Last time:
- analysis of AdaBoost: bound on error rate of its final hyp. over its input sample
- start unit on PAC learning in presence of <u>noise</u>
  - General framework

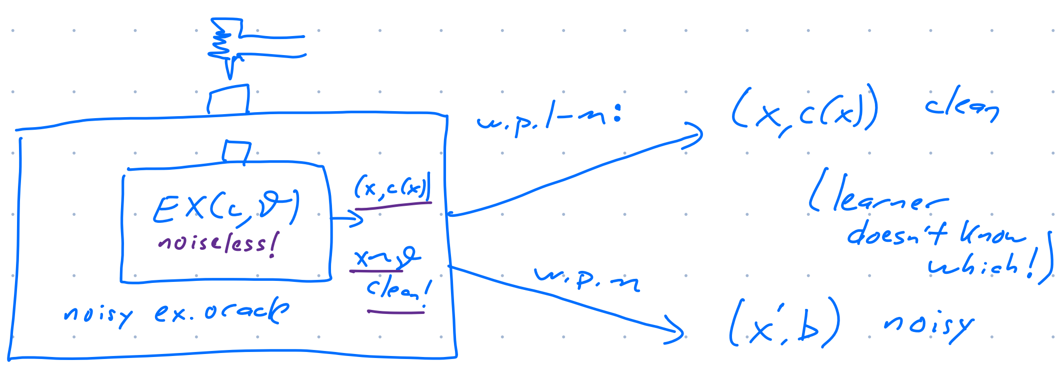
Today: Specific noise models: toy scenario
- malicious noise (hard!) ☹
- random classification noise (easy!) ☺ → RCN

- $\underbrace{\text{Learning with malicious noise}}$ is hard ☹
- Sad, sad positive result for
- Start discussion of learning with RCN
  ↳ towards new learning model: learning from STATISTICAL QUERIES

Admin: PS 4 due today, PS 5 out.

Questions?

Recall: We assume learner has access to a <u>noisy example oracle</u>: noise rate $0 \le \eta < \frac{1}{2}$



Both <u>malicious noise</u> + <u>random classification noise</u> fall in this framework:

Mal. noise: (x', b) can be arbitrary elt. of $\underline{\underline{X}} \times \underline{\underline{\{0,1\}}}$.
(no assump.; can be gen. by omniscient malevolent adversary.)

Rand. Class. Noise: get $(x, \overline{c(x)})$ (label flipped; x is not altered.)

Hi-level msg of noise unit:

- Mal. noise: challenging; • if there's mal. noise at noise rate $\eta \approx 2\varepsilon$, can't learn to acc. ε.
  - best general methods to deal w/ mal. noise: not great.

- Can usually handle RCN: we'll see a general way to convert many PAC learners to PAC learners handling RCN at rate $\eta < \frac{1}{2}$, with slowdown of only poly$\left(\frac{1}{1-2\eta}\right)$.
  ($\eta = \frac{1}{2}$: imposs. to learn.)

Simple ex (cartoon):
Learning setting w/ no examples (!), just labels.
(Noise-free data)

World 1:	World 2:
$\frac{3}{8}$ of ex +	$\frac{5}{8}$ of ex +
$\frac{5}{8}$ of ex −	$\frac{3}{8}$ " " −

With noise...?

- Mal. noise at rate $\eta = 1/5$:
  World 1: adv. can corrupt only $-$ ex. to $+$, $+$ observed data for learner:
  $3/8$ $+$ (true $+$)
  and also $\frac{1}{5} \cdot \frac{5}{8}$ (true $-$) see $+$
  See $\frac{1}{2}$ $+$, $\frac{1}{2}$ $-$ data.

  World 2: adv. can corr. only $+$ ex to $-$, $+$ again learner sees $\frac{1}{2}$ $+$, $\frac{1}{2}$ $-$ data.

  So learner can't tell which world it is.

- RCN at rate $\eta < \frac{1}{2}$:
  - world 1: learner sees $+$ with prob.

$$\underbrace{\tfrac{3}{8}}_{\text{true }+} \cdot \underbrace{(1-\eta)}_{\text{no noise}} + \underbrace{\tfrac{5}{8}}_{\text{true }-} \cdot \underbrace{\eta}_{\text{noise}} = \frac{3}{8} + \frac{1}{4}\eta$$

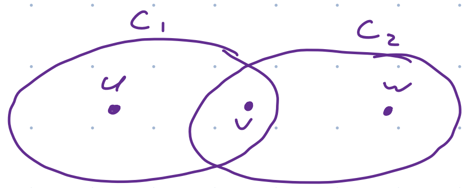
  - world 2: learner see $-$ with prob. $\frac{3}{8} + \frac{1}{4}\eta$

If $\eta < \frac{1}{2}$, this $< \frac{1}{2}$.

So with enough ex., can tell if in world 1 or 2.

Lower Bound for Learning w/ Mal. Noise

Def: $\mathcal{C}$ is distinct if $\exists\, c_1, c_2 \in \mathcal{C}$ $\ni$ 3 pts $u, v, w \in X$ s.t.



$c_1(u) = c_1(v) = 1$
$c_2(v) = c_2(w) = 1$
$c_1(w) = 0$
$c_2(u) = 0$

Thm: Let $\mathcal{C}$ be distinct. If learning w/ mal. noise at rate $\eta \geq \frac{2\varepsilon}{1+2\varepsilon}$, then no alg. can PAC learn to error $< \varepsilon$ & conf. $1-\delta > \frac{1}{2}$.

Pf: Use c_1, c_2 above.
Use dist. $\mathcal{D}$:

$$\mathcal{D}(u) = \varepsilon$$
$$\mathcal{D}(v) = 1 - 2\varepsilon$$
$$\mathcal{D}(w) = \varepsilon$$

Define adv. A_1 for target c_1,
" " A_2 " " c_2 ;

target c_1 under A_1 $\overset{\text{indist.}}{\equiv}$ target c_2 under A_2.

Adv. A_1: w.p. $\frac{\eta}{2}$, output $(u, -)$,
" " " $(w, +)$.

So : learner observes dist. of lab. ex. :

$(u,+)$	prob.	$\tau(1-\eta)$	$\}$ noiseless
$(v,+)$	"	$(1-2\tau)(1-\eta)$	
$(w,-)$	"	$\tau(1-\eta)$	
$(u,-)$	"	$\eta/2$	$\}$ noisy
$(w,+)$		$\eta/2$	

Adv. A_2: w.p. $\frac{\eta}{2}$, output $(u,+)$,
" " " $(w,-)$.

So under c_2+A_2: learner observes dist. of lab. ex.:

$(u,-)$	prob.	$\tau(1-\eta)$	$\}$ noiseless
$(v,+)$	"	$(1-2\tau)(1-\eta)$	
$(w,+)$	"	$\tau(1-\eta)$	
$(u,+)$	"	$\eta/2$	$\}$ noisy
$(w,-)$	"	$\eta/2$	

If $\tau(1-\eta)=\frac{\eta}{2}$: <u>dist's identical.</u>

$$2\tau(1-\eta)=\eta$$

$$2\tau=\eta(1+2\tau), \quad \underline{\eta=\frac{2\tau}{1+2\tau}}.$$

So can't achieve conf. $>\frac{1}{2}$ of having error $<\tau$.

---

above: bad news.

"Good" news:

How can we handle mal. noise in general?

Say have alg A that learns (PAC) if (efficiently) <u>no</u> noise.

Do the foll:

- Run A. Hope no noisy example occurred.
- Repeat "many" times.

Get $h_1, h_2, \ldots, h_T$.

- Check $h_1, \ldots, h_T$ on fresh ex; output the one that did best.

Intuition: Say $m = \#$ex. your noiseless PAC learner A needs. ($\to$ poly(n))

- If $\eta \ll \frac{1}{m}$: ($\eta \ll \varepsilon$) should work! • some h_i should be good, b/c no noise; • when we eval. how good $h_1, h_2, \ldots h_T$ are, since η tiny, adv. can't mislead too much.

Turns out: $\eta = \min\left\{\varepsilon/8, \frac{\ln m}{m}\right\}$ is "sweet spot"

if $\eta = \frac{\ln m}{m}$:

$$\left((1-x) \approx e^{-x} \quad x \text{ small}\right)$$

Pr[m ex. drawn for A are all noiseless]

$$= \left(1 - \frac{\ln m}{m}\right)^m \approx \left(e^{-\frac{\ln m}{m}}\right)^m = \frac{1}{m}$$

For $\eta = \frac{\ln m}{m}$,

Take $T = m \cdot \ln(2/\delta)$, +

$\Pr[\text{none of } T \text{ runs have noiseless data set}]$

$$= \left(1-\frac{1}{m}\right)^T = \left(1-\frac{1}{m}\right)^{m \ln 2/\delta} \approx \frac{\delta}{2}.$$

Random Classif. Noise (RCN)

Recall: rate η: w.p. • $1-\eta$, get $(x, c(x))$ $x \sim \mathcal{D}$
• η, get $(x, \overline{c(x)})$ $x \sim \mathcal{D}$.

Call this oracle $\boxed{EX^{\eta}(c, \mathcal{D})}$

Def: Alg. A PAC learns $\mathcal{C}$ with RCN if

- $\forall c \in \mathcal{C}$,
- $\forall$ dist $\mathcal{D}$,
- $\forall \eta < \frac{1}{2}$,
- $\forall \varepsilon, \delta > 0$,

Efficient: $\text{poly}\left(\frac{1}{\varepsilon}, \log\frac{1}{\delta}, n, \text{size}(c), \frac{1}{1-2\eta}\right)$

given $\eta, \varepsilon, \delta$ & access to $EX^{\eta}(c, \mathcal{D})$, A outputs an h s.t.

w.p. $\geq 1-\delta$, $\text{err}_{\mathcal{D}}(h, c) \leq \varepsilon$.

- Note 1: same goal: high acc. on noiseless example.

- Note 2: Sps true $\text{err}_{\mathcal{D}}(h, c) = \tau$. Then given (x, y) pairs $\sim EX^{\eta}(c, \mathcal{D})$, observed error rate will be

$$\Pr_{(x,y)\sim EX^{\eta}(c,\mathcal{D})}\{h(x)\neq y\} = \underbrace{\overset{\text{true error}}{\tau}}_{h\neq c}\overset{\Pr[y=c]}{(1-\eta)} + \overset{\Pr[h=c]}{(1-\tau)}\overset{\Pr[y\neq c]}{\eta}$$

$$= \tau + \eta - 2\eta\tau$$

$$= \underbrace{\eta + \tau(1-2\eta)}_{\text{observed error}}.$$

observed error $\eta + (1-2\eta)\cdot\left(\text{true error}\right)$

Next time: old alg $\rightsquigarrow$ make noise tol.

$\hookrightarrow$ STATISTICAL QUERY learning model.