CS880: Advanced Topics in Learning Theory Lecture 1 - [Unofficial]

University of Wisconsin–Madison, Fall 2023

Lecture 1, 2 - Sample Mean Estimator, Contamination Models

Lecturer: Ilias Diakonikolas *Scribed by: Deep Patel*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the instructor.*

In this course, we will be designing and analyzing statistically and computationally efficient learning algorithms in well-defined models. Specifically, we will be considering situations wherein the data is 'corrupted' with noise. Towards this, we start by analysing the properties of the classical estimator – sample mean estimator – before highlighting the fact that it's not suitable for estimation tasks when the data is 'corrupted'. We will then define various data corruption or contamination models.

1 Properties of the Sample Mean Estimator

Problem to be solved: Given N i.i.d. samples from a distribution D, find an accurate estimate, $\hat{\mu}$, of the expected value $\mu = \mathbb{E}[D]$.

As an example, say we are given $X_1, \ldots, X_N \overset{i.i.d.}{\sim} \mathcal{N}(\mu, 1)$, we can use the arithmetic mean as a statistic to estimate the mean (recall that this is as per the Maximum Likelihood principle). That is,

$$\hat{\mu} \triangleq \frac{1}{N}\left(\sum_{i=1}^{N} X_i\right)$$

Remark 1.1 (What do we want?)**.** *What we would want for any proposed estimator is the following: "As N becomes large $\Rightarrow$ $\hat{\mu}$ should be close to $\mu = \mathbb{E}[D]$ with high probability".*

For the sample mean estimator $\hat{\mu}$ defined above, we can note the following using standard concentration inequalities:

$$\mathbb{P}\left[\left|\underbrace{\frac{1}{N}\sum_{i=1}^{N} X_i - \mu}_{\mathcal{N}(0,\frac{1}{N})}\right| > \epsilon\right] \leq 2e^{-\frac{1}{2}\epsilon^2/N}$$

Alternatively, we can write the following:

$$\mathbb{E}[|\hat{\mu} - \mu|^2] = \mathrm{var}(\hat{\mu}) = \frac{1}{N} \; (\to 0 \text{ as } N \to \infty)$$

$$\Rightarrow |\hat{\mu} - \mu|^2 \leq \frac{1}{N\delta} \text{ w.p. } \geq 1 - \delta \;\; (\because \text{ By Markov's inequality})$$

Here, we made use of the Markov's inequality which is applicable for any non-negative random variable:

$$\mathbb{P}[Z \geq \lambda \mathbb{E}[Z]] \leq \frac{1}{\lambda} \;\; \forall \, \lambda > 0$$

Remark 1.2. *We can get a stronger bound – $\mathcal{O}\left(\log \frac{1}{\delta}\right)$ bound – instead of $\mathcal{O}\left(\frac{1}{\delta}\right)$ obtained above by using stronger concentration bounds.*

Remark 1.3. *We can carry out the same analysis that we did above for multivariate standard normal distributions. Given $X_1, \ldots, X_N \overset{i.i.d.}{\sim} \mathcal{N}(\mu, I_{d\times d})$, we can use the same arithmetic mean as a statistic to estimate the mean. That is,*

$$\hat{\mu} \triangleq \frac{1}{N}\left(\sum_{i=1}^{N} X_i\right)$$

Thus, $\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{1}{N} I_{d\times d}\right)$, after which we can use Markov's inequality in the same manner:

$$\mathbb{E}[\|\hat{\mu} - \mu\|_2^2] = \sum_{j=1}^{d} \mathbb{E}[(\hat{\mu}_j - \mu_j)^2] = d \cdot \frac{1}{N}$$

$$\Rightarrow \|\hat{\mu} - \mu\|_2^2 \leq \frac{d}{\delta N} \text{ w.p. } \geq 1 - \delta \ (\because \textit{By Markov's inequality})$$

Suppose we want $\|\hat{\mu} - \mu\|_2 \leq \epsilon$ w.p. $\geq 9/10$, then $N > \frac{10d}{\epsilon^2}$.

Although sample mean estimator has many desirable properties like the one we see here, it is not always the best mean estimator. For instance, even if a small fraction of the observations are contaminated, then one can make the sample mean estimate deviate by as large a value as we want it to. Throughout this course we will look at situations like these wherein a fixed, small fraction of data can be corrupted in any manner whatsoever. We will make this notion of 'contamination' mathematically precise and define different models of contamination. Having done that, we will design and analyze statistically and computationally efficient learning algorithm in well-defined models.

Remark 1.4. *In standard i.i.d. settings, we can get consistent estimators (i.e. when $n \to \infty$, error probability $\to 0$). As we will see, this is NOT true in the case of "contaminated"[1] observations.*

2 Contamination Models

Now, let's formally define what 'contamination'. In particular, let $\epsilon \in (0, 1/2)$ be the proportion of contamination. Then, roughly, the idea is that the input data contains $\lceil (1-\epsilon)n \rceil$ datapoints that come from true distribution (clean data/inliers). Therefore, $\lceil \epsilon n \rceil$ datapoints can come from anything!

Remark 2.1. *A natural question to ask is why do we assume $\epsilon < 1/2$. Intuitively, this is because the outliers will completely swallow signal from inliers otherwise.*

Let's take an example. Suppose $\epsilon = 1/2$ and we want to estimate mean of $\mathcal{N}(\mu, I)$. Then if there are two clusters consisting of 50% data each with means μ_1 and μ_2 but separated by a large distance between μ_1 and μ_2, then it's hard to say whether the mean is μ_1 or μ_2.[2]

[1] we will make these notions precise below

[2] Note that there IS something useful we can do in this case of 2-clusters – List-decodable learning – which we will cover this towards the end of this course.

2.1 Huber Contamination Model

This is the simplest contamination model and can be essentially viewed as a mixture model. It's also referred to as additive and non-adaptive contamination model.

Definition 1 (Huber Contamination Model)**.** *Given $\epsilon \in (0, 1/2)$ and a distribution D on inliers, an ϵ-Huber corrupted distribution is a distribution X of the form*

$$X = (1 - \epsilon)D + \epsilon N$$

where N is the noise distribution and it's unknown to us.

Remark 2.2. *The notation used above denotes a mixture model. That is, to draw a sample from the distribution X, we proceed as follows:*

With probability $1 - \epsilon$ draw a sample from D and with probability ϵ draw a sample from N.

Thus, roughly, the Huber model corresponds to the following procedure:

1.) First draw $\lceil (1 - \epsilon)n \rceil$ i.i.d. samples S from D.

2.) Then draw $\lceil \epsilon n \rceil$ i.i.d. samples O from N.

3.) Then randomly reorder $S \cup O$ samples and give them to the algorithm as input.

Remark 2.3. *What does non-adaptive mean here? It means that points added in Step-2 above do NOT depend on points obtained in Step-3. However, the distribution N can still be adversarially chosen based on the distribution D of the clean data.*

Definition 2 (Additive & Adaptive Contamination Model)**.** *A dataset is ϵ-additive and adaptive contaminated if it is obtained as follows:*

1.) We draw $\lceil (1 - \epsilon)n \rceil$ i.i.d. samples S from D

2.) An adversary observes the samples from Step-1 and then comes up with $\lceil \epsilon n \rceil$ outliers O.

3.) Randomly reorder $S \cup O$ samples and give them to algorithm as input.

Remark 2.4. *The final corrupted dataset is NOT a set of i.i.d. samples from ANY distribution.*

An example of additive and adaptive contaminations is data poisoning attacks in adversarial and secure ML (ICML'12)[3]. Apart from adding data points, adversary can remove or subtract clean data as well. This subtractive contamination can be either adaptive or non-adaptive just like in the case of additive contamination.

More generally, one can have general contamination wherein the adversary can both add and subtract data points. Here as well, one can have an adaptive or non-adaptive adversary.

Definition 3 (General & Non-adaptive contamination)**.** *A dataset is ϵ- general, non-adaptive contaminated if it's obtained as i.i.d. samples from the following distribution*

$$X = D - \epsilon L + \epsilon E \tag{1}$$

where X, D, L, E denote probability density functions (PDFs) of the distributions[4] and L, E are unknown noise distributions.

[3] Biggio, Battista, Blaine Nelson, & Pavel Laskov. (2012) "Poisoning attacks against support vector machines." In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1467-1474.

[4] For simplicity, we will assume that the random variables are continuous

Note that for X to be a valid PDF in Equation 1, we would require $\epsilon L \leq D$ as PDFs can't take negative values.

Remark 2.5 (Characterization via Total Variation Distance)**.** *This leads to a question as to which distributions X one can obtain in the general, non-adaptive contamination model. It turns out that these are the distributions which are close to the inlier distribution D in the sense of total variation distance. That is, the general, non-adaptive model is equivalent to sayiing that one can sample from a distribution X with $d_{TV}(X, D) \leq \epsilon$. This can be seen by using an alternate characterization of total variation – $d_{TV(X,Y)} = \inf_{A\sim X, B\sim Y} \mathbb{P}[A \neq B]$. Since the PDFs X, D differ at a set of measure $\leq \epsilon$, we have the claim that $d_{TV}(X, D) \leq \epsilon$.*

One can accordingly defined an adaptive version of general contamination.

Definition 4 (General & Adaptive (Strong) Contamination)**.** *We say that a dataset S is ϵ-general and adaptive (or ϵ-strong) corrupted from distribution D, if S is obtained as follows:*

1. *A set I of n i.i.d. samples from D is obtained*
2. *An adversary can remove up to $\lceil \epsilon n \rceil$ points from I and replace them by arbitrary points O*
3. *Randomly reorder these samples and give them to algorithm as input.*