Modern Statistical Methods

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

Course webpage:
http://www.statslab.cam.ac.uk/~rds37/modern_stat_methods.html

The field of statistics has undergone profound changes in recent decades. Firstly, the types of datasets that statisticians are asked to analyse have transformed dramatically. In the past, we typically dealt with datasets containing many observations and a modest number of carefully chosen variables. Today, by contrast, it is common to encounter datasets with thousands of variables—sometimes even far exceeding the number of observations. For instance, in genomics, we might measure the expression levels of several thousand genes but only across a few hundred tissue samples. Classical statistical methods are often simply not applicable in these "high-dimensional" settings. As the scale of data collection has expanded, so too has the scope of the questions we seek to answer. Whereas statistics was once primarily concerned with uncovering associations between variables, we are now increasingly interested in understanding the causal structure of data. And rather than focusing solely on prediction, we often aim to predict the effects of interventions. At the same time, the rapid rise of machine learning has provided us with powerful new tools. In this course, we will explore how these advances can be harnessed to tackle some of the modern statistical challenges outlined above. The selection of material is heavily biased towards my own interests, but I hope it will nevertheless give you a flavour of some of the most important recent methodological developments in statistics.

The course is divided into 4 chapters (of unequal size). Our **first chapter** will start by introducing ridge regression, a simple generalisation of ordinary least squares. Our study of this will lead us to some beautiful connections with functional analysis and ultimately one of the most successful and flexible classes of learning algorithms: kernel machines. The **second chapter** concerns the Lasso and its extensions. The Lasso has been at the centre of much of the developments that have occurred in high-dimensional statistics, and will allow us to perform regression in the seemingly hopeless situation when the number of parameters we are trying to estimate is larger than the number of observations. In the **third chapter** we will study graphical modelling and provide an introduction to the exciting field of causal inference. Where the previous chapters consider methods for relating a particular response variable to a potentially large collection of (explanatory) variables, in the third chapter, we will study how to infer relationships between the variables themselves and answer causal questions using so-called *double/debiased machine learning* approaches. In the **final chapter**, we will turn to the problem of multiple testing which concerns handling settings where we may be performing thousands of hypothesis tests at the same time.

Before we begin the main content of the course, we will briefly review two key classical statistical methods: ordinary least squares and maximum likelihood estimation. This will help to set the scene and provide a warm-up for the modern methods to come later.

Classical statistics

Ordinary least squares

Imagine data are available in the form of observations $(Y_i, x_i) \in \mathbb{R} \times \mathbb{R}^p$, $i = 1, \ldots, n$, and the aim is to infer a simple *regression function* relating the average value of a *response*, Y_i, and a collection of *predictors* or *variables*, x_i. This is an example of regression analysis, one of the most important tasks in statistics.

A *linear model* for the data assumes that it is generated according to

$$Y = X\beta^0 + \varepsilon, \tag{1}$$

where $Y \in \mathbb{R}^n$ is the vector of responses; $X \in \mathbb{R}^{n \times p}$ is the predictor matrix (or design matrix) with ith row $x_i^\top$; $\varepsilon \in \mathbb{R}^n$ represents random error; and $\beta^0 \in \mathbb{R}^p$ is the unknown vector of coefficients.

Provided $p \ll n$, a sensible way to estimate β is by ordinary least squares (OLS). This yields an estimator $\hat{\beta}^{\mathrm{OLS}}$ with

$$\hat{\beta}^{\mathrm{OLS}} := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 = (X^\top X)^{-1} X^\top Y, \tag{2}$$

provided X has full column rank.

Under the assumptions that (i) $\mathbb{E}(\varepsilon_i) = 0$ and (ii) $\mathrm{Var}(\varepsilon) = \sigma^2 I$, we have that:

- $\mathbb{E}_{\beta^0, \sigma^2}(\hat{\beta}^{\mathrm{OLS}}) = \mathbb{E}\{(X^\top X)^{-1} X^\top (X\beta^0 + \varepsilon)\} = \beta^0$.
- $\mathrm{Var}_{\beta^0, \sigma^2}(\hat{\beta}^{\mathrm{OLS}}) = (X^\top X)^{-1} X^\top \mathrm{Var}(\varepsilon) X (X^\top X)^{-1} = \sigma^2 (X^\top X)^{-1}$.

The Gauss–Markov theorem states that OLS is the best linear unbiased estimator in our setting: for any other estimator $\tilde{\beta}$ that is linear in Y (so $\tilde{\beta} = AY$ for some fixed matrix A), we have

$$\mathrm{Var}_{\beta^0, \sigma^2}(\tilde{\beta}) - \mathrm{Var}_{\beta^0, \sigma^2}(\hat{\beta}^{\mathrm{OLS}})$$

is positive semi-definite.

Maximum likelihood estimation

The method of least squares is just one way to construct as estimator. A more general technique is that of maximum likelihood estimation. Here given data $y \in \mathbb{R}^n$ that we take as a realisation of a random variable Y, we specify its density $f(y; \theta)$ up to some unknown vector of parameters $\theta \in \Theta \subseteq \mathbb{R}^d$, where Θ is the parameter space. The likelihood function is a function of θ for each fixed y given by

$$L(\theta) := L(\theta; y) = c(y) f(y; \theta),$$

where $c(y)$ is an arbitrary constant of proportionality. The maximum likelihood estimate of θ maximises the likelihood, or equivalently it maximises the log-likelihood

$$\ell(\theta) := \ell(\theta; y) = \log f(y; \theta) + \log(c(y)).$$

A key quantity in the context of maximum likelihood estimation is the *Fisher information* matrix $i(\theta) := \mathrm{Cov}_\theta(\nabla \ell(\theta))$. It can be thought of as a measure of how hard it is to estimate θ when it is the true parameter value. The Cramér–Rao lower bound states that if $\tilde{\theta}$ is an unbiased estimator of θ, then under regularity conditions,

$$\mathrm{Var}_\theta(\tilde{\theta}) - i^{-1}(\theta)$$

is positive semi-definite.

A remarkable fact about maximum likelihood estimators (MLEs) is that (under quite general conditions) they are asymptotically normally distributed, asymptotically unbiased and asymptotically achieve the Cramér–Rao lower bound.

Assume that the Fisher information matrix when there are n observations, $i^{(n)}(\theta)$ (where we have made the dependence on n explicit) satisfies $i^{(n)}(\theta)/n \to I(\theta)$ for some positive definite matrix I. Then denoting the maximum likelihood estimator of θ when there are n observations by $\hat{\theta}^{(n)}$, under regularity conditions, as the number of observations $n \to \infty$ we have

$$\sqrt{n}(\hat{\theta}^{(n)} - \theta) \xrightarrow{d} N_d(0, I^{-1}(\theta)).$$

Returning to our linear model, if we assume in addition that $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, then the log-likelihood for (β, σ^2) is

$$\ell(\beta, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2.$$

We see that the maximum likelihood estimate of β and OLS coincide. It is easy to check that

$$i(\beta, \sigma^2) = \begin{pmatrix} \sigma^{-2}X^\top X & 0 \\ 0 & n\sigma^{-4}/2 \end{pmatrix}.$$

The general theory for MLEs would suggest that approximately $\sqrt{n}(\hat{\beta} - \beta) \sim N_p(0, \sigma^2(n^{-1}X^\top X)^{-1})$; in fact it is straightforward to show that this distributional result is exact.

Chapter 1

Kernel machines

Let us revisit the linear model with

$$Y_i = \mu^0 + x_i^\top \beta^0 + \varepsilon_i.$$

Note that here we have included an explicit intercept term, for reasons that will become clear later. However, our interest will continue to centre on β^0, which quantifies the contributions of each of the predictors to the regression function. For unbiased estimators of β^0, their variance gives a way of comparing their quality in terms of squared error loss. For a potentially biased estimator, $\tilde{\beta}$, the relevant quantity is the mean-squared error (MSE),

$$\begin{aligned}\mathbb{E}_{\beta^0,\sigma^2}\{(\tilde{\beta}-\beta^0)(\tilde{\beta}-\beta^0)^\top\} &= \mathbb{E}[\{\tilde{\beta}-\mathbb{E}(\tilde{\beta})+\mathbb{E}(\tilde{\beta})-\beta^0\}\{\tilde{\beta}-\mathbb{E}(\tilde{\beta})+\mathbb{E}(\tilde{\beta})-\beta^0\}^\top] \\ &= \mathrm{Var}(\tilde{\beta}) + \{\mathbb{E}(\tilde{\beta}-\beta^0)\}\{\mathbb{E}(\tilde{\beta}-\beta^0)\}^\top,\end{aligned}$$

a sum of squared bias and variance terms. A crucial part of the optimality arguments for OLS and MLEs was *unbiasedness*. Do there exist biased methods whose variance is is reduced compared to OLS such that their overall prediction error is lower? Yes—in fact the use of biased estimators is essential in dealing with settings where the number of parameters to be estimated is large compared to the number of observations. In the first two chapters we will explore two important methods for variance reduction based on different forms of penalisation: rather than forming estimators via optimising a least squares or log-likelihood term, we will introduce an additional penalty term that encourages estimates to be shrunk towards 0 in some sense. This will allow us to produce reliable estimators that work well when classical MLEs are infeasible, and in other situations can greatly outperform the classical approaches.

1.1 Ridge regression

One way to reduce the variance of $\hat{\beta}^{\mathrm{OLS}}$ is to shrink the estimated coefficients towards 0. *Ridge regression* [Hoerl and Kennard, 1970] does this by solving the following optimisation

problem

$$(\hat{\mu}_\lambda^{\mathrm{R}}, \hat{\beta}_\lambda^{\mathrm{R}}) = \underset{(\mu,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\arg\min} \{\|Y - \mu\mathbf{1} - X\beta\|_2^2 + \lambda\|\beta\|_2^2\}.$$

Here $\mathbf{1}$ is an n-vector of 1's. We see that the usual OLS objective is penalised by an additional term proportional to $\|\beta\|_2^2$. The parameter $\lambda \geq 0$, which controls the severity of the penalty and therefore the degree of the shrinkage towards 0, is known as a *regularisation parameter* or *tuning parameter*. We have explicitly included an intercept term which is not penalised. The reason for this is that were the variables to have their origins shifted so e.g. a variable representing temperature is given in units of Kelvin rather than Celsius, the fitted values would not change. However, $X\hat{\beta}$ is not invariant under scale transformations of the variables so it is standard practice to centre each column of X (hence making them orthogonal to the intercept term) and then scale them to have ℓ_2-norm $\sqrt{n}$.

It is straightforward to show that after this standardisation of X, $\hat{\mu}_\lambda^{\mathrm{R}} = \bar{Y} := \sum_{i=1}^n Y_i/n$, and

$$\hat{\beta}_\lambda^{\mathrm{R}} = (X^\top X + \lambda I)^{-1} X^\top Y.$$

In this form, we can see how the addition of the λI term helps to stabilise the estimator. Note that when X does not have full column rank (such as in high-dimensional situations), we can still compute this estimator. On the other hand, when X does have full column rank, we have the following theorem.

Theorem 1. *For λ sufficiently small (depending on β^0 and σ^2),*

$$\mathbb{E}(\hat{\beta}^{\mathrm{OLS}} - \beta^0)(\hat{\beta}^{\mathrm{OLS}} - \beta^0)^\top - \mathbb{E}(\hat{\beta}_\lambda^{\mathrm{R}} - \beta^0)(\hat{\beta}_\lambda^{\mathrm{R}} - \beta^0)^\top$$

is positive definite.

Proof. First we compute the bias of $\hat{\beta}_\lambda^{\mathrm{R}}$. We drop the subscript λ and superscript R for convenience.

$$\begin{aligned}
\mathbb{E}(\hat{\beta}) - \beta^0 &= (X^\top X + \lambda I)^{-1} X^\top X\beta^0 - \beta^0 \\
&= (X^\top X + \lambda I)^{-1}(X^\top X + \lambda I - \lambda I)\beta^0 - \beta^0 \\
&= -\lambda(X^\top X + \lambda I)^{-1}\beta^0.
\end{aligned}$$

Now we look at the variance of $\hat{\beta}$.

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}) &= \mathbb{E}\{(X^\top X + \lambda I)^{-1} X^\top \varepsilon\}\{(X^\top X + \lambda I)^{-1} X^\top \varepsilon\}^\top \\
&= \sigma^2 (X^\top X + \lambda I)^{-1} X^\top X (X^\top X + \lambda I)^{-1}.
\end{aligned}$$

Thus $\mathbb{E}(\hat{\beta}^{\mathrm{OLS}} - \beta^0)(\hat{\beta}^{\mathrm{OLS}} - \beta^0)^\top - \mathbb{E}(\hat{\beta} - \beta^0)(\hat{\beta} - \beta^0)^\top$ is equal to

$$\sigma^2(X^\top X)^{-1} - \sigma^2(X^\top X + \lambda I)^{-1} X^\top X (X^\top X + \lambda I)^{-1} - \lambda^2 (X^\top X + \lambda I)^{-1} \beta^0 {\beta^0}^\top (X^\top X + \lambda I)^{-1}.$$

After some simplification, we see that this is equal to

$$\lambda(X^\top X + \lambda I)^{-1}[\sigma^2\{2I + \lambda(X^\top X)^{-1}\} - \lambda\beta^0{\beta^0}^\top](X^\top X + \lambda I)^{-1}.$$

Thus $\mathbb{E}(\hat{\beta}^{\mathrm{OLS}} - \beta^0)(\hat{\beta}^{\mathrm{OLS}} - \beta^0)^\top - \mathbb{E}(\hat{\beta} - \beta^0)(\hat{\beta} - \beta^0)^\top$ is positive definite for $\lambda > 0$ if and only if

$$\sigma^2\{2I + \lambda(X^\top X)^{-1}\} - \lambda\beta^0{\beta^0}^\top$$

is positive definite, which is true for $\lambda > 0$ sufficiently small (we can take $0 < \lambda < 2\sigma^2/\|\beta^0\|_2^2$). □

The theorem says that $\hat{\beta}^{\mathrm{R}}_\lambda$ outperforms $\hat{\beta}^{\mathrm{OLS}}$ provided λ is chosen appropriately. To be able to use ridge regression effectively, we need a way of selecting a good λ—we will come to this very shortly. What the theorem doesn't really tell us is in what situations we expect ridge regression to perform well. To understand that, we will turn to one of the key matrix decompositions used in statistics, the singular value decomposition (SVD).

1.1.1 Connection to principal components analysis

The singular value decomposition (SVD) is a generalisation of an eigendecomposition of a square matrix. We can factorise any $X \in \mathbb{R}^{n\times p}$ into its SVD

$$X = UDV^\top.$$

Here the $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{p\times p}$ are orthogonal matrices and $D \in \mathbb{R}^{n\times p}$ has $D_{11} \geq D_{22} \geq \cdots \geq D_{mm} \geq 0$ where $m := \min(n, p)$ and all other entries of D are zero. To compute such a decomposition requires $O(np\min(n, p))$ operations. The rth columns of U and V are known as the rth left and right singular vectors of X respectively, and D_{rr} is the rth singular value.

When $n > p$, we can replace U by its first p columns and D by its first p rows to produce another version of the SVD (sometimes known as the thin SVD). Then $X = UDV^\top$ where $U \in \mathbb{R}^{n\times p}$ has orthonormal columns (but is no longer square) and D is square and diagonal. There is an analogous version for when $p > n$.

Let us take $X \in \mathbb{R}^{n\times p}$ as our matrix of predictors and suppose $n \geq p$. Using the (thin) SVD we may write the fitted values[1] from ridge regression as follows.

$$\begin{aligned} X\hat{\beta}^{\mathrm{R}}_\lambda &= X(X^\top X + \lambda I)^{-1}X^\top Y \\ &= UDV^\top(VD^2V^\top + \lambda I)^{-1}VDU^\top Y \\ &= UD(D^2 + \lambda I)^{-1}DU^\top Y \\ &= \sum_{j=1}^{p} U_j \frac{D_{jj}^2}{D_{jj}^2 + \lambda} U_j^\top Y. \end{aligned}$$

Here we have used the notation (that we shall use throughout the course) that U_j is the jth column of U. For comparison, the fitted values from OLS (when X has full column rank) are

$$X\hat{\beta}^{\mathrm{OLS}} = X(X^\top X)^{-1}X^\top Y = UU^\top Y.$$

[1] There is a slight abuse of terminology here as we are ignoring the contribution of $\hat{\mu}^{\mathrm{R}}_\lambda$.

Both OLS and ridge regression compute the coordinates $(U_j^\top Y)_{j=1}^p$ of Y with respect to the basis of the column space of X given by the columns of U. Ridge regression then shrinks these coordinates by the factors $D_{jj}^2/(D_{jj}^2+\lambda)$; if D_{jj} is small, the amount of shrinkage will be larger.

To interpret this further, note that the SVD is intimately connected with Principal Components Analysis (PCA). Consider $v \in \mathbb{R}^p$ with $\|v\|_2 = 1$. Since the columns of X have had their means subtracted, the sample variance of $Xv \in \mathbb{R}^n$, is

$$\frac{1}{n} v^\top X^\top X v = \frac{1}{n} v^\top V D^2 V^\top v.$$

Writing $a = V^\top v$, so $\|a\|_2 = 1$, we have

$$\frac{1}{n} v^\top V D^2 V^\top v = \frac{1}{n} a^\top D^2 a = \frac{1}{n} \sum_j a_j^2 D_{jj}^2 \leq \frac{1}{n} D_{11} \sum_j a_j^2 = \frac{1}{n} D_{11}^2.$$

As $\|XV_1\|_2^2/n = D_{11}^2/n$, V_1 determines the linear combination of the columns of X which has the largest sample variance, when the coefficients of the linear combination are constrained to have ℓ_2-norm 1. $XV_1 = D_{11}U_1$ is known as the first principal component of X. Subsequent principal components $D_{22}U_2, \ldots, D_{pp}U_p$ have maximum variance D_{jj}^2/n, subject to being orthogonal to all earlier ones—see example sheet 1 for details.

Returning to ridge regression, we see that it shrinks Y most in the smaller principal components of X. Thus it will work well when most of the signal is spanned by the large principal components of X. We now turn to the problem of choosing λ.

1.2 The kernel trick

An alternative expression for the ridge regression solution is given by the following

$$\begin{aligned} X^\top (XX^\top + \lambda I_n) &= (X^\top X + \lambda I_p) X^\top \\ (X^\top X + \lambda I_p)^{-1} X^\top &= X^\top (XX^\top + \lambda I_n)^{-1} \\ (X^\top X + \lambda I_p)^{-1} X^\top Y &= X^\top (XX^\top + \lambda I_n)^{-1} Y = \hat{\beta}_\lambda^{\mathrm{R}}. \end{aligned} \tag{1.1}$$

Two remarks are in order:

- Note while $X^\top X$ is $p \times p$, $XX^\top$ is $n \times n$. Computing fitted values via the the LHS of (1.1) would require roughly $O(np^2 + p^3)$ operations. If $p \gg n$ this could be extremely costly. However, the alternative formulation

  $$X\hat{\beta}_\lambda^{\mathrm{R}} = XX^\top (XX^\top + \lambda I)^{-1} Y$$

  would only require roughly $O(n^2 p + n^3)$ operations, which could be substantially smaller.

- We see that the fitted values of ridge regression depend only on inner products $K = XX^\top$ between observations (note $K_{ij} = x_i^\top x_j$).

Now suppose that we believe the signal depends quadratically on the predictors:

$$Y_i = x_i^\top \beta + \sum_{k,l} x_{ik} x_{il} \theta_{kl} + \varepsilon_i.$$

We can still use ridge regression provided we work with an enlarged set of predictors

$$x_{i1}, \ldots, x_{ip}, x_{i1}x_{i1}, \ldots, x_{i1}x_{ip}, x_{i2}x_{i1}, \ldots, x_{i2}x_{ip}, \ldots, x_{ip}x_{ip}. \tag{1.2}$$

This will give us $O(p^2)$ predictors. Our new approach to computing fitted values would therefore have complexity $O(n^2p^2 + n^3)$, which could be rather costly if p is large.

However, rather than first creating all the additional predictors and then computing the new K matrix, we can attempt to directly compute K. To this end consider

$$\begin{aligned}(1/2 + x_i^\top x_j)^2 - 1/4 &= \left(\frac{1}{2} + \sum_k x_{ik}x_{jk}\right)^2 - \frac{1}{4} \\ &= \sum_k x_{ik}x_{jk} + \sum_{k,l} x_{ik}x_{il}x_{jk}x_{jl}.\end{aligned}$$

Observe this amounts to an inner product between vectors of the form (1.2) Thus if we set

$$K_{ij} = (1/2 + x_i^\top x_j)^2 - 1/4 \tag{1.3}$$

and plug this into the formula for the fitted values, it is *exactly* as if we had performed ridge regression with an enlarged predictor matrix

$$\Phi := \begin{pmatrix} \phi(x_1)^\top \\ \vdots \\ \phi(x_n)^\top \end{pmatrix}.$$

Now computing K using (1.3) would require only $O(p)$ operations per entry, so $O(n^2p)$ operations in total, compared to $O(n^2p^2)$ for our earlier approach.

Predictions at a new $x \in \mathbb{R}^p$ may be computed similarly. From (1.1), we have

$$\phi(x)^\top \hat{\beta}_\lambda^{\mathrm{R}} = \phi(x)\Phi^\top(\Phi\Phi^\top + \lambda I)^{-1}Y = \sum_{i=1}^n k(x, x_i)\hat{\alpha}_i$$

where $\hat{\alpha} := (K + \lambda I)^{-1}Y \in \mathbb{R}^n$.

This is a nice computational trick, but more importantly for us it serves to illustrate some general points.

- Since ridge regression only depends on inner products between observations, rather than fitting non-linear models by first mapping the original data $x_i \in \mathbb{R}^p$ to $\phi(x_i) \in \mathbb{R}^d$ (say) using some *feature map* ϕ (which could, for example introduce quadratic effects), we can instead try to directly compute $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$.
- In fact rather than thinking in terms of feature maps, we can instead try to think about an appropriate measure of similarity $k(x_i, x_j)$ between observations. Modelling in this fashion is sometimes much easier.

We will now formalise and extend what we have learnt with this example.

1.3 Kernels

We have seen how a model with quadratic effects can be fitted very efficiently by replacing the inner product matrix (known as the *Gram matrix*) $XX^\top$ in (1.1) with the matrix in (1.3). It is then natural to ask what other non-linear models can be fitted efficiently using this sort of approach.

We won't answer this question directly, but instead we will try to understand the sorts of similarity measures k that can be represented as inner products between transformations of the original data.

That is, we will study the similarity measures $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ from the input space $\mathcal{X}$ to $\mathbb{R}$ for which there exists a *feature map* $\phi : \mathcal{X} \to \mathcal{H}$ where $\mathcal{H}$ is some (real) inner product space with

$$k(x, x') = \langle \phi(x), \phi(x') \rangle. \tag{1.4}$$

Recall that an inner product space is a real vector space $\mathcal{H}$ endowed with a map $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ that obeys the following properties.

(i) Symmetry: $\langle u, v \rangle = \langle v, u \rangle$.

(ii) Linearity: for $a, b \in \mathbb{R}$ $\langle au + bw, v \rangle = a\langle u, v \rangle + b\langle w, v \rangle$.

(iii) Positive-definiteness: $\langle u, u \rangle \geq 0$ with equality if and only if $u = 0$.

Definition 1. A *positive definite kernel* or more simply a *kernel* (for brevity) k is a symmetric map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which for all $n \in \mathbb{N}$ and all $x_1, \ldots, x_n \in \mathcal{X}$, the matrix K with entries

$$K_{ij} = k(x_i, x_j)$$

is positive semi-definite.

A kernel is a little like an inner product, but need not be bilinear in general. However, a form of the Cauchy–Schwarz inequality does hold for kernels.

Proposition 2.

$$k(x, x')^2 \leq k(x, x)k(x', x').$$

Proof. The matrix

$$\begin{pmatrix} k(x,x) & k(x,x') \\ k(x',x) & k(x',x') \end{pmatrix}$$

must be positive semi-definite so in particular its determinant must be non-negative. □

First we show that any inner product of feature maps will give rise to a kernel.

Proposition 3. *k defined by $k(x,x') = \langle \phi(x), \phi(x') \rangle$ is a kernel.*

Proof. Let $x_1, \ldots, x_n \in \mathcal{X}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and consider

$$\begin{aligned} \sum_{i,j} \alpha_i k(x_i, x_j) \alpha_j &= \sum_{i,j} \alpha_i \langle \phi(x_i), \phi(x_j) \rangle \alpha_j \\ &= \left\langle \sum_i \alpha_i \phi(x_i), \sum_j \alpha_j \phi(x_j) \right\rangle \geq 0. \end{aligned}$$

□

Showing that every kernel admits a representation of the form (1.4) is slightly more involved, and we delay this until after we have studied some examples.

1.3.1 Examples of kernels

Proposition 4. *Suppose $k_1, k_2, \ldots : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are kernels.*

(i) If $a_1, a_2 \geq 0$ then $a_1 k_1 + a_2 k_2$ is a kernel.

(ii) If $\lim_{m\to\infty} k_m(x,x') =: k(x,x')$ exists for all $x, x' \in \mathcal{X}$, then k is a kernel.

(iii) The pointwise product k given by $k(x,x') := k_1(x,x')k_2(x,x')$ is a kernel.

Proof. Let $K, K_1, K_2, \ldots$ be the corresponding kernel matrices and take $\alpha \in \mathbb{R}^n$.

(i) $\alpha^\top K \alpha = a_1 \alpha^\top K_1 \alpha + a_2 \alpha^\top K_2 \alpha \geq 0$.

(ii) $\alpha^\top K \alpha = \alpha^\top \lim_{m\to\infty} K_m \alpha = \lim_{m\to\infty} \alpha^\top K_m \alpha \geq 0$.

(iii) Let X and Y be independent random vectors with $\mathrm{Var}(X) = K_1$, $\mathrm{Var}(Y) = K_2$. The the entrywise (Haddamard) procuct $K = K_1 \odot K_2$ satisfies

$$K_{ij} = \mathbb{E}(X_i X_j)\mathbb{E}(Y_i Y_j) = \mathbb{E}(X_i Y_i \, X_j Y_j) = (\mathrm{Var}(X \odot Y))_{ij},$$

and $\mathrm{Var}(X \odot Y)$ is positive semi-definite as a covariance matrix.

□

Linear kernel. $k(x,x') = x^\top x'$.

Polynomial kernel. $k(x, x') = (1 + x^\top x')^d$. To show this is a kernel, we can simply note that $1 + x^\top x'$ gives a kernel owing to the fact that 1 is a kernel and (i) of Proposition 4. Next (ii) and induction shows that k as defined above is a kernel.

Gaussian kernel. The highly popular Gaussian kernel is defined by

$$k(x, x') = \exp\bigg(- \frac{\|x - x'\|_2^2}{2h^2} \bigg).$$

For x close to x' it is large whilst for x far from x' the kernel quickly decays towards 0. The additional parameter $h > 0$ known as the *bandwidth* controls the speed of the decay to zero. Note it is less clear how one might find a corresponding feature map and indeed any feature map that represents this must be infinite dimensional.

To show that it is a kernel first decompose $\|x - x'\|_2^2 = \|x\|_2^2 + \|x'\|_2^2 - 2x^\top x'$. Note that by Proposition 3,

$$k_1(x, x') = \exp\bigg(- \frac{\|x\|_2^2}{2h^2} \bigg) \exp\bigg(- \frac{\|x'\|_2^2}{2h^2} \bigg)$$

is a kernel. Next writing

$$k_2(x, x') = \exp(x^\top x'/h^2) = \sum_{r=0}^{\infty} \frac{(x^\top x'/h^2)^r}{r!}$$

and using (i) of Proposition 4 shows that k_2 is a kernel. Finally observing that $k = k_1 k_2$ and using (ii) shows that the Gaussian kernel is indeed a kernel.

First order Sobolev kernel. Take $\mathcal{X}$ to be $[0, 1]$ and let $k(x, x') := x \wedge x' = \min(x, x')$. We have

$$k(x.x') = \int_0^1 \mathbb{1}_{[0,x]}(u) \mathbb{1}_{[0,x']}(u) du = \langle \mathbb{1}_{[0,x]}, \mathbb{1}_{[0,x']} \rangle$$

so k is a kernel by Proposition 3.

Second order Sobolev kernel. Take $\mathcal{X}$ to be $[0, 1]$ and let

$$k(x, x') := \int_0^{x \wedge x'} \int_0^{x \wedge y} (x - u)(y - u) du$$

Jaccard similarity kernel. Take $\mathcal{X}$ to be the set of all subsets of $\{1, \dots, p\}$. For $x, x' \in \mathcal{X}$ with $x \cup x' \neq \emptyset$ define

$$k(x, x') = \frac{|x \cap x'|}{|x \cup x'|}$$

and if $x \cup x' = \emptyset$ then set $k(x, x') = 1$. Showing that this is a kernel is left to the example sheet.

1.3.2 Reproducing kernel Hilbert spaces

Recall that we wish to show that each kernel k admits a representation for the form $k(x, x') = \langle \phi(x), \phi(x') \rangle$ for some feature map $\phi : \mathcal{X} \to \mathcal{H}$ where $\mathcal{H}$ is an inner product space. Before showing this, let us first consider the case where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for finite $\mathcal{X}$, so without loss of generality, $\mathcal{X} = \{1, \dots, N\}$ for some $N \in \mathbb{N}$. Then $K \in \mathbb{R}^{N \times N}$ with $K_{ij} = k(i, j)$ contains all the information about k. The eigendecomposition $K = P^\top DP$ then readily confirms what we want to show: taking $\phi(i) = D^{1/2} P_i$, we have

$$k(i, j) = K_{ij} = P_i^\top DP_j = (D^{1/2} P_i)^\top (D^{1/2} P_j) = \phi(i)^\top \phi(j).$$

Note however that representation of k through an inner product is not unique. Consider $\phi(i) = K_i$ and the weighted Euclidean inner product on the column space $\mathcal{H}$, of K given by $\langle u, v \rangle = u^\top P^\top D^+ Pu$, where D^+ has $(D^+)_{ij} = D_{ij}^{-1}$ when $D_{ij} > 0$ and 0 otherwise. Then for $\alpha \in \mathbb{R}^N$,

$$\langle \phi(i), K\alpha \rangle = (P^\top DPe_i)^\top P^\top D^+ PP^\top DP\alpha = e_i^\top K\alpha = (K\alpha)_i,$$

so

$$\langle \phi(i), \phi(j) \rangle = (K_j)_i = K_{ij}$$

as required. Note also the interesting property that the inner product of $\phi(i)$ and $K\alpha$ extracts the ith component of $K\alpha$. It is this second representation that generalises most fruitfully to the case where $\mathcal{X}$ may be infinite.

Consider now taking $\mathcal{H}_0$ to be the linear span of the functions $\{k(\cdot, x) : x \in \mathcal{X}\}$, i.e. the vector space of functions of the form

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i), \tag{1.5}$$

where $n \in \mathbb{N}$, $x_i \in \mathcal{X}$ and $\alpha_i \in \mathbb{R}$. Given another function

$$g(\cdot) = \sum_{j=1}^{m} \beta_j k(\cdot, x'_j) \tag{1.6}$$

we define their inner product to be

$$\langle f, g \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j). \tag{1.7}$$

We need to check this is well-defined as the representations of f and g in (1.5) and (1.6) need not be unique. To this end, note that

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j) = \sum_{i=1}^{n} \alpha_i g(x_i) = \sum_{j=1}^{m} \beta_j f(x'_j). \tag{1.8}$$

The first equality shows that the inner product does not depend on the particular expansion of g whilst the second equality shows that it also does not depend on the expansion of f. Thus the inner product is well-defined. We define our feature map $\phi : \mathcal{X} \to \mathcal{H}_0$ to be

$$\phi(x) = k(\cdot, x). \tag{1.9}$$

Theorem 5. *For every kernel k, the space $\mathcal{H}_0$ above is an inner product space and the feature map* (1.9) *satisfies*

$$k(x, x') = \langle \phi(x), \phi(x') \rangle. \tag{1.10}$$

Proof. First we check that with ϕ defined as in (1.9) we do have relationship (1.10). Observe that

$$\langle k(\cdot, x), f \rangle = \sum_{i=1}^{n} \alpha_i k(x_i, x) = f(x), \tag{1.11}$$

so in particular we have

$$\langle \phi(x), \phi(x') \rangle = \langle k(\cdot, x), k(\cdot, x') \rangle = k(x, x').$$

It remains to show that (1.7) is indeed an inner product. It is clearly symmetric and (1.8) shows linearity. We now need to show positive definiteness.

First note that

$$\langle f, f \rangle = \sum_{i,j} \alpha_i k(x_i, x_j) \alpha_j \geq 0 \tag{1.12}$$

by positive definiteness of the kernel. Now from (1.11),

$$f(x)^2 = (\langle k(\cdot, x), f \rangle)^2.$$

If we could use the Cauchy–Schwarz inequality on the right-hand side, we would have

$$f(x)^2 \leq \langle k(\cdot, x), k(\cdot, x) \rangle \langle f, f \rangle, \tag{1.13}$$

which would show that if $\langle f, f \rangle = 0$ then necessarily $f = 0$; the final property we need to show that $\langle \cdot, \cdot \rangle$ is an inner product. However, in order to use the traditional Cauchy–Schwarz inequality we need to first know we're dealing with an inner product, which is precisely what we're trying to show!

Although we haven't yet shown that $\langle \cdot, \cdot \rangle$ is an inner product, we do have enough information to show that it is itself a kernel. We may then appeal to Proposition 2 to obtain (1.13). With this in mind, we argue as follows. Given functions $f_1, \ldots, f_m$ and coefficients $\gamma_1, \ldots, \gamma_m \in \mathbb{R}$, we have

$$\sum_{i,j} \gamma_i \langle f_i, f_j \rangle \gamma_j = \left\langle \sum_i \gamma_i f_i, \sum_j \gamma_j f_j \right\rangle \geq 0$$

where we have used linearity and (1.12), showing that it is a kernel. □

To further discuss the space $\mathcal{H}_0$ we recall some facts from analysis. Any inner product space $\mathcal{B}$ is also a normed space: for $f \in \mathcal{B}$ we may define $\|f\|_{\mathcal{B}}^2 := \langle f, f \rangle_{\mathcal{B}}$. Recall that a Cauchy sequence $(f_m)_{m=1}^{\infty}$ in $\mathcal{B}$ has $\|f_m - f_n\|_{\mathcal{B}} \to 0$ as $n, m \to \infty$. A normed space where every Cauchy sequence has a limit (in the space) is called *complete*, and a complete inner product space is called a *Hilbert space*.

Hilbert spaces may be thought of as the (potentially) infinite-dimensional analogues of finite-dimensional Euclidean spaces. For later use we note that if V is a closed subspace of a Hilbert space $\mathcal{B}$, then any $f \in \mathcal{B}$ has a decomposition $f = u + v$ with $u \in V$ and

$$v \in V^{\perp} := \{v \in \mathcal{B} : \langle v, u \rangle_{\mathcal{B}} = 0 \text{ for all } u \in V\}.$$

Moreover, if V is finite-dimensional, then it is closed.

By adding the limits of Cauchy sequences to $\mathcal{H}_0$ (from Theorem 5) we can make create a Hilbert space. If $(f_m)_{m=1}^{\infty} \in \mathcal{H}$ is Cauchy, then since by (1.13) we have

$$|f_m(x) - f_n(x)| \leq \sqrt{k(x,x)} \|f_m - f_n\|_{\mathcal{H}},$$

we may define function $f^* : \mathcal{X} \to \mathbb{R}$ by $f^*(x) = \lim_{m \to \infty} f_m(x)$. We can check that all such f^* can be added to $\mathcal{H}_0$ to create a Hilbert space.

In fact, the completion of $\mathcal{H}_0$ is a special type of Hilbert space known as a *reproducing kernel Hilbert space* (RKHS).

Definition 2. A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ is a *reproducing kernel Hilbert space* (RKHS) if for all $x \in \mathcal{X}$, there exists $k_x \in \mathcal{B}$ such that it satisfies the reproducing property

$$f(x) = \langle k_x, f \rangle \qquad \text{for all } f \in \mathcal{B}.$$

The function

$$\begin{aligned} k : \mathcal{X} \times \mathcal{X} &\to \mathbb{R} \\ (x, x') &\mapsto \langle k_x, k_{x'} \rangle = k_{x'}(x) \end{aligned}$$

is known as the *reproducing kernel* of $\mathcal{H}$.

Note that the reproducing kernel is well-defined: if k_x and h_x satisfy the reproducing property above, then

$$\|k_x - h_x\|_{\mathcal{H}}^2 = \langle k_x, k_x - h_x \rangle - \langle h_x, k_x - h_x \rangle = (k_x - h_x)(x) - (k_x - h_x)(x) = 0.$$

To summarise what we have learnt, by Proposition 3, the reproducing kernel of any RKHS is a (positive definite) kernel, and Theorem 5 shows that to any kernel k is associated an RKHS that has reproducing kernel k. One can further show that this is the unique RKHS with k as its reproducing kernel.

Examples

Linear kernel. Here $\mathcal{H} = \{f : f(x) = \beta^{\top} x, \, \beta \in \mathbb{R}^p\}$ and if $f(x) = \beta^{\top} x$ then $\|f\|_{\mathcal{H}}^2 = \|\beta\|_2^2$.

First-order Sobolev kernel. Take $\mathcal{H}$ to be the class of almost everywhere differentiable functions $f : [0,1] \to \mathbb{R}$, with $f(0) = 0$ and $\int_0^1 (f'(u))^2 \, du < \infty$. This is an RKHS with reproducing kernel $k(x,y) = x \wedge y$ and inner product

$$\langle f, g \rangle := \int_0^1 f'(u) g'(u) \, du.$$

We can check

$$\langle f, k(\cdot, x) \rangle = \int_0^1 f'(u) \mathbb{1}_{[0,x]}(u) \, du = \int_0^x f'(u) \, du = f(x).$$

Second-order Sobolev kernel. Take $\mathcal{H}$ to be the set of differentiable functions $f : [0,1] \to \mathbb{R}$ with $f(0) = 0, f'(0) = 0$ and where f' is almost everywhere differentiable with $\int_0^1 (f''(u))^2 \, du < \infty$. For $f, g \in \mathcal{H}$ define

$$\langle f, g \rangle := \int_0^1 f''(u) g''(u) \, du.$$

Recall (see Ex. sheet) that $k_x(y) := k(x,y) := \int_0^{x \wedge y} (x-u)(y-u) \, du$ satisfies $k_x''(y) = (x-y)_+$ and for $f \in \mathcal{H}$,

$$\begin{aligned} \langle f, k_x \rangle &= \int_0^1 f''(u)(x-u)_+ \, du \\ &= [f'(u)(x-u)_+]_0^1 + \int_0^1 f'(u) \mathbb{1}_{[0,x]}(u) \, du \\ &= f(x). \end{aligned}$$

1.3.3 The representer theorem

To recap, what we have shown so far is that replacing the matrix $XX^\top$ in the definition of an algorithm by K derived form a positive definite kernel is essentially equivalent to running the same algorithm on some mapping of the original data, though with the modification that instances of $x_i^\top x_j$ become $\langle \phi(x_i), \phi(x_j) \rangle$. This corresponds to ridge regression on a predictor matrix with ith row $\phi(x_i)$ in the case where $\phi : \mathcal{X} \to \mathcal{H}$ maps into a Euclidean space, but the question remains as to how to interpret this when $\mathcal{H}$ is infinite-dimensional.

If $\mathcal{H}$ denotes the RKHS of the linear kernel, then

$$\hat{f} := \operatorname*{arg\,min}_{f \in \mathcal{H}} \left\{ \sum_{i=1}^n \{Y_i - f(x_i)\}^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\} \tag{1.14}$$

is the usual fitted regression function from ridge regression. The following theorem shows in particular that kernel ridge regression (i.e. ridge regression replacing $XX^\top$ with K) with kernel k is equivalent to the above with $\mathcal{H}$ now being the RKHS corresponding to k.

Theorem 6 (Representer theorem, [Kimeldorf and Wahba, 1970, Schölkopf et al., 2001])**.** *Let $c : \mathbb{R}^n \times \mathcal{X}^n \times \mathbb{R}^n \to \mathbb{R}$ be an arbitrary loss function, and let $J : [0, \infty) \to \mathbb{R}$ be strictly increasing. Let $x_1, \ldots, x_n \in \mathcal{X}$, $Y \in \mathbb{R}^n$. Finally, let $f \in \mathcal{H}$ where $\mathcal{H}$ is an RKHS with reproducing kernel k, and let $K_{ij} = k(x_i, x_j)$ $i, j = 1, \ldots, n$. Then $\hat{f}$ minimises*

$$Q_1(f) := c(Y, x_1, \ldots, x_n, f(x_1), \ldots, f(x_n)) + J(\|f\|_{\mathcal{H}}^2)$$

over $f \in \mathcal{H}$ iff. $\hat{f}(\cdot) = \sum_{i=1}^n \hat{\alpha}_i k(\cdot, x_i)$ and $\hat{\alpha} \in \mathbb{R}^n$ minimises Q_2 over $\alpha \in \mathbb{R}^n$ where

$$Q_2(\alpha) = c(Y, x_1, \ldots, x_n, K\alpha) + J(\alpha^\top K \alpha).$$

Proof. Let U be the linear span of $\{k(\cdot, x_i), i = 1, \ldots, n\}$. As U is finite-dimensional and hence closed, we can decompose any $f \in \mathcal{H}$ as $f = u + v$ with $u \in U$ and $v \in U^\perp$. Then

$$f(x_i) = \langle f, k(\cdot, x_i) \rangle = \langle u + v, k(\cdot, x_i) \rangle = \langle u, k(\cdot, x_i) \rangle = u(x_i).$$

Meanwhile,

$$J(\|f\|_{\mathcal{H}}^2) = J(\|u\|_{\mathcal{H}}^2 + \|v\|_{\mathcal{H}}^2) \geq J(\|u\|_{\mathcal{H}}^2),$$

with equality if and only $v = 0$. Thus in minimising Q_1, we may restrict attention to those $f \in U$, i.e., those

$$f = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$$

for some $\alpha \in \mathbb{R}^n$. But for such f,

$$\|f\|_{\mathcal{H}}^2 = \left\langle \sum_{i=1}^n \alpha_i k(\cdot, x_i), \sum_{i=1}^n \alpha_i k(\cdot, x_i) \right\rangle = \alpha^\top K \alpha,$$

and $f(x_i) = K_i^\top \alpha$ so $(f(x_i))_{i=1}^n = K\alpha$. Thus $Q_1(f) = Q_2(\alpha)$. □

Consider the result specialised the ridge regression objective. We see that (1.14) is essentially equivalent to minimising

$$\|Y - K\alpha\|_2^2 + \lambda \alpha^\top K \alpha,$$

and you may check (see example sheet 1) that the minimiser $\hat{\alpha}$ satisfies $K\hat{\alpha} = K(K + \lambda I)^{-1} Y$. Thus (1.14) is indeed an alternative way of expressing kernel ridge regression. The result also tells us how to form predictions: given a new observation x, our prediction for $f(x)$ is

$$\hat{f}(x) = \sum_{i=1}^n \hat{\alpha}_i k(x, x_i).$$

The application of the result is however not limited to ridge regression and shows that a whole host of algorithms can be 'kernelised'. For example, recall that in the classification setting where $Y_i \in \{-1, 1\}$, standard logistic regression may be motivated by assuming

$$\log \left(\frac{\mathbb{P}(Y_i = 1)}{\mathbb{P}(Y_i = -1)} \right) = \mu^0 + x_i^\top \beta^0$$

and picking $(\hat{\mu}, \hat{\beta})$ to maximise the log-likelihood. This leads to the following optimisation problem:

$$\underset{(\mu,\beta)\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \log\{1 + \exp(-Y_i(\mu + x_i^\top \beta))\}.$$

The kernelised version is then given by

$$\underset{\mu\in\mathbb{R}, f\in\mathcal{H}}{\arg\min} \left\{ \sum_{i=1}^{n} \log[1 + \exp\{-Y_i(\mu + f(x_i))\}] + \lambda \|f\|_{\mathcal{H}}^2 \right\},$$

where $\mathcal{H}$ is an RKHS (note that here we have included an unpenalised intercept term).

1.4 Kernel ridge regression

We have seen how the kernel trick allows us to solve a potentially infinite-dimensional version of ridge regression. This may seem impressive, but ultimately we should judge kernel ridge regression on its statistical properties e.g. predictive performance. Consider a setting where

$$Y_i = f^0(x_i) + \varepsilon_i, \qquad \mathbb{E}(\varepsilon) = 0,\ \mathrm{Var}(\varepsilon) = \sigma^2 I.$$

We shall assume that $f^0 \in \mathcal{H}$ where $\mathcal{H}$ is an RKHS with reproducing kernel k. Let K be the kernel matrix $K_{ij} = k(x_i, x_j)$ with eigenvalues $d_1 \geq d_2 \geq \cdots \geq d_n \geq 0$. We will see that the predictive performance depends delicately on these eigenvalues.

Let $\hat{f}_\lambda$ be the estimated regression function from kernel ridge regression with kernel k:

$$\hat{f}_\lambda = \underset{f\in\mathcal{H}}{\arg\min} \left\{ \sum_{i=1}^{n} \{Y_i - f(x_i)\}^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}.$$

Theorem 7. *The mean squared prediction error (MSPE) may be bounded above in the following way:*

$$\begin{aligned} \frac{1}{n}\mathbb{E}\left\{ \sum_{i=1}^{n} \{f^0(x_i) - \hat{f}_\lambda(x_i)\}^2 \right\} &\leq \frac{\sigma^2}{n} \sum_{i=1}^{n} \frac{d_i^2}{(d_i + \lambda)^2} + \frac{\lambda \|f^0\|_{\mathcal{H}}^2}{4n} \\ &\leq \frac{\sigma^2}{n} \frac{1}{\lambda} \sum_{i=1}^{n} \min(d_i/4, \lambda) + \frac{\lambda \|f^0\|_{\mathcal{H}}^2}{4n}. \end{aligned} \tag{1.15}$$

Proof. We know from the representer theorem that

$$\left(\hat{f}_\lambda(x_1), \ldots, \hat{f}_\lambda(x_n)\right)^\top = K(K + \lambda I)^{-1} Y.$$

You will show on the example sheet that

$$\left(f^0(x_1), \ldots, f^0(x_n)\right)^\top = K\alpha,$$

for some $\alpha \in \mathbb{R}^n$, and moreover that $\|f^0\|_{\mathcal{H}}^2 \geq \alpha^\top K\alpha$. Let the eigendecomposition of K be given by $K = UDU^\top$ with $D_{ii} = d_i$ and define $\theta = U^\top K\alpha$. We see that n times the LHS of (1.15) is

$$\begin{aligned}\mathbb{E}\|K(K+\lambda I)^{-1}(U\theta+\varepsilon) - U\theta\|_2^2 &= \mathbb{E}\|DU^\top(UDU^\top+\lambda I)^{-1}(U\theta+\varepsilon) - \theta\|_2^2 \\ &= \mathbb{E}\|D(D+\lambda I)^{-1}(\theta + U^\top\varepsilon) - \theta\|_2^2 \\ &= \|\{D(D+\lambda I)^{-1} - I\}\theta\|_2^2 + \mathbb{E}\|D(D+\lambda I)^{-1}U^\top\varepsilon\|_2^2.\end{aligned}$$

To compute the second term, we use the 'trace trick':

$$\begin{aligned}\mathbb{E}\|D(D+\lambda I)^{-1}U^\top\varepsilon\|_2^2 &= \mathbb{E}[\{D(D+\lambda I)^{-1}U^\top\varepsilon\}^\top D(D+\lambda I)^{-1}U^\top\varepsilon] \\ &= \mathbb{E}[\mathrm{tr}\{D(D+\lambda I)^{-1}U^\top\varepsilon\varepsilon^\top UD(D+\lambda I)^{-1}\}] \\ &= \sigma^2\mathrm{tr}\{D(D+\lambda I)^{-1}D(D+\lambda I)^{-1}\} \\ &= \sigma^2\sum_{i=1}^n \frac{d_i^2}{(d_i+\lambda)^2}.\end{aligned}$$

For the first term, we have

$$\|\{D(D+\lambda I)^{-1} - I\}\theta\|_2^2 = \sum_{i=1}^n \frac{\lambda^2\theta_i^2}{(d_i+\lambda)^2}.$$

Now as $\theta = DU^\top\alpha$, note that $\theta_i = 0$ when $d_i = 0$. Let D^+ be the diagonal matrix with ith diagonal entry equal to D_{ii}^{-1} if $D_{ii} > 0$ and 0 otherwise. Then

$$\sum_{i:d_i>0}\frac{\theta_i^2}{d_i} = \|\sqrt{D^+}\theta\|_2^2 = \alpha^\top KUD^+U^\top K\alpha = \alpha^\top UDD^+DU^\top\alpha = \alpha^\top K\alpha \leq \|f^0\|_{\mathcal{H}}^2.$$

By Hölder's inequality we have

$$\sum_{i=1}^n \frac{\lambda^2\theta_i^2}{(d_i+\lambda)^2} = \sum_{i:d_i>0}\frac{\theta_i^2}{d_i}\frac{d_i\lambda^2}{(d_i+\lambda)^2} \leq \max_{i=1,\ldots,n}\frac{d_i\lambda^2}{(d_i+\lambda)^2} \leq \|f^0\|_{\mathcal{H}}^2\lambda/4,$$

using the inequality $(a+b)^2 \geq 4ab$ in the final line. Finally note that

$$\frac{d_i^2}{(d_i+\lambda)^2} \leq \min\{1, d_i^2/(4d_i\lambda)\} = \min(\lambda, d_i/4)/\lambda. \qquad \square$$

To interpret this result further, it will be helpful to express it in terms of $\hat{\mu}_i := d_i/n$ (the eigenvalues of K/n) and $\gamma_n := \lambda/n$. We have

$$\frac{1}{n}\mathbb{E}\Big\{\sum_{i=1}^n\{f^0(x_i) - \hat{f}_{n\gamma_n}(x_i)\}^2\Big\} \leq \frac{\sigma^2}{\gamma_n}\frac{1}{n}\sum_{i=1}^n\min(\hat{\mu}_i/4, \gamma_n) + \|f^0\|_{\mathcal{H}}^2\gamma_n/4 =: \delta_n(\gamma). \quad (1.16)$$

Here we have treated the x_i as fixed, but we could equally well think of them as random. Consider a setup where the x_i are i.i.d. and independent of ε. If we take a further expectation on the RHS of (1.16), our result still holds true (the $\hat{\mu}_i$ are random in this setting). Ideally we would like to then replace $\mathbb{E}\min(\hat{\mu}_i/4, \gamma)$ with a quantity more directly related to the kernel k.

Mercer's theorem is helpful in this regard. This guarantees (under some mild conditions) an eigendecomposition for kernels, which recall are somewhat like infinite-dimensional analogues of symmetric positive semi-definite matrices.

Given a random variable X taking values in $\mathcal{X}$, we say a non-zero function $e \in \mathcal{H}$ is an *eigenfunction* with *eigenvalue* $\mu \in \mathbb{R}$ if

$$\mu e(x) = \mathbb{E}k(x, X)e(X).$$

Mercer's theorem states the following under mild conditions, including that $\mathbb{E}k(X, X) < \infty$:

- The set of positive eigenvalues is at most countable.
- The subspace spanned by the eigenfunctions corresponding to each positive eigenvalue has a finite dimension known as the *multiplicity* of the eigenvalue.
- Writing $(\mu_j)_{j\in J}$ (where $J = \{1, \ldots, m\}$, some m or $J \in \mathbb{N}$) for the eigenvalues counted with multiplicity, there exist corresponding eigenfunctions $(e_j)_{j\in J}$ that are orthonormal in the sense that

  $$\mathbb{E}e_j(X)e_k(X) = \mathbb{1}_{\{j=k\}}$$

  and satisfy

  $$k(x, y) = \sum_{j\in J} \mu_j e_j(x)e_j(y). \tag{1.17}$$

Lemma 8. *When* (1.17) *holds, we have for* $\gamma > 0$*,*

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} \min(\hat{\mu}_i/4, \gamma)\right) \leq \frac{1}{n}\sum_{j\in J} \min(\mu_j/4, \gamma).$$

Theorem 9. *Provided the eigendecomposition* (1.17) *holds, there exists* γ_n *such that for fixed* $\sigma^2 > 0$*,*

$$\frac{1}{n}\mathbb{E}\left\{\sum_{i=1}^{n}\{f^0(x_i) - \hat{f}_{\gamma_n}(x_i)\}^2\right\} = o(n^{-1/2}).$$

Proof. Let $\phi : [0, \infty) \to [0, \infty)$ be given by

$$\phi(\gamma) := \sum_{j\in J} \min(\mu_j, \gamma).$$

Observe that ϕ is increasing and as $\sum_{j\in J}\mu_j < \infty$, $\lim_{\gamma\downarrow 0}\phi(\gamma) = 0$ (this is clear when J is finite; otherwise note that given an arbitrary $\epsilon > 0$, there exists M such that $\sum_{j=M}^{\infty}\mu_j \le \epsilon$, but then $\phi(\gamma) \le M\gamma + \epsilon \to \epsilon$ as $\gamma \downarrow 0$). Let $\gamma_n = n^{-1/2}\sqrt{\phi(n^{-1/2})}$ so $\gamma_n = o(n^{-1/2})$. Thus for n sufficiently large $\phi(\gamma_n) \le \phi(n^{-1/2})$, whence for such n we have

$$\begin{aligned}\inf_{\gamma>0}\{\phi(\gamma)/(n\gamma) + \gamma\} &\le \frac{\phi(\gamma_n)}{n\gamma_n} + \gamma_n \\ &\le 2\sqrt{\phi(n^{-1/2})}/\sqrt{n} = o(n^{-1/2}).\end{aligned}$$

□

In specific cases, we can get faster rates.

First-order Sobolev kernel. When k is the Sobolev kernel, and considering a uniform distribution on $\mathcal{X} = [0,1]$, an eigenvalue–eigenfunction pair (μ, e) must satisfy

$$\mu e(x) = \int_0^1 \min(x,y)e(y)\,dy = \int_0^x ye(y)\,dy + x\int_x^1 e(y)\,dy,$$

so

$$\mu e'(x) = xe(x) + \int_x^1 e(y)\,dy - xe(x) = \int_x^1 e(y)\,dy, \tag{1.18}$$

hence

$$\mu e''(x) = -e(x).$$

This ODE has general solution $e(x) = a_\mu \sin(x/\sqrt{\mu}) + b_\mu \cos(x/\sqrt{\mu})$ and the boundary condition $e(0) = 0$ gives $b_\mu = 0$. Also from (1.18), we see that $\mu e'(1) = 0$, so $1/\sqrt{\mu} = \pi/2 + k\pi$ for some $k = 0, 1, 2, \ldots$. Thus the jth eigenvalue satisfies

$$\mu_j/4 = \frac{1}{\pi^2(2j-1)^2}.$$

We therefore have

$$\begin{aligned}\sum_{i=1}^{\infty}\min(\mu_i/4, \gamma_n) &\le \frac{\gamma_n}{2}\left(\frac{1}{\sqrt{\pi^2\gamma_n}} + 1\right) + \frac{1}{\pi^2}\int_{\{(\pi^2\gamma_n)^{-1/2}+1\}/2}^{\infty}\frac{1}{(2x-1)^2}dx \\ &= \sqrt{\gamma_n}/\pi + \gamma_n/2 = O(\sqrt{\gamma_n})\end{aligned}$$

as $\gamma_n \to 0$. Putting things together, we see that

$$\mathbb{E}\delta_n(\gamma_n) = O\left(\frac{\sigma^2}{n\gamma_n^{1/2}} + \gamma_n\right).$$

Thus an optimal $\gamma_n \sim (\sigma^2/n)^{2/3}$ gives an error rate of order $(\sigma^2/n)^{2/3}$.

1.5 Large-scale kernel machines

We introduced the kernel trick as a computational device that avoided performing calculations in a high or infinite dimensional feature space and, in the case of kernel ridge regression reduced computation down to forming the $n \times n$ matrix K and then inverting $K + \lambda I$. This can be a huge saving, but when n is very large, this can present serious computational difficulties. Even if p is small, the $O(n^3)$ cost of inverting $K + \lambda I$ may cause problems. What's worse, the fitted regression function is a sum over n terms:

$$\hat{f}(\cdot) = \sum_{i=1}^{n} \hat{\alpha}_i k(x_i, \cdot).$$

Even to evaluate a prediction at a single new observation requires $O(n)$ computations unless $\hat{\alpha}$ is sparse.

In recent years, there has been great interest in speeding up computations for kernel machines. We will discuss one exciting approach based on random feature expansions. Given a kernel k, the key idea is to develop a random map

$$\hat{\phi} : \mathcal{X} \to \mathbb{R}^b$$

with b small such that $\mathbb{E}\{\hat{\phi}(x)^\top \hat{\phi}(x')\} = k(x, x')$. In a sense we are trying to reverse the kernel trick by approximating the kernel using a random feature map. To increase the quality of the approximation of the kernel, we can consider

$$x \mapsto \frac{1}{\sqrt{L}}(\hat{\phi}_1(x), \ldots, \hat{\phi}_L(x)) \in \mathbb{R}^{Lb}$$

with each $(\hat{\phi}_l(x))_{l=1}^L$ being i.i.d. for each x. Let Φ be the matrix with ith row given by $(\hat{\phi}_1(x_i), \ldots, \hat{\phi}_L(x_i))/\sqrt{L}$. We may then run our learning algorithm replacing the initial matrix of predictors X with Φ. For example, when performing ridge regression, we can compute

$$(\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top Y,$$

which would require $O(nL^2b^2 + L^3b^3)$ operations: a cost linear in n. Predicting a new observation would cost $O(Lb)$.

The work of Rahimi and Recht [2007] proposes a construction of such a random mapping $\hat{\phi}$ for shift-invariant kernels, that is kernels for which there exists a function g with $k(x, x') = g(x - x')$ for all $x, x' \in \mathcal{X} = \mathbb{R}^p$. A useful property of such kernels is given by Bochner's theorem.

Theorem 10 (Bochner's theorem)**.** *Let $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ be a continuous kernel. Then k is shift-invariant if and only if there exists some $c > 0$ and distribution F on $\mathbb{R}^p$ such that when $W \sim F$*

$$k(x, x') = c\mathbb{E}e^{i(x-x')^\top W} = c\mathbb{E}\cos((x - x')^\top W).$$

To make use of this theorem, first observe the following. Let $u \sim U[-\pi, \pi]$, $x, y \in \mathbb{R}$. Then

$$2\mathbb{E}\cos(x+u)\cos(y+u) = 2\mathbb{E}\{(\cos x \cos u - \sin x \sin u)(\cos y \cos u - \sin y \sin u)\}.$$

Now as $u \stackrel{d}{=} -u$, $\mathbb{E}\cos u \sin u = \mathbb{E}\cos(-u)\sin(-u) = -\mathbb{E}\cos u \sin u = 0$. Also of course $\cos^2 u + \sin^2 u = 1$ so $\mathbb{E}\cos^2 u = \mathbb{E}\sin^2 u = 1/2$. Thus

$$2\mathbb{E}\cos(x+u)\cos(y+u) = \cos x \cos y + \sin x \sin y = \cos(x-y).$$

Given a shift-invariant kernel k with associated distribution F, suppose $W \sim F$ and let $u \sim U[-\pi, \pi]$ independently. Define

$$\hat{\phi}(x) = \sqrt{2c}\cos(W^\top x + u).$$

Then

$$\begin{aligned}\mathbb{E}\hat{\phi}(x)\hat{\phi}(x') &= 2c\mathbb{E}[\mathbb{E}\{\cos(W^\top x + u)\cos(W^\top x' + u)|W\}] \\ &= c\mathbb{E}\cos((x-x')^\top W) = k(x, x').\end{aligned}$$

As a concrete example of this approach, let us take the Gaussian kernel $k(x, x') = \exp\{-\|x - x'\|_2^2/(2h^2)\}$. Note that if $W \sim N(0, h^{-2}I)$, it has characteristic function $\mathbb{E}(e^{it^\top W}) = e^{-\|t\|_2^2/(2h^2)}$ so we may take $\hat{\phi}(x) = \sqrt{2}\cos(W^\top x + u)$.

Chapter 2

The Lasso and extensions

2.1 Model selection

Let us revisit the linear model $Y = X\beta^0 + \varepsilon$ where $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2 I$. In many modern datasets, there are reasons to believe there are many more variables present than are necessary to explain the response. Let S be the set $S = \{k : \beta_k^0 \neq 0\}$ and suppose $s := |S| \ll p$.

The MSPE of OLS is

$$\begin{aligned}
\frac{1}{n}\mathbb{E}\|X\beta^0 - X\hat{\beta}^{\mathrm{OLS}}\|_2^2 &= \frac{1}{n}\mathbb{E}\{(\beta^0 - \hat{\beta}^{\mathrm{OLS}})^\top X^\top X(\beta^0 - \hat{\beta}^{\mathrm{OLS}})\} \\
&= \frac{1}{n}\mathbb{E}[\mathrm{tr}\{(\beta^0 - \hat{\beta}^{\mathrm{OLS}})(\beta^0 - \hat{\beta}^{\mathrm{OLS}})^\top X^\top X\}] \\
&= \frac{1}{n}\mathrm{tr}[\mathbb{E}\{(\beta^0 - \hat{\beta}^{\mathrm{OLS}})(\beta^0 - \hat{\beta}^{\mathrm{OLS}})^\top\}X^\top X] \\
&= \frac{1}{n}\mathrm{tr}(\mathrm{Var}(\hat{\beta}^{\mathrm{OLS}})X^\top X) = \frac{p}{n}\sigma^2.
\end{aligned}$$

If we could identify S and then fit a linear model using just these variables, we'd obtain an MSPE of $\sigma^2 s/n$ which could be substantially smaller than $\sigma^2 p/n$. Furthermore, it can be shown that parameter estimates from the reduced model are more accurate. The smaller model would also be easier to interpret.

We now briefly review some classical model selection strategies.

Best subset regression

A natural approach to finding S is to consider all 2^p possible regression procedures each involving regressing the response on a different sets of explanatory variables X_M where M is a subset of $\{1, \ldots, p\}$. We can then pick the best regression procedure using cross-validation (say). For general design matrices, this involves an exhaustive search over all subsets, so this is not really feasible for $p > 50$.

Forward selection

This can be seen as a greedy way of performing best subsets regression. Given a target model size m (the tuning parameter), this works as follows.

1. Start by fitting an intercept only model.
2. Add to the current model the predictor variable that reduces the residual sum of squares the most.
3. Continue step 2 until m predictor variables have been selected.

2.2 The Lasso estimator

The ***L**east **a**bsolute **s**hrinkage and **s**election **o**perator (Lasso)* [Tibshirani, 1996] estimates β^0 by $\hat{\beta}^{\mathrm{L}}_\lambda$, where $(\hat{\mu}^{\mathrm{L}}, \hat{\beta}^{\mathrm{L}}_\lambda)$ minimise

$$\frac{1}{2n}\|Y - \mu\mathbf{1} - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{2.1}$$

over $(\mu, \beta) \in \mathbb{R} \times \mathbb{R}^p$. Here $\|\beta\|_1$ is the ℓ_1-norm of β: $\|\beta\|_1 = \sum_{k=1}^p |\beta_k|$.

Like ridge regression, $\hat{\beta}^{\mathrm{L}}_\lambda$ shrinks the OLS estimate towards the origin, but there is an important difference. The ℓ_1 penalty can force some of the estimated coefficients to be exactly 0. In this way the Lasso can perform simultaneous variable selection and parameter estimation. As we did with ridge regression, we can centre and scale the X matrix, so then $\hat{\mu}^{\mathrm{L}}_\lambda = \bar{Y}$. As before, our target of interest is β^0. Define

$$Q_\lambda(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1. \tag{2.2}$$

Now the minimiser(s) of $Q_\lambda(\beta)$ will also be the minimiser(s) of

$$\|Y - X\beta\|_2^2 \text{ subject to } \|\beta\|_1 \leq \|\hat{\beta}^{\mathrm{L}}_\lambda\|_1.$$

Similarly, with the Ridge regression objective, we know that $\hat{\beta}^{\mathrm{R}}_\lambda$ minimises $\|Y - X\beta\|_2^2$ subject to $\|\beta\|_2 \leq \|\hat{\beta}^{\mathrm{R}}_\lambda\|_2$.

Now the contours of the OLS objective $\|Y - X\beta\|_2^2$ are ellipsoids centred at $\hat{\beta}^{\mathrm{OLS}}$, while the contours of $\|\beta\|_2^2$ are spheres centred at the origin, and the contours of $\|\beta\|_1$ are 'diamonds' centred at 0.

The important point to note is that the ℓ_1 ball $\{\beta \in \mathbb{R}^p : \|\beta\|_1 \leq \|\hat{\beta}^{\mathrm{L}}_\lambda\|_1\}$ has corners where some of the components are zero, and it is likely that the OLS contours will intersect the ℓ_1 ball at such a corner.

2.2.1 Prediction error of the Lasso (slow rate)

A remarkable property of the Lasso is that even when $p \gg n$, it can still perform well in terms of prediction error. Suppose the columns of X have been centred and scaled (as we will always assume from now on unless stated otherwise) and assume the normal linear model

$$Y = \mu \mathbf{1} + X\beta^0 + \varepsilon \tag{2.3}$$

where $\varepsilon \sim N_n(0, \sigma^2 I)$.

Theorem 11. *Let $\hat{\beta}$ be any Lasso solution when*

$$\lambda = A\sigma\sqrt{\frac{\log(p)}{n}}.$$

With probability at least $1 - 2p^{-(A^2/2-1)}$

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leq 4A\sigma\sqrt{\frac{\log(p)}{n}}\|\beta^0\|_1.$$

Proof. From the definition of $\hat{\beta}$ we have

$$\frac{1}{2n}\|Y - \bar{Y}\mathbf{1} - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|Y - \bar{Y}\mathbf{1} - X\beta^0\|_2^2 + \lambda\|\beta^0\|_1.$$

Rearranging,

$$\frac{1}{2n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leq \frac{1}{n}\varepsilon^\top X(\hat{\beta} - \beta^0) + \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}\|_1.$$

Now $|\varepsilon^\top X(\hat{\beta} - \beta^0)| \leq \|X^\top\varepsilon\|_\infty\|\hat{\beta} - \beta^0\|_1$. Let $\Omega = \{\|X^\top\varepsilon\|_\infty/n \leq \lambda\}$. Lemma 15 below shows that $\mathbb{P}(\Omega) \geq 1 - 2p^{-(A^2/2-1)}$. Working on the event Ω, we obtain

$$\begin{aligned}
&\frac{1}{2n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leq \lambda\|\beta^0 - \hat{\beta}\|_1 + \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}\|_1, \\
&\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leq 4\lambda\|\beta^0\|_1, \qquad \text{by the triangle inequality.}
\end{aligned}$$

□

2.2.2 Concentration inequalities I

The proof of Theorem 11 relies on a lower bound for the probability of the event Ω. A union bound gives

$$\begin{aligned}
\mathbb{P}(\|X^\top\varepsilon\|_\infty/n > \lambda) &= \mathbb{P}(\cup_{j=1}^p |X_j^\top\varepsilon|/n > \lambda) \\
&\leq \sum_{j=1}^p \mathbb{P}(|X_j^\top\varepsilon|/n > \lambda).
\end{aligned}$$

Now $X_j^\top\varepsilon/n \sim N(0, \sigma^2/n)$, so if we obtain a bound on the tail probabilities of normal distributions, the argument above will give a bound for $\mathbb{P}(\Omega)$.

Motivated by the need to bound normal tail probabilities, we will briefly discuss the topic of *concentration inequalities* that provide such bounds for much wider classes of random variables. Concentration inequalities are vital for the study of many modern algorithms and in our case here, they will reveal that the attractive properties of the Lasso presented in Theorem 11 hold true for a variety of non-normal errors.

We begin our discussion with the simplest tail bound, *Markov's inequality*, which states that given a non-negative random variable W,

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

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

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

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

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

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

$$\mathbb{E} e^{\alpha W} = e^{\alpha^2 \sigma^2 / 2}. \tag{2.4}$$

Thus

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

Note that to arrive at this bound, all we required was (an upper bound on) the moment generating function (mgf) of W (2.4).

Sub-Gaussian variables

Definition 3. We say a random variable W is *sub-Gaussian* if there exists $\sigma > 0$ such that

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

for all $\alpha \in \mathbb{R}$. We then say that W is *sub-Gaussian with parameter* σ.

Proposition 12 (Sub-Gaussian tail bound)**.** *If W is sub-Gaussian with parameter σ then*

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

As well as Gaussian random variables, the sub-Gaussian class includes bounded random variables.

Lemma 13 (Hoeffding's lemma)**.** *If W takes values in $[a, b]$, then W is sub-Gaussian with parameter $(b-a)/2$.*

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

Proposition 14. *Let $(W_i)_{i=1}^n$ be a sequence of independent sub-Gaussian random variables with parameters $(\sigma_i)_{i=1}^n$ and let $\gamma \in \mathbb{R}^n$. Then $\gamma^\top W$ is sub-Gaussian with parameter $\left(\sum_i \gamma_i^2 \sigma_i^2\right)^{1/2}$.*

Proof. Wlog, we may assume $\mathbb{E}W_i = 0$ for all i. We have

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

□

We can now prove a more general version of the probability bound required for Theorem 11.

Lemma 15. *Suppose $(\varepsilon_i)_{i=1}^n$ are independent, mean-zero and sub-Gaussian with common parameter σ. Note that this includes $\varepsilon \sim N_n(0, \sigma^2 I)$. Let $\lambda = A\sigma\sqrt{\log(p)/n}$. Then*

$$\mathbb{P}(\|X^\top\varepsilon\|_\infty/n \leq \lambda) \geq 1 - 2p^{-(A^2/2-1)}.$$

Proof.

$$\mathbb{P}(\|X^\top\varepsilon\|_\infty/n > \lambda) \leq \sum_{j=1}^p \mathbb{P}(|X_j^\top\varepsilon|/n > \lambda).$$

But $\pm X_j^\top\varepsilon/n$ are both sub-Gaussian with parameter $(\sigma^2\|X_j\|_2^2/n^2)^{1/2} = \sigma/\sqrt{n}$. Thus the RHS is at most

$$2p\exp(-A^2\log(p)/2) = 2p^{1-A^2/2}.$$

□

2.2.3 Some facts from optimisation theory and convex analysis

In order to study the Lasso in detail, it will be helpful to review some basic facts from optimisation and convex analysis.

Convexity

A set $C \subseteq \mathbb{R}^d$ is *convex* if

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

Given $C \subseteq \mathbb{R}^d$, We say a function $f : C \to \mathbb{R}$ is *convex* if C is convex and

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

for all $x, y \in C$ and $t \in (0,1)$. It is *strictly convex* if the inequality is strict for all $x, y \in C$ with $x \neq y$. If a strictly convex function has a minimiser, it must be unique. In the following, $C \subseteq \mathbb{R}^d$ is a convex set.

Proposition 16. *(i) Let $f_1, \ldots, f_m : C \to \mathbb{R}$ be convex functions. Then if $c_1, \ldots, c_m \geq 0$, $c_1 f_1 + \cdots + c_m f_m : C \to \mathbb{R}$ is a convex function.*

(ii) If $f : C \to \mathbb{R}$, and $A : \mathbb{R}^m \to \mathbb{R}^d$ is an affine function (so $A(x) = Mx + b$ for $M \in \mathbb{R}^{d \times m}$ and $b \in \mathbb{R}^d$) then $g : D \to \mathbb{R}$, where $D = \{x \in \mathbb{R}^m : A(x) \in C\}$ given by $g(x) = f(A(x))$ is convex.

(iii) If $f : C \to \mathbb{R}$ is convex with C open and f is twice continuously differentiable on C, then

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

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

The Lagrangian method

Consider an optimisation problem of the form

$$\text{minimise } f(x), \text{ subject to } g(x) = 0, \qquad x \in C \subseteq \mathbb{R}^d, \tag{2.5}$$

where $g : C \to \mathbb{R}^b$. Suppose the optimal value is $c^* \in \mathbb{R}$. The Lagrangian for this problem is defined as

$$L(x, \theta) = f(x) + \theta^\top g(x)$$

where $\theta \in \mathbb{R}^b$. Note that

$$\inf_{x \in C} L(x, \theta) \leq \inf_{x \in C : g(x) = 0} L(x, \theta) = c^*$$

for all θ. The Lagrangian method involves finding a $\theta = \theta^*$ such that the minimising $x = x^*$ on the LHS satisfies $g(x^*) = 0$. This x^* must then be a minimiser in the original problem (2.5).

Subgradients

Definition 4. Given convex $C \subseteq \mathbb{R}^d$, a vector $v \in \mathbb{R}^d$ is a *subgradient* of a convex function $f : C \to \mathbb{R}$ at x if

$$f(y) \geq f(x) + v^\top (y - x) \qquad \text{for all } \; y \in C.$$

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

In order to make use of subgradients, we will require the following two facts:

Proposition 17. *Let $f : C \to \mathbb{R}$ be convex, and suppose f is differentiable at $x \in \operatorname{int}(C)$. Then $\partial f(x) = \{\nabla f(x)\}$.*

Proposition 18 (Subgradient calculus)**.** *Let $f, f_1, f_2 : \mathbb{R}^d \to \mathbb{R}$ be convex. Then*

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

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

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

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

The following easy (but key) result is often referred to in the statistical literature as the Karush–Kuhn–Tucker (KKT) conditions, though it is actually a much simplified version of them.

Proposition 19. *Given convex $f : C \to \mathbb{R}$, $x^* \in \operatorname*{arg\,min}_{x \in C} f(x)$ if and only if $0 \in \partial f(x^*)$.*

Proof.

$$\begin{aligned} f(y) \geq f(x^*) \quad \text{for all } \; y \in C &\Leftrightarrow f(y) \geq f(x^*) + 0^\top (y - x) \quad \text{for all } \; y \in C \\ &\Leftrightarrow 0 \in \partial f(x^*). \end{aligned}$$

□

Let us now compute the subdifferential of the ℓ_1-norm. First note that $\|\cdot\|_1 : \mathbb{R}^d \to \mathbb{R}$ is convex. Indeed it is a norm so the triangle inequality gives $\|tx + (1-t)y\|_1 \leq t\|x\|_1 + (1-t)\|y\|_1$. We introduce some notation that will be helpful here and throughout the rest of the course.

For $x \in \mathbb{R}^d$ and $A = \{k_1, \ldots, k_m\} \subseteq \{1, \ldots, d\}$ with $k_1 < \cdots < k_m$, by x_A we will mean $(x_{k_1}, \ldots, x_{k_m})^\top$. Similarly if X has d columns we will write X_A for the matrix

$$X_A = (X_{k_1} \cdots X_{k_m}).$$

Further in this context, by A^c, we will mean $\{1, \ldots, d\} \setminus A$. Additionally, when in subscripts we will use the shorthand $-j = \{j\}^c$ and $-jk = \{j, k\}^c$. Note these column and component

extraction operations will always be considered to have taken place first before any further operations on the matrix, so for example $X_A^\top = (X_A)^\top$. Finally, define

$$\operatorname{sgn}(x_1) = \begin{cases} -1 & \text{if } x_1 < 0 \\ 0 & \text{if } x_1 = 0 \\ 1 & \text{if } x_1 > 0, \end{cases}$$

and

$$\operatorname{sgn}(x) = (\operatorname{sgn}(x_1), \ldots, \operatorname{sgn}(x_d))^\top.$$

Proposition 20. *For $x \in \mathbb{R}^d$ let $A = \{j : x_j \neq 0\}$. Then*

$$\partial \|x\|_1 = \{v \in \mathbb{R}^d : \|v\|_\infty \leq 1 \text{ and } v_A = \operatorname{sgn}(x_A)\}$$

Proof. We can write $\|x\|_1 = \sum_{j=1}^d |e_j^\top x|$ where e_j is the jth standard basis vector. The subdifferential of the function $u \mapsto |u|$ is $[-1, 1]$ if $u = 0$ and $\{\operatorname{sgn}(u)\}$ otherwise. Thus from Proposition 18(c) we see that the subdifferential of $g_j(x) =:= |e_j^\top x|$ is

$$\partial g_j(x) = \begin{cases} \{\operatorname{sgn}(x_j) e_j\} & \text{if } x_j \geq 0 \\ \{t e_j : t \in [-1, 1]\} & \text{otherwise.} \end{cases}$$

Proposition 18(b) then gives the result. □

2.2.4 Lasso solutions

Equipped with these tools from convex analysis, we can now fully characterise the solutions to the Lasso. We have that $\hat{\beta}_\lambda^{\mathrm{L}}$ is a Lasso solution if and only if $0 \in \partial Q_\lambda(\hat{\beta}_\lambda^{\mathrm{L}})$, which is equivalent to

$$\frac{1}{n} X^\top (Y - X\hat{\beta}_\lambda^{\mathrm{L}}) = \lambda \hat{\nu},$$

for $\hat{\nu}$ with $\|\hat{\nu}\|_\infty \leq 1$ and writing $\hat{S}_\lambda = \{k : \hat{\beta}_{\lambda,k}^{\mathrm{L}} \neq 0\}$, $\hat{\nu}_{\hat{S}_\lambda} = \operatorname{sgn}(\hat{\beta}_{\lambda,\hat{S}_\lambda}^{\mathrm{L}})$.

Lasso solutions need not be unique (e.g. if X has duplicate columns), though for most reasonable design matrices, Lasso solutions will be unique. We will often tacitly assume Lasso solutions are unique in the statement of our theoretical results. It is however straightforward to show that the Lasso fitted values are unique.

Proposition 21. *Fix $\lambda \geq 0$ and suppose $\beta^{(1)}$ and $\beta^{(2)}$ are two Lasso solutions. Then $X\beta^{(1)} = X\beta^{(2)}$.*

Proof. Suppose $\beta^{(1)}$ and $\beta^{(2)}$ both give an optimal objective value of c^*. Now by strict convexity of $\|\cdot\|_2^2$,

$$\|Y - X\beta^{(1)}/2 - X\beta^{(2)}/2\|_2^2 \leq \|Y - X\beta^{(1)}\|_2^2/2 + \|Y - X\beta^{(2)}\|_2^2/2,$$