20225701 Advanced Seminar in Machine Learning **Lecturer:** Aryeh Kontorovich
Agnostic PAC lower bound July 5, 2016

These notes use the notation, definitions, and results from:

- *Bias of a random coin*[BRC] `http://tinyurl.com/bias-rand-coin` [1]
- *Pinsker's inequality*[Pinsker] `http://tinyurl.com/pol-pin` [2]

For any distribution D over $\mathcal{X} \times \{-1, 1\}$ and hypothesis class $\mathcal{H}$, define

$$h_D^* := \arg\min_{h \in \mathcal{H}} \mathrm{err}(h, D)$$

(there is no loss of generality in assuming that the minimum is achieved and unique).

For any $h \in \mathcal{H}$, define its **excess risk** under D:

$$\Delta(h, D) := \mathrm{err}(h, D) - \mathrm{err}(h_D^*, D).$$

Note that $\Delta(h, D) \geq 0$ always (why?).

Theorem 0.1 *For any hypothesis class $\mathcal{H}$ with VC-dimension $d \geq 1$ and $n \geq d/\log 3$, we have*

$$\inf_{\hat{h}_n} \sup_{D} \mathop{\mathbb{E}}_{D^n} \left[\Delta(\hat{h}_n, D) \right] \geq \sqrt{\frac{d/n}{64 \log 3}} \approx 0.1193 \sqrt{\frac{d}{n}},$$

where the infimum is over all learning algorithms $\mathcal{L}$ that map samples of size n to hypotheses $\hat{h}_n \in \mathcal{H}$ and the supremum is over all distributions over the instance space $\mathcal{X}$.

The quantity on the left-hand side of the inequality is know as the minimax expected excess risk.

Proof:

Step I: Defining the adversarial distribution. There is no loss of generality in taking $\mathcal{X} = [d] \equiv \{1, \ldots, d\}$ and $\mathcal{H} = \{-1, 1\}^{\mathcal{X}}$ (why?). For $\gamma \in \{-1, 1\}^{\mathcal{X}}$ and $b \in [0, 1]$, define the distribution $D_{b,\gamma}$ over $\mathcal{X} \times \{-1, 1\}$ as follows:

$$\mathop{\mathbb{P}}_{(X,Y) \sim D_{b,\gamma}} (X = x, Y = y) = \frac{1}{d} \left(\frac{1}{2} + \frac{y b \gamma(x)}{2} \right).$$

In words, X is drawn uniformly from $\mathcal{X} = [d]$ and conditioned on $X = x$, Y is drawn from $\mathrm{R}(b\gamma(x))$, where $\mathrm{R}(\cdot)$ is the Rademacher distribution defined in [BRC].

[1] long url: `https://www.cs.bgu.ac.il/~asml162/wiki.files/b-k-opt.pdf`
[2] long url: `https://www.cs.bgu.ac.il/~asml162/wiki.files/pollard-pinsker.pdf`

Step II: Computing the excess risk. Convince yourself that for $D_{g,\gamma}$ as defined above,

$$h^*_{D_{b,\gamma}}(x) = \arg\max_{y\in\{-1,1\}} \mathop{\mathbb{P}}_{(X,Y)\sim D_{b,\gamma}} (Y = y \,|\, X = x) = \gamma(x)$$

(this follows directly from (1) and (2) in [BRC]). Since $h^*_{D_{b,\gamma}}$ does not depend on b, we will write $h^*_\gamma := h^*_{D_{b,\gamma}}$ as a shorthand. Let us compute $\mathrm{err}(h^*_\gamma, D_{b,\gamma})$:

$$\begin{aligned}
\mathrm{err}(h^*_\gamma, D_{b,\gamma}) &= \mathop{\mathbb{P}}_{(X,Y)\sim D_{b,\gamma}} (\gamma(X) \neq Y) \\
&= \mathbb{P}(\gamma(X) = 1)\,\mathbb{P}(Y \neq 1 \,|\, \gamma(X) = 1) + \mathbb{P}(\gamma(X) = -1)\,\mathbb{P}(Y \neq -1 \,|\, \gamma(X) = -1) \\
&= \frac{1}{2} - \frac{b}{2} \qquad \text{(ex1)}
\end{aligned}$$

(exercise: verify this carefully).

Now let us compute the $\mathrm{err}(h, D_{b,\gamma})$ for an arbitrary $h \in \mathcal{H}$:

$$\begin{aligned}
\mathrm{err}(h, D_{b,\gamma}) &= \mathop{\mathbb{P}}_{(X,Y)\sim D_{b,\gamma}} (h(X) \neq Y) \\
&= \mathbb{P}\big(h(X) = h^*_\gamma(X)\big) \cdot \mathbb{P}\big(h(X) \neq Y \,|\, h(X) = h^*_\gamma(X)\big) \\
&+ \mathbb{P}\big(h(X) \neq h^*_\gamma(X)\big) \cdot \mathbb{P}\big(h(X) \neq Y \,|\, h(X) \neq h^*_\gamma(X)\big)\,.
\end{aligned}$$

We know from the above that

$$\begin{aligned}
\mathbb{P}\big(h(X) \neq Y \,|\, h(X) = h^*_\gamma(X)\big) &= \mathbb{P}\big(h^*_\gamma(X) \neq Y\big) = \frac{1}{2} - \frac{b}{2}, \\
\mathbb{P}\big(h(X) \neq Y \,|\, h(X) \neq h^*_\gamma(X)\big) &= \mathbb{P}\big(h^*_\gamma(X) = Y\big) = \frac{1}{2} + \frac{b}{2},
\end{aligned}$$

whence

$$\begin{aligned}
\mathrm{err}(h, D_{b,\gamma}) &= \mathbb{P}\big(h(X) = h^*_\gamma(X)\big)\left(\frac{1}{2} - \frac{b}{2}\right) + \mathbb{P}\big(h(X) \neq h^*_\gamma(X)\big)\left(\frac{1}{2} + \frac{b}{2}\right) \\
&= \left(\frac{1}{2} - \frac{b}{2}\right) + b\,\mathbb{P}\big(h(X) \neq h^*_\gamma(X)\big) \\
&= \mathrm{err}(h^*_\gamma, D_{b,\gamma}) + \frac{b}{d}\sum_{x\in[d]} \mathbb{1}_{\{h(x)\neq\gamma(x)\}} \qquad \text{(ex2)}
\end{aligned}$$

(exercise: verify this carefully).

We conclude that

$$\Delta(h, D_{b,\gamma}) = \frac{b}{d}\sum_{x\in[d]} \mathbb{1}_{\{h(x)\neq\gamma(x)\}}. \qquad (1)$$

Step III: Taking expectations. Let $S_n(X_i, Y_i)_{i\in[n]}$ be drawn from $D^n_{b,\gamma}$ and define the random variables N_x to be the number of occurrences of the point $x \in \mathcal{X}$ in the random sample S_n:

$$N_x = \sum_{i=1}^{n} \mathbb{1}_{\{X_i = x\}}. \tag{2}$$

Let $\mathcal{L} : S_n \mapsto \hat{h}_n \in \mathcal{H}$ be any learning algorithm, and let us lower-bound $\mathbb{E}[\Delta(\hat{h}_n, D_{b,\gamma})]$, where the expectation is both over $\gamma \sim \text{Unif}\left(\{-1,1\}^d\right)$ and $S_n \sim D^n_{b,\gamma}$:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\gamma\sim\text{Unif}\left(\{-1,1\}^d\right)} \mathop{\mathbb{E}}_{S_n\sim D^n_{b,\gamma}} \left[\Delta(\hat{h}_n, D_{b,\gamma})\right] &= \mathop{\mathbb{E}}_{\gamma} \mathop{\mathbb{E}}_{S_n} \left[\frac{b}{d} \sum_{x\in[d]} \mathbb{1}_{\{\hat{h}_n(x)\neq\gamma(x)\}}\right] \\
&= \frac{b}{d} \sum_{x\in[d]} \mathop{\mathbb{P}}_{\gamma\sim\text{Unif}\left(\{-1,1\}^d\right), \hat{h}_n\sim\mathcal{L}(S_n)} \left(\hat{h}_n(x) \neq \gamma(x)\right)
\end{aligned}$$

— note that the inner $\mathbb{P}(\cdot)$ is NOT over the input to $\hat{h}_n$ but rather over $\hat{h}_n$ itself! Let us focus on a fixed $x \in [d]$:

$$\begin{aligned}
\mathop{\mathbb{P}}_{\gamma\sim\text{Unif}\left(\{-1,1\}^d\right), \hat{h}_n\sim\mathcal{L}(S_n)} \left(\hat{h}_n(x) \neq \gamma(x)\right) &= \mathop{\mathbb{E}}_{N_x} \left[\mathop{\mathbb{P}}_{\gamma, \hat{h}_n} \left(\hat{h}_n(x) \neq \gamma(x)\right) \mid N_x\right] \\
&= \sum_{k=0}^{m} \mathbb{P}(N_x = k) \mathop{\mathbb{P}}_{\gamma, \hat{h}_n} \left(\hat{h}_n(x) \neq \gamma(x) \mid N_x = k\right),
\end{aligned}$$

where we can pull the $\mathbb{E}_\gamma$ inside because N_x and γ are independent. Recall the definition of $\text{opt}(k, b)$ from [BRC] and convince yourself that

$$\mathop{\mathbb{P}}_{\gamma\sim\text{Unif}\left(\{-1,1\}^d\right), \hat{h}_n\sim\mathcal{L}(S_n)} \left(\hat{h}_n(x) \neq \gamma(x) \mid N_x = k\right) \geq \text{opt}(k, b); \tag{ex3}$$

this is the most important step of the proof.

From here, it is easy to conclude that

$$\mathop{\mathbb{E}}_{\gamma\sim\text{Unif}\left(\{-1,1\}^d\right)} \mathop{\mathbb{E}}_{S_n\sim D^n_{b,\gamma}} \left[\Delta(\hat{h}_n, D_{b,\gamma})\right] \geq \frac{b}{d} \sum_{x\in[d]} \mathop{\mathbb{E}}_{N_x} \text{opt}(N_x, b). \tag{3}$$

Step IV: Lower-bounding opt**.** Recall from [BRC] that

$$\mathrm{opt}(k,b) = \frac{1}{2}\left(1 - \frac{1}{2}\left\|\mathrm{R}(b)^k - \mathrm{R}(-b)^k\right\|_1\right) \tag{4}$$

and from [Pinsker] that for all distributions P, Q,

$$\|P - Q\|_1 \le \sqrt{2\,\mathrm{KL}(P||Q)}. \tag{5}$$

We will apply (5) to $P = \mathrm{R}(b)^k$, $Q = \mathrm{R}(-b)^k$. From Homework 3.2(b), we know that

$$\begin{aligned}\mathrm{KL}(\mathrm{R}(b)^k||\,\mathrm{R}(-b)^k) &= k\,\mathrm{KL}(\mathrm{R}(b)||\,\mathrm{R}(-b)) \\ &= kb\log\frac{1+b}{1-b} \qquad\qquad \text{(ex4)}\end{aligned}$$

(exercise: verify the last claim). From Homework 3.3, we know that

$$\sqrt{2b\log\frac{1+b}{1-b}} \le 2b\sqrt{\log 3}, \qquad 0 \le b \le \frac{1}{2} \tag{6}$$

and hence

$$\mathrm{opt}(k,b) \ge \tfrac{1}{2}\left(1 - \sqrt{(\log 3)kb^2}\right). \tag{7}$$

Substituting into (3), we get

$$\mathop{\mathbb{E}}_{\gamma\sim\mathrm{Unif}(\{-1,1\}^d)}\mathop{\mathbb{E}}_{S_n\sim D_{b,\gamma}^n}\left[\Delta(\hat{h}_n, D_{b,\gamma})\right] \;\ge\; \frac{b}{2d}\sum_{x\in[d]}\mathop{\mathbb{E}}_{N_x}\left[1 - \sqrt{(\log 3)N_x b^2}\right].$$

Now the function $t \mapsto \left(1 - \sqrt{(\log 3)tb^2}\right)$ is convex, and furthermore, $\mathbb{E}[N_x] = n/d$. Hence, by Jensen's inequality,

$$\frac{b}{2d}\sum_{x\in[d]}\mathop{\mathbb{E}}_{N_x}(1 - \sqrt{(\log 3)N_x b^2}) \ge \frac{b}{2}\left(1 - \sqrt{\frac{(\log 3)nb^2}{d}}\right) =: B_{n,d}(b)$$

(what happened to $\frac{1}{d}\sum_x$?).

Step V: optimizing b**.** It remains specify the parameter b, which we choose to be of the form $b = c\sqrt{d/n}$, where c will be chosen so as to maximize $B_{n,d}(\cdot)$. Now

$$B_{n,d}\left(c\sqrt{d/n}\right) = \frac{1}{2}\sqrt{\frac{d}{n}}c\left(1 - c\sqrt{\log 3}\right),$$

which achieves the maximum value of $\sqrt{\frac{d/n}{64\log 3}}$ at $c = \frac{1}{\sqrt{4\log 3}}$ [(ex5): verify this!].

This proves the claim, since

$$\sup_D \mathop{\mathbb{E}}_{D^n}\left[\Delta(\hat{h}_n, D)\right] \;\ge\; \mathop{\mathbb{E}}_{\gamma}\mathop{\mathbb{E}}_{D_{b,\gamma}^n}\left[\Delta(\hat{h}_n, D_{b,\gamma})\right].$$

The condition $n \ge d/\log 3$ ensures that $b = \sqrt{\frac{d}{4(\log 3)n}} \le \frac{1}{2}$, so that the estimate in (6) applies.

■