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

# Lecture 9: Distinguishing (discrete) distributions, Various statistical distances

*Lecturer: Jasper Lee* *Scribe: Liam Glenn*

## 1 Problem Setting

Suppose we have two known discrete distributions $\mathbf{p}, \mathbf{q}$ over $[n]$ and an adversary that picks one of these distributions ($D$). We get $m$ i.i.d. samples from $D$.

We want 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 "$D = \mathbf{q}$"

We want a success probability $\geq \frac{2}{3}$

Later in the lecture, we look at the more general case of success probability $\geq 1 - \delta$

## 2 Case for $m = 1$ (Review from HW1)

The total variation distance between two probability distributions is defined as:

**Definition 1** (Definition 9.1)**.** *Given two known discrete probability distributions* $\mathbf{p}, \mathbf{q}$ *over* $[n]$*, the Total Variation distance* $d_{TV}(\mathbf{p}, \mathbf{q})$ *between* $\mathbf{p}$ *and* $\mathbf{q}$ *is defined as:*

$$d_{TV}(\mathbf{p}, \mathbf{q}) := \frac{1}{2} \sum_i |p_i - q_i| = \frac{1}{2} \|\mathbf{p} - \mathbf{q}\|_1 = \sup_{A \subseteq [n]} (\mathbf{p}(A) - \mathbf{q}(A)) .$$

[Theorem 9.2 (Variant of Neyman-Pearson Lemma)] For $m = 1$ sample from $D$. There exists an algorithm $A$ (such as the *Maximum Likelihood Tester*) such that:

$$\mathbb{P}(A = \mathbf{p} \mid D = \mathbf{p}) - \mathbb{P}(A = \mathbf{p} \mid D = \mathbf{q}) = \mathbb{P}(A = \mathbf{q} \mid D = \mathbf{q}) - \mathbb{P}(A = \mathbf{q} \mid D = \mathbf{p}) = d_{TV}(\mathbf{p}, \mathbf{q}).$$

Secondly, there does not exist an algorithm $A$ such that:

$$\mathbb{P}(A = \mathbf{p} \mid D = \mathbf{p}) - \mathbb{P}(A = \mathbf{p} \mid D = \mathbf{q}) > d_{TV}(\mathbf{p}, \mathbf{q}).$$

This also means there is no algorithm A such that the gap is greater than:

$$d_{TV}(\mathbf{p}, \mathbf{q})$$

. This implies that there is no algorithm $A$ such that both of the following hold:

- $\mathbb{P}(A = \mathbf{p} \mid D = \mathbf{p}) > \frac{1}{2} + \frac{1}{2} d_{TV}(\mathbf{p}, \mathbf{q})$
- $\mathbb{P}(A = \mathbf{q} \mid D = \mathbf{q}) > \frac{1}{2} + \frac{1}{2} d_{TV}(\mathbf{p}, \mathbf{q})$

It is, however, possible for an algorithm to satisfy one out of two (Ex: Hard code tester that always says p)
So if $d_{TV}(\mathbf{p}, \mathbf{q}) < \frac{1}{3}$, there is no algorithm that will succeed in distinguishing between two distributions with probability $\geq \frac{2}{3}$.

# 3 Case for $m > 1$

Question: Given known discrete distributions $\mathbf{p}$ and $\mathbf{q}$, what is the smallest $\mathbf{m}$ to win the game? (from beginning of lecture)

Answer: In this task, $\mathbf{m}$ is the sample complexity. Consider the problem through the lens of what we already know: $m$ i.i.d. samples from $D$ = one sample from $D^{\otimes m}$. Under this lens, $\mathbf{m}$ needs to be big enough such that $d_{TV}(\mathbf{p}^{\otimes m}, \mathbf{q}^{\otimes m}) \geq \frac{1}{3}$. Say we want to prove a lower bound on $\mathbf{m}$, we need an upper bound of $d_{TV}(\mathbf{p}^{\otimes m}, \mathbf{q}^{\otimes m})$:

**Fact 1** (Fact 9.3)**.** *For discrete probability distributions* $\mathbf{p}, \mathbf{q}$*, and for any* $m > 0$*:*

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

**Proposition 1** (Proposition 9.4)**.** *Sample Complexity* $= \Omega\left(\frac{1}{d_{TV}(\mathbf{p},\mathbf{q})}\right)$ *samples to successfully distinguish between* $\mathbf{p}, \mathbf{q}$ *with probability* $\geq \frac{2}{3}$.

Question: Is $\Omega\left(\frac{1}{d_{TV}(\mathbf{p},\mathbf{q})}\right)$ tight?
Answer: Yes. Consider Ber(0) vs Ber($d_{TV}(\mathbf{p}, \mathbf{q})$). In this case, we need $\mathrm{O}(\frac{1}{d_{TV}(\mathbf{p},\mathbf{q})})$ samples

Question: Can we do $\mathrm{O}(\frac{1}{d_{TV}(\mathbf{p},\mathbf{q})})$ samples in general?

Answer: No

**Proposition 2** (Proposition 9.5)**.** *Sample complexity* $\leq O\left(\frac{1}{d_{TV}(\mathbf{p},\mathbf{q})^2}\right)$

*Proof.* Let $A = \arg\sup_{A \subseteq [n]} (\mathbf{p}(A) - \mathbf{q}(A))$

Estimate $D(A) = \mathop{\mathbb{E}}_{x \in D}\{\mathbb{1}_A\}$ to additive error $\frac{d_{TV}(\mathbf{p},\mathbf{q})}{3}$

Now, we return the closer of $\mathbf{p}(\mathbf{A})$ and $\mathbf{q}(\mathbf{A})$

Since $d_{TV}(\mathbf{p}, \mathbf{q}) = \mathbf{p}(\mathbf{A}) - \mathbf{q}(\mathbf{A})$, if the error for our estimate of $\mathbf{D}(\mathbf{A})$ is within $\frac{d_{TV}(\mathbf{p},\mathbf{q})}{3}$, we can distinguish between $\mathbf{p}$ and $\mathbf{q}$ by returning the closer of $\mathbf{p}(\mathbf{A})$ and $\mathbf{q}(\mathbf{A})$

Our estimation of $D(A) = \mathop{\mathbb{E}}_{x \in D}\{\mathbb{1}_A\}$ to additive error $\frac{d_{TV}(\mathbf{p},\mathbf{q})}{3}$ results in the estimation of the mean of $\mathbb{1}_A$, which is a **0** to **1** Bernoulli random variable with unknown probability.

By Hoeffding, this can be done in

$$O\left(\frac{1}{d_{TV}(\mathbf{p},\mathbf{q})^2}\right) \text{ samples}$$

□

*Question: Is this bound tight?*
*Answer: Yes, in the worst case. Example: Ber* $\left(\frac{1}{2} \pm \epsilon\right)$ *needs* $\Omega\left(\frac{1}{\epsilon^2}\right)$ *samples.*

**Definition 2** (Definition 9.6(Squared Hellinger Distance)**.** *The Squared Hellinger distance is given by:*

$$d_H^2(\mathbf{p},\mathbf{q}) = \frac{1}{2}\sum_i \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 = 1 - \sum_i \sqrt{p_i q_i}.$$

*More properties:*

**Fact 2** (Fact 9.7)**.**

$$d_H^2(\mathbf{p},\mathbf{q}) \le d_{TV}(\mathbf{p},\mathbf{q}) \le \sqrt{2}\cdot d_H(\mathbf{p},\mathbf{q})$$

*Furthermore:*

$$d_H^2(\mathbf{p}^{\otimes m},\mathbf{q}^{\otimes m}) = 1 - \left(1 - d_H^2(\mathbf{p},\mathbf{q})\right)^m \le m \cdot d_H^2(\mathbf{p},\mathbf{q})$$

*Therefore:*

$$d_{TV}(\mathbf{p}^{\otimes m},\mathbf{q}^{\otimes m}) \le \sqrt{2}\cdot\sqrt{m}\cdot d_H(\mathbf{p},\mathbf{q})$$

**Proposition 3** (Proposition 9.8)**.** *Distinguishing p vs q with probability* $\ge \frac{2}{3}$ *requires*

$$\Theta\left(\frac{1}{d_H^2}\right)$$

*samples.*

*Proof.*

$$\Omega\left(\frac{1}{d_H^2}\right)$$

samples needed.

$$\text{If } m \le \frac{1}{100\cdot d_H^2}, \text{ then } d_{TV}(\mathbf{p}^{\otimes m},\mathbf{q}^{\otimes m}) \le \sqrt{2}\cdot\sqrt{m}\cdot d_H \le \frac{\sqrt{2}}{10} < \frac{1}{3}.$$

$$O\left(\frac{1}{d_H^2}\right) \text{ samples suffices.}$$

**Proof:** First, we prove the sample complexity is $\Omega(\frac{1}{d_H^2(\mathbf{p},\mathbf{q})})$

We have:
$m = \frac{1}{100 d_H^2(\mathbf{p},\mathbf{q})} \Rightarrow d_{TV}(\mathbf{p}^{\otimes m}, \mathbf{q}^{\otimes m}) \leq \sqrt{2} \cdot \sqrt{\frac{1}{100 d_H^2(\mathbf{p},\mathbf{q})}} \cdot d_H(\mathbf{p},\mathbf{q}) = \frac{\sqrt{2}}{10} < \frac{1}{3}$
We also have:

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

Now we take $m = O\left(\frac{1}{d_H^2(\mathbf{p},\mathbf{q})}\right)$, and note that $d_{TV}(\mathbf{p}^{\otimes m}, \mathbf{q}^{\otimes m}) \geq \frac{2}{3}$. This gives us two choices

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

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

□

### Succeeding with $1 - \delta$ Probability

*[Theorem 9.9] Distinguishing* $\mathbf{p}$ *vs* $\mathbf{q}$ *with probability* $\geq 1 - \delta$ *takes*

$$\Theta\left(\frac{1}{d_H^2(\mathbf{p},\mathbf{q})} \cdot \log\left(\frac{1}{\delta}\right)\right) \text{ samples.}$$

***Proof:***
***Upper Bound:*** *Use a majority vote after repeating* $\theta(\log \frac{1}{\delta})$ *times. (From Proposition 11.8)*

***Lower Bound:*** *The lower bound requires a tighter inequality of:*

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

# 4 Application: Mean Estimation

*We aim to estimate the mean of a Bernoulli distribution Bernoulli*$(p)$ *with an additive error* $\epsilon$ *and a probability of success at least* $\frac{2}{3}$*. Two methods have been discussed in previous lectures:*

1. *Computing the sample mean, which requires* $O\left(\frac{p(1-p)}{\epsilon^2}\right)$ *samples to achieve a probability of success of* $\frac{2}{3}$*.*

2. *Using the Median of Means, which requires* $O\left(\frac{p(1-p)}{\epsilon^2}\log\frac{1}{\delta}\right)$ *samples to achieve a success probability of* $1-\delta$.

*Both of these methods provide upper bounds on the sample complexity. Utilizing tools from this lecture, we can also establish corresponding lower bounds. If the mean* $p$ *can be estimated within an additive error* $\epsilon$*, then it becomes possible to differentiate between Bernoulli*$(p)$ *and Bernoulli*$(p+2\epsilon)$*. If the sample size is too small to distinguish these distributions, then the mean cannot be estimated within the desired error threshold. To simplify calculations, we instead compute the squared Hellinger distance, assuming* $\mathbf{p} = Bernoulli(p+\epsilon)$ *and* $\mathbf{q} = Bernoulli(p-\epsilon)$*, as follows:*

$$d_H^2(\mathbf{p},\mathbf{q}) = \Theta\left(\left(\sqrt{1-p+\epsilon}-\sqrt{1-p-\epsilon}\right)^2\right) + \Theta\left(\left(\sqrt{p+\epsilon}-\sqrt{p-\epsilon}\right)^2\right).$$

*For* $\epsilon < p$*, we derive:*

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

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

*Similarly, when* $p < \frac{1}{2}$*, we compute:*

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

*Thus, combining both cases, for* $p < \frac{1}{2}$ *and* $\epsilon < p$*:*

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

*We conclude with the following key results:*

1. *According to Proposition 9.8, distinguishing between the two distributions with success probability* $\frac{2}{3}$ *requires at least:*

$$\Omega\left(\frac{p(1-p)}{\epsilon^2}\right) \text{ samples.}$$

2. *By Theorem 9.9, achieving a success probability of* $1-\delta$ *demands at least:*

$$\Omega\left(\frac{p(1-p)}{\epsilon^2}\log\frac{1}{\delta}\right) \text{ samples.}$$