Formulation of the learning problem, Part 2

Maxim Raginsky

September 11, 2014

1 Agnostic (or model-free) learning

The realizable setting we have focused on in the last lecture rests on certain assumptions, which are not always warranted:

- The assumption that the target concept C^* belongs to $\mathscr{C}$ (or that the target function f^* belongs to $\mathscr{F}$) means that we are trying to fit a hypothesis to data, which are *a priori* known to have been generated by some member of the model class defined by $\mathscr{C}$ (or by $\mathscr{F}$). However, in general we may not want to (or be able to) assume much about the data generation process, and instead would like to find the best fit to the data at hand using an element of some model class of our choice.
- The assumption that the training features (or inputs) are labelled noiselessly by $\mathbf{1}_{\{x \in C^*\}}$ (or by $f(x)$) rules out the possibility of noisy measurements or observations.
- Finally, even if the above assumption were true, we would not necessarily have *a priori* knowledge of the concept class $\mathscr{C}$ (or the function class $\mathscr{F}$) containing the target concept (or function). In that case, the best we could hope for is to pick our own model class and seek the best *approximation* to the unknown target concept (or function) among the elements of that class.

The *model-free learning problem* (also referred to as the *agnostic case*), introduced by Haussler [Hau92], takes a more general decision-theoretic approach and removes the above restrictions. It has the following ingredients:

- Sets X, Y, and U
- A class $\mathscr{P}$ of probability distributions on $\mathsf{Z} \triangleq \mathsf{X} \times \mathsf{Y}$
- A class $\mathscr{F}$ of functions $f : \mathsf{X} \to \mathsf{U}$ (the *hypothesis space*)
- A *loss function* $\ell : \mathsf{Y} \times \mathsf{U} \to [0,1]$

The learning process takes place as follows. We obtain an i.i.d. sample $Z^n = (Z_1, \ldots, Z_n)$, where each $Z_i = (X_i, Y_i)$ is drawn from the same fixed but unknown $P \in \mathscr{P}$. A *learning algorithm* is a sequence $\mathscr{A} = \{A_n\}_{n=1}^{\infty}$ of mappings

$$A_n : \mathsf{Z}^n \to \mathscr{F}.$$

As before, let

$$\widehat{f}_n = A_n(Z^n) = A_n(Z_1, \ldots, Z_n) = A_n((X_1, Y_1), \ldots, (X_n, Y_n)).$$

This is the hypothesis emitted by the learning algorithm based on the *training data* Z^n. Note that, by definition, $\widehat{f}_n$ is a *random element* of the hypothesis space $\mathscr{F}$, and that it maps each point $x \in \mathsf{X}$ to a point $u = \widehat{f}_n(x) \in \mathsf{U}$. Following the same steps as in the realizable case, we evaluate the goodness of $\widehat{f}_n$ by its expected loss

$$L_P(\widehat{f}_n) \triangleq \mathbb{E}_P\big[\ell(Y, \widehat{f}_n(X))|Z^n\big] = \int_{\mathsf{X}\times\mathsf{Y}} \ell(y, \widehat{f}_n(x))P(\mathrm{d}x, \mathrm{d}y),$$

where the expectation is w.r.t. a random couple $(X, Y) \in \mathsf{Z}$ drawn according to the same P but independently of Z^n. Note that $L_P(\widehat{f}_n)$ is a random variable since so is $\widehat{f}_n$. In general, we can define the expected risk w.r.t. P for every f in our hypothesis space by

$$L_P(f) \triangleq \mathbb{E}_P[\ell(Y, f(X))] = \int_{\mathsf{X}\times\mathsf{Y}} \ell(y, f(x))P(\mathrm{d}x, \mathrm{d}y)$$

as well as the *minimum risk*

$$L_P^*(\mathscr{F}) \triangleq \inf_{f\in\mathscr{F}} L_P(f).$$

Conceptually, $L_P^*(\mathscr{F})$ is the best *possible* performance of any hypothesis in $\mathscr{F}$ when the samples are drawn from P; similarly, $L_P(\widehat{f}_n)$ is the *actual* performance of the algorithm with access to a training sample of size n. It is clear from definitions that

$$0 \le L_P^*(\mathscr{F}) \le L_P(\widehat{f}_n) \le 1.$$

The goal of learning is to guarantee that $L_P(\widehat{f}_n)$ is as close as possible to $L_P^*(\mathscr{F})$, whatever the true $P \in \mathscr{P}$ happens to be. In order to speak about this quantitatively, we need to assess the probability of getting a "bad" sample. To that end, we define, similarly to what we have done earlier, the quantity

$$r_{\mathscr{A}}(n, \varepsilon) \triangleq \sup_{P\in\mathscr{P}} P^n\left(Z^n \in \mathsf{Z}^n : L_P(\widehat{f}_n) \ge L_P^*(\mathscr{F}) + \varepsilon\right) \tag{1}$$

for every $\varepsilon > 0$. Thus, a sample $Z^n \sim P^n$ is declared to be "bad" if it leads to a hypothesis whose expected risk on an independent test point $(X, Y) \sim P$ is greater than the smallest possible loss $L_P^*(\mathscr{F})$ by at least ε. We have the following:

Definition 1. *We say that a learning algorithm for a problem* $(\mathsf{X}, \mathsf{Y}, \mathsf{U}, \mathscr{P}, \mathscr{F}, \ell)$ *is* PAC to accuracy ε *if*

$$\lim_{n\to\infty} r_{\mathscr{A}}(n, \varepsilon) = 0.$$

An algorithm that is PAC to accuracy ε *for every* $\varepsilon > 0$ *is said to be PAC. A learning problem specified by a tuple* $(\mathsf{X}, \mathsf{Y}, \mathsf{U}, \mathscr{P}, \mathscr{F}, \ell)$ *is* model-free *(or* agnostically*)* learnable *(to accuracy* ε*) if there exists an algorithm for it which is PAC (to accuracy* ε*).*

Let us look at some examples.

1.1 Function learning in the realizable case

First we show that the model-free framework contains the realizable set-up as a special case. To see this, let X be an arbitrary space and let $\mathsf{Y} = \mathsf{U} = [0, 1]$. Let $\mathscr{F}$ be a class of functions $f : \mathsf{X} \to [0, 1]$. Let $\mathscr{P}_{\mathsf{X}}$ be a family of probability distributions P_X on X. To each P_X and each $f \in \mathscr{F}$ associate a probability

distribution $P_{X,f}$ on $\mathsf{X} \times \mathsf{Y}$ as follows: let $X \sim P_X$, and let the conditional distribution of Y given $X = x$ be given by

$$P_{Y|X,f}(B|X = x) = \mathbf{1}_{\{f(x) \in B\}}$$

for all (measurable) sets $B \subseteq \mathsf{Y}$. The resulting joint distribution $P_{X,f}$ is then uniquely defined by its action on the "rectangles" $A \times B$, $A \subseteq \mathsf{X}$ and $B \subseteq \mathsf{Y}$:

$$P_{X,f}(A \times B) \triangleq \int_A P_{Y|X,f}(B|x) P_X(\mathrm{d}x) = \int_A \mathbf{1}_{\{f(x) \in B\}} P_X(\mathrm{d}x)$$

Finally, let $\mathscr{P} = \{P_{X,f} : f \in \mathscr{F}, P_X \in \mathscr{P}_{\mathsf{X}}\}$. Finally, let $\ell(y,u) \triangleq |y-u|^2$.

Now, fixing a probability distribution $P \in \mathscr{P}$ is equivalent to fixing some $P_X \in \mathscr{P}_{\mathsf{X}}$ and some $f \in \mathscr{F}$. A random element of $\mathsf{Z} = \mathsf{X} \times \mathsf{Y}$ drawn according to such a P has the form $(X, f(X))$, where $X \sim P_X$. An i.i.d. sequence $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn according to P therefore has the form

$$(X_1, f(X_1)), \ldots, (X_n, f(X_n)),$$

which is precisely what we had in our discussion of function learning in the realizable case. Next, for any $P = P_{X,f} \in \mathscr{P}$ and any other $g \in \mathscr{F}$, we have

$$\begin{aligned} L_{P_{X,f}}(g) &= \int_{\mathsf{X} \times \mathsf{Y}} |y - g(x)|^2 P_{X,f}(\mathrm{d}x, \mathrm{d}y) \\ &= \int_{\mathsf{X} \times \mathsf{Y}} \mathbf{1}_{\{y = f(x)\}} |y - g(x)|^2 P_X(\mathrm{d}x) \\ &= \int_{\mathsf{X}} |f(x) - g(x)|^2 P_X(\mathrm{d}x) \\ &= \|f - g\|^2_{L^2(P_X)}, \end{aligned}$$

which is precisely the risk $L_{P_X}(g, f)$ that we have considered in our function learning formulation earlier. Moreover,

$$L^*_{P_{X,f}} = \inf_{g \in \mathscr{F}} L_{P_{X,f}}(g) = \inf_{g \in \mathscr{F}} \|f - g\|^2_{L^2(P_X)} \equiv 0.$$

Therefore,

$$\begin{aligned} r_{\mathscr{A}}(n, \varepsilon) &= \sup_{P_{X,f} \in \mathscr{P}} P^n_{X,f}\left(Z^n \in \mathsf{Z}^n : L_{P_{X,f}}(\widehat{f}_n) \geq L^*_{P_{X,f}} + \varepsilon\right) \\ &= \sup_{P_X \in \mathscr{P}_{\mathsf{X}}} \sup_{f \in \mathscr{F}} P^n_X\left(X^n \in \mathsf{X}^n : L_P(\widehat{f}_n, f) \geq \varepsilon\right) \\ &\equiv \bar{r}_{\mathscr{A}}(n, \varepsilon, \mathscr{P}_{\mathsf{X}}). \end{aligned}$$

Thus, the function learning problem in the realizable case can be covered under the model-free framework as well.

1.2 Learning to classify with noisy labels

Consider the concept learning problem in the realizable case, except that now the labels Y_i, which in the original problem had the form $\mathbf{1}_{\{X_i \in C^*\}}$ for some target concept C^*, are *noisy*. That is, if X_i is a training feature point, then the label Y_i may be "flipped" due to chance, independently of all other X_j's, $j \neq i$.

The precise formulation of this problem is as follows. Let X be a given feature space, let $\mathscr{C}$ be a concept class on it, and let $\mathscr{P}_{\mathsf{X}}$ be a class of probability distributions on X. Suppose that Nature picks some distribution $P_X \in \mathscr{P}_{\mathsf{X}}$ of the features and some target concept $C^* \in \mathscr{C}$. The training data are generated as follows. First, an i.i.d. sample $X^n = (X_1, \ldots, X_n)$ is drawn according to some $P_X \in \mathscr{P}_{\mathsf{X}}$. Then the corresponding labels $Y_1, \ldots, Y_n \in \{0,1\}$ are generated as follows:

$$Y_i = \begin{cases} \mathbf{1}_{\{X_i \in C^*\}}, & \text{with probability } 1-\eta \\ 1 - \mathbf{1}_{\{X_i \in C^*\}}, & \text{with probability } \eta \end{cases} \qquad \text{independently of } X^n, \{Y_j\}_{j \neq i}$$

where $\eta < 1/2$ is the *classification noise rate*.

To cast this problem into the model-free framework, let $\mathsf{Y} = \mathsf{U} = \{0,1\}$, let $\mathscr{F} = \{I_C : C \in \mathscr{C}\}$, and let $\ell(y,u) = |y-u|^2$. Define a class $\mathscr{P}$ of probability distributions $\{P_{X,C} : P_X \in \mathscr{P}_{\mathsf{X}}, C \in \mathscr{C}\}$ on $\mathsf{X} \times \mathsf{Y} = \mathsf{X} \times \{0,1\}$ as follows. Let $X \sim P_X$, and for a given $C \in \mathscr{C}$ consider the conditional probability of $Y = 1$ given $X = x$. If $x \in C$, then $Y = 1$ if and only if there was no error in the label; on the other hand, if $x \notin C$, then $Y = 1$ if and only if there was an error. That is,

$$\begin{aligned} P_{Y|X,C}(1|X=x) &= (1-\eta)\mathbf{1}_{\{x \in C\}} + \eta \mathbf{1}_{\{x \in C^c\}} \\ &= (1-\eta)\mathbf{1}_{\{x \in C\}} + \eta(1 - \mathbf{1}_{\{x \in C\}}); \\ P_{Y|X,C}(0|X=x) &= 1 - P_{Y|X,C}(1|X=x) \\ &= \eta \mathbf{1}_{\{x \in C\}} + (1-\eta)(1 - \mathbf{1}_{\{x \in C\}}). \end{aligned}$$

Then for any measurable set $A \subseteq \mathsf{X}$ we will have

$$\begin{aligned} P_{X,C}(A \times \{1\}) &= \int_A P_{Y|X,C}(1|X=x) P_X(\mathrm{d}x) \\ &= \int_A \left[(1-\eta)\mathbf{1}_{\{x \in C\}} + \eta(1 - \mathbf{1}_{\{x \in C\}}) \right] P_X(\mathrm{d}x) \\ &= (1-\eta) \int_A \mathbf{1}_{\{x \in C\}} P_X(\mathrm{d}x) + \eta \int_A P_X(\mathrm{d}x) - \eta \int_A \mathbf{1}_{\{x \in C\}} P_X(\mathrm{d}x) \\ &= \eta P_X(A) + (1-2\eta) P_X(A \cap C) \end{aligned} \tag{2}$$

and similarly

$$P_{X,C}(A \times \{0\}) = (1-\eta) P_X(A) - (1-2\eta) P_X(A \cap C). \tag{3}$$

Given a hypothesis $f = I_{C'} \in \mathscr{F}$, we have

$$L_{P_{X,C}}(I_{C'}) = \int_{\mathsf{X} \times \mathsf{Y}} |y - I_{C'}(x)|^2 P_{X,C}(\mathrm{d}x, \mathrm{d}y).$$

Computing this integral is straightforward but tedious. We start by expanding it as follows:

$$\begin{aligned} &\int_{\mathsf{X} \times \mathsf{Y}} |y - I_{C'}(x)|^2 P_{X,C}(\mathrm{d}x, \mathrm{d}y) \\ &= \int_{\mathsf{X}} |0 - I_{C'}(x)|^2 P_{X,C}(\mathrm{d}x \times \{0\}) + \int_{\mathsf{X}} |1 - I_{C'}(x)|^2 P_{X,C}(\mathrm{d}x \times \{1\}) \\ &= \int_{\mathsf{X}} \mathbf{1}_{\{x \in C'\}} P_{X,C}(\mathrm{d}x \times \{0\}) + \int_{\mathsf{X}} \mathbf{1}_{\{x \in (C')^c\}} P_{X,C}(\mathrm{d}x \times \{1\}) \\ &= P_{X,C}(C' \times \{0\}) + P_{X,C}((C')^c \times \{1\}). \end{aligned}$$

Substituting the expressions (2) and (3) into the above, we get

$$\begin{aligned}
L_{P_{X,C}}(I_{C'}) &= (1-\eta)P_X(C') - (1-2\eta)P_X(C\cap C') + \eta P_X((C')^c) + (1-2\eta)P_X(C\cap (C')^c) \\
&= (1-\eta)(\underbrace{P_X(C\cap C') + P_X(C^c\cap C')}_{P_X(C')}) - (1-2\eta)P_X(C\cap C') \\
&\qquad + \eta(\underbrace{P_X(C\cap (C')^c) + P_X(C^c\cap (C')^c)}_{P_X((C')^c)}) + (1-2\eta)P_X(C\cap (C')^c) \\
&= (1-\eta)(\underbrace{P_X(C\cap (C')^c) + P_X(C^c\cap C')}_{P_X(C\triangle C')}) + \eta P_X(C\cap C') + \eta\underbrace{P(C^c\cap (C')^c)}_{P_X((C\cup C')^c)} \\
&= (1-\eta)P_X(C\triangle C') + \eta P_X(C\cap C') + \eta(1-P_X(C\cup C')) \\
&= (1-\eta)P_X(C\triangle C') + \eta - \eta(\underbrace{P_X(C\cup C') - P_X(C\cap C')}_{P_X(C\triangle C')}) \\
&= \eta + (1-2\eta)P_X(C\triangle C') \\
&\equiv \eta + (1-2\eta)L_{P_X}(C', C).
\end{aligned}$$

From this, we have

$$\begin{aligned}
L^*_{P_{X,C}}(\mathscr{F}) &= \inf_{C'\in\mathscr{C}} L_{P_{X,C}}(I_{C'}) \\
&= \eta + (1-2\eta)\inf_{C'\in\mathscr{C}} P_X(C\triangle C') \\
&= \eta,
\end{aligned}$$

where the infimum is achieved by letting $C' = C$. From this it follows that

$$L_{P_{X,C}}(C') \geq L^*_{P_{X,C}} + \varepsilon \qquad \Longleftrightarrow \qquad P_{X,C}(C\triangle C') \geq \frac{\varepsilon}{1-2\eta}$$

In other words, learning a concept to accuracy ε with noise rate η is equivalent to learning a concept to accuracy $\varepsilon/(1-2\eta)$ in the noise-free case:

$$r_{\mathscr{A}}(n,\varepsilon) = \bar{r}\left(n, \frac{\varepsilon}{1-2\eta}, \mathscr{P}_{\mathsf{X}}\right).$$

2 Empirical risk minimization

Having formulated the model-free learning problem, we must now turn to the question of how to construct PAC learning algorithms (and the related question of when a hypothesis class is PAC-learnable in the model-free setting).

We will first start with a heuristic argument and then make it rigorous. Suppose we are faced with the learning problem specified by $(\mathsf{X}, \mathsf{Y}, \mathsf{U}, \mathscr{P}, \mathscr{F}, \ell)$. Given a training set $Z^n = (Z_1, \ldots, Z_n)$, where each $Z_i = (X_i, Y_i)$ is independently drawn according to some unknown $P \in \mathscr{P}$, what should we do? The first thing to note is that, for any hypothesis $f \in \mathscr{F}$, we can approximate its risk $L_P(f)$ by the *empirical risk*

$$\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)), \tag{4}$$

whose expectation w.r.t. the distribution of Z^n is clearly equal to $L_P(f)$. In fact, since ℓ is bounded between 0 and 1, Hoeffding's inequality tells us that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\ell(Y_i,f(X_i))-L_P(f)\right|<\varepsilon \qquad \text{with probability at least } 1-2e^{-2n\varepsilon^2}.$$

We can express these statements more succinctly if we define, for each $f\in\mathscr{F}$, the function $\ell_f:\mathsf{Z}\to[0,1]$ by

$$\ell_f(z)\equiv\ell_f(x,y)\triangleq\ell(y,f(x)). \tag{5}$$

Then the empirical risk (4) is just the expectation of ℓ_f w.r.t. the empirical distribution P_{Z^n}:

$$\mathsf{P}_{Z^n}(\ell_f)=\frac{1}{n}\sum_{i=1}^{n}\ell(Y_i,f(X_i)),$$

and, since $L_P(f)=\mathbb{E}_P[\ell(Y,f(X))]=P(\ell_f)$, we will have

$$\left|\mathsf{P}_{Z^n}(\ell_f)-P(\ell_f)\right|<\varepsilon \qquad \text{with probability at least } 1-2e^{-2n\varepsilon^2}. \tag{6}$$

Now, given the data Z^n we can compute the empirical risks $\mathsf{P}_{Z^n}(\ell_f)$ for every f in our hypothesis class $\mathscr{F}$. Since (6) holds for each $f\in\mathscr{F}$ individually, we may intuitively claim that the empirical risk for each f is a sufficiently accurate estimator of the corresponding true risk $L_P(f)\equiv P(\ell_f)$. Thus, a reasonable learning strategy would be to find any $\widehat{f}_n\in\mathscr{F}$ that would *minimize* the empirical risk, i.e., take

$$\widehat{f}_n=\operatorname*{arg\,min}_{f\in\mathscr{F}}\mathsf{P}_{Z^n}(\ell_f)=\operatorname*{arg\,min}_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\ell(Y_i,f(X_i)). \tag{7}$$

The reason why we would expect something like (7) to work is as follows: if a given f^* is a minimizer of $L_P(f)=P(\ell_f)$ over $\mathscr{F}$,

$$f^*=\operatorname*{arg\,min}_{f\in\mathscr{F}}P(\ell_f),$$

then its empirical risk, $\mathsf{P}_{Z^n}(f^*)$, will be close to $L_P(f^*)=P(\ell_{f^*})=L_P^*(\mathscr{F})$ with high probability. Moreover, it makes sense to expect that, in some sense, $\widehat{f}_n$ defined in (7) would be "close" to f^*, resulting in something like

$$P(\widehat{f}_n)\approx\mathsf{P}_{Z^n}(\widehat{f}_n)\approx\mathsf{P}_{Z^n}(f^*)\approx P(f^*)$$

with high probability.

Unfortunately, this is not true in general. However, as we will now see, it is true under certain regularity conditions on the objects $\mathscr{P}$, $\mathscr{F}$, and ℓ. In order to state these regularity conditions precisely, let us define the *induced loss function class*

$$\mathscr{L}_{\mathscr{F}}\triangleq\left\{\ell_f:f\in\mathscr{F}\right\}.$$

Each $\ell_f\in\mathscr{L}_{\mathscr{F}}$ corresponds to the hypothesis $f\in\mathscr{F}$ via (5). Now, for any $n\in\mathbb{N}$ and any $\varepsilon>0$ let us define

$$q(n,\varepsilon)\triangleq\sup_{P\in\mathscr{P}}P^n\left(Z^n\in\mathsf{Z}^n:\sup_{f\in\mathscr{F}}\left|\mathsf{P}_{Z^n}(\ell_f)-P(\ell_f)\right|\geq\varepsilon\right). \tag{8}$$

For a fixed $P \in \mathscr{P}$, quantity $\sup_{f\in\mathscr{F}} \left|\mathsf{P}_{Z^n}(\ell_f) - P(\ell_f)\right|$ is the *worst-case deviation* between the empirical means $\mathsf{P}_{Z^n}(\ell_f)$ and their expectations $P(\ell_f)$ over the entire hypothesis class $\mathscr{F}$. Given P, we say that an i.i.d. sample $Z^n \in \mathsf{Z}^n$ is "bad" if there exists at least one $f \in \mathscr{F}$, for which

$$\left|\mathsf{P}_{Z^n}(\ell_f) - P(\ell_f)\right| \geq \varepsilon.$$

Equivalently, a sample is bad if

$$\sup_{f\in\mathscr{F}} \left|\mathsf{P}_{Z^n}(\ell_f) - P(\ell_f)\right| \geq \varepsilon.$$

The quantity $q(n,\varepsilon)$ then compensates for the fact that P is unknown by considering the *worst case* over the entire class $\mathscr{P}$. With this in mind, we make the following definition:

Definition 2. *We say that the induced class $\mathscr{L}_{\mathscr{F}}$ has the* uniform convergence of empirical means (UCEM) *property w.r.t. $\mathscr{P}$ if*

$$\lim_{n\to\infty} q(n,\varepsilon) = 0$$

for every $\varepsilon > 0$.

Theorem 1. *If the induced class $\mathscr{L}_{\mathscr{F}}$ has the UCEM property, then the* empirical risk minimization (ERM) *algorithm of* (7) *is PAC.*

Proof. Fix $\varepsilon, \delta > 0$. We will now show that we can find a sufficiently large $n(\varepsilon,\delta)$, such that $r_{\mathscr{A}}(n,\varepsilon) < \delta$ for all $n \geq n(\varepsilon,\delta)$, where $r_{\mathscr{A}}(n,\varepsilon)$ is defined in (1).

Let $f^* \in \mathscr{F}$ minimize the true risk w.r.t. P, i.e., $P(f^*) = L_P^*(\mathscr{F})$. For any n, we have

$$\begin{aligned} L_P(\widehat{f}_n) - L_P^* &= P\big(\ell_{\widehat{f}_n}\big) - P\big(f^*\big) \\ &= \underbrace{P\big(\ell_{\widehat{f}_n}\big) - \mathsf{P}_{Z^n}\big(\ell_{\widehat{f}_n}\big)}_{T_1} + \underbrace{\mathsf{P}_{Z^n}\big(\ell_{\widehat{f}_n}\big) - \mathsf{P}_{Z^n}\big(\ell_{f^*}\big)}_{T_2} + \underbrace{\mathsf{P}_{Z^n}\big(\ell_{f^*}\big) - P\big(\ell_{f^*}\big)}_{T_3}, \end{aligned}$$

where in the second line we have added and subtracted $\mathsf{P}_{Z^n}\big(\ell_{\widehat{f}_n}\big)$ and $\mathsf{P}_{Z^n}\big(\ell_{f^*}\big)$. We will now analyze the behavior of the three terms, T_1, T_2, and T_3. Since $\widehat{f}_n$ minimizes the empirical risk $\mathsf{P}_{Z^n}(\ell_f)$ over all $f \in \mathscr{F}$, we will have

$$T_2 = \mathsf{P}_{Z^n}\big(\ell_{\widehat{f}_n}\big) - \mathsf{P}_{Z^n}\big(\ell_{f^*}\big) \leq 0.$$

Next,

$$T_1 = P\big(\ell_{\widehat{f}_n}\big) - \mathsf{P}_{Z^n}\big(\ell_{\widehat{f}_n}\big) \leq \sup_{f\in\mathscr{F}} \left[\mathsf{P}_{Z^n}(\ell_f) - P(\ell_f)\right] \leq \sup_{f\in\mathscr{F}} \left|\mathsf{P}_{Z^n}(\ell_f) - P(\ell_f)\right|,$$

and the same upper bound holds for T_3. Hence,

$$L_P(\widehat{f}_n) - L_P^*(\mathscr{F}) \leq 2 \sup_{f\in\mathscr{F}} \left|\mathsf{P}_{Z^n}(\ell_f) - P(\ell_f)\right|. \tag{9}$$

Now, since $\mathscr{L}_{\mathscr{F}}$ has the UCEM property, we can find some sufficiently large $n_0(\varepsilon,\delta)$, such that

$$q(n,\varepsilon/2) = \sup_{P\in\mathscr{P}} P^n\left(Z^n \in \mathsf{Z}^n : \sup_{f\in\mathscr{F}} \left|\mathsf{P}_{Z^n}(\ell_f) - P(\ell_f)\right| \geq \varepsilon/2\right) < \delta, \qquad \forall n \geq n_0(\varepsilon,\delta).$$

From this it follows that, for all $n \geq n_0(\varepsilon,\delta)$, we will have

$$P^n\left(Z^n : \sup_{f\in\mathscr{F}} \left|\mathsf{P}_{Z^n}(\ell_f) - P(\ell_f)\right| \geq \varepsilon/2\right) < \delta, \qquad \forall P \in \mathscr{P}.$$

From (9), we see that

$$L_P(\widehat{f}_n) \geq L_P^*(\mathscr{F}) + \varepsilon \qquad \Longrightarrow \qquad \sup_{f\in\mathscr{F}} \left|\mathsf{P}_{Z^n}(\ell_f) - P(\ell_f)\right| \geq \varepsilon/2$$

for all n. However, for all $n \geq n_0(\varepsilon,\delta)$ the latter event will occur with probability at most δ, no matter which P is in effect. Therefore, for all $n \geq n_0(\varepsilon,\delta)$ we will have

$$\begin{aligned} r_{\mathscr{A}}(n,\varepsilon) &= \sup_{P\in\mathscr{P}} P^n\left(Z^n : L_P(\widehat{f}_n) \geq L_P^*(\mathscr{F}) + \varepsilon\right) \\ &\leq \sup_{P\in\mathscr{P}} P^n\left(Z^n : \sup_{f\in\mathscr{F}} \left|\mathsf{P}_{Z^n}(\ell_f) - P(\ell_f)\right| \geq \varepsilon/2\right) \\ &\equiv q(n,\varepsilon/2) \\ &< \delta, \end{aligned}$$

which is precisely what we wanted to show. Thus, $r_{\mathscr{A}}(n,\varepsilon) \to 0$ as $n \to \infty$ for every $\varepsilon > 0$, which means that the ERM algorithm is PAC. □

This theorem shows that the UCEM property of the induced class $\mathscr{L}_{\mathscr{F}}$ is a sufficient condition for the ERM algorithm to be PAC. Now the whole affair rests on us being able to establish the UCEM property for various "interesting" and "useful" problem specifications. This will be our concern in the lectures ahead. However, let me give you a hint of what to expect. In many cases, we will be able to show that the induced class $\mathscr{L}_{\mathscr{F}}$ is so well-behaved that the bound

$$\mathbb{E}_{P^n}\left[\sup_{f\in\mathscr{F}} \left|\mathsf{P}_{Z^n}(\ell_f) - P(\ell_f)\right|\right] \leq \frac{C_{\mathscr{F},\ell}}{\sqrt{n}} \tag{10}$$

holds for *every* P, where $C_{\mathscr{F},\ell} > 0$ is some constant that depends only on the characteristics of the hypothesis class $\mathscr{F}$ and the loss function ℓ. Since ℓ_f is bounded between 0 and 1, the function

$$g(Z^n) \triangleq \sup_{f\in\mathscr{F}} \left|\mathsf{P}_{Z^n}(\ell_f) - P(\ell_f)\right|$$

has bounded differences with constants $c_1 = \ldots = c_n = 1/n$. McDiarmid's inequality then tells us that, for any $t > 0$,

$$P^n\Big(g(Z^n) - \mathbb{E}g(Z^n) \geq t\Big) \leq e^{-2nt^2}. \tag{11}$$

Let

$$n_0(\varepsilon,\delta) \triangleq \max\left\{\frac{4C_{\mathscr{F},\ell}^2}{\varepsilon^2}, \frac{2}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right\} + 1. \tag{12}$$

Then for any $n \geq n_0(\varepsilon,\delta)$

$$\begin{aligned} P^n\Big(g(Z^n) \geq \varepsilon\Big) &= P^n\Big(g(Z^n) - \mathbb{E}g(Z^n) \geq \varepsilon - \mathbb{E}g(Z^n)\Big) \\ &\leq P^n\left(g(Z^n) - \mathbb{E}g(Z^n) \geq \varepsilon - \frac{C_{\mathscr{F},\ell}}{\sqrt{n}}\right) && \text{because of (10)} \\ &\leq P^n\left(g(Z^n) - \mathbb{E}g(Z^n) \geq \frac{\varepsilon}{2}\right) && \text{because } n > \frac{4C_{\mathscr{F},\ell}^2}{\varepsilon^2} \\ &\leq e^{-n\varepsilon^2/2} && \text{using (11) with } t = \varepsilon/2 \\ &< \delta && \text{because } n > \frac{2}{\varepsilon^2}\log\left(\frac{1}{\delta}\right) \end{aligned}$$

for *any* probability distribution P over $\mathsf{Z} = \mathsf{X} \times \mathsf{Y}$. Thus, we have derived a very important fact: If the induced loss class $\mathscr{L}_{\mathscr{F}}$ satisfies (10), then (a) it has the UCEM property, and consequently is model-free learnable using the ERM algorithm, and (b) the sample complexity is polynomial in $1/\varepsilon$ and logarithmic in $1/\delta$. Our next order of business will be to derive sufficient conditions on $\mathscr{F}$ and ℓ for something like (10) to hold.

References

[Hau92] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 95:129–161, 1992.