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

Lecture 11: Distinguishing (Discrete) Distributions

Lecturer: Jasper Lee *Scribe: Shamay Samuel*

1 Problem Setting

Suppose we have two known distributions $\mathbf{p}, \mathbf{q}$ over $[n] = \{1, 2, \cdots, n\}$ and an adversary that picks one of these distributions (let's call the choice D). We further obtain m samples from D.

The goal is here is to find an algorithm $\mathcal{A}$ such that:

- If $D = \mathbf{p}$, $\mathcal{A}$ returns "$D = \mathbf{p}$"
- If $D = \mathbf{q}$, $\mathcal{A}$ returns "$D = \mathbf{q}$"

Both cases should have success probability $\geq \frac{2}{3}$ (or more generally probability $\geq 1 - \delta$).

2 Case for $m = 1$

We first define the Total Variation distance between two probability distributions as follows

Definition 11.1 Given two discrete probability distributions $\mathbf{p}, \mathbf{q}$ over $[n]$, the Total Variation distance $d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})$ between $\mathbf{p}$ and $\mathbf{q}$ is defined as:

$$\begin{aligned} d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) :&= \frac{1}{2} \sum_{i \in [n]} |p_i - q_i| \\ &= \frac{1}{2} \|\mathbf{p} - \mathbf{q}\|_1 \\ &= \sup_{A \subseteq [n]} \mathbf{p}(A) - \mathbf{q}(A) \end{aligned}$$

We now restate some of the fundamental results concerning the Total Variation distance that we showed in HW1.

Theorem 11.2 *For the case of $m = 1$ sample from D. There exists an algorithm $\mathcal{A}$ (i.e. the Maximum Likelihood Estimator) such that:*

$$\begin{aligned} \mathbb{P}(\mathcal{A} \text{ returns } \boldsymbol{p} \mid \boldsymbol{p}) - \mathbb{P}(\mathcal{A} \text{ returns } \boldsymbol{p} \mid \boldsymbol{q}) &= \mathbb{P}(\mathcal{A} \text{ returns } \boldsymbol{q} \mid \boldsymbol{q}) - \mathbb{P}(\mathcal{A} \text{ returns } \boldsymbol{q} \mid \boldsymbol{p}) \\ &= d_{\mathrm{TV}}(\boldsymbol{p}, \boldsymbol{q}) \end{aligned}$$

Moreover, there is no algorithm $\mathcal{A}$ such that:

$$\mathbb{P}(\mathcal{A} \text{ returns } \boldsymbol{p} \mid \boldsymbol{p}) - \mathbb{P}(\mathcal{A} \text{ returns } \boldsymbol{p} \mid \boldsymbol{q}) > d_{\mathrm{TV}}(\boldsymbol{p}, \boldsymbol{q})$$

In particular, this implies that there is no algorithm $\mathcal{A}$ such that both of the following hold:

- $\mathbb{P}(\mathcal{A} \text{ returns } \boldsymbol{p} \mid \boldsymbol{p}) > \frac{1}{2} + \frac{1}{2} d_{\mathrm{TV}}(\boldsymbol{p}, \boldsymbol{q})$
- $\mathbb{P}(\mathcal{A} \text{ returns } \boldsymbol{q} \mid \boldsymbol{q}) > \frac{1}{2} + \frac{1}{2} d_{\mathrm{TV}}(\boldsymbol{p}, \boldsymbol{q})$

So if $d_{\mathrm{TV}}(\boldsymbol{p}, \boldsymbol{q}) < \frac{1}{3}$, there is no algorithm that will succeed in distinguishing between two distributions with probability $\geq \frac{2}{3}$.

3 Case for $m > 1$

We now seek to find the smallest number of samples m (i.e. sample complexity) such that we can distinguish between the distributions $\mathbf{p}, \mathbf{q}$ with success probability $\geq \frac{2}{3}$ or more generally with success probability $\geq 1 - \delta$.

We first note that taking m samples from D is equivalent to taking 1 sample from the m-fold product distribution $D^{\otimes m}$. From the previous section, we need m large enough such that $d_{\mathrm{TV}}(\mathbf{p}^{\otimes m}, \mathbf{q}^{\otimes m}) \geq \frac{1}{3}$. We will also make use of the following fact:

Fact 11.3 *For discrete probability distributions $\boldsymbol{p}, \boldsymbol{q}$, and for any $m > 0$:*

$$d_{\mathrm{TV}}(\boldsymbol{p}^{\otimes m}, \boldsymbol{q}^{\otimes m}) \leq m \cdot d_{\mathrm{TV}}(\boldsymbol{p}, \boldsymbol{q})$$

We now construct upper and lower bounds for the sample complexity m.

Proposition 11.4 *Any algorithm requires $\Omega\left(\frac{1}{d_{\mathrm{TV}}(\boldsymbol{p},\boldsymbol{q})}\right)$ samples to successfully distinguish between $\boldsymbol{p}, \boldsymbol{q}$ with probability $\geq \frac{2}{3}$.*

Proof. We have the following:

$$m = \frac{1}{100 d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})} \Rightarrow d_{\mathrm{TV}}(\mathbf{p}^{\otimes m}, \mathbf{q}^{\otimes m}) \leq \frac{1}{100} < \frac{1}{3}$$

□

Note that this bound is tight, and there is no strong lower bound. For instance, consider distinguishing between Bernoulli(0) and Bernoulli($d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})$).

Proposition 11.5 *It suffices to take $O\left(\frac{1}{d^2_{\mathrm{TV}}(\boldsymbol{p},\boldsymbol{q})}\right)$ samples to successfully distinguish between $\boldsymbol{p}, \boldsymbol{q}$ with probability $\geq \frac{2}{3}$.*

Proof. Consider $A = \arg\sup_A \mathbf{p}(A) - \mathbf{q}(A)$. Using the samples, we now estimate $D(A) = \mathbb{E}_{x \in D}[\mathbb{1}_A]$ to additive error $d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})/3$. Finally, we return the closer of $\mathbf{p}(A)$ and $\mathbf{q}(A)$.

Note that since $\mathbf{p}(A) - \mathbf{q}(A) = d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})$, if the estimate for $D(A)$ is within a error of $d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})/3$, we can definitively distinguish between $\mathbf{p}$ and $\mathbf{q}$ by return the closer $\mathbf{p}(A)$ and $\mathbf{q}(A)$.

Finally, the estimation of $\mathbb{E}_{x \in D}[\mathbb{1}_A]$ to additive error $d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})/3$ amounts to the estimation of the mean of $\mathbb{1}_A$ (i.e. a $0-1$ Bernoulli random variable with unknown probability), which can be done in $O\left(\frac{1}{d^2_{\mathrm{TV}}(\mathbf{p},\mathbf{q})}\right)$. □

Again, we note that the above bound is tight. For instance, consider distinguishing between Bernoulli($\frac{1}{2} \pm \epsilon$). We will show in the subsequent section that it takes $\Omega\left(\frac{1}{\epsilon^2}\right)$ samples to successfully distinguish between the two distributions in this example.

Notice that Propositions 11.4 and 11.5 indicate that for the $m > 1$ case, Total Variation distance cannot adequately capture the sample complexity required to distinguish between distributions, given that there is a quadratic gap between the best upper and lower bounds. We can now introduce another distance between probability distributions that can better capture the sample complexity.

Definition 11.6 Given two discrete probability distributions $\mathbf{p}, \mathbf{q}$ over $[n]$, the Squared Hellinger distance d^2_{H} between $\mathbf{p}$ and $\mathbf{q}$ is defined as:

$$d^2_{\mathrm{H}}(\mathbf{p}, \mathbf{q}) := \frac{1}{2} \sum_{i \in [n]} (\sqrt{p_i} - \sqrt{q_i})^2 = 1 - \sum_{i \in [n]} \sqrt{p_i q_i}$$

The Squared Hellinger distance has the following useful properties:

Fact 11.7 *For discrete probability distributions $\boldsymbol{p}, \boldsymbol{q}$, and for any $m > 0$:*

- $d^2_{\mathrm{H}}(\boldsymbol{p}, \boldsymbol{q}) \leq d_{\mathrm{TV}}(\boldsymbol{p}, \boldsymbol{q}) \leq \sqrt{2} d_{\mathrm{H}}(\boldsymbol{p}, \boldsymbol{q})$
- $d^2_{\mathrm{H}}(\boldsymbol{p}^{\otimes m}, \boldsymbol{q}^{\otimes m}) = 1 - (1 - d^2_{\mathrm{H}}(\boldsymbol{p}, \boldsymbol{q}))^m \leq m \cdot d^2_{\mathrm{H}}(\boldsymbol{p}, \boldsymbol{q})$

Combining the above two gives us:

$$d_{\mathrm{TV}}(\mathbf{p}^{\otimes m}, \mathbf{q}^{\otimes m}) \leq \sqrt{2} d_{\mathrm{H}}(\mathbf{p}^{\otimes m}, \mathbf{q}^{\otimes m}) \leq \sqrt{2}\sqrt{m} d_{\mathrm{H}}(\mathbf{p}, \mathbf{q})$$

We can now use the above distance to construct a tight bound on the sample complexity required to distinguish between distributions with probability $\geq \frac{2}{3}$.

Proposition 11.8 *Successfully distinguishing between $\boldsymbol{p}, \boldsymbol{q}$ requires with probability $\geq \frac{2}{3}$ takes $\Theta\left(\frac{1}{d^2_{\mathrm{H}}(\boldsymbol{p}, \boldsymbol{q})}\right)$ samples*

Proof. First we show the sample complexity is $\Omega\left(\frac{1}{d^2_{\mathrm{H}}(\mathbf{p}, \mathbf{q})}\right)$. We have the following:

$$m = \frac{1}{100 d^2_{\mathrm{H}}(\mathbf{p}, \mathbf{q})} \Rightarrow d_{\mathrm{TV}}(\mathbf{p}^{\otimes m}, \mathbf{q}^{\otimes m}) \leq \sqrt{2} \cdot \sqrt{\frac{1}{100 d^2_{\mathrm{H}}(\mathbf{p}, \mathbf{q})}} \cdot d_{\mathrm{H}}(\mathbf{p}, \mathbf{q}) = \frac{\sqrt{2}}{10} < \frac{1}{3}$$

Now we show the sample complexity is $O\left(\frac{1}{d^2_{\mathrm{H}}(\mathbf{p}, \mathbf{q})}\right)$. We have the following:

$$\begin{aligned} d_{\mathrm{TV}}(\mathbf{p}^{\otimes m}, \mathbf{q}^{\otimes m}) &\geq d^2_{\mathrm{H}}(\mathbf{p}^{\otimes m}, \mathbf{q}^{\otimes m}) \\ &= 1 - (1 - d^2_{\mathrm{H}}(\mathbf{p}, \mathbf{q}))^m \\ &\geq 1 - e^{-m d^2_{\mathrm{H}}(\mathbf{p}, \mathbf{q})} \end{aligned}$$

Now we take $m = O\left(\frac{1}{d^2_{\mathrm{H}}(\mathbf{p}, \mathbf{q})}\right)$, and note that $d_{\mathrm{TV}}(\mathbf{p}^{\otimes m}, \mathbf{q}^{\otimes m}) \geq \frac{2}{3}$. Here we have two choices to proceed:

1. We use the 1-sample d_{TV}-Tester to conclude that we need $O\left(\frac{1}{d^2_{\mathrm{TV}}(\mathbf{p}^{\otimes m}, \mathbf{q}^{\otimes m})}\right)$ samples of $D^{\otimes m}$, by Proposition 11.5
2. We observe the Maximum Likelihood Estimator for $D^{\otimes m}$ works since:

$$\begin{aligned} \mathbb{P}(MLE = \mathbf{p} \mid \mathbf{p}) &\geq \mathbb{P}(MLE = \mathbf{p} \mid \mathbf{p}) - \mathbb{P}(MLE = \mathbf{q} \mid \mathbf{p}) \\ &= d_{\mathrm{TV}}(\mathbf{p}^{\otimes m}, \mathbf{q}^{\otimes m}) \geq \frac{2}{3} \end{aligned}$$

□

Theorem 11.9 *Successfully distinguishing between $\boldsymbol{p}, \boldsymbol{q}$ with probability $\geq 1-\delta$ requires $\Theta\left(\frac{1}{d_{\mathrm{H}}^2(\boldsymbol{p},\boldsymbol{q})}\log\frac{1}{\delta}\right)$ samples.*

Proof. The upper bound follows from Proposition 11.8 and the Majority Vote algorithm. The lower bound requires the tighter inequality:

$$d_{\mathrm{H}}^2(\mathbf{p},\mathbf{q}) \geq 1-\sqrt{1-d_{\mathrm{TV}}^2(\mathbf{p},\mathbf{q})}$$

□

4 Application to Mean Estimation

Suppose we seek to estimate the mean of Bernoulli(p) with additive error ϵ with $\geq \frac{2}{3}$ probability of success. We have discussed a couple methods in previous lectures:

1. Calculating the sample mean, which requires $O\left(\frac{p(1-p)}{\epsilon^2}\right)$ samples for $\frac{2}{3}$ probability of success.
2. Calculating the Median of Means, which requires $O\left(\frac{p(1-p)}{\epsilon^2}\log\frac{1}{\delta}\right)$ samples for $1-\delta$ probability of success.

Both of these provide upper bounds for sample complexity. Using the tools from this lecture, we can show matching lower bounds as well. Note that if we can estimate mean p within additive error ϵ, then we can distinguish between Bernoulli(p) and Bernoulli($p+2\epsilon$). If the number of samples is insufficient to distinguish the above distributions, then it is impossible to estimate the mean within the required error. To simplify computation, we instead compute the Squared Hellinger distance for $\mathbf{p}$ = Bernoulli($p+\epsilon$), $\mathbf{q}$ = Bernoulli($p-\epsilon$) as follows:

$$d_{\mathrm{H}}^2(\mathbf{p},\mathbf{q}) = \Theta((\sqrt{1-p+\epsilon}-\sqrt{1-p-\epsilon})^2) + \Theta((\sqrt{p+\epsilon}-\sqrt{p-\epsilon})^2)$$

Now note that for $\epsilon < p$, we have:

$$\begin{aligned}(\sqrt{p+\epsilon}-\sqrt{p-\epsilon})^2 &= p\left(\sqrt{1+\frac{\epsilon}{p}}-\sqrt{1-\frac{\epsilon}{p}}\right)^2 \\ &= p\left(1+\Theta\left(\frac{\epsilon}{p}\right)-\left(1-\Theta\left(\frac{\epsilon}{p}\right)\right)\right)^2 \\ &= p\cdot\Theta\left(\frac{\epsilon^2}{p^2}\right) = \Theta\left(\frac{\epsilon^2}{p}\right)\end{aligned}$$

Then note for $p < \frac{1}{2}$, we have:

$$(\sqrt{1-p+\epsilon}-\sqrt{1-p-\epsilon})^2 = \Theta\left(\frac{\epsilon^2}{1-p}\right) = O\left(\frac{\epsilon^2}{p}\right)$$

So, combining the above, we have for $p < \frac{1}{2}, \epsilon < p$:

$$d_{\mathrm{H}}^2(\mathbf{p},\mathbf{q}) = \Theta\left(\frac{\epsilon^2}{p}\right) = \Theta\left(\frac{\epsilon^2}{p(1-p)}\right)$$

We can now conclude with the following statements:

1. By Proposition 11.8, we require $\Omega\left(\frac{p(1-p)}{\epsilon^2}\right)$ samples to successfully distinguish with probability $\frac{2}{3}$.

2. By Theorem 11.9, we require $\Omega\left(\frac{p(1-p)}{\epsilon^2}\log\frac{1}{\delta}\right)$ samples to successfully distinguish with probability $1-\delta$.