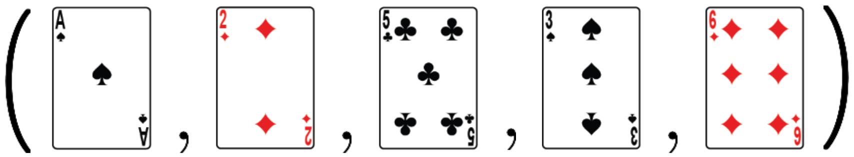
Another approach

You watch me do this trick a couple times and notice I always hand out 5 cards

Suppose you instead consider

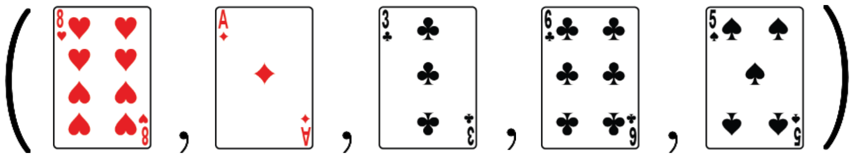
$$\mathbf{x}_1 = (0, 1, 0, 0, 1)$$

$y_1 =$



$$\mathbf{x}_2 = (1, 1, 0, 0, 0)$$

$y_2 =$



$$\vdots$$

Now, can you learn a function such that $f(\mathbf{x})$ is a reliable predictor of y?

Probability to the rescue!

Any f agreeing with the training data may be ***possible***
but that does not mean that any f is equally ***probable***

A short digression

- Suppose that Javier has a biased coin, which lands on heads with some unknown probability p
  - $\mathbb{P}[\text{heads}] = p$
  - $\mathbb{P}[\text{tails}] = 1 - p$
- Javier toss the coin n times
  - $\widehat{p} = \frac{\#\text{ of heads}}{n}$

Does $\widehat{p}$ tell us anything about p ?

What can we learn from $\widehat{p}$?

Given enough tosses (large n), we expect that $\widehat{p} \approx p$

Law of large numbers

$$\widehat{p} \to p \text{ as } n \to \infty$$

Clearly, at least in a very limited sense, we can learn something about p from observations

There is always the ***possibility*** that we are totally wrong, but given enough data, the ***probability*** should be very small

Connection to learning

Coin tosses: We want to estimate p (i.e., predict how likely a “heads” is)

Learning: We want to estimate a function $f : \mathcal{X} \to \mathcal{Y}$

Suppose we have a ***hypothesis*** h and that $\mathcal{Y}$ is discrete

Think of the $(\mathbf{x}_i, y_i)$ as a series of independent coin tosses, where the $(\mathbf{x}_i, y_i)$ are drawn from a probability distribution

- heads: our hypothesis is correct, i.e., $h(\mathbf{x}_i) = y_i$
- tails: our hypothesis is wrong, i.e., $h(\mathbf{x}_i) \neq y_i$

Define

(Population) risk: $R(h) := \mathbb{P}[h(X) \neq Y]$

Empirical risk: $\widehat{R}_n(h) := \dfrac{|\{i : h(\mathbf{x}_i) \neq y_i\}|}{n}$

Trust, but verify

The law of large numbers guarantees that as long as we have enough data, we will have that $R(h) \approx \widehat{R}_n(h)$

This means that we can use $\widehat{R}_n(h)$ to verify whether h was a good hypothesis

Unfortunately, ***verification is not learning***

- Where did h come from?
- What if $R(h)$ is large?
- How do we know if $h \approx f$, or at least, if $R(h) \approx R(f)$?
- Given many possible hypotheses, how can we pick a good one?

From coins to learning

Consider an ensemble of many hypotheses

$$\mathcal{H} = \{h_1, \ldots, h_m\}$$

If we fix a hypotheses h_j before drawing our data, then the law of large numbers tells us that $\widehat{R}_n(h_j) \to R(h_j)$

However, it is also true that for a fixed n, if m is large it can still be very likely that there is some hypothesis h_k for which $\widehat{R}_n(h_k)$ is still very far from $R(h_k)$

Example

Question 1: If I toss a fair coin 10 times, what is the probability that I get 10 heads?

Question 2: If I toss 1000 fair coins 10 times each, what is the probability that ***some*** coin will get 10 heads?

This illustrates the fundamental challenge of ***multiple hypothesis testing***

…and back to learning

If we have many hypotheses (large m), then

even though for any fixed hypothesis h_j it is likely that

$$\widehat{R}_n(h_j) \approx R(h_j)$$

it is also likely that there will be at least ***one*** hypothesis h_k where $\widehat{R}_n(h_k)$ is very different from $R(h_k)$

Can we adapt our approach to handle many hypotheses?

A first model of learning

Let's restrict our attention to binary classification

- our labels belong to $\mathcal{Y} = \{1, 0\}$ (or $\mathcal{Y} = \{+1, -1\}$)

We observe the data

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$$

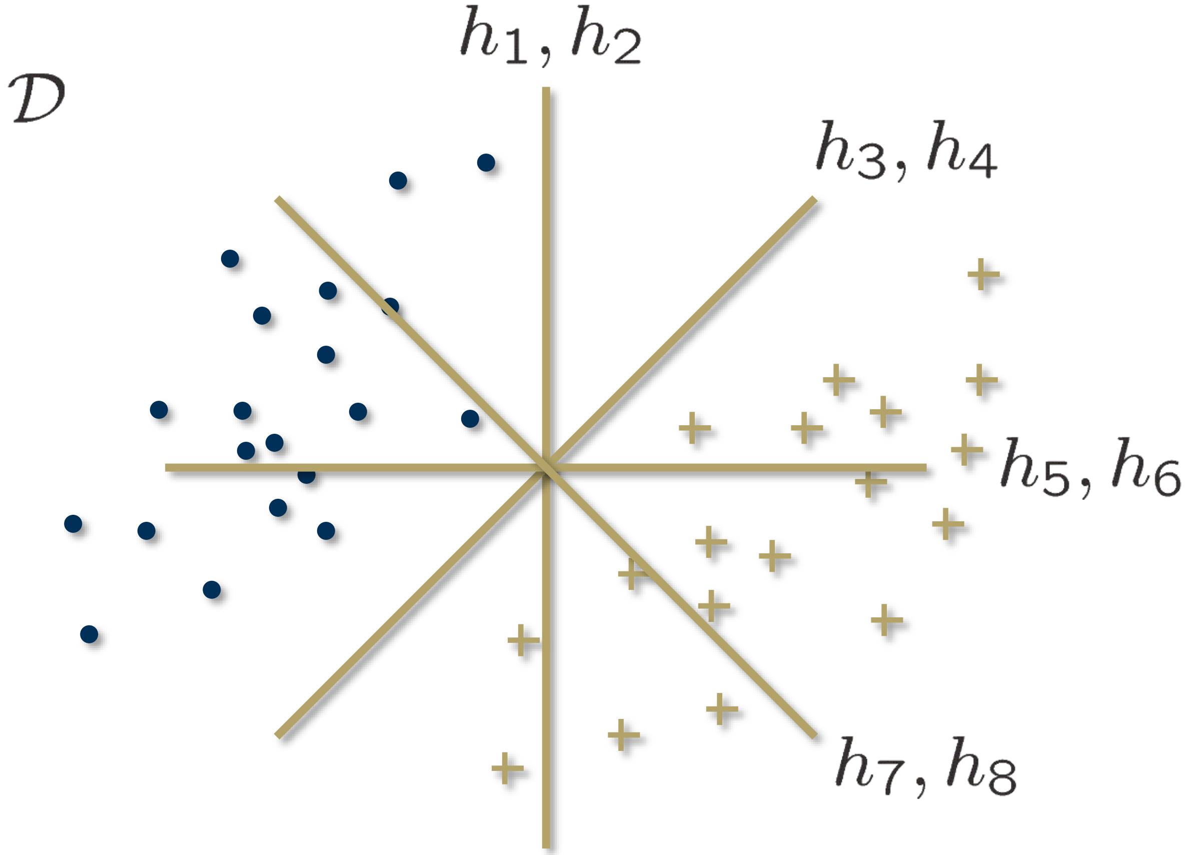
where each $\mathbf{x}_i \in \mathbb{R}^d$

Suppose we are given a list of possible hypotheses

$$\mathcal{H} = \{h_1, \ldots, h_m\}$$

From the ***training data*** $\mathcal{D}$, we would like to select the best possible hypothesis from $\mathcal{H}$

Example



$$\mathcal{H} = \{h_1, \ldots, h_8\}$$

Empirical risk

Recall our definition of ***risk*** and its empirical counterpart

Risk: $R(h_j) := \mathbb{P}[h_j(X) \neq Y]$

Empirical risk: $\widehat{R}_n(h) := \frac{|\{i : h(\mathbf{x}_i) \neq y_i\}|}{n}$

The empirical risk $\widehat{R}_n(h_j)$ gives us an estimate of the true risk $R(h_j)$, and from the law of large numbers we know that $\widehat{R}_n(h_j) \rightarrow R(h_j)$ as $n \rightarrow \infty$

We should be able to use the empirical risk to choose a good hypothesis

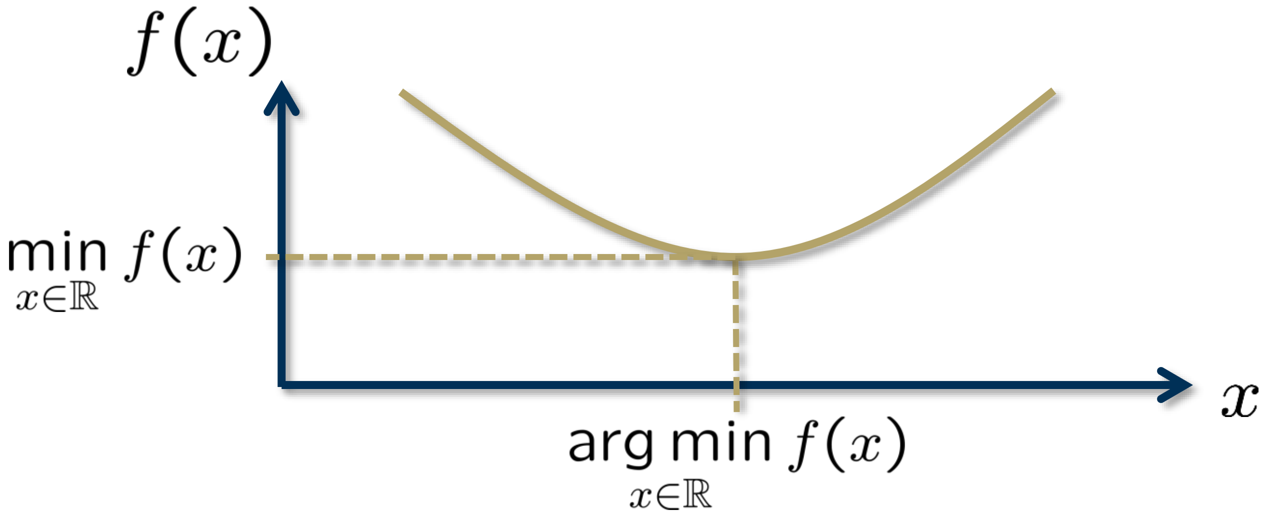
Empirical risk minimization (ERM)

We want to choose a hypothesis from $\mathcal{H}$ that achieves a small risk

Since $\widehat{R}_n(h_j)$ is supposed to be a good estimate of $R(h_j)$, an incredibly natural (and common) strategy is to pick

$$h^* = \operatorname*{arg\,min}_{h_j \in \mathcal{H}} \widehat{R}_n(h_j)$$

Aside:


$f(x)$
$\min_{x \in \mathbb{R}} f(x)$
x
$\operatorname*{arg\,min}_{x \in \mathbb{R}} f(x)$

The risk in ERM

As long as we have enough data, for any particular hypothesis h_j, we expect $\widehat{R}_n(h_j) \approx R(h_j)$

However, if m is very large, then we can also expect that there are some h_k for which $\widehat{R}_n(h_k) \ll R(h_k)$

Thus, what can we say about $R(h^*)$?

- We know that $\widehat{R}_n(h^*)$ is as small as it can be
  - this ***could*** be because $R(h^*)$ is small
  - ***or***, it could be because $\widehat{R}_n(h_k) \ll R(h_k)$ for some h_k
- Which explanation is more likely?
  - ***it depends…*** just how large is m?

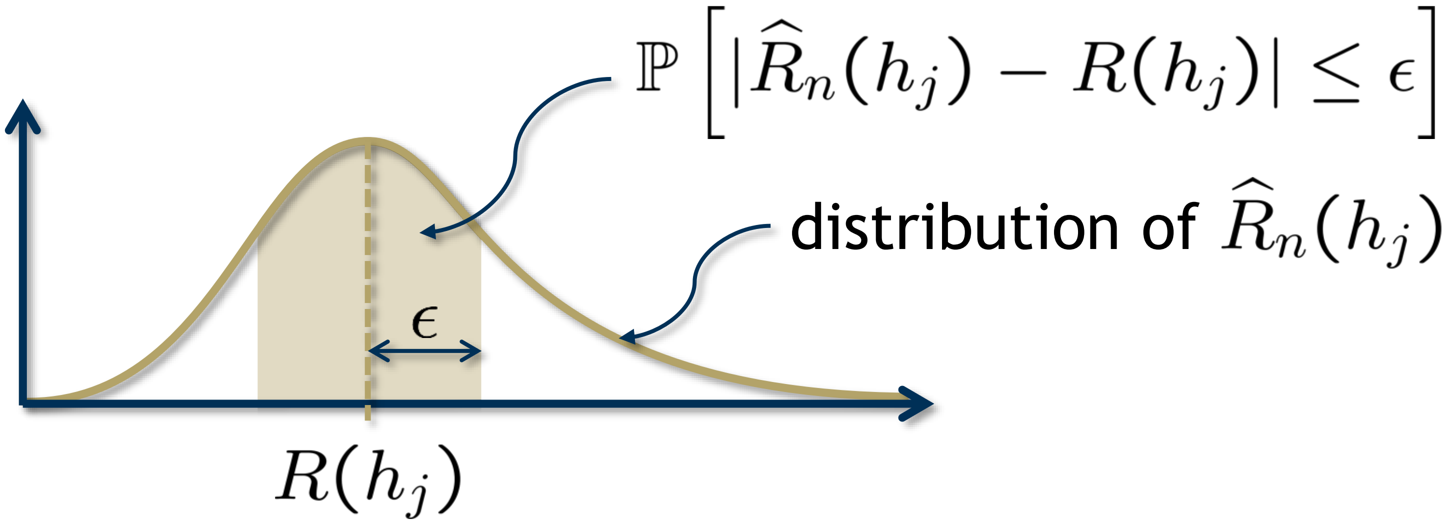
Confidence bounds

One way to provide guarantees for the ERM approach is to set m and n such that

$$|\widehat{R}_n(h_j) - R(h_j)| \leq \epsilon$$

for all j (and for some suitably small choice of ϵ)

Of course, we can never guarantee that this holds, so instead we will be concerned with the ***probability*** that $|\widehat{R}_n(h_j) - R(h_j)| \leq \epsilon$



$\mathbb{P}\left[|\widehat{R}_n(h_j) - R(h_j)| \leq \epsilon\right]$

distribution of $\widehat{R}_n(h_j)$

Too much randomness?

Ultimately, we will want to show something like

$$\mathbb{P}\left[|\widehat{R}_n(h_j) - R(h_j)| \leq \epsilon\right] \approx 1$$

for all $j = 1, \ldots, m$

What is random here?

- the training data $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$
- $\widehat{R}_n(h_1), \widehat{R}_n(h_2), \ldots, \widehat{R}_n(h_m)$, because each depends on $\mathcal{D}$
- h^*, because it depends on $\widehat{R}_n(h_1), \widehat{R}_n(h_2), \ldots, \widehat{R}_n(h_m)$

In order to tease all of this apart, let's begin by going back to just a single hypothesis h_j and studying

$$\mathbb{P}\left[|\widehat{R}_n(h_j) - R(h_j)| \leq \epsilon\right]$$

Bounding the error

We want to calculate

$$\mathbb{P}\left[|\widehat{R}_n(h_j) - R(h_j)| \leq \epsilon\right]$$

Note that $\widehat{R}_n(h_j)$ is a random variable

- we can write $\widehat{R}_n(h_j) = \frac{1}{n}\sum_{i=1}^n S_i$ where the S_i are ***Bernoulli*** random variables
- thus, $n\widehat{R}_n(h_j)$ is a ***Binomial*** random variable
- since $\mathbb{P}[S_i = 1] = \mathbb{P}[h_j(\mathbf{x}_i) \neq y_i] = R(h_j)$, we have that

$$\begin{aligned}\mathbb{E}\left[n\widehat{R}_n(h_j)\right] = \mathbb{E}\left[\sum_{i=1}^n S_i\right] &= \sum_{i=1}^n \mathbb{E}\left[S_i\right] \\ &= n\mathbb{P}\left[h_j(\mathbf{x}_i) \neq y_i\right] \\ &= nR(h_j)\end{aligned}$$

Deviation from the mean

Thus, an equivalent way to think about our problem is that we would like to calculate

$$\mathbb{P}\left[|n\widehat{R}_n(h_j) - nR(h_j)| \leq n\epsilon\right]$$

and this is just asking about the probability that a Binomial random variable will be within $n\epsilon$ of its mean

If $F(k)$ represents the cumulative distribution function (CDF) of our binomial random variable, then we can write

$$\begin{aligned}\mathbb{P}\left[|n\widehat{R}_n(h_j) - nR(h_j)| \leq n\epsilon\right]\\ = F(nR(h_j) + n\epsilon) - F(nR(h_j) - n\epsilon)\end{aligned}$$

Bounding the deviation

Unfortunately, the CDF we are interested in is given by

$$F(k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} R(h_j)^i (1 - R(h_j))^{n-i}$$

This has no nice closed form expression, and is rather unwieldy to work with and doesn't give us much intuition

Instead of calculating the probability exactly, it is enough to get a good bound of the form

$$\mathbb{P}\left[|\widehat{R}_n(h_j) - R(h_j)| \leq \epsilon\right] \geq 1 - ?$$

or equivalently

$$\mathbb{P}\left[|\widehat{R}_n(h_j) - R(h_j)| \geq \epsilon\right] \leq ?$$

Concentration inequalities

An inequality of the form

$$\mathbb{P}\left[|\widehat{R}_n(h_j) - R(h_j)| \geq \epsilon\right] \leq ?$$

tell us how a particular random variable (in this case $\widehat{R}_n(h_j)$) ***concentrates*** around its mean

There are ***many*** different concentration inequalities that give us various bounds along these lines

We will start with a very simple one, and then build up to a stronger result

Markov's inequality

The simplest of these results is ***Markov's inequality***

Let X be any nonnegative random variable.
Then for any $t \geq 0$,

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

This is cool on its own, but can be leveraged to say even more since for any strictly monotonically increasing and nonnegative-valued function ϕ

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

Chebyshev's inequality

As an example, ***Chebyshev's inequality*** states that for any random variable X,

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

Proof.

Note that $|X - \mathbb{E}[X]|$ is a nonnegative random variable. Thus we can apply Markov's inequality to obtain

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq \epsilon] = \mathbb{P}[|X - \mathbb{E}[X]|^2 \geq \epsilon^2]$$

$$\leq \frac{\mathbb{E}\left[|X - \mathbb{E}[X]|^2\right]}{\epsilon^2} = \frac{\mathrm{var}(X)}{\epsilon^2}$$

Proof of Markov (Part 1)

There is a ***simple*** proof of Markov if you know the (*super useful!*) fact that for any nonnegative random variable X

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X \geq x] dx$$

Proof. We can write $X = \int_0^X dx = \int_0^\infty 1_{\{x \leq X\}}(x) dx$ where

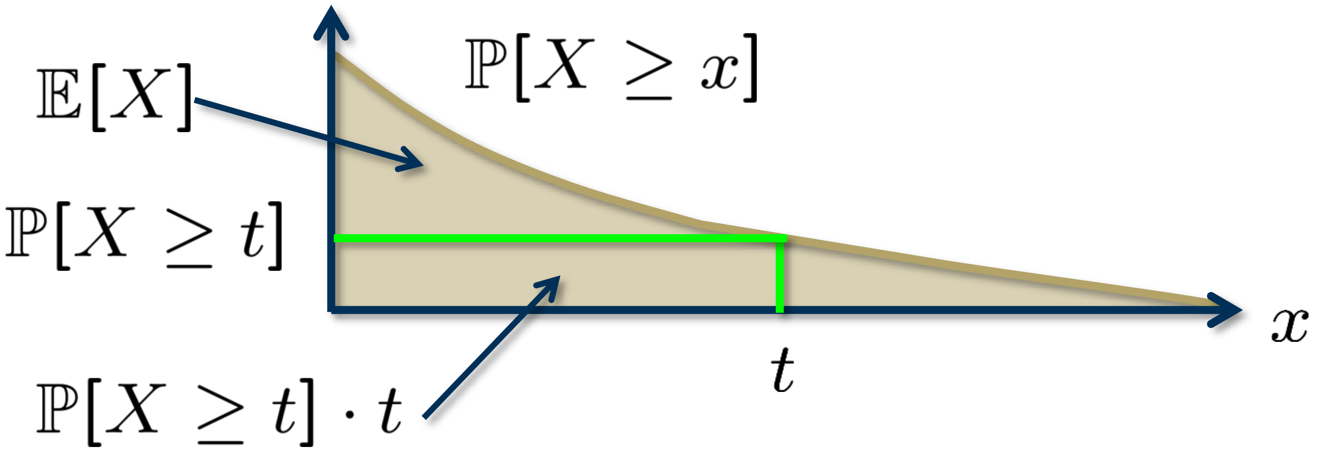
$$1_{\{A\}}(t) = \begin{cases} 1 & \text{if } t \in A \\ 0 & \text{if } t \notin A. \end{cases}$$

Thus $\mathbb{E}[X] = \mathbb{E}\left[\int_0^\infty 1_{\{x \leq X\}}(x) dx\right]$

$$= \int_0^\infty \mathbb{E}\left[1_{\{x \leq X\}}(x)\right] dx = \int_0^\infty \mathbb{P}\left[X \geq x\right] dx$$

Proof of Markov (Part 2)

We can visualize this result as



Thus, we can immediately see that we must have

$$\mathbb{E}[X] \geq \mathbb{P}[X \geq t] \cdot t$$

and hence

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

Hoeffding’s inequality

Chebyshev’s inequality gives us the kind of result we are after, but it is too *loose* to be of practical use

Hoeffding’s inequality assumes a bit more about our random variable beyond having finite variance, but gets us a much tighter and more useful result:

Let $X_1, \ldots, X_n$ be independent bounded random variables, i.e., random variables such that $\mathbb{P}[X_i \in [a,b]] = 1$ for all i

Let $S_n = \sum_{i=1}^n X_i$. Then for any $\epsilon > 0$, we have

$$\mathbb{P}\left[|S_n - \mathbb{E}[S_n]| \geq \epsilon\right] \leq 2e^{-2\epsilon^2/n(b-a)^2}$$

Chernoff's bounding method

To prove this result, we will use a similar approach as in Chebyshev's inequality

To begin consider only the upper tail inequality:

$$\begin{aligned}
\mathbb{P}[S_n - \mathbb{E}[S_n] \geq \epsilon] &= \mathbb{P}[\lambda(S_n - \mathbb{E}[S_n]) \geq \lambda\epsilon] && (\lambda > 0) \\
&= \mathbb{P}[e^{\lambda(S_n - \mathbb{E}[S_n])} \geq e^{\lambda\epsilon}] \\
&\leq \frac{\mathbb{E}\left[e^{\lambda(S_n - \mathbb{E}[S_n])}\right]}{e^{\lambda\epsilon}} && \text{(Markov)} \\
&= e^{-\lambda\epsilon}\mathbb{E}\left[e^{\lambda(X_1 - \mathbb{E}[X_1] + \cdots + X_n - \mathbb{E}[X_n])}\right] \\
&= e^{-\lambda\epsilon}\prod_{i=1}^{n}\mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}[X_i])}\right] && \text{(Independence)}
\end{aligned}$$

Hoeffding's Lemma

It is not obvious, but also not too hard to show, that

$$\mathbb{E}\left[e^{\lambda(X_i-\mathbb{E}[X_i])}\right] \leq e^{\lambda^2(b-a)^2/8}$$

(proof uses convexity and then gets a bound using a Taylor series expansion)

Plugging this in, we obtain that for any $\lambda > 0$, we have

$$\mathbb{P}[S_n - \mathbb{E}[S_n] \geq \epsilon] \leq e^{-\lambda\epsilon} e^{n\lambda^2(b-a)^2/8}$$

By setting $\lambda = 4\epsilon/n(b-a)^2$, we have

$$\begin{aligned} \mathbb{P}[S_n - \mathbb{E}[S_n] \geq \epsilon] &\leq e^{-4\epsilon^2/n(b-a)^2} e^{2\epsilon^2/n(b-a)^2} \\ &= e^{-2\epsilon^2/n(b-a)^2} \end{aligned}$$

Putting it all together

Thus, we have proven that

$$\mathbb{P}[S_n - \mathbb{E}[S_n] \geq \epsilon] \leq e^{-2\epsilon^2/n(b-a)^2}$$

An analogous argument proves

$$\mathbb{P}[\mathbb{E}[S_n] - S_n \geq \epsilon] \leq e^{-2\epsilon^2/n(b-a)^2}$$

Combined, these give

$$\mathbb{P}\left[|S_n - \mathbb{E}[S_n]| \geq \epsilon\right] \leq 2e^{-2\epsilon^2/n(b-a)^2}$$

Special case: Binomials

If the X_i are Bernoulli random variables, then S_n is a Binomial random variable and Hoeffding's inequality becomes

$$\mathbb{P}\left[|S_n - \mathbb{E}[S_n]| \geq \epsilon\right] \leq 2e^{-2\epsilon^2/n}$$

Finally going back to our original problem, this means that Hoeffding yields the bound

$$\mathbb{P}\left[|\widehat{R}_n(h_j) - R(h_j)| \geq \epsilon\right] = \mathbb{P}\left[|n\widehat{R}_n(h_j) - nR(h_j)| \geq n\epsilon\right]$$

$$\leq 2e^{-2\epsilon^2 n}$$

Multiple hypotheses

Thus, after much effort, we have that for a particular hypothesis h_j,

$$\mathbb{P}\left[|\widehat{R}_n(h_j) - R(h_j)| \geq \epsilon\right] \leq 2e^{-2\epsilon^2 n}$$

However, we are ultimately interested in h^*, not just a single hypothesis h_j

One way to argue that $|\widehat{R}_n(h^*) - R(h^*)| \leq \epsilon$ is to ensure that $|\widehat{R}_n(h_j) - R(h_j)| \leq \epsilon$ ***simultaneously*** for all j

Equivalently, we can try to bound the probability that ***any*** hypothesis h_j has an empirical risk that deviates from its mean by more than ϵ

Formal statement

We can express this mathematically as

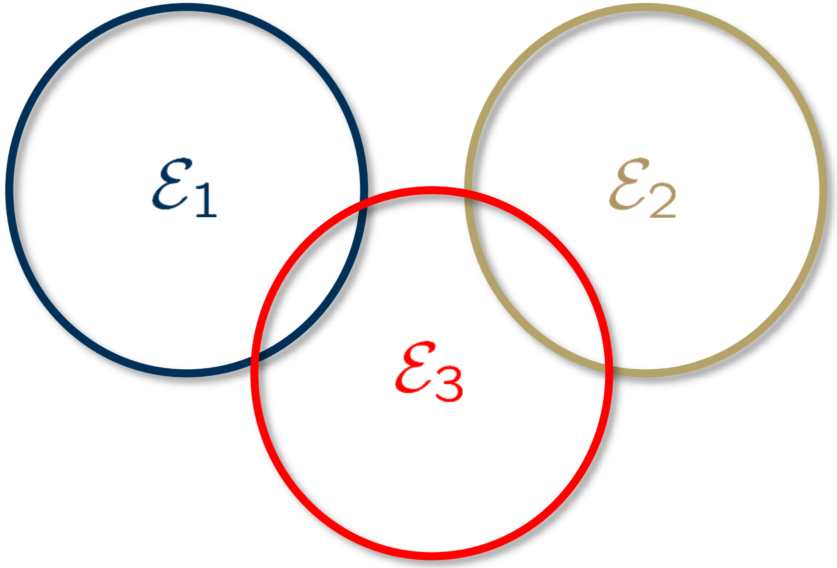
$$\begin{aligned}\mathbb{P}\left[\left|\widehat{R}_n(h^*) - R(h^*)\right| \geq \epsilon\right] \leq \mathbb{P}\Big[\quad & \left|\widehat{R}_n(h_1) - R(h_1)\right| \geq \epsilon \\ \textbf{or } & \left|\widehat{R}_n(h_2) - R(h_2)\right| \geq \epsilon \\ & \vdots \\ \textbf{or } & \left|\widehat{R}_n(h_m) - R(h_m)\right| \geq \epsilon \Big]\end{aligned}$$

We can bound this using something called the ***union bound***

Union bound

Union bound For any sequence of events $\mathcal{E}_1, \ldots, \mathcal{E}_m$

$$\mathbb{P}\left[\mathcal{E}_1 \cup \cdots \cup \mathcal{E}_m\right] \leq \mathbb{P}\left[\mathcal{E}_1\right] + \cdots + \mathbb{P}\left[\mathcal{E}_m\right]$$



The events in our case are given by

$$\mathcal{E}_j = \left|\widehat{R}_n(h_j) - R(h_j)\right| \geq \epsilon$$

Final result

$$\begin{aligned}
\mathbb{P}\left[\left|\widehat{R}_n(h^*) - R(h^*)\right| \geq \epsilon\right] \leq \mathbb{P}\Big[\quad & \left|\widehat{R}_n(h_1) - R(h_1)\right| \geq \epsilon \\
\text{or } & \left|\widehat{R}_n(h_2) - R(h_2)\right| \geq \epsilon \\
& \vdots \\
\text{or } & \left|\widehat{R}_n(h_m) - R(h_m)\right| \geq \epsilon \Big] \\
& \leq \sum_{j=1}^{m} \mathbb{P}\left[\left|\widehat{R}_n(h_j) - R(h_j)\right| > \epsilon\right] \\
& \leq \sum_{j=1}^{m} 2e^{-2\epsilon^2 n} \\
& = 2me^{-2\epsilon^2 n}
\end{aligned}$$

Interpretation

We went through all of this work to show that

$$\mathbb{P}\left[\left|\widehat{R}_n(h^*) - R(h^*)\right| \geq \epsilon\right] \leq 2me^{-2\epsilon^2 n}$$

linearly increasing

exponentially decreasing

This suggests that ERM is a reasonable approach as long as m isn't ***too*** big (i.e., $m \lessapprox e^n$)

Note that the above is equivalent to the statement that with probability at least $1 - \delta$,

$$|\widehat{R}_n(h^*) - R(h^*)| \leq \sqrt{\frac{1}{2n} \log(2m/\delta)}$$

Bounding the excess risk

Note that we would ***ideally*** actually like to choose

$$h^\sharp = \arg\min_{h_j \in \mathcal{H}} R(h_j)$$

We can also relate the performance of h^* to $h^\sharp$:

$$\begin{aligned} R(h^*) - R(h^\sharp) &= R(h^*) - \widehat{R}_n(h^*) + \widehat{R}_n(h^*) - R(h^\sharp) \\ &\leq |R(h^*) - \widehat{R}_n(h^*)| + |\widehat{R}_n(h^*) - R(h^\sharp)| \end{aligned}$$

We have already shown that with probability at least $1 - \delta$

$$|\widehat{R}_n(h^*) - R(h^*)| \leq \sqrt{\tfrac{1}{2n} \log(2m/\delta)}$$

What about $|\widehat{R}_n(h^*) - R(h^\sharp)|$?

Bounding the excess risk

We will bound $|\widehat{R}_n(h^*) - R(h^\sharp)|$ in two steps…

- $R(h^\sharp)$ cannot be too much bigger than $\widehat{R}_n(h^*)$:

  By the definition of $h^\sharp$; $R(h^\sharp) \leq R(h^*)$

  From before, we have $R(h^*) \leq \widehat{R}_n(h^*) + \sqrt{\frac{1}{2n}\log(2m/\delta)}$

  Thus $R(h^\sharp) - \widehat{R}_n(h^*) \leq \sqrt{\frac{1}{2n}\log(2m/\delta)}$

- $\widehat{R}_n(h^*)$ cannot be too much bigger than $R(h^\sharp)$:

  By the definition of h^*; $\widehat{R}_n(h^*) \leq \widehat{R}_n(h^\sharp)$

  From before, we have $\widehat{R}_n(h^\sharp) \leq R(h^\sharp) + \sqrt{\frac{1}{2n}\log(2m/\delta)}$

  Thus $\widehat{R}_n(h^*) - R(h^\sharp) \leq \sqrt{\frac{1}{2n}\log(2m/\delta)}$

The upshot

Thus,

$$\begin{aligned} R(h^*) - R(h^\sharp) &\leq |R(h^*) - \widehat{R}_n(h^*)| + |\widehat{R}_n(h^*) - R(h^\sharp)| \\ &\leq 2\sqrt{\frac{1}{2n}\log(2m/\delta)} \end{aligned}$$

Bottom line: As long as m isn't too big ($m \lessapprox e^n$) then we can be reasonably confident that $R(h^*)$ isn't too much larger than $R(h^\sharp)$

Of course, the trick in doing a good job of learning is ensure that $R(h^\sharp)$ is actually ***small***

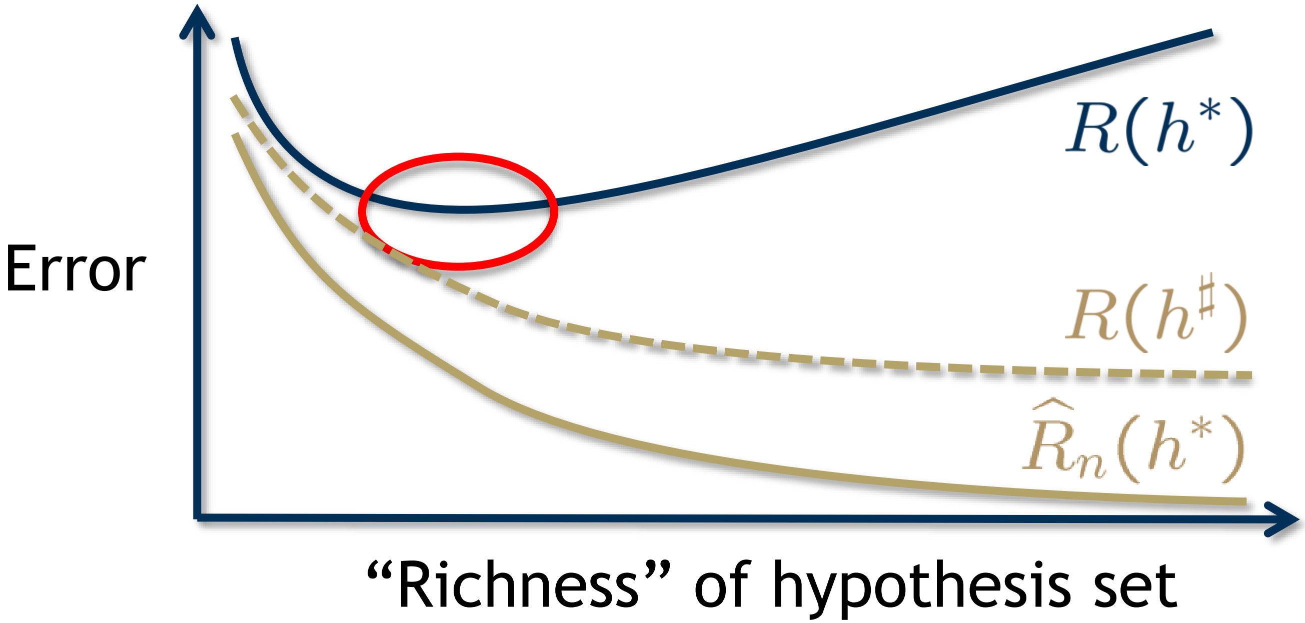
To achieve this, we need a "rich" set of possible hypotheses…

unfortunately…

Fundamental tradeoff

More hypotheses ultimately sacrifices our guarantee that $\widehat{R}_n(h^*) \approx R(h^*)$, which causes the whole argument to break

$$\text{Richer set of hypotheses} \longrightarrow \begin{cases} \widehat{R}_n(h^*) \downarrow \qquad R(h^\sharp) \downarrow \\ \widehat{R}_n(h^*) - R(h^*) \uparrow \end{cases}$$


Error
$R(h^*)$
$R(h^\sharp)$
$\widehat{R}_n(h^*)$
"Richness" of hypothesis set