# Lecture 11



**Review** * Distribution learning

- minimax lower bounds: Fano's method
- Scheffé Selector

* Relating estimation to M-ary hypothesis testing

**Agenda:** → Hypothesis testing / Distribution Testing

→ Uniformity testing

## [A] Distribution Testing Formulation

Let $\mathcal{C}$ denote a class of distributions.

Denote $d(P, \mathcal{C}) = \min_{Q \in \mathcal{C}} d(P, Q)$.

Let $\overline{\mathcal{C}}_\varepsilon$ denote the class $\{P : d(P, \mathcal{C}) \geq \varepsilon\}$.

By observing samples $X_1, \ldots, X_n$, test if the samples were generated from $P \in \mathcal{C}$ or $P \in \overline{\mathcal{C}}_\varepsilon$.

Specifically, a test $T : \mathcal{X}^n \longrightarrow \{0, 1\}$ constitutes a $(\delta_1, \delta_2)$-test

if $\forall P \in \mathcal{C}, \quad P(T(X^n) = 0) \geq 1 - \delta_1,$

and

$\forall P \in \overline{\mathcal{C}}_\varepsilon, \quad P(T(X^n) = 1) \geq 1 - \delta_2.$

Denote by $n_{\delta_1, \delta_2}$ the least $n$ such that we can find such a test $T$. For simplicity, we fix $\delta_1 = \delta_2 = \frac{1}{3}$.

Denote $n^* = n_{1/3, 1/3}$.

A few popular formulations:

Consider $\mathcal{C} \subseteq \mathcal{P}_k \equiv (k-1)$-dimensional prob. simplex.

For all our examples, $\mathcal{C}$ will be closed ($\Rightarrow$ compact).

(1) Uniformity Testing: $\mathcal{C} = \{\mathrm{unif}[k]\}$

$$n^* = \Theta\left(\frac{\sqrt{k}}{\varepsilon^2}\right)$$

(2) Identity Testing: For $P \in \mathcal{P}_k$, $\mathcal{C} = \{P\}$.

$$n^* = \Theta\left(\frac{\sqrt{k}}{\varepsilon^2}\right).$$

(3) Closeness Testing: $\mathcal{X} = [k] \times [k]$

Are two sequences $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ generated from the same distribution or $(P, Q)$ s.t. $d(P,Q) \geq \varepsilon$.

[a slight modification of the general formulation]

$$n^* = \Theta\left(\max\left\{\frac{n^{2/3}}{\varepsilon^{4/3}}, \frac{\sqrt{n}}{\varepsilon^2}\right\}\right)$$

(4) Independence testing: Observe sequence $(X^n, Y^n)$:

$\mathcal{C} = \{P_X P_Y : P_X P_Y \in \mathcal{P}_k\}$

$$n^* = \Theta\left(\frac{n}{\varepsilon^2}\right)$$

$\boxed{B}$ Uniformity Testing: Collision-based tester

→ Goldreich and Ron, "On testing expansion in bounded-degree graphs", 2000.

③

→ Batu, Fischer, Fortnow, ... "Testing random variables for independence and identity," FOCS 2001.

→ Diakonikolas, Gouleakis, Peebles, Price, "Collision-based testers are optimal for uniformity and closeness", 2016.

Key idea: Under uniform distribution, the no. of samples required to see collisions is roughly $\sqrt{k}$. (Birthday Paradox) In fact, a careful analysis will show that the uniform distribution takes the "longest" for collisions to appear. Thus, we can make a test based on no. of collisions for uniformity.

The Test: $S \equiv S(X^n) = \sum_{i<j} \mathbb{1}\{X_i = X_j\}$

$$T(X^n) = \begin{cases} \text{unif}, & S \leq \tau \\ \varepsilon\text{-away from unif}, & S > \tau, \end{cases}$$

where $\tau = \binom{n}{2} \frac{1+c.\varepsilon^2}{k}$. (N.B. → requires the knowledge of $k$)

Analysis: Notation $\|P\|_\alpha = \sum_x P(x)^\alpha$

$$\mathbb{E}[S] = \binom{n}{2} P(X_1 = X_2) = \binom{n}{2} \|P\|_2^2$$

$$\mathbb{E}[S^2] = \mathbb{E}\left[\sum_{i_1<j_1} \sum_{i_2<j_2} \mathbb{1}\{X_{i_1} = X_{j_1}, X_{i_2} = X_{j_2}\}\right]$$

$$= \mathbb{E}\Big[\sum_{\substack{\text{two}\\\text{distinct}}} \mathbb{1}_{\{X_i = X_j\}} + \sum_{\substack{\text{three}\\\text{distinct}}} \mathbb{1}_{\{X_i = X_j = X_\ell\}}$$

$$+ \sum_{\substack{i_1<j_1,\ i_2<j_2\\ \text{all distinct}}} \mathbb{1}_{\{X_{i_1}=X_{j_1}\}}\, \mathbb{1}_{\{X_{i_2}=X_{j_2}\}}\Big]$$

$$= \binom{n}{2}\|P\|_2^2 + 6\binom{n}{3}\|P\|_3^3 + \binom{n}{2}\binom{n-2}{2}\|P\|_2^4$$

$$\Rightarrow Var(S) = \binom{n}{2}\|P\|_2^2 + n(n-1)(n-2)\|P\|_3^3$$

$$- \binom{n}{2}\Big[\underbrace{\frac{n!}{2!(n-2)!} - \frac{(n-2)!}{2!(n-4)!}}_{\frac{n(n-1)-(n-2)(n-3)}{2} = \frac{4n-6}{2} = 2n-3}\Big]\|P\|_2^4$$

$$= \binom{n}{2}\left(\|P\|_2^2 - \|P\|_2^4\right) + n(n-1)(n-2)\left(\|P\|_3^3 - \|P\|_2^4\right)$$

$$\le \binom{n}{2}\|P\|_2^2 + n^3\left(\|P\|_3^3 - \|P\|_2^4\right).$$

<u>Under uniform distribution</u> $(P = \text{unif}[k])$

$$\|P\|_2^2 = \frac{1}{k}, \quad \|P\|_3^3 = \frac{1}{k^2}, \quad \mathbb{E}[S] = \binom{n}{2}\frac{1}{k}$$

Therefore, $P\left(\left|S - \binom{n}{2}\frac{1}{k}\right| > \tau\right) \le \binom{n}{2}\frac{1}{k\tau^2}$

$$\Rightarrow P\left(\left|S - \binom{n}{2}\frac{1}{k}\right| > c\sqrt{3\binom{n}{2}\frac{1}{k}}\right) \le \frac{1}{3}, \text{ for every } c > 1.$$

(1) $P\left(S \le \binom{n}{2}\frac{1}{k} + c\sqrt{3}\cdot\sqrt{\binom{n}{2}\frac{1}{k}}\right) \ge 2/3, \quad \forall c > 1.$

For $P$ s.t. $d(P, \text{unif}) > \varepsilon$

$$\frac{1}{2}\sum_{i=1}^{k}\left|P_i - \frac{1}{k}\right| > \epsilon \Rightarrow \sum_{i=1}^{k} P_i^2 - \frac{1}{k} > \frac{4\epsilon^2}{k}.$$

$$\Rightarrow \|P\|_2^2 \geq \frac{1+4\epsilon^2}{k}$$

Case 1: $\binom{n}{2}\|P\|_2^2 > n^3\left(\|P\|_3^3 - \|P\|_2^4\right)$

$$P\left(\left|S - \binom{n}{2}\|P\|_2^2\right| > \tau\right) \leq \frac{n^2\|P\|_2^2}{\tau^2}$$

$$\Rightarrow P\left(S > \binom{n}{2}\|P\|_2^2 - c\sqrt{3}\; n\|P\|_2\right) \geq \frac{2}{3}, \quad \forall c > 1.$$

$$(2) \Rightarrow P\left(S > \binom{n}{2}\frac{1}{k} + 4\binom{n}{2}\frac{\epsilon^2}{k} - c\sqrt{3}\cdot\frac{n}{\sqrt{k}}\sqrt{1+4\epsilon^2}\right) \geq \frac{2}{3}, \quad \forall c > 1.$$

Case 2: $\binom{n}{2}\|P\|_2^2 \leq n^3\left(\|P\|_3^3 - \|P\|_2^4\right)$

We will use $-\|P\|_2^4$ to remove a $\frac{1}{k^2}$ term.

$$\|P\|_3^3 - \|P\|_2^4 \leq \sum_{i=1}^{k} P_i^3 - \frac{1}{k^2}$$

$$= \sum_{i=1}^{k}\left[\left(P_i - \frac{1}{k}\right)^3 + \frac{1}{k^3} + 3\left(P_i - \frac{1}{k}\right)\frac{1}{k^2} + 3\left(P_i - \frac{1}{k}\right)^2\frac{1}{k}\right] - \frac{1}{k^2}$$

$$= \sum_{i=1}^{k}\left(P_i - \frac{1}{k}\right)^3 + \frac{3}{k}\sum_{i=1}^{k}\left(P_i - \frac{1}{k}\right)^2$$

$$\leq \|P - P_{\text{unif}}\|_2^3 + \frac{3}{k}\|P - P_{\text{unif}}\|_2^2 = g_k^2\left(\|P - P_{\text{unif}}\|_2\right)$$

(To be continued in the next lecture)