# Homework 3

Spring 25, CS 442: Trustworthy Machine Learning
**Due Fri. Apr. 25th at 23:59 CT**

Instructor: Han Zhao

**Instructions for submission** All the homework submissions should be typeset in LaTeX. For all the questions, please clearly justify each step in your derivations or proofs.

## 1 Bayes Optimal Predictor for Regression [10pts]

Let $\mu$ be a joint distribution over $X \times Y$, where $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$. In this problem we are interested in deriving the Bayes optimal predictor for regression under the mean-squared error. More specifically, we are interested in finding a function $f^*(\cdot)$ that minimizes the following expected squared loss:

$$f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim\mu}\left[(f(x) - y)^2\right],$$

where $\mathcal{F}$ is the function class that contains all possible functions from $\mathbb{R}^p$ to $\mathbb{R}$. Derive an analytical solution for $f^*$.

## 2 Tradeoff between Robustness and Accuracy [30pts]

In this problem we will show the potential tradeoff between robustness and accuracy for linear classifiers under a constructed distribution that we have seen in the lectures. Consider a binary classification task as follows. To sample a data from the distribution, we first sample uniformly at random $y \in \{+1, -1\}$, i.e., $\Pr(Y = 1) = \Pr(Y = -1) = 1/2$. Given the label of interest $y$, there are $d$ features, namely $x_1, \ldots, x_d$, whose distributions are defined as follows:

$$x_1 = \begin{cases} +y, & \text{w.p. } p \\ -y, & \text{w.p. } 1-p \end{cases}, \quad x_2, \ldots, x_d \sim \mathcal{N}(2y/\sqrt{d}, 1),$$

where $0.5 < p \leq 0.8$, $d \geq 25$ and $\mathcal{N}(2y/\sqrt{d}, 1)$ is the univariate Gaussian distribution with mean $2y/\sqrt{d}$ and variance 1.
To demonstrate the impact of enforcing robustness, we focus on using linear classifiers for classification. More specifically, let $w \in \mathbb{R}^d$, the classifier we are going to use is given by

$$f_w(x) := \text{sgn}(w^\top x).$$

### 2.1 [10pts]

Show that there exists a linear classifier $w_n \in \mathbb{R}^d$, such that the standard accuracy of $f_{w_n}(\cdot)$ on this problem is at least 0.85.
Hint: You can use the following tail bound for Gaussian distribution. Suppose $Z \sim \mathcal{N}(\mu, \sigma^2)$, then $\Pr(Z - u \leq -t\sigma) \leq \exp(-t^2/2)$ for any $t \geq 0$.

### 2.2 [15pts]

Now let's consider the $\ell_\infty$ norm attack with budget $\epsilon = \frac{4}{\sqrt{d}}$ for this problem. In particular, we are interested in finding a linear classifier $w_r \in \mathbb{R}^d$ that minimizes the following robust error under the $\ell_\infty$ ball:

$$w_r = \arg\min_w \ell_r(w) := \arg\min_w \mathbb{E}\left[\max_{\|\Delta x\|_\infty \leq \frac{4}{\sqrt{d}}} \ell_{01}(f_w(x + \Delta x), y)\right],$$

where $\ell_{01}(\hat{y}, y)$ is the 0-1 loss function which equals 0 iff $\hat{y} = y$ otherwise 1. We call $\ell_r(w)$ the *robust error* of $w$.

#### 2.2.1 [10pts]

Prove that for any $w \in \mathbb{R}^d$ such that $\exists i \geq 2$, $w_i \neq 0$, there exists $w' \in \mathbb{R}^d$ so that $\ell_r(w') < \ell_r(w)$.

#### 2.2.2 [5pts]

Based on the result in 2.2.1, find $w_r$ as well as $\ell_r(w_r)$.

### 2.3 [5pts]

Compute the standard error of $f_{w_r}(\cdot)$, i.e., $\mathbb{E}[\ell_{01}(f_{w_r}(X), Y)]$. Note: compare the standard error of this robust classifier with the one from 2.1. You will be able to see that provably there is a non-zero gap in terms of the standard accuracy between the robust classifier and the original classifier.

## 3 Certified Robustness via Mixed Integer Linear Programming [30pts]

A mixed integer linear program (MILP) over an optimization variable $x \in \mathbb{R}^d$ is an optimization problem of the following form:

$$\begin{aligned} \min_x \quad & c^\top x \\ \text{subject to} \quad & Ax = b, \\ & x \geq 0, \\ & x_i \in \mathbb{Z}, \forall i \in \mathcal{I}, \end{aligned}$$

where $\mathcal{I} \subseteq [d]$ is a subset of $\{1, \ldots, d\}$ that indicates the subset of optimization variables that are constrained to take integer values. In particular, if $|\mathcal{I}| = 0$, then the above MILP reduces to a linear program (LP); on the other hand, if $|\mathcal{I}| = d$, then it becomes a pure integer program (IP). Including

integer variables increases enormously the modeling power, at the expense of computational complexity. In fact, as we briefly discussed in class, LPs can be solved in *polynomial time* with interior-point methods (ellipsoid method, Karmarkar's algorithm, etc.). However, IP is an NP-hard problem, so there is no known polynomial-time algorithm.
In this question we will explore how to formulate the certified robustness problem of a two-layer ReLU network so that it is equivalent to solve a mixed integer linear program (under certain boundedness constraint on the input variable). More specifically, the network we are going to work with has the following form:

$$\hat{y} = \sigma\left(W_2 \cdot \mathrm{ReLU}(W_1 x)\right),$$

where $\sigma(\cdot)$ is the softmax function, i.e., $\sigma(t) = \left(\frac{\exp(t_1)}{\sum_i \exp(t_i)}, \ldots, \frac{\exp(t_k)}{\sum_i \exp(t_i)}\right)^\top \in \mathbb{R}^k$. For this network we assume the input $x \in \mathbb{R}^d$, $W_1 \in \mathbb{R}^{p \times d}$ and $W_2 \in \mathbb{R}^{k \times p}$.

### 3.1 [10pts]

As a starter, we will prove that under certain (restrictive) conditions, we can efficiently solve the following optimization problem involving a two-layer ReLU network.

$$\begin{aligned} \min_{t,x} \quad & c^\top t \\ \text{subject to} \quad & t = \mathrm{ReLU}(Ax). \end{aligned}$$

In the optimization problem above, $x \in \mathbb{R}^d$ is the input variable, $t \in \mathbb{R}^p$ is the feature vector of the hidden layer, and $c \in \mathbb{R}^p$ is the linear weight vector of the output layer. Prove that if $c \geq 0$, the above optimization problem could be solved via LP.

### 3.2 [20pts]

Given a two-layer ReLU network, recall that in order to certify robustness, it suffices if we can solve the following targeted attack problem:

$$\begin{aligned} \min_{z_1, z_2} \quad & (e_y - e_t)^\top (W_2 z_2) \\ \text{subject to} \quad & z_2 = \mathrm{ReLU}(W_1 z_1), \\ & \|z_1 - x\|_\infty \leq \epsilon, \end{aligned} \tag{1}$$

where both $e_y, e_t$ are one-hot vectors corresponding to the ground-truth and the targeted classes.

#### 3.2.1 [5pts]

In general, given network weights $W_2$ and $e_y, e_t$, could we use the same strategy as the one in 3.1 to equivalently transform the above optimization problem into an LP? Explain why.

### 3.2.2 [10pts]

Now consider the constraint $t = \text{ReLU}(x)$, under the assumption that $l \leq x \leq u$ for some constants $l \leq u$ (i.e., $x$ is bounded), by introducing a binary switching variable $a \in \{0, 1\}$, show that the following two constraints are equivalent:

$$t = \text{ReLU}(x) \iff \begin{cases} t - x & \geq 0, \\ t & \geq 0, \\ au - t & \geq 0, \\ x - (1-a)l - t & \geq 0, \\ a & \in \{0, 1\}. \end{cases}$$

### 3.2.3 [5pts]

Use the construction introduced in 3.2.2 to transform the optimization problem (1) into an MILP. How many auxiliary binary variables have been introduced in this process?

# 4 Basic Properties of Differential Privacy [10pts]

## 4.1 [5pts]

Let $M : \mathcal{X}^n \to \mathcal{Y}$ be a randomized mechanism that takes a dataset $X \in \mathcal{X}^n$ as input and outputs an element $t \in \mathcal{Y}$. Assume $X \sim X'$ are two neighboring datasets, i.e., $X$ and $X'$ only differ in one row. Show that if $M$ is $\epsilon$-differentially private for some $\epsilon > 0$, then for any pair of neighboring datasets $X \sim X'$, the total variation distance between $M(X)$ and $M(X')$ is bounded by $\epsilon$, i.e., $d_{\text{TV}}(M(X), M(X')) \leq \epsilon$.

## 4.2 [5pts]

In this problem we look at datasets that differ in multiple entries, and study the privacy guarantee of applying a differentially private mechanism over these datasets. Formally, Let $M : \mathcal{X}^n \to \mathcal{Y}$ be a randomized mechanism that takes a dataset $X \in \mathcal{X}^n$ as input and outputs an element $t \in \mathcal{Y}$. Suppose $X$ and $X'$ are two datasets of size $n$ that differ in exactly $k$ positions. Show that for any $T \subseteq \mathcal{Y}$, we have

$$\Pr(M(X) \in T) \leq \exp(k\epsilon) \cdot \Pr(M(X') \in T).$$

Note: the above inequality implies that the privacy guarantee of a differentially private mechanism decays gracefully as the distance between two datasets increase.

# 5 Laplace Mechanism in Counting [20pts]

Suppose there are $n$ binary entries in a database, and we are interested in designing an $\epsilon$-differentially private mechanism in counting the active entries in the database, for some fixed constant $\epsilon > 0$. More formally, let $X := (X_1, \ldots, X_n) \in \{0, 1\}^n$ be the entries in the database. Following the Laplace mechanism, we design the following mechanism:

$$M(X) = \frac{1}{n} \sum_{i=1}^{n} X_i + Z,$$

where $Z \sim \text{Lap}(1/n\epsilon)$ is a random variable drawn from the Laplace distribution with location and scale parameters as 0 and $1/n\epsilon$, respectively.

### 5.1 [10pts]

Using Chebyshev's inequality, show that with probability $\geq 0.95$, the following inequality holds:

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{10}{n\epsilon} \leq M(X) \leq \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{10}{n\epsilon}.$$

Recall that the error bound on the same problem we obtained using the idea of randomized response, is (roughly) on the order of $O(1/\sqrt{n}\epsilon)$. One can see that the error bound using the Laplace mechanism is quadratically better than the one from randomized response.

### 5.2 [5pts]

Derive the following equality to bound the tail probability of a Laplace random variable:

$$\Pr\left(|Z| \geq \frac{t}{n\epsilon}\right) = \exp(-t),$$

for any $t > 0$.

### 5.3 [5pts]

In fact the probabilistic guarantee we obtained in 5.1 is pessimistic: since we know exactly how the noise is injected and the distribution of the noise, we could probably say more. Using the conclusion from 5.2, show that with probability $\geq 0.95$, the following bound holds:

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{3}{n\epsilon} \leq M(X) \leq \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{3}{n\epsilon}.$$

Note: the order of the error we obtained is still $O(1/n\epsilon)$, same as the one we obtained using Chebyshev's inequality in 5.1. However, the constant dependency is better.