Last time:
- Finish learning $C = \text{intervals of } \mathbb{R}$
  $$n = \frac{3}{\varepsilon} \ln \left( \frac{1}{\delta} \right) \delta^{-\frac{1}{2}} \varepsilon^{-\delta}$$
- $OLMB \rightarrow \text{PAC conversion}$ (not all our $OLMB$ results in PAC model "for free")
- Revisit PAC learning def:
  - size of concepts
  - efficiency of eval. $h$
- Start "Chernoff bounds" for hypothesis testing

Today:
- Learning by finding consistent hypothesis from a fixed $H$
- Occam's Razor
  - General version
  - "Cardinality version"

Questions?

Motiv. CB's:
- have hyp $h$, in PAC world
- have $EX(c, \mathcal{D})$

is my $h$ good or not?

Obvious approach:
- draw $m$ ex from $EX(c, \mathcal{D})$
- eval. $h$ on each
- count frac $\hat{E} = \frac{\# \text{ of ex. where } h(x) \neq c(x)}{m}$

How good is $\hat{E}$ as estimate of $\epsilon_{\mathcal{D}}(h, c)$?
Equiv. to: \( \Pr[C \neq H] = p \quad 0 < p < 1 \).

Goal: estim. \( p \).

Estim. ( \( n \) tosses) = \( \hat{p} = \frac{\#H}{n} \). How good?

"Chernoff bounds".
"Multiplicative" version:

Let \( X_1, \ldots, X_m \) be independent \( 0/1 \) random variables, with \( \Pr[X_i = 1] = p \) for all \( i=1,\ldots,m \).

Let \( X = X_1 + \ldots + X_m \).

Then for \( 0 < \delta \leq 1 \),

\[
\Pr[X < (1-\delta)mp] \leq \exp\left(-\frac{1}{2} \delta^2 mp\right),
\]

\[
\Pr[X > (1+\delta)mp] \leq \exp\left(-\frac{1}{3} \delta^3 mp\right).
\]

Ex: Baseball team, \( n \)-game season, wins each game indep. \( \Pr[W] = \frac{1}{2} \).

\( \Pr[ \text{team finishes with } (.450 \text{ or worse record}) ] \)
$Pr\left[ X < (1 - y) \frac{\ln n}{\ln 0.9} \right] \leq e^{\frac{-1}{2} \frac{y^2}{\ln 0.9}}$.

\( r = 0.1 \)

"Additive" version:

\[ X, m, p, X_1, \ldots, X_m \text{ as above.} \]

Let \( \bar{p} = \frac{X}{m} \).

Then

\[ Pr\left[ p - \bar{p} \geq \varepsilon \right] \leq \exp(-2m\varepsilon^2) \]

\[ + \]

\[ Pr\left[ \hat{p} - p \geq \varepsilon \right] \leq \exp(-2m\varepsilon^2) \]

Ex: Doing survey. Is pineapple pizza \( \checkmark \) or \( \text{xck} \)?

How many people do you need to survey s.t. w. 95% confidence (w.p. \( > 0.95 \)) your figure is within \( \pm 3\% \) of true %?

Want to solve for \( n \): \( \varepsilon = 0.03 \)

\[ Pr\left[ p - \bar{p} \geq \varepsilon \right] \leq \exp(-2m\varepsilon^2) \]  \( \Rightarrow \)

\[ e^{-2m \cdot 0.0009} \leq \frac{1}{40} \]

\( e \leq 0.0025n \)

\[ e^{-0.0025n} \leq \frac{1}{40} \Rightarrow n = \frac{1}{0.0025} \ln(40) \]
Note: strong assumption: r.v. $X = \text{sum of indep. RV's } X = X_1 + \ldots + X_m$.

Here's a weaker but more general bound:

"Markov's inequality": Let $X$ be a nonneg. R.V. Then for any $k \geq 1$, 
$$\Pr\{X \geq k \cdot \mathbb{E}[X]\} \leq \frac{1}{k}.$$

Ex: Let $X = \# \text{ children in randomly selected US household. } \mathbb{E}[X] = 1.85$. 
$$\Pr\{X > 10\} \leq 20\%, \text{ bc } 20\% \text{ of } 10 = 2 > \frac{1}{1.85}$$

---

Learning by Finding a Consistent Hypothesis

Here's the general approach to PAC learning:
- Find a consistent hypothesis from some a priori fixed hypothesis class.

---

Theorem: Fix a c.c. $C$, hyp. class $\mathcal{H}$.
- Fix any target concept $c \in C$.
- Fix any dist $\mathcal{D}$ over $X$.
- Let $(x^i, c(x^i)), \ldots, (x^n, c(x^n))$ be drawn i.i.d. from $\mathbb{E}_X(c, \mathcal{D})$ where
\[ m \geq \frac{1}{\varepsilon} \left( \ln |\mathcal{H}| + \ln \frac{1}{\delta} \right). \]

Say \( h \in \mathcal{H} \) is bad if \( \operatorname{err}_g(h, c) > \varepsilon \). Then

\[ \Pr \left[ \text{any bad } h \in \mathcal{H} \text{ is consistent with all in lab. ex.} \right] \leq \delta. \]

**Proof:** Fix any bad \( h \).

\[ \Pr \left[ h \text{ consistent with all in lab. ex.} \right] < (1-\varepsilon) \]

(\( b/c \) \( \Pr \left[ h \text{ right on } \mathcal{S}(x, c(x)) \sim \mathcal{E}(c, \mathcal{D}) \right] < 1-\varepsilon \),

+ independence)

\( \mathcal{H} \) has \( |\mathcal{H}| \) bad hyp's. So by union bd on

\[ \Pr \left[ \text{any bad } h \in \mathcal{H} \text{ is consistent with all in lab. ex.} \right] \leq |\mathcal{H}| \cdot (1-\varepsilon). \]

Have

\[ \frac{|\mathcal{H}|}{(1-\varepsilon)} \leq |\mathcal{H}| \cdot (1-\varepsilon)^{\frac{1}{\varepsilon} \left( \ln (|\mathcal{H}|) + \ln \left( \frac{1}{\delta} \right) \right)} \]

\[ \leq |\mathcal{H}| \cdot e^{-\left( \ln (|\mathcal{H}|) + \ln \left( \frac{1}{\delta} \right) \right)} \]

\[ = |\mathcal{H}| \cdot \frac{1}{|\mathcal{H}|} \cdot \delta = \delta. \]

Interpreting/using the thm:

**Def** Fix \( C \), finite \( \mathcal{H} \). A **consistent hypothesis finder** for \( C \) using \( \mathcal{H} \) is an alg \( B \) with following property:

if \( B \) given sample \( S \) of \( m \) examples \((x', c(x'))\),....
$(x^i, c(x^i))$ where $c \in C$, then $B$ outputs $h \in \mathcal{H}$ that's consistent with $S$.

$$\langle x^i, c(x^i) \rangle, \ldots, \langle x^n, c(x^n) \rangle \xrightarrow{CHF} B \xrightarrow{h_i} h(x^i) = c(x^i) \quad i = 1, \ldots, n.$$ 

prev thesis says: if $B$ is CHF for $C$ using $\mathcal{H}$, running $B$ on $\mathcal{CHF}$

$$m = \frac{1}{\varepsilon} \left( \ln |\mathcal{H}| + \ln \frac{1}{\delta} \right)$$

many draws from $EX(\varepsilon, \delta)$ is an $(\varepsilon, \delta)$-PAC learning alg.

Application: $C =$ mon. disj. over $\{0,1\}$

$\mathcal{H} =$ "" "" "" comp. efficient!"

Our online elimination alg is a CHF for $C$ using $\mathcal{H}$.

So, by thesis, can take

$$m = \frac{1}{\varepsilon} \left( \ln (2^n) + \ln \frac{1}{\delta} \right)$$

ex, and this is PAC.

better s.c. than before, by $\sim \log(n)$ factor.

Note: big "if" here: need a CHF.

Can be (often is) the case that finding a consistent $h \in \mathcal{H}$ is computationally hard.
Happily, for mon. disj. efficient.

Note: Crucial that we had "a priori fixed" $\mathcal{H}$. Above does not mean building a lookup table is good PAC learning alg! $\not\in$ not a fixed $\mathcal{H}$.

\begin{quote}
\textbf{Occam's Razor}
"Entities should not be multiplied unnecessarily"  
- William of Occam
Should prefer short explanations to long ones.
\end{quote}

Can view above theorem:
$h$ is a "short explanation" of data.

$m$ lab. ex.: $m$ bits of data to "explain" (the labels).
any $h \in \mathcal{H}$ can be encoded with a $\# \in \{1, \ldots, |\mathcal{H}|\}$, i.e. with $\log |\mathcal{H}|$ bits.

If $m \geq \frac{1}{\varepsilon} \left( \ln |\mathcal{H}| + \ln \frac{1}{\delta} \right)$, then

indeed $h$ is a "short explanation" of the data.

Here's a "refined" ("cardinality") version of
Occam's Razor:

\[ \text{CHF} \]

Our thesis: For any \( m \), it gives a consistent \( h \in \mathcal{H} \).

Sometimes following may hold: if you give CHF a small \# \( m \) of examples, it can find a consistent \( h \) in subregion \( \mathcal{H}_m \subseteq \mathcal{H} \).

Here's another version of our CHF thesis:

"Occam's razor, cardinality version": Fix \( c \), fix finite \( \mathcal{H} \). Fix \( m, \varepsilon, \delta \).

Suppose there is a set \( \mathcal{H}_m \subseteq \mathcal{H} \) of hypotheses where

\[ m > \frac{1}{\varepsilon} \left( \ln |\mathcal{H}_m| + \ln \frac{1}{\delta} \right) \]

such that given any sample \( S \) of \( m \) ex. lab. by some \( c \in \mathcal{C} \), there is an \( h \in \mathcal{H}_m \) consistent \( w/ S \).

Let \( l \) be an \( \mathcal{A} \), which, given such an \( S \) of size \( m \), outputs an \( h \in \mathcal{H}_m \) consistent \( w/ S \).

Then running \( l \) on \( m \) ex. from \( \mathbb{E}X(c, \mathcal{D}) \) is an \((\varepsilon, \delta)\)-PAC learning alg.

Pf: as before (\( \mathcal{H}_m \) for \( \mathcal{H} \) everywhere).
Next time: use this.