DA 5001/6400 (July-Nov 2024): HW3

IIT Madras

Due Date: October 13, 2024 at 11:59 PM

Instructions

- The maximum score on the homework is 100 marks, including the bonus (extra credit) exercises.
- List the names of students you collaborated with for the homework. Also, cite any books, notes, or web resources you used for any problem. Failure to do so will be considered a violation of the honour code.
- If you used an LLM for help in one of the exercises, you must specify the name of the LLM and the exact prompt used.
- Corrections to the homework are shown in blue. Please note them.

0 Honour Code (0 marks but mandatory)

Please read the full honour code on the course webpage and write "I ACCEPT THE HONOUR CODE" in your submission. **Your submission will not be graded if you do not accept the honour code.**

1 Output Perturbation for Private Ridge Regression (Theory)

Consider the ridge regression problem, where we wish to minimize the objective

$$F(\theta) = \sum_{i=1}^{n} (x_i^\top \theta - y_i)^2 + \lambda \|\theta\|_2^2$$

where $D = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ is a dataset of n input-output pairs, $\theta \in \mathbb{R}^d$ denotes the model parameters, and λ is a regularization parameter. Recall that we can also write this in matrix form as

$$F(\theta) = \|X\theta - y\|_2^2 + \lambda \|\theta\|_2^2$$

with $X \in \mathbb{R}^{n \times d}$ as the data matrix (with x_i's as rows), and $y \in \mathbb{R}^n$ as the vector of targets.

Output perturbation returns $\theta_\star + \mathcal{N}(0, \sigma^2 I)$, where

$$\theta_\star = \arg\min_\theta F(\theta)$$

is the (unique global) minimizer of the objective θ. The goal of this exercise is to calibrate the variance σ^2 to the privacy level.

Q 1.1 Show that $\theta_\star = (X^\top X + \lambda I_d)^{-1} X^\top y$, where I_d is the $d \times d$ identity matrix. Going ahead, we will denote $H = X^\top X + \lambda I$ and $b = X^\top y$ so that $\theta_\star = H^{-1} b$.

Output perturbation is clearly an instance of the Gaussian mechanism. Thus, we only need to establish the sensitivity of $\theta_\star$: how would the minimizer change upon the addition or removal of one datapoint? We start with the case of addition:

Q 1.2 Consider a neighbhouring dataset $D' = D \cup \{(\tilde{x}, \tilde{y})\}$ for some $\tilde{x} \in \mathbb{R}^d$ and $\tilde{y} \in \mathbb{R}$. Repeat your calculations in the previous part to show that the minimizer of $\tilde{F}(\theta) = F(\theta) + (\tilde{x}^\top \theta - y)^2$ is

$$\tilde{\theta}_\star = (H + \tilde{x}\tilde{x}^\top)^{-1}(b + \tilde{y}\tilde{x})\,.$$

Q 1.3 Show that

$$\theta_\star - \tilde{\theta}_\star = H^{-1}\tilde{x}\left(\frac{\tilde{x}^\top \theta_\star - \tilde{y}}{1 + \tilde{x}^\top H^{-1}\tilde{x}}\right)\,.$$

You will find it convenient to use the following result, known variously as the Sherman-Morrison formula, the Woodbury identity, and the matrix inversion lemma:

Lemma 1. *Let $A \in \mathbb{R}^{d\times d}$ be an invertible square matrix and let $u, v \in \mathbb{R}^d$ be any vectors. Then, the matrix $A + uv^\top$ is invertible if and only if $v^\top A u \neq -1$. In this case, we have,*

$$(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1} u}\,.$$

Our next task is to bound the sensitivity $\|\tilde{\theta}_\star - \theta_\star\|_2$.

Q 1.4 Assume that $\|\tilde{x}\|_2 \leq R$. Then, show that $\|H^{-1}\tilde{x}\|_2 \leq R/\lambda$.

Q 1.5 Next, assume in addition that $|\tilde{x}^\top \theta_\star - \tilde{y}| \leq M$. Show that $\|\tilde{\theta}_\star - \theta_\star\|_2 \leq MR/\lambda$.

Q 1.6 Argue that the sensitivity for the case of removal of a datapoint $(\tilde{x}, \tilde{y}) \in D$ is also MR/λ (assuming again that $\|\tilde{x}\|_2 \leq R$ and $|\tilde{x}^\top \theta_\star - \tilde{y}| \leq M$). You need not repeat all your previous calculations; you just need to argue why the final result does not change from addition to removal.

Q 1.7 Thus, establish that output perturbation is ρ-zCDP if the variance is $\sigma^2 = \frac{M^2R^2}{2\lambda^2\rho}$.

2 Auditing Differential Privacy (Implementation)

Read through the text below carefully and understand it before implementing it.

Background Recall that DP imposes fundamental bounds on the success of an adversary in a membership inference task: $\mathsf{TPR} \leq e^\varepsilon \cdot \mathsf{FPR}$. Then, we run a membership inference attack and calculate $\varepsilon \geq \log(\mathsf{TPR}/\mathsf{FPR})$. This gives us a *lower bound* on the privacy leakage (i.e., I have exhibited X amount of leakage), in contrast with the usual approach which gives us theoretical *upper bounds* on the privacy leakage.

This approach, called **empirical privacy auditing**, has many uses. If your high probability lower bound is larger than your upper bound, then it is indicative of a bug in your implementation: see Tramèr et al. (arXiv id 2202.12219) for a real-life example. This approach can also give a true estimate of the privacy leakage (with DP upper bounds usually corresponding to unrealistic pessimistic worst cases). We will implement an auditing approach for a simulated mean estimation example. (The general approach can be used for auditing DP-SGD: you are encouraged to try it out.)

Algorithm Let D be the dataset of n examples and $\mathcal{A}$ denote the DP (randomized) algorithm you wish to audit. Further, let D_{canary} be a set of m *canary* datapoints: these are "fake" datapoints you will insert and test for membership.[1] The algorithm is as follows, and is due to Steinke et al.:

- **Input**: Randomized algorithm $\mathcal{A}$, original dataset D of size $|D| = n$, canary dataset D_{canary} of size $|D_{\mathsf{canary}}| = m$, number of positive guesses m_1 and number of negative guesses m_2 (so that number of abstenstions is $m - m_1 - m_2$).
- Sample Radamacher random variables $S_1, \dots, S_m \sim \mathsf{Rad}(1/2)$, i.e., $\mathbb{P}(S_i = 1) = 1/2 = \mathbb{P}(S_i = -1)$. Let $D^+_{\mathsf{canary}} = \{x \in D_{\mathsf{canary}} : S_i = 1\}$ and $D^-_{\mathsf{canary}} = \{x \in D_{\mathsf{canary}} : S_i = -1\}$.
- Run the algorithm $\theta \sim \mathcal{A}(D')$ with the dataset $D' = D \cup D^+_{\mathsf{canary}}$. That is, you add the "positive" canaries to your original dataset.
- Find the membership score $f_i = \mathsf{Score}(\theta, x_i)$ for each $x_i \in D_{\mathsf{canary}}$. This score is high if we believe that x_i was included in the input dataset D' and low if we believe that x_i was not inlcuded in the dataset D'. This should depend on our algorithm $\mathcal{A}$—we will specify it below.
- Assign a score $T_i \in \{-1, 0, 1\}$ for each canary $x_i \in D_{\mathsf{canary}}$ such that
  - $T_i = +1$ for m_1 largest f_i scores (guess member);
  - $T_i = -1$ for the m_2 smallest f_i scores (guess non-member);
  - $T_i = 0$ otherwise (abstain from guessing for these $m - m_1 - m_2$ canaries).
- Let $N_{\mathsf{correct}} = \sum_{i=1}^m \mathbb{I}(S_i = T_i)$ and $N_{\mathsf{total}} = \sum_{i=1}^m \mathbb{I}(T_i \neq 0) = m_1 + m_2$ denote the number of correct and total membership guesses.
- **Output**: a valid empirical lower bound on the privacy leakage which holds with probability 0.95, which is

$$\hat{\varepsilon} := \max_{\varepsilon} \left\{ \mathbb{P}\left(\mathsf{Binom}\left(N_{\mathsf{total}}, \tfrac{e^\varepsilon}{1+e^\varepsilon} \right) \geq N_{\mathsf{correct}} \right) \leq 0.05 \right\}, \tag{1}$$

  where $\mathsf{Binom}(N, p)$ denotes a Binomial random variable with N trials with probability p.[2] We can find this $\hat{\varepsilon}$ using numerical optimization (see e.g. `scipy.optimize.brentq` from our lab on DP accounting) or via binary/bisection search.

More Details (optional) In the algorithm above, S_i's denote the true membership, while T_i's denote our algorithm's guess about the membership. Clearly, DP implies that the accuracy of the algorithm cannot be too high. That intuition can be formalized into the following result:

Theorem 2. *If the randomized algorithm $\mathcal{A}$ is ε-DP, then we have for all $t \in \mathbb{R}$ that*

$$\mathbb{P}(N_{\mathsf{correct}} \geq t) \leq \mathbb{P}\left(\mathsf{Binom}\left(N_{\mathsf{total}}, \tfrac{e^\varepsilon}{1+e^\varepsilon} \right) \geq t \right).$$

[1] For a stronger audit, the canary examples have be easier to detect. Please see the project list for additional reading on privacy auditing for more information about canaries.

[2] Note that this is the same as the number of 1s obtained from N_{total} randomized response trials.

You can get a closed-form bound by using Hoeffding's inequality on the right hand side (this is an interesting exercise: do it).

The logic behind how Equation 1 is obtained from this theorem is as follows. If you have more correct guesses, then the privacy leakage ε is larger. At a given ε, you expect to see $\mathsf{Binom}\left(N_{\mathsf{total}}, \frac{e^\varepsilon}{1+e^\varepsilon}\right)$ correct guesses. What is the largest value of ε such that at least N_{correct} correct guesses out of N_{total} guesses is highly unlikely? This is the lower bound on the privacy leakage that we wish to find.

There is a technicality that we glossed over: we are returning a lower bound on ε-DP for the Gaussian mechanism. However, since the Gaussian mechanism does not satisfy ε-DP for any $\varepsilon > 0$, any ε should trivally be a correct lower bound. In general, the above auditing procedure can be extended to (ε, δ)-DP as well, but this captures the key insights.

Implementation We will consider a simple sum estimation setting:

- The input space is $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$. Our deterministic algorithm is $A(D) = \sum_{x \in D} x$.
- Since $\|x\|_2 \le 1$, its ℓ_2-sensitivity is 1. Thus, we can use the Gaussian mechanism: the randomized algorithm $\mathcal{A}(D) = \mathcal{N}(A(D), \sigma^2 I)$ satisfies (ε, δ)-DP with $\varepsilon = \sqrt{2\log(1.25/\delta)}/\sigma$. Take $\delta = 10^{-6}$ and choose σ calibrated to a certain value of ε: we will repeat this experiment for various values of ε.
- Take our dataset to be $D = \{0_d\}$ where $0_d \in \mathbb{R}^d$ is the zero vector (we will vary d below).
- We choose D_{canary} to contain m random vectors sampled uniformly from the unit sphere contain. That is, sample $u_i \sim \mathcal{N}(0, I_d)$ and set $x_i = u_i/\|u_i\|_2$ as the i^{th} canary. We will also vary m below.
- We take the score function to be
$$\mathsf{score}(\theta, x_i) = x_i^\top \theta\,.$$
Recall why this is a reasonable approach from the HW0 problem on membership inference: $\theta^\top x_i$ is zero mean only if $x_i \in D'$ is included in the "training dataset".
- Take $m_1 = m_2$ and vary this parameter between 1 and $\min\{m/2, 500\}$. Report the largest value of the empirical lower bound $\hat{\varepsilon}$ over all values of this parameter.

Questions You task is to plot the obtained empirical lower bound ε while varying the DP upper bound ε, number of canaries m, the dimension d. The details are given below:

Q 2.1 Fix $d = 10^4$, $m = 1000$. Vary σ^2 calibrated to $(\varepsilon, 10^{-6})$-DP for $\varepsilon \in \{1, 2, 4, 8, 16\}$. Plot the obtained lower bound $\hat{\varepsilon}$ vs. the theoretical upper bound ε.

Q 2.2 Fix $d = 10^4$, and σ^2 calibrated to $(16, 10^{-6})$-DP. Vary $m \in \{2^6, 2^8, 2^{10}, 2^{12}, 2^{14}\}$ and plot the lower bound $\hat{\varepsilon}$ vs. m (with m in log scale).

Q 2.3 Fix $m = 1000$, and σ^2 calibrated to $(16, 10^{-6})$-DP. Vary $d \in \{10, 10^2, 10^3, 10^4, 10^5\}$ and plot the lower bound $\hat{\varepsilon}$ vs. d (with d in log scale).

For each one, what is the trend you observe? How can you explain the trend you observe? **Hint**.[3]

[3] **Hint**: For the plot vs. m, think about the bias-variance tradeoff coming from introducing many canaries. For the plot vs. d, reason whether the membership inference task becomes easier or harder as d grows larger.