Mathematics of Machine Learning

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

1 Introduction

Consider a pair of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ 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 (two-class) *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 ℓ 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 ℓ 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* η 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 $h_0(x) = \pm 1$ 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 (classical) statistics approach to classification may attempt to model P_0 up to some unknown parameters, estimate these parameters (e.g. by maximum likelihood), 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

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

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

Non-examinable material is enclosed in *stars*.

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 variables with joint probability density function (pdf) $f_{Z,W}$ then the conditional pdf $f_{Z|W}$ of Z given W satisfies

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

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

(iv) **Best least squares predictor:** With the conditions in (iii) above, we have

$$\mathbb{E}(Z - f(W))^2 = \mathbb{E}\{Z - \mathbb{E}(Z \,|\, W)\}^2 + \mathbb{E}\{\mathbb{E}(Z \,|\, W) - f(W)\}^2. \tag{1.2}$$

Indeed, using the tower property,

$$\begin{aligned}\mathbb{E}(Z - f(W))^2 &= \mathbb{E}(Z - \mathbb{E}(Z \,|\, W) + \mathbb{E}(Z \,|\, W) - f(W))^2 \\ &= \mathbb{E}\{Z - \mathbb{E}(Z \,|\, W)\}^2 + \mathbb{E}\{\mathbb{E}(Z \,|\, W) - f(W)\}^2 \\ &\quad + 2\mathbb{E}\,\mathbb{E}[\{Z - \mathbb{E}(Z \,|\, W)\}\{\mathbb{E}(Z \,|\, W) - f(W)\} \,|\, W],\end{aligned}$$

but by 'taking out what is known', half the final term is

$$\mathbb{E}[\{\mathbb{E}(Z \,|\, W) - f(W)\} \underbrace{\mathbb{E}\{Z - \mathbb{E}(Z \,|\, W) \,|\, W\}}_{=0}] = 0.$$

Property (iv) shows that the $h : \mathcal{X} \to \mathbb{R}$ minimising $R(h)$ under squared loss is $h_0(x) = \mathbb{E}(Y \,|\, X = x)$.

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 $R(h) = \mathbb{P}(Y \neq h(X)) = \mathbb{E}\mathbb{P}(Y \neq h(X) \,|\, X)$, so $h_0(x)$ must minimise over $h(x)$

$$\begin{aligned}\mathbb{P}(Y \neq h(X) \,|\, X = x) &= \mathbb{P}(Y = 1, h(x) = -1 \,|\, X = x) + \mathbb{P}(Y = -1, h(x) = 1 \,|\, X = x) \\ &= \mathbb{P}(Y = 1 \,|\, X = x)\mathbb{1}_{\{h(x)=-1\}} + \mathbb{P}(Y = -1 \,|\, X = x)\mathbb{1}_{\{h(x)=1\}} \\ &= \mathbb{1}_{\{h(x)=-1\}}\eta(x) + \mathbb{1}_{\{h(x)=1\}}(1 - \eta(x)).\end{aligned}$$

When $\eta(x) > 1 - \eta(x)$ and so $\eta(x) > 1/2$, we must have $h_0(x) = 1$, and similarly when $\eta(x) < 1/2$, we must have $h_0(x) = -1$. If $\eta(x) = 1/2$, then the above is constant so any $h(x)$ minimises this. □

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_{new}. Define $h^* \in \operatorname*{arg\,min}_{h \in \mathcal{H}} R(h)$[2] and consider the decomposition

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

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 $Y_1, \ldots, Y_n$

[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 $\hat{h}$.

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 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*:

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

[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^*.

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 $t \geq 0$,

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

$\triangle$

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

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

- Also $W - c$ is sub-Gaussian with parameter σ for any deterministic $c \in \mathbb{R}$.

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

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

$$\begin{aligned}
\mathbb{E}e^{\alpha\varepsilon} = \frac{1}{2}(e^{-\alpha} + e^{\alpha}) &= \frac{1}{2}\left(\sum_{k=0}^{\infty}\frac{(-\alpha)^k}{k!} + \sum_{k=0}^{\infty}\frac{\alpha^k}{k!}\right) \\
&= \sum_{k=0}^{\infty}\frac{\alpha^{2k}}{(2k)!} \\
&\leq \sum_{k=0}^{\infty}\frac{\alpha^{2k}}{2^k k!} = e^{\alpha^2/2} \quad (\text{using } (2k)! \geq 2^k k!).
\end{aligned} \tag{2.3}$$

$\triangle$

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

Proof. Wlog we may assume $\mathbb{E}W = 0$. 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 $W - W' \overset{d}{=} -(W - W') \overset{d}{=} \varepsilon(W - W')$ where $\varepsilon \sim$ Rademacher with ε independent of (W, W'). (Here "$\overset{d}{=}$" 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 (2.3) 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.

The following proposition shows that somewhat analogously to how a linear combination of jointly Gaussian random variables is Gaussian, a linear combination of independent sub-Gaussian random variables is also sub-Gaussian.

Proposition 4. *Suppose $W_1, \ldots, W_n$ are independent and each W_i is sub-Gaussian with parameter σ_i. Then for $\gamma \in \mathbb{R}^n$, $\gamma^T W$ is sub-Gaussian with parameter $\left(\sum_i \gamma_i^2\sigma_i^2\right)^{1/2}$.*

Proof. Wlog we may assume $\mathbb{E}W_i = 0$.

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

□

As an application of the results above, suppose $W_1, \ldots, W_n$ are independent, and $a_i \leq W_i \leq b_i$ almost surely for all i. Then

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(W_i - \mathbb{E}W_i) \geq t\right) \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \quad \text{for } t \geq 0, \tag{2.4}$$

which is known as *Hoeffding's inequality*.

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 5. *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

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

Thus

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

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

2.2 Finite hypothesis classes

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

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

The assumption on ℓ 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, $\hat{R}(h)$ is an average of mean-zero i.i.d. quantities of the form $\ell(h(X_i), Y_i)$ taking values in $[0, M]$. 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 δ 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

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

and define

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

Then $\hat{h}^{\text{hist}}$ 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 misclassification loss) of $\hat{h}^{\text{hist}}$ is at most

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

[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 variables $(X_1, Y_1), \ldots, (X_n, Y_n)$. 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 variables 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

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

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

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

As ϵ was arbitrary, if ℓ takes values in $[0, M]$ we have

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

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 $f : \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n \to \mathbb{R}$ 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 $w_1 \in \mathcal{Z}_1, \ldots, w_n \in \mathcal{Z}_n, w_i' \in \mathcal{Z}_i$, and all $i = 1, \ldots, n$. Suppose random variables $W_1, \ldots, W_n$ taking values in $\mathcal{Z}_1, \ldots, \mathcal{Z}_n$ respectively 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 when $\mathcal{Z}_i = [a_i, b_i]$, taking $f(W_{1:n}) = \sum_i \{W_i - \mathbb{E}(W_i)\}/n$, 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

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

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

$$f(W_{1:n}) - \mathbb{E}f(W_{1:n}) = Z_n - Z_0 = \sum_{i=1}^n \underbrace{(Z_i - Z_{i-1})}_{D_i}; \tag{2.5}$$

the differences D_i play an analogous role to the individual independent random variables in the case of bounding sums. In fact, they are an example of a *martingale difference sequence*[4]:

Definition 2. A sequence of random variables $D_1, \ldots, D_n \in \mathbb{R}$ is a *martingale difference sequence* with respect to another sequence of random variables $W_0, \ldots, W_n$, if for $i = 1, \ldots, n$,

(i) $\mathbb{E}|D_i| < \infty$,

(ii) D_i is a function of $W_{0:i}$,

(iii) $\mathbb{E}(D_i \,|\, W_{0:(i-1)}) = 0$.

Example. If $D_1, \ldots, D_n$ are independent, mean zero, and satisfy (i), the sequence is a martingale difference sequence with respect to $c, D_1, \ldots, D_n$, for arbitrary constant c. $\triangle$

Example. The sequence $D_1, \ldots, D_n$ defined in (2.5) is a martingale difference sequence with respect to $c, W_1, \ldots, W_n$ for arbitrary constant c. That (ii) holds is clear. (i) certainly holds when f is bounded. That (iii) holds follows from the tower property of conditional expectation. $\triangle$

We are now in a position to prove a generalisation of Proposition 4 applicable to (weighted) averages of martingale differences.

Lemma 8. *Let $D_1, \ldots, D_n$ be a martingale difference sequence with respect to $W_0, \ldots, W_n$ such that*

$$\mathbb{E}(e^{\alpha D_i} \,|\, W_{0:(i-1)}) \leq e^{\alpha^2 \sigma_i^2/2} \quad i = 1, \ldots, n.$$

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

[4]For a more general and formal definition, see Part II *Stochastic Financial Models*.

Proof. We have

$$\begin{aligned}
\mathbb{E}\exp\Big(\alpha\sum_{i=1}^{n}\gamma_i D_i\Big) &= \mathbb{E}\,\mathbb{E}\Big\{\exp\Big(\alpha\sum_{i=1}^{n}\gamma_i D_i\Big)\,\Big|\,W_{0:(n-1)}\Big\}\\
&= \mathbb{E}\Big\{\exp\Big(\alpha\sum_{i=1}^{n-1}\gamma_i D_i\Big)\mathbb{E}(e^{\alpha\gamma_n D_n}\,|\,W_{0:(n-1)})\Big\}\\
&\le e^{\alpha^2\gamma_n^2\sigma_n^2/2}\mathbb{E}\exp\Big(\alpha\sum_{i=1}^{n-1}\gamma_i D_i\Big)\\
&\le \exp\Big(\frac{\alpha^2}{2}\sum_{i=1}^{n}\gamma_i^2\sigma_i^2\Big)\quad\text{(arguing inductively).}
\end{aligned}$$

□

The Azuma-Hoeffding inequality specialises 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 $W_0,\dots,W_n$. Suppose that 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 $A_i\le D_i\le B_i$, and $B_i-A_i\le L_i$ for a constant L_i. Then for $t\ge 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{n}D_i\ge t\right)\le\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}L_i^2}\right). \tag{2.6}$$

Proof. Conditional on $W_{0:(i-1)}$, A_i and B_i are constant. Thus we may apply Hoeffding's lemma (Lemma 3) conditionally on $W_{0:(i-1)}$ to obtain

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

The martingale difference sequence thus satisfies the hypotheses of Lemma 8. The sum $\sum_i D_i$ is sub-Gaussian with parameter $\sigma=\left(\sum_i L_i^2\right)^{1/2}/2$. 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. It is convenient to introduce $W_0\equiv w_0$ for an arbitrary constant w_0 and treat f as a function $f:\mathcal{Z}_0\times\cdots\times\mathcal{Z}_n$ where $\mathcal{Z}_0=\{w_0\}$.

Let $D_1,\dots,D_n$ be as in (2.5), so for $i=1,\dots,n$

$$D_i=\mathbb{E}\big(f(W_{0:n})\,|\,W_{0:i}\big)-\mathbb{E}\big(f(W_{0:n})\,|\,W_{0:(i-1)}\big).$$

Recall that $f(W_{0:n})-\mathbb{E}f(W_{0:n})=\sum_{i=1}^{n}D_i$.

Using the Azuma–Hoeffding inequality, it suffices to prove that $A_i\le D_i\le B_i$ almost surely where A_i and B_i are functions of $W_{0:(i-1)}$ satisfying $B_i-A_i\le L_i$ for all i, which we now do.

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

$$\begin{aligned} F_i : \mathcal{Z}_0 \times \cdots \times \mathcal{Z}_i &\to \mathbb{R} \\ (w_0, \dots, w_i) &\mapsto \mathbb{E}(f(W_{0:n}) \,|\, W_0 = w_0, \dots, W_i = w_i), \end{aligned}$$

so $D_i = F_i(W_{0:i}) - F_{i-1}(W_{0:(i-1)})$. Then define the random variables

$$\begin{aligned} A_i &:= \inf_{w_i \in \mathcal{Z}_i} F_i(W_{0:(i-1)}, w_i) - F_{i-1}(W_{0:(i-1)}) \\ B_i &:= \sup_{w_i \in \mathcal{Z}_i} F_i(W_{0:(i-1)}, w_i) - F_{i-1}(W_{0:(i-1)}), \end{aligned}$$

so A_i and B_i are functions of $W_{0:(i-1)}$. Then

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

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

$$\begin{aligned} B_i - A_i &= \sup_{w_i \in \mathcal{Z}_i} F_i(W_{0:(i-1)}, w_i) - \inf_{w_i \in \mathcal{Z}_i} F_i(W_{0:(i-1)}, w_i) \\ &= \sup_{w_i, w_i' \in \mathcal{Z}_i} \{F_i(W_{0:(i-1)}, w_i) - F_i(W_{0:(i-1)}, w_i')\} \\ &= \sup_{w_i, w_i' \in \mathcal{Z}_i} \Big\{ \mathbb{E}\big(f(W_{0:(i-1)}, w_i, W_{(i+1):n}) \,|\, W_{0:(i-1)}, W_i = w_i\big) \\ &\qquad - \mathbb{E}\big(f(W_{0:(i-1)}, w_i', W_{(i+1):n}) \,|\, W_{0:(i-1)}, W_i = w_i'\big) \Big\}. \end{aligned}$$

Now as the $W_{0:n}$ are independent, the distribution of $W_{(i+1):n}$ conditional on $W_{0:(i-1)}$ and that conditional on $W_{0:i}$ are identical, so

$$\begin{aligned} B_i - A_i &= \sup_{w_i, w_i' \in \mathcal{Z}_i} \big[\mathbb{E}\big\{ \underbrace{f(W_{0:(i-1)}, w_i, W_{(i+1):n}) - f(W_{0:(i-1)}, w_i', W_{(i+1):n})}_{\leq L_i} \,|\, W_{0:(i-1)} \big\}\big] \\ &\leq L_i. \end{aligned}$$

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

Note that from the proof above and that of the Azuma–Hoeffding inequality, we see that $f(W_{0:n})$ is a sub-Gaussian random variable with parameter $\sigma = (\sum_i L_i^2)^{1/2}/2$.

2.4 Rademacher complexity

Recall our setup: $\mathcal{H}$ is a (now possibly infinite) hypothesis class, ℓ 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

$$\mathcal{F} := \{(x, y) \mapsto -\ell(h(x), y) : h \in \mathcal{H}\}. \tag{2.7}$$

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 (2.7)).

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 variables 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*

$$\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) := \mathbb{E}\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(z_i)\right), \tag{2.8}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables. Given i.i.d. random variables $Z_1, \ldots, Z_n$ 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 (2.8) 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 obtain useful upper bounds of $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n}))$ that are uniform 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.

Below we summarise some useful properties of Rademacher complexity. Let $\mathcal{F}_1, \ldots, \mathcal{F}_m$ be classes of functions $f : \mathcal{Z} \to \mathcal{D} \subseteq \mathbb{R}$.

(i) If $\mathcal{G} = \{f_1 + f_2 : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$, then $\mathcal{R}_n(\mathcal{G}) = \mathcal{R}_n(\mathcal{F}_1) + \mathcal{R}_n(\mathcal{F}_2)$.

(ii) If $\mathcal{D} = [0, M]$, then $\mathcal{R}_n(\cup_{j=1}^m \mathcal{F}_j) \leq \max_{j=1,\ldots,m} \mathcal{R}_n(\mathcal{F}_j) + M\sqrt{2\log(m)/n}$.

We now turn to the proof of Theorem 10, 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

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

Theorem 11 (Generalisation bound based on Rademacher complexity)**.** *Let* $\mathcal{F} := \{(x, y) \mapsto \ell(h(x), y) : h \in \mathcal{H}\}$ *and suppose* ℓ *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

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

Further recall that viewing G as a function of $Z_1, \dots, Z_n$ 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

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

Hoeffding's inequality (or Bounded differences with the average function) also gives $\mathbb{P}(\hat{R}(h^*) - R(h^*) \ge t/2) \le \exp(-t^2 n/(2M^2))$. Also, noting that $\mathcal{R}_n(\mathcal{F}) = \mathcal{R}_n(-\mathcal{F})$, from Theorem 10, $\mathbb{E}G \le 2\mathcal{R}_n(\mathcal{F})$. 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 $|\mathcal{F}(z_{1:n})|$ of the function class $\mathcal{F}$ on n points $z_{1:n}$. Observe first that $|\mathcal{F}(z_{1:n})| = |\mathcal{H}(x_{1:n})|$ where $z_i = (x_i, y_i)$.

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

Proof. Let $d = |\mathcal{F}(z_{1:n})|$ and let $\mathcal{F}' := \{f_1, \dots, f_d\}$ 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 $j = 1, \dots, d$, 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 3 and Proposition 4, each W_j is sub-Gaussian with parameter $1/\sqrt{n}$. Thus we may apply Proposition 5 on the expected maximum of sub-Gaussian random variables to give the result. □

As each $h(x_i) \in \{-1, 1\}$, we always have $|\mathcal{H}(x_{1:n})| \le 2^n$. 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 $a \ne b$ (e.g. $\{a, b\} = \{-1, 1\}$) with $|\mathcal{F}| \ge 2$.

- 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 ∞ 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 $x_1, \dots, x_n$. These divide up the real line into $n+1$ intervals $(-\infty, x_1], (x_1, x_2], \dots, (x_{n-1}, x_n], (x_n, \infty)$. 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

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

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 $s(\mathcal{F}, n) \le (n+1)^{\mathrm{VC}(\mathcal{F})}$. 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*

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

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

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

**Proof of Lemma 13*.* We will prove the following stronger result. Fix $x_{1:n} \in \mathcal{X}^n$ and let x_Q for any non-empty $Q = \{i_1, \dots, i_{|Q|}\} \subseteq \{1, \dots, n\}$ be $(x_{i_1}, \dots, x_{i_{|Q|}})$. Then we claim that there are at least $|\mathcal{F}(x_{1:n})| - 1$ non-empty sets $Q \subseteq \{1, \dots, n\}$ 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 $|\mathcal{F}(x_{1:n})| = s(\mathcal{F}, n)$. As $\mathrm{VC}(\mathcal{F}) = d$, by definition no x_Q with $|Q| > d$ can be shattered, so from the claim,

$$|\mathcal{F}(x_{1:n})| - 1 \le (\#\text{ of shattered sets } x_Q) \le \sum_{i=1}^{d} \binom{n}{i}.$$

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 (which is possible as $|\mathcal{F}(x_{1:n})| \geq 2$). Then

$$|\mathcal{F}_+(x_{1:n})| + |\mathcal{F}_-(x_{1:n})| = |\mathcal{F}(x_{1:n})| = k+1.$$

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 $\mathcal{F} \supset \mathcal{F}_1, \mathcal{F}_+$. 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 $x_{Q\cup\{j\}}$ (which are distinct) can be shattered by $\mathcal{F}$. Also x_j itself is shattered by $\mathcal{F}$. Thus we see that the number of sets shattered by $\mathcal{F}$ is at least

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

thereby completing the induction step. □

Example. Let $\mathcal{X} = \mathbb{R}^p$ and consider $\mathcal{F} = \{\mathbb{1}_A : A \in \mathcal{A}\}$ where $\mathcal{A} = \big\{\prod_{j=1}^p (-\infty, a_j] : a_1, \ldots, a_p \in \mathbb{R}\big\}$. To compute $\mathrm{VC}(\mathcal{F})$, first note that the set of standard basis vectors $e_1, \ldots, e_p \in \mathbb{R}^p$ is shattered as for any $I \subseteq \{1, \ldots, p\}$, we may take $a_j = 1$ if $j \in I$ and $a_j = 0$ otherwise; then

$$e_j \in \prod_{k=1}^p (-\infty, a_k] \quad \Leftrightarrow \quad j \in I.$$

Next take $x_1, \ldots, x_{p+1} \in \mathbb{R}^p$. For each coordinate $j = 1, \ldots, p$, let $J_j = \{k : x_{kj} = \max_l x_{lj}\}$. Then there must be some x_{k^*} such that k^* is not a unique element of one of the J_j. But then for each $j = 1, \ldots, p$, there exists some x_{k_j} such that $x_{k_j j} \geq x_{k^* j}$, so for $f \in \mathcal{F}$ we can never have $f(x_{k^*}) = 0$ and $f(x_k) = 1$ for all $k \neq k^*$. Thus $\mathrm{VC}(\mathcal{F}) = p$. △

An important class of hypotheses $\mathcal{H}$ is based on functions that form a vector space. Let $\mathcal{F}$ be a vector space of functions $f : \mathcal{X} \to \mathbb{R}$, 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

$$\mathcal{H} = \{h : h(x) = \mathrm{sgn}(f(x)) \text{ where } f \in \mathcal{F}\}. \tag{2.9}$$

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

Proposition 14. *Consider hypothesis class $\mathcal{H}$ given by* (2.9) *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

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

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.

$$\sum_{i:\gamma_i>0} \gamma_i f(x_i) + \sum_{i:\gamma_i\leq 0} \gamma_i f(x_i) = 0 \quad \text{for all } f \in \mathcal{F}, \tag{2.10}$$

where wlog at least one component of γ 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 $\operatorname{sgn}(0) := -1$) as otherwise the LHS of (2.10) would be strictly positive. Thus $x_{1:d}$ cannot be shattered so $\mathrm{VC}(\mathcal{H}) \leq d - 1$ 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 sets

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

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

The intersection of an arbitrary collection of convex sets is convex, so if for each $\alpha \in I$, the set $C_\alpha \in \mathbb{R}^d$ is convex, then $\cap_{\alpha\in I} C_\alpha$ is convex (see Example Sheet 2).

Definition 5.

- For a set $S \subseteq \mathbb{R}^d$, the *convex hull* $\operatorname{conv} S$ is the intersection of all convex sets containing S.

- A point $v \in \mathbb{R}^d$ is a *convex combination* of $v_1, \dots, v_m \in \mathbb{R}^d$ if

$$v = \alpha_1 v_1 + \cdots + \alpha_m v_m$$

where $\alpha_1, \dots, \alpha_m \geq 0$ and $\sum_{j=1}^m \alpha_j = 1$.

Lemma 15. *For $S \subseteq \mathbb{R}^d$, $v \in \operatorname{conv} S$ if and only if v is a convex combination of some set of points in S.*

Proof. Let D be the set of all convex combinations of sets of points from S. We want to show $D \supseteq \operatorname{conv} S$ and $D \subseteq \operatorname{conv} S$. Showing the former is a task on Example Sheet 2; we show the latter relation $D \subseteq \operatorname{conv} S$.

Now intersections of convex sets are convex, so $\operatorname{conv} S$ is convex. Thus clearly a convex combination of any $v_1, v_2 \in S$ is in $\operatorname{conv} S$. Suppose then that for $m \geq 2$, any convex combination of m points from S is in $\operatorname{conv} S$. Take $v_1, \dots, v_{m+1} \in S$ and $\alpha_1, \dots, \alpha_{m+1} \geq 0$ with $\sum_{j=1}^{m+1} \alpha_j = 1$. Consider $v = \sum_{j=1}^{m+1} v_j \alpha_j$. If $\alpha_{m+1} = 1$, $v = v_{m+1} \in S \subseteq \operatorname{conv} S$. Otherwise, writing $t = \sum_{j=1}^m \alpha_j$, we have $t > 0$ and $\alpha_{m+1} = 1 - t$ so

$$v = t\Big(\underbrace{\frac{\alpha_1}{t} v_1 + \cdots + \frac{\alpha_m}{t} v_m}_{\substack{\in \operatorname{conv} S \text{ by the} \\ \text{induction hypothesis}}}\Big) + (1-t) v_{m+1} \in \operatorname{conv} S. \qquad \square$$

Lemma 16. *Let $S \subseteq \mathbb{R}^d$. For any linear map $L : \mathbb{R}^d \to \mathbb{R}^n$, $\operatorname{conv} L(S) = L(\operatorname{conv} S)$.*

Proof. $u \in \operatorname{conv} L(S)$ iff. there exist $v_1, \dots, v_m \in S$ and $\alpha_1, \dots, \alpha_m \geq 0$ such that $\sum_{j=1}^m \alpha_j = 1$ and

$$u = \sum_j \alpha_j L(v_j).$$

But the RHS is $L\Big(\sum_j \alpha_j v_j\Big) \in L(\operatorname{conv} S)$ and $u \in L(\operatorname{conv} S)$ iff. u takes this form. $\square$

3.2 Basic properties of convex functions

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

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

Then $-f$ is a *concave* function. It is *strictly convex* if the inequality is strict for all $x, y \in C$, $x \neq y$ 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

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

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

New convex functions from old:

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

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

(iii) Suppose $f_\alpha : C \to \mathbb{R}$ 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.

Consequences of convexity:

(iv) If f is differentiable at $x \in int(C)$ then $f(y) \geq f(x) + \nabla f(x)^T (y - x)$ for all $y \in C$. In particular, $\nabla f(x) = 0 \Rightarrow x$ minimises f.

(v) If f is a strictly convex function, then any minimiser is unique.

(vi) If $C = \operatorname{conv} D$, then $\sup_{x \in C} f(x) = \sup_{x \in D} f(x)$.

Checking convexity:

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

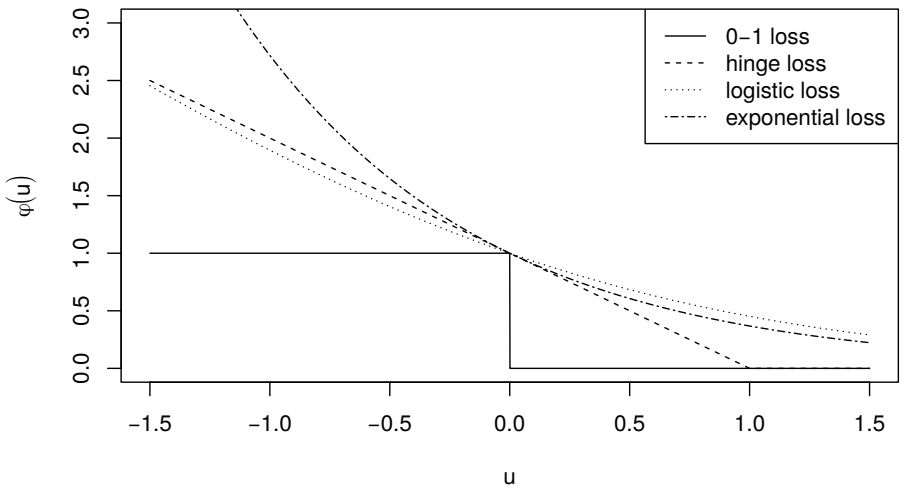
(a) f is convex iff. its Hessian matrix $H(x)$ at x is positive semi-definite for all x,

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

3.3 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:

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



(ignoring when $X_i^T\beta = 0$). The RHS is not convex and in fact not continuous due to the indicator function. If $\mathbb{1}_{(-\infty,0]}$ above were somehow replaced with a convex function, we know from Proposition 17 (i) & (ii) that the resulting objective would be a convex function of β. 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

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

  and $\phi : \mathbb{R} \to [0, \infty)$ is convex. We will refer to the corresponding risk as the ϕ-risk and denote it by R_ϕ. Note formally we will be taking $\mathcal{Y} = \mathbb{R}$ (even though the data $(Y_i)_{i=1}^n$ are in $\{-1, 1\}$).

Common choices of ϕ include the following:

- **Hinge loss:** $\phi(u) = \max(1 - u, 0)$.
- **Exponential loss:** $\phi(u) = e^{-u}$.
- **Logistic loss:** $\phi(u) = \log_2(1 + e^{-u}) = \log(1 + e^{-u})/\log(2)$.

For the strategy of using a surrogate loss to be useful, ERM with the surrogate loss should hopefully mimic using misclassification loss. For example, we would ideally like the $h_{\phi,0}$ that minimises R_ϕ (assuming it exists) to be such that $x \mapsto \text{sgn}(h_{\phi,0}(x))$ is (equivalent to)

the Bayes classifier $x \mapsto \operatorname{sgn}(\eta(x) - 1/2)$. To understand when this is the case, we introduce the following definitions.

The *conditional ϕ-risk* of h is

$$\mathbb{E}(\phi(Yh(X))|X = x) = \eta(x)\phi(h(x)) + (1 - \eta(x))\phi(-h(x)),$$

where recall $\eta(x) = \mathbb{P}(Y = 1|X = x)$. It will be helpful to consider this in terms of a generic conditional probability $\eta \in [0, 1]$ and generic value $\alpha \in \mathbb{R}$ of $h(x)$. We thus introduce

$$C_\eta(\alpha) := \eta\phi(\alpha) + (1 - \eta)\phi(-\alpha).$$

The following definition encapsulates our idea of $\operatorname{sgn} \circ h_{\phi,0}$ achieving the optimal Bayes misclassification risk, but also allows for the possibility that $\inf_h R_\phi(h)$ is not attained.

Definition 6. We say $\phi : \mathbb{R} \to [0, \infty)$ is *classification calibrated* if for any $\eta \in [0, 1]$ with $\eta \neq 1/2$,

$$\inf_{\alpha \in \mathbb{R}} C_\eta(\alpha) < \inf_{\alpha : \alpha(2\eta - 1) \leq 0} C_\eta(\alpha).$$

In words, the equation above says that the infimal generic conditional ϕ-risk is strictly less than the infimum where α (playing the role of $h(x)$) is forced to disagree in sign with the Bayes classifier. The following result tells us when the favourable case of classification calibration occurs for convex ϕ.

Theorem 18. *Let ϕ be convex. Then $\phi : \mathbb{R} \to [0, \infty)$ is classification calibrated if it is differentiable at 0 and $\phi'(0) < 0$.*

Proof. Note that C_η is convex and differentiable at 0 with

$$C'_\eta(0) = (2\eta - 1)\phi'(0).$$

Suppose $\eta > 1/2$, so $C'_\eta(0) < 0$. Then from Proposition 17 (iv),

$$C_\eta(\alpha) \geq C_\eta(0) + C'_\eta(0)\alpha \geq C_\eta(0)$$

for $\alpha \leq 0$. Also as

$$0 > C'_\eta(0) = \lim_{\alpha \downarrow 0} \frac{C_\eta(\alpha) - C_\eta(0)}{\alpha},$$

for some $\alpha^* > 0$ we have $C_\eta(\alpha^*) < C_\eta(0)$. Similarly when $\eta < 1/2$, there exists some $\alpha^* < 0$ with $C_\eta(\alpha^*) < C_\eta(0)$. Thus in both cases $\inf_{\alpha \in \mathbb{R}} C_\eta(\alpha) \leq C_\eta(\alpha^*) < \inf_{\alpha : \alpha(2\eta - 1) \leq 0} C_\eta(\alpha)$. □

We thus see that the popular choices of ϕ above are all classification calibrated.

3.4 Rademacher complexity revisited

One remaining issue is whether we can obtain guarantees on when the generalisation error measured in terms of ϕ-risk is small. Theorem 11 gives us a bound in terms of the Rademacher complexity of

$$\mathcal{F} = \{(x, y) \mapsto \phi(yh(x)) : h \in \mathcal{H}\}.$$

Our bounds for $\mathcal{R}_n(\mathcal{F})$ involving shattering coefficients and VC dimension relied heavily on the use of misclassification loss. We will need a different approach here. One useful step would be to relate $\mathcal{R}_n(\mathcal{F})$ to $\mathcal{R}_n(\mathcal{H})$ which is potentially simpler to handle. The following result, which is sometimes known as the contraction lemma, helps in this regard.

Lemma 19 (Contraction lemma)**.** *Let $r = \sup_{x\in\mathcal{X}, h\in\mathcal{H}} |h(x)|$. Suppose there exists $L \geq 0$ with $|\phi(u) - \phi(u')| \leq L|u - u'|$ for all $u, u' \in [-r, r]$, so ϕ is* Lipschitz with constant L *on $[-r, r]$. Then $\mathcal{R}_n(\mathcal{F}) \leq L\mathcal{R}_n(\mathcal{H})$.*

**Proof*.* Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X}\times\{-1, 1\}$ and let $\varepsilon_1, \ldots, \varepsilon_n$ be a sequence of i.i.d. Rademacher random variables. Then writing $z_i = (x_i, y_i)$, we have

$$\mathcal{R}(\mathcal{F}(z_{1:n})) = \mathbb{E}\left(\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i\phi(y_i h(x_i))\right).$$

Let us consider $z_{1:n}$ as fixed and, for any i, write ε_{-i} for the sequence $\varepsilon_{1:n}$ with ε_i removed. We claim that for any (suitable) function $A : \mathcal{H} \times \{-1, 1\}^{n-1}$,

$$\mathbb{E}\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\varepsilon_i\phi(y_i h(x_i)) + A(h, \varepsilon_{-i})\right) \leq \mathbb{E}\sup_{h\in\mathcal{H}}\left(\frac{L}{n}\varepsilon_i h(x_i) + A(h, \varepsilon_{-i})\right). \tag{3.1}$$

Applying this with $i = 1$ and

$$A(h, \varepsilon_{-1}) = \frac{1}{n}\sum_{i=2}^{n} \varepsilon_i\phi(y_i h(x_i)),$$

we get

$$\mathbb{E}\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\varepsilon_1\phi(y_1 h(x_1)) + \frac{1}{n}\sum_{i=2}^{n}\varepsilon_i\phi(y_i h(x_i))\right) \leq \mathbb{E}\sup_{h\in\mathcal{H}}\left(\frac{L}{n}\varepsilon_1 h(x_1) + \frac{1}{n}\sum_{i=2}^{n}\varepsilon_i\phi(y_i h(x_i))\right). \tag{3.2}$$

Next applying (3.1) with $i = 2$ and

$$A(h, \varepsilon_{-2}) = \frac{1}{n}\sum_{i=3}^{n} \varepsilon_i\phi(y_i h(x_i)) + \frac{L}{n}\varepsilon_1 h(x_1),$$

we get that the RHS of (3.2) is at most

$$\mathbb{E}\sup_{h\in\mathcal{H}}\left(\frac{L}{n}\sum_{i=1}^{2}\varepsilon_i h(x_i) + \frac{1}{n}\sum_{i=3}^{n}\varepsilon_i\phi(y_i h(x_i))\right).$$

Continuing this argument yields the result. It remains to prove the claim, which we do now. We have

$$\begin{aligned}
&\mathbb{E}\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\varepsilon_i\phi(y_ih(x_i))+A(h,\varepsilon_{-i})\,\Big|\,\varepsilon_{-i}\right)\\
=&\frac{1}{2n}\big[\sup_{h\in\mathcal{H}}\{\phi(y_ih(x_i))+nA(h,\varepsilon_{-i})\}+\sup_{h\in\mathcal{H}}\{-\phi(y_ih(x_i))+nA(h,\varepsilon_{-i})\}\big]\\
=&\frac{1}{2n}\big[\sup_{h,g\in\mathcal{H}}\{\underbrace{\phi(y_ih(x_i))-\phi(y_ig(x_i))}_{\leq L|h(x_i)-g(x_i)|}+nA(h,\varepsilon_{-i})+nA(g,\varepsilon_{-i})\}\big].
\end{aligned}$$

But by symmetry,

$$\begin{aligned}
&\sup_{h,g\in\mathcal{H}}\{L|h(x_i)-g(x_i)|+nA(h,\varepsilon_{-i})+nA(g,\varepsilon_{-i})\}\\
=&\sup_{h,g\in\mathcal{H}}[L\{h(x_i)-g(x_i)\}+nA(h,\varepsilon_{-i})+nA(g,\varepsilon_{-i})]\\
=&\sup_{h\in\mathcal{H}}\{Lh(x_i)+nA(h,\varepsilon_{-i})\}+\sup_{h\in\mathcal{H}}\{-Lh(x_i)+nA(h,\varepsilon_{-i})\}.
\end{aligned}$$

Hence

$$\mathbb{E}\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\varepsilon_i\phi(y_ih(x_i))+A(h,\varepsilon_{-i})\,\Big|\,\varepsilon_{-i}\right)\leq\mathbb{E}\sup_{h\in\mathcal{H}}\left(\frac{L}{n}\varepsilon_ih(x_i)+A(h,\varepsilon_{-i})\,\Big|\,\varepsilon_{-i}\right)$$

Taking expectations proves the claim. □

Corollary 20. *Consider the setup of Lemma 19 and suppose r is finite. Suppose ϕ is non-increasing and let $M=\phi(-r)$. Then with probability at least $1-\delta$,*

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

In order for the result above to be applicable when ϕ is e.g. one of the convex surrogates discussed earlier, we need $\mathcal{H}$ to be such that r is finite so M is finite. This will not hold for our example where $\mathcal{X}=\mathbb{R}^p$ of

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

However, if we constrain the norm of the β and $\mathcal{X}$ is a bounded subset of $\mathbb{R}^p$, we can achieve this.

3.5 ℓ_2-constraint

Suppose $\mathcal{X}=\{x\in\mathbb{R}^p:\|x\|_2\leq C\}$ and consider

$$\mathcal{H}=\{x\mapsto x^T\beta:\beta\in\mathbb{R}^p\text{ and }\|\beta\|_2\leq\lambda\}\tag{3.3}$$

for $\lambda > 0$. Then we have that for any $x_{1:n} \in \mathcal{X}^n$,

$$\begin{aligned}\hat{\mathcal{R}}(\mathcal{H}(x_{1:n})) &= \frac{1}{n}\mathbb{E}\left(\sup_{\beta:\|\beta\|_2\leq\lambda}\sum_{i=1}^n \varepsilon_i x_i^T\beta\right)\\ &\leq \frac{\lambda}{n}\mathbb{E}\Big\|\sum_{i=1}^n \varepsilon_i x_i\Big\|_2 \quad \text{(Cauchy–Schwarz)}\\ &\leq \frac{\lambda}{n}\left(\mathbb{E}\Big\|\sum_{i=1}^n \varepsilon_i x_i\Big\|_2^2\right)^{1/2},\end{aligned}$$

where the last inequality follows due to concavity of $\sqrt{\cdot}$ and Jensen's inequality. Now for $i \neq j$, $\mathbb{E}(\varepsilon_i x_i^T x_j \varepsilon_j) = 0$, so

$$\mathbb{E}\Big\|\sum_{i=1}^n \varepsilon_i x_i\Big\|_2^2 = \sum_{i=1}^n \|x_i\|_2^2 \leq nC^2.$$

Thus

$$\hat{\mathcal{R}}(\mathcal{H}(x_{1:n})) \leq \lambda C/\sqrt{n}$$

Furthermore

$$\sup_{x\in\mathcal{X},h\in\mathcal{H}} |h(x)| = \sup_{x:\|x\|_2\leq C,\beta:\|\beta\|_2\leq\lambda} x^T\beta = \lambda C.$$

Example. Take ϕ to be the hinge loss and $\mathcal{H}$ given by (3.3). Then from Corollary 20, with probability at least $1-\delta$,

$$R_\phi(\hat{h}) - R_\phi(h^*) \leq \frac{2\lambda C}{\sqrt{n}} + (\lambda C + 1)\sqrt{\frac{2\log(2/\delta)}{n}}.$$

△

3.6 *Kernel machines*

Consider the optimisation problem solved by the empirical risk minimiser with hypothesis class

$$\mathcal{H} = \left\{x \mapsto \sum_{j=1}^d \beta_j\varphi_j(x) \,:\, \beta\in\mathbb{R}^d \text{ and } \|\beta\|_2 \leq \lambda\right\}$$

and data $(X_i, Y_i)_{i=1}^n$. Consider now a Lagrangian formulation of the objective, given by

$$\operatorname*{arg\,min}_{\beta:\beta\in\mathbb{R}^d} \frac{1}{n}\ell(Y_i, (\Phi\beta)_i) + \gamma\|\beta\|_2^2. \tag{3.4}$$

Here $\gamma > 0$ is a Lagrange multiplier, and matrix $\Phi \in \mathbb{R}^{n\times d}$ has $\Phi_{ij} = \varphi_j(x_i)$ (note we have squared the ℓ_2-norm in the original constraint). If d is large, this can be a challenging optimisation problem to solve.

Consider however the projection $P \in \mathbb{R}^{d\times d}$ onto the row space of Φ, and note that $\Phi\beta = \Phi P\beta$. Meanwhile, we have that

$$\|\beta\|_2^2 = \|P\beta\|_2^2 + \|(I-P)\beta\|_2^2.$$

We conclude that any minimiser, $\hat{\beta}$, of (3.4) must satisfy $\hat{\beta} = P\hat{\beta}$, that is $\hat{\beta}$ must be in the row space of Φ. This means that we may write $\hat{\beta} = \Phi^T\hat{\alpha}$ for some $\hat{\alpha} \in \mathbb{R}^n$. Now let function $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be given by

$$k(x,x') = \sum_{j=1}^{d} \varphi_j(x)\varphi_j(x'), \tag{3.5}$$

and let $K \in \mathbb{R}^{n\times n}$ be the matrix with ijth entry $K_{ij} = k(x_i, x_j)$, so $K = \Phi\Phi^T$. Substituting $\beta = \Phi^T\alpha$ into (3.4), we see that $\hat{\alpha}$ minimises

$$\frac{1}{n}\ell(Y_i, (K\alpha)_i) + \gamma\alpha^T K\alpha \tag{3.6}$$

over $\alpha \in \mathbb{R}^n$. Note that the empirical risk minimiser evaluated at a point $x \in \mathcal{X}$ is given by

$$\sum_{j=1}^{d} \varphi_j(x)\hat{\beta}_j = \sum_{j=1}^{d} \varphi_j(x)(\Phi^T\hat{\alpha})_j = \sum_{i=1}^{n} k(x,x_i)\hat{\alpha}_i.$$

What is remarkable is that whilst the optimisation in (3.4) involves d variables, we have shown this is equivalent to (3.6) which involves n variables: this is a substantial simplification if $d \gg n$. In fact, these arguments can be generalised to the case where $d = \infty$.[5]

The function k in (3.5) is known as a *(positive-definite) kernel*. For certain families of functions $(\varphi_j)_{j=1}^d$, this can be computed very fast. For example, consider $\mathcal{X} \in \mathbb{R}^p$ and

$$\begin{aligned}(\varphi_1(x), \ldots, \varphi_d(x)) = (&x_1, \ldots, x_p,\\ &x_1x_1, \ldots, x_1x_p,\\ &x_2x_1, \ldots, x_2x_p,\\ &x_px_1, \ldots, x_px_p),\end{aligned}$$

so $d = p + p^2$; note that some functions φ_j occur twice. Naive computation of the resulting $k(x,x')$ would require summing over $O(p^2)$ terms. However note that

$$k(x,x') = \sum_{j=1}^{p} x_jx_j' + \sum_{j=1}^{p}\sum_{k=1}^{p} x_jx_kx_j'x_k' = \left(\sum_{j=1}^{p} x_jx_j' + \frac{1}{2}\right)^2 - \frac{1}{4},$$

which may be found using $O(p)$ computational operations.

3.7 ℓ_1-constraint

The ℓ_1-norm of a vector u is $\|u\|_1 := \sum_i |u_i|$ and the ℓ_∞-norm is $\|u\|_\infty := \max_i |u_i|$. Suppose now that $\mathcal{X} = \{x \in \mathbb{R}^p : \|x\|_\infty \le C\}$ and consider

$$\mathcal{H} = \{x \mapsto x^T\beta : \beta \in \mathbb{R}^p \text{ and } \|\beta\|_1 \le \lambda\}.$$

To compute the Rademacher complexity of $\mathcal{H}$, we make use of the following.

[5]This section just scratches the surface of the topic known as kernel machines: see the Part III course *Modern Statistical Methods* to learn more.

Lemma 21. *For any $A \subseteq \mathbb{R}^n$, $\hat{\mathcal{R}}(A) = \hat{\mathcal{R}}(\operatorname{conv} A)$.*

Proof. See example sheet. □

To use this, observe that if β has $\|\beta\|_1 = \lambda$, then writing

$$\beta = \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{\lambda} \operatorname{sgn}(\beta_j) e_j,$$

we see that $\beta \in \operatorname{conv} S$ where $S = \cup_{j=1}^{p} \{\lambda e_j, -\lambda e_j\}$ and e_j is the jth standard basis vector. Next if $\|\beta\|_1 \leq \lambda$, then

$$\beta = \frac{\lambda + \|\beta\|_1}{2\lambda} \underbrace{\frac{\lambda}{\|\beta\|_1}\beta}_{\in \operatorname{conv} S} + \frac{\lambda - \|\beta\|_1}{2\lambda} \underbrace{\frac{(-\lambda)}{\|\beta\|_1}\beta}_{\in \operatorname{conv} S} \in \operatorname{conv} S$$

as $\operatorname{conv} S$ is convex. Then given $x_1, \ldots, x_n$, let $L : \mathbb{R}^p \to \mathbb{R}^n$ be the linear map given by

$$L(\beta) = (x_1^T \beta, \ldots, x_n^T \beta)^T.$$

Then $\mathcal{H}(x_{1:n}) = L(\operatorname{conv} S) = \operatorname{conv} L(S)$ from Lemma 16. Thus from Lemma 21 we have

$$\begin{aligned} \hat{\mathcal{R}}(\mathcal{H}(x_{1:n})) &= \hat{\mathcal{R}}(L(S)) \\ &= \frac{\lambda}{n} \mathbb{E}\left(\max_{j=1,\ldots,p} \left| \sum_{i=1}^{n} \varepsilon_i x_{ij} \right| \right) \end{aligned}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables. Now each $\pm \sum_i \varepsilon_i x_{ij}$ is sub-Gaussian with parameter

$$\left(\sum_{i=1}^{n} x_{ij}^2 \right)^{1/2} \leq C\sqrt{n} \quad \text{(Proposition 4)}.$$

Thus from Proposition 5 we have

$$\hat{\mathcal{R}}(\mathcal{H}(x_{1:n})) \leq \frac{\lambda}{n} \times C\sqrt{n} \times \sqrt{2 \log |S|} = \frac{\lambda C}{\sqrt{n}} \sqrt{2 \log(2p)}.$$

Also

$$\sup_{x \in \mathcal{X}, h \in \mathcal{H}} |h(x)| = \sup_{x : \|x\|_\infty \leq C, \beta : \|\beta\|_1 \leq \lambda} x^T \beta = \lambda C.$$

Example. Take ϕ to be the hinge loss and $\mathcal{H}$ as above. Suppose $\mathcal{X} = [-1, 1]^p$. Then from Corollary 20, with probability at least $1 - \delta$,

$$R_\phi(\hat{h}) - R_\phi(h^*) \leq 2\lambda \sqrt{\frac{2 \log(2p)}{n}} + (\lambda + 1) \sqrt{\frac{2 \log(2/\delta)}{n}}.$$

In contrast, with $\mathcal{H}$ given by the ℓ_2-constraint (3.3) we would have a bound of order $\lambda\sqrt{p/n}$. Some notable differences are as follows.

- The dimension p contributes a factor of order $\sqrt{\log(p)}$ in the ℓ_1 constraint case versus $\sqrt{p}$ is the ℓ_2 constraint case.
- Write $\mathcal{H}_1$ and $\mathcal{H}_2$ for the ℓ_1 and ℓ_2 constrained hypothesis classes with norm constraints λ_1 and λ_2 respectively. Suppose that $\beta^0 \in \mathbb{R}^p$ is such that $h^0 : x \mapsto x^T\beta^0$ minimises R_ϕ over $\{x \mapsto x^T\beta : \beta \in \mathbb{R}^p\}$.
  - If
  $$\beta^0 = \left(\frac{1}{\sqrt{p}}, \ldots, \frac{1}{\sqrt{p}}\right)^T,$$
  in order that $\beta^0 \in \mathcal{H}_1, \mathcal{H}_2$, we require $\lambda_1 \leq \sqrt{p}$ and $\lambda_2 \geq 1$. These choices give excess risk bounds of order (treating δ as a constant)
  $$\ell_1 : \quad \sqrt{\frac{p \log p}{n}}, \qquad \ell_2 : \quad \sqrt{\frac{p}{n}}.$$
  - If
  $$\beta^0 = \Big(\underbrace{\frac{1}{\sqrt{s}}, \ldots, \frac{1}{\sqrt{s}}}_{s \text{ of these}}, 0, \ldots, 0\Big)^T,$$
  the corresponding risk bounds would be
  $$\ell_1 : \quad \sqrt{\frac{s \log p}{n}}, \qquad \ell_2 : \quad \sqrt{\frac{p}{n}}.$$

  Conclusion: If every predictor is equally important, the ℓ_2 hypothesis class will tend to perform better. If only the s predictors are important and s is small, the ℓ_1 approach can perform well.

△

3.8 Projections on to convex sets

Empirical risk minimisation (with a convex surrogate) over the ℓ_2 and ℓ_1 constraint classes discussed above involves minimising a convex function subject to the minimiser being in a convex set. In order to perform this optimisation it will be helpful to project points on to convex constraint sets.

Proposition 22. *Let $C \subseteq \mathbb{R}^d$ be a closed convex set. Then for each $x \in \mathbb{R}^d$, the minimiser of $\|x - z\|_2$ over $z \in C$ exists and is unique. Moreover writing*

$$\pi_C(x) = \text{argmin}_{z \in C} \|x - z\|_2,$$

we have that for all $x \in \mathbb{R}^d$,

$$(x - \pi_C(x))^T(z - \pi_C(x)) \leq 0 \quad \textit{for all } z \in C, \tag{3.7}$$

$$\|\pi_C(x) - \pi_C(z)\|_2 \leq \|x - z\|_2 \quad \textit{for all } z \in \mathbb{R}^d. \tag{3.8}$$

Proof. **Existence**: Let $\mu = \inf_{z\in C} \|x - z\|_2$. Write $B = \{w : \|w - x\|_2 \le \mu + 1\}$. Then

$$\inf_{z\in C} \|x - z\|_2 = \inf_{z\in C\cap B} \|x - z\|_2,$$

and the RHS is an infimum of a continuous function on a closed and bounded set, so the infimum is achieved at $\pi = \pi_C(x)$, say.
Uniqueness: For each fixed x, $z \mapsto \|x-z\|_2^2$ is a strictly convex function, so any minimiser over the convex set C must be unique (see example sheet).
(3.7): We have $(1-t)\pi + tz \in C$ for all $t \in [0,1]$, so

$$\begin{aligned}\|x - \pi\|_2^2 &\le \|x - \pi + t(\pi - z)\|_2^2 \\ &= \|x - \pi\|_2^2 - 2t(x - \pi)^T(z - \pi) + t^2\|\pi - z\|_2^2,\end{aligned}$$

whence

$$(x - \pi)^T(z - \pi) \le \frac{t}{2}\|\pi - z\|_2^2 \qquad \text{for all } t \in (0, 1].$$

Letting $t \to 0$ shows (3.7).
(3.8): From (3.7) we have

$$\begin{aligned}(x - \pi_C(x))^T(\pi_C(z) - \pi_C(x)) &\le 0 \\ (z - \pi_C(z))^T(\pi_C(x) - \pi_C(z)) &\le 0.\end{aligned}$$

Adding these we have

$$\begin{aligned}\|\pi_C(x) - \pi_C(z)\|_2^2 &\le (\pi_C(x) - \pi_C(z))^T(x - z) \\ &\le \|\pi_C(x) - \pi_C(z)\|_2\|z - x\|_2 \qquad \text{(Cauchy–Schwarz)}.\end{aligned}$$

Dividing both sides by $\|\pi_C(x) - \pi_C(z)\|_2$ thus gives the result. □

Definition 7. We call $\pi_C(x)$ above the *projection of x on C*.

3.9 Subgradients

For a convex function $f : \mathbb{R}^d \to \mathbb{R}$ differentiable at $x \in \mathbb{R}^d$, we have that

$$f(z) \ge f(x) + \nabla f(x)^T(z - x) \quad \text{for all } z \in \mathbb{R}^d,$$

so in particular there is a hyperplane passing through $(x, f(x))$ that lies below the function. This also holds true more generally at points where f may not be differentiable with $\nabla f(x)$ above replaced by a *subgradient*.

Definition 8. A vector $g \in \mathbb{R}^d$ is a *subgradient* of a convex function $f : \mathbb{R}^d \to \mathbb{R}$ at x if

$$f(z) \ge f(x) + g^T(z - x) \qquad \text{for all } \ z \in \mathbb{R}^d.$$

The set of subgradients of f at x is called the *subdifferential* of f at x and denoted $\partial f(x)$.

Proposition 23. *If $f : \mathbb{R}^d \to \mathbb{R}$ is convex, $\partial f(x)$ is non-empty for all $x \in \mathbb{R}^d$.*

**Proof*.* The set $C = \{(z, y) \in \mathbb{R}^d \times \mathbb{R} : y \geq f(z)\}$ (known as the *epigraph* of f) is closed and convex. Take a sequence $w_1, w_2, \ldots \in \mathbb{R}^{d+1}$ such that $w_k \notin C$ for each k and $w_k \to (x, f(x))$ as $k \to \infty$. Then for each k, there exists $v_k \in \mathbb{R}^{d+1}$ where

$$v_k^T w < v_k^T w_k \text{ for all } w \in C. \tag{3.9}$$

Indeed taking $v_k = w_k - \pi_C(w_k)$, from Proposition 22, we have that $v_k^T(w - \pi_C(w_k)) \leq 0$, so then

$$v_k^T w \leq v_k^T \pi_C(w_k) = v_k w_k - \|v_k\|_2^2 < v_k w_k.$$

We can rescale the v_k such that $\|v_k\|_2 = 1$, and (3.9) will be maintained. With this modification, we have that the sequence v_k lies in the closed unit ball. Thus by the Bolzano–Weierstrass theorem, there exists a convergent subsequence $v_{k_j} \to v = (-\tilde{g}, \alpha)$ as $j \to \infty$. Then in particular

$$-\tilde{g}^T z + \alpha y \leq -\tilde{g}^T x + \alpha f(x) \quad \text{for all } (z, y) \in C.$$

Clearly this is only possible if $\alpha < 0$, so dividing by $-\alpha$ and setting $g = \tilde{g}/\alpha$ and $y = f(z)$ we obtain

$$f(z) + g^T z \geq f(x) + g^T x \quad \text{for all } z. \qquad \square$$

To compute subgradients, the following facts will be helpful.

Proposition 24. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, and suppose f is differentiable at x. Then $\partial f(x) = \{\nabla f(x)\}$.*

Proof. Suppose $g \in \mathbb{R}^d$ is a subgradient of f at x. Then, for any $z \in \mathbb{R}^d$, we have

$$\nabla f(x)^T z = \lim_{t \downarrow 0} \frac{f(x + tz) - f(x)}{t} \geq g^T z.$$

In particular, taking $z = g - \nabla f(x)$, we have $\|\nabla f(x) - g\|_2^2 \leq 0$, so we must have $\nabla f(x) = g$. $\square$

Proposition 25 (Subgradient calculus)**.** *Let $f, f_1, f_2 : \mathbb{R}^d \to \mathbb{R}$ be convex. Then*

(i) $\partial(\alpha f)(x) = \{\alpha g : g \in \partial f(x)\}$ *for* $\alpha > 0$,

(ii) $\partial(f_1 + f_2)(x) = \{g_1 + g_2 : g_1 \in \partial f_1(x),\ g_2 \in \partial f_2(x)\}$.

Also if $h : \mathbb{R}^m \to \mathbb{R}$ is given by $h(x) = f(Ax + b)$ where $A \in \mathbb{R}^{d \times m}$ and $b \in \mathbb{R}^d$, then

(iii) $\partial h(x) = A^T \partial f(Ax + b)$.

Example. Consider

$$f(\beta) = \frac{1}{n} \sum_{i=1}^{n} \max(1 - y_i x_i^T \beta, 0).$$

Let $\phi(u) = \max(1-u, 0)$. Then

$$\partial\phi(u) = \begin{cases} \{0\} & \text{if } u > 1, \\ [-1, 0] & \text{if } u = 1, \\ \{-1\} & \text{if } u < 1. \end{cases}$$

By Proposition 25 (iii) writing $h_i(\beta) = \max(1 - y_i x_i^T \beta, 0)$, we have $\partial h_i(\beta) = \{-y_i x_i t : t \in [0,1]\}$ when $y_i x_i^T \beta = 1$. From Proposition 25 (i) and (ii), we see that $\partial f(\beta)$ consists of sums of the form $-\frac{1}{n}\sum_{i=1}^n y_i x_i t_i$ where each t_i may be 0, 1 or anything in $[0,1]$ depending on the value of $y_i x_i^T \beta$. △

3.10 Gradient descent

Suppose we wish to minimise a function f that is differentiable at a point β with gradient $g = \nabla f(\beta)$. A first-order Taylor expansion gives $f(z) \approx f(\beta) + g^T(z - \beta)$, so for small $\eta > 0$,

$$\min_{\delta:\|\delta\|_2=1} f(\beta + \eta\delta) \approx f(\beta) + \eta \min_{\delta:\|\delta\|_2=1} g^T\delta.$$

Thus to minimise the linear approximation of f at β, one should move in the direction of the negative gradient.

The procedure of (projected) *gradient descent* for minimising f over a closed convex set C uses this intuition to produce a sequence of iterates $\beta_1, \beta_2, \ldots$ aiming have $f(\beta_s)$ close to a minimum $f(\hat{\beta})$ for large s.

Algorithm 1 Gradient descent

Input: $\beta_1 \in C$; number of iterations $k \in \mathbb{N}$; sequence of positive step sizes $(\eta_s)_{s=1}^{k-1}$
for $s = 1$ to $k-1$ **do**
  Compute $g_s \in \partial f(\beta_s)$
  $z_{s+1} = \beta_s - \eta_s g_s$
  $\beta_{s+1} = \pi_C(z_{s+1})$
end for
return $\bar{\beta} = \frac{1}{k}\sum_{s=1}^k \beta_s$

Theorem 26. *Suppose $\hat{\beta}$ is a minimiser of convex function $f : \mathbb{R}^p \to \mathbb{R}$ over a closed convex set $C \subseteq \mathbb{R}^p$. Suppose $\sup_{\beta\in C} \|\beta\|_2 \le R < \infty$ and $\sup_{\beta\in C} \sup_{g\in\partial f(\beta)} \|g\|_2 \le L < \infty$. Then if $\eta_s \equiv \eta = 2R/(L\sqrt{k})$, the output $\bar{\beta}$ of the gradient descent algorithm above satisfies*

$$f(\bar{\beta}) - f(\hat{\beta}) \le \frac{2LR}{\sqrt{k}}.$$

Proof. We have

$$\begin{aligned} f(\beta_s) - f(\hat{\beta}) &\leq g_s^T(\beta_s - \hat{\beta}) \quad \text{(definition of subgradient)} \\ &= -\frac{1}{\eta}(z_{s+1} - \beta_s)^T(\beta_s - \hat{\beta}) \\ &= \frac{1}{2\eta}\{\|\beta_s - z_{s+1}\|_2^2 + \|\beta_s - \hat{\beta}\|_2^2 - \|z_{s+1} - \hat{\beta}\|_2^2\}. \end{aligned} \tag{3.10}$$

From Proposition 22, $\|\pi_C(z) - \pi_C(x)\|_2 \leq \|z - x\|_2$, so in particular

$$\|z_{s+1} - \hat{\beta}\|_2^2 \geq \|\beta_{s+1} - \hat{\beta}\|_2^2.$$

Using this and (3.10),

$$f(\beta_s) - f(\hat{\beta}) \leq \frac{1}{2\eta}\{\eta^2\|g_s\|_2^2 + \|\beta_s - \hat{\beta}\|_2^2 - \|\beta_{s+1} - \hat{\beta}\|_2^2\}. \tag{3.11}$$

Now $\|g_s\|_2 \leq L$. Also $\beta_1 \in C$, so by the triangle inequality, $\|\beta_1 - \hat{\beta}\|_2^2 \leq 4R^2$. Thus summing we get

$$\begin{aligned} \frac{1}{k}\sum_{s=1}^{k} f(\beta_s) - f(\hat{\beta}) &\leq \frac{\eta L^2}{2} + \frac{1}{2\eta k}\left(\|\beta_1 - \hat{\beta}\|_2^2 - \|\beta_{k+1} - \hat{\beta}\|_2^2\right) \\ &\leq \frac{\eta L^2}{2} + \frac{2R^2}{\eta k}. \end{aligned}$$

Taking the minimising $\eta = 2R/(L\sqrt{k})$ and using Jensen's inequality to give $f(\bar{\beta}) \leq \frac{1}{k}\sum_{s=1}^{k} f(\beta_s)$, we get the result. □

Example. Consider ERM with hinge loss, $\mathcal{X} = \{x \in \mathbb{R}^p : \|x\|_2 \leq C\}$ and the ℓ_2-constrained hypothesis class $\mathcal{H} = \{x \mapsto x^T\beta : \|\beta\|_2 \leq \lambda\}$. Then a subgradient of the objective function f at β takes the form

$$g = -\frac{1}{n}\sum_{i=1}^{n} y_i x_i t_i \quad \text{where } t_i \in [0, 1].$$

Thus $\|g\|_2 \leq C$ by the triangle inequality. From Theorem 26 we see that the output of gradient descent with step size $\eta = 2\lambda/(C\sqrt{k})$ satisfies $f(\bar{\beta}) - f(\hat{\beta}) \leq 2C\lambda/\sqrt{k}$. △

3.11 Stochastic gradient descent

One issue with gradient descent is that the gradients themselves may be computationally expensive to compute: in the case of ERM the gradient is a sum of n terms corresponding to each data point, and so computing the gradient typically involves a sweep over the entire dataset at each iteration.

Stochastic gradient descent can circumvent this issue in the case of minimising convex functions of the form $f(\beta) = \mathbb{E}\tilde{f}(\beta; U)$, where

- $\tilde{f} : \mathbb{R}^p \times \mathcal{U} \to \mathbb{R}$ is such that $\beta \mapsto \tilde{f}(\beta; u)$ is convex for all $u \in \mathcal{U}$,
- U is a random variable taking values in $\mathcal{U}$.

This encompasses empirical risk minimisation. Indeed let U be uniformly distributed on $\{1, \ldots, n\}$. Then the ERM objective function with $\mathcal{H} = \{h_\beta : \beta \in C\}$ may be written as

$$\frac{1}{n}\sum_{i=1}^{n} \ell(h_\beta(x_i), y_i) = \mathbb{E}\ell(h_\beta(x_U), y_U) = \mathbb{E}\tilde{f}(\beta; U).$$

Note we are thinking of the data $(x_1, y_1), \ldots, (x_n, y_n)$ as fixed; only U is random.

Algorithm 2 Stochastic gradient descent

Input: $\beta_1 \in C$; number of iterations $k \in \mathbb{N}$; sequence of positive step sizes $(\eta_s)_{s=1}^{k-1}$, i.i.d. copies $U_1, \ldots, U_{k-1}$ of U
for $s = 1$ to $k - 1$ **do**
  Compute $\tilde{g}_s \in \partial\tilde{f}(\beta_s; U_s)$ (to be interpreted as $\tilde{g}_s \in h(\beta_s)$ where $h(\beta) = \tilde{f}(\beta; U_s)$)
  $z_{s+1} = \beta_s - \eta_s\tilde{g}_s$
  $\beta_{s+1} = \pi_C(z_{s+1})$
end for
return $\bar{\beta} = \frac{1}{k}\sum_{s=1}^{k} \beta_s$

The key point to note is that computing $\tilde{g}_s$ involves just a single data point (x_{U_s}, y_{U_s}).

Theorem 27. *Suppose $\hat{\beta}$ is a minimiser of f as above over a closed convex set $C \subseteq \mathbb{R}^p$. Suppose $\sup_{\beta\in C} \|\beta\|_2 \leq R < \infty$ and $\sup_{\beta\in C} \mathbb{E}\left(\sup_{\tilde{g}\in\partial\tilde{f}(\beta;U)} \|\tilde{g}\|_2^2\right) \leq L^2 < \infty$. Then if $\eta_s \equiv \eta = 2R/(L\sqrt{k})$, the output $\bar{\beta}$ of the stochastic gradient descent algorithm above satisfies*

$$\mathbb{E}f(\bar{\beta}) - f(\hat{\beta}) \leq \frac{2LR}{\sqrt{k}}.$$

Proof. Let $g_s = \mathbb{E}(\tilde{g}_s|\beta_s)$. Then $g_s \in \partial f(\beta_s)$. Indeed we have $\tilde{f}(\beta; U_s) \geq \tilde{f}(\beta_s; U_s) + \tilde{g}_s^T(\beta - \beta_s)$ for all β. Note U_s is independent of β_s so taking expectations conditional on β_s shows $g_s \in \partial f(\beta_s)$. Then arguing as in the proof of Theorem 26,

$$\begin{aligned}
f(\beta_s) - f(\hat{\beta}) &\leq g_s^T(\beta_s - \hat{\beta}) \\
&= \mathbb{E}(\tilde{g}_s(\beta_s - \hat{\beta})|\beta_s) \\
&= -\frac{1}{\eta}\mathbb{E}\{(z_{s+1} - \beta_s)^T(\beta_s - \hat{\beta})|\beta_s\} \\
&= \frac{1}{2\eta}\mathbb{E}\{\|\beta_s - z_{s+1}\|_2^2 + \|\beta_s - \hat{\beta}\|_2^2 - \|z_{s+1} - \hat{\beta}\|_2^2|\beta_s\} \\
&\leq \frac{1}{2\eta}\mathbb{E}\{\eta^2\|\tilde{g}_s\|_2^2 + \|\beta_s - \hat{\beta}\|_2^2 - \|\beta_{s+1} - \hat{\beta}\|_2^2|\beta_s\} \text{ (Prop. 22).}
\end{aligned}$$

Taking expectations and summing we get

$$\mathbb{E}\left(\frac{1}{k}\sum_{s=1}^{k} f(\beta_s)\right) - f(\hat{\beta}) \leq \frac{\eta L^2}{2} + \frac{2R^2}{\eta k}.$$

Taking $\eta = 2R/(L\sqrt{k})$ and using Jensen's inequality we get the result. □

4 Popular machine learning methods

In the course so far, we have developed a coherent framework giving statistical and computational guarantees for a variety of procedures. However many popular machine learning methods do not fall precisely within this framework. In this last part of the course, we will briefly describe a selection of such methods in routine use today. We begin by discussing an important technique for selecting tuning parameters for machine learning methods (e.g. the λ in the cases of ℓ_1 and ℓ_2-constrained hypotheses), or more generally selecting a good classifier or regression method from among a number of competing methods.

4.1 Cross-validation

Let $H^1, \ldots, H^m$ be a collection of machine learning methods: each H^j takes as its argument i.i.d. training data $(X_i, Y_i)_{i=1}^n =: D$ and outputs a hypothesis, so $H_D^j : \mathcal{X} \to \mathbb{R}$. Given a loss function ℓ, we may ideally want to pick a j such that

$$\mathbb{E}\{\ell(H_D^j(X), Y)|D\} \tag{4.1}$$

is minimised. Here $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ is independent of D and has the same distribution as (X_1, Y_1). This j is such that conditional on the original training data, it minimises the expected loss on a new observation drawn from the same distribution as the training data.

A less ambitious goal is to find a j to minimise

$$\mathbb{E}[\mathbb{E}\{\ell(H_D^j(X), Y)|D\}] \tag{4.2}$$

where compared with (4.1), we have taken a further expectation over the training data D.

We still have no way of computing (4.2) directly, but we can attempt to estimate it. The idea of v-fold cross-validation is to split the data into v groups or folds of roughly equal size. Let D_{-k} be all the data except that in the kth fold, and let $A_k \subset \{1, \ldots, n\}$ be the observation indices corresponding to the kth fold. For each j we apply H^j to data D_{-k} to obtain hypothesis $H_{-k}^j := H_{D_{-k}}^j$. We choose the value of j that minimises

$$\mathrm{CV}(j) := \frac{1}{n}\sum_{k=1}^{v}\sum_{i \in A_k} \ell(H_{-k}^j(X_i), Y_i). \tag{4.3}$$

Writing $\hat{j}$ for the minimiser, we may take final selected hypothesis to be $H_D^{\hat{j}}$.

Note that for each $i \in A_k$,

$$\mathbb{E}\ell(H^j_{-k}(X_i), Y_i) = \mathbb{E}[\mathbb{E}\{\ell(H^j_{-k}(X_i), Y_i)|D_{-k}\}]. \tag{4.4}$$

This is precisely the expected loss in (4.2) but with training data D replaced with a training data set of smaller size. If all the folds have the same size, then $\mathrm{CV}(j)$ is an average of n identically distributed quantities, each with expected value as in (4.4). However, the quantities being averaged are not independent as they share the same data.

Thus cross-validation gives a biased estimate of the expected prediction error. The amount of the bias depends on the size of the folds, the case when the $v = n$ giving the least bias—this is known as leave-one-out cross-validation. The quality of the estimate, though, may be worse as the quantities being averaged in (4.3) will be highly positively correlated. Typical choices of v are 5 or 10.

4.1.1 *Stacking*

Cross-validation aims to allow us to choose the single best machine learning method; we could instead aim to find the best weighted combination of methods. To do this, we can attempt to minimise

$$\frac{1}{n}\sum_{k=1}^{v}\sum_{i\in A_k}\ell\left(\sum_{j=1}^{m} w_j H^j_{-k}(X_i), Y_i\right)$$

over w in the convex set

$$\{u \in \mathbb{R}^m : u_j \geq 0 \text{ for all } j\}.$$

Additional ℓ_1 or ℓ_2 constraints may be added to the set. This sort of idea is known as *stacking* and it can often outperform cross-validation.

4.2 Adaboost

Empirical risk minimisation is a technique for finding a single good hypothesis from a given hypothesis class. In an analogy with stacking, we could alternatively attempt to find a good weighted combination of hypotheses. Specifically, given an initial set $\mathcal{B}$ of classifiers $h : \mathcal{X} \to \{-1, 1\}$ such that $h \in \mathcal{B} \Rightarrow -h \in \mathcal{B}$, consider the class

$$\mathcal{H} = \left\{\sum_{m=1}^{M} \beta_m h_m : \beta_m \geq 0,\ h_m \in \mathcal{B} \text{ for } m = 1, \ldots, M\right\}.$$

The class $\mathcal{H}$ is clearly richer than $\mathcal{B}$, and the construction above turns out to be useful way of creating a more complex hypothesis class from a simpler one, with the *tuning parameter* M controlling the complexity. Performing ERM over $\mathcal{H}$, however, can be computationally challenging. The *Adaboost* algorithm can be motivated as a greedy empirical risk minimisation procedure with exponential loss. As we shall see, one attractive feature of the algorithm is that it only relies on being able to perform ERM over the simpler class $\mathcal{B}$ given different weighted versions of the data.

Adaboost first sets $\hat{f}_0$ to be the function $x \mapsto 0$ and then performs the following for $m = 1, \ldots, M$:

$$(\hat{\beta}_m,\, \hat{h}_m) = \operatorname*{arg\,min}_{\beta \geq 0, h \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^{n} \exp[-Y_i\{\hat{f}_{m-1}(X_i) + \beta h(X_i)\}]$$
$$\hat{f}_m = \hat{f}_{m-1} + \hat{\beta}_m \hat{h}_m.$$

The final classification is performed according to $\operatorname{sgn} \circ \hat{f}_M$. Let us examine the minimisation above in more detail. Set $w_i^{(m)} = n^{-1} \exp(-Y_i \hat{f}_{m-1}(X_i))$. Then

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^{n} \exp[-Y_i\{\hat{f}_{m-1}(X_i) + \beta h(X_i)\}] &= e^{\beta} \sum_{i=1}^{n} w_i^{(m)} \mathbb{1}_{\{h(X_i) \neq Y_i\}} + e^{-\beta} \sum_{i=1}^{n} w_i^{(m)} \mathbb{1}_{\{h(X_i) = Y_i\}} \\ &= (e^{\beta} - e^{-\beta}) \sum_{i=1}^{n} w_i^{(m)} \mathbb{1}_{\{h(X_i) \neq Y_i\}} + e^{-\beta} \sum_{i=1}^{n} w_i^{(m)}.\end{aligned}$$

Provided no $h \in \mathcal{B}$ perfectly classifies the data so

$$\operatorname{err}_m(h) := \frac{\sum_{i=1}^{n} w_i^{(m)} \mathbb{1}_{\{h(X_i) \neq Y_i\}}}{\sum_{i=1}^{n} w_i^{(m)}} > 0 \quad \text{for all } h \in \mathcal{B},$$

we have that

$$\hat{h}_m = \operatorname*{arg\,min}_{h \in \mathcal{B}} \operatorname{err}_m(h),$$

and $\hat{\beta}_m$ satisfies $(e^{\hat{\beta}_m} + e^{-\hat{\beta}_m})\operatorname{err}_m(\hat{h}_m) = e^{-\hat{\beta}_m}$. Letting $x = e^{\hat{\beta}_m}$ and $a = \operatorname{err}_m(\hat{h}_m)$, we have

$$\begin{aligned}(x^2 + 1)a &= 1 \\ \text{so} \quad x &= \sqrt{1/a - 1} \\ \text{i.e.} \quad \hat{\beta}_m &= \frac{1}{2} \log\left(\frac{1 - \operatorname{err}_m(\hat{h}_m)}{\operatorname{err}_m(\hat{h}_m)}\right).\end{aligned}$$

If M is large, the weighted empirical risk minimisation step to produce the $\hat{h}_m$ must be performed many times. In order for this approach to be practical, we need $\mathcal{B}$ to be such that this optimisations can be done very fast. More generally, the $\hat{h}_m$ need not be formed through ERM but may be the output of some machine learning method.

Example. Let $\mathcal{X} = \mathbb{R}^p$ and consider the class of *decision stumps*

$$\mathcal{B} = \{h_{a,j,1}(x) = \operatorname{sgn}(x_j - a),\, h_{a,j,2}(x) = \operatorname{sgn}(a - x_j) : a \in \mathbb{R},\ j = 1, \ldots, p\}.$$

To perform weighted ERM with weights $w_1, \ldots, w_n > 0$ (we have dropped the superscript m), for each $j = 1, \ldots, p$, first sort $\{X_{ij}\}_{i=1}^n$ so $X_{(1)j} < \cdots < X_{(n)j}$ (we assume these are

distinct for simplicity). Fixing j, wlog we may assume $X_{(i)j} = X_{ij} = x_i$. Now observe that (dropping the subscript m),

$$\mathrm{err}(h_{x_{k+1},j,1}) - \mathrm{err}(h_{x_k,j,1}) = Y_{k+1} w_{k+1} \Big/ \sum_l w_l.$$

Thus picking the optimal $h_{a,j,1}$ (for fixed j) amounts to picking the minimum across a sequence of cumulative sums, and similarly for $h_{a,j,2}$. This needs to be performed for each $j = 1, \ldots, p$. Assuming the sorting is performed as part of pre-processing, the weighted empirical risk minimisation has $O(np)$ computational complexity. $\triangle$

4.3 Gradient boosting

Consider the following thought experiment. Let us imagine applying gradient descent directly to minimise $R(h) = \mathbb{E}\ell(h(X), Y)$. This would involve the following steps.

1. Start with an initial guess $f_0 : \mathcal{X} \to \mathbb{R}$.
2. For $m = 1, \ldots, M$, iteratively compute

$$\begin{aligned} g_m(x) &= \frac{\partial \mathbb{E}(\ell(\theta, Y)|X = x)}{\partial \theta}\bigg|_{f_{m-1}(x)} \\ &= \mathbb{E}\left(\frac{\partial \ell(\theta, Y)}{\partial \theta}\Big|_{f_{m-1}(x)} \middle| X = x \right) \end{aligned}$$

assuming sufficient regularity conditions.

3. Update $f_m = f_{m-1} - \eta g_m$, where $\eta > 0$ is a small step length.

If we want to create a version of the 'algorithm' above that works with finite data $(X_1, Y_1), \ldots, (X_n, Y_n)$, we need to find a way of approximating the conditional expectation function

$$x \mapsto \mathbb{E}\left(\frac{\partial \ell(\theta, Y)}{\partial \theta}\Big|_{f_{m-1}(x)} \middle| X = x \right).$$

Recall from (iv) on page 3, that this minimises

$$\mathbb{E}\left(\frac{\partial \ell(\theta, Y)}{\partial \theta}\Big|_{f_{m-1}(X)} - h(X) \right)^2 \tag{4.5}$$

among all (measurable) functions $h : \mathcal{X} \to \mathbb{R}$ under suitable conditions. This observation motivates the following algorithm known as *gradient boosting*, where we try to minimise an empirical version of (4.5) using regression, thereby approximating the conditional expectation. This regression is performed using some base regression method H that takes as its argument some training data D and outputs a hypothesis $H_D : \mathcal{X} \to \mathbb{R}$. In what follows, the loss ℓ may correspond to a convex surrogate or least squares loss for example.

Algorithm 3 Gradient boosting

Input: Data $X_{1:n}$, $Y_{1:n}$; $\eta > 0$; base regression method H; stopping iteration M
Compute $\hat{\mu} = \arg\min_{\mu \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \ell(\mu, Y_i)$ and set $f_0(x) = \hat{\mu}$
for $m = 1$ to M **do**
  Compute $W_i = \frac{\partial}{\partial \theta} \ell(\theta, Y_i)|_{\theta = \hat{f}_{m-1}(X_i)}$
  Apply H to data $X_{1:n}, W_{1:n}$ to give $\hat{g}_m = H_{(X_{1:n}, W_{1:n})} : \mathcal{X} \to \mathbb{R}$
  Update $\hat{f}_m = \hat{f}_{m-1} - \eta \hat{g}_m$
end for
return $\hat{f}_M$ (or $\mathrm{sgn} \circ \hat{f}_M$ in the classification setting)

4.4 Decision trees

Gradient boosting requires a fast base regression procedure. In the setting where $\mathcal{X} = \mathbb{R}^p$, methods for fitting *decision trees* are the most popular choice. Decision trees are a generalisation of decision stumps and take the form

$$T(x) = \sum_{j=1}^{J} \gamma_j \mathbb{1}_{\{x \in R_j\}},$$

where R_j are rectangular regions that form a partition of $\mathbb{R}^p$ and the γ_j are coefficients in $\mathbb{R}$.

The regions and coefficients are typically computed from data $(X_i, Y_i)_{i=1}^n$ using the following recursive binary partitioning algorithm.

1. Input maximum number of regions J. Initialise $\hat{\mathcal{R}} = \{\mathbb{R}^p\}$.

2. For each region $R \in \hat{\mathcal{R}}$ such that $I := \{i : X_i \in R\}$ has $|I| > 1$, perform the following. For each $j = 1, \ldots, p$, let $\mathcal{S}_j$ be the set of mid-points between adjacent $\{X_{ij}\}_{i \in I}$. Find the predictor $\hat{j}_R$ and split point $\hat{s}_R$ to minimise over $j \in \{1, \ldots, p\}$ and $s \in \mathcal{S}_j$,

$$\underbrace{\min_{c_1 \in \mathbb{R}} \sum_{i \in I : X_{ij} \leq s} (Y_i - c_1)^2 + \min_{c_2 \in \mathbb{R}} \sum_{i \in I : X_{ij} > s} (Y_i - c_2)^2}_{\text{RSS on } I \text{ when splitting at } s} - \underbrace{\min_{c \in \mathbb{R}} \sum_{i \in I} (Y_i - c)^2}_{\text{RSS on } I \text{ without splitting}} . \tag{4.6}$$

   In words: For each region, we find the axis-aligned split such that the residual sum of squares (RSS) on the region is minimised.

3. Let $\hat{R}$ be the region yielding the lowest value of (4.6) and define

$$\hat{R}_1 = \{x \in \hat{R} : x_{\hat{j}_{\hat{R}}} \leq \hat{s}_{\hat{R}}\}, \quad \hat{R}_2 = \hat{R} \setminus \hat{R}_1.$$

   Refine the partition via $\hat{\mathcal{R}} \leftarrow (\hat{\mathcal{R}} \setminus \{\hat{R}\}) \cup \{\hat{R}_1, \hat{R}_2\}$.

4. Repeat steps 2 and 3 until $|\hat{\mathcal{R}}| = J$.

5. Writing $\hat{\mathcal{R}} = \{\hat{R}_1, \ldots, \hat{R}_J\}$, let $\hat{I}_j = \{i : X_i \in \hat{R}_j\}$ and

$$\hat{\gamma}_j = \frac{1}{|\hat{I}_j|} \sum_{i \in \hat{I}_j} Y_i.$$

Output $\hat{T} : \mathbb{R}^p \to \mathbb{R}$ such that $\hat{T}(x) = \sum_{j=1}^{J} \hat{\gamma}_j \mathbb{1}_{\{x \in \hat{R}_j\}}$.

4.5 Random forests

Whilst decision trees as above are a useful machine learning method in their own right, they are most useful for prediction when used in conjunction with gradient boosting or within the *Random forest* procedure which we now describe.

Consider the regression setting where $Y_i \in \mathbb{R}$ and we are using squared error loss. Let $\hat{T}_D$ be a decision tree trained on data $D := (X_i, Y_i)_{i=1}^n$. Also let $\bar{T} = \mathbb{E}\hat{T}_D$ and let (X, Y) be independent of D with $(X, Y) \stackrel{d}{=} (X_1, Y_1)$. Recall property (iv) on page 3, that for random variables $Z, W \in \mathbb{R} \times \mathcal{W}$ and $f : \mathcal{W} \to \mathbb{R}$, we have

$$\mathbb{E}\{Z - f(W)\}^2 = \mathbb{E}\{Z - \mathbb{E}(Z \,|\, W)\}^2 + \mathbb{E}\{\mathbb{E}(Z \,|\, W) - f(W)\}^2.$$

Using this, we have the following decomposition of the expected risk of $\hat{T}_D$:

$$\begin{aligned}
\mathbb{E}R(\hat{T}_D) &= \mathbb{E}[\{Y - \hat{T}_D(X)\}^2] \\
&= \mathbb{E}\{Y - \underbrace{\mathbb{E}(Y \,|\, X, D)}_{=\mathbb{E}(Y \,|\, X)}\}^2 + \mathbb{E}\{\mathbb{E}(Y \,|\, X) - \hat{T}_D(X)\}^2 \\
&= \mathbb{E}\mathrm{Var}(Y \,|\, X) + \mathbb{E}\{\hat{T}_D(X) - \underbrace{\mathbb{E}(\hat{T}_D(X) \,|\, X)}_{=\bar{T}(X)}\}^2 + \mathbb{E}\{\mathbb{E}(Y \,|\, X) - \bar{T}(X)\}^2 \\
&= \underbrace{\mathbb{E}\{\mathbb{E}(Y \,|\, X) - \bar{T}(X)\}^2}_{\text{squared bias}} + \underbrace{\mathbb{E}\mathrm{Var}(\hat{T}_D(X) \,|\, X)}_{\text{variance of the tree}} + \underbrace{\mathbb{E}\mathrm{Var}(Y \,|\, X)}_{\text{irreducible variance}}.
\end{aligned}$$

If the number of regions J used by $\hat{T}_D$ is large, some of these regions will contain only small numbers of observations in them so the corresponding coefficients $\hat{\gamma}_j$ will by highly variable and consequently $\mathbb{E}\mathrm{Var}(\hat{T}_D(X) \,|\, X)$ will tend to be large. On the other hand, the squared bias above and hence $R(\bar{T})$ may be low as a large J would allow $\bar{T}$ to approximate $x \mapsto \mathbb{E}(Y \,|\, X = x)$ well.

Random forest effectively attempts to 'estimate' $\bar{T}$ and so improve upon the variance of a single tree. If we had multiple independent datasets $D_1, \ldots, D_B$, we could form an unbiased estimate via $\sum_{b=1}^{B} \hat{T}_{D_b}$. Random forest samples the data D with replacement to form new datasets $D_1^*, \ldots, D_B^*$ and performs the following.

1. For each $b = 1, \ldots, B$, grow a decision tree $\hat{T}^{(b)} := \hat{T}_{D_b^*}$ but when searching for the best predictor to split on, randomly sample (without replacement) m_{try} of the p predictors and choose the best split from among these variables.

2. Output $f_{\text{rf}} = \frac{1}{B}\sum_{b=1}^{B} \hat{T}^{(b)}$.

The reason for sampling predictors is to try to make the $\hat{T}^{(b)}$ more independent. To see why this would be useful, suppose for $b_1 \neq b_2$ and some $x \in \mathbb{R}^p$ that $\text{Corr}(\hat{T}^{(b_1)}(x), \hat{T}^{(b_2)}(x)) = \rho \geq 0$. Then

$$\begin{aligned}\text{Var}(f_{\text{rf}}(x)) &= \frac{1}{B}\text{Var}(\hat{T}^{(1)}(x)) + \frac{\rho B(B-1)}{B^2}\text{Var}(\hat{T}^{(1)}(x)) \\ &= \frac{1-\rho}{B}\text{Var}(\hat{T}^{(1)}(x)) + \rho\text{Var}(\hat{T}^{(1)}(x)).\end{aligned}$$

Whilst the first term can be made small for large B, the second term does not depend on B, so we would like ρ to be small. The extra randomisation in the form of sampling predictors can help to achieve this, and we would expect $\text{Var}(f_{\text{rf}}(x))$ to decrease with m_{try}. On the other hand, we would expect the squared bias to increase as m_{try} is decreased.

4.6 Feedforward neural networks

In recent years, (artificial) neural networks have been shown to be very successful for a variety of learning tasks. The class of *feedforward neural networks* are based around a particular class of hypotheses $h : \mathcal{X} = \mathbb{R}^p \to \mathbb{R}$ with general form

$$h(x) = A^{(d)} \circ g \circ A^{(d-1)} \circ g \circ \cdots \circ g \circ A^{(2)} \circ g \circ A^{(1)}(x)$$

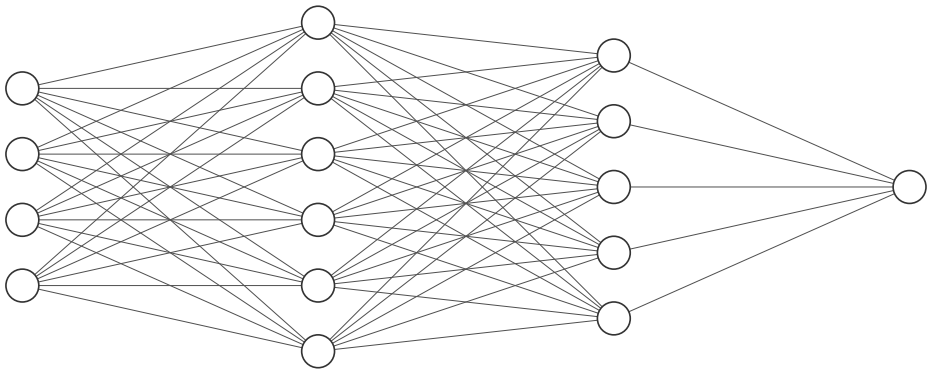
where

- d is known as the *depth* of the network;
- $A^{(k)}(v) = \beta^{(k)}v + \mu^{(k)}$ where $v \in \mathbb{R}^{m_k}$, $\beta^{(k)} \in \mathbb{R}^{m_{k+1} \times m_k}$, $\mu^{(k)} \in \mathbb{R}^{m_{k+1}}$ with $m_1 = p$ and $m_{d+1} = 1$;
- $g : \mathbb{R}^m \to \mathbb{R}^m$ applies (for any given m) a so-called *activation function* $\psi : \mathbb{R} \to \mathbb{R}$ elementwise i.e. for $v = (v_1, \ldots, v_m)^T$, $g(v) = (\psi(v_1), \ldots, \psi(v_m))^T$. The activation function is nonlinear and typical choices include

  (i) $u \mapsto \max(u, 0)$ (known as a *rectified linear unit (ReLU)*)

  (ii) $u \mapsto 1/(1 + e^{-u})$ (sigmoid).

This cascade of alternating linear and nonlinear compositions can be visualised in the form of a graph. Here we have set $h^{(0)} := x$ and for $k = 1, \ldots, d-1$, $x^{(k)} = A^{(k)}(h^{(k-1)})$, $h^{(k)} = g(x^{(k)})$. The intermediate outputs $h^{(1)}, \ldots, h^{(d-1)}$ are known as *hidden layers* and $x^{(d)} = A^{(d)}(h^{(d-1)}) = h(x)$ is sometimes known as the *output layer*. The parameters $(\beta^{(k)}, \mu^{(k)})_{k=1}^{d}$ are typically fitted to data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \{-1, 1\}$ with empirical risk minimisation using a surrogate loss ϕ. Despite the resulting optimisation being highly nonconvex, stochastic gradient descent has been shown empirically to be extremely effective in selecting good parameters. A key factor in their success has been the fact that



the gradients involved can be computed quickly due to the compositional nature of the hypotheses using the chain rule.

Suppose ϕ and ψ are differentiable. At an observation $(x, y) = (x_{U_s}, y_{U_s})$ we first compute all the intermediate quantities $h^{(l)}$ and $x^{(l)}$ given the current values of the parameters. Let $z = \phi(yh(x)) = \phi(yx^{(d)})$. We then compute, in order

$$\begin{aligned}
\frac{\partial z}{x^{(d)}} &= y\phi'(yx^{(d)}) \\
\frac{\partial z}{\partial \mu^{(d)}} = \frac{\partial z}{x^{(d)}}, &\qquad \frac{\partial z}{\partial \beta_{1k}^{(d)}} = \frac{\partial z}{x^{(d)}} h_k^{(d-1)} \qquad (4.7) \\
\frac{\partial z}{\partial h_j^{(d-1)}} &= \frac{\partial z}{\partial x^{(d)}}\, \beta_{1j}^{(d)} \\
\frac{\partial z}{\partial x_j^{(d-1)}} &= \frac{\partial z}{\partial h_j^{(d-1)}}\, \psi'(x_j^{(d-1)}) \\
\frac{\partial z}{\partial \mu_j^{(d-1)}} = \frac{\partial z}{\partial x_j^{(d-1)}}, &\qquad \frac{\partial z}{\partial \beta_{jk}^{(d-1)}} = \frac{\partial z}{\partial x_j^{(d-1)}} h_k^{(d-2)} \qquad (4.8) \\
\frac{\partial z}{\partial h_j^{(d-2)}} &= \sum_{k=1}^{m_d} \frac{\partial z}{\partial x_k^{(d-1)}}\, \beta_{kj}^{(d-1)}, \\
&\vdots
\end{aligned}$$

This process is known as *back propogation*. Note only (4.7) and (4.8) out of the equations presented above are directly used in the SGD update step; the remaining equations simply facilitate computation of the gradient with respect to the $(\beta^{(k)}, \mu^{(k)})_{k=1}^{d}$.