Lecture 16:

Hypothesis Testing

ignoring ε's
log n's

Some Problems: (Given samples of p)

Complexity (in terms of $n = |D|$)

is $p = q$ (e.g. $q = U_D$) or ε-far from q	$\sqrt{n}$
is p ε-close to q or 2ε-far from q	$\frac{n}{\log n}$
(Given samples of q) is $p = q$ or p ε-far from q	$n^{2/3}$
(Given samples of q) is p ε-close to q or 2ε-far from q	$\frac{n}{\log n}$
is p monotone or ε-far from monotone	$\sqrt{n}$
is p ε-close to monotone or ε-far from monotone	$\frac{n}{\log n}$

Other problems considered:

estimate entropy, support size

independence?

represented well via k-histogram?

monotone hazard rate

$\vdots$

A useful tool:

Given: (1) collection of distributions (via complete description) $\mathcal{H}$

(2) Samples of p such that $\underbrace{\exists\, q \in \mathcal{H}}_{\mathcal{H} \text{ contains good approximation to } p}$ for which $\text{dist}(p,q)$ is small

$\Uparrow$ strong assumption

Goal: Output $h \in \mathcal{H}$ s.t. $\text{dist}(p,h)$ small

Question:

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

Is this the same as testing closeness, uniformity?

Do lower bounds apply?

NO! guaranteed p close to some $q \in \mathcal{H}$

What we want:

Given h_1, h_2 explicit distributions

p via samples

procedure that outputs h_i that is closer to p

What if both are roughly same distance?

maybe either one is ok?

or maybe not...

More general Goal:

Given set of hypotheses $\mathcal{H}$

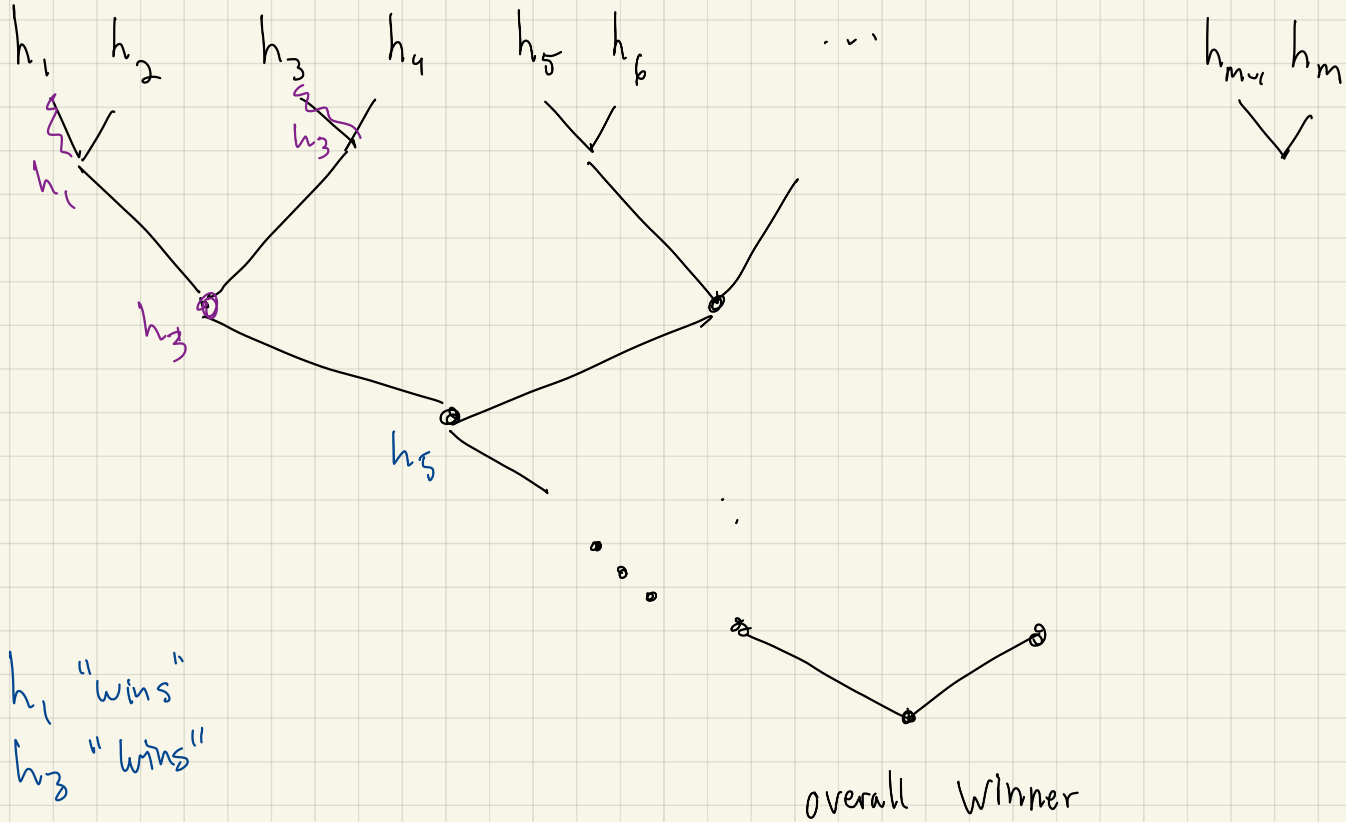
& p via samples

find $h \in \mathcal{H}$ closest to p

or pretty close to best

Find best hypothesis via "tournament"?

"winner" advances at each phase



maybe $p = h_2$

$\|p - h_1\|_1 < \varepsilon$ & h_1 "wins"

$\|p - h_3\|_1 = 2\varepsilon$ & h_3 "wins"

$\|p - h_5\|_1 = 3\varepsilon$ & h_5 "wins"

$\vdots$

overall winner $\leq O(\log n \cdot \varepsilon)$ distance from best hypothesis?

- won't use simple tournament ← instead compare every pair
- will add notion of "tie"

Output hypothesis that <u>wins</u> or <u>ties</u> every match

- hopefully there is one
- " it is close to p

A "subtool" for comparing two hypotheses:

Thm given (1) sample access to p

(2) h_1, h_2 hypothesis distributions (fully known to algorithm)

(3) accuracy parameter ε', confidence parameter δ'

then Algorithm "choose" takes $O(\log(\frac{1}{\delta'})/(\varepsilon')^2)$ samples + outputs $h \in \{h_1, h_2\}$ satisfying:

if one of h_1, h_2 has $\|h_i - p\|_1 < \varepsilon'$

then with prob $\geq 1-\delta'$, output h_j has $\|h_j - p\|_1 < 12\varepsilon'$

i.e. if both h_1, h_2 far no guarantees

if one ε'-close + one is really far ($> 12\varepsilon'$) we will output close one

if one is ε' close, other is $10\varepsilon'$ close $\rightarrow$ output either one? but both $\leq 10\varepsilon'$ close so not too bad

getting kind of complicated just to specify?

Actually a bit stronger: (focus on case where ≥ 1 close)

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 $O\left(\left(\log \frac{1}{\delta'}\right)\left(\frac{1}{\varepsilon'}\right)^2\right)$ samples & outputs $h \in \{h_1, h_2\}$ such that:

(1) If h_i more than $\underbrace{12\varepsilon'\text{-far}}_{\text{very far}}$ from p, $\underbrace{\text{unlikely}}_{2e^{-m(\varepsilon')^2/2}}$ to output h_i as winner

<u>or</u> tie

(2) If h_i more than $\underbrace{10\varepsilon'\text{-far}}_{\text{pretty far but not too bad}}$ from p, unlikely to output h_i as winner

BUT COULD TIE

Proof of subtool:

<u>idea</u>: wlog h_1 is ε'-close to p

if h_2 is close to h_1, then h_2 is pretty close to p + can win or tie

else h_2 is far from h_1, then also far from p

$\underbrace{\text{red + green}}_{\text{distance of } h_1 + h_2}$ areas are big

estimate $\Pr_{x \in p}[x \in A]$ see if it is more like h_1 or more like h_2

A = part of domain where $h_1(x) > h_2(x)$

Algorithm Choose: Input $\underset{\text{samples}}{p}, \underset{\text{explicit description}}{h_1, h_2}$

First some definitions:

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

$a_1 = h_1(A)$ ← red + blue areas

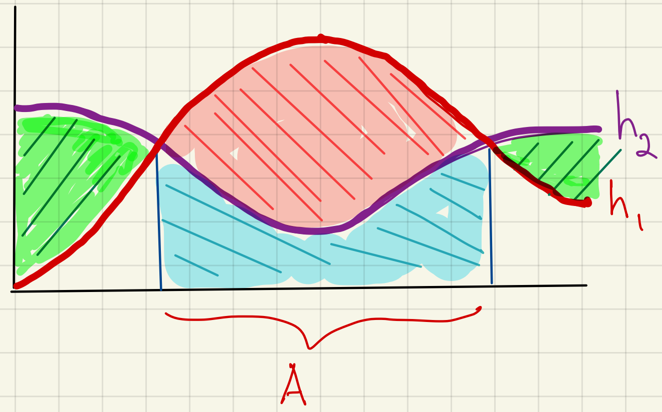
$a_2 = h_2(A)$ ← blue area

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

$\|h_1 - h_2\|_1$

$= \sum_{\substack{x \text{ s.t.} \\ h_1(x) > h_2(x)}} h_1(x) - h_2(x)$

$+ \sum_{\substack{x \text{ s.t.} \\ h_2(x) > h_1(x)}} h_2(x) - h_1(x)$



green area = red area $= a_1 - a_2$

L_1 dist = green + red $= 2 \cdot$ red $= 2(a_1 - a_2)$

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

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

3. $\alpha \leftarrow \frac{1}{m} |\{i \mid s_i \in A\}|$ ← estimates $\Pr_{x \in p}[x \in A]$

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"

$$C = \sum_x \min(h_1(x), h_2(x))$$

$$\text{green} = \sum h_2(x) - C = 1 - C$$

$$\text{red} = \sum h_1(x) - c = 1 - c$$

Why does it work?

- h_1 or h_2 is ε'-close to A (given)
- If "tie" in step 1:

  $h_1 + h_2$ are $\underbrace{10\varepsilon'}_{=2(a_1-a_2)}$-close to each other

  $\Rightarrow$ both $\le 11\varepsilon'$-close to p $\Rightarrow$ "tie" is ok output

- otherwise reach step 2:

  $$\|h_1 - h_2\|_1 > 10\varepsilon' \qquad (a_1 - a_2 > 5\varepsilon')$$

Algorithm Choose:

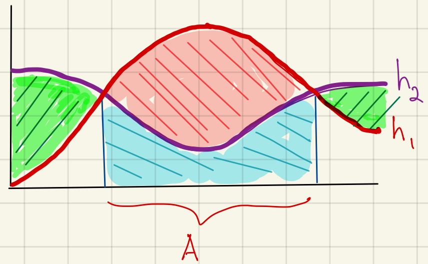
$$A = \{x \mid h_1(x) > h_2(x)\}$$
$$a_1 = h_1(A)$$
$$a_2 = h_2(A)$$
note $\|h_1 - h_2\|_1 = 2(a_1 - a_2)$

1. if $\underbrace{a_1 - a_2}_{\frac{1}{2} L_1 \text{ dist}} \le 5\varepsilon'$ declare "tie" + return h (no samples needed)
2. draw $m = \frac{2\log\frac{1}{\delta'}}{(\varepsilon')^2}$ samples $s_1 \ldots s_m$ from p
3. $\alpha \leftarrow \frac{1}{m} |\{i \mid s_i \in A\}|$
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

$$\begin{cases} \text{if } p = h_1, & E[\alpha] = a_1 \\ \text{if } p = h_2, & E[\alpha] = a_2 \end{cases}$$



green area = red area = $a_1 - a_2$

L_1-dist = green + red

blue area = a_2

blue + red area = a_1

Why does it work?

- h_1 or h_2 is ε'-close to A (given)

- If "tie" in step 1, algorithm does right thing

- Otherwise reach step 2: $\|h_1 - h_2\|_1 > 10\varepsilon'$ $(a_1 - a_2 > 5\varepsilon')$

$$E[\alpha] = \Pr_{x \in p}[x \in A] \equiv p(A)$$

assume (Chernoff) that with high prob $|\alpha - E[\alpha]| \le \frac{\varepsilon'}{2}$

h_1 assigns a_1 weight to A
h_2 " a_2 " " A

if p is ε'-close to h_1, assigns $\ge a_1 - \varepsilon'$ wt to A

$\quad$ + $\alpha \ge a_1 - \varepsilon' - \frac{\varepsilon'}{2}$ $\quad$ return h_1 whp

if p is ε'-close to h_2, assign $\le a_2 + \varepsilon'$ wt to A

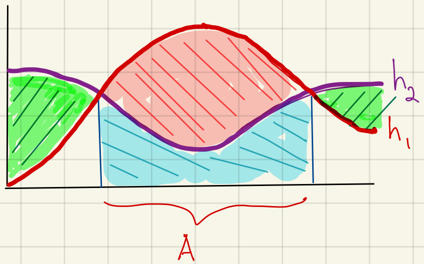
$\quad$ + $\alpha \le a_2 + \varepsilon' + \frac{\varepsilon'}{2}$ $\quad$ return h_2 whp

Algorithm Choose:

$A = \{x \mid h_1(x) > h_2(x)\}$
$a_1 = h_1(A)$
$a_2 = h_2(A)$
note $\|h_1 - h_2\|_1 = 2(a_1 - a_2)$

1. if $a_1 - a_2 \le 5\varepsilon'$ declare "tie" + return h (no samples needed)
2. draw $m = 2\frac{\log\frac{1}{\delta'}}{(\varepsilon')^2}$ samples $s_1 \dots s_m$ from p
3. $\alpha \leftarrow \frac{1}{m}|\{i \mid s_i \in A\}|$
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

$\begin{cases} \text{if } p = h_1, & E[\alpha] = a_1 \\ \text{if } p = h_2, & E[\alpha] = a_2 \end{cases}$



green area = red area = $a_1 - a_2$
L_1-dist = green + red
blue area = a_2
blue + red area = a_1

The cover method - a method for learning distributions

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

smaller set of distributions

big set of distributions

why useful? hopefully $\mathcal{C}$ much smaller than $\mathcal{D}$ & allows to approx $\mathcal{D}$

note $\mathcal{C}$ not unique

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

union bnd over $|\mathcal{C}|$ rather than $|\mathcal{D}|$

Thm $\exists$ algorithm, given $p \in \mathcal{D}$, which takes

$O\left(\frac{1}{\varepsilon^2}\log|\mathcal{C}|\right)$ samples of p + outputs $h \in \mathcal{C}$

s.t. $\|h-p\|_1 \leq 6\varepsilon$ with prob $\geq \frac{9}{10}$

Pf.

Since $p \in \mathcal{D}$, $\exists$ $q_{opt} \in \mathcal{C}$ s.t. $\|p - q_{opt}\|_1 \leq \varepsilon$

run "Choose" on p with every pair $q_1, q_2 \in \mathcal{C}$.

q_{opt}: win or tie all games

if q' is $\geq 6\varepsilon$-far from p, then $\geq 5\varepsilon$-far from q_{opt}

$\Rightarrow$ loses to q_{opt}

equivalently: if q' wins or ties all games

$\Rightarrow$ $\leq 5\varepsilon$ far from q_{opt}

$\leq 6\varepsilon$ far from p

need all matches to give correct output (estimate of α)

union bnd on $\binom{|\mathcal{C}|}{2}$ many matches ∎

Applications:

Example 1: learning distribution of a coin

domain $= \{0, 1\}$

need to learn bias

Here $\mathcal{D} = [0, 1]$ ← ~~b~~biases of coin

if use $\mathcal{C} = \{0, \frac{1}{K}, \frac{2}{K}, \frac{3}{K}, \ldots, \frac{K}{K}\}$

$\forall$ bias p let $\frac{i}{K} < p \leq \frac{i+1}{K}$

So pick $\tilde{p} \leftarrow$ closest $\frac{i}{K}$ then $|p - \tilde{p}|_1 = \frac{1}{K} \leq \varepsilon$

picking $k = \Theta(\frac{1}{\varepsilon})$

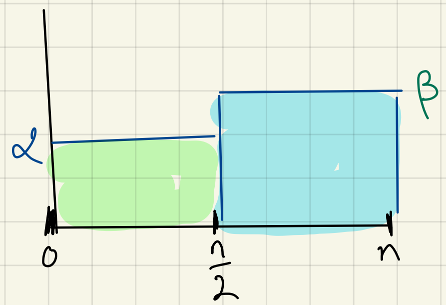
$|\mathcal{C}| = K+1 = \Theta(\frac{1}{\varepsilon})$

need $O\left(\frac{1}{\varepsilon^2} \log \frac{1}{\varepsilon}\right)$ samples

using cover method

Example 2: 2-bucket distributions

now need to specify α and β



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

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

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

Example 3: monotone distributions

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

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