Last time:

- analysis of 3-stage boosting, justification of why it gives 35.2%-error h wrt $\mathcal{D}$
- discuss ext. to "full" booster ($1-\varepsilon$ accuracy)
- " boosting over a fixed sample
- AdaBoost: simple, practical booster over a fixed sample; described alg, main ideas
  - reweight $\mathcal{D}_t \rightsquigarrow \mathcal{D}_{t+1}$ s.t. h_t is 50% acc. under $\mathcal{D}_{t+1}$
  - clever weighted MAJ vote of $h_1, \dots, h_T$ is final h.

Today:

- analysis of AdaBoost $\rightsquigarrow$ state & prove thm about its performance
- start unit on PAC learning in the presence of noise
  - general framework, partic. noise models.

Questions?

Let's consider reweighting rule: $\left(\alpha_t = \frac{1}{2} \ln \left(\frac{1-\varepsilon_t}{\varepsilon_t} \right),\ h_t, y_i \pm 1 \text{ valued} \right)$

$$\mathcal{D}_{t+1}(i) = \frac{\mathcal{D}_t(i) \cdot \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

- If $h_t(x_i) = y_i$ (right):

$$\exp(-\alpha_t \underbrace{y_i h_t(x_i)}_{=1}) = \exp(-\alpha_t)$$

$$= \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}}.$$

Tot wt on i s.t. $h_t(x_i) = y_i$ is $1-\varepsilon_t$:

mult by $\exp(-\alpha_t y_i h_t(x_i))$

$$\rightsquigarrow (1-\varepsilon_t)\sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}} = \sqrt{\varepsilon_t(1-\varepsilon_t)}$$

- If $h_t(x_i) \neq y_i$ (wrong):

$$\exp(-\alpha_t \overbrace{y_i h_t(x_i)}^{=-1}) = \exp(\alpha_t)$$
$$= \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}.$$

Tot wt on i s.t. $h_t(x_i) \neq y_i$ is ε_t :

mult by $\exp(-\alpha_t y_i h_t(x_i))$

$$\rightsquigarrow (\varepsilon_t)\sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}} = \sqrt{\varepsilon_t(1-\varepsilon_t)}$$

Thm about AdaBoost:

Suppose AdaBoost is run for T stages.
Its final hyp $H : \{x_1, \ldots, x_m\} \to \{+1, -1\}$ makes errors on $\leq$ a

$$\prod_{t=1}^{T} \sqrt{1-4\gamma_t^2} \leq \exp\left(-2\sum_{t=1}^{T} \gamma_t^2\right)$$

frac of m pts in $\{x_1, \ldots, x_m\}$. $\quad (\gamma_t = \frac{1}{2} - \varepsilon_t)$.

Cor: If each $\gamma_t \geq \gamma$, can run AdaBoost for

$T = \frac{1}{2\gamma^2} \ln\left(\frac{1}{\varepsilon}\right)$ stages, & [illegible] $\leq \varepsilon$.

($\varepsilon < \frac{1}{m} \rightarrow$ zero error on ex $x_1, \ldots, x_m$)

---

optimal!

Pf: From 3 lemmas. $\left(H(x) = \text{sign}(f(x)), \quad f(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \right)$

(L1): $\frac{1}{m} \left| \{ i \in [m] : H(x_i) \neq y_i \} \right| \leq \frac{1}{m} \cdot \sum_{i=1}^{m} \exp(-y_i f(x_i))$

(L2): $\frac{1}{m} \cdot \sum_{i=1}^{m} \exp(-y_i f(x_i)) \leq \prod_{t=1}^{T} Z_t$

(L3): $\prod_{t=1}^{T} Z_t \leq \prod_{t=1}^{T} \sqrt{1 - 4\gamma_t^2}$

Pf of (L1):

Suppose i is s.t. $H_t(x_i) \neq y_i$.

$\text{sign}(f(x_i)) \neq y_i$, so $f(x_i) \cdot y < 0$.

so

$-y_i f(x_i) > 0$, so $\exp(-y_i f(x_i)) > 1$.

And if i is s.t. $H_t(x_i) = y_i$, get 0 on LHS of $\quad\quad$ + >0 on RHS of

So $\forall i$, LHS contrib. from i $<$ RHS contrib from i. $\blacksquare$

Pf of L2: Recall def of $\mathcal{D}_{T+1}(i)$:

$$\mathcal{D}_{T+1}(i) = \mathcal{D}_{T+1}(i) = \frac{\mathcal{D}_T(i)\cdot\exp(-\alpha_T y_i h_T(x_i))}{Z_T}$$

$$= \frac{\exp(-\alpha_T y_i h_T(x_i))}{Z_T}\cdot \mathcal{D}_T(i)$$

$$= \frac{\exp(-\alpha_T y_i h_T(x_i))}{Z_T}\cdot\frac{\exp(-\alpha_{T-1} y_i h_{T-1}(x_i))}{Z_{T-1}}\cdot \mathcal{D}_{T-1}(i)$$

$$= \cdots$$

$$= \frac{\exp(-\alpha_T y_i h_T(x_i))\cdot\exp(-\alpha_{T-1} y_i h_{T-1}(x_i))\cdot\ldots\cdot\exp(-\alpha_1 y_i h_1(x_i))}{Z_T\cdot Z_{T-1}\cdots Z_1}\,\mathcal{D}_1(i)$$

$\frac{1}{m}$

$$= \frac{1}{m}\cdot\frac{\exp\left(-\sum_{t=1}^{T}\alpha_t y_i h_t(x_i)\right)}{\prod_{t=1}^{T} Z_t}$$

Sum $i=1,\ldots,m$ both sides:

$$1 = \frac{1}{m}\sum_{i=1}^{m}\frac{\exp\left(-\sum_{t=1}^{T}\alpha_t y_i h_t(x_i)\right)}{\prod_{t=1}^{T} Z_t}, \text{ i.e.}$$

$$\prod_{t=1}^{T} Z_t = \frac{1}{m} \cdot \sum_{i=1}^{m} \exp\Big(-y_i \underbrace{\sum_{t=1}^{T} \alpha_t h_t(x_i)}_{= f(x_i)}\Big)$$

Note: didn't yet use, in L1 or L2, setting of α_t

Pf of L3: We'll show $Z_t \le \sqrt{1-4\gamma_t^2}$.

$$Z_t = \sum_{i=1}^{m} D_t(i) \exp(-\alpha_t y_i h_t(x_i)) = A + B$$

$$\underbrace{\sum_{\substack{i: \\ h_t(x_i) \neq y_i}} D_t(i) \exp(-\alpha_t y_i h_t(x_i))}_{= A}$$

$$+$$

$$\underbrace{\sum_{\substack{i: \\ h_t(x_i) = y_i}} D_t(i) \exp(-\alpha_t y_i h_t(x_i))}_{= B}$$

We saw $\underbrace{\sum_{i: h_t(x_i) \neq y_i} D_t(i)}_{= \epsilon_t}$, + saw that for $i \in$ sum A,

$$\exp(-\alpha_t y_i h_t(x_i)) = \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}, \text{ so}$$

$$A = \epsilon_t \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} = \underline{\sqrt{\epsilon_t(1-\epsilon_t)}}.$$

Similarly saw $B = (1-\varepsilon_t)\sqrt{\dfrac{\varepsilon_t}{1-\varepsilon_t}} = \underline{\sqrt{\varepsilon_t(1-\varepsilon_t)}}$

So $Z_t = 2\sqrt{\varepsilon_t(1-\varepsilon_t)} = \sqrt{4\varepsilon_t(1-\varepsilon_t)}$

since $\gamma_t = \frac{1}{2} - \varepsilon_t$, have

$$\varepsilon_t = \tfrac{1}{2} - \gamma_t$$
$$2\varepsilon_t = 1 - 2\gamma_t$$
$$2(1-\varepsilon_t) = 1 + 2\gamma_t$$

$$= \sqrt{(1-2\gamma_t)(1+2\gamma_t)} = \sqrt{1-4\gamma_t^2}. \quad \blacksquare$$

$1 - x \le e^{-x}$ so $\sqrt{1-x} \le e^{-x/2}$

$x = 4\gamma_t^2$: done.

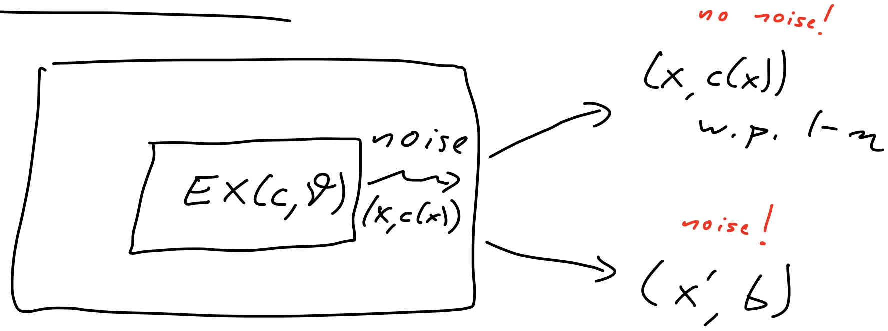
PAC Learning in the Presence of Noise

General noise framework:

still have $\mathcal{C}$, dist $\mathcal{D}$, still want

ε-acc h w.r.t. c under $\mathcal{D}$ w.p. $1-\delta$.

But now learner accesses a <u>noisy example oracle</u>: <u>noise rate η</u>



---

2 diff noise models in this framework:

① <u>Random (mis)classification noise:</u> RCN

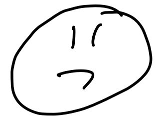
$$x' = x, \quad b = \overline{c(x)}.$$

② <u>Malicious noise:</u> (x', b) is <u>arbitrary</u>.

(Should view as gen. by omniscient malevolent adv. who "knows" $\mathcal{D}$, c, η, state of learning alg, etc.

Hi-level takeaway:

② Mal. noise very challenging: • if $\eta \approx 2\varepsilon$, can't achieve error $\leq \varepsilon$.



- Best known methods to deal w/ mal. noise: weak.

---

(1) RCN generally <u>easy</u> to deal with.

General method lets us convert many PAC algs into variants that can handle RCN, even at $\eta = 0.49$.

($\eta = \frac{1}{2}$: learning impossible.

Simple example of above:

Consider 2 worlds: (noise-free data sources)

W1	W2
$\frac{3}{8}$ + pos ex	$\frac{5}{8}$ + pos
$\frac{5}{8}$ − neg ex	$3/8$ − neg.

W1 or W2 ?

Mal noise at rate $\frac{1}{5} = \eta$:

in both worlds, can make data look

50 − 50 $\overset{''}{\eta}$

RCN at $\eta < \frac{1}{2}$:

W1: see + w.p.

(true pos) (no noise) (true neg) (noise)

$\frac{3}{8} \cdot (1-\eta) + \frac{5}{8} \cdot \eta$

$= \frac{3}{8} + \frac{\eta}{4}$.

W2: see + w.p. $\frac{5}{8} - \frac{\eta}{4}$.

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

So with enough $\approx \left(\frac{1}{(1-2\eta)^2} \right)$ ex, can

tell whether in W1 or W2.