Privacy Enhancing Technologies FS2025
Exercise Sheet 11 (version 1)

Florian Tramèr

Problem 1: Conceptual Questions. For each of the following statements, say whether it is TRUE or FALSE. Write at most one sentence to justify your answer.

(a) Let $f : \mathcal{X} \to \mathbb{R}^k$ be a neural network that, given an input $x \in \mathcal{X}$, outputs a vector $f(x) \in \mathbb{R}^k$. In a *model inversion attack*, the attacker is given $f(x)$ for some unknown input $x \in \mathcal{X}$ and tries to recover x. This attack is prevented if f is trained with DP-SGD for some ε and δ, i.e., the probability that the attacker's guess x' is correct is $\Pr[x = x'] \leq \varepsilon/|\mathcal{X}| + \delta$.

(b) Recall the global loss thresholding MI attack from lectures 31–34. We reproduce the histogram of model losses in fig. 1 for convenience. From this plot, you can deduce:

(i) For small FPR (e.g., $< 1\%$), we have that TPR $\approx$ FPR.

(ii) For small FNR (e.g., $< 1\%$), we have that TNR $\approx$ FNR.

(iii) We cannot rule out that the training algorithm satisfies ε-DP for some $\varepsilon \ll 1$.

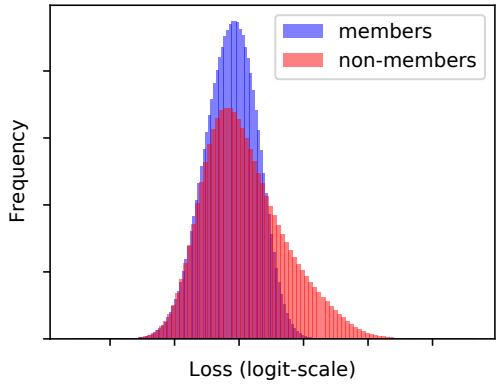

Figure 1: Histogram of the loss of a neural network trained on CIFAR-10, for samples from the training set (in red) and from the test set (in blue).

Problem 2: Homer's attack In this problem, we consider the setting from the famous paper by Homer et al. [HSR+08] on membership inference on published genomic statistics.

We consider the following simple setting:

- An individual's genomic data is a binary vector $x \in \{0,1\}^m$ of m binary features (corresponding to rare variants, or "SNPs".)
- All users' genomic data is drawn from a population distribution where each feature is independent and follows a Bernoulli distribution with probability $p \in (0,1)$ of being 1.
- We collect a set of n individuals $D = \{x^{(1)}, \ldots, x^{(n)}\}$ randomly from this distribution. We release the empirical frequency of each feature in the dataset, i.e., the vector $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_m)$ where $\hat{p}_i = \frac{1}{n}\sum_{j=1}^{n} x_i^{(j)}$.

Assume the attacker is given the vector $\hat{p}$, the true frequency p, and the genomic data x of a single individual. The attacker wants to determine whether x was in the dataset D.

(a) Let's first focus on a single feature, so $m = 1$. Formulate the null and alternative hypotheses for the membership inference problem. What are the probabilities $\Pr[\hat{p} \mid H_0, x]$ and $\Pr[\hat{p} \mid H_1, x]$ of observing $\hat{p}$ in both cases? Express these using binomial distributions $\mathrm{Bin}(n, p)$ and $\mathrm{Bin}(n-1, p)$.

Hint: Let $c = n\hat{p}$ denote the number of individuals with feature 1 in the dataset. Note that when x is in the dataset, there are $c - x$ individuals with feature 1 that were sampled from the population distribution.

(b) Show that the likelihood ratio is:

$$\frac{\Pr[\hat{p} \mid H_0, x]}{\Pr[\hat{p} \mid H_1, x]} = \left(\frac{p}{\hat{p}}\right)^x \cdot \left(\frac{1-p}{1-\hat{p}}\right)^{1-x}$$

Hint: The following weird identity may be useful:

$$\binom{n-1}{c-x} = \left(\frac{c}{n}\right)^x \left(\frac{n-c}{n}\right)^{1-x} \binom{n}{c} \quad \text{(for } x \in \{0,1\} \text{ and integers } n \geq c\text{)}$$

(c) Let's now assume there are $m \geq 2$ features, that are all independent of each other. The parameter $p = (p_1, \ldots, p_m)$ is now a vector with one Bernoulli parameter for each feature. The attacker is given the vector $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_m)$, the true frequencies p, and the genomic data $x = (x_1, \ldots, x_m)$ of a single individual.

What is the log of the likelihood ratio?

Problem 3: Packing Lower Bounds (inspired by Gautam Kamath). In this problem, we will see a different way to prove lower bounds on (pure) DP mechanisms, using a technique called *packing*.

(a) We start with a generic and intuitive bound, showing that if there are many close datasets that have very different ground-truth answers, then a private mechanism cannot have high accuracy on all of them.

Let $D_1, \ldots, D_m \in \mathcal{X}^n$ be a set of m datasets, which are at distance at most t from some fixed dataset $D \in \mathcal{X}^n$ (i.e., you need to modify at most t elements of D to get to any D_i). Let $Y_1, \ldots, Y_m \in \mathcal{Y}$ be a set of m disjoint subsets of the space $\mathcal{Y}$ (i.e., $Y_i \cap Y_j = \emptyset$ for all $i, j \in [m]$ and $\bigcup_{i=1}^{m} Y_i \subseteq \mathcal{Y}$).

Prove that if we want an ε-DP mechanism $M \colon \mathcal{X}^n \to \mathcal{Y}$ that satisfies $\Pr[M(D_i) \in Y_i] \geq \alpha$ for every $i \in [m]$, then we need to have $e^{t\varepsilon} \geq \alpha m$.

(b) Consider an algorithm $M\colon \mathcal{X}^n \to [0,1]^d$, where $\mathcal{X} = \{0,1\}^d$, which returns the mean of the dataset along each column, i.e., $M(D)_i = \frac{1}{n}\sum_{k=1}^{n}(x_k)_i$.

Use the previous bound to show that if M is ε-DP and M answers each mean query with error $< 1/2$ with probability at least 1%, then we must have $n = \Omega(d/\varepsilon)$.

(c) Is the bound in (b) tight, up to logarithmic factors? I.e., is there an ε-DP algorithm M which, given $n = \tilde{O}(d/\varepsilon)$ samples, answers all mean queries with error $< 1/2$ with probability $\geq 1\%$?

Problem 4: Auditing machine learning models. To empirically audit the privacy of a training algorithm, a common approach is to add a random example into the training set (a "canary") and then test whether an attacker can recover this example from the trained model. This works as follows. We pick a canary c uniformly at random from a set $\mathcal{U}$ of 10^6 random elements (e.g., all 6-digit numbers), and insert c into the training data D (the rest of the training data D contains no information about c). We then train a model f with DP-SGD on D. The attacker is given the trained model f, and the set of elements $\mathcal{U}$ and has to emit a guess $c' \in \mathcal{U}$ for what the canary was. The attacker guesses correctly if $c' = c$.

Suppose there exists an attacker A who guesses the correct canary with probability 1%, for any choice of the random canary c (the success probability is taken over the randomness of training the model f and running the attack). Formally:

$$\forall c \in \mathcal{U} : \Pr[A(f,\mathcal{U}) = c \mid f \text{ trained on } D + \{c\}] = 1\%$$

Compute a lower bound on the training algorithm's level of (pure) DP.

For this question, assume that two datasets D_1 and D_2 are neighboring if one dataset is obtained by adding a single example to the other (e.g, $D_1 = D_2 + \{c\}$).

References

[HSR+08] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008.