Lecture-14: Complexity theory based lower bounds

1 Lower bounds

So far we presented several upper bounds on the generalization error. In this lecture, we provide lower bounds on the generalization error of any learning algorithm in terms of the VC-dimension of the hypothesis set used. These lower bounds are shown by finding for any algorithm a *bad* distribution. In the context of the following proofs, first a lower bound is given on the expected error over the parameters defining the distributions. From that, the lower bound is shown to hold for at least one set of parameters, that is one distribution.

Definition 1.1. Consider an input space $\mathcal{X}$, binary label space $\mathcal{Y} \triangleq \{0,1\}$, a hypothesis set $H \subseteq \mathcal{Y}^{\mathcal{X}}$, a concept $c \in \mathcal{Y}^{\mathcal{X}}$, and loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \{0,1\}$ defined as $\ell(y,y') \triangleq \mathbb{1}_{\{y \neq y'\}}$ for all $y,y' \in \mathcal{Y}$. Let $h_x \in H$ be a hypothesis returned by a learning algorithm $\mathcal{A}$ for any unlabeled sample $x \in \mathcal{X}^m$. We assume that the test and training examples are sampled *i.i.d.* from the same distribution $D \in \mathcal{M}(\mathcal{X})$, then the generalization risk under algorithm $\mathcal{A}$ is defined for each $h^x \in H, f \in \mathcal{Y}^{\mathcal{X}}$ and distribution $D \in \mathcal{M}(\mathcal{X})$,

$$R_D(h^x,f) \triangleq \mathbb{E}[\ell(h^x(X),f(X)) \mid x] = \mathbb{E}_{X\sim D}\mathbb{1}_{\{h^x(X)\neq f(X)\}}.$$

Theorem 1.2 (Lower bound, realizable case). *Consider an input space $\mathcal{X}$, binary labels $\mathcal{Y} \triangleq \{0,1\}$, and hypothesis set $H \subseteq \mathcal{Y}^{\mathcal{X}}$ with $d \triangleq \text{VC-dim}(H) > 1$. Then, for any unlabeled finite sample $X : \Omega \in \mathcal{X}^m$ for $m \geqslant 1$ and any learning algorithm $\mathcal{A}$ that returns hypothesis $h^X \in H$, there exists a distribution $D \in \mathcal{M}(\mathcal{X})$ and a target function $f \in H$ such that*

$$P_{X\sim D^m}\left\{R_D(h^X,f) > \frac{d-1}{32m}\right\} \geqslant \frac{1}{100}.$$

Proof. We will show this in steps.

(a) **Construction of distribution $D \in \mathcal{M}(\mathcal{X})$.** Since $\text{VC-dim}(H) = d$, there exists a sample $\bar{x} \in \mathcal{X}^d$ that is shattered by H. We observe that all elements of this sequence are distinct and we can equivalently represent it by the set $\bar{\mathcal{X}} \triangleq \{\bar{x}_1,\dots,\bar{x}_d\} \subseteq \mathcal{X}$. Further, we have $H_{\bar{x}} \triangleq \{h_{\bar{x}} : h \in H\} = \mathcal{Y}^d$. That is, for any label sequence $y \in \mathcal{Y}^d$, there exists $h^y \in H$ such that $h^y_{\bar{x}} = y$ or equivalently $h^y(\bar{x}_i) = y_i$ for each $i \in [d]$. For any $\epsilon > 0$, we choose $D \in \mathcal{M}(\mathcal{X})$ such that the $\text{supp}(D) = \bar{\mathcal{X}}$ and so that one point $\bar{x}_d$ has very high probability $(1-8\epsilon)$, with the rest of the probability mass distributed uniformly among the other points, i.e. for any observation $W : \Omega \to \mathcal{X}$ with distribution D

$$D\{W = \bar{x}_d\} = 1 - 8\epsilon, \qquad D\{W = \bar{x}_i\} = \frac{8\epsilon}{d-1} \text{ for all } i \in [d-1].$$

(b) **Upper bounding risk under distribution D.** For any random sample $X : \Omega \to \mathcal{X}^m$ of size $m \geqslant 1$ generated *i.i.d.* for this distribution D, we have $D(\cap_{i=1}^m \{X_i \neq \bar{x}_d\}) = (8\epsilon)^m$, i.e. most samples would contain $\bar{x}_d$. Let h^X be the hypothesis returned by algorithm $\mathcal{A}$, then we can assume without loss of generality that h^X makes no error on $\bar{x}_d$. Thus, from the linearity of expectation,

$$R_D(h^X,f) = \mathbb{E}_{W\sim D}\mathbb{1}_{\{h^X(W)\neq f(W)\}} = \sum_{w\in\bar{\mathcal{X}}} \mathbb{1}_{\{h^X(w)\neq f(w)\}} D\{W=w\} \leqslant \sum_{w\in\bar{\mathcal{X}}\setminus\{\bar{x}_d\}} D\{W=w\} = 8\epsilon.$$

(c) **Lower bound on generalized risk under D.** Recall that $1 = \sum_{x\in\bar{\mathcal{X}}} \mathbb{1}_{\{W=x\}}$ almost surely for any random variable $W : \Omega \to \mathcal{X}$ sampled under $D \in \mathcal{M}(\mathcal{X})$. Thus, we can define

$$I(x) \triangleq \{i \in [m] : x \in \{\bar{x}_1,\dots,\bar{x}_{d-1}\}\}, \quad \bar{S}(x) \triangleq \{x_i : i \in I(x)\}, \quad \mathcal{S} \triangleq \{x \in \mathcal{X}^m : |I(x)| \leqslant (d-1)/2\}.$$

We observe that $\bar{S}(x) \subseteq \bar{\mathcal{X}} \setminus \{\bar{x}_d\}$, and the best that any algorithm $\mathcal{A}$ can do is return a hypothesis h^X that makes no error on *training set* $\bar{S}(X) \cup \{\bar{x}_d\}$, i.e. $h^X(w) = c(w)$ for all $w \in \bar{S}(X) \cup \{\bar{x}_d\}$. That is,

$$\mathbb{1}_{\{h^X(W)\neq c(W)\}} = \sum_{w\in\bar{\mathcal{X}}} \mathbb{1}_{\{W=w\}}\mathbb{1}_{\{h^X(w)\neq c(w)\}} \geqslant \sum_{w\in\bar{\mathcal{X}}\setminus(\bar{S}(X)\cup\{\bar{x}_d\})} \mathbb{1}_{\{W=w\}}\mathbb{1}_{\{h^X(w)\neq c(w)\}}. \tag{1}$$

We fix a sample $x \in \mathcal{S}$, and consider the uniform distribution $U \in \mathcal{M}(H_{\bar{x}})$. Applying expectation on both sides of (1), using the linearity of expectation, exchanging the order of expectations for random variables with finite support, non-negativity of indicator functions, and the definition of D, we get

$$\mathbb{E}_{h^Y \sim U} R_D(h^X, h^Y) = \mathbb{E}_{h^Y \sim U} \mathbb{E}_{W \sim D} \mathbb{1}_{\{h^X(W) \neq h^Y(W)\}} \geqslant \sum_{w \in \bar{\mathcal{X}} \setminus (\bar{S}(X) \cup \{\bar{x}_d\})} \mathbb{E}_{h^Y \sim U}[\mathbb{1}_{\{h^X(w) \neq h^Y(w)\}}] D_w$$
$$= \frac{1}{2} \sum_{w \in \bar{\mathcal{X}} \setminus (\bar{S}(X) \cup \{\bar{x}_d\})} D_w \geqslant \frac{1}{2} \frac{(d-1)}{2} \frac{8\epsilon}{(d-1)} = 2\epsilon.$$

That is, since $\bar{x}$ is shattered, algorithm $\mathcal{A}$ can essentially do no better than tossing a coin when determining the label of a point $\bar{x}_i$ not falling in the *training set* $\bar{S}(X)$.

(d) **Construction of a target function** $f \in H$**.** Since the above equation holds for all samples $x \in \mathcal{S}$, it also holds in expectation over all $X : \Omega \to \mathcal{S}$, i.e. $\mathbb{E}_{X \in \mathcal{S}} \mathbb{E}_{h^Y \sim U}[R_D(h^X, h^Y)] \geqslant 2\epsilon$. By Fubini's theorem, the order of expectations can be exchanged, and thus $\mathbb{E}_{h^Y \sim U}\left[\mathbb{E}_{X \in \mathcal{S}}[R_D(h^X, h^Y)]\right] \geqslant 2\epsilon$. This implies that $\mathbb{E}_{X \in \mathcal{S}}[R_D(h^X, f_0)] \geqslant 2\epsilon$ for at least one labeling $f_0 \in H$.

(e) **Lower bound on** $P_{X \sim D^m}\left\{R_D(h^X, f_0) \geqslant \epsilon\right\}$**.** Decomposing $\mathbb{E}_{X \in \mathcal{S}}[R_D(h^X, f_0)]$ into two parts, using upper bound $R_D(h^X, f_0) \leqslant 8\epsilon$ from part (b), and upper bounding an indicator by unity, we obtain

$$2\epsilon \leqslant \mathbb{E}_{X \in \mathcal{S}} R_D(h^X, f_0) = \mathbb{E}[R_D(h^X, f_0)(\mathbb{1}_{\{R_D(h^X, f_0) \geqslant \epsilon\}} + \mathbb{1}_{\{R_D(h^X, f_0) < \epsilon\}})] \leqslant 7\epsilon P_{X \in \mathcal{S}}\left\{R_D(h^X, f_0) \geqslant \epsilon\right\} + \epsilon.$$

Rearranging terms, we obtain $P_{X \in \mathcal{S}}\left\{R_D(h^X, f_0) \geqslant \epsilon\right\} \geqslant 1/7$. Thus, the probability over all samples $X \in \mathcal{X}^m$ can be lower bounded as

$$P_{X \sim D^m}\left\{R_D(h^X, f_0) \geqslant \epsilon\right\} \geqslant P_{X \in \mathcal{S}}\left\{R_D(h^X, f_0) \geqslant \epsilon\right\} P_{X \sim D^m}\left\{X \in \mathcal{S}\right\} \geqslant \frac{1}{7} P_{X \sim D^m}\left\{X \in \mathcal{S}\right\}.$$

(f) **Lower bound on** $P_{X \sim D^m}\left\{X \in \mathcal{S}\right\}$**.** For any sample $X \in \mathcal{X}^m$, we have $\left|\bar{S}(X)\right| = |I(X)| = \sum_{i=1}^m \mathbb{1}_{\{X_i \neq \bar{x}_d\}}$ almost surely under distribution D. Since $D_{\bar{x}_d} = 1 - 8\epsilon$, applying the multiplicative Chernoff bound for $\gamma = 1$ and $\epsilon = (d-1)/32m$, and using the fact that $d \geqslant 2$, we obtain an upper bound on the probability that more than $(d-1)/2$ points are drawn in a sample of size m

$$1 - P_{X \sim D^m}\left\{X \in \mathcal{S}\right\} = P\left\{\left|\bar{S}(X)\right| > (d-1)/2\right\} = P\left\{\frac{1}{m}\sum_{i=1}^m \mathbb{1}_{\{X_i \neq \bar{x}_d\}} \geqslant 8\epsilon(1+\gamma)\right\} \leqslant e^{-8\epsilon m \frac{\gamma^2}{3}} \leqslant e^{-1/12}.$$

We observe that $e^{-1/12} \leqslant 1 - 7\delta$ for $\delta \leqslant .01$, and hence $P_{X \sim D^m}\left\{X \in \mathcal{S}\right\} \geqslant 7\delta$.

Hence, we constructed a distribution D and hypothesis $f_0 \in H$ such that $P_{X \sim D^m}\left\{R_D(h^X, f_0) \geqslant \epsilon\right\} \geqslant \delta$. □

Remark 1. The theorem shows that for any algorithm $\mathcal{A}$, there exists a *bad* distribution over $\mathcal{X}$ and a target function f for which the error of the hypothesis returned by algorithm $\mathcal{A}$ is a constant times d/m with some constant probability. This further demonstrates the key role played by the VC-dim in learning. The result implies in particular that PAC-learning in the realizable case is not possible when the VC-dimension is infinite.

Remark 2. Note that the proof shows a stronger result than the statement of the theorem: the distribution D is selected independently of the algorithm $\mathcal{A}$.

A Concentration inequalities

Theorem A.1 (Sanov). *Consider* i.i.d. *random vector* $X : \Omega \to [0,1]^m$ *with common distribution* D *and mean* p*. Then, for any* $q \in [0,1]$*, the following inequality holds for* $\hat{p} = \frac{1}{m}\sum_{i=1}^m X_i$*,*

$$P\left\{\hat{p} \geqslant q\right\} \leqslant e^{-mD(q\|p)},$$

where $D(q\|p) \triangleq q \log \frac{q}{p} + (1-q) \log \frac{(1-q)}{(1-p)}$ *is the binary relative entropy of* p *and* q*.*

Proof. Let $t > 0$. By convexity of the function $x \mapsto e^{tx}$, the following inequality holds for all $x \in [0,1]$,

$$e^{tx} = e^{t[(1-x)0 + x1]} \leqslant 1 - x + xe^t.$$

In view of that, for any $t > 0$, we can write

$$P\{\hat{p} \geqslant q\} = P\left\{e^{tm\hat{p}} \geqslant e^{tmq}\right\} \leqslant e^{-tmq}\mathbb{E}e^{tm\hat{p}} = e^{-tmq}\prod_{i=1}^{m}\mathbb{E}e^{tX_i} \leqslant e^{-tmq}\prod_{i=1}^{m}\mathbb{E}(1 - X_i + X_i e^t) = e^{-tmq}(1 - p + pe^t)^m.$$

Now, the function $f : t \mapsto e^{-tq}(1 - p + pe^t) = (1 - p)e^{-tq} + pe^{t(1-q)}$ reaches its minimum at $t_* \triangleq \log \frac{q(1-p)}{p(1-q)}$. Plugging in this value of t in the previous inequality, we obtain

$$P\{\hat{p} \geqslant q\} \leqslant \inf_{t>0}\Big((1-p)e^{-tq} + pe^{t(1-q)}\Big)^m = ((1-p)e^{-t_*q} + pe^{t_*(1-q)})\Big)^m = (1-p)\frac{p(1-q)}{q(1-p)}^q + p$$

Plugging in this value of t in the inequality above yields $P\{\hat{p} \geqslant q\} \leqslant e^{-mD(q\|p)}$. □

Remark 3. Note that for any $0 < \epsilon \leqslant 1 - p$, with the choice $q = p + \epsilon$, the theorem implies

$$P\{\hat{p} \geqslant p + \epsilon\} \leqslant e^{-mD(p+\epsilon\|p)}.$$

This is a finer bound than Hoeffding's inequality since, by Pinsker's inequality $D(p + \epsilon\|p) \geqslant \frac{1}{2}(2\epsilon)^2 = 2\epsilon^2$. Similarly, we can derive a symmetric bound by applying the theorem to the random variables $Y_i \triangleq 1 - X_i$. Then, for any $0 < \epsilon \leqslant p$, with the choice $q = p - \epsilon$, the theorem implies

$$P\{\hat{p} \leqslant p - \epsilon\} \leqslant e^{-mD(p-\epsilon\|p)}.$$

Theorem A.2 (Multiplicative Chernoff bounds). *Consider an* i.i.d. *random vector $X : \Omega \to \mathcal{X}^m$ with unknown distribution D, mean p and support $\mathcal{X} \subseteq [0,1]$. Then, for any $\gamma \in [0, \frac{1}{p} - 1]$, the following inequality holds for $\hat{p} = \frac{1}{m}\sum_{i=1}^{m} X_i$,*

$$P\{\hat{p} \geqslant (1+\gamma)p\} \leqslant e^{-mp\frac{\gamma^2}{3}}, \qquad P\{\hat{p} \leqslant (1-\gamma)p\} \leqslant e^{-mp\frac{\gamma^2}{2}}.$$

Proof. The proof consists of deriving in each case a finer lower bound for the binary relative entropy than Pinsker's inequality. Using the inequalities $\log(1 + x) \geqslant \frac{x}{1+\frac{x}{2}}$ and $\log(1 + x) < x$, we can write

$$-D((1+\gamma)p\|p) \leqslant (1+\gamma)p\frac{-\gamma}{1+\gamma/2} + (1-p-\gamma p)\frac{\gamma p}{1-p-\gamma p} = \gamma p\Big(1 - \frac{1+\gamma}{1+\gamma/2}\Big) = -\frac{\gamma^2 p}{2+\gamma} \leqslant -\frac{\gamma^2 p}{3}.$$

Similarly, using the inequalities $(1 - x)\log(1 - x) \geqslant -x + \frac{x^2}{2}$ valid for $x \in (0,1)$ and $\log(1 - x) < -x$,

$$-D((1-\gamma)p\|p) \leqslant p(\gamma - \gamma^2/2) - \gamma p = -\frac{\gamma^2 p}{2}.$$

□