Probably Approximately Correct Learnability

Matthieu R. Bloch

Now that we have introduced a complete model for supervised learning, our objective is to show that some of the questions raised earlier have a chance of being answered. We proceed by analyzing a simplified model, which still captures the essence of the problem but is more easily amenable to analysis. We will talk about the more general setting later in the semester.

We consider the supervised learning model that consists of the following.

1. A dataset $\mathcal{D} \triangleq \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\}$
   - $\{\mathbf{x}_i\}_{i=1}^N$ drawn i.i.d. from an unknown probability distribution $P_{\mathbf{x}}$ on $\mathcal{X}$;
   - $\{y_i\}_{i=1}^N$ with $\mathcal{Y} = \{0,1\}$ (binary classification).
2. An unknown $f : \mathcal{X} \to \mathcal{Y}$, no noise.
3. A finite set of hypotheses $\mathcal{H}$, $|\mathcal{H}| = M < \infty$, denoted $\mathcal{H} \triangleq \{h_i\}_{i=1}^M$.
4. A *binary* loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+ : (y_1, y_2) \mapsto \mathbb{1}\{y_1 \neq y_2\}$.

Note that we do not specify a specific algorithm yet as we will be focusing on a more abstract learning operation.

For this model and any hypothesis $h \in \mathcal{H}$, the true risk simplifies as

$$R(h) \triangleq \mathbb{E}_{\mathbf{x}y}(\mathbb{1}\{h(\mathbf{x}) \neq y\}) = \sum_{\mathbf{x}} \sum_{y} p_{\mathbf{x},y}(\mathbf{x}, y) \mathbb{1}\{h(\mathbf{x}) \neq y\} = \mathbb{P}_{\mathbf{x}y}(h(\mathbf{x}) \neq y). \tag{1}$$

and the empirical risk becomes

$$\widehat{R}_N(h) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{h(\mathbf{x}_i) \neq y_i\}. \tag{2}$$

We will discuss this in more details later, but it is very natural for learning algorithms to attempt to minimize the empirical risk and look for a hypothesis h^* that ensures a minimal risk

$$h^* = \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{R}_N(h). \tag{3}$$

1 Sample complexity

Generalizing The first question we raised was the possibility of *generalizing* a hypothesis. Mathematically, for a specific hypothesis $h_j \in \mathcal{H}$, this means assessing how $\widehat{R}_N(h_j)$ compares to $R(h_j)$. Observe that the empirical risk in (2) is a random variable since it is a function of the data set, which is a random variable. More specifically, since every $\mathbf{x}_i$ is generated independent and identically distributed (i.i.d.), the empirical risk is actually the *sample average* of N i.i.d. variables $\mathbb{1}\{h(\mathbf{x}_i) \neq y\}$. In addition observe that

$$\mathbb{E}\Big(\widehat{R}_N(h_j)\Big) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}(\mathbb{1}\{h(\mathbf{x}_i) \neq y_i\}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_{\mathbf{x},y}(h(\mathbf{x}) \neq y_i) = R(h_j) \tag{4}$$

Therefore, the quantity $\mathbb{P}\left(\left|\widehat{R}_N(h_j) - R(h_j)\right| > \epsilon\right)$ is the probability that sample average of i.i.d. random variables differ from their mean by more than ϵ. Such bounds are extremely common in applied probability and are known as *concentration inequalities*. We will now review some of the fundamental ideas behind these bounds.

The start of most if not all concentration inequalities is Markov's lemma.

Lemma 1.1. *Let X be a non-negative real-valued random variable. Then for all $t > 0$*

$$\mathbb{P}(X \geqslant t) \leqslant \frac{\mathbb{E}(X)}{t}. \tag{5}$$

Proof. For $t > 0$, let $\mathbb{1}\{X \geqslant t\}$ be the indicator function of the event $\{X \geqslant t\}$. Then,

$$\mathbb{E}[X] \geqslant \mathbb{E}[X\mathbb{1}\{X \geqslant t\}] \geqslant t\mathbb{P}[X \geqslant t], \tag{6}$$

where the first inequality follows because the indicator function is $\{0, 1\}$-valued and X is non-negative; the second because $X \geqslant t$ whenever $\mathbb{1}\{X \geqslant t\} = 1$ and 0 else.

That was a clean and fast proof, but you may be more comfortable going back to the definition of $\mathbb{E}(X)$ to prove the result. Note that

$$\mathbb{E}(X) = \int_0^\infty x p_X(x)dx = \underbrace{\int_0^t x p_X(x)dx}_{\geqslant 0} + \int_t^\infty x p_X(x)dx \overset{(a)}{\geqslant} t\int_t^\infty p_X(x)dx \tag{7}$$

$$= t\mathbb{P}(X \geqslant t) \tag{8}$$

where (a) follows from the fact that $x \geqslant t$ in the second integral. Note that the non-negative nature of X is crucial to lower bound the first integral. ∎

By choosing $t = \epsilon\mathbb{E}(X)$ for $\epsilon > 0$ in (5), we obtain $\mathbb{P}(X \geqslant \epsilon\mathbb{E}(X)) \leqslant \frac{1}{\epsilon}$, which is consistent with the intuition that it is unlikely that a random variable takes a value very far away from its mean.

In spite of its relative simplicity, Markov's inequality is a powerful tool because it can be "boosted." For $X \in \mathcal{X} \subset \mathbb{R}$, consider $\phi : \mathcal{X} \to \mathbb{R}^+$ non-decreasing on $\mathcal{X}$ such that $\mathbb{E}(|\phi(X)|) < \infty$. Then,

$$\mathbb{P}[X \geqslant t] = \mathbb{E}[\mathbb{1}\{X \geqslant t\}] = \mathbb{E}[\mathbb{1}\{X \geqslant t\}\mathbb{1}\{\phi(X) \geqslant \phi(t)\}] \leqslant \mathbb{P}[\phi(X) \geqslant \phi(t)], \tag{9}$$

where we have used the definition of ϕ and the fact that an indicator function is upper bounded by one. Applying Markov's inequality we obtain

$$\mathbb{P}[X \geqslant t] \leqslant \frac{\mathbb{E}[\phi(X)]}{\phi(t)}, \tag{10}$$

which is potentially a better bound than (5). Of course, the difficulty is in choosing the appropriate function ϕ to make the result meaningful. The most well-known application of this concept leads to *Chebyshev's inequality*.

Lemma 1.2 (Chebyshev's inequality)**.** *Let $X \in \mathbb{R}$. Then,*

$$\mathbb{P}[|X - \mathbb{E}(X)| \geqslant t] \leqslant \frac{\mathrm{Var}(X)}{t^2}. \tag{11}$$

Proof. Define $Y \triangleq |X - \mathbb{E}(X)|$ and $\phi : \mathbb{R}^+ \to \mathbb{R}^+ : t \mapsto t^2$. Then, by the boosted Markov's inequality we obtain

$$\mathbb{P}[|X - \mathbb{E}(X)| \geqslant t] = \mathbb{P}[Y \geqslant t] \leqslant \frac{\mathbb{E}[Y^2]}{t^2} = \frac{\text{Var}(X)}{t^2}. \tag{12}$$

■

As an application of Chebyshev's inequality, we derive the weak law of large numbers.

Lemma 1.3 (Weak law of large numbers). *Let $X_i \sim p_{X_i}$ be independent with $\mathbb{E}[|X_i|] < \infty$ and $\text{Var}(X_i) < \sigma^2$ for some $\sigma^2 \in \mathbb{R}^+$. Define $Z = \frac{1}{N}\sum_{i=1}^N X_i$ for $N \in \mathbb{N}^*$. Then Z converges in probability to $\frac{1}{N}\sum_{i=1}^N \mathbb{E}(X_i)$.*

Proof. First observe that

$$\mathbb{E}[Z] = \frac{1}{N}\sum_{i=1}^N \mathbb{E}[X_i] \quad \text{and} \quad \text{Var}(Z) = \frac{1}{N^2}\sum_{i=1}^N \text{Var}(X_i). \tag{13}$$

Therefore,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^N X_i - \frac{1}{N}\sum_{i=1}^N \mathbb{E}[X_i]\right| \geqslant \epsilon\right) = \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^N X_i - \frac{1}{N}\sum_{i=1}^N \mathbb{E}[X_i]\right|^2 \geqslant \epsilon^2\right) \tag{14}$$

$$\leqslant \sum_{i=1}^N \frac{\text{Var}(X_i)}{N^2\epsilon^2} < \frac{\sigma^2}{N\epsilon^2}. \tag{15}$$

■

The weak law of large numbers is essentially stating that $\frac{1}{N}\sum_{i=1}^N X_i$ concentrates around its average. Note, however, that the convergence we proved in (15) is rather slow, on the order of $1/N$.

Let us now go back to our learning problem. Applying (15), we know that

$$\forall \epsilon > 0 \quad \mathbb{P}_{\{(\mathbf{x}_i, y_i)\}}\left(\left|\widehat{R}_N(h_j) - R(h_j)\right| \geqslant \epsilon\right) \leqslant \frac{\text{Var}(\mathbb{1}\{h_j(\mathbf{x}_1) \neq y_1\})}{N\epsilon^2} \leqslant \frac{1}{N\epsilon^2}, \tag{16}$$

where the last inequality comes from the observation that $\text{Var}(\mathbb{1}\{h_j(\mathbf{x}_1) \neq y\}) \leqslant 1$ since the indicator function is a $\{0,1\}$-valued function. Notice that the bound that we obtain is *universal* in that it does not depend on $P_{\mathbf{x}}$ anymore. This is particularly pleasing because we introduced $P_{\mathbf{x}}$ in a rather arbitrary way.

We can now compute the *sample complexity* for generalizing h_j, defined as the number of samples $N_{\epsilon,\delta}$ required to achieve $\left|\widehat{R}_N(h_j) - R(h_j)\right| \leqslant \epsilon$ with probability at least $1 - \delta$. From (16), note that we obtain

$$N_{\epsilon,\delta} \geqslant \frac{1}{\delta\epsilon^2}. \tag{17}$$

The sample complexity behavior with δ and ϵ is consistent with our intuition, the more precise we want the empirical risk to be, the more samples we need.

Learning Unfortunately, the situation is slightly bleaker than what (17) shows. What we really care about is to learn and how the empirical risk of h^* generalizes, not the empirical risk of a *given* hypothesis in $\mathcal{H}$. We therefore need to make a statement about

$$\mathbb{P}_{\{(\mathbf{x}_i,y_i)\}}\Big(\Big|\widehat{R}_N(h^*) - R(h^*)\Big| \geqslant \epsilon\Big), \tag{18}$$

which is unfortunately hard to compute explicitly because $h^* = \operatorname{argmin}_{h\in\mathcal{H}} \widehat{R}_N(h)$. We can proceed by bounding (18), noting that

$$\mathbb{P}_{\{(\mathbf{x}_i,y_i)\}}\Big(\Big|\widehat{R}_N(h^*) - R(h^*)\Big| < \epsilon\Big) \geqslant \mathbb{P}_{\{(\mathbf{x}_i,y_i)\}}\Big(\forall h_j \in \mathcal{H}\ \Big|\widehat{R}_N(h_j) - R(h_j)\Big| < \epsilon\Big) \tag{19}$$

so that

$$\mathbb{P}_{\{(\mathbf{x}_i,y_i)\}}\Big(\Big|\widehat{R}_N(h^*) - R(h^*)\Big| \geqslant \epsilon\Big) \leqslant \mathbb{P}_{\{(\mathbf{x}_i,y_i)\}}\Big(\exists h_j \in \mathcal{H}\ \Big|\widehat{R}_N(h_j) - R(h_j)\Big| \geqslant \epsilon\Big). \tag{20}$$

The quantity on the right-hand-side of (20) is still hard to analyze because the events

$$\mathcal{E}_j \triangleq \{\Big|\widehat{R}_N(h_j) - R(h_j)\Big| \geqslant \epsilon\}$$

are *not* independent since they are all functions of the same dataset. A usual trick to deal with such quantities is to use the *union bound*,

$$\mathbb{P}_{\{(\mathbf{x}_i,y_i)\}}\Big(\exists h_j \in \mathcal{H}\ \Big|\widehat{R}_N(h_j) - R(h_j)\Big| \geqslant \epsilon\Big) \leqslant \sum_{j=1}^{M} \mathbb{P}_{\{(\mathbf{x}_i,y_i)\}}\Big(\Big|\widehat{R}_N(h_j) - R(h_j)\Big| \geqslant \epsilon\Big). \tag{21}$$

Combining (20) and (21) with (16), we obtain

$$\mathbb{P}_{\{(\mathbf{x}_i,y_i)\}}\Big(\Big|\widehat{R}_N(h^*) - R(h^*)\Big| \geqslant \epsilon\Big) \leqslant \frac{M}{N\epsilon^2}, \tag{22}$$

so that the sample complexity to generalize h^* is

$$N_{\epsilon,\delta} \geqslant \frac{M}{\delta\epsilon^2}. \tag{23}$$

This is a pessimistic result, because it tells us that the number of samples in the dataset must be larger than the numbers of hypotheses in $\mathcal{H}$, which will prevent us from using large sets of hypotheses that are presumably "rich" and have a better chance of approximating the unknown function h.

We can actually improve (23) by improving upon Chebyshev's inequality and choosing a better boosting function. For instance, with $\phi : t \to t^q$ for $q \in \mathbb{N}\setminus\{0,1\}$, we have

$$\mathbb{P}[|X - \mathbb{E}(X)| \geqslant t] \leqslant \frac{\mathbb{E}[|X - \mathbb{E}(X)|^q]}{t^q}. \tag{24}$$

If $\forall q \in \mathbb{N}\setminus\{0,1\}$, $\mathbb{E}[|X - \mathbb{E}(X)|^q] < \infty$, we obtain

$$\mathbb{P}[|X - \mathbb{E}(X)| \geqslant t] \leqslant \inf_{q\in\mathbb{N}\setminus\{0,1\}} \frac{\mathbb{E}[|X - \mathbb{E}(X)|^q]}{t^q}. \tag{25}$$

This might come in handy if one has access to higher order absolute moments, but we can actually do much better.

2 Chernoff bounds

The trick to obtain exponential concentration is to boost Markov's inequality with functions of the form $\phi_\lambda : t \to e^{\lambda t}$ for $\lambda \in \mathbb{R}^+$. The resulting bounds are often known as *Chernoff bounds*. Note that for a real-valued random variable Z

$$\forall \lambda \in \mathbb{R}^+ \qquad \mathbb{P}[|Z - \mathbb{E}[Z]| \geqslant t] \leqslant \frac{\mathbb{E}(\phi_\lambda(|Z - \mathbb{E}(Z)|))}{\phi_\lambda(t)} = e^{-\lambda t}\mathbb{E}\left[e^{\lambda|Z-\mathbb{E}[Z]|}\right] \tag{26}$$

Applying the union bound

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geqslant t] = \mathbb{P}[Z - \mathbb{E}[Z] \geqslant t] + \mathbb{P}[\mathbb{E}[Z] - Z \geqslant t]. \tag{27}$$

Setting $\tilde{Z} \triangleq Z - \mathbb{E}[Z]$ or $\tilde{Z} \triangleq \mathbb{E}[Z] - Z$, the problem of deriving concentration inequalities is tantamount to studying $P[\tilde{Z} \geqslant t]$ where $\tilde{Z} \in \mathbb{R}$ is centered. We will make this assumption from now on to simplify analysis and notation without losing generality.

2.1 Concentration inequalities for sub-Gaussian random variables

If Z centered and real-valued, we have

$$\forall \lambda \in \mathbb{R}^+ \quad \mathbb{P}[Z \geqslant t] \leqslant e^{-\lambda t}\mathbb{E}[e^{\lambda Z}]. \tag{28}$$

For $\lambda \in \mathbb{R}$, $\mathbb{E}[e^{\lambda Z}]$ is the Moment Generating Function (MGF) of Z, and $\psi_Z(\lambda) \triangleq \log \mathbb{E}\left[e^{\lambda Z}\right]$ is the Cumulant Generating Function (CGF) of Z. We recall some of the properties of the CGF.

Proposition 2.1. *Let Z be centered and real-valued such that $\mathbb{E}[e^{\lambda Z}] < \infty$ for all $|\lambda| < \epsilon$ for some $\epsilon > 0$. Then, the CGF satisfies the following properties.*

1. *ψ_Z is infinitely differentiable on $]-\epsilon, \epsilon[$. In particular, $\psi'_Z(0) = \psi_Z(0) = 0$;*
2. *$\psi_Z(\lambda) \geqslant \lambda\mathbb{E}(Z) = 0$;*
3. *If $Z = \sum_{i=1}^n X_i$ with X_i independent with well defined CGFs, $\psi_Z(\lambda) = \sum_{i=1}^n \psi_{X_i}(\lambda)$.*

Proof. We skip the subtleties behind proof of differentiability, which essentially follows from the dominated convergence theorem. We will also happily swap derivatives and integrals without worrying too much. By definition, $\psi_Z(0) = \log \mathbb{E}(1) = 0$. In addition,

$$\frac{d\psi}{d\lambda}(\lambda) = \frac{\mathbb{E}\big(Ze^{\lambda Z}\big)}{\mathbb{E}(e^{\lambda Z})}, \tag{29}$$

so that $\frac{d\psi}{d\lambda}(0) = 0$ since $\mathbb{E}(Z) = 0$. For the second part, note that by Jensen's inequality

$$\psi_Z(\lambda) = \log \mathbb{E}\big(e^{\lambda Z}\big) \geqslant \mathbb{E}\big(\log e^{\lambda Z}\big) = \lambda\mathbb{E}(Z). \tag{30}$$

For the third part, we have

$$\mathbb{E}\left[e^{\lambda Z}\right] = \mathbb{E}\left[e^{\lambda \sum_{i=1}^n X_i}\right] = \mathbb{E}\left[\prod_{i=1}^n e^{\lambda X_i}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{\lambda X_i}\right] = e^{\sum_{i=1}^n \log \mathbb{E}\left[e^{\lambda X_i}\right]} \tag{31}$$

■

Since our goal is to find the best upper bound in the right-hand side of (28), it is natural to maximize over all $\lambda \in \mathbb{R}^+$ to obtain

$$\mathbb{P}[Z \geqslant t] \leqslant \exp\left(- \sup_{\lambda \in \mathbb{R}^+} \left(\lambda t - \psi_Z(\lambda)\right)\right). \tag{32}$$

Definition 2.2. *For a real-valued centered random variable Z with cumulant generating function ψ_Z, the Cramer transform of ψ_Z is ψ_Z^* defined as*

$$\forall t \in \mathbb{R}^+ \qquad \psi_Z^*(t) \triangleq \sup_{\lambda \in \mathbb{R}^+} \left(\lambda t - \psi_Z(\lambda)\right). \tag{33}$$

Note that $\psi_Z^*(t) \geqslant -\psi_Z(0) = 0$ so that $\psi_Z^*(t) \in \mathbb{R}^+$. In general, the Cramer transform only takes simple values in trivial cases. If $\forall \lambda \in \mathbb{R}_+^*$, $\psi_Z(\lambda) = \infty$ then $\psi_Z^*(t) = 0$ since $\psi_Z(0) = 0$. If $t < \mathbb{E}(Z) = 0$ then $\psi_Z^*(t) = 0$. If $t \geqslant \mathbb{E}(Z)$ then, for all $\lambda < 0$, $\lambda t - \psi_Z(\lambda) \leqslant 0$. Consequently, the bound in (32) is only useful when $t \geqslant \mathbb{E}(Z)$, in which case we can maximize over $\lambda \in \mathbb{R}$ in (33). Note that we can write $\psi_Z^*(t)$ as $\psi_Z^*(t) = \lambda_t t - \psi_Z(\lambda_t)$ with λ_t such that $\psi_Z'(\lambda_t) = t$.

Example 2.3. *Let $Z \sim \mathcal{N}(0, \sigma^2)$. Then,*

$$\mathbb{E}[e^{\lambda Z}] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{z^2}{2\sigma^2}} e^{\lambda z} dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z-\lambda\sigma^2)^2}{2\sigma^2}} e^{\frac{\lambda^2\sigma^2}{2}} dz = e^{\frac{\lambda^2\sigma^2}{2}}. \tag{34}$$

Hence $\psi_Z(\lambda) = \log e^{\frac{\lambda^2\sigma^2}{2}} = \frac{\lambda^2\sigma^2}{2}$. Then $\psi_Z'(\lambda) = \lambda\sigma^2$ so that $\psi_Z'(\lambda_t) = t \Leftrightarrow \lambda_t\sigma^2 = t \Leftrightarrow \lambda_t = \frac{t}{\sigma^2}$ and

$$\psi_Z^*(t) = \frac{t^2}{\sigma^2} - \frac{t^2}{2\sigma^2} = \frac{t^2}{2\sigma^2}. \tag{35}$$

Hence, $\mathbb{P}[Z \geqslant t] \leqslant e^{-\frac{t^2}{2\sigma^2}}$.

The pleasingly simple form of the Chernoff bound for $Z \sim \mathcal{N}(0, 1)$ stems from the simple form of the CGF. This naturally leads to the following definition.

Definition 2.4. *$Z \in \mathbb{R}$ is subgaussian if $\exists \sigma^2 \in \mathbb{R}_*^+$ such that $\forall \lambda \in \mathbb{R}$, $\psi_Z(\lambda) \leqslant \frac{\lambda^2\sigma^2}{2}$*

If Z is subgaussian then $\forall \lambda \in \mathbb{R}$, $\lambda t - \psi_Z(\lambda) \geqslant \lambda t - \frac{\lambda^2\sigma^2}{2}$. In this case

$$\psi_Z^*(t) \geqslant \sup_{\lambda \in \mathbb{R}} \left(\lambda t - \frac{\lambda^2\sigma^2}{2}\right) = \frac{t^2}{2\sigma^2}. \tag{36}$$

Consequently, proving sub-Gaussianity is a proxy for obtaining exponential concentration.

2.2 Hoeffding's inequality

As an application, we establish the celebrated *Hoeffding's inequality*. We start by proving that some variables are sub-Gaussian.

Lemma 2.5 (Hoeffding's lemma). *Let a random variable Y such that $\mathbb{E}[Y] = 0$ and $Y \in [a, b]$. Then Y is sub-Gaussian, and more specifically $\psi_Y(\lambda) \leqslant \lambda^2 \frac{(b-a)^2}{8}$.*

Proof. Recall that $\psi_Y(\lambda) = \log \mathbb{E}[e^{\lambda Y}]$. We bound $\mathbb{E}[e^{\lambda Y}]$. Since $f : x \to e^{\lambda x}$ is convex, note that we can write

$$\forall y \in [a,b] \qquad y = \underbrace{\frac{b-y}{b-a}}_{0 \leqslant \gamma \leqslant 1} a + \underbrace{\frac{y-a}{b-a}}_{1-\gamma} b. \tag{37}$$

Then $y = \gamma a + (1-\gamma)b$, hence

$$e^{\lambda y} \leqslant \gamma e^{\lambda a} + (1-\gamma)e^{\lambda b} = \frac{b-y}{b-a}e^{\lambda a} + \frac{y-a}{b-a}e^{\lambda b} \tag{38}$$

and setting $\rho \triangleq \frac{-a}{b-a}$, we obtain

$$\begin{aligned}
\mathbb{E}[e^{\lambda Y}] &\leqslant \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b} && (39)\\
&= (1-\rho)e^{\lambda a} + \rho e^{\lambda b} && (40)\\
&= \left(1 - \rho + \rho e^{\lambda(b-a)}\right) e^{-\rho\lambda(b-a)} && (41)\\
&= \exp\left(\ln\left(1 - \rho + \rho e^{\lambda(b-a)}\right) - \rho\lambda(b-a)\right) && (42)
\end{aligned}$$

Consider the function $g_\rho : x \to \ln(1-\rho+\rho e^x) - \rho x$, then $g_\rho(x) \leqslant \frac{x^2}{8}$. Hence $\mathbb{E}[e^{\lambda Y}] \leqslant \exp\left(\frac{\lambda^2(b-a)^2}{8}\right)$. ∎

Alternative proof (more tricky). Let a random variable $Y \sim p_Y$ such that $\mathbb{E}[Y] = 0$ and $Y \in [a,b]$. Then $a \leqslant Y \leqslant b$ and $\frac{a-b}{2} \leqslant Y - \frac{a+b}{2} \leqslant \frac{b-a}{2}$, so that $\left|Y - \frac{a+b}{2}\right| \leqslant \frac{b-a}{2}$ and $\text{Var}(Y) \leqslant \frac{(b-a)^2}{4}$. Define $Z \in [a,b]$ such that $p_Z(y) = \frac{e^{\lambda y}}{\mathbb{E}[e^{\lambda Y}]} p_Y(y)$. Then we also have $\text{Var}(Z) \leqslant \frac{(b-a)^2}{4}$. On the other hand

$$\begin{aligned}
\text{Var}(Z) &= \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 && (43)\\
&= \int_a^b y^2 \frac{e^{\lambda y}}{\mathbb{E}[e^{\lambda Y}]} p_Y(y)\mathrm{d}y - \left(\int_a^b y \frac{e^{\lambda y}}{\mathbb{E}[e^{\lambda Y}]} p_Y(y)\mathrm{d}y\right)^2 && (44)\\
&= \frac{\mathbb{E}[Y^2 e^{\lambda Y}]}{\mathbb{E}[e^{\lambda Y}]} - \left(\frac{\mathbb{E}[Y e^{\lambda Y}]}{\mathbb{E}[e^{\lambda Y}]}\right)^2 \leqslant \frac{(b-a)^2}{4} && (45)
\end{aligned}$$

Note that

$$\psi_Y(\lambda) = \log \mathbb{E}[e^{\lambda Y}] \qquad \psi_Y(0) = 0 \tag{46}$$

Then

$$\psi_Y'(\lambda) = \frac{\mathbb{E}[Y e^{\lambda Y}]}{\mathbb{E}[e^{\lambda Y}]} \qquad \psi_Y'(0) = 0 \tag{47}$$

and

$$\psi_Y''(\lambda) = \frac{\mathbb{E}[Y^2 e^{\lambda Y}]}{\mathbb{E}[e^{\lambda Y}]} - \frac{\mathbb{E}[Y e^{\lambda Y}]^2}{\mathbb{E}[e^{\lambda Y}]^2} = \text{Var}(Z) \leqslant \frac{(b-a)^2}{4} \tag{48}$$

From Taylor's theorem, $\exists c \in [0,\lambda]$ such that

$$\psi_Y(\lambda) = \psi_Y(0) + \lambda\psi_Y'(0) + \frac{\lambda^2}{2}\psi_Y''(c) \tag{49}$$

Therefore, $\psi_Y(\lambda) \leqslant \frac{\lambda^2(b-a)^2}{8}$. ∎

Proposition 2.6 (Hoeffding's inequality). *Consider independent random variables X_i with $\mathbb{E}[X_i] = 0$ and $X_i \in [a_i, b_i]$. Let $Y = \sum_{i=1}^n X_i$. Then*

$$\mathbb{P}\left[\sum_{i=1}^n X_i \geqslant t\right] \leqslant \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \tag{50}$$

Proof. The proof follows by combining Lemma 2.5 with Proposition 2.1 to obtain

$$\psi_Y(\lambda) = \sum_{i=1}^n \psi_{X_i}(\lambda) \leqslant \sum_{i=1}^n \frac{\lambda^2}{8}(b_i - a_i)^2 = \frac{\lambda^2 \sigma^2}{2} \tag{51}$$

with

$$\sigma^2 \triangleq \frac{1}{4}\sum_{i=1}^n (b_i - a_i)^2. \tag{52}$$

■

3 Learning may work

Let us now revisit our learning problem with Hoeffding's inequality. For a given $h_j \in \mathcal{H}$, we obtain

$$\forall \epsilon > 0 \quad \mathbb{P}_{\{(\mathbf{x}_i, y_i)\}}\left(\left|\widehat{R}_N(h_j) - R(h_j)\right| \geqslant \epsilon\right) \leqslant 2\exp\left(-2N\epsilon^2\right), \tag{53}$$

which decays *exponentially fast* with N. Consequently, following the same reasoning in Section 1, we also have

$$\mathbb{P}_{\{(\mathbf{x}_i, y_i)\}}\left(\left|\widehat{R}_N(h^*) - R(h^*)\right| \geqslant \epsilon\right) \leqslant 2M\exp\left(-2N\epsilon^2\right), \tag{54}$$

so that the sample complexity to generalize h^* is

$$N_{\epsilon,\delta} \geqslant \frac{1}{2\epsilon^2}\left(\log M + \log\frac{2}{\delta}\right). \tag{55}$$

This is a much more optimistic result. The sample complexity to generalize h^* now only depends only *logarithmically* on the number of hypotheses M. We can therefore hope to use a very large set $\mathcal{H}$ to find good approximations of f but without requiring unreasonably many samples N.

Remark 3.1. *The result is not quite ideal, because many sets $\mathcal{H}$ of practical interest (neural networks, perceptron) have $|\mathcal{H}| = \infty$. Still this should give us hope that we're doing something meaningful.*

4 PAC learnability

The last question to answer is how $R(h^*)$, the true risk of the hypothesis we pick with empirical risk minimization, compares to $R(h^\sharp)$, the true risk of the best hypothesis in the class. Upon inspection of how we derived the sample complexity with Hoeffding's inequality, note that we actually proved something much stronger than what we needed. We actually proved that the sample complexity ensures that

$$\mathbb{P}_{\{(\mathbf{x}_i, y_i)\}}\left(\forall h_j \in \mathcal{H} \left|\widehat{R}_N(h_j) - R(h_j)\right| \leqslant \epsilon\right) \geqslant 1 - \delta. \tag{56}$$

In that case, the following holds.

Lemma 4.1. *If* $\forall h_j \in \mathcal{H}$ *we have* $\left|\widehat{R}_N(h_j) - R(h_j)\right| \leqslant \epsilon$ *then* $\left|R(h^*) - R(h^\sharp)\right| \leqslant 2\epsilon$.

Proof. Note that

$$\left|R(h^*) - R(h^\sharp)\right| = \left|R(h^*) - \widehat{R}_N(h^*) + \widehat{R}_N(h^*) - R(h^\sharp)\right| \tag{57}$$

$$\leqslant \left|R(h^*) - \widehat{R}_N(h^*)\right| + \left|\widehat{R}_N(h^*) - R(h^\sharp)\right|. \tag{58}$$

By assumption, we have $\left|R(h^*) - \widehat{R}_N(h^*)\right| \leqslant \epsilon$ since $h^* \in \mathcal{H}$. In addition, by definition of $h^\sharp$ as the minimizer of the true risk,

$$R(h^\sharp) \leqslant R(h^*) \leqslant \widehat{R}_N(h^*) + \epsilon. \tag{59}$$

By definition of h^* as the minimizer of the empirical risk, we also have

$$\widehat{R}_N(h^*) \leqslant \widehat{R}_N(h^\sharp) \leqslant R(h^\sharp) + \epsilon. \tag{60}$$

so that

$$\left|\widehat{R}_N(h^*) - R(h^\sharp)\right| \leqslant \epsilon. \tag{61}$$

■

In learning theory, these ideas are formalized in terms of probably approximately correct learnability (PAC) as follows.

Definition 4.2. *A hypothesis set* $\mathcal{H}$ *is PAC learnable if there exists a function* $N_{\mathcal{H}} :]0;1[^2 \to \mathbb{N}$ *and a learning algorithm such that:*

- *for very* $\epsilon, \delta \in]0;1[$,
- *for every* $P_{\mathbf{x}}$, $P_{y|\mathbf{x}}$,
- *when running the algorithm on at least* $N_{\mathcal{H}}(\epsilon, \delta)$ *i.i.d. examples, the algorithm returns a hypothesis* $h \in \mathcal{H}$ *such that*

$$\mathbb{P}_{\mathbf{x}y}\left(\left|R(h) - R(h^\sharp)\right| \leqslant \epsilon\right) \geqslant 1 - \delta$$

The function $N_{\mathcal{H}}(\epsilon, \delta)$ is the sample complexity. Note that the definition of sample complexity is here slightly different from what we used earlier. Sample complexity is defined with respect to the true risk of $h^\sharp$, while we previously only worried about the true risk of h^*. The name probably approximately correct comes from the bound $\mathbb{P}_{\mathbf{x}y}\left(\left|R(h) - R(h^\sharp)\right| \leqslant \epsilon\right) \geqslant 1 - \delta$. In words, it says that with probability at least $1 - \delta$ (probably), the true risk incurred by h is no more than ϵ away from the best true risk (approximately correct). Note that the definition of PAC learnability is quite stringent because it requires the bound to hold *irrespective* of the what $P_{\mathbf{x}}$ and $P_{\mathbf{y}|\mathbf{x}}$ really are. All we should assume is that they exist.

Perhaps surprisingly, if you trace back everything we proved so far (check for yourself!), we have effectively already proved the following result.

Proposition 4.3. *A finite hypothesis set* $\mathcal{H}$ *is PAC learnable with the Empirical Risk Minimization algorithm and with sample complexity*

$$N_{\mathcal{H}}(\epsilon, \delta) = \lceil \frac{2\ln(2\left|\mathcal{H}\right|/\delta)}{\epsilon^2} \rceil$$

Although the caveats regarding the fact that we require $|\mathcal{H}| < \infty$ still apply, it should be comforting that we can make such a fundamental statement about learning.

Remark 4.4. *You might note that the sample complexity seems off by a factor of two compared to what we derived earlier. This is because the sample complexity as per Definition 4.2 requires the true risks of h^* and $h^\sharp$ to be close, instead of requiring the empirical risk of h^* to be close to the true risk of h^*. Proving the result of Proposition 4.3 requires you to use Lemma 4.1.*

Note, however, that this does not address the question of ensuring that the risk of the best hypothesis $h^* = \text{argmin}_{h\in\mathcal{H}} \widehat{R}_N(h)$ we find is actually small. To have a small risk, we must ensure that the hypothesis class $\mathcal{H}$ is somehow "rich enough" to have a good chance of well approximating the unknown function f. With our current analysis, the size $|\mathcal{H}|$ of the class is the proxy for the richness of the class, and although the dependence of the sample complexity on $|\mathcal{H}|$ is only logarithmic, we need many sample if the class size grows large.

In practice, the size of the dataset N is fixed, and three phenomena occur as we increase the richness of the class $\mathcal{H}$. Recall that $h^* \triangleq \text{argmin}_{h\in\mathcal{H}} \widehat{R}_N(h)$ and $h^\sharp \triangleq \text{argmin}_{h\in\mathcal{H}} R(h)$.

1. The empirical risk of h^* decreases;
2. The true risk of $h^\sharp$ decreases;
3. The true risk of h^* decreases before it increases again (the curve has a U-shape).

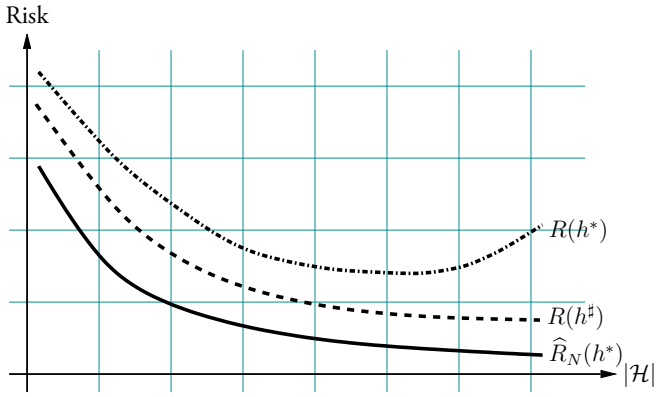

Figure 1: Evolution of risk when richness of $\mathcal{H}$ increases

In our simple learning model, the last phenomenon happens because as we increase the size of the class $|\mathcal{H}|$ for a fixed dataset size N, it becomes increasingly likely that there are hypotheses whose empirical risk is very different from their true risk. This behavior is representative of most if not all learning problems, and is summarized in Fig. 1.

One should also realize that it may not be possible to ever achieve zero risk learning. In fact, our general learning model accounts for the presence of noise through $P_{y|\mathbf{x}}$. This naturally prompts the question of what is the *smallest* risk $R(h^\sharp)$ that one can achieves and how to achieve it.

5 To go further

We have only touched upon concentration inequalities, there is an entire field of research devoted to proving such results in intricate situations. Two great references are [1], from which I borrowed most of the ideas, and [2].

To explore the topic further, recommend readings include [3, Section 1.3], [4, Chapters 2-4].

References

[1] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, 1st ed. Oxford University Press, Apr. 2016.

[2] M. Raginsky and I. Sason, "Concentration of measure inequalities in information theory, communications, and coding," *Foundations and Trends in Communications and Information Theory*, Sep. 2014.

[3] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning From Data*. AMLBook, 2012.

[4] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, May 2014.