CS880: Advanced Learning Theory **Fall 2019**

Lecture 4: Sample Complexity Lower Bounds

Lecturer: Ilias Diakonikolas *Scribes: Vasilis Kontonis*

1 Recap

Denote by $\mathcal{U}_n$ the uniform distribution supported on n elements. In the previous lectures we showed how to design an algorithm that with $m = O(\sqrt{n}/\epsilon^2)$ samples from a distribution D, with probability 99%, outputs "YES" if $D = \mathcal{U}_n$ and "NO" if $d_{\mathrm{TV}}(D, \mathcal{U}_n) \geq \epsilon$. What happens if we ignore this and feed our algorithm the same number of samples m from a distribution D that is closer to $\mathcal{U}_n$ e.g. $d_{\mathrm{TV}}(\mathcal{U}_n, D) = \epsilon/1000$? The answer is simply that the behavior of our algorithm is unspecified in this regime. The above guarantee requires *a priori* that $d_{\mathrm{TV}}(\mathcal{U}_n, D) \geq \epsilon$.

The above testing problem consists of two classes: the null hypothesis $D = \mathcal{U}_n$ and the alternative hypothesis $d_{\mathrm{TV}}(D, \mathcal{U}_n) \geq \epsilon$. The null hypothesis only consists of one distribution: it is a *simple hypothesis*. On the other hand, the alternative hypothesis contains infinitely many distributions and is called a *composite hypothesis*.

Another important problem is how much one can relax the assumption that the null hypothesis $D = \mathcal{U}_n$ is *exactly* a uniform distribution. One possible variant is to test distributions that are very close to uniform, i.e. in total variation distance $O(\epsilon/\sqrt{n})$ vs distributions that are ϵ-far from being uniform. Notice that now the null hypothesis is also composite because infinitely many distributions are close to uniform distribution. It is left as an exercise to show that the collision based tester, that we analyzed for the simple null hypothesis also handles this case.

A more challenging problem is whether we can relax the null hypothesis to contain all distributions in total variation less than $\epsilon/2$. This turns out to require almost linearly many samples, $O(n/(\log(n)\epsilon^2))$.

2 Lower Bounds

The main question that we will consider in this lecture is the following sample complexity lower bound.

Question 1. *How many samples are* ***necessary*** *to learn in total variation distance ϵ a discrete distribution supported on n elements?*

The standard way of proving sample complexity lower bounds is by using information theoretic arguments based on shared (mutual) information. Instead of using information theory, we will present an 'ad-hoc' argument showing that $\Omega(n/\epsilon^2)$ samples are needed to learn a discrete distribution supported on n-elements.

A standard way of proving sample complexity lower bounds is by discretizing the space of distributions and then providing a lower bound for the (easier) problem of learning a distribution in this discretized space. A standard notion that is used in this context is that of an ϵ-packing. Although, this notion can be defined for general metric spaces, for concreteness, we will give the definition for the metric space of distributions with total variation as metric.

Definition 2 (ϵ-packing)**.** *Let $\mathcal{D}$ be a set of distributions. A set $\mathcal{D}_\epsilon \subseteq \mathcal{D}$ is an ϵ-packing of $\mathcal{D}$, if for any pair $P, Q \in \mathcal{D}_\epsilon$ it holds $d_{\mathrm{TV}}(P, Q) \geq \epsilon$.*

Now, we show that learning a distribution from a class $\mathcal{D}$ is at least as hard as *choosing* a hypothesis from a finite 2ϵ-packing of $\mathcal{D}$. More precisely, fix a 2ϵ-packing $\mathcal{S}$ of cardinality k and consider the following random experiment

1. Pick one distribution D uniformly at random from $\mathcal{S}$. Let X be the random variable that corresponds to the index of the chosen distribution D.

2. Given m samples from the chosen distribution D can you find the index X?

It is now clear that if we have an algorithm that, with m samples, can learn a distribution from the class $\mathcal{D}$ we can use it to also solve the above problem. Simply run this algorithm to get a distribution $\widetilde{D}$ (this does not necessarily belongs in $\mathcal{S}$) that is ϵ-close to the target D and then choose the distribution S^* of $\mathcal{S}$ that is closest to $\widetilde{D}$. Using the fact that $\mathcal{S}$ is a 2ϵ-packing and the triangle inequality we see that apart from S^* all other distributions of $\mathcal{S}$ are at least ϵ-far from $\widetilde{D}$. Thus, S^* is the unique distribution of $\mathcal{S}$ inside the ball of radius ϵ around $\widetilde{D}$ and therefore is the correct answer to the previous problem. Intuitively it is clear that we would like to "pack" as many distributions as possible in our set because usually more distributions make it harder to find the one that generated the samples. The difficulty is that the more distributions ones wants to "pack" the harder it gets to have all the pairwise distances be at least ϵ. Notice, however, that there are pretty large packings whose corresponding testing problem is trivial. Consider, for example, the family of n discrete distributions each of which assigns probability mass 1 to element i. All these distribution are in total variation distance 1 from each other but with 1 sample we can always identify the one that was picked in the above random experiment.

2.1 Distinguishing Coins

Before we answer 1 we will first consider the fundamental task of distinguishing two Bernoulli random variables, i.e. two coins. As we will see the initial problem of learning discrete distributions can in fact be "reduced" to the following question.

Question 3. *How many samples are* ***necessary*** *to distinguish between a coin whose "heads" probability is* $1/2 + \epsilon/2$ *from a coin whose "heads" probability is* $1/2 - \epsilon/2$*?*

The following lemma connects the hardness of choosing between two distributions with their total variation.

Lemma 4 (Testing Error and Total Variation)**.** *Let* P_0, P_1 *be two distributions on a domain* S*. Let* $\mathcal{A}$ *be a (deterministic) algorithm that given* k *samples from* P_i *with* $i = 0, 1$*, decides whether the samples come from* P_0 *or* P_1*, that is* $\mathcal{A}$ *finds* i *with probability at least* $1 - \delta, \delta < 1/2$*. Then* $d_{\mathrm{TV}}(P_0^k, P_1^k) \geq 1 - 2\delta$*.*

Proof. Our algorithm gets a sample vector $(X_1, \ldots, X_k) \in S^k$. We partition the domain of the samples S^k as $S_0 = \{x \in S^k : \mathcal{A}(x) = 0\}$, $S_1 = \{x \in S^k : \mathcal{A}(x) = 1\}$. So S_0 resp. S_1 contain all the sample vectors where our algorithm answers 0 resp. 1. Since the error probability of $\mathcal{A}$ is at most δ we have

$$\mathbb{P}_{X \sim P_1^k}[\mathcal{A}(X) = 0] = \mathbb{P}_{X \sim P_1^k}[X \in S_0] \leq \delta \qquad \mathbb{P}_{X \sim P_0^k}[\mathcal{A}(X) = 1] = \mathbb{P}_{X \sim P_0^k}[X \in S_1] \leq \delta$$

Adding these inequalities we obtain

$$\begin{aligned} 2\delta &\geq \mathbb{P}_{X \sim P_1^k}[X \in S_0] + \mathbb{P}_{X \sim P_0^k}[X \in S_1] \\ &= 1 + \mathbb{P}_{X \sim P_1^k}[X \in S_0] - \mathbb{P}_{X \sim P_0^k}[X \in S_0] \end{aligned}$$

Rearranging and using the definition of total variation, we get,

$$\begin{aligned} d_{\mathrm{TV}}(P_0^k, P_1^k) &= \sup_{W \subseteq S} (\mathbb{P}_{X \sim P_0^k}[X \in W] - \mathbb{P}_{X \sim P_1^k}[X \in W]) \\ &\geq \mathbb{P}_{X \sim P_0^k}[X \in S_0] - \mathbb{P}_{X \sim P_1^k}[X \in S_0] \\ &\geq 1 - 2\delta \end{aligned}$$

□

Remark 1. *Observe that Lemma 4 also implies that if the distribution that generates the samples is chosen uniformly at random, with* $V \in \{0, 1\}$ *being the chosen index and* $X \sim P_V^k$ *the random variable corresponding to the sample, then*

$$\mathbb{P}_{V,X}[\mathcal{A}(X) \neq V] \geq \frac{1}{2} - \frac{d_{\mathrm{TV}}(P_0^k, P_1^k)}{2}$$

We prove the following lemma for discrete distributions although the proof for general distributions is identical.

Fact 5. *Let P, Q be two distributions on $[n]$ and let P^k, Q^k be the corresponding product distributions on $[n]^k$. Then,*

$$d_{\mathrm{TV}}(P^k, Q^k) \leq k \; d_{\mathrm{TV}}(P, Q)$$

Proof. We have

$$\begin{aligned}
d_{\mathrm{TV}}(P^k, Q^k) &= \sum_{x_1,\ldots,x_k \in [n]^k} \left| \prod_{i=1}^{k} p(x_i) - \prod_{i=1}^{k} q(x_i) \right| \\
&= \sum_{x_1,\ldots,x_k \in [n]^k} \left| \prod_{i=1}^{k} p(x_i) - \prod_{i=1}^{k-1} p(x_i) q(x_k) + \prod_{i=1}^{k-1} p(x_i) q(x_k) - \prod_{i=1}^{k} q(x_i) \right| \\
&\leq \sum_{x_1,\ldots,x_{k-1} \in [n]^{k-1}} \left(\prod_{i=1}^{k-1} p(x_i) \sum_{x_k \in [n]} |p(x_k) - q(x_k)| + \sum_{x_k \in [n]} q(x_k) \left| \prod_{i=1}^{k-1} p(x_i) - \prod_{i=1}^{k-1} q(x_i) \right| \right) \\
&\leq d_{\mathrm{TV}}(P^{k-1}, Q^{k-1}) + d_{\mathrm{TV}}(P, Q)
\end{aligned}$$

□

We are now able to prove a (loose) lower bound for the sample complexity of question 3. Let $\mathcal{A}$ be an algorithm that with k samples distinguish the coins $\mathrm{B}(1/2-\epsilon/2)$ and $\mathrm{B}(1/2+\epsilon/2)$ with probability at least $9/10$. Then, combining Lemma 4 and Fact 5 we obtain that $k\epsilon \geq 8/10$ which implies that $k = \Omega(1/\epsilon)$. As we have already seen, the upper bound for this problem is $O(1/\epsilon^2)$; the lower bound we proved is not sharp.

If P, Q are arbitrary distribution then the bound of Fact 5 is tight. If P, Q are Bernoulli we can do much better. We will use a useful inequality that is true in general, Pinsker's inequality. This inequality upper bounds the total variation distance by Kullback-Leibler divergence.

Definition 6 (Kullback-Leibler Divergence)**.** *Let p, q be two probability mass functions supported on n elements. Then*

$$D_{\mathrm{KL}}(p \| q) = \sum_{i=1} p(i) \log \frac{p(i)}{q(i)}$$

We can now state Pinsker's inequality.

Fact 7 (Pinsker's Inequality)**.** $d_{\mathrm{TV}}(P, Q) \leq \sqrt{\frac{D_{\mathrm{KL}}(P \| Q)}{2}}$

Using the above fact we can prove a better upper on the total variation distance of Bernoulli product distributions.

Fact 8. *Let $P = \mathrm{B}(1/2 + \epsilon/2)$, $Q = \mathrm{B}(1/2 - \epsilon/2)$ be two Bernoulli distributions. Then*

$$d_{\mathrm{TV}}(P^k, Q^k) \leq \sqrt{k}\frac{\epsilon}{\sqrt{1-\epsilon}}$$

Proof. We first bound the KL-divergence of P, Q

$$\begin{aligned} D_{\mathrm{KL}}(P\|Q) &= (1/2 + \epsilon/2)\log\left(\frac{1+\epsilon}{1-\epsilon}\right)(1/2 - \epsilon/2)\log\left(\frac{1-\epsilon}{1+\epsilon}\right) \\ &= \epsilon\log\left(\frac{1+\epsilon}{1-\epsilon}\right) = \epsilon(\log(1+\epsilon) - \log(1-\epsilon)) \\ &= \epsilon\int_{-\epsilon}^{\epsilon}\frac{1}{1+x}dx \leq \frac{2\epsilon^2}{1-\epsilon}. \end{aligned}$$

Using Pinksker's inequality, Fact 7, we have

$$d_{\mathrm{TV}}(P^k, Q^k) \leq \sqrt{\frac{D_{\mathrm{KL}}(P^k\|Q^k)}{2}} = \sqrt{k\frac{D_{\mathrm{KL}}(P\|Q)}{2}} \leq \sqrt{k\frac{\epsilon^2}{1-\epsilon}}$$

□

We are now ready to prove our tight (up to constant factors) lower bound for testing Bernoulli distributions. Combining Lemma 4 and Fact 8 we obtain the answer to Question 3.

Lemma 9 (Testing Bernoulli Distributions)**.** *Let $\epsilon \in (0, 1/2)$. Let V be a uniform random variable $V \in \{0, 1\}$. Let $B_0 = B(1/2 - \epsilon/2)$, $B_1 = B(1/2 - \epsilon/2)$ and let X be a vector of k samples from B_V. Let $\mathcal{A}$ be an algorithm such that $\mathbb{P}[\mathcal{A}(X) \neq V] \leq 1/9$. Then $\mathcal{A}$ must draw at least $1/(4\epsilon^2)$ samples, that is $k = \Omega(1/\epsilon^2)$.*

2.2 Learning Discrete Distributions

We now define the family $\mathcal{S}$ of distributions that we will use to prove our lower bound. The family contains $2^{n/2}$ distributions all of which are in distance ϵ from the uniform distribution over n elements. More specifically, assume without loss of generality that the size of the support n is even and split he elements of the support in $n/2$ pairs of

consecutive elements. For the pair $\{i, i+1\}$ we throw a fair random coin $Z_i \in \{-1, +1\}$ and set the probabilities of the elements $i, i+1$ to be

$$p_i = \frac{1 - \epsilon Z_i}{n} \qquad p_{i+1} = \frac{1 + \epsilon Z_i}{n}$$

We call bucket i the subset $\{i, i+1\}$ of the support. Therefore, if $Z \in \{-1, 1\}^{n/2}$ is the vector of the results of the random coins $Z_1, \ldots, Z_{n/2}$ the corresponding distribution is the vector of n probabilities $p_Z = (p_1, \ldots, p_n)$. This is indeed a family of $2^{n/2}$ distributions (each one corresponding to a point of the hypercube $\{-1, 1\}^{n/2}$. Observe that this family is not an ϵ-packing since two distributions may only differ in one pair and in that case the total variation distance between them is $2\epsilon/n$ which is tiny compared to ϵ. However, this family contains a large enough number of distributions that are sufficiently far. In fact, we do not need to explicitly construct a packing for our lower bound to work.

To see intuitively why this family works, observe first that we can restrict our algorithm to output a distribution that belongs to the constructed family. Indeed, let $p_Z \in \mathcal{S}$ be the distribution that generates the samples and let $D_{\mathcal{A}}$ be the output distribution of $\mathcal{A}$. The guarantee of the learning algorithm is that $d_{\mathrm{TV}}(D_{\mathcal{A}}, p_Z) \leq \epsilon$. We can always find a distribution $p_{Z'}$ that belongs to the family $\mathcal{S}$ and is ϵ-close to $D_{\mathcal{A}}$ (p_Z is one such distribution). From the triangle inequality we have that $d_{\mathrm{TV}}(p_{Z'}, p_Z) \leq 2\epsilon$. Therefore, learning p_Z essentially means learning the vector of coins $Z \in \{-1, +1\}^{n/2}$. Moreover, observe that the contribution of each z_i to the total variation between our guess and the true distribution is exactly $2\epsilon/n$. Therefore, unless we correctly learn at least a constant fraction of the $n/2$ coins our guess will not be ϵ-close to p_Z. Since, we want to identify the bias Z_i that corresponds to each pair, observe that conditional that a sample falls in bucket i its distribution is the mixture $1/2B((1+\epsilon)/2)+1/2B((1-\epsilon)/2)$. In the previous section we saw that to distinguish these two Bernoulli distributions we roughly need $\Omega(1/\epsilon^2)$ samples. Since the probability that a sample lands in bin i is $2/n$ we conclude that $\Omega(n/\epsilon^2)$ are needed to identify a constant fraction of Z_i's.

Theorem 10. *Let $\epsilon \in (0, 1/2)$. Let $\mathcal{A}$ be an algorithm that draws k samples from a discrete distribution D supported on $[n]$ and with probability at least $9/10$ outputs a distribution at total variation distance at most ϵ from D. Then $k = \Omega(n/\epsilon^2)$.*

Proof. For simplicity we assume that n is even. As we already discussed we can assume that the output of the learning algorithm is a vector $W \in \{-1, +1\}^{n/2}$. The total variation distance between p_W and another distribution p_Z, $Z \in \{-1, +1\}^{n/2}$ is

$$d_{\mathrm{TV}}(p_Z, p_W) = \frac{2\epsilon}{n} \sum_{i=1}^{n} \mathbb{1}\{W_i \neq Z_i\}$$

Moreover, since each coin Z_i depends only on the samples that fall in bucket i, we can assume that the algorithm uses only the samples that fall in bucket i to determine the value of Z_i, i.e. the coordinate W_i of the output vector only depends on those samples. Now assuming that we choose Z uniformly at random from $\{-1,+1\}^{n/2}$, and then draw a sample $X \in [n]^k$ from p_Z, we will bound the expected total variation distance between p_W and p_Z. We also define the k random variables $B_\ell \in [n/2]$ that corresponds to the bucket that sample X_ℓ fell, $B_\ell = \lfloor X_\ell/2 \rfloor$.

$$\begin{aligned}\mathbb{E}\left[d_{\mathrm{TV}}(p_Z, p_W)\}\right] &= \frac{2\epsilon}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{1}\{W_i(X) \neq Z_i\}\right] \\ &= \frac{2\epsilon}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\{W_i(X) \neq Z_i\}|B_1,\ldots,B_k\right]\right]\end{aligned}$$

We have that $W_i(X)$ only depends on the samples that fall in bucket i, that is $B_j = i$, and that the conditional distribution of any X_j given B_j is a Bernoulli distribution supported on B_j, B_{j+1} with corresponding probabilities $(1-\epsilon Z_{B_j})/2, (1-\epsilon Z_{B_j})/2$. Therefore, using Lemma 9 (or more precicely Fact 8 and Lemma 4) we obtain

$$\mathbb{E}\left[\mathbb{1}\{W_i(X) \neq Z_i\}|B_1,\ldots,B_k\right] \geq \frac{1}{2} - \sqrt{2}\epsilon\,\sqrt{k_i},$$

where k_i is the number of samples that landed in bucket i, $k_i = \sum_{j=1}^{k}\mathbb{1}\{B_j = i\}$. We have $\mathbb{E}[k_i] = k/(n/2) = 2k/n$ since each B_j is a uniform distribution over $[n/2]$. Using Jensen's inequality and the fact that $x \mapsto \sqrt{x}$ is concave we have that $\mathbb{E}[\sqrt{k_i}] \leq \sqrt{E[k_i]} = \sqrt{2k/n}$, and thus

$$\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\{W_i(X) \neq Z_i\}|B_1,\ldots,B_k\right]\right] \geq \frac{1}{2} - 2\epsilon\,\sqrt{k/n}$$

Overall, $\mathbb{E}\left[d_{\mathrm{TV}}(p_Z, p_W)\}\right] \leq \epsilon(1 - 4\epsilon\sqrt{k/n})$. Thus, to make the expected total variation smaller than $\epsilon/2$, we need $k \geq n/(32\epsilon^2)$.

□

3 Further Readings

Variants of the basic hypothesis testing lemmas that we proved can be found in Sections 2.3, 2.4, 4.2, 4.3 of Bar-Yossef's thesis [BY02]. A survey of the formal methods for proving sample complexity lower bounds is [Yu97]. Chapter 2 of Tsybakov's book [Tsy08] also covers sample complexity lower bound techniques.

References

[BY02] Ziv Bar-Yossef. *The Complexity of Massive Data Set Computations*. PhD thesis, Berkeley, CA, USA, 2002. AAI3183783.

[Tsy08] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[Yu97] Bin Yu. *Assouad, Fano, and Le Cam*, pages 423–435. Springer New York, New York, NY, 1997.