CSCI 1951-W Sublinear Algorithms for Big Data **Fall 2020**

Lecture 12: Learning Discrete Distributions

Lecturer: Jasper Lee *Scribe: Qianfan Chen*

0 Overview

Last class, we talked about the task of distinguishing between 2 known distributions. For today and for the next few lectures, we will talk about sublinear algorithms where the main objects being tested are probability distributions.

The first question is, what does "linear" mean in "sublinear algorithms" in this context? Today, we will talk about how we can learn a discrete distribution with "linearly" many samples.

1 Problem Setting

We have an unknown probability distribution D over the set $\{1, \ldots, n\}$, and we get m samples $x_1, \ldots x_m$ from D, assumed to be iid.

Tasks we might be interested in:

- Learn (approximately) the distribution D;
- Test if D has property $\mathcal{P}$ vs ϵ-far from $\mathcal{P}$;
- Estimate functions or parameters of D (we have already seen examples of this, for example mean estimation)

We care about the sample complexity m required.

What can we do with linearly many (linear as to the size of the domain; so, $O(n)$) samples? Today, we will show that we can learn D to within ϵ in total variation distance. Specifically, our algorithm will output estimation $\hat{D}$ such that $d_{\mathrm{TV}}(\hat{D}, D) \leq \epsilon$ with probability at least $1 - \delta$, using $\Theta\left(\frac{n+\log\frac{1}{\delta}}{\epsilon^2}\right)$ samples. In this lecture we will show both the upper and lower bound. Note that the $\log\frac{1}{\delta}$ term here is additive, not multiplicative. This means we are not just repeating an algorithm $\log\frac{1}{\delta}$ many times, which would not be the optimal strategy (even if we figure out how to combine the output distributions from multiple runs).

2 Upper Bound

Algorithm 12.1 *Take $x_1, \ldots, x_m$ drawn iid from $\hat{D}$, with $m = O\left(\frac{n+\log\frac{1}{\delta}}{\epsilon^2}\right)$. Return the empiricial distribution*

$$\hat{D}_i = \frac{|\{x_j = i\}|}{m}$$

Theorem 12.2 *Algorithm 12.1, on input $m = O\left(\frac{n+\log\frac{1}{\delta}}{\epsilon^2}\right)$ samples, return $\hat{D}$ with $d_{\mathrm{TV}}(\hat{D}, D) \leq \epsilon$ with probability at least $1 - \delta$.*

Note: when we say "with probability at least $1-\delta$", the randomness comes from the random sampling of $x_1, \ldots, x_m$; though if we use some other algorithm with randomization within the algorithm, that could also add to the randomness.

Hint for the proof: $m = O\left(\frac{1}{\epsilon^2} \log \frac{1}{\frac{\delta}{2^n}}\right)$

Proof. Observe that, by the definition of total variation distance,

$$d_{\mathrm{TV}}(\hat{D}, D) \geq \epsilon \iff \exists S \subset [n] \text{ } s.t. \text{ } \hat{D}(S) - D(S) \geq \epsilon$$

We could take the union bound over all 2^n possible $S \subseteq [n]$.
Fix S, consider $s_i = \mathbb{1}\{x_i \in S\}$, the indicator variable for whether x_i is in the set S, then s_i is drawn from Bernoulli($D(S)$).
Apply Hoeffding's:

$$\mathbb{P}\left(\hat{D}(S) - D(S) \geq \epsilon\right) = \mathbb{P}\left(\frac{1}{m}\sum s_i \geq \mathbb{E}\left[\frac{1}{m}\sum s_i\right] + \epsilon\right) \leq e^{-\Theta\left(m\epsilon^2\right)} = \frac{\delta}{2^n}$$

By union bound over all S,

$$\mathbb{P}\left(d_{\mathrm{TV}}(\hat{D}, D) < \epsilon\right) = \mathbb{P}\left(\forall S \subset [n],\ \hat{D}(S) - D(S) < \epsilon\right) \geq 1 - \delta$$

□

Note: similar to JL analysis, we are union bounding over a lot of failure events that happen with tiny probability. This analysis gives the same additive $\log \frac{1}{\delta}$ term.

3 Lower Bound

First question: how to prove a lower bound $\Omega\left(\frac{n+\log \frac{1}{\delta}}{\epsilon^2}\right)$, which is a sum of two terms?

Idea: Prove two lower bounds $\Omega\left(\frac{n}{\epsilon^2}\right)$ and $\Omega\left(\frac{\log \frac{1}{\delta}}{\epsilon^2}\right)$, which would give us a lower bound of

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

We first prove the lower bound $\Omega\left(\frac{\log \frac{1}{\delta}}{\epsilon^2}\right)$.

Consider the problem of distinguishing Bernoulli($\frac{1}{2} - \epsilon$) vs Bernoulli($\frac{1}{2} + \epsilon$) on elements $\{1, 2\} \subseteq [n]$ with probability at least $1 - \delta$.
Recall that

$$d_{\mathrm{H}}^2\left(B\left(\frac{1}{2} + \epsilon\right), B\left(\frac{1}{2} - \epsilon\right)\right) = O\left(\epsilon^2\right)$$

and apply Theorem 11.9, we get that successfully distinguishing between Bernoulli($\frac{1}{2} - \epsilon$) vs Bernoulli($\frac{1}{2} + \epsilon$) on elements $\{1, 2\} \subseteq [n]$ with probability at least $1 - \delta$ requires $\Omega\left(\frac{\log \frac{1}{\delta}}{\epsilon^2}\right)$

samples. Since $d_{\mathrm{TV}}\left(B\left(\frac{1}{2}+\epsilon\right), B\left(\frac{1}{2}-\epsilon\right)\right) = 2\epsilon$, we know that it requires at least $\Omega\left(\frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$ samples to learn $\hat{D}$ to within ϵ of D with probability at least $1-\delta$.

We then prove the lower bound $\Omega\left(\frac{n}{\epsilon^2}\right)$.

Assume n is even (since we are proving a lower bound, it is fine if we just prove it for every even n), consider the following class of distributions:

$$p_{2i} = \frac{1-100\epsilon z_i}{n}$$

$$p_{2i+1} = \frac{1+100\epsilon z_i}{n}$$

with $z_i \in \{1,-1\}$ for each $i \le \frac{n}{2}$.

Each distribution could be identified by a vector $\mathbf{z}$ of length $\frac{n}{2}$. There are $2^{\frac{n}{2}}$ different possible values for $\mathbf{z}$, and therefore there are $2^{\frac{n}{2}}$ different possible distributions in the class defined above.

Intuition: we need to learn $\ge 99\%$ for the z_i's to be within ϵ in total variation distance, since for each z_i that we get wrong, it contributes $\frac{200\epsilon}{n}$ to the total variation distance.

Further intuition: conditioned on the "bucket" $b_i = \{2i, 2i+1\}$, we get Bernoulli($\frac{1-50\epsilon z_i}{2}$). So we need $\Omega(\frac{1}{\epsilon^2})$ samples learn each z_i. However, sample falls into bucket b_i with probability $O(\frac{1}{n})$ (so, in expectation, need $\Omega(\frac{n}{\epsilon^2})$ samples to learn for a bucket). Also, we need to learn this for $\ge 99\% \cdot \frac{n}{2}$ many i's. Formalizing this is trickier.

Lemma 12.3 *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 Bernoulli$\left(\frac{1}{2}\right)$. We want to analyze the number of wrong coordinates in $\mathbf{w} = A(Y_1, \ldots, Y_n)$, that is, $\sum_{\text{bucket } i} \mathbb{1}\{w_i \ne z_i\}$.

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

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

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

Note that the sum $\sum \mathbb{1}\{w_i \ne z_i\}$ 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 12.4

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

Assuming **Claim 12.4**, 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 12.4** is correct.

Proof of **Claim 12.4**:

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 Bernoulli $\left(\frac{1-100\epsilon z_i}{2}\right)$ 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 12.5 *Pick $q = \frac{1\pm 100\epsilon}{2}$ uniformly (denoted as q_+, q_-, respectively). Take m samples iid. from* Bernoulli(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 12.5:
By Theorem 11.1, we know

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

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)

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

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 12.6 *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 12.3 and the lower bound $\Omega\left(\frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$ which we proved earlier. □

4 DKW Inequality

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

Definition 12.7 (Kolmogorov Distance) ℓ_∞ distance between the CDFs

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

Theorem 12.8 (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.*