ECS 189A Sublinear Algorithms for Big Data **Fall 2024**

Lecture 10: Learning discrete distribution, upper and lower bound

Lecturer: Jasper Lee *Scribe: Dongmin Wu*

Sample Complexity in Learning Distributions

Setting:
Given an unknown distribution D defined over a finite domain $[n]$. We can obtain a collection of independent and identically distributed (i.i.d.) samples $X_1, X_2, \ldots, X_m$ drawn from D.

- One goal is to **learn** an approximate distribution of D, so that the total variation distance (d_{TV}) from D is within ϵ.
- Another potential task is to perform **property testing** on the distribution D, such as testing whether D satisfies certain properties.
- We can also estimate parameters, functionals, or statistics of D.

Sample Complexity for Learning D:
We can approximately learn D within total variation distance ϵ using $m = \Theta\left(\frac{n+\log\frac{1}{\delta}}{\epsilon^2}\right)$ samples, with probability at least $1-\delta$.

$$m = \Theta\left(\frac{n+\log\frac{1}{\delta}}{\epsilon^2}\right)$$

Here, the sample size m grows linearly with the domain size n and includes an **additive** $\log\frac{1}{\delta}$ term, rather than a multiplicative term.
Note: This is similar to the sample size calculation that appears in the Johnson-Lindenstrauss Lemma.

Upper Bound on Sample Complexity

Algorithm 10.1 Empirical Distribution Estimation

Input: i.i.d. samples $X_1, X_2, \ldots, X_m$ from distribution D **Sample size:** $m = \Theta\left(\frac{n+\log\frac{1}{\delta}}{\epsilon^2}\right)$
Construct the empirical distribution $\hat{D}$ where each probability $\hat{D}_i$ is defined as:

$$\hat{D}_i = \frac{\#\{X_j = i\}}{m}$$

Output: An estimate $\hat{D}$ of the true distribution D

Theorem 10.2. *Algorithm 10.1, given input $m = O\left(\frac{n+\log\frac{1}{\delta}}{\epsilon^2}\right)$ samples, returns an estimated distribution $\hat{D}$ such that the total variation distance between $\hat{D}$ and D satisfies*

$$d_{TV}(\hat{D}, D) \leq \epsilon$$

with probability at least $1-\delta$.

Note: The $O\left(\frac{n+\log\frac{1}{\delta}}{\epsilon^2}\right)$ sample complexity is linear in n, with an additive $\log\frac{1}{\delta}$ term. The sample complexity expression can also be interpreted as: $\Theta\left(\frac{1}{\epsilon^2}\log\frac{1}{\frac{\delta}{2^n}}\right)$, where $\log\frac{2^n}{\delta} = \log 2^n + \log\frac{1}{\delta}$.

Proof. Observe that $d_{TV}(\hat{D}, D) \geq \epsilon$ if and only if there exists some subset $S \subseteq [n]$ such that $\hat{D}(S) - D(S) \geq \epsilon$.

We want to show that, with probability at least $1-\delta$,

$$\forall S \subseteq [n], \quad \hat{D}(S) - D(S) < \epsilon.$$

We will apply a union bound over all $S \subseteq [n]$.

- Fix a subset $S \subseteq [n]$:

$$\hat{D}(S) = \frac{1}{m}\sum_j \mathbb{1}\{X_j \in S\}.$$

  Then, the probability that $\hat{D}(S) - D(S) \geq \epsilon$ is

$$\mathbb{P}\left(\hat{D}(S) - D(S) \geq \epsilon\right) = \mathbb{P}\left(\frac{1}{m}\sum_j \mathbf{1}\{X_j \in S\} - \mathbb{E}[\hat{D}(S)] \geq \epsilon\right)$$

  By Hoeffding's inequality, this is bounded by

$$\leq e^{-\Theta(m\epsilon^2)}.$$

- Determining the Sample Size: we choose

$$m = \Theta\left(\frac{1}{\epsilon^2}\log\frac{1}{\delta/2^n}\right).$$

- Applying the Union Bound: By the union bound, we have

$$\mathbb{P}\left(d_{TV}(\hat{D}, D) \geq \epsilon\right) = \mathbb{P}\left(\exists S \subseteq [n] : \hat{D}(S) - D(S) \geq \epsilon\right) \leq 2^n \cdot \frac{\delta}{2^n} = \delta.$$

□

Note: The additive $\log\frac{1}{\delta}$ term (as opposed to multiplicative) is due to the use of a union bound over many events, similar to the approach in Johnson-Lindenstrauss Lemma.

Lower Bound on Sample Complexity

We want to show that $\Omega\left(\frac{n+\log\frac{1}{\delta}}{\epsilon^2}\right)$ samples are necessary for learning the distribution within total variation distance ϵ.

Approach To prove this lower bound, we split it into two parts: 1. Show that $\Omega\left(\frac{n}{\epsilon^2}\right)$ samples are necessary. 2. Show that $\Omega\left(\frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$ samples are necessary.

These two bounds together imply a lower bound of

$$\Omega\left(\max\left(\frac{n}{\epsilon^2}, \frac{\log\frac{1}{\delta}}{\epsilon^2}\right)\right) \geq \Omega\left(\frac{n+\log\frac{1}{\delta}}{\epsilon^2}\right).$$

Proof. 1. For the Term $\Omega\left(\frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$: - Consider the problem of distinguishing between two Bernoulli distributions: Bernoulli $\left(\frac{1}{2}-\epsilon\right)$ and Bernoulli $\left(\frac{1}{2}+\epsilon\right)$. - By Theorem 9.9, we have:

$$d_H^2(Ber(\frac{1}{2}\pm\epsilon)) = \Theta(\epsilon^2).$$

- This means that $\Omega\left(\frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$ samples are required to distinguish these two distributions with probability at least $1-\delta$.

2. For the Term $\Omega\left(\frac{n}{\epsilon^2}\right)$: - we show that to learn the distribution over a domain of size n requires at least $\Omega\left(\frac{n}{\epsilon^2}\right)$ samples.

Combining these results, we conclude that

$$\Omega\left(\frac{n+\log\frac{1}{\delta}}{\epsilon^2}\right)$$

samples are necessary for learning the distribution within total variation distance ϵ with high probability.

□

Strategy for Lower Bound on Sample Complexity

To establish a lower bound of $\Omega\left(\frac{n}{\epsilon^2}\right)$ for sample complexity, we consider the following class of distributions.

1. Define the probabilities:

$$p_{2i} = \frac{1-100\epsilon z_i}{n}, \quad p_{2i+1} = \frac{1+100\epsilon z_i}{n}, \quad z_i \in \{\pm 1\}.$$

Each distribution $P_{\underline{Z}}$ is identified by an $\frac{n}{2}$-length vector $Z \in \{\pm 1\}^{\frac{n}{2}}$.

2. Intuition: To learn the distribution $P_{\underline{Z}}$ to within ϵ in total variation distance, one must learn at least 99% of the z_i's. If even a small fraction (e.g., 1%) is incorrect, it contributes significantly to the total variation distance.

Informal Analysis

Consider a fixed "bucket" $B_i = \{Y_{2i}, Y_{2i+1}\}$. Conditioning on B_i learning z_i is equivalent to distinguishing between two cases:

$$\text{Ber}\left(\frac{1-100\epsilon}{2}\right) \text{ vs. } \text{Ber}\left(\frac{1+100\epsilon}{2}\right).$$

To distinguish between these two Bernoulli distributions with high probability, we need $\Omega\left(\frac{1}{\epsilon^2}\right)$ samples.

However, a sample falls in B_i with probability $\frac{2}{n}$, so overall we need $\Omega\left(\frac{n}{\epsilon^2}\right)$ samples to learn the distribution.

Formal analysis

Lemma 10.3. *Learning a distribution in the above class with probability at least $\frac{2}{3}$ requires $\Omega(\frac{n}{\epsilon^2})$ samples.*

Proof. Consider an arbitrary algorithm A outputting $P_{\mathbf{w}}$ or just $\mathbf{w}$, where $\mathbf{w}$ is a vector of length $\frac{n}{2}$ in the form of $\mathbf{z}$ defined above.

Claim: Without loss of generality, A depends only on histogram

$$Y_i = \sum_j \mathbb{1}\{x_j = i\}$$

Proof of claim: consider an algorithm A' that takes the histogram, generates a random ordering of samples based on the histogram, and feed it into A. $A's$ input has exactly the same distribution as $D^{\otimes m}$.

Consider drawing $\mathbf{z}$ uniformly at random, i.e. each z_i is drawn iid from Ber $\left(\frac{1}{2}\right)$. We want to analyze the number of wrong coordinates in $\mathbf{w} = A(Y_1, \ldots, Y_n)$, that is, $\sum_{\text{bucket } i} \mathbb{1}\{w_i \neq z_i\}$.

Note: z_i is random, $\{x_i\}$ are random even conditioning on $\mathbf{z}$, and $\mathbf{w}$ might be random even conditioned on $(x_1, \ldots, x_m)$.

We want to prove that

$$\mathbb{P}\left(\sum \mathbb{1}\{w_i \neq z_i\} > 0.01 \cdot \frac{n}{2}\right) > \frac{1}{3}$$

$$\iff \mathbb{P}\left(\#\text{ correct coordinates } > 0.99 \cdot \frac{n}{2}\right) < \frac{2}{3}$$

Note that the sum $\sum \mathbb{1}\{w_i \neq z_i\}$ is not a sum of independent terms, so we can't use any of the exponential tail bounds that we've seen before. The reason why it is not a sum of independent terms is that:

- w_i might depend on samples from buckets other than the ith bucket.
- The buckets themselves are correlated. In particular, any two distinct buckets $i \neq j$ are not independent. This is because the total samples need to sum up to m. (next week we will see a trick called Poissonisation that resolves this issue)

Goal: show that the expected number of correct coordinates $\approx \frac{1}{2} \cdot \frac{n}{2}$ for $m = \frac{1}{100} \cdot \frac{n}{\epsilon^2}$ (which means the number of incorrect coordinates will also be roughly a half). Then by Markov's we will be able to show that

$$\mathbb{P}\left(\# \text{ correct coordinates} > 0.99 \cdot \frac{n}{2}\right) \leq \frac{\frac{1}{2}}{0.99} \leq \frac{2}{3}$$

We compute

$$\mathbb{E}\left[\sum_i \mathbb{1}\{w_i \neq z_i\}\right] = \sum_i \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\{w_i \neq z_i\} \mid B_1, B_2, \ldots, B_{\frac{n}{2}}\right]\right]$$

where $B_i =$ the number of samples in bucket $i = Y_{2i} + Y_{2i+1}$.

Claim 10.4.

$$\mathbb{E}\left[\mathbb{1}\{w_i \neq z_i\} \mid B_1, B_2, \ldots, B_{\frac{n}{2}}\right] \geq \frac{1}{2} - O(\epsilon) \cdot \sqrt{B_i}$$

Assuming **Claim 10.4**, then we can compute the lower bound proof:

$$\begin{aligned}
\mathbb{E}\left[\sum_i \mathbb{1}\{w_i \neq z_i\}\right] &\geq \sum_i \frac{1}{2} - O(\epsilon) \cdot \mathbb{E}\sqrt{B_i} \\
&= \frac{n}{4} - O(\epsilon) \cdot \sum_i \mathbb{E}\sqrt{B_i} \\
&\geq \frac{n}{4} - O(\epsilon) \cdot \sum_i \sqrt{\mathbb{E}\, B_i} \qquad \text{by Jensen's} \\
&= \frac{n}{4} - O(\epsilon) \cdot \sum_i \sqrt{\frac{2m}{n}} \\
&= n\left(\frac{1}{4} - O(\epsilon) \cdot \sqrt{\frac{2m}{n}}\right)
\end{aligned}$$

If $m = \frac{n}{O(\epsilon^2)}$, then last line $\approx \frac{n}{4} = \frac{1}{2} \cdot \frac{n}{2}$, then we are done, by applying Markov's, as stated earlier.

So what remains is to show that **Claim 10.4** is correct.

Proof of **Claim 10.4**:

Rewrite

$$\mathbb{E}\left[\mathbb{1}\{w_i \neq z_i\} \mid B_1, B_2, \ldots, B_{\frac{n}{2}}\right]$$

further as

$$\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\{w_i \neq z_i\} \mid B_1, B_2, \ldots, B_{\frac{n}{2}}, Z_{-i}, \text{samples outside bucket } i\right]\right]$$

where $\mathbf{z}_{-i}$ means all z_j with $j \neq i$, and the outer expectation is over $\mathbf{z}_{-i}$, samples outside bucket i, conditioned on $B_1, \ldots, B_{\frac{n}{2}}$.

Fix $\mathbf{z}_{-i}$, samples outside bucket i, and B_i, then algorithm A just takes B_i samples in bucket

i and outputs the vector $\mathbf{w}$ (we only care about w_i, and in particular we want a lower bound for $\mathbb{P}(w_i \neq z_i)$). In other words, A takes B_i samples from Ber $\left(\frac{1-100\epsilon z_i}{2}\right)$ and outputs w_i, hoping that $w_i = z_i$. This is similar to distinguishing between coin flips of two distributions, except that this time z_i is uniformly drawn, instead of adversarially picked.

Thus, it suffices to prove the following claim:

Claim 10.5. *Pick $q = \frac{1\pm 100\epsilon}{2}$ uniformly (denoted as q_+, q_-, respectively). Take m samples iid. from $\text{Ber}(q)$ (m corresponds to B_i in previous parts). Then for any algorithm A',*

$$\mathbb{P}\left(A'(\textit{samples}) \neq q\right) \geq \frac{1}{2} - O(\epsilon) \cdot \sqrt{m}$$

Proof of Claim 10.5:
By Theorem 11.1, we know

$$\mathbb{P}\left(A' = q_+ \,|\, q = q_+\right) - \mathbb{P}\left(A' = q_+ \,|\, q = q_-\right) \leq d_{\text{TV}}\left(\text{Ber}(q_+)^{\otimes m}, \text{Ber}(q_-)^{\otimes m}\right)$$

L.H.S. $=$

$$1 - \mathbb{P}\left(A' = q_- \,|\, q = q_+\right) - \mathbb{P}\left(A' = q_+ \,|\, q = q_-\right)$$

R.H.S. $\leq$ (by Fact 11.7)

$$\sqrt{m} \cdot d_{\text{H}}\left(\text{Ber}(q_+), \text{Ber}(q_-)\right) = \sqrt{m} \cdot O(\epsilon)$$

Now we have

$$\begin{aligned}\frac{1}{2}\left(1 - \sqrt{m} \cdot O(\epsilon)\right) &\leq \frac{1}{2}\left(\mathbb{P}\left(A' = q_- \,|\, q = q_+\right) + \mathbb{P}\left(A' = q_+ \,|\, q = q_-\right)\right) \\ &= \mathbb{P}\left(A' \neq q \,|\, q = \text{Unif}\{q_\pm\}\right)\end{aligned}$$

which is exactly what we are trying to show.

□

Theorem 10.6. *Any algorithm learning discrete distributions over $[n]$ to within total variation distance error ϵ with probability at least $1-\delta$ requires $\Omega\left(\frac{n+\log\frac{1}{\delta}}{\epsilon^2}\right)$ samples.*

Proof. Apply Lemma 10.3 and the lower bound $\Omega\left(\frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$ which we proved earlier. □

DKW Inequality

We will end with stating the DKW Inequality, which concerns learning a distribution in *Kolmogorov distance.*

Definition 10.7 (Kolmogorov Distance)**.** ℓ_∞ distance between the CDFs

$$d_{\text{K}}(\mathbf{p}, \mathbf{q}) = \sup_x \left|\mathbf{p}\left(-\infty, x\right] - \mathbf{q}\left(-\infty, x\right]\right|$$

Theorem 10.8 (DKW Inequality)**.** *Given any distribution* $\mathbf{p}$ *on* $\mathbb{R}$ *(not necessarily discrete), consider*

$$\hat{\mathbf{p}}_{\mathbf{m}} = m\text{-sample empirical CDF}$$

Then

$$\mathbb{P}\left(d_{\mathrm{K}}\left(\hat{\mathbf{p}}_{\mathbf{m}}, \mathbf{p}\right) > \epsilon\right) \leq 2e^{-2me^2}$$

So to learn $\mathbf{p}$ *within* ϵ *in* d_k*, we only need* $O\left(\frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$ *samples.*