Homework 2

December 5, 2019

Problem 1: **Trying out adversarial examples.** We're obviously not going to grade this one, but if you have the resources, try making an adversarial example for a standard neural network! This library is a good starting point: `https://github.com/MadryLab/robustness`.

Problem 2: **Certifying classical learning algorithms.** For the following binary classification models, give an efficient (i.e. polynomial time) certification algorithm. If you can't find an exact one, give a convex relaxation, as best as you can:

- The classifier is a linear model, i.e. $f(X) = \mathrm{sgn}(\langle X, \theta \rangle)$ for some $\theta \in \mathbb{R}^d$, and the perturbation set is $\mathcal{P}_{2,\varepsilon}$.
- The classifier is a linear model, i.e. $f(X) = \mathrm{sgn}(\langle X, \theta \rangle)$ for some $\theta \in \mathbb{R}^d$, and the perturbation set is $\mathcal{P}_{\infty,\varepsilon}$.
- The classifier is a linear model with a polynomial kernel, i.e. $f(X) = \mathrm{sgn}(p(X))$, where p is a degree k polynomial, for some constant k, and the perturbation set is $\mathcal{P}_{2,\varepsilon}$.
- The classifier is a nearest neighbor classifier, i.e. we have two datasets S_0, S_1, and our classifier is $f(X) = \arg\min_i \min_{x' \in S_i} \|x - x_i\|_2^2$, and the perturbation set is $\mathcal{P}_{2,\varepsilon}$. Here efficient means time which is polynomial in the dimension and in the size of S_0 and S_1.

Problem 3: **Histograms with approximate DP.** Let $f : \mathcal{X}^n \to \mathbb{R}^{\mathcal{S}}$ be a histogram over a potentially infinite range $\mathcal{S}$. Consider the following algorithm. Given $x \in \mathcal{X}^n$, construct $a \in \mathbb{R}^{\mathcal{S}}$ as follows:

- If $f(x)_S = 0$ for some $S \in \mathcal{S}$, then set $a_S = 0$.
- If $f(x)_S \neq 0$, then:
  (a) Set $a'_S = f(x)_S + \mathrm{Lap}(\frac{2}{\varepsilon n})$.
  (b) If $a'_S < \frac{2\ln(2/\delta)}{\varepsilon n} + 1/n$, then let $a_S = 0$. Otherwise, let $a_S = a'_S$.

(a) Show that with probability ≥ 0.99, $\|f(x) - a\|_\infty \lesssim \frac{\log 1/\delta}{\varepsilon}$.

(b) Show that the map $x \mapsto a$ is (ε, δ)-differentially private.
Hint: Let x, x' be adjacent datasets, and decompose $\mathcal{S}$ into four subsets, namely, sets S where both $f(x)_S$ and $f(x')_S$ are non-zero, sets S where only one of $f(x)_S, f(x')_S$ are nonzero, and sets where both are zero. What can you say about the privacy guarantee on each type of set?