6 Appendix

6.1 An Example to Show Training Stability is not a Direct Consequence of Differential Privacy

We now present an example to illustrate that training stability is a property of the training algorithm and not a direct consequence of differential privacy. We present a problem and two α-differentially private training algorithms which approximately optimize the same function; the first algorithm is based on exponential mechanism, and the second on a maximum of Laplace random variables mechanism. We show that while both provide α-differential privacy guarantees, the first algorithm does not satisfy training stability while the second one does.

Let $i \in \{1, \ldots, l\}$, and let $f : \mathcal{X}^n \times \mathbf{R} \to [0, 1]$ be a function such that for all i and all datasets D and D' of size n that differ in the value of a single individual, $|f(D, i) - f(D', i)| \leq \frac{1}{n}$.

Consider the following training and validation problem. Given a sensitive dataset D, the private training procedure A outputs a tuple $(i^*, t_1, \ldots, t_l)$, where i^* is the output of the $\alpha/2$-differentially private exponential mechanism [20] run to approximately maximize $f(D, i)$, and each t_i is equal to $f(D, i)$ plus an independent Laplace random variable with standard deviation $\frac{2l}{\alpha n}$. For any validation dataset V, the validation score $q((i^*, t_1, \ldots, t_l), V) = t_{i^*}$.

It follows from standard results that A is α-differentially private. Moreover, A can be represented by a tuple $\mathcal{T}_A = (\mathcal{G}_A, F_A)$, where $\mathcal{G}_A$ is the following density over sequences of real numbers of length $l + 1$:

$$\mathcal{G}_A(r_0, r_1, \ldots, r_l) = \mathbf{1}_{0 \leq r_0 \leq 1} \cdot \frac{1}{2^l} e^{-(|r_1| + |r_2| + \ldots + |r_l|)}$$

Thus $\mathcal{G}_A$ is the product of the uniform density on $[0, 1]$ and l standard Laplace densities. Consider the following map E_0. For $r \in [0, 1]$, let

$$E_0(r) = i, \quad \text{if } \frac{\sum_{j<i} e^{n\alpha f(D,j)/4}}{\sum_j e^{n\alpha f(D,j)/4}} \leq r \leq \frac{\sum_{j \leq i} e^{n\alpha f(D,j)/4}}{\sum_j e^{n\alpha f(D,j)/4}}$$

In other words, $E_0(r)$ is the map that converts a random number r drawn from the uniform distribution on $[0, 1]$ to the $\alpha/2$-differentially private exponential mechanism distribution that approximately maximizes $f(D, i)$. Given a $l + 1$-tuple $R = (R_0, R_1, \ldots, R_l)$, F_A is now the following map:

$$F_A(D, \alpha, R) = \left(E(R_0), f(D, 1) + \frac{2lR_1}{\alpha n}, f(D, 2) + \frac{2lR_2}{\alpha n}, \ldots, f(D, l) + \frac{2lR_l}{\alpha n} \right)$$

Let $l = 2$ and D and D' be two datasets that differ in the value of a single individual. Suppose it is the case that $f(D, 1) = 1$, $f(D, 2) = \frac{1}{2}$ and $f(D', 1) = 1 - \frac{1}{n}$, $f(D', 2) = \frac{1}{2} + \frac{1}{n}$. Observe that for D, the exponential mechanism picks 1 with probability $\frac{e^{n\alpha/4}}{e^{n\alpha/4} + e^{n\alpha/8}}$, and 2 with probability $\frac{e^{n\alpha/8}}{e^{n\alpha/4} + e^{n\alpha/8}}$, where as for D', it picks 1 with probability $\frac{e^{(n-1)\alpha/4}}{e^{(n-1)\alpha/4} + e^{(n+2)\alpha/8}}$ and 2 with probability $\frac{e^{(n+2)\alpha/8}}{e^{(n-1)\alpha/4} + e^{(n+2)\alpha/8}}$. Thus, if R_0 lies in the interval $[\frac{e^{(n-1)\alpha/4}}{e^{(n-1)\alpha/4} + e^{(n+2)\alpha/8}}, \frac{e^{n\alpha/4}}{e^{n\alpha/4} + e^{n\alpha/8}}]$, then, $F_A(D, \alpha, R) = t_1$ whereas $F_A(D', \alpha, R) = t_2$. When n is large enough, with high probability, $|t_1 - t_2| \geq \frac{1}{3}$; thus, the training stability condition does not hold for A for $\beta_1 = o(n)$ and $\delta < \frac{e^{n\alpha/8}(e^{\alpha/2} - 1)}{(e^{n\alpha/8} + 1)(e^{n\alpha/8} + e^{\alpha/2})}$.

Consider a different algorithm A' which computes $t_1, \ldots, t_l$ first, and then outputs the index i^* that maximizes t_{i^*}. Then A' can be represented by a tuple $\mathcal{T}_{A'} = (\mathcal{G}_{A'}, F_{A'})$, where $G_{A'}$ is a density over sequences of real numbers of length l as follows:

$$\mathcal{G}_A(r_1, \ldots, r_l) = \frac{1}{2^l} e^{-(|r_1| + \ldots + |r_l|)}$$

and $F_{A'}$ is the map:

$$F_{A'}(D, \alpha, R) = \left(\text{argmax}_i (f(D, i) + \frac{lR_i}{\alpha n}), f(D, 1) + \frac{lR_1}{\alpha n}, f(D, 2) + \frac{lR_2}{\alpha n}, \ldots, f(D, l) + \frac{lR_l}{\alpha n} \right)$$

For the same value of $R_1, \ldots, R_l$, if $i^* = i$ on input dataset D and if $i^* = i'$ on input dataset D', then, $|f(D,i) - f(D,i')| \leq \frac{1}{n}$; this implies that

$$|q(F_{A'}(D,\alpha,R),V) - q(F_{A'}(D',\alpha,R),V)| = |t_i - t_{i'}| = |f(D,i) - f(D',i')| \leq \frac{1}{n}$$

with probability 1 over $\mathcal{G}_{A'}$. Thus the training stability condition holds for $\beta_1 = 1$ and $\delta = 0$.

6.2 Output Perturbation Algorithm

We present the output perturbation algorithm for regularized linear classification.

Algorithm 4 Output Perturbation for Differentially Private Linear Classification

1: **Inputs:** Regularization parameter λ, training set $T = \{(x_i, y_i), i = 1, \ldots, n\}$, privacy parameter α.
2: Let $\mathcal{G}$ be the following density over $\mathbf{R}^d$: $\rho_{\mathcal{G}}(r) \propto e^{-\|r\|}$. Draw $R \sim \mathcal{G}$.
3: Solve the convex optimization problem:

$$w^* = \text{argmin}_{w \in \mathbf{R}^d} \frac{1}{2}\lambda\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} \ell(w, x_i, y_i) \tag{4}$$

4: Output $w^* + \frac{2}{\lambda\alpha n}R$.

6.3 Case Study: Histogram Density Estimation

Our second case study is developing an end-to-end differentially private solution for histogram-based density estimation. In density estimation, we are given n samples $x_1, \ldots, x_n$ drawn from an unknown density f, and our goal is to build an approximation $\hat{f}$ to f. In a histogram density estimator, we divide the range of the data into equal-sized bins of width h; if n_i out of n of the input samples lie in bin i, then $\hat{f}$ is the density function: $\hat{f}(x) = \sum_{i=1}^{1/h} \frac{n_i}{hn} \cdot \mathbf{1}(x \in \text{Bin } i)$.

A critical parameter while constructing the histogram density estimator is the bin size h. There is much theoretical literature on how to choose h – see [15, 24] for surveys. However, the choice of h is usually data-dependent, and in practice, the optimal h is often determined by building a histogram density estimator for a few different values of h, and selecting the one which has the best performance on held-out validation data.

The most popular measure to evaluate the quality of a density estimator is the L_2-distance or the Integrated Square Error (ISE) between the density estimate and the true density:

$$\|\hat{f} - f\|_2 = \int_x (\hat{f}(x) - f(x))^2 \mathrm{d}x = \int_x f^2(x)\mathrm{d}x + \int_x \hat{f}^2(x)\mathrm{d}x - 2\int_x f(x)\hat{f}(x)\mathrm{d}x \tag{5}$$

f is typically unknown, so the ISE cannot be computed exactly. Fortunately it is still possible to *compare* multiple density estimates based on this distance. The first term in the right hand side of Equation 5 depends only on f, and is equal for all $\hat{f}$. The second term is a function of $\hat{f}$ only and can thus be computed. The third term is $2\mathbb{E}_{x\sim f}[\hat{f}(x)]$, and even though it cannot be computed exactly without knowledge of f, we can estimate it based on a held out validation dataset. Thus, given a density estimator $\hat{f}$ and a validation dataset $V = \{z_1, \ldots, z_m\}$, we will use the following function to evaluate the quality of $\hat{f}$ on V:

$$q(\hat{f}, V) = -\int_x \hat{f}^2(x)\mathrm{d}x + \frac{2}{m}\sum_{i=1}^{m} \hat{f}(z_i) \tag{6}$$

A higher value of q indicates a smaller distance $\|\hat{f} - f\|^2$, and thus a higher quality density estimate. For other measures, see [5].

In the sequel, we assume that the data lies in the interval $[0, 1]$ and that this interval is known in advance. For ease of notation, we also assume without loss of generality that $\frac{1}{h}$ is an integer. For

ease of exposition, we confine ourselves to one-dimensional data, although the general techniques can be easily extended to higher dimensions. Given n samples and a bin size h, several works, including [6, 18, 25, 26, 19, 27, 13] have shown different ways of constructing and sampling from differentially private histograms. The most basic approach is to construct a non-private histogram and then add Laplace noise to each cell, followed by some post-processing. Algorithm 5 presents a variant of a differentially private histogram density estimator due to [18] in our framework.

Algorithm 5 Differentially Private Histogram Density Estimator

1: **Inputs:** Bin size h (such that $1/h$ is an integer), data $T = \{x_1, \ldots, x_n\}$, privacy parameter α.
2: **for** $i = 1, \ldots, \frac{1}{h}$ **do**
3: Draw R_i independently from the standard Laplace density: $\rho_{\mathcal{G}}(r) = \frac{1}{2}e^{-|r|}$.
4: Let $I_i = \left[\frac{i-1}{h}, \frac{i}{h}\right)$. Define: $n_i = \sum_{j=1}^{n} \mathbf{1}(x_j \in I_i)$, and let $\tilde{n}_i = \max\left(0, n_i + \frac{2R_i}{\alpha}\right)$.
5: **end for**
6: Let $\tilde{n} = \sum_i \tilde{n}_i$. Return the density estimator: $\hat{f}(x) = \sum_{i=1}^{1/h} \frac{\tilde{n}_i}{h\tilde{n}} \cdot \mathbf{1}(x \in I_i)$

The following theorem shows stability guarantees on the differentially private histogram density estimator described in Algorithm 5.

Theorem 6 (Stability of Private Histogram Density Estimator) *Let $H = \{h_1, \ldots, h_k\}$ be a set of bin sizes, and let $h_{\min} = \min_i h_i$. For any fixed δ, if the sample size $n \geq 1 + \frac{2\ln(4k/\delta)}{\alpha\sqrt{h_{\min}}}$, then, the validation score q in Equation 6 is $(\beta_1, \beta_2, \frac{\delta}{k})$-Stable with respect to Algorithm 5 and H for: $\beta_1 = \frac{6}{(1-\nu)h_{\min}}$, $\beta_2 = \frac{2}{h_{\min}}$, where: $\nu = \frac{2\ln(4k/\delta)}{n\alpha\sqrt{h_{\min}}}$.*

6.4 Proofs of Theorems 1, 2 and 3

We now present the proofs of Theorems 1, 2 and 3. Our proofs involve ideas similar to those in the analysis of the multiplicative weights update method for answering a set of linear queries in a differentially private manner [12].

Let $\mathcal{A}(D)$ denote the output of Algorithm 1 when the input is a sensitive dataset $D = (T, V)$, where T is the training part and V is the validation part. Let $D' = (T', V)$ where T and T' differ in the value of a single individual, and let $D'' = (T, V')$ where V and V' differ in the value of a single individual. The proof of Theorem 1 is a consequence of the following two lemmas.

Lemma 1 *Suppose that the conditions in Theorem 1 hold. Then, for all $D = (T, V)$, all $D' = (T', V)$, such that T and T' differ in the value of a single individual, and for any set of outcomes S:*

$$\Pr(\mathcal{A}(D) \in S) \leq e^{\alpha_2} \Pr(\mathcal{A}(D') \in S) + \delta \tag{7}$$

Lemma 2 *Suppose that the conditions in Theorem 1 hold. Then, for all $D = (T, V)$, all $D'' = (T, V')$ such that V and V' differ in the value of a single individual, and for any set of outcomes S,*

$$\Pr(\mathcal{A}(D) \in S) \leq e^{\alpha_2} \Pr(\mathcal{A}(D'') \in S) + \delta \tag{8}$$

PROOF: (Of Lemma 1) Let $S = (I, C)$, where $I \subseteq [k]$ is a set of indices and $C \subseteq \mathcal{C}$. Let E be the event that all of $R_1, \ldots, R_k$ lie in the set Σ. We will first show that conditioned on E, for all i, it holds that:

$$\Pr(i^* = i | D, E) \leq e^{\alpha_2} \Pr(i^* = i | D', E) \tag{9}$$

Since $\Pr(E) \geq 1 - \delta$, from the conditions in Theorem 1, for any subset I of indices, we can write:

$$\begin{aligned} \Pr(i^* \in I | D) &\leq \Pr(i^* \in I | D, E) \Pr(E) + (1 - \Pr(E)) \\ &\leq e^{\alpha_2} \Pr(i^* \in I | D', E) \Pr(E) + \delta \\ &\leq e^{\alpha_2} \Pr(i^* \in I, E | D') + \delta \\ &\leq e^{\alpha_2} \Pr(i^* \in I | D') + \delta \end{aligned} \tag{10}$$

We will now prove Equation 9. For this purpose, we adopt the following notation. We use the notation $Z_{\setminus i}$ to denote the random variables $Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_k$ and $z_{\setminus i}$ to denote the set of values $z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_k$. We also use the notation $h(\cdot)$ to represent the density induced on the random variables $Z_1, \ldots, Z_k$ by Algorithm 1. In addition, we use the notation R to denote the vector $(R_1, \ldots, R_k)$. We first fix a value $z_{\setminus i}$ for $Z_{\setminus i}$, and a value of R such that $R_1, \ldots, R_k$ all lie in Σ, and consider the ratio of probabilities:

$$\frac{\Pr(i^* = i | Z_{\setminus i} = z_{\setminus i}, D, R)}{\Pr(i^* = i | Z_{\setminus i} = z_{\setminus i}, D', R)}$$

Observe that this ratio of probabilities is equal to:

$$\frac{\Pr(Z_i + q(F(T, \theta_i, \alpha_1, R_i), V) \geq \sup_{j \neq i} z_j + q(F(T, \theta_j, \alpha_1, R_j), V))}{\Pr(Z_i + q(F(T', \theta_i, \alpha_1, R_i), V) \geq \sup_{j \neq i} z_j + q(F(T', \theta_j, \alpha_1, R_j), V))}$$

which is in turn equal to:

$$\frac{\Pr(Z_i \geq \sup_{j \neq i} z_j + q(F(T, \theta_j, \alpha_1, R_j), V) - q(F(T, \theta_i, \alpha_1, R_i), V))}{\Pr(Z_i \geq \sup_{j \neq i} z_j + q(F(T', \theta_j, \alpha_1, R_j), V) - q(F(T', \theta_i, \alpha_1, R_i), V))}$$

Observe that from the stability condition,

$$\begin{aligned} & |(q(F(T, \theta_j, \alpha_1, R_j), V) - q(F(T, \theta_i, \alpha_1, R_i), V)) - (q(F(T', \theta_j, \alpha_1, R_j), V) - q(F(T', \theta_i, \alpha_1, R_i), V))| \\ \leq \; & |q(F(T, \theta_j, \alpha_1, R_j), V) - q(F(T', \theta_j, \alpha_1, R_j), V')| + |q(F(T, \theta_i, \alpha_1, R_i), V) - q(F(T', \theta_i, \alpha_1, R_i), V)| \\ \leq \; & \frac{2\beta_1}{n} \leq 2\beta \end{aligned}$$

Thus, the ratio of the probabilities is at most the ratio $\Pr(Z_i \geq \gamma) / \Pr(Z_i \geq \gamma + 2\beta)$ where $\gamma = \sup_{j \neq i} z_j + q(F(T, \theta_j, \alpha_1, R_j), V) - q(F(T, \theta_i, \alpha_1, R_i), V)$, which is at most e^{α_2} by properties of the exponential distribution. Thus, we have established that for all $z_{\setminus i}$, for all R in Σ^k,

$$\Pr(i^* = i | Z_{\setminus i} = z_{\setminus i}, D, R) \leq e^{\alpha_2} \cdot \Pr(i^* = i | Z_{\setminus i} = z_{\setminus i}, D', R)$$

Equation 9 follows by integrating over $z_{\setminus i}$ and R. The lemma follows. □

PROOF:(Of Lemma 2) Let $S = (I, C)$, where $I \subseteq [k]$ is a set of indices and $C \subseteq \mathcal{C}$. Let E be the event that all of $R_1, \ldots, R_k$ lie in Σ. We will first show that conditioned on E, for all i, it holds that:

$$\Pr(i^* = i | D, E) \leq e^{\alpha_2} \Pr(i^* = i | D'', E) \tag{11}$$

Since $\Pr(E) \geq 1 - \delta$, from the conditions in Theorem 1, for any subset I of indices, we can write:

$$\begin{aligned} \Pr(i^* \in I | D) &\leq \Pr(i^* \in I | D, E) \Pr(E) + (1 - \Pr(E)) \\ &\leq e^{\alpha_2} \Pr(i^* \in I | D'', E) \Pr(E) + \delta \\ &\leq e^{\alpha_2} \Pr(i^* \in I, E | D'') + \delta \\ &\leq e^{\alpha_2} \Pr(i^* \in I | D'') + \delta \end{aligned} \tag{12}$$

We will now focus on showing Equation 11. We first consider the case when event E holds, that is, $R_j \in R$, for $j = 1, \ldots, k$. In this case, the stability definition and the conditions of the theorem imply that for all $\theta_j \in \Theta$,

$$|q(F(T, \theta_j, \alpha_1, R_j), V) - q(F(T, \theta_j, \alpha_1, R_j), V')| \leq \frac{\beta_2}{m} \leq \beta \tag{13}$$

In what follows, we use the notation $Z_{\setminus i}$ to denote the random variables $Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_k$ and $z_{\setminus i}$ to denote the set of values $z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_k$. We also use the notation $h(\cdot)$ to represent the density induced on the random variables $Z_1, \ldots, Z_k$ by Algorithm 1. In addition, we use the notation R to denote the vector $(R_1, \ldots, R_k)$. We first fix a value $z_{\setminus i}$ for $Z_{\setminus i}$, and a value of R such that E holds, and consider the ratio of probabilities:

$$\frac{\Pr(i^* = i | Z_{\setminus i} = z_{\setminus i}, D, R)}{\Pr(i^* = i | Z_{\setminus i} = z_{\setminus i}, D'', R)}$$

Observe that this ratio of probabilities is equal to:

$$\frac{\Pr(Z_i + q(F(T,\theta_i,\alpha_1,R_i),V) \geq \sup_{j\neq i} z_j + q(F(T,\theta_j,\alpha_1,R_j),V))}{\Pr(Z_i + q(F(T,\theta_i,\alpha_1,R_i),V') \geq \sup_{j\neq i} z_j + q(F(T,\theta_j,\alpha_1,R_j),V'))}$$

which is in turn equal to:

$$\frac{\Pr(Z_i \geq \sup_{j\neq i} z_j + q(F(T,\theta_j,\alpha_1,R_j),V) - q(F(T,\theta_i,\alpha_1,R_i),V))}{\Pr(Z_i \geq \sup_{j\neq i} z_j + q(F(T,\theta_j,\alpha_1,R_j),V') - q(F(T,\theta_i,\alpha_1,R_i),V'))}$$

Observe that from Equation 13,

$$|(q(F(T,\theta_j,\alpha_1,R_j),V) - q(F(T,\theta_i,\alpha_1,R_i),V)) - (q(F(T,\theta_j,\alpha_1,R_j),V') - q(F(T,\theta_i,\alpha_1,R_i),V'))| \leq \frac{2\beta_2}{m} \leq 2\beta$$

Thus, the ratio of the probabilities is at most the ratio $\Pr(Z_i \geq \gamma)/\Pr(Z_i \geq \gamma + 2\beta)$ for $\gamma = \sup_{j\neq i} z_j + q(F(T,\theta_j,\alpha_1,r_j),V) - q(F(T,\theta_i,\alpha_1,r_i),V)$, which is at most e^{α_2} by properties of the exponential distribution. Thus, we have established that when $R \in \Sigma^k$, for all j,

$$\frac{\Pr(i^* = i | Z_{\setminus i} = z_{\setminus i}, D, R)}{\Pr(i^* = i | Z_{\setminus i} = z_{\setminus i}, D'', R)} \leq e^{\alpha_2}$$

Thus for any such R, we can write:

$$\frac{\Pr(i^* = i | D, R)}{\Pr(i^* = i | D'', R)} = \frac{\int_{z_{\setminus i}} \Pr(i^* = i | Z_{\setminus i} = z_{\setminus i}, D, R) h(z_{\setminus i}) dz_{\setminus i}}{\int_{z_{\setminus i}} \Pr(i^* = i | Z_{\setminus i} = z_{\setminus i}, D'', R) h(z_{\setminus i}) dz_{\setminus i}} \leq e^{\alpha_2}$$

Equation 11 now follows by integrating R over E. □

PROOF:(Of Theorem 1) The proof of Theorem 1 follows from a combination of Lemmas 1 and 2. □

PROOF:(Of Theorem 2) The proof of Theorem 2 follows from privacy composition; Theorem 1 ensures that Step (2) of Algorithm 2 is (α_2, δ)-differentially private; moreover the training procedure $\mathcal{T}$ is α_1-differentially private. The theorem follows by composing these two results. □

PROOF:(Of Theorem 3) Observe that:

$$\Pr\left(q(h_{i^*},V) < \max_{1\leq i\leq k} q(h_i,V) - \frac{2\beta\log(k/\delta_0)}{\alpha_2}\right) \leq \Pr\left(\exists j \text{ s.t. } Z_j \geq \frac{\log(k/\delta_0)}{\alpha_2}\right)$$

By properties of the exponential distribution, for any fixed j, $\Pr(Z_j \geq \frac{\log(k/\delta_0)}{\alpha_2}) \leq \frac{\delta_0}{k}$. Thus the theorem follows by an Union Bound. □

6.5 Proof of Theorem 4

PROOF: (Of Theorem 4 for Output Perturbation) Let T and T' be two training sets which differ in a single labelled example ((x_n, y_n) vs. (x'_n, y'_n)), and let $w^*(T)$ and $w^*(T')$ be the solutions to the regularized convex optimization problem in Equation 1 when the inputs are T and T' respectively. We observe that for fixed λ, α and R,

$$F(T,\lambda,\alpha,R) - F(T',\lambda,\alpha,R) = w^*(T) - w^*(T')$$

When the training sets are T and T', the objective functions in the regularized convex optimization problems are both λ-strongly convex, and they differ by $\frac{1}{n}(\ell(w,x_n,y_n) - \ell(w,x'_n,y'_n))$. Combining this fact with Lemma 1 of [3], and using the fact that ℓ is 1-Lipschitz, we have that for all λ and R,

$$\|F(T,\lambda,\alpha,R) - F(T',\lambda,\alpha,R)\| \leq \frac{2}{\lambda n}$$

Since g is L-Lipschitz, this implies that for any fixed validation set V, and for all λ, α and R,

$$|q(F(T,\lambda,\alpha,R),V) - q(F(T',\lambda,\alpha,R),V)| \leq \frac{2L}{\lambda n} \tag{14}$$

Now let V and V' be two validation sets that differ in the value of a single labelled example $(\bar{x}_m, \bar{y}_m)$. Since $g \geq 0$ for all inputs, for any such V and V', and for a fixed Λ, α and R, $|q(F(T,\lambda,\alpha,R),V) - q(F(T,\lambda,\alpha,R),V')| \leq \frac{g_{\max}}{m}$, where

$$g_{\max} = \sup_{(x,y)\in\mathcal{X}} g(F(T,\lambda,\alpha,R),x,y)$$

By definition, $g_{\max} \leq g^*$. Moreover, as g is L-Lipschitz,

$$g_{\max} \leq L \cdot \|F(T,\lambda,\alpha,R)\|$$

Now, let E be the event that $\|R\| \leq d\log(dk/\delta)$. From Lemma 4 of [3], $\Pr(E) \geq 1-\delta/k$. Thus, provided E holds, we have that:

$$\|F(T,\lambda,\alpha,R)\| \leq \|w^*\| + \frac{d\log(dk/\delta)}{\lambda\alpha n} \leq \frac{1}{\lambda} + \frac{d\log(dk/\delta)}{\lambda\alpha n} = \frac{1}{\lambda}\left(1 + \frac{d\log(dk/\delta)}{n\alpha}\right)$$

where the bound on $\|w^*\|$ follows from an application of Lemma 1 of [3] on the functions $\frac{1}{2}\lambda\|w\|^2$ and $\frac{1}{2}\lambda\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n}\ell(w,x_i,y_i)$. This implies that provided E holds, for all training sets T, and for all λ,

$$|q(F(T,\lambda,\alpha,R),V) - q(F(T,\lambda,\alpha,R),V')| \leq \frac{L}{\lambda m}\left(1 + \frac{d\log(dk/\delta)}{n\alpha}\right) \tag{15}$$

The theorem now follows from a combination of Equations 14 and 15, and the definition of g^*. □

PROOF: (Of Theorem 4 for Objective Perturbation) Let T and T' be two training sets which differ in a single labelled example (x_n, y_n). We observe that for a fixed R and λ, the objective of the regularized convex optimization problem in Equation 2 differs in the term $\frac{1}{n}(\ell(w,x_n,y_n) - \ell(w,x'_n,y'_n))$. Combining this with Lemma 1 of [3], and using the fact that ℓ is 1-Lipschitz, we have that for all λ, α, R,

$$\|F(T,\lambda,\alpha,R) - F(T',\lambda,\alpha,R)\| \leq \frac{2}{\lambda n}$$

Since g is L-Lipschitz, this implies that for any fixed validation set V, and for all λ and r,

$$|q(F(T,\lambda,\alpha,R),V) - q(F(T',\lambda,\alpha,R),V)| \leq \frac{2L}{\lambda n} \tag{16}$$

Now let V and V' be two validation sets that differ in the value of a single labelled example $(\bar{x}_m, \bar{y}_m)$. Since $g \geq 0$, for any such V and V', $|q(F(T,\lambda,\alpha,R),V) - q(F(T,\lambda,\alpha,R),V')| \leq \frac{g_{\max}}{m}$, where

$$g_{\max} = \sup_{(x,y)\in\mathcal{X}} g(F(T,\lambda,\alpha,R),x,y)$$

By definition $g_{\max} \leq g^*$. Moreover, as g is L-Lipschitz,

$$g_{\max} \leq L \cdot \|F(T,\lambda,\alpha,R)\|$$

Let E be the event that $\|R\| \leq d\log(dk/\delta)$. From Lemma 4 of [3], $\Pr(E) \geq 1-\delta/k$. Thus, provided E holds, we have that:

$$\|F(T,\lambda,\alpha,R)\| \leq \frac{1 + \|R\|/(\alpha n)}{\lambda} \leq \frac{1}{\lambda}\left(1 + \frac{d\log(dk/\delta)}{n\alpha}\right)$$

This implies that provided E holds, for all training sets T, and for all λ,

$$|q(F(T,\lambda,\alpha,R),V) - q(F(T,\lambda,\alpha,R),V')| \leq \frac{L}{\lambda m}\left(1 + \frac{d\log(dk/\delta)}{n\alpha}\right) \tag{17}$$

The theorem now follows from a combination of Equations 16 and 17, and the definition of g^*. □

6.6 Proof of Theorem 6

Lemma 3 (Concentration of Sum of Laplace Random Variables) *Let $Z_1, \ldots, Z_s$ be $s \geq 2$ iid standard Laplace random variables, and let $Z = Z_1 + \ldots + Z_s$. Then, for any θ,*

$$\Pr(Z \geq \theta) \leq \left(1 - \frac{1}{s}\right)^{-s} e^{-\theta/\sqrt{s}} \leq 4e^{-\theta/\sqrt{s}}$$

PROOF: The proof follows from using the method of generating functions. The generating function for the standard Laplace distribution is: $\psi(X) = \mathbb{E}[e^{tX}] = \frac{1}{1-t^2}$, for $|t| \leq 1$. As $Z_1, \ldots, Z_s$ are independently distributed, the generating function for Z is $\mathbb{E}[e^{tZ}] = (1-t^2)^{-s}$. Now, we can write:

$$\begin{aligned} \Pr(Z \geq \theta) &= \Pr(e^{tZ} \geq e^{t\theta}) \\ &\leq \frac{\mathbb{E}[e^{tZ}]}{e^{t\theta}} = e^{-t\theta} \cdot (1-t^2)^{-s} \end{aligned}$$

Plugging in $t = \frac{1}{\sqrt{s}}$, we get that:

$$\Pr(Z \geq \theta) \leq \left(1 - \frac{1}{s}\right)^{-s} e^{-\theta/\sqrt{s}}$$

The lemma follows by observing that for $s \geq 2$, $(1 - \frac{1}{s})^s \geq \frac{1}{4}$. $\square$

PROOF: (Of Theorem 6) Let $V = \{z_1, \ldots, z_m\}$ be a validation dataset, and let V' be a validation dataset that differs from V in a single sample (z_m vs z'_m). We use the notation R to denote the sequence of values $R = (R_1, R_2, \ldots, R_{1/h})$. Given an input sample T, a bin size h, a privacy parameter α, and a sequence R, we use the notation $\hat{f}_{T,h,\alpha,R}$ to denote the density estimator $F(T, h, \alpha, R)$. For all such T, all h, all α and all R, we can write:

$$\begin{aligned} |q(F(T,h,\alpha,R),V) - q(F(T,h,\alpha,R),V')| &= \frac{2}{m}(\hat{f}_{T,h,\alpha,R}(z_m) - \hat{f}_{T,h,\alpha,R}(z'_m)) \\ &\leq \frac{2}{m} \cdot \frac{\max_i \tilde{n}_i}{h\tilde{n}} \leq \frac{2}{mh} \end{aligned} \tag{18}$$

For a fixed value of h, we define the following event E:

$$\sum_{i=1}^{1/h} R_i \geq -\frac{\ln(4k/\delta)}{\sqrt{h}}$$

Using the symmetry of Laplace random variables and Lemma 3, we get that $\Pr(E) \geq 1 - \delta/k$. We observe that provided the event E holds,

$$\tilde{n} \geq n - \sum_{i=1}^{1/h} R_i \geq n - \frac{2\ln(4k/\delta)}{\alpha\sqrt{h}} \geq n(1-\nu) \tag{19}$$

Let T and T' be two input datasets that differ in a single sample (x_n vs x'_n). We fix a bin size h, a value of α, and a sequence R, and for these fixed values, we use the notation $\tilde{n}_i$ and $\tilde{n}'_i$ to denote the value of $\tilde{n}_i$ in Algorithm 5 when the inputs are T and T' respectively. Similarly, we use $\tilde{n} = \sum_i \tilde{n}_i$ and $\tilde{n}' = \sum_i \tilde{n}'_i$.

For any V, we can write:

$$\begin{aligned} q(F(T,h,\alpha,R),V) - q(F(T',h,\alpha,R),V) &= \frac{2}{m}\sum_{j=1}^{m}(\hat{f}_{T,h,\alpha,R}(z_j) - \hat{f}_{T',h,\alpha,R}(z_j)) \\ &- \sum_{i=1}^{1/h} h \cdot \left(\frac{\tilde{n}_i^2}{h^2\tilde{n}^2} - \frac{\tilde{n}_i'^2}{h^2\tilde{n}'^2}\right) \end{aligned} \tag{20}$$

We now look at bounding the right hand side of Equation 20 term by term. Suppose T' is obtained rom T by moving a single sample x_n from bin a to bin b in the histogram. Then, depending on the relative values of $\tilde{n}_a$ and $\tilde{n}_b$, there are four cases:

1. $\tilde{n}'_a = \tilde{n}_a - 1, \tilde{n}'_b = \tilde{n}_b + 1$. Thus $\tilde{n}' = \tilde{n}$.
2. $\tilde{n}'_a = \tilde{n}_a = 0, \tilde{n}'_b = \tilde{n}_b + 1$. Thus $\tilde{n}' = \tilde{n} + 1$.
3. $\tilde{n}'_a = \tilde{n}_a - 1, \tilde{n}'_b = \tilde{n}_b = 0$. Thus $\tilde{n}' = \tilde{n} - 1$.
4. $\tilde{n}'_a = \tilde{n}_a = 0, \tilde{n}'_b = \tilde{n}_b = 0$. Thus $\tilde{n}' = \tilde{n}$.

In the fourth case, $\hat{f}_{T,h,\alpha,R} = \hat{f}_{T',h,\alpha,R}$, and thus the right hand side of Equation 20 is 0. Moreover, the second and the third cases are symmetric. We thus focus on the first two cases.

In the first case, the first term in the right hand side of Equation 20 can be written as:

$$\begin{aligned}\Big|\frac{2}{m}\cdot\sum_{j=1}^{m}\sum_{i=1}^{1/h}\mathbf{1}(z_j\in I_i)\cdot\left(\frac{\tilde{n}_i}{h\tilde{n}}-\frac{\tilde{n}'_i}{h\tilde{n}'}\right)\Big| &= \Big|\frac{2}{m}\cdot\sum_{j=1}^{m}\sum_{i=1}^{1/h}\mathbf{1}(z_j\in I_i)\cdot\frac{\tilde{n}_i-\tilde{n}'_i}{h\tilde{n}}\Big| \\ &\leq \frac{2}{m}\cdot m\cdot\frac{1}{h\tilde{n}}\leq\frac{2}{h\tilde{n}}\end{aligned}$$

The second term on the right hand side of Equation 20 can be written as:

$$\begin{aligned}\Big|\sum_{i=1}^{1/h}\left(\frac{\tilde{n}_i^2}{h\tilde{n}^2}-\frac{\tilde{n}_i'^2}{h\tilde{n}'^2}\right)\Big| &= \frac{\tilde{n}_a^2+\tilde{n}_b^2-(\tilde{n}_a-1)^2-(\tilde{n}_b+1)^2}{h\tilde{n}^2} \\ &= \Big|\frac{2\tilde{n}_a-2\tilde{n}_b-2}{h\tilde{n}^2}\Big|\leq\frac{2}{h\tilde{n}}\end{aligned}$$

where the last step follows from the fact that $\tilde{n}'_b = \tilde{n}_b + 1 \leq \tilde{n}$. Thus, for the first case, the right hand side of Equation 20 is at most $\frac{4}{h\tilde{n}}$.

We now consider the second case. The first term on the right hand side of Equation 20 can be written as:

$$\begin{aligned}&\Big|\frac{2}{m}\cdot\sum_{j=1}^{m}\sum_{i=1}^{1/h}\mathbf{1}(z_j\in I_i)\cdot\left(\frac{\tilde{n}_i}{h\tilde{n}}-\frac{\tilde{n}'_i}{h\tilde{n}'}\right)\Big| \\ =\ &\Big|\frac{2}{mh}\cdot\sum_{j=1}^{m}\sum_{i=1}^{1/h}\mathbf{1}(z_j\in I_i)\cdot\left(\frac{\tilde{n}_i}{\tilde{n}}-\frac{\tilde{n}'_i}{\tilde{n}+1}\right)\Big| \\ \leq\ &\frac{2}{hm}\cdot m\cdot\frac{1}{\tilde{n}(\tilde{n}+1)}\cdot\max(|\tilde{n}_i(\tilde{n}+1)-\tilde{n}_i\tilde{n}|,|\tilde{n}_i(\tilde{n}+1)-\tilde{n}(\tilde{n}_i+1)|) \\ \leq\ &\frac{2}{h}\cdot\frac{1}{\tilde{n}(\tilde{n}+1)}\cdot\max(|\tilde{n}_i|,|\tilde{n}-\tilde{n}_i|)\leq\frac{2}{h(\tilde{n}+1)}\end{aligned}$$

where the last step follows from the fact that $\max(|\tilde{n}_i|, |\tilde{n} - \tilde{n}_i|) \leq \tilde{n}$. The second term on the right hand side of Equation 20 can be written as:

$$\begin{aligned}\Big|\sum_{i=1}^{1/h}\left(\frac{\tilde{n}_i^2}{h\tilde{n}^2}-\frac{\tilde{n}_i'^2}{h\tilde{n}'^2}\right)\Big| &= \sum_{i\neq b}\left(\frac{\tilde{n}_i^2}{h\tilde{n}^2}-\frac{\tilde{n}_i^2}{h(\tilde{n}+1)^2}\right)+\Big|\frac{\tilde{n}_b^2}{h\tilde{n}^2}-\frac{(\tilde{n}_b+1)^2}{h(\tilde{n}+1)^2}\Big| \\ &= \frac{2\tilde{n}+1}{h\tilde{n}^2(\tilde{n}+1)^2}\cdot\sum_{i\neq b}\tilde{n}_i^2+\Big|\frac{(\tilde{n}_b-\tilde{n})(2\tilde{n}_b\tilde{n}+\tilde{n}+\tilde{n}_b)}{h\tilde{n}^2(\tilde{n}+1)^2}\Big| \\ &\leq \frac{2\tilde{n}+1}{h(\tilde{n}+1)^2}+\frac{\tilde{n}\cdot 2\tilde{n}(\tilde{n}+1)}{h\tilde{n}^2(\tilde{n}+1)^2}\leq\frac{4}{h(\tilde{n}+1)}\end{aligned}$$

Thus, in the second case, the right hand side of Equation 20 is at most $\frac{6}{h(\tilde{n}+1)}$. We observe that the third case is symmetric to the second case, and thus we can carry out very similar calculations in the third case to show that the right hand side is at most $\frac{6}{h\tilde{n}}$. Thus, we have that for any T and T', provided the event E holds,

$$|q(F(T,h,\alpha,R),V)-q(F(T',h,\alpha,R),V)|\leq\frac{6}{h\tilde{n}} \tag{21}$$

The theorem now follows by combining Equation 21 with Equation 19. □

6.7 Proof of Theorem 5

Lemma 4 *(Parallel construction) Let $\mathbb{A} = \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k\}$ be a list of k independently randomized functions, and let $\mathcal{A}_i$ be α_i-differentially private. Let $\{D_1, D_2, \ldots, D_k\}$ be k subsets of a set D such that $i \neq j \implies D_i \cap D_j = \emptyset$. Algorithm $\mathcal{B}(D, \mathbb{A}) = (\mathcal{A}_1(D_1), \mathcal{A}_2(D_2), \ldots, \mathcal{A}_k(D_k))$ is $\max_{1 \leq i \leq k} \alpha_i$-differentially private.*

PROOF: Let D, D' be two datasets such that their symmetric difference contains one element. We have that

$$\frac{P(\mathcal{B}(D, \mathbb{A}) \in S)}{P(\mathcal{B}(D', \mathbb{A}) \in S)} = \frac{P(\mathcal{B}(D, \mathbb{A}) \in S_1 \times \cdots \times S_k)}{P(\mathcal{B}(D', \mathbb{A}) \in S_1 \times \cdots \times S_k)} = \frac{P(\mathcal{A}_1(D_1) \in S_1) \cdots P(\mathcal{A}_k(D_k) \in S_k)}{P(\mathcal{A}_1(D'_1) \in S_1) \cdots P(\mathcal{A}_k(D'_k) \in S_k)} \tag{22}$$

by independence of randomness in the $\mathcal{A}_i$. Since $i \neq j \implies D_i \cap D_j = \emptyset$, there exists at most one index j such that $D_j \neq D'_j$. If j does not exist, (22) reduces to $e^0 \leq e^{\max_{1 \leq i \leq k} \alpha_i}$. Let j exist, then

$$\frac{P(\mathcal{B}(D, \mathbb{A}) \in S)}{P(\mathcal{B}(D', \mathbb{A}) \in S)} = \frac{P(\mathcal{A}_j(D_j) \in S_j)}{P(\mathcal{A}_j(D'_j) \in S_j)} \leq e^{\alpha_j} \leq e^{\max_{1 \leq i \leq k} \alpha_i},$$

which concludes the proof. □

PROOF: (Theorem 5) We begin by separating task (a) of producing the f_i in step 1. from the task (b) of computing e_i in step 2. and selecting i^* in step 3.

From the parallel construction Lemma 4 it follows that (a) in $dataSplit$ is α-differentially private. From standard composition of privacy it follows that (a) in $alphaSplit$ is α-differentially private.

Task (b) is for both $alphaSplit$ and $dataSplit$ an application of the exponential mechanism [20], which for choosing with a probability proportional to $\epsilon(-e_i)$ yields $2\epsilon\Delta$-differential privacy, where Δ is the sensitivity of e_i. Since a single change in V can change the number of errors any fixed classifier can make by at most $1 = \Delta$, we get that task (b) is α-differentially private for $\epsilon = \alpha/2$.

If T and V are disjoint, we get by parallel construction that both $alphaSplit$ and $dataSplit$ yield α-differential privacy. If T and V are not disjoint, by standard composition of privacy we get that both $alphaSplit$ and $dataSplit$ yield 2α-differential privacy.

In $Random$, the results of step 2. in task (b) are never used in step 3. Step 3 is done without looking at the input data and does not incur loss of differential privacy. We can therefore simulate $Random$ by first choosing i^* uniformly at random, and then computing f_i at α-differential privacy, which by standard privacy composition is α-differentially private. □

6.8 Experimental selection of regularizer index

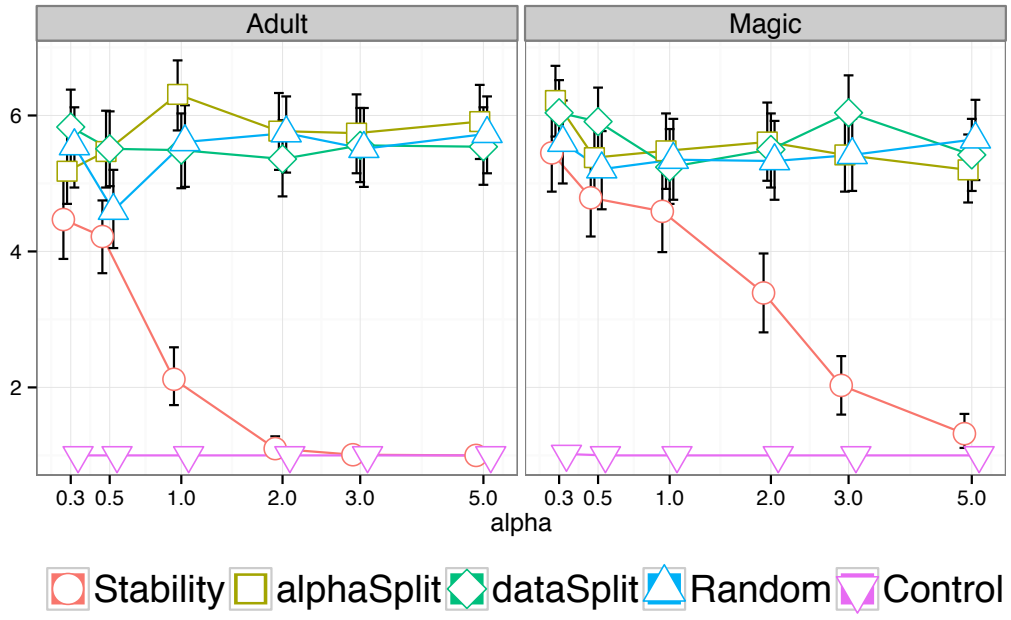

Figure 2: A summary of 10 times 10-fold cross-validation selection of regularizer index i into Θ for different privacy levels α. Each point in the figure represents a summary of 100 data points. The error bars indiciate a boot-strap sample estimate of the 95% confidence interval of the mean. A small amount of jitter was added to positions on the x-axes to avoid over-plotting.