Empirical Risk Minimization

In the last lecture, we saw how the classification problem was solved in two completely different frameworks.

When we have a perfect probability model, that is we know that joint distribution $f_{X,Y}(x, y)$ of the data points X and response variables Y, we know exactly what the best classifier is going to be (the Bayes classifier), and we don't need any training data to figure it out.

In contrast, the nearest-neighbor classifier let the data do all the talking. Not only was the classification rule derived from the data, but the algorithm itself relied on the data in a very explicit way.

Many problems in statistical learning fall somewhere in between these two extremes. We observe a series of (data points, responses) $(X_n = \boldsymbol{x}_n, Y_n = y_n)$, and we want to discover a relationship between them. Rather than assume that we have intimate knowledge of the distribution between X and Y, and deriving a single, analytical answer, we directly test a *family* of relationships (or **hypotheses**, as they are called in statistics), and use the one that works the best.

Let's make all of this a little more precise. The data points X are random vectors in $\mathbb{R}^D$, and the response variables Y are random variables in some subset $\mathcal{Y}$ of $\mathbb{R}$. We want to find a mapping $h : \mathbb{R}^D \to \mathcal{Y}$ such that

$$h(X) \approx Y.$$

When this rule is learned from examples, this is the prototypical "supervised learning" problem.

We have seen two examples where we have assumed perfect knowledge of the joint distribution $f_{X,Y}(\boldsymbol{x}, y)$, and then derived a clean,

optimal answer to this problem. The first was the MMSE estimator, where we predict Y from X using the conditional mean $\mathrm{E}[Y|X]$. The other was the Bayes classifier, where we selected the *a posteriori* most likely value of Y given X. Both of these results were derived by writing down exactly what we wanted, setting up an optimization problem, then solving it.

Generalizing this procedure gives us a clear path for moving from relying on knowledge of $f_{X,Y}(\boldsymbol{x}, y)$ to being completely data-driven. The first thing to do is to write down exactly how we will measure how well h explains the relationship between X and Y. This is done with a **loss function** $\ell : \mathbb{R}^D \times \mathcal{Y} \to \mathbb{R}$:

$$\text{loss for } h \text{ at } (\boldsymbol{x}, y) \quad = \quad \ell(h(\boldsymbol{x}), y).$$

This, for example, could be the squared-error,

$$\ell(h(\boldsymbol{x}), y) = (h(\boldsymbol{x}) - y)^2,$$

or an indicator[1] telling us whether $h(\boldsymbol{x})$ is different than y,

$$\ell(h(\boldsymbol{x}), y) = \begin{cases} 0, & h(\boldsymbol{x}) = y, \\ 1, & h(\boldsymbol{x}) \neq y, \end{cases},$$

or any one of a number of other things.

The loss function characterizes the performance of h at a single point. The **risk** of a mapping h is simply its performance averaged over all points:

$$R(h) = \mathrm{E}[\ell(h(X), Y)].$$

[1] This is called the "0/1" loss. It really only makes sense as a performance measure when $\mathcal{Y}$ is discrete.

Obviously, to compute $R(h)$, we need to know the joint distribution $f_{X,Y}(\boldsymbol{x}, y)$. Given a joint distribution on (X, Y), it is a fair question to ask which h *minimizes the risk*.

Pop Quiz: Suppose

$$\ell(h(\boldsymbol{x}), y) = (h(\boldsymbol{x}) - y)^2,$$

and we know the joint distribution $f_{X,Y}(\boldsymbol{x}, y)$. What is the solution to

$$\underset{h:\mathbb{R}^D \to \mathcal{Y}}{\text{minimize}} \;\; \mathrm{E}[\ell(h(X), Y)] \;\; ?$$

Answer:

$$h(\boldsymbol{x}) =$$

Here is another familiar example. Suppose that $\mathcal{Y} = \{1, \ldots, K\}$ and we use the 0/1 loss

$$\ell(h(\boldsymbol{x}), y) = \begin{cases} 0, & h(\boldsymbol{x}) = y, \\ 1, & h(\boldsymbol{x}) \neq y. \end{cases}$$

So when $h(\boldsymbol{x}) = y$, this means we have assigned the right "class" to $\boldsymbol{x}$. If we know the joint distribution $f_{X,Y}(\boldsymbol{x}, y)$, then the optimal classifier under this metric can be computed by solving

$$\underset{h:\mathbb{R}^D \to \mathcal{Y}}{\text{minimize}} \;\; \mathrm{E}[\ell(h(X), Y)].$$

We have already seen that mapping which minimizes the risk is

$$h^*(\boldsymbol{x}) = \arg\max_{k \in \mathcal{Y}} \;\; \mathrm{P}\left(Y = k | X = \boldsymbol{x}\right).$$

We gave this the name "Bayes classifier".

For those two loss functions, we were able to compute analytical solutions. In practice, even if we know the joint distribution $f_{X,Y}(\boldsymbol{x}, y)$, there might not be a closed form solution for a particular loss function ℓ. In this case, we will need to solve the optimization program numerically to find the best mapping h. To make this tractable, we will have to discretize/parameterize the set of hypotheses that we search over. This inherently limits our search to some set $\mathcal{H}$ of candidate mappings.

In general, the process of **risk minimization** takes a known joint distribution $f_{X,Y}(\boldsymbol{x}, y)$, a loss function ℓ, and a set of candidate mappings (a set of hypotheses) $\mathcal{H}$ and solves

$$\underset{h \in \mathcal{H}}{\text{minimize}} \quad \mathrm{E}[\ell(h(X), Y)]$$

Examples of $\mathcal{H}$ that we might use are

- All functions that lie in the linear span of some set of basis functions $\psi_1(\boldsymbol{x}), \ldots, \psi_N(\boldsymbol{x})$. This means that $\mathcal{H}$ is a finite-dimensional linear subspace of $L_2(\mathbb{R}^D)$. Formally, we could specify $\mathcal{H}$ by

$$\mathcal{H} = \left\{ h \in L_2(\mathbb{R}^D) \ : \ h(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i \psi_i(\boldsymbol{x}), \quad \text{for some } \alpha_i \in \mathbb{R} \right\}$$

- All functions in a reproducing kernel Hilbert space. In this case, $\mathcal{H}$ will be infinite dimensional, but we have seen how solving optimization programs in this setting can still be tractable (e.g. kernel regression). An example of how $\mathcal{H}$ might be formalized

in this setting would be

$$\mathcal{H} = \Bigg\{ h \in L_2(\mathbb{R}^D) \ : \ h(\boldsymbol{x}) = \sum_{i=1}^{M} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i), \\ \text{for some } \ M \in \mathbb{Z}, \ \boldsymbol{x}_i \in \mathbb{R}^D, \ \alpha_i \in \mathbb{R} \text{ and kernel } k(\cdot,\cdot) \Bigg\}$$

- All binary valued functions that take the values $(-1, 1)$ on either side of a hyperplane:

$$\mathcal{H} = \left\{ h \ : \ h(\boldsymbol{x}) = \text{sign}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b\right), \ \boldsymbol{w} \in \mathbb{R}^D, \ b \in \mathbb{R} \right\}.$$

Note the important fact that $\mathcal{H}$ in general consists of an infinite number of mappings, even when the set is parameterized by a finite number of parameters (as in the first and third examples above).

Learning h from data

In practice, it is often the case that we have no idea what $f_{X,Y}(\boldsymbol{x}, y)$ is. But we might have a decent number of examples, or "training data", $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$.

One strategy would then be to use this data to estimate the joint pdf $f_{X,Y}(\boldsymbol{x}, y)$, then proceed as above. To get a reliable estimate, you would of course need some kind of parameterization of the joint pdf. You would estimate these parameters (using MLE, etc), and then compute the h has optimally low risk under this estimated model. This is a very classical way of thinking that goes under the name "plug-in methods".

But there is a better way, and why this way is better was perhaps not fully appreciated until the 1970s or 1980s — relatively recent, in the long arc of the history of statistics. The idea is simple: we replace the "true" risk

$$R(h) = \mathrm{E}[\ell(h(X), Y)],$$

with the **empirical risk**

$$\hat{R}_N(h) = \frac{1}{N} \sum_{n=1}^{N} \ell(h(\boldsymbol{x}_n), y_n),$$

then choose a hypothesis by solving[2]

$$\underset{h \in \mathcal{H}}{\text{minimize}} \ \hat{R}_N(h) \quad = \quad \underset{h \in \mathcal{H}}{\text{minimize}} \ \frac{1}{N} \sum_{n=1}^{N} \ell(h(\boldsymbol{x}_n), y_n).$$

For a fixed hypothesis h, the weak law of large numbers tells us that when the observations $(X_n, Y_n) = (\boldsymbol{x}_n, y_n)$ are independent,

$$\frac{1}{N} \sum_{n=1}^{N} \ell(h(\boldsymbol{x}_n), y_n) \ \rightarrow \ \mathrm{E}[\ell(h(X), Y)],$$

and so minimizing $\hat{R}_N(h)$ becomes the same as minimizing $R(h)$.

Now we can start to see the wisdom of this strategy, as estimating the mean of a random variable from a finite number of observations is orders of magnitude easier and more reliable than estimating a full joint density function. We are also making absolutely no assumptions about what the underlying distribution is, other than the fact that the observations are independent.

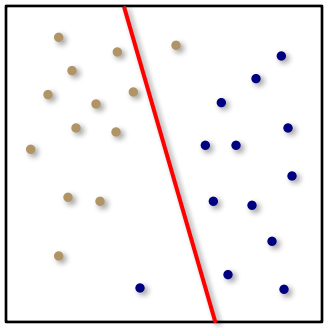
[2] Of course, the $1/N$ in the second expression can be dropped without changing the answer.

Example: Fitting a linear classifier. Say we are given points $(\boldsymbol{x}_n, y_n)$ with $\boldsymbol{x}_n \in \mathbb{R}^D$ and $y_n \in \{0, 1\}$. We want to find a linear classification rule. The natural loss here is 0/1:

$$\ell(h(\boldsymbol{x}), y) = 1_{h(\boldsymbol{x})=y} = \begin{cases} 0, & h(\boldsymbol{x}) = y, \\ 1, & h(\boldsymbol{x}) \neq y. \end{cases}$$

The empirical risk of a given h is simply the percentage of points that are mislabeled under the rule:

$$\hat{R}_N(h) = \frac{1}{N} \sum_{n=1}^{N} 1_{h(\boldsymbol{x}_n)=y_n}$$



The set of classifiers we are considering,

$$\mathcal{H} = \left\{ h \ : \ h(\boldsymbol{x}) = \text{sign}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b\right), \ \boldsymbol{w} \in \mathbb{R}^D, \ b \in \mathbb{R} \right\},$$

has $D + 1$ parameters, the vector $\boldsymbol{w} \in \mathbb{R}^D$ and the scalar $b \in \mathbb{R}$ — although, this is redundant in that we get exactly the same function by doubling $\|\boldsymbol{w}\|_2$ and halving b.

The empirical risk minimizer is simply the classifier that makes the fewest mistakes on the training data. This is straightforward conceptually, but actually finding this linear classifier on a general data set is computationally hard (i.e. NP hard).

Example: Linear regression. Now we are given points $(\boldsymbol{x}_n, y_n)$ with $y_n \in \mathbb{R}$, and want to fit a function to them. If we consider the class of linear functionals on $\mathbb{R}^D$,

$$\mathcal{H} = \left\{ h \ : \ h(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x} + b, \ \text{ for some } \boldsymbol{w} \in \mathbb{R}^D, b \in \mathbb{R} \right\},$$

and take the loss function to be the squared error,

$$\ell(h(\boldsymbol{x}), y) = (y - h(\boldsymbol{x}))^2,$$

then ERM is equivalent to solving the problem

$$\underset{\boldsymbol{w} \in \mathbb{R}^D, b \in \mathbb{R}}{\text{minimize}} \ \sum_{n=1}^{N} (y_n - \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x} - b)^2.$$

We have met an old friend; we know the solution to the above is

$$\begin{bmatrix} \hat{\boldsymbol{w}} \\ b \end{bmatrix} = (\boldsymbol{A}^{\mathrm{T}} \boldsymbol{A})^{-1} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{y},$$

where

$$\boldsymbol{A} = \begin{bmatrix} x_1[1] & x_1[2] & \cdots & x_1[D] & 1 \\ x_2[1] & x_2[2] & \cdots & x_2[D] & 1 \\ \vdots & & \ddots & & \vdots \\ x_N[1] & x_2[N] & \cdots & x_N[D] & 1 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}.$$

So ... does empirical risk minimization work?

To make this question more precise, let h^* be the truly-best hypothesis in our set, which could be found by minimizing the true risk

$$h^* = \arg\min_{h \in \mathcal{H}} \ \mathrm{E}[\ell(h(X), Y)]$$

if we had access to the joint distribution of (X, Y). This is, by definition, the best performing hypothesis on future unlabeled data. Now let $\hat{h}$ he the data-driven empirical risk minimizer

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \ \frac{1}{N} \sum_{n=1}^{N} \ell(h(\boldsymbol{x}_n), y_n)$$

which can be computed (at least in theory) without any knowledge of the joint distribution. Is the future performance of $\hat{h}$ comparable to h^*? That is, we know that

$$\mathrm{E}[\ell(\hat{h}(X), Y)] = R(\hat{h}) \ \geq \ R(h^*) = \mathrm{E}[\ell(h^*(X), Y)],$$

but is it at least close?

The answer is "yes, if you have enough data". What "enough data" means depends heavily on the classifier set $\mathcal{H}$, or more precisely, some measure of the *complexity* of $\mathcal{H}$. Making this statement precise is what is known as the **theory of generalization**, and requires a tremendous amount of complicated mathematics.

It's good that we are not scared of complicated mathematics ...

A first look at generalization

Let's see how this works when

1. $\mathcal{Y} = \{0, 1\}$ (i.e. binary classification),
2. we use the 0/1 loss, and
3. the hypothesis set $\mathcal{H}$ is *finite*.

The last assumption does not fit into any of the problems we have discussed so far — even the space of linear classifiers is infinite. But still, working out the generalization result in full is very instructive.

For your reference, here is a summary of the notation:

h^* = best possible classifier, found by minimizing true risk $R(h)$

$\hat{h}$ = best empirical classifier, found by minimizing emp. risk $\hat{R}_N(h)$

$R(h^*)$ = true risk (or *generalization error*) of the best classifier,
this is the ultimate performance limit we can expect

$R(\hat{h})$ = generalization error of ERM,
this is the quantity we want to compare to $R(h^*)$

$\hat{R}_N(h^*)$ = empirical risk of best classifier,
this how well h^* fits the actual data we observed

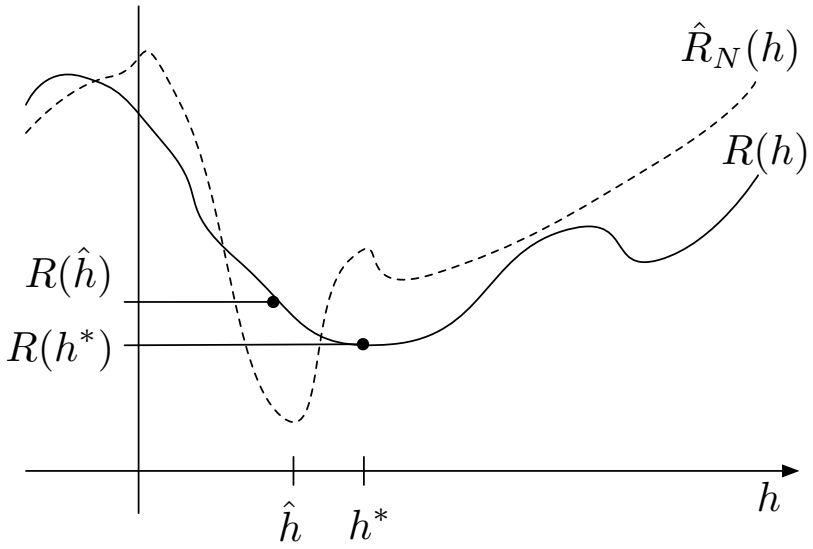
$\hat{R}_N(\hat{h})$ = empirical risk of ERM,
this is the "best fit" classifier to the actual data.

By definition, we know that

$$R(h^*) \leq R(\hat{h}), \quad \text{and} \quad \hat{R}_N(\hat{h}) \leq \hat{R}_N(h^*).$$

We are interested in the difference $R(\hat{h}) - R(h^*)$ — this tells us how the performance of the classifier found through ERM compares to the best possible classifier in $\mathcal{H}$.

Here is a picture that illustrates what is going on:



Both the empirical risk $\hat{R}_N$ and the true risk R are functions h (the classifier being considered). Of course, the mapping $h : \mathbb{R}^D \to \mathcal{Y}$ will usually be parameterized (in terms of basis expansion coefficients, weights, etc) — the above graph has one parameter (changing along the horizontal axis), but in general there will be many.

The $\hat{R}_N(h)$ curve above (the dotted line) of course depends on the data $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$. Since this data is random, you can think of this curve as a random function. The $R(h)$ curve (the black line) is not random; it is the true risk (or generalization error) for each of the classifiers being considered. It is the mean of the random function.

Fixed hypothesis

We will start by getting a feel for how well we can assess the risk for a particular classifier. With h fixed, we will be looking for a bound on $\hat{R}_N(h) - R(h)$. We can compute $\hat{R}_N(h)$ from the data, but $R(h)$ is unknown.

At this point, it is critical to realize that $\hat{R}_N(h)$ is a random variable, as it depends on the data $(\boldsymbol{x}_n, y_n)$ which is random. So our bounds will be probabilistic; we want something of the form

$$\mathrm{P}\left(|\hat{R}_N(h) - R(h)| \leq \epsilon\right) \;\geq\; ??,$$

or

$$\mathrm{P}\left(|\hat{R}_N(h) - R(h)| \geq \epsilon\right) \;\leq\; ??.$$

In both cases, the bound will depend on ϵ (as well as the number of data points N); in the first case, we are looking for the right hand side to be close to 1, in the second case, we are looking for the right hand side to be close to 0.

To get the bound, we will show that $\hat{R}_N(h)$ is a sum of independent random variables (this is easy), then show that $\mathrm{E}[\hat{R}_N(h)] = R(h)$ (also easy), and then develop a general-purpose probabilistic tail bound that quantifies how such a sum concentrates around its mean (this is hard).

We start by re-writing the empirical risk as a sum of independent random variables. Let

$$S_n = \begin{cases} 1, & h(\boldsymbol{x}_n) \neq y_i, \\ 0, & h(\boldsymbol{x}_n) = y_i. \end{cases}$$

Since the $(\boldsymbol{x}_n, y_n)$ are independent and identically distributed, the S_n are independent Bernoulli (i.e. binary-valued) random variables

with

$$\mathrm{P}\left(S_n = 1\right) = \mathrm{P}\left(h(\boldsymbol{x}_n) \neq y_i\right), \quad \mathrm{P}\left(S_n = 0\right) = 1 - \mathrm{P}\left(h(\boldsymbol{x}_n) \neq y_n\right).$$

A simple calculation reveals that

$$\mathrm{E}[S_n] = R(h).$$

By construction,

$$\hat{R}_N(h) = \frac{1}{N}\sum_{n=1}^{n} S_n, \tag{1}$$

and so

$$\mathrm{E}[\hat{R}_N(h)] = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}[S_n] = R(h).$$

We are left with the question: How close is the sum of independent random variables $\frac{1}{N}\sum_n S_n$ to its mean?

An answer to this question is given by the Hoeffding inequality:

Hoeffding Inequality. Let $X_1, \ldots, X_N$ be independent random variables that are bounded, meaning $a \leq X_n \leq b$ with probability 1. Let $Z_N = \sum_{n=1}^{N} X_n$. Then for any $\epsilon \geq 0$,

$$\mathrm{P}\left(|Z_N - \mathrm{E}[Z_N]| \geq \epsilon\right) \;\leq\; 2e^{-2\epsilon^2/N(b-a)^2}. \tag{2}$$

Applying this to $\hat{R}_N(h)$ in (1), with $a = 0, b = 1$, we have

$$\mathrm{P}\left(|N\hat{R}_N(h) - NR(h)| \geq N\epsilon\right) \;\leq\; 2e^{-2N\epsilon^2},$$

and so

$$\mathrm{P}\left(|\hat{R}_N(h) - R(h)| \geq \epsilon\right) \;\leq\; 2e^{-2N\epsilon^2}. \tag{3}$$

This gives us insight into how the performance of **one single** classification rule generalizes. What we want is some assurance that the one we judge to be the best, by performing ERM on the data, will be close to the best choice we could have made. We will get this assurance by developing a similar probability bound that holds **uniformly** over all classifiers in $\mathcal{H}$.

How close is the empirical minimizer to the true minimizer?

We have linked the performance of a single, fixed classifier to its true risk; this bound depended on the amount of data N that we have seen. Now we will use this result to get a **uniform** bound on how far the empirical risk deviates from the true risk for all $h \in \mathcal{H}$. Our analysis will be limited to the case where $\mathcal{H}$ is finite:

$$|\mathcal{H}| = M.$$

By re-arranging the main result (3) from the previous section, we see that with probability at least $1 - \delta$,

$$|\hat{R}_N(h) - R(h)| \leq \sqrt{\frac{1}{2N}\log(2/\delta)}.$$

But since our decision on which classifier was the best depended on the empirical risk of all of the classifiers in $\mathcal{H}$, we would like to make sure that there empirical performance was somewhat near their ideal performance. That is, we want to show that

$$\max_{h\in\mathcal{H}} |\hat{R}_N(h) - R(h)| \;\leq\; \epsilon, \tag{4}$$

with probability at least $1-\delta$ for some appropriate choice of ϵ and δ. We want to fill in the right hand side of

$$\mathrm{P}\left(\max_{h\in\mathcal{H}}|\hat{R}_N(h)-R(h)|>\epsilon\right) \;\leq\; ???.$$

We do this by applying the **union bound** to our expression for a single classifier. Recall the following fact from basic probability theory. If $\mathcal{A}_1,\ldots,\mathcal{A}_M$ are arbitrary events, then the probability of at least one of them occurring is less than the sum of their individual probabilities:

$$\mathrm{P}\left(\mathcal{A}_1 \text{ or } \mathcal{A}_2 \text{ or} \cdots \mathcal{A}_M\right) \;\leq\; \mathrm{P}\left(\mathcal{A}_1\right)+\mathrm{P}\left(\mathcal{A}_2\right)+\cdots+\mathrm{P}\left(\mathcal{A}_M\right).$$

As you know, the bound above holds with equality when the sets $\mathcal{A}_m$ are disjoint.

We can rewrite the event of interest as

$$\begin{aligned}\left\{\max_{h\in\mathcal{H}}|\hat{R}_N(h)-R(h)|>\epsilon\right\} = \Big\{|\hat{R}_N(h_1)-R(h_1)|>\epsilon\Big\} \text{ or}\\ \Big\{|\hat{R}_N(h_2)-R(h_2)|>\epsilon\Big\} \text{ or}\\ \vdots \qquad\qquad \vdots\\ \Big\{|\hat{R}_N(h_M)-R(h_M)|>\epsilon\Big\}.\end{aligned}$$

Thus

$$\begin{aligned}\mathrm{P}\left(\max_{h\in\mathcal{H}}|\hat{R}_N(h)-R(h)|>\epsilon\right) &\leq \sum_{j=1}^{M}\mathrm{P}\left(|\hat{R}_N(h_j)-R(h_j)|>\epsilon\right)\\ &\leq 2Me^{-2N\epsilon^2}.\end{aligned} \tag{5}$$

When the bound (4) holds, we can relate the generalization performance of the empirical risk minimizer $\hat{h}$ to the performance of the best possible choice h^*. We have[3]

$$\begin{aligned} R(\hat{h}) - R(h^*) &= R(\hat{h}) - \hat{R}_N(\hat{h}) + \hat{R}_N(\hat{h}) - R(h^*) \\ &\leq |R(\hat{h}) - \hat{R}_N(\hat{h})| + |\hat{R}_N(\hat{h}) - R(h^*)| \end{aligned}$$

The first term above is immediately controlled by (4). For the second term, we combine (4) with optimality of h^* and $\hat{h}$ in two different ways. Since h^* is the minimizer of the true risk,

$$R(h^*) \;\leq\; R(\hat{h}) \;\leq\; \hat{R}_N(\hat{h}) + \epsilon,$$

and since $\hat{h}$ is the minimizer of the empirical risk,

$$\hat{R}_N(\hat{h}) \;\leq\; \hat{R}_N(h^*) \;\leq\; R(h^*) + \epsilon.$$

Combining the two statements above gives us $|\hat{R}_N(\hat{h}) - R(h^*)| \leq \epsilon$, and so

$$\max_{h \in \mathcal{H}} |\hat{R}_N(h) - R(h)| \;\leq\; \epsilon \quad \Rightarrow \quad R(\hat{h}) - R(h^*) \;\leq\; 2\epsilon.$$

Putting it all together gives us our main result:

$$\mathrm{P}\left(R(\hat{h}) - R(h^*) > \epsilon\right) \;\leq\; 2Me^{-N\epsilon^2/2}.$$

This means that with probability at least $1 - \delta$,

$$R(\hat{h}) - R(h^*) \;\leq\; \sqrt{\frac{2}{N}\left(\log M + \log(2/\delta)\right)}.$$

[3]Note that $R(\hat{h}) - R(h^*)$ will always be positive.

ERM with finite $\mathcal{H}$. Let $\mathcal{H}$ be a set of classifiers with finite size $|\mathcal{H}| = M$. We are presented with N iid labeled data points $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$. Let $\hat{h}$ be the empirical risk minimizer,

$$\hat{h} = \arg\min_{h\in\mathcal{H}} \hat{R}_N(h) = \arg\min_{h\in\mathcal{H}} \frac{1}{N}\sum_{n=1}^N S_n(h),$$

where

$$S_n(h) = \begin{cases} 0, & h(\boldsymbol{x}_n) = y_n, \\ 1, & h(\boldsymbol{x}_n) \neq y_n. \end{cases}$$

Let h^* be the true risk minimizer

$$h^* = \arg\min_{h\in\mathcal{H}} R(h) = \arg\min_{h\in\mathcal{H}} \mathrm{P}\left(h(X) \neq Y\right).$$

Then with probability exceeding $1-\delta$

$$R(\hat{h}) - R(h^*) \;\leq\; \sqrt{\frac{2}{N}\left(\log M + \log(2/\delta)\right)}.$$

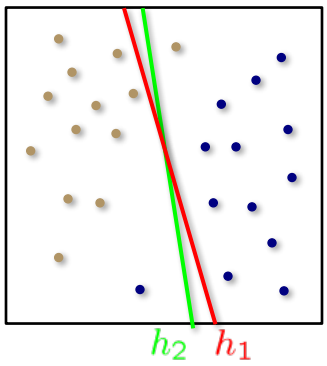
Complexity of $\mathcal{H}$

The results in the last section show that the number of samples we need to have the empirical risk minimizer $\hat{h}$ be on par with the best classifier grows like $\log M$. This can be interpreted as the *complexity* of the set $\mathcal{H}$ — as more hypothesis are added, the complexity grows.

But of course, we are often interested in hypothesis sets that have infinite size — even the simple set of linear classifiers on $\mathbb{R}^D$ has

an infinite number of elements. The generalization bound in the previous section tells us nothing the situation where $M = \infty$.

The problem with the analysis above is the application of the union bound in (5) above (where the max comes outside the probability as a sum). This bound is tight when the performance of each h is on the training data is approximately independent, but of course we know that is not the case — the performance of classifiers with very similar parameters are of course related:



For infinite sets of hypotheses, the argument requires something more refined than a union bound, this in turn leads to a more refined notion of the complexity of a set of mappings $\mathcal{H}$. In this analysis, the log M is replaced by something called the **VC dimension** of the set[4] $\mathcal{H}$. The definition of VC dimension is outside the scope of the course, but qualitatively it is a measure of the **statistical complexity** of $\mathcal{H}$.

[4]VC stands for Vapnik-Chervonenkis, the two people who birthed this theory.

Summary: Empirical risk minimization

- $X \in \mathbb{R}^D$ is a random ("feature") vector and $Y \in \mathcal{Y}$ is a random ("response") variable.
- We want to determine a *rule* h for predicting Y given $X = \boldsymbol{x}$.
- The loss $\ell(h(\boldsymbol{x}), y)$ quantifies the price I pay for being wrong.
- We want to choose the "best" function h from the class $\mathcal{H}$
- **Risk minimization**:
  When we know $f_{X,Y}(\boldsymbol{x}, y)$, we choose h by solving

$$\underset{h \in \mathcal{H}}{\text{minimize}} \;\; \mathrm{E}[\ell(h(X), Y)]$$

- We don't know $f_{X,Y}(\boldsymbol{x}, y)$, but are given examples $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$
- **Empirical risk minimization**:
  Replace the expectation by the sample mean and optimize

$$\underset{h \in \mathcal{H}}{\text{minimize}} \;\; \frac{1}{N} \sum_{n=1}^{N} \ell(h(\boldsymbol{x}_n), y_n)$$

- As $N \to \infty$,

$$\frac{1}{N} \sum_{n=1}^{N} \ell(h(\boldsymbol{x}_n), y_n) \;\; \to \;\; \mathrm{E}[\ell(h(X), Y)]$$

  uniformly over all $h \in \mathcal{H}$, so the solutions will be close

The last statement above might be considered the "fundamental theorem of machine learning" (or at least of supervised learning).

Technical Details: Proof of the Hoeffding Ineq.

We start with a basic question: how close is a single random variable X to its mean? This question is answered by applying the following basic result from probability theory.

Markov inequality. Let X be any non-negative random variable. Then for any $t \geq 0$,

$$\mathrm{P}\left(X \geq t\right) \;\leq\; \frac{\mathrm{E}[X]}{t}.$$

We encountered (and proved) the Markov inequality earlier in the course notes, when we were talking about the Weak Law of large numbers.

The Markov inequality actually tells us much more than what is in the box above. It is easily extended by realizing that for any function $\phi(x)$ which is non-negative and strictly monotonically increasing,

$$\mathrm{P}\left(X \geq t\right) = \mathrm{P}\left(\phi(X) \geq \phi(t)\right).$$

We now have any number of ways to modify the bound, as

$$\mathrm{P}\left(X \geq t\right) \;\leq\; \frac{E[\phi(X)]}{\phi(t)},$$

for any such ϕ. Moreover, the above holds for general random variables X, as we only need $\phi(X) \geq 0$ to apply Markov.

A **Chernoff bound** is simply an application of Markov with $\phi(t) = e^{\lambda t}$ for some $\lambda > 0$:

$$\mathrm{P}\left(X \geq t\right) \;\leq\; e^{-\lambda t}\,\mathrm{E}[e^{\lambda X}].$$

This is particularly useful when X is a sum of independent random variables. For instance, suppose that $Z_1, Z_2, \ldots, Z_N$ are iid random variables. Then the Chernoff bound on their sum is

$$\begin{aligned}
\mathrm{P}\left(Z_1 + \cdots + Z_N \geq t\right) &\leq e^{-\lambda t}\, \mathrm{E}[e^{\lambda(Z_1+\cdots+Z_N)}] \\
&= e^{-\lambda t}\, \mathrm{E}[e^{\lambda Z_1} e^{\lambda Z_2} \cdots e^{\lambda Z_N}] \\
&= e^{-\lambda t}\, \mathrm{E}[e^{\lambda Z_1}]\, \mathrm{E}[e^{\lambda Z_2}] \cdots \mathrm{E}[e^{\lambda Z_N}] \quad \text{(independence)} \\
&= e^{-\lambda t} \left(\mathrm{E}[e^{\lambda Z_1}]\right)^N \quad \text{(identically dist.)}.
\end{aligned}$$

Thus we can get a tail bound on the sum by looking at moment generating function (mgf) of one of the terms. Recall that the mgf is the Laplace transform of the density:

$$\mathrm{mgf}_Z(\lambda) = \mathrm{E}[e^{\lambda Z}] = \int e^{\lambda z} f_Z(z)\, dz.$$

To get (2), Hoeffding proved the following lemma:

> Let Z be a random variable that falls in the interval $[a, b]$ with probability 1. Then
>
> $$\mathrm{E}\left[e^{\lambda(Z - \mathrm{E}[Z])}\right] \;\leq\; e^{-\lambda^2 (b-a)^2/8},$$
>
> for all $\lambda > 0$.

Proof of this is not so straightforward, but in the end it just relies on the convexity of the function $e^{\lambda t}$ combined with the Taylor theorem. The proof is done nicely on Wikipedia[5].

[5] https://en.wikipedia.org/wiki/Hoeffding's_inequality

Now if $Z_1, Z_2, \ldots, Z_N$ are iid and fall in $[a, b]$, we have

$$\mathrm{P}\left(\sum_{n=1}^{N} Z_n - \mathrm{E}[Z_n] > t\right) \;\leq\; e^{-\lambda t} e^{N\lambda^2 (b-a)^2/8}, \quad \text{for all } \lambda > 0.$$

The value of λ that minimizes the right hand side above is

$$\lambda = \frac{4t}{N(b-a)^2},$$

and so plugging this in a simplifying gives us

$$\mathrm{P}\left(\sum_{n=1}^{N} Z_n - \mathrm{E}[Z_n] > t\right) \;\leq\; e^{-2t^2/N(b-a)^2}.$$