Introduction to Statistical Learning Theory

MIT 15.097 Course Notes
Cynthia Rudin

Credit: A large part of this lecture was taken from an introduction to learning theory of Bousquet, Boucheron, Lugosi

Now we are going to study, in a probabilistic framework, the properties of learning algorithms. At the beginning of the semester, I told you that it was important for our models to be "simple" in order to be able to *generalize*, or *learn* from data. I didn't really say that precisely before, but in this lecture I will.

Generalization = Data + Knowledge

Finite data cannot replace knowledge. Knowledge allows you to choose a simpler set of models.

Perhaps surprisingly, there is no one universal right way to measure simplicity or complexity of a set of models - simplicity is not an absolute notion. But we'll give several precise ways to measure this. And we'll precisely show how our ability to learn depends on the simplicity of the models. So we'll make concrete (via proof) this philosophical argument that learning somehow needs simplicity.

In classical statistics, the number of parameters in the model is the usual measure of complexity. Here we'll use other complexity measures, namely the Growth Function and VC dimension (which is a beautiful combinatorial quantity), covering number (the one I usually use), and Rademacher average.

Assumptions

Training and test data are drawn iid from the same distribution. If there's no relationship between training and test, there's no way to learn of course. (That's like trying to predict rain in Africa next week using data about horse-kicks in the Prussian war) so we have to make some assumption.

Each learning algorithm encodes specific knowledge (or a specific assumption, perhaps about what the optimal classifier must look like) and works best when this assumption is satisfied by the problem to which it is applied.

Notation

Input space $\mathcal{X}$, output space $\mathcal{Y} = \{-1, 1\}$, unknown distribution D on $\mathcal{X} \times \mathcal{Y}$. We observe m iid pairs $\{(x_i, y_i)\}_{i=1}^m$ drawn iid from D. The goal is to construct a function $f : \mathcal{X} \to \mathcal{Y}$ that predicts y from x.

We would like the true risk to be as small as possible, where the *true risk* is:

$$R^{\text{true}}(f) := \mathbb{P}_{(X,Y)\sim D}(f(X) \neq Y) = \mathbb{E}_{(X,Y)\sim D}[\mathbf{1}_{f(X)\neq Y}].$$

Did you recognize this nice thing that comes from the definition of expectation and probability? We can flip freely between notation for probability and expectation.

$$\mathbb{P}_{Z\sim D}(Z = \text{blah}) = \sum_{\text{outcomes}} \mathbf{1}_{[\text{outcome=blah}]}\mathbb{P}_{Z\sim D}(Z = \text{outcome}) = \mathbb{E}_{Z\sim D}\mathbf{1}_{[Z=\text{blah}]}.$$

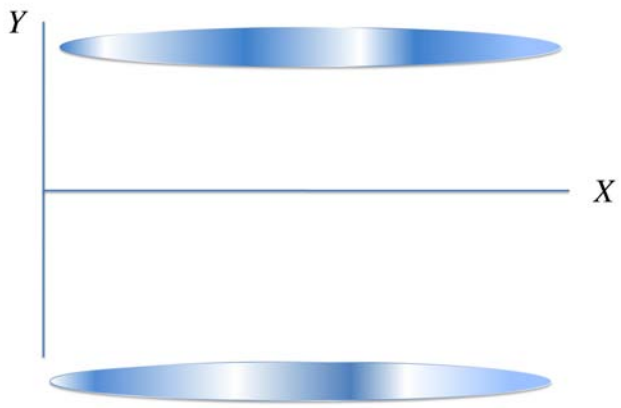
We introduce the *regression function*

$$\eta(x) = \mathbb{E}_{(X,Y)\sim D}(Y|X = x)$$

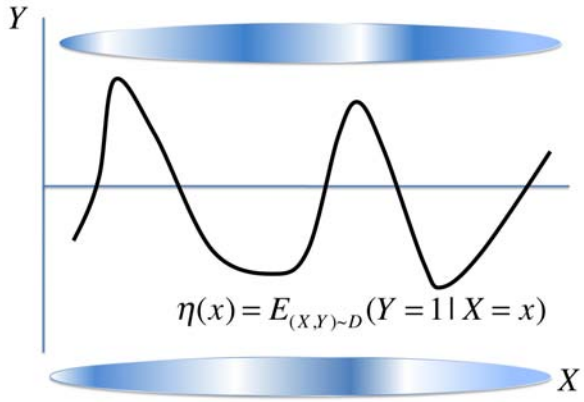
and the *target function* (or Bayes classifier)

$$t(x) = \text{sign } \eta(x).$$

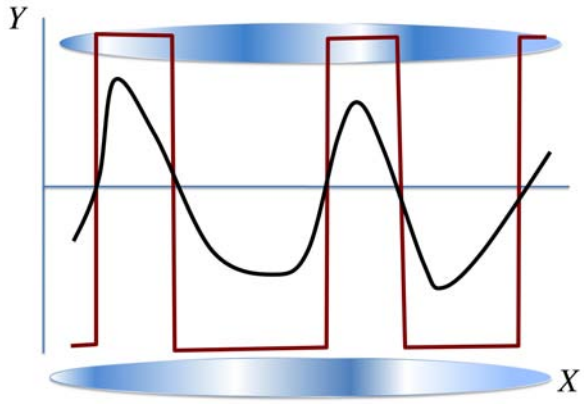
Think of the distribution D, which looks sort of like this:



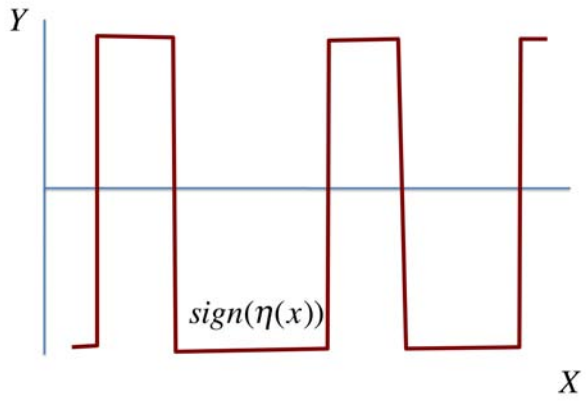
Here's the function η:



Now take the sign of it:



And that's t:



The target function achieves the minimum risk over all possible measurable functions:

$$R^{\text{true}}(t) = \inf_f R^{\text{true}}(f).$$

We denote the value $R^{\text{true}}(t)$ by R^*, called the *Bayes Risk*.

Our goal is to identify this function t but since D is unknown, we cannot evaluate t at any x.

The *empirical risk* that we can measure is:

$$R^{\text{emp}}(f) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{[f(x_i) \neq y_i]}.$$

Algorithm

Most of the calculations don't depend on a specific algorithm, but you can think of using regularized empirical risk minimization.

$$f_m \in \text{argmin}_{f \in \mathcal{F}} R^{\text{emp}}(f) + C\|f\|^2$$

for some norm. The regularization term will control the complexity of the model to prevent overfitting. The class of functions that we're working with is $\mathcal{F}$.

Bounds

Remember, we can compute f_m and $R^{\text{emp}}(f_m)$, but we cannot compute things like $R^{\text{true}}(f_m)$.

The algorithm chooses f_m from the class of functions $\mathcal{F}$. Let us call the best function in the class f^*, so that

$$R^{\text{true}}(f^*) = \inf_{f \in \mathcal{F}} R^{\text{true}}(f).$$

Then, I would like to know how far $R^{\text{true}}(f_m)$ is from R^*. How bad is the function we chose, compared to the best one, the Bayes Risk?

$$\begin{aligned} R^{\text{true}}(f_m) - R^* &= [R^{\text{true}}(f^*) - R^*] + [R^{\text{true}}(f_m) - R^{\text{true}}(f^*)] \\ &= \text{Approximation Error} + \text{Estimation Error} . \end{aligned}$$

The Approximation Error measures how well functions in $\mathcal{F}$ can approach the target (it would be zero if $t \in \mathcal{F}$). The Estimation Error is a random quantity (it depends on data) and measures how close is f_m to the best possible choice in $\mathcal{F}$.

Draw Approximation Error and Estimation Error

Figuring out the Approximation Error is usually difficult because it requires knowledge about the target, that is, you need to know something about the distribution D. In Statistical Learning Theory, generally there is no assumption made about the target (such as its belonging to some class). This is probably the main reason why this theory is so important - it does not require any knowledge of the distribution D.

Also, even if the empirical risk converges to the Bayes risk as m gets large (the algorithm is *consistent*), it turns out that the convergence can be arbitrarily slow if there is no assumption made about the regularity of the target. On the other hand, the rate of convergence of the Estimation Error can be computed without any such assumption. We'll focus on the Estimation Error for this class.

We would really like to understand how bad the true risk of our algorithm's output, $R^{\text{true}}(f_m)$, could possibly be. We want this to be as small as possible of

course. We'll consider another way to look at $R^{\text{true}}(f_m)$:

$$R^{\text{true}}(f_m) = R^{\text{emp}}(f_m) + [R^{\text{true}}(f_m) - R^{\text{emp}}(f_m)], \tag{1}$$

where remember we can measure $R^{\text{emp}}(f_m)$.

We could upper bound the term $R^{\text{true}}(f_m) - R^{\text{emp}}(f_m)$, to make something like this:

$$R^{\text{true}}(f_m) \leq R^{\text{emp}}(f_m) + \text{Stuff}(m, \mathcal{F}).$$

The "Stuff" will get more interesting as this lecture continues.

A Bound for One Function f

Let's define the loss g corresponding to a function f. The loss at point (x, y) is:

$$g(x) = \mathbf{1}_{f(x) \neq y}.$$

Given $\mathcal{F}$, define the *loss class*, which contains all of the loss functions coming from $\mathcal{F}$.

$$\mathcal{G} = \{g : (x, y) \to \mathbf{1}_{f(x) \neq y} : f \in \mathcal{F}\}.$$

So g doesn't look at predictions f, instead it looks at whether the predictions were correct. Notice that $\mathcal{F}$ contains functions with range in $\{-1, 1\}$ while $\mathcal{G}$ contains functions with range $\{0, 1\}$.

There's a bijection between $\mathcal{F}$ and $\mathcal{G}$. You can go from an f to its g by $g(x, y) = \mathbf{1}_{f(x) \neq y}$. You can go from a g to its f by saying that if $g(x, y) = 1$ then set $f(x) = -y$, otherwise set $f(x) = y$. We'll use the g notation whenever we're bounding the difference between an empirical average and its mean because the notation is slightly simpler.

Define this notation:

$$P^{\text{true}} g = \mathbb{E}_{(X,Y) \sim D}[g(X, Y)] \quad \text{(true risk again)}$$

$$P^{\text{emp}} g = \frac{1}{m} \sum_{i=1}^{m} g(X_i, Y_i) \quad \text{(empirical risk again)}$$

so that we have another way to write the true risk and empirical risk directly in terms of the loss. P^{emp} is called the *empirical measure* associated to the

training sample. It just computes the average of a function at the training points. Remember, we are interested in the difference between the true risk and empirical risk, same thing as in the right side of (1), which we're going to upper bound:

$$P^{\text{true}} g_m - P^{\text{emp}} g_m. \tag{2}$$

(g_m is the loss version of f_m.)

Hoeffding's Inequality

For convenience we'll define $Z_i = (X_i, Y_i)$ and $Z = (X, Y)$, and probabilities will be taken with respect to $Z_1 \sim D, ..., Z_m \sim D$ which we'll write $\mathbf{Z} \sim D^m$.

Let's rewrite the quantity we're interested in, for a general g this time:

$$P^{\text{true}} g - P^{\text{emp}} g = \mathbb{E}_{\mathbf{Z} \sim D^m}[g(Z)] - \frac{1}{m} \sum_{i=1}^{m} g(Z_i).$$

It's a difference between an empirical mean and its expectation. By the law of large numbers we know asymptotically that the mean converges to the expectation in probability. So with probability 1, with respect to $\mathbf{Z} \sim D^m$,

$$\lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} g(Z_i) = \mathbb{E}_{\mathbf{Z} \sim D^m}[g(Z)].$$

So with enough data, the empirical risk is a good approximation to its true risk.

There's a quantitative version of the law of large numbers when variables are bounded:

Theorem 1 (Hoeffding). *Let $Z_1 ... Z_m$ be m iid random variables, and h is a bounded function, $h(Z) \in [a, b]$. Then for all $\epsilon > 0$ we have:*

$$\mathbb{P}_{\mathbf{Z} \sim D^m} \left[\left| \frac{1}{m} \sum_{i=1}^{m} h(Z_i) - \mathbb{E}_{\mathbf{Z} \sim D^m}[h(Z)] \right| \geq \epsilon \right] \leq 2 \exp\left(-\frac{2m\epsilon^2}{(b-a)^2} \right).$$

The probability that the empirical average and expectation are far from each other is small. Let us rewrite the formula to better understand its consequences.

Let the right hand side be δ, so

$$\delta = 2\exp\left(-\frac{2m\epsilon^2}{(b-a)^2}\right).$$

Then if I solve for ϵ, I get:

$$\epsilon = (b-a)\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

So Hoeffding's inequality, applied to the function g becomes:

$$\mathbb{P}_{\mathbf{Z}\sim D^m}\left[|P^{\text{emp}}g - P^{\text{true}}g| \geq (b-a)\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right] \leq \delta.$$

There's a technique called "inversion" that we'll use a lot.

Inversion

Using inversion, we get that with probability at least $1-\delta$:

$$|P^{\text{emp}}g - P^{\text{true}}g| \leq (b-a)\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

The expression above is "2-sided" in that there's an absolute value on the left. This considers whether $P^{\text{emp}}g$ is larger than $P^{\text{true}}g$ or smaller than it. There's also 1-sided versions of Hoeffding's inequality where we look at deviations in one direction or the other, for instance here is a 1-sided version of Hoeffding's:

$$\mathbb{P}_{\mathbf{Z}\sim D^m}\left[\mathbb{E}_{\mathbf{Z}\sim D^m}[h(Z)] - \frac{1}{m}\sum_{i=1}^{m} h(Z_i) \geq \epsilon\right] \leq \exp\left(-\frac{2m\epsilon^2}{(b-a)^2}\right).$$

If we again set the right side to δ and solve for ϵ, and invert, we get that with probability at least $1-\delta$,

$$\mathbb{E}_{\mathbf{Z}\sim D^m}[h(Z)] - \frac{1}{m}\sum_{i=1}^{m} h(Z_i) \leq (b-a)\sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

Here is the inverted one applied to g: with probability at least $1-\delta$,

$$P^{\text{true}}g - P^{\text{emp}}g \leq (b-a)\sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

Moving the empirical term to the right, we have that with probability at least $1-\delta$,

$$P^{\text{true}}g \leq P^{\text{emp}}g + (b-a)\sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

Remember that g is the loss, so $g(Z) = \mathbf{1}_{f(X)\neq Y}$ and that way we have an upper bound for the true risk, which we want to be small.

This expression seems very nice, but guess what? It doesn't apply when f (i.e., g) comes from any reasonable learning algorithm!

Why not?

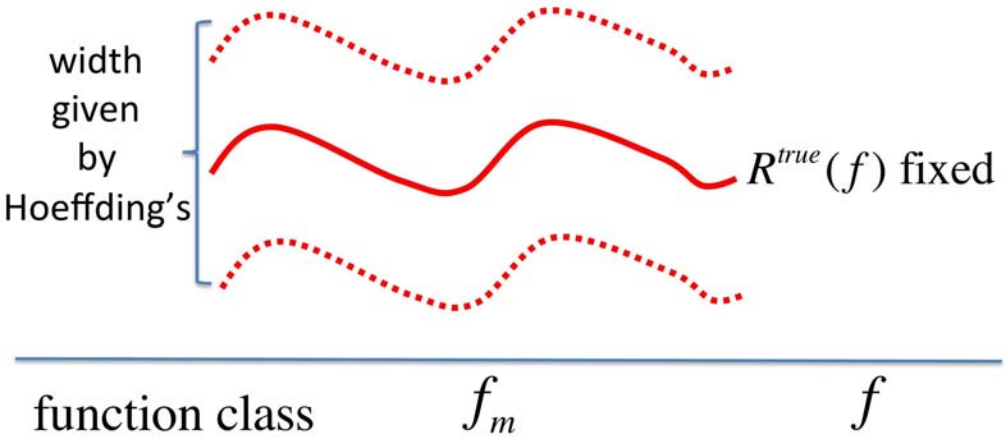
Limitations

The result above says that for each fixed function $g \in \mathcal{G}$, there is a set S of "good" samples $z_1, ..., z_m$, for which

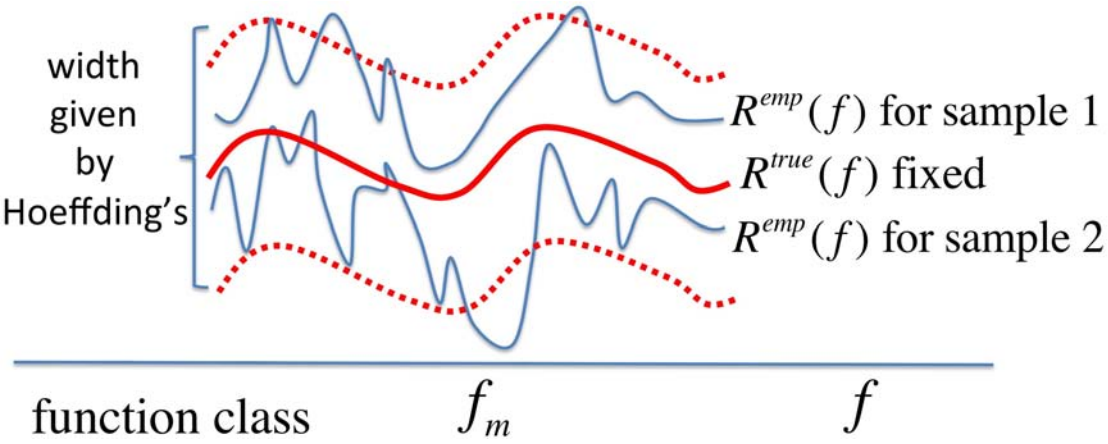
$$P^{\text{true}}g - P^{\text{emp}}g \leq (1-0)\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

and this set of samples has measure $\mathbb{P}_{\mathbf{Z}\sim D^m}[\mathbf{Z} \in S] \geq 1-\delta$. However, these sets may be different for different functions g. In other words, for the sample S we actually observe, there's no telling how many of the functions in $\mathcal{G}$ will actually satisfy this inequality!

This figure might help you understand. Each point on the x-axis is a different function. The curve marked $R^{\text{true}}(f)$ is the true risk, which is a constant for each f since it involves the whole distribution (and not a sample).



If you give me a sample and a function f, I can calculate $R^{\text{emp}}(f)$ for that sample, which gives me a dot on the plot. So, for each sample we get a different curve on the figure. For each f, Hoeffding's inequality makes sure that most of the time, the $R^{\text{emp}}(f)$ curves lie within a small distance of $R^{\text{true}}(f)$, though we don't know which ones. In other words, for an observed sample, only some of the functions in $\mathcal{F}$ will satisfy the inequality, not all of them.



But remember, our algorithms choose f_m *knowing* the data. They generally try to minimize the regularized $R^{\text{emp}}(f)$. Consider drawing a sample S, which corresponds to a curve on the figure. Our algorithm could (on purpose) meander along that curve until it chooses a f_m that gives a small value of R^{emp}. This value could be very far from $R^{\text{true}}(f_m)$. This could definitely happen, and if there are more f's to choose from (if the function class is larger), then this happens more easily - uh oh! In other words, if $\mathcal{F}$ is large enough, one can find, somewhere along the axis, a function f for which the difference between the two curves $R^{\text{emp}}(f)$ and $R^{\text{true}}(f)$ will be very large.

We don't want this to happen!

Uniform Bounds

We really need to make sure our algorithm doesn't do this - otherwise it will never generalize. That's why we're going to look at *uniform deviations* in order to upper bound (1) or (2):

$$R^{\text{true}}(f_m) - R^{\text{emp}}(f_m) \leq \sup_{f \in \mathcal{F}} (R^{\text{true}}(f) - R^{\text{emp}}(f))$$

where we look at the worst deviation over all functions in the class.

Let us construct a first uniform bound, using Hoeffding's inequality and the union bound. Define:

$$C_j = \{z_1, ..., z_m : P^{\text{true}} g_j - P^{\text{emp}} g_j \geq \epsilon\}.$$

This set contains all the "bad" samples, those for which the bound fails for function g_j. From Hoeffding's Inequality, for each j,

$$\mathbb{P}_{\mathbf{Z} \sim D^m}[\mathbf{Z} \in C_j] \leq \delta.$$

Consider two functions g_1 and g_2. Say we want to measure how many samples are "bad" for either one of these functions or the other. We're going to use the *union bound* to do this, which says:

$$\mathbb{P}[C_1 \cup C_2] \leq \mathbb{P}[C_1] + \mathbb{P}[C_2] \leq 2\delta,$$

the probability that we hit a bad sample for either g_1 or g_2 is

$$\leq$$

prob to hit a bad sample for C_1 + prob to hit a bad sample for C_2.

More generally, the union bound is:

$$\mathbb{P}[C_1 \cup ... \cup C_N] \leq \sum_{j=1}^{N} \mathbb{P}[C_j] \leq N\delta$$

So this is a bound on the probability that our chosen sample will be bad for any of the functions $g_1, ..., g_N$. So we get:

$$\begin{aligned}
&\mathbb{P}_{\mathbf{Z}\sim D^m}[\exists g \in \{g_1, ..., g_N\} : P^{\text{true}}g - P^{\text{emp}}g \geq \epsilon] \\
&\leq \sum_{j=1}^{N} \mathbb{P}_{\mathbf{Z}\sim D^m}[P^{\text{true}}g_j - P^{\text{emp}}g_j \geq \epsilon] \\
&\leq \sum_{j=1}^{N} \exp(-2m\epsilon^2) \quad \boxed{\text{Where did this come from?}} \\
&= N\exp(-2m\epsilon^2).
\end{aligned}$$

If we define a new δ so we can invert:

$$\delta := N\exp(-2m\epsilon^2)$$

and solve for ϵ, we get:

$$\epsilon = \sqrt{\frac{\log N + \log\frac{1}{\delta}}{2m}}.$$

Plugging that in and inverting, we find that with probability at least $1-\delta$,

$$\forall g \in \{g_1, ..., g_N\} : P^{\text{true}}g - P^{\text{emp}}g \leq \sqrt{\frac{\log N + \log\frac{1}{\delta}}{2m}}.$$

Changing g's back to f's, we've proved the following:

Theorem. (Hoeffding + Union Bound)
For $\mathcal{F} = \{f_1...f_N\}$, for all $\delta > 0$ with probability at least $1-\delta$,

$$\forall f \in \mathcal{F}, \;\; R^{\text{true}}(f) \leq R^{\text{emp}}(f) + \sqrt{\frac{\log N + \log\frac{1}{\delta}}{2m}}.$$

Just to recap the reason why this bound is better than the last one, if we know our algorithm only picks functions from a finite function class $\mathcal{F}$, we now have a bound that can be applied to f_m, even though it depends on the data.

Note the main difference with plain Hoeffding's inequality is the extra $\log N$ term on the right hand side. This term is the one saying we want N bounds to hold simultaneously.

Estimation Error

Let's say we're doing empirical risk minimization, that is, f_m is the minimizer of the empirical risk R^{emp}.

We can use the theorem above (combined with (1)) to get an upper bound on the Estimation Error. Start with this:

$$R^{\text{true}}(f_m) \;=\; R^{\text{true}}(f_m) - R^{\text{true}}(f^*) + R^{\text{true}}(f^*)$$

Then we'll use the fact that $R^{\text{emp}}(f^*) - R^{\text{emp}}(f_m) \geq 0$. Why is that?

We'll add that to the expression above:

$$\begin{aligned} R^{\text{true}}(f_m) &\leq [R^{\text{emp}}(f^*) - R^{\text{emp}}(f_m)] + R^{\text{true}}(f_m) - R^{\text{true}}(f^*) + R^{\text{true}}(f^*) \\ &= R^{\text{emp}}(f^*) - R^{\text{true}}(f^*) - R^{\text{emp}}(f_m) + R^{\text{true}}(f_m) + R^{\text{true}}(f^*) \\ &\leq |R^{\text{true}}(f^*) - R^{\text{emp}}(f^*)| + |R^{\text{true}}(f_m) - R^{\text{emp}}(f_m)| + R^{\text{true}}(f^*) \\ &\leq 2 \sup_{f \in \mathcal{F}} |R^{\text{true}}(f) - R^{\text{emp}}(f)| + R^{\text{true}}(f^*). \end{aligned}$$

We could use a 2-sided version of the theorem (with an extra factor of 2 somewhere) that with probability $1 - \delta$, that first term is bounded by the square root term in the theorem. Specifically, we know that with probability $1 - \delta$:

$$R^{\text{true}}(f_m) \leq 2\sqrt{\frac{\log N + \log \frac{2}{\delta}}{2m}} + R^{\text{true}}(f^*).$$

Actually, if you think about it, both terms in the right hand side depend on the size of the class $\mathcal{F}$. If this size increases, the first term will increase, and the second term will decrease. Why?

Summary and Perspective

- Generalization requires knowledge (like restricting f to lie in a restricted class $\mathcal{F}$).
- The error bounds are valid with respect to the repeated sampling of training sets.

- For a fixed function f, for most of the samples,
$$R^{\text{true}}(f) - R^{\text{emp}}(f) \approx 1/\sqrt{m}.$$
- For most of the samples if the function class if finite, $|\mathcal{F}| = N$,
$$\sup_{g \in \mathcal{G}} [R^{\text{true}}(g) - R^{\text{emp}}(g)] \approx \sqrt{\log N/m}.$$
The extra term is because we choose f_m in a way that changes with the data.
- We have the Hoeffding + Union Bound Theorem above, which bounds the worst difference between empirical risk and true risk among functions in the class.

There are several things that could be improved. For instance Hoeffding's inequality only uses the boundedness of the functions, not their variance, which is something we won't deal with here. The supremum over $\mathcal{F}$ of $R^{\text{true}}(f) - R^{\text{emp}}(f)$ is not necessarily what the algorithm would choose, so the upper bound could be loose. The union bound is in general loose, because it is as bad as if all the $f_j(Z)$'s are independent.

Infinite Case: VC Dimension

Here we'll show how to extend the previous results to the case where the class $\mathcal{F}$ is infinite.

We'll start with a simple refinement of the union bound that allows to extend the previous results to the (countably) infinite case.
Recall that by Hoeffding's inequality for a single function g, for each $\delta > 0$, where possibly we could choose δ depending on g, which we write $\delta(g)$, we have:
$$\mathbb{P}_{\mathbf{Z} \sim D^m} \left[P^{\text{true}} g - P^{\text{emp}} g \geq \sqrt{\frac{\log \frac{1}{\delta(g)}}{2m}} \right] \leq \delta(g).$$
Hence if we have a countable set $\mathcal{G}$, the union bound gives:
$$\mathbb{P}_{\mathbf{Z} \sim D^m} \left[\exists g \in \mathcal{G} : P^{\text{true}} g - P^{\text{emp}} g \geq \sqrt{\frac{\log \frac{1}{\delta(g)}}{2m}} \right] \leq \sum_{g \in \mathcal{G}} \delta(g).$$

If we choose the $\delta(g)$'s so that they add up to a constant total value δ, that is, $\delta(g) = \delta\, p(g)$ where $\sum_{g\in\mathcal{G}} p(g) = 1$, then the right hand side is just δ and we get the following with inversion: with probability at least $1-\delta$,

$$\forall g \in \mathcal{G}, P^{\text{true}} g \leq P^{\text{emp}} g + \sqrt{\frac{\log \frac{1}{p(g)} + \log \frac{1}{\delta}}{2m}}.$$

If $\mathcal{G}$ is finite with size N, and we take a uniform $p(g) = \frac{1}{N}$, we get the $\log N$ term as before.

General Case

When the set $\mathcal{G}$ is uncountable, the previous approach doesn't work because $p(g)$ is a density, so it's 0 for a given g and the bound will be vacuous. We'll switch back to the original class $\mathcal{F}$ rather than the loss class for now. The general idea is to look at the function class's behavior on the sample. Given $z_1, ..., z_m$, we consider

$$\mathcal{F}_{z_1,...,z_m} = \{f(z_1), ..., f(z_m) : f \in \mathcal{F}\}.$$

$\mathcal{F}_{z_1,...,z_m}$ is the set of ways the data $z_1, ..., z_m$ are classified by functions from $\mathcal{F}$. Since the functions f can only take two values, this set will always be finite, no matter how big $\mathcal{F}$ is.

Definition (Growth Function) The growth function is the maximum number of ways into which m points can be classified by the function class:

$$S_{\mathcal{F}}(m) = \sup_{(z_1,...,z_m)} |\mathcal{F}_{z_1,...,z_m}|.$$

Intuition for Growth Function and Example of Halfplanes

We defined the growth function in terms of the initial class $\mathcal{F}$ but we can do the same with the loss class $\mathcal{G}$ since there's a 1-1 mapping, so we'll get $S_{\mathcal{G}}(m) = S_{\mathcal{F}}(m)$.

This growth function can be used as a measure of the 'size' of a class of functions as demonstrated by the following result:

Theorem-GrowthFunction (Vapnik-Chervonenkis) *For any $\delta > 0$, with probability at least* $1 - \delta$ with respect to a random draw of the data,

$$\forall f \in \mathcal{F} \;\; R^{\text{true}}(f) \leq R^{\text{emp}}(f) + 2\sqrt{2\frac{\log S_{\mathcal{F}}(2m) + \log \frac{4}{\delta}}{m}}$$

(proof soon).

This bound shows nicely that simplicity implies generalization. The simpler the function class, the better the guarantee that R^{true} will be small. In the finite case where $|\mathcal{F}| = N$ (we have N possible classifiers), we have $S_{\mathcal{F}}(m) \leq N$ (at worst we use up all the classifiers when we're computing the growth function). So this bound is always better than the one we had before (except for the constants).

But we need to figure out how to compute $S_{\mathcal{F}}(m)$. We'll do that using VC dimension.

VC dimension

Since $f \in \{-1, 1\}$, it is clear that $S_{\mathcal{F}}(m) \leq 2^m$.

If $S_{\mathcal{F}}(m) = 2^m$ there is a data set of size m points such that $\mathcal{F}$ can generate any classification on these points (we say $\mathcal{F}$ *shatters* the set).

The VC dimension of a class $\mathcal{F}$ is the size of the largest set that it can shatter.

Definition. (VC dimension) *The VC dimension of a class $\mathcal{F}$ is the largest m such that*

$$S_{\mathcal{F}}(m) = 2^m.$$

What is the VC dimension of halfplanes in 2 dimensions?

Can you guess the VC dimension of halfplanes in d dimensions?

In the example, the number of parameters needed to define the half space in $\mathbf{R}^d$ is the number of dimensions, d. So a natural question to ask is whether the VC dimension is related to the number of parameters of the function class. In other

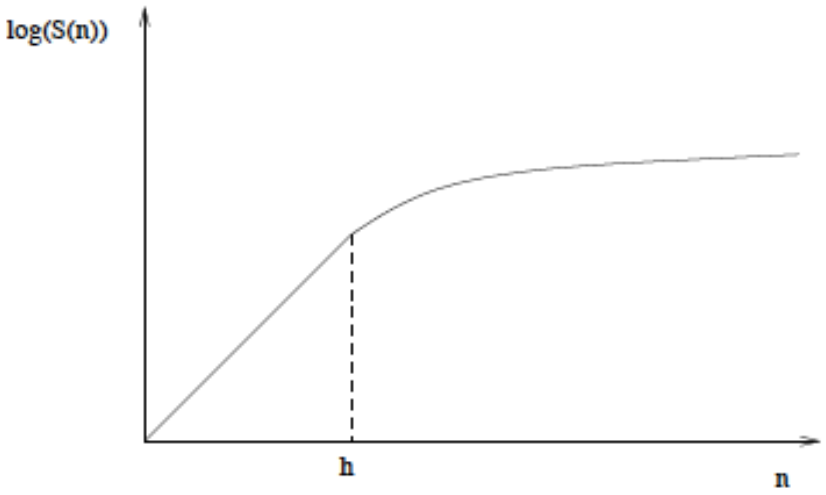
words, VC dimension is supposed to measure complexity of a function class - does it just basically measure the number of parameters?

Is the VC dimension always close to the number of parameters?

So how can VC dimension help us compute the growth function? Well, if a class of functions has VC dim h, then we know that we can shatter m examples when $m \leq h$, and in that case, $S_{\mathcal{F}}(m) = 2^m$. If $m > h$, then we know we can't shatter the points, so $S_{\mathcal{F}}(m) < 2^m$ otherwise.

This doesn't seem very helpful perhaps, but actually an intriguing phenomenon occurs for $m \geq h$, shown below.

The plot below shows for $m \geq h$ (where we can't shatter) the number of ways we can classify - that's the growth function. The growth function which is exponential up until the VC dimension, becomes polynomial afterwards!



Typical behavior of the log growth function.

This is captured by the following lemma.

Lemma. (Vapnik and Chervonenkis, Sauer, Shelah) *Let $\mathcal{F}$ be a class of*

functions with finite VC dimension h. Then for all $m \in \mathbb{N}$,

$$S_{\mathcal{F}}(m) \leq \sum_{i=0}^{h} \binom{m}{i} \qquad \boxed{\text{Intuition}}$$

and for all $m \geq h$

$$S_{\mathcal{F}}(m) \leq \left(\frac{em}{h}\right)^{h}.$$

Using this lemma for $m \geq h$ along with Theorem-GrowthFunction, we get:

Theorem VC-Bound. If $\mathcal{F}$ has VC dim h, and for $m \geq h$, with prob. at least $1 - \delta$,

$$\forall f \in \mathcal{F} \;\; R^{\text{true}}(f) \leq R^{\text{emp}}(f) + 2\sqrt{2\frac{h \log \frac{2em}{h} + \log \frac{4}{\delta}}{m}}.$$

What is important to remember from this result is that the difference between the true and empirical risk is at most of order

$$\sqrt{\frac{h \log m}{m}}.$$

Before we used VC dim, the bound was infinite, i.e., vacuous!

Recap

Why is Theorem VC-Bound important? It shows that limiting the complexity of the class of functions leads to better generalization. An interpretation of VC dim and growth functions is that they measure the "effective" size of the class, that is, the size of the projection of the class onto finite samples. This measure doesn't just count the number of functions in the class, but depends on the geometry of the class, that is, the projections onto the possible samples. Also since the VC dimension is finite, our bound shows that the empirical risk will converge uniformly over the class $\mathcal{F}$ to the true risk.

Back to Margins

How is it that SVM's limit the complexity? Well, the choice of kernel controls the complexity. But also the margin itself controls complexity. There is a set of

linear classifiers called "gap-tolerant classifiers" that I won't define precisely (it gets complicated) that require a margin of at least Δ between points of the two different classes. The points are also forced to live inside a sphere of diameter D. So the class of functions is fairly limited, since they not only need to separate the points with a margin of Δ, but also we aren't allowed to move the points outside of the sphere.

"Theorem" VC-Margin. (Vapnik) *For data in* $\mathbf{R}^d$*, the VC dimension* h *of (linear) gap-tolerant classifiers classifiers with gap* Δ *belong to a sphere of diameter* D*, is bounded by the inequality:*

$$h \leq \min\left(\left\lceil \frac{D^2}{\Delta^2} \right\rceil, d\right) + 1.$$

So the VC dimension (of the set of functions that separate points with some margin) is less than 1/margin. If we have a large margin, we necessarily have a small VC-dimension.

What does this say about halfspaces in $\mathbf{R}^d$?
(Think about the VC dimension example we did earlier.)

Symmetrization

We'll do the proof of Theorem-GrowthFunction. The key ingredient is the *symmetrization lemma*. We'll use what's called a "ghost sample" which is an extra (virtual) data set $Z'_1, ..., Z'_m$. Denote P'^{emp} the corresponding empirical measure.

(Lemma-Symmetrization) *For any* $t > 0$*, such that* $mt^2 \geq 2$*,*

$$\mathbb{P}_{\mathbf{Z}\sim D^m}\left[\sup_{g\in\mathcal{G}}(P^{\text{true}} - P^{\text{emp}})g \geq t\right] \leq 2\mathbb{P}_{\mathbf{Z}\sim D^m, \mathbf{Z}'\sim D^m}\left[\sup_{g\in\mathcal{G}}(P'^{\text{emp}} - P^{\text{emp}})g \geq t/2\right].$$

That is, if we can bound the difference between the behavior on one sample versus another, it gives us a bound on the behavior of a sample with respect to the true risk.

Proof. Let g_m be the function achieving the supremum in the lhs term, which depends on $Z_1, ...Z_m$. Think about the event that: $(P^{\text{true}} - P^{\text{emp}})g_m \geq t$ (the

sample's loss is far from the true loss) and $(P^{\text{true}} - P'^{\text{emp}})g_m < t/2$ (the ghost sample's loss is close to the true loss). If this event were true, it sort of means that things didn't generalize well for $Z_1, ..., Z_m$ but that they did generalize well for $Z'_1, ..., Z'_m$. If we can show that this event happens rarely, then the ghost sample can help us. Again, the event that we want to happen rarely is $(P^{\text{true}} - P^{\text{emp}})g_m \geq t$ and $(P^{\text{true}} - P'^{\text{emp}})g_m < t/2$.

$$\begin{aligned}
&\mathbf{1}_{(P^{\text{true}} - P^{\text{emp}})g_m \geq t} \mathbf{1}_{(P^{\text{true}} - P'^{\text{emp}})g_m < t/2} \\
&= \mathbf{1}_{(P^{\text{true}} - P^{\text{emp}})g_m \geq t \text{ and } (P^{\text{true}} - P'^{\text{emp}})g_m < t/2} \\
&= \mathbf{1}_{(P^{\text{true}} - P^{\text{emp}})g_m \geq t \text{ and } (P'^{\text{emp}} - P^{\text{true}})g_m > -t/2} \\
&\leq \mathbf{1}_{(P^{\text{true}} - P^{\text{emp}} + P'^{\text{emp}} - P^{\text{true}})g_m > t - t/2 = t/2} = \mathbf{1}_{(-P^{\text{emp}} + P'^{\text{emp}})g_m > t/2}.
\end{aligned}$$

The inequality came from the fact that the event on the second last line ($(P^{\text{true}} - P^{\text{emp}})g_m \geq t$ and $(P'^{\text{emp}} - P^{\text{true}})g_m > -t/2$) implies the event on the last line ($(P^{\text{true}} - P^{\text{emp}} + P'^{\text{emp}} - P^{\text{true}})g_m > t - t/2$), so the event on the last line could happen more often.

Taking expectations with respect to the second sample, and using the trick to change expectation into probability,

$$\begin{aligned}
&\mathbf{1}_{(P^{\text{true}} - P^{\text{emp}})g_m \geq t} \mathbb{P}_{\mathbf{Z}' \sim D^m}[(P^{\text{true}} - P'^{\text{emp}})g_m < t/2] \\
&\leq P_{\mathbf{Z}' \sim D^m}[(P'^{\text{emp}} - P^{\text{emp}})g_m > t/2]. \qquad (3)
\end{aligned}$$

Do you remember Chebyshev's Inequality? It says $\mathbb{P}[|X - \mathbb{E}X| \geq t] \leq \text{Var}X/t^2$. We'll apply it now, to that second term on the left, inverted:

$$\mathbb{P}_{\mathbf{Z}' \sim D^m}[(P^{\text{true}} - P'^{\text{emp}})g_m \geq t/2] \leq \frac{4\text{Var}g_m}{mt^2}.$$

I hope you'll believe me when I say that any random variable that has range $[0, 1]$ has variance less than or equal to $1/4$. Hence,

$$\mathbb{P}_{\mathbf{Z}' \sim D^m}[(P^{\text{true}} - P'^{\text{emp}})g_m \geq t/2] \leq \frac{1}{mt^2}.$$

Inverting back, so that it looks like the second term on the left of (3) again:

$$\mathbb{P}_{\mathbf{Z}' \sim D^m}[(P^{\text{true}} - P'^{\text{emp}})g_m < t/2] \geq 1 - \frac{1}{mt^2}.$$

Multiplying both sides by $\mathbf{1}_{(P^{\text{true}} - P^{\text{emp}})g_m \geq t}$ I get back to the left of (3):

$$\begin{aligned}
\mathbf{1}_{(P^{\text{true}} - P^{\text{emp}})g_m \geq t} \left(1 - \frac{1}{mt^2}\right) &\leq \mathbf{1}_{(P^{\text{true}} - P^{\text{emp}})g_m \geq t} \mathbb{P}_{\mathbf{Z}' \sim D^m}[(P^{\text{true}} - P'^{\text{emp}})g_m < t/2] \\
&\leq \mathbb{P}_{\mathbf{Z}' \sim D^m}[(P'^{\text{emp}} - P^{\text{emp}})g_m > t/2] \quad \text{from (3)}.
\end{aligned}$$

Taking the expectation with respect to the first sample, the term

$$\mathbb{E}_{\mathbf{Z}\sim D^m}\mathbf{1}_{(P^{\text{true}}-P^{\text{emp}})g_m\geq t} \qquad \text{becomes} \qquad \mathbb{P}_{\mathbf{Z}\sim D^m}[(P^{\text{true}}-P^{\text{emp}})g_m\geq t].$$

And now we get:

$$\mathbb{P}_{\mathbf{Z}\sim D^m}[(P^{\text{true}}-P^{\text{emp}})g_m\geq t]\left(1-\frac{1}{mt^2}\right)\leq \mathbb{P}_{\mathbf{Z}'\sim D^m\mathbf{Z}\sim D^m}[(P'^{\text{emp}}-P^{\text{emp}})g_m> t/2]$$

$$\mathbb{P}_{\mathbf{Z}\sim D^m}[(P^{\text{true}}-P^{\text{emp}})g_m\geq t]\leq \left(\frac{1}{1-\frac{1}{mt^2}}\right)\mathbb{P}_{\mathbf{Z}'\sim D^m\mathbf{Z}\sim D^m}[(P'^{\text{emp}}-P^{\text{emp}})g_m> t/2].$$

Only one more step, which uses our assumption $mt^2\geq 2$.

$$\begin{aligned} mt^2 &\geq 2 \\ \frac{1}{mt^2} &\leq \frac{1}{2} \\ 1-\frac{1}{mt^2} &\geq 1-\frac{1}{2}=\frac{1}{2} \\ \left(\frac{1}{1-\frac{1}{mt^2}}\right) &\leq 2 \end{aligned}$$

Plug:

$$\begin{aligned} \mathbb{P}_{\mathbf{Z}\sim D^m}[(P^{\text{true}}-P^{\text{emp}})g_m\geq t] &\leq 2\mathbb{P}_{\mathbf{Z}'\sim D^m\mathbf{Z}\sim D^m}[(P'^{\text{emp}}-P^{\text{emp}})g_m> t/2] \\ &\leq 2\mathbb{P}_{\mathbf{Z}\sim D^m,\mathbf{Z}'\sim D^m}\left[\sup_{g\in\mathcal{G}}(P'^{\text{emp}}-P^{\text{emp}})g> t/2\right]. \end{aligned}$$

We have an upper bound by changing the strict inequality ">" to a "≥." Then the result is the same as the statement of the lemma. ■

Remember, we're still in the middle of proving Theorem-GrowthFunction. The symmetrization is just a step in that proof. This symmetrization lemma allows us to replace the expectation $P^{\text{true}}g$ by an empirical average over the ghost sample. As a result, the proof will only depends on the *projection* of the class $\mathcal{G}$ on the double sample

$$\mathcal{G}_{Z_1\ldots Z_m,Z'_1\ldots Z'_m},$$

which contains finitely many different vectors. In other words, an element of this set is just the vector $[g(x_1),...,g(x_m)]$, and there are finitely many possibilities for vectors like this. So we can use the union bound that we used for the finite

case. The other ingredient that we need to prove Theorem-GrowthFunction is this one:

$$\mathbb{P}_{\mathbf{Z}\sim D^m \mathbf{Z}'\sim D^m}[P^{\text{emp}}g - P'^{\text{emp}}g \geq t] \leq 2e^{-mt^2/2}. \tag{4}$$

This one comes itself from a mix of Hoeffding's with the union bound:

$$\begin{aligned}
&\mathbb{P}_{\mathbf{Z}\sim D^m \mathbf{Z}'\sim D^m}[P^{\text{emp}}g - P'^{\text{emp}}g \geq t] \\
&= \mathbb{P}_{\mathbf{Z}\sim D^m \mathbf{Z}'\sim D^m}[P^{\text{emp}}g - P^{\text{true}}g + P^{\text{true}}g - P'^{\text{emp}}g \geq t] \\
&\leq \mathbb{P}_{\mathbf{Z}\sim D^m}[P^{\text{emp}}g - P^{\text{true}}g \geq t/2] + \mathbb{P}_{\mathbf{Z}'\sim D^m}[P^{\text{true}}g - P'^{\text{emp}}g \geq t/2] \\
&\leq e^{-2m(t/2)^2} + e^{-2m(t/2)^2} \\
&= 2e^{-mt^2/2}.
\end{aligned}$$

We just have to put the pieces together now:

$$\begin{aligned}
&\mathbb{P}_{\mathbf{Z}\sim D^m}[\sup_{g\in\mathcal{G}}(P^{\text{true}} - P^{\text{emp}})g \geq t] \\
&\leq 2\mathbb{P}_{\mathbf{Z}\sim D^m \mathbf{Z}'\sim D^m}[\sup_{g\in\mathcal{G}}(P'^{\text{emp}} - P^{\text{emp}})g \geq t/2] \quad \text{Lemma-Symmetrization} \\
&= 2\mathbb{P}_{\mathbf{Z}\sim D^m \mathbf{Z}'\sim D^m}[\sup_{g\in\mathcal{G}_{Z_1,\ldots,Z_m,Z'_1,\ldots,Z'_m}}(P'^{\text{emp}} - P^{\text{emp}})g \geq t/2] \quad \text{(restrict to data)} \\
&\leq 2\sum_{g\in\mathcal{G}_{Z_1,\ldots,Z_m,Z'_1,\ldots,Z'_m}} \mathbb{P}_{\mathbf{Z}\sim D^m \mathbf{Z}'\sim D^m}[(P'^{\text{emp}} - P^{\text{emp}})g \geq t/2] \quad \boxed{\text{(union bound)}} \\
&\leq 2\sum_{g\in\mathcal{G}_{Z_1,\ldots,Z_m,Z'_1,\ldots,Z'_m}} 2e^{-m(t/2)^2/2} \\
&= 4e^{-mt^2/8}\sum_{g\in\mathcal{G}_{Z_1,\ldots,Z_m,Z'_1,\ldots,Z'_m}} 1 \\
&= 4S_{\mathcal{G}}(2m)\ e^{-mt^2/8}.
\end{aligned}$$

And using inversion,

$$\mathbb{P}_{\mathbf{Z}\sim D^m}[\sup_{g\in\mathcal{G}}(P^{\text{true}} - P^{\text{emp}})g \leq t] \geq 1 - 4S_{\mathcal{G}}(2m)\ e^{-mt^2/8}.$$

Letting $\delta = 4S_{\mathcal{G}}(2m)\ e^{-mt^2/8}$, solving for t yields:

$$t = \sqrt{\frac{8}{m}\log\frac{4S_{\mathcal{G}}(2m)}{\delta}}$$

Plug:

$$\mathbb{P}_{\mathbf{Z}\sim D^m}\left[\sup_{g\in\mathcal{G}}(P^{\text{true}}-P^{\text{emp}})g \le 2\sqrt{2\frac{\log S_{\mathcal{G}}(2m)+\log\frac{4}{\delta}}{m}}\right] \ge 1-\delta.$$

So, with probability at least $1-\delta$,

$$\forall g\in\mathcal{G} \quad (P^{\text{true}}-P^{\text{emp}})g \le 2\sqrt{2\frac{\log S_{\mathcal{G}}(2m)+\log\frac{4}{\delta}}{m}}.$$

That's the result of Theorem-GrowthFunction. ■

Other kinds of capacity measures

One important aspect of VC dimension is that it doesn't depend on D, so it is a *distribution independent* quantity. The growth function is also distribution independent. That's nice in some ways, because it allows us to get bounds that don't depend on the problem at hand: the same bound holds for any distribution. Although this may seem like an advantage, it could also be a drawback since, as a result, the bound may be loose for most distributions.

It turns out there are several different quantities that are distribution dependent, which we can use in generalization bounds.

VC-entropy

One quantity is called the (annealed) VC entropy. Recall the notation $|\mathcal{G}_{z_1,\ldots,z_m}|$ which is the number of ways we can correctly/incorrectly classify $z_1,\ldots,z_m$.

$$\text{VC-entropy}_G(m) := \log \mathbb{E}_{\mathbf{Z}\sim D^m}[|\mathcal{G}_{Z_1,\ldots,Z_m}|].$$

If the VC-entropy is large, it means that a lot of the time, there are a lot of different ways to classify m data points. So the capacity of the set of functions is somehow large. There's a bound for the VC-entropy that's very similar to the one for Theorem-GrowthFunction (which I won't go into here).

How does the VC-entropy relate to the Growth Function?

Covering Number

Another quantity is called the covering number. Covering numbers can be defined in several different ways, but we'll just define one of them. Let's start by endowing the function class $\mathcal{G}$ with the following (random) metric:

$$d_m(g, g') = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{[g(Z_i) \neq g'(Z_i)]},$$

the fraction of times they disagree on the (random) sample.

Also denote $B(g_j, \epsilon)$ as ball of radius ϵ around g_j, using the metric d_m. In other words, $B(g_j, \epsilon)$ contains all the functions in $\mathcal{G}$ that are within distance ϵ of g_j according to our metric. We say that a set $g_1, ..., g_N$ *covers* $\mathcal{G}$ at radius ϵ if:

$$\mathcal{G} \subset \bigcup_{j=1}^{N} B(g_j, \epsilon)$$

that is, if $\mathcal{G}$ is contained within the collection of balls centered at the g_j's. We then define the covering number.

Definition. The *covering number* of $\mathcal{G}$ at radius ϵ with respect to d_m, denoted by $N(\mathcal{G}, \epsilon, m)$ is the minimum size of a cover of radius ϵ.

Illustration

(Remember of course that since there's a bijection between $\mathcal{F}$ and $\mathcal{G}$ that $N(\mathcal{G}, \epsilon, m) = N(\mathcal{F}, \epsilon, m)$).

If the covering number is finite, it means we can approximately represent $\mathcal{G}$ by a finite set of functions that cover $\mathcal{G}$. This allows us to use the (finite) union bound. Basically, the proof technique involves showing that things don't change too much within an ϵ ball, so we can characterize the whole ball of functions by its center, then union bound over the centers. This kind of proof technique is my favorite for creating generalization bounds - just because to me it seems more straightforward (of course this is highly subjective).

A typical result (stated without proof) is:

Theorem-Covering. For any $t > 0$,

$$\mathbb{P}_{\mathbf{Z}\sim D^m}[\exists f \in \mathcal{F} : R^{\text{true}}(f) \geq R^{\text{emp}}(f) + t] \leq 8\mathbb{E}_{\mathbf{Z}\sim D^m}[N(\mathcal{F}, t, m)]e^{-mt^2/128}.$$

We can relate the covering number to the VC dimension.

Lemma (Haussler). Let $\mathcal{F}$ be a class of VC dimension h. Then for all $\epsilon > 0$, all m, and any sample,

$$N(\mathcal{F}, \epsilon, m) \leq Ch(4e)^h \epsilon^h.$$

(where C is a constant). One thing about this result is that the upper bound does not depend on the sample size m. Probably it is a loose upper bound, but it's nice to be able to get this independent of m!

Rademacher Averages

Another way to measure complexity is to see how well functions from the class $\mathcal{F}$ can classify random noise. So if I arbitrarily start flipping the labels, how well can functions from my class fit those arbitrary labels. If the functions from $\mathcal{F}$ can fit those arbitrary labels really well, then $\mathcal{F}$ must have high complexity. That's the intuition behind the Rademacher complexity measure that we're going to introduce.

I'm going to introduce some notation,

$$R_m g := \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(Z_i)$$

where the σ_i's are independent $\{+1, -1\}$-valued random variables with probability $1/2$ of taking either value. You can think of them as random coin flips. Denote $\mathbb{E}_{\boldsymbol{\sigma}}$ the expectation taken with respect to the σ_i's.

Definition. (Rademacher Averages) For a class $\mathcal{G}$ of functions, the Rademacher average is defined as:

$$\mathcal{R}(\mathcal{G}) := \mathbb{E}_{\boldsymbol{\sigma},\mathbf{Z}} \sup_{g\in\mathcal{G}} R_m g.$$

In other words, for each set of flips of the coin, you consider them as labels for your data, and find a function from $\mathcal{G}$ that matches them the best. If you can match a lot of random coin flips really well using functions from $\mathcal{G}$, then $\mathcal{G}$ has high complexity.

Here's a bound involving Rademacher averages:

Theorem-RademacherBound. *With probability at least* $1-\delta$,

$$\forall g \in \mathcal{G}, R^{\text{true}}(g) \leq R^{\text{emp}}(g) + 2\mathcal{R}(\mathcal{G}) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

The proof of this requires a powerful tool called a concentration inequality. Actually, Hoeffding's inequality is a concentration inequality, so you've seen one already. Concentration inequalities say that as m increases, the empirical average is concentrated around the expectation. McDiarmid's concentration inequality is my favorite, and it generalizes Hoeffding's in that applies to functions that depend on m iid random variables:

Theorem. (McDiarmid's Inequality) *Assume for all* $i = 1, ..., m$

$$\sup_{z_1,...,z_m,z_i'} |F(z_1, ..., z_i, ..., z_m) - F(z_1, ..., z_i', ..., z_m)| \leq c$$

then for all $\epsilon > 0$,

$$\mathbb{P}[|F - \mathbb{E}[F]| \geq \epsilon] \leq 2\exp\left(-\frac{2\epsilon^2}{mc^2}\right).$$

This is a Hoeffding-like inequality that holds for any function of m variables, as long as replacing one of those variables by another one won't allow the function to change too much.

Proof of Theorem-RademacherBound

We'll follow two steps:

- *concentration* to relate $\sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp}}g)$ to its expectation,

- *symmetrization* to relate the expectation to the Rademacher average.

We want to use McDiarmid's inequality on $\sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp}}g)$, so we'll need to show that if we modify one training example, it doesn't change too much. Denote $P^{\text{emp,i}}$ as the empirical measure obtained by replacing z_i by z_i' of the sample. The following holds:

$$|\sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp}}g) - \sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp,i}}g)| \leq \sup_{g\in\mathcal{G}}|P^{\text{emp,i}}g - P^{\text{emp}}g|. \quad (5)$$

This isn't too hard to check. For instance, let's say that the first term is larger than the second so we can remove the absolute value. Then say g^* achieves $\sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp}}g)$. Then

$$\begin{aligned}
&\sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp}}g) - \sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp,i}}g) \\
&= (P^{\text{true}}g^* - P^{\text{emp}}g^*) - \sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp,i}}g) \\
&\leq (P^{\text{true}}g^* - P^{\text{emp}}g^*) - (P^{\text{true}}g^* - P^{\text{emp,i}}g^*) \\
&= P^{\text{emp,i}}g^* - P^{\text{emp}}g^* \leq \sup_{g\in\mathcal{G}}(P^{\text{emp,i}}g - P^{\text{emp}}g)
\end{aligned}$$

And if the second term is larger than the first, we do an identical calculation. So (5) holds.

$P^{\text{emp,i}}g - P^{\text{emp}}g$ is a difference of two averages, since remember $P^{\text{emp}}g$ is $\frac{1}{m}\sum_{i=1}^m g(Z_i)$ and $P^{\text{emp,i}}g$ is the same thing except that the ith term got replaced with $g(Z_i')$.

$$|P^{\text{emp,i}}g - P^{\text{emp}}g| = \frac{1}{m}|g(Z_i') - g(Z_i)| \leq \frac{1}{m}$$

where we used that $g \in \{0, 1\}$ for the last inequality.

This means from (5) that we have

$$|\sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp}}g) - \sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp,i}}g)| \leq \frac{1}{m}, \quad (6)$$

the function $F = \sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp}}g)$ can't change more than $\frac{1}{m}$ when we fiddle with one of the Z_i's. This means we can directly apply McDiarmid's inequality to it with $c = \frac{1}{m}$.

$$\mathbb{P}_{\mathbf{Z}\sim D^m}[|\sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp}}g) - \mathbb{E}_{\mathbf{Z}\sim D^m}[\sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp}}g)]| \geq \epsilon] \leq 2\exp\left(-\frac{2\epsilon^2}{m\frac{1}{m^2}}\right).$$

That's the first part of the proof (the concentration part).

Now for symmetrization, which we'll use to prove that the expected difference between the true and empirical risks is bounded by twice the Rademacher average.

Lemma.

$$\mathbb{E}_{\mathbf{Z}\sim D^m} \sup_{g\in\mathcal{G}}[P^{\text{true}}g - P^{\text{emp}}g] \leq 2\mathbb{E}_{\boldsymbol{\sigma},\mathbf{Z}} \sup_{g\in\mathcal{G}} R_m g = 2R(\mathcal{G})$$

To prove it, we introduce a ghost sample and it's corresponding measure P'^{emp}. We're going to use that $\mathbb{E}_{\mathbf{Z}\sim D^m} P'^{\text{emp}}g = P^{\text{true}}g$ in the second line below.

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Z}\sim D^m} \sup_{g\in\mathcal{G}}[P^{\text{true}}g - P^{\text{emp}}g] \\
&= \mathbb{E}_{\mathbf{Z}\sim D^m} \sup_{g\in\mathcal{G}}[\mathbb{E}_{\mathbf{Z}'\sim D^m}[P'^{\text{emp}}g - P^{\text{emp}}g]] \\
&\leq \mathbb{E}_{\mathbf{Z}\sim D^m \mathbf{Z}'\sim D^m} \sup_{g\in\mathcal{G}}[P'^{\text{emp}}g - P^{\text{emp}}g] \quad \text{(uses Jensen's Inequality, sup is convex)} \\
&= \mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m}(g(Z_i') - g(Z_i))\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma},\mathbf{Z},\mathbf{Z}'}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i(g(Z_i') - g(Z_i))\right] \quad \boxed{\text{Why?}} \\
&\leq \mathbb{E}_{\boldsymbol{\sigma},\mathbf{Z}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i g(Z_i')\right] + \mathbb{E}_{\boldsymbol{\sigma},\mathbf{Z}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m}-\sigma_i g(Z_i)\right] \\
&= 2\mathbb{E}_{\boldsymbol{\sigma},\mathbf{Z}} \sup_{g\in\mathcal{G}} R_m g.
\end{aligned}$$

The last step uses that $\sigma_i g(Z_i)$ and $-\sigma_i g(Z_i)$ have the same distribution. ■

Let's put the two parts of the proof of Theorem-RademacherBound together. The first part said:

$$\mathbb{P}_{\mathbf{Z}\sim D^m}[|\sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp}}g) - \mathbb{E}_{\mathbf{Z}\sim D^m}[\sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp}}g)]| \geq \epsilon] \leq 2\exp\left(-\frac{2\epsilon^2}{1/m}\right).$$

The second part said:

$$\mathbb{E}_{\mathbf{Z}\sim D^m} \sup_{g\in\mathcal{G}}[P^{\text{true}}g - P^{\text{emp}}g] \leq 2\mathbb{E}_{\boldsymbol{\sigma},\mathbf{Z}} \sup_{g\in\mathcal{G}} R_m g = 2R(\mathcal{G})$$

Let's fiddle with the first part. First, let

$$\delta := 2\exp\left(-\frac{2\epsilon^2}{1/m}\right) \quad \text{so that} \quad \epsilon = \sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

So the first part looks like

$$\mathbb{P}_{\mathbf{Z}\sim D^m}\left[\left|\sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp}}g) - \mathbb{E}_{\mathbf{Z}\sim D^m}\left[\sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp}}g)\right]\right| \geq \sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right] \leq \delta.$$

Using inversion, with probability at least $1-\delta$,

$$\left|\sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp}}g) - \mathbb{E}_{\mathbf{Z}\sim D^m}\left[\sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp}}g)\right]\right| \leq \sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

I can remove the absolute value and the left hand side only gets smaller. Then,

$$\sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp}}g) \leq \mathbb{E}_{\mathbf{Z}\sim D^m}\left[\sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp}}g)\right] + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

Using the second part,

$$\sup_{g\in\mathcal{G}}(P^{\text{true}}g - P^{\text{emp}}g) \leq 2\mathcal{R}(\mathcal{G}) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

Writing it all together, with probability at least $1-\delta$, for all $g \in \mathcal{G}$,

$$P^{\text{true}}g \leq P^{\text{emp}}g + 2\mathcal{R}(\mathcal{G}) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}},$$

and that's the statement of Theorem-RademacherBound.■

Relating the Rademacher Complexity of the Loss Class and the Initial Class

Theorem RademacherBound is nice, but it only applies to the class $\mathcal{G}$, and we also care about class $\mathcal{F}$, so we need to relate the Rademacher average of $\mathcal{G}$ to

that of class $\mathcal{F}$. We do this as follows, using the fact that σ_i and $Y_i\sigma_i$ have the same distribution.

$$\begin{aligned}
\mathcal{R}(\mathcal{G}) &= \mathbb{E}_{\boldsymbol{\sigma},\mathbf{Z}}\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\mathbf{1}_{[f(X_i)\neq Y_i]}\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma},\mathbf{Z}}\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\frac{1}{2}(1-Y_i f(X_i))\right] \quad \boxed{\text{Why?}} \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma},\mathbf{Z}}\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i Y_i f(X_i)\right] = \frac{1}{2}\mathcal{R}(\mathcal{F}). \quad \boxed{\text{Why?}}
\end{aligned}$$

Ok so we've related the two classes' Rademacher averages. So, with probability at least $1-\delta$,

$$\forall f\in\mathcal{F},\quad R^{\text{true}}f \leq R^{\text{emp}}f + \mathcal{R}(\mathcal{F}) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

Computing the Rademacher Averages

Let's assess the difficulty of actually computing Rademacher averages. Start here:

$$\begin{aligned}
\frac{1}{2}\mathcal{R}(\mathcal{F}) &= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma},\mathbf{Z}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(X_i)\right] \\
&= \frac{1}{2} + \mathbb{E}_{\boldsymbol{\sigma},\mathbf{Z}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}-\frac{1-\sigma_i f(X_i)}{2}\right] \\
&= \frac{1}{2} - \mathbb{E}_{\boldsymbol{\sigma},\mathbf{Z}}\left[\inf_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\frac{1-\sigma_i f(X_i)}{2}\right] \\
&= \frac{1}{2} - \mathbb{E}_{\boldsymbol{\sigma},\mathbf{Z}}\left[\inf_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_{f(X_i)\neq\sigma_i}\right]. \quad (1/2 \text{ comes through the inf})
\end{aligned}$$

Look closely at that last expression. The thing on the inside says to find the empirical risk minimizer, where the labels are the σ_i's. This expression shows that computing $\mathcal{R}(\mathcal{F})$ is not any harder than computing the expected empirical

risk minimizer over random labels. So we could design a hypothetical procedure, where we generate the σ_i's randomly, and minimize the empirical error in $\mathcal{G}$ with respect to the labels σ_i. (Of course we technically also have to do this over the possible X_i's, but there are some ways to get around this that I won't go into here.)

It's also true that

$$\mathcal{R}(\mathcal{F}) \leq 2\sqrt{\frac{h \log \frac{em}{h}}{m}}$$

So we could technically use the Rademacher complexity bound to get another VC-bound.

So we're done! We showed formally that limiting the complexity leads to better generalization. We gave several complexity measures for a set of models (Growth Function, VC-dimension, VC-entropy, Covering Number, Rademacher averages).

15.097 Prediction: Machine Learning and Statistics
Spring 2012