Last time: "boosting over fixed sample"
- Ada Boost: $h_i$ has error $\frac{1}{2}$ under $P_{x_i}$, (like $h_i$ under $P_x$ in Schapire's booster)
- This time: prove that saying gives $h_i \rightarrow$ acc final $H$ over $S$.
- New unit: learning w/ noise. KV chap 5 (start every learning and misclass noise)

\[ \frac{1}{m} \sum_{i \in [m]} | \{i : e \neq y_i \} : H(x_i) \neq y_i, \frac{1}{2} \sum_{i \in [m]} \sqrt{1 - 4 \varepsilon_i^2} \] 
- error of $H$ on $S$
- small if $\varepsilon_i$'s decent.

Cor: Suppose each $\varepsilon_i \geq \gamma > 0$. Can run Ada Boost for
\[ T = \frac{1}{2 \gamma} \ln \left( \frac{1}{\gamma} \right) \] stages & RHS $\leq \varepsilon$.
- CB-like quantity...
- optimal: every booster, given only $\gamma$-adv. weak learner needs
\[ \Omega \left( \frac{1}{\gamma^2} \log \frac{1}{\gamma} \right) \] stages, for $\varepsilon$-acc.
- boosting "can't be parallelized" i.e. need this many sequential stages.

Pf: 3 claims.

Claim 1: \[ \frac{1}{m} \sum_{i \in [m]} | \{i : e \neq y_i \} : H(x_i) \neq y_i, \frac{1}{2} \sum_{i \in [m]} \exp (-y_i f(x_i)) \]
- Recall $f(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$, $H(x) = \text{sign}(f(x))$

Pf: $\frac{1}{m}$ (some $I$'s). Consider $i$ s.t. $H(x_i) \neq y_i$.
- $\exists \gamma i$. $\exists i$. Mean $H(x) = -1$, i.e. $\text{sign}(f(x)) = -1$, i.e. $f(x) < 0$, so
- $\exp(-y_i f(x_i)) = \exp(\text{pos}) > 1$. So each $\frac{1}{m}$ contr. to LHS
- matched by contr. of $\frac{1}{m}$ to RHS. So ineq. holds.

Claim 2: \[ \frac{1}{m} \sum_{i \in [m]} \exp(-y_i f(x_i)) = \prod_{t=1}^{T} Z_t \]
- Pf of Claim 2: use rec. def. of $\sum_{t=1}^{T} Z_t$ in terms of $\sum_{t=1}^{T} \varepsilon_t$.
Last distr:

\[ \mathbb{D}_{T+1}(i) = \frac{\exp(-\alpha_T y, h_T(x_i))}{Z_T} \cdot \mathbb{D}_T(i) \]

prod of T ratios

\[ = \left( \frac{\exp(-\alpha_T y, h_T(x_i))}{Z_T} \right) \cdot \ldots \cdot \frac{\exp(-\alpha_T y, h_T(x_i))}{Z_T} \cdot \mathbb{D}_{T-1}(i) \]

\[ \mathbb{D}_T(i) = \frac{1}{m} \cdot \sum_{i=1}^{m} \exp \left( -\frac{T}{\varepsilon_c} \left( \frac{T}{\varepsilon_c} \cdot f(x_i) \right) \right) \]

\[ D_{T+1}(i) \text{ a distr sums to } Z. \]

Last claim: Need to show \( Z_c = \sqrt{1 - \varepsilon_c^2} \).

\[ P_c: \quad Z_c = \frac{1}{m} \sum_{i=1}^{m} \mathbb{D}_T(i) \exp(-\varepsilon_c y, h_c(x_i)) \]

\[ = \frac{1}{m} \sum_{i=1}^{m} \mathbb{D}_c(i) \exp(-\varepsilon_c y, h_c(x_i)) + \sum_{i: h_c(x_i) \neq y} \mathbb{D}_c(i) \exp(-\varepsilon_c y, h_c(x_i)) \]

\( \varepsilon_c = \frac{1}{2} (1 - \varepsilon_c) \)

Last time: saw \( A = \left( \sum_{i: h_c(x_i) \neq y} \mathbb{D}_c(i) \right) \cdot \sqrt{\frac{1 - \varepsilon_c}{\varepsilon_c}} = \varepsilon_c \cdot \sqrt{\frac{1 - \varepsilon_c}{\varepsilon_c}} = \sqrt{\varepsilon_c (1 - \varepsilon_c)} \)

\[ = \varepsilon_c \]

\[ + B = \left( \sum_{i: h_c(x_i) = y} \mathbb{D}_c(i) \right) \cdot \sqrt{\frac{\varepsilon_c}{1 - \varepsilon_c}} = \left( 1 - \varepsilon_c \right) \cdot \sqrt{\frac{\varepsilon_c}{1 - \varepsilon_c}} = \sqrt{\varepsilon_c (1 - \varepsilon_c)}. \]

So \( Z_c = \sqrt{\frac{\varepsilon_c (1 - \varepsilon_c)}{2}} \cdot \sqrt{2 \varepsilon_c (1 - \varepsilon_c)} = \sqrt{2 \varepsilon_c (1 - \varepsilon_c)} \).
\[ 2\varepsilon_t = 1 - 2\varepsilon_{t-1} \quad \text{so} \quad \downarrow = \sqrt{(1-2\varepsilon_t)/(1+2\varepsilon_t)} = \sqrt{1-4\varepsilon_t^2}. \]

**New unit: Learning in Presence of Noise**

*KV book Chap. 5.*

*Motiv.:* PAC model, avg. noise. New noise process.

*Setup:* New learner accesses a noisy version of the ex oracle:

- w/ prob. \(\varepsilon\), indp. each ex rec’d is noisy \((\hat{x}, b)\):
  - w. prob. \(1-\varepsilon\), get \((x, c(x))\) noisel.
  - w. prob. \(\varepsilon\), get \((\hat{x}, b)\).
- noisy ex.

*Goal:* unchanged. Still want e-acc h w.r.t. \(\mathcal{D}\), w/ prob \(\frac{1}{2}\).

*Diff. assumptions about \((\hat{x}, b) \leftrightarrow \text{diff. noise models.}*

1. Random Classification Noise (RCN): \((\hat{x}, b): x = \hat{x}, b = c(x)\) (flips label w.p. \(\varepsilon\))

2. Malicious Noise: \((\hat{x}, b)\) is arbitrary. \(\hat{x} \in \mathcal{X}, b \in \{0, 1\}\)

   Should think of \((\hat{x}, b)\) as being gen. by mal. omniscient adv. who knows \(\mathcal{D}, c, \varepsilon, \text{state of learning alg. etc.}\)

Others can be studied too: eg. random attrib. noise:

\(X = \{0, 1\}^n, (\hat{x}, b): b = c(x), \text{but each } \hat{x}\)

is obt. by indp. flip \(x\), w/ prob. \(p\).
H-level take away:
(2) Mal. noise: very hard to handle. We’ll show:
if have mal. noise at $n \geq 2\epsilon$, imposs. to learn to error. Best methods known to handle
mal. noise: very weak. 😞

(1) RCN: 😊 For most C’s, can handle $\epsilon = 0.49\%$

General method: many PAC ab’s $\rightarrow$ noise-tol.
RCN ab’s.


$\epsilon = \frac{1}{2}$, RCN: imposs. to learn: Each ex. is
just labeled by 😃.

Just for intuition: Consider 2 diff. poss.
for noise-free data.

$S_1$: $+ + + + - - - - - - - -$ $\frac{5}{8}$ pos $\frac{3}{8}$ neg.

$S_2$: $+ + + + + - - - - - -$ $5/8$ pos $3/8$ neg.

- Mal. noise, rate $\frac{1}{5} = \frac{\epsilon}{2}$:
in $S_1$, when $+$, unchanged
when $-$, flip to $+$. What learner sees:
in $S_2$, when $-$, unchanged
when $+$, flip to $-$.

What learner sees:
\[ \frac{1}{2} \text{ pos} \quad \frac{1}{2} \text{ neg} \quad \frac{1}{2} \text{ pos} \quad \frac{1}{2} \text{ neg} \]

impossible for learner to tell if \( S_1 \) or \( S_2 \).

- RCN rate \( \eta < \frac{1}{2} \) (\( \eta = 0.49 \)):
  \[ S_1 \text{ freq of } + \text{ will be } \frac{3}{8} (1-\eta) + \frac{5}{8} \eta = \frac{3}{8} + \frac{1}{4} \eta \]
  \[ \text{freq obs} = \frac{5}{8} - \frac{1}{4} \eta \]

- \( S_2 \) freq of obs + is \( \frac{5}{8} - \frac{1}{4} \eta \), freq obs - is \( \frac{3}{8} + \frac{1}{4} \eta \).

\[ \eta < \frac{1}{2} : \frac{3}{8} + \frac{1}{4} \eta < \frac{1}{2}, \quad \eta > \frac{1}{2}. \]

So for \( \eta < \frac{1}{2} \), can dist. \( S_1 \) from \( S_2 \) (cost \( \approx \frac{1}{1-2\eta} \)).

Sample to get very high acc. est. of freq of obs + \( \tilde{s} \):
  if \( \frac{1}{2} \), say \( S_1 \) +; if \( \frac{1}{2} \), say \( S_2 \).

Moreover, if known \( \eta \), given good est. of \( \frac{3}{8} + \frac{1}{4} \eta \), can use \( \epsilon \) to "solve for \( \frac{3}{8} \)." Can learn char. of orig. source from noisy data.

\underline{Lower Bound for Mal. Noise}

Any modestly nontriv. \( \epsilon \) too much mal. noise makes \( 1 - \epsilon \) acc. learn. impossible.
Def: A c.c. $E$ is distinct if $\exists c, c_2 \in E, u, v, w \in X$ s.t. $c_1 \cap c_2 = \emptyset, u, v, w \in c_1, w \notin c_2, v, w \in c_2, u \notin c_2$.

Thm: Let $E$ be a distinct c.c. Then no alg can PAC learn to acc $E$ with conf 0.49 in pres. of nul. noise at noise rate $\eta$ if $\eta > \frac{2e}{1 + e}$.

Idea: $S_1, S_2$. $S_1$: target conc. $c$, certain $D_1$; adv. acts one way.

$S_2$: target conc. $c_2$, certain $D_2$; adv. act diff. way.

Learner "see" exact same thing in $S_1$ + $S_2$. Impossible to tell which one she's in, thus $\varepsilon$-acc. $\gamma$ prob. $> \frac{1}{2}$ impossible. [KV chapter 5]