Distribution Testing

Δίνονται 2 κατανομές p, q

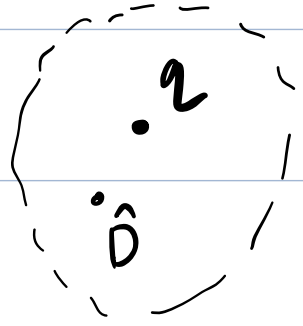
και sample access από μια άγνωστη D

Ξέρουμε ότι είτε $D = p$

είτε $D = q$

Πόσα δείγματα χρειαζόμαστε για να διακρίνουμε

αυτές τις περιπτώσεις;



- Approach 1

Μαθαίνω $\hat{D}$ τ.ω. $d_{TV}(D, \hat{D}) < \frac{\varepsilon}{2}$

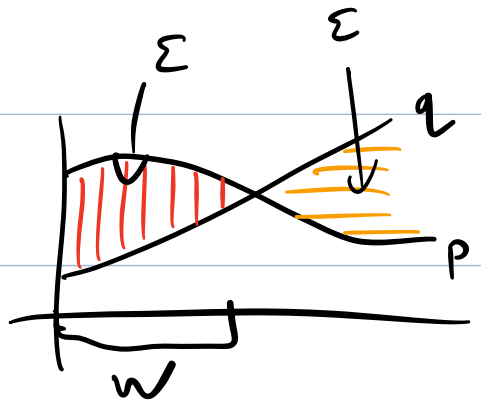
με πιθανότητα $1-\delta$

Χρειάζομαι $m = \Omega\left(\frac{n + \log(1/\delta)}{\varepsilon^2}\right)$

- Approach 2

$$d_{TV}(p,q) = \varepsilon$$

$$\| \quad \sum_i (p_i - q_i)_+$$



$$W = \{ i \in [n] : p_i > q_i \}$$

$$p(w) = q(w) + \varepsilon$$

$$D(w) = p(w)$$

$$\| \quad \eta$$

$$D(w) = q(w)$$

Μετράω το ποσοστό δειγμάτων της D που έπεσαν στο W

$$|\hat{d} - D(w)| < \frac{\varepsilon}{2}$$

Χρειάζομαι μόνο $O(1/\varepsilon^2)$ δείγματα

Αν $p = D$

$$|\hat{d} - p(w)| < \frac{\varepsilon}{2}$$

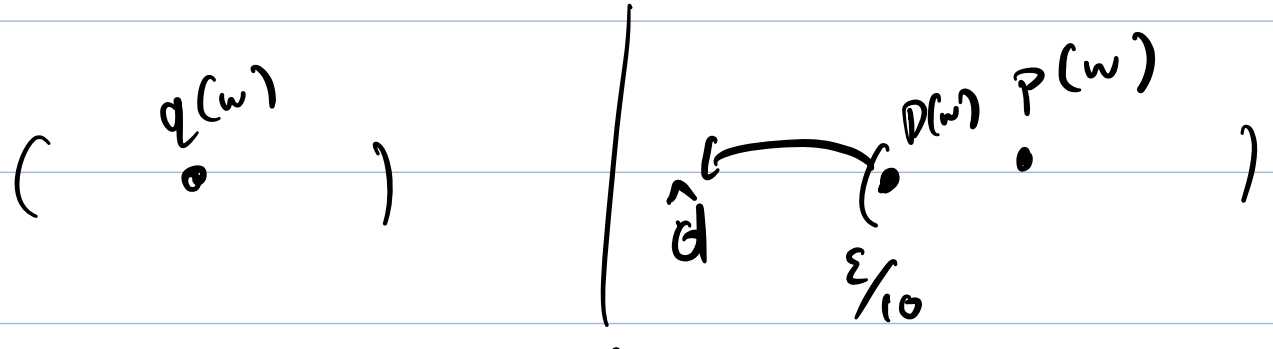
$$|\hat{d} - q(w)| = |\hat{d} - p(w) - \varepsilon|$$
$$\geq \varepsilon - |\hat{d} - p(w)|$$
$$> \varepsilon/2$$

Έστω ότι η D είναι μόνο $\varepsilon/10$ κοντά στην p ή q δλδ

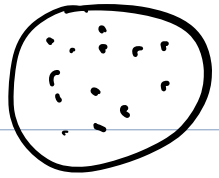
$$d_{TV}(D, p) < \frac{\varepsilon}{10} \quad \text{ή} \quad d_{TV}(D, q) < \frac{\varepsilon}{10}$$

$$d_{TV}(p, q) = \varepsilon$$

$$|p(w) - D(w)| < \frac{\varepsilon}{10}$$



Από Testing σε Learning



Οικογένεια κατανομών $\mathcal{D}$

ε-cover: Για μια οικογένεια κατανομών $\mathcal{D}$ ένα υποσύνολο $\mathcal{D}_\varepsilon \subseteq \mathcal{D}$ λέγεται ε-cover αν για κάθε $p \in \mathcal{D}$ $\exists\, p' \in \mathcal{D}_\varepsilon$ ώστε $d_{TV}(p, p') \leq \varepsilon$

π.χ. $D = \{ Ber(p) : p \in [0,1] \}$

$D_\varepsilon = \{ Ber(p) : p \in \{\varepsilon, 2\varepsilon, 3\varepsilon, 4\varepsilon, \ldots\} \}$

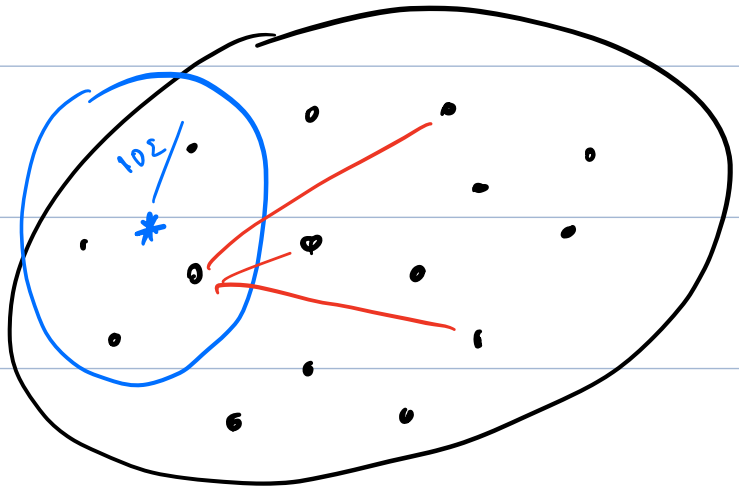
$|D_\varepsilon| = \frac{1}{\varepsilon}$

π.χ. $D = \{ p \in [0,1]^n : \sum_{i=1}^{n} p_i = 1 \}$

$D_\varepsilon = \{ p \in \{\frac{\varepsilon}{n}, 2\frac{\varepsilon}{n}, 3\frac{\varepsilon}{n}, \ldots\}^n : \Sigma p_i = 1 \}$

$|D_\varepsilon| \leq \left(\frac{n}{\varepsilon}\right)^n \cdot \cdot 1 \cdot \cdot 1 \cdot \cdot 1 \cdot \cdot 1$

$\binom{\frac{n}{\varepsilon}}{n} \simeq \left(\frac{1}{\varepsilon}\right)^n$



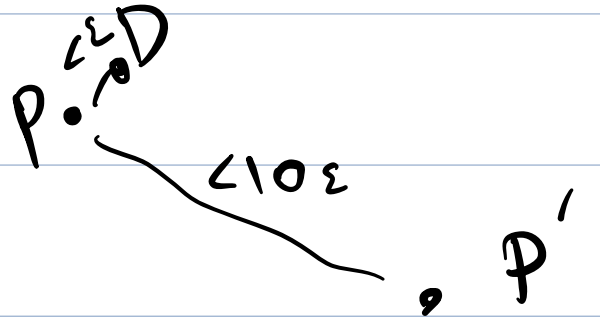
Tournament :

Τρέχω τον testing αλγόριθμο για κάθε ζεύγος κατανομών

- Νικητής είναι όποιος κερδίζει όλες τις κατανομές σε απόσταση 10 ε

Έστω P που έχει

$$d_{TV}(P, D) < \varepsilon$$



Αν δοκιμάσω έναν τέτοιο νικητή P'

$$d_{TV}(P', D) < 11\varepsilon$$

Η πιθανότητα σφάλματος $\leq |D_\Sigma|^2 \delta$

Αν θέλω συνολικά $\leq \delta'$

θέτω $\delta = \frac{\delta'}{|D_\Sigma|^2}$

$|D_\Sigma|^2$

Συνολικά το πλήθος των δειγμάτων που χρειάζομαι είναι

$$\Theta\left(\frac{1}{\varepsilon^2}\log(1/\delta)\right) \quad \text{δηλαδή}$$

$$\Theta\left(\frac{1}{\varepsilon^2}\log\left(|D_\Sigma|^2/\delta'\right)\right)$$

$$\Theta\left(\frac{\log|D_\Sigma|}{\varepsilon^2} + \frac{\log(1/\delta')}{\varepsilon^2}\right)$$

Για διακριτές κατανομές σε n στοιχεία

$$|D_\varepsilon| \leq \left(\frac{1}{\varepsilon}\right)^n$$

$$\Theta\left(\frac{\log|D_\varepsilon|}{\varepsilon^2} + \frac{\log(1/\delta)}{\varepsilon^2}\right)$$

$$=$$

$$\Theta\left(\frac{n\log(1/\varepsilon)}{\varepsilon^2} + \frac{\log(1/\delta)}{\varepsilon^2}\right)$$

$$=$$

$$\tilde{\Theta}\left(\frac{n}{\varepsilon^2} + \frac{\log(1/\delta)}{\varepsilon^2}\right)$$

$$\tilde{\Theta}(f(n,m)) \equiv \Theta(f(n,m)\log^{O(1)} f(n,m))$$

Distribution Testing

Έστω ότι μου δίνεται μια αγνωστη q και έχω δείγματα από μια p

Πρέπει να διακρίνω

- $p = q$
- $d_{TV}(p, q) > \varepsilon$

Γενικά χρειάζομαι $O\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ δείγματα για να λύσουμε το πρόβλημα

Uniformity Testing $p = \left(\frac{1}{n}, \dots, \frac{1}{n}\right)$ ή

$d_{TV}(p, (\frac{1}{n}, \dots, \frac{1}{n})) > \varepsilon$

Συνεπάγεται $\frac{\|p\|_{2/3}}{\varepsilon^2} = \left(\sum p_i^{2/3} \right)^{3/2} \leq \sqrt{n}$

δείγματα για γενικό p

Αλγόριθμος

Με m δείγματα από την p

$C \leftarrow$ # collisions

δηλ $C = \sum_{i<j}^{m} 1\{x_i = x_j\}$

Αν $C < \frac{1 + \varepsilon^2/2}{n} \binom{m}{2}$

τότε "Uniform"

Αλλιώς "οχι Uniform"

- Αναμενο πλήθος Collisions

$$E[C] = \sum_{i<j} Pr[x_i = x_j]$$

$$= \sum_{i<j} \sum_{k=1}^{n} Pr[x_i = k \wedge x_j = k]$$

$$= \sum_{i<j} \sum_{k=1}^{n} p_k^2$$

$$= \sum_{k=1}^{n} p_k^2 \cdot \binom{m}{2}$$

$$= \|p\|_2^2 \binom{m}{2}$$

$$\|p\|_2^2 \geq \frac{\|p\|_1^2}{n} = \frac{1}{n}$$

$(=$ όταν p uniform$)$

Έστω ότι το p έχει

$$d_{TV}(p, U_n) > \varepsilon$$

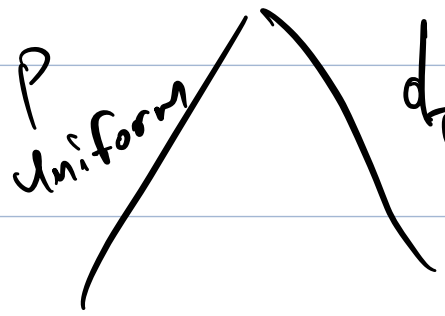
$$\|p - U_n\|_1 > \varepsilon$$

$$\Downarrow$$

$$\|p - U_n\|_2^2 > \frac{\varepsilon^2}{n}$$

$$\|p\|_2^2 = \|U_n\|_2^2 + \|p - U_n\|_2^2$$

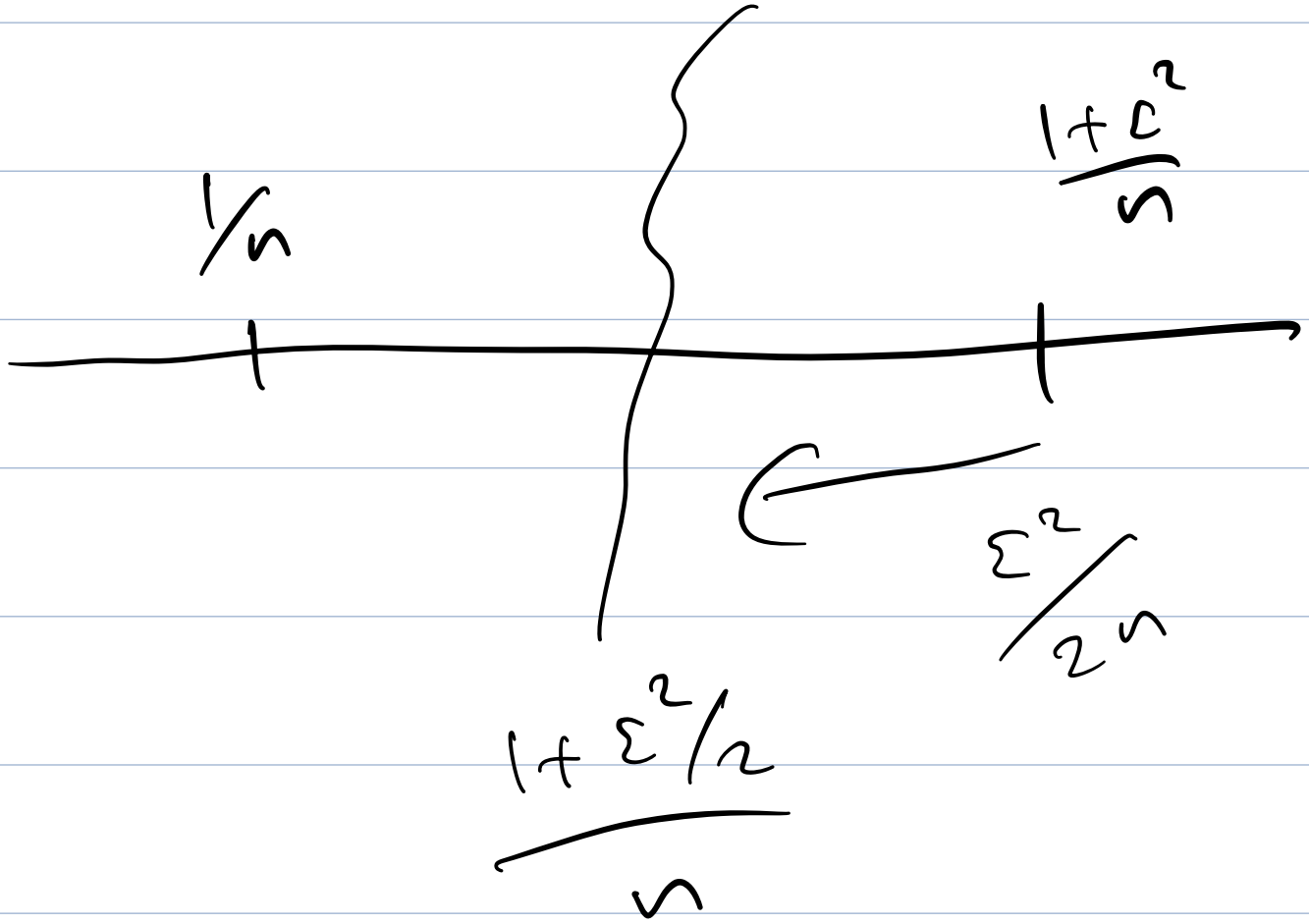
$$\geq \frac{1}{n} + \frac{\varepsilon^2}{n}$$



$d_{TV}(p, U_n) > \varepsilon$

$E[C] > \binom{n}{2} \frac{1+\varepsilon^2}{n}$

$E[C] = \binom{n}{2} \frac{1}{n}$



Ano Chebyshev

$$\Pr\left[\left|C - E[C]\right| \geq 2\sqrt{\mathrm{Var}[C]}\right] \leq \frac{1}{4}$$

$$\mathrm{Var}\left[\frac{C}{\binom{m}{2}}\right] \leq \frac{\varepsilon^2}{4n}$$

$$\mathrm{Var}[C] \leq 2\, m^3 \|p\|_2^3$$

$$\mathrm{Var}\left[\frac{C}{\binom{m}{2}}\right] \leq O\left(\frac{\|p\|_2^3}{m}\right)$$

$$\|p\|_2^3 \leq \frac{\varepsilon^4}{}$$

$$\frac{}{m} - \frac{}{16n^2}$$

$$m \geq O\left(\frac{n^2 \|p\|_2^3}{\varepsilon^4}\right) \simeq O\left(\frac{\sqrt{n}}{\varepsilon^4}\right)$$