Last time:
- VCDIM(C) as lower bound on best poss. mistake bound of any OCMB alg. for C
- Predicting from Expert Advice
  - Weighted Majority (WM) alg

rebranded noise-tolerant HA

Today:
- Randomized Weighted Majority alg (RWM)
- Start next unit:

Probably Approximately Correct (PAC) Learning Model

  - motivation
  - basic definition
  - learning C = intervals over $X = \mathbb{R}$

Questions?

Recall WM: suppose $q_0 = 51$, $q_1 = 49$

WM will definitely predict 0.

"Predictable... needlessly!

Predicting q_1 would also have "ensured progress" if mistake...

Plausible Variant: in situation, toss a biased coin...

Randomized WM Alg: ($0 \le \beta < 1$) again, N experts.

- Initialize $w_i = 1 \quad \forall i \in [N]$.
- At each trial, expert i predicts z_i (need not be 0/1 value!)
- RWM: outputs each z_i w.p. w_i / W,
  $W = w_1 + \dots + w_n$.
- Given correct outcome ℓ, for each i s.t. $z_i \neq \ell$,
  set $w_i \leftarrow w_i \cdot \beta$.

Thm: Assume seq. of outcomes fixed in advance.
(like "obliv. adv." assump. for RHA).
If best expert in pool makes m mistakes, then
$\mathbb{E}[\#\text{mistakes RWM makes}]$ is

$$M \le \frac{m \cdot \ln(1/\beta) + \ln N}{1 - \beta}.$$

Ex: $\beta = \frac{1}{2} \to 1.39m + 2\ln N$

$\beta = 3/4 \to 1.15m + 4\ln N$

$\beta = 1-\varepsilon,\ \varepsilon \to 0$: bd $\to m + \frac{1}{\varepsilon} \cdot \ln N$.

Pf: Consider any fixed seq. of T trials.
Let

F_i = frac. of tot. wt. at trial i that's on wrong pred. for trial i.

$\Pr[\text{RWM makes mist. on trial } i] = F_i$.

So $M = \mathbb{E}[\#\text{mistakes RWM makes}] = \sum_{i=1}^{T} F_i$.

Before i^{th} ex: tot wt W is

$W = \overbrace{(1-F_i)W}^{\text{wt on right pred}} + \overbrace{F_i W}^{\text{wrong}}$

After i^{th} ex: new W is $(1-F_i)W + \beta F_i W$
$= W \cdot (1-(1-\beta)F_i)$.

So final W is $N \cdot \prod_{i=1}^{T} (1-(1-\beta)F_i)$.

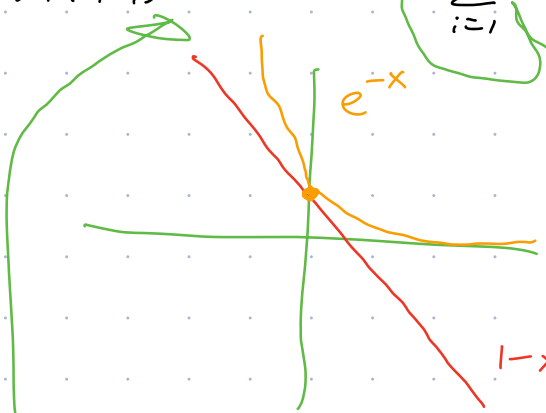
As before, W is also $\geq \beta^m$ (from best expert).

So $\beta^m \leq W = N \prod_{i=1}^{T} (1-(1-\beta)F_i$ ∎

$\ln$: $m \ln \beta \leq \ln N + \sum_{i=1}^{T} \ln(1-(1-\beta)F_i)$.

$-$: $m \ln \frac{1}{\beta} \geq -\ln N - \sum_{i=1}^{T} \ln(1-(1-\beta)F_i)$

Recall:



$1-x \leq e^{-x}$ $\forall x$

so

$\ln(1-x) \leq -x$

so

$x \leq -\ln(1-x)$

Use this: ($x = (1-\beta)F_i$ each elt of sum):

$m \ln \frac{1}{\beta} \geq -\ln N + \sum_{i=1}^{T} (1-\beta)F_i$, so

---

$$\frac{m \ln \frac{1}{\beta} + \ln N}{1-\beta} \geq \sum_{i=1}^{T} F_i = M.$$

PAC Learning

(new model!)

Drawbacks of OLMB:

- assumes any ex seq. possible... adversarial!
- MB crit. measures perf. from very beginning: no "training" period.
- MB guarantee can be unsatisfying: if haven't reached the alg's MB, can't say anything abt being right on next example.

New model: independently identically distributed

- assume all ex $x \in X$ received for learning are i.i.d. drawn from some fixed, unknown, arbitrary distrib. $\mathcal{D}$ over X.

Each ex: $(x, c(x))$ $x \sim \mathcal{D}$.

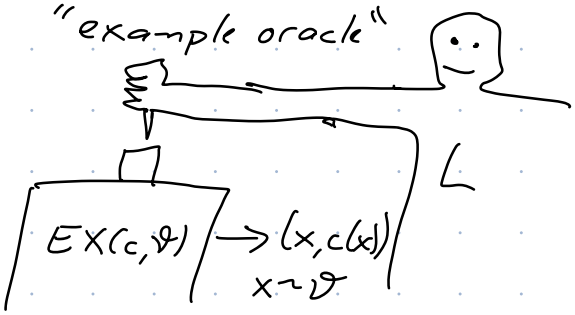
- "batch" model: get data set of, train on that. Alg. outputs just one hyp h based on

- Performance guarantee we want:

w/ high prob., hyp h should do well on future ex $(x, c(x))$, $x \sim \mathcal{D}$. same distrib.

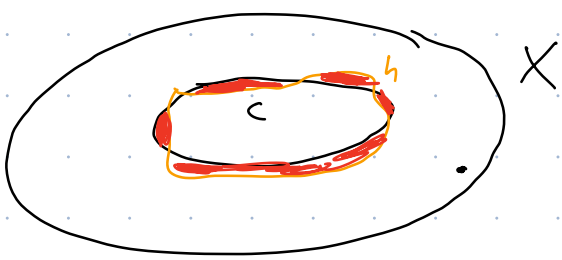
PAC framework for learning concept class $\mathcal{C}$ using a hypoth. class $\mathcal{H}$:

- "Knows" $\mathcal{C}$, doesn't know target concept. $c \in \mathcal{C}$
- Learner can access "example oracle" $EX(c, \mathcal{D})$



- Learner does computation, + outputs some $h \in \mathcal{H}$.

Def: Let $h, c : X \to \{0,1\}$, let $\mathcal{D}$ be a dist. over X.
Then
$$err_{\mathcal{D}}(h,c) := \Pr_{x \sim \mathcal{D}}[h(x) \neq c(x)]$$ is "error of h on c under $\mathcal{D}$"



$\Pr_{\mathcal{D}}[\text{shaded}] = \text{error}$

- Shouldn't expect to achieve $err_{\mathcal{D}}(h,c) = 0$: $\mathcal{D}$ may put small but nonzero wt on parts of X...

- Shouldn't expect to achieve small $er_{\mathcal{D}}(h,c)$ with probability 1: could have atypical training set of examples, & then may have high error.

Crit. for success: whp, achieve low error hyp.

Def: (Prelim. Def. of PAC Learning)

"Alg. A PAC learns $\mathcal{C}$ using $\mathcal{H}$" means:

- for any $c \in \mathcal{C}$ (target concept),
- for any dist. $\mathcal{D}$ over X (distr.),
- for any input param's $0 < \varepsilon, \delta < 1$ (error / accuracy; failure prob / confidence param)

if A is given ε, δ, & access to $EX(c, \mathcal{D})$, A outputs some $h \in \mathcal{H}$ s.t. with prob. $\geq 1-\delta$ (over draws from $EX(c,\mathcal{D})$ & any internal randomness of A), have

$$er_{\mathcal{D}}(h,c) \leq \varepsilon.$$

Usually write $m = m(\varepsilon, \delta)$ for # examples; "sample complexity" of A.

We say A is "efficient" if runtime of A is $\text{poly}(\frac{1}{\varepsilon}, \frac{1}{\delta})$.

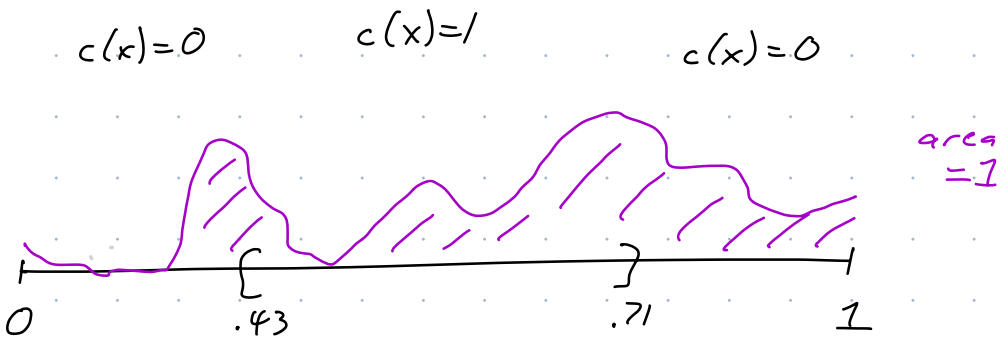
usually $X = \{0,1\}^n$ or $\mathbb{R}^n$.

$\text{poly}(n, \frac{1}{\varepsilon}, \frac{1}{\delta})$.

If $\mathcal{H} = \mathcal{C}$, we say A is "proper"

Ex: PAC learning intervals

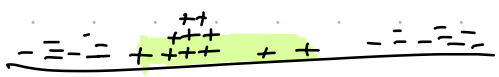
$X = [0,1]$ $\quad \mathcal{C}$ = all intervals $[a,b]$



Say target c is $[a,b]$.

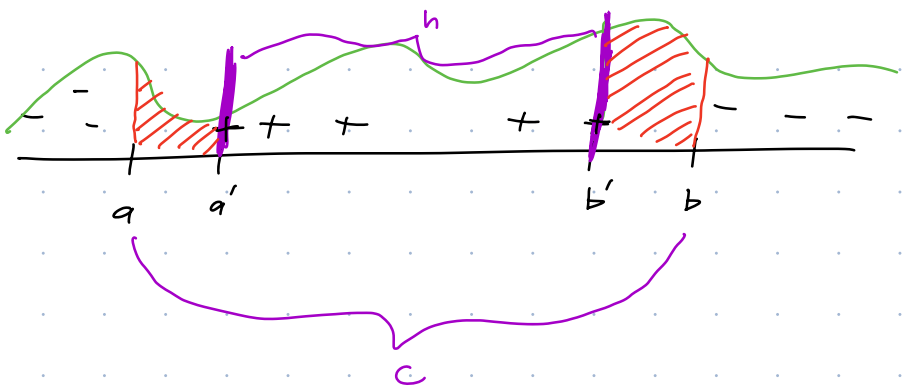
Here's an alg A:

- draw m examples.



- Let a' = leftmost pos ex in m samples.
  b' = rightmost pos ex

Output $h = [a', b']$.



$er_g(h,c)$ = area of 2 shaded regions.

Next time: analyze this, show that for $m = O(\frac{1}{\varepsilon} \cdot \ln \frac{1}{\delta})$, w. prob. $1-\delta$, $er_g(h,c) \leq \varepsilon$.