Lecture 12

Review

For testing if $X_1, \ldots, X_n$ are generated by $P = \text{unif}[k]$ or some $Q \in \mathcal{P}_k$ s.t. $d(P, Q) \geq \varepsilon$, the least no. of samples $n^* \; (= n_{1/3 - 1/3})$ needed satisfy

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

Today: - Complete the proof of upper bound

- Lower bound

$\boxed{A}$ Analysis of Collision-based tester (contd.)

Recall $S = \sum_{i<j} \mathbb{1}_{\{X_i = X_j\}}$

$$\mathbb{E}_P[S] = \binom{n}{2} \|P\|_2^2$$

②

$$\mathbb{E}_P[S^2] = \binom{n}{2} \|P\|^2 + n(n-1)(n-2)\|P\|_3^3$$

$$\Rightarrow \mathrm{Var}(S) \leq n^2 \|P\|^2 + n^3\left(\|P\|_3^3 - \|P\|_2^4\right)$$

For uniform

With large prob.,

$$S \leq \binom{n}{2}\frac{1}{k} + \sqrt{3\,\frac{n^2}{k}}$$

For a P s.t. $d(P,\mathrm{unif}) \geq \varepsilon$

With large prob.,

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

Observations: (i) $\|P\|_2^2 = \|P-\mathrm{unif}\|_2^2 + \frac{1}{k}$

(ii) $\|P-\mathrm{unif}\|_2^2 \geq \frac{\|P-\mathrm{unif}\|_1^2}{k}$

$$= \frac{4\,d^2(P,\mathrm{unif})}{k}$$

Thus, with large prob.,

$$S \geq \binom{n}{2}\frac{1}{k} + \binom{n}{2}\|P-\mathrm{unif}\|_2^2 - \sqrt{3\left(n^2\|P\|^2 + n^3\left(\|P\|_3^3 - \|P\|_2^4\right)\right)}$$

Therefore, it suffices to have

(a) $\frac{n^2}{k} d^2(P, \text{unif}) \gg \sqrt{\frac{n^2}{k}}$

$$\Leftrightarrow n \gg \frac{\sqrt{k}}{d^2(P, \text{unif})} \Leftarrow n \gg \frac{\sqrt{k}}{\varepsilon^2};$$

and (b) $n^2 \|P - \text{unif}\|_2^2 \gg n \sqrt{\|P\|_2^2 + n\left(\|P\|_3^3 - \|P\|_2^4\right)}$.

Note that

$$\|P\|_2^2 + n\left(\|P\|_3^3 - \|P\|_2^4\right)$$
$$\leq \|P - \text{unif}\|_2^2 + \frac{1}{k} + n \|P - P_{\text{unif}}\|_2^3 + \frac{3n}{k} \|P - P_{\text{unif}}\|_2^2$$

Thus, it suffices to have

$$n \gg \max\left\{ \frac{1}{\|P - \text{unif}\|_2}, \frac{1}{\sqrt{k}\, \|P - \text{unif}\|_2^2}, \sqrt{\frac{n}{\|P - P_{\text{unif}}\|_2}}, \sqrt{\frac{n}{k} \frac{1}{\|P - P_{\text{unif}}\|_2^2}} \right\}$$

$$\Leftarrow n \gg \max\left\{ \frac{\sqrt{k}}{\varepsilon}, \frac{\sqrt{k}}{\varepsilon^2}, \sqrt{\frac{nk}{\varepsilon}}, \sqrt{\frac{nk}{\varepsilon^2}} \right\}.$$

$$\Leftrightarrow n \gg \frac{\sqrt{k}}{\varepsilon^2}.$$

[B] Lower Bound

We shall prove $n^* = \Omega\left(\frac{\sqrt{k}}{\varepsilon^2}\right)$

(1) Let's show a simpler result first,

namely, $n^* = \Omega(\sqrt{k})$

Lemma (Birthday Paradox)

For a uniform distribution on $[k]$,

for $n \ll \sqrt{k}$, with large prob., no. repetitions occur.

Proof Sketch. Poisson approximation

Let $N \sim Poi(n)$. Consider N iid samples $X_1, \ldots, X_N$ from $P \in \mathcal{P}_k$. Then,

(a) $N_x := \sum_{i=1}^{N} \mathbb{1}_{\{X_i = x\}} \equiv$ # of times x occurs in X^N, $x \in \mathcal{X}$.

$\{N_x\}$ are independent for different x,

with $N_x \sim Poi(nP_x)$.

Let $M_2 = \sum_x \mathbb{1}_{\{N_x \geq 2\}}$, i.e., no. of symbols appearing 2 or more times.

Then,

$$\mathbb{P}(M_2 > 0) = \mathbb{P}(M_2 \geq 1)$$

$$\leq \mathbb{E}[M_2] = \sum_x \mathbb{P}(N_x \geq 2)$$

$$= \sum_x \frac{(nP_x)^2}{2} e^{-nP_x}$$

For $P \equiv \text{unif}[k]$,

$$\mathbb{P}(M_2 > 0) \leq k \cdot e^{-n/k} \; \frac{n^2}{k^2} \leq \frac{n^2}{k}$$

(b) Removing Poisson approx.

$$\mathbb{P}(M_2 > 0) \geq \mathbb{P}\left(N \in \left[\frac{n}{2}, 2n\right]\right) \mathbb{P}\left(M_2 > 0 \,\middle|\, N = \frac{n}{2}\right)$$

$$\geq \left(1 - e^{-cn}\right) \cdot \mathbb{P}\left(M_2 > 0 \,\middle|\, N = \frac{n}{2}\right)$$

$\Rightarrow$ For $\frac{n}{2}$ samples, ~~no~~ repetitions occur with prob. $\leq \frac{n^2}{k} (1 - e^{-cn})^{-1}$.

Therefore, no repetitions occur w.p. ≈ 1 if $n \ll \sqrt{k}$

Now the proof of lower bound proceeds as follows:

Consider the distributions $Q_A = \text{unif}\{A\}$ for subsets A with $|A| = k/2$. Each Q_A is at a distance $d(\text{unif}[k], Q_A) = \frac{1}{2}$.

Let $\bar{Q}^{(n)} = \frac{1}{\binom{k}{k/2}} \sum_{A : |A| = k/2} Q_A^{\otimes n}$.

By symmetry, a sufficient statistic to distinguish unif from $\bar{Q}^{(n)}$ is the so called profile, i.e, the vector $\Phi = (\Phi_1, \ldots, \Phi_n)$ where Φ_i = # of symbols appearing i times.

But for $n \ll \sqrt{k}$, by Birthday Paradox, the profiles under unif$[k]$ and $\bar{Q}^{(n)}$ are exactly $\Phi = (n, 0, \ldots, 0)$ with large prob.

The proof is completed by noting that any test for $\varepsilon \leq \frac{1}{2}$ can distinguish $\bar{Q}^{(n)}$ from unif$[k]$.

(2) Paninski's lower bound

Consider now the family $\{Q_z\}_{z \in \{-1,1\}^{k/2}}$ defined as follows: under Q_z, the elements $2i$ and $2i+1$, $0 \le i \le \frac{k-1}{2}$, have masses $\frac{1+2\varepsilon z_i}{k}$ and $\frac{1-2\varepsilon z_i}{k}$, resp.

Thus, $d(Q_z, \text{unif}[k]) = \varepsilon$ for every $z \in \{-1,1\}^{k/2}$.

Let $\bar{Q}^{(n)} = \frac{1}{2^{k/2}} \sum_z Q_z^{\otimes n}$.

We want to bound $d(\text{unif}[k]^{\otimes n}, \bar{Q}^{(n)})$.

Aside: $d^2(P,Q) \le D(P\|Q)$ (if $P \ll Q$, $Q \ll P$)

$$= \mathbb{E}_P\left[\log \frac{P(X)}{Q(X)}\right]$$

$$\le \mathbb{E}_P\left[\left|\frac{P(X)}{Q(X)} - 1\right|\right]$$

Alternatively,

$$4d^2(P,Q) = \mathbb{E}_Q\left[\left|\frac{P(X)}{Q(X)} - 1\right|\right]$$

$$\le \sqrt{\mathbb{E}_Q\left[\left(\frac{P(X)}{Q(X)} - 1\right)^2\right]}.$$

Q. Which of these bounds do you prefer?

For our constuction,

$$\frac{Q_z^{\otimes n}(\underline{x})}{P(\underline{x})} = \prod_{i=1}^{n} \left[1+g(x_i, z)\right]$$

where $g(x_i, z) = \begin{cases} 2\varepsilon z_j & \text{if } x_i = 2j \\ -2\varepsilon z_j & \text{if } x_i = 2j+1. \end{cases}$

Thus,

$$\left(\sum_z 2^{-k/2} \frac{Q_z^{\otimes n}(\underline{x})}{P(\underline{x})} - 1\right)^2$$

$$= \left(2^{-k/2} \sum_z \left[1+\sum_i g(x_i, z) + \sum_{i_1<i_2} g(x_{i_1}, z) g(x_{i_2}, z) + \dots\right] - 1\right)^2$$

$$= 2^{-k} \sum_{z, z'} \left(\sum_{i_1, i_2} g(x_{i_1}, z) g(x_{i_2}, z') + \sum_{i_1, i_2 \le i_3} g(x_{i_1}, z) g(x_{i_2}, z) g(x_{i_3}, z') + \dots \right)$$

Under P, $\mathbb{E}\left[g(X_i, z)\right] = 0$ and $X_1, \dots, X_n$ are iid. Thus, only the terms involving pairs $g(X_i, z)\, g(X_i, z')$ remain on taking $\mathbb{E}_P$.

Therefore,

$$\mathbb{E}_{p^{\otimes n}}\left[\left(\frac{\bar{Q}^{\otimes n}}{p^{\otimes n}}(x^n)-1\right)^2\right]$$

$$= 2^{-k}\sum_{z,z'}\left[\sum_j H_j(z,z') + \sum_{j>j'} H_j(z,z')H_{j'}(z,z') + \dots\right],$$

where

$$H_j(z,z') = \mathbb{E}_p\left[g(X_j,z)\,g(x_j,z')\right]$$

$$= \frac{1}{k}\sum_{i=1}^{k/2} g(2i,z)\,g(2i,z') + g(2i+1,z)\,g(2i+1,z')$$

$$= \frac{8\varepsilon^2}{k}\sum_{i=1}^{k/2} z_i \cdot z_i'.$$

Thus, the left-side above is bounded by

$$2^{-k}\sum_{z,z'}\prod_{j=1}^{n}\left(1+H_j(z,z')\right) - 1$$

$$\leq \mathbb{E}_{z,z'}\left[e^{\sum_{j=1}^{n} H_j(z,z')}\right] - 1$$

$$\leq \mathbb{E}_{zz'}\left[e^{\frac{8n\varepsilon^2}{k}\sum_{i=1}^{k/2} z_i z_i'}\right] - 1$$

$$\leq e^{C\cdot\frac{n^2\varepsilon^4}{k}} - 1 \Rightarrow \boxed{\frac{n^2\varepsilon^4}{k} \geq \text{constant}}$$