Lecture 16:

Hypothesis Testing

A useful tool: Hypothesis Testing

Given collection of distributions, $\mathcal{H}$, at least one (← via complete description)
has high accuracy for describing p ← given via samples
output one of collection that is close to p.

How many samples in terms of $|\mathcal{H}|$ + domain size?

Why is this different than testing closeness, uniformity?
Do we have the same lower bounds?

NO

Since p is guaranteed to be close to some $q \in \mathcal{H}$, all bets are off!!

A "subtool": allows comparing two hypothesis

Thm Given sample access to p
Given h_1, h_2 hypothesis distributions (fully known to algorithm)
Given accuracy parameter ε', confidence δ'
Algorithm "Choose" takes $O(\log(1/\delta')/(\varepsilon')^2)$ samples + outputs $h \in \{h_1, h_2\}$. If one of h_1, h_2 has $\|h_i - p\|_1 < \varepsilon'$
Then with prob $\geq 1-\delta'$, output h_j has $\|h_j - p\|_1 \leq 12\varepsilon'$

- if both h_1, h_2 far, no guarantees
- if one is close, you output something pretty close.

Actually, will prove something stronger:

Thm p given via samples

h_1, h_2 fully known & p is ε'-close to at least one of h_1, h_2

ε', δ' given

Algorithm "Choose" takes $O(\log(1/\delta')(1/\varepsilon')^2)$ samples

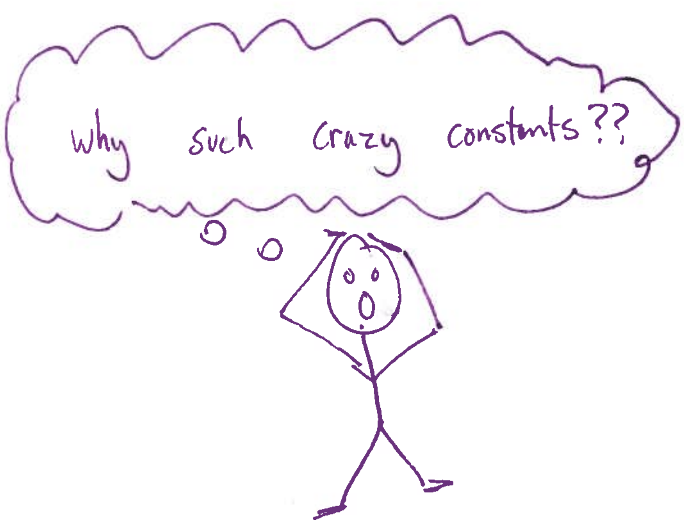
& outputs $h \in \{h_1, h_2\}$ satisfying:

(1) if h_i more than $12\varepsilon'$-far from p, unlikely to output it as winner or tie

  ($12\varepsilon'$: very bad; unlikely: $2e^{-m\varepsilon'^2/2}$)

(2) if h_i more than $10\varepsilon'$-far, unlikely to output as winner

  ($10\varepsilon'$: not that bad; as winner: might tie but won't win)



Proof of "Subtool":

idea: wlog h_1 is ε'-close,

if h_2 is $10\varepsilon'$-close, then either output ok as "winner" or "tie"

else, if h_2 is not $10\varepsilon'$-close but is $12\varepsilon'$-close, ok to "tie" or output h_1

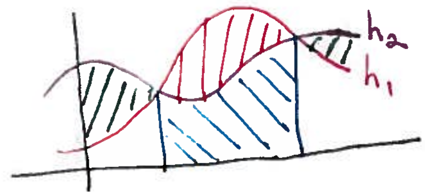
else, h_2 is $12\varepsilon'$-far from h_1 & $11\varepsilon'$-far from p

so samples from p will fall in "difference" between h_1 & h_2 & will output h_1

Algorithm Choose:

Input p, h_1, h_2 (p: samples; h_1, h_2: explicit description)

First some definitions:



$$A = \{x \mid h_1(x) > h_2(x)\}$$

$$a_1 = h_1(A)$$

$$a_2 = h_2(A)$$

note $\|h_1 - h_2\|_1 = 2(a_1 - a_2)$

factor of 2 in constants →

green area = red area

L_1 dist = green + red

red area = $a_1 - a_2$

blue area = a_2

blue + red area = a_1

this is important! ⇒

1. if $a_1 - a_2 \le 5\varepsilon'$ declare "tie" & return h_1 (no samples needed)

additive factor in constants ↗

2. draw $m = 2 \cdot \frac{\log \frac{1}{\delta'}}{(\varepsilon')^2}$ samples $s_1 \ldots s_m$ from p

3. $\alpha \leftarrow \frac{1}{m} \left| \{ i \mid s_i \in A \} \right|$

if $p = h_1$, $E[\alpha] = a_1$

if $p = h_2$, $E[\alpha] = a_2$

another additive factor in constants (will see breakdown on next page) →

4. if $\alpha > a_1 - \frac{3}{2}\varepsilon'$ return h_1

   else if $\alpha < a_2 + \frac{3}{2}\varepsilon'$ return h_2

   else declare "tie" & return h_1

Why does it work?

$E[\alpha] = p(A)$

- if reach step 2, whp (via Chernoff) $|\alpha - E[\alpha]| \le \frac{\varepsilon'}{2}$

  if $\|p - h_1\|_1 > 12\varepsilon'$ or $\|p - h_2\|_1 > 12\varepsilon'$ then since other is $\le \varepsilon'$ distance, $\|h_1 - h_2\|_1 > 11\varepsilon'$

  so will reach step 2

  if p ε'-close to h_1, whp $\alpha > a_1 - \varepsilon' - \frac{\varepsilon'}{2}$ (↑ from closeness to h_1; ↖ sampling error)

  so output h_1

  else, p is $12\varepsilon'$ far from h_1 but ε'-close to h_2

  whp $\alpha < a_2 + \varepsilon' + \frac{\varepsilon'}{2}$ (affects constants)

- if one of h_1, h_2 ε'-close from and other is $> 10\varepsilon'$ far but not $12\varepsilon'$ far $\Rightarrow$ return h_2 whp

  if $a_1 - a_2 \le 5\varepsilon'$ then declares draw, so neither are declared "winner"

  else $\|h_1 - h_2\| > 9\varepsilon'$ far + similar reasoning shows that medium far will not win (in fact, will lose)

  recall: $\|h_1 - h_2\|_1 = 2(a_1 - a_2)$

- if both are $10\varepsilon'$-close, might output h_1, h_2 or "tie"

The Cover Method

a method for learning distributions

def $\mathcal{C}^{\mathcal{D}}$ is a **ε-cover** of $\mathcal{D}$ if $\forall\, p \in \mathcal{D}$ $\exists\, q \in \mathcal{C}^{\mathcal{D}}$ s.t. $\|p-q\|_1 \le \varepsilon$

$\mathcal{C}^{\mathcal{D}}$ ↑ set of distributions (smaller)

$\mathcal{D}$ ↑ set of distributions (big)

Why useful? hopefully $\mathcal{C}$ is much smaller than $\mathcal{D}$ – allows us to "approx" $\mathcal{D}$

note $\mathcal{C}$ not unique

Big improvement: union bnd over size of $\mathcal{C}^{\mathcal{D}}$ not $\mathcal{D}$!! ⟹

Thm $\exists$ algorithm, given $p \in \mathcal{D}$, which takes $O\left(\frac{1}{\varepsilon^2}\log|\mathcal{C}^{\mathcal{D}}|\right)$ samples of p & outputs $h \in \mathcal{C}^{\mathcal{D}}$ s.t. $\|h-p\|_1 \le 6\varepsilon$ with prob $\ge 9/10$

Pf.

since $p \in \mathcal{D}$, $\exists\, q \in \mathcal{C}^{\mathcal{D}}$ s.t. $\|p-q\|_1 \le \delta$

(but there could be more than 1) ⟸ we just need to find one, not even required to return p

will run Choose on p with every pair $q_1, q_2 \in \mathcal{C}^{\mathcal{D}}$

if q doesn't win all of its "matches" then it loses to someone that is not so bad

Furthermore can show that whp there is a q' s.t. (it is $\le 10\varepsilon$ far)

q' wins or ties all matches. (best q never loses, any one that ties it is $\le 10\varepsilon$ far)

need all matches to give correct output – union bound on $\binom{|\mathcal{C}|}{2}$ matches ∎

The cover method

Example 1: learning distribution of a coin

domain = $\{0,1\}$

need to learn bias

Here $\mathcal{D} = \mathbb{R}$ ← describes bias of coin

if use $\mathcal{C} = \{0, \frac{1}{k}, \frac{2}{k}, \ldots, \frac{k-1}{k}, 1\}$

then $\forall$ bias p, let $\frac{i}{k} \leq p \leq \frac{i+1}{k}$

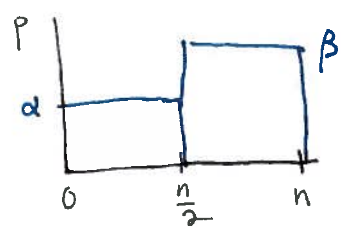
then picking $\tilde{p} = \frac{i}{k}$ gives $\|p-\tilde{p}\|_1 = \left|\frac{i}{k} - p\right| + \left|(1-\frac{(i+1)}{k}) - (1-p)\right|$

$\leq \frac{2}{k}$

so using $k = \Theta(\frac{1}{\varepsilon})$ gives $\|p-\tilde{p}\|_1 \leq \varepsilon$

$|\mathcal{C}| = k+1 = \Theta(\frac{1}{\varepsilon})$, # samples needed by cover method is $O(\frac{1}{\varepsilon^2} \cdot \log \frac{1}{\varepsilon})$

Example 2: 2-bucket distributions



now need to specify α and β

so $\mathcal{C} = \{ (\frac{i}{k}, \frac{j}{k}) \mid i,j \in \{0 \ldots k\} \}$

$|\mathcal{C}| = \Theta((\frac{1}{\varepsilon})^2)$

samples is $O(\frac{1}{\varepsilon^2} \log \frac{1}{\varepsilon})$

Example 3: monotone distributions

Birge $\Rightarrow$ $\mathcal{C} = \{ (\frac{i_1}{k}, \ldots, \frac{i_{(\log n/\varepsilon)}}{k}) \mid i_1, i_2, \ldots \in \{0 \ldots k\} \}$

$|\mathcal{C}| = \Theta\left(\frac{1}{\varepsilon}^{(\log n)/\varepsilon} \right) \Rightarrow$ # samples is $O(\frac{1}{\varepsilon^3} \log n \cdot \log \frac{1}{\varepsilon})$