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 $\boldsymbol{\theta}_{priv}$ compared to $\widehat{\boldsymbol{\theta}}$.

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

Proof. Now,

$$\mathcal{L}(\boldsymbol{\theta}_{priv}) - \mathcal{L}(\widehat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\Big(\ell(\langle\boldsymbol{\theta}_{priv}, \boldsymbol{x}_i\rangle; y_i) - \ell(\langle\widehat{\boldsymbol{\theta}}, \boldsymbol{x}_i\rangle; y_i)\Big).$$

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

$$\mathcal{L}(\boldsymbol{\theta}_{priv}) - \mathcal{L}(\widehat{\boldsymbol{\theta}}) \leq \frac{1}{n}\sum_{i=1}^{n} L|\langle\boldsymbol{\theta}_{priv} - \widehat{\boldsymbol{\theta}}, \boldsymbol{x}_i\rangle| \leq \frac{L}{n}\sum_{i=1}^{n}|\langle\boldsymbol{b}, \boldsymbol{x}_i\rangle|.$$

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

$$\mathop{\mathbb{E}}_{\boldsymbol{b}}\left[\mathcal{L}(\boldsymbol{\theta}_{priv}) - \mathcal{L}(\widehat{\boldsymbol{\theta}})\right] \leq \frac{L}{n}\sum_{i=1}^{n} E_{\boldsymbol{b}}\left[|\langle\boldsymbol{b}, \boldsymbol{x}_i\rangle|\right] \leq \frac{L\sigma}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i\|_2 \leq LR_2\sigma.$$

Hence Proved. ☐

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

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

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 $J(\boldsymbol{\theta})$, $\tilde{J}(\boldsymbol{\theta})$, $\widehat{\boldsymbol{\theta}}$, λ and the loss function ℓ be defined as above. Then, the following holds $\forall\boldsymbol{\theta}\in\mathbb{R}^p$ (with probability at least $1-\gamma$):*

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

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

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

That is,

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

Theorem now follows by using $\sigma = \frac{(LR_2)\sqrt{\log(1/\delta)+\epsilon}}{\lambda\epsilon}$, by setting $\lambda = \frac{LR_2\sqrt{n}}{\|\boldsymbol{\theta}^*\|_2}$ 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 $\widehat{\boldsymbol{\theta}}$ obtained by solving (8). To this end, we form the Lagrangian of (8):

$$\mathcal{L}(\boldsymbol{\theta}, \nu) = \frac{1}{n}\sum_{i=1}^{n}\ell(\langle\boldsymbol{x}_i, \boldsymbol{\theta}\rangle; y_i) + \frac{\lambda}{n}\sum_{j=1}^{p}\theta_j\log(\theta_j) + \frac{\nu}{n}\Big(\sum_i\theta_i - 1\Big) \tag{14}$$

Now, using optimality conditions:

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

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

$$\widehat{\theta}_j = \exp\left(-\frac{\nu^*}{\lambda} - 1 - \frac{1}{\lambda}\sum_i\ell'(\langle\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}\rangle; y_i)\boldsymbol{x}_i^j\right),$$

$$\exp\left(\frac{\nu^*}{\lambda}\right) = \sum_{r\in[p]}\exp\left(-1 - \frac{1}{\lambda}\sum_{i\in[n]}\ell'(\langle\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}\rangle; y_i)\boldsymbol{x}_i^r\right),$$

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

That is,

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

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

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

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

Now, first consider the following:

$$\frac{\exp\left(-\frac{1}{\lambda}\sum_i \ell'(\langle \boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}\rangle; y_i)\boldsymbol{x}_i^j\right)}{\exp\left(-\frac{1}{\lambda}\ell'(\langle \boldsymbol{x}'_1, \widehat{\boldsymbol{\theta}}'\rangle; y'_1)\boldsymbol{x}_1'^j - \frac{1}{\lambda}\sum_{i=2}^n \ell'(\langle \boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}'\rangle; y_i)\boldsymbol{x}_i^j\right)}$$
$$= \exp\Bigg(-\frac{1}{\lambda}\ell'(\langle \boldsymbol{x}_1, \widehat{\boldsymbol{\theta}}\rangle; y_1)\boldsymbol{x}_1^j + \frac{1}{\lambda}\ell'(\langle \boldsymbol{x}'_1, \widehat{\boldsymbol{\theta}}'\rangle; y'_1)\boldsymbol{x}_1'^j$$
$$+\frac{1}{\lambda}\sum_{i=2}^n \left(\ell'(\langle \boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}\rangle; y_i) - \ell'(\langle \boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}'\rangle; y_i)\right)\boldsymbol{x}_i^j\Bigg)$$
$$\leq \exp\left(\frac{2LR_\infty}{\lambda} + \frac{n{R_\infty}^2 L_g\|\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}'\|_1}{\lambda}\right) = A, \tag{17}$$

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

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

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

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

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

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

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

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

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

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

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

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

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 $m = \left(\frac{\epsilon\lambda}{\log(1/\delta)}\right)^2 \left(32 + \frac{16n{R_\infty}^2}{\lambda}L_g\right)^{-2}$ 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 $\boldsymbol{\theta}_{priv}$, the following is true.*

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

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

Proof. Recall that,

$$\boldsymbol{\theta}_{priv} = \frac{1}{m}\sum_s \boldsymbol{e}_{a_s},$$

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

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

$$\mathbb{E}_{\boldsymbol{\theta}_{priv}}\left[\langle \boldsymbol{x}, \boldsymbol{\theta}_{priv}\rangle\right] = \mathbb{E}_{a_s}[\langle \boldsymbol{x}, \boldsymbol{e}_{a_s}\rangle] = \langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}\rangle$$

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

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

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

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

Now,

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

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

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

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

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

□