Lecture 2: Total variation, statistical models, and lower bounds

October 2, 2019

In the first lecture we saw two algorithms for robustly learning the mean of a univariate Gaussian with known covariance. One of these two algorithms, the median, was able to achieve an asymptotic error of $O(\varepsilon)$, given an ε-corrupted set of samples from the Gaussian. The second, the truncated mean, attains error $O(\sqrt{\varepsilon})$ when the distribution has bounded second moment. In this lecture, we will show that in fact both error rates are optimal for their respective problems. To do so we will need to introduce a statistical notion of ε-corruption, and some notions of distances between distributions which will be very useful throughout this class.

1 Total variation distance

As it turns out, the notion of learning in the presence of gross corruption is very intimately related to the notion of *total variation* distance. For any probability distribution P and any event A, let $P(A) = \Pr_{X\sim P}[X \in A]$ denote the probability mass that P puts on A.

Definition 1.1 (Total variation distance)**.** Let P, Q be two probability distributions over a shared probability space Ω. Then, the *total variation distance* between them, denoted $\mathrm{d_{TV}}(P, Q)$, is given by

$$\mathrm{d_{TV}}(P, Q) = \sup_{A\subseteq\Omega} |P(A) - Q(A)| \,. \tag{1}$$

It is left to the reader to check that this is in fact a metric over probability distributions. There are many equivalent formulations of total variation distance that we will use interchangeably throughout this class. Before we state them, we first recall the definition of *coupling*, an important concept in probability theory:

Definition 1.2 (Coupling)**.** Let P, Q be two probability distributions over probability spaces Ω_1, Ω_2, respectively. A random variable $Z = (X, Y)$ on $\Omega_1 \times \Omega_2$ is a *coupling* of P and Q if the marginal distribution of X is distributed as P and the marginal distribution of Y is distributed as Q.

We now have the following theorem. It is stated for distributions over $\mathbb{R}^d$ but also holds more generally:

Theorem 1.1. *Let P, Q be two probability distributions over $\mathbb{R}^d$, with probability distribution functions p, q, respectively. Then:*

(i) We have

$$\mathrm{d_{TV}}(P, Q) = \sup_{A\subseteq\mathbb{R}^d} P(A) - Q(A) = \sup_{A\subseteq\mathbb{R}^d} Q(A) - P(A) \,. \tag{2}$$

(ii) We have

$$\mathrm{d_{TV}}(P, Q) = \frac{1}{2}\int |p(x) - q(x)|dx \,. \tag{3}$$

(iii) We have $\mathrm{d_{TV}}(P, Q) = \inf_{(X,Y)} \Pr[X \neq Y]$, where the supremum is taken over all (X, Y) which are couplings of P and Q.

Proof. We first prove (i). To do so, we observe that the set A which achieves the supremum in the definition of total variation distance is given by the set

$$S_+ = \{x \in \mathbb{R}^d : p(x) \geq q(x)\} ,$$

as well as the set

$$S_- = \{x \in \mathbb{R}^d : q(x) < p(x)\} .$$

It is then easily verified that S_+ also achieves the supremum for the middle expression in (2), and S_- achieves the supremum for the last expression in (2). Moreover, observe that

$$\begin{aligned}\int |p(x) - q(x)|dx &= \int_{S^+} p(x) - q(x)dx + \int_{S^-} q(x) - p(x)dx \\ &= P(S^+) - Q(S_+) + Q(S_-) - P(S_-) = 2\,\mathrm{d_{TV}}(P, Q) ,\end{aligned}$$

which proves (ii). We now prove (iii). If $\mathrm{d_{TV}}(P, Q) = 1$, the claim is obvious (why?). Thus without loss of generality we may assume that $\mathrm{d_{TV}}(P, Q) < 1$ (Notice that $\mathrm{d_{TV}}$ is always bounded between 0 and 1). We first demonstrate a coupling (X, Y) so that $\Pr[X \neq Y] = \mathrm{d_{TV}}(P, Q)$. We do so as follows. Let $r(x) = p(x)$ for $x \in S_-$, and $r(x) = q(x)$ for $x \in S^+$. This function is well-defined, non-negative, and moreover $r(x) \leq p(x)$ and $r(x) \leq q(x)$ for all $x \in \mathbb{R}^d$. Moreover, define $p_1(x) = p(x) - r(x)$ and $q_1(x) = q(x) - r(x)$. Then, these two functions are also non-negative. We have $\int p_1(x)dx = \int q_1(x)dx = \mathrm{d_{TV}}(P, Q)$, and

$$1 - \int r(x)dx = \int p(x) - r(x)dx = \int_{S_+} p(x) - q(x)dx = \mathrm{d_{TV}}(P, Q) .$$

Therefore $\frac{1}{1-\mathrm{d_{TV}}(P,Q)} r(x)$ is a valid pdf for a distribution R, and $\frac{1}{\mathrm{d_{TV}}(P,Q)} p_1(x)$ and $\frac{1}{\mathrm{d_{TV}}(P,Q)} q_1(x)$ are valid pdfs for probability distributions P_1 and Q_1, respectively. The coupling is now given as follows: we draw X, Y as follows: with probability $1 - \mathrm{d_{TV}}(P, Q)$, we sample $Z \sim R$, and let $X = Y = Z$. Otherwise, we sample $X \sim P_1$ and $Y \sim P_2$. Then it is clear that $\Pr[X \neq Y] = \mathrm{d_{TV}}(P, Q)$, and it is a simple calculation to check that this is indeed a valid coupling.

We have now shown that $\mathrm{d_{TV}}(P, Q) \geq \inf_{(X,Y)} \Pr[X \neq Y]$. We now show the opposite inequality. Let (X, Y) be any valid coupling of P and Q, and let $\alpha = \Pr[X \neq Y]$. Consider the four events $E_{++}, E_{+-}, E_{-+}, E_{--}$, defined by

$$E_{ab} = \{X \in S_a, Y \in S_-\} , \qquad a, b \in \{+, -\} .$$

Observe that $\Pr[E_{+-}] \leq \alpha$. Then we have

$$\begin{aligned}\mathrm{d_{TV}}(P, Q) &= P(S_+) - Q(S_+) = (\Pr[E_{++}] + \Pr[E_{+-}]) - (\Pr[E_{++}] + \Pr[E_{-+}]) \\ &= \Pr[E_{+-}] - \Pr[E_{-+}] \leq \alpha .\end{aligned}$$

Thus $\mathrm{d_{TV}}(P, Q) \leq \inf_{(X,Y)} \Pr[X \neq Y]$, which completes the proof. □

2 Models of corruption

Let us now more formally define some notions of corruption that we will consider rather extensively for the next several classes.

2.1 Replacement corruption

The strongest form of adversary, and the one that we will spend most of our time on, is the *full* or *replacement* corruption model. Roughly speaking, given a dataset $X_1, \ldots, X_n$, the full adversary is allowed to inspect the data points, then change an ε-fraction of these points arbitrarily. We will call such a set of points ε-corrupted:

Definition 2.1 (ε-corruption)**.** A set of points S of size n is *ε-corrupted* from a distribution D if it is generated via the following process:

1. First, n samples $Y_1, \ldots, Y_n$ are drawn i.i.d. from D.
2. Then, these points are given to an adversary, and the adversary is allowed to inspect them.
3. Based on the inspection, the adversary is allowed to change an ε-fraction of the Y_i arbitrarily.
4. The corrupted set of points are then returned to us in any order.

Equivalently, we say that S is ε-corrupted if $S = S_{\text{good}} \cup S_{\text{bad}} \setminus S_r$ where S_{good} is a set of n i.i.d. samples from D, we have $S_r \subset S_{\text{good}}$, and $|S_{\text{bad}}| = |S_r| = \varepsilon n$. Given such a set S, we say $(S_{\text{good}}, S_{\text{bad}}, S_r)$ is the *canonical* decomposition of S.

In a slight abuse of notation, we will often simply write $S = S_{\text{good}} \cup S_{\text{bad}} \setminus S_r$, and it should be understood that we always mean the canonical decomposition. Note that while this canonical decomposition is very useful from an analytic perspective, it is unknown to the algorithm. As an additional shorthand, if $|S| = n$, we will often use identify S with it $[n]$, that is, we will identify $S = \{X_1, \ldots, X_n\}$ with its indices. For instance, we will often write $\widehat{\mu} = \frac{1}{n}\sum_{i \in S} X_i$ to be the empirical mean of the points in S. We will similarly index into $S_{\text{good}}, S_{\text{bad}}$, and S_r. The meaning should hopefully be clear from context.

This adversary is an *adaptive* adversary, that is, it is allowed to inspect the data before choosing the corruptions. Alternatively, we can consider *non-adaptive* versions of the adversary. It turns out that this naturally corresponds to a statistical notion of corruption, coming from the TV distance. In particular, Theorem 1.1 (iii) says that independent samples from a distribution P can be well-simulated by ε-corrupted samples from another distribution with small total-variation distance to P. Formally:

Corollary 2.1. *Let P, Q be so that $\mathrm{d}_{\mathrm{TV}}(P, Q) = \varepsilon$, and let $c > 0$. Let $(X_1, \ldots, X_n)$ be independent samples from P. With probability at least $1 - \exp(-\Omega(c^2 \varepsilon n))$, these are $(1 + c)\varepsilon$-corrupted samples from Q.*

Proof. For each i, let (X_i, Y_i) be the coupling between P and Q so that $\Pr[X_i \neq Y_i] = \mathrm{d}_{\mathrm{TV}}(P, Q)$. Then, by a Chernoff bound, with probability $1 - \exp(-\Omega(c^2 \varepsilon n))$, we have that the number of indices i so that $X_i \neq Y_i$ is at most $(1 + c)\varepsilon n$. Since the Y_i are independent draws from Q, this completes the proof. □

This motivates the following statistical model of corruption:

Definition 2.2. We say that a set of samples $(X_1, \ldots, X_n)$ are an *ε-obliviously corrupted* set of samples from P if the samples are drawn independently from Q, where $\mathrm{d}_{\mathrm{TV}}(P, Q) \leq \varepsilon$.

In this language, the above corollary states that, up to sub-constant factors in the loss, and exponentially small success probabilities, ε-corruption can simulate ε-oblivious corruption.

It is a natural question as whether or not the two models of corruption are equivalent. It is not hard to see that ε-corruption is, at least formally, strictly stronger than ε-oblivious corruption (why?). Intuitively, ε-oblivious corruption corresponds to the model where the adversary must (in a distributional sense) specify the corruptions they wish to make to the samples, ahead of time, by specifying the corrupting distribution. In other words, the corruptions that the adversary chooses can only depend on the distribution of the data, not the data points themselves. This motivates the name of the corruption model. However, for the purposes of this class, and indeed, more or less all of the results in the field, the two models are essentially equivalent.

The main utility of introducing this statistical notion of corruption will be for demonstrating lower bounds. We begin by noting that the following problem is almost vacuously impossible: given two distributions P, Q with $\mathrm{d}_{\mathrm{TV}}(P, Q) = \varepsilon$, and $(X_1, \ldots, X_n)$, decide if $(X_1, \ldots, X_n)$ are ε-obliviously corrupted from P or from Q. Combining this with the above corollary yields:

Corollary 2.2. *Let P, Q be so that $\mathrm{d}_{\mathrm{TV}}(P, Q) = \varepsilon$, and let $c > 0$. Consider the following distinguishing problem: given $X_1, \ldots, X_n$, distinguish if they are $(1 + c)\varepsilon$-corrupted set of samples from P or Q. Then, no algorithm can succeed at this distinguishing task except with probability $1 - \exp(-\Omega(c^2 \varepsilon n))$.*

2.2 Additive corruption

In the full corruption model we are considering, the adversary is allowed to change an ε-fraction of points arbitrarily. This can be thought of as a two-step process. First, the adversary removes an ε-fraction of the good points, then they add in an ε-fraction of corrupted points.

A weaker notion of corruption is that of additive noise, which is what we considered in the previous lecture. Morally speaking, in such models of corruption, the adversary is only allowed to add corruptions, but cannot delete points. We can define this both in the adaptive and in the oblivious setting. In the adaptive case, the definition is straightforward:

Definition 2.3 (Additive corruption)**.** Let P be a distribution over $\mathbb{R}^d$ and let $\varepsilon > 0$. We say that a set of points $X_1, \ldots, X_n$ is *ε-additively corrupted* from P if it is generated through the following process:

(i) First, $(1-\varepsilon)n$ points are drawn i.i.d. from P.

(ii) Then, an adversary is allowed the inspect these points, and can add εn points arbitrarily to this set.

(iii) The points are then returned in an arbitrary order.

As with the general corruption model, there is also a distributional model of additive corruption:

Definition 2.4 (Oblivious additive corruption)**.** Let P be a distribution over $\mathbb{R}^d$ and let $\varepsilon > 0$. We say that a set of points $X_1, \ldots, X_n$ is *ε-obliviously additively corrupted* from P if they are drawn independently from Q, where $Q = (1-\varepsilon)P + \varepsilon N$ for some arbitrary distribution N.

We leave it to the reader to demonstrate that the same sorts of simulation properties between additive and oblivious additive noise exist as in the general model we considered above. It is also a useful exercise to find instances where additive noise cannot simulate the full oblivious adversary.

As a historical note, this model of corruption was the first to be considered by statisticians in the 60s, specifically Huber [1]. As a result, this is also known as *Huber's contamination model* in the statistics literature.

3 Lower bounds for robustly learning Gaussians

We now return to the problem we considered in the first lecture, namely, learning the mean of a Gaussian, and we use our new machinery to demonstrate that in fact the median is optimal. To do so, it now suffices to prove the following:

Theorem 3.1. *Let $\varepsilon > 0$ be sufficiently small, let $\sigma > 0$, and let μ_1, μ_2 be so that $|\mu_1 - \mu_2| = \sigma\varepsilon$. Then*

$$\mathrm{d_{TV}}(\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)) = \left(\frac{1}{\sqrt{2\pi}} + o(1)\right)\varepsilon\ .$$

Before we prove this theorem, note that this theorem in conjunction with Corollary 2.2 implies that $\Omega(\varepsilon\sigma)$ is impossible to improve upon for the problem of learning the mean of a univariate Gaussian.

Proof of Theorem 3.1. Since TV distance is scale-invariant, we may assume without loss of generality that $\sigma = 1$. Now let p_1 be the pdf of $\mathcal{N}(\mu_1, 1)$, and similarly p_2 be the pdf of $\mathcal{N}(\mu_2, 1)$, and assume without loss of generality that $\mu_1 \le \mu_2$. By explicit calculation, we have that $p_1(x) \ge p_2(x)$ if and only if $x \le (\mu_1 + \mu_2)/2$. Therefore

$$\mathrm{d_{TV}}(\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)) = \int_{-\infty}^{(\mu_1+\mu_2)/2} p_1(x) - p_2(x)dx = \Pr_{X\sim\mathcal{N}(0,1)}[X \in [-\alpha, \alpha]]\ ,$$

where $\alpha = (\mu_2 - \mu_1)/2 = \varepsilon/2$. For ε small, this quantity is $\left(\frac{1}{\sqrt{2\pi}} + o(1)\right)\varepsilon$, as claimed. □

This lower bound, as written, only holds in the (full) oblivious adversary setting. However, the analysis of the median we had before was for the additive adversary setting, so this doesn't quite complete the picture for this problem. In the homework, we will do so: we will first show that the median also works with replacement corruptions, and we will see a lower bound for additive corruption.

4 Lower bounds for robustly learning the mean with bounded second moment

Here we also show that the truncated mean is optimal up to constants for the other setting considered in the previous lecture. In particular, we will show:

Theorem 4.1. *There exist two distributions D_1, D_2 so that:*

(i) both distributions have variance at most σ^2

(ii) we have $D_2 = (1-\varepsilon)D_1 + \varepsilon N$ for some distribution N

(iii) if μ_1, μ_2 are the means for D_1, D_2, respectively, then $|\mu_1 - \mu_2| > \sigma\sqrt{\varepsilon}$.

In particular, this implies that samples from D_2 are ε-obliviously corrupted samples from D_1. Since they could also be ε-corrupted samples from D_2 itself, no algorithm can achieve error for this problem better than the gap between the means, which is $\gtrsim \sqrt{\varepsilon}$.

Proof of Theorem 4.1. By scaling the distributions, it will suffice to demonstrate this for $\sigma = 1$. The two distributions will be as follows: D_1 is simply the point mass at 0, and D_2 is $(1-\varepsilon)D_1 + \varepsilon N$, where N is a point mass at $1/\sqrt{\varepsilon}$. Clearly by construction, (ii) holds. We now check (i) and (iii). The distribution D_1 has mean 0 and variance 0. The distribution D_2 has mean $\sqrt{\varepsilon}$ and variance $1-\varepsilon$. These things together imply (i) and (iii), by a straightforward calculation. □

References

[1] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.