- PS3 due tonight
- Midterm in a week

- PS4 out next week (after midterm)

Last time:

- weak vs. strong PAC learning

γ-Weak learning assumption (γ-WLA) for a data set $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (X \times \{-1,1\})^m$

- equivalence of γ-WLA to linear separability in a certain feature space

  ↑ in dim $|H|$ where H is the hypothesis class used by W.L.

Today:

- the actual reduction from strong learning to weak learning (all done "in-sample")
  i.e., boosting algorithm

- "out-of-sample" properties

γ-Weak learning assumption (γ-WLA) for a data set $S = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \{-1,1\})^n$:

$\exists$ algo. WL

For any reweighting $p = (p_1, \ldots, p_n) \in \Delta^{n-1}$,

$WL(p)$ produces $h : \mathcal{X} \to \{-1,1\}$ with

$$\text{acc}(h, p) := \sum_{i=1}^{n} p_i \cdot \mathbb{1}\{h(x_i) = y_i\} \geq \frac{1+\gamma}{2}$$

Let $\mathcal{H} \subseteq \{-1,1\}^{\mathcal{X}}$ be the hypothesis class used by W.L.

$\phi(x) = (h(x))_{h \in \mathcal{H}}$

$\mathcal{H} = \{h_1, \ldots, h_d\}$, $d = |\mathcal{H}|$.

Our goal: Produce a $\hat{h} : \mathcal{X} \to \{-1,1\}$

s.t. $\text{acc}(\hat{h}, \text{uniform}) = 1$

$$\max_{q \in \Delta^{d-1}} \min_{p \in \Delta^{n-1}} p^T A q = \min_{p \in \Delta^{n-1}} \max_{q \in \Delta^{d-1}} p^T A q \geq \frac{1+\gamma}{2}$$

$$A_{ij} = \mathbb{1}\{h_j(x_i) = y_i\}$$

Boosting as Hedge

Input: $S = ((x_1, y_1), \dots, (x_n, y_n)) \in (\mathcal{X} \times \{-1, 1\})^n$

WL guaranteed to exist by γ-WLA.

Init: $p_1 = (\frac{1}{n}, \dots, \frac{1}{n}) \in \Delta^{n-1}$

For $t = 1, 2, \dots, T$:

- Use W.L. w/ $p_t = (p_{t,1}, \dots, p_{t,n}) \in \Delta^{n-1}$ as input to get $h_t : \mathcal{X} \to \{-1, 1\}$
- Define loss vector $l_t = (l_{t,1}, \dots, l_{t,n}) \in \{0, 1\}$

$$l_{t,i} = \mathbb{1}\{h_t(x_i) = y_i\}$$

- New weighting

$$p_{t+1,i} = \frac{p_{t,i} \exp(-\eta l_{t,i})}{Z_{t+1}} \qquad \forall i \in [n]$$

where Z_{t+1} is the normalizer so $p_{t+1} \in \Delta^{n-1}$.

Return $\hat{h}(x) = \text{sign}(\hat{g}(x))$ ← LTF in the feature space defined by ϕ.

$$\hat{g}(x) = \frac{1}{T} \sum_{t=1}^{T} h_t(x)$$

Theorem about Hedge tells us: $\forall u \in \Delta^{n-1}$

$$\frac{1+\gamma}{2} \le \frac{1}{T} \sum_{t=1}^{T} \langle p_t, l_t \rangle \le \frac{1}{T} \sum_{t=1}^{T} \langle u, l_t \rangle + \frac{\ln(n)}{\eta T} + \frac{\eta}{2}$$

$$\frac{1}{T} \sum_{t=1}^{T} \langle p_t, l_t \rangle = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n} p_{t,i} \cdot \mathbb{1}\{h_t(x_i) = y_i\} \ge \frac{1+\gamma}{2}.$$

$\langle p_t, l_t \rangle$ = $\text{acc}(h_t, p_t)$

$\text{acc}(h_t, p_t) \ge \frac{1+\gamma}{2}$ by γ-WLA

Let's plug-in $u = e_i$

$$\frac{1}{T} \sum_{t=1}^{T} \langle e_i, l_t \rangle = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{h_t(x_i) = y_i\}$$

$$= \frac{1}{T} \sum_{t=1}^{T} \frac{1 + y_i h_t(x_i)}{2}$$

$$= \frac{1}{2} + \frac{1}{2T} y_i \sum_{t=1}^{T} h_t(x_i)$$

$$= \frac{1}{2} + \frac{1}{2} y_i \hat{g}(x_i)$$

$\forall i \in [n]$

$$\gamma \le y_i \hat{g}(x_i) + \frac{2\ln(n)}{\eta T} + \eta$$

$$y_i \hat{g}(x_i) \ge \gamma - \frac{2\ln(n)}{\eta T} - \eta$$

Choose $\eta := \frac{\gamma}{2}$, $T > \frac{8 \ln n}{\gamma^2}$

then

$y_i \hat{g}(x_i) > 0$. So $\hat{h}(x_i) = \text{sign}(\hat{g}(x_i)) = y_i \quad \forall i$

Slightly different analysis: $\forall u \in \Delta^{n-1}$

$$\frac{1}{T} \sum_{t=1}^{T} \langle p_t, l_t \rangle \le \frac{1}{T} \sum_{t=1}^{T} \langle u, l_t \rangle + \frac{RE(u, p_1)}{\eta T} + \frac{\eta}{2}$$

u = uniform on $M = \{i \in [n] \mid y_i \hat{g}(x_i) \le 0\}$

$$RE(u, p_1) = \sum_{i=1}^{n} u_i \ln \frac{u_i}{1/n} = \ln\left(\frac{n}{|M|}\right)$$

$$\frac{1+\gamma}{2} \le \sum_{i=1}^{n} u_i \cdot \left(\frac{1}{2} + \frac{1}{2} y_i \hat{g}(x_i)\right) + \frac{\ln\left(\frac{n}{|M|}\right)}{\eta T} + \frac{\eta}{2}$$

$$= \frac{1}{2} + \frac{1}{2} \sum_{i=1}^{n} u_i \, y_i \hat{g}(x_i) + \frac{\ln\left(\frac{n}{|M|}\right)}{\eta T} + \frac{\eta}{2}$$

≤ 0

$$\gamma \le \frac{2\ln(n)}{\eta T} - \frac{2\ln(|M|)}{\eta T} + \eta$$

$$\ln|M| \le \ln(n) - \frac{(\gamma - \eta)\eta T}{2} \qquad \eta = \frac{\gamma}{2}$$

$$= \ln(n) - \frac{\gamma^2 T}{8}$$

$$|M| \le n \cdot \exp(-\gamma^2 T / 8)$$

Out-of-sample properties of $\hat{h}$ from boosting.

W.L. uses hypothesis class $\mathcal{H}$
- Let's say it has VC dim d_{WL}

Boosting uses

$$\text{Majority}_{\mathcal{H},T} = \left\{ x \mapsto \text{sign}\left(\frac{1}{T}\sum_{t=1}^{T} h_t(x)\right) \;\middle|\; h_1,\dots,h_T \in \mathcal{H} \right\}$$

Prop $\text{VCdim}\left(\text{Majority}_{\mathcal{H},T}\right) \le O\left(d_{WL} \cdot T \log T\right)$

Pf: $S \subset X$, $m = |S|$

$2^m = \left|\text{Maj}_{\mathcal{H},T}[S]\right|$ if S is shattered by $\text{Maj}_{\mathcal{H},T}$

$\le |\mathcal{H}[S]|^T$

$\le \binom{m}{\le d_{WL}}^T$

$\le \left(\frac{me}{d_{WL}}\right)^{d_{WL}\cdot T}$ ... $m \le O\left(d_{WL} \cdot T \log T\right)$

Using "uniform convergence" technology, can derive sample complexity bounds

$T = O\left(\frac{\log n}{\gamma^2}\right)$ is sufficient to get $\text{acc}(\hat{h}) = 1$

$$n = \frac{\text{VCdim}\left(\text{Maj}_{\mathcal{H},T}\right)}{\epsilon}\log\frac{1}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta} = \tilde{O}\left(\frac{d_{WL}}{\epsilon\gamma^2}\right)$$

$O(d_{WL} \cdot T(\log T))$