PROBABLY APPROXIMATELY CORRECT LEARNABILITY

MATTHIEU R BLOCH

A SIMPLER SUPERVISED LEARNING PROBLEM

Consider a special case of the general supervised learning problem

1. 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 unknown* $P_\mathbf{x}$ on $\mathcal{X}$
   - $\{y_i\}_{i=1}^N$ labels with $\mathcal{Y} = \{0, 1\}$ (binary classification)
2. Unknown $f : \mathcal{X} \to \mathcal{Y}$, no noise.
3. Finite set of hypotheses $\mathcal{H}$, $|\mathcal{H}| = M < \infty$
   - $\mathcal{H} \triangleq \{h_i\}_{i=1}^M$
4. Binary loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+ : (y_1, y_2) \mapsto \mathbf{1}\{y_1 \neq y_2\}$

In this very specific case, the true risk simplifies

$$R(h) \triangleq \mathbb{E}_{\mathbf{x}y}[\mathbf{1}\{h(\mathbf{x}) \neq y\}] = \mathbb{P}_{\mathbf{x}y}(h(\mathbf{x}) \neq y)$$

The empirical risk becomes

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

CAN WE LEARN?

Our objective is to find a hypothesis $h^* = \text{argmin}_{h\in\mathcal{H}} \widehat{R}_N(h)$ that ensures a small risk

For a *fixed* $h_j \in \mathcal{H}$, how does $\widehat{R}_N(h_j)$ compares to $R(h_j)$?

Observe that for $h_j \in \mathcal{H}$

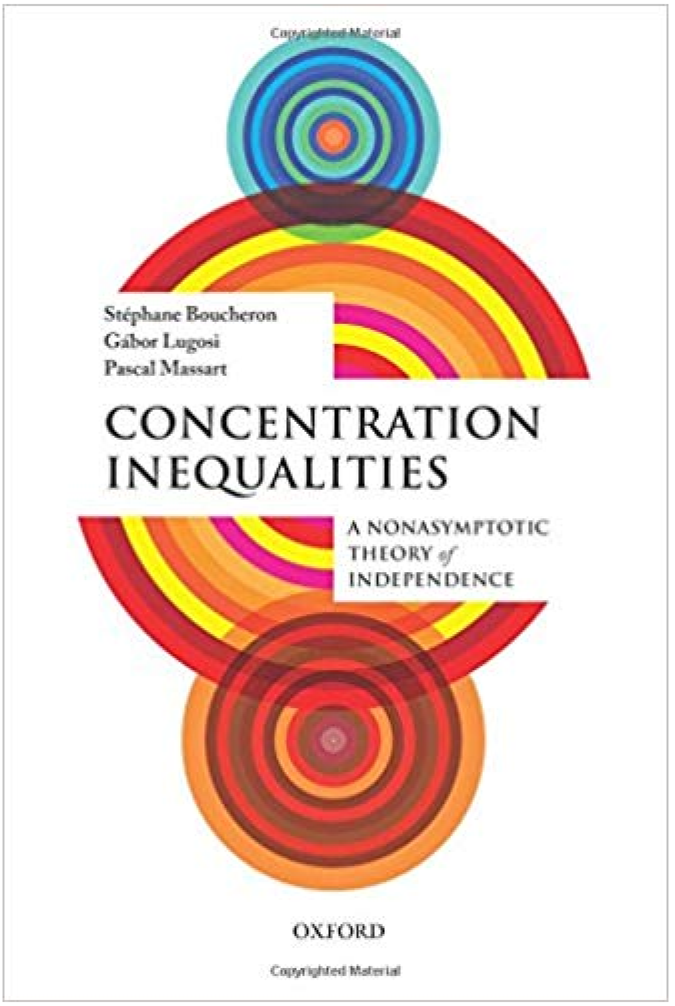
- The empirical risk is a sum of iid random variables

$$\widehat{R}_N(h_j) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{h_j(\mathbf{x}_i) \neq y_i\}$$

- $\mathbb{E}\left[\widehat{R}_N(h_j)\right] = R(h_j)$

$\mathbb{P}\left(\left|\widehat{R}_N(h_j) - R(h_j)\right| > \epsilon\right)$ is a statement about the deviation of a normalized sum of iid random variables from its mean

We're in luck! Such bounds, a.k.a, known as *concentration inequalities*, are a well studied subject



CONCENTRATION INEQUALITIES: BASICS

Lemma (Markov's inequality)

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

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Lemma (Chebyshev's inequality)

Let X be a real-valued random variable. Then for all $t > 0$

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

Proposition (Weak law of large numbers)

Let $\{X_i\}_{i=1}^N$ be i.i.d. real-valued random variables with finite mean μ and finite variance σ^2. Then

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^N X_i - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{N\epsilon^2} \qquad \lim_{N\to\infty} \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^N X_i - \mu\right| \geq \epsilon\right) = 0.$$

Proof:

① $$\mathbb{E}(X) = \int_0^{+\infty} x\, p_X(x)\, dx = \underbrace{\int_0^{t} x\, p_X(x)\, dx}_{\geq 0} + \underbrace{\int_t^{\infty} x\, p_X(x)\, dx}_{\geq t \int_t^{+\infty} p_X(x)\, dx \,=\, t\,\mathbb{P}(X \geq t)} \qquad (t > 0)$$

$$\geq t\,\mathbb{P}(X \geq t)$$

② $$\mathbb{1}\{X \geq t\} \in \{0,1\}$$

$$\mathbb{E}(X) \geq \mathbb{E}\big[X \underbrace{\mathbb{1}\{X \geq t\}}_{\in \{0,1\}}\big] \geq t\,\mathbb{E}\big[\mathbb{1}\{X \geq t\}\big] = t\,\mathbb{P}(X \geq t) \qquad \blacksquare$$

Proof: We can boost Markov's inequality

Assume $X \in \mathcal{X} \subset \mathbb{R}$. Consider $\phi : \mathcal{X} \longrightarrow \mathbb{R}^+$ non decreasing s.t. $\mathbb{E}[|\phi(X)|] < \infty$.

$$\mathbb{P}(X \geq t) = \mathbb{E}\left[\mathbb{1}\{X \geq t\}\right]$$

$$= \mathbb{E}\left[\mathbb{1}\{X \geq t\}\mathbb{1}\{\phi(X) \geq \phi(t)\}\right]$$

always 1 if $X \geq t$

$$\leq \mathbb{E}\left[\mathbb{1}\{\phi(X) \geq \phi(t)\}\right]$$

$$= \mathbb{P}(\phi(X) \geq \phi(t)) \leq \frac{\mathbb{E}(\phi(X))}{\phi(t)}$$

boosting

Application: $Y \triangleq |X - \mathbb{E}(X)|$ $\qquad \mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \dfrac{\mathbb{E}(|X - \mathbb{E}(X)|)}{t}$

$\phi : \mathbb{R}_+ \longrightarrow \mathbb{R}_+ : x \longmapsto x^2$

Hence $\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \dfrac{\mathbb{E}(|X - \mathbb{E}(X)|^2)}{t^2} = \dfrac{\mathrm{Var}(X)}{t^2}$ $\qquad \square$

Proof: Define $Z \triangleq \frac{1}{N}\sum_{i=1}^{N} X_i$

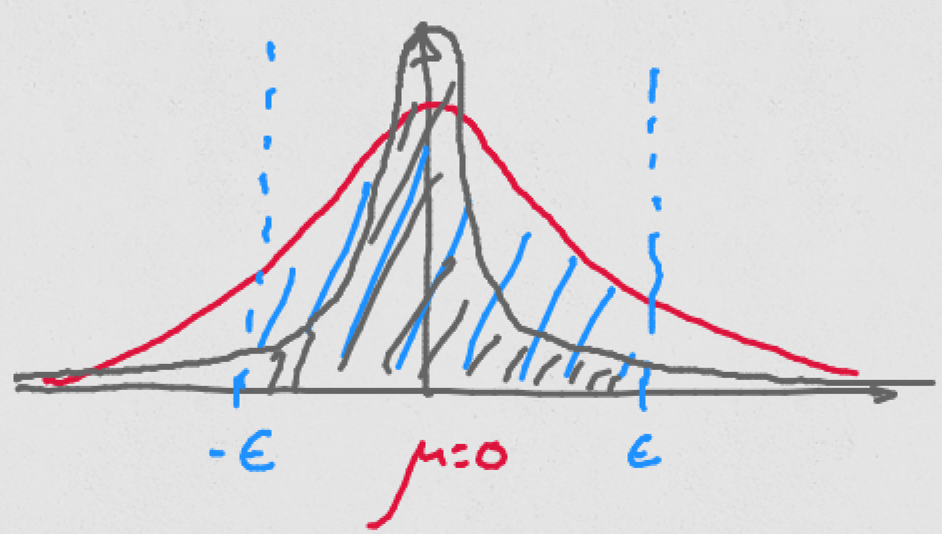
$$\mathbb{E}[Z] = \frac{1}{N}\sum_{i=1}^{N} \underbrace{\mathbb{E}(X_i)}_{\mu} = \mu$$

$$\mathrm{Var}(Z) = \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{1}{N^2}\mathrm{Var}\left(\sum_{i=1}^{N} X_i\right) = \frac{1}{N^2}\sum_{i=1}^{N} \underbrace{\mathrm{Var}(X_i)}_{\sigma^2} = \frac{\sigma^2}{N}$$

Apply Chebyshev's inequality:

$$\mathbb{P}(|Z-\mu| \geq t) \leq \frac{\sigma^2}{Nt^2} \quad \text{and} \quad \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{N\epsilon^2}$$

∎



BACK TO LEARNING

By the law of large number, we know that

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

Given enough data, we can *generalize*

How much data? $N = \frac{1}{\delta\epsilon^2}$ to ensure $\mathbb{P}\left(\left|\widehat{R}_N(h_j) - R(h_j)\right| \geq \epsilon\right) \leq \delta$.

That's not quite enough! We care about $\widehat{R}_N(h^*)$ where $h^* = \mathrm{argmin}_{h\in\mathcal{H}} \widehat{R}_N(h)$

- If $M = |\mathcal{H}|$ is large we should expect the existence of $h_k \in \mathcal{H}$ such that $\widehat{R}_N(h_k) \ll R(h_k)$

$$\mathbb{P}\left(\left|\widehat{R}_N(h^*) - R(h^*)\right| \geq \epsilon\right) \leq \mathbb{P}\left(\exists j : \left|\widehat{R}_N(h_j) - R(h_j)\right| \geq \epsilon\right)$$

$$\mathbb{P}\left(\left|\widehat{R}_N(h^*) - R(h^*)\right| \geq \epsilon\right) \leq \frac{M}{N\epsilon^2}$$

If we choose $N \geq \lceil \frac{M}{\delta\epsilon^2} \rceil$ we can ensure $\mathbb{P}\left(\left|\widehat{R}_N(h^*) - R(h^*)\right| \geq \epsilon\right) \leq \delta$.

- That's a lot of samples!

CONCENTRATION INEQUALITIES: NOT SO BASIC

We can obtain *much* better bounds than with Chebyshev

> **Lemma (Hoeffding's inequality)**
>
> Let $\{X_i\}_{i=1}^N$ be i.i.d. real-valued zero-mean random variables such that $X_i \in [a_i; b_i]$ with $a_i < b_i$. Then for all $\epsilon > 0$
>
> $$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^N X_i\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2N^2\epsilon^2}{\sum_{i=1}^N (b_i - a_i)^2}\right).$$

In our learning problem

$$\forall \epsilon > 0 \quad \mathbb{P}\left(\left|\widehat{R}_N(h_j) - R(h_j)\right| \geq \epsilon\right) \leq 2\exp(-2N\epsilon^2)$$

$$\forall \epsilon > 0 \quad \mathbb{P}\left(\left|\widehat{R}_N(h^*) - R(h^*)\right| \geq \epsilon\right) \leq 2M\exp(-2N\epsilon^2)$$

We can now choose $N \geq \lceil \frac{1}{2\epsilon^2}\left(\ln \frac{2M}{\delta}\right) \rceil$
M can be quite large (almost exponential in N) and, with enough data, we can generalize h^*.

How about learning $h^\sharp \triangleq \operatorname{argmin}_{h\in\mathcal{H}} R(h)$?

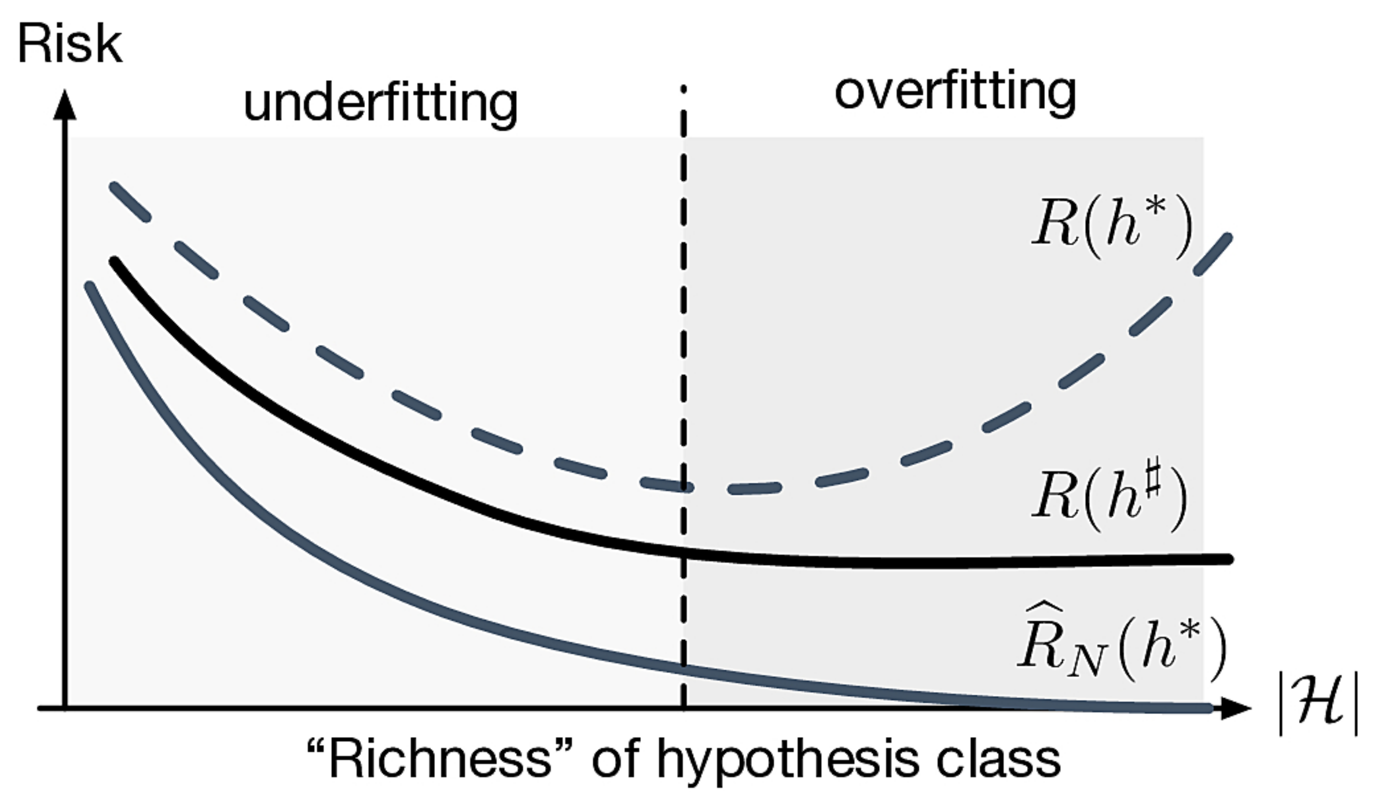
LEARNING CAN WORK!

Lemma.

If $\forall j \in \mathcal{H} \left|\widehat{R}_N(h_j) - R(h_j)\right| \leq \epsilon$ then $\left|R(h^*) - R(h^\sharp)\right| \leq 2\epsilon.$

How do we make $R(h^\sharp)$ small?

- Need bigger hypothesis class $\mathcal{H}$! (could we take $M \to \infty$?)
- Fundamental trade-off of learning



Proof: $$|R(h^*) - R(h^\#)| = |R(h^*) - \hat{R}_N(h^*) + \hat{R}_N(h^*) - R(h^\#)|$$
$$\leq \underbrace{|R(h^*) - \hat{R}_N(h^*)|}_{(1)} + \underbrace{|\hat{R}_N(h^*) - R(h^\#)|}_{(2)}$$

$\forall j \; |\hat{R}_N(h_j) - R(h_j)| \leq \epsilon$ hence $|\hat{R}_N(h^*) - R(h^*)| \leq \epsilon$ b/c $h^* \in \mathcal{H}$ (1)

By def of $h^\#$ $R(h^\#) \leq R(h^*) \leq \hat{R}_N(h^*) + \epsilon$ $(*)$

Similarly, by def h^* $\hat{R}_N(h^*) \leq \hat{R}_N(h^\#) \leq R(h^\#) + \epsilon$ b/c $h^\# \in \mathcal{H}$ $(**)$

Hence $|\hat{R}_N(h^*) - R(h^\#)| \leq \epsilon$ (2)

Therefore $|R(h^*) - R(h^\#)| \leq 2\epsilon$ ∎

PROBABLY APPROXIMATELY CORRECT LEARNABILITY

Definition. (PAC learnability)
A hypothesis set $\mathcal{H}$ is (agnostic) 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

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

The function $N_{\mathcal{H}}(\epsilon, \delta)$ is called *sample complexity*

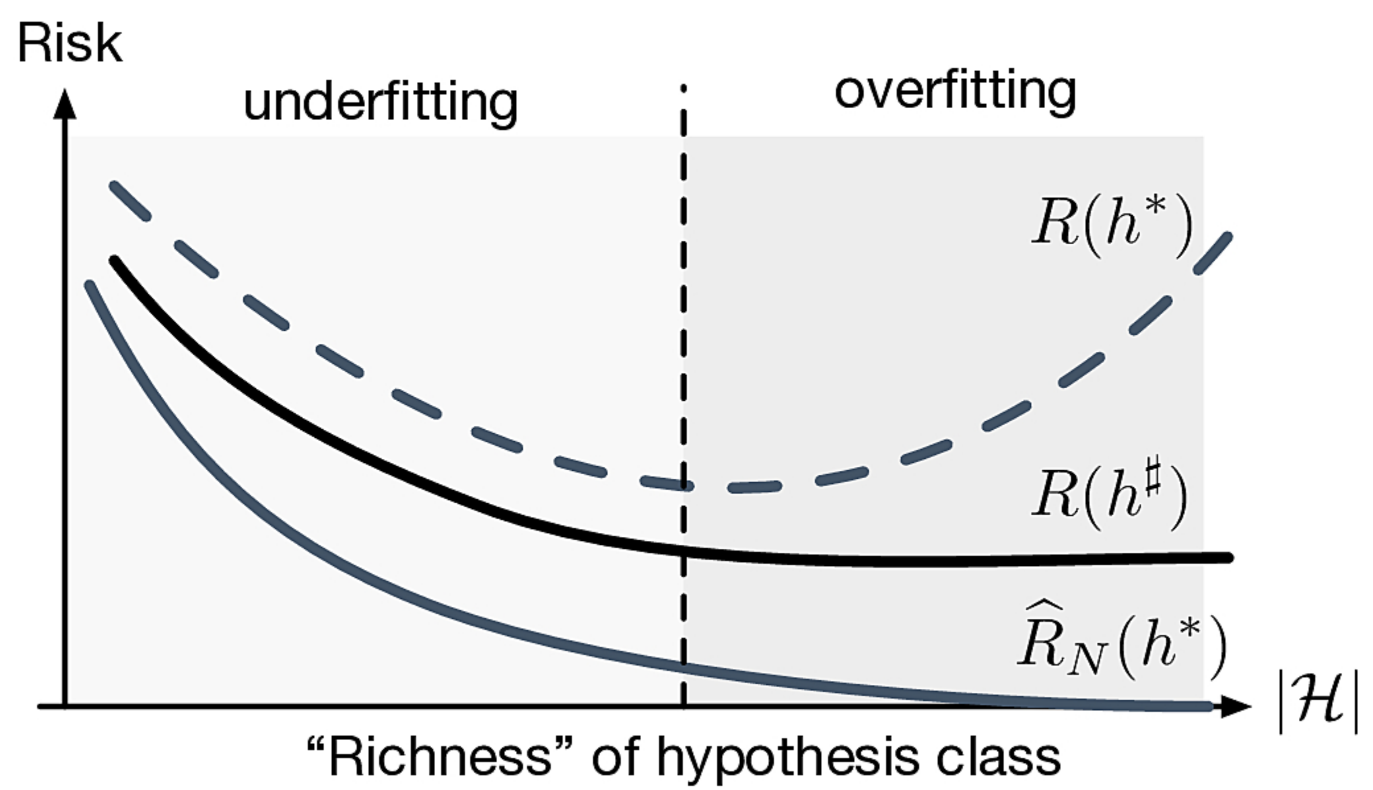
We have effectively already proved the following result

Proposition.

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

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

WHAT IS A GOOD HYPOTHESIS SET?



Ideally we want $|\mathcal{H}|$ small so that $R(h^*) \approx R(h^\sharp)$ and get lucky so that $R(h^*) \approx 0$

In general this is *not* possible

- Remember, we usually have to learn $P_{y|\mathbf{x}}$, not a function f

Questions

- What is the optimal binary classification hypothesis class?
- How small can $R(h^*)$ be?