A useful tool: Hypothesis Testing

Given collection of distributions $\mathcal{H}$, at least one has high accuracy for describing $p$—given via samples. Cut out 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 hypotheses

**Thm**

Given sample access to $p$

Given $h_1, h_2$ hypothesis distributions (fully known to algorithm)

Given accuracy parameter $\varepsilon'$, confidence $\delta'$

Algorithm "Choose" takes $O(\log(1/\delta')/(\varepsilon')^2)$ samples + outputs $h \in \mathcal{H}$.

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

Then with prob $\geq 1 - \delta'$, output $h_j$ has $||h_j - p|| < \varepsilon'$
Actually, will prove something stronger:

**Thm**

$p$ given via samples

$h_1, h_2$ fully known

$\varepsilon', \delta'$ 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

\[2e^{-m\varepsilon'^2/2} \iff \text{tie}\]

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

\[\geq 2e^{-m\varepsilon'/2} \iff \text{very bad}\]

\[\leq 2e^{-m\varepsilon'/2} \iff \text{not that bad}\]
Proof of "Subtool":

idea: if \( h_1 \) is \( \varepsilon' \)-close, show will output \( h_1 \), why
else if \( h_1 \) is \( 12\varepsilon' \)-far, show will not output \( h_1 \)
  will output \( h_2 \)
else (\( h_1 \) is not \( 12\varepsilon' \)-far, but not \( \varepsilon' \)-close, so \( h_2 \) must be \( \varepsilon' \)-close)
  we don't know what will happen,
  but either way we are golden (either \( h_1 \) or \( h_2 \) are that bad)

Algorithm Choose:

\[ \text{input } p, h_1, h_2 \]

First some definitions:

\[ A = \mathbb{R} \times \mathbb{R} \setminus h_1(x) > h_2(x) \]
\[ a_1 = h_1(A) \]
\[ a_2 = h_2(A) \]
note \( \| h_1 - h_2 \|_1 = 2(a_1 - a_2) \)

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

2. draw \( m = 2. \log \frac{4}{\delta} \) samples \( s_1 \ldots s_m \) from \( p \)

3. \( \alpha \leftarrow \frac{1}{m} \left| \{ s_i \in A \} \right| \)

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 reached step 2, w.h.p. (via Chernoff) \( |\alpha - E[\alpha]| \leq \frac{\varepsilon}{2} \)

If \( \|p - h_1\|_1 > 12\varepsilon' \) then since other is \( \leq \varepsilon' \) distance, or \( \|p - h_2\|_1 > 12\varepsilon' \) \( \|h_1 - h_2\|_1 > 12\varepsilon' \)

So will reach step 2.

If \( p \varepsilon' \)-close to \( h_1 \), w.h.p. \( \alpha > a_1 - \varepsilon' - \frac{3\varepsilon'}{2} \)

So output \( h_1 \)

Else, \( p \) is \( 12\varepsilon' \)-far from \( h_1 \) but \( \varepsilon' \)-close to \( h_2 \)

w.h.p. \( \alpha > a_2 + \varepsilon' + \frac{3\varepsilon'}{2} \)

If \( h_1 \) or \( h_2 \) \( \geq 10\varepsilon' \) far but not \( 12\varepsilon' \) far \( \Rightarrow \) return \( h_2 \) w.h.p.

if \( p_1 - p_2 \leq 5\varepsilon' \) then declares draw, so neither are declared "winner"

Else \( \|h_1 - h_2\|_1 > 9\varepsilon' \) far

Similar reasoning shows that medium far will not win (in fact, will lose).
The Cover Method

A method for learning distributions

\[ \mathcal{C} \text{ is a } \delta \text{-cover of } \mathcal{D} \text{ if } \forall p \in \mathcal{D} \]
\[ \exists q \in \mathcal{C} \text{ s.t. } \| p - q \|_1 \leq \delta \]

Why useful?

- Hopefully \( \mathcal{C} \) is much smaller than \( \mathcal{D} \) - allows us to approximate
- Note \( \mathcal{C} \) not unique

Then, \( \exists \) algorithm, given \( p \in \mathcal{D} \), which takes
\[ O(\frac{1}{\delta^2} \log |\mathcal{C}|) \] samples of \( p \) outputs \( h \in \mathcal{C} \)
\[ \text{s.t. } \| h - p \|_1 \leq 6\delta \text{ with prob } \geq 9/10 \]

Proof:

- Since \( p \in \mathcal{D} \), \( \exists q \in \mathcal{C} \) s.t. \( \| p - q \|_1 \leq \delta \)
  (but there could be more than 1)

will run Choose on \( p \) with every pair \( q_1, q_2 \in \mathcal{C} \)
if \( q \) doesn't win all of its "matches," then it loses to someone that is not so bad

Furthermore, can show that when there is a \( q' \) s.t.
\[ q' \text{ wins or ties all matches.} \]
The cover method

Example 1: learning distribution of a coin

domain = \{0, 1\}^3
need to learn bias
Here \(\mathcal{D} = R\)
if use \(C = \{0, \frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}, 1\}\)
then A bias \(p\), let \(\frac{A}{K} \leq p \leq \frac{A+1}{K}\)
then picking \(\hat{p} = \frac{A}{K}\) gives \(\|p - \hat{p}\|_1 = \left|\frac{A}{K} - p\right| + \frac{1}{K} \leq \frac{2}{K}\)

so using \(k = \Theta\left(\frac{1}{\epsilon}\right)\) gives \(\|p - \hat{p}\|_1 \leq \epsilon\)
\(|\mathcal{E}| = k+1 = \Theta\left(\frac{1}{\epsilon}\right)\)
# samples needed by cover method is \(O\left(\frac{1}{\epsilon^2} \cdot \log \frac{1}{\epsilon}\right)\)

Example 2: 2-bucket distributions

now need to specify \(\alpha\) and \(\beta\)
so \(C = \{0, \frac{1}{K}, \frac{2}{K}\} \cup \{0, \frac{1}{K} \ldots, \frac{K-1}{K}\}\)
\(|\mathcal{E}| = \Theta\left(\frac{1}{\epsilon^2}\right)\)
# samples is \(O\left(\frac{1}{\epsilon^2} \cdot \log \frac{1}{\epsilon}\right)\)

Example 3: monotone distributions

Birge \(\Rightarrow C = \{0, \frac{1}{K}, \frac{2}{K}, \ldots, \frac{\log n}{K}\} \cup \{0, \frac{1}{K} \ldots, \frac{K-1}{K}\}\)
\(|\mathcal{E}| = \Theta\left(\frac{1}{\epsilon (\log n)^2}\right) \Rightarrow \# \text{ samples is } O\left(\frac{1}{\epsilon^3} \log n \cdot \log \frac{1}{\epsilon}\right)\)