MA3K1: Mathematics of machine learning

Clarice Poon[1],

February 3, 2025

[1]Mathematical Institute, Zeeman Building, University of Warwick, Coventry CV4 7AL, UK, clarice.poon@warwick.ac.uk

Some excellent resources

We will largely follow the book of Martin Lotz in the previous years, which is available on Moodle. Note that Martin's book has more material than will be covered in this module. These lecture notes will be updated as term progresses.

Some additional resources that these notes were based on are:

- Mathematics of machine learning lecture notes by Rajen Shah http://www.statslab.cam.ac.uk/~rds37/teaching/machine_learning/notes.pdf
- Foundations of machine learning by Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. https://cs.nyu.edu/~mohri/mlbook/
- Learning from first principles by Francis Bach https://www.di.ens.fr/~fbach/ltfp_book.pdf
- Mathematical foundations of data science by Gabriel Peyré. https://mathematical-tours.github.io/book-sources/chapters-pdf/machine-learning.pdf. There are also a series of computational exercises on the numerical tours website. http://www.numerical-tours.com/python/
- For coding, some useful packages are sklearn, pytorch or JAX.

Chapter 1

Statistical learning

1.1 Classification and regression

Suppose we are given a pair of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with joint probability measure P_0. We will call X the *input* vector or vector of features and call Y the *output* or response. To give some examples of what X, Y might represent:

- X is disease factors (age, genetic indicators, BMI, ...) and Y is the disease status.
- X is an email and Y is an indicator on whether an email is spam or not.
- X is the number of bedrooms or other features in a randomly selected house, and Y is the price.

Goal: Learn a function $h : \mathcal{X} \to \mathcal{Y}$ to *predict* the random variable Y from X. We call h a *hypothesis* or *estimator*.

Regression vs classification
If $Y \in \mathbb{R}$ is quantitative, then this learning problem is called a *regression* problem. If Y is categorical e.g. $\{\text{Yes}, \text{No}\}$, $\{-1, 1\}$ or $\{0, 1, \ldots, 9\}$, then we call this problem a *classification* problem.

The loss function To measure the quality of a prediction/hypothesis h, we introduce a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. To give a couple of examples of common loss functions

- For regression, one can consider $\ell(h(x), y) = (h(x) - y)^2$
- For classification, one can consider the *unit loss* $\ell(h(x), y) = \begin{cases} 0 & h(x) = y \\ 1 & h(x) \neq y \end{cases}$.

The goal is to choose a hypothesis (measurable function) h that minimises the **risk**: [1]

The Risk

$$R(h) \coloneqq \int_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \ell(h(x), y) dP_0(x, y) = \mathbb{E}[\ell(h(X), Y)]. \tag{1.1}$$

Remark 1. For the unit loss above, we can write $\ell(z, y) = \mathbf{1}\{z \neq y\}$ and

$$R(h) = \mathbb{E}[\mathbf{1}\{h(X) \neq Y\}] = \mathbb{P}(h(X) \neq Y). \tag{1.2}$$

Remark 2. In the case of regression, one can show that the h that minimizes $R(h)$ with $\ell(z, y) = (z - y)^2$ is

$$h(x) = \int y dP_0(x, y) = \mathbb{E}[Y|X = x].$$

We call h the **regression function**.

1.2 The Bayes classifier

Goal: Characterize the classifier that minimizes the misclassification risk (1.2).

We will assume throughout that $Y \in \{0, 1\}$. We will call $h_* \coloneqq \operatorname{argmin}_h R(h)$ the *Bayes classifier* [2]. In this section, we will characterize the Bayes classifier. Before embarking on this, we first review conditional expectation.

1.2.1 Conditional expectation

Let $Z \in \mathbb{R}$ and $W = (W_1, \ldots, W_d)^\top \in \mathbb{R}^d$ be random variables with joint probability density function $f_{z,w}$. Then, the conditional probability density function $f_{z|w}$ of Z given W satisfies

$$f_{z|w}(z|w) = \begin{cases} f_{z,w}(z, w)/f_w(w) & f_w(w) \neq 0 \\ 0 & f_w(w) = 0 \end{cases}$$

with $f_w(w) = \int f_{z,w}(z, w)dz$ (marginal pdf of W).

If $\mathbb{E}|Z| < \infty$, then the conditional expectation of Z given $W = w$ is

$$g(w) \coloneqq \mathbb{E}[Z|W = w] = \int z f_{z|w}(z|w)dz.$$

[1] Note that a probability space consists of a triplet $(\mathcal{X}, \mathcal{F}, P_0)$ where $\mathcal{X}$ is a set, $\mathcal{F}$ is a sigma algebra (that is, $\mathcal{F}$ is a set of subsets of $\mathcal{X}$, $\emptyset, \mathcal{X} \in \mathcal{F}$ and $\mathcal{F}$ is closed to complement and countable unions), $P_0 : \mathcal{F} \to [0, 1]$ is a probability measure. We write $\int g(x)dP_0(x)$ to denote the integration of a function $g : \mathcal{X} \to \mathbb{R}$ against the measure P_0. Just think of this as $\mathbb{E}[g(X)]$ for a random variable $X \in \mathcal{X}$. If P_0 has probability density function f_0, then $dP_0(x) = f(x)dx$; while if P_0 is a discrete measure, this is a sum over probabilities. See Chapter 1 of Probability: Theory and Examples by R. Durrett for more details. https://services.math.duke.edu/ rtd/PTE/pte.html

[2] When we write $h_* \coloneqq \operatorname{argmin}_h R(h)$, we mean the function h such that $R(h_*) = \inf_h R(h)$ (assuming this exists)

We also write $\mathbb{E}[Z|W]$ for the random variable $g(W)$. Some useful properties of conditional expectation are as follows:

- If Z, W are independent, then $\mathbb{E}[Z|W] = \mathbb{E}[Z]$. If U is a random variable such that W is independent of (Z, U), then $\mathbb{E}[Z|U, W] = \mathbb{E}[Z|U]$.
- Tower property: Let $f : \mathbb{R}^d \to \mathbb{R}^m$ be a (measurable) function. Then,

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

  Taking f to be the constant function, we obtain $\mathbb{E}[\mathbb{E}[Z|W]] = \mathbb{E}[Z]$.
- Taking out what is known: If $\mathbb{E}[Z^2] < \infty$ and $g : \mathbb{R}^d \to \mathbb{R}$ satisfies $\mathbb{E}[g(W)^2] < \infty$, then

$$\mathbb{E}[g(W)Z|W] = g(W)\mathbb{E}[Z|W]$$

- Best least squares estimator $\mathbb{E}[(Z-g(W))^2] = \mathbb{E}[(Z-\mathbb{E}[Z|W])^2]+\mathbb{E}[(\mathbb{E}[Z|W]-g(W))^2]$
- Conditional Jensen [3]: Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function and $\mathbb{E}[|f(Z)|] < \infty$. Then, $\mathbb{E}[f(Z)|W] \geqslant f(\mathbb{E}[Z|W])$.

1.2.2 Characterization of the Bayes classifier

Define $\eta(x) \coloneqq \mathbb{P}(Y = 1|X = x)$. This is called the *regression* function.

Proposition 1. *The Bayes classifier h_* is given by*

$$h_*(x) = \begin{cases} 1 & \eta(x) > \frac{1}{2} \\ 0 & \eta(x) \leqslant \frac{1}{2} \end{cases}.$$

Remark 3.
- Note that $\mathbb{E}[Y|X] = 1 \cdot \mathbb{P}(Y = 1|X) + 0 \cdot \mathbb{P}(Y = 0|X) = \mathbb{P}(Y = 1|X)$. So, $\eta(x) = \mathbb{E}[Y|X = x]$.
- We can also write $h_*(x) = \operatorname{argmax}_y \mathbb{P}(Y = y|X = x)$, so h_* is the 'maximum a-posteriori estimator'.
- If $\eta(x) = \frac{1}{2}$, then we can equally take $h_*(x) = \pm 1$ and achieve the same misclassification error.

Proof. First note that

$$R(h) = \mathbb{E}[\mathbf{1}\{h(X) \neq Y\}] = \mathbb{E}[\mathbb{E}[\mathbf{1}\{h(X) \neq Y\}|X]].$$

[3] The 'standard' Jensen inequality is: if $f : \mathbb{R} \to \mathbb{R}$ is a convex function and $\mathbb{E}[|f(Z)|] < \infty$, then, $\mathbb{E}[f(Z)] \geqslant f(\mathbb{E}[Z])$.

Let us consider the term inside the expectation:

$$\begin{aligned}\mathbb{E}[\mathbf{1}\{h(X) \neq Y\}|X] &= \mathbb{E}[\mathbf{1}\{h(X)=1, Y=0\} + \mathbf{1}\{h(X)=0, Y=1\}|X] \\ &\overset{\text{take out the known}}{=} \mathbf{1}\{h(X)=1\}\mathbb{E}[\mathbf{1}\{Y=0\}|X] + \mathbf{1}\{h(X)=0\}\mathbb{E}[\mathbf{1}\{Y=1\}|X] \\ &= \mathbf{1}\{h(X)=1\}\mathbb{P}[Y=0|X] + \mathbf{1}\{h(X)=0\}\mathbb{P}[Y=1|X] \\ &= \mathbf{1}\{h(X)=1\}(1-\eta(X)) + \mathbf{1}\{h(X)=0\}\eta(X).\end{aligned}$$

How can we choose h to minimize this? Suppose $1-\eta(x) > \eta(x)$, that is, $\eta(x) < \frac{1}{2}$, then this is minimized with $h(x) = 0$. If $1-\eta(x) < \eta(x)$, then this is minimized with $h(x) = 1$. If $\eta(x) = 1-\eta(x)$, then the above is constant and we can take $h(x) = 0$ or $h(x) = 1$.

□

Remark 4. From the proof, we can observe that

$$\begin{aligned}R(h_*) &= \mathbb{E}[\mathbf{1}\{h_*(X)=1\}(1-\eta(X)) + \mathbf{1}\{h_*(X)=0\}\eta(X)] \\ &= \mathbb{E}[\min\{\eta(X), 1-\eta(X)\}] \leqslant \frac{1}{2}\end{aligned}$$

If X, Y are independent, then $\eta(X) = \mathbb{P}(Y=1|X) = \mathbb{P}(Y=1)$. So,

$$R(h_*) = \min\{\mathbb{P}(Y=1), \mathbb{P}(Y=0)\}.$$

Example 1 (Computing Bayes classifier for a probability mixture model). Consider the following problem: We are given a list of house prices from 2 different neighbourhoods, call them 0 and 1. The goal is to predict the neighbourhood from the price. We will assume that the prices in each neighbourhood follow some Gaussian distribution, say ρ_0 for neighbourhood 0 and ρ_1 for neighbourhood 1. We also assume that there are more houses in one neighbourhood than the other.

To model this, let $\mathcal{X} \times \mathcal{Y} = \mathbb{R} \times \{0,1\}$. Assume that $Y \sim \text{Ber}(q)$ so that $\mathbb{P}(Y=1) = q$ and $\mathbb{P}(Y=0) = 1-q$. We assume that the conditional density of X given Y is

$$f_{X|Y=1} = \rho_1 \quad \text{and} \quad f_{X|Y=0} = \rho_0.$$

The pdf of X is

$$f_X = f_{X|Y=1}\mathbb{P}(Y=1) + f_{X|Y=0}\mathbb{P}(Y=0) == q\rho_1(x) + (1-q)\rho_0(x).$$

In this case,

$$\eta(x) = \mathbb{P}(Y=1|X=x) = \frac{\mathbb{P}(X=x|Y=1)\mathbb{P}(Y=1)}{\mathbb{P}(X=x)} = \frac{q\rho_1(x)}{(1-q)\rho_0(x) + q\rho_1(x)}.$$

So, $\eta(x) > \frac{1}{2}$ iff $\rho_1(x)/\rho_0(x) > (1-q)/q$. The Bayes classifier is

$$h_*(x) = \begin{cases} 1 & \rho_1(x)/\rho_0(x) > (1-q)/q \\ 0 & \text{otherwise} \end{cases}.$$

Important practical consideration In practice, we observe only the data points (X_i, Y_i) for $i = 1, \dots, n$ to learn from and the probability distribution P_0 is **unknown**. The classical approach is to either estimate P_0 by constructing some histogram on the data; or assume some probability model (e.g. mixture of Gaussians) and estimate these parameters (means and standard deviations) from the data, before computing the Bayes classifier $\hat{h}_*$.

In the following, we take a different approach: we will define a class of functions $\mathcal{H}$ from which we will pick the best classifier based on the given data.

Some possible choices of $\mathcal{H}$ in the context of regression are

- $\mathcal{H} = \left\{h \,;\, h(x) = b + w^\top x, \quad b \in \mathbb{R}, w \in \mathbb{R}^p\right\}$
- $\mathcal{H} = \left\{h \,;\, h(x) = b + \sum_{j=1}^d w_j \varphi_j(x),\ w \in \mathbb{R}^d, b \in \mathbb{R}\right\}$ for a given set of basis functions $\varphi_j : \mathcal{X} \to \mathbb{R}$.

For classification, we can compose with the sign function. e.g.

$$\mathcal{H} = \left\{h \,;\, h(x) = \operatorname{sign}(b + w^\top x),\ b \in \mathbb{R}, w \in \mathbb{R}^p\right\}.$$

1.3 Empirical risk minimization and hypothesis classes

Empirical risk Recall that $R(h) = \mathbb{E}[\ell(h(X), Y)]$ where the expectations is over (X, Y) from some probability distribution P_0. Suppose we observe n samples $(X_i, Y_i) \overset{iid}{\sim} P_0$, $i = 1, \dots, n$. The ***empirical risk*** is

$$\hat{R}(h) = \frac{1}{n} \sum_{i=1}^n \ell(h(X_i), Y_i) \tag{1.3}$$

and the ***empirical risk minimizer*** (ERM) is

$$\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} \hat{R}(h). \tag{1.4}$$

Clearly, $\mathbb{E}[\hat{R}(h)] = R(h)$.

Example 2 (Linear regression). Consider the regression setting

$$\mathcal{X} = \mathbb{R}^p, \quad \mathcal{Y} = \mathbb{R}, \qquad \ell(z, y) = (z - y)^2$$

with $\mathcal{H}$ being linear functions [4]:

$$\mathcal{H} = \{x \mapsto b + x^\top w,\ w \in \mathbb{R}^p, b \in \mathbb{R}\}.$$

[4] We will often use the notation $w^\top x \coloneqq \langle w, x\rangle \coloneqq \sum_{i=1}^n w_i x_i$ for vectors $w, x \in \mathbb{R}^n$. Note that since we work with real vectors, this is symmetric: $w^\top x = x^\top w = \langle x, w\rangle = \langle w, x\rangle$.

The ERM is $\hat{h}(x) = x^\top \hat{w} + \hat{b}$ where $(\hat{b}, \hat{w}) = \text{argmin}_{w \in \mathbb{R}^p, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n (Y_i - X_i^\top w - b)^2$. In this case, this is a least squares problem

$$(\hat{w}, \hat{b}) = \text{argmin}_{w \in \mathbb{R}^p, b \in \mathbb{R}} \left\| M \begin{pmatrix} w \\ b \end{pmatrix} - Y_{1:n} \right\|^2$$

where $Y_{1:n} = (Y_j)_{j=1}^n \in \mathbb{R}^n$ and $M \in \mathbb{R}^{n \times (p+1)}$ is the matrix with rows $(X_i^\top, 1)^\top$. One can show that

$$\begin{pmatrix} \hat{w} \\ \hat{b} \end{pmatrix} = (M^\top M)^{-1} M^\top Y_{1:n}.$$

Example 3 (K-nearest neighbours). This is a method for classification and suppose $Y \in \{-1, 1\}$. The idea is that input points that are close to each other should have similar output.

1-NN Given training data (X_i, Y_i) for $i = 1, \ldots, n$ and a new input $\bar{X}$, let

$$i' = \text{argmin}_{i=1,\ldots,n} \left\| \bar{X} - X_i \right\|.$$

The nearest neighbour estimator is $h(\bar{X}) = Y_{i'}$. This is however very sensitive to noise (e.g. if some of the examples have been mislabelled).

K-NN The idea of K-nearest neighbours is to consider the K training points nearest to $\bar{X}$ and assign it the most common label. Define the vectors of distances from $\bar{X}$ to the data by $d_{\bar{X}} = (\left\| \bar{X} - X_i \right\|)_{i=1}^n$. Let $I_{\bar{X}}$ index k smallest entries of $d_{\bar{X}}$.

The K-NN estimator is

$$h(\bar{X}) = \text{sign} \left(\frac{1}{K} \sum_{i' \in I_{\bar{X}}} Y_{i'} \right).$$

This is a 'vote' on the class by the K nearest neighbours and we let K be odd to break ties.

1.4 Bias-variance trade-off

In principle, the larger the hypothesis class, the better we can estimate the Bayes estimator; on the other hand, performing empirical risk minimization over a larger hypothesis class can result in high sensitivity to noise (recall the previous example where the 1-NN neighbour is sensitive to noise in the training data).

To explain this trade-off, suppose we are given

- a dataset $D = \{(X_i, Y_i)\}_{i=1,\dots,n}$ of points drawn iid from P_0.
- Denote the expected output by $\bar{y}(x) = \mathbb{E}[Y|X = x]$.
- Denote the hypothesis learned from the dataset by $h_D(x)$.
- Denote the expected hypothesis is $\bar{h}(x) \coloneqq \mathbb{E}_{D\sim P_0^n}[h_D(x)]$.

Note that when considering $\mathbb{E}[(h_D(X) - Y)^2]$, the randomness is over X, Y, D. To clarify which random variable the expectation is over, we will sometimes indicate this with a subscript on the expectation $\mathbb{E}$ below: We can decompose the expected test error as follows

$$\begin{aligned}\mathbb{E}(h_D(X) - Y)^2 &= \mathbb{E}(h_D(X) - \bar{y}(X) + \bar{y}(X) - Y)^2 \\ &= \mathbb{E}(h_D(X) - \bar{y}(X))^2 + \mathbb{E}(\bar{y}(X) - Y)^2\end{aligned}$$

by the least squares property of conditional expectation, recalling that $\bar{y}(x) = \mathbb{E}[Y|X = x]$. We call $\mathbb{E}(\bar{y}(X) - Y)^2$ the *noise term*. The term $\mathbb{E}(h_D(X) - \bar{y}(X))^2$ can be decomposed as follows:

$$\begin{aligned}\mathbb{E}[(h_D(X) - \bar{y}(X))^2] &= \mathbb{E}[(h_D(X) - \bar{h}(X) + \bar{h}(X) - \bar{y}(X))^2] \\ &= \mathbb{E}[(h_D(X) - \bar{h}(X))^2] + \mathbb{E}[(\bar{h}(X) - \bar{y}(X))^2] \\ &\quad + 2\underbrace{\mathbb{E}[(h_D(X) - \bar{h}(X))(\bar{h}(X) - \bar{y}(X))]}_{=0 \text{ since } \mathbb{E}_D[h_D(X)]=\bar{h}(X)}.\end{aligned}$$

We can therefore write

$$\mathbb{E}(h_D(X) - Y)^2 = \underbrace{\mathbb{E}(h_D(X) - \bar{h}(X))^2}_{\text{variance}} + \underbrace{\mathbb{E}(\bar{h}(X) - \bar{y}(X))^2}_{\text{bias}} + \underbrace{\mathbb{E}(\bar{y}(X) - Y)^2}_{\text{noise}}.$$

For example, if $Y = f(X) + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, then $\bar{y}(X) = f(X)$ and the noise term is $\mathbb{E}[(\bar{y}(X) - Y)^2] = \sigma^2$. If $\mathcal{H}$ is more 'complex', then we expect the variance to be larger (more data is needed to approximate the expected classifier); on the other hand, less 'complexity' in $\mathcal{H}$ results in a larger bias. Hence, there is a *trade-off* between bias and variance.

1.4.1 K nearest neighbours

Suppose the random variables X, Y are related by $Y = h_*(X) + \varepsilon$ where $\mathbb{E}\varepsilon = 0$ and $\mathbb{E}\varepsilon^2 = \sigma^2$. Note that $\mathbb{E}[Y|X = x] = h_*(x)$.

Given data $D \coloneqq \{(X_i, Y_i)\}_{i=1}^n$, consider the estimator

$$h_{D,K}(x) \coloneqq \frac{1}{K} \sum_{j \in I_x} Y_j$$

where I_x is the index of the K points in $\{X_i\}_{i=1}^n$ that are closest to x.

Then,

$$\begin{aligned}
&\mathbb{E}[(Y - h_{D,K}(X))^2] = \mathbb{E}[(Y - h_*(X) + h_*(X) - h_{D,K}(X))^2] \\
&= \underbrace{\mathbb{E}[(Y - h_*(X))^2]}_{\mathbb{E}\varepsilon^2 = \sigma^2} + \mathbb{E}[(h_*(X) - h_{D,K}(X))^2]] \\
&= \sigma^2 + \mathbb{E}[(h_*(X) - h_{D,K}(X))^2]
\end{aligned}$$

Observe that

$$\mathbb{E}[h_{D,K}(x)|X_{1:n}] = \mathbb{E}[\frac{1}{K} \sum_{j \in I_x} h_*(X_j) + \varepsilon_j | X_{1:n}] = \frac{1}{K} \sum_{j \in I_x} h_*(X_j).$$

Taking expectation of $(h_*(x) - h_{D,K}(x))^2$, *conditional on* X_i, we obtain

$$\begin{aligned}
&\mathbb{E}_{Y_1,\ldots,Y_n|X_1,\ldots,X_n} \left[(h_*(x) - h_{D,K}(x))^2\right] \\
&= \mathbb{E}_{Y_1,\ldots,Y_n|X_1,\ldots,X_n} \left[\left(h_*(x) - \frac{1}{K} \sum_{j \in I_x} h_*(X_j) + \frac{1}{K} \sum_{j \in I_x} h_*(X_j) - h_{D,K}(x)\right)^2\right] \\
&= \mathbb{E}_{Y_1,\ldots,Y_n|X_1,\ldots,X_n} \left[(h_*(x) - \frac{1}{K} \sum_{j \in I_x} h_*(X_j))^2 + (\frac{1}{K} \sum_{j \in I_x} h_*(X_j) - Y_j)^2\right] \\
&= \underbrace{(h_*(x) - \frac{1}{K} \sum_{j \in I_x} h_*(X_j))^2}_{\text{bias}} + \underbrace{\frac{\sigma^2}{K}}_{\text{variance}}
\end{aligned}$$

To summarise,

$$\mathbb{E}[(Y - h_{D,K}(X))^2] = \underbrace{\sigma^2}_{\text{noise}} + \underbrace{\mathbb{E}[(h_*(X) - \frac{1}{K} \sum_{j \in I_x} h_*(X_j))^2]}_{\text{bias}} + \underbrace{\frac{\sigma^2}{K}}_{\text{variance}}.$$

Notice that for small K, $\frac{1}{K} \sum_{j \in I_x} h_*(X_j)$ can potentially provide a better fit to h_* and for large K, $\frac{1}{K} \sum_{j \in I_x} h_*(X_j)$ resembles the population mean. As we increase K, the bias will typically increase (with a bias towards the population mean), but on the other hand, the variance decreases as K increases.

1.4.2 Cross-validation

As we saw in the above example, there is a 'sweet spot' for the choice of K in the nearest neighbours estimator. Given a dataset D, suppose we have different estimators trained on the dataset D, $h_{k,D} : \mathcal{X} \to \mathcal{Y}$, for $k = 1, \ldots, K$. For example,

- each $h_{k,D}$ is a nearest neighbours estimators and k is the number of neighbours.
- We have some basis functions $\varphi_i : \mathcal{X} \to \mathbb{R}^d$ and each h_k is constructed from the first k basis functions. i.e. $h_{k,D}(x) = \sum_{j \leqslant k} w_j \varphi_j(x)$ where w_j are some coefficients learned from data.

Which estimator should we choose? We want to pick the estimator h_{k_*} such that

$$R(h_{k,D}) = \mathbb{E}[\ell(h_k(X), Y)|D].$$

is minimized. Or find the parameter k_* that minimizes

$$\mathbb{E}[R(h_{k_*,D})] = \min_k \mathbb{E}[\mathbb{E}[\ell(h_k(X), Y)|D]].$$

However, in practice, the data distribution and hence the expected loss is not know and this risk cannot be computed.

Idea Given a dataset D, hold out part of this data $\tilde{D} \subset D$ to use as *validation data* and use $D \setminus \tilde{D}$ as training data: that is, consider the estimators $\hat{h}(x) = h_{D \setminus \tilde{D}}(x)$ constructed from data $D \setminus \tilde{D}$, and take the estimator that minimizes the empirical error on $\tilde{D}$:

$$\sum_{(x,y) \in \tilde{D}} (\hat{h}(x) - y)^2.$$

V-fold One can also partition the data into V (often equally sized) sets. Each set V_i can be treated as the hold-out dataset to be used for validation and we can compute the hold-out error E_i. We can then choose the best parameter based on the average hold out errors: $\frac{1}{V} \sum_{i=1}^{V} E_i$. Typical choices of V are 5 or 10.

1.5 Excess risk

In the previous section, we demonstrated the trade-off in the mean squared error of an estimator by looking at $\mathbb{E}[(Y - h(X))^2]$ and made comparisons against the expected estimator $\mathbb{E}_D h_D(x)$ or $\mathbb{E}_{Y_{1:N}|X_{1:N}} h_D(x)$. For classification, these expected estimators are not classifiers in general. In this section, we will look at a related notion by comparing $R(\hat{h})$ the risk of the empirical estimator $\hat{h} = \operatorname{argmin}_{h \in \mathcal{H}} \hat{R}(h)$ against $R(h_*)$ where $h^* = \operatorname{argmin}_h R(h)$ is the Bayes classifier.

Definition 1 (Excess risk). *Given an estimator $\hat{h}$, the* ***excess risk*** *is given by*

$$\mathcal{E} := R(\hat{h}) - R(h^*)$$

where h^* is the Bayes classifier.

No free lunch Is it possible to construct an algorithm that learns optimally for all probability distributions? The answer is *no*. There are several "No free lunch" theorems in theoretical machine learning, telling us that learning without assumptions is impossible. Below is one such theorem.

Theorem 2 (No free lunch theorem, non-examinable). *Assume $\mathcal{X}$ is infinite. Fix $n \in \mathbb{N}$ and let $\varepsilon > 0$. Consider a binary classification problem with unit loss and $\mathcal{Y} = \{0, 1\}$. Let $\hat{h}_n$ constructed from random data $D^n = \{(X_i, Y_i) \overset{iid}{\sim} P_0\}_{i=1}^n$. Then, there exists a probability distribution P_0 on $\mathcal{X} \times \{0, 1\}$ such that*

$$\mathbb{E}_{D^n \overset{iid}{\sim} P_0^n} \mathcal{E}(\hat{h}_n) \geqslant \frac{1}{2} - \varepsilon.$$

This essentially says that is is impossible to construct a sequence of classifiers $\hat{h}_n$ such that the excess risk $\mathcal{E}(\hat{h}_n)$ converges to 0 uniformly for all distributions P_0.

Key questions to be addressed in this section

- How does the complexity of $\mathcal{H}$ influence $\mathcal{E}$?
- How does a change in the number of data points n influence $\mathcal{E}$?

Example 4 (Learning rectangles). Suppose the input space is $\mathcal{X} \subset \mathbb{R}^2$ and $\mathcal{Y} = \{0, 1\}$. We want to learn the indicator function $\mathbf{1}_B$ where B is a closed rectangle. We have access to the iid random data $\{(X_i, Y_i)\ ;\ i = 1, \ldots, n\}$, where $X_i \in \mathcal{X}$ and Y_i indicates whether $X_i \in B$ or not.

Given these n data points, suppose we compute a candidate classifier $\hat{h} = \mathbf{1}_{\hat{B}} : \mathbb{R}^2 \to \mathcal{Y}$ where $\hat{B}$ is the smallest rectangle containing the points that have been labelled as being inside B, i.e. $\{X_i\ ;\ Y_i = 1\}$. We would like to quantify the risk:

$$R(\hat{h}) = \mathbb{P}(\hat{h}(X) \neq Y) = \mathbb{P}(\hat{h}(X) \neq \mathbf{1}_B(X)).$$

In particular, how many samples do we need to ensure that

$$\mathbb{P}(R(\hat{h}) \leqslant \varepsilon) \geqslant 1 - \delta? \tag{1.5}$$

Note that $\hat{h}$ depends on the dataset and hence, $R(\hat{h})$ is a random variable of the dataset.

To understand (1.5), we first observe that

$$R(\hat{h}) = \mathbb{P}(\hat{h}(X) \neq \mathbf{1}_B(X)) = \mathbb{P}((X \in \hat{B} \text{ and } X \notin B) \text{ or } (X \notin \hat{B} \text{ and } X \in B))$$
$$= \mathbb{P}(X \notin \hat{B} \text{ and } X \in B) = \mathbb{P}(X \in B \setminus \hat{B}),$$

where the last line follows because $\hat{B} \subseteq B$. Assume that $\mathbb{P}(X \in B) > \varepsilon$ (otherwise, we trivially have $R(\hat{h}) = \mathbb{P}(X \in B) \leqslant \varepsilon$).

Suppose $B = [a, b] \times [c, d]$. Let

- $a < a'$ be the smallest number such that $\mathbb{P}(X \in R_1) \geqslant \varepsilon/4$ where $R_1 \coloneqq [a, a'] \times [c, d]$;
- $b' < b$ be the largest number such that $\mathbb{P}(X \in R_2) \geqslant \varepsilon/4$ with $R_2 \coloneqq [b', b] \times [c, d]$;
- $c < c'$ be the smallest number such that $\mathbb{P}(X \in R_3) \geqslant \varepsilon/4$ with $R_3 \coloneqq [a, b] \times [c, c']$;
- $d' < d$ be the largest number such that $\mathbb{P}(X \in R_4) \geqslant \varepsilon/4$ with $R_4 \coloneqq [a, b] \times [d', d]$.

If $\hat{B} \cap R_i \neq \emptyset$ for all i, then $\mathbb{P}(X \in B \setminus \hat{B}) \leqslant \varepsilon$: To see this, if $B = [\hat{a}, \hat{b}] \times [\hat{c}, \hat{d}]$, then we must have $\hat{a} \leqslant a'$, $\hat{b} \geqslant b'$, $\hat{c} \leqslant c'$, $\hat{d} \geqslant d'$. So, $\mathbb{P}(X \in B \setminus \hat{B}) \leqslant \mathbb{P}(X \in \cup_{i=1}^4 R_i^o) \leqslant \varepsilon$, where R_i^o is R_i with its innermost edge removed (indeed, $R_1^o = [a, a') \times [c, d]$ and if $\mathbb{P}(R_1^o) > \varepsilon/4$, then there exists $a'' < a'$ such that $\mathbb{P}([a, a''] \times [c, d]) \geqslant \varepsilon/4$ which is a contradiction, a similar argument can be applied for the other rectangles). [5]

So, since the event that "$\hat{B} \cap R_i \neq \emptyset$ for all i" is contained in the event that "$R(\hat{h}) \leqslant \varepsilon$", we have

$$\mathbb{P}(R(\hat{h}) \leqslant \varepsilon) \geqslant \mathbb{P}(\hat{B} \cap R_i \neq \emptyset, \quad i = 1 \ldots, 4) = 1 - \mathbb{P}(\exists i \in \{1, 2, 3, 4\}, \quad \hat{B} \cap R_i = \emptyset).$$

By the union bound,

$$\mathbb{P}(\exists i \in \{1, 2, 3, 4\}, \quad \hat{B} \cap R_i = \emptyset) \leqslant \sum_{i=1}^{4} \mathbb{P}(\hat{B} \cap R_i = \emptyset)$$

Now for each i, $\hat{B} \cap R_i = \emptyset$ if and only if $X_k \notin R_i$ for all $k = 1, \ldots, n$, since $\mathbb{P}(X \notin R_i) \leqslant 1 - \varepsilon/4$, we have by independence $\mathbb{P}(\forall k, \ X_k \notin R_i) = \prod_{k=1}^n \mathbb{P}(X_k \notin R_i) \leqslant (1 - \varepsilon/4)^n$. Therefore,

$$\mathbb{P}(\exists i \in \{1, 2, 3, 4\}, \quad \hat{B} \cap R_i = \emptyset) \leqslant 4(1 - \varepsilon/4)^n \leqslant 4e^{-n\varepsilon/4}$$

where we used the fact that $1 - x \leqslant e^{-x}$. We therefore have

$$\mathbb{P}(R(\hat{h}) \leqslant \varepsilon) \geqslant 1 - 4(1 - \varepsilon/4)^n \geqslant 1 - 4e^{-n\varepsilon/4} \geqslant 1 - \delta,$$

provided that

$$n \geqslant \frac{4}{\varepsilon} \log\left(\frac{4}{\delta}\right).$$

[5] You should note that that this result does not depend on the probability distribution of X, and this distribution can even be supported on a discrete set of points. If we assume that X is absolutely continuous, i.e., it has an integrable probability density function so that $\mathbb{P}(X \in R) = \int_R f(x)dx$ for some function f, then the proof simplifies and we can set $\mathbb{P}(X \in R_i) = \varepsilon/4$ instead of handling the inequalities when choosing a', b', c', d'.

In the previous example, the hypothesis class $\mathcal{H}$ is the set of indicator functions of rectangles, the true classifier is $\mathbf{1}_B$ is contained inside $\mathcal{H}$, so $R(h_*) = 0$. In general, the true classifier is not necessarily contained in $\mathcal{H}$ and $R(h_*) > 0$. In the following section, we will carry out a systematic study of the excess risk.

1.5.1 Decomposition of excess risk

Decomposition of the excess risk

$$R(\hat{h}) - R(h_*) = \underbrace{R(\hat{h}) - \inf_{h\in\mathcal{H}} R(h)}_{\text{Estimation error}} + \underbrace{\inf_{h\in\mathcal{H}} R(h) - R(h^*)}_{\text{approximation error}}$$

This can be viewed as a kind of variance/bias separation. As the size of $\mathcal{H}$ increases, the approximation error will decrease, but the estimation error increases. The size of $\mathcal{H}$ should therefore depend on n to avoid overfitting (choosing too large $\mathcal{H}$ will lead to overfitting).

The approximation error The approximation error fundamentally requires some assumption on the Bayes classifier h^* and in particular, the underlying probability distribution P_0.

To given an example of a bound on the approximation error, assume that the loss $z \mapsto \ell(y, z)$ is L-Lipschitz continuous. Then,

$$R(h) - R(h') = \mathbb{E}[\ell(Y, h(X)) - \ell(Y, h'(X))] \leqslant L\mathbb{E}\,|h(X) - h'(X)|$$

Assume that $h^*(x) = w_*^\top \varphi(x)$ for some 'feature' vector $\varphi(x)$. For example,

$$\varphi(x) = (1, x_1, \cdots, x_p, x_1^2, x_1x_2, \cdots, x_p^2, \cdots),$$

so h^* is a polynomial function. Suppose $\mathcal{H} = \big\{h_w(x) = w^\top\varphi(x)\ ;\ \|w\| \leqslant K\big\}$. Then,

$$\inf_{h\in\mathcal{H}} R(h) - R(h^*) \leqslant L \inf_{\|w\|\leqslant K} \mathbb{E}[|\langle\varphi(x),\, w - w_*\rangle|] \leqslant L\mathbb{E}\,\|\varphi(x)\| \max(\|w_*\| - K, 0)$$

which is zero if $\|w_*\| \leqslant K$. Note that in the above, we use the fact that

$$\inf_{\|w\|\leqslant K} \|w - w^*\| = \max(\|w_*\| - K, 0).$$

1.6 The estimation error

In this section, we consider the estimation error $R(\hat{h}) - \inf_{h\in\mathcal{H}} R(h)$ of the **empirical risk minimizer (ERM)** $\hat{h}$, where we recall the empirical risk $\hat{R}$ and ERM $\hat{h}$ defined in (1.3)

and (1.4). Note that $\hat{h}$ and $\hat{R}$ *depend* on some iid dataset $D = \{(X_i, Y_i)\}_{i=1}^n$. Intuitively, the greater n is, the smaller the estimation error and the number of samples should also depend on the size of $\mathcal{H}$. In this section, we will make this precise using concentration inequalities.

Let $\bar{h} = \operatorname{argmin}_{h\in\mathcal{H}} R(h)$ (assume for simplicity that this exists, otherwise, we can repeat our analysis for $\bar{h}$ such that $R(\bar{h}) \leqslant \inf_{h\in\mathcal{H}} R(h) + \varepsilon$).

We begin with the following observation

$$R(\hat{h}) - R(\bar{h}) = R(\hat{h}) - \hat{R}(\hat{h}) + \underbrace{\hat{R}(\hat{h}) - \hat{R}(\bar{h})}_{\leqslant 0 \text{ since } \hat{h} \text{ is an ERM}} + \hat{R}(\bar{h}) - R(\bar{h}) \tag{1.6}$$

$$\leqslant \sup_{h\in\mathcal{H}} R(h) - \hat{R}(h) + \hat{R}(\bar{h}) - R(\bar{h}) \leqslant 2 \sup_{h\in\mathcal{H}} \left|R(h) - \hat{R}(h)\right| \tag{1.7}$$

Note that $\sup_{h\in\mathcal{H}} \left|R(h) - \hat{R}(h)\right|$ is a random variable on the data set D. Ideally, we would like to upper bound the probability

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} \left|R(h) - \hat{R}(h)\right| > t\right) \leqslant \delta. \tag{1.8}$$

i.e. with probability at least $1-\delta$, $\sup_{h\in\mathcal{H}} \left|R(h) - \hat{R}(h)\right| \leqslant t$. Our goal now is to understand how δ and t depend on the number of samples n.

A naive approach

To motivate the need for the machinery of concentration inequalities, let us first 'guess' of how one might naively estimate (1.8). If $|\mathcal{H}| < \infty$, then

$$\mathbb{P}(\max_{h\in\mathcal{H}} \left|R(h) - \hat{R}(h)\right| > t) \leqslant \mathbb{P}\left(\bigcup_{h\in\mathcal{H}} \{\left|R(h) - \hat{R}(h)\right| > t\}\right) \leqslant \sum_{h\in\mathcal{H}} \mathbb{P}\left(\left|R(h) - \hat{R}(h)\right| > t\right).$$

For each $h \in \mathcal{H}$,

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

Since $\{Z_i\}_i$ are mean-zero iid random variables, by the central limit theorem, $\sqrt{n}(R(h) - \hat{R}(h))$ behaves like a Gaussian random variable (distributed like $\mathcal{N}(0, \operatorname{Var}[\ell(h(X_i), Y_i)])$ [6] The main issue here is that the CLT is an asymptotic result as $n \to \infty$ and does not give quantitative bounds on (1.8); moreover, we want a uniform result over all $h \in \mathcal{H}$.

To bound (1.8), one typically deploys *concentration inequalities* which are a collection of results that give non-asymptotic and quantitative tail bounds. We will look at one such inequality, known as Hoeffding's inequality.

[6] The central limit states that for any sequence of random variables X_i with mean μ and variance σ^2, $\frac{1}{n}\sum_{i=1}^n X_i - \mu \implies \mathcal{N}(0, \sigma^2)$, where $\implies$ denotes weak convergence: $Z_n \implies Z$ means that $\mathbb{P}(Z_n \leqslant t) \to \mathbb{P}(Z \leqslant t)$ for all t.

1.6.1 Tools from probability

Markov's inequality We begin with the simple tail bound, Markov's inequality. Let W be a non-negative random variable. Then, for all $t > 0$,

$$\mathbb{P}(W \geqslant t) \leqslant \frac{\mathbb{E}[W]}{t}. \tag{1.9}$$

This is a simple consequence of: $t\mathbf{1}\{W \geqslant t\} \leqslant W$ for all $t \geqslant 0$. Taking expectation of both sides yields Markov's inequality.

The Chernoff bound states that for any real valued random variable W and $t \in \mathbb{R}$,

$$\mathbb{P}(W \geqslant t) \leqslant \inf_{\alpha > 0} \exp(-\alpha t)\, \mathbb{E}\exp(\alpha W). \tag{1.10}$$

The function $\alpha \mapsto \mathbb{E}\exp(\alpha W)$ is called the *moment generating function* (MGF) of W.

This is a consequence of Markov's inequality: Let $\varphi : \mathbb{R} \to (0, \infty)$ be any strictly increasing function. Then, given any random variable W and $t \in \mathbb{R}$,

$$\mathbb{P}(W \geqslant t) = \mathbb{P}(\varphi(W) \geqslant \varphi(t)) \leqslant \frac{\mathbb{E}[\varphi(W)]}{\varphi(t)}.$$

Taking $\varphi(t) = \exp(\alpha t)$ for $\alpha > 0$ yields the Chernoff bound.

Example 5. Let $W \sim \mathcal{N}(0, \sigma^2)$. The MGF has a closed form expression $\mathbb{E}[e^{\alpha W}] = e^{\alpha^2\sigma^2/2}$. It follows that for all $t \geqslant 0$,

$$\mathbb{P}(W \geqslant t) \leqslant \inf_{\alpha > 0} e^{\alpha^2\sigma^2/2 - \alpha t} = \exp(-t^2/(2\sigma^2)),$$

where the final equality is because the infimum is achieved at $\alpha = t/\sigma^2$.

Sub-Gaussian random variables From the previous example, we see that it is straightforward to apply the Chernoff bound provided we have an upper bound for the MGF of W. A particularly useful class of random variables is the following:

Definition 2. *We say that a random variable W is sub-Gaussian with parameter $\sigma > 0$ if*

$$\mathbb{E}e^{\alpha(W - \mathbb{E}(W))} \leqslant e^{\alpha^2\sigma^2/2}, \qquad \forall \alpha \in \mathbb{R}.$$

An immediate consequence of this definition and the Chernoff bound is the following result which bounds the tail probability of a subGaussian rv [7].

[7] In fact, one can also establish a converse statement to Proposition 3: If

$$\mathbb{P}(|W - \mathbb{E}[W]| \geqslant t) \leqslant \beta \exp(-t^2\kappa)) \tag{1.11}$$

for some $\beta, \kappa > 0$, then there exists $c > 0$ such that $\mathbb{E}[\exp(\alpha(W - \mathbb{E}(W))] \leqslant \exp(c\alpha^2)$. So, (1.11) is sometimes used as the definition of being sub-Gaussian. The proof of this result is slightly more involved and omitted (see Foucart and Rauhut, *A Mathematical Introduction to Compressive Sensing*, Prop 7.24 if you are interested).

Proposition 3. *If W is sub-Gaussian with parameter $\sigma > 0$, then*

$$\mathbb{P}(W - \mathbb{E}(W) \geqslant t) \leqslant e^{-t^2/(2\sigma^2)}, \qquad \forall t \geqslant 0.$$

Remark 5. If W is sub-Gaussian with parameter $\sigma > 0$, then

- W is also sub-Gaussian with parameter σ' for all $\sigma' \geqslant \sigma$.
- $-W$ is also sub-Gaussian with parameter σ. Using the Chernoff bound on $-W$ yields

  $$\mathbb{P}(-W + \mathbb{E}[W] \geqslant t) \leqslant \exp(-t^2/(2\sigma^2)).$$

  This implies that $\mathbb{P}(|W - \mathbb{E}[W]| \geqslant t) \leqslant 2\exp(-t^2/(2\sigma^2))$.
- $W - c$ is also sub-Gaussian for any constant $c \in \mathbb{R}$.

Example 6. A Rademacher random variable ε takes value ± 1 with equal probability. This is sub-Gaussian random with parameter 1:

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

where we use the fact that $(2k)! \geqslant 2^k k!$.

More generally, the following result shows that bounded random variables are sub-Gaussian:

Lemma 4 (Hoeffding's lemma)**.** *If W takes values in $[a, b]$, then W is sub-Gaussian with parameter $\sigma = (b - a)/2$.*

Analogous to the result that a linear combination of jointly Gaussian random variables is Gaussian, a linear combination of independent sub-Gaussian r.v. is sub-Gaussian.

Proposition 5. *Suppose $W_1, \ldots, W_n$ are independent and each W_i is sub-Gaussian with parameter σ_i. Then, for any $\gamma \in \mathbb{R}^n$, $\sum_{i=1}^n \gamma_i W_i$ is sub-Gaussian with parameter $(\sum_i \gamma_i^2\sigma_i^2)^{1/2}$.*

Proof. WLOG, assume $\mathbb{E}[W_i] = 0$. Then

$$\mathbb{E}\exp\left(\alpha\sum_i \gamma_i W_i\right) = \prod_{i=1}^n \mathbb{E}[\exp(\alpha\gamma_i W_i)] \leqslant \exp\left(\alpha^2\sum_i \gamma_i^2\sigma_i^2/2\right).$$

□

This leads to **Hoeffding's inequality**:

Corollary 6 (Hoeffding's inequality)**.** *If $W_1, \dots, W_n$ are independent and satisfy $a_i \leqslant W_i \leqslant b_i$ for all i. Then, for all $t \geqslant 0$, $Z := \frac{1}{n}\sum_{i=1}^n W_i$ satisfies*

$$\mathbb{P}\left(Z - \mathbb{E}Z \geqslant t\right) \leqslant \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{1.12}$$

and

$$\mathbb{P}\left(|Z - \mathbb{E}Z| \geqslant t\right) \leqslant 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{1.13}$$

Proof. From the previous proposition, each W_i is sub-Gaussian with parameter $(b_i - a_i)/2$ and $Z := \frac{1}{n}\sum_{i=1}^n W_i$ is sub-Gaussian with parameter $n^{-1}\sqrt{\sum_{i=1}^n (b_i - a_i)^2/4}$. By Proposition 3, the result follows. □

1.6.2 Finite hypothesis classes

Theorem 7. *Assume that $\ell(h(X), Y)$ is takes values in $[0, L]$ almost surely for all $h \in \mathcal{H}$. Let $\bar{h} = \operatorname{argmin}_{h\in\mathcal{H}} R(h)$. Suppose $\mathcal{H}$ is finite. Then, with probability at least $1-\delta$, the ERM $\hat{h}$ satisfies*

$$R(\hat{h}) - R(\bar{h}) \leqslant L\sqrt{\frac{2\log(|\mathcal{H}|) + 2\log(\delta^{-1})}{n}}.$$

Proof. On the event $\hat{h} = \bar{h}$, we have $R(\hat{h}) - R(\bar{h}) = 0$, so $\mathbb{P}(R(\hat{h}) - R(\bar{h}) > t) = \mathbb{P}(R(\hat{h}) - R(\bar{h}) > t, \hat{h} \neq \bar{h})$. Recall from (1.6) that $R(\hat{h}) - R(\bar{h}) \leqslant R(\hat{h}) - \hat{R}(\hat{h}) + \hat{R}(\bar{h}) - \hat{R}(\bar{h})$, so

$$\mathbb{P}(R(\hat{h}) - R(\bar{h}) > t) \leqslant \mathbb{P}\left(R(\hat{h}) - \hat{R}(\hat{h}) > t/2, \hat{h} \neq \bar{h}\right) + \mathbb{P}\left(\hat{R}(\bar{h}) - R(\bar{h}) > t/2\right).$$

By applying Hoeffding's inequality (1.12) to the second term with r.v. $W_i = \mathbb{E}[\ell(\bar{h}(X_i), Y_i)] - \ell(\bar{h}(X_i), Y_i)$, we obtain

$$\mathbb{P}\left(\hat{R}(\bar{h}) - R(\bar{h}) > t/2\right) \leqslant \exp\left(-nt^2/(2L^2)\right).$$

When $\hat{h} \neq \bar{h}$,

$$R(\hat{h}) - \hat{R}(\hat{h}) \leqslant \max_{h\in\mathcal{H}\setminus\{\bar{h}\}} R(h) - \hat{R}(h).$$

So, by the union bound and Hoeffding's inequality,

$$\begin{aligned}\mathbb{P}\left(R(\hat{h}) - \hat{R}(\hat{h}) > t/2, \hat{h} \neq \bar{h}\right) &\leqslant \sum_{h\in\mathcal{H}\setminus\{\bar{h}\}} \mathbb{P}(R(h) - \hat{R}(h) > t/2) \\ &\leqslant (|\mathcal{H}| - 1)\exp(-nt^2/(2L^2)).\end{aligned}$$

Therefore,

$$\mathbb{P}(R(\hat{h}) - R(\bar{h}) > t) \leqslant |\mathcal{H}|\exp(-nt^2/(2L^2)) \leqslant \delta$$

provided that $\frac{2L^2\left(\log(|\mathcal{H}|) + \log(\delta^{-1})\right)}{n} \leqslant t^2$.

□

Remark 6. Following the above argument, one can also show that with probability at least $1-\delta$, **for all** $h \in \mathcal{H}$ (see problem sheet 3),

$$R(h) \leqslant \hat{R}(h) + \sqrt{\frac{L^2 \log\left(|\mathcal{H}|/\delta\right)}{2n}}$$

In particular, this also holds for $\hat{h}$. From a practical point of view, the RHS evaluated at the ERM $\hat{h}$ can be easily evaluated given the data $\{(X_i, Y_i)\}_{i=1}^n$ and gives an upper bound for $R(\hat{h}) = \mathbb{E}[\ell(h(X), Y)]$ which cannot be computed unless the underlying probability distribution is known. This idea is exploited for model selection (i.e. how too choose $\mathcal{H}$) in techniques called **structural risk minimization**.

Example 7. Let $\mathcal{X} = [0,1)^2$ and $\mathcal{Y} = \{-1, 1\}$. Suppose we want to learn a classifier given data $(X_i, Y_i) \overset{iid}{\sim} P_0$. Partition $\mathcal{X}$ into m^2 disjoint squares of length $1/m$: $R_1, \ldots, R_m$ of the form $[r/m, (r+1)/m) \times [s/m, (s+1)/m)$ for $r, s = 0, \ldots, m-1$. Let

$$\mathcal{H} = \left\{ h(x) = \sum_{j=1}^{m^2} s_j \mathbf{1}_{R_j}(x) \; ; \; s_j \in \{-1, 1\} \right\}.$$

Note that $\mathcal{H}$ consists of 2^{m^2} classifiers. The ERM over $\mathcal{H}$ is $\hat{h}(x) = \sum_{j=1}^{m^2} \bar{Y}_j \mathbf{1}_{R_j}(x)$ where $\bar{Y}_j = \text{sign}\left(\sum_{X_i \in R_j} Y_i\right)$ (the most common sign of the X_i's falling in R_j). The above theorem tells us that

$$R(\hat{h}) - R(\bar{h}) \leqslant \sqrt{\frac{2m^2 \log(2) + 2\log(\delta^{-1})}{n}}$$

1.6.3 Rademacher complexity

In this section, we tackle the case of **infinite hypothesis classes** $|\mathcal{H}| = \infty$. To bound the estimation error, we will make use of the notion of Rademacher complexity, a way of quantifying how large a hypothesis class is.

Recall that for $\bar{h} = \operatorname{argmin}_{h\in\mathcal{H}} R(h)$, $R(\hat{h}) - R(\bar{h}) \leqslant R(\hat{h}) - \hat{R}(\hat{h}) + \hat{R}(\bar{h}) - R(\bar{h})$. Taking expectation on both sides, we obtain

$$\mathbb{E}[R(\hat{h}) - R(\bar{h})] \leqslant \mathbb{E}G \quad \text{where} \quad G \coloneqq \left(\sup_{h\in\mathcal{H}} R(h) - \hat{R}(h)\right).$$

In this section, we will simply focus on obtaining a bound $\mathbb{E}G$ and hence bound the expected risk $\mathbb{E}[R(\hat{h}) - R(\bar{h})]$ [8].

First note that writing $Z_i = (X_i, Y_i)$ for $i \in [n]$, and consider the set of functions

$$\mathcal{F} = \{(x, y) \mapsto \ell(h(x), y) \; ; \; h \in \mathcal{H}\},$$

we can write $G = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[f(Z_i)] - f(Z_i)$. We want to somehow characterize how complex the behviours of $\mathcal{F}$ can be on the dataset Z_i.

In general, we have the following definition:

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

- *$\mathcal{F}(z_{1:n}) \coloneqq \{(f(z_1), \ldots, f(z_n)) \; ; \; f \in \mathcal{F}\} \subset \mathbb{R}^n$ is a collection of vectors giving the 'behaviours' of $\mathcal{F}$ on $z_{1:n}$.*
- *the empirical Rademacher complexity is*

$$\hat{\mathfrak{R}}(\mathcal{F}(z_{1:n})) \coloneqq \frac{1}{n}\mathbb{E}\left(\sup_{w\in\mathcal{F}(z_{1:n})} \varepsilon^\top w\right) = \frac{1}{n}\mathbb{E}\left(\sup_{f\in\mathcal{F}} \sum_i \varepsilon_i f(z_i)\right)$$

  where $\varepsilon = (\varepsilon_i)_{i=1}^n$ is a vector consisting of iid Rademacher random variables. Intuitively, this is quantifying how closely aligned $f(z_{1:n})$ is to random labels.

- *Given random variables $Z_1, \ldots, Z_n$, we can view $\hat{\mathfrak{R}}(\mathcal{F}(Z_{1:n})) = \frac{1}{n}\mathbb{E}\left(\sup_{f\in\mathcal{F}} \sum_i \varepsilon_i f(z_i) | Z_{1:n}\right)$ as a random variable. The Rademacher complexity is*

$$\mathfrak{R}_n(\mathcal{F}) \coloneqq \mathbb{E}\left(\hat{\mathfrak{R}}(\mathcal{F}(Z_{1:n}))\right).$$

Example 8. Let's look at two extreme cases

[8] Using further concentration inequality tools (see Remark 8), one can convert this to a high probability bound, i.e. $\mathbb{P}(R(\hat{h}) - R(\bar{h}) > \varepsilon) \leqslant \delta$ as in the previous section, but this is beyond the scope of this module.

- Suppose $\mathcal{F} = \{f\}$ is a singleton, then the set $\mathcal{F}(z_{1:n}) = \{(f(z_i)_{i=1}^n\}$ is also a singleton vector and

$$\hat{\mathfrak{R}}(\mathcal{F}(z_{1:n})) = \frac{1}{n}\mathbb{E}_\varepsilon\left(\sum_i \varepsilon_i f(z_i)\right) = \sum_i \frac{1}{2}f(z_i) - \frac{1}{2}f(z_i) = 0.$$

  So, $\mathfrak{R}_n(\mathcal{F}) = 0$.

- Suppose $\mathcal{F}$ consists of all possible classifiers taking values in $\mathcal{Y} = \{-1, 1\}$. Then, $\mathcal{F}(z_{1:n})$ is the set of all possible vectors of length n whose entries take values ± 1 (this is a set of 2^n vectors). So,

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

  since inside the supremum, we can choose f such that $f(z_i) = \varepsilon_i$.

Theorem 8. *Let $\mathcal{F}$ be a class of real-valued functions $f : \mathcal{Z} \to \mathbb{R}$ and let $Z_1, \ldots, Z_n$ be iid random variables taking values in $\mathcal{Z}$. Then,*

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

Remark 7. The LHS depends in a complicated way on the distribution of Z_i. However, $\hat{\mathfrak{R}}(\mathcal{F}(z_{1:n}))$ is independent of P_0 and if we obtain a uniform bound in $z_{1:n} \in \mathcal{Z}^n$ then we have a bound on $\mathfrak{R}_n(\mathcal{F})$.

An immediate consequence of Theorem 8 is the following result, which is the main result of this section, showing that the **estimation error is controlled by the Rademacher complexity.**

Theorem 9. *Let $\mathcal{F} \coloneqq \{(x, y) \mapsto \ell(h(x), y)\ ;\ h \in \mathcal{H}\}$. Then,*

$$\mathbb{E}\left[R(\hat{h}) - R(\bar{h})\right] \leqslant 2\mathfrak{R}_n(\mathcal{F}).$$

Example 9 (Linear models). If $\ell(\cdot, y)$ is G-Lipschitz, then it is possible to show that $\mathcal{F} \coloneqq \{(x, y) \mapsto \ell(h(x), y)\ ;\ h \in \mathcal{H}\}$ satisfies $\mathfrak{R}_n(\mathcal{F}) \leqslant G\mathfrak{R}_n(\mathcal{H})$. [9]

[9] The proof is not entirely trivial, see for example Prop 4.3 of https://www.di.ens.fr/~fbach/ltfp_book.pdf

Consider the linear models $\mathcal{H} = \{h_w(x) = w^\top \varphi(x) \; ; \; \|w\| \leqslant K\}$. Then:

$$\begin{aligned}
\mathfrak{R}_n(\mathcal{H}) = \mathbb{E}_{x,\varepsilon}\left[\sup_{\|w\|\leqslant K} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle \varphi(x_i),\, w\rangle\right] &\leqslant \frac{K}{n}\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i \varphi(x_i)\right\| \\
&\leqslant \frac{K}{n}\sqrt{\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i \varphi(x_i)\right\|^2\right]} \qquad \text{by Jensen's inequality} \\
&= \frac{K}{n}\sqrt{\mathbb{E}\left[\langle \sum_{i=1}^n \varepsilon_i \varphi(x_i),\, \sum_{j=1}^n \varepsilon_j \varphi(x_j)\rangle\right]}.
\end{aligned}$$

Note that the expectation above is over the samples x_i and also an expectation over the Rademacher random variables ε. To clarify, I will write $\mathbb{E}_\varepsilon$ too denote the expectation over ε and $\mathbb{E}_x$ to denote the expectation over $x_{1:n}$:

$$\begin{aligned}
\mathbb{E}\left[\langle \sum_{i=1}^n \varepsilon_i \varphi(x_i),\, \sum_{j=1}^n \varepsilon_j \varphi(x_j)\rangle\right] &= \sum_{i,j=1}^n \mathbb{E}_x[\langle \varphi(x_i),\, \varphi(x_j)\rangle]\mathbb{E}_\varepsilon[\varepsilon_i \varepsilon_j] \\
&= \sum_{i=1}^n \mathbb{E}_x[\langle \varphi(x_i),\, \varphi(x_j)\rangle] \qquad \text{since } \mathbb{E}\,[\varepsilon_i \varepsilon_j] = \delta_{ij} \\
&= n\mathbb{E}[\|\varphi(x)\|^2] \qquad \text{since } x_i \text{ are iid.}
\end{aligned}$$

So, the above result tell us that the expected estimation error satisfies

$$\mathbb{E}\left[R(\hat{h}) - \inf_{h\in\mathcal{H}} R(h)\right] \leqslant \frac{K\sqrt{\mathbb{E}\left[\|\varphi(x)\|^2\right]}}{\sqrt{n}}.$$

This is independent of the dimension of $\varphi(x)$!

Nonexaminable proofs and additional remarks

Proof of Theorem 8 . The proof of this result is via a symmetrization technique. This is based on the idea that a random variable X is called symmetric, if X and $-X$ have the same distribution. The crucial observation for symmetrization is that a symmetric random vector X and the random vector εX, where ε is a Rademacher random variable independent of X, have the same distribution. This is a powerful technique to pass from a sum of arbitrary independent random variables to a Rademacher sum.

Let $(Z'_1, \ldots, Z'_n)$ be an independent copy of $(Z_1, \ldots, Z_n)$. Due to this independence,

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \mathbb{E}f(Z_i) - f(Z_i) &= \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \mathbb{E}[f(Z'_i) - f(Z_i)|Z_{1:n}] \\
&\leqslant \mathbb{E}[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(Z'_i) - f(Z_i)|Z_{1:n}],
\end{aligned}$$

where we used the fact that for any collection of r.v. V_t, $\sup_t \mathbb{E}[V_t] \leqslant \mathbb{E}[\sup_t V_t]$.

Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables and also independent of $(Z_{1:n}, Z'_{1:n})$. Then,

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

where the last equality follows because $\varepsilon_i \overset{d}{=} -\varepsilon_i$. So,

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

□

Remarks on proof of Theorem 9. Recall that

$$\begin{aligned}\mathbb{E}R(\hat{h}) - R(\bar{h}) &\leqslant \mathbb{E}\sup_{h\in\mathcal{H}} R(h) - \hat{R}(h) \\ &= \mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n (\mathbb{E}(f(Z_i)) - f(Z_i)) \leqslant 2\mathfrak{R}_n(\mathcal{F}).\end{aligned} \tag{1.14}$$

Note also that since $\varepsilon_{1:n} \overset{d}{=} -\varepsilon_{1:n}$ for Rademacher random variables, $\mathfrak{R}_n(\mathcal{F}) = \mathfrak{R}_n(-\mathcal{F})$ and we also have

$$\mathbb{E}\sup_{h\in\mathcal{H}}\left(R(h) - \hat{R}(h)\right) \leqslant 2\mathfrak{R}_n(\mathcal{F}) \quad \text{and} \quad \mathbb{E}\sup_{h\in\mathcal{H}}\left(\hat{R}(h) - R(h)\right) \leqslant 2\mathfrak{R}_n(\mathcal{F}). \tag{1.15}$$

□

Remark 8 (Tail bounds, non-examinable). One can convert the expectation bounds from (1.15) into probability tail bounds. If $\ell(h(X), Y)$ takes values in $[0, L]$ almost surely for all $h \in \mathcal{H}$, then one can use another concentration inequality called **McDiarmid's inequality** to bound the deviation of $G := \sup_{h\in\mathcal{H}}\left(R(h) - \hat{R}(h)\right)$ from its mean to show that with probability at least $1-\delta$,

$$\sup_{h\in\mathcal{H}} R(h) - \hat{R}(h) \leqslant 2\mathfrak{R}_n(\mathcal{F}) + L\sqrt{\frac{2\log(\delta^{-1})}{n}}$$

Using (recall the computation in (1.7), $R(\hat{h}) - \inf_{h\in\mathcal{H}} R(h) \leqslant \sup_{h\in\mathcal{H}} R(h) - \hat{R}(h) + \hat{R}(\bar{h}) - R(\bar{h})$, you can check that, with probability $1-\delta$, the estimation error satisfies

$$R(\hat{h}) - \inf_{h\in\mathcal{H}} R(h) \leqslant 2\mathfrak{R}_n(\mathcal{F}) + 2L\sqrt{\frac{2\log(2\delta^{-1})}{n}}.$$

End of: Nonexaminable proofs and additional remarks

1.6.4 Vapnik-Chernovenkis (VC) dimension

We have reduced the problem of bounding the expected risk to bounding the Rademacher complexity of $\mathcal{F}$ (the composition of our loss function with elements in $\mathcal{H}$). We now focus on the setting of **binary classification** where $\mathcal{Y} = \{-1, 1\}$. One way to study Rademacher complexities is via VC dimension.

Idea: Given n data points $z_{1:n}$, count the number of behaviours of the function class $\mathcal{F}$ on $z_{1:n}$, i.e. $|\mathcal{F}(z_{1:n})|$.

We first make the observation that the number of vectors in $\mathcal{F}(z_{1:n})$ is no greater than the number of vectors in $\mathcal{H}(x_{1:n}) = \{(h(x_i))_{i=1}^n \,;\, h \in \mathcal{H}\} \subset \mathbb{R}^n$. Indeed, $|\mathcal{F}(z_{1:n})| \leqslant |\mathcal{H}(x_{1:n})|$ for $z_i = (x_i, y_i)$ since $(\ell(h(x_i), y_i)_{i=1}^n \neq (\ell(h'(x_i), y_i)_{i=1}^n$ implies $(h(x_i))_{i=1}^n \neq (h'(x_i))_{i=1}^n$.

Lemma 10. *[Massart's Lemma] Let $\mathcal{F} := \{(x, y) \mapsto \ell(h(x), y) \,;\, h \in \mathcal{H}\}$ and assume that ℓ takes values in $[0, 1]$. Then,*

$$\hat{\mathfrak{R}}(\mathcal{F}(z_{1:n})) \leqslant \sqrt{2\log(|\mathcal{F}(z_{1:n})|)/n}.$$

We will make use of the following bound (the proof is based on Jensen's inequality, see problem sheet 3):

Proposition 11. *Suppose $W_1, \ldots, W_d$ are mean zero and sub-Gaussian with parameter $\sigma > 0$ (not necessarily independent). Then,*

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

Proof of Lemma 10. Let $d = |\mathcal{F}(z_{1:n})|$ and $\mathcal{F}' = \{f_1, \ldots, f_d\}$ such that $\mathcal{F}(z_{1:n}) = \mathcal{F}'(z_{1:n})$. For $j = 1, \ldots, d$, let $W_j = \frac{1}{n}\sum_{i=1}^n \varepsilon_i f_j(z_i)$. Then,

$$\hat{\mathfrak{R}}(\mathcal{F}(z_{1:n})) = \mathbb{E} \max_{j=1}^n W_j.$$

Since ε_j are sub-Gaussian with parameter 1 and $|f_j(z_i)| \leqslant 1$, W_j is sub-Gaussian with parameter $\sqrt{\sum_{i=1}^n \frac{1}{n^2}} = n^{-1/2}$ (linear combination of subGaussian rv are subGaussian by Prop 5). It follows by Prop 11 that $\mathbb{E}\max_{j=1}^d W_j \leqslant \frac{\sqrt{2\log(d)}}{\sqrt{n}}$. □

This lemma is interesting provided that $|\mathcal{H}(x_{1:n})|$ grows slower than exponential in n (e.g. polynomially n^p), then $\hat{\mathfrak{R}}(\mathcal{F}(z_{1:n}))$ will converge to 0 as n increases. Note that since $h(x_i) \in \{-1, 1\}$, the maximum number of distinct behaviours is $|\mathcal{H}(x_{1:n})| \leqslant 2^n$.

Definition 4. *Let $\mathcal{H}$ be a class of functions $h : \mathcal{X} \to \{a, b\}$ for $a \neq b$ and $|\mathcal{H}| \geqslant 2$.*

- *$\mathcal{H}$ shatters $x_{1:n} \in \mathcal{X}^n$ if $|\mathcal{H}(x_{1:n})| = 2^n$. In words, every possible labelling can be achieved on the data $x_{1:n}$.*

- $\Pi_{\mathcal{H}}(n) = \max_{x_{1:n} \in \mathcal{X}^n} |\mathcal{H}(x_{1:n})|$ *is the shattering coefficient or growth function. This is the maximum number of behaviours on any* n *points using* $\mathcal{H}$*.*

- *The VC dimension* $\mathrm{VC}(\mathcal{H})$ *is the largest integer* n *such that* $x_{1:n}$ *is shattered by* $\mathcal{H}$*, or* ∞ *if none exist.*

$$\mathrm{VC}(\mathcal{H}) = \sup\left\{n \in \mathbb{N} \,;\, \Pi_{\mathcal{H}}(n) = 2^n\right\}.$$

Example 10.
- If $\mathcal{H}$ consists of every possible classifier, then any set of n distinct points can be shattered, $\Pi_{\mathcal{H}}(n) = 2^n$ and $\mathrm{VC}(\mathcal{H}) = \infty$.

- If $\mathcal{H} = \{h_1, h_2\}$ where h_1 takes constant 1 and h_2 takes constant value -1, then $x_{1:n}$ is shattered if and only if $n = 1$, $\Pi_{\mathcal{H}}(n) = 2$ for all n, and $\mathrm{VC}(\mathcal{H}) = 1$. More generally, if $\mathcal{H}$ is finite, then $\Pi_{\mathcal{H}}(n) \leqslant |\mathcal{H}|$ for all n. So, $x_{1:n}$ cannot be shattered if $2^n > |\mathcal{H}|$. We therefore have $\mathrm{VC}(|\mathcal{H}|) \leqslant \log_2(|\mathcal{H}|)$.

Remark 9 (How do we check the VC dimension?). Observer that if $m \leqslant n$ and $|\mathcal{H}(x_{1:n})| = 2^n$, then $|\mathcal{H}(x_{1:m})| = 2^m$. So, to show that $\mathrm{VC}(\mathcal{H}) = n$, we need to

i) find a set of n data-points $x_{1:n}$ that is shattered by $\mathcal{H}$

ii) show that no set of $n+1$ points $x_{1:(n+1)}$ can be shattered.

Note that we always assume that $x_{1:n}$ consist of n distinct points as it cannot be shattered if there are repeated points.

Example 11. Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \left\{ h_{a,b} \,;\, a, b \in \mathbb{R},\ h_{a,b}(x) = \begin{cases} 1 & x \in [a,b) \\ 0 & x \notin [a,b) \end{cases} \right\}$. Given

$$x_1 < \cdots < x_n$$

and we have $n+1$ intervals $[x_i, x_{i+1})$.

How many distinct vectors [10] *are there in* $\{(h(x_i))_{i=1}^n \,;\, h \in \mathcal{H}\}$*?* If a, a' are in the same interval and b, b' are in the same interval, then $h_{a,b}(x_i) = h_{a',b'}(x_i)$ for all i. There are $n+1$ intervals in which we can place a and $n+1$ intervals to place b, so there are *at most* $(n+1)^2$ functions $h_{a,b}$ that give distinct behaviour on $x_{1:n}$. So, $\Pi_{\mathcal{H}}(n) \leqslant (n+1)^2$.

Note that $x_{1:2}$ can be shattered since letting $a' < a < x_1 < b < x_2 < c$, we have $h_{a',a}(x_{1:2}) = \binom{0}{0}$, $h_{a,b}(x_{1:2}) = \binom{1}{0}$, $h_{a',c}(x_{1:2}) = \binom{1}{1}$ and $h_{b,c}(x_{1:2}) = \binom{0}{1}$. Note however that $x_{1:3}$ (WLOG, assume $x_1 < x_2 < x_3$) cannot be shattered since there is no function $h_{a,b} \in \mathcal{H}$ that can satisfy $h_{a,b}(x_{1:3}) = (1,0,1)^\top$ since this implies $x_1, x_3 \in [a,b)$ but $x_2 \notin [a,b)$.

In the previous example, we observe that $\Pi_{\mathcal{H}}(n) \leqslant (n+1)^{\mathrm{VC}(\mathcal{H})}$. This is true more generally:

Lemma 12 (Sauer-Shelah Lemma)**.** *Let* $\mathcal{H}$ *be a class of finite* VC *dimension* d*. Then,* $\Pi_{\mathcal{H}}(n) \leqslant (n+1)^d$.

[10] We only make a crude upper bound here to demonstrate that growth function is at most polynomial in n, I will leave it to you in problem sheet 3 to work through this example in detail.

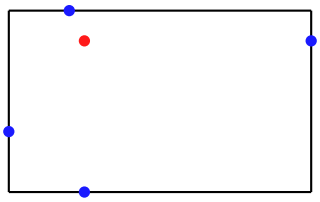


Figure 1.1: The VC dimension of indicator functions on rectangles is 4. You can check that any 4 points can be shatter, but no 5 points can be shattered. In this figure, you see that one cannot achieve the given labelling of red and blue points using any rectangle.

Note that by the definition of VC dimension, $\Pi_{\mathcal{H}}(n) < 2^n$ for all $n > d$. But could we have $\Pi_{\mathcal{H}}(n) < 1.8^n$ or $\Pi_{\mathcal{H}}(n) < 2^n - 1$? The answer is no. Beyond d, the growth of $\Pi_{\mathcal{H}}(n)$ is radically different as it is polynomial.

To summarize, $|\mathcal{H}(x_{1:n})| \leqslant \Pi_{\mathcal{H}}(n) \leqslant (n+1)^d$ where $d = \mathrm{VC}(\mathcal{H})$. So,

$$\mathbb{E}[R(\hat{h}) - \inf_{h\in\mathcal{H}} R(h)] \underbrace{\leqslant}_{Thm.9} 2\mathfrak{R}_n(\mathcal{F}) \underbrace{\leqslant}_{Lem.10} 2\sqrt{\frac{2\mathrm{VC}(\mathcal{H})\log(n+1)}{n}}$$

which converges to 0 as $n \to \infty$.

Example 12. Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0,1\}$ and $\mathcal{H} := \{\mathbf{1}_R \,;\, R = [a,b]\times[c,d], \quad a < b, \quad c < d\}$. The VC dimension of $\mathcal{H}$ is 4. Given 4 points $x_1, \ldots, x_4$ arranged in the shape of a diamond, there are 16 possible labellings all of which can be achieved by some $\mathbf{1}_R$ (try this). However, for any 5 points, let R_0 be the smallest rectangle containing $x_{1:5}$. Then each edge contains one point x_i (if not, make the rectangle smaller). If $x_1, \ldots, x_4$ lie on distinct edges, any rectangle that contains $x_1, \ldots, x_4$ must contain R_0 and hence x_5. Assign $+1$ to $x_1, \ldots, x_4$ and 0 to x_5: this labelling cannot be realised by any $\mathbf{1}_R$. See figure 1.1

Proof of the Sauer-Shelah Lemma. Let $x_1, \ldots, x_n \in \mathcal{X}$. Given any index set $Q \subset \{1, \ldots, n\}$, let $x_Q := \{x_i \,;\, i \in Q\}$. Let us first assume that there exist at least $|\mathcal{H}(x_{1:n})| - 1$ non empty sets Q such that $\mathcal{H}$ shatters x_Q. **If this is true,** then letting $x_{1:n}$ be such that $|\mathcal{H}(x_{1:n}| = \Pi_{\mathcal{H}}(n)$, we have

$$|\mathcal{H}(x_{1:n})| - 1 \leqslant \text{number of sets } Q \text{ shattered by } \mathcal{H} \leqslant \sum_{i=1}^{\min(d,n)} \binom{n}{i},$$

where, we sum only up to $\min(d,n)$ because no x_Q with $|Q| \geqslant d$ can be shattered by our initial assumption that $\mathcal{H}$ is of finite VC dimension d. Without loss of generality, assume that $n \geqslant d$, then the result immediately follows:

$$\sum_{i=1}^{d} \binom{n}{i} \leqslant n + n^2 + \ldots + n^d \leqslant (n+1)^d - 1.$$

It remains to show the above claim that there exist at least $|\mathcal{H}(x_{1:n})| - 1$ non empty sets Q such that $\mathcal{H}$ shatters x_Q. We will show this by induction: Assume that the functions in $\mathcal{H}$ takes values in $\{-1, 1\}$. If $|\mathcal{H}(x_{1:n})| = 1$, then the claim is trivially true. Assume that the claim is true for all $n \in \mathbb{N}$ and $|\mathcal{H}(x_{1:n})| \leqslant k$ for some $k \geqslant 1$.

Suppose that $|\mathcal{H}(x_{1:n})| = k + 1$. Since $|\mathcal{H}(x_{1:n})| \geqslant 2$, there exists x_j and $h, h' \in \mathcal{H}$ such that $h(x_j) = 1$ and $h'(x_j) = -1$. Define

$$\mathcal{H}_+ = \{h \in \mathcal{H} \,;\, h(x_j) = 1\} \quad \text{and} \quad \mathcal{H}_- = \{h \in \mathcal{H} \,;\, h(x_j) = -1\}.$$

These sets are non-empty and $h(x_{1:n}) \neq h'(x_{1:n})$ for all $h \in \mathcal{H}+$, $h' \in \mathcal{H}_-$. So,

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

In particular, $|\mathcal{H}_+(x_{1:n})|, |\mathcal{H}_-(x_{1:n})| \leqslant k$.

Let $\mathcal{X}_+, \mathcal{X}_-$ be sets of subvectors x_Q that are shattered by $\mathcal{H}_+$ and $\mathcal{H}_-$ respectively. By assumption,

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

If $x_Q \in \mathcal{X}_- \cup \mathcal{X}_+$, then x_Q can be shattered by $\mathcal{H} \supset \mathcal{H}_+, \mathcal{H}_-$ and none of the subvectors in $\mathcal{X}_+ \cup \mathcal{X}_-$ can contain x_j as $\mathcal{H}_+$ and $\mathcal{H}_-$ are constant on x_j. On the other hand, when $x_Q \in \mathcal{X}_- \cap \mathcal{X}_+$, both x_Q and $x_{Q\cup\{j\}}$ can be shattered by $\mathcal{H}$. Also, x_j is itself shattered by $\mathcal{H}$. So, the number of sets shattered by $\mathcal{H}$ is at least

$$|\mathcal{X}_- \cup \mathcal{X}_+| + 1 + |\mathcal{X}_- \cap \mathcal{X}_+| = 1 + |\mathcal{X}_+| + |\mathcal{X}_-| \geqslant k = |\mathcal{H}(x_{1:n})| - 1$$

as required. □

Chapter 2

Optimization

2.1 Preliminaries

Consider the following optimization problem

$$\inf_{w\in\mathbb{R}^p} f(w),$$

where $f:\mathbb{R}^p\to\mathbb{R}$. The goal is therefore to devise an "efficient" (iterative) algorithm to approximate a minimizer when it exists. The hope is that each iteration has low computational cost and we construct a sequence w_k that converges to a minimizer.

Motivating examples

- Learning linear models for the regression problem: For $w\in\mathbb{R}^p$, let $h_w(x)\coloneqq x^\top w$. The empirical risk minimization problem is to solve, for given data $(X_i,Y_i)\in\mathbb{R}^{p+1}$ and $i=1,\dots,n$ the following problem

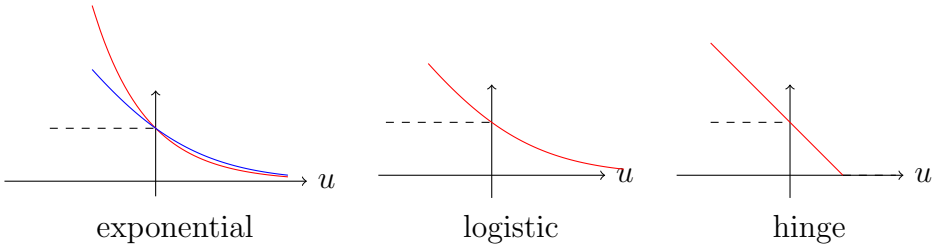
$$\min_{w\in\mathbb{R}^p} f(w),\quad \text{where}\quad f(w)=\frac{1}{n}\sum_{i=1}^n (Y_i-h_w(X_i))^2=\frac{1}{n}\left\|X_{1:n}^\top w-Y_{1:n}\right\|^2,$$

  where $X_{1:n}\in\mathbb{R}^{p\times n}$ is the matrix whose kth column is X_k and $Y_{1:n}=(Y_j)_{j=1}^n\in\mathbb{R}^n$.

- Consider the classification problem where $\mathcal{Y}=\{-1,1\}$. The empirical risk is

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\operatorname{sign}(X_i^\top w)\neq Y_i}=\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{(-\infty,0]}(Y_iX_i^\top w)\coloneqq f(w).$$

  The difficulty with the above loss is that it is non-differentiable, e.g. if w is such that $Y_iX_i^\top w\ll 0$ for all i, then $f(w)=1$ and a small perturbation of w will leave $f(w)$ unchanged – which direction should we move w to decrease $f(w)$? It is therefore desirable to consider a *smooth* approximation $\mathbf{1}_{(-\infty,0]}$, denoted by ℓ and solve $\min_{w\in\mathbb{R}^p}\sum_{i=1}^n \ell(Y_iX_i^\top w)$ instead. Some common choices are



exponential logistic hinge

i) The logistic loss $\ell(u) = \frac{\log(1+\exp(-u))}{\log(2)}$.

ii) The exponential loss $\ell(u) = \exp(-u)$.

iii) The hinge loss $\ell(u) = \max(1-u, 0)$.

Definition 5. *We say that $w^* \in \mathbb{R}^p$ is*

- *a global minimizer if for all $w \in \mathbb{R}^p$, $f(w^*) \leqslant f(w)$.*
- *a local minimizer if there exists an open neighbourhood U of w^* such that $f(w^*) \leqslant f(w)$ for all $w \in U$.*
- *a strict local minimizer if there exists an open neighbourhood U of w^* such that $f(w^*) < f(w)$ for all $w \in U \setminus \{w^*\}$.*
- *an isolated local minimizer if there exists an open neighbourhood U of w^* in which w^* is the only local minimizer.*

In general, f might not have a minimizer, e.g. $f(w) = -w^2$ with $\inf_w f(w) = -\infty$. Another example is $f(w) = \exp(-w)$ with $\inf_w f(w) = 0$ but no minimizer exists (issue is that f does not grow at $\pm\infty$). There might be multiple minimizers, e.g. $f(w) = -2*w^2+w^4$ or $f(w) = \max(w^4 - 2, 0)$.

Typically, to show that f has a minimizer, it is sufficient to show that there exists a bounded closed set $\Omega \subset \mathbb{R}^p$ such that $\inf_{w\in\mathbb{R}^p} f(w) = \inf_{w\in\Omega} f(w)$ and f is continuous (or lower-semicontinuous) on this set. We will not delve into these details but assume for the rest of this chapter a minimizer exists. We will denote the set of minimizers by

$$\operatorname{argmin}(f) := \left\{ w \in \mathbb{R}^p \; ; \; f(w) = \inf_{w\in\mathbb{R}^p} f(w) \right\}.$$

2.1.1 Convexity

We study the case where f is a convex (and differentiable) function. Convex functions are more straightforward to minimize, since all local minimizers are global minimizers. Moreover, in the convex setting (at least when f is smooth), there exists efficient algorithms to find a global minimizer.

Definition 6. *We say that $C \subset \mathbb{R}^p$ is convex if for all $v, w \in C$ and all $\lambda \in [0,1]$, $\lambda v + (1-\lambda)w \in C$. We say that $f : S \to \mathbb{R}$ is*

- *convex if $S \subset \mathbb{R}^p$ is convex and for all $v, w \in S$ and all $\lambda \in [0,1]$,*

$$f(\lambda v + (1-\lambda)w) \leqslant \lambda f(v) + (1-\lambda) f(w).$$

  We say that f is strictly convex if for all $v, w \in S$ and all $\lambda \in (0,1)$ with $v \neq w$, $f(\lambda v + (1-\lambda)w) < \lambda f(v) + (1-\lambda) f(w)$.

- *concave if $-f$ is convex.*

Examples of convex functions

- $f(x) = x^3$ is convex in $\mathbb{R}_{\geqslant 0}$ but not convex on $\mathbb{R}$.
- Linear functions $f(x) = ax + b$ is convex on $\mathbb{R}$.
- Note that convex functions are not necessarily smooth, e.g. $f(x) = |x|$ is convex but not differentiable at 0.

Convexity is stable under many transformations

- If f, g are convex, then $\alpha f + \beta g$ is convex for all $\alpha, \beta \geqslant 0$.
- If f, g are convex, then $\max(f, g)$ is convex.
- If $g : \mathbb{R}^q \to \mathbb{R}$ is convex and $B \in \mathbb{R}^{q \times p}$ and $b \in \mathbb{R}^q$, then $f(x) = g(Bx + b)$ is convex.

Theorem 13. *Let $f : \mathbb{R}^p \to \mathbb{R}$ be convex. Then any local minimizer of f is a global minimizer. Moreover, if f is strictly convex and has a minimizer, then it is unique.*

Proof. For the first statement, let v be a local minimizer and assume that it is not a global minimizer. So, there exists $w \in \mathbb{R}^p$ such that $f(w) < f(v)$. For all $\lambda \in [0,1]$, let $v_\lambda = \lambda w + (1-\lambda)v = v + \lambda(w - v)$. Given any open neighbourhood U of v, $v_\lambda \in U$ for all λ sufficiently small, but $f(v_\lambda) \leqslant \lambda f(w) + (1-\lambda) f(v) < f(v)$ contradicts the assumption that v is a local minimizer.

We prove the second statement by contradiction: suppose u, v are distinct global minimizers and f is strictly convex, then

$$\inf_w f(w) \leqslant f(\frac{1}{2}(u+v)) < \frac{1}{2} f(u) + \frac{1}{2} f(v) = \inf_w f(w).$$

□

Gradients It is useful to characterize convexity through differentiability. Recall that if f is differentiable along each axis, then

$$\nabla f(w) = \left(\frac{\partial f(w)}{\partial w_1}, \quad \frac{\partial f(w)}{\partial w_2}, \quad \cdot, \frac{\partial f(w)}{\partial w_p} \right)^\top \in \mathbb{R}^p.$$

Note that $\nabla f : \mathbb{R}^p \to \mathbb{R}^p$. The partial derivatives are defined by

$$\frac{\partial f(w)}{\partial w_k} = \lim_{t \to 0} \frac{f(w + t\delta_k) - f(w)}{t}$$

where δ_k is the vector whose kth entry is 1 and all other entries are zero.

Recall also that f is said to be differentiable at w if

$$\lim_{v \to 0} \frac{f(w+v) - f(w) - \langle \nabla f(w),\, v \rangle}{\|v\|} = 0$$

We can equivalently write $f(w+v) = f(w) + \langle \nabla f(w),\, v \rangle + o(\|v\|)$.

Finally, the Hessian of f (if it exists) is denoted by $\nabla^2 f(w) = \left(\frac{\partial f}{\partial w_i \partial w_j}(w) \right)_{i,j=1}^{p} \in \mathbb{R}^{p \times p}$.

Theorem 14 (Characterization of convexity via differentiability)**.** *i) If f is differentiable, then f is convex if and only if for all $v, w \in \mathbb{R}^p$, $f(w) \geqslant f(v) + \nabla f(v)^\top (w - v)$.*

ii) If f is twice-differentiable, then f is convex if and only if $\nabla^2 f(v)$ is positive semi-definite for all v. If moreover $\nabla^2 f(v)$ is positive definite for all v, then f is strictly convex.

Recall: A is positive semidefinite if $v^\top A v \geqslant 0$ for all v (and positive definite if $\geqslant$ is replaced by $>$).

Proof. We prove i). Assume that f is convex. Then, for all $\lambda \in [0,1]$, $f((1-\lambda)v + \lambda w) \leqslant (1-\lambda) f(v) + \lambda f(w)$. Rearranging yields

$$\frac{f(v + \lambda(w - v)) - f(v)}{\lambda} \leqslant f(w) - f(v). \tag{2.1}$$

The LHS converges to $\langle \nabla f(v),\, w - v \rangle$ as $\lambda \to 0$. So, $\langle \nabla f(v),\, w - v \rangle \leqslant f(w) - f(v)$ as required.

For the converse, let $v_\lambda = (1-\lambda)v + \lambda v'$.

$$f(v) \geqslant f(v_\lambda) + \langle \nabla f(v_\lambda),\, v - v_\lambda \rangle = f(v_\lambda) + \lambda \langle \nabla f(v_\lambda),\, v - v' \rangle \tag{2.2}$$
$$f(v') \geqslant f(v_\lambda) + \langle \nabla f(v_\lambda),\, v' - v_\lambda \rangle = f(v_\lambda) - (1-\lambda) \langle \nabla f(v_\lambda),\, v - v' \rangle. \tag{2.3}$$

Multiplying (2.2) by $(1-\lambda)$ and (2.3) by λ and summing the two inequalities imply that

$$(1-\lambda) f(v) + \lambda f(v') \geqslant f(v_\lambda).$$

The proof of ii) is left as an exercise.

□

In general, it is known that $f : \mathbb{R}^p \to \mathbb{R}$ be a differentiable function and w^* is a minimizer of f, then $\nabla f(w^*) = 0$. Indeed, By Taylor expansion, $f(w^*) \leqslant f(w^* + tu) = f(w^*) + t\langle \nabla f(w^*), u\rangle + o(t)$. So, $\langle \nabla f(w^*), u\rangle \geqslant 0$. This is true for all u and in particular, taking $-u$, we have $\langle \nabla f(w^*), u\rangle = 0$. So, $\nabla f(w^*)$.

However, the converse if false: $f(x) = -x^2$ has $\nabla f(0) = 0$ but 0 is a maximizer. Or, $f(x) = x^3$ has $\nabla f(x) = 0$ but 0 is not a local minimizer. The following propositions shows that for *convex* and differentiable functions, $\nabla f(w^*) = 0$ is equivalent to w^* being a minimizer.

Proposition 15 (First order optimality condition in the unconstrained case)**.** *Suppose that f is convex and differentiable. Then, $w^* \in \operatorname{argmin}(f)$ if and only if $\nabla f(w^*) = 0$.*

Proof. We already know that w^* is a minimizer implies that $\nabla f(w^*) = 0$. For the converse, assume that $\nabla f(w^*) = 0$. By convexity, $f(w) \geqslant f(w^*) + \langle \nabla f(w^*), w - w^*\rangle = f(w^*)$ holds for all w. So, w^* is a minimizer. □

Proposition 16 (First order optimality condition in the constrained case)**.** *Let f be convex and differentiable and consider $\min_{w\in\mathcal{F}} f(w)$ for some convex set $\mathcal{F}$. Then, w^* is a minimizer if and only if*

$$\forall w \in \mathcal{F}, \quad \nabla f(w^*)^\top (w - w^*) \geqslant 0.$$

Proof. We first assume that $\nabla f(w^*)^\top (w - w^*) \geqslant 0$. Since f is convex,

$$f(w) \underbrace{\geqslant}_{f \text{ convex}} f(w^*) + \langle \nabla f(w^*), w - w^*\rangle \underbrace{\geqslant}_{\text{assump.}} f(w^*).$$

This holds for all $w \in \mathcal{F}$, so w^* is a minimizer.

For the converse direction, assume that $w^* \in \mathcal{F}$ is a minimizer but $\nabla f(w^*)^\top (w - w^*) < 0$ for some $w \in \mathcal{F}$. Let $\lambda \in [0, 1]$ and define $v(\lambda) = (1 - \lambda)w^* + \lambda w \in \mathcal{F}$. Then, letting $g(\lambda) = f(v(\lambda))$,

$$g'(0) = \langle \nabla f(w^*), w - w^*\rangle < 0.$$

This implies that g is decreasing at $\lambda = 0$ and hence, for all λ sufficiently small, $f(v(\lambda)) < f(v(0)) = f(w^*)$. Contradiction. □

Remark 10. If $\mathcal{F} = \mathbb{R}^p$, then $\nabla f(w^*)^\top v \geqslant 0$ for all v, so $\nabla f(w^*) = 0$. This is the first order optimality condition in the unconstrained setting.

Example 13 (Least squares). Let $f(w) = \|Aw - b\|^2$ for $A \in \mathbb{R}^{n\times p}$ and $b \in \mathbb{R}^n$. Then, $\nabla f(w) = 2A^\top (Aw - b)$, indeed,

$$f(w + \varepsilon) = \|Aw - b + A\varepsilon\|^2 = f(w) + 2\langle \varepsilon, A^\top (Aw - b)\rangle + \|A\varepsilon\|^2 .$$

So, $f(w + \varepsilon) = f(w) + 2\langle \varepsilon, A^\top (Aw - b)\rangle + o(\|\varepsilon\|)$ and $\nabla f(w) = 2A^\top (Aw - b)$. From Proposition 15, w^* is a minimizer if and only if $A^\top Aw^* = A^\top b$. If $A^\top A$ is invertible (i.e. $\ker(A) = \{0\}$), then $w^* = (A^\top A)^{-1} A^\top y$.

2.2 Constrained optimization

We now consider the following constrained optimization problem

$$\min_{w\in\mathbb{R}^p} f_0(w) \quad \text{subject to} \quad Aw = b \quad \text{and} \quad \forall i = 1,\dots,m,\ f_i(w) \leqslant 0 \tag{P}$$

where $A \in \mathbb{R}^{n\times p}$, $b \in \mathbb{R}^p$ and $f_i : \mathbb{R}^p \to \mathbb{R}$. We assume that f_i are convex functions. The feasibility set of (P) is

$$\mathcal{F} \coloneqq \{w \in \mathbb{R}^p \ ;\ \forall i = 1,\dots,m,\ \ f_i(w) \leqslant 0 \quad \text{and} \quad Aw = b\}\,.$$

One can show that $\mathcal{F}$ is a convex set. If f_i and f are also linear functions, then $\mathcal{F}$ is a polyhedron and (P) is called a *linear programming* problem.

Example 14. Let $\mathcal{H} \coloneqq \left\{h_w(x) \coloneqq w^\top\varphi(x)\ ;\ \|w\| \leqslant K\right\}$. Then, the ERM problem is

$$\min_{\|w\|\leqslant K} f_0(w), \quad \text{where} \quad f_0(w) \coloneqq \frac{1}{n}\sum_{i=1}^n \ell(h_w(X_i), Y_i)$$

We can place this in the form (P) by letting $m = 1$ and $f_1(w) = \sum_i w_i^2 - K^2$.

Example 15. Consider the so-called support vector machine (SVM): Suppose we are given data (X_i, Y_i) with $Y_i \in \{-1, 1\}$. The SVM is

$$\min_{w,b} \|w\| \quad \text{s.t.} \quad \forall i = 1,\dots,N, \quad Y_i(w^\top X_i + b) \geqslant 1.$$

We will derive this problem later, but it essentially finds the hyperplane that separates positively and negatively labelled data points that has the largest margin.

2.2.1 Duality

In this section, we introduce the idea of duality. In optimization, one can often view the problem from two alternative perspectives: the primal problem and the dual problem. These alternative perspectives can sometimes lead to more efficient solvers and different intuitions on an optimization problem.

Definition 7 (Duality)**.**

- *The* ***Lagrange function*** *of* (P) *is defined for all* $w \in \mathbb{R}^p$, $\xi \in \mathbb{R}^n$ *and* $\nu \in \mathbb{R}^m_{\geqslant 0}$ *by*

$$L(w, \xi, \nu) \coloneqq f_0(w) + \xi^\top(Aw - b) + \sum_{k=1}^m \nu_k f_k(w).$$

  The vectors ξ, ν *are called the* ***Lagrange multipliers****.*

- *The* ***Lagrange dual function*** *is defined by*

$$D(\xi, \nu) \coloneqq \inf_{w\in\mathbb{R}^p} L(w, \xi, \nu).$$

  If $w \mapsto L(w, \xi, \nu)$ *is unbounded from below, then we write* $D(\xi, \nu) = -\infty$.

We observe two key properties of the dual function (which then motivates the two definitions that follow:

- **The dual function is concave:** Note that $(\xi, \nu) \mapsto D(w, \xi, \nu)$ is linear in ξ and ν, and since $D(\xi, \nu)$ is the pointwise infimum of affine functions, D is always concave, even if f_i are not convex.
- The dual function $D(\xi, \nu)$ provides a **lower bound** for $\inf_{w \in \mathcal{F}} f_0(w)$: If $w \in \mathcal{F}$ then $Aw - b = 0$ and $f_j(w) \leqslant 0$ for all $j = 1, \ldots, m$. So, $\xi^\top (Aw - b) + \sum_{i=1}^m \nu_i f_i(w) \leqslant 0$ and hence,
$$D(\xi, \nu) \leqslant L(w, \xi, \nu) \leqslant f_0(w), \qquad \forall w \in \mathcal{F}.$$
Taking the inf on the RHS implies that $D(\xi, \nu) \leqslant \inf_{w \in \mathcal{F}} f_0(w)$.

Definition 8 (The dual problem)**.** *The dual optimization problem to* (P) *is*
$$\max_{\xi \in \mathbb{R}^n, \nu \in \mathbb{R}^m} D(\xi, \nu) \quad s.t. \quad \nu \geqslant 0 \tag{$\mathcal{D}$}$$

Definition 9 (Weak and strong duality)**.** *Let* (ξ^*, ν^*) *maximize* ($\mathcal{D}$) *and* w^* *minimize* (P)*. We always have* $D(\xi^*, \nu^*) \leqslant f_0(w^*)$ *and this is called* ***weak duality****. If we have* $D(\xi^*, \nu^*) = f_0(w^*)$*, then we call this* ***strong duality****.*

Example 16. Consider
$$\min_{x,y} \exp(-x), \quad \text{subject to} \quad h(x, y) \leqslant 0$$
where $h(x, y) = \frac{x^2}{y}$ has domain $\{(x, y) \ ; \ y > 0\}$. Note that h is convex (check its Hessian and show that it is positive definite). The primal problem has minimizers $(0, y)$ for any $y > 0$. We in particular have $\min_{x,y} \{\exp(-x) \ ; \ h(x, y) \leqslant 0\} = 1$.

For the dual problem,
$$L(x, y, z) = \exp(-x) + z h(x, y)$$
and for all $z \geqslant 0$,
$$D(z) = \inf \left\{\exp(-x) + z x^2/y \ ; \ y > 0\right\} = 0$$
So,
$$\sup_{z \geqslant 0} D(z) = 0.$$
There is no strong duality here.

Some natural questions are

- When does strong duality hold? This will be addressed by Slater's theorem below.
- If we have strong duality, how are ξ^*, ν^*, w^* related? The conditions relating these quantities are call the KKT conditions.

Theorem 17 (Slater's constraint quantification)**.** *Assume that* f_i *are convex for* $i = 0, \ldots, m$ *with* $\mathrm{dom}(f_0) = \mathbb{R}^p$*. If there exists* $w \in \mathbb{R}^p$ *such that* $Aw = b$ *and* $f_\ell(w) < 0$ *for all* $\ell = 1, \ldots, m$*, then strong duality holds for* (P) *(i.e. there exists a point in the interior of* $\mathcal{F}$*).*

Proof. See Section 5.3.2 of Convex Optimization by Boyd and Vandenberghe. □

The Karush-Kuhn-Tucker (KKT) conditions Suppose that (w^*, ξ^*, ν^*) optimize (P) and ($\mathcal{D}$) and strong duality holds. Then,

$$f_0(w^*) = D(\xi^*, \nu^*) = \inf_{w \in \mathbb{R}^p} f_0(w) + \sum_{i=1}^m \nu_i^* f_i(w) + \langle \xi^*,\, Aw - b \rangle \leqslant f_0(w^*) + \sum_{i=1}^m \nu_i^* f_i(w^*) \leqslant f_0(w^*).$$

So, all the inequalities are equalities and we have

$$\sum_{i=1}^m \nu_i^* f_i(w^*) = 0 \implies \forall i = 1, \ldots, m,\ \nu_i^* f_i(w^*) = 0$$

This is called the complementary slackness condition. Moreover, since $w^* \in \operatorname{argmin}_w L(w, \xi^*, \nu^*)$, we have $\nabla_w L(w^*, \xi^*, \nu^*) = 0$.

Theorem 18 (Karush-Kuhn-Tucker)**.** *If w^* and (ξ^*, ν^*) solve (P) and ($\mathcal{D}$) and strong duality holds, then the following KKT conditions hold:*

$$\begin{aligned} f_i(w^*) \leqslant 0, \quad Aw^* = b, \quad \nu^* \geqslant 0 \\ \nu_j^* f_j(w^*) = 0, \quad \forall j = 1, \ldots, m \\ \nabla f(w^*) + \sum_i \nu_i \nabla f_i(w^*) + A^\top \xi^* = 0 \end{aligned}$$

If (P) is convex and the Slater conditions hold, then any (w^*, ξ^*, ν^*) that satisfy the KKT conditions are solutions to (P) and ($\mathcal{D}$).

Example 17. Consider

$$\min_x \|x\|^2 \quad \text{s.t.} \quad Ax = b.$$

If there exists x such that $Ax = b$, then strong duality holds. In this case,

$$L(x, z) = \|x\|^2 - \langle Ax - b,\, z \rangle$$

and $2x - A^\top z = 0$

$$D(z) = \min_x \|x\|^2 - \langle Ax - b,\, z \rangle = -\frac{1}{4} \left\|A^\top z\right\|^2 + \langle b,\, z \rangle.$$

The dual problem is

$$\max_z \frac{1}{4} \left\|A^\top z\right\|^2 + \langle b,\, z \rangle.$$

Example 18.

$$D(\xi, q) = \min_w -\frac{1}{2} \|q\|^2 + \xi(\|w\|^2 - K) + \langle q,\, Xw - b \rangle$$

By the optimality condition,

$$-\frac{1}{2\xi} X^\top q = w$$

$$D(\xi, q) = \min_w -\frac{1}{2} \|q\|^2 - \xi K - \langle q,\, b \rangle - \frac{1}{4\xi} \left\|X^\top q\right\|^2$$

So the dual problem is

$$\max_{\xi > 0} \max_{q \in \mathbb{R}^n} -\frac{1}{2} \|q\|^2 - \xi K - \langle q,\, b \rangle - \frac{1}{4\xi} \left\|X^\top q\right\|^2.$$

2.3 Support vector machines

Given data (X_i, Y_i) with $Y_i \in \{-1, 1\}$ and $X_i \in \mathbb{R}^p$, consider the problem of finding a linear classifier $h(x) = \text{sign}(w^\top x + b)$ for $w \in \mathbb{R}^p$ and $b \in \mathbb{R}$. Note that

$$h(X_i) = Y_i \iff Y_i(w^\top X_i + b) > 0.$$

In particular, we would like to separate the data points X_i corresponding to label $+1$ from the data points corresponding to label -1 using a hyperplane $w^\top x + b$. Assuming that this data set is linearly separable, there are infinitely many separating hyperplanes. So, which one should we choose?

Intuitively, we should choose the one that maximizes the distance to the closest data point in both classes: Denote the hyperplane by $H \coloneqq \left\{x \; ; \; w^\top x + b = 0\right\}$. The margin for the hyperplane H and dataset $\{X_i\}_i$ is

$$\gamma \coloneqq \min \left\{ \|x - X_i\| \; ; \; i = 1, \dots, n, \; x \in H \right\}.$$

2.3.1 Distance to the hyperplane

Given $x \in \mathbb{R}^p$, compute $\min_{z \in H} \|x - z\|$.

Geometric argument Observe that the projection of x onto the hyperplane H is of the form $z = x + tw$ for some $t \in \mathbb{R}$, since w is orthogonal to the hyperplane. Since we require that z is on the hyperplane,

$$-b = z^\top w = x^\top w + t \|w\|^2 \iff t = \frac{-b - x^\top w}{\|w\|^2}.$$

So,

$$\min_{z \in H} \|x - z\| = |t| \, \|w\| = \frac{\left|b + x^\top w\right|}{\|w\|}.$$

Dual approach Just for illustrations purposes, we can also compute $\min_{z \in H} \|x - z\|$ using the duality theory we developed in the previous section. First write as the optimization problem

$$\min_z \|z - x\|^2 \quad \text{s.t.} \quad w^\top z + b = 0.$$

The dual problem is defined for $\xi \in \mathbb{R}$ by

$$\max_\xi D(\xi) = \min_z \|z - x\|^2 + \xi(w^\top z + b).$$

Note that there is strong duality since the primal problem is convex and there exists z such that $w^\top z = -b$. The minimizer z^* in $D(\xi)$ satisfies $2(z^* - x) + \xi w = 0 \implies z^* = -\frac{1}{2}\xi w + x$. So,

$$D(\xi) = -\frac{1}{4}\xi^2 \|w\|^2 + \xi(b + w^\top x).$$

The maximizer ξ^* to D satisfies

$$\nabla D(\xi^*) = -\xi^* \|w\|^2 /2 + b + w^\top x = 0. \tag{2.4}$$

So, $\xi^* = \frac{2(b+w^\top x)}{\|w\|^2}$. The primal solution therefore satisfies

$$z^* = x - \frac{(b + w^\top x)}{\|w\|^2} w \implies \|z^* - x\| = \min_{z \in H} \|z - x\|^2 = \frac{\left|b + w^\top x\right|}{\|w\|}$$

2.3.2 Primal formulation of SVM

The margin of H is therefore

$$\gamma(w, b) = \min_{i=1}^{n} \frac{\left|b + w^\top X_i\right|}{\|w\|}. \tag{2.5}$$

The maximum margin classifier (SVM) Motivated the margin formula in (2.5), we want to find w, b to maximize

$$\max_{w,b} \min_{i=1}^{n} \frac{\left|b + w^\top X_i\right|}{\|w\|} \quad \text{s.t.} \quad \forall i,\ Y_i(w^\top X_i + b) \geqslant 0.$$

Note that $\gamma(tw, tb) = \gamma(w, b)$ for all $t \in \mathbb{R} \setminus \{0\}$. Due to this scale invariance, we can fix $\min_{i=1}^n \left|b + w^\top X_i\right| = 1$ and consider

$$\text{argmax}_{w,b} \frac{1}{\|w\|} \quad \text{s.t.} \quad \forall i,\ Y_i(w^\top X_i + b) \geqslant 0 \quad \text{and} \quad \left|b + w^\top X_i\right| \geqslant 1.$$

One can check that the constraint is equivalent to imposing $Y_i(w^\top X_i + b) \geqslant 1$ and hence, we arrive at the problem

$$\text{argmin}_{w,b} \|w\| \quad \text{s.t.} \quad \forall i,\ Y_i(w^\top X_i + b) \geqslant 1. \tag{2.6}$$

This is called the support vector machine and the data points for which $Y_i(w^\top X_i + b) = 1$ are the support vectors – removal of these points will lead the separating hyperplane to change.

2.3.3 The dual problem for SVM

Note that (2.6) is a convex optimization problem. If we assume that there exists a linear separating hyperplane, that is there exists w, b such that for all i,

$$(w^\top X_i + b)Y_i > 1,$$

then Slater's conditions hold and there is strong duality.

The Lagrange function is defined for $\xi \in \mathbb{R}^n_{\geqslant 0}$ and $(w, b) \in \mathbb{R}^{p+1}$ by

$$L(w, b, \xi) = \frac{1}{2} \|w\|^2 + \sum_{i=1}^{n} \xi_i \left(1 - (w^\top X_i + b) Y_i\right)$$

To work out the dual problem for (2.6), first note that if $\langle Y, \xi \rangle \neq 0$, then

$$D(\xi) = \inf_{w,b} L(w, b, \xi) = \inf_{w} \frac{1}{2} \|w\|^2 + \sum_{i=1}^{n} \xi_i \left(1 - (w^\top X_i) Y_i\right) + \inf_{b} -b\langle Y, \xi \rangle = -\infty.$$

Moreover, if $\langle Y, \xi \rangle = 0$ and (w, b) minimize $\min_{w,b} L(w, b, \xi)$, then they satisfy the first order optimality conditions

$$\partial_w L(w, b, \xi) = w - \sum_i \xi_i X_i Y_i = 0 \quad \text{and} \quad \partial_b L(w, b, \xi) = -\sum_i \xi_i Y_i = 0.$$

The dual function therefore satisfies

$$D(\xi) = \begin{cases} -\frac{1}{2} \left\| \sum_i \xi_i X_i Y_i \right\|^2 + \sum_i \xi_i & \text{if } Y^\top \xi = 0, \\ -\infty & \text{otherwise.} \end{cases}$$

The dual problem [1] is therefore

$$\max_{\xi} -\frac{1}{2} \left\| \sum_{i=1}^{n} \xi_i X_i Y_i \right\|^2 + \sum_i \xi_i \qquad \text{s.t.} \quad Y^\top \xi = 0 \quad \text{and} \quad \forall i,\ \xi_i \geqslant 0.$$

Once we solve the dual problem to obtain the maximizer ξ^*, we can retrieve the optimal w^* as $w^* = \sum_i \xi_i^* X_i Y_i$ and the optimal b is defined using the complementary slackness condition

$$\xi_i (1 - Y_i (w^\top X_i + b)) = 0.$$

One simply chooses $\xi_i \neq 0$ and solve for b. Note that $\xi_i \neq 0$ only if $1 = Y_i(w^\top X_i + b))$. The points for which $\xi_i \neq 0$ are called the support points and indicates whether the data point X_i (called the support vector) contributes to the decision boundary.

2.3.4 Non-exact separation

In the case where the data cannot be linearly separated, the constraints in (2.6) are infeasible. To handle this, one modify (2.6) to allow for a small number of *mistakes*. In particular, introduce slack variables $s_i \geqslant 0$ that measure the amount of misfit:

$$\text{argmin}_{w,b,s} \|w\|^2 + \mu \sum_i s_i \quad \text{s.t.} \quad Y_i (w^\top X_i + b) \geqslant 1 - s_i, \quad s_i \geqslant 0 \tag{2.7}$$

[1] Popular methods for solving this include projected gradient descent, where one performs gradient descent and project onto the constraint set at each iteration

The parameter μ controls the amount of misfit. Note that this can be rewritten as

$$\begin{aligned}&\operatorname{argmin}_{w,b,s} \|w\|^2 + \mu \sum_i s_i \quad \text{s.t.} \quad s_i \geqslant \max\left(0, 1 - Y_i(w^\top X_i + b)\right)\\ &= \operatorname{argmin}_{w,b} \|w\|^2 + \mu \sum_i \max\left(0, 1 - Y_i(w^\top X_i + b)\right).\end{aligned}$$

2.3.5 Nonlinear decision boundaries

As mentioned, it is not always possible to linearly separate the data into two classes. One can add in the slack variables as in the previous section, but this still leads to linear decision boundaries. To allow for nonlinear decision boundaries, instead of linear models $h(x) = w^\top x$ for some vector $w \in \mathbb{R}^p$, one can define a 'feature map' $\Phi : \mathcal{X} \to \mathbb{R}^d$ and write $\Phi(x) = (\varphi_j(x))_{j=1}^d$ for some basis functions $\varphi_j : \mathcal{X} \to \mathbb{R}$. The idea is that the feature map φ lifts the data points x into a possibly larger dimensional space in which linear separation becomes possible. We consider the hypothesis class

$$\mathcal{H} = \left\{h(x) = w^\top \Phi(x) + b \,;\; w \in \mathbb{R}^d, b \in \mathbb{R}\right\}.$$

This allows us to learn *nonlinear* models. For example, $\mathcal{X} = \mathbb{R}^2$ and consider the feature map $x = (x_1, x_2) \mapsto \Phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)^\top$. Then, $h(x) = w^\top \Phi(x) + b$ is a polynomial of degree 2.

Remark 11. Observe that $h(x) = w^\top \Phi(x) + b$ is a *nonlinear* function in x but it is linear in the parameters (w, b).

The previous derivations for SVM also extend immediately to this setting, where instead of separating the data points $(X_i)_{i=1}^n$, we separate the 'lifted' points $(\Phi(X_i))_{i=1}^n$. This leads to the following primal and corresponding dual problems

$$\operatorname{argmin}_{w \in \mathbb{R}^d, b \in \mathbb{R}} \|w\| \quad \text{s.t.} \quad Y_i(w^\top \Phi(X_i) + b) \geqslant 1. \tag{2.8}$$

and dual problem

$$\max_{\xi \in \mathbb{R}^n} -\frac{1}{2}\left\|\sum_i \xi_i \Phi(X_i) Y_i\right\|^2 + \sum_i \xi_i \qquad \text{s.t.} \quad Y^\top \xi = 0 \quad \text{and} \quad \xi \geqslant 0. \tag{2.9}$$

Kernels

Note that if we define the **kernel** $k(x, x') = \Phi(x)^\top \Phi(x')$, then (2.9) is equivalent to

$$\max_{\xi \in \mathbb{R}^n} -\frac{1}{2}\langle (\xi \odot Y),\, K_n(\xi \odot Y)\rangle + \sum_i \xi_i \qquad \text{s.t.} \quad Y^\top \xi = 0 \quad \text{and} \quad \xi \geqslant 0. \tag{2.10}$$

where we denoted the matrix $K_n \coloneqq (k(X_i, X_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}$ and the vector $\xi \odot Y \coloneqq (\xi_i Y_i)_{i=1}^n$.

The classifier $h(x) = \Phi(x)^\top w^* + b_*$ where w^* and b^* solve the primal problem can also be expressed purely in terms of the kernel function. In particular, since $w^* = \sum_i \xi_i^* \Phi(X_i) Y_i$, we have

$$h(x) = \sum_i \xi_i^* Y_i k(X_i, x) + b^*,$$

and b^* can be computed by choosing some i such that $\xi_i \neq 0$ and solving

$$0 = \xi_i^*(1 - Y_i((w^*)^\top \Phi(X_i) + b) = \xi_i^*(1 - Y_i(\sum_j \xi_j^* k(X_j, X_i) Y_j + b).$$

Observe that one can directly compute $\hat{h}$ and the coefficients c_i in terms of the kernel function. In practice, Φ might embed the data into a very high (or even infinite) dimensional space, but since the dual problem only requires evaluating the kernel, there is no need to handle the Φ explicitly.

Polynomial kernels Some common kernels are the linear kernel $k(x, x') = x^\top x'$, or the polynomial kernels $k(x, x') = (x^\top x' + 1)^d$ for $d \in \mathbb{N}$. Consider the case of $p = 2$ and $d = 2$, then

$$k(x, z) = (x^\top z + 1)^2 = (\sum_{i=1}^2 x_i z_i + 1)^2 = \sum_{i,j=1}^2 (x_i x_j)(z_i z_j) + \sum_{i=1}^2 (\sqrt{2} x_i)(\sqrt{2} z_i) + 1$$

and the corresponding feature vector is

$$\varphi(x) = \begin{pmatrix} x_1 x_1 & x_1 x_2 & x_2 x_1 & x_2 x_2 & \sqrt{2} x_1 & \sqrt{2} x_2 & 1 \end{pmatrix}$$

In general, if $x \in \mathbb{R}^p$, $k(x, z) = (x^\top z + 1)^d$ corresponds to a features space of dimension $\binom{p+d}{d}$ corresponding to all monomials up to order d. Despite working in $\mathcal{O}(p^d)$ dimensional space, computing $K(x, z)$ is $\mathcal{O}(p)$.

Gaussian kernel Another popular choice is the Gaussian kernel

$$k(x, x') = \exp(-\left\|x - x'\right\|^2 / (2\sigma^2)), \quad \text{where} \quad \sigma > 0$$

(also called the radial basis function kernel). The associated feature map in this case is infinite dimensional: For simplicity, let $\sigma = 1$. Then,

$$k(x, y) = \exp(-\left\|y\right\|^2 / 2) \exp(-\left\|x\right\|^2 / 2) \exp(x^\top y) = \exp(-\left\|y\right\|^2 / 2) \exp(-\left\|x\right\|^2 / 2) \sum_{n=0}^\infty \frac{(x^\top y)^n}{n!}$$

$$= \exp(-\left\|y\right\|^2 / 2) \exp(-\left\|x\right\|^2 / 2) \sum_{n=0}^\infty \sum_{k_1 + \cdots k_p = n} \frac{1}{k_1! \cdots k_p!} \prod_{j=1}^p (x_j y_j)^{k_j}$$

where we used the multinomial theorem $(z_1 + \cdots z_p)^n = \sum_{k_1+\cdots+k_p=n} \frac{n!}{k_1!\cdots k_p!} \prod_{j=1}^p z_j^{k_j}$. So, we can let

$$\Phi(x) = \left(\exp(-\|x\|^2/2) \frac{1}{\sqrt{k_1! \cdots k_p!}} \prod_{j=1}^p x_j^{k_j} \right)_{\substack{n \in \mathbb{N}_0 \\ k_1+\cdots+k_p=n}}.$$

Note however that the feature map Φ associated to k is not unique. Another possibility is to let $\Phi : \mathcal{X} \to L^2(\mathbb{R})$ map into the function space of functions where $\int |f(x)|^2 dx < \infty$, with $\Phi(x) = \sqrt{(2\sigma)/\pi} \exp(-2\sigma^2 \|x - \cdot\|^2)$.

Remark 12. Note from the above discussion that

$$h(x) = \sum_{i=1}^n a_j K(X_i, x) + b$$

for some $a_j \in \mathbb{R}$ and $b \in \mathbb{R}$. If we consider the *sigmoid kernels* where

$$K(x, x') = \tanh(\alpha(x^\top x') + \beta), \quad \text{where} \quad \alpha, \beta > 0,$$

this actually looks like a simple neural network with the inner weights being the data points X_i!

Practical considerations

- In general, one needs to choose the kernel. The choice is not always obvious, but polynomial and Gaussian kernels are popular choices.
- Recall the issues of trade-off between approximation error and estimation error. Choosing a very high dimensional feature map will lead to better approximation error, but worse estimation error in general. In practice, you would choose the d in the polynomial kernel or the σ in the Gaussian kernel using cross validation.
- In general, kernel methods have at least $\mathcal{O}(n^2)$ complexity since you need to evaluate the kernel matrix K_n of all pairwise interactions. There are approaches such as using randomized feature maps (see Rahimi and Recht, Random Features for Large-Scale Kernel Machines, NIPS 2007)) to deal with large dimensions, but in general, kernel SVM tends to work best with small or medium sized datasets

2.4 Gradient descent

Iterative methods In general, we cannot compute in closed form $w^* \in \operatorname{argmin} f$. Even if a minimizer exists, it might be computationally infeasible to directly compute w^* (e.g. for the least squares example, computing w^* requires inverting a $n \times n$ linear system). We will typically construct a sequence $w_k \in \mathbb{R}^p$ such that $w_k \to w^*$ as $k \to \infty$. Generally, we will initialize with some $w_0 \in \mathbb{R}^p$ and define iteratively $w_{k+1} = w_k + p_k$ where p_k should be chosen such that $f(w_{k+1}) < f(w_k)$.