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 is known as the (two-class) *classification* setting. 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. We refer to Y as the output, or response, and X as the input and its components as predictors or variables.

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 in the machine learning literature 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, loss ℓ given by the *misclassification error* is particularly relevant:

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

In this context h is also referred to as a *classifier*. In regression settings, the use of *squared error* $\ell(h(x), y) = (h(x) - y)^2$ is common, and we will take this to be the case unless specified otherwise. 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[1]. For a deterministic h, $R(h) = \mathbb{E}\ell(h(X), Y)$.

Recall that the function h that minimises the risk in a regression setting is $x \mapsto \mathbb{E}(Y \mid X = x)$, which we refer to as the *regression function*.

A classifier h_0 that minimises the misclassification risk is known as a *Bayes classifier*, and its risk is called the *Bayes risk*. A key function in the classification context is

$$\eta(x) := \mathbb{P}(Y = 1 \mid X = x),$$

which is also known as the regression function here.

[1] Note that this is a different definition from the 'risk' you may have seen in *Principles of Statistics*.

Proposition 1. *A Bayes classifier h_0 is given by*[2]

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

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 $R(\hat{h})$ or $\mathbb{E}R(\hat{h})$ 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. 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 $\mathcal{H}$ in the context of regression include for instance

- $\mathcal{H} = \{h : h(x) = \mu + x^\top \beta \text{ where } \mu \in \mathbb{R},\ \beta \in \mathbb{R}^p\}$;
- $\mathcal{H} = \left\{h : h(x) = \mu + \sum_{j=1}^d \varphi_j(x)\beta_j \text{ where } \mu \in \mathbb{R},\ \beta \in \mathbb{R}^d\right\}$ for a given set of what are known in this context as *basis functions* $\varphi_1, \ldots, \varphi_d : \mathcal{X} \to \mathbb{R}$;
- $\mathcal{H} = \left\{h : h(x) = \sum_{j=1}^d w_j \varphi_j(x) \text{ where } w \in \mathbb{R}^d,\ \varphi_j \in \mathcal{B}\right\}$ for a given class $\mathcal{B}$ of functions $\varphi : \mathcal{X} \to \mathbb{R}$.

In the classification setting, we may consider versions of the above composed with the sgn function e.g. $\mathcal{H} = \{h : h(x) = \operatorname{sgn}(\mu + x^\top \beta) \text{ where } \mu \in \mathbb{R},\ \beta \in \mathbb{R}^p\}$.

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

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.

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

Recall that if $Z \in \mathbb{R}$ and $W = (W_1, \ldots, W_d)^\top \in \mathbb{R}^d$ are random variables with joint probability density function (pdf) $f_{Z,W}$ with respect to measure μ, 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

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

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 $\mathbb{E}(Z \,|\, W) = \mathbb{E}Z$. If additionally for a random variable U, W is independent of (Z, U), then $\mathbb{E}(Z \,|\, U, W) = \mathbb{E}(Z \,|\, U)$.

(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, 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) **Fixing what is known:** We have

$$\begin{aligned} &\mathbb{E}\{f(W_1, \ldots, W_d) \,|\, W_1 = w_1, \ldots, W_r = w_r\} \\ =&\mathbb{E}\{f(w_1, \ldots, w_r, W_{r+1}, \ldots, W_d) \,|\, W_1 = w_1, \ldots, W_r = w_r\}, \end{aligned}$$

provided the r.h.s. is well-defined. In particular, if $\mathbb{E}Z^2 < \infty$ and $g : \mathbb{R}^d \to \mathbb{R}$ is such that $\mathbb{E}[\{g(W)\}^2] < \infty$, then $\mathbb{E}\{g(W)Z \,|\, W\} = g(W)\mathbb{E}(Z \,|\, W)$, a property sometimes referred to as 'taking out what is known'.

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

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

Indeed, using the tower property,

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

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

$$\mathbb{E}[\{\mathbb{E}(Z \mid W) - g(W)\} \underbrace{\mathbb{E}\{Z - \mathbb{E}(Z \mid W) \mid W\}}_{=0}] = 0.$$

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

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

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

1.2 Bayes risk

Proof of Proposition 1. We have $R(h) = \mathbb{P}(Y \neq h(X)) = \mathbb{E}\mathbb{P}(Y \neq h(X) \mid X)$, so $h_0(x)$ must minimise over $h(x)$

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

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 $h \in \mathcal{H}$:

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

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

Example 1. Consider the regression setting with $\mathcal{Y} = \mathbb{R}$, squared error loss and $\mathcal{H} = \{x \mapsto \mu + x^\top \beta \text{ for } \mu \in \mathbb{R},\ \beta \in \mathbb{R}^p\}$. Then empirical risk minimisation is equivalent to ordinary least squares, i.e. we have

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

We can consider applying this more generally where

$$\mathcal{H} = \left\{ x \mapsto \sum_{j=1}^{d} \varphi_j(x)\beta_j \text{ where } \beta \in \mathbb{R}^d \right\}$$

and $\varphi_j : \mathbb{R}^p \to \mathbb{R}$ for $j = 1, \dots, d$. For instance in the case where $p = 1$, we could have $\varphi_j(x) = x^{j-1}$. Then forming matrix $\Phi \in \mathbb{R}^{n\times d}$ with entries $\Phi_{ij} = \varphi_j(X_i)$ assumed to be of full column rank, and writing $\varphi(x) = (\varphi_1(x), \dots, \varphi_d(x))$, we have that the ERM $\hat{h} : x \mapsto \hat{\beta}^\top \varphi(x)$ where

$$\hat{\beta} = (\Phi^\top \Phi)^{-1}\Phi^\top Y_{1:n} \tag{1.3}$$

and $Y_{1:n} := (Y_1, \dots, Y_n)^\top$. △

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 ERM $\hat{h}$ to predict a new data point $(X, Y) \sim P_0$ given only knowledge of X. To understand the competing factors that drive this sort of quantity, it is helpful to consider the case of squared error loss where, as we shall see, this may be related to a sum of bias and variance terms.

1.4 Bias–variance tradeoff

Let us consider $\hat{h} = \hat{h}_D$ trained on data $D = (X_i, Y_i)_{i=1}^n$ formed of iid copies of an independent random pair (X, Y). We first consider its expected performance in terms of squared error at X. To this end, it is helpful to introduce

$$\bar{h} : x \mapsto \mathbb{E}(\hat{h}_D(x)),$$

i.e. the average over the training data of $\hat{h}_D$, and the related function

$$\widetilde{h}_{X_{1:n}} : x \mapsto \mathbb{E}(\hat{h}_D(x) \,|\, X_{1:n}).$$

Recall property (iv) of conditional expectations, 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

$$\begin{aligned}
&\mathbb{E}[\{Y - \hat{h}_D(X)\}^2 \,|\, X] \\
&= \mathbb{E}[\{Y - \underbrace{\mathbb{E}(Y \,|\, X, D)}_{=\mathbb{E}(Y \,|\, X)}\}^2 \,|\, X] + \mathbb{E}[\{\mathbb{E}(Y \,|\, X) - \hat{h}_D(X)\}^2 \,|\, X] \\
&= \mathrm{Var}(Y \,|\, X) + \mathbb{E}[\{\hat{h}_D(X) - \underbrace{\mathbb{E}(\hat{h}_D(X) \,|\, X)}_{=\bar{h}(X)}\}^2 \,|\, X] + \mathbb{E}[\{\mathbb{E}(Y \,|\, X) - \bar{h}(X)\}^2 \,|\, X]. \qquad (1.4)
\end{aligned}$$

Thus, taking expectations:

$$\mathbb{E}R(\hat{h}_D) = \underbrace{\mathbb{E}\{\mathbb{E}(Y \,|\, X) - \bar{h}(X)\}^2}_{\text{squared bias}} + \underbrace{\mathbb{E}\mathrm{Var}(\hat{h}_D(X) \,|\, X)}_{\text{variance of } \hat{h}} + \underbrace{\mathbb{E}\mathrm{Var}(Y \,|\, X)}_{\text{irreducible variance}}. \qquad (1.5)$$

If $\hat{h}$ were an ERM over class $\mathcal{H}$, we would expect a rich class of hypotheses to result in a smaller squared bias term. However, the variance would likely increase as empirical risk minimisation may fit to the realised $Y_1, \ldots, Y_n$ closely and so $\hat{h}_D$ would be very sensitive to the training data D.

To see this tradeoff more clearly, it is instructive to consider a related decomposition to (1.4) involving $\widetilde{h}$: we have

$$\mathbb{E}[\{Y - \hat{h}_D(X)\}^2 \,|\, X = x] = \mathbb{E}\{\mathbb{E}(Y \,|\, X = x) - \widetilde{h}_{X_{1:n}}(x)\}^2 + \mathbb{E}\{\hat{h}_D(x) - \widetilde{h}_{X_{1:n}}(x)\}^2 + \mathrm{Var}(Y \,|\, X = x).$$

We examine the middle term in more detail, and consider the special case where $\hat{h}_D$ is the ERM of Example 1 given by (1.3), that is $\hat{h}_D(x) = \varphi(x)^\top(\Phi^\top\Phi)^{-1}\Phi^\top Y_{1:n}$ with $\varphi(x) \in \mathbb{R}^d$. To facilitate our analysis, let us assume that $\mathrm{Var}(Y \,|\, X = x) =: \sigma^2$ is constant in x. Then we have

$$\begin{aligned}
&\mathbb{E}[\{\hat{h}_D(x) - \widetilde{h}_{X_{1:n}}(x)\}^2 \,|\, X_{1:n}] \\
&= \mathbb{E}[\{\varphi(x)^\top(\Phi^\top\Phi)^{-1}\Phi^\top(Y_{1:n} - \mathbb{E}(Y_{1:n} \,|\, X_{1:n}))\}^2 \,|\, X_{1:n}] \\
&= \varphi(x)^\top(\Phi^\top\Phi)^{-1}\Phi^\top\mathbb{E}[\{Y_{1:n} - \mathbb{E}(Y_{1:n} \,|\, X_{1:n})\}\{Y_{1:n} - \mathbb{E}(Y_{1:n} \,|\, X_{1:n})\}^\top \,|\, X_{1:n}]\Phi(\Phi^\top\Phi)^{-1}\varphi(x).
\end{aligned}$$

Note that by property (i) of conditional expectations, $\mathbb{E}(Y_j \,|\, X_{1:n}) = \mathbb{E}(Y_j \,|\, X_j)$ and also,

$$\begin{aligned}
\mathbb{E}[\{Y_j - \mathbb{E}(Y_j \,|\, X_j)\}\{Y_k - \mathbb{E}(Y_k \,|\, X_k)\} \,|\, X_{1:n}] &= \mathbb{E}[\{Y_j - \mathbb{E}(Y_j \,|\, X_j)\}\{Y_k - \mathbb{E}(Y_k \,|\, X_k)\} \,|\, X_j, X_k] \\
&= \mathbb{E}(Y_jY_k \,|\, X_j, X_k) - \mathbb{E}(Y_j \,|\, X_j)\mathbb{E}(Y_k \,|\, X_k),
\end{aligned}$$

using the tower property in the final line. Now if $j \neq k$,

$$\begin{aligned}
\mathbb{E}(Y_jY_k \,|\, X_j, X_k) &= \mathbb{E}\{\mathbb{E}(Y_jY_k \,|\, Y_j, X_j, X_k) \,|\, X_j, X_k\} \qquad \text{(tower property)} \\
&= \mathbb{E}\{Y_j\mathbb{E}(Y_k \,|\, Y_j, X_j, X_k) \,|\, X_j, X_k\} \qquad \text{(taking out what is known)} \\
&= \mathbb{E}\{Y_j\mathbb{E}(Y_k \,|\, X_k) \,|\, X_j, X_k\} \qquad \text{(property (i))} \\
&= \mathbb{E}(Y_j \,|\, X_j)\mathbb{E}(Y_k \,|\, X_k) \qquad \text{(taking out what is known and (i))}.
\end{aligned}$$

Thus $\mathbb{E}[\{Y_{1:n} - \mathbb{E}(Y_{1:n} \mid X_{1:n})\}\{Y_{1:n} - \mathbb{E}(Y_{1:n} \mid X_{1:n})\}^\top \mid X_{1:n}] = \sigma^2 I$, and so

$$\mathbb{E}[\{\hat{h}_D(x) - \widetilde{h}_{X_{1:n}}(x)\}^2 \mid X_{1:n}] = \sigma^2 \varphi(x)^\top (\Phi^\top \Phi)^{-1} \varphi(x).$$

Consider now averaging this over the training points $x = X_1, \ldots, X_n$. Noting that $\varphi(X_i)$ is the ith row of Φ, we may compute, using the 'trace trick' (and that trace is invariant to cyclic permutations),

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \sigma^2 \mathrm{tr}\{\varphi(X_i)^\top (\Phi^\top \Phi)^{-1} \varphi(X_i)\} &= \frac{\sigma^2}{n} \mathrm{tr}\Bigg(\underbrace{\sum_{i=1}^n \varphi(X_i)\varphi(X_i)^\top}_{=\Phi^\top\Phi} (\Phi^\top \Phi)^{-1} \Bigg) \\ &= \frac{\sigma^2 d}{n}.\end{aligned}$$

Thus the variance term increases linearly with d, while the squared bias should decrease when adding further basis functions φ_j.

At least two questions may arise at this stage: how should we choose the number of basis functions in practice in order to obtain a small expected risk? And, particularly in multivariate settings, what are sensible ways of choosing the basis functions themselves? We turn to the first of these questions next.

1.5 Cross-validation

The question of selecting the appropriate number of basis functions in a linear regression may be seen as a special case of the following problem: given a number of competing machine learning methods, select from these (ideally) the best one i.e. one that trades off bias and variance most favourably. In the case of linear regression, each regression using a given set of basis functions may be thought of as one of the competing methods.

Now let $\hat{h}^{(1)}, \ldots, \hat{h}^{(m)}$ be a collection of machine learning methods: for instance $\hat{h}^{(j)}$ could correspond to performing linear regression using basis functions $\varphi_1, \ldots, \varphi_j$. Each $\hat{h}^{(j)}$ takes as its argument i.i.d. training data $(X_i, Y_i)_{i=1}^n =: D \in (\mathcal{X} \times \mathcal{Y})^n$ and outputs a hypothesis, so $\hat{h}^{(j)}_D : \mathcal{X} \to \mathbb{R}$. Given a loss function ℓ with associated risk R, we may ideally want to pick a $\hat{h}^{(j)}$ such that the risk

$$R(\hat{h}^{(j)}_D) = \mathbb{E}\{\ell(\hat{h}^{(j)}_D(X), Y) \mid D\} \tag{1.6}$$

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 $\hat{h}^{(j)}$ 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 the expected risk

$$\mathbb{E}R(\hat{h}^{(j)}_D) = \mathbb{E}[\mathbb{E}\{\ell(\hat{h}^{(j)}_D(X), Y) \mid D\}] \tag{1.7}$$

where compared with (1.6), we have taken a further expectation over the training data D.

We still have no way of computing (1.7) 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 $\hat{h}^{(j)}$ to data D_{-k} to obtain hypothesis $\hat{h}^{(j)}_{-k} := \hat{h}^{(j)}_{D_{-k}}$. We choose the value of j that minimises

$$\mathrm{CV}(j) := \frac{1}{n} \sum_{k=1}^{v} \sum_{i \in A_k} \ell(\hat{h}^{(j)}_{-k}(X_i), Y_i). \tag{1.8}$$

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

Note that for each $i \in A_k$,

$$\mathbb{E}\ell(\hat{h}^{(j)}_{-k}(X_i), Y_i) = \mathbb{E}[\mathbb{E}\{\ell(\hat{h}^{(j)}_{-k}(X_i), Y_i)|D_{-k}\}]. \tag{1.9}$$

This is precisely the expected loss in (1.7) 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 (1.9). 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 $v = n$ typically 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 (1.8) will tend to be positively correlated. Typical choices of v are 5 or 10.

2 Popular machine learning methods I

2.1 Decision trees

We now have a way to select an appropriate subset of basis functions to use from a larger collection, but how should we choose this collection in the first place? Decision trees (also known as regression trees in the regression context we study here; there are also variants for classification which we will not discuss) form a highly popular class of methods for doing this in a data-driven fashion.

Regression trees use a set of basis functions consisting of indicator functions on rectangular regions and take the form

$$T(x) = \sum_{j=1}^{J} \gamma_j \mathbb{1}_{R_j}(x); \tag{2.1}$$

here 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. We now split one of the regions in $\hat{\mathcal{R}}$ using an axis aligned split such that a particular splitting criterion is minimised. In the regression case, it often makes sense to aim to minimise the overall residual sum of squares (RSS) as follows.

   (a) 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 $s \in \mathcal{S}_j$,

   $$\underbrace{\min_{\gamma_L \in \mathbb{R}} \sum_{i \in I : X_{ij} \le s} (Y_i - \gamma_L)^2 + \min_{\gamma_R \in \mathbb{R}} \sum_{i \in I : X_{ij} > s} (Y_i - \gamma_R)^2}_{\text{RSS on } I \text{ when splitting at } s} - \underbrace{\min_{c \in \mathbb{R}} \sum_{i \in I} (Y_i - c)^2}_{\text{RSS on } I \text{ without splitting}} . \tag{2.2}$$

   (b) Let $\hat{R}$ be the region yielding the lowest value of (2.2) and define

   $$\hat{R}_L = \{x \in \hat{R} : x_{\hat{j}_{\hat{R}}} \le \hat{s}_{\hat{R}}\}, \quad \hat{R}_R = \hat{R} \setminus \hat{R}_L.$$

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

3. Repeat step 2 until $|\hat{\mathcal{R}}| = J$.

4. 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 $\hat{T}(x) = \sum_{j=1}^J \hat{\gamma}_j \mathbb{1}_{\{x \in \hat{R}_j\}}$.

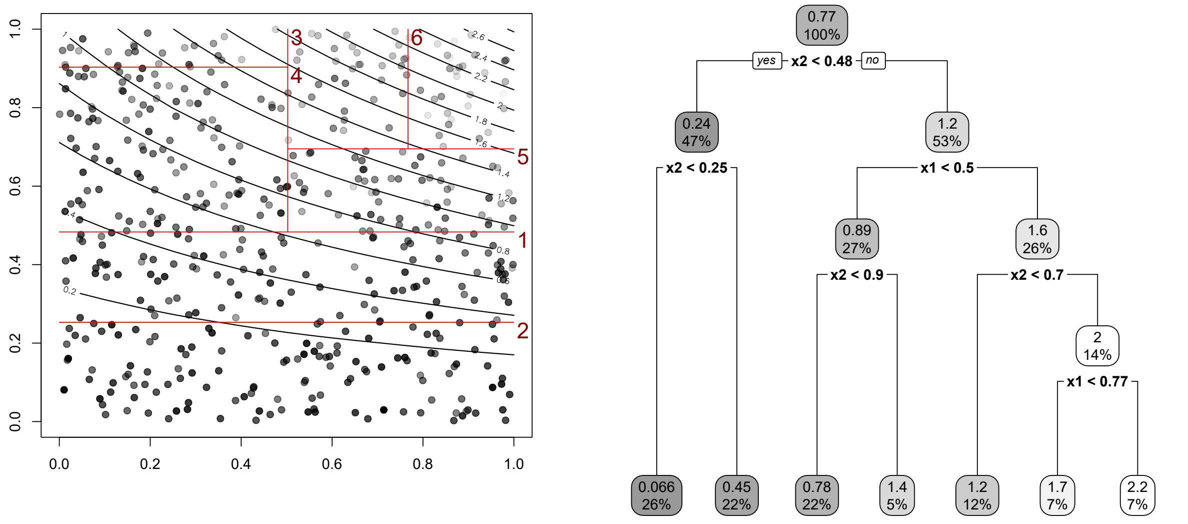
Note that $\hat{T}$ is the ERM over the class of functions

$$\left\{ T : T(x) = \sum_{j=1}^J \gamma_j \mathbb{1}_{\hat{R}_j}(x) \ : \gamma \in \mathbb{R}^J \right\},$$

with the regions $\hat{R}_1, \ldots, \hat{R}_J$ fixed. Note that although the regions were constructed in a data-driven fashion, they were chosen greedily to minimise the RSS at each stage. Thus in general, the fitted $\hat{T}$ will not coincide the the RSS-minimising function of the form (2.1).

The fitted $\hat{T}$ can be conveniently visualised in terms of a tree as indicated in Figure 1b. The regions $\hat{R}_j$ correspond to the so-called *leaves*, those bottom nodes with only a single edge emanating from them.

At first sight, it might appear that the minimisation in 2 (a) is computationally intensive as it involves both a loop over j and for each $s \in \mathcal{S}_j$ performing a least squares regression.



(a) Rectangular regions constructed using the regression tree algorithm fitted to a dataset with two predictors with numbers indicating the order in which the splits were made. Also shown are the contours of the true regression function $\mathbb{E}(Y \mid X = x)$.

(b) Visualisation of the fitted regression tree. The percentages give the proportion of data in the corresponding region and also given is the average of the responses corresponding to those points.

To see how the computations may be arranged efficiently, let us consider, for notational simplicity, the first split, so $I = \{1, \dots, n\}$, and where $p = 1$.

Suppose that the $\{X_i\}_{i=1}^n$ are sorted so $X_1 < X_2 < \cdots < X_n$. The minimisation problem above is equivalent to finding m to minimise $Q_m + P_m$ where

$$Q_m := \min_{\gamma_L \in \mathbb{R}} \sum_{i \le m} (Y_i - \gamma_L)^2 \qquad \text{and} \qquad P_m := \min_{\gamma_R \in \mathbb{R}} \sum_{i > m} (Y_i - \gamma_R)^2.$$

Note that

$$Q_m = \sum_{i \le m} \left(Y_i - \frac{1}{m} \sum_{i \le m} Y_i \right)^2 = \sum_{i \le m} Y_i^2 - \frac{1}{m} \left(\sum_{i \le m} Y_i \right)^2,$$

with a similar decomposition for P_m. Thus

$$P_m + Q_m = \sum_{i=1}^n Y_i^2 - \frac{1}{m} \left(\sum_{i \le m} Y_i \right)^2 - \frac{1}{n-m} \left(\sum_{i > m} Y_i \right)^2.$$

As the first term does not depend on m, we may equivalently maximise

$$\frac{1}{m} \left(\sum_{i \le m} Y_i \right)^2 + \frac{1}{n-m} \left(\sum_{i > m} Y_i \right)^2$$

over m. Let $A_m := \sum_{i \le m} Y_i$ and $B_m := \sum_{i > m} Y_i$. Then $A_{m+1} = A_m + Y_{m+1}$ and $B_{m+1} = B_m - Y_{m+1}$. Thus all $A_1, \dots, A_{n-1}$ and $B_1, \dots, B_{n-1}$ may be computed in $O(n)$ operations.

Thus we may compute the display above for all $m = 1, \ldots, n-1$ in $O(n)$ operations, and hence we may minimise it over m with the same cost.

In order to use a decision tree is practice, one must choose the number of regions J: a large J might result in overfitting, while a small J may result in a large bias. Choosing J may be done via cross-validation. An alternative (typically preferred) approach is to grow a very large tree, and then collapse regions together according to a pruning strategy; we do not discuss this here.

2.2 Random forests

Whilst decision trees as above are a useful machine learning method in their own right, they have a few disadvantages:

- The piecewise constant estimated regression functions they fit, while useful for visualisation purposes (see Figure 1b) might not always deliver the best prediction error particularly when the true regression function varies smoothly with the predictors.
- The process of building a tree is greedy and unstable. As a consequence, small changes in the training data may lead to a very different tree; that is a fitted tree can have high variance (over the training data).

The *Random forest* procedure is a highly successful algorithm that aims to remedy these two deficiencies, though as we shall see, it does sacrifice interpretability of the fitted regression function.

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 $\bar{T}$ be given by $\bar{T}(x) = \mathbb{E}\hat{T}_D(x)$ and let (X, Y) be independent of D with $(X, Y) \overset{d}{=} (X_1, Y_1)$.

Recall the decomposition of the expected risk (1.5) in Section 1.4:

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

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

One 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 $\mathrm{Corr}(\hat{T}^{(b_1)}(x), \hat{T}^{(b_2)}(x)) = \rho \geq 0$. Then

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

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 $\mathrm{Var}(f_{\mathrm{rf}}(x))$ to decrease[3] with m_{try}. On the other hand, we would expect the squared bias to increase as m_{try} is decreased. An appropriate value of m_{try} may be selected using cross-validation.

3 Statistical learning theory

In a regression setting, using OLS with a set of d basis functions as in Example 1 to give $\hat{h}_D$ (where $D = (X_{1:n}, Y_{1:n})$ is the training data) yields

$$\mathbb{E}R(\hat{h}_D) - \mathbb{E}R(\widetilde{h}_{X_{1:n}}) \approx \frac{\sigma^2 d}{n}, \tag{3.1}$$

assuming $\sigma^2 := \mathrm{Var}(Y \,|\, X = x)$ is constant in x (see example sheet and the discussion in Section 1.4).

Our goal now is to study a roughly analogous quantity to the LHS of (3.1) in the classification setting. For an ERM $\hat{h}$ over a class $\mathcal{H}$, in general, $x \mapsto \mathbb{E}(\hat{h}_D(x) \,|\, X_{1:n})$ will not be a classifier. Instead, we may compare the risk or expected risk of $\hat{h}_D = \hat{h}$ to

$$h^* := \operatorname*{arg\,min}_{h \in \mathcal{H}} R(h),$$

the best[4] hypothesis in $\mathcal{H}$.

The quantity $R(\hat{h}) - R(h^*)$ is sometimes known as the *excess risk*. Some questions of interest are:

- How does the 'complexity' of $\mathcal{H}$ influence the excess risk?
- How does a change in the size n of the data affect the excess risk?

[3] In actual fact, the story may not be quite so simple as it is not entirely clear how $\mathrm{Var}(\hat{T}^{(1)}(x))$ will behave as m_{try} is varied.

[4] 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 to follow will carry through. In fact similar reasoning is applicable to the ERM $\hat{h}$.

Statistical learning theory is the branch of machine learning devoted to these sorts of considerations and in this course we aim to provide an introduction to some of the key ideas in this area. Our starting point is the following decomposition of the excess risk:

$$\begin{aligned} R(\hat{h}) - R(h^*) &= R(\hat{h}) - \hat{R}(\hat{h}) + \underbrace{\hat{R}(\hat{h}) - \hat{R}(h^*)}_{\leq 0} + \hat{R}(h^*) - R(h^*) \\ &\leq \sup_{h \in \mathcal{H}} \{R(h) - \hat{R}(h)\} + \hat{R}(h^*) - R(h^*). \end{aligned}$$

We wish to bound either the tail probability or the expectation of the excess risk. To motivate the developments to follow, consider the former case, for which it would be helpful to upper bound

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} \{R(h) - \hat{R}(h)\} > t\right)$$

for a given $t \geq 0$. Consider, for the time being, the setting where $|\mathcal{H}|$ is finite; ultimately we would like to tackle the case where $|\mathcal{H}|$ is infinite. A *union bound* gives

$$\begin{aligned} \mathbb{P}\left(\max_{h \in \mathcal{H}} \{R(h) - \hat{R}(h)\} > t\right) &= \mathbb{P}(\cup_{h \in \mathcal{H}} \{R(h) - \hat{R}(h) > t\}) \\ &\leq \sum_{h \in \mathcal{H}} \mathbb{P}(R(h) - \hat{R}(h) > t). \end{aligned} \tag{3.2}$$

Now for each fixed $h \in \mathcal{H}$,

$$R(h) - \hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} [\mathbb{E}\{\ell(h(X_i), Y_i)\} - \ell(h(X_i), Y_i)]$$

is an average of n i.i.d. mean-zero random variables. The central limit theorem (CLT) would suggest that $\sqrt{n}\{R(h) - \hat{R}(h)\}$ should behave like a $N(0, \mathrm{Var}(\ell(h(X_1), Y_1)))$-distributed random variable. However, in order to make use of this to bound (3.2), we would need a *uniform* limiting result for all $h \in \mathcal{H}$. In order to trade off bias and variance favourably, we may wish to increase the complexity of $\mathcal{H}$, i.e. the size of $|\mathcal{H}|$, for large n, so it is not at all clear that such a uniform result should hold. Moreover, in order for (3.2) to be small, we would need to consider t fairly large, so we would need such a limiting result to provide a good approximation in the far right tail of the distribution of $\sqrt{n}\{R(h) - \hat{R}(h)\}$. Such desiderata go far beyond what is offered by the CLT, and instead we turn to concentration inequalities, an important area of probability theory that (for example) can provide nonasymptotic tail bounds that mimic what we would have liked to obtain from the CLT, for averages of certain types of independent random variables.

3.1 Sub-Gaussianity and Hoeffding's inequality

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

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

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

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

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

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

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

$$\mathbb{E}e^{\alpha W} = e^{\alpha^2\sigma^2/2}. \tag{3.3}$$

Thus for $t \geq 0$,

$$\mathbb{P}(W \geq t) \leq \inf_{\alpha>0} e^{\alpha^2\sigma^2/2-\alpha t} = e^{-t^2/(2\sigma^2)}. \tag{3.4}$$

△

Note that to arrive at this bound, all we required was (an upper bound on) the moment generating function (mgf) of W (3.3). 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 (3.4) 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 (3.4) 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 3. A *Rademacher* random variable ε takes values $\{-1, 1\}$ with equal probability. It is sub-Gaussian with parameter $\sigma = 1$:

$$\begin{aligned}\mathbb{E}e^{\alpha\varepsilon} = \frac{1}{2}(e^{-\alpha} + e^{\alpha}) &= \frac{1}{2}\left(\sum_{k=0}^{\infty}\frac{(-\alpha)^k}{k!} + \sum_{k=0}^{\infty}\frac{\alpha^k}{k!}\right) \\ &= \sum_{k=0}^{\infty}\frac{\alpha^{2k}}{(2k)!} \\ &\leq \sum_{k=0}^{\infty}\frac{\alpha^{2k}}{2^k k!} = e^{\alpha^2/2} \quad \text{(using } (2k)! \geq 2^k k!\text{)}. \qquad (3.5)\end{aligned}$$

△

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

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

Proof. We will prove a weaker result here with $\sigma = b - a$. Let W' be an independent copy of W. We have

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

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

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

We now apply our previous result (3.5) 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 $W_1, \dots, W_n$ are independent and each W_i is sub-Gaussian with parameter σ_i. Then for $\gamma \in \mathbb{R}^n$, $\gamma^\top W$ is sub-Gaussian with parameter $\Big(\sum_i \gamma_i^2 \sigma_i^2\Big)^{1/2}$.*

Proof.

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

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

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (W_i - \mathbb{E}W_i) \ge t\right) \le \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \quad \text{for } t \ge 0, \tag{3.6}$$

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.

Proposition 5. *Suppose $W_1, \dots, W_d$ are all mean-zero and sub-Gaussian with parameter $\sigma > 0$ (but are not necessarily independent). Then*

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

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

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

Now

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

Thus

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

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

3.2 Finite hypothesis classes

Recall that

$$\begin{aligned} R(\hat{h}) - R(h^*) &= R(\hat{h}) - \hat{R}(\hat{h}) + \underbrace{\hat{R}(\hat{h}) - \hat{R}(h^*)}_{\leq 0} + \hat{R}(h^*) - R(h^*) \\ &\leq \sup_{h \in \mathcal{H}} \{R(h) - \hat{R}(h)\} + \hat{R}(h^*) - R(h^*). \end{aligned}$$

In the case where $\mathcal{H}$ is finite, Proposition 5 can be used to obtain a bound on the expected excess risk. We can also obtain the following tail bound.

Theorem 6. *Consider the classification setting with misclassification loss. Suppose $\mathcal{H}$ is finite. Then with probability at least $1-\delta$, the ERM $\hat{h}$ satisfies*

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

Proof. For each h, $\hat{R}(h)$ is an average of i.i.d. quantities of the form $\ell(h(X_i), Y_i)$ taking values in $[0,1]$. 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

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

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 first 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,

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

Thus

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

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

Example 4. Consider a simple classification setting with $X_i \in [0,1)^2$. Let us partition $[0,1)^2$ into m^2 disjoint squares $R_1, \dots, R_{m^2} \subset [0,1)^2$ of the form $[r/m, (r+1)/m) \times [s/m, (s+1)/m)$ for $r, s = 0, \dots, m-1$. Let

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

and define the 'histogram classifier'

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

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 $R_1, \dots, R_{m^2}$. The result above tells us that the generalisation error of $\hat{h}^{\text{hist}}$ satisfies

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

[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.] △

3.3 Rademacher complexity

Whilst a union bound in combination with Hoeffding's inequality and Proposition 5 were enough to give bounds on the risk and expected risk respectively in the case where $|\mathcal{H}| < \infty$, we will need some new ideas to tackle the case where $|\mathcal{H}| = \infty$. In this course, we will consider bounds on the expected risk; in fact, using a more powerful concentration inequality, the so-called bounded differences inequality, this analysis may be extended to give high probability bounds on the risk as well, but we do not pursue this here.

Recall our setup: $\mathcal{H}$ is a (now possibly infinite) hypothesis class and we have

$$\mathbb{E}R(\hat{h}) - R(h^*) \leq \mathbb{E}G \qquad \text{where} \qquad G := \sup_{h \in \mathcal{H}} \{R(h) - \hat{R}(h)\}.$$

Let us write $Z_i = (X_i, Y_i)$ for $i = 1, \dots, n$ and

$$\mathcal{F} := \{(x, y) \mapsto -\ell(h(x), y) : h \in \mathcal{H}\}. \tag{3.7}$$

Then we have

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

The following definitions apply to a general function class $\mathcal{F}$ not necessarily coming from (3.7).

Definition 2. Let $\mathcal{F}$ be a class of real-valued functions $f : \mathcal{Z} \to \mathbb{R}$ and let $z_1, \dots, z_n \in \mathcal{Z}$.

- Let
$$\mathcal{F}(z_{1:n}) := \{(f(z_1), \dots, f(z_n)) : f \in \mathcal{F}\},$$
be the class of 'behaviours' of $\mathcal{F}$ on $z_{1:n}$.

- Define the *empirical Rademacher complexity*
$$\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) := \mathbb{E}\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(z_i)\right), \tag{3.8}$$
where $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. Rademacher random variables. Note that $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n}))$ is well-defined in that the right-hand side of (3.8) only depends on $\mathcal{F}(z_{1:n})$.

  Given i.i.d. random variables $Z_1, \dots, Z_n$ taking values in $\mathcal{Z}$, we sometimes view the empirical Rademacher complexity as a random variable:
$$\hat{\mathcal{R}}(\mathcal{F}(Z_{1:n})) := \mathbb{E}\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(Z_i) \,\middle|\, Z_{1:n}\right).$$
  Some intuition: Consider a classification problem with inputs $Z_1, \dots, Z_n$ and *completely random* labels $\varepsilon_1, \dots, \varepsilon_n$. The empirical Rademacher complexity then captures how closely aligned the 'predictions' $f(Z_i)$ are to the random labels.

- Define the *Rademacher complexity* of $\mathcal{F}$, $\mathcal{R}_n(\mathcal{F}) := \mathbb{E}\hat{\mathcal{R}}(\mathcal{F}(Z_{1:n}))$.

Theorem 7. *Let $\mathcal{F}$ be a class of real-valued functions $f : \mathcal{Z} \to \mathbb{R}$ and let $Z_1, \dots, Z_n$ be i.i.d. random variables taking values in $\mathcal{Z}$. Then*
$$\mathbb{E}\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \{f(Z_i) - \mathbb{E}f(Z_i)\}\right) \leq 2\mathcal{R}_n(\mathcal{F}).$$

Before we prove Theorem 7, let us reflect on what it might achieve. Considering our main problem of bounding $\mathbb{E}G$, a key challenge is that it depends in a complicated way on the unknown P_0. Key point: $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n}))$ does not depend on P_0, and 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. We now turn to the proof of Theorem 7, which uses a symmetrisation technique.

Proof of Theorem 7. Let us introduce an independent copy $(Z_1', \dots, Z_n')$ of $(Z_1, \dots, Z_n)$. We have
$$\begin{aligned}\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \{f(Z_i) - \mathbb{E}f(Z_i)\} &= \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\{f(Z_i) - f(Z_i') \mid Z_{1:n}\} \text{ (independence of } Z_{1:n} \text{ and } Z_{1:n}') \\ &\leq \mathbb{E}\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \{f(Z_i) - f(Z_i')\} \,\middle|\, Z_{1:n}\right).\end{aligned}$$

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

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

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

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

An immediate consequence of Theorem 7 is the following bound on the expected risk (note we do not require the loss to be misclassification loss here).

Theorem 8 (Expected risk bound based on Rademacher complexity)**.** *Let* $\mathcal{F} := \{(x, y) \mapsto \ell(h(x), y) : h \in \mathcal{H}\}$. *Then*

$$\mathbb{E} R(\hat{h}) - R(h^*) \leq 2\mathcal{R}_n(\mathcal{F}).$$

Proof. From the above and Theorem 7, we have $\mathbb{E} R(\hat{h}) - R(h^*) \leq 2\mathcal{R}_n(-\mathcal{F})$, where $-\mathcal{F} = \{-f : f \in \mathcal{F}\}$. However, as for Rademacher random variables $\varepsilon_1, \ldots, \varepsilon_n$, $\varepsilon_{1:n} \overset{d}{=} -\varepsilon_{1:n}$, we have $\mathcal{R}_n(-\mathcal{F}) = \mathcal{R}_n(\mathcal{F})$. □

3.4 VC dimension

All we need to do in order to bound the expected risk 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 $|\mathcal{F}(z_{1:n})|$ of behaviours 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 $z_i = (x_i, y_i)$. Indeed, there is a bijection

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

Lemma 9. *We have* $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) \leq \sqrt{2 \log(|\mathcal{F}(z_{1:n})|)/n} = \sqrt{2 \log(|\mathcal{H}(x_{1:n})|)/n}$.

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

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

where $\varepsilon_{1:n}$ are i.i.d. Rademacher random variables. Then $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) = \mathbb{E}\max_j W_j$. By Lemma 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 $|\mathcal{H}(x_{1:n})| \leq 2^n$. Considering the result above, an interesting case then is when $|\mathcal{H}(x_{1:n})|$ is growing slower than exponentially in n, e.g. growing polynomially in n.

Definition 3. Let $\mathcal{H}$ be a class of functions $h : \mathcal{X} \to \{a, b\}$ with $a \neq b$ (e.g. $\{a, b\} = \{-1, 1\}$) with $|\mathcal{H}| \geq 2$.

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

If $|\mathcal{H}(x_{1:n})| = 2^n$, then $|\mathcal{H}(x_{1:m})| = 2^m$ for all $1 \leq m \leq n$. Thus to show $\mathrm{VC}(\mathcal{H}) = n$ we need to (i) exhibit an $x_{1:n}$ that is shattered (usually the easier part), and (ii) show that no $x_{1:n+1}$ can be shattered. Note that as no set of points that are not distinct can be shattered, we can always restrict attention to sets of distinct points.

Example 5. Let $\mathcal{X} = \mathbb{R}$ and consider $\mathcal{H} = \{h_{a,b} : h_{a,b}(x) = \mathbb{1}_{[a,b)}(x) : a, b, \in \mathbb{R}\}$. Consider n distinct points $x_1, \ldots, x_n$. These divide up the real line into $n + 1$ intervals $(-\infty, x_1], (x_1, x_2], \ldots, (x_{n-1}, x_n], (x_n, \infty)$. Now if a and a' are in the same interval, and b and b' are in the same interval, then $(h_{a,b}(x_i))_{i=1}^n = (h_{a',b'}(x_i))_{i=1}^n$. Thus every possible behaviour $(h_{a,b}(x_i))_{i=1}^n$ can be obtained by picking one of the $n + 1$ intervals for each of a and b, so

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

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

It is a bit tedious to determine the shattering coefficient individually for each $\mathcal{H}$ and see whether it grows polynomially; we would like a more streamlined approach. Observe that in the previous example, we have $s(\mathcal{H}, n) \leq (n + 1)^{\mathrm{VC}(\mathcal{H})}$. 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 10 (Sauer–Shelah)**.** *Let $\mathcal{H}$ be a class with finite VC dimension d. Then*

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

What is striking about this result is that whilst we know from the definition that for all $n > d$, $s(\mathcal{H}, n) < 2^n$, it is not immediately obvious that we cannot have $s(\mathcal{H}, n) = 2^n - 1$, or $s(\mathcal{H}, n) = 1.8^n$ for $n > d$. The result shows that beyond d the growth of $s(\mathcal{H}, n)$ is radically different in that it is polynomial. The important consequence of this is that from Lemma 9 we have

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

where $\mathcal{F}$ is the loss class associated with $\mathcal{H}$.

**Proof of Lemma 10*.* We will prove the following stronger result. Fix $x_{1:n} \in \mathcal{X}^n$ and let x_Q for any non-empty $Q = \{i_1, \ldots, i_{|Q|}\} \subseteq \{1, \ldots, n\}$ be $(x_{i_1}, \ldots, x_{i_{|Q|}})$. Then we claim that there are at least $|\mathcal{H}(x_{1:n})| - 1$ non-empty sets $Q \subseteq \{1, \ldots, n\}$ such that $\mathcal{H}$ 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 $|\mathcal{H}(x_{1:n})| = s(\mathcal{H}, n)$. As $\mathrm{VC}(\mathcal{H}) = d$, by definition no x_Q with $|Q| > d$ can be shattered, so from the claim,

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

But then assuming wlog that $n \geq d$, we have

$$\begin{aligned}\sum_{i=1}^{d} \binom{n}{i} &\leq n + n^2 + \cdots + n^d \\ &\leq n^d + \binom{d}{1} n^{d-1} + \binom{d}{2} n^{d-2} + \cdots + \binom{d}{d-1} n + \binom{d}{d} - 1 = (n+1)^d - 1.\end{aligned}$$

It remains to prove the claim, which we do by induction on $|\mathcal{H}(x_{1:n})|$. Wlog assume the functions in $\mathcal{H}$ map to $\{-1, 1\}$. The claim when $|\mathcal{H}(x_{1:n})| = 1$ 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 $\mathcal{H}$ with $|\mathcal{H}(x_{1:n})| \leq k$. We will show the result holds at $k+1$. Take any $n \in \mathbb{N}$, $x_{1:n} \in \mathcal{X}^n$ and $\mathcal{H}$ with $|\mathcal{H}(x_{1:n})| = k+1$. Let x_j be such that $\mathcal{H}_+ := \{h \in \mathcal{H} : h(x_j) = 1\}$ and $\mathcal{H}_- := \{f \in \mathcal{H} : h(x_j) = -1\}$ are both non-empty (which is possible as $|\mathcal{H}(x_{1:n})| \geq 2$). Then

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

Let $\mathcal{X}_-$ and $\mathcal{X}_+$ be the sets of subvectors x_Q that are shattered by $\mathcal{H}_-$ and $\mathcal{H}_+$ 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 $\mathcal{H} \supset \mathcal{H}_1, \mathcal{H}_+$. 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 $h(x_j)$ 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 $\mathcal{H}$. Also x_j itself is shattered by $\mathcal{H}$. Thus we see that the number of sets shattered by $\mathcal{H}$ 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 6. Let $\mathcal{X} = \mathbb{R}^p$ and consider $\mathcal{H} = \{\mathbb{1}_A : A \in \mathcal{A}\}$ where $\mathcal{A} = \big\{ \prod_{j=1}^p (-\infty, a_j] : a_1, \dots, a_p \in \mathbb{R}\big\}$. To compute $\mathrm{VC}(\mathcal{H})$, first note that the set of standard basis vectors $e_1, \dots, e_p \in \mathbb{R}^p$ is shattered as for any $I \subseteq \{1, \dots, p\}$, 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 $x_1, \dots, x_{p+1} \in \mathbb{R}^p$ and let π_j be the jth coordinate function, so $\pi_j(x_i) = x_{ij}$, where x_{ij} is the jth component of x_i. Then by the pigeonhole principle, there must be some x_{k^*} that is not the unique maximiser of any of the π_j over $x_1, \dots, x_{p+1}$. But then for each $j = 1, \dots, p$, there exists some x_{k_j} such that $x_{k_j j} \geq x_{k^* j}$, so for $h \in \mathcal{H}$ we can never have $h(x_{k^*}) = 0$ and $h(x_k) = 1$ for all $k \neq k^*$. Thus $\mathrm{VC}(\mathcal{H}) = p$. △

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

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

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

$$\mathcal{H} = \{h : h(x) = \mathrm{sgn}(f(x)) \text{ where } f \in \mathcal{F}\}. \tag{3.9}$$

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

Proposition 11. *Consider hypothesis class $\mathcal{H}$ given by* (3.9) *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.

$$\sum_{i:\gamma_i > 0} \gamma_i f(x_i) + \sum_{i:\gamma_i \leq 0} \gamma_i f(x_i) = 0 \quad \text{for all } f \in \mathcal{F}, \tag{3.10}$$

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 $\mathrm{sgn}(0) := -1$) as otherwise the LHS of (3.10) would be strictly positive. Thus $x_{1:d}$ cannot be shattered so $\mathrm{VC}(\mathcal{H}) \leq d - 1$ as required. □

4 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.

4.1 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 4.

- 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 $v_1, \ldots, v_m \in \mathbb{R}^d$ if

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

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

Lemma 12. *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 $v_1, \ldots, v_{m+1} \in S$ and $\alpha_1, \ldots, \alpha_{m+1} \geq 0$ 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

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

Lemma 13. *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. □

4.2 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

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

Then $-f$ is a *concave* function. It is *strictly convex* if the inequality is strict for all $x, y \in C$, $x \neq y$ and $t \in (0,1)$. For example, any norm $\|\cdot\| : \mathbb{R}^d \to [0, \infty)$ is convex as by the triangle inequality, for all $t \in (0,1)$,

$$\|(1-t)x + ty\| \leq \|(1-t)x\| + \|ty\| = (1-t)\|x\| + t\|y\|.$$

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 $t > 0$ sufficiently small, then by convexity

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

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

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

Proposition 14. *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) For all $M \in \mathbb{R}$, *the sublevel set* $\{x \in C : f(x) \leq M\}$ *is convex.*

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

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

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

Checking convexity:

(viii) 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.

4.3 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:

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

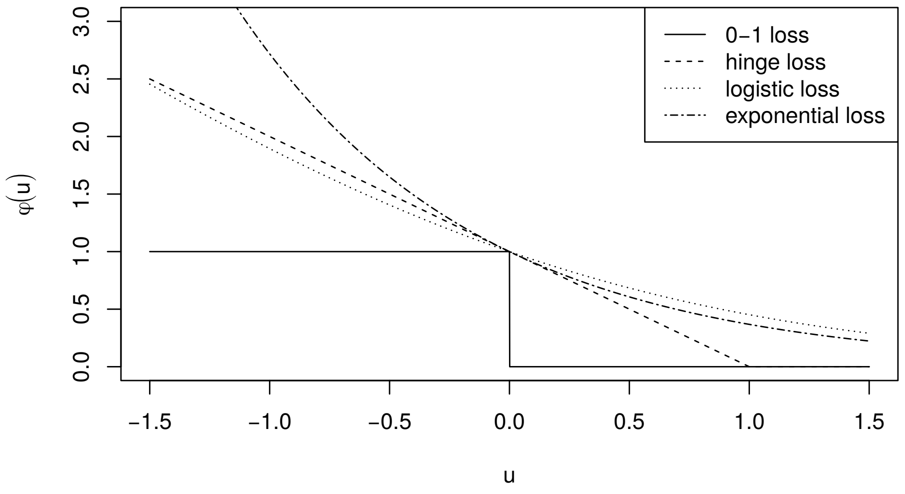
(ignoring when $X_i^\top \beta = 0$). The RHS is not convex and in fact not continuous due to the indicator function. If $\mathbb{1}_{(-\infty,0]}$ above were somehow replaced with a convex function, we know from Proposition 14 (i) & (ii) that the resulting objective would be a convex function of β. The minimising $\hat{\beta}$ may still be able to deliver classification performance via $x \mapsto \operatorname{sgn}(x^\top \hat{\beta})$ 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 $x \mapsto \operatorname{sgn}(h(x))$.

- We will consider loss functions $\ell : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ of the form

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



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

$$\hat{R}_\phi(h) := \frac{1}{n}\sum_{i=1}^n \phi(Y_i h(X_i))$$

for the corresponding empirical ϕ-risk.

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 $x \mapsto \text{sgn}(\eta(x) - 1/2)$. 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 $\text{sgn} \circ h_{\phi,0}$ achieving the optimal Bayes misclassification risk, but also allows for the possibility that $\inf_h R_\phi(h)$ is not attained.

Definition 5. 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 15. *Let $\phi : \mathbb{R} \to [0, \infty)$ be convex. Then ϕ 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 14 (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 $C_\eta(\alpha^*) < C_\eta(0) \leq \inf_{\alpha \leq 0} C_\eta(\alpha)$. Next note that $C_{1/2+\theta}(\alpha) = C_{1/2-\theta}(-\alpha)$ for $\theta \in [0, 1/2]$, so similarly when $\eta < 1/2$, there exists some $\alpha^* < 0$ with $C_\eta(\alpha^*) < C_\eta(0) \leq \inf_{\alpha \geq 0} C_\eta(\alpha)$. 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.

4.4 Rademacher complexity revisited

One remaining issue is whether we can obtain guarantees on when the expected ϕ-risk is small. Theorem 8 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 16 (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 for any $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{-1, 1\}$, writing $z_i = (x_i, y_i)$, we have $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) \leq L\hat{\mathcal{R}}(\mathcal{H}(x_{1:n}))$, so in particular $\mathcal{R}_n(\mathcal{F}) \leq L\mathcal{R}_n(\mathcal{H})$.*

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

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

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 $A : \mathcal{H} \times \{-1, 1\}^{n-1}$,

$$\mathbb{E} \sup_{h \in \mathcal{H}} \left(\frac{1}{n}\varepsilon_i \phi(y_i h(x_i)) + A(h, \varepsilon_{-i})\right) \le \mathbb{E} \sup_{h \in \mathcal{H}} \left(\frac{L}{n}\varepsilon_i h(x_i) + A(h, \varepsilon_{-i})\right). \tag{4.1}$$

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

$$\mathbb{E} \sup_{h \in \mathcal{H}} \left(\frac{1}{n}\varepsilon_1 \phi(y_1 h(x_1)) + \frac{1}{n}\sum_{i=2}^{n} \varepsilon_i \phi(y_i h(x_i))\right) \le \mathbb{E} \sup_{h \in \mathcal{H}} \left(\frac{L}{n}\varepsilon_1 h(x_1) + \frac{1}{n}\sum_{i=2}^{n} \varepsilon_i \phi(y_i h(x_i))\right). \tag{4.2}$$

Next applying (4.1) 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 (4.2) 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

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

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

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

Taking expectations proves the claim. □

Corollary 17. *Consider the setup of Lemma 16. Then with $\hat{h}$ the ERM with ϕ-loss (so in a classification setup, the corresponding classifier would be* $\operatorname{sgn} \circ \hat{h}$*), and $h^* \in \arg\min_{h \in \mathcal{H}} R_\phi(h)$,*

$$\mathbb{E} R_\phi(\hat{h}) - R_\phi(h^*) \leq 2L\mathcal{R}_n(\mathcal{H}).$$

In order for the result above to be applicable, we need $\mathcal{H}$ to be such that $\mathcal{R}_n(\mathcal{H})$ is finite. This will not necessarily hold for our example where $\mathcal{X} = \mathbb{R}^p$ of

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

However, if we constrain the norm of the β and $\mathcal{X}$ is a bounded subset of $\mathbb{R}^p$, we can achieve this. Such considerations lead to two hugely important classes of machine learning methods, those based on constraining the ℓ_2-norm and those constraining the ℓ_1-norm.

4.5 ℓ_2-constraint

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

$$\mathcal{H} = \{x \mapsto x^\top \beta : \beta \in \mathbb{R}^p \text{ and } \|\beta\|_2 \leq \lambda\} \tag{4.3}$$

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

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

where the last inequality follows due to concavity of $\sqrt{\cdot}$ and Jensen's inequality. Now for $i \neq j$, $\mathbb{E}(\varepsilon_i x_i^\top x_j \varepsilon_j) = 0$, so

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

Thus

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

In fact, more generally if $\mathcal{X} = \mathbb{R}^p$ but $\mathbb{E}\|X_i\|_2^2 \leq C$, we have

$$\mathcal{R}_n(\mathcal{H}) \leq \frac{\lambda C}{\sqrt{n}}.$$

Example 7. Take ϕ to be the hinge loss and $\mathcal{H}$ given by (4.3); the corresponding ERM $\hat{h}$ is then known as the *support vector classifier*. Note that the hinge loss is Lipschitz with constant 1. From Corollary 17,

$$\mathbb{E}R_\phi(\hat{h}) - R_\phi(h^*) \leq \frac{2\lambda C}{\sqrt{n}}$$

△

4.6 ℓ_1-constraint

The ℓ_1-norm of a vector u is $\|u\|_1 := \sum_i |u_i|$. Suppose now that

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

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

Lemma 18. *For any $A \subseteq \mathbb{R}^n$, $\hat{\mathcal{R}}(A) = \hat{\mathcal{R}}(\operatorname{conv} A)$.*

Proof. See example sheet. □

Note that here

$$\hat{\mathcal{R}}(A) := \mathbb{E}\left(\sup_{a \in A} \frac{1}{n}\sum_{i=1}^n \varepsilon_i a_i\right),$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables.

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

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

Then $\mathcal{H}(x_{1:n}) = L(\operatorname{conv} S) = \operatorname{conv} L(S)$ from Lemma 13. Thus from Lemma 18 we have

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

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables. Now by Proposition 4, each $\pm\sum_i \varepsilon_i x_{ij}$ is sub-Gaussian with parameter

$$\left(\sum_{i=1}^n x_{ij}^2\right)^{1/2}.$$

Thus from Proposition 5 we have

$$\hat{\mathcal{R}}(\mathcal{H}(x_{1:n})) \leq \frac{\lambda}{n} \times \left(\sum_{i=1}^n x_{ij}^2\right)^{1/2} \times \sqrt{2\log|S|} = \left(\frac{1}{n}\sum_{i=1}^n x_{ij}^2\right)^{1/2} \frac{\lambda}{\sqrt{n}}\sqrt{2\log(2p)}.$$

Now if $\mathcal{X} = [-C, C]^p$, we have (using $\mathbb{E}(U^2) \geq (\mathbb{E}U)^2$) that

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

Example 8. Take ϕ to be the hinge loss and $\mathcal{H}$ as above. Suppose $\mathcal{X} = [-1, 1]^p$. Then from Corollary 17,

$$\mathbb{E}R_\phi(\hat{h}) - R_\phi(h^*) \leq 2\lambda\sqrt{\frac{2\log(2p)}{n}}.$$

In contrast, with $\mathcal{H}$ given by the ℓ_2-constraint (4.3) 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 $h_0 : x \mapsto x^\top\beta^0$ minimises R_ϕ over $\{x \mapsto x^\top\beta : \beta \in \mathbb{R}^p\}$.
  - If
  $$\beta^0 = \left(\frac{1}{\sqrt{p}}, \ldots, \frac{1}{\sqrt{p}}\right)^\top,$$
  in order that $\beta^0 \in \mathcal{H}_1, \mathcal{H}_2$, we require $\lambda_1 \geq \sqrt{p}$ and $\lambda_2 \geq 1$. These choices give expected excess risk bounds of order
  $$\ell_1 : \quad \sqrt{\frac{p\log p}{n}}, \qquad \ell_2 : \quad \sqrt{\frac{p}{n}}.$$
  - If
  $$\beta^0 = \Big(\underbrace{\frac{1}{\sqrt{s}}, \ldots, \frac{1}{\sqrt{s}}}_{s \text{ of these}}, 0, \ldots, 0\Big)^\top,$$
  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.

△

4.7 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 19. *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*

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

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

$$(x - \pi_C(x))^\top (z - \pi_C(x)) \leq 0 \quad \textit{for all } z \in C, \tag{4.4}$$

$$\|\pi_C(x) - \pi_C(z)\|_2 \leq \|x - z\|_2 \quad \textit{for all } z \in \mathbb{R}^d. \tag{4.5}$$

Proof. **Existence**: Let $\mu = \inf_{z \in C} \|x - z\|_2$. Write $B = \{w : \|w - x\|_2 \leq \mu + 1\}$. 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).
(4.4): We have $(1 - t)\pi + tz \in C$ for all $t \in [0, 1]$, so

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

whence

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

Letting $t \to 0$ shows (4.4).
(4.5): From (4.4) we have

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

Adding these we have

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

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

Definition 6. We call $\pi_C(x)$ above the *projection of x on C*.

4.8 Subgradients

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

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

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

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

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

Proposition 20. *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

$$v_k^\top w < v_k^\top w_k \text{ for all } w \in C. \tag{4.6}$$

Indeed taking $v_k = w_k - \pi_C(w_k)$, from Proposition 19, we have that $v_k^\top (w - \pi_C(w_k)) \leq 0$, so then

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

We can rescale the v_k such that $\|v_k\|_2 = 1$, and (4.6) 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

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

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

$$f(z) + g^\top z \geq f(x) + g^\top x \quad \text{for all } z.$$ □

To compute subgradients, the following facts will be helpful.

Proposition 21. *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

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

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

Proposition 22 (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) $\partial h(x) = \{A^\top g : g \in \partial f(Ax + b)\}$.

Example 9. Consider

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

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 22 (iii) writing $h_i(\beta) = \max(1 - y_i x_i^\top \beta, 0)$, we have $\partial h_i(\beta) = \{-y_i x_i t : t \in [0, 1]\}$ when $y_i x_i^\top \beta = 1$. From Proposition 22 (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 $y_i x_i^\top \beta$. △

4.9 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 $f(z) \approx f(\beta) + g^\top (z - \beta)$, so for small $\eta > 0$,

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

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 to 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 23. *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 $\sup_{\beta \in C} \|\beta\|_2 \leq R < \infty$ and $\sup_{\beta \in C} \sup_{g \in \partial f(\beta)} \|g\|_2 \leq L < \infty$. Then if $\eta_s \equiv \eta = 2R/(L\sqrt{k})$, the output $\bar{\beta}$ of the gradient descent algorithm above satisfies*

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

Proof. We have

$$\begin{aligned} f(\beta_s) - f(\hat{\beta}) &\leq g_s^\top(\beta_s - \hat{\beta}) \quad \text{(definition of subgradient)} \\ &= -\frac{1}{\eta}(z_{s+1} - \beta_s)^\top(\beta_s - \hat{\beta}) \\ &= \frac{1}{2\eta}\{\|\beta_s - z_{s+1}\|_2^2 + \|\beta_s - \hat{\beta}\|_2^2 - \|z_{s+1} - \hat{\beta}\|_2^2\}. \end{aligned} \tag{4.7}$$

From Proposition 19, $\|\pi_C(z) - \pi_C(x)\|_2 \leq \|z - x\|_2$, so in particular

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

Using this and (4.7),

$$f(\beta_s) - f(\hat{\beta}) \leq \frac{1}{2\eta}\{\eta^2\|g_s\|_2^2 + \|\beta_s - \hat{\beta}\|_2^2 - \|\beta_{s+1} - \hat{\beta}\|_2^2\}. \tag{4.8}$$

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 10. Consider ERM with hinge loss, $\mathcal{X} = \{x \in \mathbb{R}^p : \|x\|_2 \leq C\}$ and the ℓ_2-constrained hypothesis class $\mathcal{H} = \{x \mapsto x^\top \beta : \|\beta\|_2 \leq \lambda\}$. 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 23 we see that the output of gradient descent with step size $\eta = 2\lambda/(C\sqrt{k})$ satisfies $f(\bar{\beta}) - f(\hat{\beta}) \leq 2C\lambda/\sqrt{k}$. △

4.10 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 24. *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 $\tilde{f}(\beta;U_s) \geq \tilde{f}(\beta_s;U_s) + \tilde{g}_s^\top(\beta - \beta_s)$ 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 23,

$$\begin{aligned}
f(\beta_s) - f(\hat{\beta}) &\leq g_s^\top(\beta_s - \hat{\beta}) \\
&= \mathbb{E}(\tilde{g}_s(\beta_s - \hat{\beta}) \,|\, \beta_s) \\
&= -\frac{1}{\eta}\mathbb{E}\{(z_{s+1} - \beta_s)^\top(\beta_s - \hat{\beta}) \,|\, \beta_s\} \\
&= \frac{1}{2\eta}\mathbb{E}\{\|\beta_s - z_{s+1}\|_2^2 + \|\beta_s - \hat{\beta}\|_2^2 - \|z_{s+1} - \hat{\beta}\|_2^2 \,|\, \beta_s\} \\
&\leq \frac{1}{2\eta}\mathbb{E}\{\eta^2\|\tilde{g}_s\|_2^2 + \|\beta_s - \hat{\beta}\|_2^2 - \|\beta_{s+1} - \hat{\beta}\|_2^2 \,|\, \beta_s\} \text{ (Prop. 19).}
\end{aligned}$$

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. □

5 Popular machine learning methods II

5.1 Adaboost

Empirical risk minimisation is a technique for finding a single good hypothesis from a given hypothesis class. Alternatively, we could attempt to find a good weighted combination of hypotheses. Specifically, given a base 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 base class $\mathcal{B}$, and the construction above turns out to be a 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 over $\mathcal{H}$ 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.

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

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

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

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

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

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

Assuming that for each j, we know the order of the $\{X_{ij}\}_{i=1}^n$, weighted ERM over this class may be performed in $O(np)$ operations (see example sheet). △

5.2 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

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

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

$$x \mapsto \mathbb{E}\biggl(\frac{\partial \ell(\theta, Y)}{\partial \theta}\bigg|_{f_{m-1}(x)}\bigg| X = x\biggr).$$

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

$$\mathbb{E}\biggl(\frac{\partial \ell(\theta, Y)}{\partial \theta}\bigg|_{f_{m-1}(X)} - h(X)\biggr)^2 \tag{5.1}$$

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 (5.1) using regression, thereby approximating the conditional expectation. This regression is performed using some base regression method that takes as input some training data D and outputs a hypothesis $\hat{h}_D : \mathcal{X} \to \mathbb{R}$. In what follows, the loss ℓ may correspond to a differentiable convex surrogate or least squares loss for example. When ℓ is least squares loss, we have

$$W_i = \frac{\partial}{\partial \theta}(Y_i - \theta)^2\bigg|_{\theta = \hat{f}_{m-1}(X_i)} = -2(Y_i - \hat{f}_{m-1}(X_i)),$$

so W_i is proportional to the negative residuals.

In the next step of Algorithm 3, we regress these residuals back onto our $X_{1:n}$ using regression method $\hat{h}$.

Example 12. When ℓ is least squares, one choice of $\hat{h}_D$ is the ERM over the class

$$\{x \mapsto \mu + x_j\beta : \mu \in \mathbb{R},\ \beta \in \mathbb{R},\ j = 1, \ldots, p\}$$

given data D. Gradient boosting then amounts to repeatedly regressing the residuals onto the predictor that has the largest (in absolute value) sample correlation with them, and adding η times the fitted regression function to the current fit $\hat{f}_m$ (see example sheet).

Algorithm 3 Gradient boosting

```
**Input:** Data $X_{1:n}$, $Y_{1:n}$; $\eta > 0$; base regression method $\hat{h}$; 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 $\hat{h}$ to data $X_{1:n}, W_{1:n}$ to give $\hat{g}_m = \hat{h}_{(X_{1:n}, W_{1:n})} : \mathcal{X} \to \mathbb{R}$
    Update $\hat{f}_m = \hat{f}_{m-1} - \eta\hat{g}_m$
**end for**
**return** $\hat{f}_M$ (or $\text{sgn} \circ \hat{f}_M$ in the classification setting)
```

5.3 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

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

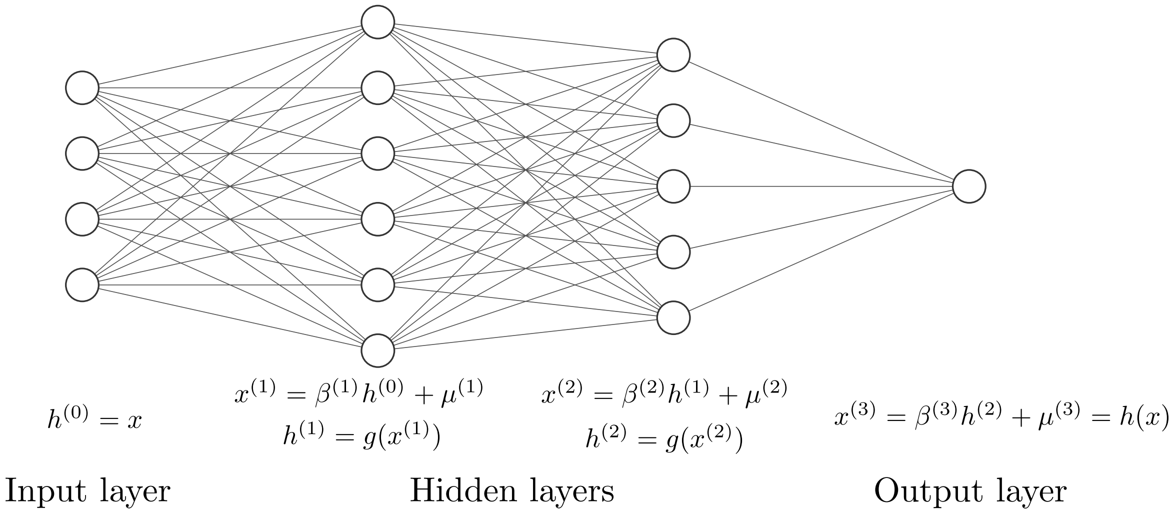
where

- 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 $v = (v_1, \ldots, v_m)^\top$, $g(v) = (\psi(v_1), \ldots, \psi(v_m))^\top$. 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.


$h^{(0)} = x$
$x^{(1)} = \beta^{(1)} h^{(0)} + \mu^{(1)}$
$h^{(1)} = g(x^{(1)})$
$x^{(2)} = \beta^{(2)} h^{(1)} + \mu^{(2)}$
$h^{(2)} = g(x^{(2)})$
$x^{(3)} = \beta^{(3)} h^{(2)} + \mu^{(3)} = h(x)$
Input layer
Hidden layers
Output layer

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

$$\frac{\partial z}{\partial x^{(d)}} = y\phi'(yx^{(d)})$$

$$\frac{\partial z}{\partial \mu^{(d)}} = \frac{\partial z}{\partial x^{(d)}}, \qquad \frac{\partial z}{\partial \beta^{(d)}_{1k}} = \frac{\partial z}{\partial x^{(d)}} h^{(d-1)}_k \tag{5.2}$$

$$\frac{\partial z}{\partial h^{(d-1)}_j} = \frac{\partial z}{\partial x^{(d)}} \, \beta^{(d)}_{1j}$$

$$\frac{\partial z}{\partial x^{(d-1)}_j} = \frac{\partial z}{\partial h^{(d-1)}_j} \, \psi'(x^{(d-1)}_j)$$

$$\frac{\partial z}{\partial \mu^{(d-1)}_j} = \frac{\partial z}{\partial x^{(d-1)}_j}, \qquad \frac{\partial z}{\partial \beta^{(d-1)}_{jk}} = \frac{\partial z}{\partial x^{(d-1)}_j} h^{(d-2)}_k \tag{5.3}$$

$$\frac{\partial z}{\partial h^{(d-2)}_j} = \sum_{k=1}^{m_d} \frac{\partial z}{\partial x^{(d-1)}_k} \, \beta^{(d-1)}_{kj},$$

$$\vdots$$

This process is known as *back propogation.* Note only (5.2) and (5.3) 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 $(\beta^{(k)}, \mu^{(k)})^d_{k=1}$.