# Mathematics of Machine Learning

Rajen D. Shah r.shah@statslab.cam.ac.uk

## 1 Introduction

Consider a pair of random elements $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with joint distribution $P_0$, where $X$ is to be thought of as an input or vector of predictors, and $Y$ as an output or response. For instance $X$ may represent a collection of disease risk factors (e.g. BMI, age, genetic indicators etc.) for a subject randomly selected from a population and $Y$ may represent their disease status; or $X$ could represent the number or bedrooms and other facilities in a randomly selected house, and $Y$ could be its price. In the former case we may take $\mathcal{Y} = \{-1, 1\}$, and this setting, known as the *classification* setting, will be of primary interest to us in this course. The latter case where $Y \in \mathbb{R}$ is an instance of a *regression* setting. We will take $\mathcal{X} = \mathbb{R}^p$ unless otherwise specified.

It is of interest to predict the random $Y$ from $X$; we may attempt to do this via a (measurable) function $h : \mathcal{X} \to \mathcal{Y}$, known as a *hypothesis*. To measure the quality of such a prediction we will introduce a *loss* function

$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}.$$

In the classification setting we typically take $\ell$ to be the *misclassification error*

$$\ell(h(x), y) = \begin{cases} 0 & \text{if } h(x) = y, \\ 1 & \text{otherwise.} \end{cases}$$

Penalise by 1 if wrong category selected. Treating all incorrect categories the same.

In this context $h$ is also referred to as a *classifier*. In regression settings the *squared error* $\ell(h(x), y) = (h(x) - y)^2$ is common. We will aim to pick a hypothesis $h$ such that the *risk*

Why is squared loss common? Why not $(h(x)-y)^\alpha$ w/ $\alpha$ a hyperparameter? Differentiability is an advantage over absolute cost.

$$R(h) := \int_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \ell(h(x), y) dP_0(x, y)$$

is small. For a deterministic $h$, $R(h) = \mathbb{E}\ell(h(X), Y)$ In what follows we will take $\ell$ and $R$ to be the misclassification loss and risk respectively, unless otherwise stated.

A classifier $h_0$ that minimises the misclassification risk is known as a *Bayes classifier*, and its risk is called the *Bayes risk*. Define the *regression function* $\eta$ by

$$\eta(x) := \mathbb{P}(Y = 1 | X = x).$$

**Proposition 1.** *A Bayes classifier $h_0$ is given by*[1]

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

[1] When $\eta(x) = 1/2$, we can equally well take $h_0 = \pm 1$ and achieve the same misclassification error.

In most settings of interest, the joint distribution $P_0$ of $(X, Y)$, which determines the optimal $h$, will be unknown. Instead we will suppose we have i.i.d. copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of the pair $(X, Y)$, known as *training data*. Our task is to use this data to construct a classifier $\hat{h}$ such that $R(\hat{h})$ is small. **Important point:** $R(\hat{h})$ is a random variable depending on the random training data:

$$R(\hat{h}) = \mathbb{E}(\ell(h(X), Y)|X_1, Y_1, \ldots, X_n, Y_n).$$

A statistical approach to classification may attempt to model $P_0$ up to some unknown parameters, estimate these parameters, and thereby obtain an estimate of the regression function (or the conditional expectation in the case of least squares—see below). We will take a different approach and assume that we are given a class $\mathcal{H}$ of hypotheses from which to pick our $\hat{h}$. Possible choices of $\mathcal{H}$ include for instance

- $\mathcal{H} = \{h : h(x) = \text{sgn}(\mu + x^T\beta) \text{ where } \mu \in \mathbb{R},\ \beta \in \mathbb{R}^p\}$;
- $\mathcal{H} = \left\{h : h(x) = \text{sgn}\left(\mu + \sum_{j=1}^d \varphi_j(x)\beta_j\right) \text{ where } \mu \in \mathbb{R},\ \beta \in \mathbb{R}^d\right\}$ for a given *dictionary* of functions $\varphi_1, \ldots, \varphi_d : \mathcal{X} \to \mathbb{R}$.

**Technical note:** In this course we will take $\text{sgn}(0) = -1$. (It does not matter much whether we take $\text{sgn}(0) = \pm 1$, but we need to specify a choice in order that the $h$ defined above are classifiers.)

## 1.1 Brief review of conditional expectation

For many of the mathematical arguments in this course we will need to manipulate conditional expectations.

Recall that if $Z \in \mathbb{R}$ and $W \in \mathbb{R}^d$ are random elements with joint probability density function (pdf) $f_{Z,W}$ then the conditional pdf $f_{Z|W}$ of $Z$ given $W$ satisfies

$$f_{Z|W}(z|w) = \begin{cases} f_{Z,W}(z,w)/f_W(w) & \text{if } f_W(w) = 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $f_W$ is the marginal pdf of $W$. When one or more of $Z$ and $W$ are discrete we typically work with probability mass functions.

Suppose $\mathbb{E}|Z| < \infty$. Then the conditional expectation function $\mathbb{E}(Z|W = w)$ is given by

$$g(w) := \mathbb{E}(Z|W = w) = \int z f_{Z|W}(z|w) dz. \tag{1.1}$$

We write $\mathbb{E}(Z|W)$ for the random variable $g(W)$ (note this is a function of $W$, not $Z$).

This is not a fully general definition of conditional expectation (for that see the Stochastic Financial Models course) and we will not use it. We will however make frequent use of the following properties of conditional expectation.

(i) **Role of independence:** If $Z$ and $W$ are independent, then $\mathbb{E}(Z|W) = \mathbb{E}Z$. (Recall: $Z$ and $W$ being independent means $\mathbb{P}(Z \in A, W \in B) = \mathbb{P}(Z \in A)\mathbb{P}(W \in B)$ for all measurable $A \subseteq \mathbb{R}$, $B \subseteq \mathbb{R}^d$)

(ii) **Tower property:** Let $f : \mathbb{R}^d \to \mathbb{R}^m$ be a (measurable) function. Then

$$\mathbb{E}\{\mathbb{E}(Z|W)|f(W)\} = \mathbb{E}\{Z|f(W)\}.$$

In particular, $\mathbb{E}\{\mathbb{E}(Z|W)|W_1, \ldots, W_m\} = \mathbb{E}(Z|W_1, \ldots, W_m)$ for $m \leq d$. Taking $f \equiv c \in \mathbb{R}$ and using (i) gives us that $\mathbb{E}\{\mathbb{E}(Z|W)\} = \mathbb{E}(Z)$ (as $f(W)$ is a constant it is independent of any random variable).

(iii) **Taking out what is known:** If $\mathbb{E}Z^2 < \infty$ and $f : \mathbb{R}^d \to \mathbb{R}$ is such that $\mathbb{E}[\{f(W)\}^2] < \infty$ then $\mathbb{E}\{f(W)Z|W\} = f(W)\mathbb{E}(Z|W)$.

Probabilistic results can be 'applied conditionally', for example:

**Conditional Jensen.** Recall that $f : \mathbb{R} \to \mathbb{R}$ is a convex function if

$$tf(x) + (1-t)f(y) \geq f\big(tx + (1-t)y\big) \quad \text{for all } x, y \in \mathbb{R} \text{ and } t \in (0,1).$$

The conditional version of *Jensen's inequality* states that if $f : \mathbb{R} \to \mathbb{R}$ is convex and random variable $Z$ has $\mathbb{E}|f(Z)| < \infty$, then

$$\mathbb{E}\big(f(Z)|W\big) \geq f\big(\mathbb{E}(Z|W)\big).$$

## 1.2 Bayes risk

*Proof of Proposition 1.* We have

$$\begin{aligned} R(h) &= \frac{1}{4}\mathbb{E}\{(Y - h(X))^2\} \\ &= \frac{1}{4}\mathbb{E}\{(Y - \mathbb{E}(Y|X) + \mathbb{E}(Y|X) - h(X))^2\} \\ &= \frac{1}{4}\underbrace{\mathbb{E}\{(Y - \mathbb{E}(Y|X))^2\}}_{\text{Independent of } h} + \frac{1}{4}\mathbb{E}\{(\mathbb{E}(Y|X) - h(X))^2\} + \frac{1}{2}\underbrace{\mathbb{E}\{(Y - \mathbb{E}(Y|X))(\mathbb{E}(Y|X) - h(X))\}}_{=0}. \end{aligned}$$

But

$$\begin{aligned} &\mathbb{E}\{(Y - \mathbb{E}(Y|X))(\mathbb{E}(Y|X) - h(X))\} \\ &= \mathbb{E}\,\mathbb{E}\{(Y - \mathbb{E}(Y|X))(\mathbb{E}(Y|X) - h(X))|X\} \quad \text{(tower property)} \\ &= \mathbb{E}\big[\underbrace{(\mathbb{E}(Y|X) - h(X))}_{\text{since only depends on } X}\underbrace{\mathbb{E}\{(Y - \mathbb{E}(Y|X))|X\}}_{=0}\big] \quad \text{(taking out what is known)} \\ &= 0. \end{aligned}$$

Thus minimising $R(h)$ is equivalent to minimising $\mathbb{E}\{(\underbrace{\mathbb{E}(Y|X)}_{\in\{-1,1\}} - h(X))^2\}$. We therefore get $h_0(X) = \operatorname{sgn}(\mathbb{E}(Y|X))$. □

The proof also shows that the risk under least squares loss is minimised by taking $h(x) = \mathbb{E}(Y|X = x)$ (provided $\mathbb{E}Y^2 < \infty$).

## 1.3 Empirical risk minimisation

*Empirical risk minimisation* replaces the expectation over the unknown $P_0$ in the definition of the risk with the empirical distribution, and seeks to minimise the resulting objective over $h \in \mathcal{H}$:

$$\hat{R}(h) := \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i), \qquad \hat{h} \in \arg\min_{h\in\mathcal{H}} \hat{R}(h).$$

$\hat{R}(h)$ is the *empirical risk* or *training error* of $h$.

**Example.** Consider the regression setting with $\mathcal{Y} = \mathbb{R}$, squared error loss and $\mathcal{H} = \{x \mapsto \mu + x^T\beta$ for $\mu \in \mathbb{R},\ \beta \in \mathbb{R}^p\}$. Then empirical risk minimisation is equivalent to ordinary least squares, i.e. we have

This is as one should expect.

We are relying that $\hat{R}$ is similar to $R$

$$\hat{h}(x) = \hat{\mu} + \hat{\beta}^T x \quad \text{where } (\hat{\mu}, \hat{\beta}) \in \arg\min_{(\mu,\beta)\in\mathbb{R}\times\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu - X_i^T\beta)^2.$$

A good choice for the class $\mathcal{H}$ will result in a low *generalisation error* $R(\hat{h})$. This is a measure of how well we can expect the empirical risk minimiser (ERM) $\hat{h}$ to predict a new data point $(X_{\text{new}}, Y_{\text{new}}) \sim P_0$ given only knowledge of $X_{\text{new}}$. Define $h^* \in \arg\min_{h\in\mathcal{H}} R(h)$[2] and consider the decomposition

Can get more data to fix this.

i.e. $h^*$ is error-minimising within our class of functions.

$$R(\hat{h}) - R(h_0) = \underbrace{R(\hat{h}) - R(h^*)}_{\text{stochastic error / excess risk}} + \underbrace{R(h^*) - R(h_0)}_{\text{approximation error}}.$$

Nothing beats $h_0$ it's Bayes

To improve this, we need to consider a different class of functions than $\mathcal{H}$

Clearly a richer class $\mathcal{H}$ will decrease the approximation error. However, it will tend to increase the stochastic error as empirical risk minimisation will fit to the realised $Y_1, \ldots, Y_n$ too closely and result in poor generalisation. There is thus a tradeoff between the stochastic error due to the complexity of the class $\mathcal{H}$, and its approximation error.

i.e. richer class of functions leads to overfitting.

We will primarily study the stochastic term or *excess risk*[3], and aim to provide bounds on this in terms of the complexity of $\mathcal{H}$. Recall that whilst for a fixed $h \in \mathcal{H}$, $R(h)$ is deterministic, $R(\hat{h})$ is a random variable. The bounds we obtain will be of the form "with probability at least $1 - \delta$,

$$R(\hat{h}) - R(h^*) \leq \epsilon."$$

[2] If there is no $h^*$ that achieves the associated infimum, we can consider an approximate minimiser with $R(h^*) < \inf_{h\in\mathcal{H}} R(h) + \epsilon$ for arbitrary $\epsilon > 0$ and all our analysis will carry through. Similar reasoning is applicable to $\hat{h}$.

[3] Sometimes "excess risk" is used for $R(\hat{h}) - R(h_0)$. However since we are considering $\mathcal{H}$ to be fixed in advance for much of the course, we will use excess risk to refer to the risk relative to that of $h^*$.

# 2 Statistical learning theory

Consider the following decomposition of the excess risk:

$$R(\hat{h}) - R(h^*) = \underbrace{R(\hat{h}) - \hat{R}(\hat{h})}_{\text{concentration}} + \underbrace{\hat{R}(\hat{h}) - \hat{R}(h^*)}_{\leq 0} + \underbrace{\hat{R}(h^*) - R(h^*)}_{\text{concentration}}$$
$$\leq R(\hat{h}) - \hat{R}(\hat{h}) + \hat{R}(h^*) - R(h^*).$$

since by def$^n$ $\hat{h}$ is best on the observed data

i.e. how close are variables to their expectation compared to the size of the sample size $n$

Note that $R(h^*)$ is an average of $n$ i.i.d. random variables, each with expectation $R(h^*)$. To bound $\hat{R}(h^*) - R(h^*)$ we will consider the general problem of how random variables concentrate around their expectation, a problem which is the topic of an important area of probability theory concerning *concentration inequalities*. The term $R(\hat{h}) - \hat{R}(\hat{h})$ is more complicated as $\hat{R}(\hat{h})$ is not a sum of i.i.d. random variables, but we will see extensions of techniques for the simpler case may be used to tackle this.

## 2.1 Sub-Gaussianity and Hoeffding's inequality

We will apply inequalities to RVs in general, but we want to use them for $\hat{R}(h^*) - R(h^*)$

We begin our discussion of concentration inequalities with the simplest tail bound, *Markov's inequality*. Let $W$ be a non-negative random variable. Taking expectations of both sides of $t\mathbb{1}_{\{W \geq t\}} \leq W$ for $t > 0$, we obtain after dividing through by $t$

instant proof!

$$\mathbb{P}(W \geq t) \leq \frac{\mathbb{E}(W)}{t}.$$

This immediately implies that given a strictly increasing function $\varphi : \mathbb{R} \to [0, \infty)$ and any random variable $W$,

$\varphi$ increasing

$$\mathbb{P}(W \geq t) = \mathbb{P}\big(\varphi(W) \geq \varphi(t)\big) \leq \frac{\mathbb{E}(\varphi(W))}{\varphi(t)}.$$

Applying this with $\varphi(t) = e^{\alpha t}$ $(\alpha > 0)$ yields the so-called *Chernoff bound*:

$$\mathbb{P}(W \geq t) \leq \inf_{\alpha > 0} e^{-\alpha t} \mathbb{E} e^{\alpha W}.$$

**Example.** Consider the case when $W \sim N(0, \sigma^2)$. Recall that

$$\mathbb{E} e^{\alpha W} = e^{\alpha^2 \sigma^2 / 2}. \tag{2.1}$$

Thus

$$\mathbb{P}(W \geq t) \leq \inf_{\alpha > 0} e^{\alpha^2 \sigma^2 / 2 - \alpha t} = e^{-t^2/(2\sigma^2)}. \tag{2.2}$$

Note that to arrive at this bound, all we required was (an upper bound on) the moment generating function (mgf) of $W$ (2.1). This motivates the following definition.

**Definition 1.** We say a random variable $W$ is *sub-Gaussian* with parameter $\sigma > 0$ if

$$\mathbb{E}e^{\alpha(W-\mathbb{E}W)} \leq e^{\alpha^2\sigma^2/2} \quad \text{for all } \alpha \in \mathbb{R}.$$

From (2.2) we immediately have the following result.

**Proposition 2.** *If $W$ is* sub-Gaussian *with parameter $\sigma > 0$, then*

$$\mathbb{P}(W \geq t) \leq e^{-t^2/(2\sigma^2)} \quad \textit{for all } t \geq 0.$$

Note that if $W$ is sub-Gaussian with parameter $\sigma > 0$, then

- it is also sub-Gaussian with parameter $\sigma'$ for any $\sigma' \geq \sigma$;
- $-W$ is also sub-Gaussian with parameter $\sigma > 0$. This means we have from (2.2) that

$$\mathbb{P}(|W - \mathbb{E}W| \geq t) = \mathbb{P}(W - \mathbb{E}W \geq t) + \mathbb{P}(-(W - \mathbb{E}W) \geq t) \leq 2e^{-t^2/(2\sigma^2)}.$$

This is useful since we want to know how much RVs deviate from their mean values.

"Any linear combination of Sub-Gaussians is Sub-Gaussian"

**Proposition 3.** *Suppose $W_1, \ldots, W_n$ are independent and each $W_i$ is sub-Gaussian with parameter $\sigma_i$ and has mean $\mu_i$. Then for $\gamma \in \mathbb{R}^n$, $\gamma^T W$ is sub-Gaussian with parameter* $\left(\sum_i \gamma_i^2\sigma_i^2\right)^{1/2}$.

Sub-Gaussian is helpful since we can bound the RVs and their sums/products by the variance of Gaussians which are well known distributions.

*Proof.*

$$\begin{aligned}
\mathbb{E}\exp\left(\alpha\sum_{i=1}^n \gamma_i W_i\right) &= \prod_{i=1}^n \mathbb{E}\exp(\alpha\gamma_i W_i) \\
&\leq \prod_{i=1}^n \exp(\alpha^2\gamma_i^2\sigma_i^2/2) \\
&= \exp\left(\alpha^2\sum_{i=1}^n \gamma_i^2\sigma_i^2/2\right).
\end{aligned}$$

□

Combining with Proposition 2 we obtain

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (W_i - \mu_i) \geq t\right) \leq \exp\left(-\frac{n^2t^2}{2\sum_{i=1}^n \sigma_i^2}\right) \quad \text{for } t \geq 0. \tag{2.3}$$

As well as implying concentration around the mean, the bound on the mgf satisfied by sub-Gaussian random variables also offers a bound on the expected maximum of $d$ sub-Gaussians. We do not need the following result at this stage, but will make use of it later.

**Proposition 4.** *Suppose $W_1, \ldots, W_d$ are all mean-zero and sub-Gaussian with parameter $\sigma > 0$ (but are not necessarily independent). Then*

$$\mathbb{E}\max_j W_j \leq \sigma\sqrt{2\log(d)}.$$

*Proof.* Let $\alpha > 0$. By convexity of $x \mapsto \exp(\alpha x)$ and Jensen's inequality we have

$$\exp(\alpha \mathbb{E} \max_j W_j) \leq \mathbb{E} \exp(\alpha \max_j W_j) = \mathbb{E} \max_j \exp(\alpha W_j).$$

Now

$$\mathbb{E} \max_{j=1,\dots,d} \exp(\alpha W_j) \leq \sum_{j=1}^{d} \mathbb{E} \exp(\alpha W_j) \leq d e^{\alpha^2 \sigma^2/2}.$$

Since maximum ≤ sum of them.

Since they are $d$ sub-Gaussian RVs

Thus

$$\mathbb{E} \max_j W_j \leq \frac{\log(d)}{\alpha} + \frac{\alpha \sigma^2}{2}.$$

← Applying Jensen's ineq.

Optimising over $\alpha > 0$ yields the result. □

Gaussian random variables are sub-Gaussian, but the sub-Gaussian class is much broader than this.

a.152

**Example.** A *Rademacher* random variable $\varepsilon$ takes values $\{-1, 1\}$ with equal probability. It is sub-Gaussian with parameter $\sigma = 1$:

This is just a trivial calculation.

$$\begin{aligned}
\mathbb{E} e^{\alpha \varepsilon} = \frac{1}{2}(e^{-\alpha} + e^{\alpha}) &= \frac{1}{2}\left( \sum_{k=0}^{\infty} \frac{(-\alpha)^k}{k!} + \sum_{k=0}^{\infty} \frac{\alpha^k}{k!} \right) \\
&= \sum_{k=0}^{\infty} \frac{\alpha^{2k}}{(2k)!} \\
&\leq \sum_{k=0}^{\infty} \frac{\alpha^{2k}}{2^k k!} = e^{\alpha^2/2} \quad (\text{using } (2k)! \geq 2^k k!). \qquad (2.4)
\end{aligned}$$

Recall that we are interested in the concentration properties of $\mathbb{1}_{\{h(X_i) \neq Y_i\}} - \mathbb{P}(h(X_i) \neq Y_i)$, which in particular is bounded.

**Lemma 5** (Hoeffding's lemma)**.** *If $W$ is mean-zero and takes values in $[a, b]$, then $W$ is sub-Gaussian with parameter $\sigma = (b-a)/2$.*

*Proof.* We will prove a weaker result here with $\sigma = b - a$; see the Example sheet for a proof with $\sigma = (b-a)/2$. Let $W'$ be an independent copy of $W$. We have

We prove for $\sigma = (b-a)$, $\sigma = (b-a)/2$ is Exercise for the example sheet.

$$\begin{aligned}
\mathbb{E} e^{\alpha W} &= \mathbb{E} e^{\alpha(W - \mathbb{E}W')} \qquad \text{This is a symmetrication argument} \\
&= \mathbb{E} e^{\mathbb{E}\{\alpha(W - W')|W\}} \quad \text{using } \mathbb{E}(W') = \mathbb{E}(W'|W) \text{ and } \mathbb{E}(W|W) = W \\
&\leq \mathbb{E} e^{\alpha(W - W')} \quad \text{(Jensen conditional on } W \text{ and tower prop.).}
\end{aligned}$$

Idea: If $W$ was s.t. $W \overset{d}{=} -W$ then $W \overset{d}{=} \varepsilon W$ so could run the argument for Rademacher RV with $\varepsilon W$ instead of $\varepsilon$, conditional on $W$

Now $W - W' \overset{d}{=} -(W - W') \overset{d}{=} \varepsilon(W - W')$ where $\varepsilon \sim$ Rademacher with $\varepsilon$ independent of $(W, W')$. Thus

$$\mathbb{E} e^{\alpha W} \leq \mathbb{E} e^{\alpha \varepsilon (W - W')} = \mathbb{E}\{\mathbb{E}(e^{\alpha \varepsilon (W - W')} | W, W')\}.$$

We now apply our previous result (2.4) conditionally on $(W - W')$ to obtain

$$\mathbb{E}e^{\alpha W} \leq \mathbb{E}e^{\alpha^2(W-W')^2/2} \leq \mathbb{E}e^{\alpha^2(b-a)^2/2}$$

as $|W - W'| \leq b - a$. □

The introduction of an independent copy $W'$ and a Rademacher random variable here is an example of a *symmetrisation argument*; we will make use of this technique again later in the course. As an application of the result above, suppose $W_1, \ldots, W_n$ are independent, mean-zero and $a_i \leq W_i \leq b_i$ almost surely for all $i$. Then

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} W_i \geq t\right) \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right), \quad (2.5)$$

Applying the result about summing sub-Gaussian RVs.

which is known as *Hoeffding's inequality*.

We are now in a position to bound $R(\hat{h}) - R(h^*)$ when $\mathcal{H}$ is finite.

## 2.2 Finite hypothesis classes

**Theorem 6.** *Suppose $\mathcal{H}$ is finite and $\ell$ takes values in $[0, M]$. Then with probability at least $1 - \delta$, the ERM $\hat{h}$ satisfies*

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

The assumption on $\ell$ includes as a special case misclassification loss. However the extra generality will prove helpful later in the course.

*Proof.* Recall that

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

$\hat{h}$ is empirical risk minimiser

risk minimiser within hypothesis class

Now for each $h$, $\hat{R}(h)$ is an average of mean-zero i.i.d. quantities of the form $\ell(h(X_i), Y_i) - \mu$ with $\mu = \mathbb{E}\ell(h(X_i), Y_i)$ which take values in $[-\mu, M - \mu]$. Now for $t > 0$,

$$\begin{aligned}\mathbb{P}(R(\hat{h}) - R(h^*) > t) &= \mathbb{P}(R(\hat{h}) - R(h^*) > t, \hat{h} \neq h^*) \\ &\leq \mathbb{P}(R(\hat{h}) - \hat{R}(\hat{h}) > t/2, \hat{h} \neq h^*) + \mathbb{P}(\hat{R}(h^*) - R(h^*) > t/2)\end{aligned}$$

We can immediately apply Hoeffding's inequality to obtain

$$\mathbb{P}(\hat{R}(h^*) - R(h^*) \geq t/2) \leq \exp(-nt^2/(2M^2)).$$

However the complicated dependence among the summands in $\hat{R}(\hat{h})$ prevents this line of attack for bounding the first term. To tackle this issue, we note that when $\hat{h} \neq h^*$,

$$R(\hat{h}) - \hat{R}(\hat{h}) \leq \max_{h \in \mathcal{H}_-} R(h) - \hat{R}(h),$$

where $\mathcal{H}_- := \mathcal{H} \setminus \{h^*\}$. We then have using a union bound,

$$\begin{aligned}\mathbb{P}(\max_{h\in\mathcal{H}_-} R(h) - \hat{R}(h) \geq t/2) &= \mathbb{P}(\cup_{h\in\mathcal{H}_-} R(h) - \hat{R}(h) \geq t/2) \\ &\leq \sum_{h\in\mathcal{H}_-} \mathbb{P}(R(h) - \hat{R}(h) \geq t/2) \\ &\leq |\mathcal{H}_-| \exp(-nt^2/(2M^2)).\end{aligned}$$

Note this is meaningful only when $|\mathcal{H}| < \infty$.

Thus

$$\mathbb{P}(R(\hat{h}) - R(h^*) > t) \leq |\mathcal{H}| \exp(-nt^2/(2M^2))$$

Writing $\delta := |\mathcal{H}| \exp(-nt^2/(2M^2))$ and then expressing $t$ in terms of $\delta$ gives the result. □

**Example.** Consider a simple classification setting with $X_i \in [0,1)^2$. Let us divide $[0,1)^2$ into $m^2$ disjoint squares $R_1, \ldots, R_{m^2} \subset [0,1)^2$ of the form $[r/m, (r+1)/m) \times [s/m, (s+1)/m)$ for $r, s = 0, \ldots, m-1$. Let

$$\bar{Y}_j = \operatorname{sgn}\left( \sum_{i: X_i \in R_j} Y_i \right)$$

and define

$$h^{\text{hist}}(x) = \sum_{j=1}^{m} \bar{Y}_j \mathbb{1}_{R_j}(x).$$

Then $h^{\text{hist}}$ is equivalent to the ERM over hypothesis class $\mathcal{H}$ consisting of the $2^{m^2}$ classifiers each corresponding to a way of assigning labels in $\{-1, 1\}$ to each of the regions $R_1, \ldots, R_{m^2}$. The result above tells us that the generalisation error (with mislassification loss) of $h^{\text{hist}}$ is at most

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

Increasing $m$ gives a finer histogram grid.

If $m$ is too large this will overfit & give a bad classifier

[In fact it can be shown that the approximation error $R(h^*) - R(h_0) \to 0$ if $m \to \infty$ for any given $P_0$. Combining with the above, we then see that choosing e.g. $m = n^{1/3}$ we can approach the Bayes risk for $n$ sufficiently large.]

Whilst a union bound and Hoeffding's inequality sufficed to give us a bound in the case where $\mathcal{H}$ is finite, to handle the more common setting where $\mathcal{H}$ is infinite, we will need more sophisticated techniques. Our approach will be to view the key quantity

We will usually have an infinite hypothesis class, eg decision trees, linear regression, SVMs, NNs, etc. basically everything.

$$G(X_1, Y_1, \ldots, X_n, Y_n) := \sup_{h\in\mathcal{H}} R(h) - \hat{R}(h)$$

as a function $G$ of the i.i.d. random elements $(X_1, Y_1), \ldots, (X_n, Y_n)$. We currently only have at our disposal concentration inequalities where $g$ takes the form of an average; however $G$ will in general clearly be much more complex. Intuitively though, the key property of the empirical average that results in concentration is that the individual contributions of each

Since there are lots of them.
of the random elements is not too large. Can we show that our $G$ would, despite having an intractable form, nevertheless share this property in common with the empirical average?

Given data $(x_1, y_1), \ldots (x_n, y_n)$ and $\epsilon > 0$, let $\tilde{h} \in \mathcal{H}$ be such that

$$G(x_1, y_1, \ldots, x_n, y_n) < R(\tilde{h}) - \hat{R}(\tilde{h}) + \epsilon.$$

That it isn't affected much by the fluctuation of a single data point.

Now consider perturbing (wlog) the first pair of arguments of $G$. We have

$$\begin{aligned}
&G(x_1, y_1, \ldots, x_n, y_n) - G(x_1', y_1', x_2, y_2, \ldots, x_n, y_n) \\
&< R(\tilde{h}) - \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \tilde{h}(x_i)) - \sup_{h \in \mathcal{H}} \left( R(h) - \frac{1}{n}\ell(y_1', h(x_1')) - \frac{1}{n}\sum_{i=2}^{n} \ell(y_i, h(x_i)) \right) + \epsilon \\
&\leq \frac{1}{n}\ell(y_1, \tilde{h}(x_1)) - \ell(y_1', \tilde{h}(x_1')) + \epsilon.
\end{aligned}$$

$\hat{R}(\tilde{h})$

As $\epsilon$ was arbitrary, if $\ell$ takes values in $[0, M]$ we have

We crudely bound $\ell$

$$G(x_1, y_1, \ldots, x_n, y_n) - G(x_1', y_1', x_2, y_2, \ldots, x_n, y_n) \leq M/n.$$

We thus seek a concentration inequality for multivariate functions where arbitrary perturbations of a single argument change the output by a bounded amount.

## 2.3 Bounded differences inequality

The result we are going to aim for is the so-called Bounded differences inequality. Let us adopt the notation that for a sequence of vectors $a_s, a_{s+1}, a_{s+2}, \ldots$ (where the starting index $s$ can be e.g. 0 or 1), $a_{j:k}$ for $j \leq k$ is the subsequence $a_j, \ldots, a_k$.

**Theorem 7** (Bounded differences inequality)**.** *Let $f : \mathbb{R}^{d \cdot n} \to R$ satisfy a* bounded differences property *such that*

$$f(w_1, \ldots, w_{i-1}, w_i, w_{i+1}, \ldots, w_n) - f(w_1, \ldots, w_{i-1}, w_i', w_{i+1}, \ldots, w_n) \leq L_i,$$

*for all $w_1, \ldots, w_n, w_i' \in \mathbb{R}^d$, and all $i = 1, \ldots, n$. Suppose random vectors $W_1, \ldots, W_n$ taking values in $\mathbb{R}^d$ are independent. Then*

$$\mathbb{P}\big(f(W_{1:n}) - \mathbb{E}f(W_{1:n}) \geq t\big) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} L_i^2}\right)$$

Note that taking $d = 1$, $f(W_{1:n}) = \sum_i \{W_i - \mathbb{E}(W_i)\}/n$ and assuming the $W_i$ take values in $[a, b]$, we recover Hoeffding's inequality.

To motivate the proof, consider the sequence of random variables given by $Z_0 = \mathbb{E}f(W_{1:n})$, $Z_n = f(W_{1:n})$ and

$$Z_i = \mathbb{E}(f(W_{1:n})|W_{1:i}) \quad \text{for } i = 1, \ldots, n-1. \tag{2.6}$$

Note that in the special case where $f(W_{1:n}) = \sum_i W_i$ and $\mathbb{E}W_i = 0$, we have $Z_k - Z_0 = \sum_{i=1}^k W_i$. Our approach centres on the telescoping decomposition

$$f(W_{1:n}) - \mathbb{E}f(W_{1:n}) = Z_n - Z_0 = \sum_{i=1}^n \underbrace{(Z_i - Z_{i-1})}_{D_i};$$

the differences $D_i$ play an analogous role to the individual independent random variables in the case of bounding sums. These considerations help to motivate the definition of a *martingale*[4], which encapsulates the key features of the $Z_i$ above that will allow our arguments to go through.

**Definition 2.** A sequence of random variables $Z_0, Z_1, \ldots, Z_n$ is a *martingale sequence* with respect to another sequence of random vectors $W_0, W_1, \ldots, W_n \in \mathbb{R}^d$ if

(i) $\mathbb{E}|Z_i| < \infty$ for $i = 0, \ldots, n$,

(ii) $Z_i$ is a function of $W_{0:i}$ for $i = 0, \ldots, n$,

(iii) $\mathbb{E}(Z_i | W_{0:(i-1)}) = Z_{i-1}$ $i = 1, \ldots, n$.

We call $D_i := Z_i - Z_{i-1}$ a *martingale difference sequence* with respect to $W_0, W_1, \ldots, W_n$. We also extend the definitions to infinite sequences.

**Example.** Consider a random walk with $Z_0 = 0$, $Z_i = Z_{i-1} \pm 1$ each with probability $1/2$. Then $(Z_i)_{i \geq 1}$ is just a sequence of partial sums of i.i.d. Rademacher random variables and is a martingale sequence with respect to $(Z_{i-1})_{i \geq 1}$. Now suppose $Z_i = Z_{i-1}$ if $Z_i = z$ for some $z \in \mathbb{Z}$ but with the dynamics governed by the random walk otherwise. We still have

$$\mathbb{E}(Z_i | Z_{0:(i-1)}) = z \mathbb{1}_{\{Z_{i-1} = z\}} + Z_{i-1} \mathbb{1}_{\{Z_{i-1} \neq z\}} = Z_{i-1}.$$

Note that in this case the corresponding martingale difference sequence is not i.i.d. and exhibits some dependence.

**Example.** The sequence $Z_{0:n}$ defined earlier via (2.6) is an example of a *Doob martingale*. Formally it is a martingale sequence with respect to $W_{0:n}$ where we may set $W_0$ to an arbitrary constant. That (ii) holds is clear. (i) certainly holds when $f$ is bounded, but holds more generally provided $\mathbb{E}|f(W_{1:n})| < \infty$:

$$\mathbb{E}|Z_i| = \mathbb{E}|\mathbb{E}\{f(W_{1:n}) | W_{1:i}\}| \leq \mathbb{E}|f(W_{1:n})| < \infty$$

using Jensen's inequality conditional on $W_{1:i}$ applied to the convex function $|\cdot|$, and then the tower property. That (iii) holds follows from the tower property of conditional expectation.

We are now in a position to prove a generalisation of (2.3) applicable to averages of martingale differences.

[4]For a more general and formal definition of a martingale, see the Stochastic Financial Models course.

**Lemma 8.** *Let $D_1, \dots, D_n$ be a martingale difference sequence with respect to $W_0, \dots, W_n \in \mathbb{R}^d$ such that*

$$\mathbb{E}(e^{\alpha D_i} | W_{0:(i-1)}) \leq e^{\alpha^2 \sigma_i^2 / 2} \quad \textit{almost surely.}$$

*Let $\gamma \in \mathbb{R}^n$ and write $D = (D_1, \dots, D_n)^T$. Then $\gamma^T D$ is sub-Gaussian with parameter $\left(\sum_i \gamma_i^2 \sigma_i^2\right)^{1/2}$.*

*Proof.* We have

$$\begin{aligned}
\mathbb{E} \exp\left(\alpha \sum_{i=1}^n \gamma_i D_i\right) &= \mathbb{E} \exp\left(\alpha \sum_{i=1}^{n-1} \gamma_i D_i\right) \mathbb{E}(e^{\alpha \gamma_n D_n} | W_{0:(n-1)}) \\
&\leq e^{\alpha^2 \gamma_n^2 \sigma_n^2 / 2} \mathbb{E} \exp\left(\alpha \sum_{i=1}^{n-1} \gamma_i D_i\right) \\
&\leq \exp\left(\frac{\alpha^2}{2} \sum_{i=1}^n \gamma_i^2 \sigma_i^2\right) \quad \text{(by induction).}
\end{aligned}$$

□

The Azuma-Hoeffding inequality applies the above result to the case of bounded random variables.

**Theorem 9** (Azuma–Hoeffding)**.** *Let $D_1, \dots, D_n$ be a martingale difference sequence with respect to $W_0, \dots, W_n \in \mathbb{R}^d$. Suppose the following holds for each $i = 1, \dots, n$: there exist random variables $A_i$ and $B_i$ that are functions of $W_{0:(i-1)}$ such that $A_i \leq D_i \leq B_i$ almost surely and $B_i - A_i \leq L_i$ almost surely for a constant $L_i$. Then*

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n D_i \geq t\right) \leq \exp\left(-\frac{2n^2 t^2}{\sum_{i=1}^n L_i^2}\right). \tag{2.7}$$

*Proof.* Conditional on $W_{0:(i-1)}$, $A_i$ and $B_i$ are constant. Thus we may apply Hoeffding's Lemma (Lemma 5) conditionally on $W_{0:(i-1)}$ to obtain

$$\mathbb{E}(e^{\alpha D_i} | W_{0:(i-1)}) \leq e^{\alpha^2 L_i^2 / 2} \quad \text{almost surely.}$$

The martingale difference sequence thus satisfies the hypotheses of Lemma 8. The average $\sum_i D_i / n$ is sub-Gaussian with parameter $\sigma = \left(\sum_i L_i^2\right)^{1/2} / n$. The result then follows from the sub-Gaussian tail bound (Proposition 2). □

We are finally ready to prove the Bounded differences inequality.

*Proof of Theorem 7.* Let $D_1, \dots, D_n$ be the Martingale difference sequence associated with the Doob martingale, so

$$D_i = \mathbb{E}(f(W_{1:n}) | W_{1:i}) - \mathbb{E}(f(W_{1:n}) | W_{1:i-1}).$$

Recall that $f(W_{1:n}) - \mathbb{E}f(W_{1:n}) = \sum_{i=1}^n D_i$. Using the Azuma–Hoeffding inequality, it suffices to prove that $A_i \leq D_i \leq B_i$ almost surely where $A_i$ and $B_i$ are functions of $W_{1:(i-1)}$ satisfying $B_i - A_i \leq L_i$ for all $i$, which we now do.

Let us define for each $i = 1, \ldots, n$, functions

$$\begin{aligned} F_i : \mathbb{R}^{d \cdot i} &\to \mathbb{R} \\ (\underbrace{w_1}_{\in \mathbb{R}^d}, \ldots, w_i) &\mapsto \mathbb{E}(f(W_{1:n}) | W_1 = w_1, \ldots, W_i = w_i). \end{aligned}$$

Then define the random variables

$$\begin{aligned} A_i &:= \inf_{w_i \in \mathbb{R}^d} F_i(W_{1:(i-1)}, w_i) - \mathbb{E}(f(W) | W_{1:(i-1)}) \\ B_i &:= \sup_{w_i \in \mathbb{R}^d} F_i(W_{1:(i-1)}, w_i) - \mathbb{E}(f(W) | W_{1:(i-1)}). \end{aligned}$$

Note that $A_i$ and $B_i$ are functions of $W_{1:(i-1)}$. Furthermore

$$\begin{aligned} D_i - A_i &= F_i(W_{1:i}) - \inf_{w_i \in \mathbb{R}^d} F_i(W_{1:(i-1)}, w_i) \geq 0 \\ D_i - B_i &= F_i(W_{1:i}) - \sup_{w_i \in \mathbb{R}^d} F_i(W_{1:(i-1)}, w_i) \leq 0, \end{aligned}$$

so $A_i \leq D_i \leq B_i$. Also

$$\begin{aligned} B_i - A_i &= \sup_{w_i \in \mathbb{R}^d} F_i(W_{1:(i-1)}, w_i) - \inf_{w_i \in \mathbb{R}^d} F_i(W_{1:(i-1)}, w_i) \\ &= \sup_{w_i, w_i' \in \mathbb{R}^d} \{F_i(W_{1:(i-1)}, w_i) - F_i(W_{1:(i-1)}, w_i')\} \\ &= \sup_{w_i, w_i' \in \mathbb{R}^d} \{\mathbb{E}(f(W_{1:(i-1)}, w_i, W_{(i+1):n}) | W_{1:(i-1)}, W_i = w_i) \\ &\qquad - \mathbb{E}(f(W_{1:(i-1)}, w_{i'}, W_{(i+1):n}) | W_{1:(i-1)}, W_i = w_{i'})\}. \end{aligned}$$

Now as the $W_{1:n}$ are independent, the distribution of $W_{(i+1):n}$ conditional on $W_{1:(i-1)}$ and $W_i = w_i$ is equal to its unconditional distribution, and the same holds when conditioning on $W_i = w_{i'}$. Thus

$$\begin{aligned} B_i - A_i &= \sup_{w_i, w_i' \in \mathbb{R}^d} [\mathbb{E}\{f(W_{1:(i-1)}, w_i, W_{(i+1):n}) - f(W_{1:(i-1)}, w_{i'}, W_{(i+1):n}) | W_{1:(i-1)}\}] \\ &\leq \mathbb{E}[\sup_{w_i, w_i' \in \mathbb{R}^d} \{f(W_{1:(i-1)}, w_i, W_{(i+1):n}) - f(W_{1:(i-1)}, w_{i'}, W_{(i+1):n})\} | W_{1:(i-1)}] \leq L_i. \end{aligned}$$

Note we have used the fact that for any collection of random variables $V_t$, $\sup_{t'} \mathbb{E}V_{t'} \leq \mathbb{E} \sup_t V_t$; this may easily be verified by removing the supremum over $t'$ and noting that the resulting inequality must hold for all $t'$. We have have verified all the conditions of the Azuma–Hoeffding inequality which may now be applied to give the result. □

## 2.4 Rademacher complexity

Recall our setup: $\mathcal{H}$ is a (now possibly infinite) hypothesis class, $\ell$ takes values in $[0, M]$ are we are aiming to bound the right-hand side of

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

where $G := \sup_{h \in \mathcal{H}} \{R(h) - \hat{R}(h)\}$. The Bounded differences inequality provides a means to bound $G - \mathbb{E}G$, but in order to make use of this, we must find a way of bounding $\mathbb{E}G$. Let us write $Z_i = (X_i, Y_i)$ for $i = 1, \ldots, n$ and

$$\mathcal{F} := \{(x, y) \mapsto -\ell(y, h(x)) : h \in \mathcal{H}\}. \tag{2.8}$$

Then we have

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

We will prove the following result which applies for a general function class $\mathcal{F}$ (not necessarily coming from (2.8)).

**Theorem 10.** *Let $\mathcal{F}$ be a class of real-valued functions $f : \mathcal{Z} \to \mathbb{R}$ and let $Z_1, \ldots, Z_n$ be i.i.d. random elements taking values in $\mathcal{Z}$. Then*

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

*where $\mathcal{R}_n(\mathcal{F})$ is the* Rademacher complexity *of $\mathcal{F}$ defined by*

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

*Here $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables independent of $Z_{1:n}$.*

Some intuition: Consider a classification problem with inputs $Z_1, \ldots, Z_n$ and *completely random* labels $\varepsilon_1, \ldots, \varepsilon_n$. The Rademacher complexity then captures how closely aligned the 'predictions' $f(Z_i)$ are to the random labels.

Before we prove Theorem 10, let us reflect on what it might achieve. Considering our main problem of bounding $\mathbb{E}G$, a key challenge is that it depends strongly and in a complicated way on the unknown $P_0$. To understand the potential advantages of Rademacher complexity, it is helpful to introduce the following.

**Definition 3.** Let $\mathcal{F}$ be a class of real-valued functions $f : \mathcal{Z} \to \mathbb{R}$ and let $z_1, \ldots, z_n \in \mathcal{Z}$. Writing

$$\mathcal{F}(z_{1:n}) = \{(f(z_1), \ldots, f(z_n)) : f \in \mathcal{F}\}$$

define the *empirical Rademacher complexity*

$$\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) := \mathbb{E}\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(z_i)\right), \tag{2.9}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables. Given i.i.d. random elements $Z_1, \ldots, Z_n$ taking values in $\mathcal{Z}$, we sometimes view the empirical Rademacher complexity as a random variable:

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

Note that $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n}))$ is well-defined in that the right-hand side of (2.9) only depends on $\mathcal{F}(z_{1:n})$, the behaviours of the functions in $\mathcal{F}$ on the fixed set of points $z_{1:n}$.

Key point: $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n}))$ does not depend on $P_0$. It is conceivable that we could upper bound $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n}))$ uniformly in $z_{1:n} \in \mathcal{Z}^n$. We then immediately get a bound on $\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\{\hat{\mathcal{R}}(\mathcal{F}(Z_{1:n}))\}$ that is independent of $P_0$.

We now turn to the proof of the result, which uses a symmetrisation technique.

*Proof of Theorem 10.* Let us introduce an independent copy $(Z'_1, \ldots, Z'_n)$ of $(Z_1, \ldots, Z_n)$. We have

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\{f(Z_i) - \mathbb{E}f(Z_i)\} &= \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{f(Z_i) - f(Z'_i)|Z_{1:n}\} \text{ (independence of } Z_{1:n} \text{ and } Z'_{1:n}) \\
&\leq \mathbb{E}\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\{f(Z_i) - f(Z'_i)\}\middle| Z_{1:n}\right).
\end{aligned}$$

Now let $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. Rademacher random variables, independent of $Z_{1:n}$ and $Z'_{1:n}$. Then

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\{f(Z_i) - f(Z'_i)\} &\stackrel{d}{=} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\{f(Z_i) - f(Z'_i)\} \\
&\leq \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(Z_i) + \sup_{g\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\{-\varepsilon_i g(Z_i)\} \\
&\stackrel{d}{=} \sup_{f\in\mathcal{F}} \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i f(Z_i).
\end{aligned}$$

Taking expectations and putting things together, we obtain the the result. □

**Theorem 11** (Generalisation bound based on Rademacher complexity)**.** *Let* $\mathcal{F} := \{(x, y) \mapsto \ell(y, h(x)) : h \in \mathcal{H}\}$ *and suppose* $\ell$ *takes values in* $[0, M]$*. With probability at least* $1 - \delta$*,*

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

*Proof.* Let $G := \sup_{h\in\mathcal{H}} R(h) - \hat{R}(h)$ and recall that

$$R(\hat{h}) - R(h^*) \leq G + \hat{R}(h^*) - R(h^*) = (G - \mathbb{E}G) + \mathbb{E}G + \hat{R}(h^*) - R(h^*).$$

Further recall that viewing $G$ as a function of $Z_1, \ldots, Z_n$ where $Z_i = (X_i, Y_i)$, it satisfies a bounded differences property with constants $L_i = M/n$. Thus the Bounded differences inequality gives us that

$$\mathbb{P}(G - \mathbb{E}G \geq t/2) \leq \exp(-t^2 n/(2M^2)).$$

Hoeffding's inequality (or Bounded differences with the average function) also gives $\mathbb{P}(\hat{R}(h^*) - R(h^*) \geq t/2) \leq \exp(-t^2 n/(2M^2))$. Also, noting that $\mathcal{R}_n(\mathcal{F}) = \mathcal{R}_n(-\mathcal{F})$, from Theorem 10, $\mathbb{E}G \leq 2\mathcal{R}_n(\mathcal{F})$. Thus taking $t = M\sqrt{2\log(1/\delta)/n}$ gives the result. □

## 2.5 VC dimension

All we need to do in order to bound the generalisation error is to obtain bounds on the Rademacher complexity. There are various ways of tackling this problem in general. Here, we will explore an approach suited to the classification setting with misclassification loss and $\mathcal{F} := \{(x, y) \mapsto \ell(y, h(x)) : h \in \mathcal{H}\}$. Our bounds will be in terms of the number of behaviours of the function class on $n$ points $|\mathcal{F}(z_{1:n})|$. Observe first that $|\mathcal{F}(z_{1:n})| = |\mathcal{H}(x_{1:n})|$ where $z_i = (x_i, y_i)$.

**Lemma 12.** *We have* $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) \leq \sqrt{2\log(|\mathcal{F}(z_{1:n})|)/n} = \sqrt{2\log(|\mathcal{H}(x_{1:n})|)/n}$.

*Proof.* Let $d = |\mathcal{F}(z_{1:n})|$ and let $\mathcal{F}' := \{f_1, \ldots, f_d\}$ be such that $\mathcal{F}(z_{1:n}) = \mathcal{F}'(z_{1:n})$ (so each $f_j$ has a unique behaviour on $z_{1:n}$). For $j = 1, \ldots, d$, let

$$W_j = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f_j(z_i),$$

where $\varepsilon_{1:n}$ are i.i.d. Rademacher random variables. Then $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) = \mathbb{E}\max_j W_j$. By Lemma 5 and Proposition 3, each $W_j$ is sub-Gaussian with parameter $1/\sqrt{n}$. Thus we may apply Proposition 4 on the expected maximum of sub-Gaussian random variables to give the result. □

As each $h(x_i) \in \{-1, 1\}$, we always have $|\mathcal{H}(x_{1:n})| \leq 2^n$. Considering the result above, an interesting case then is when $|\mathcal{H}(x_{1:n})|$ is growing slower than exponentially in $n$, e.g. growing polynomially in $n$.

# Mathematics of Machine Learning

Rajen D. Shah r.shah@statslab.cam.ac.uk

## 1 Introduction

Consider a pair of random elements $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with joint distribution $P_0$, where $X$ is to be thought of as an input or vector of predictors, and $Y$ as an output or response. For instance $X$ may represent a collection of disease risk factors (e.g. BMI, age, genetic indicators etc.) for a subject randomly selected from a population and $Y$ may represent their disease status; or $X$ could represent the number or bedrooms and other facilities in a randomly selected house, and $Y$ could be its price. In the former case we may take $\mathcal{Y} = \{-1, 1\}$, and this setting, known as the *classification* setting, will be of primary interest to us in this course. The latter case where $Y \in \mathbb{R}$ is an instance of a *regression* setting. We will take $\mathcal{X} = \mathbb{R}^p$ unless otherwise specified.

It is of interest to predict the random $Y$ from $X$; we may attempt to do this via a (measurable) function $h : \mathcal{X} \to \mathcal{Y}$, known as a *hypothesis*. To measure the quality of such a prediction we will introduce a *loss* function

$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}.$$

In the classification setting we typically take $\ell$ to be the *misclassification error*

$$\ell(h(x), y) = \begin{cases} 0 & \text{if } h(x) = y, \\ 1 & \text{otherwise.} \end{cases}$$

In this context $h$ is also referred to as a *classifier*. In regression settings the *squared error* $\ell(h(x), y) = (h(x) - y)^2$ is common. We will aim to pick a hypothesis $h$ such that the *risk*

$$R(h) := \int_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \ell(h(x), y) dP_0(x, y)$$

is small. For a deterministic $h$, $R(h) = \mathbb{E}\ell(h(X), Y)$. In what follows we will take $\ell$ and $R$ to be the misclassification loss and risk respectively, unless otherwise stated.

A classifier $h_0$ that minimises the misclassification risk is known as a *Bayes classifier*, and its risk is called the *Bayes risk*. Define the *regression function* $\eta$ by

$$\eta(x) := \mathbb{P}(Y = 1 | X = x).$$

**Proposition 1.** *A Bayes classifier $h_0$ is given by*[1]

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

[1]When $\eta(x) = 1/2$, we can equally well take $h_0 = \pm 1$ and achieve the same misclassification error.

In most settings of interest, the joint distribution $P_0$ of $(X, Y)$, which determines the optimal $h$, will be unknown. Instead we will suppose we have i.i.d. copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of the pair $(X, Y)$, known as *training data*. Our task is to use this data to construct a classifier $\hat{h}$ such that $R(\hat{h})$ is small. **Important point:** $R(\hat{h})$ is a random variable depending on the random training data:

$$R(\hat{h}) = \mathbb{E}(\ell(\hat{h}(X), Y)|X_1, Y_1, \ldots, X_n, Y_n).$$

A statistical approach to classification may attempt to model $P_0$ up to some unknown parameters, estimate these parameters, and thereby obtain an estimate of the regression function (or the conditional expectation in the case of least squares—see below). We will take a different approach and assume that we are given a class $\mathcal{H}$ of hypotheses from which to pick our $\hat{h}$. Possible choices of $\mathcal{H}$ include for instance

- $\mathcal{H} = \{h : h(x) = \text{sgn}(\mu + x^T\beta) \text{ where } \mu \in \mathbb{R},\ \beta \in \mathbb{R}^p\}$;
- $\mathcal{H} = \left\{h : h(x) = \text{sgn}\left(\mu + \sum_{j=1}^d \varphi_j(x)\beta_j\right) \text{ where } \mu \in \mathbb{R},\ \beta \in \mathbb{R}^d\right\}$ for a given *dictionary* of functions $\varphi_1, \ldots, \varphi_d : \mathcal{X} \to \mathbb{R}$.

**Technical note:** In this course we will take $\text{sgn}(0) = -1$. (It does not matter much whether we take $\text{sgn}(0) = \pm 1$, but we need to specify a choice in order that the $h$ defined above are classifiers.)

## 1.1 Brief review of conditional expectation

For many of the mathematical arguments in this course we will need to manipulate conditional expectations.

Recall that if $Z \in \mathbb{R}$ and $W \in \mathbb{R}^d$ are random elements with joint probability density function (pdf) $f_{Z,W}$ then the conditional pdf $f_{Z|W}$ of $Z$ given $W$ satisfies

$$f_{Z|W}(z|w) = \begin{cases} f_{Z,W}(z, w)/f_W(w) & \text{if } f_W(w) = 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $f_W$ is the marginal pdf of $W$. When one or more of $Z$ and $W$ are discrete we typically work with probability mass functions.

Suppose $\mathbb{E}|Z| < \infty$. Then the conditional expectation function $\mathbb{E}(Z|W = w)$ is given by

$$g(w) := \mathbb{E}(Z|W = w) = \int z f_{Z|W}(z|w) dz. \tag{1.1}$$

We write $\mathbb{E}(Z|W)$ for the random variable $g(W)$ (note this is a function of $W$, not $Z$).

This is not a fully general definition of conditional expectation (for that see the Stochastic Financial Models course) and we will not use it. We will however make frequent use of the following properties of conditional expectation.

(i) **Role of independence:** If $Z$ and $W$ are independent, then $\mathbb{E}(Z|W) = \mathbb{E}Z$. (Recall: $Z$ and $W$ being independent means $\mathbb{P}(Z \in A, W \in B) = \mathbb{P}(Z \in A)\mathbb{P}(W \in B)$ for all measurable $A \subseteq \mathbb{R}$, $B \subseteq \mathbb{R}^d$)

(ii) **Tower property:** Let $f : \mathbb{R}^d \to \mathbb{R}^m$ be a (measurable) function. Then

$$\mathbb{E}\{\mathbb{E}(Z|W)|f(W)\} = \mathbb{E}\{Z|f(W)\}.$$

In particular, $\mathbb{E}\{\mathbb{E}(Z|W)|W_1, \ldots, W_m\} = \mathbb{E}(Z|W_1, \ldots, W_m)$ for $m \leq d$. Taking $f \equiv c \in \mathbb{R}$ and using (i) gives us that $\mathbb{E}\{\mathbb{E}(Z|W)\} = \mathbb{E}(Z)$ (as $f(W)$ is a constant it is independent of any random variable).

(iii) **Taking out what is known:** If $\mathbb{E}Z^2 < \infty$ and $f : \mathbb{R}^d \to \mathbb{R}$ is such that $\mathbb{E}[\{f(W)\}^2] < \infty$ then $\mathbb{E}\{f(W)Z|W\} = f(W)\mathbb{E}(Z|W)$.

Probabilistic results can be 'applied conditionally', for example:

**Conditional Jensen.** Recall that $f : \mathbb{R} \to \mathbb{R}$ is a convex function if

$$tf(x) + (1-t)f(y) \geq f\big(tx + (1-t)y\big) \quad \text{for all } x, y \in \mathbb{R} \text{ and } t \in (0,1).$$

The conditional version of *Jensen's inequality* states that if $f : \mathbb{R} \to \mathbb{R}$ is convex and random variable $Z$ has $\mathbb{E}|f(Z)| < \infty$, then

$$\mathbb{E}\big(f(Z)|W\big) \geq f\big(\mathbb{E}(Z|W)\big).$$

## 1.2 Bayes risk

*Proof of Proposition 1.* We have

$$\begin{aligned} R(h) &= \frac{1}{4}\mathbb{E}\{(Y - h(X))^2\} \\ &= \frac{1}{4}\mathbb{E}\{(Y - \mathbb{E}(Y|X) + \mathbb{E}(Y|X) - h(X))^2\} \\ &= \frac{1}{4}\mathbb{E}\{(Y - \mathbb{E}(Y|X))^2\} + \frac{1}{4}\mathbb{E}\{(\mathbb{E}(Y|X) - h(X))^2\} + \frac{1}{2}\mathbb{E}\{(Y - \mathbb{E}(Y|X))(\mathbb{E}(Y|X) - h(X))\}. \end{aligned}$$

But

$$\begin{aligned} &\mathbb{E}\{(Y - \mathbb{E}(Y|X))(\mathbb{E}(Y|X) - h(X))\} \\ =\,&\mathbb{E}\,\mathbb{E}\{(Y - \mathbb{E}(Y|X))(\mathbb{E}(Y|X) - h(X))|X\} \quad \text{(tower property)} \\ =\,&\mathbb{E}\big[(\mathbb{E}(Y|X) - h(X))\mathbb{E}\{(Y - \mathbb{E}(Y|X))|X\}\big] \quad \text{(taking out what is known)} \\ =\,&0. \end{aligned}$$

Thus minimising $R(h)$ is equivalent to minimising $\mathbb{E}\{(\mathbb{E}(Y|X) - h(X))^2\}$. We therefore get $h_0(X) = \operatorname{sgn}(\mathbb{E}(Y|X))$. □

The proof also shows that the risk under least squares loss is minimised by taking $h(x) = \mathbb{E}(Y|X = x)$ (provided $\mathbb{E}Y^2 < \infty$).

## 1.3 Empirical risk minimisation

*Empirical risk minimisation* replaces the expectation over the unknown $P_0$ in the definition of the risk with the empirical distribution, and seeks to minimise the resulting objective over $h \in \mathcal{H}$:

$$\hat{R}(h) := \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i), \qquad \hat{h} \in \operatorname*{arg\,min}_{h\in\mathcal{H}} \hat{R}(h).$$

$\hat{R}(h)$ is the *empirical risk* or *training error* of $h$.

**Example.** Consider the regression setting with $\mathcal{Y} = \mathbb{R}$, squared error loss and $\mathcal{H} = \{x \mapsto \mu + x^T\beta \text{ for } \mu \in \mathbb{R},\ \beta \in \mathbb{R}^p\}$. Then empirical risk minimisation is equivalent to ordinary least squares, i.e. we have

$$\hat{h}(x) = \hat{\mu} + \hat{\beta}^T x \quad \text{where } (\hat{\mu}, \hat{\beta}) \in \operatorname*{arg\,min}_{(\mu,\beta)\in\mathbb{R}\times\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu - X_i^T\beta)^2.$$

A good choice for the class $\mathcal{H}$ will result in a low *generalisation error* $R(\hat{h})$. This is a measure of how well we can expect the empirical risk minimiser (ERM) $\hat{h}$ to predict a new data point $(X_{\text{new}}, Y_{\text{new}}) \sim P_0$ given only knowledge of $X_{\text{new}}$. Define $h^* \in \operatorname*{arg\,min}_{h\in\mathcal{H}} R(h)$[2] and consider the decomposition

$$R(h) - R(h_0) = \underbrace{R(h) - R(h^*)}_{\text{stochastic error}} + \underbrace{R(h^*) - R(h_0)}_{\text{approximation error}}.$$

Clearly a richer class $\mathcal{H}$ will decrease the approximation error. However, it will tend to increase the stochastic error as empirical risk minimisation will fit to the realised $Y_1, \ldots, Y_n$ too closely and result in poor generalisation. There is thus a tradeoff between the stochastic error due to the complexity of the class $\mathcal{H}$, and its approximation error.

We will primarily study the stochastic term or *excess risk*[3], and aim to provide bounds on this in terms of the complexity of $\mathcal{H}$. Recall that whilst for a fixed $h \in \mathcal{H}$, $R(h)$ is deterministic, $R(\hat{h})$ is a random variable. The bounds we obtain will be of the form "with probability at least $1 - \delta$,

$$R(\hat{h}) - R(h^*) \leq \epsilon."$$

[2]If there is no $h^*$ that achieves the associated infimum, we can consider an approximate minimiser with $R(h^*) < \inf_{h\in\mathcal{H}} R(h) + \epsilon$ for arbitrary $\epsilon > 0$ and all our analysis will carry through. Similar reasoning is applicable to $\hat{h}$.

[3]Sometimes "excess risk" is used for $R(\hat{h}) - R(h_0)$. However since we are considering $\mathcal{H}$ to be fixed in advance for much of the course, we will use excess risk to refer to the risk relative to that of $h^*$.

# 2 Statistical learning theory

Consider the following decomposition of the excess risk:

$$\begin{aligned} R(\hat{h}) - R(h^*) &= \underbrace{R(\hat{h}) - \hat{R}(\hat{h})}_{\text{concentration}} + \underbrace{\hat{R}(\hat{h}) - \hat{R}(h^*)}_{\leq 0} + \underbrace{\hat{R}(h^*) - R(h^*)}_{\text{concentration}} \\ &\leq R(\hat{h}) - \hat{R}(\hat{h}) + \hat{R}(h^*) - R(h^*). \end{aligned}$$

Note that $\hat{R}(h^*)$ is an average of $n$ i.i.d. random variables, each with expectation $R(h^*)$. To bound $\hat{R}(h^*) - R(h^*)$ we will consider the general problem of how random variables concentrate around their expectation, a problem which is the topic of an important area of probability theory concerning *concentration inequalities*. The term $R(\hat{h}) - \hat{R}(\hat{h})$ is more complicated as $\hat{R}(\hat{h})$ is not a sum of i.i.d. random variables, but we will see how extensions of techniques for the simpler case may be used to tackle this.

## 2.1 Sub-Gaussianity and Hoeffding's inequality

We begin our discussion of concentration inequalities with the simplest tail bound, *Markov's inequality*. Let $W$ be a non-negative random variable. Taking expectations of both sides of $t\mathbb{1}_{\{W \geq t\}} \leq W$ for $t > 0$, we obtain after dividing through by $t$

$$\mathbb{P}(W \geq t) \leq \frac{\mathbb{E}(W)}{t}.$$

This immediately implies that given a strictly increasing function $\varphi : \mathbb{R} \to [0, \infty)$ and any random variable $W$,

$$\mathbb{P}(W \geq t) = \mathbb{P}\big(\varphi(W) \geq \varphi(t)\big) \leq \frac{\mathbb{E}(\varphi(W))}{\varphi(t)}.$$

Applying this with $\varphi(t) = e^{\alpha t}$ $(\alpha > 0)$ yields the so-called *Chernoff bound*:

$$\mathbb{P}(W \geq t) \leq \inf_{\alpha > 0} e^{-\alpha t} \mathbb{E} e^{\alpha W}.$$

**Example.** Consider the case when $W \sim N(0, \sigma^2)$. Recall that

$$\mathbb{E} e^{\alpha W} = e^{\alpha^2 \sigma^2 / 2}. \tag{2.1}$$

Thus for $t \geq 0$,

$$\mathbb{P}(W \geq t) \leq \inf_{\alpha > 0} e^{\alpha^2 \sigma^2 / 2 - \alpha t} = e^{-t^2/(2\sigma^2)}. \tag{2.2}$$

Note that to arrive at this bound, all we required was (an upper bound on) the moment generating function (mgf) of $W$ (2.1). This motivates the following definition.

**Definition 1.** We say a random variable $W$ is *sub-Gaussian* with parameter $\sigma > 0$ if

$$\mathbb{E}e^{\alpha(W-\mathbb{E}W)} \leq e^{\alpha^2\sigma^2/2} \quad \text{for all } \alpha \in \mathbb{R}.$$

From (2.2) we immediately have the following result.

**Proposition 2.** *If $W$ is* sub-Gaussian *with parameter $\sigma > 0$, then*

$$\mathbb{P}(W - \mathbb{E}W \geq t) \leq e^{-t^2/(2\sigma^2)} \quad \textit{for all } t \geq 0.$$

Note that if $W$ is sub-Gaussian with parameter $\sigma > 0$, then

- it is also sub-Gaussian with parameter $\sigma'$ for any $\sigma' \geq \sigma$;
- $-W$ is also sub-Gaussian with parameter $\sigma > 0$. This means we have from (2.2) that

$$\mathbb{P}(|W - \mathbb{E}W| \geq t) \leq \mathbb{P}(W - \mathbb{E}W \geq t) + \mathbb{P}(-(W - \mathbb{E}W) \geq t) \leq 2e^{-t^2/(2\sigma^2)}.$$

**Proposition 3.** *Suppose $W_1, \ldots, W_n$ are independent and each $W_i$ is sub-Gaussian with parameter $\sigma_i$ and has mean 0. Then for $\gamma \in \mathbb{R}^n$, $\gamma^T W$ is sub-Gaussian with parameter $\left(\sum_i \gamma_i^2\sigma_i^2\right)^{1/2}$.*

*Proof.*

$$\begin{aligned}
\mathbb{E}\exp\left(\alpha\sum_{i=1}\gamma_i W_i\right) &= \prod_{i=1}\mathbb{E}\exp(\alpha\gamma_i W_i) \\
&\leq \prod_{i=1}^{n}\exp(\alpha^2\gamma_i^2\sigma_i^2/2) \\
&= \exp\left(\alpha^2\sum_{i=1}^{n}\gamma_i^2\sigma_i^2/2\right). \qquad \square
\end{aligned}$$

Combining with Proposition 2 we obtain

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(W_i - \mathbb{E}W_i) \geq t\right) \leq \exp\left(-\frac{n^2t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right) \quad \text{for } t \geq 0. \tag{2.3}$$

As well as implying concentration around the mean, the bound on the mgf satisfied by sub-Gaussian random variables also offers a bound on the expected maximum of $d$ sub-Gaussians. We do not need the following result at this stage, but will make use of it later.

**Proposition 4.** *Suppose $W_1, \ldots, W_d$ are all mean-zero and sub-Gaussian with parameter $\sigma > 0$ (but are not necessarily independent). Then*

$$\mathbb{E}\max_j W_j \leq \sigma\sqrt{2\log(d)}.$$

*Proof.* Let $\alpha > 0$. By convexity of $x \mapsto \exp(\alpha x)$ and Jensen's inequality we have

$$\exp(\alpha \mathbb{E} \max_j W_j) \leq \mathbb{E} \exp(\alpha \max_j W_j) = \mathbb{E} \max_j \exp(\alpha W_j).$$

Now

$$\mathbb{E} \max_{j=1,\ldots,d} \exp(\alpha W_j) \leq \sum_{j=1}^{d} \mathbb{E} \exp(\alpha W_j) \leq d e^{\alpha^2 \sigma^2 / 2}.$$

Thus

$$\mathbb{E} \max_j W_j \leq \frac{\log(d)}{\alpha} + \frac{\alpha \sigma^2}{2}.$$

Optimising over $\alpha > 0$ yields the result. □

Gaussian random variables are sub-Gaussian, but the sub-Gaussian class is much broader than this.

**Example.** A *Rademacher* random variable $\varepsilon$ takes values $\{-1, 1\}$ with equal probability. It is sub-Gaussian with parameter $\sigma = 1$:

$$\begin{aligned}
\mathbb{E} e^{\alpha \varepsilon} &= \frac{1}{2}(e^{-\alpha} + e^{\alpha}) = \frac{1}{2}\left( \sum_{k=0}^{\infty} \frac{(-\alpha)^k}{k!} + \sum_{k=0}^{\infty} \frac{\alpha^k}{k!} \right) \\
&= \sum_{k=0}^{\infty} \frac{\alpha^{2k}}{(2k)!} \\
&\leq \sum_{k=0}^{\infty} \frac{\alpha^{2k}}{2^k k!} = e^{\alpha^2/2} \quad \text{(using } (2k)! \geq 2^k k!\text{)}. \qquad (2.4)
\end{aligned}$$

Recall that we are interested in the concentration properties of $\mathbb{1}_{\{h(X_i) \neq Y_i\}} - \mathbb{P}(h(X_i) \neq Y_i)$, which in particular is bounded.

**Lemma 5** (Hoeffding's lemma)**.** *If $W$ is mean-zero and takes values in $[a, b]$, then $W$ is sub-Gaussian with parameter $\sigma = (b - a)/2$.*

*Proof.* We will prove a weaker result here with $\sigma = b - a$; see the Example sheet for a proof with $\sigma = (b - a)/2$. Let $W'$ be an independent copy of $W$. We have

$$\begin{aligned}
\mathbb{E} e^{\alpha W} &= \mathbb{E} e^{\alpha(W - \mathbb{E} W')} \\
&= \mathbb{E} e^{\mathbb{E}\{\alpha(W - W') | W\}} \quad \text{using } \mathbb{E}(W') = \mathbb{E}(W'|W) \text{ and } \mathbb{E}(W|W) = W \\
&\leq \mathbb{E} e^{\alpha(W - W')} \quad \text{(Jensen conditional on } W \text{ and tower prop.)}.
\end{aligned}$$

Now $W - W' \stackrel{d}{=} -(W - W') \stackrel{d}{=} \varepsilon(W - W')$ where $\varepsilon \sim$ Rademacher with $\varepsilon$ independent of $(W, W')$. (Here "$\stackrel{d}{=}$" means "equal in distribution".) Thus

$$\mathbb{E} e^{\alpha W} \leq \mathbb{E} e^{\alpha \varepsilon (W - W')} = \mathbb{E}\{\mathbb{E}(e^{\alpha \varepsilon (W - W')} | W, W')\}.$$

We now apply our previous result (2.4) conditionally on $(W - W')$ to obtain

$$\mathbb{E}e^{\alpha W} \leq \mathbb{E}e^{\alpha^2(W-W')^2/2} \leq \mathbb{E}e^{\alpha^2(b-a)^2/2}$$

as $|W - W'| \leq b - a$. □

The introduction of an independent copy $W'$ and a Rademacher random variable here is an example of a *symmetrisation argument*; we will make use of this technique again later in the course. As an application of the result above, suppose $W_1, \ldots, W_n$ are independent, mean-zero and $a_i \leq W_i \leq b_i$ almost surely for all $i$. Then

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} W_i \geq t\right) \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \quad \text{for } t \geq 0, \tag{2.5}$$

which is known as *Hoeffding's inequality*.

We are now in a position to bound $R(\hat{h}) - R(h^*)$ when $\mathcal{H}$ is finite.

## 2.2 Finite hypothesis classes

**Theorem 6.** *Suppose $\mathcal{H}$ is finite and $\ell$ takes values in $[0, M]$. Then with probability at least $1 - \delta$, the ERM $\hat{h}$ satisfies*

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

The assumption on $\ell$ includes as a special case misclassification loss. However the extra generality will prove helpful later in the course.

*Proof.* Recall that

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

Now for each $h$, $\hat{R}(h) - R(h)$ is an average of mean-zero i.i.d. quantities of the form $\ell(h(X_i), Y_i) - \mu$ with $\mu = \mathbb{E}\ell(h(X_i), Y_i)$ which take values in $[-\mu, M - \mu]$. For $t > 0$,

$$\begin{aligned}\mathbb{P}(R(\hat{h}) - R(h^*) > t) &= \mathbb{P}(R(\hat{h}) - R(h^*) > t, \hat{h} \neq h^*) \\ &\leq \mathbb{P}(R(\hat{h}) - \hat{R}(\hat{h}) > t/2, \hat{h} \neq h^*) + \mathbb{P}(\hat{R}(h^*) - R(h^*) > t/2)\end{aligned}$$

We can immediately apply Hoeffding's inequality to the second term to obtain

$$\mathbb{P}(\hat{R}(h^*) - R(h^*) \geq t/2) \leq \exp(-nt^2/(2M^2)).$$

However the complicated dependence among the summands in $\hat{R}(\hat{h})$ prevents this line of attack for bounding the first term. To tackle this issue, we note that when $\hat{h} \neq h^*$,

$$R(\hat{h}) - \hat{R}(\hat{h}) \leq \max_{h \in \mathcal{H}_-} R(h) - \hat{R}(h),$$

where $\mathcal{H}_- := \mathcal{H} \setminus \{h^*\}$. We then have using a union bound,

$$\begin{aligned}\mathbb{P}(\max_{h\in\mathcal{H}_-} R(h) - \hat{R}(h) \geq t/2) &= \mathbb{P}(\cup_{h\in\mathcal{H}_-} R(h) - \hat{R}(h) \geq t/2) \\ &\leq \sum_{h\in\mathcal{H}_-} \mathbb{P}(R(h) - \hat{R}(h) \geq t/2) \\ &\leq |\mathcal{H}_-| \exp(-nt^2/(2M^2)).\end{aligned}$$

Thus

$$\mathbb{P}(R(\hat{h}) - R(h^*) > t) \leq |\mathcal{H}| \exp(-nt^2/(2M^2)).$$

Writing $\delta := |\mathcal{H}| \exp(-nt^2/(2M^2))$ and then expressing $t$ in terms of $\delta$ gives the result. $\square$

**Example.** Consider a simple classification setting with $X_i \in [0,1)^2$. Let us divide $[0,1)^2$ into $m^2$ disjoint squares $R_1, \ldots, R_{m^2} \subset [0,1)^2$ of the form $[r/m, (r+1)/m) \times [s/m, (s+1)/m)$ for $r, s = 0, \ldots, m-1$. Let

$$\bar{Y}_j = \mathrm{sgn}\Big(\sum_{i: X_i \in R_j} Y_i\Big)$$

and define

$$\hat{h}^{\mathrm{hist}}(x) = \sum_{j=1}^{m^2} \bar{Y}_j \mathbb{1}_{R_j}(x).$$

Then $\hat{h}^{\mathrm{hist}}$ is equivalent to the ERM over hypothesis class $\mathcal{H}$ consisting of the $2^{m^2}$ classifiers each corresponding to a way of assigning labels in $\{-1, 1\}$ to each of the regions $R_1, \ldots, R_{m^2}$. The result above tells us that the generalisation error (with misclassification loss) of $\hat{h}^{\mathrm{hist}}$ is at most

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

[In fact it can be shown that the approximation error $R(h^*) - R(h_0) \to 0$ if $m \to \infty$ for any given $P_0$. Combining with the above, we then see that choosing e.g. $m = n^{1/3}$ we can approach the Bayes risk for $n$ sufficiently large.]

Whilst a union bound and Hoeffding's inequality sufficed to give us a bound in the case where $\mathcal{H}$ is finite, to handle the more common setting where $\mathcal{H}$ is infinite, we will need more sophisticated techniques. Our approach will be to view the key quantity

$$G(X_1, Y_1, \ldots, X_n, Y_n) := \sup_{h\in\mathcal{H}} \{R(h) - \hat{R}(h)\}$$

as a function $G$ of the i.i.d. random elements $(X_1, Y_1), \ldots, (X_n, Y_n)$. We currently only have at our disposal concentration inequalities where $g$ takes the form of an average; however $G$ will in general clearly be much more complex. Intuitively though, the key property of the empirical average that results in concentration is that the individual contributions of each

of the random elements is not too large. Can we show that our $G$ would, despite having an intractable form, nevertheless share this property in common with the empirical average?

Given data $(x_1, y_1), \dots (x_n, y_n)$ and $\epsilon > 0$, let $\tilde{h} \in \mathcal{H}$ be such that

$$G(x_1, y_1, \dots, x_n, y_n) < R(\tilde{h}) - \hat{R}(\tilde{h}) + \epsilon.$$

Now consider perturbing (wlog) the first pair of arguments of $G$. We have

$$\begin{aligned}
&G(x_1, y_1, \dots, x_n, y_n) - G(x_1', y_1', x_2, y_2, \dots, x_n, y_n) \\
&< R(\tilde{h}) - \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \tilde{h}(x_i)) - \sup_{h \in \mathcal{H}} \left( R(h) - \frac{1}{n}\ell(y_1', h(x_1')) - \frac{1}{n}\sum_{i=2}^{n} \ell(y_i, h(x_i)) \right) + \epsilon \\
&\leq \frac{1}{n}\{\ell(y_1', \tilde{h}(x_1')) - \ell(y_1, \tilde{h}(x_1))\} + \epsilon.
\end{aligned}$$

As $\epsilon$ was arbitrary, if $\ell$ takes values in $[0, M]$ we have

$$G(x_1, y_1, \dots, x_n, y_n) - G(x_1', y_1', x_2, y_2, \dots, x_n, y_n) \leq M/n.$$

We thus seek a concentration inequality for multivariate functions where arbitrary perturbations of a single argument change the output by a bounded amount.

## 2.3 Bounded differences inequality

The result we are going to aim for is the so-called Bounded differences inequality. Let us adopt the notation that for a sequence of vectors $a_s, a_{s+1}, a_{s+2}, \dots$ (where the starting index $s$ can be e.g. 0 or 1), $a_{j:k}$ for $j \leq k$ is the subsequence $a_j, \dots, a_k$.

**Theorem 7** (Bounded differences inequality)**.** *Let $f : \mathbb{R}^{d \cdot n} \to \mathbb{R}$ satisfy a* bounded differences property *such that*

$$f(w_1, \dots, w_{i-1}, w_i, w_{i+1}, \dots, w_n) - f(w_1, \dots, w_{i-1}, w_i', w_{i+1}, \dots, w_n) \leq L_i,$$

*for all $w_1, \dots, w_n, w_i' \in \mathbb{R}^d$, and all $i = 1, \dots, n$. Suppose random vectors $W_1, \dots, W_n$ taking values in $\mathbb{R}^d$ are independent. Then for $t \geq 0$,*

$$\mathbb{P}\big(f(W_{1:n}) - \mathbb{E}f(W_{1:n}) \geq t\big) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} L_i^2}\right)$$

Note that taking $d = 1$, $f(W_{1:n}) = \sum_i \{W_i - \mathbb{E}(W_i)\}/n$ and assuming the $W_i$ take values in $[a_i, b_i]$, we recover Hoeffding's inequality.

To motivate the proof, consider the sequence of random variables given by $Z_0 = \mathbb{E}f(W_{1:n})$, $Z_n = f(W_{1:n})$ and

$$Z_i = \mathbb{E}(f(W_{1:n})|W_{1:i}) \quad \text{for } i = 1, \dots, n-1. \tag{2.6}$$

Note that in the special case where $f(W_{1:n}) = \sum_i W_i$ and $\mathbb{E}W_i = 0$, we have $Z_k - Z_0 = \sum_{i=1}^k W_i$. Our approach centres on the telescoping decomposition

$$f(W_{1:n}) - \mathbb{E}f(W_{1:n}) = Z_n - Z_0 = \sum_{i=1}^n \underbrace{(Z_i - Z_{i-1})}_{D_i};$$

the differences $D_i$ play an analogous role to the individual independent random variables in the case of bounding sums. These considerations help to motivate the definition of a *martingale*[4], which encapsulates the key features of the $Z_i$ above that will allow our arguments to go through.

**Definition 2.** A sequence of random variables $Z_0, Z_1, \ldots, Z_n$ is a *martingale sequence* with respect to another sequence of random vectors $W_0, W_1, \ldots, W_n \in \mathbb{R}^d$ if

(i) $\mathbb{E}|Z_i| < \infty$ for $i = 0, \ldots, n$,

(ii) $Z_i$ is a function of $W_{0:i}$ for $i = 0, \ldots, n$,

(iii) $\mathbb{E}(Z_i|W_{0:(i-1)}) = Z_{i-1}$ $i = 1, \ldots, n$.

We call $D_i := Z_i - Z_{i-1}$ a *martingale difference sequence* with respect to $W_0, W_1, \ldots, W_n$. We also extend the definitions to infinite sequences.

**Example.** Consider a random walk with $Z_0 = 0$, $Z_i = Z_{i-1} \pm 1$ each with probability $1/2$. Then $(Z_i)_{i\geq 1}$ is just a sequence of partial sums of i.i.d. Rademacher random variables and is a martingale sequence with respect to $(Z_{i-1})_{i\geq 1}$. Now suppose $Z_i = Z_{i-1}$ if $Z_i = z$ for some $z \in \mathbb{Z}$ but with the dynamics governed by the random walk otherwise. We still have

$$\mathbb{E}(Z_i|Z_{0:(i-1)}) = z\mathbb{1}_{\{Z_{i-1}=z\}} + Z_{i-1}\mathbb{1}_{\{Z_{i-1}\neq z\}} = Z_{i-1}.$$

Note that in this case the corresponding martingale difference sequence is not i.i.d. and exhibits some dependence.

**Example.** The sequence $Z_{0:n}$ defined earlier via (2.6) is an example of a *Doob martingale*. Formally it is a martingale sequence with respect to $W_{0:n}$ where we may set $W_0$ to an arbitrary constant. That (ii) holds is clear. (i) certainly holds when $f$ is bounded, but holds more generally provided $\mathbb{E}|f(W_{1:n})| < \infty$:

$$\mathbb{E}|Z_i| = \mathbb{E}|\mathbb{E}\{f(W_{1:n})|W_{1:i}\}| \leq \mathbb{E}|f(W_{1:n})| < \infty$$

using Jensen's inequality conditional on $W_{1:i}$ applied to the convex function $|\cdot|$, and then the tower property. That (iii) holds follows from the tower property of conditional expectation.

We are now in a position to prove a generalisation of Proposition 3 applicable to averages of martingale differences.

[4]For a more general and formal definition of a martingale, see the Stochastic Financial Models course.

**Lemma 8.** *Let $D_1, \dots, D_n$ be a martingale difference sequence with respect to $W_0, \dots, W_n \in \mathbb{R}^d$ such that*

$$\mathbb{E}(e^{\alpha D_i} | W_{0:(i-1)}) \leq e^{\alpha^2 \sigma_i^2 / 2} \quad \textit{almost surely.}$$

*Let $\gamma \in \mathbb{R}^n$ and write $D = (D_1, \dots, D_n)^T$. Then $\gamma^T D$ is sub-Gaussian with parameter $(\sum_i \gamma_i^2 \sigma_i^2)^{1/2}$.*

*Proof.* We have

$$\begin{aligned}
\mathbb{E} \exp\left(\alpha \sum_{i=1}^n \gamma_i D_i\right) &= \mathbb{E} \exp\left(\alpha \sum_{i=1}^{n-1} \gamma_i D_i\right) \mathbb{E}(e^{\alpha \gamma_n D_n} | W_{0:(n-1)}) \\
&\leq e^{\alpha^2 \gamma_n^2 \sigma_n^2 / 2} \mathbb{E} \exp\left(\alpha \sum_{i=1}^{n-1} \gamma_i D_i\right) \\
&\leq \exp\left(\frac{\alpha^2}{2} \sum_{i=1}^n \gamma_i^2 \sigma_i^2\right) \quad \text{(by induction).}
\end{aligned}$$

□

The Azuma-Hoeffding inequality applies the above result to the case of bounded random variables.

**Theorem 9** (Azuma–Hoeffding)**.** *Let $D_1, \dots, D_n$ be a martingale difference sequence with respect to $W_0, \dots, W_n \in \mathbb{R}^d$. Suppose the following holds for each $i = 1, \dots, n$: there exist random variables $A_i$ and $B_i$ that are functions of $W_{0:(i-1)}$ such that $A_i \leq D_i \leq B_i$ almost surely and $B_i - A_i \leq L_i$ almost surely for a constant $L_i$. Then for $t \geq 0$,*

$$\mathbb{P}\left(\sum_{i=1}^n D_i \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n L_i^2}\right). \tag{2.7}$$

*Proof.* Conditional on $W_{0:(i-1)}$, $A_i$ and $B_i$ are constant. Thus we may apply Hoeffding's Lemma (Lemma 5) conditionally on $W_{0:(i-1)}$ to obtain

$$\mathbb{E}(e^{\alpha D_i} | W_{0:(i-1)}) \leq e^{\alpha^2 (L_i/2)^2 / 2} \quad \text{almost surely.}$$

The martingale difference sequence thus satisfies the hypotheses of Lemma 8. The sum $\sum_i D_i$ is sub-Gaussian with parameter $\sigma = (\sum_i L_i^2)^{1/2}/2$. The result then follows from the sub-Gaussian tail bound (Proposition 2). □

We are finally ready to prove the Bounded differences inequality.

*Proof of Theorem 7.* Let $D_1, \dots, D_n$ be the Martingale difference sequence associated with the Doob martingale, so $D_1 = \mathbb{E}(f(W_{1:n})|W_1) - \mathbb{E}f(W_{1:n})$ and for $i = 2, \dots, n$,

$$D_i = \mathbb{E}(f(W_{1:n})|W_{1:i}) - \mathbb{E}(f(W_{1:n})|W_{1:i-1}).$$

Recall that $f(W_{1:n}) - \mathbb{E}f(W_{1:n}) = \sum_{i=1}^{n} D_i$. Using the Azuma–Hoeffding inequality, it suffices to prove that $A_i \leq D_i \leq B_i$ almost surely where $A_i$ and $B_i$ are functions of $W_{1:(i-1)}$ satisfying $B_i - A_i \leq L_i$ for all $i$, which we now do.

Let us define for each $i = 1, \ldots, n$, functions

$$\begin{aligned} F_i : \mathbb{R}^{d \cdot i} &\to \mathbb{R} \\ (\underbrace{w_1}_{\in \mathbb{R}^d}, \ldots, w_i) &\mapsto \mathbb{E}(f(W_{1:n}) | W_1 = w_1, \ldots, W_i = w_i). \end{aligned}$$

Then define the random variables

$$\begin{aligned} A_1 &:= \inf_{w_1 \in \mathbb{R}^d} F_1(w_1) - \mathbb{E}f(W_{1:n}), \quad A_i := \inf_{w_i \in \mathbb{R}^d} F_i(W_{1:(i-1)}, w_i) - \mathbb{E}(f(W_{1:n}) | W_{1:(i-1)}) \quad (i \geq 2) \\ B_1 &:= \sup_{w_1 \in \mathbb{R}^d} F_1(w_1) - \mathbb{E}f(W_{1:n}), \quad B_i := \sup_{w_i \in \mathbb{R}^d} F_i(W_{1:(i-1)}, w_i) - \mathbb{E}(f(W_{1:n}) | W_{1:(i-1)}) \quad (i \geq 2). \end{aligned}$$

Note that setting $W_0 \equiv 0$, $A_i$ and $B_i$ are functions of $W_{0:(i-1)}$. Furthermore for $i = 2, \ldots, n$

$$\begin{aligned} D_i - A_i &= F_i(W_{1:i}) - \inf_{w_i \in \mathbb{R}^d} F_i(W_{1:(i-1)}, w_i) \geq 0 \\ D_i - B_i &= F_i(W_{1:i}) - \sup_{w_i \in \mathbb{R}^d} F_i(W_{1:(i-1)}, w_i) \leq 0, \end{aligned}$$

so $A_i \leq D_i \leq B_i$. Also

$$\begin{aligned} B_i - A_i &= \sup_{w_i \in \mathbb{R}^d} F_i(W_{1:(i-1)}, w_i) - \inf_{w_i \in \mathbb{R}^d} F_i(W_{1:(i-1)}, w_i) \\ &= \sup_{w_i, w_i' \in \mathbb{R}^d} \{F_i(W_{1:(i-1)}, w_i) - F_i(W_{1:(i-1)}, w_i')\} \\ &= \sup_{w_i, w_i' \in \mathbb{R}^d} \{\mathbb{E}(f(W_{1:(i-1)}, w_i, W_{(i+1):n}) | W_{1:(i-1)}, W_i = w_i) \\ &\qquad - \mathbb{E}(f(W_{1:(i-1)}, w_i', W_{(i+1):n}) | W_{1:(i-1)}, W_i = w_i')\}. \end{aligned}$$

Now as the $W_{1:n}$ are independent, the distribution of $W_{(i+1):n}$ conditional on $W_{1:(i-1)}$ and $W_i = w_i$ is equal to its unconditional distribution, and the same holds when conditioning on $W_i = w_{i'}$. Thus

$$\begin{aligned} B_i - A_i &= \sup_{w_i, w_i' \in \mathbb{R}^d} [\mathbb{E}\{\underbrace{f(W_{1:(i-1)}, w_i, W_{(i+1):n}) - f(W_{1:(i-1)}, w_i', W_{(i+1):n})}_{\leq L_i} | W_{1:(i-1)}\}] \\ &\leq L_i. \end{aligned}$$

Analogous properties hold when $i = 1$.

We have have verified all the conditions of the Azuma–Hoeffding inequality which may now be applied to give the result. □

## 2.4 Rademacher complexity

Recall our setup: $\mathcal{H}$ is a (now possibly infinite) hypothesis class, $\ell$ takes values in $[0, M]$ are we are aiming to bound the right-hand side of

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

where $G := \sup_{h \in \mathcal{H}} \{R(h) - \hat{R}(h)\}$. The Bounded differences inequality provides a means to bound $G - \mathbb{E}G$, but in order to make use of this, we must find a way of bounding $\mathbb{E}G$. Let us write $Z_i = (X_i, Y_i)$ for $i = 1, \ldots, n$ and

$$\mathcal{F} := \{(x, y) \mapsto -\ell(h(x), y) : h \in \mathcal{H}\}. \tag{2.8}$$

Then we have

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

We will prove the following result which applies for a general function class $\mathcal{F}$ (not necessarily coming from (2.8)).

**Theorem 10.** *Let $\mathcal{F}$ be a class of real-valued functions $f : \mathcal{Z} \to \mathbb{R}$ and let $Z_1, \ldots, Z_n$ be i.i.d. random elements taking values in $\mathcal{Z}$. Then*

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

*where $\mathcal{R}_n(\mathcal{F})$ is the* Rademacher complexity *of $\mathcal{F}$ defined by*

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

*Here $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables independent of $Z_{1:n}$.*

Some intuition: Consider a classification problem with inputs $Z_1, \ldots, Z_n$ and *completely random* labels $\varepsilon_1, \ldots, \varepsilon_n$. The Rademacher complexity then captures how closely aligned the 'predictions' $f(Z_i)$ are to the random labels.

Before we prove Theorem 10, let us reflect on what it might achieve. Considering our main problem of bounding $\mathbb{E}G$, a key challenge is that it depends strongly and in a complicated way on the unknown $P_0$. To understand the potential advantages of Rademacher complexity, it is helpful to introduce the following.

**Definition 3.** Let $\mathcal{F}$ be a class of real-valued functions $f : \mathcal{Z} \to \mathbb{R}$ and let $z_1, \ldots, z_n \in \mathcal{Z}$. Writing

$$\mathcal{F}(z_{1:n}) = \{(f(z_1), \ldots, f(z_n)) : f \in \mathcal{F}\}$$

define the *empirical Rademacher complexity*

$$\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) := \mathbb{E}\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(z_i)\right), \tag{2.9}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables. Given i.i.d. random elements $Z_1, \ldots, Z_n$ taking values in $\mathcal{Z}$, we sometimes view the empirical Rademacher complexity as a random variable:

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

Note that $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n}))$ is well-defined in that the right-hand side of (2.9) only depends on $\mathcal{F}(z_{1:n})$, the behaviours of the functions in $\mathcal{F}$ on the fixed set of points $z_{1:n}$.

Key point: $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n}))$ does not depend on $P_0$. It is conceivable that we could upper bound $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n}))$ uniformly in $z_{1:n} \in \mathcal{Z}^n$. We then immediately get a bound on $\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\{\hat{\mathcal{R}}(\mathcal{F}(Z_{1:n}))\}$ that is independent of $P_0$.

We now turn to the proof of the result, which uses a symmetrisation technique.

*Proof of Theorem 10.* Let us introduce an independent copy $(Z_1', \ldots, Z_n')$ of $(Z_1, \ldots, Z_n)$. We have

$$\begin{aligned}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\{f(Z_i) - \mathbb{E}f(Z_i)\} &= \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\{f(Z_i) - f(Z_i')|Z_{1:n}\} \text{ (independence of } Z_{1:n} \text{ and } Z_{1:n}')\\ &\leq \mathbb{E}\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\{f(Z_i) - f(Z_i')\}\middle| Z_{1:n}\right).\end{aligned}$$

Note we have used the fact that for any collection of random variables $V_t$, $\sup_{t'} \mathbb{E}V_{t'} \leq \mathbb{E}\sup_t V_t$; this may easily be verified by removing the supremum over $t'$ and noting that the resulting inequality must hold for all $t'$. Now let $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. Rademacher random variables, independent of $Z_{1:n}$ and $Z_{1:n}'$. Then

$$\begin{aligned}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\{f(Z_i) - f(Z_i')\} &\stackrel{d}{=} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\{f(Z_i) - f(Z_i')\}\\ &\leq \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(Z_i) + \sup_{g\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\{-\varepsilon_i g(Z_i)\}.\end{aligned}$$

Noting that $\varepsilon_{1:n} \stackrel{d}{=} -\varepsilon_{1:n}$, we have

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\{f(Z_i) - f(Z_i')\}\right) \leq \mathbb{E}\left(\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i f(Z_i)\right) = 2\mathcal{R}_n(\mathcal{F}). \qquad \square$$

**Theorem 11** (Generalisation bound based on Rademacher complexity)**.** *Let* $\mathcal{F} := \{(x, y) \mapsto \ell(h(x), y) : h \in \mathcal{H}\}$ *and suppose* $\ell$ *takes values in* $[0, M]$*. With probability at least* $1 - \delta$*,*

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

*Proof.* Let $G := \sup_{h\in\mathcal{H}} R(h) - \hat{R}(h)$ and recall that

$$R(\hat{h}) - R(h^*) \leq G + \hat{R}(h^*) - R(h^*) = (G - \mathbb{E}G) + \mathbb{E}G + \hat{R}(h^*) - R(h^*).$$

Further recall that viewing $G$ as a function of $Z_1, \ldots, Z_n$ where $Z_i = (X_i, Y_i)$, it satisfies a bounded differences property with constants $L_i = M/n$. Thus the Bounded differences inequality gives us that

$$\mathbb{P}(G - \mathbb{E}G \geq t/2) \leq \exp(-t^2 n/(2M^2)).$$

Hoeffding's inequality (or Bounded differences with the average function) also gives $\mathbb{P}(\hat{R}(h^*) - R(h^*) \geq t/2) \leq \exp(-t^2 n/(2M^2))$. Also, noting that $\mathcal{R}_n(\mathcal{F}) = \mathcal{R}_n(-\mathcal{F})$, from Theorem 10, $\mathbb{E}G \leq 2\mathcal{R}_n(\mathcal{F})$. Thus taking $t = M\sqrt{2\log(2/\delta)/n}$ gives the result. □

## 2.5 VC dimension

All we need to do in order to bound the generalisation error is to obtain bounds on the Rademacher complexity. There are various ways of tackling this problem in general. Here, we will explore an approach suited to the classification setting with misclassification loss and $\mathcal{F} := \{(x, y) \mapsto \ell(h(x), y) : h \in \mathcal{H}\}$. Our bounds will be in terms of the number of behaviours of the function class on $n$ points $|\mathcal{F}(z_{1:n})|$. Observe first that $|\mathcal{F}(z_{1:n})| = |\mathcal{H}(x_{1:n})|$ where $z_i = (x_i, y_i)$.

**Lemma 12.** *We have* $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) \leq \sqrt{2\log(|\mathcal{F}(z_{1:n})|)/n} = \sqrt{2\log(|\mathcal{H}(x_{1:n})|)/n}$.

*Proof.* Let $d = |\mathcal{F}(z_{1:n})|$ and let $\mathcal{F}' := \{f_1, \ldots, f_d\}$ be such that $\mathcal{F}(z_{1:n}) = \mathcal{F}'(z_{1:n})$ (so each $f_j$ has a unique behaviour on $z_{1:n}$). For $j = 1, \ldots, d$, let

$$W_j = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f_j(z_i),$$

where $\varepsilon_{1:n}$ are i.i.d. Rademacher random variables. Then $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) = \mathbb{E}\max_j W_j$. By Lemma 5 and Proposition 3, each $W_j$ is sub-Gaussian with parameter $1/\sqrt{n}$. Thus we may apply Proposition 4 on the expected maximum of sub-Gaussian random variables to give the result. □

As each $h(x_i) \in \{-1, 1\}$, we always have $|\mathcal{H}(x_{1:n})| \leq 2^n$. Considering the result above, an interesting case then is when $|\mathcal{H}(x_{1:n})|$ is growing slower than exponentially in $n$, e.g. growing polynomially in $n$.

**Definition 4.** Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \{a, b\}$ with $a \neq b$ (e.g. $\{a, b\} = \{-1, 1\}$) with $|\mathcal{F}| \geq 2$.

- We say $\mathcal{F}$ *shatters* $x_{1:n} \in \mathcal{X}^n$ if $|\mathcal{F}(x_{1:n})| = 2^n$.
- Define also $s(\mathcal{F}, n) := \max_{x_{1:n} \in \mathcal{X}^n} |\mathcal{F}(x_{1:n})|$; this is known as the *shattering coefficient*.
- The *VC dimension* $\mathrm{VC}(\mathcal{F})$ is the largest integer $n$ such that some $x_{1:n}$ is shattered by $\mathcal{F}$, or $\infty$ if no such $n$ exists. Equivalently, $\mathrm{VC}(\mathcal{F}) = \sup\{n \in \mathbb{N} : s(\mathcal{F}, n) = 2^n\}$.

**Example.** Let $\mathcal{X} = \mathbb{R}$ and consider $\mathcal{F} = \{f_{a,b} : f_{a,b}(x) = \mathbb{1}_{[a,b)}(x) : a, b, \in \mathbb{R}\}$. Consider $n$ distinct points $x_1, \ldots, x_n$. These divide up the real line into $n + 1$ intervals $(-\infty, x_1], (x_1, x_2], \ldots, (x_{n-1}, x_n], (x_n, \infty)$. Now if $a$ and $a'$ are in the same interval, and $b$ and $b'$ are in the same interval, then $(f_{a,b}(x_i))_{i=1}^n = (f_{a',b'}(x_i))_{i=1}^n$. Thus every possible behaviour $(f_{a,b}(x_i))_{i=1}^n$ can be obtained by picking one of the $n + 1$ intervals for each of $a$ and $b$, so

$$s(\mathcal{F}, n) \leq (n + 1)^2.$$

Now consider $\mathrm{VC}(\mathcal{F})$. Any $x_{1:2}$ can be shattered, but with three points $x_1 < x_2 < x_3$, we can never have $f(x_1) = f(x_3) = 1$ but $f(x_2) = 0$. Thus $\mathrm{VC}(\mathcal{F}) = 2$.

It is a bit tedious to determine the shattering coefficient individually for each $\mathcal{F}$ and see whether it grows polynomially; we would like a more streamlined approach. Observe that in the previous example, we have $s(\mathcal{F}, n) \leq (n + 1)^{\mathrm{VC}(\mathcal{F})}$. The usefulness of the VC dimension, named after its inventors Vladmir Vapnik and Alexey Chervonenkis, is due to the remarkable fact that this is true more generally. The result below is known as the Sauer–Shelah lemma.

**Lemma 13** (Sauer–Shelah)**.** *Let $\mathcal{F}$ be a class with finite VC dimension $d$. Then*

$$s(\mathcal{F}, n) \leq \sum_{i=0}^{d} \binom{n}{i} \leq (n + 1)^d.$$

What is striking about this result is that whilst we know from the definition that for all $n > d$, $s(\mathcal{F}, n) < 2^n$, it is not immediately obvious that we cannot have $s(\mathcal{F}, n) = 2^n - 1$, or $s(\mathcal{F}, n) = 1.8^n$ for $n > d$. The result shows that beyond $d$ the growth of $s(\mathcal{F}, n)$ is radically different in that it is polynomial. The important consequence of this is that from Lemma 12 we have

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

**Proof of Lemma 13*.* This proof is non-examinable. We will prove the following stronger result. Fix $x_{1:n} \in \mathcal{X}^n$ and let $x_Q$ for any non-empty $Q \subseteq \{1, \ldots, n\}$ be $(x_{i_1}, \ldots, x_{i_{|Q|}})$ if $Q = \{i_1, \ldots, i_{|Q|}\}$. Then we claim that there are at least $|\mathcal{F}(x_{1:n})| - 1$ non-empty sets $Q \subseteq \{1, \ldots, n\}$ such that $\mathcal{F}$ shatters $x_Q$.

That this implies the statement of the lemma may be seen from the following reasoning. Take $x_{1:n}$ to be such that $|\mathcal{F}(x_{1:n})| = s(\mathcal{F}, n)$ and (for a contradiction) such that

$$|\mathcal{F}(x_{1:n})| > \sum_{i=0}^{d} \binom{n}{i} = \sum_{i=1}^{d} \binom{n}{i} + 1.$$

Then as the right-hand side is one more than the number of non-empty subsets $Q \subseteq \{1, \ldots, n\}$ of size at most $d$, we must have that some $x_Q$ with $Q > d$ is shattered by $\mathcal{F}$, but this contradicts $\mathrm{VC}(\mathcal{F}) = d$.

It remains to prove the claim, which we do by induction on $|\mathcal{F}(x_{1:n})|$. Wlog assume the functions in $\mathcal{F}$ map to $\{-1, 1\}$. The claim when $|\mathcal{F}(x_{1:n})| = 1$ is clearly true (the statement is vacuous in this case). Now take $k \geq 1$ and suppose the result is true for all $n \in \mathbb{N}$ and $x_{1:n} \in \mathcal{X}^n$ and $\mathcal{F}$ with $|\mathcal{F}(x_{1:n})| \leq k$. We will show the result holds at $k+1$. Take any $n \in \mathbb{N}$, $x_{1:n} \in \mathcal{X}^n$ and $\mathcal{F}$ with $|\mathcal{F}(x_{1:n})| = k+1$. Let $x_j$ be such that $\mathcal{F}_+ := \{f \in \mathcal{F} : f(x_j) = 1\}$ and $\mathcal{F}_- := \{f \in \mathcal{F} : f(x_j) = -1\}$ are both non-empty. Then $|\mathcal{F}_+(x_{1:n})|$ and $|\mathcal{F}_-(x_{1:n})|$ sum to $|\mathcal{F}(x_{1:n})| = k+1$ and they both are non-zero.

Let $\mathcal{X}_-$ and $\mathcal{X}_+$ be the sets of subvectors $x_Q$ that are shattered by $\mathcal{F}_-$ and $\mathcal{F}_+$ respectively. By the induction hypothesis, $|\mathcal{X}_-| + |\mathcal{X}_+| \geq k+1$. Clearly if $x_Q \in \mathcal{X}_- \cup \mathcal{X}_+$, $x_Q$ can be shattered by $\mathcal{F}$. Now none of the subvectors in $\mathcal{X}_- \cup \mathcal{X}_+$ can have $x_j$ as a component as then the subvector could not be shattered (each subfamily of hypotheses has all $f(x_j)$ taking the same value). But then when $x_Q \in \mathcal{X}_- \cap \mathcal{X}_+$, it must be the case that both $x_Q$ and $x_{Q \cup \{j\}}$ can be shattered by $\mathcal{F}$. Thus we see that the number of sets shattered by $\mathcal{F}$ is at least

$$|\mathcal{X}_- \cup \mathcal{X}_+| + |\mathcal{X}_- \cap \mathcal{X}_+| = |\mathcal{X}_-| + |\mathcal{X}_+| = k+1,$$

thereby completing the induction step. □

An important class of hypotheses $\mathcal{H}$ is based on functions that form a vector space. Let $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ be a vector space of functions e.g. consider $\mathcal{X} = \mathbb{R}^p$ and

$$\mathcal{F} = \{x \mapsto x^T \beta : \beta \in \mathbb{R}^p\}.$$

From $\mathcal{F}$ form a class of hypotheses

$$\mathcal{H} = \{h : h(x) = \mathrm{sgn}(f(x)) \text{ where } f \in \mathcal{F}\}. \tag{2.10}$$

The following Proposition bounds the VC dimension of $\mathcal{H}$.

**Proposition 14.** *Consider hypothesis class $\mathcal{H}$ given by* (2.10) *where $\mathcal{F}$ is a vector space of functions. Then*

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

*Proof.* Let $d = \dim(\mathcal{F}) + 1$ and take $x_{1:d} \in \mathcal{X}^d$. We need to show that $x_{1:d}$ cannot be shattered by $\mathcal{H}$. Consider the linear map $L : \mathcal{F} \to \mathbb{R}^d$ given by

$$L(f) = (f(x_1), \ldots, f(x_d)) \in \mathbb{R}^d.$$

The rank of $L$ is at most $\dim(\mathcal{F}) = d-1 < d$. Therefore, there must exist non-zero $\gamma \in \mathbb{R}^d$ orthogonal to everything in the image $L(\mathcal{F})$ i.e.

$$\sum_{i:\gamma_i>0} \gamma_i f(x_i) + \sum_{i:\gamma_i\leq 0} \gamma_i f(x_i) = 0 \quad \text{for all } f \in \mathcal{F}, \tag{2.11}$$

where wlog at least one component of $\gamma$ is strictly positive. Let $I_+ = \{i : \gamma_i > 0\}$ and $I_- = \{i : \gamma_i \leq 0\}$. Then it is not possible to have

$$h(x_i) = 1 \Rightarrow f(x_i) > 0 \text{ for all } i \in I_+,$$
$$h(x_i) = -1 \Rightarrow f(x_i) \leq 0 \text{ for all } i \in I_-,$$

(recall we are taking $\text{sgn}(0) := -1$) as otherwise the LHS of (2.11) would be strictly positive. Thus $x_{1:d}$ cannot be shattered so $\text{VC}(\mathcal{H}) \leq d$ as required. □

# 3 Computation for empirical risk minimisation

The results of the previous section have given us a good understanding of the theoretical properties of the ERM $\hat{h}$ corresponding to a given hypothesis class. We have not yet discussed whether $\hat{h}$ can be computed in practice, and how to do so; these questions are the topic of this chapter.

For a general hypothesis class $\mathcal{H}$, computation of the ERM $\hat{h}$ can be arbitrarily hard. Things simplify greatly if computing $\hat{h}$ may be equivalently phrased in terms of minimising a convex function over a convex set.

## 3.1 Basic properties of convex functions

Recall that a set $C \subseteq \mathbb{R}^d$ is *convex* if

$$x, y \in C \Rightarrow (1-t)x + ty \in C \qquad \text{for all } t \in (0,1).$$

In the following, let $C \subseteq \mathbb{R}^d$ be a convex set. A function $f : C \to \mathbb{R}$ is *convex* if

$$f\big((1-t)x + ty\big) \leq (1-t)f(x) + tf(y) \quad \text{for all } x, y \in C \text{ and } t \in (0,1).$$

Then $-f$ is a *concave* function. It is *strictly convex* if the inequality is strict for all $x, y \in \mathbb{R}^d$, $x \neq y$ and $t \in (0,1)$.

Convex functions exhibit a "local to global phenomenon": for example local minima are necessarily global minima. Indeed, if $x \in C$ is a local minimum, so for all $y \in C$, $f((1-t)x + ty) \geq f(x)$ for all $t$ sufficiently small, then by convexity

$$f(x) \leq f((1-t)x + ty) \leq (1-t)f(x) + tf(y),$$

so $f(x) \leq f(y)$ for all $y \in C$. On the other hand, non-convex functions can have many local minima whose objective values are far from the global minimum, which can make them very hard to optimise.

We collect together several useful properties of convex functions in the following proposition.

**Proposition 15.** *In the following, let $C \subseteq \mathbb{R}^d$ be a convex set and let $f : C \to \mathbb{R}$ be a convex function.*

*(i) Let $g : C \to \mathbb{R}$ be a convex function. Then if $a, b \geq 0$, $af + bg$ is a convex function.*

*(ii) Let $A \in \mathbb{R}^{d \times m}$ and $b \in \mathbb{R}^d$. Then $g : C \to \mathbb{R}$ given by $g(x) = f(Ax - b)$ is a convex function.*

*(iii) Suppose $f_\alpha : C \to \mathbb{R}^d$ is convex for all $\alpha \in I$ where $I$ is some index set, and define $g(x) := \sup_{\alpha \in I} f_\alpha(x)$. Then*

*(a) $D := \{x \in C : g(x) < \infty\}$ is convex and*

*(b) function $g$ restricted to $D$ is convex.*

*(iv) If $f$ is differentiable at $x \in int(C)$ then $f(y) \geq f(x) + \nabla f(x)^T (y - x)$ for all $y \in C$.*

*(v) If $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable then*

*(a) $f$ is convex iff. its Hessian matrix $H(x)$ at $x$ is positive semi-definite for all $x$,*

*(b) $f$ is strictly convex if $H(x)$ is positive definite for all $x$.*

## 3.2 Convex surrogates

In the classification setting, one problem with using misclassification loss is that the ERM optimisation can be intractable for many hypothesis classes. For example, taking $\mathcal{H}$ based on half-spaces, the ERM problem minimises over $\beta \in \mathbb{R}^p$ the following objective:

$$\sum_{i=1}^{n} \mathbb{1}_{\{\text{sgn}(X_i^T \beta) \neq Y_i\}} \approx \sum_{i=1}^{n} \mathbb{1}_{(-\infty, 0]}(Y_i X_i^T \beta)$$

(ignoring when $X_i^T \beta = 0$). The RHS is not convex and in fact not differentiable due to the indicator function. If $\mathbb{1}_{(-\infty,0]}$ above were somehow replaced with a convex function, we know from Proposition 15 (ii) that the resulting objective would be a convex function of $\beta$. The minimising $\hat{\beta}$ may still be able to deliver classification performance via $x \mapsto \text{sgn}(x^T \hat{\beta})$ that is comparable to that of the ERM provided the convex function is a sufficiently good approximation to an indicator function.

These considerations motivate the following changes to the classification framework that we have been studying thus far.

- Rather than performing ERM over a set of classifiers, let us consider a family $\mathcal{H}$ of functions $h : \mathcal{X} \to \mathbb{R}$. Each $h \in \mathcal{H}$ determines a classifier via $x \mapsto \text{sgn}(h(x))$.
- We will consider loss functions $\ell : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ of the form

$$\ell(h(x), y) = \phi(yh(x))$$