Mathematics of Machine Learning

Arthur Conmy*

Part II, Lent Term 2021

These notes are based on lectures given (virtually) by Dr R. Shah in Lent term 2021. Credit is also due to Evan Chen for the style file for these notes[1].

§1 Introduction

The course will be divided into three parts:

- Statistical learning theory (including empirical risk minimization).
- Computation (including (stochastic) gradient descent).
- Popular methods in practise.

The course will move from a more theoretical background to practical things. There is a significant gap between the theory and the practise in machine learning, as will be seen in the course.

§1.1 Conditional Expectation

The results stated here are true subject to certain convergence conditions. II Stochastic Financial Models and II Probability and Measure deal with such things more formally. For our purposes, we will use the results as tools to develop theory.

Definition 1.1 (Conditional density)**.** Let Z and W be random variables with joint density $f(z, w)$, and let $f_W(w)$ be the marginal density of W (integrate over all z values). Then the **conditional density** of z given w is

$$f_{Z|W}(z, w) = \begin{cases} f(z, w)/f_W(w) & \text{where } f_W(w) \neq 0. \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Definition 1.2 (Conditional expectation)**.** In the same notation as above, we define the *random variable* $\mathbb{E}[Z \mid W]$ as

$$\mathbb{E}[Z \mid W] = \int z f_{Z|W}(z, W) \mathrm{d}z. \tag{2}$$

Note that $\mathbb{E}[Z \mid W]$ is a function of a random variable (W), so is itself a random variable. It is not simply a number, as ordinary expectations are.

We can interpret it as 'the function of W that's our best guess for Z given only the information contained in W'[2]. This interpretation gives some intuition behind the next

*Please send any corrections and/or feedback to asc70@cam.ac.uk.

[1] Available here: https://github.com/vEnhance/dotfiles/blob/master/texmf/tex/latex/evan/evan.sty.

[2] see https://dynalist.io/d/bx3GM7El5D_PsHTOvxgJlTyW

result, which generalises what we did in IA.

Theorem 1.3 (General tower property)
Let $f : \mathbb{R}^d \to \mathbb{R}^m$ and Z and W be random variables. Then

$$\mathbb{E}[\mathbb{E}[Z \mid W] \mid f(W)] = \mathbb{E}[Z \mid f(W)]. \tag{3}$$

Theorem 1.4 (Taking out what is known)
Let f be real-valued. Then

$$\mathbb{E}[f(W)Z \mid W) = f(W)\mathbb{E}[Z \mid W] \tag{4}$$

Theorem 1.5 (Best least squares predictor)
The following holds:

$$\mathbb{E}\left[Z - f(W)\right]^2 = \mathbb{E}\left[Z - \mathbb{E}\left[Z \mid W\right]\right]^2 + \mathbb{E}\left[f(W) - \mathbb{E}\left[Z \mid W\right]\right]^2 \tag{5}$$

We will write out the extended details of how to apply our conditional expectation theory to get this result.

Proof. Initially, add the obvious term to the LHS:

$$\begin{aligned}\mathbb{E}\left[Z - f(W)\right]^2 &= \mathbb{E}\left[Z - \mathbb{E}\left[Z \mid W\right] + \mathbb{E}\left[Z \mid W\right] - f(W)\right]^2 \\ &= \mathbb{E}\left[Z - \mathbb{E}\left[Z \mid W\right]\right]^2 + \mathbb{E}\left[f(W) - \mathbb{E}\left[Z \mid W\right]\right]^2 \\ &\quad -2\mathbb{E}\left[Z - \mathbb{E}\left[Z \mid W\right]\right]\mathbb{E}\left[f(W) - \mathbb{E}\left[Z \mid W\right]\right].\end{aligned}$$

So we need to show that $\mathbb{E}\left[Z - \mathbb{E}\left[Z \mid W\right]\right]\mathbb{E}\left[f(W) - \mathbb{E}\left[Z \mid W\right]\right] = 0$.
We do this by using the tower property to insert a condition on W:

$$\begin{aligned}&\mathbb{E}\left[Z - \mathbb{E}\left[Z \mid W\right]\right]\mathbb{E}\left[f(W) - \mathbb{E}\left[Z \mid W\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[Z - \mathbb{E}\left[Z \mid W\right]\right]\mathbb{E}\left[f(W) - \mathbb{E}\left[Z \mid W\right]\right] \mid W\right].\end{aligned}$$

From here we can pull out the latter expectation term, since it's a function of W, by (1.4). The former term left inside the expectation is 0, since

$$\begin{aligned}\mathbb{E}\left[Z - \mathbb{E}\left[Z \mid W\right] \mid W\right] &= \mathbb{E}\left[Z \mid W\right] - \mathbb{E}\left[\mathbb{E}\left[Z \mid W\right] \mid W\right] \\ &= \mathbb{E}\left[Z \mid W\right] - \mathbb{E}\left[Z \mid W\right] = 0\end{aligned}$$

using the tower property once more.

□

We'll introduce terminology in the next section that means that this result is saying

that the hypothesis $h : \mathcal{X} \to \mathbb{R}$ minimising $R(h)$ under squared error loss is $h_0(x) = \mathbb{E}[Y \mid X = x]$, and nothing deeper than this (by considering $Z = Y$ and $X = W$).

Theorem 1.6 (Conditional Jensen)
Given convex $f : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[f(Z) \mid W] \geq f(\mathbb{E}[Z|W]). \tag{6}$$

To remember which way round the inequality goes, do a sanity check of W constant and $f(x) = x^2$ (variance is non-negative).

§1.2 Terminology

We will set up definitions in order to solve classification problems like MNIST or spam detection.

Consider a pair of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with joint distribution P_0. We call X the **input** or **predictor** and Y the **output** or **response**.

Our goal is to predict Y from X. We do this via a **hypothesis**[4] $h : \mathcal{X} \to \mathcal{Y}$, and measure the quality of the prediction using a **loss function** $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

We can be in the **classification** setting where $\mathcal{Y} = \{-1, 1\}$, and typically ℓ is the **misclassification loss**, or '0-1 loss' $\ell(h(x), y) = \mathbb{1}[h(x) \neq y]$. Here, we refer to h as a **classifier**.

Alternatively, we can be in the **regression** setting, where $\mathcal{Y} = \mathbb{R}$ and typically ℓ is the squared error: $\ell(h(x), y) = (h(x) - y)^2$.

Our aim is to pick h with small **risk**

$$R(h) = \mathbb{E}[\ell(h(X), Y) \mid h]. \tag{7}$$

We have the conditioning over h as we consider the classifier to be fixed; h will be generally constructed from some random data, and this takes that into account.

A classifier h_0 that minimises the 0-1 risk is called a **Bayes classifier**. Its associated risk is the **Bayes risk**.

Define the **regression function** η as

$$\eta(x) = \mathbb{P}(Y = 1 \mid X = x). \tag{8}$$

Note that in practise we don't have 'access' to η, since to know η we need to know the joint distribution to evaluate it. When minimising Bayes risk, we're essentially minimising *population-wide* risk. We'll go on to study empiricial risk minimisation (ERM), where we minimise risk with respect to our training data, or data we observe empirically.

[4] the use of the term is unlike the use in 'hypothesis testing' from statistics.

Theorem 1.7

A Bayes Classifier is given by

$$h_0(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ -1 & \text{otherwise.} \end{cases} \tag{9}$$

Proof. This is not a deep result. The 0-1 risk is just the dumb function '0 if we were right, 1 if we were wrong' so the risk, given $X = x$ is

$$\begin{aligned} R(h(x)) &= \mathbb{1}[h(x) = 1]\,\mathbb{P}[Y = -1 \mid X = x] + \mathbb{1}[h(x) = -1]\,\mathbb{P}[Y = 1 \mid X = x] \\ &= \mathbb{1}[h(x) = 1]\,(1 - \eta(x)) + \mathbb{1}[h(x) = -1]\,\eta(x). \end{aligned}$$

Now we want to minimise this, so just do casework on $\eta(x) < \frac{1}{2}$, $\eta(x) = \frac{1}{2}$ and $\eta(x) > \frac{1}{2}$. In each case our Bayes classifier will be optimal. □

In reality, we have **training data** that is a set of iid copies (X_1, Y_1), (X_2, Y_2), $\cdots$, (X_n, Y_n) of (X, Y). We want to use this to construct $\hat{h}$ minimising $R(\hat{h})$. Therefore the conditioning in (7) will be on this training data, when considering $R(\hat{h})$.

The classical statistical approach is to model P_0 using a parametric family. In this approach, we need estimate these unknown parameters.

The machine learning approach is that we're given a class $\mathcal{H}$ from which we will then pick $\hat{h}$. In this approach, we will need an algorithm for picking $\hat{h}$.

Definition 1.8 (Sign)**.** sgn is the sign (± 1) of a real, and we define $\text{sgn}(0) = -1$ in this course.

For this reason, in the two-class classification problem, we will label the classes with $+1$ and -1 rather than 0 and 1, as seen elsewhere.

Example 1.9 (Examples of $\mathcal{H}$)

$$\mathcal{H} = \{x \mapsto \text{sgn}(\mu + x^T\beta)\} \tag{10}$$

where $\mu \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$ is one example of a class. Another is

$$\mathcal{H} = \left\{ x \mapsto \text{sgn}\left(\sum_j \phi_j(x) w_j \right) \right\} \tag{11}$$

where $w \in \mathbb{R}^d$ and $\phi_j \in \mathbb{B}$ for a given class of functions $\mathbb{B} = \{f : \mathcal{X} \to \mathbb{R}\}$. Note that these classes have, in general, many degrees of freedom.

§2 Statistical Learning Theory

As alluded to earlier, minimising population-wide risk is generally not practical nor possible when applying ML methods. So almost all ML methods aim to minimise empirical risk:

Definition 2.1 (Empirical Risk)**.** Empirical risk or **training error** is the expectation of the loss $\ell(h(X), Y)$ where (X, Y) follows the *empirical* distribution (which will generally (always?) be some number of iid samples from P_0):

$$\hat{R}(h) = \frac{1}{n} \sum_i \ell(h(X_i), Y_i) \tag{12}$$

Given some class $\mathcal{H}$ of hypotheses, the argmin over this set with respect to this empirical risk is called the empirical risk minimiser.

We've used a lot of words, but there are immediate examples that this setup allows us to describe:

Example 2.2 (Least squares regression from IB Statistics)

In this familiar setting, our class of hypotheses is

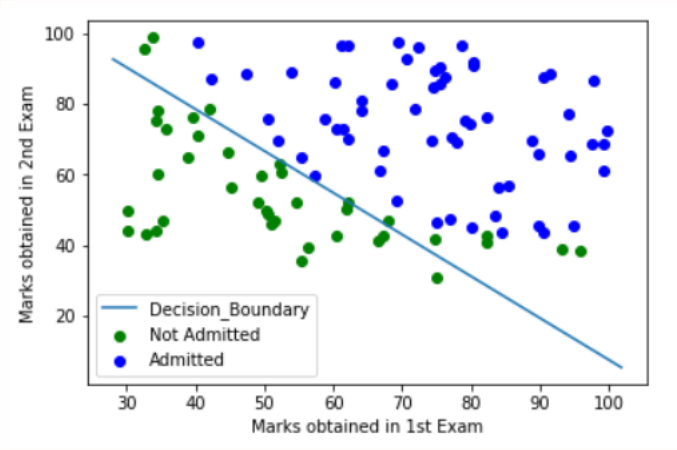
$$\mathcal{H} = \{x \mapsto (\mu + x^T \beta)\} \tag{13}$$

where $\mu \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$, and our loss function is squared error

$$\ell(X_i, Y_i) = (Y_i - \mu - X_i^T \beta)^2. \tag{14}$$

Example 2.3 (0-1 classification)

If $\mathcal{Y} = \{-1, +1\}$, $\mathcal{H} = \{x \mapsto \text{sgn}\mu + x^T\beta$ (see (10)) and our loss is 0-1 loss, then what we're doing here is finding a hyperplane that divides space into two regions, one for each of -1 and 1, and the empirical loss will be the number of 1s in the -1 region plus the number of -1s in the 1 region, all over n. Visually:



The hyperplane in this case is known as the *decision boundary*.

Definition 2.4 (Aitch, aitch and aitch)**.** Let $\hat{h} \in \mathcal{H}$ be the hypothesis minimising empirical risk.

Let $h^* \in \mathcal{H}$ be the hypothesis minimising (population) risk over all $\mathcal{H}$.

Let h_0 be the function minimising (population) risk over all functions $h : \mathcal{X} \mapsto \mathcal{Y}$.

Note that h_0, the 'God hypothesis' won't be a perfect predictor, since our data comes from P_0, an inherently random distribution. We can now consider

$$R(\hat{h}) - R(h_0) = \underbrace{R(\hat{h}) - R(h^*)}_{\substack{\textbf{Excess risk} \text{ from our} \\ \text{choice within hypothesis class.}}} + \underbrace{R(h^*) - R(h_0)}_{\substack{\textbf{Appoximation error} \text{ from our} \\ \text{choice of hypothesis class.}}} \tag{15}$$

where the terms have the given interpretations. Furthermore, the first term can be seen as a measure of how much we overfit to the training data, and the latter term how much we underfit the underlying distribution by choosing too restrictive a $\mathcal{H}$.

Write

$$R(\hat{h}) - R(h^*) = R(\hat{h}) - \hat{R}(\hat{h}) + \underbrace{\hat{R}(\hat{h}) - \hat{R}(h^*)}_{\leq 0} + \hat{R}(h^*) - R(h^*) \tag{16}$$

$$\leq R(\hat{h}) - \hat{R}(\hat{h}) + \hat{R}(h^*) - R(h^*). \tag{17}$$

where the middle term is non-positive since we chose $\hat{h}$ to be optimal on the training data.

We've interpreted all these terms already, except $\hat{R}(h^*)$. This is the risk of the optimal classifier with respect to all $\mathcal{H}$, but when it can only see the training data. So it is another measure of how much we are overfitting.

§2.1 Sub-Gaussianity and Hoeffding's Inequality

Recall Markov's inequality from 1A Probability. Note that this is a simple inequality, but has a nice consequence (on an example sheet from that course).

Theorem 2.5 (Chernoff Bound)

Let W be a random variable and $\alpha > 0$. Then

$$\mathbb{P}[W \geq t] \leq e^{-\alpha t}\mathbb{E}\left[e^{\alpha W}\right] \tag{18}$$

Proof. For any increasing function $\phi : \mathbb{R} \to [0, \infty)$, $W \geq t$ implies $\phi(W) \geq \phi(t)$, so in this case

$$\mathbb{P}[W \geq t] \leq \mathbb{P}\left[e^{\alpha W} \geq e^{\alpha t}\right] \leq e^{-\alpha t}\mathbb{E}\left[e^{\alpha W}\right] \tag{19}$$

by Markov's inequality, as required. □

This is important since this has introduced an MGF.

Example 2.6 (A tail bound for Gaussian random variables)

Suppose $W \sim N(0, \sigma^2)$, so has MGF $\mathbb{E}\left[e^{\alpha W}\right] = \exp\left(\frac{1}{2}\alpha^2\sigma^2\right)$. Then for $t > 0$,

$$\mathbb{P}[W \geq t] \leq \exp\left(-\alpha t + \frac{1}{2}\alpha^2\sigma^2\right). \tag{20}$$

Taking the infinimum over all $\alpha > 0$, we get

$$\mathbb{P}[W \geq t] \leq \exp\left(-t/2\sigma^2\right). \tag{21}$$

This derivation essentially uses bounding with MGFs. This motivates

Definition 2.7 (Sub-Gaussian)**.** A random variable W is **sub-Gaussian** with parameter $\sigma > 0$ if

$$\mathbb{E}\left[\exp\left(\alpha(W - \mathbb{E}\left[W\right])\right] \leq \exp\left(\frac{1}{2}\alpha^2\sigma^2\right) \tag{22}$$

for every $\alpha \in \mathbb{R}$.

i.e. after normalising to mean 0, the MGF of W is always less than the MGF of a variance $N(0, \sigma^2)$ random variable.

Example 2.8 (Basic properties of sub-Gaussian random variables)
Let W be a sub-Gaussian random variable with parameter $\sigma > 0$. Then

- W is sub-Gaussian for all $\sigma' \geq \sigma$.
- $-W$ is also sub-Gaussian.
- $\mathbb{P}\left[W - \mathbb{E}\left[W\right] \geq t\right] \leq \exp\left(-t^2/2\sigma^2\right)$.
- $\mathbb{P}\left[|W - \mathbb{E}\left[W\right]| \geq t\right] \leq 2\exp\left(-t^2/2\sigma^2\right)$

Proof. The first two remarks are immediate. To see the third, apply the Chernoff bound since it directly introduces an MGF. The fourth is then a corollary of the third, using the second.

The third property is called the *sub-Gaussian tail bound*, and we will repeatedly use it. □

Definition 2.9. A *Rademacher random variable* ε takes values ± 1 with equal probability.

Theorem 2.10
A Rademacher random variable is sub-Gaussian with $\sigma = 1$.

Proof. Directly compute

$$\begin{aligned}\mathbb{E}\left[e^{\alpha\varepsilon}\right] = \frac{1}{2}(e^{\alpha} + e^{-\alpha}) &= \sum_{k=0}^{\infty} \frac{\alpha^{2k}}{(2k)!} \\ &\leq \sum_{k=0}^{\infty} \left(\frac{\alpha^2}{2}\right)^k \frac{1}{k!} \\ &= e^{\alpha^2/2}.\end{aligned}$$

□

How does this relate to ERM? This brings us onto the first deeper result of the course.

Theorem 2.11 (Hoeffding's Lemma)

If W takes values in an interval $[a, b]$, then W is sub-Gaussian with parameter $(b-a)/2$.

Proof. We will prove the weaker result with $\sigma = b - a$, using a technique called *symmetrisation*. The full result will be on the example sheet, using a *change of measure* argument[a].

To check whether a random variable is sub-Gaussian we normalize so the mean is zero, so WLOG assume $\mathbb{E}[W] = 0$.

We want to use the results already established for Rademacher random variables, and to do this we cook up W' an independent copy of W so that the random variable $W - W'$ is symmetric about 0, meaning that $\varepsilon(W - W')$ and $W - W'$ have the same distribution (written $W - W' \overset{d}{=} \varepsilon(W - W')$). We write

$$\mathbb{E}\left[e^{\alpha\varepsilon(W-W')} \mid W, W'\right] \leq \exp\left(\alpha(W-W')^2/2\right) \leq \exp\left(\alpha^2(b-a)^2/8\right), \tag{23}$$

where conditioning on W, W' effectively fixes these random variables, allowing us to use the Rademacher result.

We relate this back to W by noting that the familiar linearity of expectation property

$$\mathbb{E}[X + Y \mid Z] = \mathbb{E}[X \mid Z] + \mathbb{E}[Y \mid Z] \tag{24}$$

(where all the expectations are defined) holds, so we have that

$$\alpha(W - \mathbb{E}[W]) = \mathbb{E}\left[\alpha(W - W') \mid W\right]. \tag{25}$$

Therefore

$$\begin{aligned}
\mathbb{E}[\exp(\alpha W)] &= \mathbb{E}\left[\exp\left(\alpha(W - \mathbb{E}\left[W'\right])\right)\right] && (26)\\
&= \mathbb{E}\left[\exp\left(\mathbb{E}\left[\alpha(W - W') \mid W\right]\right)\right] && (27)\\
&\leq \mathbb{E}\left[\exp\left(\alpha(W - W')\right)\right]. && (28)
\end{aligned}$$

where in the last step we apply conditional Jensen, and then simplify via the tower property in the degenerate case conditioning on some constant random variable, to go from to expectation signs to just one.

Finally, we can use the tower property once more to write the equality

$$\mathbb{E}\left[\exp\left(\alpha(W - W')\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\exp\left(\alpha\varepsilon(W - W')\right) \mid W, W'\right]\right] \tag{29}$$

where to be specific, we choose the f in the tower property statement to be some degenerate (e.g constant) function to complete the argument (in essence we omit something like '... | 42]' at the end of both of the above expectation expressions). We've recovered (23), so we're done.

□

[a] see [2], also.

Theorem 2.12

Suppose $W_1, \dots, W_n$ are independent and each W_i is sub-Gaussian with parameters $\sigma_i > 0$. Then for all $\gamma \in \mathbb{R}^n$, $\gamma^T W$ is sub-Gaussian with parameter

$$\sqrt{\sum_i \gamma_i^2 \sigma_i^2}. \tag{30}$$

Proof. As before, WLOG $\mathbb{E}[W_i] = 0$. Then consider the MGF of $\gamma^T W$, which factorises via independence.

□

Theorem 2.13 (Hoeffding's Inequality)

Suppose $W_1, \dots, W_n$ are independent and bounded random variables, with $a_i \leq W_i \leq b_i$.

$$\mathbb{P}\left[\frac{1}{n}\sum_i (W_i - \mathbb{E}[W_i]) \geq t\right] \leq \exp\left(\frac{-2n^2t^2}{\sum_i (a_i - b_i)^2}\right) \tag{31}$$

Proof. This is in the form of (2.12), so apply (2.8). □

Theorem 2.14 (Upper bound on sub-Gaussian, mean zero random variables)

Suppose $W_1, \dots, W_d$ are sub-Gaussian random variables with mean 0 and parameter $\sigma > 0$. Then

$$\mathbb{E}[\max W] \leq \sigma\sqrt{2\log d}. \tag{32}$$

Proof. We see sub-Gaussian things, so we gravitate towards an MGF. For $\alpha > 0$,

$$\begin{aligned} \mathbb{E}[\max(\exp(\alpha W))] &= \mathbb{E}[\exp(\alpha \max W)] && (33)\\ &\geq \exp(\alpha \mathbb{E}[\max W]) && (34) \end{aligned}$$

by applying Jensen. So we can concern ourselves with that initial expression, and crudely bound the maximum with a sum (when we can't make further progress because we make almost no distributional assumptions in this course, we will often do this 'union bound' trick):

$$\mathbb{E}[\max(\exp(\alpha W))] \leq \sum \mathbb{E}[\exp(\alpha W)] \leq d\exp\left(\alpha^2\sigma^2/2\right) \tag{35}$$

now this works for any α, and after rearranging and finding the α that gives us the sharpest inequality, we get the result.

□

It's worth noting that while this result makes no distributional assumptions on the W_i, it is 'more striking' in some sense when the W_i are independent; where the variables are well correlated, the max won't differ too much from, say, W_1.

§2.2 Finite Hypotheses Classes

We have developed enough theory to prove an important bound on the risk of the ERM.

Theorem 2.15 (Finite hypothesis class excess risk bound)

Suppose $\mathcal{H}$ is finite and ℓ takes values in $[0, M]$. Then with probability at least $1-\delta$, the ERM $\hat{h}$ satisfies

$$R(\hat{h}) - R(h^*) \leq M\sqrt{\frac{2(\log|\mathcal{H}| + \log(1/\delta))}{n}}. \tag{36}$$

Proof. In the notation of Hoeffding's inequality (2.13), we have $a_i = 0$ and $b_i = M$, and the RHS of the bound of that inequality is therefore $\exp\left(-2nt^2/M^2\right)$.

Recall we can write

$$R(\hat{h}) - R(h^*) = R(\hat{h}) - \hat{R}(\hat{h}) + \underbrace{\hat{R}(\hat{h}) - \hat{R}(h^*)}_{\leq 0} + \hat{R}(h^*) - R(h^*) \tag{37}$$

(where the inequality holds since $\hat{h}$ is optimal over the training data).

Let $t > 0$. Then

$$\mathbb{P}\left[R(\hat{h}) - R(h^*) > t\right] = \mathbb{P}\left[R(\hat{h}) - R(h^*) > t, \hat{h} \neq h^*\right] \tag{38}$$

$$\leq \mathbb{P}\left[R(\hat{h}) - \hat{R}(h^*) > t/2, \hat{h} \neq h^*\right] + \mathbb{P}\left[\hat{R}(h^*) - R(h^*) > t/2\right] \tag{39}$$

since the two last events imply the former event. The latter term is actually pretty much in the form of Hoeffding's inequality, specifically

$$\mathbb{P}\left[\hat{R}(h^*) - R(h^*) > t/2\right] = \mathbb{P}\left[\frac{1}{n}\sum \ell(h^*(X_i), Y_i) - \mathbb{E}\left[\ell(h^*(X_i), Y_i)\right] > t/2\right] \tag{40}$$

$$\leq \exp\left(-\frac{nt^2}{2M^2}\right) \tag{41}$$

by that result. For the first term, we have to get a bit messy (as we have already done by excluding h^* as we did) to be able to get a clean final form, and introduce $\mathcal{H}^- = \mathcal{H} \setminus \{h^*\}$. When $\hat{h} \in \mathcal{H}^-$, $R(\hat{h}) - \hat{R}(\hat{h}) \leq \max_{h\in\mathcal{H}^-} R(h) - \hat{R}(h)$. Now we know nothing about h^* and therefore have to use crude union bounding once more:

$$\mathbb{P}\left[R(\hat{h}) - \hat{R}(h^*) > t/2, \hat{h} \neq h^*\right] \leq \mathbb{P}\left[\max_{h\in\mathcal{H}^-} R(h) - \hat{R}(h) > t/2\right] \tag{42}$$

$$= \mathbb{P}\left[\bigcup_{h\in\mathcal{H}^-} \{R(h) - \hat{R}(h) > t/2\}\right] \tag{43}$$

$$\leq \sum_{h\in\mathcal{H}^-} \mathbb{P}\left[R(h) - \hat{R}(h) > t/2\right] \tag{44}$$

$$\leq (|\mathcal{H}| - 1)\exp\left(-\frac{nt^2}{2M^2}\right) \tag{45}$$

So

$$\mathbb{P}\left[R(\hat{h}) - R(h^*) > t\right] \leq |\mathcal{H}|\exp\left(-nt^2/2M^2\right) = \delta \tag{46}$$

and we get the result by rearranging this for t. □

Note that a similar bound could be found by applying the central limit theorem, since our setup involves a bunch of iid random variables. However, our result is not asymptotic, unlike the limit theorems we've seen before.

This result is something to be positive about: in loose terms, even if $\mathcal{H}$ is pretty large, we don't need that much training data to ensure our ERM has low excess risk.

Example 2.16 (The histogram classifier)

Consider the classification setting $\mathcal{X} = [0,1)^2$. Divide $[0,1)^2$ into m^2 disjoint squares $R_0, \ldots, R_{m^2-1}$ where

$$R_{im+j} = \left[\frac{i}{m}, \frac{i+1}{m}\right) \times \left[\frac{j}{m}, \frac{j+1}{m}\right). \tag{47}$$

Also let

$$\bar{Y_j} = \operatorname{sgn} \sum_{i: X_i \in R_j} Y_i \tag{48}$$

i.e. these output what the majority of the points in each square are (there's annoying mismatch between -1 and 1, and 0 and 1 here). And finally

$$\hat{h}^{\text{hist}}(x) = \sum_{j=0}^{m^2-1} \hat{Y}_j \mathbb{1}\left[x \in R_j\right] \tag{49}$$

i.e. classify based on plurality of training data in the region that the test data lands in.

With $\mathcal{H}$ being the size 2^{m^2} set of hypotheses classifying based on which $1/m$ square a point falls in, we can bound the risk with (2.15).

It can be shown that we approach the Bayes classifier in this scenario, as we increase m (in some limit scenario).

§2.3 Infinite Hypotheses Classes

To handle the case where the hypothesis class is infinite, we turn maxima into suprema: let

$$G(Z_1, ..., Z_n) = \sup_{h \in \mathcal{H}} R(h) - \hat{R}(h) \tag{50}$$

where $Z_i = (X_i, Y_i)$. What is more problematic here is that our usual union bounding trick will fail miserably.

However a key property that we used before is that when we have a bunch of iid random variables $Z_1, ..., Z_n \in \mathcal{X} \times \mathcal{Y}$ and $\varepsilon > 0$ let $\tilde{h} \in \mathcal{H}$ be an 'ε-good' hypothesis in the sense that

$$G(z_1, ..., z_n) < R(\tilde{h}) - \hat{R}(\tilde{h}) + \varepsilon \tag{51}$$

Consider perturbing WLOG the first argument of G. Then (writing out everything...)

$$G(z_1, ..., z_n) - G(z_1', z_2, ..., z_n) < \frac{1}{n}\{\ell(y_1', \tilde{h}(x_1')) - \ell(y_1, \tilde{h}(x_1))\} + \varepsilon \tag{52}$$

which is a formalisation of the intuitively clear idea that individual data points do not affect the 'global' gap in excess risk too much (due to the factor of $1/n$).

In fact if the loss takes values in $[0, M]$, and $\varepsilon > 0$ is arbitrary,

$$|G(z_1, ..., z_n) - G(z_1, ..., z_i', ..., z_n)| \leq \frac{M}{n}. \tag{53}$$

for an perturbed i. Such an inequality is called a **bounded differences property**.

§2.4 Bounded Differences Inequality

For our next result, we will need the following notation and definition:

For a sequence $a_s, a_{s+1}, ...$ write $a_{j:k}$ for the **subsequence** $a_j, ..., a_k$.

Definition 2.17 (Martingale difference sequence)**.** A sequence of random variables $D_1, ..., D_n$ is a **martingale difference sequence** with respect to another sequence of random variables $W_0, ..., W_n$ if, for $1 \leq i \leq n$,

1. $\mathbb{E}[|D_i|] < \infty$.
2. D_i is a function of $W_{0:i}$.
3. $\mathbb{E}[D_i \mid W_{0:i-1}] = 0$.

Example 2.18

If $D_1, ..., D_n$ are independent and mean zero and satisfy the first property, then they are a martingale difference sequence with respect to $c, D_1, ..., D_n$ where c is a deterministic constant.

We will first need two preliminary results to prove the bounded differences inequality

Theorem 2.19 (2.12 for martingale random variables)

Let $D_1, ..., D_n$ be a martingale difference sequence with respect to $W_0, ..., W_n$ such that

$$\mathbb{E}\left[\exp\left(\alpha D_i\right) \mid W_{0:i-1}\right] \leq \exp\left(\frac{1}{2}\alpha^2\sigma_i^2\right) \tag{54}$$

holds for all α and all i. Also let $\gamma \in \mathbb{R}^n$. Then $\sum_i \gamma_i D_i$ is sub-Gaussian with parameter $\sqrt{\sum_i \sigma_i^2\gamma_i^2}$.

Proof. By the tower property

$$\mathbb{E}\left[\exp\left(\alpha\sum_{i=1}^{n}\gamma_i D_i\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\exp\left(\alpha\sum_{i=1}^{n-1}\gamma_i D_i\right)\exp\left(\alpha\gamma_n D_n\right) \mid W_{0:n-1}\right]\right] \tag{55}$$

and now we can use taking out what is known since $D_{0:n-1}$ is a function of $W_{0:n-1}$, which will leave the last term in the inner expectation as something our assumptions gave us control over:

$$\leq \mathbb{E}\left[\exp\left(\alpha\sum_{i=1}^{n-1}\gamma_i D_i\right)\exp\left(\frac{1}{2}\alpha^2\gamma_n^2\sigma_n^2\right)\right] \tag{56}$$

and at this point we can pull out that second factor and apply the same tower trick a further $n-1$ times to get

$$\leq \prod_{i=1}^{n}\exp\left(\frac{1}{2}\alpha^2\gamma_i^2\sigma_i^2\right) = \exp\left(\frac{1}{2}\alpha^2\sum_{i=1}^{n}\gamma_i^2\sigma_i^2\right). \tag{57}$$

□

Theorem 2.20 (Azuma-Hoeffding)

Let $D_1, ..., D_n$ be a martingale difference sequence with respect to $W_0, ..., W_n$. Suppose also that for each i we have random variables A_i and B_i that are bounds for D_i ($A_i \leq D_i \leq B_i$) that differ by at most L_i where the L_is are constant, and that A_i and B_i are also functions of $W_{0:i-1}$. Then for $t \geq 0$

$$\mathbb{P}\left[\sum_{i=1}^{n} D_i \geq t\right] \leq \exp\left(-2t^2 \Big/ \sum_i L_i^2\right). \tag{58}$$

Proof. Conditional on $W_{0:i-1}$, A_i and B_i are fixed, so D_i is a bounded random variable and hence we're in the setting where we can apply Hoeffding's lemma with this conditioning. Therefore we have the moment generating function bound

$$\mathbb{E}\left[e^{\alpha D_i} \mid W_{0:i-1}\right] \leq \exp\left(\frac{1}{2}\alpha^2 \left(\frac{L_i}{2}\right)^2\right) \tag{59}$$

which means we can apply (2.19) to deduce that $\sum_{i=1}^{n} D_i$ is sub-Gaussian with parameter $\frac{1}{2}\sqrt{\sum_i L_i^2}$. So we can apply the sub-Gaussian tail bound and we get the desired result.

□

Theorem 2.21 (Bounded differences inequality)

Let $f : \mathcal{Z}_1 \times ... \times \mathcal{Z}_n \to \mathbb{R}$ satisfy a bounded differences property

$$f(w_{1:n}) - f(w_{1:i-1}, w_i', w_{i+1:n}) \leq L_i \tag{60}$$

$\forall w_1 \in \mathcal{Z}_1, ..., w_i' \in \mathcal{Z}_i, ..., w_n \in \mathcal{Z}_n$ where $1 \leq i \leq n$. Suppose random variables $W_1 \in \mathcal{Z}_1, ..., W_n \in \mathcal{Z}_n$ are independent. Then

$$\mathbb{P}\left[f(W_{1:n}) - \mathbb{E}\left[f(W_{1:n})\right] \geq t\right] \leq \exp\left(-2t^2 \Big/ \sum_i L_i^2\right). \tag{61}$$

Proof. Introduce the deterministic random variable $W_0 = c$ (c arbitrary and constant). Then we can turn our expression into a sum of a bunch of random variables as follows:

$$f(W_{1:n}) - \mathbb{E}\left[f(W_{1:n})\right] = \sum_i \underbrace{\mathbb{E}\left[f(W_{1:n}) \mid W_{0:i}\right] - \mathbb{E}\left[f(W_{1:n}) \mid W_{0:i-1}\right]}_{D_i}. \tag{62}$$

Since this will telescope, and the first term comes out through conditional expectation. Now D_i is a martingale difference sequence with respect to $W_0, ..., W_n$, since checking the definition,

1. f satisfies the bounded differences inequality so itself must be bounded (it will vary by at most $\sum L_i$).
2. Is clear.
3. Follows from the tower property, since $W_{0:i-1}$ is a function of itself, and also a function of $W_{0:i}$.

We now want to cook up A_i and B_i in the notation of Azuma-Hoeffding. Let $\mathcal{Z}_0 = \{c\}$. Define

$$F_i : \mathcal{Z}_0 \times ... \times \mathcal{Z}_i \to \mathbb{R} \tag{63}$$

by $(w_0, ..., w_i) \mapsto \mathbb{E}\left[f(W_{0:n}) \mid W_{0:i} = w_{0:i}\right]$ and so we have

$$D_i = F_i(W_{0:i}) - F_{i-1}(W_{0:i}). \tag{64}$$

Now force $A_i \leq D_i \leq B_i$ in the most blatant way possible:

$$A_i = \inf_{w_i} F_i(W_{0:i-1}, w_i) - F_{i-1}(W_{0:i-1}) \tag{65}$$

and let B_i be the corresponding supremum (recall these are supposed to be functions of $W_{0:i-1}$!). We now have to do a careful check that $B_i - A_i \leq L_i$ by a lot of symbol pushing: to begin note that

$$B_i - A_i = \sup_{w_i, w_i'} F_i(W_{0:i-1}, w_i) - F_i(W_{0:i-1}, w_i') \tag{66}$$

and writing the F expressions out and bringing the expectations together (using independence of all the $W_{0:i}$ as well as the Martingale independence of the future $W_{i+1:n}$ from the past, we get a quantity that's always at most L_i.

□

This is a generalisation of Hoeffding's inequality, by taking $\mathcal{Z}_i = [a_i, b_i]$ and $f(W_{1:n}) = \frac{1}{n}\sum W_i$, where the boundedness of the random variables in Hoeffding's inequality give the bounded differences property.

§2.5 Rademacher Complexity

We still want to develop some more tools in the $|\mathcal{H}| = \infty$ case.

Recall the inequality (17). We can now introduce the supremum quantity G (50) to further write

$$R(\hat{h}) - R(h^*) \leq R(\hat{h}) - \hat{R}(\hat{h}) + \hat{R}(h^*) - R(h^*) \tag{67}$$

$$\leq \underbrace{G - \mathbb{E}[G]}_{\text{controllable by bounded differences (2.21)}} + \underbrace{\hat{R}(h^*) - R(h^*)}_{\text{controllable by Hoeffding's (2.13)}} + \mathbb{E}[G]. \tag{68}$$

so we need to figure out how to control $\mathbb{E}[G]$.

Now define the set of functions (indexed by hypotheses)

$$\mathcal{F} = \{(x, y) \overset{f}{\mapsto} -\ell(h(x), y) | h \in \mathcal{H}\} \tag{69}$$

(check the order of terms in G to verify that we do need a minus sign here) so that we can rewrite

$$G = \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}[f(Z_i)]. \tag{70}$$

We're going to develop a result that controls terms like that in the summand.

For each f we get a mean zero random variable, and we're taking a supremum over these random variables. We've already established a result about the maximum of some mean zero *sub-Gaussian* random variables (2.14); $\mathbb{E}[\max W_i] \leq \sigma\sqrt{2\log d}$. But here d is infinite, so we can't (yet) apply it. So, more technology!

Definition 2.22 (More Rademacher terminology)**.** Let $\mathcal{F}$ be a class of functions $f : \mathcal{Z} \to \mathbb{R}$ and let $z_1, ..., z_n \in \mathcal{Z}$.

Then

$$\mathcal{F}(Z_{1:n}) = \{(f(Z_1), ..., f(Z_n)) | f \in \mathcal{F}\} \tag{71}$$

is the set of 'behaviours' that can be exhibited on $Z_{1:n}$.

Then define the **empirical Rademacher complexity** is

$$\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(z_i)\right]. \tag{72}$$

Given iid $Z_{1:n}$, we sometimes view the empirical Rademacher complexity as a random variable via conditioning:

$$\hat{\mathcal{R}}(\mathcal{F}(Z_{1:n})) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(Z_i) \mid Z_{1:n}\right]. \tag{73}$$

Finally, define the **Rademacher complexity** of $\mathcal{F}$ as $\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\left[\hat{R}(\mathcal{F}(Z_{1:n}))\right]$.

Theorem 2.23

Let $\mathcal{F}$ be a class of functions $f : \mathcal{Z} \to \mathbb{R}$ and let $Z_1, ..., Z_n \in \mathcal{Z}$ be iid random variables. Then

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}\left[f(Z_i)\right]\right] \leq 2\mathcal{R}_n(\mathcal{F}). \tag{74}$$

Proof. Symmetrization.

Let $Z'_{1:n}$ be an independent copy of $Z_{1:n}$ (I don't think there is any assumption that the Z are all distributed in the same way (as training data probably would be). This is a more general situation). We can use this to 'bring the expectation out' of the inner term being expected over as follows, and crucially we end up with a symmetric distribution within the expectation:

$$\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}\left[f(Z_i)\right] = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[f(Z_i) - f(Z'_i)|Z_{1:n}\right] \tag{75}$$

$$\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(Z_i) - f(Z'_i)|Z_{1:n}\right] \tag{76}$$

where in the first equality we use that $\mathbb{E}\left[X|X\right] = X$ and $\mathbb{E}\left[X|Y\right] = \mathbb{E}\left[X\right]$ where X and Y are independent, and in the inequality we use $\sup \mathbb{E}\left[-\right] \leq \mathbb{E}\left[\sup -\right]$ (because each individual expectation is at most the expectation of the sup, so then we can take sup).

Let $\varepsilon_{1:n}$ be iid Rademachers. Magically,

$$\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(Z_i) - f(Z'_i) =^d \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i(f(Z_i) - f(Z'_i)). \tag{77}$$

Now take expectation of the above inequality, and break up the sup as follows:

$$\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(Z_i) + \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(Z'_i)\right] \leq 2\mathcal{R}_n(\mathcal{F}) \tag{78}$$

where we can turn the negative in positive due to the ε_i being symmetric. □

Theorem 2.24

Let $\mathcal{F} = \{(x, y) \mapsto \ell(h(x), y) | h \in \mathcal{H}\}$ and suppose ℓ takes values in $[0, M]$. Then with probability at least $1 - \delta$,

$$R(\hat{h}) - R(h^*) \leq 2\mathcal{R}_n(\mathcal{F}) + M\sqrt{2\log(2/\delta)/n}. \tag{79}$$

Proof. Let $G = \sup_{h\in\mathcal{H}} R(h) - \hat{R}(h)$. We have our old friend, the bound

$$R(\hat{h}) - R(h^*) \leq G - \mathbb{E}[G] + \hat{R}(h^*) - R(h^*) + \mathbb{E}[G] \tag{80}$$

and now we know that that last term is at most $2\mathcal{R}_n(\mathcal{F})$. Now viewing G as a function of all the Z_i, it satisfies a bounded differences property with $L_i = M/n$ (53). So we can apply the bounded differences inequality directly:

$$\mathbb{P}[G - \mathbb{E}[G] \geq t/2] \leq \exp\left(-\frac{nt^2}{2M^2}\right). \tag{81}$$

Applying Hoeffding (or the generalisation; bounded differences with the average function) gives that

$$\mathbb{P}\left[\hat{R}(h^*) - R(h^*) \geq t/2\right] \leq \exp\left(-\frac{nt^2}{2M^2}\right). \tag{82}$$

Now we can pick the t values such that

$$\delta = 2\exp\left(-\frac{nt^2}{2M^2}\right) \tag{83}$$

and then the result falls out.

□

Recall the inequality (79), that involves the Rademacher complexity $\mathcal{R}_n(\mathcal{F})$.

The Rademacher complexity is computed by assigning uniformly random ± 1 variables to all the empirical losses and then taking this sup, and hence loosely is a measure of how 'wild' the various ± 1 strings we can get from $(f(z_1), ..., f(z_n))$ are (since if these are all similar, this expected value will be lower).

Theorem 2.25

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(Z_i)\right] \tag{84}$$

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\left[\hat{R}(\mathcal{F}(Z_{1:n}))\right] = \mathbb{E}\left[\underbrace{\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(Z_i) | Z_{1:n}\right]}_{\hat{R}(\mathcal{F}(Z_{1:n}))}\right] = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(Z_i)\right] \tag{85}$$

by using the degenerate case of the tower property, and (73). □

Recall that if $W_{1:d}$ are mean zero σ-sub-Gaussian random variables then $\mathbb{E}[\max W] \leq$

$\sigma\sqrt{\log d}$. We can apply that to the inner expectation since the conditioning essentially fixes the $f(Z_i)$, and then we simply have a linear combination of Rademachers, which is sub-Gaussian by (2.12). Having done this conditioning, we are essentially concerned with bounding the size of the set of ± 1 strings mentioned above.

§2.6 VC Dimension

Consider the classification setting with $0-1$ loss. We aim to bound $\mathcal{R}_n(\mathcal{F})$, where the function class is the set of loss maps again (69).

Note that if $x_{1:n} \in \mathcal{X}^n$ and class labels $y_{1:n} \in \{-1,+1\}^n$ and $z_i = (x_i, y_i)$ then $|\mathcal{F}(z_{1:n})| = |\mathcal{H}(z_{1:n})|$, because there's a bijection

$$(\ell(h(x_i), y_i))_{i=1}^n \leftrightarrow (h(x_i))_{i=1}^n \tag{86}$$

since the y_is are fixed and so we can just cross check these terms. This is essentially a bijection of multisets however, since two different hypotheses could be identical on the training data.

Theorem 2.26

$$\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) \leq \sqrt{\frac{2\log|\mathcal{F}(z_{1:n})|}{n}}. \tag{87}$$

Let $|\mathcal{F}(z_{1:n})| = d$ and let $\mathcal{F}' = \{f_1, ..., f_d\} \subseteq \mathcal{F}$ be the set of *distinct* behaviours.

Given Rademacher random variables $\varepsilon_{1:n}$ set $W_j = \frac{1}{n}\sum_{i=1}^n \varepsilon_i f_j(z_i)$. Then $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) = \mathbb{E}[\max_j W_j]$.

Each ε_i is sub-Gaussian with parameter 1, so we can apply the linear combination of sub-Gaussians result; W_j is sub-Gaussian with parameter $\sqrt{\frac{1}{n^2}\sum_{i=1}^n f_j^2(z_i)} \leq \frac{1}{\sqrt{n}}$ (the bound because we're dealing with $0-1$ loss). Also $\mathbb{E}[W_j] = 0$, and so by (2.14) we're done. □

Essentially now we've reduced the complicated world of probability to the more beautiful world of combinatorics, since this bound involves the size of a set.

How can we bound this count of the number of behaviours?

Definition 2.27. Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \{a, b\}$ and $|\mathcal{F}| \geq 2$. We say that f **shatters** $x_{1:n}$ if $|\mathcal{F}(x_{1:n})| = 2^n$ i.e we get all the behaviours. Also define the **shattering coefficient** $s(\mathcal{F}, n) = \max_{x_{1:n}\in\mathcal{X}^n} |\mathcal{F}(x_{1:n})|$. Note that this must be bounded above by 2^n.

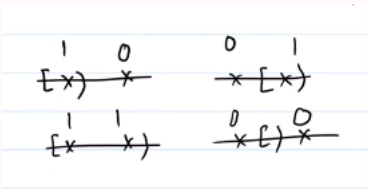
We need $s(\mathcal{F}, n)$ to be bounded by something better than 2^n since plugged into (2.26) this would give an n independent bound.

Definition 2.28. Define the **VC dimension** $VC(\mathcal{F})$ to be the largest n such that there exists $x_{1:n}$ shattered by $\mathcal{F}$, or if no such n exists define $VC(\mathcal{F}) = +\infty$.

Example 2.29 (Shattering coefficient bound and VC dimension where $|\mathcal{H}| = \infty$)
Suppose $\mathcal{F} = \{\mathbb{1}[[a,b)] \,|\, a \leq b \in \mathbb{R}\}$, the set of $[a,b)$ windows, an infinite function class.

To compute the shattering coefficient of this class, consider $x_1 < x_2 < ... < x_n$ which divide the real line into $n+1$ intervals. Then our length n binary strings will turn out to be the same iff a and a' are in the same interval, and b and b' are in the same interval. So $s(\mathcal{F}, n) \leq \binom{n+1}{2} + 1$, where the $+1$ comes from choosing some really dumb interval that gives us an all zero string.

We can shatter all two points:



But can't shatter distinct $x_1 < x_2 < x_3$ since the binary string $\{1,0,1\}$ can't be generated by our simple interval indicator class. So $VC(\mathcal{F}) = 2$.

The lesson here is that despite being an infinite hypothesis class, $\mathcal{F}$ has a asymptotically sharp shattering coefficient bound.

Actually, we have the surprising result

Theorem 2.30 (Sauer-Shelan)
Let $\mathcal{F}$ have finite VC dimension d. Then

$$s(\mathcal{F}, n) \leq (n+1)^d. \tag{88}$$

Proof. Non-examinable. Combinatorial? □

Corollary 2.31

$$\mathcal{R}_n(\mathcal{F}) \leq \sqrt{\frac{2VC(\mathcal{F})\log(n+1)}{n}} \tag{89}$$

(recall 2.26 and shattering definition).

Example 2.32

Let $\mathcal{X} = \mathbb{R}^p$ and consider the class $\mathcal{F} = \{\mathbb{1}[A] \mid A \in \mathcal{A}\}$ where

$$\mathcal{A} = \left\{ \prod_{j=1}^{p} (-\infty, a_j] \mid a_1, ..., a_p \in \mathbb{R} \right\}. \tag{90}$$

We claim that $VC(\mathcal{F}) = p$.

After unpacking definitions, it is clear that we can shatter n points.

It's slightly less easy to see that we can't shatter $p+1$ points. After checking the small cases for n, it's clear that we have some point which is not 'extreme' in some direction (or at very least, is not *uniquely* extreme in some direction). Then we can cook up a binary string with 0 in the place of this entry, and 1 everywhere else, and we can't get this behaviour.

Theorem 2.33

Let $\mathcal{F}$ be a vector space of functions. Then we can consider the class of classifiers $\mathcal{H} = \{\text{sgn} \circ f \mid f \in \mathcal{F}\}$. Then

$$VC(\mathcal{H}) \leq \dim \mathcal{F}. \tag{91}$$

Proof. We can first note that this generalises the previous result (I think I could cook up things to make this work with some zero product thing).

Let $d = \dim \mathcal{F} + 1$ and take $x_{1:d} \in \mathcal{X}^d$. We need show that $x_{1:d}$ cannot be shattered by $\mathcal{H}$.

Consider the linear map $L : \mathcal{F} \to \mathbb{R}^d$ defined by $f \mapsto (f(x_1), ..., f(x_d))$. Then the image dimension is at most $d-1$ by rank-nullity. From here, take $\gamma \neq 0$ orthogonal to this image space. Then break γ down in to its positive and non-negative components; let $\gamma_i > 0$ for all $i \in I_+$ and $\gamma_i < 0$ for all $i \in I_-$. Then

$$\sum_{i \in I_+} \gamma_i f(x_i) + \sum_{i \in I_-} \gamma_i f(x_i) = 0 \tag{92}$$

holds for all $f \in \mathcal{F}$. Then the behaviour where $f(x_i) = \pm 1$ on $I_\pm$ cannot be observed since if so, the LHS of (92) would be positive. □

Example 2.34

Consider $\mathcal{X} = [0,1)^2$ and $\mathcal{F}$ to be the set of polynomials of degree at most d, and define $\mathcal{H}$ as the set of signs of these polynomials as seen in (2.33).

Then by stars and bars, $\dim \mathcal{F} = \binom{d+2}{2}$. So if $d = 5$ then $VC(\mathcal{H}) \leq 21$ and previous results (namely, (79) and the VC bound (2.6)) imply

$$R(\hat{h}) - R(h^*) \leq 2\sqrt{\frac{2 \times 42 \log(n+1)}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}}. \tag{93}$$

Comparing to the histogram classifier (2.16), with the finite hypothesis class bound, we had

$$R(\hat{h}) - R(h^*) \leq \sqrt{\frac{2m^2 \log 2 + 2\log(1/\delta)}{n}} \tag{94}$$

where *importantly*, the two h^*s are different: we're dealing with two different hypothesis classes.

with probability at least $1 - \delta$.

and with that example over, we finish the most stats-heavy part of the course. Phew!

§3 Computation for ERM

Computation of the ERM with 0-1 loss can be hard (even NP-hard in some scenarios?!).

We will try and turn this hard problem into a convex optimisation problem.

§3.1 Basic properties of convex sets

Definition 3.1. $C \subset \mathbb{R}^d$ is **convex** if all lines segments lie entirely in C.

Example 3.2 (Basic properties of convex sets)

Intersections of convex sets are themselves convex.

Definition 3.3. The **convex hull** of $S \subset \mathbb{R}^d$, written $\text{conv}S$ is the intersection of all convex sets containing S. By the above, it is unsurprisingly convex.

Definition 3.4. $v \in \mathbb{R}^d$ is a **convex combination** of $v_1, ..., v_m \in \mathbb{R}^d$ if the α are non-negative and sum to 1 and

$$v = \alpha_1 v_1 + ... + \alpha_m v_m. \tag{95}$$

Lemma 3.5 (Convex combinations ... are what you think they are)
For $S \subset \mathbb{R}^d$, $v \in \text{conv}S$ iff v is a convex combination of some set of points in S.

Proof. Let D be the set of all convex combinations of points from S.
Then $\text{conv}S \subset D$ is on the example sheet.
To show $D \subset \text{conv}S$ induct on the number m of non-zero α terms appearing in (95). The case $m = 1$ is clear. Then for $m + 1$ non-zero terms WLOG making the indices nice,

$$v = \alpha_1 v_1 + ... + \alpha_{m+1} v_{m+1} = t\left(\frac{\alpha_1}{t} v_1 + ... + \frac{\alpha_m}{t} v_m\right) + (1-t) v_{m+1}. \tag{96}$$

and terms here must indeed lie in $\text{conv}S$ by its convexity. □

References

[1] Rajen D. Shah (2021), *Mathematics of Machine Learning*, http://www.statslab.cam.ac.uk/~rds37/teaching/machine_learning/notes_current.pdf.

[2] Philippe Rigollet, *18.657: Mathematics of Machine Learning*, https://ocw.mit.edu/courses/mathematics/18-657-mathematics-of-machine-learning-fall-2015/lecture-notes/MIT18_657F15_LecNote.pdf