CSC 665: Information-theoretic lower bounds of PAC sample complexity

Chicheng Zhang

September 26, 2019

In the last lecture, we show that finite VC dimension is sufficient for distribution-free agnostic PAC learnability. For a hypothesis class $\mathcal{H}$ of VC dimension d, for all data distributions, ERM has an agnostic PAC sample complexity $O\left(\frac{1}{\epsilon^2}(d\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta})\right)$. [1]

In this lecture, to complement the learnability result, given $\mathcal{H}$ of VC dimension d, we show that *any learning algorithm* must consume at least $\Omega\left(\frac{1}{\epsilon^2}(d+\ln\frac{1}{\delta})\right)$ samples to achieve agnostic PAC learning guarantee. Moreover, if $\mathcal{H}$ has infinite VC dimension, any learning algorithm is unable to achieve distribution-free PAC learning. The latter fact implies that finite VC dimension is *necessary* for distribution-free PAC learnability.

Theorem 1. *For any hypothesis class $\mathcal{H}$ such that $\mathrm{VC}(\mathcal{H}) \geq d$, and any learning algorithm $\mathcal{A}$, and any $\epsilon, \delta \in (0, \frac{1}{8})$, there exists a distribution D over $\mathcal{X} \times \{-1, +1\}$, such that when a set S of $m = \frac{1}{16\epsilon^2}(\frac{d}{200}+\ln\frac{1}{16\delta})$ examples is drawn iid from D, with probability at least δ,*

$$\mathrm{err}(\hat{h}, D) - \min_{h\in\mathcal{H}} \mathrm{err}(h, D) > \epsilon,$$

where $\hat{h} = \mathcal{A}(S)$ is the output of learning algorithm.

Remark. Note well the order of quantifiers on algorithm $\mathcal{A}$ and distribution D. It may be tempting to show a theorem that says "for every $\mathcal{H}$, ϵ and δ, there is a distribution D such that for any algorithm $\mathcal{A}$, $\mathcal{A}$ fail to satisfy (ϵ, δ)-agnostic PAC learning guarantee." Unfortunately this is impossible. Suppose the distribution D is chosen, then there is a trivial algorithm $\mathcal{A}$ that satisfies the agnostic PAC learning guarantee - outputting $h^\star = \mathrm{argmin}_{h\in\mathcal{H}} \mathrm{err}(h, D)$.

We show the theorem in the following two lemmas.

Lemma 1. *Suppose the setting is the same as that of Theorem 1. There exists a distribution D such that, if m, the size of S is at most $\frac{1}{8\epsilon^2}\ln\frac{1}{16\delta}$, then with probability at least δ,*

$$\mathrm{err}(\hat{h}, D) - \min_{h\in\mathcal{H}} \mathrm{err}(h, D) > \epsilon.$$

Lemma 2. *Suppose the setting is the same as that of Theorem 1. There exists a distribution D such that, if m, the size of S is at most $\frac{d}{1600\epsilon^2}$, then with probability at least $1/4$,*

$$\mathrm{err}(\hat{h}, D) - \min_{h\in\mathcal{H}} \mathrm{err}(h, D) > \epsilon.$$

To see why the two lemmas together imply the theorem, consider two cases. When $\frac{d}{200} \geq \ln\frac{1}{16\delta}$, by Lemma 2, $\mathcal{A}$ will fail to satisfy agnostic PAC guarantee with $m = \frac{1}{16\epsilon^2}(\frac{d}{200}+\ln\frac{1}{16\delta}) \leq \frac{d}{1600\epsilon^2}$ training examples. Similarly, when $\frac{d}{200} < \ln\frac{1}{16\delta}$, by Lemma 1, $\mathcal{A}$ will fail to satisfy agnostic guarantee with $m = \frac{1}{16\epsilon^2}(\frac{d}{200}+\ln\frac{1}{16\delta}) \leq \frac{1}{8\epsilon^2}\ln\frac{1}{16\delta}$ training examples.

[1] In fact, the sample complexity can be sharpened to $O\left(\frac{1}{\epsilon^2}(d+\ln\frac{1}{\delta})\right)$ by an advanced technique called chaining (see Section 27.2 of [1]).

1 Proof of Lemma 1: an introduction to Le Cam's method

Le Cam's method [2] is a systematic way to prove information theoretic lower bounds. It is based on the following thought experiment. Suppose we are given two possible distributions $P_i, i \in \{\pm 1\}$ over the observation space $\mathcal{O}$ (where each draw from the distribution results in an observation O in $\mathcal{O}$). Our task is to guess the identity of i given O, i.e. output a $\hat{i}$ based on O (we can think of $\hat{i} = f(O)$, where f encodes our thought process). If P_{+1} and P_{-1} are close, then there exists at least one distribution P_i, under which our guess $\hat{i}$ would be wrong with decent probability.

(It may be helpful to think of P_{+1} and P_{-1} as two possible "scientific hypotheses", and O is an scientific experiment we conduct. Our task is to tell which hypothesis is the ground truth.) If you are familar with hypothesis testing in statistics, this is exactly the same setting: we would like to show that no matter what test we use, the sum of type I and type II errors would be large so long as the two hypotheses are close to each other.

We will use the shorthand that $\mathbb{P}_i$ (resp. $\mathbb{E}_i$) denotes $\mathbb{P}_{O\sim P_i}$ (resp. $\mathbb{E}_{O\sim P_i}$).

Lemma 3 (Le Cam's method)**.** *Suppose f is a mapping from $\mathcal{O}$ to $\{-1,+1\}$. Then for at least one of i in $\{-1,+1\}$,*

$$\mathbb{P}_i(f(O)\neq i) = \mathbb{E}_i \mathbf{1}(f(O)\neq i) \geq \frac{1}{2}\sum_{o\in\mathcal{O}} \min(P_{-1}(o), P_{+1}(o)).$$

Remark. The right hand side is often written as $\|P_{-1}\wedge P_{+1}\|_1$, measuring the similarity between two distributions. Generally, if we have two distributions Q_1 and Q_2, we have:

$$\begin{aligned}
\|Q_1\wedge Q_2\|_1 &= \sum_{o\in\mathcal{O}} \min\big(Q_1(o), Q_2(o)\big) \\
&= \sum_{o\in\mathcal{O}} \frac{Q_1(o)+Q_2(o)}{2} - \frac{|Q_1(o)-Q_2(o)|}{2} \\
&= 1 - \sum_{o\in\mathcal{O}} \frac{|Q_1(o)-Q_2(o)|}{2} \\
&= 1 - \frac{1}{2}\|Q_1-Q_2\|_1.
\end{aligned}$$

As a sanity check, if $Q_1 = Q_2$, $\|Q_1\wedge Q_2\|_1 = 1$ and $\|Q_1 - Q_2\|_1 = 0$; on the other extreme, if Q_1 and Q_2 have disjoint support, then $\|Q_1\wedge Q_2\|_1 = 0$ and $\|Q_1 - Q_2\|_1 = 2$.

Suppose I is chosen uniformly at random from $\{\pm 1\}$. What is the function $f^\star$ that minimizes $\mathbb{P}(f(O)\neq I)$? Think of the problem as a binary classification problem, where (feature,label) pair (O, I) comes from a joint distribution we have full knowledge about. Given O, we would like to classify O as either $+1$ or -1 to minimize the error.

If you have studied probabilistic machine learning, you now can see that $f^\star$ is the Bayes classifier:

$$f^\star(o) = \begin{cases} +1 & \mathbb{P}(I=+1|O=o)\geq \frac{1}{2} \\ -1 & \text{otherwise} \end{cases}$$

Why does this function minimize the error rate? Observe that for any function f,

$$\mathbb{P}(f(O)\neq I) = \sum_{o\in\mathcal{O}} \mathbb{P}(O=o)\left(\mathbb{P}(I=-1|O=o)\mathbf{1}(f(o)=+1) + \mathbb{P}(I=-1|O=o)\mathbf{1}(f(o)=-1)\right],$$

if we would like to choose f that minimize $\mathbb{P}(f(O)\neq I)$, it suffices for us to decide for each o, whether $f(o)$ should take value -1 or $+1$. Therefore, the f that minimizes the error will choose to predict $\operatorname{argmax}_{i\in\{\pm1\}} \mathbb{P}(I = i|O=o)$, which is equivalent to $f^\star(o)$.

This means that we can calculate $\mathbb{P}(f(O) \neq I)$ explicitly. In addition,

$$\mathbb{P}(f(O) \neq I) = \frac{1}{2}\left(\mathbb{P}_{+1}(f(O) \neq +1) + \mathbb{P}_{-1}(f(O) \neq -1)\right) \leq \max_i \mathbb{P}_i(f(O) \neq i), \tag{1}$$

so a lower bound of $\mathbb{P}(f(O) \neq I)$ implies a lower bound of $\max_i \mathbb{P}_i(f(O) \neq i)$.

Let us now formalize the ideas above.

Proof. Suppose I is chosen uniformly from $\{\pm 1\}$, and given I, O is drawn from $\mathbb{P}_I$. Then for any function f,

$$\begin{aligned}
\mathbb{P}(f(O) \neq I) &\geq \mathbb{P}(f^\star(O) \neq I) \\
&= \frac{1}{2}\left(\mathbb{P}_{-1}(f^\star(O) = +1) + \mathbb{P}_{+1}(f^\star(O) = -1)\right) \\
&= \frac{1}{2}\left(\sum_{o: P_{+1}(o) \geq P_{-1}(o)} P_{-1}(o) + \sum_{o: P_{-1}(o) > P_{+1}(o)} P_{+1}(o)\right) \\
&= \frac{1}{2}\sum_{o \in \mathcal{O}} \min\left(P_{-1}(o), P_{+1}(o)\right) \quad \square
\end{aligned}$$

Le Cam's method is a statement about hypothesis testing. How can Le Cam's method be useful in sample complexity lower bounds? It turns out that we can construct a pair of learning problems, such that in order to ensure PAC learning on both problems, solving a variant of hypothesis testing is *necessary*.

The construction. Suppose that x_0 is an unlabeled example, $\mathcal{H}$ contains two classifiers h_{+1} and h_{-1}, such that $h_i(z_0) = i$ for both $i \in \{-1, +1\}$. Define an unlabeled distribution D_X such that $\mathbb{P}_{D_X}(x = z_0) = 1$. For $i \in \{\pm 1\}$, define

$$D_i(y|z_0) = \begin{cases} \frac{1}{2} + i\epsilon, & y = +1, \\ \frac{1}{2} - i\epsilon, & y = -1. \end{cases}$$

In other words,

$$D_i(y|z_0) = \begin{cases} \frac{1}{2} + \epsilon, & y = i, \\ \frac{1}{2} - \epsilon, & y = -i. \end{cases}$$

In addition, D_{+1} (resp. D_{-1}) are specified by the marginal D_X and the $D_{+1}(y|x)$ (resp. $D_{-1}(y|x)$) described above.

Here, we can think of the observations O are the training examples S, where given i, S is drawn from D_i^m (m iid draws from distribution D_i).

Lemma 4. *Suppose training sample size $m \leq \frac{1}{8\epsilon^2} \ln \frac{1}{16\delta}$. Then, there exists $i \in \{-1, +1\}$ such that*

$$\mathbb{P}_i\left(\mathrm{err}(\hat{h}, D_i) - \min_{h \in \mathcal{H}} \mathrm{err}(h, D_i)\right) > \delta.$$

Proof. We show the lemma in two steps.

Step 1: reducing PAC learning to hypothesis testing. $\hat{h}$ induces a "guess" on the hypothesis index i, that is,

$$\hat{i} = \hat{h}(z_0).$$

Note that as $\hat{h} = \mathcal{A}(S)$ is a function of training examples S, $\hat{i}$ can also be written as a function of S - we use f to denote that function.

For a classifier h, it is easy to see its error rate on D_i is: $\mathrm{err}(h, D_i) = \frac{1}{2} - \epsilon + 2\epsilon \mathbf{1}(h(z_0) \neq i)$. In addition, under D_i, $\min_{h \in \mathcal{H}} \mathrm{err}(h, D_i) = \frac{1}{2} - \epsilon$, attained by a classifier $h \in \mathcal{H}$ such that $h(z_0) = i$. This implies the following relationship on the events:

$$\Big\{ f(S) \neq i \Big\} = \Big\{ \hat{h}(z_0) \neq i \Big\} \subseteq \Big\{ \mathrm{err}(\hat{h}, D_i) - \min_{h \in \mathcal{H}} \mathrm{err}(h, D_i) > \epsilon \Big\}$$

So proving the lemma reduces to showing that for at least one i in $\{\pm 1\}$, $\mathbb{P}_i(f(S) \neq i) > \delta$, as this would immediately imply $\mathbb{P}_i(\mathrm{err}(\hat{h}, D_i) - \min_{h \in \mathcal{H}} \mathrm{err}(h, D_i) > \epsilon) \geq \mathbb{P}_i(f(S) \neq i) > \delta$.

Step 2: applying Le Cam's method. Invoking Lemma 3, we have that there exists i, $\mathbb{P}_i(\hat{I} \neq i) \geq \frac{1}{2} \| P_{-1} \wedge P_{+1} \|_1$. We shall show a lower bound on the right hand side.

$$\begin{aligned} \| P_{-1} \wedge P_{+1} \|_1 &= \frac{1}{2} \sum_{o \in \mathcal{O}} \min(P_{-1}(o), P_{+1}(o)) \\ &= \frac{1}{2} \sum_{S \in (\{z_0\} \times \{\pm 1\})^n} \min\big(P_{-1}(S), P_{+1}(S)\big) \end{aligned} \tag{2}$$

Step 3: reducing distribution similarity to binomial tail lower bound. Given a set $S = (z_0, y_1), \ldots, (z_0, y_m)$, how shall we reason about $P_{-1}(S)$, the probability of seeing dataset S when examples from S are drawn iid from D_{-1}? Denote by $m_+(S)$ the number of $+1$'s in y. Then,

$$P_{-1}(S) = \left(\frac{1}{2} - \epsilon \right)^{m_+(S)} \left(\frac{1}{2} + \epsilon \right)^{m - m_+(S)}.$$

Symmetrically,

$$P_{+1}(S) = \left(\frac{1}{2} + \epsilon \right)^{m_+(S)} \left(\frac{1}{2} - \epsilon \right)^{m - m_+(S)}.$$

Therefore, $P_{+1}(S) \geq P_{-1}(S)$ iff $n_+(S) \geq \frac{n}{2}$. Therefore, the right hand side of Equation (2) can be written as:

$$\begin{aligned} & \frac{1}{2} \left(\sum_{S : m_+(S) \geq \frac{m}{2}} P_{-1}(S) + \sum_{S : m_+(S) < \frac{m}{2}} P_{+1}(S) \right) \\ = \; & \frac{1}{2} \left(\mathbb{P}_{-1}(m_+(S) \geq \frac{m}{2}) + \mathbb{P}_{+1}(m_+(S) < \frac{m}{2}) \right) \\ \geq \; & \frac{1}{2} \mathbb{P}_{-1}(m_+(S) \geq \frac{m}{2}). \end{aligned} \tag{3}$$

Now, let us look closely at the probability that $\mathbb{P}_{-1}(m_+(S) \geq \frac{m}{2})$. It can be seen that under P_{-1}, $m_+(S)$ is the sum of m iid Bernoulli($\frac{1}{2} - \epsilon$) random variables (i.e. binomial distribution with m trials and success probability $\frac{1}{2} - \epsilon$). Our task is to lower bound its right tail probability, that is, the probability the empirical mean exceeds $\frac{1}{2}$. [2]

We invoke Slud's Inequality from probability theory:

Fact 1. *Suppose* $X \sim \mathrm{B}(n, \frac{1}{2} - \epsilon)$. *Then,*

$$\mathbb{P}(X \geq \frac{n}{2}) \geq \frac{1}{2} (1 - \sqrt{1 - \exp\Big\{ - \frac{4n\epsilon^2}{1 - 4\epsilon^2} \Big\}}).$$

[2] This is an anti-concentration result, in contrast to the concentration inequalities we have shown in the first few lectures.

Continuing Equation (3), with the choice of $m \le \frac{1}{8\epsilon^2} \ln \frac{1}{16\delta}$, we have that $\exp\left\{-\frac{4m\epsilon^2}{1-4\epsilon^2}\right\}$ is at least 16δ, therefore, Slud's Inequality implies that the right hand side of Equation (3) is lower bounded by

$$\begin{aligned}\frac{1}{4}(1-\sqrt{1-\exp\left\{-\frac{4m\epsilon^2}{1-4\epsilon^2}\right\}}) &\ge \frac{1}{4}(1-\sqrt{1-16\delta}) \\ &\ge \frac{1}{4}(1-\sqrt{(1-8\delta)^2}) \\ &\ge \frac{1}{4}\cdot 8\delta > \delta.\end{aligned}$$

This concludes the proof of the lemma. □

2 Proof of Lemma 2: Assouad's method

Assouad's method is a generalization of Le Cam's method, showing information-theoretic lower bounds on testing more than two hypotheses. Suppose we are given 2^d possible distributions $P_\tau, \tau \in \{\pm 1\}^d$ over the observation space $\mathcal{O}$ (where each draw from the distribution results in an observation O in $\mathcal{O}$). Our task is to guess the identity of τ given O. Different from the last section where we are concerned with the probability that our guess $\hat{\tau}$ does not agree with the true τ, here we assign a *loss function* measuring the difference between $\hat{\tau}$ and τ:

$$\ell(\hat{\tau}, \tau) = \sum_{j=1}^{d} \mathbf{1}(\hat{\tau}_j \neq \tau_j).$$

Here we use the Hamming loss, which counts the number of coordinates the two vectors differ.

We would like to show that if the P_τ's are close to each other (in certain sense), then for any tester f there exists at least one τ such that under P_τ, the expectation of $\ell(\hat{\tau}, \tau)$ will be large.

We call $\tau \overset{j}{\sim} \tau'$ if τ and τ' only differ in their j-th coordinate, and call $\tau \sim \tau'$ if τ and τ' only differ in one coordinate.

Similar to Le Cam's method, we will use the shorthand that $\mathbb{P}_\tau$ (resp. $\mathbb{E}_\tau$) denotes $\mathbb{P}_{O\sim P_\tau}$ (resp. $\mathbb{E}_{O\sim P_\tau}$).

Lemma 5 (Assouad's method)**.** *For any collection of functions $f = (f_1, \ldots, f_d)$, $f_i : \mathcal{O} \to \{\pm 1\}$, there exists at least one τ in $\{\pm 1\}^d$, such that*

$$\mathbb{E}_\tau \ell(f(O), \tau) \ge \frac{d}{2} \cdot \min_{\tau, \tau' : \tau \sim \tau'} \|P_\tau \wedge P'_\tau\|_1.$$

We defer the proof to the end of this section. We now discuss the implication of this lemma to agnostic PAC learning.

The construction. As $\mathrm{VC}(\mathcal{H}) = d$, we can find d examples that $z_1, \ldots, z_d$ that are shattered by $\mathcal{H}$. That is, for any $\tau \in \{\pm 1\}^d$, there exists a h_τ in $\mathcal{H}$ such that $(h(z_1), \ldots, h(z_d)) = \tau$.

Define an unlabeled distribution D_X as uniform over $\{z_1, \ldots, z_d\}$. For $\tau \in \{\pm 1\}^d$, define

$$D_\tau(y|z_i) = \begin{cases} \frac{1}{2} + 2\tau_i\epsilon, & y = +1, \\ \frac{1}{2} - 2\tau_i\epsilon, & y = -1. \end{cases}$$

In other words,

$$D_\tau(y|z_i) = \begin{cases} \frac{1}{2} + 2\epsilon, & y = \tau_i, \\ \frac{1}{2} - 2\epsilon, & y = \tau_i. \end{cases}$$

For every $\tau \in \{\pm 1\}^d$, D_τ is specified by the marginal D_X and the $D_\tau(y|x)$ described above.

Lemma 6. *Suppose training sample size $m \le \frac{d}{1600\epsilon^2}$. Then, there exists $\tau \in \{-1,+1\}^d$ such that*

$$\mathbb{P}_\tau \left(\mathrm{err}(\hat{h}, D_\tau) - \min_{h \in \mathcal{H}} \mathrm{err}(h, D_\tau) \right) > \frac{1}{4}.$$

Proof. We prove the lemma in multiple steps.

Step 1: reducing PAC learning to hypothesis testing. Suppose the learner outputs a classifier $\hat{h} = \mathcal{A}(S)$. We can convert $\hat{h}$ to a hypothesis tester $\hat{\tau} = (h(z_1), \ldots, h(z_d))$. Note that $\hat{\tau}$ can be written as $f(S)$ for some function f. We observe that under distribution D_τ, the error of a classifier h is

$$\begin{aligned} \mathrm{err}(h, D_\tau) &= \sum_{j=1}^{d} D_\tau(z_j) \left(D_\tau(\tau_j | z_j) \mathbf{1}(h(z_j) \neq \tau_j) + D_\tau(-\tau_j | z_j) \mathbf{1}(h(z_j) = \tau_j) \right) \\ &= (\frac{1}{2} - 2\epsilon) + \frac{4\epsilon}{d} \cdot \sum_{j=1}^{d} \mathbf{1}(h(z_j) \neq \tau_j). \end{aligned}$$

Therefore, the optimal classifier h in $\mathcal{H}$ under D_τ is h_τ, which has an error rate of $\frac{1}{2} - 2\epsilon$. Moreover, for general classifier h, we have the following relationship between its excess error and the Hamming loss of its corresponding hypothesis tester:

$$\mathrm{err}(h, D_\tau) - \min_{h \in \mathcal{H}} \mathrm{err}(h, D_\tau) = \frac{4\epsilon}{d} \ell(\hat{\tau}, \tau). \tag{4}$$

Step 2: Applying Assouad's method. By Lemma 5 (recall that $\hat{\tau}$ can be written as $f(S)$ for some function f), along with Equation (4), there exists a τ in $\{\pm 1\}^d$, such that

$$\mathbb{E}_\tau[\mathrm{err}(\hat{h}, D_\tau) - \min_{h \in \mathcal{H}} \mathrm{err}(h, D_\tau)] \ge 2\epsilon \min_{\tau, \tau' : \tau \sim \tau'} \| P_\tau \wedge P'_\tau \|_1. \tag{5}$$

Now the task comes down to lower bounding $\| P_\tau \wedge P'_\tau \|_1$ for all neighboring pairs τ and τ'.

Step 3: Bounding the ℓ_1 distance using KL divergence. For a neighboring pair τ and τ', suppose they differ at coordinate j. What can we say about $\| P_\tau \wedge P'_\tau \|_1$? We first recall that

$$\| P_\tau \wedge P'_\tau \|_1 = 1 - \frac{1}{2} \| P_\tau - P_{\tau'} \|_1.$$

Now, recall that in the calibration exercise, we have shown that

$$\| P_\tau - P_{\tau'} \|_1 \le \sqrt{2 \, \mathrm{KL}(P_\tau, P_{\tau'})}.$$

Now, by Lemma 7 (as we will see shortly),

$$\mathrm{KL}(P_\tau, P_{\tau'}) \le \frac{48 m \epsilon^2}{d}.$$

With the choice of $m \le \frac{d}{1600\epsilon^2}$, we have that

$$\mathrm{KL}(P_\tau, P_{\tau'}) < \frac{1}{32},$$

which implies that

$$\| P_\tau \wedge P'_\tau \|_1 > 1 - \frac{1}{2} \cdot \frac{1}{4} = \frac{7}{8}.$$

THe above inequality, in conjunction with Equation (5), implies that

$$\mathbb{E}_\tau[\mathrm{err}(\hat{h}, D_\tau) - \min_{h \in \mathcal{H}} \mathrm{err}(h, D_\tau)] > \frac{7}{4}\epsilon. \tag{6}$$

Step 4: High expected error implies high error with decent probability. Now, define random variable $W \triangleq \mathrm{err}(\hat{h}, D_\tau) - \min_{h\in\mathcal{H}} \mathrm{err}(h, D_\tau)$. By Equation (4), W lies in $[0, 4\epsilon]$. Suppose for the sake of contradiction that $\mathbb{P}_\tau(W > \epsilon) \le \frac{1}{4}$, then

$$\begin{aligned}\mathbb{E}_\tau[W] &\le \mathbb{E}_\tau[W\mathbf{1}(W > \epsilon) + W\mathbf{1}(W \le \epsilon)] \\ &\le 4\epsilon\mathbb{P}_\tau(W > \epsilon) + \epsilon \cdot (1 - \mathbb{P}_\tau(W > \epsilon)) \\ &\le \epsilon + 3\epsilon\mathbb{P}_\tau(W > \epsilon) \le \frac{7}{4}\epsilon,\end{aligned}$$

contradition. Therefore, under P_τ, with probability $> \frac{1}{4}$, the excess error of $\hat{h}$ is at least ϵ. □

Lemma 7. *For τ and τ' in $\{\pm 1\}^d$ such that $\tau \sim \tau'$,*

$$\mathrm{KL}(P_\tau, P_{\tau'}) \le \frac{48m\epsilon^2}{d}.$$

Proof. Let us expand $\mathrm{KL}(P_\tau, P_{\tau'})$:

$$\begin{aligned}\mathrm{KL}(P_\tau, P_{\tau'}) &= \sum_{(x_1,y_1),\ldots,(x_m,y_m)\in(\{z_1,\ldots,z_d\}\times\{\pm1\})^m} P_\tau((x_1,y_1),\ldots,(x_m,y_m)) \ln \frac{P_\tau((x_1,y_1),\ldots,(x_m,y_m))}{P_{\tau'}((x_1,y_1),\ldots,(x_m,y_m))} \\ &= \sum_{(x_1,y_1),\ldots,(x_m,y_m)\in(\{z_1,\ldots,z_d\}\times\{\pm1\})^m} P_\tau((x_1,y_1),\ldots,(x_m,y_m)) \sum_{i=1}^m \ln \frac{D_\tau(x_i,y_i)}{D_{\tau'}(x_i,y_i)} \\ &= \mathbb{E}_{S\sim D_\tau^m}[\sum_{i=1}^m \ln \frac{D_\tau(X_i,Y_i)}{D_{\tau'}(X_i,Y_i)}] \\ &= m\mathbb{E}_{(X,Y)\sim D_\tau} \ln \frac{D_\tau(X,Y)}{D_{\tau'}(X,Y)} \\ &= m\,\mathrm{KL}(D_\tau, D_{\tau'}),\end{aligned}$$

where the first equality is from the definition of the KL divergence between two distributions; the second equality uses the fact that as the examples of S are independent, $P_\tau((x_1,y_1),\ldots,(x_m,y_m)) = \prod_{i=1}^m D_\tau(x_i,y_i)$; the third equality follows from viewing $\sum_{i=1}^m \ln \frac{D_\tau(x_i,y_i)}{D_{\tau'}(x_i,y_i)}$ as a function of $(x_1,y_1),\ldots,(x_m,y_m)$ and using the definition of expectation; the fourth equality is from linearity of expecation, and the fact that all (X_i,Y_i)'s come from the same distribution D_τ; the last inequality is again from the definition of KL divergence.

Note that $D_\tau(x,y)$ and $D_{\tau'}(x,y)$ only differs when $x = z_j$, specifically:

$$\ln \frac{D_\tau(x,y)}{D_\tau(x,y)} = \ln \frac{1/d \cdot D_\tau(y|x)}{1/d \cdot D_{\tau'}(y|x)} = \begin{cases} \ln \frac{1/2+2\epsilon}{1/2-2\epsilon}, & x = z_j, y = \tau_j \\ \ln \frac{1/2-2\epsilon}{1/2+2\epsilon}, & x = z_j, y = -\tau_j \\ 0, & x \ne z_j \end{cases}$$

Therefore,

$$\mathrm{KL}(D_\tau, D_{\tau'}) = \sum_{(x,y)} D_\tau(x,y) \ln \frac{D_\tau(x,y)}{D_{\tau'}(x,y)} = \frac{1}{d}\left(\frac{1}{2}+2\epsilon\right)\ln\frac{1/2+2\epsilon}{1/2-2\epsilon} + \left(\frac{1}{2}-2\epsilon\right)\ln\frac{1/2-2\epsilon}{1/2+2\epsilon} = \frac{1}{d}\,\mathrm{kl}\left(\frac{1}{2}+2\epsilon, \frac{1}{2}-2\epsilon\right).$$

The lemma is concluded in light of Lemma 8:

$$\mathrm{KL}(P_\tau, P_{\tau'}) = m\,\mathrm{KL}(D_\tau, D_{\tau'}) \le \frac{48m\epsilon^2}{d}.$$ □

Lemma 8. *For $\epsilon \in (0, \frac{1}{8})$, we have*

$$\mathrm{kl}\left(\frac{1}{2}+2\epsilon, \frac{1}{2}-2\epsilon\right) \leq 48\epsilon^2.$$

Proof. First, observe that

$$\mathrm{kl}\left(\frac{1}{2}+2\epsilon, \frac{1}{2}-2\epsilon\right) = \left(\frac{1}{2}+2\epsilon\right)\ln\frac{1/2+2\epsilon}{1/2-2\epsilon} + \left(\frac{1}{2}-2\epsilon\right)\ln\frac{1/2-2\epsilon}{1/2+2\epsilon} = 4\epsilon(\ln(1+4\epsilon) - \ln(1-4\epsilon).$$

Now, $\ln(1+4\epsilon) \leq 4\epsilon$. In addition,

$$-\ln(1-4\epsilon) = \sum_{i=1}^{\infty}\frac{(4\epsilon)^i}{i} \leq \sum_{i=1}^{\infty}(4\epsilon)^i = \frac{4\epsilon}{1-4\epsilon} \leq 8\epsilon.$$

The lemma follows by algebra. □

2.1 Proof of Lemma 5

For j in $\{1,\ldots,d\}$, define $P_{j,+}$ to be the uniform mixture of all P_τ's such that $\tau_j = 1$. Formally,

$$P_{j,+}(o) = \frac{1}{2^{d-1}}\sum_{\tau:\tau_j=+1} P_\tau(o).$$

Similarly, define $P_{j,-}$ as the uniform mixture of all P_τ's such that $\tau_j = -1$.

We first show the following simple lemma.

Lemma 9. *For every $jin\{1,\ldots,d\}$,*

$$\|P_{j,+} \wedge P_{j,-}\|_1 \geq \min_{\tau,\tau':\tau\sim\tau'} \|P_\tau \wedge P'_\tau\|.$$

Proof. Recall that $\|P_{j,+} \wedge P_{j,-}\|_1$ can be written in the following more intuitive form:

$$\|P_{j,+} \wedge P_{j,-}\|_1 = 1 - \frac{1}{2}\|P_{j,+} - P_{j,-}\|_1.$$

Now, denote by τ^j the vector that differs with τ at coordinate j, we have

$$\begin{aligned}
\|P_{j,+} - P_{j,-}\|_1 &= \|\frac{1}{2^{d-1}}(\sum_{\tau:\tau_j=+1} P_\tau - \sum_{\tau:\tau_j=-1} P_{\tau'})\|_1 \\
&= \|\frac{1}{2^{d-1}}(\sum_{\tau:\tau_j=+1} P_\tau - P_{\tau^j})\|_1 \\
&\leq \frac{1}{2^{d-1}}\sum_{\tau:\tau_j=+1}\|P_\tau - P_{\tau^j}\|_1 \\
&\leq \max_{\tau:\tau_j=+1}\|P_\tau - P_{\tau^j}\|_1 \\
&\leq \max_{\tau,\tau':\tau\sim\tau'}\|P_\tau - P_{\tau'}\|_1,
\end{aligned}$$

where the first inequality is from triangle inequality; the second inequality is by replacing each term with the max; the third inequality is from that $\tau \sim \tau^j$. Therefore,

$$\begin{aligned}\|P_{j,+} \wedge P_{j,-}\|_1 &\geq 1 - \frac{1}{2} \max_{\tau,\tau':\tau\sim\tau'} \|P_\tau - P_{\tau'}\|_1 \\ &= \min_{\tau,\tau':\tau\sim\tau'} (1 - \frac{1}{2}\|P_\tau - P_{\tau'}\|_1) \\ &= \min_{\tau,\tau':\tau\sim\tau'} \|P_\tau \wedge P_{\tau'}\|_1.\end{aligned}$$

□

Lemma 5 now follows straightforwardly. Consider a random index T drawn uniformly at random from $\{\pm 1\}^d$. We will show that f has a large expected loss. Specifically:

$$\begin{aligned}\mathbb{E}_{T\sim U(\{\pm1\}^d),O\sim P_T} \ell(f(O), T) &= \mathbb{E} \sum_{j=1}^{d} \mathbf{1}(f_j(O) \neq T_j) \\ &= \sum_{j=1}^{d} \mathbb{P}_{I\sim \mathrm{U}(\{\pm1\}),O\sim P_{j,I}}(f_j(O) \neq I) \\ &\geq \sum_{j=1}^{d} \frac{1}{2}\|P_{j,+} \wedge P_{j,-}\| \\ &\geq \frac{d}{2} \cdot \min_{\tau,\tau':\tau\sim\tau'} \|P_\tau \wedge P'_\tau\|,\end{aligned}$$

where the first equality is from the definition of ℓ; the second equality is from linearity of expectation, and the fact that we can alternatively view O as generated by the follwing process: first draw an $I \sim \mathrm{U}(\{\pm 1\})$, then draw O from $P_{j,I}$; the first inequality is from Le Cam's Lemma (Lemma 3); the second inequality is from Lemma 9.

Therefore, there exists at least one τ in $\{\pm 1\}^d$, such that

$$\mathbb{E}_\tau \ell(f(O), \tau) \geq \frac{d}{2} \cdot \min_{\tau,\tau':\tau\sim\tau'} \|P_\tau \wedge P'_\tau\|_1. \quad \square$$

References

[1] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[2] Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.