# Lecture # 24: Learning in the presence of adversaries

*Instructor: Aditya Bhaskara* *Scribe: Meysam Alishahi*

**CS 5966/6966: Theory of Machine Learning**

*April 12th, 2022*

**Abstract**

In this lecture, we discuss some high-level ideas regarding learning in the presence of adversaries.

## 1 Introduction

The regular strategy in learning is to use data to train a model and then do the inference using the trained model. We assume that there is an adversary who wants to affect the learning process in a harmful way. The adversary can interrupt the process in either training time or test time. It is worth noting that the adversary takes advantage of obtainable model information and uses it to create malicious attacks.

## 2 Adversarial machine learning strategies

In most machine learning techniques, we have the assumption that training and test data are generated from the same statistical distribution (IID). It is a reasonable assumption that during the preparation of real-world data, the adversaries may interfere in violating that statistical assumption. We can categorize adversarial machine learning strategies in the two following aspects.

- **Training Time interference.** In this scenario, we assume that the adversary can corrupt a small fraction of inputs. For example, (s)he can add some new data points to the training set (in unsupervised learning like clustering) or mislabels some of the data points (in supervised learning like classification).
- **Test time interference.** In this scenario, we supposed that the adversary evaluates the model on inputs with "imperceptible error" added. It can be viewed as having different input distribution and test distribution.

## 3 Noise in Training Data

### 3.1 *Benigh Noise*

- As a very common assumption, we assume that the given data is perturbed by a small amount of noise (i.e., mean zero noise following standard Gaussian distribution). This added noise is not designed to harm

the model. For instance, in regression, we assume that the label could be noisy:

$$y = \langle a, x \rangle + \eta \qquad \eta \sim \mathcal{N}(0, \varepsilon).$$

In classification, we might have some data points mislabeled due to noise, or in clustering, we could observe data points perturbed by an unknown noise.

- Another less common assumption is that some (small fraction) of the samples (uniformly random subset) from the underlying distribution can be arbitrarily (badly) corrupted (Huber's contamination model). A basic problem, under this assumption, is to estimate robustly the mean of the underlying distribution. One can imagine that under this assumption the median cannot change a lot so it can be still a reasonable choice for estimating the mean.

### 3.2 *Adversarial noise (data poisoning)*

In this scenario, the data points might be corrupted in a structured way while the noise is random. Indeed, the adversary is aware of the model we use and can adapt his (her) strategy as most effectively as possible considering the properties of the model. As a simple case, finding the mean of a Gaussian with given covariance can be easily misled. In this situation, even basic problems such as robust mean estimation for Gaussian data and robust PCA turn out to be very difficult. Since, lots of problems can be reduced to mean estimation, lots of works in this area are on robust mean estimation. For instance, solving some loss-min problems can be reduced (in practice) to mean estimation (a gradient descent algorithm can be viewed as a sequence of mean estimation steps). This problem can formally be defined as follows:

**Mean estimation problem.** There is an unknown distribution $\mathcal{D}$ with bounded variance and we are given $x_1, \ldots, x_N \sim \mathcal{D}^\varepsilon$ and the goal is to estimate the mean of $\mathcal{D}$. Here $\mathcal{D}^\varepsilon$ means that the $\varepsilon$ fraction of the data is corrupted by the adversary.

## 4 Classic Algorithms

- **Mean estimation in low dimensions (median vs mean).** If only a small fraction of data (sat 1%) is corrupted, then although the mean could change dramatically, the median usually won't change a lot so it could be considered a good estimator for the mean.
- **Inlier pursut.** The key idea in this algorithm is the fact that the inliers reinforce one another. Roughly speaking, assume that we have a guess for mean and variance. We check for consistency and throw away the extremes (say 1% of the data). Now, we can re-estimate our guesses for mean and variance.

## 5 Problem of dimensionality

In the high dimensional mean estimation problem, we have $n$ clean data from distribution in $d$ dimensions (say Gaussian with mean $\mu$ and covariance matrix $\Sigma$). Also, assume that $\varepsilon n$ points of the data are replaced with

some adversary chosen points. The goal is to recover the parameters $\mu'$ and $\Sigma'$ so that $\mathcal{N}(\mu, \Sigma) \approx \mathcal{N}(\mu', \Sigma')$. It is known that using exponential time, this can de be done to $O(\varepsilon)$ error assuming that $n$ is polynomially large enough. In 1-dimensional case, it is a fairly easy problem. For $d$-dimensional case and Gaussian distributions, a result by Diakonikolas et al. (2016) and Lai et al. (2016) indicates that there is an algorithm that can efficiently recover the mean to the error which is rough $O(\varepsilon\sqrt{\log 1/\varepsilon})$. This algorithm can be extended to arbitrary distributions as well.