbb{E}_{a_s} [\langle \mathbf{x}, e_{a_s} \rangle] = \langle \mathbf{x}, \widehat{\theta} \rangle$$

Furthermore, $|\langle \mathbf{x}, \mathbf{e}_{a_s} \rangle| \leq \|\mathbf{x}\|_\infty = R_\infty$. Therefore by Hoeffding's inequality, with probability at least $1 - \gamma$,

$$|\langle \mathbf{x}, \boldsymbol{\theta}_{priv} \rangle - \langle \mathbf{x}, \hat{\boldsymbol{\theta}} \rangle| = O\left(\frac{R_\infty \log(1/\gamma)}{\sqrt{m}}\right).$$

Observing $|\langle \mathbf{x}, \boldsymbol{\theta}_{priv} \rangle - \langle \mathbf{x}, \hat{\boldsymbol{\theta}} \rangle|$ is universally bounded by R_∞ , and setting $\gamma = \frac{1}{\sqrt{m}}$, we have

$$\mathbb{E}_{\boldsymbol{\theta}_{priv}} \left[|\langle \mathbf{x}, \boldsymbol{\theta}_{priv} \rangle - \langle \mathbf{x}, \hat{\boldsymbol{\theta}} \rangle| \right] = O\left(\frac{R_\infty \log m}{\sqrt{m}}\right).$$

Now,

$$\begin{aligned} & \mathbb{E}_{\boldsymbol{\theta}_{priv}} \left[\mathbb{E}_{\mathbf{x} \sim Dist} \left[|\langle \mathbf{x}, \boldsymbol{\theta}_{priv} \rangle - \langle \mathbf{x}, \hat{\boldsymbol{\theta}} \rangle| \right] \right] = \\ & \mathbb{E}_{\mathbf{x} \sim Dist} \left[\mathbb{E}_{\boldsymbol{\theta}_{priv}} \left[|\langle \mathbf{x}, \boldsymbol{\theta}_{priv} \rangle - \langle \mathbf{x}, \hat{\boldsymbol{\theta}} \rangle| \right] \right] \\ & \leq \max_{\mathbf{x} \in \mathcal{X}} \mathbb{E}_{\boldsymbol{\theta}_{priv}} \left[|\langle \mathbf{x}, \boldsymbol{\theta}_{priv} \rangle - \langle \mathbf{x}, \hat{\boldsymbol{\theta}} \rangle| \right] = O\left(\frac{R_\infty \log m}{\sqrt{m}}\right) \end{aligned}$$

Therefore, with probability at least 9/10 over the randomness of $\boldsymbol{\theta}_{priv}$, we have

$$\mathbb{E}_{(\mathbf{x}, y) \sim Dist} [\ell(\langle \mathbf{x}, \boldsymbol{\theta}_{priv} \rangle; y) - \ell(\langle \mathbf{x}, \hat{\boldsymbol{\theta}} \rangle; y)] = O\left(\frac{LR_\infty \log m}{\sqrt{m}}\right).$$

Now, using standard uniform convergence bound of (Shalev-Shwartz et al., 2009; Kakade et al., 2008), we get:

$$\begin{aligned} & \mathbb{E}_{(\mathbf{x}, y) \sim Dist} [\ell(\langle \mathbf{x}, \hat{\boldsymbol{\theta}} \rangle; y) - \ell(\langle \mathbf{x}, \boldsymbol{\theta}^* \rangle; y)] = \\ & O\left(\frac{LR_\infty \log m}{\sqrt{m}} + \frac{\lambda}{n} \log p + \frac{(LR_\infty)^2}{\lambda}\right). \end{aligned}$$

□